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

**Daniel Szelogowski**

University of Wisconsin - Whitewater

1

# Outline of presentation

- Introduction
- Related Work
- Technical Approach
- Implementation
- Evaluation
- Discussion
- Conclusion

2

# Introduction

- Current research overwhelmingly lacks study on emotional analysis of singing voices
  - Most recent studies have aimed to classify a pre-recorded audio clip of one emotion of spoken (and sung) voices
  - No real-time feedback tool
  - Little testing with noisy data
    - 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
- **Biofeedback** is scientifically proven to strengthen "mind-to-motor" coordination and help to influence production toward improvement
  - Machine learning is the key to creating such a tool at a non-enterprise level
  - By implementing a visual analysis tool for recognizing sung emotions in real-time, much like pitch and rhythm feedback apps, we can create a tool that would be extremely useful for musicians, actors, teachers, and eventually full ensembles
  - Emotional expression is hard! The voice alone has many features which exhibit emotion, including the breath and its fullness/depth, the volume, pressure, and stability of the voice, among others
    - This is instinctual for most people; replicating the same ability in artificial intelligence is challenging

3

# Introduction

- The most likely candidate for building an emotion-detecting system in the field of machine learning is the **neural network (NN)**
  - A form of artificial intelligence that seeks to replicate biological learning in the brain through the development of neurons and synaptic connections
  - One specialized form of NN, known as a **Convolutional Neural Network (CNN)**, is a type of deep-learning neural network primarily applied to analyzing visual imagery for image/video recognition and classification, and natural language processing
    - This type of artificial intelligence is inspired by biological processes, wherein the connectivity pattern between neurons resembles the organization of the animal visual cortex
    - To train these types of networks, a large, consistent dataset is especially necessary for creating a reliable and accurate prediction and classification system
      - Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): database of 24 professional actors (12 male, 12 female) containing audio and video recordings of 8 different emotions (neutral, calm, happy, sad, angry, fearful, disgust, surprise) spoken and sung with face and voice expressions

4

# Related Work

- A great amount of experimentation and analysis has been performed on speech emotion recognition – not only from a machine learning perspective but also purely scientific through acoustic and psychological studies with valence (intrinsic attractiveness/aversiveness) and activation (arousal, stimulation response) analysis
  - **Spectral analysis** – a type of visual-based audio analysis used in measuring the distribution of acoustic energy across frequencies
    - Great for biofeedback
    - Provides a means of either analyzing energy distribution in speech/sung sound (through the **Fast Fourier Transform, FFT**), including voice harmonics, or estimating the vocal tract filter that shaped the sound (**Linear Predictive Coding, LPC**)
      - Applying the inverse FFT to an audio signal returns the **cepstrum** of the spectral data, or "spectrum of a spectrum" allowing us to measure how the power of the signal is distributed over frequency and its density through the **power cepstrum**.
        - On a short-term basis, known as the **Mel-frequency cepstrum (MFC)**
        - Measures the sound and vibration of a signal
        - **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
          - Our ears perform FFT to transform audio information into sound!

5

# Related Work

- Recognizing emotions in normal speech is a process carried out by the amygdala, a region of the brain within the medial temporal lobe involved in emotional processes as part of the limbic system: the body's own neural network that handles various aspects of memory and emotion
  - The amygdala responds to two properties of emotions: valence, measured as positive (attractive) or negative (aversive), and intensity, measured from low to high
  - Thus, a CNN is the most effective machine learning model to represent the amygdala given its classification ability
  - To recreate the analytical process of the amygdala within the CNN, one way to measure the emotional properties of sound is through the coefficients of the MFC measurements of an audio clip, known as Mel-frequency cepstral coefficients (MFCCs) which also assists in better representing compressed audio on the Mel scale
- One aspect that may potentially affect the outcome of the CNN is the noise of the data it trains and/or tests on
  - This model was trained on noiseless data to ensure keep information received from the MFCCs of the audio clean and reliable
    - This is an unrealistic expectation for a biofeedback app and is difficult to achieve from a field scenario
  - Vocal isolation continues to be an ongoing battle of its own, especially for noisy data such as an accompanied voice (even with complex machine learning algorithms)
    - Recognizing emotional expression and coloring has also proven to be equally, if not more challenging than vocal isolation
- Recognizing emotional expression and coloring has also proven to be equally, if not more challenging than vocal isolation
  - Recent attempts have included analyzing the acoustic features of a vocal signal and deriving a result manually through mathematical analysis, rather than a NN
  - Another recent approach was performed by the measurement of Signal-to-Noise Ratio (SNR) levels and their correlation to emotional valence, comparing mathematically predicted emotion values to actual perceived emotions in humans

