Data Article

Correlated lip motion and voice audio data

Marco Colasito\textsuperscript{a}, Jeremy Straub\textsuperscript{b,\ast}, Pratap Kotala\textsuperscript{b}

\textsuperscript{a} Department of Computer Science, Valdosta State University, United States
\textsuperscript{b} Department of Computer Science, North Dakota State University, United States

article info

Article history:
Received 4 September 2018
Received in revised form 9 October 2018
Accepted 15 October 2018
Available online 18 October 2018

abstract

This data set is comprised of correlated audio and lip movement data in multiple videos of multiple subjects reading the same text. It was collected to facilitate the development and validation of algorithms used to train and test a compound biometric system that consists of lip-motion and voice recognition. The data set is a collection of videos of volunteers reciting a fixed script that is intended to be used to train software to recognize voice and lip-motion patterns. A second video is included of the individual reciting a shorter phrase, which is designed to be used to test the recognition functionality of the system. The recordings were collected in a controlled, indoor setting with a 4K professional-grade camcorder and adjustable, LED lights.

\& 2018 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Specifications table

| Subject area          | Computer Science |
|-----------------------|------------------|
| More specific subject area | Video processing; speech and facial recognition |
| Type of data          | Digital video files |
| How data was acquired | Video recordings using a Sony 4K Handycam AX100 |
| Data format           | Raw |
| Experimental factors  | N/A |

\ast Corresponding author.

E-mail address: jeremy.straub@ndsu.edu (J. Straub).

https://doi.org/10.1016/j.dib.2018.10.043
2352-3409\copyright 2018 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).
Experimental features

Videos of the subjects’ head and upper shoulders while speaking. The dataset includes two videos per subject with all subjects speaking the same script. One is a longer script consisting of multiple sentences and the second is a single sentence.

Data source location

Fargo, North Dakota, USA

Data accessibility

Mendeley Data

Related research article

Value of the data

- Data was collected in a controlled environment facilitating analysis and minimizing confounding factors.
- Includes multiple individuals with similar and dissimilar characteristics facilitating system testing and analysis of complex algorithms.
- Two videos are included, per subject, for use as training and validation sets.
- Can be used for training and testing of speech recognition and lip-motion recognition.

1. Data

Human speech recognition is used for numerous applications including transcription [1] and commanding computer workstations [2]. In support of these applications, techniques have been developed to overcome key technical challenges such as assessing recognition system performance confidence [3] and recognizing speech, despite background noise [4]. However, advances in speech synthesis have resulted in an inability to tell human speech from automated speech, in some applications [5]. Identifying that a human is speaking [6] can be a form of automated attack prevention. Identifying a particular human who is speaking could be used for authentication or as a factor in a multi-factor authentication system.

An alternate approach to recognizing humans and speech is by analyzing the human face. Lip reading has been performed by both software systems [7] and humans [8], with automated systems outperforming humans in some instances [8]. Facial recognition can also be used to identify humans [9], directly; however, other techniques [6] may be needed to verify that the is live, and not pre-recorded, input designed to fool a system. Facial imagery analysis has been used to aid speech recognition work [10] for synchronized audio and video feeds and, in particular, facilitate speech recognition in high-noise environments [11]. The analysis of facial imagery has also been used to assess other characteristics such as emotion and stress levels [12]. Problematically, creating synchronized audio and video lip movements has also been a study of work for some time [13] and efforts in this area may undermine the ability to validate subject liveness via combined audio and video analysis.

Sengupta et al. [14] proposed a technique for lip reading as part of authentication and others [15–17] have performed related speaker recognition and authentication work. Work is also ongoing in the audio speech recognition and lip reading domains, independent from security applications.

In support of the development and analysis of the aforementioned and other speech recognition algorithms, a variety of data sets have been collected. These have included data sets collected from pre-existing video footage [18] as well as specialty data set collected in application area environments, such as automobiles [19].

This dataset provides videos of 20 male subjects, with ages ranging from 18 to 26, which can be used for speech recognition, lip reading and combined speech and lip reading applications. Two videos were collected per subject. In the first, the subject read the FitnessGram Pacer Test script and in the second they said an arbitrary sentence. The data was collected in a controlled and well-lit environment.
environment with minimal environmental noise. This data was originally collected to train and test a compound biometric system that consists of lip-motion and voice authentication. The training videos averaged between 30 and 45 s and the testing videos averaged between 5 and 8 s each.

2. Experimental design, materials, and methods

This section discusses the procedures and equipment that was used to collect the video data. First, the equipment and configuration of the collection environment are discussed. Then, the data collection procedures are discussed.

2.1. Equipment and configuration

A 4K Sony Handycam AX100 was used to record the videos. A single Yongnuo YN600L light, set at 40% brightness, was used to illuminate the subject’s face. A standard projector screen was used as the background for all videos. Two Neewer LED500LRC lights were used as background lighting to eliminate any shadows that may be cast onto the projector screen by the lighting used to illuminate the subject’s face.

