Eyes and Ears Together: New Task for Multimodal Spoken Content Analysis

Yasufumi Moriya[1], Ramon Sanabria [2], Florian Metze [2], Gareth J. F. Jones[1]
[1]Dublin City University, Dublin, Ireland
[2]Carnegie Mellon University, Pittsburgh, PA, USA
{yasufumi.moriya,gareth.jones}@adaptcentre.ie,{ramons,fmetze}@cs.cmu.edu

ABSTRACT

Human speech processing is often a multimodal process combining audio and visual processing. Eyes and Ears Together proposes two benchmark multimodal speech processing tasks: (1) multimodal automatic speech recognition (ASR) and (2) multimodal co-reference resolution on the spoken multimedia. These tasks are motivated by our desire to address the difficulties of ASR for multimedia spoken content. We review prior work on the integration of multimodal signals into speech processing for multimedia data, introduce a multimedia dataset for our proposed tasks, and outline these tasks.

1 INTRODUCTION

Human use of natural language for communication is grounded in real world entities, concepts, and activities. The importance of the real world in language interpretation is illustrated in [18], where participants were presented with a picture of an apple on a towel, a towel without an apple, and a box. When they heard the sentence "put the apple on the towel", their gaze moved to the towel without an apple, before the reader finished the complete sentence "put the apple on the towel in the box". Another example of visual grounding in language understanding is the McGurk effect [9]. When participants were exposed to the voiced alveolar stop ("da") sound, and to a video, whose lip movement indicates the voiced bilabial stop ("ba") sound, they perceived the bilabial sound rather than the alveolar sound. Furthermore, it is reported that the presence of a speaker's face facilitates speech comprehension [19]. These experiments all demonstrate that human language processing is affected by the context provided in visual signals.

Despite those findings, research on automatic speech recognition (ASR) has generally focused only on audio signals, even if the use of visual and contextual information could be considered (e.g., in multimedia data where the audio is accompanied by a video data stream and metadata). However, high word error rates (WERs) of 30-40% are often reported for ASR of multimedia data from contemporary sources such as YouTube videos and TV shows [1, 8]. More recent work on ASR for YouTube videos illustrates that much lower WERs are possible [17], but the use of 100k hours of data for system development is not feasible in many situations.

This paper presents two proposed multimodal speech processing benchmark tasks motivated by the multimodal nature of human language processing and the practical difficulty of automated spoken data processing. The reminder of the paper is organised as follows: Section 2 reviews previous work on multimodal processing of spoken multimedia data. Section 3 presents an audio-visual dataset suitable for our proposed tasks. Section 4 introduces our potential tasks for spoken multimedia understanding. Section 5 provides concluding remarks.

MediaEval'18, 29-31 October 2018, Sophia Antipolis, France

2 PRIOR WORK

Figure 1: Comparison of the Grid corpus for AVSR (left) to the CMU "How-to" corpus for multimodal ASR (right).

The integration of visual information into ASR systems is a long-standing topic of investigation in the field of audio-visual speech recognition (AVSR). Motivated by the McGurk effect [9], AVSR aims to build noise-robust ASR systems by incorporating lip movement into recognition of phonemes [15]. The most recent approach to AVSR employs a multimodal deep neural network (DNN) to fuse visual lip movement with audio features [13].

Although AVSR is known to be effective on noisy audio conditions, application of the AVSR is limited to situations, where a speaker frontal face is visible to enable lip movement features to be extracted. Figure 1 shows a comparison of AVSR (Grid corpus) [2] with multimodal ASR (CMU "How-to" corpus) [3, 16] [1]. As shown in the Grid corpus example, constraints on an AVSR dataset are: (1) the presence of a speaker mouth region, and (2) precise synchronisation of a visual signal with speech. Multimodal ASR can exploit any available contextual information to improve ASR accuracy.

Recent work has begun to explore the use of more general multimodal information in ASR. Figure 2 demonstrates a basic framework for the integration of contextual information into ASR using a DNN acoustic model [4] and a recurrent neural network (RNN) language model with long short-term memory (LSTM) [5, 10]. In this framework, a convolutional neural network (CNN) model extracts a fixed-length image feature vector from the video frame within the time region of each utterance. The image feature vector is concatenated with the audio feature vector for the DNN acoustic model. Alternatively, the image feature vector can be taken as

[1]The corpus was also used in the 2018 Jelinek workshop: https://www.clsp.jhu.edu/workshops/18-workshop/

the input of the first token of the RNN language model, before the model reads embedded word tokens of the utterance. It should be noted that the contextual feature vector does not need to be extracted from a video frame, but can be taken from any feature that represents the environment of the utterance being spoken. Typically, the DNN acoustic model is used for ASR with a weighted finite-state transducer [11], and the RNN language model re-scores n-best hypotheses generated in the ASR decoding step.

