Project Notes:

Project Title: Using Machine Learning Techniques to Improve Computer-Based Assessment of Musical Performances
Name: Sumanth Sura

Note Well: There are NO SHORT-cuts to reading journal articles and taking notes from them. Comprehension is paramount. You will most likely need to read it several times so set aside enough time in your schedule.

Contents:

Knowledge Gaps:

This list provides a brief overview of the major knowledge gaps for this project, how they were resolved and where to find the information.

Knowledge Gap	Resolved By	Information is located	Date resolved
How to incorporate video data into musical assessment	Unresolved	N/A	N/A
How neudesin is formed	It is NENF +/+ → derived from neurons	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4436896/	08/29
How to generate spectrograms from DTFTs	They are collated together.	W. Jenkins and M. Desai, "The discrete frequency Fourier transform," in IEEE Transactions on Circuits and Systems, vol. 33, no. 7, pp. 732-734, July 1986, doi: 10.1109/TCS.1986.1085978.	12/20

Literature Search Parameters:

These searches were performed between (Start Date of reading) and XX/XX/2019.
List of keywords and databases used during this project.

Database/search engine	Keywords	Summary of search
IEEE	“Music analysis”	Found documents on musical document layout analysis. This information will be important for interpreting the “ideal input that the user needs”.
IEEE	“Song recognition”	Yielded results about recognizing emotion in music. This fits right up the alley of my project because it quantifies the less concrete aspects of music.
IEEE	“Activation functions”	Gave information about hyperbolic tangent activation functions (tanh) compared to sigmoid functions. Choosing an efficient activation function is essential for having a fast-running model.

Article TEMPLATE

Article notes should be on separate sheets

KEEP THIS BLANK AND USE AS A TEMPLATE

Source Title	
Source citation (APA Format)	
Original URL	
Source type	
Keywords	
Summary of key points + notes (include methodology)	
Research Question/Problem/ Need	
Important Figures	
VOCAB: (w/definition)	
Cited references to follow up on	
Follow up Questions	

Article Notes #1: Perceived Intensity and Discrimination Ability for Lingual Electrotactile Stimulation Depends on Location and Orientation of Electrodes

Article notes should be on separate sheets

Source Title	https://www.frontiersin.org
Source citation (APA Format)	Moritz Jr., J., Turk, P., Williams, J. D., & Stone-Roy, L. M. (1AD, January 1). Perceived intensity and discrimination ability for lingual electrotactile stimulation depends on location and orientation of electrodes. Frontiers. Retrieved August 21, 2022, from https://www.frontiersin.org/articles/10.3389/fnhum.2017.00186/full
Original URL	https://www.frontiersin.org/articles/10.3389/fnhum.2017.00186/full
Source type	Academic paper
Keywords	Sensory substitution, tongue, somatosensation, receptive fields, bioengineering, human, psychophysics
Summary of key points + notes (include methodology)	Researchers took 25 participants and subjected them to different vibrations on different parts of the tongue. They noted that the anterior medial region of the tongue was most noticeable when electronically excited. It is also the most intensely stimulated. The tongue stimulation device was created for this experiment. It is a black box with electronic sensors on one end and wires attached to two sides on the other end.
Research Question/Problem/ Need	Can humans detect electric pulses on different parts of the tongue, and what position/orientation gives the most effective response? This experiment's purpose is to help develop the method of enhancing other senses through stimulation
Important Figures	Joel Moritz Jr. - he works in robotics at Colorado State University
VOCAB: (w/definition)	somatosensation - the brain's ability to sense physical outside stimuli Innervation - the state of containing nerves Electrotactile stimulation - the process of passing an electric current to the body (this is the action being done the entire experiment)

Cited references to follow up on	- Avivi-Arber, L., Martin, R., Lee, J. C., and Sessle, B. J. (2011). Face sensorimotor cortex and its neuroplasticity related to orofacial sensorimotor functions. Arch. Oral Biol. 56, 1440–1465. doi: 10.1016/j.archoralbio.2011. 04.005 - Bach-y-Rita, P. (2004). Tactile sensory substitution studies. Ann. N Y Acad. Sci. 1013, 83–91. doi: 10.1196/annals.1305.006 Bach-y-Rita, P., Collins, C. C., Saunders, F. A., White, B., and Scadden, L. (1969). Vision substitution by tactile image projection. Nature 221, 963–964.
Follow up Questions	- Where does the applied electric current flow, and what path does it take to the brain? - Is this path longer in the anterior medial region of the tongue? - Is there a connection between the hemisphere of the tongue that is stimulated more strongly and the handedness of the participant?

Article Notes #2: Neudesin is involved in anxiety behavior: structural and neurochemical correlates

Article notes should be on separate sheets

Source Title	https://www.frontiersin.org
Source citation (APA Format)	Novais, A., Ferreira, A. C., Marques, F., Pêgo, J. M., Cerqueira, J. J., David-Pereira, A., … Sousa, J. C. (2013). Neudesin is involved in anxiety behavior: structural and neurochemical correlates. Frontiers in Behavioral Neuroscience, 7. doi:10.3389/fnbeh.2013.00119
Original URL	https://www.frontiersin.org/articles/10.3389/fnbeh.2013.00119/full
Source type	Academic Journal
Keywords	neudesin, anxiety, dopamine, ventral hippocampus, bed nucleus of the stria terminalis
Summary of key points + notes (include methodology)	The scientists used chemically altered mice (Nenf +/+) which was achieved through cross-breeding to ensure the existence of the Nenf protein positive gene. They also used mice with the Nenf -/- gene (normal mice). They tested several behaviors (their urgency to eat, their behavior in light vs. dark environments, and how quickly they reacted to sudden stimuli. Here were each of the experiments done on the mice: - **Open field**: The mice were placed in a 43.2cm side length square room and given 5 minutes to explore it. Researchers recorded total distance traveled, and time/distance spent in the center vs the boundaries. - **Forced swim test**: The mice swam in identical conditions on two different days and their behavior was recorded. They were placed in a glass cylinder at a depth of 30cm in 24°C conditions. - **Morris water maze**: This task tested the mice's spatial reference memory. They were placed in a circular white pool and a plexiglas platform was placed inside each quadrant. There were 4 trials each day (1 for each quadrant) for 4 days. On the fifth day they switched the position of the goal to the opposite hemisphere and noticed how long it took for the mice to adjust. - **Elevated plus maze**: This tested anxiety by placing the mouse on a raised platform and allowing them to either go into two opposite light rooms or two opposite dark rooms. The mice was placed at the top and was allowed to explore the

	space for 5 minutes. - **Acoustic startle**: The mice were placed into a chamber, and a high frequency speaker was placed 33cm above the chamber. The mice got habituated to the scenario for 24 hours, and then a sound was played. The mice's movements were captured by a piezoelectric element, corresponding to a given startle reflex. A higher startle reflex corresponds to a higher level of anxiety. - **Light/dark box**: The mice were placed between two compartments, one light and one dark, and then allowed to travel between the two for 10 minutes. The amount of time spent in each half was recorded by an infrared system. - **Novelty suppressed feeling**: The mice were not allowed to eat for 24hrs and then were placed in the corner of a compartment. After 10 minutes, a single pellet of food was placed in the center of the compartment. Then, the mice were given predetermined amounts of food and then their intake was checked after 5, 10, and 15 minutes to check how strong their appetite was.
Research Question/Problem/ Need	How does the protein neudesin affect behavior (particularly motion) and cognitive ability? This experiment's purpose is to try to find the main chemical processes that cause anxiety. Determining these proteins will give scientists an idea of how to prevent them and lessen the effects of chronic anxiety in humans.
Important Figures	Ashley Novais - studies Microfabrication and Exploratory nanotechnology at International Iberian Nanotechnology Laboratory
VOCAB: (w/definition)	Ventral hippocampus - part of the brain that controls emotional/anxiety behavior Neudesin - referred to as "Nenf", long form = neuron derived neurotrophic factor, has a mass of 21000 amu and has 171 amino acids. (Source: Kimura 2005) Piezoelectric - the ability of a material to produce an electric current when under mechanical stress
Cited references to follow up on	- Kimura, I., Yoshioka, M., Konishi, M., Miyake, A., and Itoh, N. (2005). Neudesin, a novel secreted protein with a unique primary structure and neurotrophic activity. J. Neurosci. Res. 79, 287–294. doi: 10.1002/jnr.20356 (← "Kimura 2005") - Kimura, I., Nakayama, Y., Konishi, M., Terasawa, K., Ohta, M., Itoh, N., et al. (2012). Functions of MAPR (membrane-associated progesterone receptor) family members as heme/steroid-binding proteins. Curr. Protein Pept. Sci. 13, 687–696. doi: 10.2174/1389203711209070687

Follow up Questions	- Does neudesin change the shape of the hippocampus and the amygdala permanently? Are any changes to the brain permanent/is there any long-term effect? - What additional proteins are used alongside neudesin to cause anxiety? What would be the result of these experiments if 2 or more of these proteins along with neudesin were tested? - How does neudesin affect the mice’s interaction with other mice? Can the presence of neudesin in neurons affect human social behavior?

Article Notes #3: How the brain constructs the world

Source Title	ScienceDaily
Source citation (APA Format)	Scuola Internazionale Superiore di Studi Avanzati. (2018, February 9). How the brain constructs the world: Insights into processes underlying the integration and storage of sensory signals. ScienceDaily. Retrieved December 20, 2022 from www.sciencedaily.com/releases/2018/02/180209100639.htm
Original URL	https://www.sciencedaily.com/releases/2018/02/180209100639.htm
Source type	General Article
Keywords	Neuroscience, brain, senses
Summary of key points + notes (include methodology)	This article talks about recent developments (as of 2018) into how the brain takes sensory input and creates that input into our perspective of the outside world. The main discovery has to do with the usage of the posterior parietal cortex. This part of the brain merges the signals from different sensory inputs. I would imagine that it is the logic part of the brain that would connect, say, vision and hearing. For example, upon seeing a violin (a visual input) I might be reminded of a particular concerto (an auditory memory). Scientists also believe that the posterior parietal cortex controls the memory part of the brain. A study was conducted with mice involving interacting with an object via touch, sight, and both. The researchers concluded that the mice used neural connections that related to the object, not what sense they were using. They also noticed (as expected) that the mice that used both touch and sight got a stronger neural connection. In another study, published in Nature, Athena Akrami and co-authors studied the way that sensory memories are stored in the brain. When you hear a stimuli and move toward it, for example, your brain has to update how close it thinks you are to that source. Akrami has been researching the pathways in the brain that store the current location of the stimuli, and how often it changes. This article talks about recent developments (as of 2018) into how the brain takes sensory input and creates that input into our perspective of the outside world. The main discovery has to do with the usage of the posterior parietal cortex. This part of the brain merges the signals from different sensory inputs. I would imagine that it is the logic part of the brain that would connect, say, vision and hearing. For example, upon seeing a violin (a visual input) I might be reminded of a particular concerto (an auditory memory). Scientists also believe that the posterior parietal cortex controls the memory

	part of the brain. A study was conducted with mice involving interacting with an object via touch, sight, and both. The researchers concluded that the mice used neural connections that related to the object, not what sense they were using. They also noticed (as expected) that the mice that used both touch and sight got a stronger neural connection. In another study, published in Nature, Athena Akrami and co-authors studied the way that sensory memories are stored in the brain. When you hear a stimuli and move toward it, for example, your brain has to update how close it thinks you are to that source. Akrami has been researching the pathways in the brain that store the current location of the stimuli, and how often it changes.
Research Question/Problem/ Need	How does the brain process phenomena, and in turn, how is this processing system connected to memory?
Important Figures	Nader Nikbakht - Brain Research @ MIT. (postdoctoral research associate)
VOCAB: (w/definition)	*Posterior parietal cortex* - region of the brain associated with special awareness and eye movements.
Cited references to follow up on	Nader Nikbakht, Azadeh Tafreshiha, Davide Zoccolan, Mathew E. Diamond. Supralinear and Supramodal Integration of Visual and Tactile Signals in Rats: Psychophysics and Neuronal Mechanisms. Neuron, 2018; 97 (3): 626 DOI: 10.1016/j.neuron.2018.01.003 Athena Akrami, Charles D. Kopec, Mathew E. Diamond, Carlos D. Brody. Posterior parietal cortex represents sensory history and mediates its effects on behaviour. Nature, 2018; DOI: 10.1038/nature25510
Follow up Questions	1. How is this perception affected in blind people? 2. Other than sight, how do other senses use this short-term memory? Is the recording log in the human brain stronger in some senses than others? 3. How does Alzheimer’s alter the system proposed in the paper?

