Challenges and Solutions in the Charting of Dynamic Emotion Evolution

The Second International Conference on AI-based Systems and Services
AISyS 2025, September 28, 2025 to October 02, 2025 - Lisbon, Portugal

Clement H.C. Leung
Chinese University of Hong Kong, Shenzhen
clementleung@cuhk.edu.cn

香港中文大學(深圳)
The Chinese University of Hong Kong, Shenzhen

Clement LEUNG

- FULL PROFESSORSHIPS at
  - University of London, UK; National University of Singapore; Chinese University of Hong Kong, Shenzhen, China; Hong Kong Baptist University; Victoria University, Australia
- Two US patents, five books and over 150 research articles
- Program Chair, Keynote Speaker, Panel Expert of major International Conferences
- Editorial Board of ten International Journals
- Listed in Who's Who in the World and Great Minds of the 21st Century
- Fellow of the International Academy, Research, and Industry Association, Fellow of the British Computer Society, Fellow of the Royal Society of Arts, Chartered Engineer

Different Forms of Intelligence

Computation Intelligence

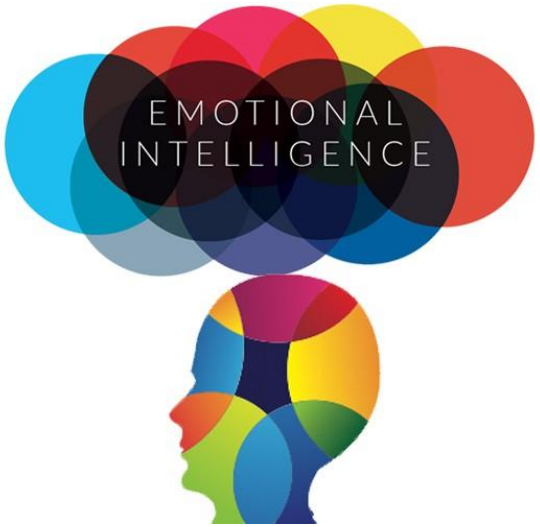
(high performance computation)

Perception Intelligence

(face recognition, speech recognition)

Cognition Intelligence

(emotion recognition, text sentiment)

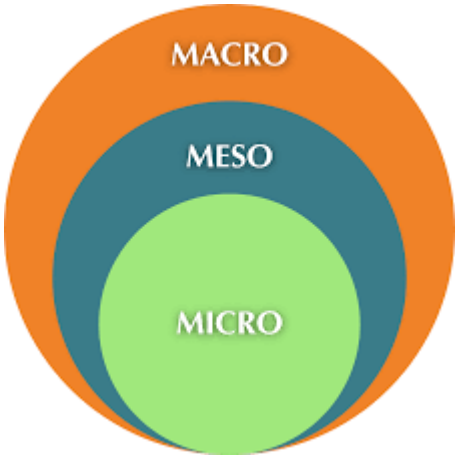


Emotion State Prediction

- Safety-critical domains require stable emotional states
  - pilots, surgeons, drivers
- Dynamic emotions affect wellbeing, UX, de-escalation, and risk
- New ingredients: LLM zero-shot multimodal reasoning
  - probabilistic macro-dynamics (Semi-Markov Models)
- We need models that read signals and forecast trajectories under interventions

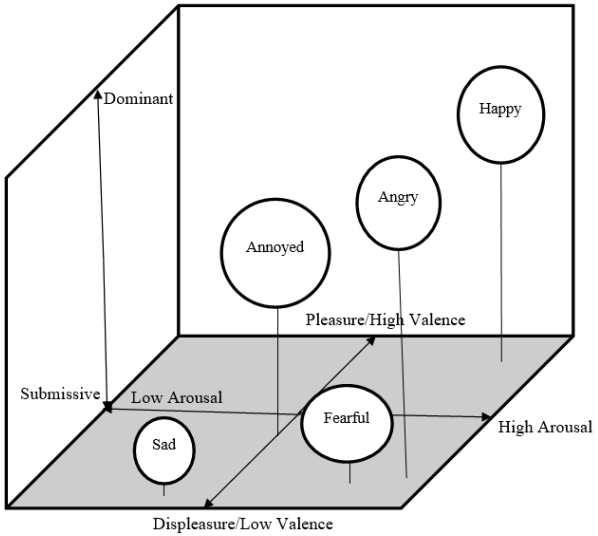
Dynamic Emotion Changes

- Trajectories of emotions over time
- Time scales
  - micro-level
    - ms–s
  - meso-level
    - turns/segments
  - macro-level
    - sessions/days