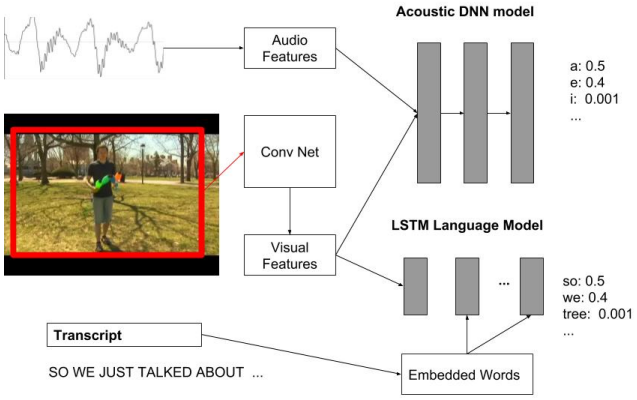

Figure 2: Framework for integration of visual features into ASR. A convolutional neural network (CNN) extracts a fixed-length vector from a video frame, which is appended to either an audio feature vector or fed to the neural language model before reading embedded word tokens.

A number of interesting findings have been reported in the existing work on multimodal ASR. Gupta et al. extracted object features and scene features from a video frame randomly chosen from within the time range of each utterance [3]. They used these features to adapt the DNN acoustic model and the RNN language model. Scene features were particularly effective in improving recognition of utterances being spoken outside. It is likely that enabling the acoustic model to know that the audio input may contain background noise implicitly transforms the audio features into a cleaner representation. Moriya and Jones investigated whether video titles can provide the RNN language model with background context of each video [12]. They represented each video title as the average of embedded words in the title, and found that the adapted model predicted "keywords" of a video better than the non-adapted model (i.e., "fish" in a fishing video).

Huang et al. conducted a new line of work that connects speech transcription with vision. In [7], they propose a method to align entities in a video with actions that produce the entities. Their goal was to jointly resolve linguistic ambiguities (e.g,. "oil mixed with salt" can be referred to as "the mixture"), and visual ambiguities (e.g., "yogurt" can look similar to "dressing"). This approach was further extended to a multimodal co-reference resolution system which links entities in a video with the objects in a transcription, and even with referring expressions (e.g., "it") [6]. Their system was evaluated on the YouCook2 dataset, a collection of unstructured cooking videos [20]. Although spoken transcriptions are accompanied by video data, the transcriptions are simplified to imperative sentences and do not represent real utterances that are actually spoken in videos. We believe that this may form an interesting new task to analyse environments (video) of utterances (speech) being spoken.

3 DATASET

This section outlines the CMU "How-to" corpus [16]. The corpus contains instruction videos from *YouTube*, speech transcriptions and various types of meta-data (e.g., video titles, video description, the number of likes). An example image from the corpus is shown in Figure 1. Audio conditions of videos vary, e.g. some of the videos are recorded outdoors with background noise present. The corpus was used for experiments in [3] and [12]. Two different setups of the corpus are provided: 480 hours of audio and 90 hours of audio. In the both setups, development and test partition remain the same. In [12], symbols and numbers in the transcription were removed or expanded to words. In addition, regions of transcription that are likely to be a mismatch with audio were rejected. For this reason, the experimental results in [3] and [12] are not directly comparable. We propose the creation of a standardised version of the corpus for development of common multimodal ASR task.

4 TASK DESCRIPTION

We propose two tasks for investigation with spoken multimedia content: multimodal ASR and multimodal co-reference resolution.

4.1 Multimodal ASR

Multimodal ASR is a conventional ASR task that focuses on the use of multimodal signals in ASR with effectiveness measured using standard WER. The two main goals of this task are: (1) identifying visual or contextual features that contribute to the improvement of ASR systems; (2) exploring suitable ASR system architectures for better exploitation of visual or contextual features. The former encourages participants to explore alternative features available in videos and meta-data in ASR, e.g., temporal features. The latter aims to explore unconventional architectures for ASR systems, e.g., use of a end-to-end neural architecture [14].

4.2 Multimodal Co-reference Resolution

The goal of multimodal co-reference resolution aims to bridge the gap between the speech modality and the visual modality. We plan to provide participants with ASR transcription containing pronouns and referred objects appearing in a video with the task of resolving the pronouns. Effectiveness may be measured using F1 scores, as in [6]. Such resolution may find utility in reducing WER in second pass ASR decoding.

