INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Caption Alignment for Low Resource Audio-Visual Data

Vighnesh Reddy Konda[1], *Mayur Warialani*[1], *Rakesh Prasanth Achari*[1], *Varad Bhatnagar*[1], *Jayaprakash Akula*[1], *Preethi Jyothi*[1], *Ganesh Ramakrishnan*[1], *Gholamreza Haffari*[2] *and Pankaj Singh*[1]

[1]Indian Institute of Technology Bombay, India
[2]Monash University, Australia
{vighnesh,mayurwarialani}@cse.iitb.ac.in

Abstract

Understanding videos via captioning has gained a lot of traction recently. While captions are provided alongside videos, the information about where a caption aligns within a video is missing, which could be particularly useful for indexing and retrieval. Existing work on learning to infer alignments has mostly exploited visual features and ignored the audio signal. Video understanding applications often underestimate the importance of the audio modality. We focus on how to make effective use of the audio modality for temporal localization of captions within videos. We release a new audio-visual dataset that has captions time-aligned by (i) carefully listening to the audio and watching the video, and (ii) watching only the video. Our dataset is audio-rich and contains captions in two languages, English and Marathi (a low-resource language). We further propose an attention-driven multimodal model, for effective utilization of both audio and video for temporal localization. We then investigate (i) the effects of audio in both data preparation and model design, and (ii) effective pretraining strategies (Audioset, ASR-bottleneck features, PASE, *etc.*) handling low-resource setting to help extract rich audio representations.

Index Terms: multimodal models, low-resource audio-visual corpus, caption alignment for videos

1. Introduction

Rooted in video understanding, temporally localizing captions within videos is a relatively new and challenging task where sentences are provided alongside videos, and the task involves predicting start and end times where the sentence best aligns with the video [1, 2, 3, 4]. An established approach to tackle the alignment problem is to extract frame-level video features, and compare their similarity with sentence level features. This is based on the idea that, in some latent space, the most similar video features will be closest to the sentence features. However, these techniques do not exploit the multimodal nature of videos and ignore the audio modality altogether.

In this paper, we aim to improve performance of temporal localization in videos by *incorporating audio* in an effective way. Even for existing datasets, the audio modality may benefit sentence alignment annotations, *e.g.* for ActivityNet [5] where the ground truth sentence alignments were created by largely ignoring the audio modality. Henceforth, we will refer to sentence as textual content. What if the ground truth alignments were instead created in an audio-sensitive and not an audio-agnostic manner? What is the effect of the language of the audio speech and that of the caption on the quality of the alignment? What are the learnings from existing datasets that can be leveraged for a new language? We investigate these questions through a new dataset MALTA_{av}[1] (see Figure 1) which we make available[2] through this work, and a new attention-based model that leverages both video and audio.

Our work makes contributions on three main fronts:

1. Data: We present a new multilingual, richly-annotated dataset MALTA_{av}. The ground truth of MALTA_{av} is generated by instructing 10 annotators to pay close attention to the audio *and* the visual streams while aligning the sentence captions with the video. We observe that this process is a lot more intensive than the video-driven and largely audio-agnostic alignment process that has been employed to create erstwhile datasets. We empirically quantify this slowdown to be by a factor of 3 by also having another subset of annotators align captions with a subset of our videos in MALTA_{av} by ignoring audio (as is typically done in benchmark datasets). We refer to this subset as MALTA_{v} and use it only for evaluation purposes.

2. Model: Our attention-based multimodal architecture MALTA is based on language specific pretraining of the audio modality, and mutual co-attention between the three audio, video and text modalities for their effective combination.

3. Pretrained Audio Features: We examine the role of pretrained audio features within our architecture. We take a detailed look at various audio representations and investigate how they interact with other modalities in the low-resource setting.