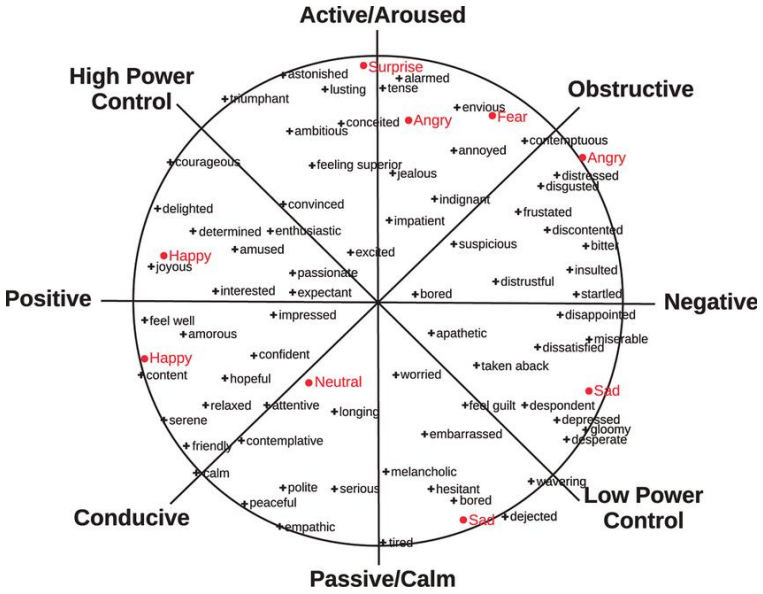


Discrete and Dimensional Emotion Models

- **Emotion Theory**
  - Discrete Emotion (Six basic emotions, or Nine basic emotions, OCC model)
  - Dimensional Emotion (Arousal-Valence, Pleasure-Arousal-Dominance, Circumplex)



Burden, David, and Maggi Savin-Baden. *Virtual humans: Today and tomorrow*. Chapman and Hall/CRC, 2019.



Scherer's circumplex model (Scherer, 2005)

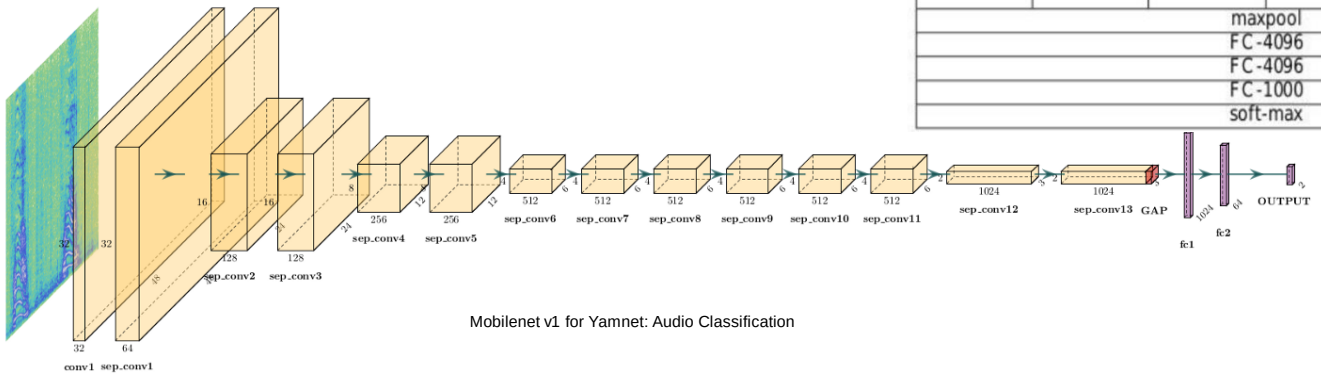
Emotion Recognition Using Neural Nets

- Emotion Recognition from single modality (e.g., Speech, Text)
  - Speech Emotion Recognition (CNN, LSTM, CNN-LSTM, BERT, etc.)
  - Text Sentiment (LSTM, BERT, universal language model fine-tuning,

VGG-19 Network

Figure from 'Very Deep Convolutional Networks For Large-Scale Image Recognition"

