Emotion Recognition of the Singing Voice: Toward a Real-Time Analysis Tool for Singers

Abstract. Current computational-emotion research has focused on applying acoustic properties to analyze how emotions are perceived mathematically or used in natural language processing machine learning models. With most recent interest being in analyzing emotions from the spoken voice, little experimentation has been performed to discover how emotions are recognized in the singing voice – both in noiseless and noisy data (i.e., data that is either inaccurate, difficult to interpret, has corrupted/distorted/nonsense information like actual noise sounds in this case, or has a low ratio of usable/unusable information). Not only does this ignore the challenges of training machine learning models on more subjective data and testing them with much noisier data, but there is also a clear disconnect in progress between advancing the development of convolutional neural networks and the goal of emotionally cognizant artificial intelligence. By training a new model to include this type of information with a rich comprehension of psycho-acoustic properties, not only can models be trained to recognize information within extremely noisy data, but advancement can be made toward more complex biofeedback applications – including creating a model which could recognize emotions given any human information (language, breath, voice, body, posture) and be used in any performance medium (music, speech, acting) or psychological assistance for patients with disorders such as BPD, alexithymia, autism, among others. This paper seeks to reflect and expand upon the findings of related research and present a stepping-stone toward this end goal.

Keywords:

Biofeedback	*A technique that involves monitoring a person's physiological state and feeding information about it back to that person*
Neural Network (NN) Not just for computer scientists anymore!	*A form of artificial intelligence that seeks to replicate biological learning in the brain through the development of neurons and synaptic connections*
Convolutional Neural Network (CNN)	*A type of deep-learning neural network primarily applied to analyzing visual imagery for image/video recognition and classification, and natural language processing*
Spectral Analysis	*A type of visual-based audio analysis used in measuring the distribution of acoustic energy across frequencies*
Fast Fourier Transform (FFT)	*An algorithm for analyzing energy distribution in speech/sung sound, including voice harmonics; computes a sequence's discrete Fourier transform (DFT)/inverse (IDFT)*
Linear Predictive Coding (LPC)	*A means of estimating the vocal tract filter (spectral envelope) that shaped the sound of a spoken/sung voice; used mostly in audio signal processing and speech processing*
[Power] Cepstrum	*"Spectrum of a spectrum" of spectral data, for analyzing sound & vibration of a signal to measure how the power of the signal is distributed over frequency & its density*
Mel Scale	*A scale relating a sound's perceived frequency to its actual frequency, which can be used to scale a sound close to how it would be perceived by the human ear*
Mel-frequency Cepstral Coefficients (MFCCs)	*Coefficients of Mel-frequency Cepstrals (MFCs); can be used to represent power cepstrum data on a short-term basis to measure the emotional properties of sound*
[Psychoacoustic] Valence	*Intrinsic attractiveness/aversiveness of an event, object or situation; measured as positive (attractive) or negative (aversive)*
[Psychoacoustic] Activation	*Arousal, stimulation response, measured by intensity from low to high; the stimulation of the cerebral cortex into a state of general wakefulness, or attention*
Epoch	*Evolutions where the dataset is split into training and testing portions to be evaluated before beginning the next epoch*
[Model] Weight	*The best parameters for training a model over time; transforms input data within the network's hidden layers*
[Model] Loss	*Also known as the cost or objective function, used to find the best weights; a quantitative measure of how much the predictions differ from the actual output (label)*
Confusion Matrix	*A specific table layout (graph) that allows visualization of the performance of an algorithm; used to compare actual versus predicted labels in classification learning*

- Daniel Szelogowski –

Emotion Recognition of the Singing Voice: Toward a Real-Time Analysis Tool for Singers

Resources:

- Full code, paper, and presentation: https://github.com/danielathome19/Sung-EmotioNN-Detector
- Research data overview: https://youtu.be/f9hs8TYyBxU
- Model and prototype tool demonstration: https://youtu.be/dsruK0GctG4

Sample Figures:

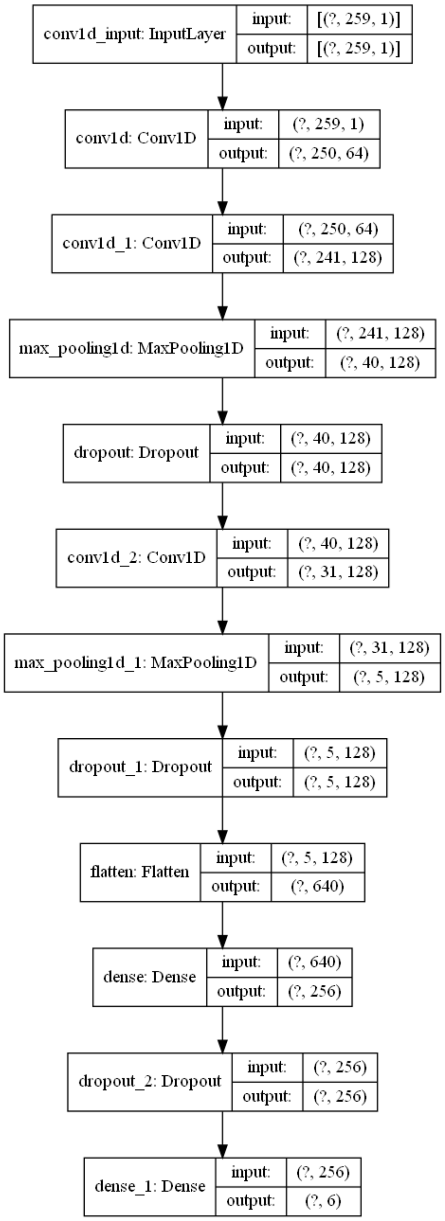


Figure 10: Layer diagram of the EmotioNN model (input order)

