1. Appendix

Appendix A.1: Definition of Digital Linguistic Cues of Depression

Feature #	**A1: Use of first-person singular pronouns**
Description/ Explanation	The ratio (R_{PRP}) between the number of first-person singular ($count_{PRP}$) and the number of total words ($total\_words$). $R_{PRP} = \frac{count_{PRP}}{total\_words}$ (1) First person singular pronouns examples are: I, me, my, mine, myself (Garside, Leech, and Sampson 1987, p. 167).
Feature #	**A2: Use of negatively valanced words**
Description/ Explanation	The ratio (R_{NVW}) between negatively valanced words ($count_{NVW}$) and the number of total words ($total\_words$) (Yao, 2019). $R_{NVW} = \frac{count_{NVW}}{total\_words}$ (2) Sentiment analysis is used to infer negatively valanced words, such as, “Hurt”, “Tears”, “Hate”, “Sleep”, “Worry”, “Pain”, “Alone”, “Sad”, “Sadness”, “Angry”, “Anger”, “Depressed” and “Depression”.
Feature #	**A3: Explicit mentions of treatment of depression**
Description/ Explanation	The ratio (R_{ETD}) between of use of words specifically related to treatment of depression ($count_{ETD}$) and the number of total words ($total\_words$). $R_{ETD} = \frac{count_{ETD}}{total\_words}$ (3) Words specifically related to depression: "side effects", "therapy", "treatment", "pills", "psychotherapy", "medication", etc (Al-Mosaiwi M and Johnstone, 2018).
Feature #	**A4: Use of absolutist words**
Description/ Explanation	The ratio (R_{AW}) between of use of absolutist words ($count_{AW}$) and and the number of total words ($total\_words$). $R_{AW} = \frac{count_{AW}}{total\_words}$ (4) For the definition of an absolutist word, we adopt the definition by *Al-Mosaiwi and Johnstone, 2018.* Examples of such words are: “Always”, “Never”, “All the time”, “Everything”, “Nothing”.
Feature #	**A5: Focusing on the past**
Description/ Explanation	The ratio (R_{PW}) between of use of verbs in past tense (VBD) or verbs in past participle (VBN) ($count_{PW}$) and the number of total words ($total\_words$) (American Psychiatric Association, 2015). $R_{PW} = \frac{count_{PW}}{total\_words}$ (5)

<table>
<tr><td></td><td>VBN and VBD are one of the Penn Treebank POS (XPOS) tags.

VBD examples: dipped, pleaded, swiped, soaked, tidied, convened, halted, registered, cushioned, exacted, snubbed, …

VBP examples: predominate, wrap, resort, sue, twist, spill, cure, lengthen, brush, terminate, appear, tend, stray, glisten, obtain.</td></tr>
<tr><td>**Feature #**</td><td>**A6: Low complexity of the language**</td></tr>
<tr><td>**Description/ Explanation**</td><td>Speech Complexity is defined by Syntactic Complexity and Lexical Complexity.

Syntactic complexity *(A6.1), also* called syntactic maturity or linguistic complexity, signifies ‘the range of forms that surface in language production and the degree of sophistication of such forms’ (Ortega, 2003, p. 492). *Syntactic complexity* can be represented as a function of quantitative variables *Sentence length (A6.1.2)* and *Sentence complexity (A6.1.2)* (Lu et al., 2019).

Lexical complexity (A6.2) (or lexical richness) measures the richness of vocabulary in writings. It can be presented as a function of quantitative variables of *Lexical diversity (A6.2.1), Lexical sophistication (A6.2.2) and Lexical density (A6.2.3)* (Lu et al., 2019).</td></tr>
<tr><td colspan="2">**A6*.1.1: Sentence length (MSL)*** measures the number of words in a sentence. An average sentence length is 24.9 words with a standard deviation of 11.6, and usually 126.7 words comprise a paragraph. MSL >25 represents a high complex speech (Ortega, 2003, p. 492).

