The 44th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)

Speech Emotion Recognition Using Multi-hop Attention Mechanism

[1]Seunghyun Yoon, **[1]Seokhyun Byun**, **[2]Subhadeep Dey** and **[1]Kyomin Jung**

SEOUL NATIONAL UNIVERSITY

RESEARCH INSTITUTE

Index

- Problem to Solve
- Related Works & Limitations

- Proposed Model: Multi-hop Attention
- Implementation Details
- Empirical Results
- Conclusion

Research Problem

Speech Emotion Recognition

Exploiting textual and acoustic data of an utterance for the speech emotion classification task

Related Work: Single modality

- **Using Regional Saliency for Speech Emotion Recognition**, Aldeneh, et., al., ICASSP-17

- **CNN based model**
- Achieve up to **60.7%** WA in IEMOCAP dataset

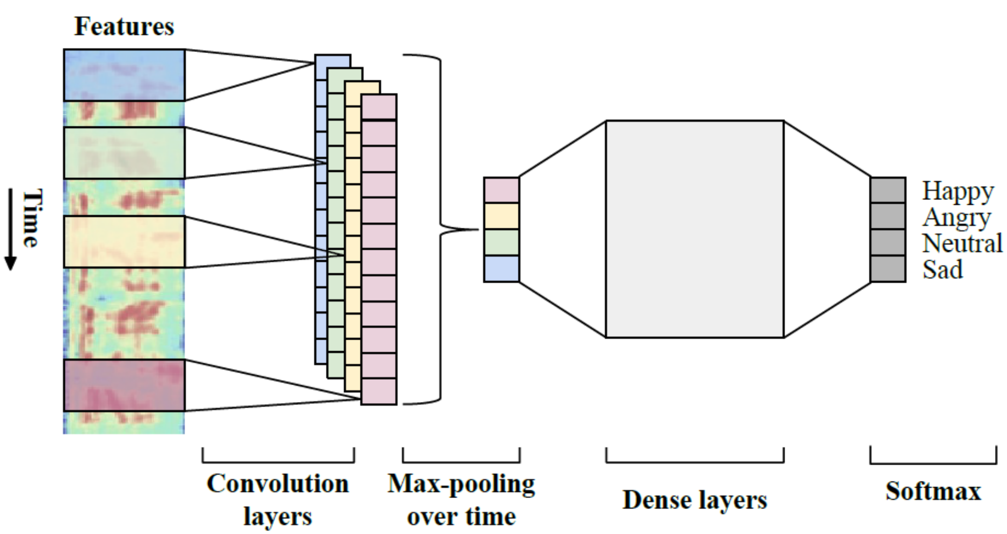
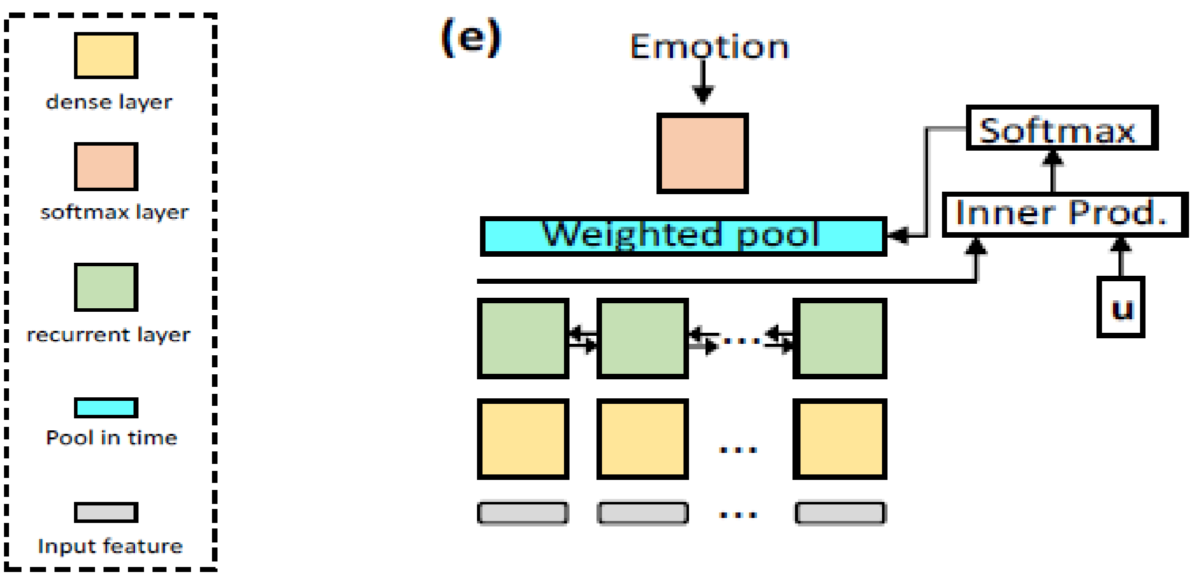


Fig. 1. Network Overview (four filters shown).

