

# INF642 Lab: Socio-emotional Embodied Conversational Agents

## *Upper Facial Gesture Generation based on Speech Prosody*

**Lab Instructor:** Mireille Fares

**Contact:** mf.upmc@gmail.com

### General Description

*Speech* is the “mirror of the soul”: it can reflect our emotional states and feelings. Our most personal experience can be expressed by the mean of speech, in multiple degree of variation, rendering each expression a unique act. *Prosody* refers to all suprasegmental aspects of speech [2]. It involves pitch, duration, amplitude, and voice quality that are used to perform lexical contrasts and convey meaning.

Human-machine interactions can become as natural as human-human interactions by employing speech with prosodic expressiveness, that is, speech with emotional content [1]. The emotional content of speech can be conveyed by manipulating the acoustic as well as prosody parameters (f0, duration, and energy).

One of the key challenges in designing *Embodied Conversational Agents* is to produce human-like gestural and visual prosody expressivity. The goal of the three lab sessions is to develop a model that can predict expressive upper facial gestures of a virtual agent, based on acoustic and prosodic speech signals. The three lab sessions will tackle the following:

- Lab 1: Acoustic and visual data preparation and preprocessing
- Lab 2: Developing a deep learning model that can predict upper facial expressions based on acoustic and prosodic voice features
- Lab 3: Evaluating and experimenting the developed model

**N.B.:** You are strongly encouraged to do the **Bonus** parts.

### References

- [1] Carlos Monzo, Ignasi Iriondo, and Joan Claudi Socoró. 2014. Voice quality modelling for expressive speech synthesis. *The Scientific World Journal* 2014 (2014).
- [2] Yi Xu. 2019. Prosody, tone and intonation. *The Routledge handbook of phonetics* (2019), 314–356.

# INF642 : Socio-emotional Embodied Conversational Agents

## Lab 1: Acoustic and Visual Data Preprocessing

15/01/2020 – 1:30 pm

**Note:** The deadline to submit this lab is : 20/01/2020 midnight. You need to submit a report, explaining every step of your work, as well as your source code, and results. Add all the deliverables in a ZIP file, save it with your first and last name and send it to [mf.upmc@gmail.com](mailto:mf.upmc@gmail.com)

**Objective(s):** The first objective is to get the dataset and extract the features that we need for training our model. After extracting the features, we will do some preprocessing to clean our data. Another objective is to get familiarized with OpenSmile and OpenFace.

### Dataset

We are going to use TEDx talks as a dataset from which we will extract voice acoustic and prosodic features, as well as the upper facial expressions of the speakers.

Download the URLs of the talks from:

[https://drive.google.com/file/d/1luLNLv7Hr8LQI4SU2\\_Y730dnQGfc3pYY/view?usp=sharing](https://drive.google.com/file/d/1luLNLv7Hr8LQI4SU2_Y730dnQGfc3pYY/view?usp=sharing)

We need to have two formats for each talk, the first one is the “.mp4” format, and the second one is the Waveform Audio File format “.wav”

Download the TEDx talks videos (extension: ‘.mp4’) using “youtube-dl” (or any other tool if you prefer).

Convert the “.mp4” files to “.wav” files using ‘ffmpeg’ (<https://www.howtoforge.com/tutorial/ffmpeg-audio-conversion/>)

### OpenSmile Tool

OpenSmile is an open-source toolkit for audio feature extraction and classification of speech and music signals. We are going to use this tool to extract voice acoustic and prosodic features from the dataset that we have. Follow these steps :

1. Go to <https://www.audeering.com/opensmile/> and download OpenSmile
2. Open “README.md”, and follow the quick start section to build the required files
3. Whether you’re on Windows or Linux, in this lab you are going to use the executable/binary ‘SMILExtract’ that you will find in “./build/progrsrc/smilextract” after doing what is asked in the previous step
4. To know more about OpenSmile, check out this link:  
<https://audeering.github.io/opensmile/about.html#capabilities>

### OpenSmile – Extracting prosodic and acoustic features

We are going to use the following prosodic and acoustic features: **Fundamental Frequency (F0), Mel Frequency, Jitter, Shimmer, and Harmonic to Noise Ratio (HNR)**

To extract them from our .wav files, we must first create a configuration file to specify the features that we need to extract. You will find examples of configuration files in ‘config’ folder.















After creating and editing your configuration file, you will have to use ‘SMILEExtract’ along with your configuration file to extract the required features. The following command in Linux is an example of extracting the features from a .wav file and saving them in a .csv :

```
./SMILEExtract -C MyConf.conf -I _1VpOweDio8.wav -O _1VpOweDio8.csv;
```

You must extract all the features for all the dataset. (One way to do it, is by creating and executing a .sh file that contains the commands for all the .wav files)

## OpenFace Tool

The work in this and the following labs will be based on *Facial Action Coding System (FACS)* [1], a system that describes the facial movements based on 44 Action Units (AUs). The AUs that represent eyebrows and eyelids movements are AU01, AU02, AU04, AU05, AU06, AU07 and AU09. However, we are going to use the first 6 action units and disregard AU09.

