Multiresolution and Multimodal Speech Recognition with Transformers

Georgios Paraskevopoulos **Srinivas Parthasarathy** **Aparna Khare** **Shiva Sundaram**
Amazon Lab126
geopar@central.ntua.gr, {parsrini,apkhare,sssundar}@amazon.com

Abstract

This paper presents an audio visual automatic speech recognition (AV-ASR) system using a Transformer-based architecture. We particularly focus on the scene context provided by the visual information, to ground the ASR. We extract representations for audio features in the encoder layers of the transformer and fuse video features using an additional crossmodal multihead attention layer. Additionally, we incorporate a multitask training criterion for multiresolution ASR, where we train the model to generate both character and subword level transcriptions. Experimental results on the How2 dataset, indicate that multiresolution training can speed up convergence by around 50% and relatively improves word error rate (WER) performance by upto 18% over subword prediction models. Further, incorporating visual information improves performance with relative gains upto 3.76% over audio only models. Our results are comparable to state-of-the-art Listen, Attend and Spell-based architectures.

1 Introduction

Automatic speech recognition is a fundamental technology used on a daily basis by millions of end-users and businesses. Applications include automated phone systems, video captioning and voice assistants providing an intuitive and seemless interface between users and end systems. Current ASR approaches rely solely on utilizing audio input to produce transcriptions. However, the wide availability of cameras in smartphones and home devices acts as motivation to build AV-ASR models that rely on and benefit from multimodal input.

Traditional AV-ASR systems focus on tracking the user's facial movements and performing lipreading to augment the auditory inputs (Potamianos et al., 1997; Mroueh et al., 2015; Tao and Busso, 2018). The applicability of such models in real world environments is limited, due to the need for accurate audio-video alignment and careful camera placement. Instead, we focus on using video to contextualize the auditory input and perform multimodal grounding. For example, a basketball court is more likely to include the term "lay-up" whereas an office place is more likely include the term "layoff". This approach can boost ASR performance, while the requirements for video input are kept relaxed (Caglayan et al., 2019; Hsu et al., 2019). Additionally we consider a multiresolution loss that takes into account transcriptions at the character and subword level. We show that this scheme regularizes our model showing significant improvements over subword models. Multitask learning on multiple levels has been previously explored in the literature, mainly in the context of CTC (Sanabria and Metze, 2018; Krishna et al., 2018; Ueno et al., 2018). A mix of seq2seq and CTC approaches combine word and character level (Kremer et al., 2018; Ueno et al., 2018) or utilize explicit phonetic information (Toshniwal et al., 2017; Sanabria and Metze, 2018).

Modern ASR systems rely on end-to-end, alignment free neural architectures, i.e. CTC (Graves et al., 2006) or sequence to sequence models (Graves et al., 2013; Zhang et al., 2017). The use of attention mechanisms significantly improve results in (Chorowski et al., 2015) and (Chan et al., 2016). Recently, the success of transformer architectures for NLP tasks (Vaswani et al., 2017; Devlin et al., 2019; Dai et al., 2019) has motivated speech researchers to investigate their efficacy in end-to-end ASR (Karita et al., 2019b). Zhou et. al., apply an end-to-end transformer architecture for Mandarin Chinese ASR (Zhou et al., 2018). Speech-Transformer extends the scaled dot-product attention mechanism to 2D and achieves competitive results for character level recognition (Dong et al., 2018; Karita et al., 2019a). Pham et. al. introduce the idea of stochastically deactivating layers dur-

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 2381–2387

ing training to achieve a very deep model (Pham et al., 2019). A major challenge of the transformer architecture is the quadratic memory complexity as a function of the input sequence length. Most architectures employ consecutive feature stacking (Pham et al., 2019) or CNN preprocessing (Dong et al., 2018; Karita et al., 2019b) to downsample input feature vectors. Mohamed et al. (2019) use a VGG-based input network to downsample the input sequence and achieve learnable positional embeddings.