Related Work: Single modality

- **Automatic Speech Emotion Recognition Using Recurrent Neural Networks with Local Attention**, Mirsamadi et., al., ICASSP-17

- **RNN based model with Attention mechanism**
- Achieve up to **63.5%** WA in IEMOCAP dataset

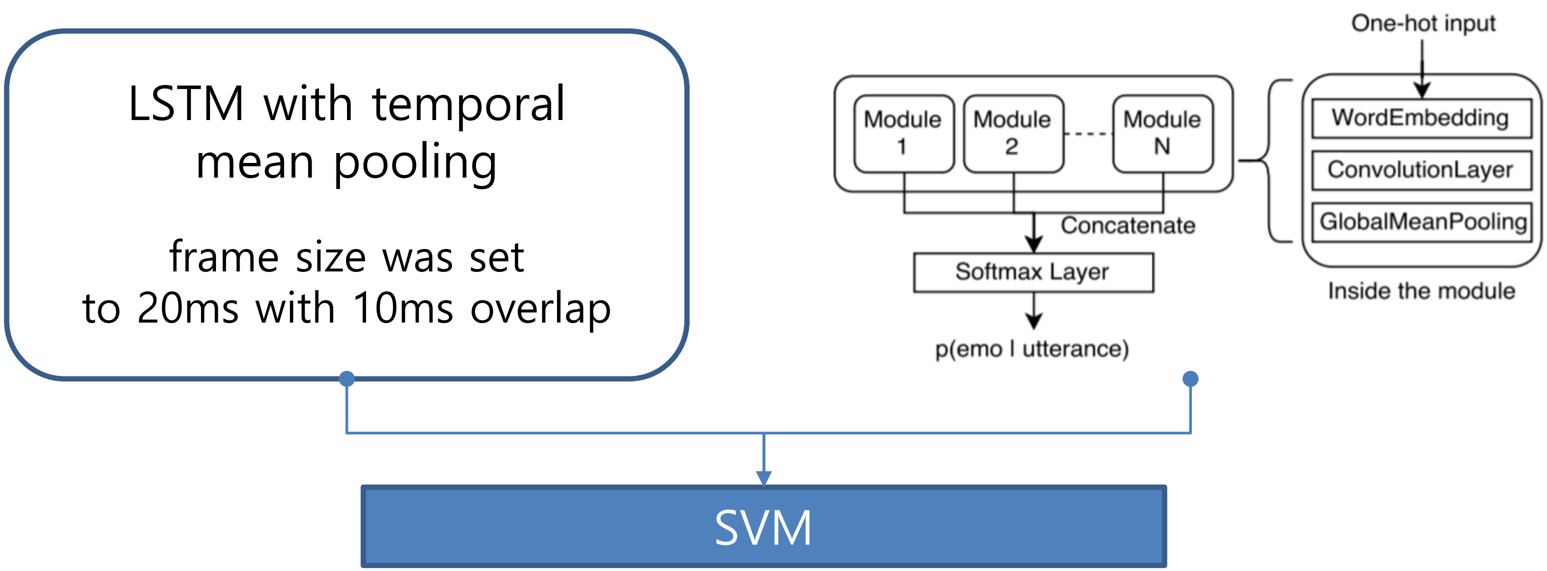


Related Work: Multi modality

- **Deep Neural Networks for Emotion Recognition Combining Audio and Transcripts**, Cho et., al., Interspeech-18

- **Combine acoustic information and conversation transcripts**
- Achieve up to **64.9%** WA in IEMOCAP dataset

Acoustic system *Multi-resolution CNN for **transcripts***



Related Work: Multi modality

- **Multimodal Speech Emotion Recognition Using Audio and Text**, Yoon et., al., SLT-18

- **End-to-end** training
- Achieve up to **71.8%** WA in IEMOCAP dataset

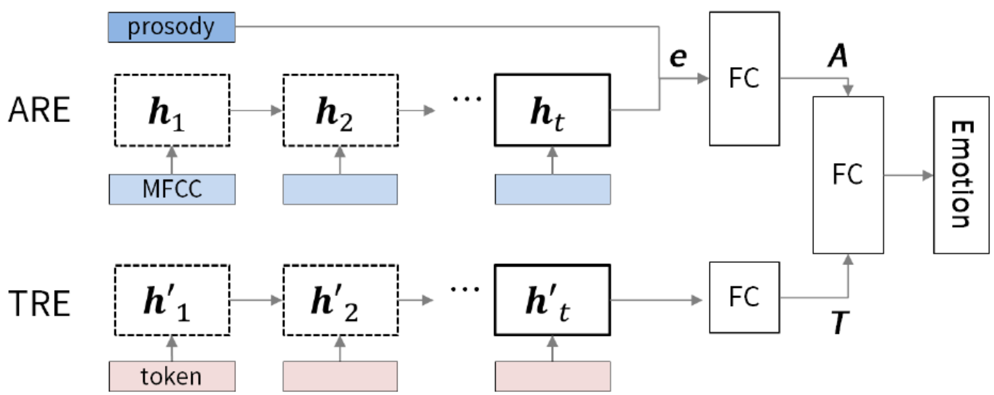


Fig. 1. Multimodal dual recurrent encoder. The upper part shows the ARE, which encodes audio signals, and the lower part shows the TRE, which encodes textual information.

