Apprentissage pour l'interaction humain/machine

TP - Socio-emotional Embodied Conversational Agents

Joel CABRERA
angel_joel.joel@etu.sorbonne-universite.fr

Contents

1 Data Preparation and Preprocessing

1.1 Data Collection

The audiovisual data were sourced from TEDx talks on YouTube and downloaded using `yt-dlp` to ensure access to high-quality streams. The frame rate was standardized to 30 fps using `FFmpeg`, and the audio tracks were extracted and converted to `.wav` format for subsequent analysis.

1.2 Acoustic Feature Extraction (OpenSmile)

Acoustic features were extracted using the OpenSmile toolkit, focusing on the Fundamental Frequency (F_0), which represents the vibration rate of the vocal folds and is closely related to perceived pitch [1]. Figure 1 shows raw F_0 trajectories for three TEDx talks. Distinct F_0 ranges reflect individual speaking styles and expressiveness. Abrupt spikes correspond to unvoiced segments or pitch detection errors, motivating the application of median filtering to smooth discontinuities while preserving prosodic structure.

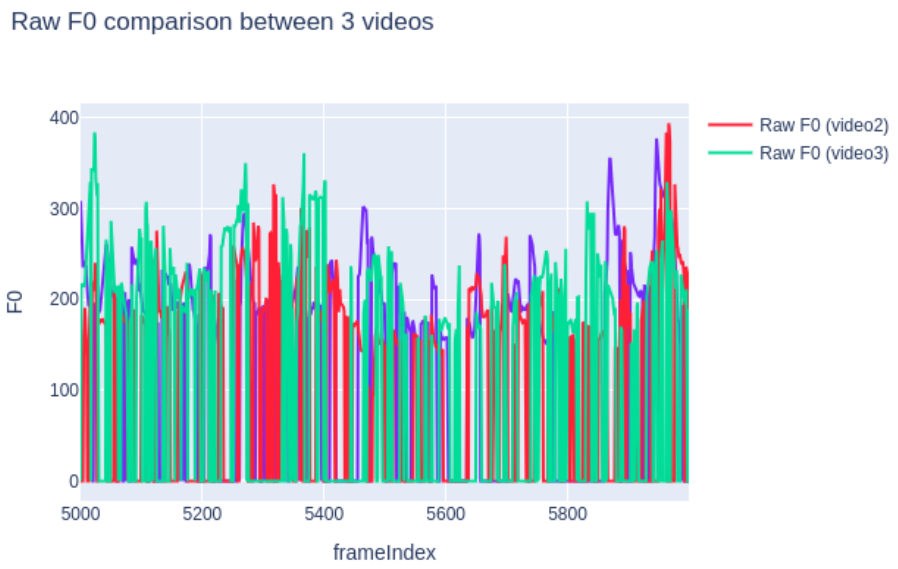


Figure 1: Raw F_0 trajectories for three TEDx talks over frames 5000–6000, showing speaker-specific pitch fluctuations.

1.3 Facial Feature Extraction (OpenFace)

Facial data were obtained using OpenFace, which provides facial keypoints and Action Units (AUs) defined by the Facial Action Coding System (FACS) [2]. AUs describe specific muscle activations (e.g., AU01: inner brow raiser, AU04: brow lowerer) that can be used to model facial expressions objectively. Signals were smoothed using a median filter with a kernel size of 7 frames (230 ms at 30 fps), effectively removing noise while preserving expression peaks (Figure 2).

Temporal alignment between OpenSmile and OpenFace data was achieved using timestamps as a common key. Frames with unsuccessful detections (`Success = 0`) or low confidence (`Confidence < 0.5`) were discarded. As shown in Figure 3, the peak in AU01_r corresponds to an observed eyebrow lift, confirming the consistency between visual and numerical data.

