DAIC-WOZ Depression Database

This database is part of a larger corpus, the Distress Analysis Interview Corpus (DAIC) (Gratch et al.,2014), that contains clinical interviews designed to support the diagnosis of psychological distress conditions such as anxiety, depression, and post-traumatic stress disorder. These interviews were collected as part of a larger effort to create a computer agent that interviews people and identifies verbal and nonverbal indicators of mental illness (DeVault et al., 2014). Data collected include audio and video recordings and extensive questionnaire responses; this part of the corpus includes the Wizard-of-Oz interviews, conducted by an animated virtual interviewer called Ellie, controlled by a human interviewer in another room. Data has been transcribed and annotated for a variety of verbal and non-verbal features.

Data description

The package includes 189 folders of sessions 300-492. Certain sessions have been excluded for technical reasons (see below). Data are grouped by session.

Pack\

- 300_P
- 301_P
- …
- 492_P
- util
- documents
- train_split.csv
- dev_split.csv
- test_split.csv

Excluded sessions: 342,394,398,460
Included sessions with special notes:

- 373 – there is an interruption around 5:52-7:00, the confederate enters the room to fix a minor technical issue, the session continuous and completes .
- 444 – there is an interruption around 4:46-6:27, the participant's phone rings and the confederate enters the room to help them turn it off. Session continuous and completes.
- 451,458,480 – sessions are technically complete, but missing Ellie (the virtual human) part of the transcripts. Participant transcripts are still included, but without the interviewer questions.

- 402 – video recording is cut ~2min before the end of the conversation.

train_split_Depression_AVEC2017.csv: This file comprises participant IDs, PHQ8 (Kroenke et al., 20009) Binary labels (PHQ8 Scores >= 10), PHQ8 Scores, and participant gender, and single responses for every question of the PHQ8 questionnaire for the official train split. PHQ8 refers to the patient health questionnaire. Details are provided in the documentation folder file: scherer_etal2015_VowelSpace.pdf.

dev_split_Depression_AVEC2017.csv: This file comprises participant IDs, PHQ8 Binary labels, PHQ8 Scores, and participant gender, and single responses for every question of the PHQ8 questionnaire for the official development split.

test_split_Depression_AVEC2017.csv: This file comprises participant IDs and participant gender for the official test split.

Every session folder includes the following files (where XXX is session number, for example XXX=301 in folder 301_P).

```
XXX_P\

        XXX_CLNF_features.txt
        XXX_ CLNF _features3D.txt
        XXX_ CLNF _gaze.txt
        XXX_ CLNF _hog.bin
        XXX_ CLNF _pose.txt
        XXX_ CLNF _AUs.csv
        XXX_AUDIO.wav
        XXX_COVAREP.csv
        XXX_FORMANT.csv
        XXX_TRANSCRIPT.csv
```

Utility files shared:

```
        util\
                runHOGread_example.m
                Read_HOG_files.m
```

File description and feature documentation

This section documents the specific files that are shared for each session. Files that come from the same software are grouped by the software.

1. CLNF framework output

T. Baltrušaitis, P. Robinson, L-P. Morency. OpenFace: an open source facial behavior analysis toolkit in IEEE Winter Conference on Applications of Computer Vision (WACV), 2016 (http://ieeexplore.ieee.org/abstract/document/7477553/)

Link: https://github.com/TadasBaltrusaitis/OpenFace

Files:

- XXX.CLNF_features.txt:
  68 2D points on the face. The file format is as follows
  "frame, timestamp(seconds), confidence, detection_success, x0, x1,…, x67, y0, y1,…, y67". Points are in pixel coordinates.

- XXX_CLNF_AUs.csv :
  "*frame, timestamp, confidence, success, AU01_r, AU02_r, AU04_r, AU05_r, AU06_r, AU09_r, AU10_r, AU12_r, AU14_r, AU15_r, AU17_r, AU20_r, AU25_r, AU26_r, AU04_c, AU12_c, AU15_c, AU23_c, AU28_c, AU45_c*". The values indicated with "_r" are regression outputs for each action unit and "_c" are binary labels reflecting 1 action unit is present or 0 not present. Action units:
  https://en.wikipedia.org/wiki/Facial_Action_Coding_System

- XXX.CLNF_features3D.txt:
  68 3D points on the face. The file format is as follows
  "frame, timestamp(seconds), confidence, detection_success, X0, X1,…, X67, Y0, Y1,…, Y67, Z0, Z1,…, Z67". The points are in millimeters in world coordinate space, with camera being at (0,0,0) and the axes aligned to the camera

- XXX.CLNF_gaze.txt:
  "*frame, timestamp(seconds), confidence, detection_success, x_0, y_0, z_0, x_1, y_1, z_1, x_h0, y_h0, z_h0, x_h1, y_h1, z_h1*"
  The gaze is output as 4 vectors, first two vectors are in world coordinate space describing the gaze direction of both eyes, the second two vectors describe the gaze in

head coordinate space (so if the eyes are rolled up, the vectors will indicate up even if the head is turned or tilted)

- XXX.CLNF_hog.bin:
  HOG in a binary file format using the Felzenswalb's HoG on the aligned 112x112 area of the face. This results in a 4464 vector per frame. The way it is stored is a byte stream with every frame being: "*num_cols, num_rows, num_channels, valid_frame, 4464d vector*". In the util folder there is a function "Read_HOG_files.m" from the CLM framework to read HOG binary format into a matlab matrix.

