SEOUL NATIONAL UNIVERSITY
Department of Electrical and Computer Engineering

Multimodal Speech Emotion Recognition Using Audio and Text

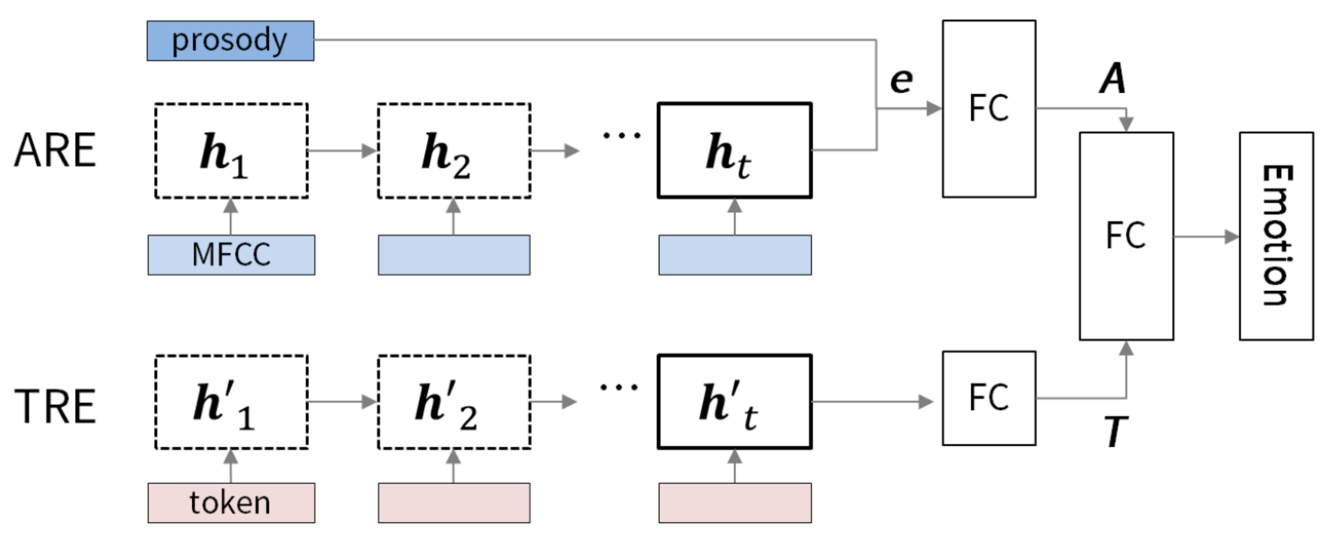
Seunghyun Yoon, Seokhyun Byun, and Kyomin Jung

Goal of our Research

- We propose a novel deep dual recurrent encoder model that utilizes text data and audio signals simultaneously to obtain a better understanding of speech data.
- Our proposed model outperforms previous state-of-the-art methods in assigning data to one of four emotion categories (i.e., angry, happy, sad and neutral) as reflected by accuracies ranging from 68.8% to 71.8%.

Model

- **Multimodal Dual Recurrent Encoder (MDRE)**: The upper part shows the audio recurrent encoder (**ARE**), which encodes audio signals, and the lower part shows the text recurrent encoder (**TRE**), which encodes textual information.



