# Multimodal Probabilistic Learning of Human Communication
# Homework 1

**Due date:** Friday, September 30th, 11pm PST
**Instructions:** Please send your homework responses directly on USC Blackboard in PDF format. You can decide to include separate figures (also in PDF format) by zipping them together.

**QUESTIONS**

Please answer the following three questions by analyzing the YouTube videos and their annotations:

1. Watch all videos and take notes while watching. Pay attention to both nonverbal and verbal behavior alike. Subjectively note what types of behavior correlates with positive/negative/neutral opinions and write a paragraph about your findings. (~1/2 page)
2. We provide you with annotations of three independent "experts" that rated each spoken utterance with labels negative (-1) neutral (0) and positive (+1). Use Krippendorf's alpha as a parameter and measure of the inter-rater agreement or reliability. Briefly discuss your findings and discuss the alpha values. (~3-4 sentences)
3. Confirm your suspicions and subjective observations (from 1.) with some statistical analysis. Choose about 4-5 audiovisual cues (do not limit your analysis to just one modality) and visualize them with respect to the three classes (positive, negative, neutral). The ground truth labels are provided by the majority vote of the three independent annotators. Use boxplots for the visualization of the data and confirm tendencies with t-tests or ANOVA and analyze possible significant findings. Write a paragraph about your findings and discuss them. (~1/2 page)

   To perform the statistical analysis required in Question 3, you are supposed to segment the visual (or audio features) based on the segments defined in the sentimentAnnotation_rev_v02.xlsx file, perform some summarization per segment (e.g., mean or standard deviation) and group these observations based on their polarity (negative, neutral and positive). For the purpose of the t-test, you probably want to focus on negative vs positive (but other pairs could also be interesting).

**DATASET**

We collected 48 videos from the social media web site YouTube. The dataset is created in such a way that it is not based on one particular topic. The videos were found using the following keywords: opinion, review, product review, best perfume, toothpaste, war, job, business, cosmetics review, camera review, baby product review, I hate, I like.

The dataset has 20 female and 28 male speakers randomly selected from youtube.com, with their age ranging approximately from 14-60 years. Although from different ethnic backgrounds (e.g., Caucasian, African-American, Hispanic, Asian), all speakers expressed themselves in English. The videos are converted to .mp4 format with a standard size of 360x480. The length of the videos varies from 2-5 minutes.

All videos are pre-processed to address the following issues: introductory titles and multiple topics. Many videos on YouTube contain an introductory sequence where a title is shown, sometime accompanied with a visual animation. As a simple way to address this issue, the first 30 seconds of each video is removed.

The second issue is related to multiple topics. Videos posted on social networks can address more than one topic. For example, a person can start by talking about the new movie she recently saw and then switch to a new (or related) topic such as food served in movie theaters. To simply address this issue, all video sequences are normalized to be 30 seconds in length.

The dataset can be downloaded here:
https://www.dropbox.com/s/m7yk3wuzpjvq85j/Homework_1.zip?dl=0

**ACOUSTIC FEATURES**

- the acoustic features are provided in the folder: AcousticFeatures; all the features are sampled at 100Hz i.e. every 10 ms a new value; the .feat files are csv files and the columns are ordered as follows:
    - Voiced-Unvoiced: values {1, 0}; if the value is 0 the speech signal is not voiced, which means the vocal folds are not vibrating. fundamental frequency or pitch therefore is not valid in those parts
    - PeakSlope: [1, -1] usually a negative value; is a measure of the tenseness of the voice; higher negative values indicate tenser voices
    - NAQ: [0, 1]; measure of the tenseness of the voice; higher values indicate tenser voices
    - fundamental frequency; [50, 300]; the frequency of the vocal fold vibration
    - energy; in dB; higher negative values are less loud samples
    - energy slope; the first derivative of energy; the change of energy
    - spectral stationarity; a measure of the monotonicity of the speech; higher in more monotonous speech

**VISUAL FEATURES:**

Each video file was processed using 4 software: OKAO, CLM, GAVAM and SHORE. We describe below the output of each software. It is important to note that the framerates (number of frame/image per second) are not constant between videos since they were downloaded directly from YouTube. YouTube do not resample all videos at a constant framerate. Instead, they publish the same framerate used to record the original video. Our clipped videos also kept the original framerates. The annotation file sentimentAnnotation.xlsx should contain the framerate for each video.