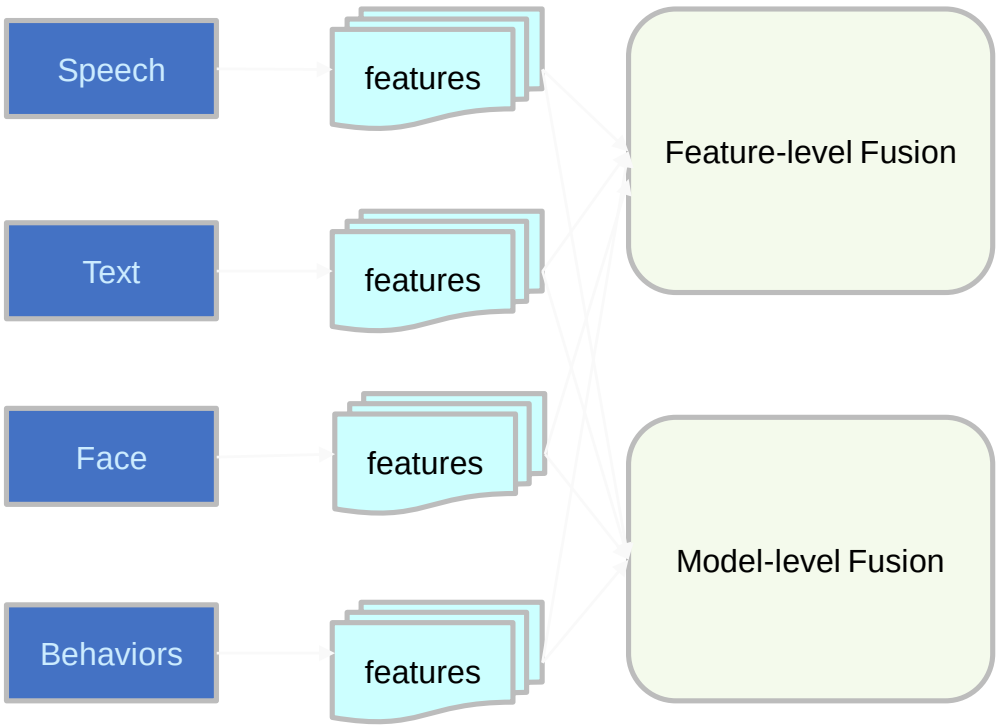
ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
input (224 × 224 RGB image)					
conv3-64	conv3-64 LRN	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 conv1-256	conv3-256 conv3-256	conv3-256 conv3-256 conv3-256 conv3-256
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 conv1-512	conv3-512 conv3-512	conv3-512 conv3-512 conv3-512 conv3-512
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 conv1-512	conv3-512 conv3-512	conv3-512 conv3-512 conv3-512 conv3-512
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					



Mobilenet v1 for Yamnet: Audio Classification

Multimodal Emotions

- Joint Representation
- Multimodal fusion
- Co-learning
- Generative adversarial network



Feature Extraction from Speech Audio

To consider the temporal characteristic, a large speech collection is represented by

$$X = \{x_1, x_2, x_3, \ldots, x_N\},$$

where $x_i \in F \times T, F \text{ and } T \text{ denote for dimensionality of spectrogram}$

The goal is to learn an embedding $g := r^{F \times T} \rightarrow r^d$, such that

$$\left|\left|\boldsymbol{g}(x_i) - \boldsymbol{g}(x_j)\right|\right| <= ||\boldsymbol{g}(x_i) - \boldsymbol{g}(x_k)||, \text{ when } |i - j| < = |i - k|$$

We learn a triplet loss function

$$\Theta(z) = \sum_{i=1}^{N} [\|g(x_i) - g(x_j)\|_2^2 + \|g(x_i) - g(x_k)\|_2^2 + \delta]$$