- XXX.CLNF_pose.txt:
  "*frame_number, timestamp(seconds), confidence, detection_success, X, Y, Z, Rx, Ry, Rz*"
  Pose is an output of 6 numbers, X,Y,Z are the position coordinates and Rx,Ry,Rz the head rotation coordinates. Position is in world coordinates in millimeters and rotation is in radians and in Euler angle convention (to get to a proper rotation matrix the following is used R = Rx * Ry * Rz).

All .txt files include the appropriate headers. Each line represents results for a frame. "Confidence" is a measure in [0,1] representing the confidence of the tracking.

2. Audio file

Audio file: XXX_AUDIO.wav (scrubbed)

Audio recordings of head mounted microphone (Sennheiser HSP 4-EW-3) at 16kHz. Audio file might contain small amounts of bleed-over of virtual interviewer; use transcript files to alleviate this issue when processing. Identifiable utterances are scrubbed from the audio recordings, i.e. the waveform is zeroed out during the respective times; use transcript files and keyword "scrubbed_entry" to spot these instances. Scrubbed entries are also zeroed out in the feature files.

3. Transcript file

XXX_TRANSCRIPT.csv (scrubbed)

Transcription conventions

- Upper case words in the corpus but it is very rare. If they are present the upper case has no significance, except for it being a location name for example.
- Incomplete words should are annotated as follows:

If speech is cut-off, put down the complete intended word, followed by a comment with the part that was actually pronounced in angle brackets: people <peop>. The comment is just for human readers; the reason for transcribing a whole word is to avoid confusing the processing modules by training them on non-words.

Unrecognizable words are indicated as 'xxx'

Speech overlap is indicated by overlapping time stamps

- Transcript files are "tab separated" files.
- Transcriptions of virtual interviewer above Participant ID 363 were generated automatically and contain unique identifiers of utterances before the content of the utterance is provided in brackets.
  For example: `165.854  166.324  Ellie  yeah3 (yeah)`

Detailed transcription manual is provided in documents directory.

4. Audio features

Audio features are extracted using the COVAREP toolbox (v. 1.3.2) available at: https://github.com/covarep/covarep

Files:

- XXX_COVAREP.csv (scrubbed): The following features are extracted:
  - All audio features (including Formants; see below) are every 10ms. Hence, the audio features are sampled at 100Hz.
  - F0, VUV, NAQ, QOQ, H1H2, PSP, MDQ, peakSlope, Rd, Rd_conf, MCEP_0-24, HMPDM_0-24, HMPDD_0-12
  - Descriptions for each of these features can be found on the COVAREP website and in the provided COVAREP publication. In addition, detailed information about the exact procedures of feature extraction are provided in the respective publications cited in the COVAREP scripts provided via github.
  - One important aspect is that VUV (voiced/unvoiced) provides a flag ({0,1}) if the current segment is voiced or unvoiced. ***In unvoiced case, i.e. VUV = 0, the vocal folds are detected to not be vibrating, hence values such as F0, NAQ, QOQ, H1H2, PSP, MDQ, peakSlope, and Rd should not be utilized.***
  - Scrubbed entries are set to zeros.
- XXX_FORMANT.csv (scrubbed): Contains the first 5 formants, i.e. the vocal tract resonance frequencies, that are tracked throughout the interview.
  - Scrubbed entries are set to zeros.

References

Jonathan Gratch, Ron Artstein, Gale Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, David Traum, Skip Rizzo, Louis-Philippe Morency, "*The Distress Analysis Interview Corpus of human and computer interviews*", Proceedings of Language Resources and Evaluation Conference (LREC), 2014

DeVault, D., Artstein, R., Benn, G., Dey, T., Fast, E., Gainer, A., Georgila, K., Gratch, J., Hartholt, A., Lhommet, M., Lucas, G., Marsella, S., Morbini, F., Nazarian, A., Scherer, S., Stratou, G., Suri, A., Traum, D., Wood, R., Xu, Y., Rizzo, A., and Morency, L.-P. (2014). "*SimSensei kiosk: A virtual human interviewer for healthcare decision support*". In Proceedings of the 13th International Conference on Autonomous Agents and Multiagent Systems (AAMAS'14), Paris

Degottex, G.; Kane, J.; Drugman, T.; Raitio, T.; and Scherer, S., *COVAREP - A collaborative voice analysis repository for speech technologies*. In Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2014), pages 960-964, 2014.

Kroenke K, Strine TW, Spitzer RL, Williams JB, Berry JT, Mokdad AH. The PHQ-8 as a measure of current depression in the general population. Journal of affective disorders. 2009 Apr 30;114(1):163-73.