6

# Approach

- The backend system presented in this research will consist of three primary components:
  - CNN model
  - Vocal isolator
  - WAV-file audio recorder/splitter
- CNN will be trained on the RAVDESS dataset which contains 1,012 sung text examples across the various actors
  - Expressions: neutral, calm, happy, angry, sad, or fear
- **WAV-file** system will record the user's microphone as a WAV file and feature a system to divide an audio file into user-defined seconds-long segment files for evaluation and real-time biofeedback testing, live and historically
- Vocal isolator will use spectral data obtained from an audio file to separate voice from any accompaniment and/or background noise through FFT and filtering
- Accuracy is difficult due to the lack of datasets designed for such testing, but approximately 75% accuracy is a reasonable goal

7

# Implementation

- Architecture of the model built in the Python programming language using Keras/Tensorflow libraries
  - Librosa library to extract the Mel spectrogram and obtain the cepstral data
  - Following initial training, **hyperparameter** tuning (parameters with values that control a model's learning) to obtain the most optimized model
- After training on all voice actor audio clips for a user-defined number of **epochs** (evolutions where the dataset is split into training and testing portions to be evaluated before the next epoch), the NN will be prepared to predict new audio files outside of the training set
  - Only the best model is saved, i.e., when an epoch finds a new maximum accuracy (improvement) score
- Vocal isolation method: **REPET-SIM** method – small FFT window overlap and converting non-local filters into soft masks using Wiener filtering
  - Audio file converted into a spectrogram, filter applied to aggregate and constrain similar audio frames, reduce the bleed of the vocal and instrumental/accompaniment (noise) masks, then separate the masks into background (noise) and foreground (voice) spectrums
    - After separating spectrums, the foreground will contain the newly filtered audio with vocals

8

# Implementation

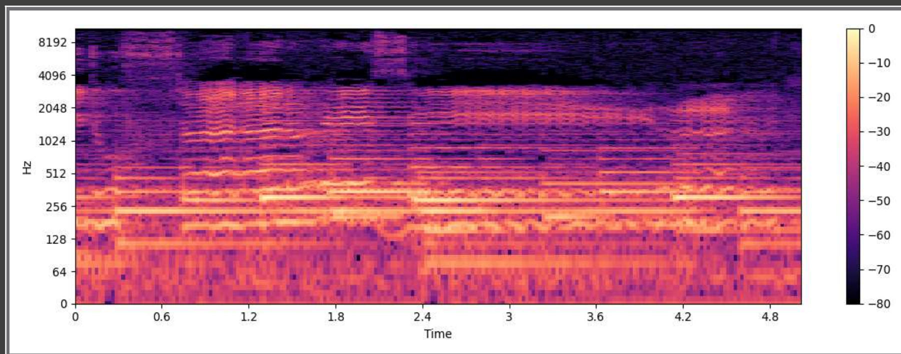


**Figure 1:** Spectral analysis of a baritone voice singing "Allerseelen" by Richard Strauss with piano accompaniment across five seconds. The "hot" waves on the graph, seen especially between 128-1024 Hz, represent the vocal frequencies.

9

# Implementation

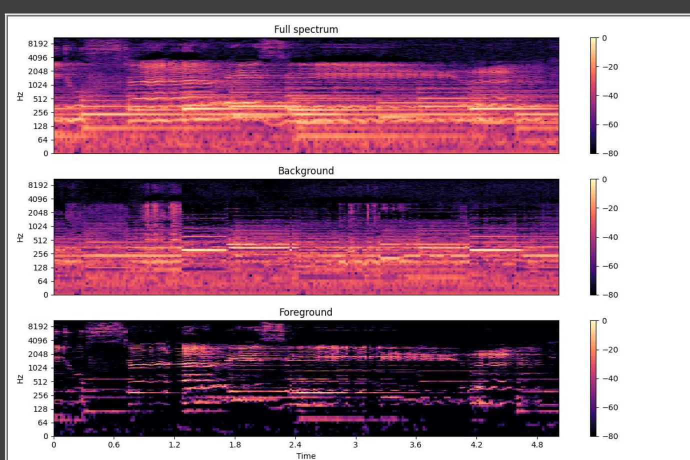


**Figure 6:** Spectral analysis of a baritone voice singing "Allerseelen" by Richard Strauss. The background layer represents noise and the foreground layer represents the voice.

