THEMES INFERRED AUDIO-VISUAL CORRESPONDENCE LEARNING

Runze Su1,2∗, Fei Tao1, Xudong Liu3, Haoran Wei1,4, Xiaorong Mei3, Zhiyao Duan1, Lei Yuan1, Ji Liu1, Yuying Xie2

1Y-tech Lab, Kuaishou Technology, Bellevue, USA
2Department of Statistics and Probability, Michigan State University, USA
3Kwai Ads Platform, Beijing, China
4Department of ECE, University of Texas at Dallas, USA

ABSTRACT
The applications of short-term user generated video (UGV), such as snapchat, youtube short-term videos, booms recently, raising lots of multimodal machine learning tasks. Among them, learning the correspondence between audio and visual information from videos is a challenging one. Most previous work of the audio-visual correspondence (AVC) learning only investigated on constrained videos or simple settings, which may not fit the application of UGV. In this paper, we proposed new principles for AVC and introduced a new framework to set sight on the themes of videos to facilitate AVC learning. We also released the KW AI-AD-AudVis corpus which contained 85432 short advertisement videos (around 913 hours) made by users. We evaluated our proposed approach on this corpus and it was able to outperform the baseline by 23.15% absolute difference.

Index Terms— audio-visual correspondence, multimodal signal processing, deep learning

1. INTRODUCTION
Recently, the applications of short-term user generated video (UGV) booms fast, such as youtube short-term videos, snapchat and Kwai. Audio-visual correspondence (AVC) learning, which can tell whether or how well the audio and visual information matches in the video, is able to bring benefits to these applications. It can recommend audio or visual streams to users given the other modality for more “like”, evaluate the quality of the short-term video for pushing to users, and build better high-level representation of videos for other uses.

Efforts have been made on the AVC learning. However, most previous works have limitations, mainly lying on two shortages: the task setting was simple, such as matching background audio to single image [1]; and the approaches relied on simple assumption that audio and visual information should be similar in some projected space [2]. These shortages may fail the systems on UGVs, whose information is complex and complicated. One example can illustrate this: user may combine a series of cheerful wedding ceremony photos and low-spirited style music to present the theme marriage is the tomb of love – spoken by Giacomo Casanova.

To tackle with the complexity of the UGVs, we propose theme inferred AVC (Ti-AVC) learning to introduce theme information in AVC learning. This idea is based on the following thoughts. All UGVs convey authors’ ideas (video themes), which may consist of several aspects. Each aspect is reflected by one or more modalities. Since the theme can contain complex semantic information, the aspects may disagree with each other. Simply measuring agreement or similarity of the modalities cannot tell how they match. The matched audio and visual information should follow two principles: 1) they need to convey a desired theme together; 2) there should be relation between them when present the theme. For the first principle, we designed a new way to inject the theme information into AVC learning. Since it is not clear how to represent the theme, we adopted the video tags to model theme indirectly in this paper. For the second principle, we follow conventional idea and adopted a state-of-the-art framework to model the relationship.

To evaluate our proposed idea, we collected 85432 UGVs from Kwai, a popular short-term video app in China. All the collected videos are advertisement (ads) uploaded by commercial advertiser, whose video quality is under control to some level. We will publish the dataset as the extension of KWAI-AD [3] dataset. It is shown that our proposed approach is able to gain 23.15% improvement (absolute value) compared with a state-of-the-art AVC learning framework.

We summarize the contribution of this paper below: 1) We introduced new principles for AVC learning task and proposed the first theme inferred audio-visual correspondence (Ti-AVC) framework which is suitable for UGVs. It outperformed the state-of-the-art baseline by 23.15% absolute difference. 2) We publish the first audio-visual dataset grouped by contents based on short-term ads video. 3) We interpret the hidden values of the model to analyze the modality contribution in AVC.

∗This work was done while the author was an intern at Kwai Y-tech lab.
2. RELATED WORK

Researchers cast much attention on reciprocity between audio and visual information on various tasks [4]. Transfer learning was proposed to convey information across modalities [5, 6]. But the correspondence between modalities was not modeled in these works.