Multimodal grounding for ASR systems has been explored in (Caglayan et al., 2019), where a pretrained RNN-based ASR model is finetuned with visual information through Visual Adaptive Training. Sterpu et al. (2018) propose a seq2seq model based on RNNs for lip-reading that performs cross-modal alignment of face tracking and audio features through an attention mechanism. Furthermore, Hsu et al. (2019) use a weakly supervised semantic alignment criterion to improve ASR results when visual information is present. Multimodal extensions of the transformer architecture have also been explored. These extensions mainly fuse visual and language modalities in the fields of Multimodal Translation and Image Captioning. Most approaches focus on using the scaled dot-product attention layer for multimodal fusion and cross-modal mapping. Afouras et al. (2018) present a transformer model for AV-ASR targeted for lip-reading in the wild tasks. It uses a self attention block to encode the audio and visual dimension independently. A decoder individually attends to the audio and video modalities producing character transcriptions. In comparison our study uses the video features to provide contextual information to our ASR. Libovickỳ et al. (2018) employ two encoder networks for the textual and visual modalities and propose four methods of using the decoder attention layer for multimodal fusion, with hierarchical fusion yielding the best results. Yu et al. (2019) propose an encoder variant to fuse deep, multi-view image features and use them to produce image captions in the decoder. Le et al. (2019) use cascaded multimodal attention layers to fuse visual information and dialog history for a multimodal dialogue system. Tsai et al. (2019) present Multimodal Transformers, relying on a deep pairwise cascade of cross-modal attention mechanisms to map between modalities for multimodal sentiment analysis.

In relation to the previous studies, the main contributions of this study are a) a fusion mechanism for audio and visual modalities based on the cross-modal scaled-dot product attention, b) an end to end training procedure for multimodal grounding in ASR and c) the use of a multiresolution training scheme for character and subword level recognition in a seq2seq setting without relying on explicit phonetic information. We evaluate our system in the 300 hour subset of the How2 database (Sanabria et al., 2018), achieving relative gains up to 3.76% with the addition of visual information. Further we show relative gains of 18% with the multiresolution loss. Our results are comparable to state-of-the-art ASR performance on this database.

2 Proposed Method

Our transformer architecture uses two transformer encoders to individually process acoustic and visual information (Fig. 1). Audio frames are fed to the first set of encoder layers. We denote the space of the encoded audio features as the audio space $\mathbb{A}$. Similarly, video features are projected to the video space $\mathbb{V}$ using the second encoder network. Features from audio and visual space are passed through a tied feed forward layer that projects them into a common space before passing them to their individual encoder layers respectively. This tied embedding layer is important for fusion as it helps align the semantic audio and video spaces. We then use a cross-modal attention layer that maps projected video representations to the projected audio space (Section 2.1). The outputs of this layer are added to the original audio features using a learnable parameter α to weigh their contributions. The fused features are then fed into the decoder stack followed by dense layers to generate character and subword outputs. For multiresolution predictions (Section 2.2), we use a common decoder for both character and subword level predictions, followed by a dense output layer for each prediction. This reduces the model parameters and enhances the regularization effect of multitask learning.

2.1 Cross-modal Attention

Scaled dot-product attention operates by constructing three matrices, K, V and Q from sequences of inputs. K and V may be considered keys and values in a "soft" dictionary, while Q is a query that contextualizes the attention weights. The attention mechanism is described in Eq. 1, where σ denotes