Where $\delta \; is \; non - negative \; margin \; hyperparameter$

Text Embedding Extraction

Given a sequence of text extracted from speech

(1) Map the tokens of the initial input sequence to an embedding space

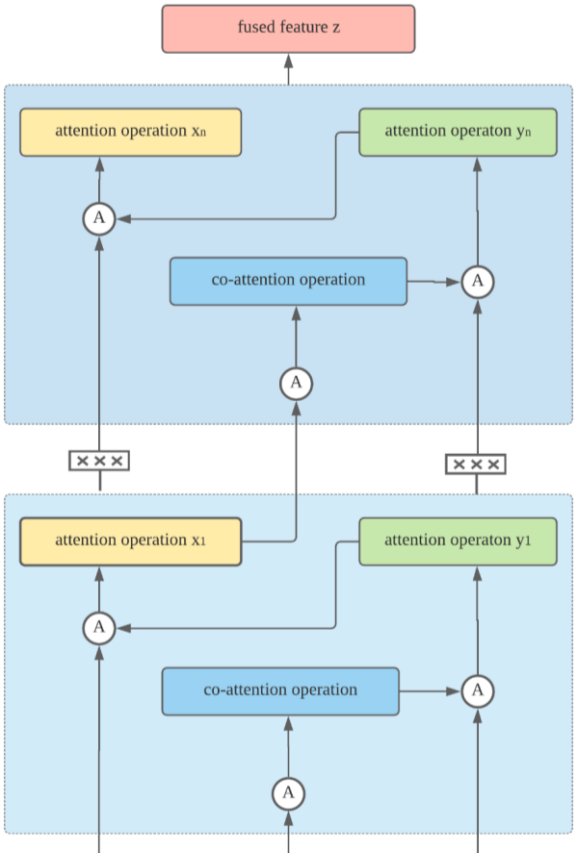
(2) Input the embedded sequence to the encoder layer

(a) The encoder layer is composed of a stack of blocks

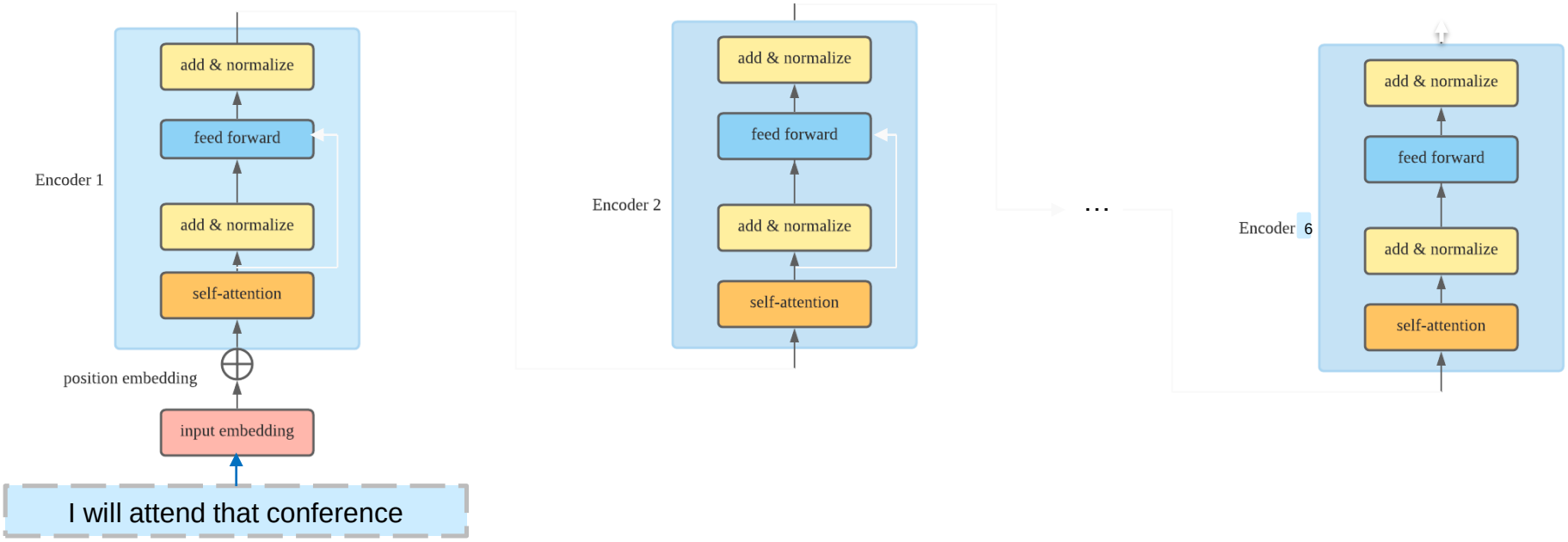
(b) Each block contains self-attention followed by feed-forward