$L^3$ net [7] was proposed to explicitly model AVC. It used several sub-networks to perform inputs processing and modalities fusion. Relying on the max-pooling layer in the fusion sub-network, the $L^3$ net had flexible framework that was able to take sequential or single input. It showed state-of-the-art performance and therefore was widely adopted later: [5, 8, 9] adopted $L^3$ net to perform sound localization task. [10, 11] deployed pre-trained $L^3$ net as embedding extractor. [12] applied $L^3$ net framework to learn AVC based on the emotion from audio and visual streams. This work was evaluated on a new released dataset that contained audio and visual emotion information. Other approaches were also proposed to improve the AVC. Gating mechanism [13] was applied to filter uncorresponding information in audio and visual streams. Dual attention matching [9] added attention to both audio and visual inputs to predict their event sequential localization relevance between modalities. Elastic multi-way network [14] designed loss function with the distance between samples and an anchor point to evaluate correspondence. [15] relied on bimodal recurrent neural network to learn the temporal correspondence information in a data-driven fashion. Unsupervised methods such as video audio correspondence are also investigated such as audio-visual deep clustering model [16]. Most of the approaches focused on modeling the similarity between modalities and showed decent performance. However, most of the approaches were evaluated on constrained dataset. AVC learning on unconstrained data is still a complicated and difficult task. [17, 18].

The approaches should be developed and evaluated on more general and unconstrained audio-visual dataset, like UGVs dataset. Youtube-8M [19] is one of the most popular UGVs dataset, which covers various themes (i.e. tags). However, the video quality is not controlled intentionally. Also, the video duration is too long and not suitable for the case of short-term videos. On the other side, the Moments in Time [20] contains 1,000,000 3-second videos, which is too short. Flickr-SoundNet [1] is an unconstrained dataset, however it only has single image with background sound track. The shortage of good quality short-term UGVs inspired us to collect a new data, whose details will be introduced later.

The approaches should be developed and evaluated on more general and unconstrained audio-visual dataset, like UGVs dataset. Youtube-8M [19] is one of the most popular UGVs dataset, which covers various themes (i.e. tags). However, the video quality is not controlled intentionally. Also, the video duration is too long and not suitable for the case of short-term videos. On the other side, the Moments in Time [20] contains 1,000,000 3-second videos, which is too short. Flickr-SoundNet [1] is an unconstrained dataset, however it only has single image with background sound track. The shortage of good quality short-term UGVs inspired us to collect a new data, whose details will be introduced later.

3. KWAI-AD-AUDVIS DATASET

In this study, we developed our framework on KWAI-AD-AudVis dataset. It is consisted of 85,432 ads videos (around 913 hours) from the China popular short-term video app, Kwai. The videos were made and uploaded by commercial advertisers rather than personal users. The reason to use the ads videos lied on two folds: 1) the source guarantees the videos under control to some level, such as high-resolution pictures and intentionally designed scene; 2) the ads videos mimic the style of the ones uploaded by personal users, as they are played in between the personal videos in Kwai app. It can be seen as a quality controlled UGVs dataset.

The dataset was collected in two batches (Batch-1 is our preliminary work. Its industry categories is subset of Batch-2), coming with the label of ads industry category. The videos were randomly picked from a pool. The pool was formed by selecting the ads from several contiguous days. Half of the selected ads had click through rate (CTR) in top 30000 within that day and the other half had CTR in bottom 30000. It should be noticed that the released dataset is a subset of the pool. The statistics of the duration are summarized in Table 1. The audio track had 2 channels (we mixed to mono channel in the study) and was sampled at 44.1 kHz, while the visual track had resolution of $1280 \times 720$ and was sampled at $25$ frame per second (FPS). This dataset is a extension of the KWAI-AD corpus [3]. It is not only suitable for tasks in multimodal learning area, but also for ones in ads recommendation.

The details and data of KWAI-AD-AudVis can be accessed through Zenodo [1]. It shows that the ads videos have three main characteristics: 1) The videos may have very inconsistent information in visual or audio streams. For example, the video may play a drama-like story at first, and then present the product introduction, whose scenes are very different. 2) The correspondence between audio and visual streams is not clear. For instance, similar visual objects (e.g. talking salesman) come with very different audio streams. 3) The relationship between audio and video varies in different industries. For example, game or E-commerce ads will have very different styles. These characteristics make the dataset suitable yet challenging for our study about the AVC learning.