1.4 Tool Limitations

Even though OpenFace provides reliable results overall, there are still segments where facial detection fails or yields low-confidence predictions. These issues may arise from low video resolution, rapid head movements, or suboptimal lighting conditions. Furthermore, as the model is pre-trained,

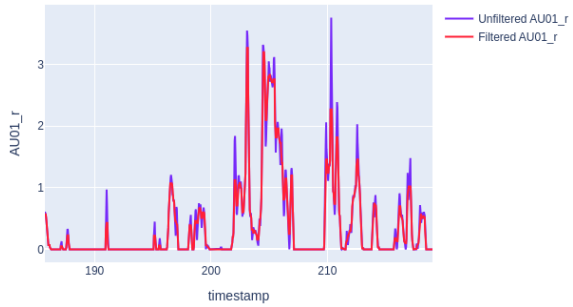
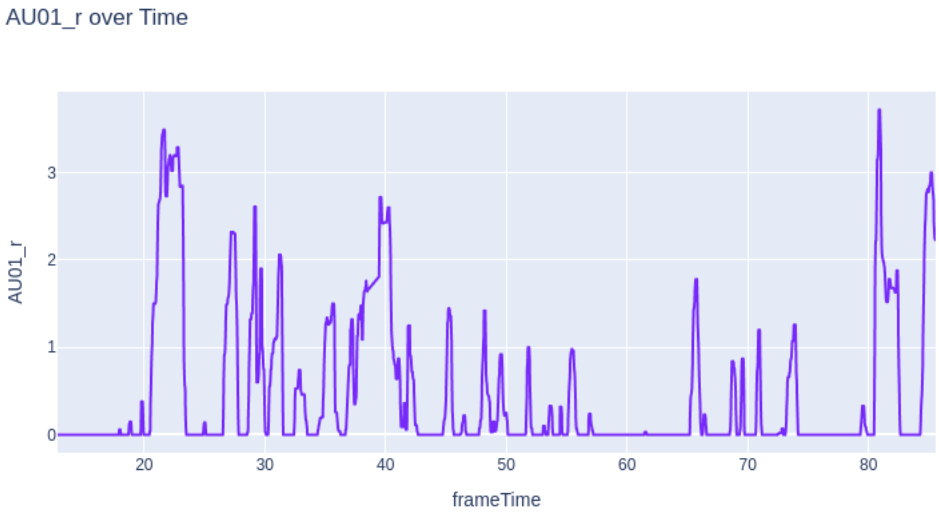


Figure 2: Example of AU01_r before and after applying a median filter.



(a) Temporal evolution of AU01_r (inner brow raiser) across the video segment. A peak is observed around 1:20, indicating strong eyebrow activation.

(b) Frame before activation (neutral eyebrows).

(c) Frame at 1:20 showing a strong AU01 activation (eyebrows raised).

Figure 3: Visualization of AU01_r dynamics and corresponding video frames. The sudden rise in AU01_r around 1:20 matches the visual observation of the speaker's eyebrow movement.

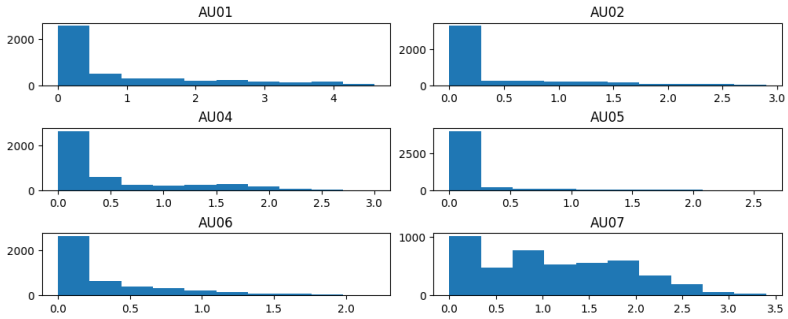
it may exhibit biases and perform less accurately on faces with diverse skin tones, medical conditions, or atypical features. Another practical challenge is the installation and dependency management process, which can be time-consuming and sometimes frustrating for non-expert users. In my case, most difficulties occurred during installation on Ubuntu 24.04, mainly due to GCC version incompatibilities.