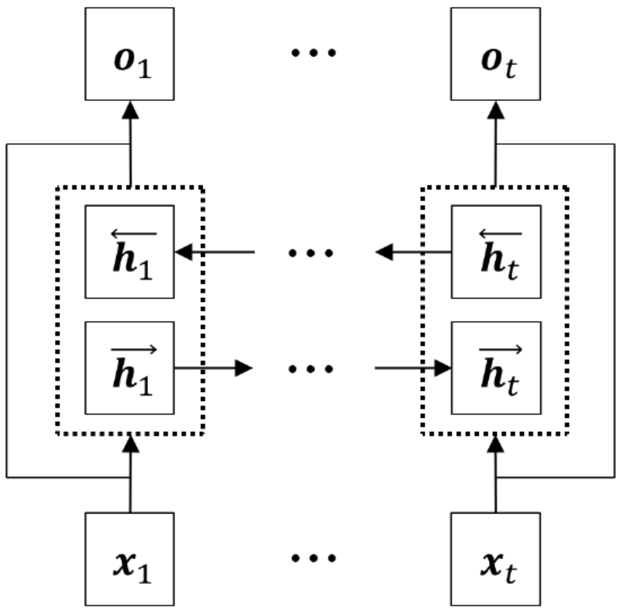
Bidirectional Recurrent Encoder (BRE)

- Audio-BRE
  - Recurrent Encoder for audio modality

- Features
  - Bidirectional
  - Residual Connection

$$\overrightarrow{\mathbf{h}}_t = f_\theta(\overrightarrow{\mathbf{h}}_{t-1}, \overrightarrow{\mathbf{x}}_t) + \overrightarrow{\mathbf{x}_t},$$
$$\overleftarrow{\mathbf{h}}_t = f'_\theta(\overleftarrow{\mathbf{h}}_{t+1}, \overleftarrow{\mathbf{x}}_t) + \overleftarrow{\mathbf{x}_t},$$
$$\mathbf{o}_t = [\overrightarrow{\mathbf{h}}_t; \overleftarrow{\mathbf{h}}_t],$$
$$\mathbf{o}_t^A = [\mathbf{o}_t; \mathbf{p}]$$

$\boldsymbol{x_t}$: audio feature
$\mathbf{p}$: prosodic feature vector



BRE model

Bidirectional Recurrent Encoder (BRE)

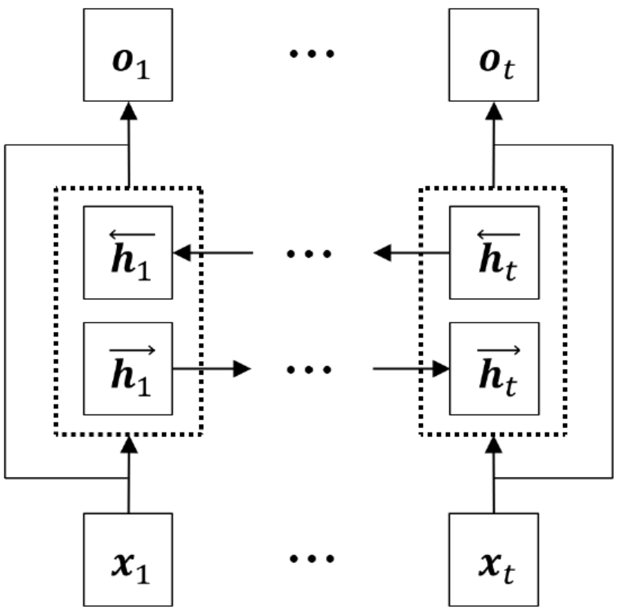
- Text-BRE
  - Recurrent Encoder for textual modality

- Tokenize textual information
  - I'm happy to hear the story

  → I 'm happy to hear the story

$$\overrightarrow{\mathbf{h}}_t = f_\theta(\overrightarrow{\mathbf{h}}_{t-1}, \overrightarrow{\mathbf{x}}_t) + \overrightarrow{\mathbf{x}_t},$$
$$\overleftarrow{\mathbf{h}}_t = f'_\theta(\overleftarrow{\mathbf{h}}_{t+1}, \overleftarrow{\mathbf{x}}_t) + \overleftarrow{\mathbf{x}_t},$$
$$\mathbf{o}_t^T = [\overrightarrow{\mathbf{h}}_t; \overleftarrow{\mathbf{h}}_t]$$

