Calculating people's attraction based on their facial expressions, body movement, voice tone and words spoken

Mitrea Ioan David

November 16, 2025

1 Data

1.1 Data Source

The dataset used in this research consists of publicly available YouTube videos showing first-time interactions between two people, primarily in speed-dating contexts. These videos typically contain short conversations (3–5 minutes) where participants decide whether they would like to go on a second date. This final decision serves as the ground-truth label (*positive attraction* = 1, *no attraction* = 0).

This approach is supported by previous research showing that real-time recordings of natural interactions can reliably capture social and emotional cues [1]. Furthermore, speed-dating scenarios have been used to study romantic interest and body-movement synchrony [2], as well as trait-rated attraction in virtual dating environments [3].

1.2 Modalities Collected

Following multimodal affective-computing research [4, 5, 6], four complementary modalities are extracted:

- **Facial expressions**: frame-level emotional and micro-expression features.
- **Body movements**: upper-body posture, gesture frequency, and motion energy.
- **Voice tone**: prosodic signals—pitch, intensity, vibrato, and spectral shape.
- **Spoken language**: utterance segmentation and semantic embeddings extracted from speech-to-text transcripts.

These modalities align with the theoretical view that attraction is multimodal, influenced simultaneously by appearance, vocal qualities, movement, and language [7, 8, 9].

1.3 Data Preprocessing

1.3.1 Video Preprocessing

Videos are sampled at fixed intervals (1 s, 2 s, and 5 s) to test different temporal granularities. Black frames—introduced by video editing—are removed from the *visual stream* but **audio is preserved**. This ensures that vocal features remain accessible even when visual frames are invalid.

Feature extraction follows affective computing best practices [5], and motion cues are computed analogously to speed-dating movement studies [2].

1.3.2 Audio Preprocessing

Audio is diarized into two speakers using pretrained separation models, following multimodal speech guidelines [10]. This yields:

- Speaker A prosody features
- Speaker B prosody features
- Speaker A transcript embeddings
- Speaker B transcript embeddings

This separation is crucial because the model computes attraction *per individual*, not jointly.

1.3.3 Text Processing

Speech is transcribed using a transformer-based ASR system, as recommended in recent ASR surveys [10]. Sentences are encoded using modern semantic embeddings to capture warmth, interest, and affective vocabulary.

1.4 Labels

The label for each interaction is extracted from the explicit verbal response of the participants ("yes/no" to a second date). These labels serve as ground truth for binary attraction classification.

Since no prior multimodal dataset explicitly measures attraction, our contribution constitutes the first dataset of its kind, bridging multimodal emotion recognition [4] and romantic-attraction analysis [9].

2 Methodology

2.1 Overview

The proposed approach is a multimodal sequential model that integrates facial expressions, body movement, vocal tone, and linguistic features to predict

attraction during first-time interactions. The model follows the principle that emotions and social signals evolve temporally and are influenced by previously expressed states [11, 12].

2.2 Model Architecture

2.2.1 Parallel Feature Extractors

Each modality is processed through a dedicated encoder operating at the same sampling rate:

- **Facial Expression Encoder**: a CNN-based feature extractor, following multimodal emotion-recognition architectures [4].
- **Body Movement Encoder**: a motion-vector and pose-estimation module using standard affective-computing visual pipelines [5].
- **Voice Tone Encoder**: an audio CNN or Transformer extracting prosodic features, following speech emotion research [10].
- **Language Encoder**: Transformer-based embeddings of the transcript, similar to multimodal sentiment-analysis frameworks [13].

2.2.2 Fusion and Alignment

All modality embeddings at time t are concatenated into a unified multimodal vector:

$$x_t = [x_t^{\text{face}}, x_t^{\text{body}}, x_t^{\text{voice}}, x_t^{\text{text}}].$$

Temporal alignment is performed using:

- strict timestamp matching (baseline),
- sliding-window alignment as proposed in SWRR [14],
- tensor fusion for cross-modal interaction [13].

This allows us to evaluate the sensitivity of the model to synchronization choices, as recommended in multimodal fusion surveys [6].

2.3 Sequential Modeling

A recurrent neural network (RNN, LSTM, or GRU) processes the multimodal sequence:

$$h_t = f(h_{t-1}, x_t),$$

where h_t represents the latent *interaction state* at time t, reflecting gradual emotional buildup, consistent with theories of situated emotion [11].

Following [12], an LSTM architecture is preferred due to its capacity to model long-term dependencies without vanishing gradients.