2. Related Work

To match the query and video frame candidates, one approach is to map the visual features of the frame candidates and the textual feature of the caption into a shared space and measure their semantic similarity. This is the basis of Moment Context Network [3] and Cross-modal Temporal Regression Localize [6]. Most relevant to our work is Attention Based Location Regression (ABLR) [7] which uses a multimodal co-attention mechanism to identify the relevant video frames based on an encoding of the caption. [2] proposes a reinforcement learning based agent to progressively regulate the temporal grounding boundaries. [8] proposed Moment Alignment Network, which unifies the candidate frame encoding and temporal structural reasoning into a single-shot feed-forward network.

Several techniques have been employed to leverage information from both audio and visual modalities for the task of caption generation. [9, 10] leveraged multimodality within an encoder-decoder model and obtained a boost in performance. [11, 12] also used speech features from the audio modality to

[1]multi-**M**odal **A**nd multi-**L**ingual **T**emporal sentence **A**lignment.

[2]The dataset and the codebase will be available for download at https://www.cse.iitb.ac.in/~malta/, the extended version of our paper can be found at https://www.cse.iitb.ac.in/~malta/malta-extended.pdf

http://dx.doi.org/10.21437/Interspeech.2020-3157

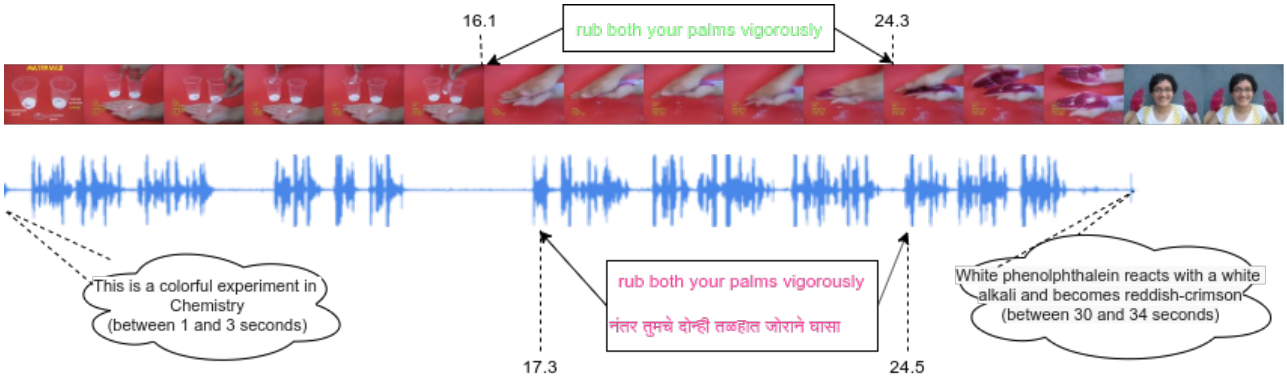

Figure 1: *Illustrating temporal sentence localization based on audio-visual features for a video from our dataset, showing annotations specific to* MALTA$_{av}$ *(in pink) and* MALTA$_{v}$ *(in green). Please note for the sake of the readers, that the two call outs at the beginning and end are English translations of the original Marathi speech.*

gain further improvements. On other tasks such as video event classification, [13, 14, 15] have shown improvements by using audio features along with visual features. As for the use of multiple modalities for caption alignment, there is no specific prior work that has come to our attention.

3. The Architecture

Our architecture is designed to enable multi-modal co-attention across important audio and visual segments in the video on the one hand and words in the sentence on the other hand. We achieve this by scaffolding our architecture MALTA on Attention Based Location Regression (ABLR) [7]. It is an end-to-end architecture to convert video and sentence inputs to the temporal coordinates in the output. MALTA comprises three main components as depicted in Figure 2: (i) context-dependent feature encoding of the input audio, video streams and sentence, (ii) multi-modal co-attention interaction highlighting important audio, visual segments in the video and words in the sentence, and (iii) attention based output prediction which can directly regress the temporal coordinates of the target video.