$$MSL = \frac{\sum_{i-1}^{N} SLi}{N} \qquad (6)$$
where $\sum_{i-1}^{N} SLi$ represents the number words in all sentences and N represents the number of sentences in each discourse sequence.</td></tr>
<tr><td colspan="2">**A6*.1.2: Sentence complexity (Clause Ratio)*** measures the number of sentence phrases per sentence, including features such as the number of clauses (i.e., a structure with a subject and a finite verb) (avg=2.45 for medium complexity in spoken interaction (Kormos, 2011)).

$$CaluseRatio = \frac{n_{all\_clauses}}{n_{all\_senteces}} \qquad (7)$$</td></tr>
<tr><td colspan="2">**A6*.2.1: Lexical diversity (TTR)*** refers to the number of different word stems (i.e., word types) used in the text. This indicator is usually measured by the ratio between distinct word stems ($N_{distinct\_types}$) and number of words (N) in a sentence. This indicator is usually measured by the ratio between distinct word stems (T) and number of words (N) in a sentence (Ellis and Yuan, 2004).

$$TTR = \frac{T}{N} \qquad (8)$$
To identify number of unique word types, we compare word’s *treebank-specific POS (XPOS) tags* and *feats* (i.e., universal morphological features that distinguish additional lexical and grammatical properties of words and are not covered by the POS tags)</td></tr>
</table>

A6*.2.2: Lexical sophistication (MVL)* signifies the degree of sophistication of lexical items (i.e., nouns, verbs, adjectives, and adverbs). It is usually measured by the average length of tokens in each paper, eq. (9), or the coverage of a certain vocabulary list, which can reflect the cognitive complexity for both writers and readers (avg=16.25 characters per item for high complexity) (Ferris, 1994).

A6*.2.3: Lexical density* (L_d) is defined as the proportion of lexical items by the total number of words (avg = above 65 for high complexity. It is calculated by counting the ratio of lexical items (N_{lex}) and all tokens (N) in each discourse sequence (recording), based on their part of speech (Ellis & Yuan, 2004):

$$L_d = \frac{N_{lex}}{N}\ x\ 100 \qquad (10)$$

Appendix A2.: Definition of Digital Speech Cues of Depression

Feature #	**B1: Speaking rate**
Description/ Explanation	Speaking rate (SR_S) is expressed in words per minute (wpm): $SR_S = \frac{total\_words}{total\_speech\_duration}$ (12) where $total\_speech\_duration$ is measured in minutes. The average SR_S is measured at roughly 150 wpm. Speech at rates 100 wpm or below is considered as slow (Barnard D., 2022).
Feature #	**B2: Engagement in verbal communication**
Description/ Explanation	Engagement in verbal communication is defined as “the process by which two (or more) participants establish, maintain and end their perceived connection.” or, more concretely, how interested and attentive they are towards a conversation Yu, Aoki, and Woodruff 2004). One can attempt to capture an instantaneous notion of conversational participants’ feelings by analysing speech as it passes through the voice channel, i.e., by analysing temporal characteristics (pitch, intensity, formants), by analysing prosody, spectral features and voice quality. Overall, we extract the following features to evaluate the engagement in verbal communication: • Pitch • Speech Intensity • Formant Parselmouth library is used to extract the low-level acoustic features (Jadoul et al., 2018).

B2.1: Pitch represents periodicity candidates as a function of time. It is sampled into a number of frames cantered around equally spaced times. The implemented algorithm detects acoustic periodicity based on autocorrelation method (Boersma, 1993) defined as:

$$r_x(\tau) \approx \frac{r_\alpha(\tau)}{r_w(\tau)} \qquad (13)$$

where the autocorrelation $r_x(\tau)$ of the original signal segment is estimated by dividing the autocorrelation of the windowed signal $r_\alpha(\tau)$ with the autocorrelation of the window $r_w(\tau)$. The window length is set to 75 Hz.