10

# Implementation

a)
b)

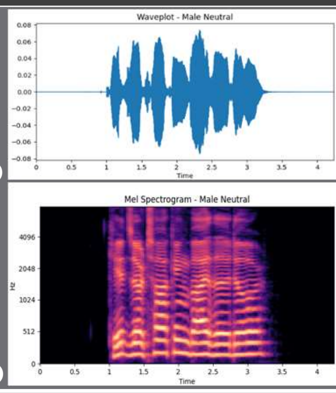


**Figure 2:** Audio extraction of a 'neutral' male voice as waveform **(a)** & Mel spectrogram **(b)**

Model: "sequential"

| Layer (type) | Output Shape | Param # |
|---|---|---|
| conv1d (Conv1D) | (None, 250, 64) | 704 |
| conv1d_1 (Conv1D) | (None, 241, 128) | 82048 |
| max_pooling1d (MaxPooling1D) | (None, 40, 128) | 0 |
| dropout (Dropout) | (None, 40, 128) | 0 |
| conv1d_2 (Conv1D) | (None, 31, 128) | 163968 |
| max_pooling1d_1 (MaxPooling1 | (None, 5, 128) | 0 |
| dropout_1 (Dropout) | (None, 5, 128) | 0 |
| flatten (Flatten) | (None, 640) | 0 |
| dense (Dense) | (None, 256) | 164096 |
| dropout_2 (Dropout) | (None, 256) | 0 |
| dense_1 (Dense) | (None, 6) | 1542 |

Total params: 412,358
Trainable params: 412,358
Non-trainable params: 0

**Figure 3:** EmotioNN model architecture as CNN

11

# Evaluation

- The final model was trained to 2,000 epochs, achieving approximately 73% accuracy from the test data
- The **loss** from the model – also known as the cost or objective function, used to find the best parameters, known as **weights** for the model – was much greater in the test validation than the training validation
  - While accuracy scaled relatively proportionally between testing and training, loss diverged away from minimization compared to the training model and continued to grow exponentially over time
  - While this may appear problematic, an entropy of approximately 1.8 out of six classes gives a loss of 0.3, or roughly 30% – inverse to the accuracy value
  - To see the actual loss based on how the model predicts classes of emotions, a **confusion matrix** can be used to compare actual versus predicted labels
  - While this level of accuracy is decent, a much more complex architecture is necessary to train a better performing model in the future, especially including training on acoustic properties among the MFCCs

12

# Evaluation

a) b) c)

