The Emotional Voices Database: Towards Controlling the Emotion Dimension in Voice Generation Systems

Adaeze Adigwe Noé Tits Kevin El Haddad Sarah Ostadabbas Thierry Dutoit

Overview

Database:

- intended to be open-sourced (English part already is)
- for synthesis and generation purpose
- male and female actors in English and a male actor in French
- 5 emotion classes

Experiments:

- emotional voice conversion
- categorical emotional TTS
- control of emotional intensity in TTS

Data description

Type of data	Audio, text and emotion category
How data was acquired	Audio recorded in 1 anechoic chamber of the University of Mons and 2 different anechoic chambers of the Northeastern University campus.
Data format	Segmented in sentences, associated with transcriptions (CMU-Artic/SIWIS), classified in emotional categories
Experimental features	Recordings of sentences uttered by 2 male and 2 female speakers in 5 different emotions, making a total of 7000 sentences
Data accessibility	https://github.com/numediart/EmoV-DB

Speaker	Gender	Language	Neutral	Amused	Angry	Sleepy	Disgust
Spk-Je	Female	English	417	222	523	466	189
Spk-Bea	Female	English	373	309	317	520	347
Spk-Sa	Male	English	493	501	468	495	497
Spk-Jsh	Male	English	302	298	-	263	-
Spk-No	Male	French	317	-	273	-	-

Table 1: Gender and language of recorded sentences of/from each actor/speaker and amount of utterances segmented per speaker and per emotion. All speakers were recorded in all emotions, the - sign only signifies that the corresponding data were not segmented yet.

Experiments

Emotional Voice Conversion

Use of Merlin Toolkit

- Acoustic feature extraction with the WORLD vocoder (source and target)
- DTW to align features
- Regression with DNN of 6 layers of 1024 tangent units

Pair	Spk-Bea	Spk-Sa	Spk-No
neutral-neutral	96%	90%	98%
neutral-angry	78%	71%	83%

Table 2: Percentage of angry and neutral speech styles being accurately classified.

Categorical Emotional TTS

Use of DCTTS (tensorflow implementation)

- pre-training on LJ-Speech
- fine-tuning towards the neutral voice of one of the actresses
- fine-tuning towards each emotion class of the same speaker

	Intelligibility	Confidence
Amused	2.01 ± 0.24	2.00 ± 0.27
Angry	2.76 ± 0.25	2.10 ± 0.28
Disgusted	2.17 ± 0.27	2.27 ± 0.30
Neutral	3.60 ± 0.26	3.59 ± 0.24
Sleepy	2.59 ± 0.28	3.29 ± 0.26

Table 3: MOS test results of synthesized files

Control of Emotional Intensity in TTS

Modified version of DCTTS that takes an encoding of the emotion category at the input. We concatenate encodings with character embeddings.

- one-hot encoding is used during training
- at synthesis stage, we can modify the intensity of an emotion category by inputting other codes. We chose these constraints: the sum must be one
- Demonstration

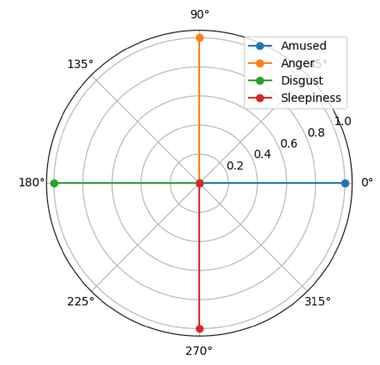

Figure 1: Demo

Future Works

- Perception Tests for the last experiment
- Multi-speaker model (For now we use the data from only one speaker)
- Synthesis with non-verbal expressions