Speech Emotion Recognition Using Multi-hop Attention Mechanism

Seunghyun Yoon, Seokhyun Byun, Subhadeep Dey and Kyomin Jung
Index

• Problem to Solve
• Related Works & Limitations

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

- LSTM with temporal mean pooling
- Frame size was set to 20ms with 10ms overlap

**Multi-resolution CNN for transcripts**

- One-hot input
  - WordEmbedding
  - ConvolutionLayer
  - GlobalMeanPooling

- Inside the module
  - Softmax Layer
  - $p(\text{emo I utterance})$

- SVM
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

\[
\vec{h}_t = f_\theta(\vec{h}_{t-1}, \bar{x}_t) + x_t, \\
\hat{h}_t = f'_\theta(\hat{h}_{t+1}, \bar{x}_t) + \hat{x}_t, \\
o_t = [\vec{h}_t; \hat{h}_t], \\
o_t^A = [o_t; p]
\]

\(x_t\) : audio feature  
\(p\) : 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
  - \( \rightarrow \) I ’m happy to hear the story

\[
\begin{align*}
\mathbf{\bar{h}}_t &= f_{\theta}(\mathbf{\bar{h}}_{t-1}, \mathbf{\bar{x}}_t) + \mathbf{\bar{x}}_t, \\
\mathbf{\hat{h}}_t &= f'_{\theta}(\mathbf{\hat{h}}_{t-1}, \mathbf{\hat{x}}_t) + \mathbf{\hat{x}}_t, \\
\mathbf{o}_t^T &= [\mathbf{\bar{h}}_t; \mathbf{\hat{h}}_t]
\end{align*}
\]

\( x_t \): textual feature
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 : $H^1$

$$a_i = \frac{\exp\left( (o_{last}^A)^T o_i^T \right)}{\sum_i \exp\left( (o_{last}^A)^T o_i^T \right)}, \quad (i = 1, \ldots, t)$$

$$H^1 = \sum_i a_i o_i^T, \quad H = [H^1; o_{last}^A].$$
### ② Multi-hop Attention (MHA)

- **Second Hop**
  - **Context**: Updated textual information
  - **Aggregate**: Audio information
  - **Result**: $H^2$

\[
a_i = \frac{\exp\left((H_1)^T o_i^A\right)}{\sum_i \exp\left((H_1)^T o_i^A\right)}, \quad (i = 1, \ldots, t)
\]

\[
H^2 = \sum_i a_i o_i^A, \quad H = [H^1; H^2],
\]
Multi-hop Attention (MHA)

- **Third Hop**

- **Context**: Updated audio information
- **Aggregate**: Textual information
- **Result**: $H^3$

\[
a_i = \frac{\exp \left( (H_2)^T o_i^T \right)}{\sum_i \exp \left( (H_2)^T o_i^T \right)}, \quad (i = 1, \ldots, t)
\]

\[
H^3 = \sum_i a_i o_i^T, \quad H = [H^3; H^2],
\]
Optimization

• Objective: classification

• Compute distribution of the predicted probability

• Cross-entropy loss

\[
\hat{y}_c = \text{softmax}( (H)^T W + b ),
\]

\[
\mathcal{L} = - \log \prod_{i=1}^{N} \sum_{c=1}^{C} y_{i,c} \log(\hat{y}_{i,c}),
\]
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 derivate
    • Maximum step: 750 (mean + std) → 120-dims
  • **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%

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

8% (0.646 → 0.698)
Results

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

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

6.5% (0.718 → 0.765)
Results

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

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

6.6% (0.698 → 0.652)
Results

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

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

4.6% (0.765 → 0.730)
### Results

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

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

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]

[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)

![Diagram](image.png)
Error Analysis

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

6% 20% 15% 13%
Error Analysis

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

![Heatmaps](attachment:image.png)
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