A learned emotion space for emotion recognition and emotive speech synthesis

Zack Hodari[1,2], Simon King[1]

zack.hodari@ed.ac.uk, Simon.King@ed.ac.uk

[1]The Centre for Speech Technology Research (CSTR), The University of Edinburgh, UK
[2]EPSRC Centre for Doctoral Training in Data Science, The University of Edinburgh, UK

Introduction

Most emotion recognition research focusses on two descriptions of emotion, both of these have flaws;

- Categorical (happy, sad, angry, neutral) **Too coarse** to describe real emotion
- Dimensional (arousal, valence, dominance) Open to interpretation → **unreliable** annotations

We address this in three stages;

- Learning an abstract **emotion space** with MTL
- Using **stimulation** to improve interpretability
- Evaluation using **expressive SPSS** voices

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 1: Performance classifying; happy, sad, angry, neutral

Model	Inputs	Accuracy
Random	N/A	24.14%
Most common	N/A	33.00%
LSTM	eGeMAPS LLDs	43.17%
TD-CNN	Spectrogram	58.94%
DNN	eGeMAPS functionals	**72.77%**
RNN-ELM [4]	MFCCs, F_0, VUV, zero-crossings	63.89%
CNN-MKL [5]	ComParE 2016, video, word2vec	**76.85%**