Article Notes #4: The brain predicts our perception of the outside world

Source Title	University of Liège
Source citation (APA Format)	X, S. (2007, July 10). The brain predicts our perception of the outside world. Medical Xpress - medical research advances and health news. Retrieved December 20, 2022, from https://medicalxpress.com/news/2007-07-brain-perception-world.html
Original URL	https://medicalxpress.com/news/2007-07-brain-perception-world.html
Source type	General Article
Keywords	Brain, fMRI, awareness, neurology
Summary of key points + notes (include methodology)	The human brain anticipates our perception of the outside world. For example, it is capable of predicting if we are going to perceive tactile stimulation of weak intensity or, on the contrary, if a more intense stimulation will be perceived more or less painfully. A study was published in July 2007 that outlines an experiment conducted by Dr. Melanie Boly and Dr. Steven Laureys. It shows that the spontaneous activity measured in certain parts of the brain have a direct influence over our conscious perception and our perception of the intensity of pain. 'Our brain is never really at rest, but science does not have a good understanding of how the spontaneous and continuous activity of our neurons influences our perception of the world. Our study contributes to lifting a corner of the veil over these mechanisms', the researchers state. The researchers used functional magnetic resonance imaging (fMRI) to conduct this study. The most fascinating part of this study is how it shows when humans are more sensitive to pain than other times. There are two parts to awareness: awakening (to the self, on the inside) and awareness of the environment. The researchers figured out that the neural connections that connect awareness to the self and awareness to the environment are different. This is why emotions are processed differently over time. The emotional effect that sleep has on our brain is powerful, since we can experience our most emotional events without feeling the emotion behind them. We can realize the logical effects of what

	happens in our daily lives, and make sense of how we can calm down our emotions. People who are deprived of sleep have greater trouble keeping track of their emotions, and consequently people who get sufficient sleep are better at controlling their emotions.
Research Question/Problem/ Need	How does the human brain interpret past experiences and make predictions on future behavior?
Important Figures	Dr. Steven Laureys - Belgian neurologist - studies human consciousness
VOCAB: (w/definition)	*fMRI* - functional magnetic resonance imaging - its purpose is to determine where blood flow is occurring in the brain to detect brain activity. Blood flows in different regions of the brain when a person is doing different things.
Cited references to follow up on	Owen, A. M., Coleman, M. R., Boly, M., Davis, M. H., Laureys, S., & Pickard, J. D. (2006). Detecting awareness in the vegetative state. *Science (New York, N.Y.)*, *313*(5792), 1402. https://doi.org/10.1126/science.1130197
Follow up Questions	1. What is the accuracy of using fMRI scans for understanding how the brain processes events? 2. How are intensity of emotions observed? Is it necessarily true that the amount of, say, serotonin flooding through our brains determine the extremity of happiness that a human feels? 3. Is it possible to redirect blood flow to different parts of the brain to make a person feel different things, once we detect finer patterns in blood flow that determine mood/behavior?

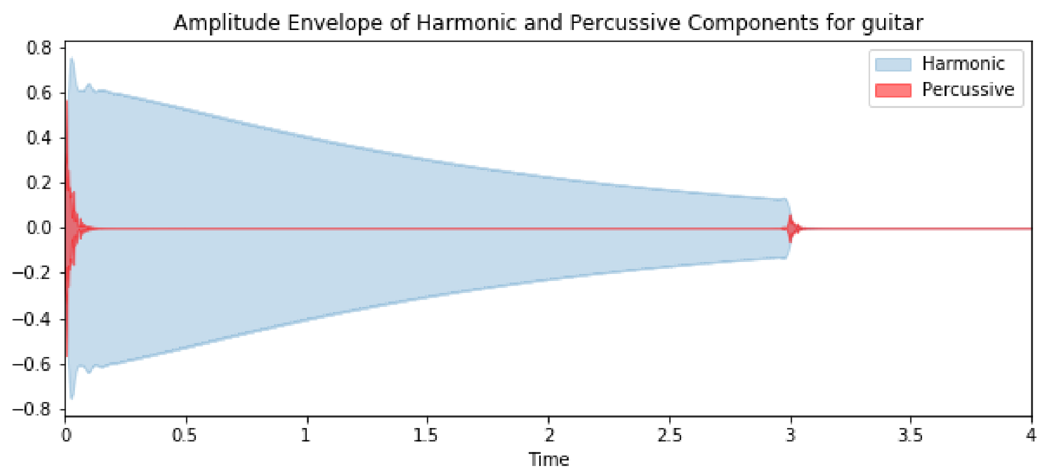
Article Notes #5: Automatic assessment of violin performance using dynamic time warping classification

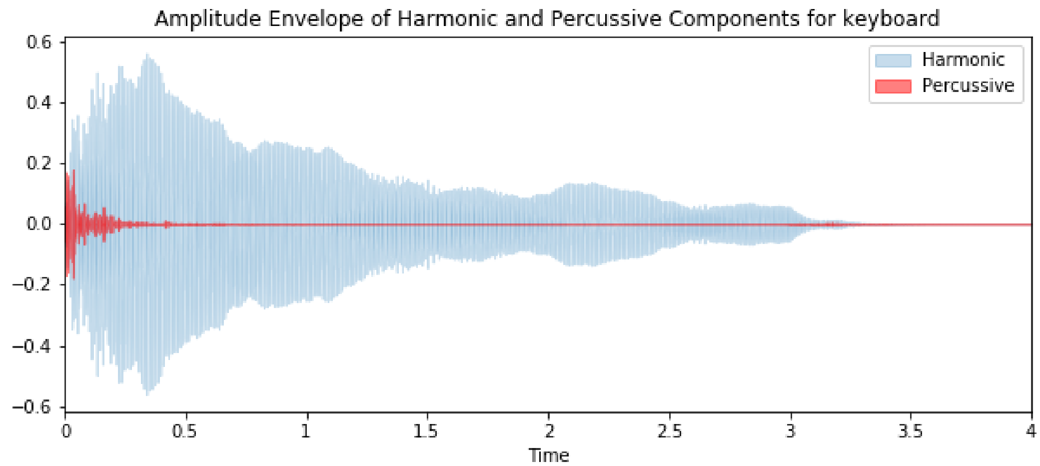
Source Title	https://ieeexplore.ieee.org
Source citation (APA Format)	Giraldo, S., Ortiga, A., Perez, A., Ramirez, R., Waddell, G., & Williamon, A. (n.d.). *Automatic assessment of violin performance using Dynamic Time Warping Classification*. IEEE Xplore. Retrieved September 27, 2022, from https://ieeexplore.ieee.org/document/8404556
Original URL	https://ieeexplore.ieee.org/document/8404556
Source type	Journal Article
Keywords	Violin, Automation, AI, machine learning, dynamic time warping
Summary of key points + notes (include methodology)	Most musical performance AI leaves to chance the tone quality and the style, which is sometimes highly debated and a computer cannot come up with intelligent information on this topic.
Research Question/Problem/ Need	How is it possible to get computers to understand different styles of music? - Phrasing (how can a computer detect phrases) - Style (how can the computer determine the emotion with which the player is playing? I. This goes back to the phrasing. If there are smooth transitions between phrases, then it might be a more smooth piece, but if the phrase changes are accented then it might indicate a mood change, etc. - How to determine when a given recording is better than an original recording? Have an objective baseline What are the criteria for an objective baseline?
Important Figures	Sergio Giraldo
VOCAB: (w/definition)	Semantic labeling: “the process of mapping attributes in data sources to classes in an ontology” (src: Pham - “Semantic labeling: a

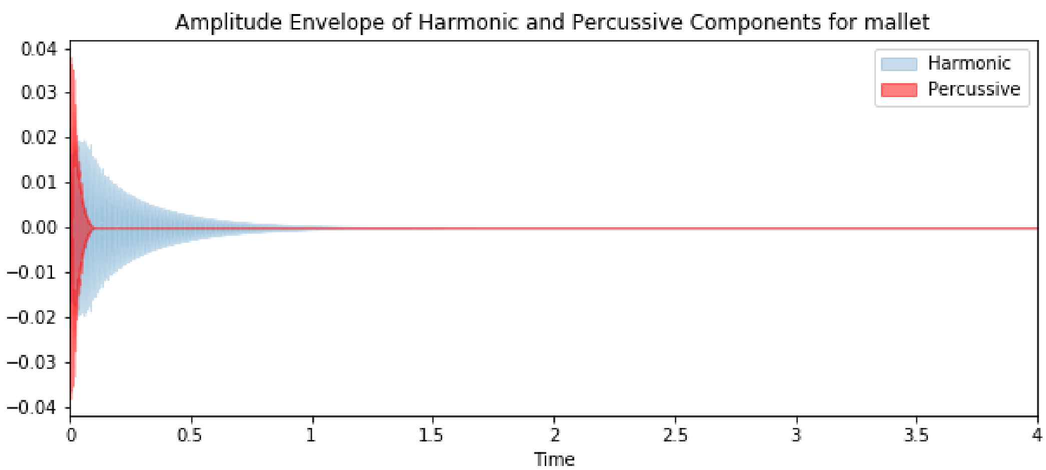
	domain-independent approach)
Cited references to follow up on	C. Dittmar, E. Cano, J. Abeßer, and S. Grollmisch, “Music information retrieval meets music education,” in Dagstuhl Follow-Ups, vol. 3. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2012. C. Raphael, “Music plus one and machine learning,” in Proceedings of the 27th International Conference on Machine Learning (ICML-10), 2010, pp. 21–28.
Follow up Questions	1. How can one use video data to enhance this? 2. What if a given recording is better than the master recording? 3. How does one determine the instrument used? Should the output be invalid?

https://medium.com/@nadimkawwa/can-we-guess-musical-instruments-with-machine-learning-afc8790590b8

Acoustic instruments have wider amplitude envelopes.




Amplitude Envelope of Harmonic and Percussive Components for keyboard
Harmonic
Percussive
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
0
0.5
1
1.5
2
2.5
3
3.5
4
Time


Amplitude Envelope of Harmonic and Percussive Components for mallet
Harmonic
Percussive
0.04
0.03
0.02
0.01
0.00
−0.01
−0.02
−0.03
−0.04
0
0.5
1
1.5
2
2.5
3
3.5
4
Time

Article Notes #6: Audio Music Monitoring: Analyzing Current Techniques for Song Recognition and Identification

Source Title	https://link.springer.com/
Source citation (APA Format)	Senevirathna, E. D. N. W., & Jayaratne, L. (2015, November 26). *Audio Music Monitoring: Analyzing current techniques for song recognition and identification - GSTF Journal on Computing (JOC)*. SpringerLink. Retrieved October 17, 2022, from https://link.springer.com/article/10.7603/s40601-014-0015-7
Original URL	https://link.springer.com/article/10.7603/s40601-014-0015-7
Source type	Journal Article
Keywords	Audio fingerprint; features extraction; wavelets; broadcast monitoring; Audio classification; Audio identification
Summary of key points + notes (include methodology)	. Music is attached to people since the day they were born. When a music repository grows, people face lots of challenges such as finding a song quickly, categorizing, organizing and even listening again when they want etc. Because of this, people tend to find electronic solutions. To index music, most of the researchers use content based information retrieval mechanism since content based classification doesn't need any additional information rather than audio features embedded to it. As well as it is the most suitable way to search music, when users don't know the metadata attached to it, like the author of the song.
Research Question/Problem/ Need	- How can AI recognize the cover versions of a song and then still return the same output as the original song? - Can the program output the right value when given only a small portion of the song? - How will the system be able to overcome factors like external

	noise? - Can programs determine the name of the piece when multiple versions exist? - If the audio clip is pulled from a source that contains extra footage, like an advertisement playing before/after the song on a radio, can it ignore that data and only give us the song?
Important Figures	E.D. Nishan W. Senevirathna, Lakshman Jayaratne
VOCAB: (w/definition)	**Wavelet** (*n*) - a wave-like oscillation that begins at 0, increases/decreases in amplitude, and heads back to 0.
Cited references to follow up on	- T. Zhang and C.-C. J. Kuo, “Hierarchical system for content-based audio classification and retrieval,” in Photonics East (ISAM, VVDC, IEMB), 1998, pp. 398-409,International Society for Optics and Photonics, 1998. - P. Cano, “Content-Based Audio Search from Fingerprinting to Semantic Audio Retrieval,” Ph.D. Dissertation. UPF, 2007.
Follow up Questions	4. What is the extent a file can be distorted until the algorithm stops working? 5. If one song is based off of another (e.g. the way that “7 rings” by Ariana Grande uses the same melody as “Favorite Things”), will the program output both in the beginning? 6. If pitchless noises are coming in, then how does it affect the bass noise coming in?