3. Sure on this Shining Night – Morten Lauridsen (split every 20 seconds)[33]

Non-Isolated Vocals		Isolated Vocals	
0	calm	0	calm
20	**calm**	20	**happy**
40	calm	40	calm
60	calm	60	calm
80	sad	80	sad
100	calm	100	calm
120	happy	120	happy
140	**sad**	140	**fear**
160	angry	160	angry
180	calm	180	calm
200	calm	200	calm
220	**calm**	220	**sad**
240	calm	240	calm
260	**calm**	260	**sad**

Table 4: Comparison of isolated versus non-isolated vocals split into 20-second-long segments

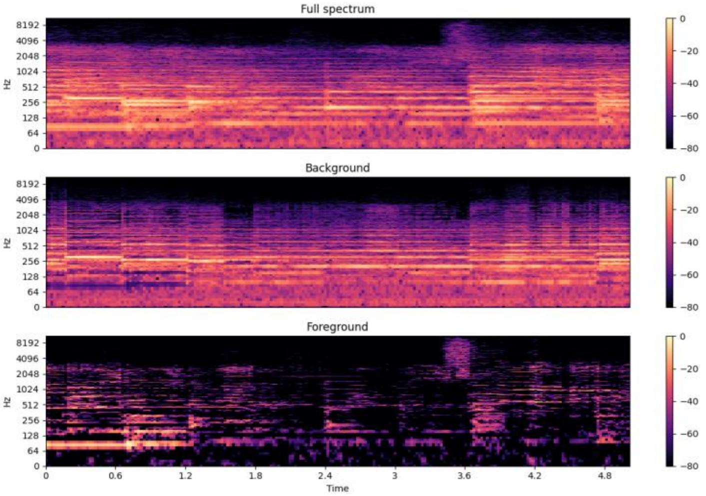
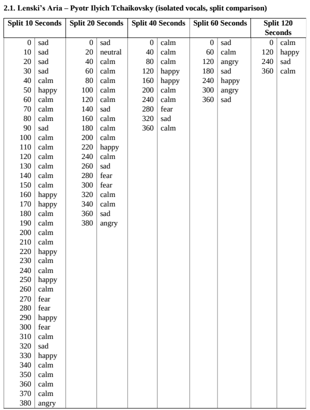


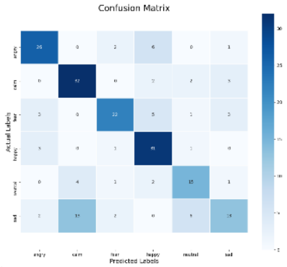
Figure 8: Spectral analysis of an SATB choir singing “Sure on this Shining Night” by Lauridsen (5 seconds)

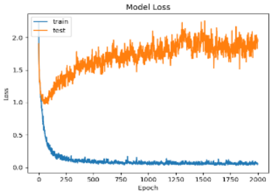


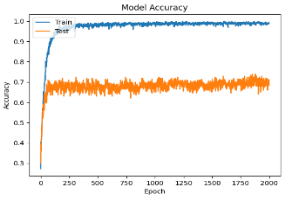
2.1. Lenski's Aria – Pyotr Ilyich Tchaikovsky (isolated vocals, split comparison)

Split 10 Seconds		Split 20 Seconds		Split 40 Seconds		Split 60 Seconds		Split 120 Seconds	
0	sad	0	sad	0	calm	0	sad	0	calm
10	sad	20	neutral	40	calm	60	calm	120	happy
20	sad	40	calm	80	calm	120	angry	240	sad
30	sad	60	calm	120	happy	180	sad	360	calm
40	calm	80	calm	160	happy	240	happy		
50	happy	100	calm	200	calm	300	angry		
60	calm	120	calm	240	calm	360	sad		
70	calm	140	sad	280	fear				
80	calm	160	calm	320	sad				
90	sad	180	calm	360	calm				
100	calm	200	calm						
110	calm	220	happy						
120	calm	240	calm						
130	calm	260	sad						
140	calm	280	fear						
150	calm	300	fear						
160	happy	320	calm						
170	happy	340	calm						
180	calm	360	sad						
190	calm	380	angry						
200	calm								
210	calm								
220	happy								
230	calm								
240	calm								
250	happy								
260	calm								
270	fear								
280	fear								
290	happy								
300	fear								
310	calm								
320	sad								
330	happy								
340	calm								
350	calm								
360	calm								
370	calm								
380	angry								

Table 3: Comparison of isolated vocals split into various segment lengths







- Daniel Szelogowski –