(c) Residual skip from self-attention layer and feed-forward

(d) Dropout within the feed-forward network

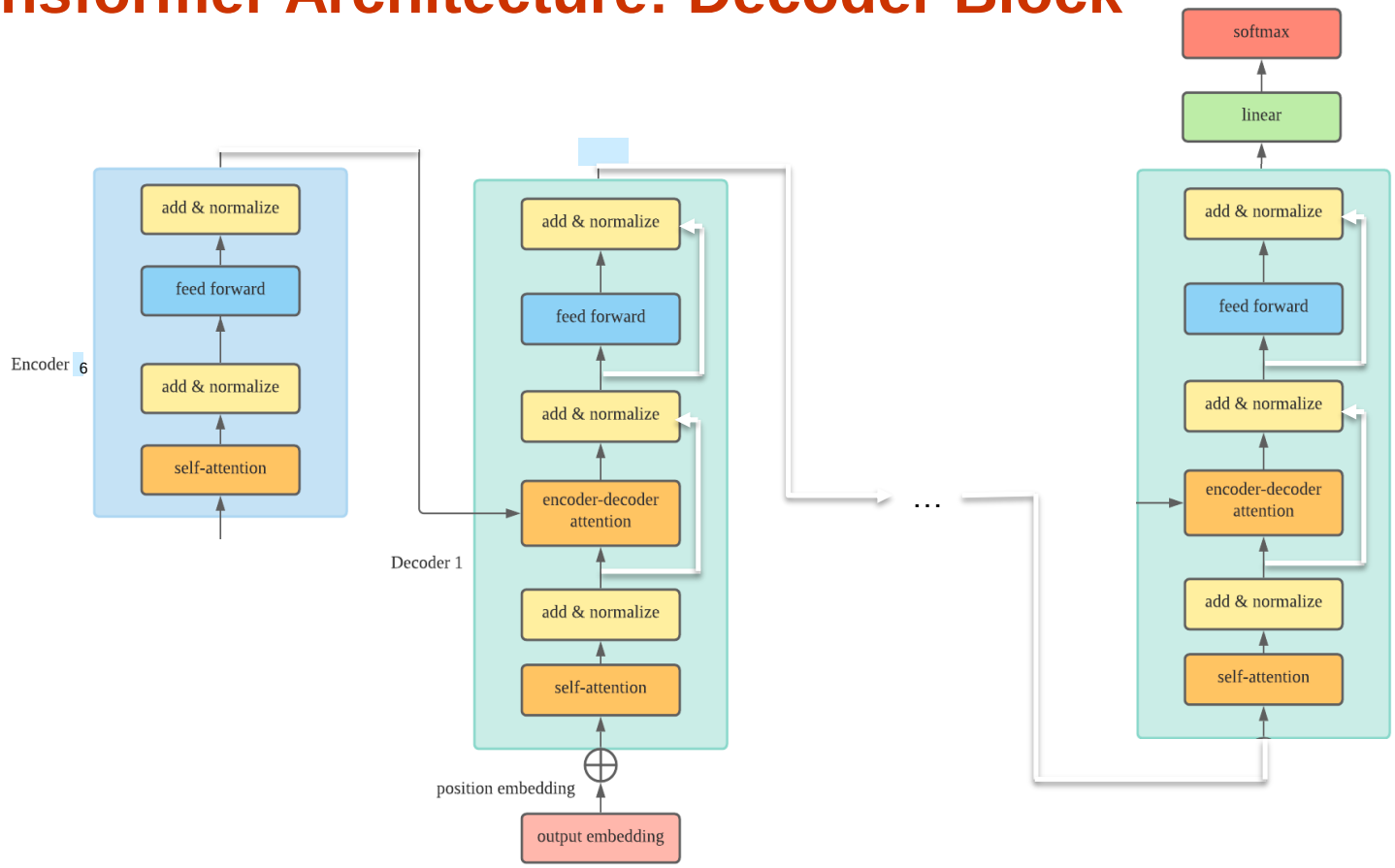


Transformer Architecture: Encoder Block



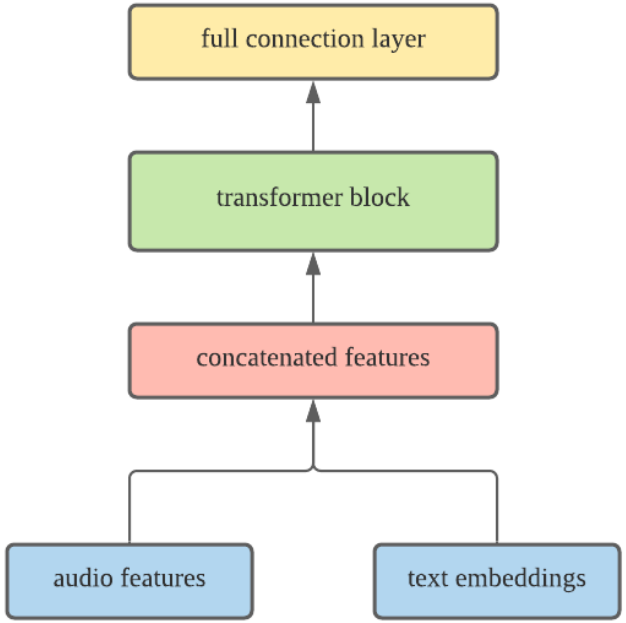
- The encoder block uses the self-attention mechanism
- Residual connections is added, and layer normalization LayerNorm(x + Sublayer(x))
- Encoder is stacked, and the output of last encoder block is the input of decoder