1. **OKAO:** Please look at the file OKAO_info.pdf for details about the OKAO features.
2. **CLM:** The output Video with the 66 CLM-Z Points tracked, in the .avi format.
    a. The CLMZ output file containing the corresponding frame number of the frame being processed in the first column, followed by the coordinates of the 66 points (so, 66 x 2 columns) and the last column is the measure of the CLM uncertainty.
3. **GAVAM:** The GAVAM output file is organized as follows :
    a. The first column is the frame being processed.

b. The second column is the indicator of the time of the occurrence of the particular frame in milliseconds. So say a value of 100 means, it occurs at the 100th millisecond.
c. The third column indicates whether the GAVAM or the CLM pose estimate is used. (1-GAVAM, 2-CLM).
d. The fourth column is the displacement of the face w.r.t X-axis (measured by the displacement of the normal to the frontal view of the face in the X-direction).
e. The fifth column is the displacement of the face w.r.t Y-axis.
f. The sixth column is the displacement of the face w.r.t Z-axis.
g. The seventh column is the angular displacement of the face w.r.t X-axis (measured by the angular displacement of the normal to the frontal view of the face with the X-axis).
h. The eighth column is the angular displacement of the face w.r.t Y-axis.
i. The ninth column is the angular displacement of the face w.r.t Z-axis.
j. The tenth column is the measure of the confidence.

4. **SHORE:** The SHORE output file is pretty self-revealing. The first column states the frame being processed. The other parameters are quite easy to interpret. Few of the other columns are e.g. the type of object identified, may be a "face". Again, another might indicate an estimate of age, while another one might state the bounding box of the mouth.

Note: The visual features were extracted with these 4 different software. For the purpose of this homework, you should suppose that each of these software create an output file which starts at frame 1 and ignore any extra frame at the end of the result files. In other words, if for example OKAO returns 3 extra rows (representing 3 more frames) of results when compared with GAVAM, you can ignore these 3 extra rows.

**TRANSCRIPTIONS**

All video clips are manually transcribed to extract spoken words as well as the start time of each spoken utterance. The Transcriber software was used to perform this task. The transcription is done using only the audio track of each video clip. In other words, the transcriptions are done without the visual information.

To load a Transcriber file in Matlab, you can use the file readTrans.m found in the \matlab directory of this homework. The Transcriber files are simple text files where each line start with a timestamp followed by the spoken words. Pauses are expressed using the backslash symbol. You can find more details about the transcription scheme below:

INTONATION

| | | |
|---|---|---|
| . | i'm done . | falling intonation |
| ? | why ? | rising intonation |

PAUSE

| | |
|---|---|
| / | micro pause (less than 150 miliseconds) |
| // | pause ( .150 - 1 second) |
| /// | pause ( 1 second) + |

PRONUNCIATION

| | | |
|---|---|---|
| ( ) | (uncertain word) | uncertain interpretation |
| (xxx) | unintelligible or untranscribable utterance | |

| | | |
|---|---|---|
| ex_a_c_tly | | slower or emphasized |
| : | i li::ke it | lengthening; the more colons, the more elongation |
| - | jona- | incomplete word |

VOLUME

| | | |
|---|---|---|
| CAPS | WATCH OUT | louder speech relative to the adjacent speech |
| * * | *some secret* | softer speech or whisper |

LAUGHTER

| | |
|---|---|
| ((laughter)) | for a lot of laughter. |

OTHER

| | |
|---|---|
| (( )) | ((knock on door interruption)) |
| | additional comments by transcriber |

**SENTIMENT ANNOTATIONS**

All spoken utterances were first manually segmented. Then three separate annotators were asked to judge the sentiment expressed during each utterance: 0 for neutral utterances, -1 for negative utterances and 1 for positive utterances. The file sentimentAnnotations.xls contains the list of sentiment annotations for each annotator (columns 6, 7 and 8). The second and third columns of this file contains the start and end time of each utterance. The fourth column contains the framerate of the video (which is different for most videos). The columns 5 and 6 contains the same start and end points of the segments but translated for your convenience in frame number. A frame number of 1 means that it's the first frame in all 4 visual feature result files (OKAO, GAVAM, CLM and SHORE). The final column contains the majority vote, which can be used as final label for each utterance.

**USEFUL INFORMATION**

- You can load comma separated value (CSV) files with matlab using functions like: csvread, importdata, readtable (Use the built in documentation in matlab if you need detailed information: >> doc csvread)
- boxplots can be drawn using the boxplot function and statistical tests with ttest command and anova respectively.
- if in doubt… ask us or other students using email or the blackboard! (we might be more responsive via email though)