Article Notes #7: Robust Speech/Music Classification in Audio Documents

Article notes should be on separate sheets

Source Title	7th International Conference on Spoken Language Processing, ICSLP2002
Source citation (APA Format)	Pinquier, J., Rouas, J.-L., & André-Obrecht, R. (2002). Robust speech / music classification in Audio Documents. 7th International Conference on Spoken Language Processing (ICSLP 2002). https://doi.org/10.21437/icslp.2002-551
Original URL	https://www.isca-speech.org/archive/icslp_2002/pinquier02_icslp.htm
Source type	Journal Article
Keywords	Segmentation, classification, audio, music, modulation, entropy, soundtrack

Summary of key points + notes (include methodology)	The researchers created algorithms of checking each of 3 unusual features of audio files – 4Hz modulation energy, entropy modulation, and entropy duration – and combined them into one general algorithm to solve the research need. (NOTE: A fourth method is used, *segments duration*, but it is not included in the final product. Its accuracy rate is also below 80%, so it is not effective.) Features already in use before this research paper: - Zero crossing rate and spectral centroid used for discerning speech from unintelligible noise - Spectral flux used to detect harmonic continuity in music - Speech processing focused on "cepstral features" I. Gaussian Mixture Models II. k-nearest neighbors III. Hidden Markov Models The first method is **4 Hz modulation energy**. In speech, the modulation energy around a Hz period is higher. (This is an observation made based on speech patterns.) Music is more consistent, and therefore has smaller modulation energy. To calculate the modulation energy, the "classical procedure" was applied: - Segment signal in 16ms frames - MFC (Mel Frequency Coefficients) extracted - Filter centered on 4Hz to extract sound from those time periods - Energy summed and normalized by mean energy on frame - Modulation computed (variance of filtered energy in dB [on 1sec of signal to improve efficiency]) Speech has more modulation energy, so the larger output is the speech. The second method is **entropy modulation**. Music is more "orderly" than speech. Another way to think about this metric is to think about the consistency of the audio recording. For instance, in most music, there is a beat, which means that there is typTo mathematically define this "order", the researchers used an equation that maps signal entropy: $H = \sum_{i=1}^{k} -p_i \log_2 p_i$, where p_i represents the probability of event *i*, and k is the number of 16ms segments of the audio file. Entropy modulation is much higher for speech than for music. GRAPHICAL JUSTIFICATION FOR THE FIRST 2 METHODS

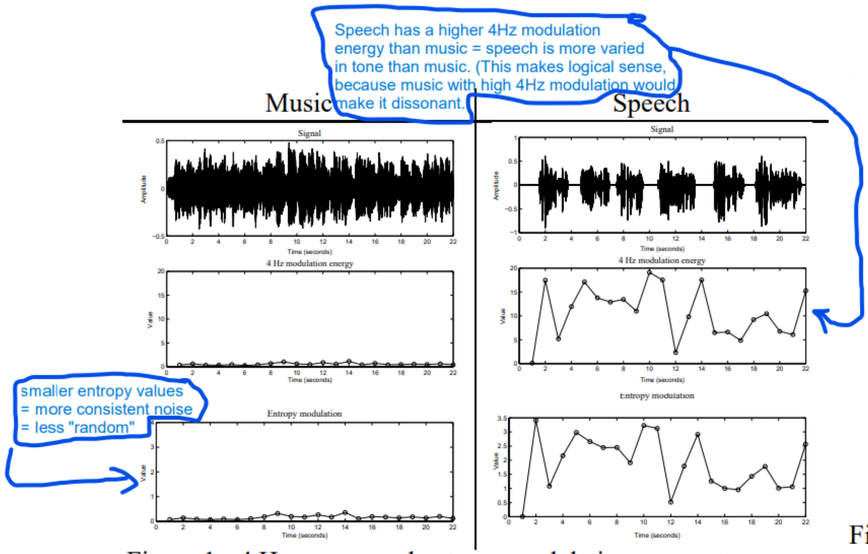


Figure 1 - 4 Hz energy and entropy modulation parameters for music and speech.

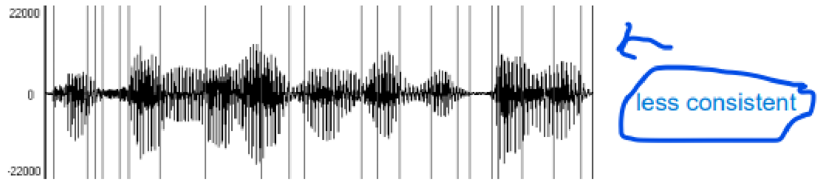
The third (and final) method is **entropy duration**. It observes how long consistent audio is present in the file. Music generally has longer segments, whereas speech has short segments (typically vowels are the segments). The more rapid segments there are, the more likely it is that the audio is speech.

To mathematically model the consistent vs. inconsistent parts of a recording, a probability density function, p(g), was used. The consistent parts are part of a continuous function while the inconsistent parts are 0.

$$p(g) = \sqrt{\frac{\lambda}{2\pi g^3}} \cdot e^{\frac{-\lambda(g-\mu)^2}{2\mu^2 g}} \qquad g \geq 0$$

$$p(g) = 0 \qquad g \leq 0,$$

where μ = mean value of g and $\frac{\mu^3}{\lambda}$ is the variance of g.

	 Figure 2a - Segmentation on about 1 second of speech.  Figure 2b - Segmentation on about 1 second of music.
Research Question/Problem/ Need	How can the accuracy of technology that discerns speech and music be improved?
Important Figures	Julien Pinquier, Regine Andre-Obrecht
VOCAB: (w/definition)	*Development corpus*: Large set of data used in testing an algorithm. In our case, the corpus was a TV movie soundtrack. Since this soundtrack contains lots of dialogue and music, it offers a sufficiently large testing size for the researchers' algorithm. *Spectral flux:* Variation in spectral magnitude *MFC:* Mel Frequency Coefficients, used to determine the power spectrum (where the strongest input of signal is)
Cited references to follow up on	M. Franz, J. Scott McCarley, T. Ward and W. Zhu, Topics styles in IR and TDT: Effect on System Behavior, EUROSPEECH'2001, Scandinavia, pp. 287-290, September 2001. J.L. Gauvain, L. Lamel and G. Adda, Audio partitioning and transcription for broadcast data indexation, CBMI'99, Toulouse, pp. 67-73, October 1999. S. Rossignol, X. Rodet, J. Soumagne, J.L. Collette and P. Depalle, Automatic characterization of musical signals: feature extraction and temporal segmentation, Journal of New Music Research, 2000, pp. 1-16.
Follow up Questions	1. How does this system work on rap? Rap is a combination of music and talking, so how does the system discern this? 2. Does the system detect speaking when one is speaking with a slur? For example, if a sample was recorded of a drunken individual speaking and the audio was all connected, can the system detect that it is not music?

	3. How can the algorithm be made more efficient? 4. Are there other “unusual” aspects of audio files that can be researched?

Article Notes #8: Audio Recognition using Mel Spectrograms and Convolution Neural Networks

Article notes should be on separate sheets

Source Title	Noiselab@UCSD (University of California San Diego)
Source citation (APA Format)	Leitner, B.Z., & Thornton, S. (2019). Audio Recognition using Mel Spectrograms and Convolution Neural Networks.
Original URL	http://noiselab.ucsd.edu/ECE228_2019/Reports/Report38.pdf
Source type	Journal Article
Keywords	Machine Learning; Audio Recognition; Mel Spectrogram; CNNs
Summary of key points + notes (include methodology)	The researchers tested and compared two approaches to analyzing spectrograms – a self-developed convolutional neural network (or CNN) and a transfer learning model adapted from the VVG19 neural network (19 layers deep, available on Matlab for anyone to use). The main efforts on the paper were to perform more preprocessing before inputting the audio into the CNNs, because this seems to be the main problem with current CNNs used to classify audio. This preprocessing is especially important for the VVG19 adaptation model. VVG19, as shown in Section IV, was designed originally for image processing. Neural networks have weights and biases for every pixel (VVG19 uses 128x128px images), and VVG19 is ideally using weights that scan every corner of the image. The researchers were able to change these weights to focus on the frequencies in the images which mattered. For example, an audio clip of someone speaking would not need to consider frequencies at, say, 8000Hz, because no human can speak with that high of a frequency. **Results**: The deep CNN model (built from the ground up) had an accuracy rate of 88.9%. The transfer learning model (weighted VGG19 model) had an accuracy rate of 88.5%, which is exceptionally close to the first. The VVG19 (no weight) model had the worst results, with accuracy of 82.9%. It is remarkable how the researchers could edit an algorithm for

	analyzing images and adapt it to one for audio, given how little data they had. (The size of the data was the main limitation the researchers had.
Research Question/Problem/ Need	How can the algorithms of conventional neural networks be improved to better classify audio spectrograms?
Important Figures	Muhammad Turab, who worked on the paper *Investigating Multi-Feature Selection and Ensembling for Audio Classification*, heavily influenced this paper. More specifically, Leitner, Zhang, and Thornton used the same equation to convert frequencies from the raw audio into frequencies inputted into the spectrogram frequencies. They wanted the frequencies going into the spectrogram to be within the range of human hearing (20 to 20k Hz), so they both used the following equation to convert from frequency in Hz to value for the Mel Spectrogram: $m = 2595 \log_{10}(1 + \frac{f}{700})$
VOCAB: (w/definition)	*Spectrogram* (n.) – a 2D model used for representing audio data. Time lies on a linear scale on the x-axis, frequencies lie on a logarithmic scale on the y-axis, and amplitudes (magnitude of the sound waves) at each frequency and time frame are represented as colors (bright colors mean high frequency and dark colors mean low frequency).
Cited references to follow up on	[1] S. Chu, S. Narayanan and C. J. Kuo, "Environmental Sound Recognition with Time–Frequency Audio Features," in IEEE Transactions on Audio, Speech, and Language Processing, vol. 17, no. 6, pp. 1142-1158, Aug. 2009. [2] “The Nature of Sound.” The Physics Hypertextbook. [3] Kilshore Prahallad, “Spectrogram, Cepstrum and MelFrequency Analysis,” Carnegie Mellon University. [4] Homburg, Helge, et al. "A Benchmark Dataset for Audio Classification and Clustering." ISMIR. Vol. 2005. 2005. [5] Piczak, Karol J. "Environmental sound classification with convolutional neural networks." 2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP). IEEE, 2015.

	[6] Hershey, Shawn, et al. "CNN architectures for large-scale audio classification." 2017 ieee international conference on acoustics, speech and signal processing (icassp). IEEE, 2017. [7] Salamon, Justin, and Juan Pablo Bello. "Deep convolutional neural networks and data augmentation for environmental sound classification." IEEE Signal Processing Letters 24.3 (2017): 279-283. [8] Becker, Sören, et al. "Interpreting and explaining deep neural networks for classification of audio signals." arXiv preprint arXiv:1807.03418 (2018). [9] Eduardo Fonseca, Jordi Pons, Xavier Favory, Frederic Font, Dmitry Bogdanov, Andres Ferraro, Sergio Oramas, Alastair Porter, and Xavier Serra. "Freesound Datasets: A Platform for the Creation of Open Audio Datasets." In Proceedings of the International Conference on Music Information Retrieval, 2017. [10] Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li, YFCC100M: The New Data in Multimedia Research, Commun. ACM, 59(2):64–73, January 2016 [11] Kaggle Freeaudio Tagging 2019, https://www.kaggle.com/c/freesound-audio-tagging2019/overview/evaluation
Follow up Questions	1. How could the quality of the image (e.g. using 256x256px images instead of 128x128px images) decrease training time? 2. Is it possible to run both the deep CNN and the Transfer Learning model at the same time? Would this produce a more accurate result? 3. CNNs specifically tailored for image processing, like VVG19, are good at recognizing shapes. Can these shape recognition systems be useful for recognizing different techniques, like vibrato (looks like a sine wave), staccato, (looks like dashes), etc.?

Article Notes #9: Musical Instrument Detection

(Detecting instrumentation in polyphonic musical signals on a frame-by-frame basis)

Source Title	*Center for Computer Research in Music and Acoustics* (2006)
Source citation (APA Format)	Sell, G., Mysore, G. J., & Chon, S. H. (2006). Musical instrument detection. *Center for Computer Research in Music and Acoustics*.
Original URL	http://www.cucugi.us/1.pdf
Source type	Journal Article
Keywords	Instrument, detection algorithm, music, sounds, acoustics
Summary of key points + notes (include methodology)	The researchers used binary comparisons to classify instruments. They had 3 different groups(piano, violin; clarinet, piano, violin; cello, clarinet, flute, piano, violin). 4 audio samples were taken from 13 different recordings for each instrument. 40 of the samples were training data; the other 12 were test data. 4 different models were used: - Logistic - KNN - SVM-Linear - SVM-Gaussian They used limited models that ran with part of the data, and they also used full models that covered all the data. Results: They were able to classify 2 instruments with an accuracy of 79.46%, 3 instruments with accuracy 51.69%, and 5 instruments with accuracy 42.44%.
Research Question/Problem/ Need	How can an algorithm be developed that can differentiate between different instruments?
Important Figures	Gregory Sell - works at Otter.ai (collaboration + note-taking app) Gautham J. Mysore - works at Adobe Research (homepage: https://www.gauthamjmysore.com/)
VOCAB: (w/definition)	*Supervised machine learning model:* A type of machine learning model that takes in labeled data and serves as a prediction tool for labeling new data.