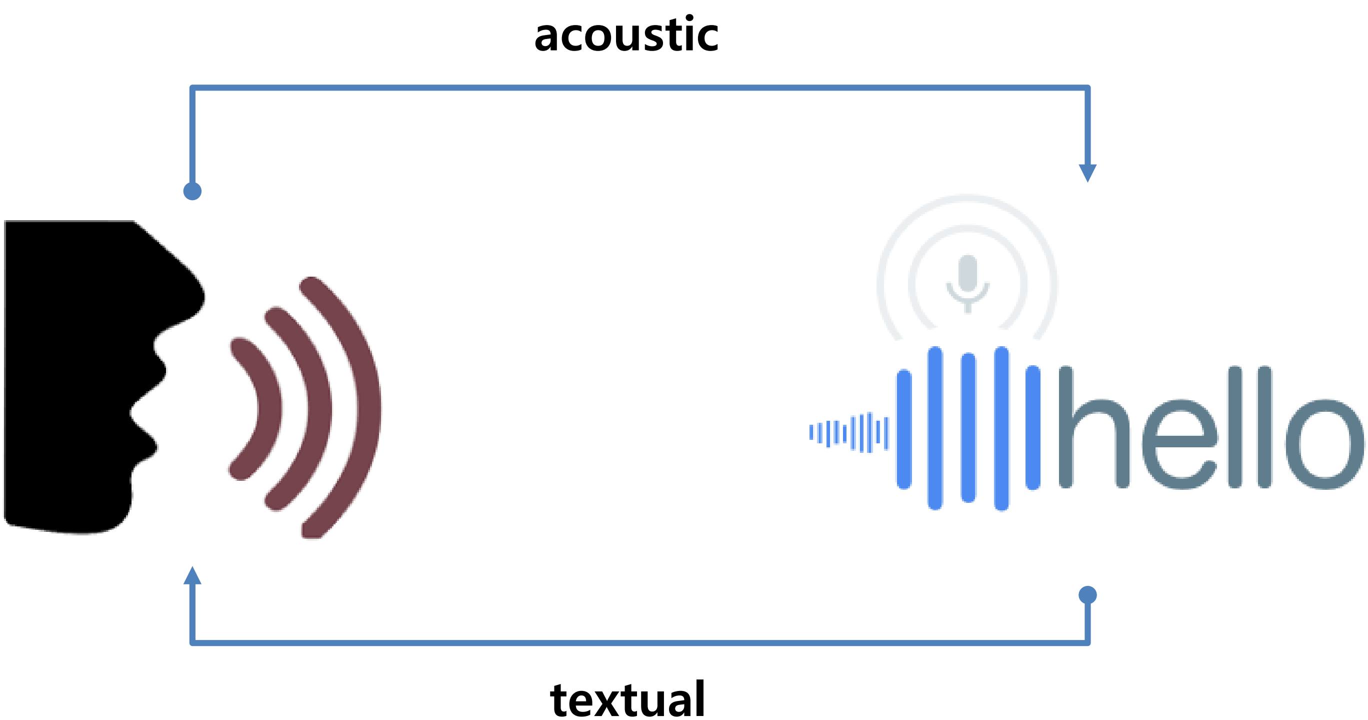
x_t : textual feature



BRE model

Multi-hop Attention (MHA)

- **Motivated by human behavior**
  - **Contextual Understanding from an iterative process**



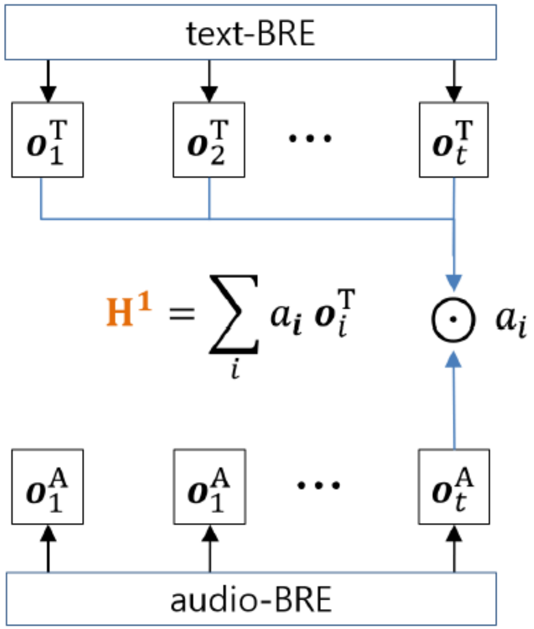
① Multi-hop Attention (MHA)

- **First Hop**

- **Context : Audio information**
- **Aggregate : Textual information**
- **Result : $\mathbf{H}^{1}$**

$$a_i = \frac{\exp(\ (\mathbf{o}_{\text{last}}^{A})^{\intercal}\ \mathbf{o}_i^{T}\ )}{\sum_i \exp(\ (\mathbf{o}_{\text{last}}^{A})^{\intercal}\ \mathbf{o}_i^{T}\ )},\ (i = 1, ..., t)$$

$$\mathbf{H}^{1} = \sum_i a_i\ \mathbf{o}_i^{T},\quad \mathbf{H} = [\mathbf{H}^{1}; \mathbf{o}_{\text{last}}^{A}].$$



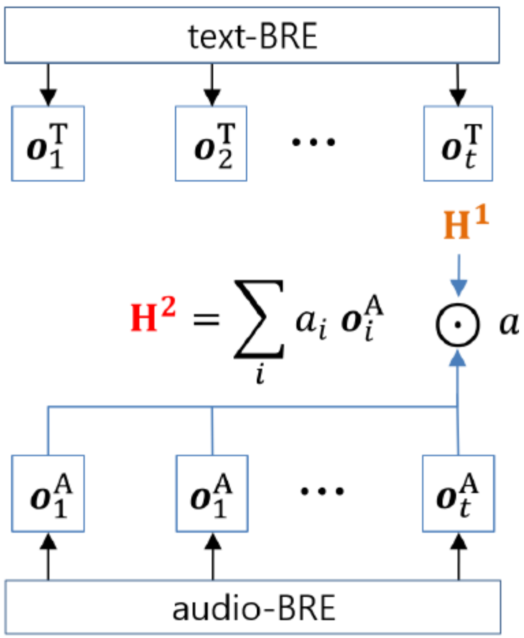
② Multi-hop Attention (MHA)

