LRS3-TED: a large-scale dataset for visual speech recognition

Triantafyllos Afouras, Joon Son Chung, Andrew Zisserman

Visual Geometry Group, Department of Engineering Science,
University of Oxford, UK

{afourast, joon, az}@robots.ox.ac.uk

Abstract

This paper introduces a new multi-modal dataset for visual and audio-visual speech recognition. It includes face tracks from over 400 hours of TED and TEDx videos, along with the corresponding subtitles and word alignment boundaries. The new dataset is substantially larger in scale compared to other public datasets that are available for general research.

Index Terms: lip reading, visual speech recognition, large-scale, dataset

1. Introduction

Visual speech recognition (or *lip reading*) is a very challenging task, and a difficult skill for a human to learn. In the recent years, there has been significant progress [1, 2, 3, 4] in the performance of automated lip reading due to the application of deep neural network models and the availability of large scale datasets. However, most of these datasets are subject to some restrictions (e.g. LRW [5] or the LRS2-BBC [6] cannot be used by industrial research labs) and this has meant that it is difficult to compare the performance of one lip-reading system to another, as there is no large scale common benchmark dataset. Our aim in releasing the LRS3-TED dataset is to provide such a benchmark dataset, and one that is larger in size compared to any available dataset in this field.

The `LRS3-TED` dataset can be downloaded from http://www.robots.ox.ac.uk/~vgg/data/lip_reading.

2. LRS3-TED dataset

The dataset consists of over 400 hours of video, extracted from 5594 TED and TEDx talks in English, downloaded from YouTube.

The cropped face tracks are provided as `.mp4` files with a resolution of 224×224 and a frame rate of 25 fps, encoded using the h264 codec. The audio tracks are provided as single-channel 16-bit 16kHz format, while the corresponding text transcripts, as well as the alignment boundaries of every word are included in plain text files.

The dataset is organized into three sets: *pre-train*, *train-val* and *test*. The first two overlap in terms of content but the last is completely independent. The statistics for each set are given in Table 1.

2.1. Data collection

We use a multi-stage pipeline for automatically generating the large-scale dataset for audio-visual speech recognition. Using this pipeline, we have been able to collect hundreds of hours of spoken sentences and phrases along with the corresponding facetrack.

We start from the TED and TEDx videos that are available on their respective YouTube channels. These videos were se-

lected for mutliple reasons: (1) a wide range of speakers appears in the videos, unlike movies or dramas with a fixed cast; (2) shot changes are less frequent, therefore there are more full sentences with continuous facetracks; (3) the speakers usually talk without interruption, allowing us to obtain longer face tracks. TED videos have previously been used for audio-visual datasets for these reasons [9].

The pipeline is based on the methods described in [1, 6], but we give a brief sketch of the method here.

Video preparation. We use a CNN face detector based on the Single Shot MultiBox Detector (SSD) [10] to detect face appearances in the individual frames.

The time boundaries of a shot are determined by comparing color histograms across consecutive frames [11], and within each shot, face tracks are generated from face detections based on their positions.

Audio and text preparation. Only the videos providing english subtitles created by humans were used. The subtitles in the YouTube videos are broadcast in sync with the audio only at sentence-level, therefore the Penn Phonetics Lab Forced Aligner (P2FA) [12] is used to obtain a word-level alignment between the subtitle and the audio signal. The alignment is double-checked against an off-the-shelf Kaldi-based ASR model.

AV sync and speaker detection. In YouTube or broadcast videos, the audio and the video streams can be out of sync by up to around one second, which can introduce temporal offsets between the videos and the text labels (aligned to the audio). We use a two-stream network (SyncNet) described in [13] to synchronise the two streams. The same network is also used to determine which face's lip movements match the audio, and if none matches, the clip is rejected as being a voice-over.

Sentence extraction. The videos are divided into individual sentences/ phrases using the punctuations in the transcript. The sentences are separated by full stops, commas and question marks. The sentences in the *train-val* and *test* sets are clipped to 100 characters or 6 seconds.

