Audio-Visual TED Corpus: Enhancing the TED-LIUM Corpus with Facial Information, Contextual Text and Object Recognition

Guan-Lin Chao*

Electrical and Computer Engineering
Carnegie Mellon University
guanlinchao@cmu.edu

Chih Chi Hu*

Electrical and Computer Engineering
Carnegie Mellon University
chihhu@cmu.edu

Bing Liu

Electrical and Computer Engineering
Carnegie Mellon University
liubing@cmu.edu

John Paul Shen

Electrical and Computer Engineering
Carnegie Mellon University
jpshen@cmu.edu

Ian Lane

Electrical and Computer Engineering
Language Technologies Institute
Carnegie Mellon University
lane@cmu.edu

ABSTRACT

We present a variety of new visual features in extension to the TED-LIUM corpus. We re-aligned the original TED talk audio transcriptions with official TED.com videos. By utilizing state-of-the-art models for face and facial landmarks detection, optical character recognition, object detection and classification, we extract four new visual features that can be used for Large-Vocabulary Continuous Speech Recognition (LVCSR) systems, including facial images, landmarks, text, and objects in the scenes. The facial images and landmarks can be used in combination with audio for audio-visual acoustic modeling where the visual modality provides robust features in adverse acoustic environments. The contextual information, i.e. extracted text and detected objects in the scene can be used as prior knowledge to create contextual language models. Experimental results showed the efficacy of using visual features on top of acoustic features for speech recognition in overlapping speech scenarios.

*Equal contribution.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org. *UbiComp/ISWC '19 Adjunct, September 9–13, 2019, London, United Kingdom* © 2019 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-6869-8/19/09...\$15.00

<https://doi.org/10.1145/3341162.3344861>

KEYWORDS

audio-visual speech recognition; speech recognition corpus; multi-modal interaction

ACM Reference Format:

Guan-Lin Chao, Chih Chi Hu, Bing Liu, John Paul Shen, and Ian Lane. 2019. Audio-Visual TED Corpus: Enhancing the TED-LIUM Corpus with Facial Information, Contextual Text and Object Recognition. In *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and the 2019 International Symposium on Wearable Computers (UbiComp/ISWC '19 Adjunct)*, September 9–13, 2019, London, United Kingdom. ACM, New York, NY, USA, 6 pages. <https://doi.org/10.1145/3341162.3344861>

1 INTRODUCTION

Adverse acoustic environments have been a long-standing challenge for automatic speech recognition systems, where the existence of background speech signals and noise drastically deteriorates recognition accuracy. To mitigate this problem, researchers in audio-visual speech recognition proposed to use the visual information from the speech, which provides correlated and supplementary information on top of the acoustic signal and is robust against environmental acoustic noise.

The state of the art audio-visual speech recognition systems use deep learning methods to transcribe the audio-visual speech input. Training such deep neural network models relies on large-scale training data to avoid serious overfitting. As a result, several audio-visual speech recognition corpora have been introduced, including TULIPS1 [14], XM2VTSDB [13], CUAVE [17], AVICAR [11], GRID [2], IV2 [18], MOBIO [12], MODALITY [3], etc.

However, many previous corpora suffer from one or more of the following limitations [3]: 1) speech utterances are isolated words (recognition targets being individual words,

Table 1: The Audio-Visual TED corpus

Characteristics	Train	Dev	Test
Number of talks	1495	8	11
Number of videos	1444	8	11
Video duration (hour)	303.7	1.5	2.5
% of frames with faces	60.0%	62.7%	70.0%
Key frames	327883	3112	5862
Words extracted	181609	1612	2374
Objects detected	34006	226	708

instead of full sentences from continuous speech), 2) limited vocabulary (recognition target vocabulary being limited to digits, alphabet or simple commands), and 3) limited size of speaker population (previous largest speaker population being 300 distinct speakers speaking 15 French sentences from IV2 corpus [18]). Therefore, an audio-visual Large-Vocabulary Continuous Speech Recognition (LVCSR) corpus with a rich set of speakers is essential as training material for the audio-visual speech recognition recognizers to be deployed in real world applications.

In this paper, we introduce audio-visual TED corpus, an enhancement of audio-only TED-LIUM v2 corpus [22] with new features, mainly focusing on visual features extracted directly from matching TED talk videos. The enhancements include:

- **facial information:** addition of frame-by-frame extraction of facial images and facial landmarks of speakers, which can be combined with the original audio data for audio-visual speech recognition in low signal-to-noise ratio acoustic environments, using the techniques of face detection and facial landmark estimation.
- **visual contextual information:** addition of visual contextual information, by recognizing text and detecting objects that appear in the video key frames, using state-of-the-art techniques of optical character recognition, object detection and object classification. Taking into account the contextual information will be useful for language model rescoring as there is a significant probability for the text or objects that appear in the video to be correlated to the speech.
- **talk meta data:** addition of talk meta data, including talk titles and abstracts, speakers' names, occupations and short biographies. This textual meta data can provide prior information for language modeling.