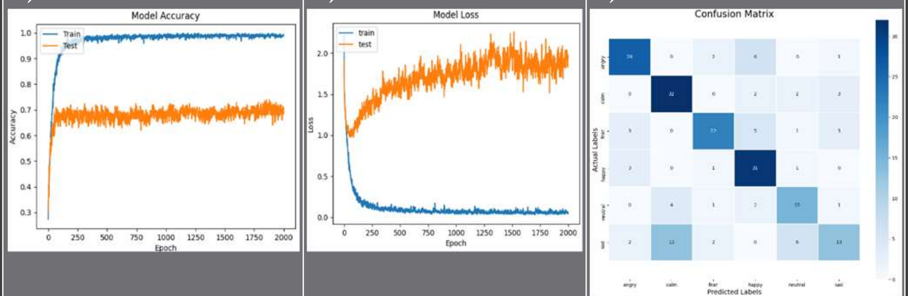


**Figure 4:** Final trained model after 2,000 epochs: accuracy **(a)**, loss **(b)**, and confusion matrix **(c)**

13

# Evaluation

| Non-Isolated Vocals (10 seconds) | | Isolated Vocals (10 seconds) | | Non-Isolated Vocals (20 seconds) | | Isolated Vocals (20 seconds) | |
|---|---|---|---|---|---|---|---|
| 0 | calm | 0 | calm | 0 | calm | 0 | calm |
| 10 | **calm** | 10 | **sad** | 20 | **sad** | 20 | **fear** |
| 20 | **calm** | 20 | **sad** | 40 | calm | 40 | calm |
| 30 | calm | 30 | calm | 60 | sad | 60 | sad |
| 40 | calm | 40 | calm | 80 | **calm** | 80 | **sad** |
| 50 | **calm** | 50 | **sad** | 100 | calm | 100 | calm |
| 60 | **calm** | 60 | **sad** | 120 | calm | 120 | calm |
| 70 | **calm** | 70 | **sad** | 140 | **sad** | 140 | **calm** |
| 80 | **calm** | 80 | **sad** | 160 | calm | 160 | calm |
| 90 | calm | 90 | calm | 180 | calm | 180 | calm |
| 100 | calm | 100 | calm | 200 | angry | 200 | angry |
| 110 | **fear** | 110 | **happy** | 220 | **happy** | 220 | **angry** |
| 120 | calm | 120 | calm | 240 | **fear** | 240 | **calm** |
| 130 | calm | 130 | calm | | | | |
| 140 | calm | 140 | calm | | | | |
| 150 | happy | 150 | happy | | | | |
| 160 | calm | 160 | calm | | | | |
| 170 | **calm** | 170 | **sad** | | | | |
| 180 | **neutral** | 180 | **sad** | | | | |
| 190 | happy | 190 | happy | | | | |
| 200 | angry | 200 | angry | | | | |
| 210 | **happy** | 210 | **calm** | | | | |
| 220 | happy | 220 | happy | | | | |
| 230 | **calm** | 230 | **sad** | | | | |
| 240 | **fear** | 240 | **calm** | | | | |

**Example output:**
Der Erlkönig – Schubert (split every 10 and 20 seconds)

**Table 6:** Comparison of isolated versus non-isolated vocals split into 10- and 20-second-long segments

14

# Discussion

- While the accuracy of the model may appear questionable, most results are surprisingly correct – albeit subjectively in some instances
  - Tests were performed in segments (rather than the entire audio file) with and without isolating the vocals from the accompaniment of various art songs, including one choral piece and one purely instrumental piece; however, making these changes did not appear to make a visible difference in how the model predicted various segments of a song
- Vocals, both isolated and accompanied, showed nearly identical results so long as the segment contains a voice – silence was typically the only unidentical factor and this may be for one of two reasons in particular:
  - Either the silence contained residual harmony (even minuscule) from the accompaniment or noise and the isolated version of the file did not, or,
  - The file is comprised of just the accompaniment which became silent after isolating the vocals
  - As well, changing the length of the segments also did not appear to make a difference in the perceived emotion either, as the smaller splits only added to the precision of the model's accuracy
  - Of course, as expected, the model did not fare well attempting to classify emotions of instrumental music given its lack of training and more precise **psychoacoustical** parameters
- One major flaw of the model appears to be its evolutionary progression; once the model achieves approximately 70% accuracy, it seldom improves
  - In an attempt to train the CNN 10,000 epochs, the model eventually failed after over 7500 epochs and its accuracy flatlined at less than 15% accuracy and never recovered, though the loss continued to change over time
  - Strangely, these weights were saved as the "new best" by the model anyway, and the CNN appeared to predict "angry" for every given piece of data
  - Additionally, in the final model, classifying in real-time also causes the fully trained (73% accurate) model to predict "angry" as well, but classifying these files later on yields normal results