2.4 Output Layer

After the final frame T, the model produces a global attraction score:

$$\hat{y} = \sigma(W h_T + b),$$

classifying the interaction as "attracted" (1) or "not attracted" (0).

2.5 Validation and Evaluation

2.5.1 Comparison Baselines

Because no prior work explicitly predicts attraction from all four modalities, we compare our model with:

- unimodal baselines (face-only, body-only, voice-only, text-only),
- Tensor Fusion Network [13],
- sliding-window attention models [14].

This tests whether multimodal fusion provides measurable performance gains, as suggested by multimodal-attraction studies [7, 8].

2.5.2 Evaluation Metrics

- Accuracy
- Precision, recall, F1
- ROC-AUC
- Temporal consistency (smoothness of predictions)
- Cross-modal correlation scores

2.5.3 Experimental Variants

To validate robustness, we vary:

- **frame rate**: 1 s, 2 s, 5 s
- **sequence length**: full interaction vs. early-segment prediction
- **fusion strategies**: early vs. late fusion vs. tensor fusion
- **alignment methods**: strict vs. sliding window

2.6 Mathematical Model

For each modality $m \in \{face, \text{body}, \text{voice}, \text{text}\}$ and time step t:

$$x_t^m = E_m(I_t^m)$$

All modality embeddings are fused as:

$$x_t = x_t^{\text{face}} + x_t^{\text{body}} + x_t^{\text{voice}} + x_t^{\text{text}}$$

The RNN updates its hidden state via:

$$h_t = \text{LSTM}(x_t, h_{t-1})$$

Final attraction probability:

$$P(y = 1 \mid x_{1:T}) = \sigma(W h_T + b)$$

Loss function (binary cross-entropy):

$$\mathcal{L} = -\left[y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})\right]$$

This provides a complete mathematical formulation of the proposed experimental model.

References

[1] K. K. Haidet, J. Tate, D. Divirgilio-Thomas, A. Kolanowski, and M. B. Happ. Methods to improve reliability of video-recorded behavioral data. *Research in Nursing & Health*, 32(4):465–474, 2009.

[2] A. Chang, H. E. Kragness, W. Tsou, D. J. Bosnyak, A. Thiede, and L. J. Trainor. Body sway predicts romantic interest in speed dating. *Social Cognitive and Affective Neuroscience*, 16(1–2):185–192, 2021.

[3] J. E. French, L. J. Bolton, and A. L. Meltzer. Virtual speed dating: Utilizing online-meeting platforms to study initial attraction and relationship formation. *Personal Relationships*, 31(2):420–444, 2024.

[4] K. Kraack. A multimodal emotion recognition system: Integrating facial expressions, body movement, speech, and spoken language. *arXiv preprint*, 2023.

[5] A. Sherstinsky. Facial expression and body gesture emotion recognition: A systematic review on the use of visual data in affective computing. *IEEE Transactions on Affective Computing*, 13(4):1648–1670, 2022.

[6] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. 2017.

[7] Megan Nicole Williams and Coren Lee Apicella. A test of multimodal communication in humans using 881 judgements of men and women's physical, vocal, and olfactory attractiveness. *Heliyon*, 9(6):e16895, 2023.

[8] A. Schirmer, M. Franz, L. Krismann, V. Nöring, M. Große, M. Mahmut, and I. Croy. Attraction in every sense: How looks, voice, movement and scent draw us to future lovers and friends. *British Journal of Psychology*, 116(3):684–701, 2025.

[9] Theresa DiDonato and Brett Jakubiak. Romantic attraction. In *The Science of Romantic Relationships*, pages 123–157. Cambridge University Press, 2023.

[10] H. Ahlawat, N. Aggarwal, and D. Gupta. Automatic speech recognition: A survey of deep learning techniques and approaches. *ACM Computing Surveys*, 55(7):1–34, 2022.

[11] L. A. M. Lebois, C. D. Wilson-Mendenhall, W. K. Simmons, L. F. Barrett, and L. W. Barsalou. Learning situated emotions. *Neuropsychologia*, 145:106637, 2018.

[12] R. M. Schmidt. Recurrent neural networks (rnns): A gentle introduction and overview. *arXiv preprint arXiv:1912.05911*, 2019.

[13] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. 2017.

[14] Z. Zhao, T. Gao, H. Wang, and B. W. Schuller. Swrr: Feature map classifier based on sliding window attention and high-response feature reuse for multimodal emotion recognition. pages 2433–2437, 2023.