Learning interpretable control dimensions for speech synthesis by using external data

Zack Hodari, Oliver Watts, Srikanth Ronanki, Simon King
The Centre for Speech Technology Research, University of Edinburgh, United Kingdom
{zack.hodari, oliver.watts, srikanth.ronanki, simon.king}@ed.ac.uk

Introduction
There are many aspects of speech that we might want to control when creating text-to-speech systems. We present a general method that enables control of arbitrary aspects of speech, which we demonstrate on emotion control.

Controllable SPSS

Waveform

eGeMAPS features

Text

Linguistic features

Neural network

Categorical emotions

Acoustic and Duration model

Dimensional emotions

Figure 1: Proposed controllable SPSS system for emotion control. Left – Emotion recognition model trained on external data. Right – Predicted labels used as auxiliary features in a SPSS voice.

Datasets

External data – IEMOCAP, 12 hours of dyadic conversations from 10 actors, with categorical and continuous emotion labels.

TTS data – Blizzard Challenge 2017 dataset, contains 6.5 hours of expressive speech from a British female speaker.

Label prediction
Using the emotion recognition model (Figure 1) trained on IEMOCAP, we predict labels using the TTS dataset to provide annotations for training a TTS voice.

Figure 2: Demonstration of $F_0$ variation as control is changed

Table 1:

<table>
<thead>
<tr>
<th>Objective metric</th>
<th>MCD (dB)</th>
<th>BAP (dB)</th>
<th>log F0 (RMSE)</th>
<th>VUV (error %)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNN-B (baseline)</td>
<td>5.650</td>
<td>0.075</td>
<td>51.209</td>
<td>7.451</td>
</tr>
<tr>
<td>DNN-C (with control)</td>
<td>5.719</td>
<td>0.076</td>
<td>50.624</td>
<td>7.551</td>
</tr>
</tbody>
</table>

Listening tests

Table 2: Confusion matrix for the forced-choice emotion classification task; accuracy for each emotion is in bold face

<table>
<thead>
<tr>
<th>Correct class</th>
<th>Angry</th>
<th>Happy</th>
<th>Neutral</th>
<th>Sad</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted class</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Angry</td>
<td>30%</td>
<td>51%</td>
<td>13%</td>
<td>7%</td>
</tr>
<tr>
<td>Happy</td>
<td>36%</td>
<td>13%</td>
<td>29%</td>
<td>22%</td>
</tr>
<tr>
<td>Neutral</td>
<td>10%</td>
<td>15%</td>
<td>66%</td>
<td>10%</td>
</tr>
<tr>
<td>Sad</td>
<td>10%</td>
<td>4%</td>
<td>30%</td>
<td>56%</td>
</tr>
</tbody>
</table>

Mean accuracy 41%

Figure 3: Pairwise preference ratios and 95% confidence interval

DNN-C = 0.493
DNN-B = 0.507
DNN-C = 0.552
DNN-R = 0.448