4. PROPOSED APPROACHES

4.1. Data and Feature

In this study, we used KWAI-AD-AudVis dataset to develop our AVC learning framework. To reduce the data size and training workload, we used our in-house key-frame extractor to extract 8 images from each video to represent the visual information. The resulting images did not follow the original order in the video. Audio tracks were extracted as same as in original videos. The visual and audio information are

| Batch ID | Dura Mean | Dura Std | Ind Categories |
|----------|-----------|----------|----------------|
| Batch-1  | 39.85     | 18.70    | 8              |
| Batch-2  | 37.99     | 18.23    | 19             |

Table 1. KWAI-AD2 dataset statistics. The "Dura Mean" stands for average duration. The "Dura Std" is the standard deviation of the duration. "Ind Categories" shows the number of industry categories.

1 https://zenodo.org/record/4023390#X12Dr5NKgUE
pre-processed through Mobilenetv2 \cite{21} and VGGish \cite{22}. Embedding from top layers of the two pre-trained was fed to our proposed system.

4.2. Theme Inferred AVC learning System

Figure 1 shows the diagram of our proposed approach, theme-inferred audio-visual correspondence (Ti-AVC) learning framework. It consisted of two parts, a theme-learning (TL) model and a correspondence-learning (CL) model.

For the TL model, we were inspired by the $L^3$ net and designed a similar network to $L^3$ net, except its task was theme prediction (in this study, it is ads industry category prediction). It took the audio and visual embedding as input. It consisted of three sub-networks. Two sub-network processed the input of single modality separately, and the third one processed the fused information. The audio sub-network had a time distributed dense layer, an LSTM layer and a self-attention layer. The output is a 128-dimensional vector. The visual sub-network had a time series dense layer. The output was a sequence of 8 128-dimensional vectors. The output from audio sub-network were repeated and concatenated to each vector from visual sub-network. The concatenated embedding was fed into the fusion sub-network. The fusion sub-network had two 1-D convolutional neural networks (CNN), a max-pooling layer and 2 fully connected (FC) layers to predict themes.

Once the TL model was trained, we fixed it as embedding extractor to extract three types of information for the CL model: audio embedding from top-layer of the audio sub-network, visual embedding from top-layer of the visual sub-network and the predicted theme. We concatenated the predicted theme with the theme ground truth to form the complete theme information, which was injected into the CL model with audio and visual embedding. By adding the true theme, we expect the CL model to learn how the input audio and visual embedding performed in predicting the theme. This was following the principle (1) we mentioned in section 1. In this study, the theme ground-truth corresponded to the visual modalities. The CL model has similar architecture as the fusion sub-network in the TL model. We intended to use both of the theme prediction and ground-truth for telling how the two modalities represent desired theme. The CL model was expected to capture two points: 1) how the desired theme was presented; 2) how the modalities related to each other. These two points corresponded to the two principles we proposed in Section 1. The correspondence result was eventually predicted based on the two points. It should be noticed that it was not cheating to have the theme ground-truth in CL model, since we assumed this information available during AVC learning.

5. EXPERIMENT AND ANALYSIS

5.1. Experiment Setup

We used the original videos from KWAI-AD-AudVis dataset as positive samples, where we assumed audio and visual information matched with each other. Negative samples were generated by pairing audio and visual tracks from different videos. We generated same number of negative samples as positive ones. The dataset was partitioned to 80%, 10% and 10% for training, validation and testing respectively. We applied Adam as optimizer and set a learning rate for both models in all experiments as 0.0001, batch size as 8. We also set an early stopper, which monitors loss on validation set. If validation loss doesn’t decrease for 5 epochs, the training would stop. We used ads industries categories as theme information in this study.