3.1. Input Feature Representations

To extract video features, the video is first clipped into segments that are then encoded into dense video representations using the well-known C3D network [16]. These feature vectors are subsequently passed as input to a bidirectional LSTM-based encoder(single layer) and further transformed by a linear layer applied to its hidden states. The audio modality is encoded using different representations that we detail further in Section 4. Representations for the captions are derived from the final hidden state of a bidirectional LSTM that takes a sequence of FastText word embeddings as input.

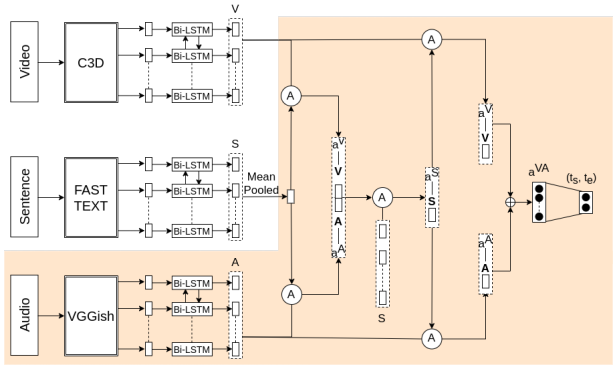

Figure 2: *Our proposed* MALTA *architecture. We denote the final attention features by* $\boldsymbol{a}_A$ *for the audio, and* $\boldsymbol{a}_V$ *for the video.*

3.2. Co-Attention of Audio, Video and Text Modalities

We consider sentence-video and sentence-audio interactions independently and compute attention distributions over the video/audio modalities using co-attention. We use the sentence to learn attention on both video and audio modalities separately and then concatenate both attended features to further attend to the sentence. We use the attended sentence features to attend once again to the audio and video modalities separately. We finally sum the attention distributions over both video and audio modalities, normalize it and use the resulting distribution to regress the temporal coordinates of the sentence within the video (see Figure 2).

3.3. Training Objective

The predicted start and end times[3], $\hat{\tau}_i^s$ and $\hat{\tau}_i^e$, are obtained using the sum of final audio and video attention weights ($\boldsymbol{a}_{VA}$) and directly regressing the temporal coordinates: we use a linear interpolation of two losses, L_{reg} and L_{cal}, to supervise the prediction of temporal coordinates of a sentence within a video.

$$L_{\text{reg}} = \sum_{i=1}^{N} [R(\bar{\tau}_i^s - \hat{\tau}_i^s) + R(\bar{\tau}_i^e - \hat{\tau}_i^e)] \quad (1)$$

$$L_{\text{cal}} = -\sum_{i=1}^{N} \frac{\sum_{j=1}^{M} \delta_{i,j} \log\left(a_{V_i,j} * a_{A_i,j}\right)}{\sum_{j=1}^{M} \delta_{i,j}} \quad (2)$$

where $R(\cdot)$ is a smooth L1 function [7], $\delta_{i,j} = 1$ if the j^{th} segment in V_i is within the ground truth interval and 0 otherwise. Here, $a_{V,j}$ denotes the relative importance of video in the j-th clip for the given sentence. The caliberation loss is similar to the one in [7], with the audio attention included.

[3]The ground-truth start and end times are normalized by the duration of the video. That is, $(\bar{\tau}_i^s, \bar{\tau}_i^e) = (\frac{\tau_i^s}{d_i}, \frac{\tau_i^e}{d_i})$

4. Experimental Results

We attempt to answer the following questions through our experiments. (i) Do we consistently benefit from attending to multiple modalities? (ii) What is the effect of use of different modalities when the ground truth sentence alignments are created in an audio-video-driven (as against only video-driven) manner? (iii) What is the effect of the language of the speech in the audio and the language of the sentence captions on the quality of the alignment output? Owing to space constraints here, we defer some details of the above as well as additional investigations (e.g. how does performance vary using MALTA when we deliberately manipulate videos to have incongruent audio?) to the extended version on our website.