	Regression model (for machine learning): A type of supervised learning model where the output value caps at a certain value. The logistic model starts off like an exponential, reaches a maximum growth rate at half its carrying capacity, and then starts slowing down until it reaches its maximum carrying capacity. Letting t, K, and P represent time, carrying capacity, and current population, we have that the equation of the differential is $\frac{dP}{dt} = kP(1 - \frac{P}{K})$, and that the equation of the population over time is $P(t) = \frac{K}{1+Ae^{-kt}}$, where $A = \frac{K-P_0}{P_0}$. *K-NN model:* K-nearest neighbors model. ML model that classifies data based on K of its nearest neighbors. *SVN:* A model that differentiates two different groups with points having n dimensions (with each coordinate representing a single quality of the datum). LINEAR: A line is drawn that separates both sample labeled groups (therefore, an SVN is a supervised learning model). GAUSSIAN: A Gaussian curve is used to separate the groups. Oftentimes,
Cited references to follow up on	[1] Juha Parhankangas Esa Alhoniemi, Johan Himberg and Juha Vesanto. SOM Toolbox. available online at http://www.cis.hut.fi/projects/somtoolbox, 2000-5. [2] Gokhan Bakir Jason Weston, Andre Elisseeff and Fabian Sinz. The Spider. available online at http://www.kyb.tuebingen.mpg.de/bs/people/spider/, 2006. [3] G. Richard S. Essid and B. David. Instrument recognition in polyphonic music. In Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing, 2005.
Follow up Questions	1. How does the model filter background noise? 2. How does the model handle cases where multiple instruments are being played simultaneously (e.g. a violin + piano duet)? 3. This paper was published in ‘06. Has the technology on these methods improved in recent years? If, so, how? How would any changes have an effect on the methodology/results?

Article Notes #10: Music Information Retrieval by Detecting Mood via Computational Media Aesthetics

Source Title	Proceedings IEEE/WIC International Conference on Web Intelligence (WI 2003)
Source citation (APA Format)	Yazhong Feng, Yueting Zhuang and Yunhe Pan, "Music information retrieval by detecting mood via computational media aesthetics," Proceedings IEEE/WIC International Conference on Web Intelligence (WI 2003), 2003, pp. 235-241, doi: 10.1109/WI.2003.1241199.
Original URL	https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=1241199
Source type	Conference Proceeding
Keywords	Music, mood, classification, detection, computational media aesthetics
Summary of key points + notes (include methodology)	Goal: Classify popular music (sampled popular music from a corpus of 353 samples) into one of four mood categories (happiness, anger, sadness, and fear) based on two metrics: tempo and articulation. Methodology: 1. Calculate tempo (using a multiple-agent approach) 2. Calculate relative tempo 3. Find mean (μ) and standard deviation (σ) of average silence ratio 4. BP neural network classifier used to detect mood Computational Media Aesthetics are aimed at filling the gap between the complexity of “emotion” and the shallowness of musical data. The researchers related tempo and articulation with certain emotions. For example, fast music played with staccato articulation was considered to be happy. RESULTS: The model was most accurate at recognizing happiness and anger, the models with the fastest tempos (presumably from 150 to 250). The percent accuracy of the model for detecting happy and angry music were 86% and 83%, respectively.
Research Question/Problem/ Need	How can music be classified into different moods? Can mood/emotion be retrieved by analyzing certain qualities in the music?

Important Figures	Yazhong Feng - in general, studies models of large datasets (mostly water)
VOCAB: (w/definition)	*BP* - backpropagation. The process that most neural networks use to train themselves. Backpropagation works by evaluating a cost function (a measure of how “badly” the network is doing), and then calculates the negative gradient, which is a measure of how to change the parameters so that the function is decreased as rapidly as possible.
Cited references to follow up on	[1] Ghias A., Logan,J. Camberlin,D. and Smith, B. C. “Query by humming: Musical information retrieval in an audio database”, Proc. ACM Int. Conf. On Multimedia, ACM, San Fancisco, CA, 1995, pp. 231–236. [2] Feng,Y. Z., Zhuang,Y. T. and Pan,Y. H. “A hierarchical approach: query large music database by acoustic input”, Proc. SIGIR, July 2002, pp. 441-442. [3] Kosugi, N., Nishihara, Y., Sakata, T., Yamamuro, M., and Kushima, K. “A practical query-by humming system for a large music database”, Proc of the ACM MM2000, Marina del Ray, CA, 2000, pp. 333-342. [4] Feng, Y. Z., Zhuang, Y. T., Pan, Y .H. “Popular music retrieval by independent component analysis”, ISMIR 2002 Conf. Proc., October 2002, pp. 281-282 [5] Liu, M. and Wan, C. “A study of content-based retrieval of mp3 music objects”, Proc. of the Int’l conf on Information and knowledge management, Atlanta, Georgia, 2001, ACM, pp. 506-511. [6] Yang, C. “The MACSIS Acoustic indexing framework for music retrieval: an experimental study”, Proc. of third international conference on music information retrieval (ISMIR2002), Paris, France, 2002, pp. 53-62. [7] Cooper, M. and Foote, J. “Automatic Music Summarization via Similarity Analysis”, Proc. of third international conference on music information retrieval (ISMIR2002), Paris, France, 2002, pp. 81-85. [8] Logan, B. and Salomon, A. “A music similarity function based on signal analysis”, Proc. of IEEE International Conference on Multimedia and Expo (ICME), August 2001. [9] Rauber, A., Pampalk, E., Merkl, D. “Content-based music indexing and organization”, Proc. SIGIR, July 2002. [10] Pickens, J. “A survey of feature selection techniques for music information retrieval”, Proc. SIGIR, September, 2001. [11] Rolland, P. Y., Raskinis, G., Ganascia, J. G. “Musical content-based retrieval: an overview of the Melodiscov approach and system”, ACM Multimedia (1) 1999, pp.81-84. [12] Goldberg, D., Nichols, D. , Oki, B. and Terry, D., “Using

Collaborative Filtering to Weave an Information Tapestry", Communications of the ACM, Vol. 35(12), December 1992, pp. 61-70.
[13] Burke, R., "Hybrid Recommender Systems: Survey and Experiments", User modeling and User-adapted Interaction, 2002.
[14] Dorai, C. and Venkatesh, S., "Computational Media Aesthetics: Finding meaning beautiful", IEEE Multimedia, 8(4), October-December, 2001, pp. 10-12.
[15] Truong, B.T., Venkatesh, S., Dorai, C., "Application of computational media aesthetics methodology to extracting color semantics in film", ACM Multimedia, Juan Les Pins, France, 2002.
[16] Adams B., C. Dorai, and S. Venkatesh, "Automated film rhythm extraction for scene analysis", IEEE International Conference on Multimedia and Expo, Tokyo, Japan, August, 2001.
[17] Adams B., C. Dorai, and S. Venkatesh, "Towards automatic extraction of expressive elements from motion pictures: Tempo", IEEE International Conference on Multimedia and Expo, volume II, New York City, USA, July 2000, pp. 641–645.
[18] Dellaert, F., Polzin, T., Waibel, A. "Recognizing Emotion In Speech", Proc. ICSLP '96.
[19] Huber, R., Batliner, A., Buckow, J., Nöth, E., Warnke, V., Niemann, H. "Recognition of Emotion in a Realistic Dialogue Scenario".,Proc. ICSLP'2000.
[20] Tzanetakis, G. and Cook, P. "Musical Genre Classification of Audio Signals", IEEE Transaction on Speech and Audio Processing, vol.10, July 2002, pp. 293-302.
[21] Farnsworth, P.R., "The Social Psychology of Music", Holt, Rinehart and Winston, New York, 1958.
[22] Thayer, R. E. The biopsychology of mood and arousal. New York: Oxford University Press, 1989.
[23] Zillmann, D. "Mood management in the context of selective exposure theory", In M. E. (Ed.), Communication yearbook 23, Thousand Oaks, CA: Sage, pp. 103-123.
[24] Juslin, P.N., "Cue utilization in communication of emotion in music performance: Relating performance to perception", J. Experimental Psychology, 26, 2000, pp. 1797-1813.
[25] Dixon, S. "A lightweight multi-agent musical beat tracking system", Proceedings of the Pacific Rim International Conference on Artificial Intelligence, Melbourne, Australia, 2000, pp. 778-788.
[26] Scheirer, E., "Tempo and beat analysis of acoustic musical signals", Journal of the Acoustical Society of America, 103 (1), 1998, pp. 588-601.

	[27] Goto, M. and Muraoka, Y. “An audio-based real-time beat tracking system and its applications”, Proc. of the International Computer Music Conference, Computer Music Association, San Francisco CA, 1998. [28] Foote, J. “The beat spectrum: a new approach to rhythm analysis”, IEEE International Conference on Multimedia & Expo, Tokyo, Japan, 2001.
Follow up Questions	1. How can the model be extended to different moods? 2. Is it possible to pick out the emotion in the singer’s voice? How can we tell when the singer sounds happy or angry? 3. How fast does the model learn in terms of accuracy? How significantly would the model improve if given 3,000 samples instead of 300?

Article Notes #11: Automatic Detection of Audio Problems for Quality Control in Digital Music Distribution

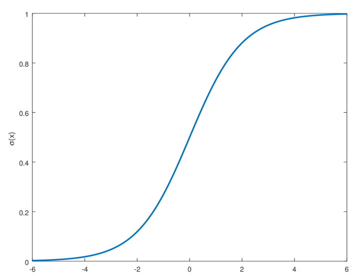
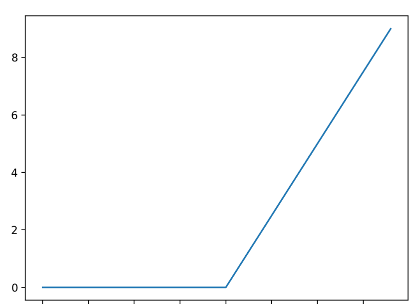
Source Title	e-Repositori upf. Congressos (Departament de Tecnologies de la Informació i les Comunicacions)
Source citation (APA Format)	AL. Alonso-Jimé, P. nez, L. Joglar-Ongay, and X. Serra, D. Bogdanov, "Automatic Detection of Audio Problems for Quality Control in Digital Music Distribution," Paper 10205, (2019 March.). doi:
Original URL	https://repositori.upf.edu/bitstream/handle/10230/37041/alonso_aes_automatic.pdf?sequence=1&isAllowed=y
Source type	Speech/Conference communication
Keywords	Stereo, correction, margins, algorithms, preprocessing
Summary of key points + notes (include methodology)	The researchers broke up their analysis into a classification problem. To do this, they first recognized the main problems that occur when creating and recording music. The ones addressed in the study were: ● **Incorrect margins.** The space between the beginning and end of a recording. If these values differ significantly in size, or if they are too long/short, these errors are classified as incorrect margins. ● **Phase/Stereo Problems.** If the stereo matches the beat exactly, it's a "false stereo". The sine wave of the stereo and the audio must be similar (e.g. can be approximated by the same sine wave). ● **Audio artifacts.** These include random-seeming parts of the recording which do not contribute to the overall performance, so they make it less smooth. These include things like bursts, discontinuities, gaps, and clicks. ● **Low frequency humming**: Low frequencies that cannot be heard or felt, typically those below 50Hz, take up space and it would be more efficient to remove them if they do not contribute to the music. ● **Loudness problems.** This essentially boils down to the recording being loud or soft. RESULTS: The largest mistakes found in the recordings were