- 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

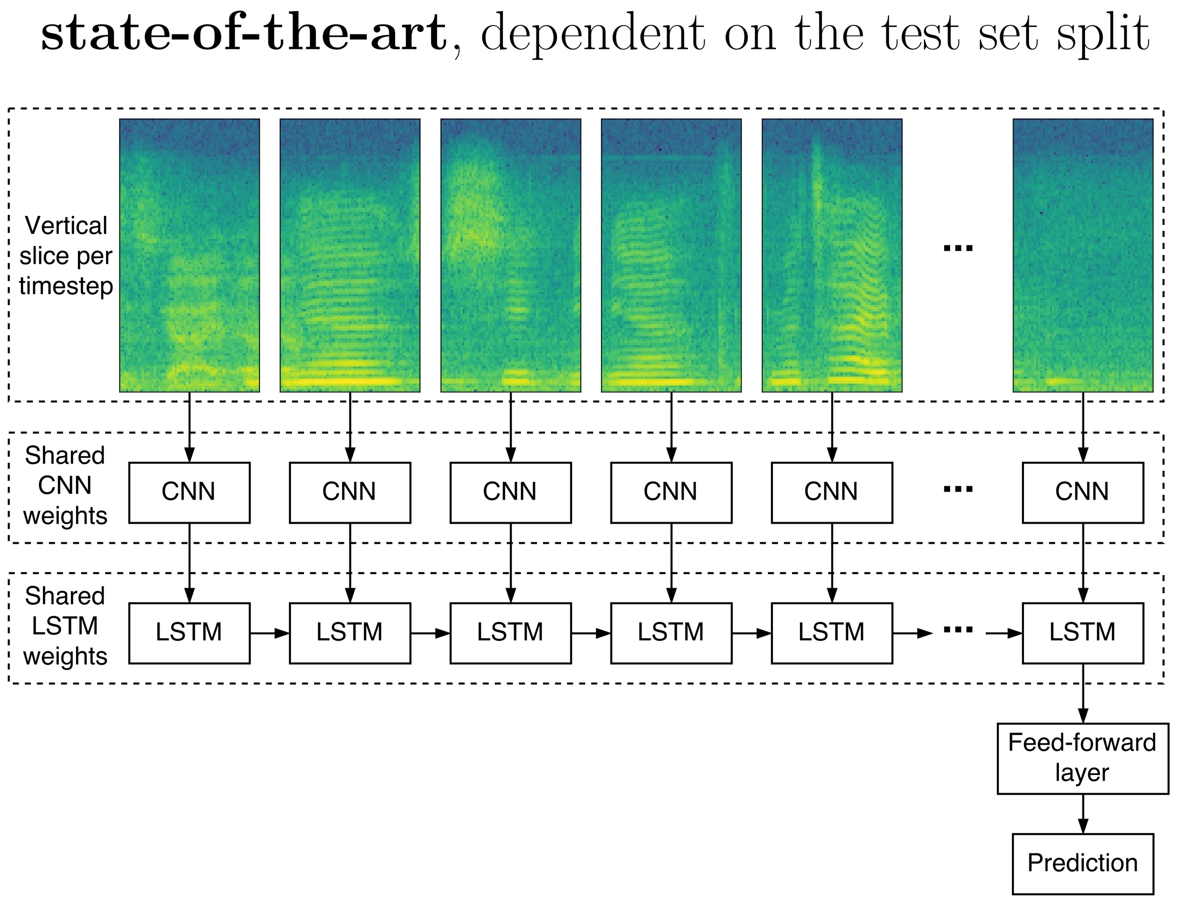


Figure 1: Time-distributed CNN architecture

Emotion space

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

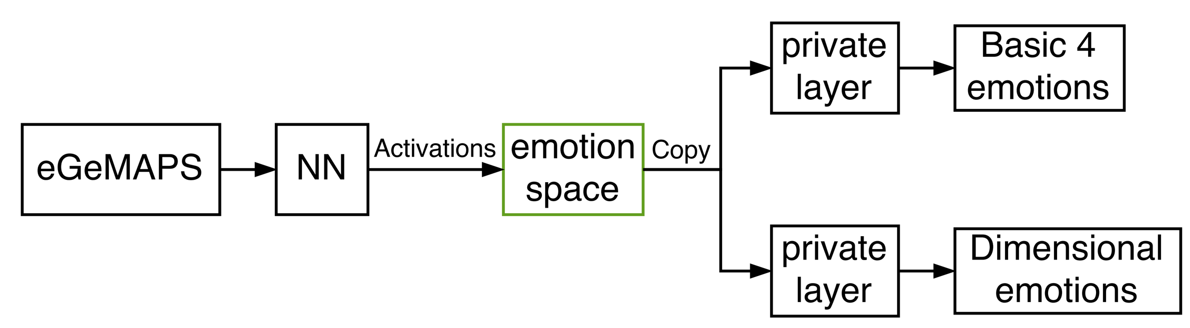
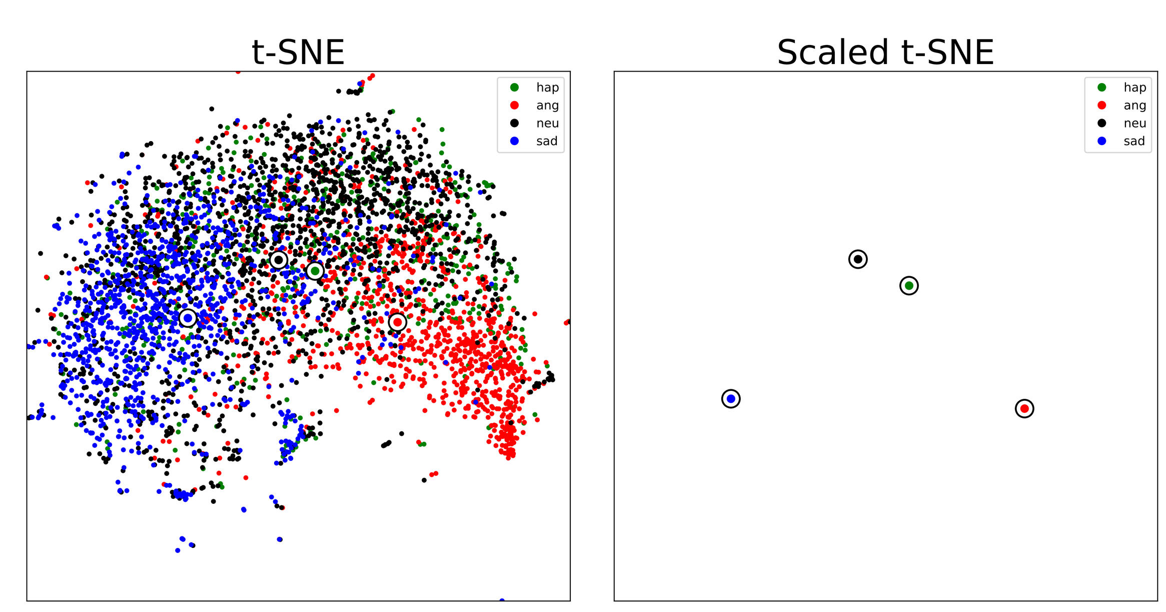