2 Model Development for Acoustic-based Facial Expression Prediction

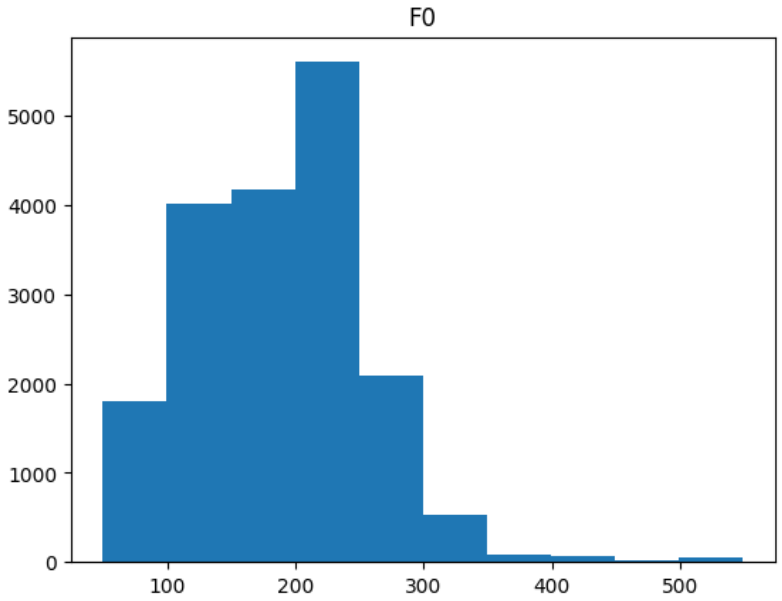
2.1 Data processing

If we observe the distribution of the action units in the Figure 4a, the first observation is that most AU values have a distribution with a strong tail on the left, meaning that the majority of values are small and close to 0. Values far from 0 may correspond to noise. However, AU07 (lid tightener) presents a wider spread, which makes sense given that eye tightening is frequent in natural facial expressions.

In contrast, F_0 values are much higher as we can see in Figure 4b, with an average around 220 Hz. Its distribution resembles a Gaussian shape, though we can also observe a few large values (around 500 Hz) with low occurrence.



(a) Distribution of Action Unit (AU) intensities. Most AUs have low activation values, except AU07 which shows higher average intensity.



(b) Distribution of fundamental frequency (F_0) values, centered around 220 Hz with a roughly Gaussian shape.

Figure 4: Distributions of visual (AU) and acoustic (F_0) features used for model training.

To reduce the effect of outliers, we clipped the data to retain the central 90% of values, keeping those between the 0.05 and 0.95 quantiles. We then normalized the data to the range [0,1]

and quantized it. For quantization, we selected non-padded values (mask > 0) and computed the rounded product with $(B-1)$, using $B = 16$. Since this quantized version is the model output, we also implemented an inverse function to recover the original scale. As shown in Figure 5, the reconstructed distributions closely match the originals, although some fine details are lost, particularly for AU07 due to the discrete resolution of $B = 16$.

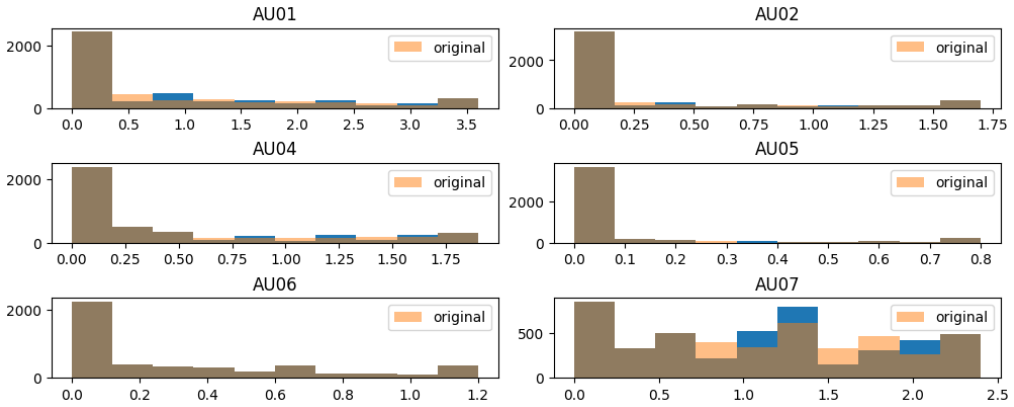


Figure 5: Comparison of original and reconstructed AU distributions after quantization. The recovered signals maintain the general shape of the originals, with minor deviations in high-intensity ranges.

2.2 Model Development

The goal is to model the mapping from acoustic prosody to upper-face Action Units (AUs): given a sequence of F_0 values, the model predicts six AU token sequences {AU01, AU02, AU04, AU05, AU06, AU07}.