We built two baselines for comparison. We made two changes on the TL model to form our first baseline (named as “baseline-1”): 1) we replaced the theme prediction task by correspondence prediction; 2) we double the number of all trainable layers in the fusion sub-network to have same parameter number as our proposed approach. This baseline was very similar to $L^3$ net. The second baseline (named as “baseline-2”) had exactly same architecture as baseline-1, except we also input theme ground-truth to the fusion sub-network concatenated with the modalities embedding. This made the comparison fair since the system also got theme information like the proposed approach. For the proposed approach, we made two training strategies and therefore had two systems. We named the system that trained TL and CL models separately as “Ti-AVC”, while we named the one jointly trained (i.e. a multitask learning system with TL and CL tasks) as “joint Ti-AVC”. We kept all systems same number of parameters.

5.2. Experiment Results

The accuracy AUC score of each system is shown in Table 2. The baseline-1, which had similar architecture to $L^3$ net,
| Model          | Match AUC |
|----------------|-----------|
| Baseline-1     | 55.58%    |
| Baseline-2     | 74.52%    |
| Joint Ti-A VC  | 77.88%    |
| Ti-A VC        | 78.73%    |

Table 2. Summary of experiment results.

had random-guess results (we had same amount of positive and negative samples). The baseline-2, which was same as baseline-1 except it took theme as extra input, could outperform baseline-1 by 18.94%. This verified our hypothesis that theme information was necessary in AVC learning on UGVs. Both of our proposed approaches was able to beat the baselines (by at least 3.36% absolute difference) and the Ti-A VC achieved the best performance. Since the TL and CL models were trained separately in Ti-A VC, it indicated that properly injecting the information how the audio and visual modalities presented the desired theme could improve the performance of correspondence learning task. We would like to emphasize the Ti-A VC is flexible in application, because the TL model can be fixed as embedding extractor and the label ground-truth provided in CL model can be obtained from either modality (in this study, we made it follow visual modality).

We would like to check whether adding theme ground-truth in CL model provided a “cheating” shortcut for AVC task. We evaluated the Ti-A VC framework within each category and listed the accuracy AUC in Figure 2. This evaluation is equivalent to eliminate the help from providing theme information, since all the testing candidates came from the same category and had the same theme ground-truth. We compare the results with the baseline-1, which doesn’t include theme information neither. The result shows that the Ti-A VC framework can outperform the baseline in 15 categories out of 19. Especially, we notice all the categories with most samples outperformed the baseline. This result indicates that the proposed framework can help improve the correspondence learning even without theme information.

5.3. Contribution Analysis

To further verify the rationality of our proposed approach, we analyzed the contribution of each input in CL model. We define the contribution in equation 1, where $W_i$ is the weight of the first layer connecting the $i_{th}$ input and $X_i$ is the $i_{th}$ input, $I$ is the input type (audio, visual, predicted theme and true theme).

$$
\text{Contribution}^l = \sum |W_{i}^l \cdot X_{i}^l| \quad (1)
$$

We listed the computed proportion of the inputs for the matched pairs in Table 3. It showed that audio modalities has the most portion contribution (58.78%). The theme information took up 10.38%, where the predicted and true ones were close (4.52% and 5.86%). This result indicated both of them

![Fig. 2. AUC and sample counts per ADs category. The green bar represents the AUC, whose scale axis is on left; the blue bar represents the number of samples, whose scale axis is on right. The red line is the baseline-1 accuracy AUC.](image)

| Vision | Audio | Predicted Themes | True Themes |
|--------|-------|------------------|-------------|
| 30.85% | 58.78%| 4.52%            | 5.86%       |

Table 3. Modal contributions calculated from a batch of positive audio-visual pairs and a batch of negative audio-visual pairs.

could not be neglected, which verified our proposed principles for AVC in section and the capability of the proposed approach.

6. CONCLUSION AND FUTURE WORK

In this paper, we proposed new principles in audio-visual correspondence learning task on users generated videos, which introduced theme information in AVC task. We proposed a new framework to perform the AVC task under unconstrained scenarios. To evaluate the proposed approach, we also collected and released the KWAI-AD-AudVis corpus, consisting of 85432 short-term videos (around 913 hours).