AU1  Inner brow raiser	AU2  Outer brow raiser	AU4  Brow Lowerer	AU5  Upper lid raiser	AU6  Cheek raiser
AU7  Lid tighten	AU9  Nose wrinkle	AU12  Lip corner puller	AU15  Lip corner depressor	AU17  Chin raiser
AU23  Lip tighten	AU24  Lip presser	AU25  Lips part	AU27  Mouth stretch	

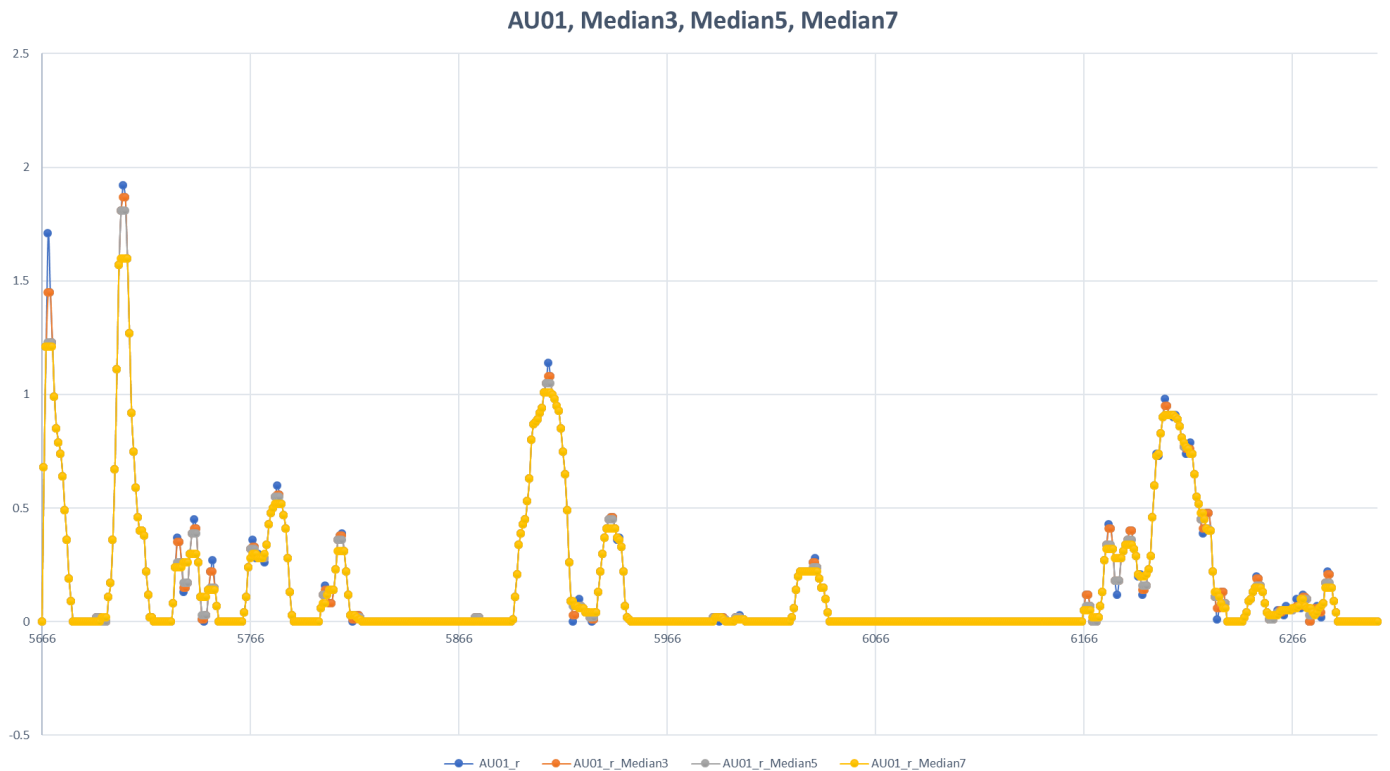
You need to extract the AUs intensities features using OpenFace [2] for all the videos (.mp4) of our TEDx dataset.

First, go to <https://github.com/TadasBaltrusaitis/OpenFace>, download OpenFace. To extract the Action Units you will need to use the executable ‘FeatureExtraction.exe’. Read the documentation to understand how it functions.

In OpenFace, each generated AU intensity is given a “Success” score (0 or 1), and a “Confidence” level (between 0 and 1). After extracting the features, you need to apply a **median filter** with a window size of your choice (choose a good one to remove the noises, without eliminating the peaks), to remove noises and deal with the cases where the confidence is low.

Plot some of raw and filtered data for the AUs, to illustrate how you have eliminated the noise, and add these plots to your report.

One example of a plot is illustrated in the following picture. This plot shows the results of a median filter applied on AU01 values, with different window size values (3, 5 and 7)



After applying the median filter, you need to remove the frames where the success is equal to zero, from the generated OpenFace files. You also need to remove their corresponding frames from OpenSmile generated files, so that for a specific frame with a success =1, we have the action units and their corresponding acoustic and prosodic features.

## References

- [1] Rosenberg Ekman. 1997. What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS). Oxford University Press, USA.  
<https://www.cs.cmu.edu/~face/facs.htm>
- [2] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. 2016. Openface: an open source facial behavior analysis toolkit. In 2016 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 1–10.