Input representation. The F_0 signal is preprocessed, normalized to $[0, 1]$, and passed through a learnable `F0_encoder` that produces frame-wise embeddings of dimension $d_{\text{model}} = 64$. A sinusoidal positional encoding is then added to retain temporal order.

Positional encoding. Since the Transformer architecture does not have any inherent notion of order, we include a sinusoidal positional encoding (PE) to provide each time step with an explicit temporal reference. This encoding injects information about the relative position of frames into the embedding space, allowing the model to capture temporal dependencies such as pitch variations and rhythm patterns in the F_0 contour. The positional encoding is added element-wise to the output of the `F0_encoder` before passing it to the Transformer encoder. Formally, each dimension of the encoding is defined as:

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{2i/d_{\text{model}}}}\right), \quad PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{2i/d_{\text{model}}}}\right),$$

where pos is the frame index and i is the dimension. This continuous representation helps the network generalize to sequences of different lengths and to interpolate smoothly between time steps.

Encoder. A 4-layer Transformer encoder (8 heads, `batch_first=True`) processes the acoustic embeddings and outputs a memory tensor $X \in R^{B\times T\times 64}$. A source key padding mask is applied to ignore invalid or missing acoustic frames.

Target representation. Each AU intensity is quantized into B discrete bins, and special symbols `<SOS>`, `<EOS>`, and `<PAD>` are added (16,17 and 18 respectively). This transforms AU regression into a sequence prediction problem over a vocabulary of size V.

Decoders. The model includes six independent Transformer decoders (one per AU), each composed of 4 layers and 8 attention heads. At time step t, a decoder takes the acoustic memory X and the previously generated AU tokens to predict logits over the AU vocabulary. Key padding masks are used to ignore padded elements, while a causal mask prevents future information leakage during training.

Training objective. The model is trained with a token-level cross-entropy loss, averaged over time steps and across all six AUs, while ignoring `<PAD>` tokens:

$$\mathcal{L} = \frac{1}{6} \sum_{k=1}^{6} \frac{1}{|\mathcal{T}_k|} \sum_{t \in \mathcal{T}_k} \mathrm{CE}\big(p_\theta^{(k)}(y_t \mid y_{<t}, X),\, y_t\big),$$

where $\mathcal{T}_k$ denotes the valid positions of AU k. Teacher forcing[1] is used during training.

Inference. At inference time, the model first encodes the F_0 sequence and then performs greedy decoding for each AU independently, starting from `<SOS>` and stopping when `<EOS>` is reached or the target length is exceeded. This approach ensures efficient and stable generation given the small vocabulary size.

Limitations. As we saw before, quantization introduces reconstruction error, and F_0 alone does not capture the full complexity of facial motion. Temporal misalignments or frames with low confidence may also degrade context information, which is why proper masking is essential during both training and inference.

[1] Teacher forcing is a training strategy where, at each decoding step, the model receives the ground-truth output from the previous time step instead of its own prediction, which helps stabilize and accelerate convergence.

3 Evaluation, Experiments, and Virtual Agent Simulation

3.1 Training Setup and Observations

The data were structured as tensors with dimensions **[B,C,S]**, where **B** is the batch size, **C** the number of AU channels, and **S** the sequence length. For instance, a batch shape of `[32, 3, 50]` corresponds to 32 samples, each containing 3 AU sequences of 50 frames.

The model was trained using a Transformer with `D_MODEL = 64`, `NHEAD = 8`, and four encoder–decoder layers. We used a batch size of 32, a learning rate of `1e-4`, and a small weight decay (`5e-5`) for regularization. The loss function was the cross-entropy loss with label smoothing (0.05), ignoring the `<PAD>` token.

One crucial aspect during training was the correct use of padding masks. Without masking padded positions in both the encoder and decoder, the model incorrectly learned from invalid time steps, leading to unstable or meaningless predictions. Applying masks during training and excluding padding tokens from the loss was therefore essential to achieve proper convergence.

Throughout training, we observed that as the number of epochs increased, the model tended to generate longer AU sequences before predicting the `<EOS>` token. However, this also caused the validation loss to increase after around 200 epochs (Figure 6). This behavior suggests a trade-off: longer generated sequences capture more temporal information but accumulate more token-level errors, which penalizes the validation score. For this reason, we selected checkpoints trained up to roughly 200 epochs, where this balance was optimal.