Datasets: We conduct experiments on our newly constructed MALTA_{av} as well as two standard benchmarks, namely Charades-STA [6] and ActivityNet [5]. MALTA_{av} consists of simple video tutorials of two types: (i) TFT_{av} that describes the creation of scientific toys from waste material[4](ii) ATMA_{av} that features farmers describing and demonstrating organic farming techniques. Both video collections have speakers in the background narrating the process in Marathi. These videos are rich in both video and audio content. TFT_{av} consists of 492 videos, with an average length of 80 seconds and around 7 sentences describing every video in each of two languages, *viz.*, Marathi and English, along with background speech in Marathi. On the other hand, ATMA_{av} is relatively smaller, consisting of 95 videos, with an average length of 111 seconds and around 18 sentences describing every video in a single language, *viz.*, Marathi, accompanied by background speech in Marathi. We show results from our experiments on TFT_{av} and ATMA_{av} separately.

Charades-STA [6] contains 16128 clip-sentence pairs; we created training/test splits containing 12408/3720 pairs, respectively. ActivityNet [5] is significantly larger containing 20K videos and 100K sentences annotated with start and end times. We used the publicly-available train set for training and the validation set to evaluate our models.

Implementation Details: Videos in ActivityNet, Charades-STA and MALTA_{av} were split into 8922:4369 , 5338:1334 and 389:103 clips for training and testing, respectively. We extracted 4096-dimensional C3D features for each dataset to serve as the video features and 128-dimensional audio features were extracted using VGG. Bidirectional LSTM layers with a hidden state size of 256 were used for each modality. We used the Adam optimizer to train MALTA with a learning rate of 0.001. Following the metrics adopted in prior work for temporal localization of sentences in videos [6], for each sentence, we calculate the Intersection over Union (IoU) between the predicted and ground truth temporal coordinates. "IoU $= \alpha$" denotes the percentage of the sentence queries which have an IoU larger than α.

Audio Representations. We investigate the following:

• VGG features [17]: These are extracted from a pretrained network trained on AudioSet consisting of audio events [18].

• PASE features: PASE [19] is a pretrained speech model consisting of multiple workers that are jointly trained to optimize seven different speech-driven self-supervised tasks, including regression tasks that involve predicting the waveform, MFCC [20] and prosody features and binary discrimination tasks that differentiate between positive and negative samples based on an anchor utterance. We do not make use of the speech labels while extracting PASE features.

• ASR-bnf features: We used the Kaldi toolkit [21] to train a state-of-the-art time delay neural network (TDNN) acoustic model on roughly 100 hours of weakly labelled Marathi spoken tutorial data.[5] The TDNN model has 12 layers with a 128-dimensional bottleneck layer before the penultimate layer. We decoded Marathi speech from the videos in both TFT_{av} and ATMA_{av} using this trained network and extracted bottleneck features.

[4]We downloaded these videos from `http://www.arvindguptatoys.com/toys-from-trash.php` and obtained consent from the content creator

	Activity-Net		Charades	
MODEL	IoU= .5	IoU= .7	IoU= .5	IoU= .7
A-only	0.3373	0.1705	0.3583	0.1537
V-only	0.3571	0.1786	0.3611	0.1462
ABLR [7]	0.3571	0.1786	0.3611	0.1462
MALTA	0.3636	0.1872	0.3650	0.1490

Table 1: *Results on Activity-Net and Charades using a single modality(A-ONLY, V-ONLY) and multiple modalities (ABLR, MALTA).*

4.1. Single Modality

In order to analyze the importance of combining modalities (question (i)), we first investigate systems that only consider co-attention between a single modality (video or audio) and the sentence. In Table 1, we report results on the two existing benchmark datasets and in Table 2 we present results on TFT_{av}; A-ONLY refers to using only the audio VGG features, and V-ONLY refers to using just the C3D video features. Given the smaller size of our dataset, we report mean IoUs and standard deviations computed across five different random seeds for TFT_{av}, and show results from the best-performing seed for Activity-Net and Charades in Table 1. We observe that V-ONLY outperforms A-ONLY on ActivityNet. On Charades and TFT_{av}, A-ONLY is better than V-ONLY, with the margin being larger for TFT_{av}. This further validates our claim that TFT_{av} is content-rich in the audio modality.