- Second Hop

- Context : Updated textual information
- Aggregate : Audio information
- Result : $\mathbf{H}^2$

$$a_i = \frac{\exp(\ (\mathbf{H}_1)^\intercal\ \mathbf{o}_i^A\ )}{\sum_i \exp(\ (\mathbf{H}_1)^\intercal\ \mathbf{o}_i^A\ )},\ (i = 1, ..., t)$$

$$\mathbf{H}^2 = \sum_i a_i\ \mathbf{o}_i^A,\ \ \mathbf{H} = [\mathbf{H}^1; \mathbf{H}^2],$$



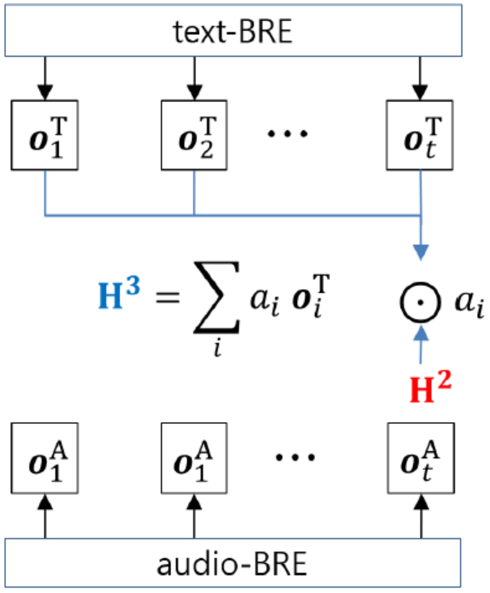
③ Multi-hop Attention (MHA)

- **Third Hop**

- **Context : Updated audio information**
- **Aggregate : Textual information**
- **Result : $\mathbf{H}^3$**

$$a_i = \frac{\exp(\ (\mathbf{H}_2)^\intercal\ \mathbf{o}_i^T\ )}{\sum_i \exp(\ (\mathbf{H}_2)^\intercal\ \mathbf{o}_i^T\ )},\ (i = 1, ..., t)$$

$$\mathbf{H}^3 = \sum_i a_i\ \mathbf{o}_i^T,\ \ \mathbf{H} = [\mathbf{H}^3; \mathbf{H}^2],$$



$\mathbf{H} = f\ (\mathbf{H}^3,\ \mathbf{H}^2)$

Optimization

- **Objective : classification**

- **Compute distribution of the predicted probability**
- **Cross-entropy loss**

$$\hat{y}_c = \mathrm{softmax}((\mathbf{H})^\intercal\ \mathbf{W} + \mathbf{b}\ ),$$

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

Dataset

- **Interactive Emotional Dyadic Motion Capture (IEMOCAP)**
  - **Five sessions** of utterances between two speakers (one male and one female)
  - Total 10 unique speakers participated

- **Environment setting**
  - **1,636 happy, 1,084 sad, 1,103 angry and 1,708 neutral**
  - “**excitement**” → merge with “**happiness**”
  - **10-fold** cross-validation

Implementation Details

- **Audio data**
  - **MFCC features (using Kaldi)**
    - frame size 25 ms at a rate of 10 ms with the Hamming window
    - concatenate it with its first, second order derivates → **120**-dims
    - Maximum step: 750 (mean + std)
  - **Prosodic features (using OpenSMILE)**
    - **35**-dims

- **Textual data**
  - **Ground-truth transcript form IEMOCAP dataset**
  - **ASR-processed transcript* (WER 5.53%)**

***Google Cloud Speech API**

Results

- **Textual information vs Acoustic information**
  - **text-BRE** shows higher performance than that of **audio-BRE** by 8%

Model	Modality	WA	UA
Ground-truth transcript			
E_vec-MCNN-LSTM [18]	A+T	0.649	0.659
MDRE [7]	A+T	0.718	-
audio-BRE (ours)	A	0.646	0.652
text-BRE (ours)	T	0.698	0.703
MHA-1 (ours)	A+T	**0.756**	**0.765**
MHA-2 (ours)	A+T	**0.765**	**0.776**
MHA-3 (ours)	A+T	0.740	0.753
ASR-processed transcript			
text-BRE-ASR (ours)	T	0.652	0.658
MHA-2-ASR (ours)	A+T	0.730	0.739

8% (0.646 → 0.698)

Results

- **Comparison with best baseline model**
  - **MHA-2** outperformed the **MDRE** by 6.5%

Model	Modality	WA	UA
Ground-truth transcript			
E_vec-MCNN-LSTM [18]	A+T	0.649	0.659
MDRE [7]	A+T	0.718	-
audio-BRE (ours)	A	0.646	0.652
text-BRE (ours)	T	0.698	0.703
MHA-1 (ours)	A+T	**0.756**	**0.765**
MHA-2 (ours)	A+T	**0.765**	**0.776**
MHA-3 (ours)	A+T	0.740	0.753
ASR-processed transcript			
text-BRE-ASR (ours)	T	0.652	0.658
MHA-2-ASR (ours)	A+T	0.730	0.739

6.5% (0.718 → 0.765)

Results

- **ASR-processed transcript**
  - performance degradation in **text-BRE-ASR** by 6.6%