15

# Discussion

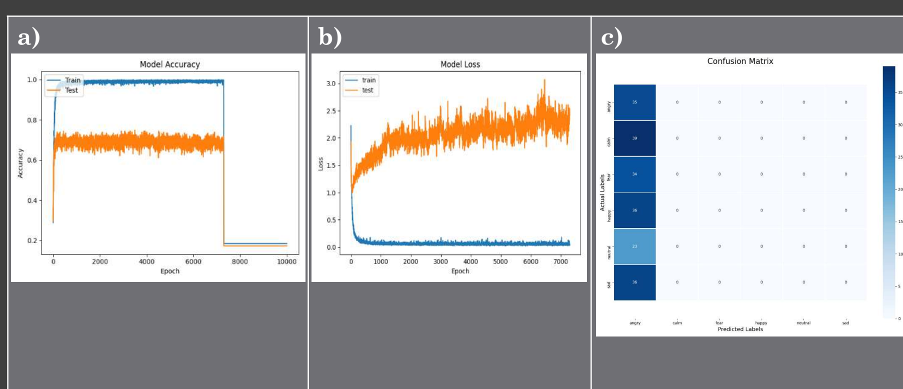


**Figure 5:** Early trained model to 10,000 epochs: accuracy **(a)**, loss **(b)**, and confusion matrix **(c)**

16

# Conclusion

- This paper focused on the creation of a CNN with the ability to accurately recognize emotions in the singing voice using MFCCs as the primary feature, as well as laying out the developmental plans for a future mobile (or desktop) application that utilizes this type of model for providing biofeedback
  - By training the model on pure vocal audio, the model was able to create a working classification memory even when data is noisy or includes instrumental accompaniment
  - Ideally, the model should be nearly or just as accurate without vocal isolation as with it – this appeared to be the case even with the current final model, fortunately
- In the future, this type of model and architecture could be expanded upon and utilized as the backend system of a much more complex piece of software – hopefully used in a biofeedback app as intended, providing visual feedback of both real-time and historical data for use in private music lessons
  - With finer tuning and the inclusion of more acoustic properties, this model has the potential to become much more accurate in a more refined architecture
  - Hopefully, a larger sung-emotion dataset will be created as well soon
  - Eventually, more training may be done on the analysis of choral music and, when a dataset permits, analyzing instrumental music both individually and for ensembles

17

# Conclusion

- With a much more expansive and precise dataset, a future model could also be trained to recognize a wider array of emotions in both singing and speaking voices, as well as recognizing emotions in the breath, language (text – through **natural language processing**), face, and potentially body language, as a means of creating a near-true neural network amygdala
- Given a more accurate and reliable, this technology could serve voice teachers and professionals alike as a means of training emotional expression with live feedback – especially in the case of rehearsing a very expressive work such as an aria or art song
  - With a reliable model, a teacher may have a mobile application which provides live feedback during a lesson or ensemble rehearsal, or an individual singer may use the biofeedback during practice sessions.

18

# References


- Beethoven: Coriolan Overture, Op. 62 (with Score). YouTube. YouTube, 2019. https://www.youtube.com/watch?v=lHM8uekMxls.
- Der Erlkönig: Franz Schubert, Philippe Sly: Bass-Baritone, Maria Fuller: Piano. YouTube. YouTube, 2011. https://www.youtube.com/watch?v=jZxzz-N3oxM.
- Dietrich Fischer-Dieskau; "Allerseelen"; Richard Strauss. YouTube. YouTube, 2021. https://www.youtube.com/watch?v=c9xBSTAHcW4.
- Edwin, Robert. “I Second That Emotion.” NATS - Journal of Singing 44, no. 4 (March 1988): 30–1.
- Erickson, Heidi M. “Mobile Apps and Biofeedback in Voice Pedagogy.” NATS - Journal of Singing 77, no. 4 (February 17, 2021): 485–99.
- Eyben, Florian, Gláucia L. Salomão, Johan Sundberg, Klaus R Scherer, and Björn W Schuller. “Emotion in the Singing Voice—A Deeper Look at Acoustic Features in the Light of Automatic Classification.” EURASIP Journal on Audio, Speech, and Music Processing 2015, no. 1 (June 30, 2015). https://doi.org/10.1186/s13636-015-0057-6.
- FitzGerald, Derry. “Vocal Separation Using Nearest Neighbours and Median Filtering.” IET Irish Signals and Systems Conference (ISSC 2012), June 28, 2012. https://doi.org/10.1049/ic.2012.0225.
- Helding, Lynn. “Emotion and Empathy: How Voice Can Save the Culture.” NATS - Journal of Singing 73, no. 4 (March 2017): 429–33.
- IBM Cloud Education. “What Are Convolutional Neural Networks?” IBM, October 20, 2020. https://www.ibm.com/cloud/learn/convolutional-neural-networks.
- Insight, Analytics. “Speech Emotion Recognition (SER) through Machine Learning.” Analytics Insight, February 24, 2021. https://www.analyticsinsight.net/speech-emotion-recognition-ser-through-machine-learning/.
- Kosaka, Muriel. “Speech_Emotion_Recognition.” GitHub, August 4, 2020. https://github.com/mkosaka1/Speech_Emotion_Recognition.
- Livingstone, Steven R., and Frank A. Russo. “The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A Dynamic, Multimodal Set of Facial and Vocal Expressions in North American English.” PLOS ONE 13, no. 5 (May 16, 2018). https://doi.org/10.1371/journal.pone.0196391.
- McCoy, Scott. The Basics of Voice Science & Pedagogy. Gahanna, OH: Inside View Press, 2020.
- Miller, Richard. “Learning to Portray Emotion.” NATS - Journal of Singing 57, no. 5 (May 2001): 31–3.
- Norton, Michael P., and Dennis G. Karczub. Fundamentals of Noise and Vibration Analysis for Engineers. Cambridge: Cambridge Univ. Press, 2007.
- Parada-Cabaleiro, Emilia, Maximilian, Schmitt, Anton, Batliner, Simone, Hantke, Giovanni, Costantini, Klaus, Scherer, and Björn, Schuller. “Identifying Emotions in Opera Singing: Implications of Adverse Acoustic Conditions.” 19th International Society for Music Information Retrieval Conference. Paris, France, 2018.