We note another interesting trend in the IoUs in Table 2. IoU=α exhibits a decreasing trend in standard deviations with increasing values of α. Given the smaller dataset sizes of TFT_{av}, ATMA_{av} and the higher variance at smaller α's, we report IoU=0.5 and IoU=0.7 in all subsequent experiments for TFT_{av} and ATMA_{av}.

MODEL	IoU= .5	IoU= .7
A-ONLY	0.1529 ± 0.005	0.0557 ± 0.001
V-ONLY	0.1321 ± 0.004	0.0486 ± 0.002

Table 2: *Results on the* TFT_{av} *dataset*

4.2. Combining Modalities

Next we investigate the question of whether attending to multiple modalities helps. In Table 1, we report the performance of our multimodal MALTA on ActivityNet and Charades. We find consistent improvements in performance using MALTA over ABLR [7], which is a near state-of-the-art system on Charades. On ActivityNet, we obtain fairly significant improvements in performance at $\alpha = 0.5$ and $\alpha = 0.7$. The result on ActivityNet [7] is different from the numbers originally reported as we were unable to download roughly 1000 videos (that are no longer available) and hence could not use them during training.

[5]Available from: `https://spoken-tutorial.org/`

A-feat	IoU= .5	IoU= .7
None	0.1321 ± 0.004	0.0485 ± 0.002
VGG	0.1420 ± 0.002	0.0485 ± 0.005
MFCC	0.1425 ± 0.006	0.0439 ± 0.006
PASE-TFT$_{av}$ scratch	0.1387 ± 0.006	0.0474 ± 0.003
PASE-spk scratch	0.1375 ± 0.006	0.0496 ± 0.002
PASE-TFT$_{av}$ finetuned	0.1450 ± 0.005	0.0459 ± 0.003
PASE-spk finetuned	0.1459 ± 0.005	0.0484 ± 0.005
PASE-TFT$_{av}$ +spk finetuned	0.1478 ± 0.006	0.0462 ± 0.005
ASR-bnf	0.1550 ± 0.005	0.0545 ± 0.005

Table 3: *Results on* TFT$_{av}$ *with multimodal coattention comparing different audio representations. The first row corresponds to the* V-ONLY *model.*

A-feat	IoU= .5	IoU= .7
VGG	0.0476 ± 0.012	0.0065 ± 0.002
MFCC	0.0388 ± 0.003	0.0112 ± 0.004
PASE-ATMA$_{av}$ scratch	0.0382 ± 0.006	0.0147 ± 0.006
PASE-spk finetuned	0.0392 ± 0.007	0.0157 ± 0.003

Table 4: *Results on the* ATMA$_{av}$ *dataset.*

4.3. Comparing Audio Representations & Pretraining

In Table 3, we present results on the TFT$_{av}$ using different audio representations and training strategies with our multimodal model MALTA. VGG [17] has been described earlier in Section 4. MFCC [20] features are standard speech features. "tft" refers to the data in TFT$_{av}$ and "spk" refers to the weakly labeled spoken tutorial data. "scratch" indicates that the PASE model was trained starting from randomly initialized weights and "finetuned" indicates that we started with a pretrained PASE model which was further trained with the specified dataset. ASR-bnf refers to the bottleneck features extracted from the ASR model detailed in Section 4.

We make the following three main observations: 1) Starting from a pretrained PASE model and further fine-tuning it is consistently a better strategy than training a PASE model from scratch, especially given the relatively small size of the datasets used for fine-tuning. 2) All the finetuned PASE features are better than simple MFCC features. 3) ASR-bnf features outperform all the other features by a clear margin. These features were extracted from an ASR system and hence are most phonetically aware among all the representations. We also present results on ATMA$_{av}$ in Table 4 which exhibit similar trends as TFT$_{av}$ for IoU= 0.7 and demonstrate the transferability of PASE features. (The slightly different trend at IoU= 0.5 is possibly because ATMA$_{av}$ is roughly one-fifth the size of TFT$_{av}$.)