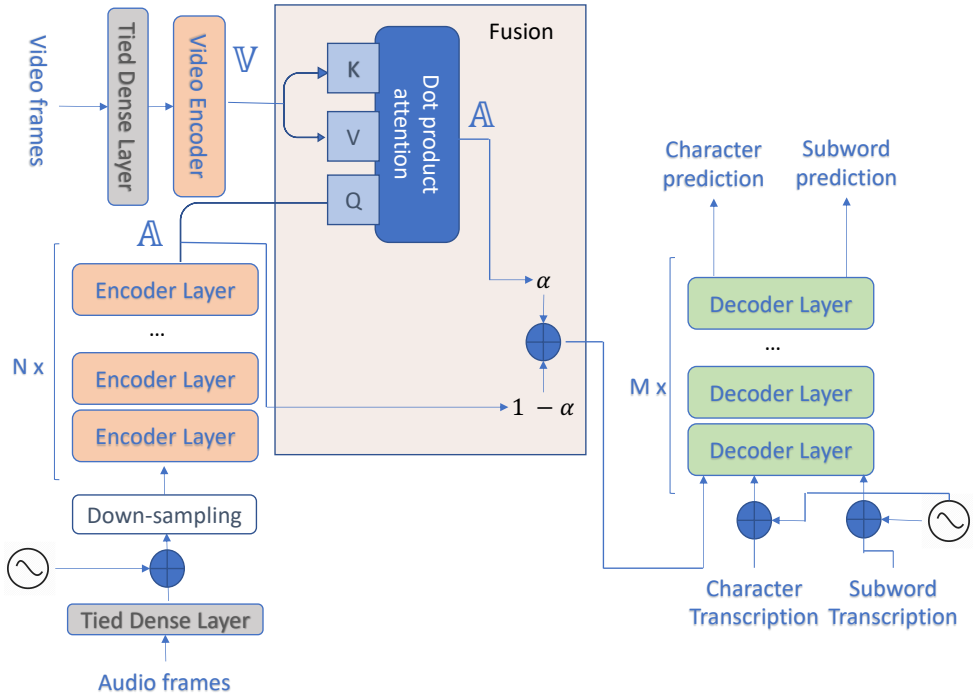

Figure 1: Overall system architecture. A cross-modal scaled dot-product attention layer is used to project the visual data into the audio feature space followed by an additive fusion.

the softmax operation.

$$Y = \sigma(KQ^T)V \quad (1)$$

The case where K, V and Q are constructed using the same input sequence consists a self-attention mechanism. We are interested in cross-modal attention, where K and V are constructed using inputs from one modality $\mathbb{M}_1$, video in our case (Fig. 1) and Q using another modality $\mathbb{M}_2$, audio. This configuration as an effective way to map features from $\mathbb{M}_1$ to $\mathbb{M}_2$ (Tsai et al., 2019). Note, that such a configuration is used in the decoder layer of the original transformer architecture (Vaswani et al., 2017) where targets are attended based on the encoder outputs.

2.2 Multiresolution training

We propose the use of a multitask training scheme where the model predicts both character and subword level transcriptions. We jointly optimize the model using the weighted sum of character and subword level loss, as in Eq. 2:

$$L = \gamma * L_{subword} + (1 - \gamma) * L_{character} \quad (2)$$

where γ is a hyperparameter that controls the importance of each task.

The intuition for this stems from the reasoning that character and subword level models perform different kinds of mistakes. For character prediction, the model tends to predict words that sound phonetically similar to the ground truths, but are syntactically disjoint with the rest of the sentence. Subword prediction, yields more syntactically correct results, but rare words tend to be broken down to more common words that sound similar but are semantically irrelevant. For example, character level prediction may turn "*old-fashioned*" into "*old-fashioning*", while subword level turns the sentence "*ukuleles are different*" to "*you go release are different*". When combining the losses, subword prediction, which shows superior performance is kept as the preliminary output, while the character prediction is used as an auxiliary task for regularization.

