Introduction

Most emotion recognition research focuses on two descriptions of emotion, both of these have flaws:

- **Categorical** (happy, sad, angry, neutral)
- **Dimensional** (arousal, valence, dominance)

To be too coarse to describe real emotion.

Datasets

- **IEMOCAP** dataset [1] contains 12 hours of scripted and improvised dyadic interactions from 10 actors. Each utterance has categorical and dimensional labels from 3 annotators.
- **Usborne** children’s audiobook dataset, used in Blizzard 2017 [2], contains 6.5 hours of expressive speech from a British female speaker.

Emotion recognition

Standard categorical emotion recognition on IEMOCAP. Using narrowband spectrogram, or the minimalistic acoustic parameter set, eGeMAPS [3]

<table>
<thead>
<tr>
<th>Model</th>
<th>Inputs</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>N/A</td>
<td>24.14%</td>
</tr>
<tr>
<td>Most common</td>
<td>N/A</td>
<td>33.00%</td>
</tr>
<tr>
<td>LSTM</td>
<td>eGeMAPS LLDs</td>
<td>43.17%</td>
</tr>
<tr>
<td>TD-CNN</td>
<td>Spectrogram</td>
<td>58.94%</td>
</tr>
<tr>
<td>DNN</td>
<td>eGeMAPS functionals</td>
<td>72.77%</td>
</tr>
<tr>
<td>RNN-ELM</td>
<td>MFCCs, F0, VUV, zero-crossings</td>
<td>63.89%</td>
</tr>
<tr>
<td>CNN-MKL</td>
<td>ComParE 2016, video, word2vec</td>
<td>76.85%</td>
</tr>
</tbody>
</table>

- **LSTM** - recurrent neural network; ongoing work
- **TD-CNN** - time-distributed CNN; ongoing work
- **DNN** for 4-class speech-only IEMOCAP; result is state-of-the-art, dependent on the test set split

Stimulation [6]

- Regularisation method that encourages high activations surrounding points in a prior map
- Prior map is a layout of classes on a unit-grid
- t-SNE embedding used as prior map (Figure 3b)
- Stimulation improves interpretability of emotion

Emotion space

- Multi-task learning (MTL) to train emotion space
- Emotion space is the final shared layer’s activations

```
<table>
<thead>
<tr>
<th>eGeMAPS</th>
<th>NN</th>
<th>emotion space</th>
</tr>
</thead>
<tbody>
<tr>
<td>private layer</td>
<td></td>
<td></td>
</tr>
<tr>
<td>basic emotions</td>
<td></td>
<td></td>
</tr>
<tr>
<td>dimensional emotions</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```

Emotive speech synthesis

- **DNN** synthesis using Merlin toolkit [7]
- **Style adaptation using auxiliary features**
- eGeMAPS - 88 acoustic parameters from waveform
- Dimensional - 3-dimensional emotion description
- eGrid - emotion space, stimulated in a 16 x 16 grid
- Categorical - 4-class emotion description
- Non-emotive - no auxiliary features

Listening test

- MUSHRA listening test, 16 screens, 20 participants
- Copy synthesis reference: 100 rating for all samples

Cross-corpus prediction

Create auxiliary features for SPSS style adaptation;

- Use recognition model trained on IEMOCAP
- From Usborne data, predict emotion space, categorical and dimensional labels

![Figure 1: Time-distributed CNN architecture](image1)

![Figure 2: MTL architecture showing emotion space](image2)

![Figure 3: t-SNE embedding of eGeMAPS features for IEMOCAP](image3)

![Figure 4: Visualisation of activations without & with stimulation](image4)

![Figure 5: Distribution of Usborne categorical emotion predictions](image5)

![Figure 6: Ranksum test. Median rating & 95% confidence interval](image6)

Conclusion

- To mitigate issues with existing emotion descriptions, we learn an emotion space using MTL
- Stimulation is added to improve interpretability
- Evaluation is performed with a perceptual test

References