Multimodal Emotion Detection with Transfer Learning and State Space Model

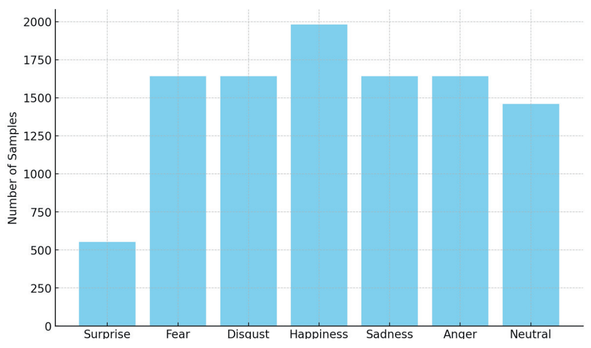
Zanxu Wang[1], Homayoon Beigi[1][2]
Columbia University[1], NY, Recognition Technologies, Inc.[2], NY
zw2864@columbia.edu, beigi@recotechnologies.com

Overview

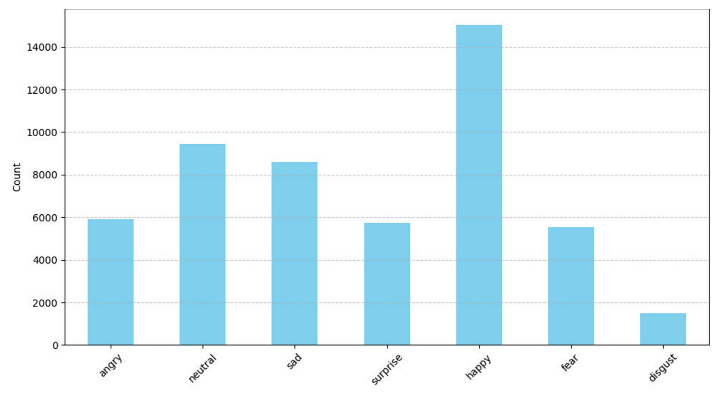
- A **multi-stage hierarchical approach** to multi-modal emotion recognition in conversational contexts.
- leveraged features from **speaker recognition, speech recognition, face recognition**, and a **sentence transformer**.
- Integrated diverse unimodal datasets across text, audio, and facial expressions.
- Fusion methods with **Mamba** based state space models.
- Achieve a promising **64.40%** weighted accuracy on MELD with speaker, face, text.

Unimodal Datasets

- **Audio**: Crema-D, RAVDESS, SAVEE, TESS



- **Face**: JAFFE，CK+, RAF-DB，FER2013



- **Text**: Balanced sampling from 5 sources with each emotion 6.2k utterances.

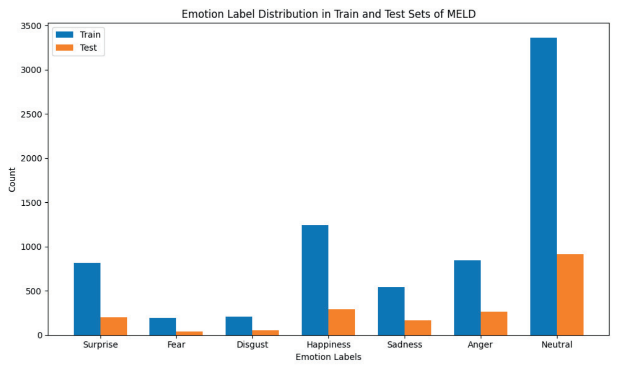
Name	anger	disgust	fear	joy	neutral	sadness	surprise
Crowdflower (2016)	Yes	-	-	Yes	Yes	Yes	Yes
Emotion Dataset, Elvis et al. (2018)	Yes	-	Yes	Yes	-	Yes	Yes
GoEmotions, Demszky et al. (2020)	Yes	Yes	Yes	Yes	Yes	Yes	Yes
ISEAR, Vikash (2018)	Yes	Yes	Yes	Yes	-	Yes	-
SemEval-2018, EI-reg, Mohammad et al. (2018)	Yes	-	Yes	Yes	-	Yes	-