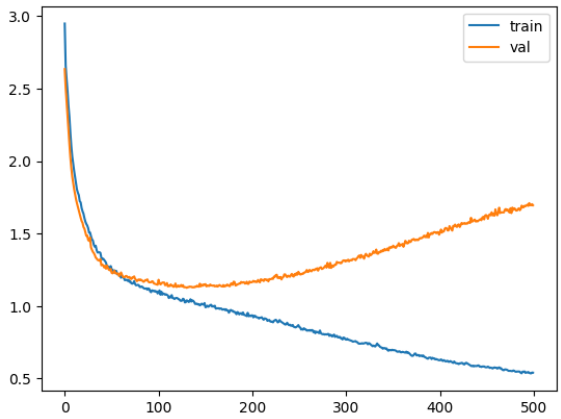


Figure 6: Training vs. Validation losses

During inference, the model generates each AU sequence until the `<EOS>` token is reached. Since these sequences vary in length, an auxiliary function was used to pad them to a fixed target length (`TARGET_T = 49`) so that all predictions share the same shape for evaluation and visualization.

3.2 Results and Visual Analysis

Figures 7–9 compare predicted and ground-truth (GT) trajectories. For AU01 and AU02, the model captures the main rises between frames 60–220 and short peaks near frame 400, indicating sustained eyebrow elevation followed by brief activations. Predicted curves are slightly noisier than GT, which leads to small threshold-crossing mismatches (spurious on/off around 0.5). This behaviour is consistent with (a) the discrete quantization used as target, (b) acoustic-only conditioning (using F_0), and (c) residual alignment noise. A simple post-processing (e.g., short median filter or hysteresis with a minimum active duration) reduces spurious flips while preserving true peaks.

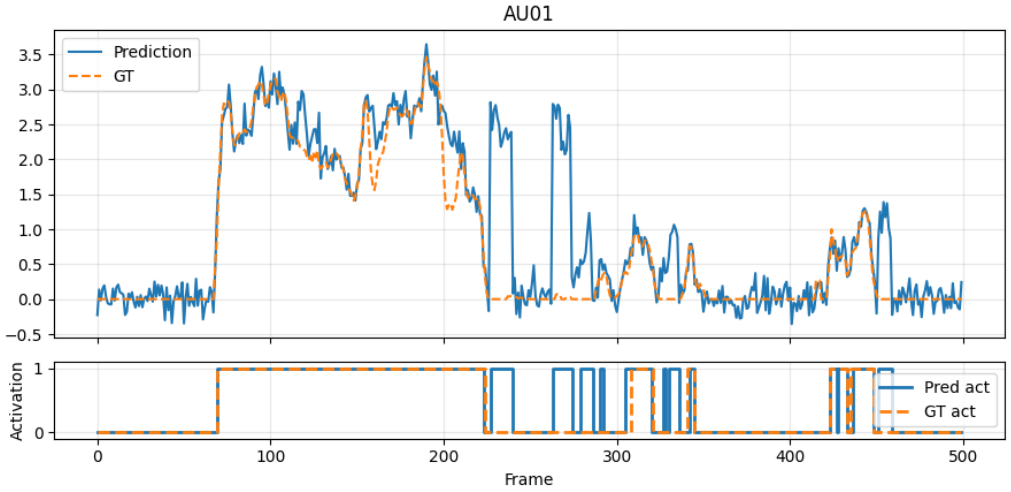


Figure 7: AU01 (inner brow raiser): Predicted vs. GT intensities (top) and binarized activations (bottom). The model tracks the long activation (frames 60–220), with mild high-frequency noise causing extra short activations.