Figure 2: MTL architecture showing emotion space

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



(a) t-SNE embedding (b) Scaled class means

Figure 3: t-SNE embedding of eGeMAPS features for IEMOCAP

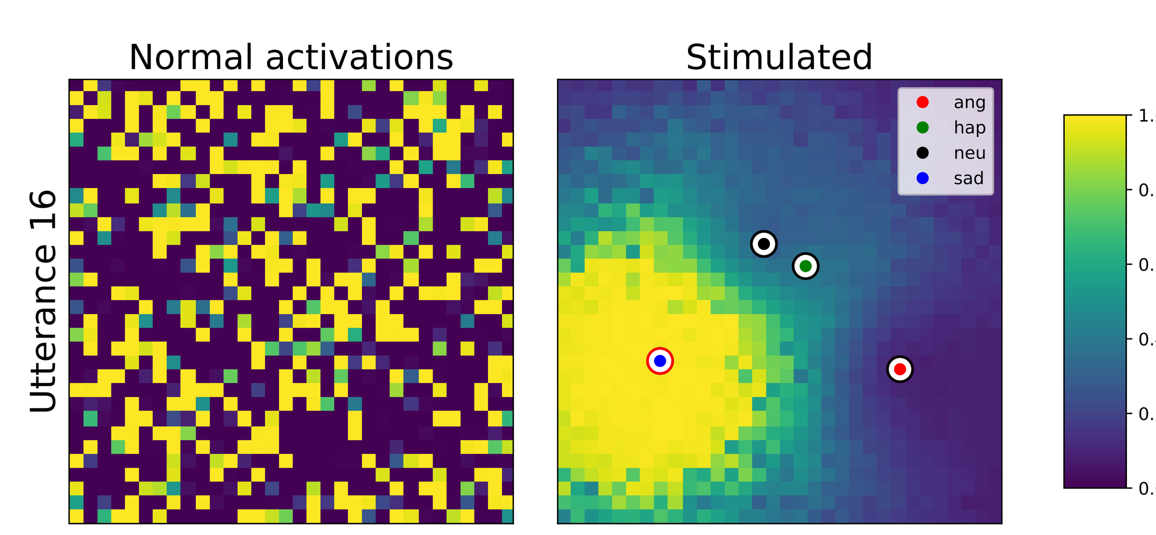


Figure 4: Visualisation of activations without & with stimulation

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

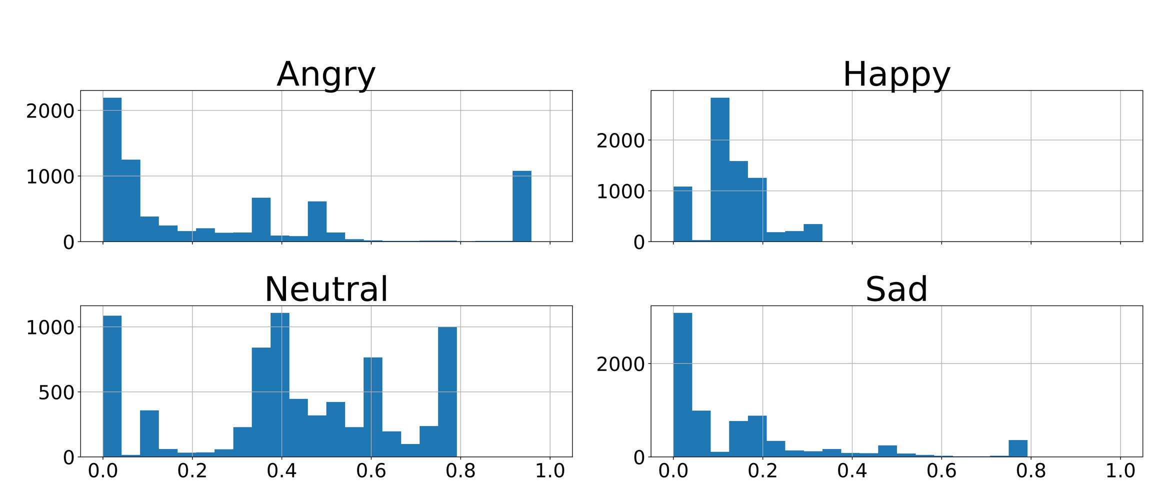


Figure 5: Distribution of Usborne categorical emotion predictions

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

Table 2: Objective results of trained DNN synthesis voices

	Objective metric			
	MCD (dB)	BAP (dB)	log F_0 (RMSE)	VUV (error %)
eGeMAPS	**5.631**	**0.314**	**44.356**	**14.254**
Dimensional	5.850	0.327	50.439	14.864
eGrid	5.825	0.327	51.420	15.211
Categorical	5.820	0.324	52.372	14.493
Non-emotive	5.845	0.329	52.846	14.768

Listening test

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

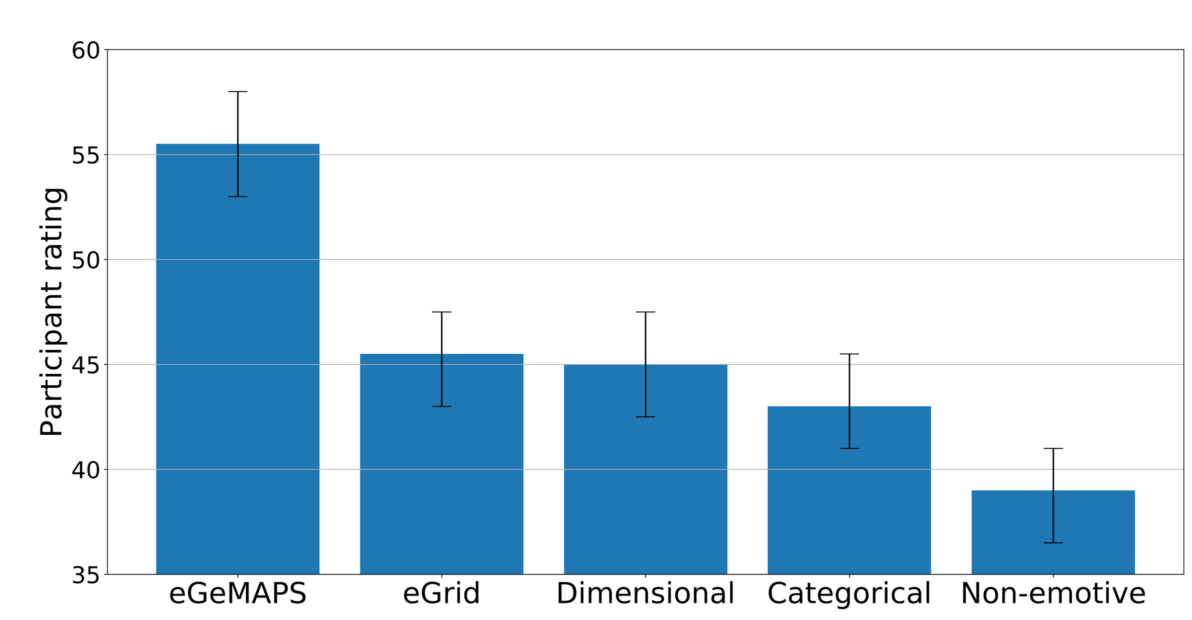


Figure 6: Ranksum test. Median rating & 95% confidence interval

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

[1] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. IEMOCAP: Interactive emotional dyadic motion capture database, 2008.

[2] Simon King and Vasilis Karaiskos. The blizzard challenge 2016.

[3] Florian Eyben, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, et al. The geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing, 2016.

[4] Jinkyu Lee and Ivan Tashev. High-level feature representation using recurrent neural network for speech emotion recognition, 2015.

[5] Soujanya Poria, Iti Chaturvedi, Erik Cambria, and Amir Hussain. Convolutional MKL based multimodal emotion recognition and sentiment analysis, 2016.

[6] Shawn Tan, Khe Chai Sim, and Mark Gales. Improving the interpretability of deep neural networks with stimulated learning, 2015.

[7] Zhizheng Wu, Oliver Watts, and Simon King. Merlin: An open source neural network speech synthesis system, 2016.