Table 2: Metadata for Eric Mead's The Magic of the Placebo

Title: Eric Mead: The magic of the placebo
Abstract: Sugar pills, injections of nothing – studies show that, more often than you'd expect, placebos really work. At TEDMED, magician Eric Mead does a trick to prove that, even when you know something's not real, you can still react as powerfully as if it is. (Warning: This talk is not suitable for viewers who are disturbed by needles or blood.)
Speaker: Eric Mead
Occupation: Magician
Short biography: Eric Mead is a prolific magician, mentalist and comedian who worked his way up from doing magic on the street to appearing at exclusive events around the world.

2 OVERVIEW OF THE AUDIO-VISUAL TED CORPUS

The TED-LIUM v2 corpus. The TED-LIUM v2 corpus was an enhancement of the original TED-LIUM corpus [21]. Both are audio-based corpora dedicated for Large-Vocabulary Continuous Speech Recognition (LVCSR). TED-LIUM v2 extended v1 with additional talks and aligned transcripts, as well as monolingual text data for language modeling. In TED-LIUM v2, there are a total 1495 audio talks and transcripts covering 1242 unique speakers, a pronunciation dictionary with 159848 entries, and 6 English text corpora selected from the publicly available WMT 2013 machine translation evaluation campaign for language modeling.

The Audio-Visual TED corpus. We were able to download the videos from TED.com for 1444 out of the original 1495 talks in the TED-LIUM v2 train set, and completed feature extraction on these talks. All talks within the original dev set and test set have visual features. In Table 1 we highlight the statistics of various features of the Audio-Visual TED corpus. We also collected textual meta data from the TED.com website. For each talk, the meta data includes 1) talk title, 2) talk abstract, 3) speaker name, 4) speaker occupation, and 5) short speaker biography. An example of a talk meta data is shown in Table 2.

3 VISUAL FEATURE EXTRACTION

We focus on four features extracted from videos of the original TED talks, including facial images and landmarks of the speaker, text recognition and object detection. Visual feature are especially useful in complementing acoustic features when there is a relatively high level of background noise.

Figure 1: Face and Facial Landmark Detection from Eric Mead's TED Talk: The magic of the placebo.

While we do not distinguish between the frames of the speaker, the audience, and presentation slides, we have basic detection to determine the applicability of each feature. For example, we do not provide facial landmarks when the speaker is not present. Text recognition is mostly applicable on slides and pictures, while object detection can be applied throughout the video.

Audio Video Alignment

The videos downloaded from TED.com all have the same opening music. However, the audios in the TED-LIUM corpus come in 5 different kinds of opening music. Thus a subset of TED-LIUM audios, as well as the transcripts, are not aligned with the TED.com videos in terms of timestamp. Because it is crucial for the audio-visual features to be aligned for speech recognition, we extracted the audios from the TED.com videos to make sure our videos and audios match in timestamp, replaced the original TED-LIUM audios with the extracted audios, and modified the TED-LIUM transcripts accordingly. The subset of transcripts that needed to be modified to account for the audio-video misalignment can be determined by the opening music of the original TED-LIUM audios. Among the original 1463 TED-LIUM talks whose audio have 5 different kinds of opening music, 975 talks which correspond to a certain type of opening music are not aligned with their transcripts, while the rest of are. We found a 3.8 seconds offset to be reasonable. To automatically determine the type of opening music from TED-LIUM audios, we extracted the statistical acoustic features of the first 2 seconds of each audio and perform classification.

Facial Images

We used a pre-trained Convolutional Neural Network (CNN) for face detection and a correlation tracker for video face tracking from the dlib library [9]. The dlib's CNN face detector was trained on the face images from a collection of publicly available datasets, including ImageNet [23], AFLW

Figure 2: Object detection on a key frame from Michael Specter's TED Talk: The danger of science denial.

[10], Pascal VOC [6], the VGG face dataset [16], WIDER FACE [28], and FaceScrub [15]. The correlation tracker is an implementation of [4], which tracks the position of a moving object in a sequence of video frames.