B2.2: *Speech Intensity* is represented by an intensity contour calculated around linearly spaced time points in dB SPL: $\mathrm{t_i} = \mathrm{t_1} + (\mathrm{i} - 1)\mathrm{dt}$. Speech intensity is accepted as an important prosodic parameter not only for the speech perception but also for other psychoacoustic studies (Yu et al., 2004).

The implementation of speech intensity consists first squaring the values of signal and then convolving with a Gaussian window (Kaiser-20; sidelobes below -190 dB; with effective duration=3.2/minimum pitch). When dB averaging method is used, mean speech intensity between $\mathrm{t_1}$ and $\mathrm{t_2}$:

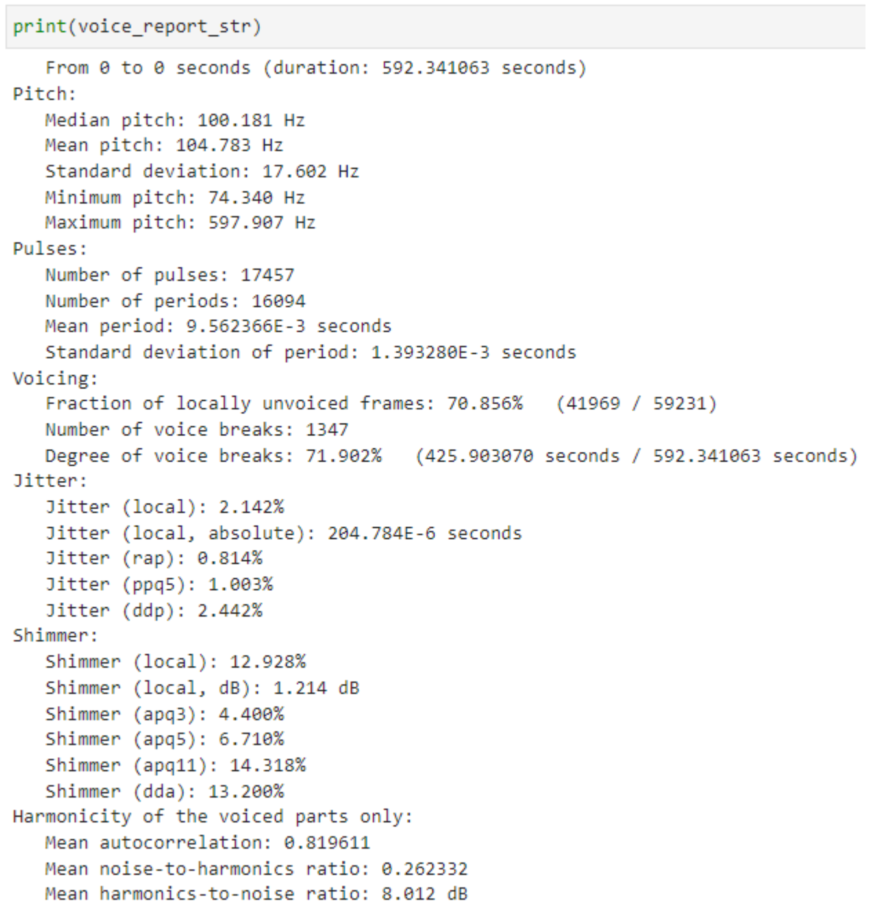
$$\frac{1}{(t_2-t_1)}\int_{t_1}^{t_2} x(t)dt \qquad (14)$$

B2.3: Formant is the measure for the spectral structure of the signal as a function of time (in the units of Hertz or Bark). It shows the local maximum or broad peak in the spectrum (Yu et al., 2004). In the implementation of formants, signal is sampled as equally spaced times with frequency and bandwidth information. Numbering of formants starts from 0 and specify the formant's ordinal number.