$$\mathbf{h}_t = \mathrm{GRU}\,(\mathbf{h}_{t-1}, \mathbf{x}_t)$$

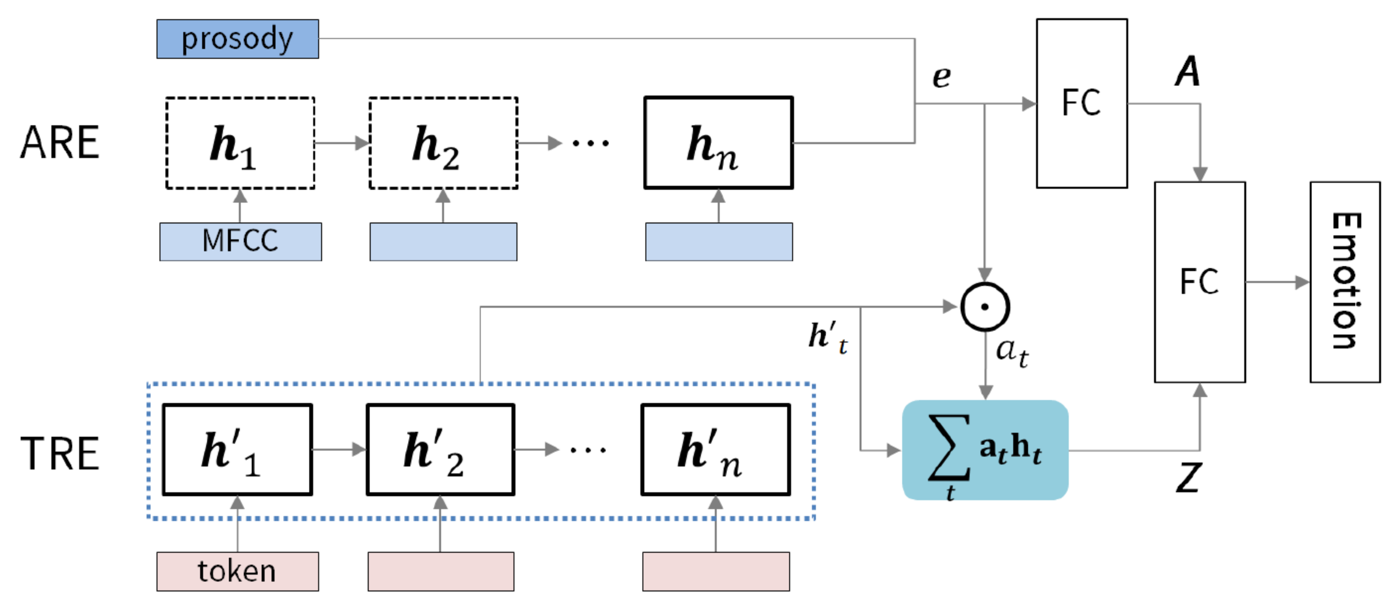
$$\mathbf{A} = \mathrm{FullyConnected}\,(\mathbf{e}), \qquad \mathbf{T} = \mathrm{FullyConnected}\,(\mathbf{h}'_{last})$$

$$\hat{y}_i = \mathrm{softmax}\,(\mathrm{concat}\,(\mathbf{A}, \mathbf{T})^{\intercal}\mathbf{M} + \mathbf{b})$$

$$\mathcal{L} = -\log \prod_{i=1}^{N} \sum_{c=1}^{C} y_{i,c} \log\,(\hat{y}_{i,c})$$

- **Audio Features:** A total 39 MFCC feature set that includes 12 MFCC parameters (1-12) from the 26 Mel-frequency bands and log-energy parameters, 13 delta and 13 acceleration coefficients. The frame size is set to 25 ms at a rate of 10 ms. The prosodic features are composed of 35 features (F0, voicing probability, loudness contours) extracted using the OpenSMILE toolkit.
- **Text Features:** The vocabulary size of the dataset is 3,747. The word-embedding matrix is initialized from the Glove and fine-tuned while training.

- **Multimodal Dual Recurrent Encoder with Attention (MDREA):** The weighted sum of the sequence of the hidden states of the **TRE**, $\mathbf{h}'_t$, is taken using the attention weight a_t; a_t is calculated as the dot product of the final encoding vector of the **ARE**, $\mathbf{e}$, and $\mathbf{h}'_t$.