(December 2018)

- de Pinto, Marco Giuseppe, Marco Polignano, Pasquale Lops, and Giovanni Semeraro. “Emotions Understanding Model from Spoken Language Using Deep Neural Networks and Mel-Frequency Cepstral Coefficients.” 2020 IEEE Conference on Evolving and Adaptive Intelligent Systems (EAIS), June 23, 2020. https://doi.org/10.1109/eais48028.2020.9122698.
- “Power Spectral Density: What Is It and How Is It Measured?” Safe Load Testing Technologies, January 4, 2021. https://www.safeloadtesting.com/power-spectral-density/.
- Rafii, Zafar and Bryan Pardo. “Music/Voice Separation Using the Similarity Matrix.” 13th International Society for Music Information Retrieval (ISMIR) Conference, 2012.
- Rolando Villazon - Kuda, Kuda - Lensky's Aria. YouTube. YouTube, 2007. https://www.youtube.com/watch?v=L2_GBhFtM2U.
- Salomão, Gláucia L., Johan Sundberg, and Klaus Scherer. “Emotional Coloring of the Singing Voice.” Pan-European Voice Conference: Pevoc 11. Firenze University Press, 2015. (August 31, 2015).
- Salzman, C. Daniel. "Amygdala." Encyclopedia Britannica. https://www.britannica.com/science/amygdala.
- Singh, Jyotika. “An Introduction to Audio Processing and Machine Learning Using Python.” Opensource.com, September 19, 2019. https://opensource.com/article/19/9/audio-processing-machine-learning-python.
- Sure on the Shining Night (Morten Lauridsen). YouTube. YouTube, 2018. https://www.youtube.com/watch?v=PgiKeRt0IUo.
- Szelogowski, Daniel J. “Sung-EmotioNN-Detector.” Github, April 5, 2020. https://github.com/danielathome19/Sung-EmotioNN-Detector.
- Szelogowski, Daniel J. Emotion Detection in the Singing Voice Art Song Research Data Overview. YouTube. YouTube, 2021. https://www.youtube.com/watch?v=f9hs8TYyBxU.
- Wah, Lin Edward Kin. “Singing Voice Analysis in Popular Music Using Machine Learning Approaches,” 2018.
- Wah, Lin Edward Kin, B. T. Balamurali, Enyan Koh, Simon Lui, and Dorien Herremans. “Singing Voice Separation Using a Deep Convolutional Neural Network Trained by Ideal Binary Mask and Cross Entropy.” Neural Computing and Applications 32, no. 4 (December 13, 2018): 1037–50. https://doi.org/10.1007/s00521-018-3933-z.
- Wood, Sidney. “Spectral Analysis.” Welcome to SWPhonetics, November 12, 2013. https://swphonetics.com/praat/tutorials/spectral-analysis/.
- Xu, Min, Ling-Yu Duan, Jianfei Cai, Liang-Tien Chia, Changsheng Xu, and Qi Tian. “HMM-Based Audio Keyword Generation.” Advances in Multimedia Information Processing - PCM 2004, 2004, 566–74. https://doi.org/10.1007/978-3-540-30543-9_71.


19

Thank you!

20