Feature #	**B3: Voiced Speech and Pauses**
Description/ Explanation	Unvoiced speech is nonperiodic, random-like sounds, caused by air passing through a narrow constriction of the vocal tract as when consonants are spoken. In normal speech roughly two thirds are voiced. The use of unvoiced speech, i.e., SR_U is calculated as: $SR_U = \frac{N_{Uframes}}{total\_frames}$ (16) where $N_{Uframes}$ represents the total number of unvoiced frames. Voiced and unvoiced frames are separated by silence region which has no speech output means that no excitation supplied to the vocal tract. On the hand, silence region is very important for the understanding and intelligibility of speech. The duration of the silence region categorizes the sound (Rabiner and Schafer, 2011)Kliknite ali tapnite tukaj, če želite vnesti besedilo.. Voiced, silence and duration features of the audio were extracted by using Myprosody Praat script[1]. In voiced regions, Because of not having speech, silence region is identified as having low energy and low amplitude of the signal. Based on Zechner et al. 2009, we automatically extract the following features: • number of pause frames • the duration of pauses • voiced speech frames • the duration of pauses • balance = the duration of voiced speech / original duration • percentage of voiced speech • percentage of pausing (silence) duration
Feature #	**B4: Low Articulation rate**

[1] https://shahabks.github.io/myprosody/, last accessed September 2022

Description/ Explanation	Mean Articulation Rate (MAR) is a prosodic feature, defined as a measure of rate of speaking . All unvoiced speech, including pauses > 250ms, are excluded from the speech. MAR is calculated by considering the average articulation rate of 10-20 syllables, whole word, segments, i.e (Cosyns et al., 2018). $MAR = avg(\sum_1^j AR)$ (17) where j represents the number of segments with complete words that contain at least 10 and at most 20 syllables and avg represents the average function. For each such segment the Articulation Rate (AR) is then calculated as: $AR = \frac{n_{syllables}}{segement\_duration}$ (18) A typical speaking rate for English is between 4 syllables per second. The value depends on the language and culture setting (Kuperman et al., 2021).
Feature #	**B5: Decreased voice quality**
Description/ Explanation	The basics of voice quality with modern approach were given by Kreiman and Gerratt, 2011. The overall quality (or timbre) of a sound is traditionally defined as "that attribute of auditory sensation in terms of which a listener can judge that two sounds similarly presented and having the same loudness and pitch are dissimilar". Voice quality is the consequences of different combination of respiratory regions such as vocal folds, lips, jaw, tongue, and soft palate. Thus, this phenomenon is multidimensional and requires a more complex approach. From the technical point of view, there are different measures related with voice quality such as fundamental frequency (F0), amplitude instability, jitter, shimmer, variation, and harmonics to noise ratio. Having harshness, hoarseness, or vocal roughness voice – also used for pathological voice recognition- are connected with fundamental frequency perturbation (or jitter), amplitude perturbation (shimmer) and F0 variation . The amount of noise in the voice can be measured with harmonics to noise ratio measure and it shows the correlation between normal and decreased voice quality (Kreiman et al., 2003). Jitter in acoustics refers to random fluctuations in the timing or phase of a sound signal. It can be thought of as a form of noise that introduces irregularities or variations in the signal's timing. Generally, lower jitter values are desirable, as they indicate a more stable and accurate signal. Jitter is often measured using statistical methods, such as Peak-to-Peak Jitter: The difference between the maximum and minimum timing variations. Shimmer is another type of acoustic distortion that refers to random fluctuations in the amplitude or intensity of a sound signal. It can be thought of as a variation in the loudness or volume of the sound over time. In the implementation, all voice quality measures are extracted via voice report property of the Parselmouth python library for the specified region of the sound. The technical definitions and formulas can be found in Teixeira et al., 2013 . An example screenshot of the voice report:



```
print(voice_report_str)
```

```
   From 0 to 0 seconds (duration: 592.341063 seconds)
Pitch:
   Median pitch: 100.181 Hz
   Mean pitch: 104.783 Hz
   Standard deviation: 17.602 Hz
   Minimum pitch: 74.340 Hz
   Maximum pitch: 597.907 Hz
Pulses:
   Number of pulses: 17457
   Number of periods: 16094
   Mean period: 9.562366E-3 seconds
   Standard deviation of period: 1.393280E-3 seconds
Voicing:
   Fraction of locally unvoiced frames: 70.856%   (41969 / 59231)
   Number of voice breaks: 1347
   Degree of voice breaks: 71.902%   (425.903070 seconds / 592.341063 seconds)
Jitter:
   Jitter (local): 2.142%
   Jitter (local, absolute): 204.784E-6 seconds
   Jitter (rap): 0.814%
   Jitter (ppq5): 1.003%
   Jitter (ddp): 2.442%
Shimmer:
   Shimmer (local): 12.928%
   Shimmer (local, dB): 1.214 dB
   Shimmer (apq3): 4.400%
   Shimmer (apq5): 6.710%
   Shimmer (apq11): 14.318%
   Shimmer (dda): 13.200%
Harmonicity of the voiced parts only:
   Mean autocorrelation: 0.819611
   Mean noise-to-harmonics ratio: 0.262332
   Mean harmonics-to-noise ratio: 8.012 dB
```