Given an input image, the CNN face detector outputs multiple bounding boxes of possible face locations, and we chose the largest bounding box to be the speaker's face, and we applied the correlation tracker to track the speaker face bounding box on the subsequent frames. The reason of using a face tracker is that it is less time-consuming than performing face detection on every single frame, and that it takes into account the detected speaker face in the preceding frames, which is a useful prior information for moving object tracking. Whenever the confidence of the tracking target was lower than a threshold in a certain frame, face detection was performed again on that frame, and then tracking would be performed in its subsequent frames. To standardize the facial image dimensions across different talks and speakers, for each of the video frames that contain a speaker face, we extracted a square region centered at the face bounding box, resized it to 100px by 100px, and saved it as the facial image. An example of a detected face is shown in Figure 1.

Facial Landmarks

Once a face bounding box was detected from a given image, we used a pre-trained facial landmark detector of the OpenCV library [1] to find 68 facial landmarks, including 10 landmarks for eye brows, 12 for eyes, 9 for nose, 20 for mouth, 17 for jaw. The landmark detector implements the facial alignment algorithm proposed in [20], and was trained on the ibug 300-W dataset [24]. We applied the facial landmark detector on each frame where a face is detected, and stored the landmarks as a list of x-y coordinates, corresponding to the landmark positions in the original video frames. An example of detected facial landmarks is shown in Figure 1.

Generating Key Frames

The original TED videos are generated at 30 frames per second (FPS). With a total of 207 hours in the TED-LIUM data set, there are at least 22,400,000 frames. Extracting visual features is a very time consuming process, and often during frame transitions, images are blurred such that detection and recognition do not yield useful results in addition to those extracted from key frames alone. With the goal of incorporating and reusing the visual feature extraction within Kaldi, we found that extracting information from key frames alone was enough and greatly reduced the extraction process run time. Thus for facial detection and facial landmarks, we evaluated each frame, while for text recognition and object detection we only evaluated at key frames (intra frames).

Key frames were generated through the use of ffmpeg [5]. For the whole TED-LIUM data set we reduced the number of frames from 33,000,000 to under 340,000 frame, which is close to 100x reduction, and resulted in visual feature extraction at an average of a single key frame every 3 seconds. Features generated for key frames were then matched with utterances based on the start and end timestamps.

Object - Object Detection and Classification

To provide a list of objects in key frames, we utilized the state-of-the-art object detection model YOLO9000 [19] to analyze the objects in each key frame. TED talks cover broad and powerful ideas that are worth spreading, thus we see a large variety of objects in the talks. An example of such an image is shown in Figure 2. With joint training on object detection and classification of multiple data sets, YOLO9000 can predict object classes without labeled ground truth data, and is capable of detecting over 9000 categories in real-time.

We leveraged pre-trained weights for YOLO9000 released by the original authors with a class threshold of 0.5 and hierarchical threshold of 0.5. This yielded descriptive and on point object results that vary from cats and dogs to rain forests and cyclists.

Text - Optical Character Recognition (OCR)

For text extraction, commonly known as optical character recognition (OCR), we utilized ImageMagick [27] and Tesseract Open Source Engine version 4.0 [26] for all key frames. Within ImageMagick, we leveraged its textcleaner and textdeskew to pre-process the key frames before feeding it into the Tesseract Open Source Engine version 4.0 leveraging long short-term memory (LSTM) units [8]. Textcleaner first converts images to gray scale, then enhances the brightness in order to perform local adaptive thresholding for a clear background, and finally sharpens pixels and blurs for clarity. Textdeskew first finds the Fast Fourier Transform (FFT) spectrum of an image in order to locate its major peaks,

Table 3: Word Error Rate comparisons of single-speaker scenarios

feature	WER (%)
audio	14.7
audio + visual context	14.3
audio + visual context + face embedding	13.8

Table 4: Word Error Rate comparisons of two-speaker scenarios

feature	WER (%)
audio	64.1
audio + visual context	62.2
audio + visual context + face embedding	55.9

and then calculates its rotation angle to result in up-right printed text. To achieve robust results, we fed pre-processed results from textcleaner, textdeskew, and a combination of the textcleaner and textdeskew, into the Tesseract OCR engine. Results from the Tesseract OCR Engine above 80% confidence threshold were retained.

Figure 3 shows an example of text processing from Tom Wujec's talk "Build a Tower, Build a Team". Figure 3a and 3d are original key frames. Figure 3b highlights the default Tesseract binarization image where fine text is lost. In Figure 3e, the slight rotation of text is corrected using textdeskew. Figure 3c and 3f show the pre-processed image after textcleaner which retains and sharpens all the text in the frame.