3 Experimental Setup

We conduct our experiments on the How2 instructional videos database (Sanabria et al., 2018). The dataset consists of 300 hours of instructional videos from the YouTube platform. These videos depict people showcasing particular skills and have high variation in video/audio quality, camera angles and duration. The transcriptions are mined from the YouTube subtitles, which contain a mix of automatically generated and human annotated transcriptions. Audio is encoded using 40 mel-filterbank coefficients and 3 pitch features with a frame size

Input handling	Recognition level	WER
Filtering	Character	33.0
Filtering	Subword	29.7
Chunking	Character	31.3
Chunking	Subword	29.9
Stacking	Character	28.3
Stacking	Subword	26.1
Stacking	MR	**21.3**

Table 1: Results for different methods of input filtering for different prediction resolutions. *MR* stands for multiresolution.

of 10 ms, yielding 43-dimensional feature vectors. The final samples are segments of the original videos, obtained using word-level alignment. We follow the video representation of the original paper (Caglayan et al., 2019), where a 3D ResNeXt-101 architecture, pretrained on action recognition, is used to extract 2048D features (Hara et al., 2018). Video features are average pooled over the video frames yielding a single feature vector. For our experiments, we use the train, development and test splits proposed by (Sanabria et al., 2018), which have sizes 298.2 hours, 3.2 hours and 3.7 hours respectively.

Our model consists of 6 encoder layers and 4 decoder layers. We use transformer dimension 480, intermediate ReLU layer size 1920 and 0.2 dropout. All attention layers have 6 attention heads. The model is trained using Adam optimizer with learning rate 10^{-3} and 8000 warmup steps. We employ label smoothing of 0.1. We weigh the multitask loss with $\gamma = 0.5$ which gives the best performance. A coarse search was performed for tuning all hyperparameters over the development set. For character-level prediction, we extract 41 graphemes from the transcripts. For subword-level prediction, we train a SentencePiece tokenizer (Kudo and Richardson, 2018) over the train set transcriptions using byte-pair encoding and vocabulary size 1200. For decoding we use beam search with beam size 5 and length normalization parameter 0.7. We train models for up to 200 epochs and the model achieving the best loss is selected using early stopping. Any tuning of the original architecture is performed on the development split. No language model or ensemble decoding is used in the output.

4 Results and Discussion

One of the challenges using scaled dot-product attention is the quadratic increase of layerwise memory complexity as a function of the input sequence length. This issue is particularly prevalent in ASR tasks, with large input sequences. We explore three simple approaches to work around this limitation. First, we filter out large input sequences ($x > 15s$), leading to loss of 100 hours of data. Second we, chunk the input samples to smaller sequences, using forced-alignment with a conventional DNN-HMM model to find pauses to split the input and the transcriptions. Finally, we stack 4 consecutive input frames into a single feature vector, thus reducing the input length by 4. Note that this only reshapes the input data as the dimension of our input is increased by the stacking process [1]. Results for the downsampling techniques for character and subword level predictions are summarized in Table 1. We observe that subword-level model performs better than the character level (upto 10% relative) in all settings. This can be attributed to the smaller number of decoding steps needed for the subword model, where error accumulation is smaller. Furthermore, we see that the naive filtering of large sequences yields to underperforming systems due to the large data loss. Additionally, we see that frame stacking has superior performance to chunking. This is not surprising as splitting the input samples to smaller chunks leads to the loss of contextual information which is preserved with frame stacking. We evaluate the proposed multiresolution training technique with the frame stacking technique, observing a significant improvement(18.3%) in the final WER. We thus observe that predicting finer resolutions as an auxiliary task can be used as an effective means of regularization for this sequence to sequence speech recognition task. Furthermore, we have empirically observed that when training in multiple resolutions, models can converge around 50% faster than single resolution models.

Next, we evaluate relative performance improvement obtained from utilizing the visual features (Table 2). We observe that incorporating visual information improves ASR results. Our AV-ASR system yields gains $> 3\%$ over audio only models for both subword and multiresolution predictions. Finally, we observe that while the Listen, Attend and Spell-based architecture of (Caglayan et al., 2019) is slightly stronger than the transformer model, the gains from adding visual information

[1] We tried to use the convolutional architecture from (Mohamed et al., 2019), but it failed to converge in our experiments, possibly due to lack of data