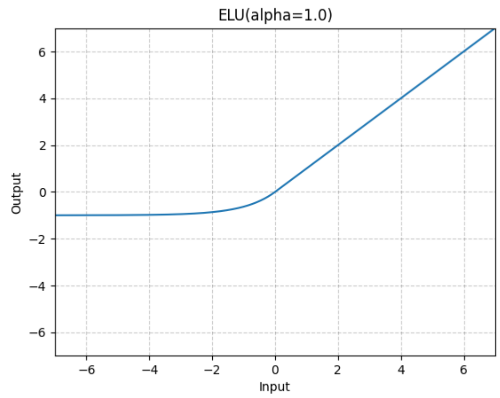
	humming (62.63%), clipping (49.09%), and noise bursts(24.77%).
Research Question/Problem/ Need	Develop an algorithm that can recognize audio inconsistencies or issues to more efficiently tidy up tracks for release.
Important Figures	Pablo Alonso-Jiménez - PhD student working for the MTG (Music Technology Group) at UPF Barcelona (Universitat Pompeu Fabra Barcelona)
VOCAB: (w/definition)	*PCC* - Pearson's correlation coefficient. Measures linear correlation between two sets of data. Formula: $r = \frac{\sum (x_i - x)(y_i - y)}{\sqrt{\sum (x_i - x)^2 \sum (y_i - y)^2}}$, used commonly in statistics.
Cited references to follow up on	[1] Dmitry Bogdanov, Nicolas Wack, Emilia Gómez, Sankalp Gulati, Perfecto Herrera, O. Mayor, Gerard Roma, Justin Salamon, J. R. Zapata, and Xavier Serra. Essentia: An audio analysis library for music information retrieval. In International Society for Music Information Retrieval Conference (ISMIR'13), pages 493–498, Curitiba, Brazil, 04/11/2013 2013. [2] Earl Vickers. The loudness war: background, speculation, and recommendations. In Audio Engineering Society Convention 129. Audio Engineering Society, Nov 2010. [3] BS.1387-1, ITU-R. Method for objective measurements of perceived audio quality, 2001. first edition. [4] Rainer Huber and Birger Kollmeier. Pemo-q—a new method for objective audio quality assessment using a model of auditory perception. IEEE Transactions on audio, speech, and language processing, 14(6):1902–1911, 2006. [5] James M Kates and Kathryn H Arehart. The Hearing-Aid Speech Quality Index (HASQI). Journal of the Audio Engineering Society, 58(5):363–381, 2010. [6] Paul Kendrick, Iain R Jackson, Francis F Li, BM Fazenda, TJ Cox, et al. Perceived audio quality of sounds degraded by non-linear distortions and single-ended assessment using hasqi. Journal of the Audio Engineering Society, 63(9):698–712, 2015. [7] Ignasi Adell Arteaga. Automatic detection of audio defects in personal music collections. Master's thesis, Universitat Pompeu Fabra, 2016. [8] Rudolf Mühlbauer. Automatic audio defect detection, 2010.

[9] Lawrence R Rabiner and Ronald W Schafer. Digital processing of speech signals, volume 100. Prentice-hall Englewood Cliffs, NJ, 1978.
[10] Saeed V. Vaseghi. Advanced digital signal processing and noise reduction, chapter 13, pages 351–353. John Wiley & Sons, fourth edition, 2008.
[11] George Turin. An introduction to matched filters. IRE transactions on Information theory, 6(3):311–329, 1960.
[12] Ryan Laney. Automatic detection of flaws in recorded music using wavelet fingerprinting, 2011.
[13] Matthias Brandt and Joerg Bitzer. Automatic detection of hum in audio signals. Journal of the Audio Engineering Society, 62(9):584–595, 2014.
14] Howard A. Chinn, D.K. Gannett, and R.M. Morris. A new standard volume indicator and reference level. Proceedings of the IRE, 28(1):1–17, 1940.
[15] Bob Katz. Integrated approach to metering, monitoring, and leveling practices, part 1: Two-channel metering. Journal of the Audio Engineering Society, 48(9):800–809, 2000.
[16] TECH 3343, EBU. Practical guidelines for production of programmes in accordance with EBU-R 128, 2011. Version 2.0.
[17] Fabian Begnert, Håkan Ekman, and Jan Berg. Difference between the ebu r-128 meter recommendation and human subjective loudness perception. In Audio Engineering Society Convention 131. Audio Engineering Society, 2011.
[18] BS.1770-4, ITU-R. Algorithms to measure audio programme loudness and true-peak audio level, 2015. fourth edition.
[19] Hugh Robjohns. What are my phase-correlation meters telling me? https://www.soundonsound.com/sound-advice/q-what-are-my-phase-correlation-meters-telling-me, 2016. accessed on 2019-01-09.
[20] Yariv Ephraim and David Malah. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. IEEE Transactions on acoustics, speech, and signal processing, 32(6):1109–1121, 1984.
[21] Justin Salamon and Emilia Gómez. Melody extraction from polyphonic music signals using pitch contour characteristics. IEEE Transactions on Audio, Speech, and Language Processing, 20(6):1759–1770, 2012.

	[22] TD1004.1.15-10, AES. Recommendation for loud- ness of audio streaming and network file playback, 2015. first edition.
Follow up Questions	1. Where does the technology most likely fail? 2. Could the process be done in reverse? If someone wanted to "olden" a recording, could they add negative elements to it? 3. How can the model be improved by today's technology?

Article Notes #12: Activation Functions: Comparison of Trends in Practice and Research for Deep Learning

Source Title	arXiv.org (https://arxiv.org/abs/1811.03378)
Source citation (APA Format)	Nwankpa, C., Ijomah, W., Gachagan, A., & Marshall, S. (2018). Activation Functions: Comparison of Trends in Practice and Research for Deep Learning. Computer Vision and Pattern Recognition .
Original URL	https://arxiv.org/pdf/1811.03378
Source type	Journal Article
Keywords	Machine learning, deep learning, activation functions, ReLU, sigmoid, tanh, neural networks, learning algorithms
Summary of key points + notes (include methodology)	There are 3 main types of machine learning activation functions (AFs): I. *Sigmoid function.* Equation: $\sigma(x) = \frac{1}{1+e^{-x}}$ Graph:  II. *ReLU (Rectified Linear Unit).* Equation = $f(x) = \max(0, x)$.  Graph: III. *ELU (Exponential Linear Unit)* Equation:

	$$f(x) = \begin{cases} x, x > 0 \\ \alpha \cdot e^x - 1, x \le 0 \end{cases}$$  Graph: RESULTS: The researchers concluded that since ReLUs were featured in 10 out of 11 winners for an image processing algorithm, ReLU was the most effective. These models had more accuracy when doing backpropagation.
Research Question/Problem/ Need	Determine the trends between the different AFs seen in research literature (specifically those in deep learning research).
Important Figures	Chigozie Enyinna Nwankpa - Tutor- SBS Administration
VOCAB: (w/definition)	Activation function - a function used to determine whether or not neurons are active after calculating the weighted sum of the activations of a previous layer. It's a squash function (the sigmoid is consistently called this) that squishes the real number line into a smaller range.
Cited references to follow up on	[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436–444, 2015. [Online]. Available: https: //doi.org/10.1038/nature14539 [2] N. Jones, "Computer science: The learning machines,," Nature, vol. 505, no. 7482, pp. 146–148, 1 2014. [3] L. Deng, "A tutorial survey of architectures, algorithms, and applications for deep learning," APSIPA Transactions on Signal and Information Processing, vol. 3, 2014. [Online]. Available: https://doi.org/10.1017/atsip.2013.9 [4] M. Z. Alom, T. M. Taha, C. Yakopcic, S. Westberg, M. Hasan, B. V. Esesn, and V. K. Asari, "The History Began from AlexNet: A Comprehensive Survey on Deep Learning Approaches," arXiv, 2018. [Online]. Available: http://arxiv.org/abs/1803.01164 [5] I. Goodfellow, Y. Bengio, and A. Courville, "Deep learning." MIT Press, 2016. [Online]. Available: http://www.deeplearningbook.org [6] S. Agatonovic-Kustrin and R. Beresford, "Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research," Journal of Pharmaceutical and Biomedical Analysis, vol. 22, no. 5, pp. 717–727, 2000. [Online]. Available: https://doi.org/10.1016/S0731-7085 [7] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation Applied to Handwritten Zip Code Recognition," Neural Computation, vol. 1, no. 4, pp. 541–551, 1989. [Online]. Available: https://doi.org/10.1162/neco.1989.1.4.541 [8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," 2012. [Online]. Available:

http://citeseerx.ist.psu.edu/viewdoc/summary
[9] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” arXiv, 2015. [Online]. Available:
https://arxiv.org/pdf/1409.1556.pdf
[10] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,,” in
Conference on Computer Vision and Pattern Recognition (CVPR), vol. 2015. IEEE, 2015, pp. 1–9.
18
[11] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” in IEEE Conf. Comput. Vis. Pattern Recognit. IEEE, 2016,
pp. 770–778. [Online]. Available: https://doi.org/10.1109/CVPR.2016.90
[12] G. Huang, Y. Sun, Z. Liu, D. Sedra, and K. Q. Weinberger, “Deep networks with stochastic depth,” ArXiv, vol. 9908, no. LNCS, pp. 646–661, 2016.
[Online]. Available: 10.1007/978-3-319-46493-0_39;http://arxiv.org/abs/1603.09382
[13] J. Turian, J. Bergstra, and Y. Bengio, “Quadratic features and deep architectures for chunking,” in Proceedings of Human Language Technologies:
The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics,, vol. Companion Volume:, 2009, pp.
245–248. [Online]. Available: https://dl.acm.org/citation.cfm
[14] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and undefined R. Salakhutdinov, “Dropout: A Simple Way to Prevent Neural Networks from
Overfitting,,” Journal of Machine Learning Research, vol. 15, pp. 1929–1958, 2014.
[15] D. J. Matas, “All you need is a good init,,” ArXiv, 2016. [Online]. Available:
http://arxiv.org/abs/1511.06422
[16] P. Luo, X. Wang, W. Shao, and Z. Peng, “Understanding Regularization in Batch Normalization,,” arXiv, 2018. [Online]. Available:
http://export.arxiv.org/abs/1809.00846
[17] S. Ioffe and C. Szegedy, “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift,,” in Proceedings of the 32nd
International Conference on Machine Learning, PMLR . JMLR, 2015, pp. 448–456.
[18] X. Glorot and Y. Bengio, “Understanding the difficulty of training deep feedforward neural networks,” in International Conference on Artificial Intelligence
and Statistics, vol. 9. PLMR, 2010, pp. 249–256. [Online]. Available:
http://proceedings.mlr.press/v9/glorot10a/glorot10a.pdf?hc_location=ufi
[19] P. Ramachandran, B. Zoph, and Q. V. Le, “Searching for Activation Functions,” ArXiv, 2017. [Online]. Available: 1710.05941;http:
//arxiv.org/abs/1710.05941
[20] X. Glorot, A. Bordes, and Y. Bengio, “Deep Sparse Rectifier Neural Networks,” in International Conference on Machine Learning, 2011. [Online].
Available: http://proceedings.mlr.press/v15/glorot11a/glorot11a.pdf
[21] Y. Bengio, P. Simard, and P. Frasconi, “Learning long-term dependencies with gradient descent is difficult,” IEEE Transactions on Neural Networks,
vol. 5, no. 2, pp. 157–166, 1994. [Online]. Available: https://doi.org/10.1109/72.279181
[22] R. Pascanu, T. Mikolov, and Y. Bengio, “On the difficulty of training recurrent neural networks,” in Proceedings of the 30th International Conference
on International Conference on Machine Learning, vol. 28, 2013, pp. III–1310. [Online]. Available:
https://dl.acm.org/citation.cfm
[23] J. Bergstra, D. Yamins, and D. D. Cox, “Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision
Architectures,,” in Machine Learning research, 2013. [Online]. Available:
http://proceedings.mlr.press/v28/bergstra13.pdf
[24] C. Bircanoglu and N. Arica, “A comparison of activation functions in artificial neural networks,,” in Signal Processing and Communications Applications
Conference (SIU), vol. 26. IEEE‘, 2018, pp. 1–4.
[25] K. He, X. Zhang, S. Ren, and J. Sun, “Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification,” arXiv, 2015.
[Online]. Available: http://arxiv.org/abs/1502.01852
[26] S. M. Noor, J. Ren, S. Marshall, and K. Michael, “Hyperspectral Image Enhancement and Mixture Deep-Learning Classification of Corneal Epithelium
Injuries,” Sensors. Multidisciplinary Digital Publishing Institute, vol. 17, no. 11, p. 2644, 2017.
[27] T. N. Sainath, B. Kingsbury, G. Saon, H. Soltau, A. Mohamed, G. Dahl, and B. Ramabhadran, “Deep Convolutional Neural Networks for Large-scale
Speech Tasks,” Neural Networks, vol. 64, pp. 39–48, 2015. [Online]. Available:
https://doi.org/10.1016/J.NEUNET.2014.08.005
[28] A. Graves, A. Mohamed, and undefined G. Hinton, “Speech recognition with deep recurrent neural networks,,” in 2013 IEEE International Conference
on Acoustics, Speech and Signal Processing. IEEE, 2013, pp. 6645–6649.
[29] V. Badrinarayanan, A. Kendall, and R. Cipolla, “SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation,” arXiv, 2015.
[Online]. Available: http://arxiv.org/abs/1511.00561
[30] R. Hu, P. Dollár, K. He, T. Darrell, and R. Girshick, “Learning to Segment Every Thing,,” arXiv, 2017. [Online]. Available:
10.1109/CVPR.2018.00445;http://arxiv.org/abs/1711.10370
[31] A. Karpathy and L. Fei-Fei, “Deep Visual-Semantic Alignments for Generating Image Descriptions,” in Computer Vision and Pattern Recognition
(CVPR). IEEE, 2015, pp. 3128–3137. [Online]. Available:
https://www.cv-foundation.org/openaccess/content_cvpr_2015/html/Karpathy_Deep_Visual-