Model	Modality	WA	UA
Ground-truth transcript			
E_vec-MCNN-LSTM [18]	A+T	0.649	0.659
MDRE [7]	A+T	0.718	-
audio-BRE (ours)	A	0.646	0.652
text-BRE (ours)	T	0.698	0.703
MHA-1 (ours)	A+T	**0.756**	**0.765**
MHA-2 (ours)	A+T	**0.765**	**0.776**
MHA-3 (ours)	A+T	0.740	0.753
ASR-processed transcript			
text-BRE-ASR (ours)	T	0.652	0.658
MHA-2-ASR (ours)	A+T	0.730	0.739

6.6% (0.698 → 0.652)

Results

- **ASR-processed transcript**
  - performance degradation in **MHA-2-ASR** by 4.6%

Model	Modality	WA	UA
Ground-truth transcript			
E_vec-MCNN-LSTM [18]	A+T	0.649	0.659
MDRE [7]	A+T	0.718	-
audio-BRE (ours)	A	0.646	0.652
text-BRE (ours)	T	0.698	0.703
MHA-1 (ours)	A+T	**0.756**	**0.765**
MHA-2 (ours)	A+T	**0.765**	**0.776**
MHA-3 (ours)	A+T	0.740	0.753
ASR-processed transcript			
text-BRE-ASR (ours)	T	0.652	0.658
MHA-2-ASR (ours)	A+T	0.730	0.739

4.6% (0.765 → 0.730)

Results

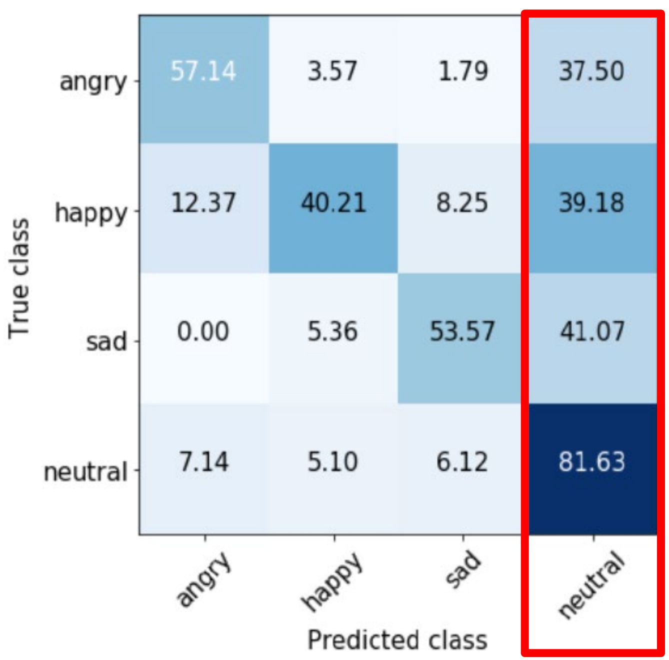
- **ASR-processed vs ground-truth**
  - **MHA-2** still outperformed the **MDRE** by 1.6%

Model	Modality	WA	UA
Ground-truth transcript			
E_vec-MCNN-LSTM [18]	A+T	0.649	0.659
MDRE [7]	A+T	0.718	-
audio-BRE (ours)	A	0.646	0.652
text-BRE (ours)	T	0.698	0.703
MHA-1 (ours)	A+T	**0.756**	**0.765**
MHA-2 (ours)	A+T	**0.765**	**0.776**
MHA-3 (ours)	A+T	0.740	0.753
ASR-processed transcript			
text-BRE-ASR (ours)	T	0.652	0.658
MHA-2-ASR (ours)	A+T	0.730	0.739

1.6% (0.718 → 0.730)

Error Analysis

- **Audio-BRE**
  - Most of the emotion labels are frequently misclassified as *"neutral"*
  - Supporting the claims in [7, 25]



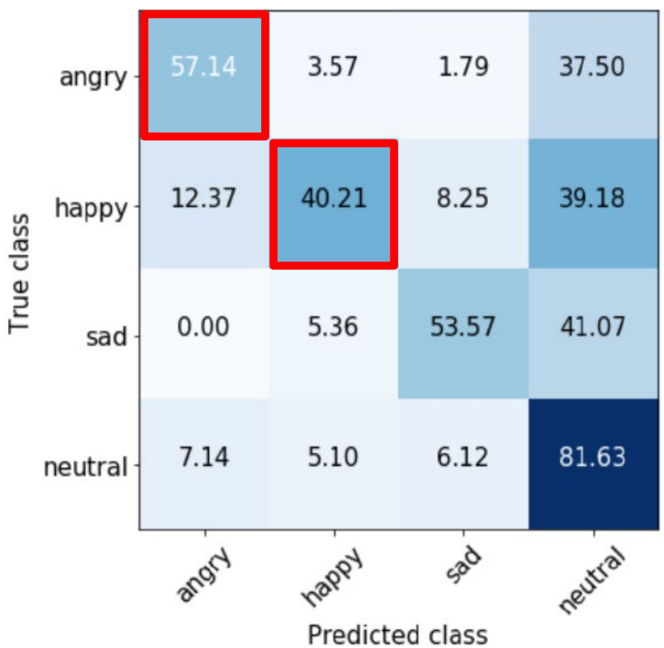
(a) audio-BRE

[7] Multimodal speech emotion recognition using audio and text, Yoon et. al., SLT-18

