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

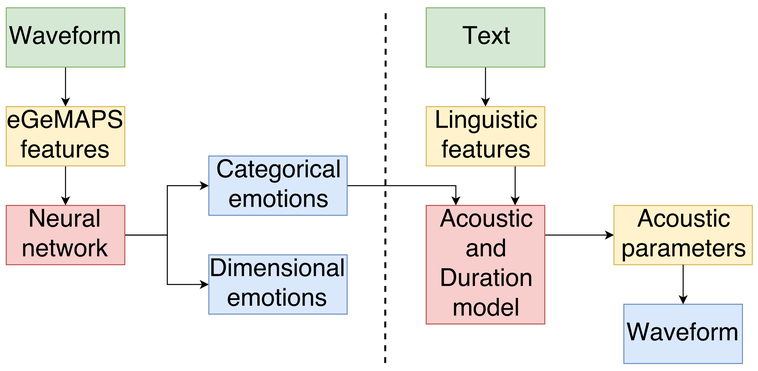


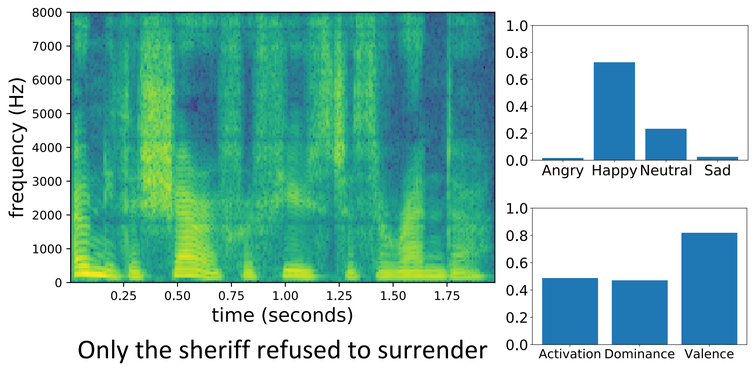
Figure 1: Proposed controllable SPSS system for emotion control.
Left − Emotion recognition model trained on external data.
Right − Predicted labels used as auxiliary features in a SPSS voice.

Datasets

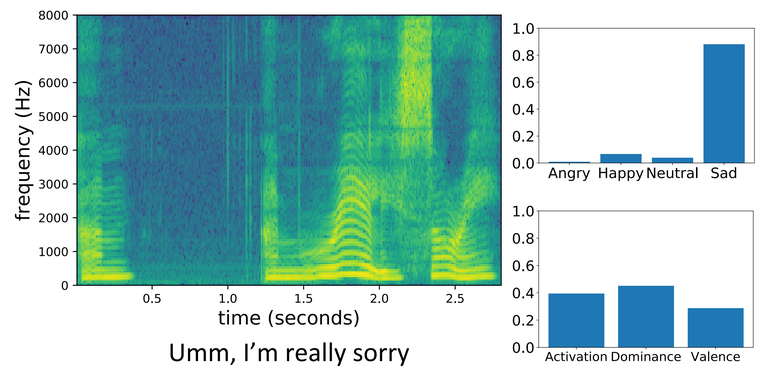
External data − IEMOCAP, 12 hours of dyadic conversations from 10 actors, with categorical and continuous emotion labels.
TTS data − Blizzard Challenge 2017 dataset, contains 6.5 hours of expressive speech from a British female speaker.

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



Only the sheriff refused to surrender



Umm, I'm really sorry

Objective evaluation

	Objective metric			
	MCD (dB)	BAP (dB)	log F0 (RMSE)	VUV (error %)
DNN-B (baseline)	**5.650**	**0.075**	51.209	**7.451**
DNN-C (with control)	5.719	0.076	**50.624**	7.551

Table 1: Objective results with and without control vectors

Listening tests

Correct class	Predicted class: Angry	Happy	Neutral	Sad
Angry	**30%**	51%	13%	7%
Happy	36%	**13%**	29%	22%
Neutral	10%	15%	**66%**	10%
Sad	10%	4%	30%	**56%**
	Mean accuracy 41%			

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

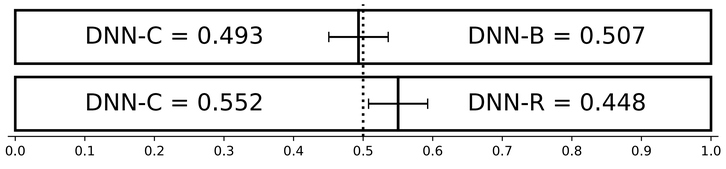


Figure 3: Pairwise preference ratios and 95% confidence interval

Figure 2: Demonstration of F_0 variation as control is changed