Features	Level	WER	⇑ over audio
Audio	Subword	26.1	-
Audio + ResNeXt	Subword	25.0	3.45%
Audio	MR	**21.3**	-
Audio + ResNeXt	MR	20.5	3.76%
Audio (B)	Subword	**19.2**	-
Audio + ResNext (B)	Subword	**18.4**	3.13%

Table 2: Comparison of audio only ASR models versus AVASR models with ResNeXt image features. *MR* stands for multiresolution. *(B)* shows the results for the LAS model (Caglayan et al., 2019)

Missing input handling	WER
Zeros	23.1
Gaussian Noise σ=0.2	22.6
Gating visual input α=0	22.8

Table 3: Experimental evaluation of AV-ASR model for handling missing visual input. Here σ denotes the standard deviation of the noise

is consistent across models. It is important to note that our models are trained end-to-end with both audio and video features.

An important question for real-world deployment of multimodal ASR systems is their performance when the visual modality is absent. Ideally, a robust system satisfactorily performs when the user's camera is off or in low light conditions. We evaluate our AV-ASR systems in the absence of visual data with the following experiments - a) replace visual feature vectors by zeros b) initialize visual features with gaussian noise with standard deviation 0.2 c) tweak the value α to 0 on inference, gating the visual features completely. Table 3 shows the results for the different experiments. Results indicate gating visual inputs works better than zeroing them out. Adding a gaussian noise performs best which again indicates the limited availability of data. Overall, in the absence of visual information, without retraining, the AV-ASR model relatively worsens by 6% compared to audio only models.

5 Conclusions

This paper explores the applicability of the transformer architecture for multimodal grounding in ASR. Our proposed framework uses a crossmodal dot-product attention to map visual features to audio feature space. Audio and visual features are then combined with a scalar additive fusion and used to predict character as well as subword transcriptions. We employ a novel multitask loss that combines the subword level and character losses. Results on the How2 database show that a) multiresolution losses regularizes our model producing significant gains in WER over character level and subword level losses individually b) Adding visual information results in relative gains of 3.76% over audio model's results validating our model.

Due to large memory requirements of the attention mechanism, we apply aggressive preprocessing to shorten the input sequences, which may hurt model performance. In the future, we plan to alleviate this by incorporating ideas from sparse transformer variants (Kitaev et al., 2020; Child et al., 2019). Furthermore, we will experiment with more ellaborate, attention-based fusion mechanisms. Finally, we will evaluate the multiresolution loss on larger datasets to analyze it's regularizing effects.

References

Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. 2018. Deep audio-visual speech recognition. *IEEE transactions on pattern analysis and machine intelligence*.

Ozan Caglayan, Ramon Sanabria, Shruti Palaskar, Loic Barraul, and Florian Metze. 2019. Multimodal grounding for sequence-to-sequence speech recognition. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8648–8652. IEEE.

William Chan, Navdeep Jaitly, Quoc V. Le, and Oriol Vinyals. 2016. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4960–4964.

Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. 2019. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*.

Jan Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. 2015. Attention-based models for speech recognition. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, page 577–585, Cambridge, MA, USA. MIT Press.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th*

Annual Meeting of the Association for Computational Linguistics, pages 2978–2988.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Linhao Dong, Shuang Xu, and Bo Xu. 2018. Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5884–5888. IEEE.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, page 369–376, New York, NY, USA. Association for Computing Machinery.

Alex Graves, Abdel-rahman Mohamed, and Geoffrey E. Hinton. 2013. Speech recognition with deep recurrent neural networks. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE.

Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. 2018. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6546–6555.

Wei-Ning Hsu, David Harwath, and James Glass. 2019. Transfer learning from audio-visual grounding to speech recognition. *Proc. Interspeech 2019*, pages 3242–3246.

Shigeki Karita, Nelson Enrique Yalta Soplin, Shinji Watanabe, Marc Delcroix, Atsunori Ogawa, and Tomohiro Nakatani. 2019a. Improving Transformer-Based End-to-End Speech Recognition with Connectionist Temporal Classification and Language Model Integration. In *Proc. INTERSPEECH*, pages 1408–1412.