Feature #	**B6: Monotonous speech**
Description/ Explanation	In the early studies (Fry, 1958), pitch was used as a main parameter of stress and, intensity and duration of the voice were evaluated as important as pitch for the speech. In the recent years, lack of linguistic stress, reduced variation in loudness, lower pitch variability and change rate (compared to non-depressed) are specified as distinctive cues of monotonous speech or stressful voice (Marmar et al., 2019). All measures are obtained by Parselmouth and Myprosody libraries. The implementation of pitch, intensity and duration are the same as explained in the previous features.

Appendix A3: Definition of Digital Visual Cues of Depression

Feature #	**C1: Occurrence of facial expressions**
Description/ Explanation	In the research field of emotions there are several theories, which can be organized in three main classes: affect-program theories, constructionist theories, and appraisal theories. Even though they differ in many points, there is an agreement on the following characteristics or components of emotion: they are reflected in changes in subjective feelings, accompanied by a cognitive component, they guide us toward action, and are expressed in behavior (e.g., facial expression) and physiology. They usually follow an antecedent event and are directed at an object (e.g., situation or person) (Lange et al., 2020).

<table>
<tr>
<td></td>
<td>Mapping of action units onto basic emotion categories (anger, disgust, fear, happy, sadness and surprise) was described in a systematic review study (Clark et al., 2020). Those mapping was achieved according to the FACS Investigators' Guide (Ekman et al., 2002). by breaking down a single facial expression into different combinations of AUs (9 AUs in upper face+18 AUs in the lower face) using a finite number of rules.
<table>
<tr><th>Basic Emotion</th><th>Mapping to Combinations of AUs</th></tr>
<tr><td>Anger</td><td>4+5+7+10+22+23+25/26
4+5+7+10+23+25/26
4+5+7+17+23/24
4+5+7+23/24
4+5/7
17+24</td></tr>
<tr><td>Disgust</td><td>9/10+17
9/10+16+25/26
9/10</td></tr>
<tr><td>Fear</td><td>1+2+4
1+2+4+5+20+25/26/27
1+2+4+5+25/26/27
1+2+4+5
1+2+5+25/26/27
5+20+25/26/27
5+20
20</td></tr>
<tr><td>Happy</td><td>12
6+12</td></tr>
<tr><td>Sadness</td><td>1+4
1+4+11/15
1+4+15+17
6+15
11+17
1</td></tr>
<tr><td>Surprise</td><td>1+2+5+26/27
1+2+5
1+2+26/27
5+26/27</td></tr>
</table>
(A/B means "either A or B")