Motivation

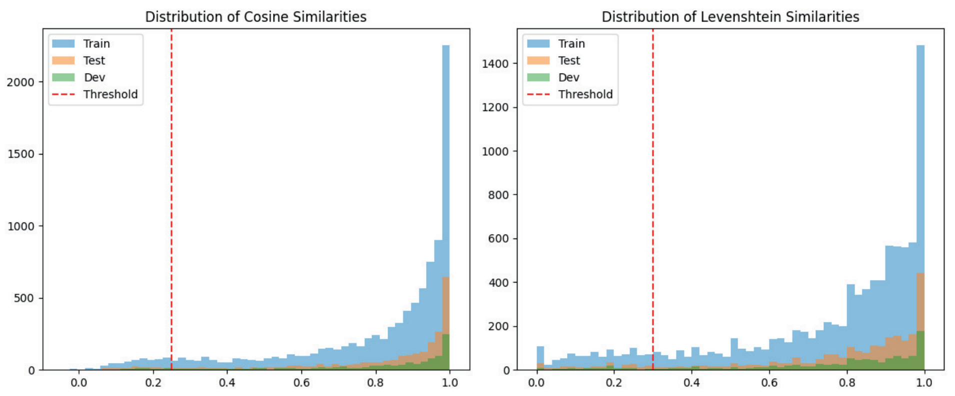
- **Multi-modal approach**: Unified emotion labels across all modalities
- **Unimodal datasets**: Comprehensive emotion foundation per modality
- **Speaker & Speech features**: 352 speakers, complementary acoustic cues
- **Fusion Strategy**: Inspired by I-vectors in ASR
- **Mamba Block**: Linear time complexity for efficient processing

Mulimodal Dataset: MELD

Audio: Extract audio from multiple channels among videos.



Text: Re-transcribe audio with Whisper, filter out misaligned utterance through 2 metrics.



Importance of Multimodal Cues*

Utterance: *"Become a drama critic!"*
Emotion: *Joy* **Sentiment:** *Positive*

Text	Audio	Visual
Ambiguous	Joyous tone	Smiling Face

Utterance: *"Great, now he is waving back"*
Emotion: *Disgust* **Sentiment:** *Negative*

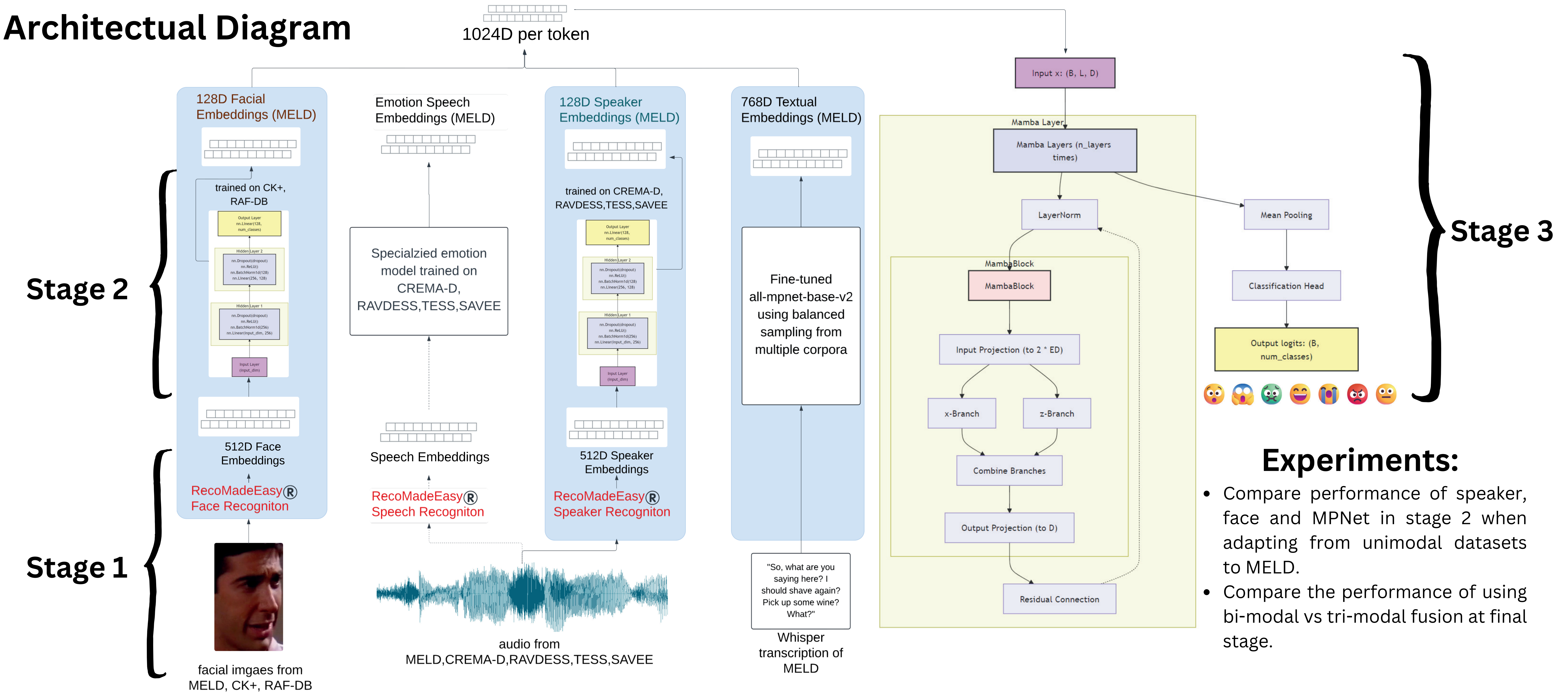
Text	Audio	Visual
Positive/Joy	Flat tone	Frown