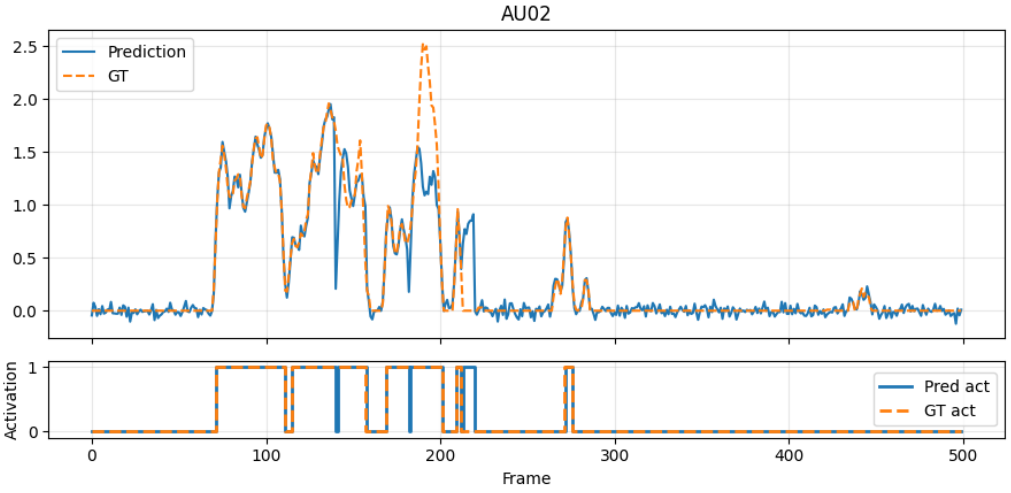


Figure 8: AU02 (outer brow raiser): Strong agreement with GT in both timing and amplitude; minor noise near the 0.5 threshold explains occasional activation mismatches.

3.3 Objective Evaluation of Predicted Action Units

The performance of the model on the test set was assessed using four quantitative metrics: Root Mean Squared Error (RMSE), Pearson Correlation Coefficient (PCC), Activity Hit Ratio (AHR), and Non-Activity Hit Ratio (NAHR).

Root Mean Squared Error (RMSE). RMSE measures the average magnitude of the prediction error between the predicted (p_t) and ground truth (g_t) AU intensities:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(p_t - g_t)^2}$$

Lower values indicate better precision in the prediction of AU intensity.

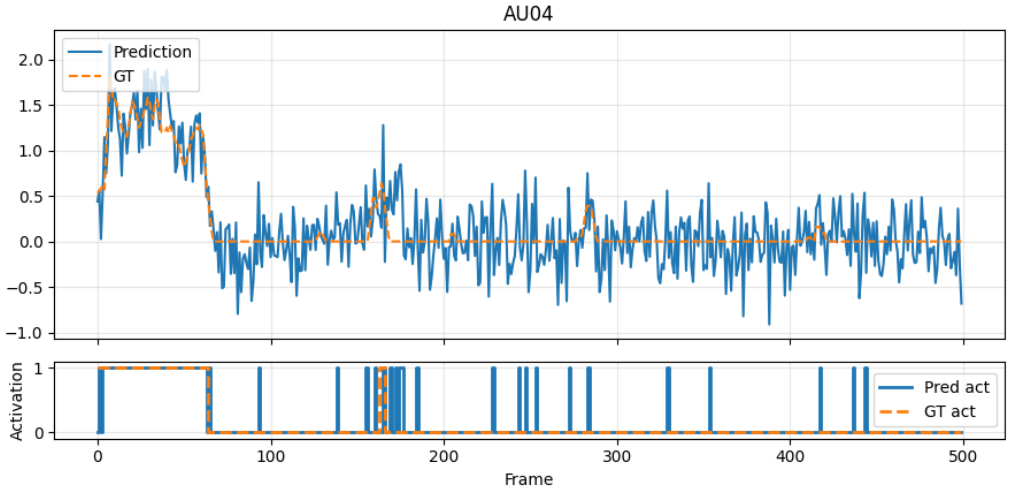


Figure 9: AU04 (brow lowerer): Higher activity at the beginning and sparse events later. The prediction is noisier and close to the decision boundary, yielding intermittent on/off around short segments.

Pearson Correlation Coefficient (PCC). PCC evaluates how well the predicted signal follows the shape and temporal dynamics of the ground truth:

$$\mathrm{PCC} = \frac{\sum_{t=1}^{N}(p_t - \bar{p})(g_t - \bar{g})}{\sqrt{\sum_{t=1}^{N}(p_t - \bar{p})^2}\sqrt{\sum_{t=1}^{N}(g_t - \bar{g})^2}}$$

where $\bar{p}$ and $\bar{g}$ represent the mean of the predicted and ground truth sequences. Values close to 1 denote strong positive correlation.