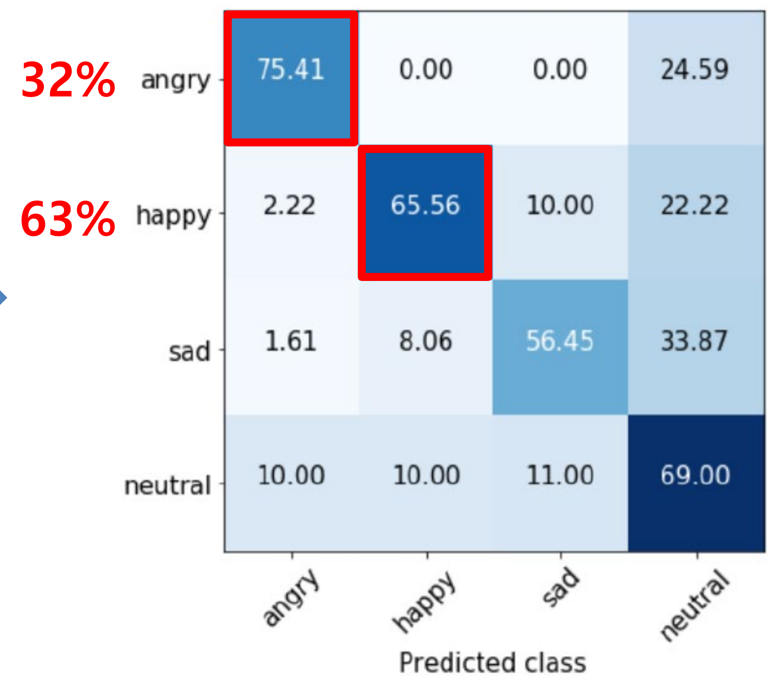
[25] Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech, Neumann et. al., Interspeech-17

Error Analysis

- **Text-BRE**
  - *"angry"* and "*happy*" are correctly classified by 32% (57.14 to 75.41) and 63% (40.21 to 65.56)



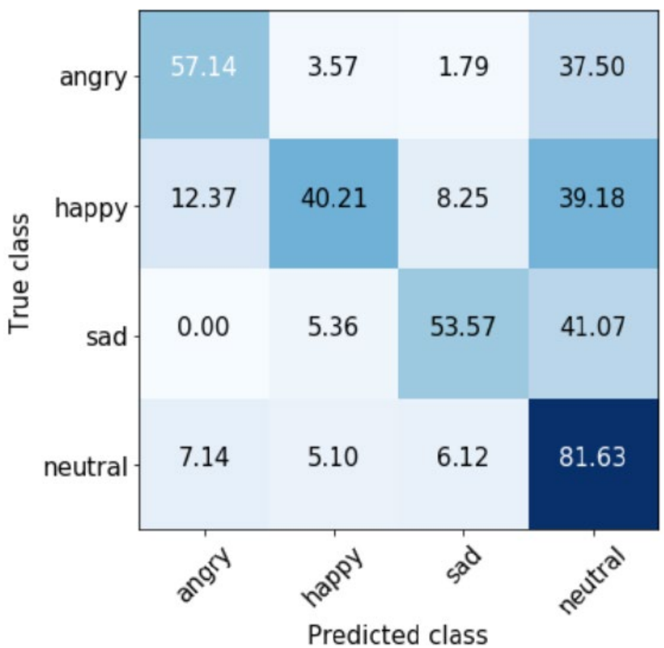
(a) audio-BRE



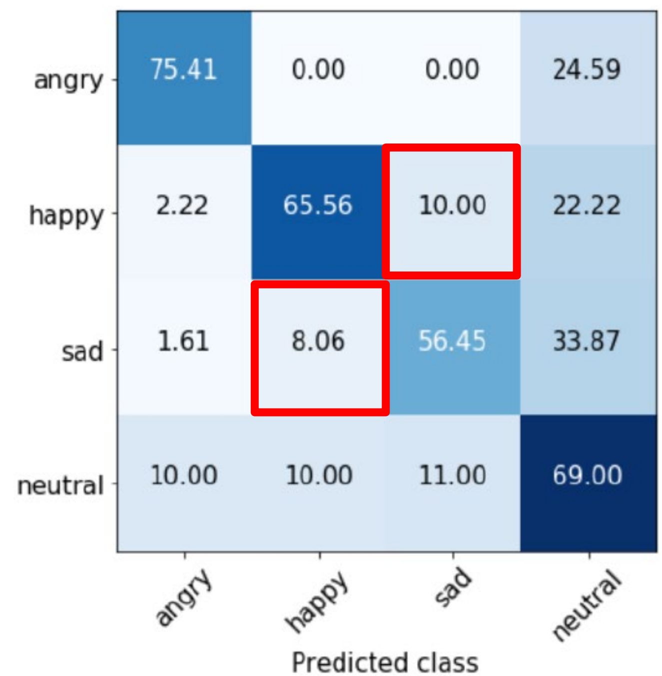
(b) text-BRE

Error Analysis

- **Text-BRE**
  - Incorrectly predicted instances of the "***happy***" as "**sad"** in 10%
  - even though these emotional states are opposites of one another