Transformer Architecture: Decoder Block

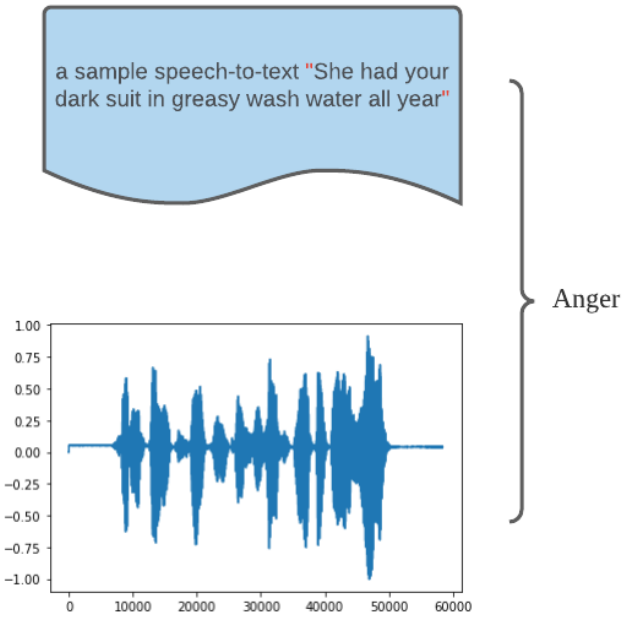


- The decoder block uses encoder-decoder attention
- The embeddings use features from input and partial output

Joint Emotion Representation



Multimodal fusion of speech audio and text on embeddings extracted from transfer learning



Sample speech with anger emotion

Emotion Categories

- FER-2013 representation of emotions
  - 0=Angry
  - 1=Disgust
  - 2=Fear
  - 3=Happy
  - 4=Sad
  - 5=Surprise
  - 6=Neutral

Our Modified Categories

- We augment the categories to eight
  - Joy
  - Trust
  - Surprise
  - Anticipation
  - Sadness
  - Disgust
  - Anger
  - Fear

State Reduction

- To facilitate practical application - e.g. duty roster compilation - the eight states can be reduced to two operational categories
- Positive (proceed)
  - Joy
  - Trust
  - Surprise
  - Anticipation

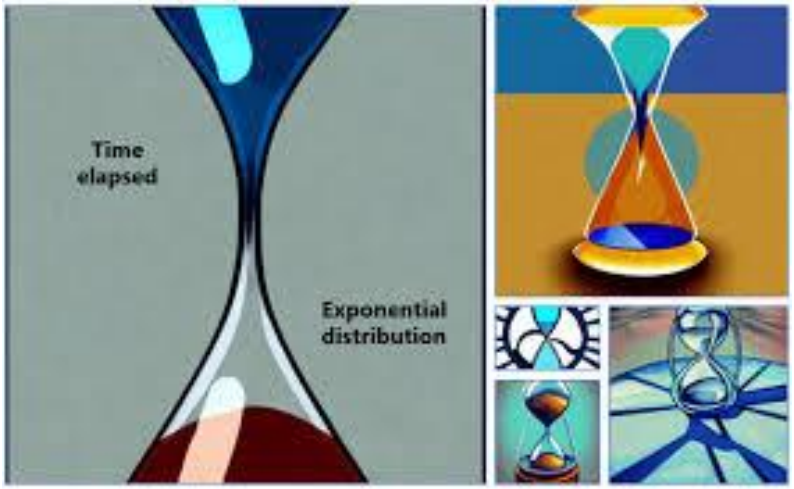
- Negative (denied)
  - Sadness
  - Disgust
  - Anger
  - Fear

Emotion Stability

- Emotion stability measures the propensity of changes in emotion states
- Parameter λ controls persistence/volatility of states
  - Small λ indicates high stability in emotions which tends to persist
  - Large λ means highly volatile emotions, with emotion fluctuations occurring frequently