Openface library (Baltrusaitis et al., 2016) provides two different descriptions for action units: presence/occurrence (if AU is visible in the face, 0 or 1) and intensity (how intense is the AU (minimal to maximal) on a 5-point scale).
The occurrence of each emotion was calculated according to each possible AU combination as shown in the table above. For this hand-crafted feature, the number of frames for each facial expression were calculated.</td>
</tr>
<tr><td><b>Feature #</b></td><td><b>C2: Intensity of facial expressions</b></td></tr>
<tr><td><b>Description/ Explanation</b></td><td>Average Intensity of each emotion was calculated according to each possibility AU combination as shown in the table above (Clark et al., 2020). Calculation was done by taking the average intensity values of all related emotion's AUs of one of the possibilities that shows the intensity greater than 1.</td></tr>
</table>

	Additionally, average intensities of positive emotions (happy) and negative emotions (sadness, surprise, anger, fear and disgust) were provided. Also, total emotion variability was calculated as the total number of emotions which have average intensity value greater than 1.
Feature #	**C3: Occurrence and emotional variability**
Description/ Explanation	Occurrence of positive, negative (sadness, surprise, anger, fear and disgust) and neutral (frames that have an intensity equal or lower than the minimum intensity of all AUs) expressions was calculated according to each possible AU combination as shown in the table above. For this hand-crafted feature, the number of frames for each facial expression were calculated and an average intensity based on maximal intensity within the single expression frames is calculated.
Feature #	**C4: Eyebrow movements**
Description/ Explanation	The presence of the eyebrow movement is determined according to the related action units (AU1- Inner Brow Raiser, AU2- Outer Brow Raiser and AU4- Brow Lowerer) on the face (Wood et al., 2015). We obtained those action units via Openface library and calculated the intensity values of each action unit (frames that have intensity value higher than 1 (either AU1 or AU2 or AU4) marked as moving eyebrow frame). For this hand-crafted feature, the average intensity of moving eyebrow frames and the number of moving eyebrow frames were calculated.
Feature #	**C5: Gaze aversion and downward gaze**
Description/ Explanation	Openface library uses the fast and accurate method for eye-gaze estimation. The estimation of eye gaze and detection of eyelids, iris, and the pupil are obtained with a Constrained Local Neural Field (CLNF) landmark detector (Baltrušaitis et al., 2013). CLNF landmark detector was trained by the SynthesEyes training dataset (Wood et al., 2015). Computation of the eye gaze vector is done for each eye and by using the pupil and eye location to obtain the 3D camera coordinates. The estimated gaze vector is calculated from the 3D eyeball center to the pupil location. Openface provides gaze information with 4 vectors: two vectors are in real world coordinate space describing the gaze direction of both eyes; other two vectors describe the gaze in head coordinate space. For our study, Eye gaze feature was calculated from the gaze vectors of Openface output file according to the angle difference in y direction: Magnitude: $\lvert F\rvert = \sqrt{F_x^2 + F_y^2 + F_z^2}$ (19) Direction: $\cos Q_x = \frac{F_x}{F}, \cos Q_y = \frac{F_y}{F}, \cos Q_z = \frac{F_z}{F}$ (20) Angle difference of consecutive frames in degree was calculated by inverse cosine function for y component and looking down decision is taken as: If the difference of gaze angle in y directions between consecutive frames is less than 0, then this means that person is looking down.
Feature #	**C6: Frowns**
Description/ Explanation	The presence of the lip press is determined according to the related action unit (AU4-Brow Lowerer) on the face (Ekman and Friesen,1978; Pantic and Rothkrantz, 2000). We obtained the action unit AU4 via Openface library and calculated the intensity value (frames that have intensity value higher than 1) marked as frowning frame).