Our proposed approach was able to outperform a state-of-the-art AVC framework by 23.15% (absolute difference) in accuracy AUC. We also showed that the proposed approach could still outperform the baseline even without theme information. Besides, the proposed framework would be flexible in real application because the TL model can be fixed and the theme information can correspond to either modality. This study only focused on learning correspondence between audio and visual modalities. We directly concatenated the embedding of the modalities in the framework. The future work lies on applying more sophisticated fusion strategy and further analyzing how the modality correlate with each other in matching.
7. REFERENCES

[1] Y. Aytar, C. Vondrick, and A. Torralba, “Soundnet: Learning sound representations from unlabeled video,” in Advances in neural information processing systems, 2016, pp. 892–900.

[2] R. Arandjelovic and A. Zisserman, “Look, listen and learn,” in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 609–617.

[3] H. Chen, G. Ding, X. Liu, Z. Lin, J. Liu, and J. Han, “Imram: Iterative matching with recurrent attention memory for cross-modal image-text retrieval,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 12 655–12 663.

[4] F. Tao and C. Busso, “End-to-end audiovisual speech recognition system with multitask learning,” IEEE Transactions on Multimedia, 2020.

[5] H. Zhao, C. Gan, A. Rouditchenko, C. Vondrick, J. McDermott, and A. Torralba, “The sound of pixels,” in Proceedings of the European conference on computer vision (ECCV), 2018, pp. 570–586.

[6] F. Tao and C. Busso, “Aligning audiovisual features for audiovisual speech recognition,” in IEEE International Conference on Multimedia and Expo (ICME 2018), San Diego, CA, USA, July 2018, pp. 1–6.

[7] Y. Aytar, C. Vondrick, and A. Torralba, “See, hear, and read: Deep aligned representations,” arXiv preprint arXiv:1706.00932, 2017.

[8] R. Arandjelovic and A. Zisserman, “Objects that sound,” in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 435–451.

[9] Y. Wu, L. Zhu, Y. Yan, and Y. Yang, “Dual attention matching for audio-visual event localization,” in Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 6292–6300.

[10] J. Cramer, H.-H. Wu, J. Salamon, and J. P. Bello, “Look, listen, and learn more: Design choices for deep audio embeddings,” in ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019, pp. 3852–3856.

[11] S.-W. Chung, J. S. Chung, and H.-G. Kang, “Perfect match: Improved cross-modal embeddings for audio-visual synchronisation,” in ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019, pp. 3965–3969.

[12] G. Verma, E. G. Dhekane, and T. Guha, “Learning affective correspondence between music and image,” in ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019, pp. 3975–3979.

[13] F. Tao and C. Busso, “Gating neural network for large vocabulary audiovisual speech recognition,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 26, no. 7, pp. 1286–1298, July 2018.

[14] R. Wang, H. Huang, X. Zhang, J. Ma, and A. Zheng, “A novel distance learning for elastic cross-modal audiovisual matching,” in 2019 IEEE International Conference on Multimedia & Expo Workshops (ICM EW). IEEE, 2019, pp. 300–305.

[15] F. Tao and C. Busso, “End-to-end audiovisual speech activity detection with bimodal recurrent neural models,” ArXiv e-prints (arXiv:1809.04553), pp. 1–11, September 2018.

[16] R. Lu, Z. Duan, and C. Zhang, “Audio–visual deep clustering for speech separation,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, no. 11, pp. 1697–1712, 2019.

[17] H. Zhu, M. Luo, R. Wang, A. Zheng, and R. He, “Deep audio-visual learning: A survey,” arXiv preprint arXiv:2001.04758, 2020.

[18] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, “Multimodal machine learning: A survey and taxonomy,” IEEE transactions on pattern analysis and machine intelligence, vol. 41, no. 2, pp. 423–443, 2018.

[19] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan, “Youtube-8m: A large-scale video classification benchmark,” arXiv preprint arXiv:1609.08675, 2016.

[20] M. Monfort, A. Andonian, B. Zhou, K. Ramakrishnan, S. A. Bargal, T. Yan, L. Brown, Q. Fan, D. Gutfreund, C. Vondrick et al., “Moments in time dataset: one million videos for event understanding,” IEEE transactions on pattern analysis and machine intelligence, vol. 42, no. 2, pp. 502–508, 2019.