Semi-Markov Model

- Emotion changes in random time intervals
- Emotion states persists in accordance with a duration govern by probability
  - Exponentially distributed length of interval
- Semi-Markov Model with independent identical distribution for all states between transitions



Semi-Markov Model

- When emotion changes occur, the state transitions follow a Markov Chain (Embedded Markov Chain)
- The duration of staying in a particular emotion state is exponentially distributed
  - A Semi-Markov model can be likened to a frog on a lily pond who jumps from lily pad to lily pad (different states - memoryless) after some random duration

Embedded Markov Chain

The transition probability of the Embedded Markov Chain is an 8x8 matrix

$$P = \begin{bmatrix} p_{11} & p_{12} & p_{13} & p_{14} & p_{15} & p_{16} & p_{17} & p_{18} \\ p_{21} & p_{22} & p_{23} & p_{24} & p_{25} & p_{26} & p_{27} & p_{28} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ p_{81} & p_{82} & p_{83} & p_{84} & p_{85} & p_{86} & p_{87} & p_{88} \end{bmatrix}$$

Embedded Markov Chain

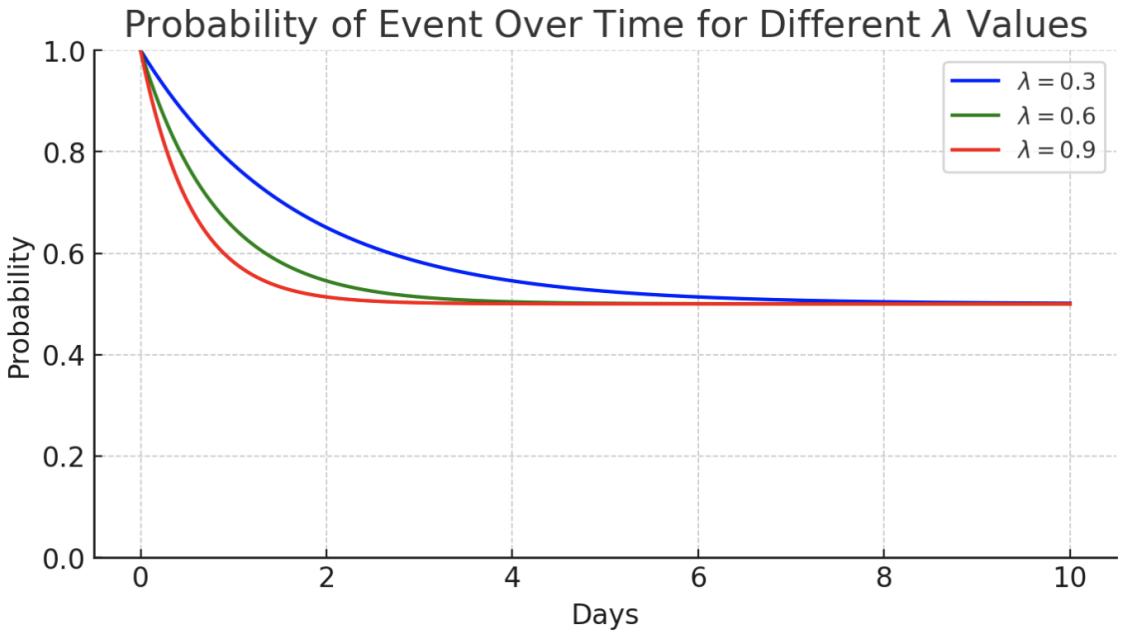
For example, when the emotion is Joy (E_1), the transition probabilities are:

$$p_{1j} = \begin{cases} 0.5, & \text{if } j = 1 \text{ (remain in Joy)} \\ 0.15, & \text{if } j = 2 \text{ (transition to Trust)} \\ 0.15, & \text{if } j = 3 \text{ (transition to Surprise)} \\ 0.1, & \text{if } j = 4 \text{ (transition to Anticipation)} \\ 0.05, & \text{if } j = 5 \text{ (transition to Sadness)} \\ 0.02, & \text{if } j = 6 \text{ (transition to Disgust)} \\ 0.02, & \text{if } j = 7 \text{ (transition to Anger)} \\ 0.01, & \text{if } j = 8 \text{ (transition to Fear)} \end{cases}$$