Semantic_Alignments_2015_CVPR_paper.html
[32] P. O. Pinheiro and R. Collobert, “Recurrent Convolutional Neural Networks for Scene Labeling,” in International Conference on Machine Learning,
2014. [Online]. Available: http://proceedings.mlr.press/v32/pinheiro14.pdf
[33] O. Vinyals, L. Kaiser, T. Koo, S. Petrov, I. Sutskever, and G. Hinton, “Grammar as a Foreign Language,” arXiv, 2014. [Online]. Available:
10.1146/annurev.neuro.26.041002.131047;http://arxiv.org/abs/1412.7449
[34] Y. Liu and J. Zhang, “Deep Learning in Machine Translation,” in in Deep Learning in Natural Language Processing. Singapore: Springer, 2018, pp.
147–183.
[35] S. O. Arik, M. Chrzanowski, A. Coates, G. Diamos, A. Gibiansky, Y. Kang, and M. Shoeybi, “Deep Voice: Real-time Neural Text-to-Speech,” arXiv,
2017. [Online]. Available: http://arxiv.org/abs/1702.07825
19
[36] W. Ping, K. Peng, A. Gibiansky, S. O. Arik, A. Kannan, S. Narang, and J. Miller, “Deep Voice 3: Scaling Text-to-Speech with Convolutional
Sequence Learning,” in International Conference on Learning Representations - ICLR, vol. 79, 2018, pp. 1094–1099. [Online]. Available:
https://doi.org/10.1001/jam
[37] S. Albarqouni, C. Baur, F. Achilles, V. Belagiannis, S. Demirci, and N. Navab, “AggNet: Deep Learning From Crowds for Mitosis Detection in Breast
Cancer Histology Images,,” IEEE Trans. Med. Imaging, vol. 35, no. 5, pp. 1313–1321, 5 2016.
[38] D. Wang, A. Khosla, R. Gargeya, H. Irshad, and A. H. Beck, “Deep Learning for Identifying Metastatic Breast Cancer,” arXiv, 2016. [Online].
Available: 1606.05718;http://arxiv.org/abs/1606.05718
[39] A. A. Cruz-Roa, J. A. Ovalle, A. Madabhushi, and F. G. Osorio, “A Deep Learning Architecture for Image Representation,” Springer, pp. 403–410,
2013. [Online]. Available: https://doi.org/10.1007/978-3-642-40763-5_50
[40] L. Lazimul and D.Binoy, “Fingerprint liveness detection using convolutional neural network and fingerprint image enhancement,” 2017. [Online].
Available: 10.1109/ICECDS.2017.8389533
[41] H. Jung and Y. S. Heo, “Fingerprint liveness map construction using convolutional neural network,” Electronics Letters, vol. 54, no. 9, pp. 564–566,
2018. [Online]. Available: 10.1049/el.2018.0621
[42] A. Grover, A. Kapoor, and E. Horvitz, “A Deep Hybrid Model for Weather Forecasting.” New York, New York, USA: ACM Press, 2015, pp.
379–386. [Online]. Available: https://doi.org/10.1145/2783258.2783275
[43] M. Hossain, B. Rekabdar, S. J. Louis, and S. Dascalu, Forecasting the weather of Nevada: A deep learning approach, 2015. [Online]. Available:
https://doi.org/10.1109/IJCNN.2015.7280812
[44] A. L. Xar, Y. Demir, C. Neyt, and G. Zelis, “Object recognition and detection with deep learning for autonomous driving applications,,” Modeling and
Simulation International, vol. 93, no. 9, pp. 759–769, 2017. [Online]. Available: 10.1177/0037549717709932
[45] C. Chen, A. Seff, A. Kornhauser, and J. Xiao, DeepDriving: Learning Affordance for Direct Perception in Autonomous Driving, 2015. [Online].
Available:
https://www.cv-foundation.org/openaccess/content_iccv_2015/html/Chen_DeepDriving_Learning_Affordance_ICCV_2015_paper.html
[46] B. Karlik and A. Vehbi, “Performance Analysis of Various Activation Functions in Generalized MLP Architectures of Neural Networks,”
International Journal of Artificial Intelligence and Expert Systems (IJAE), vol. 1, no. 4, pp. 111–122, 2011. [Online]. Available:
http://www.cscjournals.org/library/manuscriptinfo.php
[47] D. L. Elliott, “A Better Activation Function for Artifical Neural Networks,” 1993. [Online]. Available: https://drum.lib.umd.edu/handle/1903/5355
[48] J. Han and C. Moraga, “The influence of the sigmoid function parameters on the speed of backpropagation learning,,” in Natural to Artificial Neural
Computation. IWANN 1995. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 1995, pp. 195–201.
[49] R. M. Neal, “Connectionist learning of belief networks,” Artificial Intelligence, vol. 56, no. 1, pp. 71–113, 1992. [Online]. Available:
https://doi.org/10.1016/0004-3702(92)90065-6
[50] M. Courbariaux, Y. Bengio, and J. P. David, “BinaryConnect: Training Deep Neural Networks with binary weights during propagations.” arXiv, pp.
3123–3131, 2015. [Online]. Available: https://arxiv.org/pdf/1511.00363.pdf
[51] S. Elfwing, E. Uchibe, and K. Doya, ““Sigmoid-Weighted Linear Units for,” arXiv, 2017. [Online]. Available: http://arxiv.org/abs/1702.03118
[52] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, “Language Modeling with Gated Convolutional Networks,,” arXiv, 2017.
[53] A. Maas, A. Hannun, and A. Ng, “Rectifier Nonlinearities Improve Neural Network Acoustic Models,” in International Conference on Machine Learning
(icml), 2013.
[54] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, “Natural Language Processing (Almost) from Scratch,,” Journal of
Machine Learning Research, vol. 12, pp. 2493–2537, 2011.
[55] M. Lin, Q. Chen, and S. Yan, “Network In Network,” arXiv, 2013. [Online]. Available: http://arxiv.org/abs/1312.4400

[56] P. Le and W. Zuidema, “Compositional Distributional Semantics with Long Short Term Memory,” Association for Computational Linguistics, 2015.
[Online]. Available: https://doi.org/10.18653/v1/S15-1002
[57] V. Nair and G. E. Hinton, “Rectified linear units improve restricted boltzmann machines,” Haifa, 2010, pp. 807–814. [Online]. Available:
https://dl.acm.org/citation.cfm
[58] M. D. Zeiler, M. Ranzato, R. Monga, M. Mao, K. Yang, Q. V. Le, and G. E. Hinton, “On rectified linear units for speech processing,” in International
Conference on Acoustics, Speech and Signal Processing. IEEE, 2013, pp. 3517–3521, IEEE.
https://doi.org/10.1109/ICASSP.2013.6638312.
[59] G. E. Dahl, T. N. Sainath, and G. E. Hinton, “Improving deep neural networks for LVCSR using rectified linear units and dropout,” in International
Conference on Acoustics, Speech and Signal Processing. IEEE, 2013.
[60] M. D. Zeiler and R. Fergus, “Visualizing and Understanding Convolutional Networks,” arXiv, 2013.
[Online]. Available: http://arxiv.org/abs/1311.2901
[61] B. Xu, N. Wang, H. Kong, T. Chen, and M. Li, “Empirical Evaluation of Rectified Activations in Convolution Network,” arXiv, 2015. [Online].
Available: https://arxiv.org/abs/1505.00853
[62] X. Jin, C. Xu, J. Feng, Y. Wei, J. Xiong, and S. Yan, “Deep Learning with S-shaped Rectified Linear Activation Units,,” arXiv, pp. 1737–1743, 2015.
[Online]. Available: https://arxiv.org/pdf/1512.07030.pdf
20
[63] C. Dugas, Y. Bengio, F. Bélisle, C. Nadeau, and R. Garcia, “Incorporating Second-Order Functional Knowledge for Better Option Pricing,” in Advances
in Neural Information Processing Systems (NIPS), vol. 13, 2001, pp. 472–478. [Online]. Available:
https://papers.nips.cc/paper/1920-incorporating-
second-order-functional-knowledge-for-better-option-pricing
[64] H. Zheng, Z. Yang, W. Liu, J. Liang, and Y. Li, “Improving deep neural networks using softplus units,” in International Joint Conference on Neural
Networks (IJCNN), vol. 2015. IEEE, 2015, pp. 1–4.
[65] A. Senior and X. Lei, “Fine context, low-rank, softplus deep neural networks for mobile speech recognition,” in International Conference on Acoustics,
Speech and Signal Processing (ICASSP). IEEE, 2014, pp. 7644–7648, IEEE.
https://doi.org/10.1109/ICASSP.2014.6855087.
[66] D. A. Clevert, T. Unterthiner, and S. Hochreiter, “Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs),” arXiv, 2015.
[Online]. Available: http://arxiv.org/abs/1511.07289
[67] L. Trottier, P. Giguere, and B. Chaib-draa, “Parametric Exponential Linear Unit for Deep Convolutional Neural Networks,” arXiv, 2018. [Online].
Available: https://arxiv.org/pdf/1605.09332v1.pdf
[68] G. Klambauer, T. Unterthiner, A. Mayr, and S. Hochreiter, “Self-Normalizing Neural Networks,” arXiv, 2017. [Online]. Available:
https://arxiv.org/pdf/1706.02515v5.pdf
[69] J. Lehman, J. Chen, J. Clune, and K. O. Stanley, “Safe mutations for deep and recurrent neural networks through output gradients.” New York, New
York, USA: ACM Press, 2018, pp. 117–124. [Online]. Available: https://doi.org/10.1145/3205455.3205473
[70] I. J. Goodfellow, D. Warde-Farley, M. Mirza, A. Courville, and Y. Bengio, “Maxout networks,” 2013, p. 1319. [Online]. Available:
https://dl.acm.org/citation.cfm
[71] L. Tóth, “Phone recognition with hierarchical convolutional deep maxout networks,,” EURASIP J. Audio, Speech, Music Process, vol. 2015, no. 1, p. 25,
12 2015.
[72] M. Basirat and P. M. Roth, “The Quest for the Golden Activation Function,” arXiv, 2018. [Online].
Available: https://arxiv.org/pdf/1808.00783.pdf
[73] Y. LeCun, “Backpropagation Applied to Handwritten Zip Code Recognition,,” Neural Computing, vol. 1, no. 4, pp. 541–551, 1989.
[74] A. M. Saxe, J. L. McClelland, and S. Ganguli, “Exact solutions to the nonlinear dynamics of learning in deep linear neural networks,” arXiv, 2013.
[Online]. Available: http://arxiv.org/abs/1312.6120
[75] J. Fu, H. Zheng, and T. Mei, “Look closer to see better: Recurrent attention convolutional neural network for fine-grained image
recognition,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2017, pp. 4476–4484. [Online]. Available:
https://doi.org/10.1109/CVPR.2017.476
[76] G. Huang, Z. Liu, L. der Maaten, and K. Q. Weinberger, “Densely Connected Convolutional Networks,” in IEEE Conference on Computer Vision and
Pattern Recognition (CVPR). IEEE, 2017, pp. 2261–2269. [Online]. Available:
https://doi.org/10.1109/CVPR.2017.243
[77] G. Larsson, M. Maire, and G. Shakhnarovich, “FractalNet: Ultra-Deep Neural Networks without Residuals,” arXiv, 2016. [Online]. Available:
http://arxiv.org/abs/1605.07648
[78] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, “Aggregated residual transformations for deep neural networks,” in IEEE Conference on Computer
Vision and Pattern Recognition (CVPR). IEEE, 1 2017, pp. 5987–5995. [Online]. Available:
https://doi.org/10.1109/CVPR.2017.634
[79] V. Golkov, A. Dosovitskiy, J. I. Sperl, M. I. Menzel, M. Czisch, P. Sämann, and D. Cremers, “q-Space Deep Learning: Twelve-Fold Shorter

	and Model-Free Diffusion MRI Scans," IEEE Transactions on Medical Imaging, vol. 35, no. 5, pp. 1344–1351, 2016. [Online]. Available: https://doi.org/10.1109/TMI.2016.2551324 [80] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications,," arXiv, 2017. [Online]. Available: http://arxiv.org/abs/1704.04861 [81] Z. Deng, Z. Wang, and undefined S. Wang, "Stochastic area pooling for generic convolutional neural network,," Front. Artif. Intell. Appl, vol. 285, pp. 1760–1761, 2016. [82] D. Pedamonti, "Comparison of non-linear activation functions for deep neural networks on MNIST classification task,," arXiv, 2018. [Online]. Available: http://arxiv.org/abs/1804.02763
Follow up Questions	1. How could the activation functions be combined? 2. Is there a way to do machine learning on the activation function itself? 3. If ReLU is the more simplified model, why is it more effective than the sigmoid function? It seems counterintuitive.