4 EXPERIMENTS

We evaluated the Audio-Visual TED corpus by comparing speech recognition systems with three combinations of input modalities: audio-only, audio + visual context, and audio + visual context + face embedding, under two acoustic scenarios: single-speaker and overlapping speech (two-speaker). In the single-speaker scenario, the original utterances were used as audio input. In the two-speaker scenario, we simulated overlapping speech of two speakers in a cocktail party by mixing each of the original utterance with a background utterance randomly selected from another speaker with equal energy on the single acoustic channel.

Audio-visual Speech Recognition System

Across all the experiments of different combinations of input modalities and acoustic scenarios, we used the same architectures for acoustic and language models. The Kaldi recipe for training the aforementioned models will be released with the corpus.

(a) The original frame showcasing a variety of text in different sizes, colors, and fonts.

(b) The image after applying Tesseract's default binarization with text lost in the process.

(c) The image after applying textcleaner with small text still preserved and sharpened correctly detects all text including the small bottom row of text.

(d) The original frame showcasing text from a distance with a rotation angle.

(e) The image after applying textdeskew to correct the slight rotation of the text given the perspective angle of this image.

(f) The image after applying textcleaner with clear text correctly extracts the words "Marshmallow Challenge".

Figure 3: Text Extraction: Pre-processing steps prior to applying the Tesseract OCR Engine on key frames from Tom Wujec's TED Talk: Build a Tower, Build a Team. Figure 3a shows an image of a slide while Figure 3d shows a mixture of speaker, background, and slides.

Features

The audio features are 40-bin MFCC features with 10 ms window shift and 50% overlap extracted from original utterances (single-speaker) or mixed utterances (two-speaker), followed by unit mean and variance normalization per speaker. The visual context features are grayscale pixel values of face images resized to 60px by 60px. A face embedding is a static vector per speaker, which was extracted by feeding a speaker's face images to a FaceNet face recognition model [25] and averaging the output embeddings. We used a pre-trained FaceNet model which was trained on the MS-Celeb-1M dataset [7]. For frames with no speaker face present, we used all-zero values as the face images and face embeddings.

Results

Table 3 shows the Word Error Rate (WER) in the single-speaker scenarios. Table 4 shows the WER in the two-speaker scenarios. We observed that in the single-speaker scenarios, the use of visual features on top of acoustic features did not make a significant difference in WER. On the other hand, in simulated overlapping speech scenarios, the introduction of

visual context features on top of acoustic features reduced the WER from 64.1% to 62.2%, and the face embedding features further improved the WER to 55.9%. We suspect this finding to be caused by the resolution of visual information. The resolution of the face images may not be high enough for lipreading. In other words, it is challenging for the speech recognizer to recognize the visemes from the mouth region in the face images. On the other hand, the face images and face embeddings may still be informative enough for identifying a speaker's attributes, such as gender or age, which are useful cues to discern a speaker's voice from overlapping speech.

5 CONCLUSION

In this work, we present the Audio-Visual TED corpus, an enhancement of the audio-only TED-LIUM v2 corpus which was released in 2014. By re-aligning the original TED talk audio transcriptions with official TED.com videos and utilizing state of the art models in computer vision and deep learning, we extract four new visual features that can be used to enhance the performance of Large-Vocabulary Continuous Speech Recognition systems. The facial images and

landmarks can be used for audio-visual speech recognition in adverse acoustic environments where the signal-to-noise ratio is low. Visual contextual information from text extraction, object detection, as well as textual talk meta data can be used to create contextual language models. Our experiments showed that in simulated overlapping speech scenarios where the speech of target and background speakers are mixed on the same acoustic channel with equal energy, using face image features on top of acoustic features reduced the WER from 64.1% to 62.2%, and the face embedding features further improved the WER to 55.9%, while it did not introduce a significant difference to the 13.7% WER for single-speaker scenarios.