Emotion Dynamic Changes

- From Positive State to Positive State

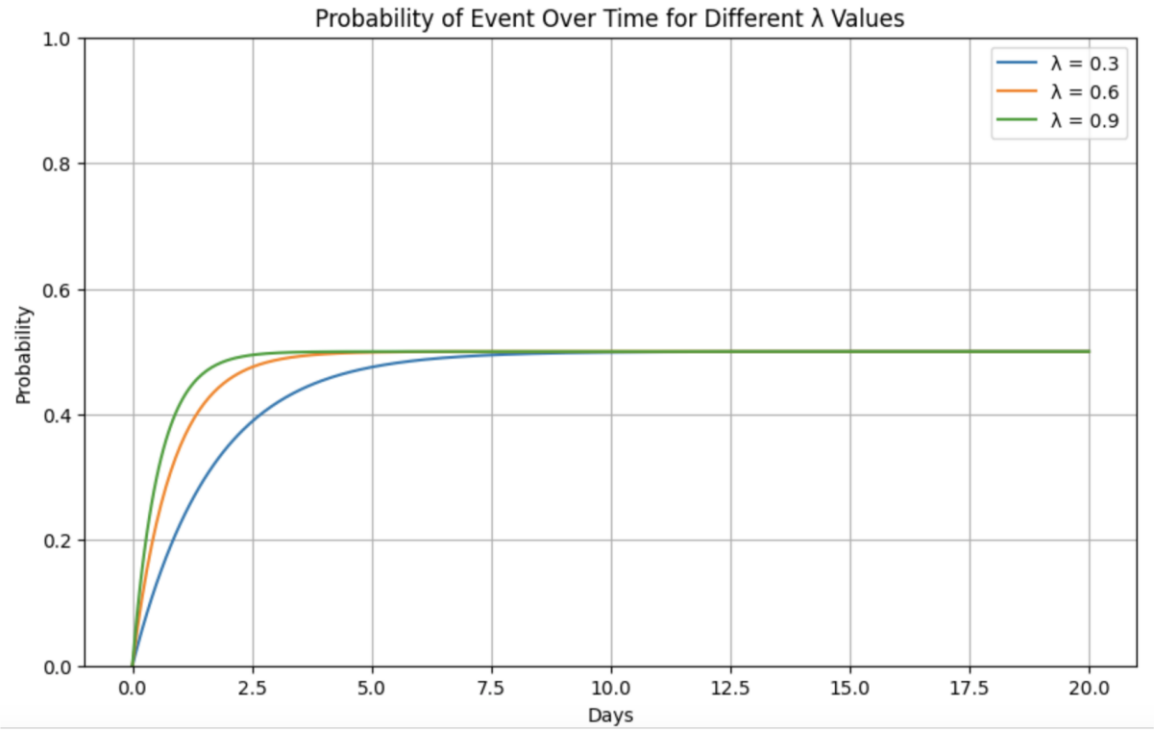
$$P[S(t) = 1|S(0) = 1] = e^{-\lambda t}[1 + \frac{(\lambda t)^2}{2!} + \frac{(\lambda t)^4}{4!} ... + ...]$$
$$= e^{-\lambda t} \cosh \lambda t$$



Emotion Dynamic Changes

- From Positive State to Negative State

$$P[S(t) = -1|S(0) = 1]] = e^{-\lambda t}[1 + \frac{(\lambda t)^3}{3!} + \frac{(\lambda t)^5}{5!} ... + ...]$$
$$= e^{-\lambda t} \sinh \lambda t$$



Emotion State Transition Dynamics

The probability of emotion E_i remaining the same at time t is:

$$P_{\text{stay},E_i}(t) = P_{E_i}(t) \cdot e^{-\lambda_i t} \cosh(\lambda_i t)$$

The probability of emotion E_i transitioning is:

$$P_{\text{trans},E_i}(t) = P_{E_i}(t) \cdot [1 - e^{-\lambda_i t} \cosh(\lambda_i t)]$$

The adjusted emotion probability is:

$$\tilde{P}_{E_i}(t) = P_{\text{stay},E_i}(t) + \sum_{j \neq i} P_{\text{trans},E_j}(t) \cdot p_{ji}$$

where p_{ji} is the probability of transitioning from emotion E_j to emotion E_i.

Challenge Map

- Ground truth over time
  - subjectivity, lag, annotator drift
  - dataset guidelines vs human perception
- Temporal granularity
  - static images vs conversational dynamics
  - need mapping across micro→meso→macro
- Cross-modal asynchrony
  - text/audio/video/physiology peak at different times
- Personalization/culture
  - expression of disgust/neutral/surprise varied
- Non-stationarity
  - λ and P may change across contexts and time


Thank
you!