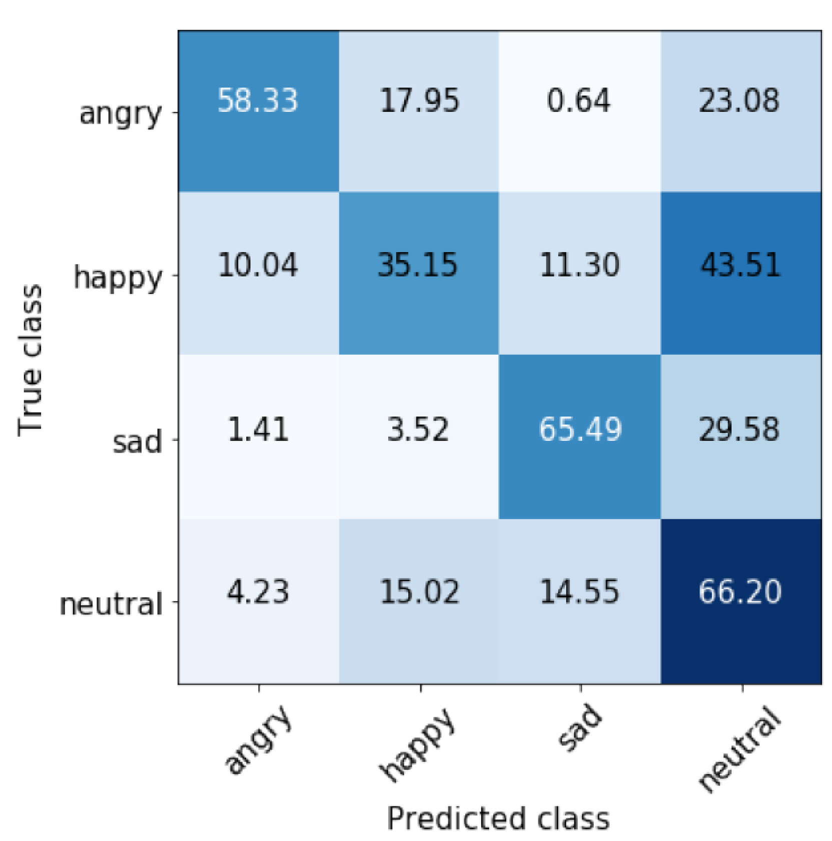
$$\mathrm{a}_t = \frac{\exp(\mathbf{e}^{\intercal}\mathbf{h}'_t)}{\sum_t \exp(\mathbf{e}^{\intercal}\mathbf{h}'_t)}, \quad \mathbf{Z} = \sum_t \mathrm{a}_t\, \mathbf{h}'_t, \quad \hat{y}_i = \mathrm{softmax}\,(\mathrm{concat}\,(\mathbf{Z}, \mathbf{A})^{\intercal}\mathbf{M} + \mathbf{b})$$

Empirical Results

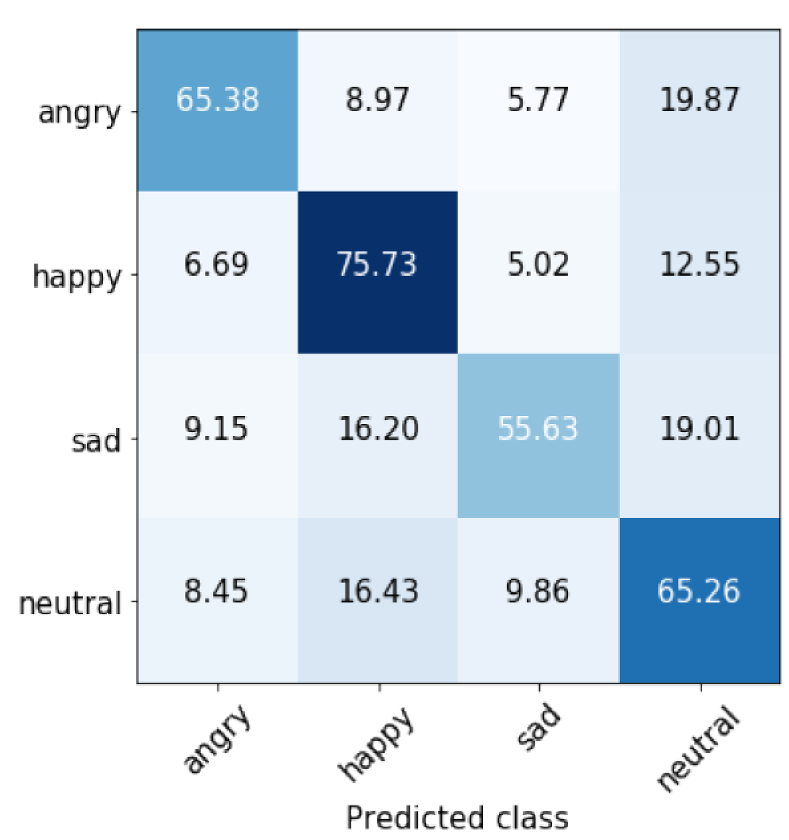
- **Dataset:** Interactive Emotional Dyadic Motion Capture (IEMOCAP). Following by the previous research, the final dataset contains a total of 5,531 utterances (1,636 happy, 1,084 sad, 1,103 angry, 1,708 neutral).
- **Comparison with the state-of-the-art methods:** The top 2 best performing models are marked in bold. The "-ASR" models are trained with processed transcripts from the Google Cloud Speech API (WER 5.53%).