	For this hand-crafted feature, the average intensity of frowning frames and the number of frowning frames were calculated.
Feature #	**C7: Head movement and turns**
Description/ Explanation	Openface library provides head pose (translation and orientation) information for 3 directions (up-down, left-right and z-direction). Head pose estimation is done mainly by using the 3D representation of facial landmarks (output of Openface library) and projecting those landmarks onto the frame via orthographic camera projection. This projection provides accurate estimation by using perspective problem of n points (de la Rosa et al., 2013) . We obtained all the head pose values (up-down direction-pose_Ry, left right direction-pose_Rx and z direction pose_Ry) via Openface library and filtered the values (frames that have a value higher than absolute (0.3)) marked as head moving frame). For this hand-crafted feature, the number of head moving frames were calculated for all directions.
Feature #	**C8: Rotational energy of Head Movements**
Description/ Explanation	Rotational energy shows the energy of a moving object. The differences from the regular kinetic energy (for objects moving along a straight line) are to measure the energy of rotating objects and to use the angular velocity instead of a regular speed: $K = \frac{1}{2} I w^2$ (21) where: K is the rotational kinetic energy (Joules); I is the moment of inertia of the object (characteristic property of a rigid body) (kg*m²) and ω is the angular velocity of the body (radians per second) (Yoganandan et al., 2009). In our case, the total rotational energy is equal to the sum of rotational energies of all three directions (x, y, z). And rotational coordinates were used to compute the angular velocity in radians for all directions. Moment of inertia was calculated according to specimen-specific data from Plaga et al. (2005) as $I = m\, x r^2$. We used the mean of all participants (Iyy (kg-cm2): MOI = 148.44) to use in our study. As a last step, we classified the rotational energy values into 3 energy levels for observing the different energy levels: • If the total rotational energy is between 50 and 500, we marked that frame as frame with small rotational energy. • If the total rotational energy is between 500 and 5000, we marked that frame as frame with large rotational energy. • If the total rotational energy is greater 5000, we marked that frame as frame with high rotational energy.
Feature #	**C9: Mouth Movement**
Description/ Explanation	The presence of mouth movement is determined according to the related action units (AU12-Lip Corner Pull, AU15-Lip Corner Depress, AU20-Lip stretcher, AU23-Lip Tighten, AU25-Lips Part and AU26-Jaw Drop) on the face (Lien et al., 2000). We obtained those action units via Openface library and calculated the intensity values of each action unit (frames that have intensity value higher than 1 (AU12 or AU15 or AU20 or AU23 or AU25 or AU26) marked as moving mouth

	frame). For this hand-crafted feature, the average intensity of moving mouth frames and the number of moving mouth frames were calculated.
Feature #	**C10: Lip Press**
Description/ Explanation	The presence of the lip press is determined according to the related action unit (AU23-Lip Tightener) on the face (Ekman and Friesen,1978; Pantic and Rothkrantz, 2000). Although there are other action units related with lip pressing (AU24- Lip Pressor or AU28- Lip Suck), those action units are not provided by OpenFace library. We obtained the action unit AU23 via Openface library and calculated the intensity value (frames that have intensity value higher than 1) marked as lip pressed frame). For this hand-crafted feature, the average intensity of lip pressed frames and the number of lip pressed frames were calculated.
Feature #	**C11: Down-Angled Mouth Corners**
Description/ Explanation	The presence of the down-angled mouth corners is determined according to the related action units (AU15-Lip Corner Depressor and AU20-Lip stretcher) on the face (Lien et al., 2000). We obtained those action units via Openface library and calculated the intensity values of each action unit (frames that have intensity value higher than 1 (either AU15 or AU20) marked as down-angled mouth frame). For this hand-crafted feature, the average intensity of down-angled mouth frames and the number of down-angled mouth frames were calculated.
Feature #	**C12: Smile Intensity and Frequency**
Description/ Explanation	The presence of smiling is determined according to the combination of related action units (AU6- Cheek Raiser and AU12- Lip Corner Puller) on the face (Park et al., 202. We obtained those action units via Openface library and calculated the intensity values of each action unit (frames that have intensity value higher than 1 (both AU6 and AU12) marked as smiling frame). For this hand-crafted feature, the average intensity of smiling frames and the number of smiling frames were calculated. In addition, we also calculated the length of consecutive frames that have a smiling appearance and their total occurrence number during the conversation.