4.4. Skylines for Audio Modality

Our design of MALTA is reinforced in two skyline experiments wherein (i) we use ground truth based 'hard' attention (instead of attentions inferred from MALTA) to regress the temporal coordinates for TFT$_{av}$ and (ii) we use transcriptions for the speech in TFT$_{av}$ as input instead of audio features. These transcriptions are derived using Google's ASR API for Marathi. Table 5 shows results from both these skyline experiments. ASR-based transcription is expected to serve as a skyline because we expect the Marathi transcriptions from Google's API to be largely accurate, in which case the sentences are expected to have significant n-gram overlap with the speech transcriptions.

MODEL	IoU= .5	IoU= .7
Hard-Attention	0.6526	0.5364
Transcript	0.1710	0.0736
Video+Transcript	0.1730	0.0689

Table 5: *Skyline results on* TFT$_{av}$ *using hard attention and Google transcriptions for Marathi speech*

4.5. Correctly Analysing Gains from Audio

As a first step toward answering question (ii), i.e., assessing the importance of deriving ground truth alignments using both audio and video modalities as opposed to just the video modality, we compute the overlap between the video driven annotations on TFT$_{v}$ (instance of MALTA$_{v}$ illustrated in Figure 1) with the more ideal, audio-video driven annotations on TFT$_{av}$. We find IoU=0.1 to be 0.71 and IoU=0.7 to be 0.19. The IOU is not that high, reinforcing our claim that alignment using only the video modality may not be very accurate. In Table 6, we illustrate the approximate assessment of the improvement of MALTA over V-ONLY on a less accurately aligned dataset such as TFT$_{v}$. On the other hand, TFT$_{av}$ more faithfully represents the gains obtained by MALTA compared to V-ONLY (*c.f.* Table 3).

MODEL	IoU= .5	IoU= .7
V-ONLY	0.0906	0.0313
MALTA	0.0953	0.0375

Table 6: *Results using* MALTA *over* V-ONLY *on* TFT$_{v}$

4.6. Cross-Lingual Evaluation

Here, we address our final question of interest: How does MALTA perform when speech in the videos is in a language (Marathi) that is different from the captions' language (English)? With English captions, using V-ONLY on MALTA$_{av}$, IoU= .5 is 0.143 and IoU= .7 is 0.045, while using MALTA with ASR-bnf features yields IoU= .5 and IoU= .7 values of 0.149 and 0.057, respectively. Even with the mismatch in language, at higher α's, we see an improvement using the audio modality with MALTA over using just video.

5. Conclusion

We present a new dataset MALTA$_{av}$ and an attention-based model MALTA for localizing sentences/captions in videos that leverages both audio and video modalities and that can generalize to new and possibly low-resource language settings. We study a state-of-the-art model, as well as our MALTA on existing monolingual, video-heavy benchmarks as well as on our new dataset and observe clear advantages of using MALTA that leverages the audio modality. We also present a detailed investigation of different audio representations as well as pretraining, that gives us insights on how best to capture information from this modality for the alignment task.

6. Acknowledgements

We are grateful to IBM Research, India (specifically the IBM AI Horizon Networks - IIT Bombay initiative) for their support and sponsorship.