Model	WAP
ACNN [31]	0.561
LLD RNN-attn [26]	0.635
RNN(prop.)-ELM [34]	0.628
3CNN-LSTM10H [20]	0.688
ARE	0.546 ±0.009
TRE	0.635 ±0.018
MDRE	**0.718 ±0.019**
MDREA	0.690 ±0.019
TRE-ASR	0.593 ±0.022
MDRE-ASR	**0.691 ±0.019**
MDREA-ASR	0.677 ±0.013

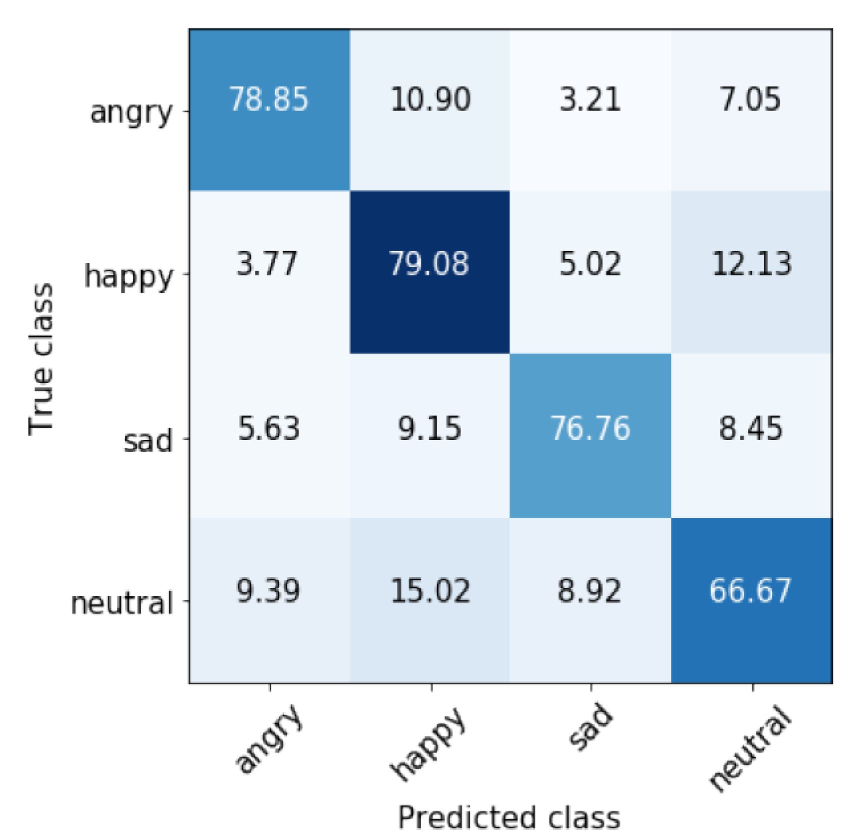
- **Error Analysis:** Confusion matrix of each model.



(a) ARE



(b) TRE



(c) MDRE

paper, slide → http://david-yoon.github.io