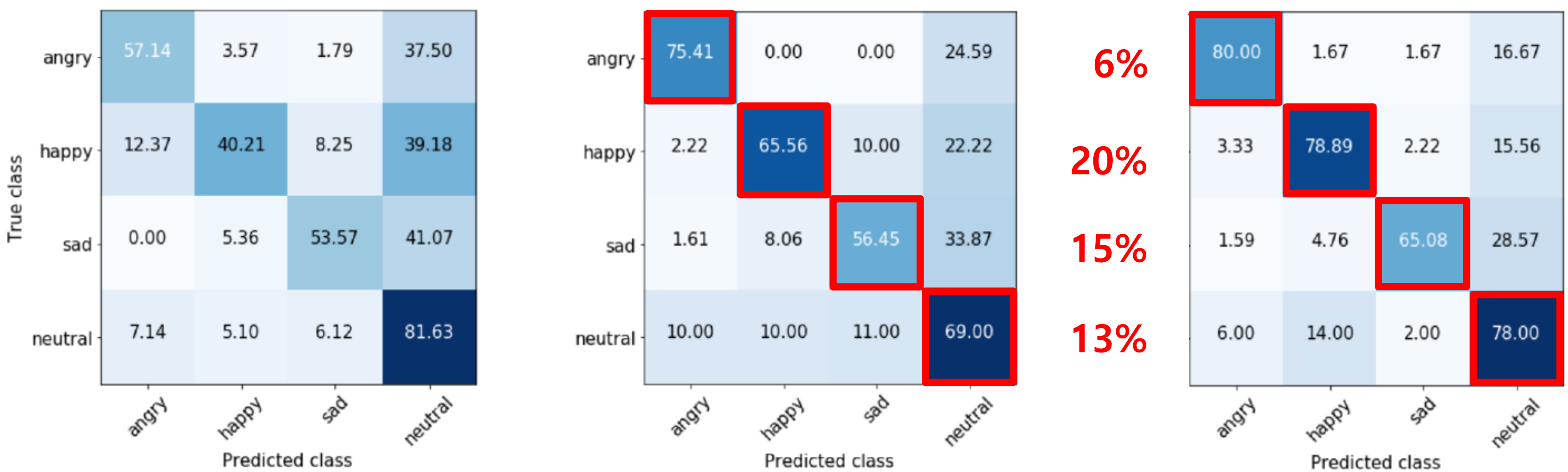
(a) audio-BRE



(b) text-BRE

Error Analysis

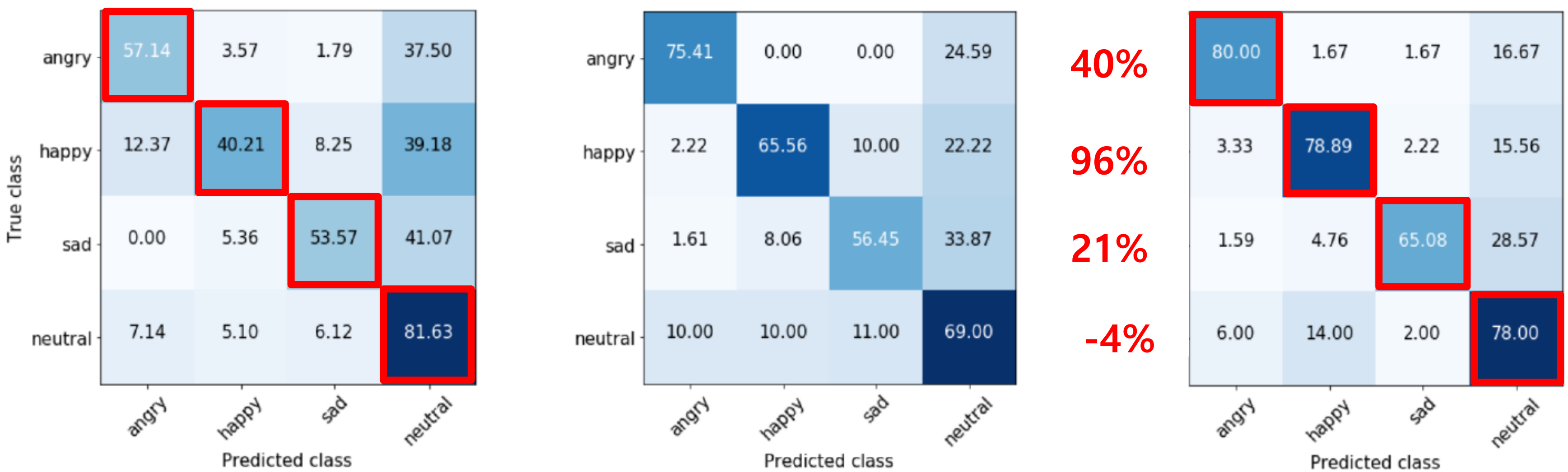
- **MHA-2**
  - Benefits from strengths of **audio-BRE** and **text-BRE**
  - Significant performance gain for all predictions (vs **text-BRE**)



(a) audio-BRE (b) text-BRE (c) MHA-2

Error Analysis

- **MHA-2**
  - Benefits from strengths of **audio-BRE** and **text-BRE**
  - Significant performance gain for all predictions (vs **audio-BRE**)



(a) audio-BRE (b) text-BRE (c) MHA-2

Conclusion

We study how to recognize speech emotion

- **PROPOSE** multi-hop attention model to combine acoustic and textual data for speech emotion recognition task
- **SHOW** proposed model outperforms the best baseline system
- **TEST** with ASR-processed transcripts and show the reliability of the proposed system in the practical scenario where the ground-truth transcripts are not available

Thank you