[21] S. Abu-Ellail, 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 2017 ieee international conference on acoustics, speech and signal processing (icassp). IEEE, 2017, pp. 131–135.
8. REFERENCES

[1] Y. Aytar, C. Vondrick, and A. Torralba, “Soundnet: Learning sound representations from unlabeled video,” in Advances in neural information processing systems, 2016, pp. 892–900.

[2] R. Arandjelovic and A. Zisserman, “Look, listen and learn,” in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 609–617.

[3] H. Chen, G. Ding, X. Liu, Z. Lin, J. Lui, and J. Han, “Imram: Iterative matching with recurrent attention memory for cross-modal image-text retrieval,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 12655–12663.

[4] F. Tao and C. Busso, “End-to-end audiovisual speech recognition system with multitask learning,” IEEE Transactions on Multimedia, 2020.

[5] H. Zhao, C. Gan, A. Rouditchenko, C. Vondrick, J. McDermott, and A. Torralba, “The sound of pixels,” in Proceedings of the European conference on computer vision (ECCV), 2018, pp. 570–586.

[6] F. Tao and C. Busso, “Aligning audiovisual features for audiovisual speech recognition,” in IEEE International Conference on Multimedia and Expo (ICME 2018), San Diego, CA, USA, July 2018, pp. 1–6.

[7] Y. Aytar, C. Vondrick, and A. Torralba, “See, hear, and read: Deep aligned representations,” arXiv preprint arXiv:1706.00932, 2017.

[8] R. Arandjelovic and A. Zisserman, “Objects that sound,” in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 435–451.

[9] Y. Wu, L. Zhu, Y. Yan, and Y. Yang, “Dual attention matching for audio-visual event localization,” in Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 6292–6300.

[10] J. Cramer, H.-H. Wu, J. Salamon, and J. P. Bello, “Look, listen, and learn more: Design choices for deep audio embeddings,” in ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019, pp. 3852–3856.

[11] S.-W. Chung, J. S. Chung, and H.-G. Kang, “Perfect match: Improved cross-modal embeddings for audiovisual synchronisation,” in ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019, pp. 3965–3969.

[12] G. Verma, E. G. Dhekane, and T. Guha, “Learning affective correspondence between music and image,” in ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019, pp. 3975–3979.

[13] F. Tao and C. Busso, “Gating neural network for large vocabulary audiovisual speech recognition,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 26, no. 7, pp. 1286–1298, July 2018.

[14] R. Wang, H. Huang, X. Zhang, J. Ma, and A. Zheng, “A novel distance learning for elastic cross-modal audiovisual matching,” in 2019 IEEE International Conference on Multimedia & Expo Workshops (ICMEW). IEEE, 2019, pp. 300–305.

[15] F. Tao and C. Busso, “End-to-end audiovisual speech activity detection with bimodal recurrent neural models,” ArXiv e-prints (arXiv:1809.04553), pp. 1–11, September 2018.

[16] R. Lu, Z. Duan, and C. Zhang, “Audio–visual deep clustering for speech separation,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, no. 11, pp. 1697–1712, 2019.

[17] H. Zhu, M. Luo, R. Wang, A. Zheng, and R. He, “Deep audio-visual learning: A survey,” arXiv preprint arXiv:2001.04758, 2020.

[18] T. Baltrusaitis, C. Ahuja, and L.-P. Morency, “Multimodal machine learning: A survey and taxonomy,” IEEE transactions on pattern analysis and machine intelligence, vol. 41, no. 2, pp. 423–443, 2018.

[19] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan, “Youtube-8m: A large-scale video classification benchmark,” arXiv preprint arXiv:1609.08675, 2016.

[20] M. Monfort, A. Andonian, B. Zhou, K. Ramakrishnan, S. A. Bargal, T. Yan, L. Brown, Q. Fan, D. Gutfreund, C. Vondrick et al., “Moments in time dataset: one million videos for event understanding,” IEEE transactions on pattern analysis and machine intelligence, vol. 42, no. 2, pp. 502–508, 2019.

[21] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, “Mobilenetv2: Inverted residuals and linear bottlenecks,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 4510–4520.

[22] 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 2017 ieee international conference on acoustics, speech and signal processing (icassp). IEEE, 2017, pp. 131–135.