Shigeki Karita, Xiaofei Wang, Shinji Watanabe, Takenori Yoshimura, Wangyou Zhang, Nanxin Chen, Tomoki Hayashi, Takaaki Hori, Hirofumi Inaguma, Ziyan Jiang, Masao Someki, Nelson Enrique Yalta Soplin, and Ryuichi Yamamoto. 2019b. A comparative study on transformer vs RNN in speech applications. In *IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2019, Singapore, December 14-18, 2019*, pages 449–456. IEEE.

Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. 2020. Reformer: The efficient transformer. In *International Conference on Learning Representations*.

Jan Kremer, Lasse Borgholt, and Lars Maaløe. 2018. On the inductive bias of word-character-level multitask learning for speech recognition.

Kalpesh Krishna, Shubham Toshniwal, and Karen Livescu. 2018. Hierarchical multitask learning for ctc-based speech recognition. *arXiv preprint arXiv:1807.06234*.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71.

Hung Le, Doyen Sahoo, Nancy Chen, and Steven Hoi. 2019. Multimodal transformer networks for end-to-end video-grounded dialogue systems. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5612–5623.

Jindřich Libovickỳ, Jindřich Helcl, and David Mareček. 2018. Input combination strategies for multi-source transformer decoder. In *Proc. 3rd Conference on Machine Translation*, pages 253–260.

Abdelrahman Mohamed, Dmytro Okhonko, and Luke Zettlemoyer. 2019. Transformers with convolutional context for asr. *CoRR*.

Youssef Mroueh, Etienne Marcheret, and Vaibhava Goel. 2015. Deep multimodal learning for audio-visual speech recognition. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2130–2134. IEEE.

Ngoc-Quan Pham, Thai-Son Nguyen, Jan Niehues, Markus Müller, and Alex Waibel. 2019. Very deep self-attention networks for end-to-end speech recognition. *Proc. Interspeech 2019*, pages 66–70.

Gerasimos Potamianos, Eric Cosatto, Hans Peter Graf, and David B Roe. 1997. Speaker independent audio-visual database for bimodal asr. In *Proc. European Tutorial and Research Workshop on Audio-Visual Speech Processing*, pages 65–68.

Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. 2018. How2: a large-scale dataset for multimodal language understanding. In *Proceedings of the Workshop on Visually Grounded Interaction and Language (ViGIL)*. NeurIPS.

Ramon Sanabria and Florian Metze. 2018. Hierarchical multitask learning with ctc. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 485–490. IEEE.

George Sterpu, Christian Saam, and Naomi Harte. 2018. Attention-based audio-visual fusion for robust automatic speech recognition. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, pages 111–115.

Fei Tao and Carlos Busso. 2018. Aligning audiovisual features for audiovisual speech recognition. In *2018 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE.

Shubham Toshniwal, Hao Tang, Liang Lu, and Karen Livescu. 2017. Multitask learning with low-level auxiliary tasks for encoder-decoder based speech recognition. *Proc. Interspeech 2017*, pages 3532–3536.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6558–6569.

Sei Ueno, Hirofumi Inaguma, Masato Mimura, and Tatsuya Kawahara. 2018. Acoustic-to-word attention-based model complemented with character-level ctc-based model. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5804–5808. IEEE.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Jun Yu, Jing Li, Zhou Yu, and Qingming Huang. 2019. Multimodal transformer with multi-view visual representation for image captioning. *IEEE Transactions on Circuits and Systems for Video Technology*.

Yu Zhang, William Chan, and Navdeep Jaitly. 2017. Very deep convolutional networks for end-to-end speech recognition. pages 4845–4849.

Shiyu Zhou, Linhao Dong, Shuang Xu, and Bo Xu. 2018. Syllable-based sequence-to-sequence speech recognition with the transformer in mandarin chinese. *Proc. Interspeech 2018*, pages 791–795.