Activity Hit Ratio (AHR) and Non-Activity Hit Ratio (NAHR). After normalizing AU intensities to the range $[0, 1]$, a threshold of 0.5 is used to separate active and non-active frames. AHR and NAHR measure the model's accuracy in detecting both activated and non-activated states:

$$\mathrm{AHR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \qquad \mathrm{NAHR} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}$$

where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives respectively. Higher AHR and NAHR indicate better detection performance.

Results. Figure 10–12 summarize the objective results for AU01, AU02, and AU04 respectively. Each figure combines four plots: time-series comparison (GT vs. Pred), residual error, scatter correlation, and activation hit ratios.

Discussion. Overall, the results confirm that the model captures the temporal dynamics of upper facial expressions effectively from acoustic cues. AU02 achieves the best alignment and lowest error, suggesting that outer brow movements are most consistently correlated with prosodic features. AU01 exhibits slightly higher amplitude variance and occasional over-activation, while AU04 remains well predicted but with a noisier residual pattern. The consistently high PCC values across all AUs show that the model learns the correct temporal structure of facial motion, even if some small intensity biases persist.

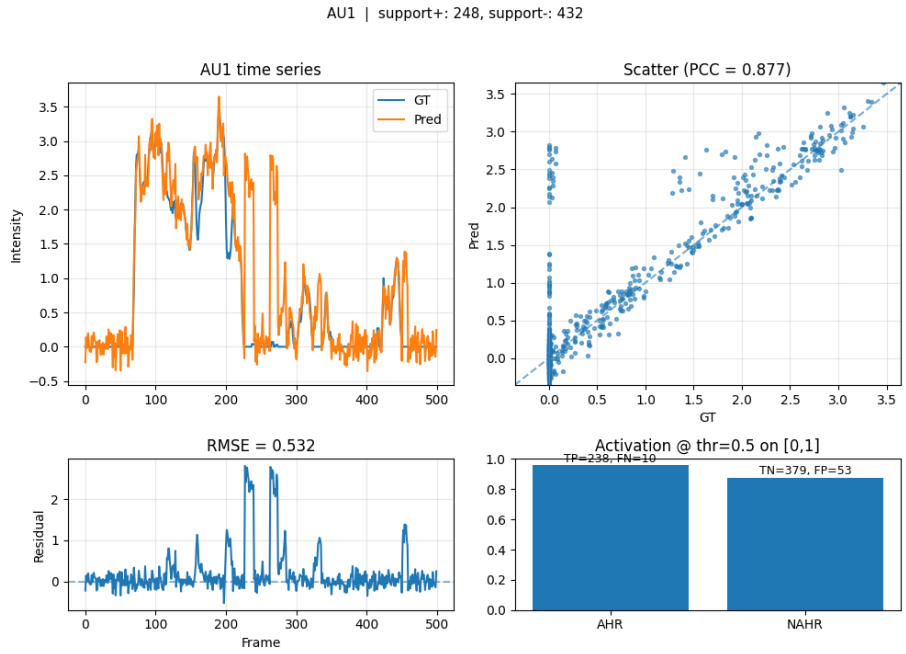


Figure 10: Evaluation results for AU01 (inner brow raiser). The predicted signal follows the overall temporal shape of the ground truth but tends to over-activate during neutral frames. RMSE = 0.53, PCC = 0.877, AHR/NAHR = 0.960 / 0.877.