The *train-val* and *test* sets are divided by videos (extracted from disjoint sets of original videos). Although we do not explicitly label the identities, it is unlikely that there are many identities that appear in both training and test sets, since the speakers do not generally appear on TED programs repeatedly. This is in contrast to the LRW and the LRS2-BBC datasets that are based on regular TV programs, hence the same characters are likely to appear in common from one episode to the next.

The *pre-train* set is more extensive, as it contains videos spanning the full duration of the face track, along with the corresponding subtitles. It is extracted from the same set of original YouTube videos as the *train-val* set. However, these videos may be shorter or longer than the full sentences included in the *train-val* and *test* sets, and are annotated with the alignment boundaries of every word.

Dataset	Source	Split	Dates	# Spk.	# Utt.	Word inst.	Vocab	# hours
GRID [7]	-	-	-	51	33,000	165k	51	27.5
MODALITY [8]	-	-	-	35	5,880	8,085	182	31
LRW [5]	BBC	Train-val	01/2010 - 12/2015	-	514k	514k	500	165
		Test	01/2016 - 09/2016	-	25k	25k	500	8
LRS2-BBC [6]	BBC	Pre-train	01/2010 - 02/2016	-	96k	2M	41k	195
		Train-val	01/2010 - 02/2016	-	47k	337k	18k	29
		Test	03/2016 - 09/2016	-	1,243	6,663	1,693	0.5
		Text-only	01/2016 - 02/2016	-	8M	26M	60k	-
LRS3-TED	TED & TEDx (YouTube)	Pre-train	-	5,543	132k	4.2M	52k	444
		Train-val	-	4,004	32k	358k	17k	30
		Test	-	451	1,452	11k	2,136	1
		Text-only	-	5,543	1.2M	7.2M	57k	-

Table 1: *A comparison of publicly available lip reading datasets. Division of training, validation and test data; and the number of utterances, number of word instances and vocabulary size of each partition.* **Utt:** *Utterances.*

3. Conclusion

In this document, we have briefly described the LRS3-TED audio-visual corpus. The dataset is useful for many applications including lip reading, audio-visual speech recognition, video-driven speech enhancement, as well as other audio-visual learning tasks. [6] reports the performance of some of the latest lip reading models on this dataset.

4. References

[1] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip reading sentences in the wild," in *Proc. CVPR*, 2017. 1

[2] Y. M. Assael, B. Shillingford, S. Whiteson, and N. de Freitas, "Lipnet: Sentence-level lipreading," *arXiv preprint arXiv:1611.01599*, 2016. 1

[3] T. Stafylakis and G. Tzimiropoulos, "Combining residual networks with LSTMs for lipreading," in *Interspeech*, 2017. 1

[4] B. Shillingford, Y. Assael, M. W. Hoffman, T. Paine, C. Hughes, U. Prabhu, H. Liao, H. Sak, K. Rao, L. Bennett *et al.*, "Large-scale visual speech recognition," *arXiv preprint arXiv:1807.05162*, 2018. 1

[5] J. S. Chung and A. Zisserman, "Lip reading in the wild," in *Proc. ACCV*, 2016. 1, 2

[6] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Deep audio-visual speech recognition," in *arXiv*, 2018. 1, 2

[7] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006. 2

[8] A. Czyzewski, B. Kostek, P. Bratoszewski, J. Kotus, and M. Szykulski, "An audio-visual corpus for multimodal automatic speech recognition," *Journal of Intelligent Information Systems*, pp. 1–26, 2017. 2

[9] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein, "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation," *CoRR*, vol. abs/1804.03619, 2018. 1

[10] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *Proc. ECCV*. Springer, 2016, pp. 21–37. 1

[11] R. Lienhart, "Reliable transition detection in videos: A survey and practitioner's guide," *International Journal of Image and Graphics*, Aug 2001. 1

[12] J. Yuan and M. Liberman, "Speaker identification on the scotus corpus," *Journal of the Acoustical Society of America*, vol. 123, no. 5, p. 3878, 2008. 1

[13] J. S. Chung and A. Zisserman, "Out of time: automated lip sync in the wild," in *Workshop on Multi-view Lip-reading, ACCV*, 2016. 1