Article Notes #13: On the Studies of Syllable Segmentation and Improving MFCCs for Automatic Birdsong Recognition

Source Title	2008 IEEE Asia-Pacific Services Computing Conference
Source citation (APA Format)	C. -H. Chou, P. -H. Liu and B. Cai, "On the Studies of Syllable Segmentation and Improving MFCCs for Automatic Birdsong Recognition," 2008 IEEE Asia-Pacific Services Computing Conference, 2008, pp. 745-750, doi: 10.1109/APSCC.2008.6.
Original URL	https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4780764
Source type	Conference Proceedings
Keywords	Birdsong, Mel Frequency Cepstral Coefficients, Classification, Music, Nature, Birds
Summary of key points + notes (include methodology)	Birdsongs are categorized by four main elements (note, syllable, phrase, song), but mostly use syllables. So, the methodology here is to use a syllable recognition algorithm. This happened by using MFCCs (Mel Frequency Cepstral Coefficients). MFCCs were historically used for detecting vibrations in earthquakes. However, it was proven to be effective in distinguishing between certain types of syllables, so it was adapted for speech recognition. The researchers used a bird vocalization database containing 420 bird species. RESULTS: The models were able to improve on the state of current technology by 65.93% and 54.27% for two different models.
Research Question/Problem/ Need	How can birdsong recognition be improved?
Important Figures	Chih-Hsun Chou - biologist who works with genetic algorithms
VOCAB: (w/definition)	*Mel Frequency Cepstral Coefficients*: A marker of audio recordings adapted from spectrograms that tell the “shape” of a noise
Cited references to follow up on	1] C.K. Catchpole and P.J.B. Slater, Bird Song: Biological Themes and Variations, Cambridge University Press, 1995.

[2] S.E. Anderson, A.S. Dave and D. Margoliash, “Template-based automatic recognition of birdsong syllables from continuous recordings,” Journal of the Acoustical Society of America, vol. 100, no. 2, pp. 1209-1219, 1996.
[3] S.A. Selouani, M. Kardouchi, E. Hervet and D. Roy, “Automatic birdsong recognition based on autoregressive time-delay neural networks,” in Proc. ICSC Congr. Comput. Intell. Methods Appl., pp. 1-6, 2005.
[4] P. Somervuo, “A. Härmä, Bird song recognition based on syllable pair histograms,” in Proc. IEEE Intern. Conf. Acoustics, Speech, and Signal Processing, vol. 5, pp. V-825-8, 2004.
[5] A.L. McIlraith and H.C. Card, “Bird song identification using artificial neural networks and statistical analysis,” in Proc. Canadian Conf. Electrical and Computer Engineering, vol. 1, pp. 63-66, 1997.
[6] A.L. McIlraith and H.C. Card, “A comparison of backpropagation and statistical classifiers for bird identification,” in Proc. IEEE Intern. Conf. Neural Networks, vol. 1, pp. 100-104, 1997.
[7] A.L. McIlraith and H.C. Card, “Birdsong recognition using backpropagation and multivariate statistics,” IEEE Trans. Signal Processing, vol. 45, no. 11, pp. 2740-2748, 1997.
[8] L.R. Rabiner and M.R. Sambur, “An algorithm for determining the endpoints of isolated utterances”, Bell Syst. Tech. J., vol.54, pp. 297-315, Feb. 1975.
[9] S.N He, and J.B Yu, “A novel Chinese continuous speech endpoint detection method based on time domain features of the word structure,” IEEE Intern. Conf. Communications, Circuits and Systems and West Sino Expositions, vol. 2, pp. 992-996, July 2002.
[10] W.J. Zhang, and J.Y. Xie, “Endpoint detection based on MDL using subband speech satisfied auditory model,” IEEE Intern. Conf. Neural Networks and Signal Processing, vol. 2, pp. 892-895, Dec. 2003.
[11] S.N. He and J.B. Yu, “The performance analysis of Chinese speech endpoint detection based on continuous multi sub-band spectral features,” Proc. IEEE Intern. Conf. Communications, Circuits and Systems and West Sino Expositions, vol. 2, pp. 997-1002, July 2002.
[12] S.E. Bou-Ghazale and K. Assaleh, “A robust endpoint detection of speech for noisy environments with application to automatic speech recognition,” Proc. IEEE Intern. Conf. Acoustics, Speech, and Signal Processing, vol. 4, pp. 3808-3811, May 2002.
[13] J.A. Haigh, and J.S. Mason , “Robust voice activity

	detection using cepstral features," Proc. IEEE TENCON , pp. 321-324, 1993. [14] B.F. Wu, and K.C. Wang, "Robust Endpoint Detection Algorithm Based on the Adaptive Band-Partitioning Spectral Entropy in Adverse Environments," IEEE Trans. Speech and Audio Processing, pp. 762-775, Sept. 2005. [15] H.G. Xu, H.S. Li, J. Liu and R.S. Liu, "Endpoint detection algorithm for Mandarin digit recognition using DSP," in Proc. 6th Intern. Conf. Signal Processing, vol. 1, pp. 548-551, Aug. 2002. [16] L.R. Rabiner and M.R. Sambur, "An algorithm for determining the endpoints of isolated utterances," The Bell System Technical Journal, vol. 54, no. 2, pp. 297-315, 1975. [17] K. Kitayama, M. Goto, K. Itou and T. Kobayashi, "Speech Stgarter: Noise-robust endpoint detection by using filled pauses," Proc. 8th European Conf. Speech Communication and Technology (Eurospeech 2003), pp. 1237-1240, Setp. 2003. [18] J.A. Kogan and D. Margoliash, "Automated recognition of bird song elements from continuous recordings using dynamic time warping and hidden Markov models: A comparative study," Journal of the Acoustical Society of America, vol. 103, no. 4, pp. 2185-2195, 1998. [19] A. Selin, J. Turunen and J.T. Tanttu, "Wavelets in recognition of bird sounds," EURASIP Journal on Advances in Signal Processing, vol. 2007, Article ID 51806, 9 pages. [20] C. Kwan et al., "An automated acoustic system to monitor and classify birds," EURASIP Journal on Applied Signal Processing, vol. 2006, Article ID 96706, pp. 1-19. [21] S. Fagerlund, "Bird species recognition using support vector machines," EURASIP Journal on Advances in Signal Processing, vol. 2007, Article ID 38637, 8 pages. [22] S.Y. Kung, Digital Neural Networks, Prentice-Hall, Inc., 1993. [23] T. Kabaya and M. Matsuda, The Songs & Calls of 420 Birds in Japan, SHOGAKUKAN Inc., Tokyo, 2001
Follow up Questions	1. Could this technique be extended to the calls of other animals? 2. How would the results be different if there was noise cancellation preprocessing used? 3. Can the tune of a partly finished birdsong be finished? What would happen if this predicted "data" was fed into the algorithm?

Article Notes #14: Comparison between Resilient and Standard Back Propagation Algorithms Efficiency in Pattern Recognition

Source Title	Journal of Scientific & Engineering Research
Source citation (APA Format)	Mushgil, H. M., Alani, H. A., & George, L. E. (2015). Comparison between resilient and standard back propagation algorithms efficiency in pattern recognition. International Journal of Scientific & Engineering Research, 6(3), 773-778.
Original URL	https://www.researchgate.net/profile/Hanaa-Mohammed-8/publication/337257452_Comparison_between_Resilient_and_Standard_Back_Propagation_Algorithms_Efficiency_in_Pattern_Recognition/links/5dcd68984585156b35120e45/Comparison-between-Resilient-and-Standard-Back-Propagation-Algorithms-Efficiency-in-Pattern-Recognition.pdf
Source type	Journal Article
Keywords	Pattern Recognition, Neural Network, Back Propagation, Resilient BP, Pattern mode, Batch mode, Local minimum
Summary of key points + notes (include methodology)	Rprop backpropagation uses a more advanced version of backpropagation in that they use adapted parameters that change with every iteration. The upside is that the process converges much quicker than typical backpropagation. The downside to doing this is that the process is less efficient. The researchers wanted to figure out where the balance between efficiency and time complexity is. One of the biggest hurdles in gradient descent and backpropagation is the Flat-Spot Problem. Since the step of a backpropagation cycle is proportional to the steepness of the gradient, then flat regions which are not the bottom of the model take a long time to get through. This adds significant runtime to the model. RESULTS: Backpropagation batch mode took an average of 1.6 hr, and took about 1,500 trials to get to an error of below 10%.

	However, backpropagation pattern mode took 0.13hr, and only took 125 trials to achieve good accuracy (less than or equal to 10% error). Rprop backpropagation took 0.21hr and needed 300 trials to achieve good accuracy.
Research Question/Problem/ Need	Activation functions like the sigmoid have smaller and smaller slopes the further away one strays from 0. This can lead to gradients with small magnitudes even when the weights and biases are very far from their destination. This makes backpropagation very inefficient. Need: What type of backpropagation must be used so that the computer makes reasonable-sized gradient steps and has shorter runtime?
Important Figures	Hanaa Mohammed Mushgil - works with Artificial Intelligence. Is a member of the faculty of the Science/Computer Science department at Al-Nahrain University
VOCAB: (w/definition)	Backpropagation - *the process by which a machine learning algorithm calculates the negative gradient of its cost function (i.e. how to change the algorithm's weights and biases so that the program reduces error as quickly as possible)* Node - *a "neuron" in a neural network - contains a number either preset by the data itself or calculated through an activation function*
Cited references to follow up on	1] Amit G., Y. P. Kosta, Gaurang P., Chintan G.," Initial Classification Through Back Propagation In a Neural Network Following Optimization Through GA to Evaluate the Fitness of an Algorithm", International Journal of Computer Science & Information Technology (IJCSIT), Vol. 3, No 1, Feb 2011. [2] Bogdan M. W.," Neural Network Architectures and Learning ", IEEE, 2003. [3] Bo Y. , Ya-Dong W., Xiao-Hong S., Lijuan W. ," Solving Flat-Spot Problem In Back-Propagation Learning Algorithm Based On Magnified Error ", Proceedings of the Third International Conference on Machine Learning and Cybernetics, Shanghai, 26-29 August 2004, IEEE, 2004. [4] Devendra K. C., " Soft Computing Techniques and its Applications in Electrical Engineering", Springer-Verlag Berlin Heidelberg, 2008. [5] D. Randall Wilson Tony R. Martinez ," The Inefficiency of Batch Training for Large Training Sets ",Proceedings of the International Joint Conference on Neural Networks, IJCNN,Vol. II,

	pp. 113-117, July 2000. [6] Hamed A., Saeid S., Karim M.," Improving the Neural Network Training for Face Recognition using Adaptive Learning Rate, Resilient Back Propagation and Conjugate Gradient Algorithm ", International Journal of Computer Applications, (0975 – 8887) Volume 34– No.2, November 2011. [7] Hanaa Mohammed M., Dr. Loay E. George , Dr. Ban N. Dhannoon,"Agents Technology Based Cooperative Neural Networks for General Pattern Classification Model", International Journal of Scientific & Engineering Research, Volume 5, Issue 3, March-2014. [8] Iftikhar A., M.A Ansari, Sajjad M.,"Performance Comparison between Backpropagation Algorithms Applied to Intrusion Detection in Computer Network Systems ", 9th WSEA International Conference on Neural Networks (NN'08), Sofia, Bulgaria, May 2-4, 2008. [9] Insung J., Gi-Nam W.," Pattern Classification of Back-Propagation Algorithm Using Exclusive Connecting Network ", International Journal of Electrical and Electronics Engineering, 2008. [10] Kritika G., Sandeep K.," Implementation of Resilient Backpropagation & Fuzzy Clustering Based Approach for Finding Fault Prone Modules in Open Source Software Systems ", International Journal of Research in Engineering and Technology (IJRET), Vol. 1 No. 1, 2012. [11] Norhamreeza Binti Abdul Hamid," The Effect Of Adaptive Parameters On The Performance Of Back Propagation", PhD dissertation, Faculty of Computer Science and Information Technology ,University Tun Hussein Onn Malaysia,2012. [12] Sandhyahe S.," Neural networks from applied sciences and engineering ", Taylor & Francis Group, 2007.
Follow up Questions	1. Could typical backpropagation and Rprop be combined? 2. How pronounced would these differences become as the number of input nodes skyrocketed? This would be important for image processing. 3. How could the buffer value for the sigmoid function (which was mentioned to be 0.1 in the study) be changed so that backprop runs more efficiently?