**Figure reproduced from the MELD paper.*

Face Processing:

- Relabel videos for 352 distinct speakers.
- Sample utterances per speaker.
- Extract faces using YOLO-v8 @1 fps.
- Manual inspection to filter out incorrect detections
- Generate speaker embeddings with FaceNet.
- Extract faces for all videos using segment timestamps.
- Identify speaking individual by comparing extracted faces to speaker embeddings.

Architectual Diagram


1024D per token
128D Facial Embeddings (MELD)
trained on CK+, RAF-DB
Output Layer nn.Linear(128, num_classes)
Hidden Layer 2
nn.Dropout(dropout) nn.ReLU() nn.BatchNorm1d(128) nn.Linear(256, 128)
Hidden Layer 1
nn.Dropout(dropout) nn.ReLU() nn.BatchNorm1d(256) nn.Linear(input_dim, 256)
Input Layer (input_dim)
512D Face Embeddings
RecoMadeEasy® Face Recogniton
facial imgaes from MELD, CK+, RAF-DB
Emotion Speech Embeddings (MELD)
Specialzied emotion model trained on CREMA-D, RAVDESS,TESS,SAVEE
Speech Embeddings
RecoMadeEasy® Speech Recogniton
audio from MELD,CREMA-D,RAVDESS,TESS,SAVEE
128D Speaker Embeddings (MELD)
trained on CREMA-D, RAVDESS,TESS,SAVEE
Output Layer nn.Linear(128, num_classes)
Hidden Layer 2
nn.Dropout(dropout) nn.ReLU() nn.BatchNorm1d(128) nn.Linear(256, 128)
Hidden Layer 1
nn.Dropout(dropout) nn.ReLU() nn.BatchNorm1d(256) nn.Linear(input_dim, 256)
Input Layer (input_dim)
512D Speaker Embeddings
RecoMadeEasy® Speaker Recogniton
768D Textual Embeddings (MELD)
Fine-tuned all-mpnet-base-v2 using balanced sampling from multiple corpora
"So, what are you saying here? I should shave again? Pick up some wine? What?"
Whisper transcription of MELD
Input x: (B, L, D)
Mamba Layer
Mamba Layers (n_layers times)
LayerNorm
MambaBlock
MambaBlock
Input Projection (to 2 * ED)
x-Branch
z-Branch
Combine Branches
Output Projection (to D)
Residual Connection
Mean Pooling
Classification Head
Output logits: (B, num_classes)
Stage 1
Stage 2
Stage 3

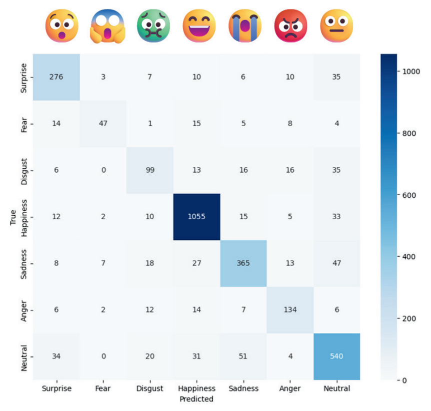
Experiments:

- Compare performance of speaker, face and MPNet in stage 2 when adapting from unimodal datasets to MELD.
- Compare the performance of using bi-modal vs tri-modal fusion at final stage.

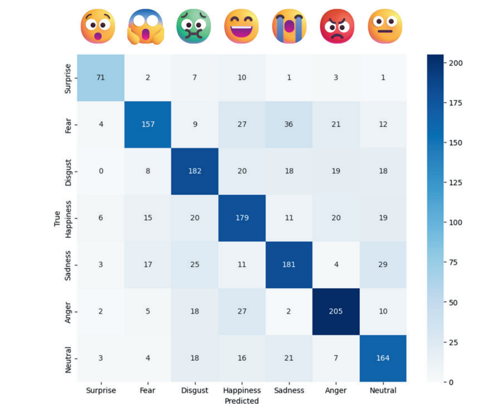
Results

Stage 2 Performance of Face, Speaker and Text model on unimodal datasets

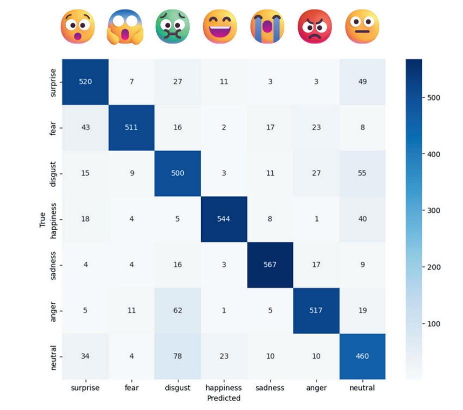
Face MLP
Trained on CK+, RAF-DB(train)
Test on CK+, RAF-DB(test) ACC: 81%