REFERENCES

- [1] Gary Bradski. 2000. The OpenCV Library. *Dr. Dobb's Journal of Software Tools* (2000).
- [2] Martin Cooke, Jon Barker, Stuart Cunningham, and Xu Shao. 2006. An audio-visual corpus for speech perception and automatic speech recognition. *Journal of the Acoustical Society of America* (2006).
- [3] Andrzej Czyzewski, Bożena Kostek, Piotr Bratoszewski, Jozef Kotus, and Marcin Szykalski. 2017. An audio-visual corpus for multimodal automatic speech recognition. *Journal of Intelligent Information Systems* (2017).
- [4] Martin Danelljan, Gustav Häger, Fahad Khan, and Michael Felsberg. 2014. Accurate scale estimation for robust visual tracking. In *British Machine Vision Conference (BMVC)*.
- [5] The FFmpeg developers. 2019. ffmpeg version 2.8.11-0. <http://ffmpeg.org/>. Accessed: 2019-06-14.
- [6] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. 2015. The Pascal Visual Object Classes Challenge: A Retrospective. *International Journal of Computer Vision (IJCV)* (2015).
- [7] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. 2016. MS-Celeb-1M: A Dataset and Benchmark for Large-Scale Face Recognition. In *European Conference on Computer Vision (ECCV)*.
- [8] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation* (1997).
- [9] Davis E. King. 2009. Dlib-ml: A Machine Learning Toolkit. *Journal of Machine Learning Research* (2009).
- [10] Martin Koestinger, Paul Wohlhart, Peter M Roth, and Horst Bischof. 2011. Annotated Facial Landmarks in the Wild: A Large-scale, Real-world Database for Facial Landmark Localization. In *ICCV Workshop on Benchmarking Facial Image Analysis Technologies*.
- [11] Bowon Lee, Mark Hasegawa-Johnson, Camille Goudeseune, Suketu Kamdar, Sarah Borys, Ming Liu, and Thomas Huang. 2004. AVICAR: Audio-visual speech corpus in a car environment. In *Eighth International Conference on Spoken Language Processing*.
- [12] Christopher McCool, Sebastien Marcel, Abdenour Hadid, Matti Pietikäinen, Pavel Matejka, Jan Cernocký, Norman Poh, Josef Kittler, Anthony Larcher, Christophe Levy, Driss Matrouf, Jean-François Bonastre, Phil Tresadern, and Timothy Cootes. 2012. Bi-modal person recognition on a mobile phone: using mobile phone data. In *International Conference on Multimedia and Expo Workshops*.
- [13] Kieron Messer, Jiri Matas, Josef Kittler, Juergen Luetttin, and Gilbert Maitre. 1999. XM2VTSDB: The extended M2VTS database. In *International Conference on Audio and Video-based Biometric Person Authentication*.
- [14] Javier R Movellan. 1995. Visual speech recognition with stochastic networks. In *Advances in Neural Information Processing Systems*.
- [15] Hong-Wei Ng and Stefan Winkler. 2014. A Data-Driven Approach to Cleaning Large Face Datasets. In *International Conference on Image Processing (ICIP)*.
- [16] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, et al. 2015. Deep Face Recognition. In *British Machine Vision Conference (BMVC)*.
- [17] Eric K Patterson, Sabri Gurbuz, Zekeriya Tufekci, and John N Gowdy. 2002. CUAVE: A new audio-visual database for multimodal human-computer interface research. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.
- [18] Dijana Petrovska-Delacrétaz, Sylvie Lelandais, Joseph Colineau, Liming Chen, Bernadette Dorizzi, M Ardabilian, E Krichen, M-A Mellakh, A Chaari, S Guerfi, et al. 2008. The IV 2 Multimodal Biometric Database (Including Iris, 2D, 3D, Stereoscopic, and Talking Face Data), and the IV 2-2007 Evaluation Campaign. In *International Conference on Biometrics: Theory, Applications and Systems*.
- [19] Joseph Redmon and Ali Farhadi. 2017. YOLO9000: Better, Faster, Stronger. In *Computer Vision and Pattern Recognition (CVPR)*.
- [20] Shaoqing Ren, Xudong Cao, Yichen Wei, and Jian Sun. 2014. Face Alignment at 3000 fps via Regressing Local Binary Features. In *Computer Vision and Pattern Recognition (CVPR)*.
- [21] Anthony Rousseau, Paul Deléglise, and Yannick Estève. 2012. TED-LIUM: an Automatic Speech Recognition dedicated corpus. In *International Conference on Language Resources and Evaluation (LREC)*.
- [22] A Rousseau, P Deléglise, and Y Estève. 2014. Enhancing the TED-LIUM Corpus with Selected Data for Language Modeling and More TED Talks. In *International Conference on Language Resources and Evaluation (LREC)*.
- [23] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)* (2015).
- [24] Christos Sagonas, Epameinondas Antonakos, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 2016. 300 Faces In-the-Wild Challenge: Database and Results. *Image and Vision Computing* (2016).
- [25] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. FaceNet: A Unified Embedding for Face Recognition and Clustering. In *Computer Vision and Pattern Recognition (CVPR)*.
- [26] Ray Smith. 2007. An Overview of the Tesseract OCR Engine. In *International Conference on Document Analysis and Recognition (ICDAR)*.
- [27] Fred Weinhaus. 2019. ImageMagick. <http://www.fmwconcepts.com/imagemagick/>. Accessed: 2019-06-14.
- [28] Shuo Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. 2016. WIDER FACE: A Face Detection Benchmark. In *Computer Vision and Pattern Recognition (CVPR)*.