5 CONCLUSION

This paper presents potential tasks for multimodal spoken content analysis. The motivation for the use of multimodal grounding in ASR arises from the multimodal nature of human language understanding and from the poor performance of ASR systems, when applied to multimedia data. Section 2 highlights existing work on integration of multimodal signals into ASR. Section 3 introduces a multimodal dataset suitable for use in the proposed tasks, and Section 4 outlines details of two proposed tasks.

REFERENCES

[1] P. Bell, M. J. F. Gales, T. Hain, J. Kilgour, P. Lanchantin, X. Liu, A. McParland, S. Renals, O. Saz, M. Wester, and P. C. Woodland. 2015. The MGB challenge: Evaluating multi-genre broadcast media recognition. In *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. 687–693.

[2] M. Cooke, J. Barker, S. Cunningham, and X. Shao. 2006. An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America* 120, 5 (2006), 2421–2424. https://doi.org/10.1121/1.2229005

[3] A. Gupta, Y. Miao, L. Neves, and F. Metze. 2017. Visual features for context-aware speech recognition. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 5020–5024.

[4] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury. 2012. Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. *IEEE Signal Processing Magazine* 29, 6 (Nov 2012), 82–97. https://doi.org/10.1109/MSP.2012.2205597

[5] S Hochreiter and J Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–80.

[6] D. A. Huang, S. Buch*, L. Dery, A. Garg, L. Fei-Fei, and J. C. Niebles. 2018. Finding "It": Weakly-Supervised, Reference-Aware Visual Grounding in Instructional Videos. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*. 5948–5957.

[7] D. A. Huang, J. J. Lim., L. Fei-Fei, and J. C. Niebles. 2017. Unsupervised Visual-Linguistic Reference Resolution in Instructional Videos. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*. 2183–2192.

[8] H. Liao, E. McDermott, and A. Senior. 2013. Large scale deep neural network acoustic modeling with semi-supervised training data for YouTube video transcription. In *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. 368–373.

[9] H. McGurk and J. MacDonald. 1976. Hearing lips and seeing voices. *Nature* 264, 5588 (1976), 746–748.

[10] T. Mikolov, M. Karafiat, L. Burget, J. Cernocky, and S. Khudanpur. 2010. Recurrent neural network based language model. In *Proceedings of Interspeech*. 1045–1048.

[11] M. Mohri, F. Pereira, and M. Riley. 2002. Weighted finite-state transducers in speech recognition. *Computer Speech & Language* 16, 1 (2002), 69–88.

[12] Y. Moriya and G. J. F. Jones. 2018. LSTM language model adaptation with images and titles for multimedia automatic speech recognition. In *(to appear) Workshop on Spoken Language Technology (SLT)*.

[13] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng. 2011. Multimodal deep learning. In *Proceedings of the International Conference on Machine Learning (ICML)*. 689–696.

[14] S. Palaskar, R. Sanabria, and F. Metze. 2018. End-to-End Multimodal Speech Recognition. In *International Conference on Acoustic, Speech and Signal Processing (ICASSP)*. 5774–5778.

[15] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. W. Senior. 2003. Recent advances in the automatic recognition of audiovisual speech. *Proc. IEEE* 91, 9 (Sept 2003), 1306–1326. https://doi.org/10.1109/JPROC.2003.817150

[16] R. Sanabria, O. Caglayan, S. Palaskar, D. Elliott, L. Barrault, L. Specia, and F Metze. 2018. How2: A large-scale dataset for multimodal language understanding. In *(to appear) Proceedings of Neural Information Processing Systems (NIPS)*.

[17] H. Soltau, H. Liao, and H Sak. 2016. Neural speech recognizer: acoustic-to-word LSTM model for large vocabulary speech recognition. *arXiv* (2016). http://arxiv.org/abs/1610.09975

[18] M. K. Tanenhaus, M. J. Spivey-Knowlton, K. M. Eberhard, and J. C Sedivy. 1995. Integration of visual and linguistic information in spoken language comprehension. *Science* 268, 5217 (1995), 1632–1634.

[19] V. van Wassenhove, K. W. Grant., and D. Poeppel. 2005. Visual speech speeds up the neural processing of auditory speech. *Proceedings of the National Academy of Sciences* 102, 4 (2005), 1181–1186. https://doi.org/10.1073/pnas.0408949102

[20] L. Zhou, C. Xu, and J. J. Corso. 2018. Towards Automatic Learning of Procedures from Web Instructional Videos. In *AAAI*. 7590–7598.