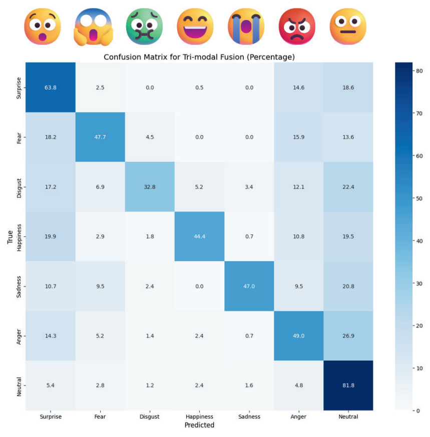
Speaker MLP
Trained on CREMA-D, TESS...(train)
Test on CREMA-D, TESS...(test) ACC: 68%



MPNet V2
Trained on ISEAR, Crowdflower...(train)
Test on ISEAR, Crowdflower...(test), ACC: 83%

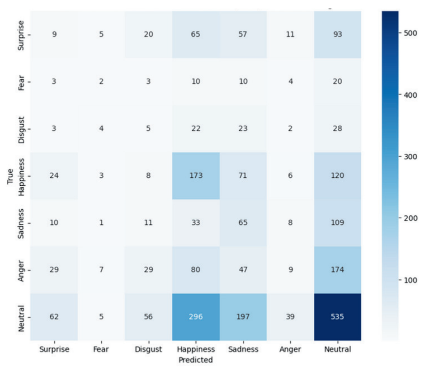


Final Stage Fusion with SSM:
Trained on MELD(train)
Tested on MELD(test), ACC:64%

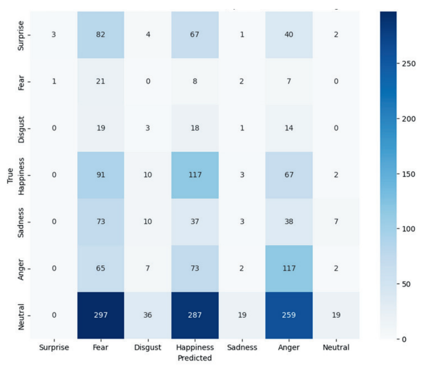


Stage 2 Inferance of Face, Speaker and Text model on MELD_test

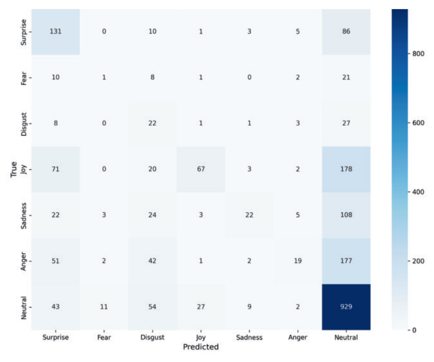
Trained on CK+, RAF-DB(train)
Test on MELD(test), ACC: 31%

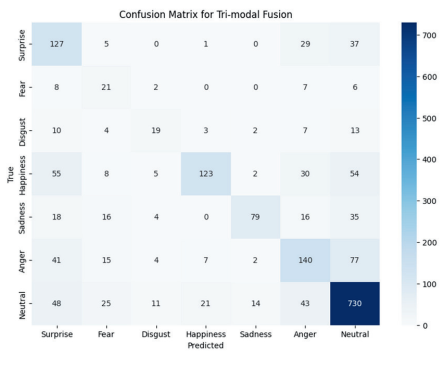


Trained on CREMA-D, TESS...(train)
Test on MELD(test), ACC: 18%



Trained on ISEAR, Crowdflower...(train)
Test on MELD(test), ACC: 55%





Final Stage Bi-modal vs Tri-modal Fusion Accuracy train on MELD_train, test on MELD_test

Metric	Text+Speaker	Text+Face	Text+Speaker+Face
Accuracy	60.21%	61.48%	64.40%
Weighted Avg F1 Score	60.33%	61.62%	64.53%

Future Work

Replace MLP in stage 2 with Kolmogorov-Arnold Networks:

- Utilize b-spline activation function for enhanced non-linearity
- Aim to capture more complex emotional patterns

Incorporate speech recognition model features:

- Add sequential information to complement speaker recognition
- Explore Mamba architecture in stage 2

Integrate IEMOCAP dataset:

- Investigate domain adaptation between MELD and IEMOCAP