The Handycam AX100 was placed on a tripod that was positioned approximately four feet and three inches from the subject, directly in front of the subject with the subject facing the camcorder and the projector screen behind the subject. The light for facial illumination was placed to the right of the camera. The locations of the subject and pieces of equipment are depicted in Fig. 1. Fig. 2 shows a subject being recorded. Both the camera and facial illumination light were positioned at the height of the middle of the subject’s face. The subject was located approximately five feet and three inches in front of the projector screen.

The Neewer LED500LRC background lights were positioned four feet and six inches to the left and three feet to the right of the subject, from the perspective of the camcorder. The projector screen was six feet tall and was elevated two feet and eight inches from the ground.

![Diagram of subject, camera and lighting position.](image-url)
2.2. Data collection procedures

Twenty male subjects ranging in age between 18 and 26 were selected to participate. Each subject was given a copy of the script and instructed to stand on the center marker on the carpet, facing the camcorder. They were instructed to watch the camcorder and when they see that the red recording light is lit, to begin reciting the training script while keeping their head level and mouth visible to the camera.

After finishing the training script recording, a second video was recorded of the subject reading the testing script, with the subject again instructed to start reading when they see that the red recording light is lit. Each video was manually reviewed to ensure that there was not disruptive
background or extraneous noise or other interference. Three frames from the recording are shown in Fig. 3, which also depicts how the subject’s face and lips can be identified from the video.

Acknowledgments

The collection of this data was supported by the United States National Science Foundation (NSF award # 1757659). Thanks is given to Collin Gros and William Clemons for their help in collecting data. Facilities and some equipment used for the collection of this data were provided by the North Dakota State University Institute for Cyber Security Education and Research and the North Dakota State University Department of Computer Science.

Transparency document. Supplementary material

Transparency data associated with this article can be found in the online version at https://doi.org/10.1016/j.dib.2018.10.043.

References

[1] F. Seide, G. Li, D. Yu, Conversational speech transcription using context-dependent deep neural networks, Interspeech (2011) 437–440.
[2] Continuous speech recognition of text and commands, 1998.
[3] H. Jiang, Confidence measures for speech recognition: a survey, Speech Commun. 45 (4) (2005) 455–470.
[4] Y. Gong, Speech recognition in noisy environments: a survey, Speech Commun. 16 (3) (1995) 261–291.
[5] P.L. De Leon, M. Pucher, J. Yamagishi, I. Hernaez, I. Saratxaga, Evaluation of speaker verification security and detection of HMM-based synthetic speech, IEEE Trans. Audio Speech Lang. Process. 20 (8) (2012) 2280–2290.
[6] G. Chetty, M. Wagner, Automated lip feature extraction for liveness verification in audio-video authentication, Proc. Image Vis. Comput. (2004).
[7] R. Bowden, et al., Recent developments in automated lip-reading, vol. 8901, 2013, p. 89010J.
[8] S. Hilder, R. Harvey, B.-J. Theobald, Comparison of human and machine-based lip-reading, in: Proceedings of the International Conference on Audio-Visual Speech Processing, 2009.
[9] N. Kar, M. Kant, D. Barma, A. Saha, Study of implementing automated attendance system using face recognition technique, Int. J. Comput. Commun. Eng. 1 (2) (2012) 100–103.
[10] A. Torfi, S.M. Iranmanesh, N. Nasrabi, J. Dawson, 3D convolutional neural networks for cross audio-visual matching recognition, IEEE Access 5 (2017) 22081–22091.
[11] M. Guoying Zhao, Barnard, M. Pietikainen, Lipreading with local spatiotemporal descriptors, IEEE Trans. Multimed. 11 (7) (2009) 1254–1265.
[12] Voice, lip-reading, face and emotion stress analysis, fuzzy logic intelligent camera system, 2003.
[13] J.P. Lewis, F.I. Parke, J.P. Lewis, F.I. Parke, Automated lip-synch and speech synthesis for character animation, in: Proceedings of the SIGCHI/GI Conference on Human Factors in Computing Systems and Graphics Interface – CHI’87, vol. 17(SI), 1987, pp. 143–147.
[14] S. Sengupta, A. Bhattacharya, P. Desai, A. Gupta, Automated lip reading technique for password authentication, Int. J. Appl. Inf. Syst. 4 (3) (2012) 18–24.
[15] J. Luetten, N.A. Thacker, S.W. Beet, Speaker identification by lipreading, in: Proceedings of the Fourth International Conference on Spoken Language Processing, ICSP’96, vol. 1, pp. 62–65.
[16] M.-I. Faraj, J. Bigun, Audio-visual person authentication using lip-motion from orientation maps, Pattern Recognit. Lett. 28 (11) (2007) 1368–1382.
[17] S.-H. Leung, S.-L. Wang, W.-H. Lau, Lip image segmentation using fuzzy clustering incorporating an elliptic shape function, IEEE Trans. Image Process. 13 (1) (2004) 51–62.
[18] J.S. Chung, A. Zisserman, Lip reading in the wild, in: Proceedings of the Asian Conference on Computer Vision, 2016, pp. 87–103.
[19] B. Lee, et al., AVICAR: audio-visual speech corpus in a car environment, in: Proceedings of the 8th International Conference on Spoken Language Processing, 2004.