7. References

[1] J. Chen, X. Chen, L. Ma, Z. Jie, and T.-S. Chua, "Temporally grounding natural sentence in video," in *Proceedings of the*

2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium: Association for Computational Linguistics, Oct.-Nov. 2018, pp. 162–171. [Online]. Available: https://www.aclweb.org/anthology/D18-1015

[2] D. He, X. Zhao, J. Huang, F. Li, X. Liu, and S. Wen, "Read, watch, and move: Reinforcement learning for temporally grounding natural language descriptions in videos," *CoRR*, vol. abs/1901.06829, 2019. [Online]. Available: http://arxiv.org/abs/1901.06829

[3] L. A. Hendricks, O. Wang, E. Shechtman, J. Sivic, T. Darrell, and B. C. Russell, "Localizing moments in video with temporal language," *CoRR*, vol. abs/1809.01337, 2018. [Online]. Available: http://arxiv.org/abs/1809.01337

[4] M. Liu, X. Wang, L. Nie, Q. Tian, B. Chen, and T.-S. Chua, "Cross-modal moment localization in videos," 10 2018, pp. 843–851.

[5] R. Krishna, K. Hata, F. Ren, F. Li, and J. C. Niebles, "Dense-captioning events in videos," *CoRR*, vol. abs/1705.00754, 2017. [Online]. Available: http://arxiv.org/abs/1705.00754

[6] J. Gao, C. Sun, Z. Yang, and R. Nevatia, "TALL: temporal activity localization via language query," *CoRR*, vol. abs/1705.02101, 2017. [Online]. Available: http://arxiv.org/abs/1705.02101

[7] T. M. Y. Yuan and W. Zhu, "To find where you talk: Temporal sentence localization in video with attention based location regression. aaai," 2019.

[8] D. Zhang, X. Dai, X. Wang, Y. Wang, and L. S. Davis, "MAN: moment alignment network for natural language moment retrieval via iterative graph adjustment," *CoRR*, vol. abs/1812.00087, 2018. [Online]. Available: http://arxiv.org/abs/1812.00087

[9] V. Ramanishka, A. Das, D. H. Park, S. Venugopalan, L. A. Hendricks, M. Rohrbach, and K. Saenko, "Multimodal video description," in *Proceedings of the 24th ACM International Conference on Multimedia*, ser. MM '16. ACM, 2016, pp. 1092–1096.

[10] Q. Jin, J. Liang, and X. Lin, "Generating natural video descriptions via multimodal processing." in *Proceedings of Interspeech*, 2016, pp. 570–574.

[11] Q. Jin, J. Chen, S. Chen, Y. Xiong, and A. Hauptmann, "Describing videos using multi-modal fusion," in *Proceedings of the 24th ACM international conference on Multimedia*, 2016, pp. 1087–1091.

[12] C. Hori, T. Hori, T. K. Marks, and J. R. Hershey, "Early and late integration of audio features for automatic video description," in *Proceedings of ASRU*, 2017, pp. 430–436.

[13] Y.-G. Jiang, Z. Wu, J. Tang, Z. Li, X. Xue, and S.-F. Chang, "Modeling multimodal clues in a hybrid deep learning framework for video classification," *IEEE Transactions on Multimedia*, vol. 20, no. 11, pp. 3137–3147, 2018.

[14] X. Long, C. Gan, G. de Melo, X. Liu, Y. Li, F. Li, and S. Wen, "Multimodal keyless attention fusion for video classification," in *Proceedings of AAAI*, 2018.

[15] V. Vielzeuf, S. Pateux, and F. Jurie, "Temporal multimodal fusion for video emotion classification in the wild," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, 2017, pp. 569–576.

[16] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 221–231, 2013. [Online]. Available: https://doi.org/10.1109/TPAMI.2012.59

[17] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold *et al.*, "Cnn architectures for large-scale audio classification," in *Proceedings of ICASSP*, 2017, pp. 131–135.

[18] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 776–780.

[19] *Multi-task self-supervised learning for Robust Speech Recognition*, 01 2020.

[20] F. Zheng, G. Zhang, and Z. Song, "Comparison of different implementations of mfcc," *Journal of Computer science and Technology*, vol. 16, no. 6, pp. 582–589, 2001.

[21] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," 2011, iEEE Catalog No.: CFP11SRW-USB. [Online]. Available: http://infoscience.epfl.ch/record/192584