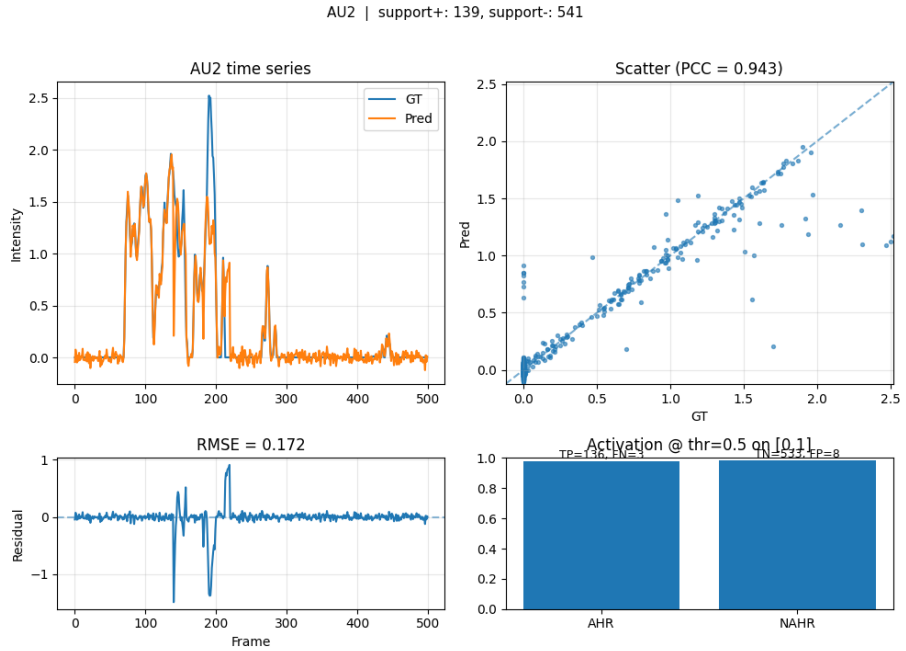


Figure 11: Evaluation results for AU02 (outer brow raiser). Very strong correlation with the ground truth (PCC = 0.943) and low RMSE = 0.17. The model detects nearly all activations (AHR = 0.978) with few false positives (NAHR = 0.985).

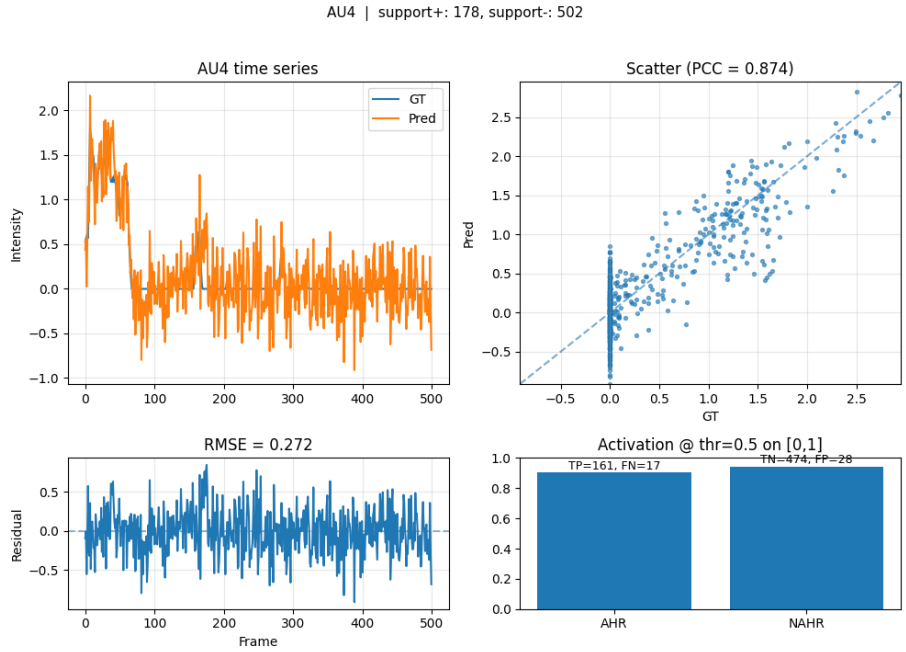


Figure 12: Evaluation results for AU04 (brow lowerer). Predictions are coherent and follow the trend of the ground truth with RMSE = 0.27 and PCC = 0.874. AHR/NAHR = 0.904 / 0.944 indicate balanced activation detection performance.

References

[1] Daniel Hirst and Céline De Looze. "Fundamental Frequency and Pitch". In: *The Cambridge Handbook of Phonetics*. Ed. by Rachael-Anne Knight and JaneEditors Setter. Cambridge Handbooks in Language and Linguistics. Cambridge University Press, 2021, pp. 336–361.

[2] Paul Ekman and Wallace V. Friesen. *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Palo Alto, CA: Consulting Psychologists Press, 1978.