VidTIMIT Dataset Documentation

Conrad Sanderson
conrad.sanderson at nicta dot com dot au

Version: 30 Nov 2009

License

The VidTIMIT dataset is Copyright © 2001-2009 Conrad Sanderson. Distribution and research usage of this dataset is permitted under the following conditions:

1. This notice is left intact and not modified in any way.

2. The dataset is provided as is. There is no warranty as to the fitness for any particular purpose.

3. The author of the dataset is not responsible for any direct or indirect losses resulting from the use of the dataset.

4. Any publication (e.g. conference paper, journal article, technical report, book chapter, book) resulting from the usage of the VidTIMIT dataset must cite the following paper:

   - C. Sanderson and B.C. Lovell.
     Multi-Region Probabilistic Histograms for Robust and Scalable Identity Inference.
     Lecture Notes in Compuer Science (LNCS), Vol. 5558, pp. 199-208, 2009.

Overview

The VidTIMIT dataset is comprised of video and corresponding audio recordings of 43 volunteers (19 female and 24 male), reciting short sentences. It can be useful for research on topics such as automatic lip reading, multi-view face recognition, multi-modal speech recognition and person identification/verification.

The dataset was recorded in 3 sessions, with a mean delay of 7 days between Session 1 and 2, and 6 days between Session 2 and 3. The delay between sessions allows for changes in the voice, hair style, make-up, clothing and mood (which can affect the pronunciation), thus incorporating attributes which would be present during the deployment of a verification system. Additionally, the zoom factor of the camera was randomly perturbed after each recording.

The sentences were chosen from the test section of the NTIMIT corpus[1]. There are ten sentences per person. The first six sentences (sorted alpha-numerically by filename) are assigned to Session 1. The next two sentences are assigned to Session 2 with the remaining two to Session 3. The first two sentences for all persons are the same, with the remaining eight generally different for each person. The mean duration of each sentence is 4.25 seconds, or approximately 106 video frames (using 25 fps).

A typical example[2] of the sentences used is in Table 1. There is complete correspondence of the subject IDs between VidTIMIT and NTIMIT (and hence the recited sentences).

In addition to the sentences, each person performed an extended *head rotation* sequence in each session, which allows for extraction of profile and 3D information. The sequence consists of the person moving their head to the left, right, back to the center, up, then down and finally return to center. The recording was done in a noisy office environment (mostly computer fan noise) using a broadcast quality digital video camera. For almost all the recordings the lighting setup[3] was as follows:

---

[1] C. Jankowski, A. Kalyanswamy, S. Basson and J. Spitz, "NTIMIT: A Phonetically Balanced, Continuous Speech Telephone Bandwidth Speech Database", Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP), Albuquerque, 1990, Vol. 1, pp. 109-112. NTIMIT can be obtained from the Linguistic Data Consortium (www.ldc.upenn.edu).

[2] Copyright restrictions on the NTIMIT corpus prevent the list of all sentences used in VidTIMIT.

[3] An exception to the lighting setup occurs for Session 3 of *mbdg0* and Session 2 of *mjar0*. For these cases the incandescent lamp was switched off.

Section ID	Sentence ID	Sentence text
Session 1	sa1	She had your dark suit in greasy wash water all year
	sa2	Don't ask me to carry an oily rag like that
	si1398	Do they make class-biased decisions?
	si2028	He took his mask from his forehead and threw it, unexpectedly, across the deck
	si768	Make lid for sugar bowl the same as jar lids, omitting design disk
	sx138	The clumsy customer spilled some expensive perfume
Session 2	sx228	The viewpoint overlooked the ocean
	sx318	Please dig my potatoes up before frost
Session 3	sx408	I'd ride the subway, but I haven't enough change
	sx48	Grandmother outgrew her upbringing in petticoats

Table 1: Typical example of sentences used in the VidTIMIT database

1. Standard overhead fluorescent tubes, like in most office environments. The lights were covered with A4 size white office paper in order to diffuse the light – this reduced the glare on the face and top of the head.

2. An incandescent lamp in front of the person (just below the camera). The lamp was covered with a sheet of A4 size white office paper.

The video of each person is stored as a numbered sequence of JPEG images with a resolution of 384×512 pixels (rows $\times$ columns). A quality setting of 90% was used during the creation of the JPEG images. The corresponding audio[4] is stored as a mono, 16 bit, 32 kHz WAV file.

The VidTIMIT dataset is comprised of 44 files (including this documentation) taking up about 3 Gb. Each zip archive is for one person (e.g. *felc0.tar*) and has the following internal structure:

subjectID / audio / sentenceID.wav
subjectID / video / sentenceID / ###

where *sentenceID* is the head rotation or sentence identifier (e.g. *sx396*) and *###* is a three digit[5] frame number (e.g. 037). Each frame is stored as a JPEG image (note that there is no *.jpg* extension). There is no audio for the head rotation sequences.

Acknowledgements

---

[4]The audio was recorded using the camera's microphone.

[5]The three digit frame number limitation does not apply to the *head2* and *head3* sequences of *mrgg0*.