Article Notes #15: Musical genre classification of audio signals

Source Title	IEEE Transactions on Speech and Audio Processing
Source citation (APA Format)	G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," in IEEE Transactions on Speech and Audio Processing, vol. 10, no. 5, pp. 293-302, July 2002, doi: 10.1109/TSA.2002.800560.
Original URL	https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=1021072
Source type	Journal Article
Keywords	Audio classification, beat analysis, feature extraction, musical genre classification, wavelets
Summary of key points + notes (include methodology)	The researchers extracted timbral texture features (common in speech recognition). They used the STFT algorithm and found MFCCs. Spectral rolloff differed between genres, as they were different shapes. RESULTS: Through their methodology, the researchers were able to achieve 61% accuracy for the classification of ten different musical genres.
Research Question/Problem/ Need	Classify audio in different genres (based on properties such as instrumentation, rhythm, and pitch range) automatically through a computer algorithm.
Important Figures	George Tzanetakis - works on acoustics and music theory applications. Professor of Computer Science, Faculty of Engineering, University of Victoria
VOCAB: (w/definition)	*STFT* - Short-Time Fourier Transform, used to find the Fast Fourier Transform (amplitude vs. time to amplitude vs. frequency) over a short period of time. STFTs at small time frames throughout an entire recording can be collated to achieve a spectrogram.

Cited references to follow up on	[1] F. Pachet and D. Cazaly, “A classification of musical genre,” in Proc. RIAO Content-Based Multimedia Information Access Conf., Paris, France, Mar. 2000. [2] S. Davis and P. Mermelstein, “Experiments in syllable-based recognition of continuous speech,” IEEE Trans. Acoust., Speech, Signal Processing, vol. 28, pp. 357–366, Aug. 1980. [3] J. Saunders, “Real time discrimination of broadcast speech/music,” in Proc. Int. Conf. Acoustics, Speech, Signal Processing (ICASSP), 1996, pp. 993–996. [4] E. Scheirer and M. Slaney, “Construction and evaluation of a robust multifeature speech/music discriminator,” in Proc. Int. Conf. Acoustics, Speech, Signal Processing (ICASSP), 1997, pp. 1331–1334. [5] D. Kimber and L. Wilcox, “Acoustic segmentation for audio browsers,” in Proc. Interface Conf., Sydney, Australia, July 1996. [6] T. Zhang and J. Kuo, “Audio content analysis for online audiovisual data segmentation and classification,” Trans. Speech Audio Processing, vol. 9, pp. 441–457, May 2001. [7] A. L. Berenzweig and D. P. Ellis, “Locating singing voice segments within musical signals,” in Proc. Int. Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA) Mohonk, NY, 2001, pp. 119–123. [8] E. Wold, T. Blum, D. Keislar, and J. Wheaton, “Content-based classifi- cation, search, and retrieval of audio,” IEEE Multimedia, vol. 3, no. 2, 1996. [9] J. Foote, “Content-based retrieval of music and audio,” Multimed. Storage Archiv. Syst. II, pp. 138–147, 1997. [10] G. Li and A. Khokar, “Content-based indexing and retrieval of audio data using wavelets,” in Proc. Int. Conf. Multimedia Expo II, 2000, pp. 885–888. [11] S. Li, “Content-based classification and retrieval of audio using the nearest feature line method,” IEEE Trans. Speech Audio Processing,

vol. 8, pp. 619–625, Sept. 2000.
[12] E. Scheirer, “Tempo and beat analysis of acoustic musical signals,” J.
Acoust. Soc. Amer., vol. 103, no. 1, p. 588, 601, Jan. 1998.
[13] M. Goto and Y. Muraoka, “Music understanding at the beat level: Real-time beat tracking of audio signals,” in Computational Auditory Scene Analysis, D. Rosenthal and H. Okuno, Eds. Mahwah, NJ: Lawrence Erlbaum, 1998, pp. 157–176.
[14] J. Laroche, “Estimating tempo, swing and beat locations in audio record-
ings,” in Proc. Int. Workshop on Applications of Signal Processing to Audio and Acoustics WASPAA, Mohonk, NY, 2001, pp. 135–139.
[15] J. Seppänen, “Quantum grid analysis of musical signals,” in Proc. Int.
Workshop on Applications of Signal Processing to Audio and Acoustics
(WASPAA) Mohonk, NY, 2001, pp. 131–135.
[16] J. Foote and S. Uchihashi, “The beat spectrum: A new approach to
rhythmic analysis,” in Proc. Int. Conf. Multimedia Expo., 2001.
[17] G. Tzanetakis, G. Essl, and P. Cook, “Automatic musical genre classifi-
cation of audio signals,” in Proc. Int. Symp. Music Information Retrieval
(ISMIR), Oct. 2001.
[18] L. Rabiner and B. H. Juang, Fundamentals of Speech Recogni-
tion. Englewood Cliffs, NJ: Prentice-Hall, 1993.
[19] B. Logan, “Mel frequency cepstral coefficients for music modeling,” in
Proc. Int. Symp. Music Information Retrieval (ISMIR), 2000.
[20] S. G. Mallat, A Wavelet Tour of Signal Processing. New York: Aca-
demic, 1999.
[21] I. Daubechies, “Orthonormal bases of compactly supported wavelets,”
Commun. Pure Appl. Math, vol. 41, pp. 909–996, 1988.
[22] T. Tolonen and M. Karjalainen, “A computationally efficient multip-
itch analysis model,” IEEE Trans. Speech Audio Processing, vol. 8, pp.
708–716, Nov. 2000.
[23] G. Tzanetakis, G. Essl, and P. Cook, “Audio analysis using the discrete
wavelet transform,” in Proc. Conf. Acoustics and Music Theory Appli-
cations, Sept. 2001.
[24] M. A. Bartsch and G. H. Wakefield, “To catch a chorus: Using chroma-

	based representation for audio thumbnailing," in Proc. Int. Workshop on Applications of Signal Processing to Audio and Acoustics Mohonk, NY, 2001, pp. 15–19. [25] R. N. Shepard, "Circularity in judgments of relative pitch," J. Acoust. Soc. Amer., vol. 35, pp. 2346–2353, 1964. [26] J. Pierce, "Consonance and scales," in Music Cognition and Comput- erized Sound, P. Cook, Ed. Cambridge, MA: MIT Press, 1999, pp. 167–185. [27] J.-J. Aucouturier and M. Sandler, "Segmentation of musical signals using hidden Markov models," in Proc. 110th Audio Engineering Society Convention, Amsterdam, The Netherlands, May 2001. 28] G. Tzanetakis and P. Cook, "Multifeature audio segmentation for browsing and annotation," in Proc. Workshop Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, NY, 1999. [29] R. Duda, P. Hart, and D. Stork, Pattern Classification. New York: Wiley, 2000. [30] D. Perrot and R. Gjerdigen, "Scanning the dial: An exploration of fac- tors in identification of musical style," in Proc. Soc. Music Perception Cognition, 1999, p. 88, (abstract). [31] D. Pye, "Content-based methods for the management of digital music," in Proc. Int. Conf Acoustics, Speech, Signal Processing (ICASSP), 2000. [32] G. Tzanetakis and P. Cook, "Sound analysis using MPEG compressed audio," in Proc. Int. Conf. Acoustics, Speech, Signal Processing (ICASSP), Istanbul, Turkey, 2000. [33] , "Marsyas: A framework for audio analysis," Organized Sound, vol. 4, no. 3, 2000.
Follow up Questions	1. How accurate would the model extend to other genres? 2. Could this be combined with instrument detection algorithms to better classify genres? 3. What happens when the song contains elements of multiple genres? Does the program predict both, or is it trained to make a confident decision?

Patent Notes #1: Speech based user recognition

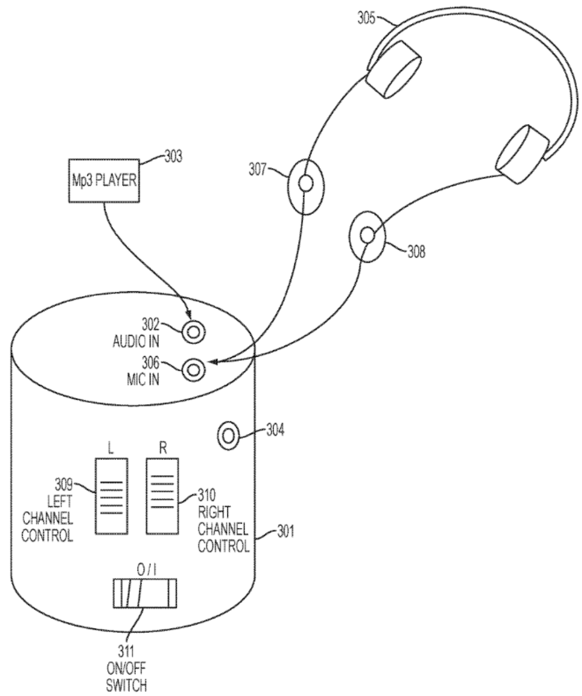
Source Title	United States Patent and Trademark Office
Source citation (APA Format)	Matsoukas , S., Khare, A., Krishnamoorthy, V., Somashekhar, S., & Mandal, A. (2022, March 8). SPEECH BASED USER RECOGNITION.
Original URL	https://patentimages.storage.googleapis.com/f1/05/6d/d6ac90996ac015/US11270685.pdf
Source type	Patent
Keywords	Network, Artificial Intelligence, Classification, Audio, Human voice
Summary of key points + notes (include methodology)	Their process was: - Receiving training data from the voice of a user - Connect the training data with the user. - Combine the training data itself with the user label - Patterns were determined by some metrics: I. Consistency of tone of voice (i.e. how “robotic” one’s voice is) II. Vocal range (Hz) III. Amplitude changes between time frames. (For example, a high positive change in the final time frames might indicate a question.) - Calculate confidence of model - Rerun this secondary data (the confidence) into the algorithm so that it is more accurate - Deploy the software on a mobile device
Research Question/Problem/ Need	How can an algorithm be generated that can distinguish between different users’ voices?

Important Figures	Spyridon Matsoukas - scientist at Amazon. Specializes in ASR algorithms and improving natural language processing.
VOCAB: (w/definition)	*ASR* - Automatic Speech Recognition - the ability for a computer to extract words being spoken and understand
Cited references to follow up on	U.S. PATENT DOCUMENTS 9,711,148 B1 * 7/2017 Sharifi GIOL 17/24 9,729,821 B1 * 8/2017 Fineberg H04W 4/029 2011/0243449 A1 * 10/2011 Hannuksela GIOL 17/00 382/190 2013/0311168 A1 * 11/2013 Li G06F 40/30 704/9 2014/0016835 A1 * 1/2014 Song GO6K 9/00892 382/118 2014/0244072 A1 * 8/2014 Vaghefinazari G01C 21/3608 701/2
Follow up Questions	1. Could Amazon combine their technology with that of musical assessment to check how well users sing? 2. Could Amazon use the voice data to create new voices that can be used for different purposes (e.g. functioning as a morning alarm)? 3. Is there any way to reduce training time and make the technology take up less (data) space?

Patent Notes #2: Active noise control and customized audio system

Source Title	United States Patent and Trademark Office
Source citation (APA Format)	Benattar, B. D. (2020, July 9). Active noise control and customized audio system.
Original URL	https://patentimages.storage.googleapis.com/c6/5f/ac/23afe97f9cdcda/US20200221220A1.pdf
Source type	Patent
Keywords	Ambient noise, filtration, listening devices, audio, sound

Summary of key points + notes (include methodology)	- Goal is to enhance an audio user's environment
	- Used ultrasound to determine location of source
	- Developed algorithms that detect the nature of the noise
	  I. Analyzed gap length and implications for character of the noise
	  II. Consistency of the noise in terms of frequency, amplitude, and gap length (uses data from I)
	- Diminished or strengthened ambient/primary noise so that the user can experience either
	- Control direction of noise, which matched with headphones. One can increase/decrease the directional audio from the east(right ear) or from the west (left ear).

	 FIG. 3
Research Question/Problem/ Need	Control certain qualities of an audio file, like increase/decrease ambient noise in a recording. Consequently, the technology must also be able to pick up on ambient vs. primary noise.
Important Figures	Benjamin D. Benattar - previously managing director at J. P. Morgan - inventor
VOCAB: (w/definition)	*ANC* - Active noise cancellation. The practice of headphones "canceling" ambient noise by transmitting the same frequency that the ambient signal is transmitting. The frequencies then cancel each other out.
Cited references to follow up on	U.S. Pat. No. 6,462,808 B2
Follow up Questions	1. What would be the effect of hooking up multiple machines consecutively? 2. Is it possible to distribute the noise through more outlets? How precise is the directionality system on the system? 3. Is there any way to create a smaller version of the model that works on earbuds?