About emotions and other things

Title: A computer science approach to the definition of emotions through audio features

Mariana Rodríguez Castañeda

[1]Universidad Nacional Autónoma de México

[2]Ms. Sc. Digital Signal Processing
Mexico City, Mexico.

mariana.rodriguez.c@comunidad.unam.mx

Abstract. *In this work, a deep dive into the selection of audio features is undertaken. Much has been accomplished regarding emotion recognition systems, but a question persists without an answer: the selection of features for the task has been ignored, and the focus on gathering more data and building larger architectures has garnered all the attention. However, in recent years, a reduction in computation is necessary. Consequently, this work presents a new methodology for the collection of meaningful data for the task of emotion recognition, employing genetic algorithms and dimensionality reduction techniques.*

1. General Information

In recent years, deep learning algorithms have introduced various approaches to emotion classification. In the audio domain, several representations can be utilized to train models. However, some of the primary issues with current architectures include high computational requirements, complexity, and the lack of interpretability of the data they entail. This work focuses on addressing these problems by concentrating the attention on pre-processing the database to emphasize features that effectively represent emotions. Resulting in inputs with reduced complexity and more significant features for the task. This approach leads to less noise in the data and a better understanding of emotions through their features.

This document describes the study of audio and the pursuit of a more interpretable definition of emotion from a computer science perspective. The definition of sound serves as a starting point, and its examination as a signal involves disciplines such as physics, electronics, and music, this last one providing clear examples of our interactions with them. Then, the focus shifts to the information that can be extracted from the explanation and understanding of the most relevant audio features. Emotions are approached as the interrogatory; first, a short review of literature is presented to identify intersections between disciplines, ultimately leading to the proposal of an algorithm to render emotions more interpretable from a computer science framework.

1.1. Introduction to Sound

Sound, as a physical phenomenon, involves the alternation of compression and rarefaction of the air. Our eardrum responds to the waveform of the sound propagated through the air, known as traveling waves or acoustic waves. Electronics introduced concepts such as frequency, sampling, and quantization what allow the technology of recording and playback of sound.

The human ear can perceive frequencies ranging from 20Hz to 20KHz. The fundamental frequency for men when speaking is around 120 Hz, while for women it's approximately 210 Hz[1]. The frequency of the sound to be captured is crucial due to the sampling rate. According to Nyquist's theorem, the sampling frequency must be at least twice the highest frequency to be sampled for accurate signal reconstruction. Common sampling rates include 44.1KHz for compact discs and 48KHz for digital video, but higher resolutions such as 192KHz are available for high-resolution audio [2].

Once the sound has been recorded, computer science introduced sound processing through coding and applications like digital workstations. More recently, artificial intelligence has gained significant recognition and importance in our community, from music recommendation systems and music improvisation, to the detection of COVID through coughing [3] or stomach noise detection[4]. The exploration of sound in the last decade has profoundly impacted our daily routine.

2. Sound Representations

2.1. Sheet Representation

One of the oldest representations of sound is found in music. The sheet music representation consists of a set of music notations for the player to interpret. It includes notations such as the key signature, articulation marks, and the duration of the notes. Each instrument has its own set of notations, and in the case of transposing instruments, they may have their own rules, which is why each player must learn a specific notation. Generally, the sheet representation is useful in human-computer interaction, as it is often synchronized with lyrics and sound for the implementation of cross-modal music information retrieval and data analysis[5].

2.2. Symbolic Representation

Another common representation in music is the piano roll representation. It consists of the eighty-eight keys on a piano, with the duration represented by the size of a hole at the position of the respective key. The MIDI representation has its own standard; it follows the same concept as the piano roll but encodes the information representing only performance indications allowing for digital interpretations. MusicXML is also a coding representation, unlike the performance indication aspect of the MIDI standard. MusicXML encodes information from the music sheet and its musical symbols [6].

2.3. Audio Representation

In this work, the audio representation is implemented, similar to other representations, sound is described with a set of parameters. In this case, the description of the waveform is crucial, as it depicts how air pressure changes over time. The variables of amplitude, period, duration, and phase are sufficient to describe it, where the inverse of the period equals the frequency in Hz. More complex sounds, as Fourier explained with the theory behind the Fourier transform, involve an addition of different frequencies with different waveforms, thus requiring a set of different descriptions.

When this sound is sampled, an array of one dimension for each channel is created, storing the normalized energy at a certain sample rate. This is called raw audio and is a representation of audio in the time domain. This is useful for extracting features related to duration, such as how long a key is played or the envelope of a signal. The envelope of a signal can be described in four main stages: firstly, when it starts, it's usually represented

as a positive slope and is called the attack phase. Then, there is the decay phase, during which the slope changes to negative for a brief moment until it reaches a stable value. The time the steady sound remains with a slope of zero is called the sustain phase. Lastly, the release phase starts when the sound lowers and its slope decays. The release phase is the final stage of the envelope and ends when it reaches zero energy.

Audio can also be analyzed in the frequency domain, where the focus is on the frequencies contained within the sound. While this representation may lose some meaning in the time domain, it provides different and important information about the sound, such as the weight of each component. For instance, variations in energy within the frequencies that compose a note are known as partials or harmonics. These frequencies accompany the fundamental frequency due to resonance, and their energy variations give each instrument a distinct sound for the same note, contributing to the instrument's unique timbre.

In light of the explanations regarding the time and frequency domains, analyzing frequency results in the loss of all time-related information, and vice versa. However, with the assistance of image, a combination of both domains can be studied simultaneously. This representation is known as a spectrogram, which samples the signal over time in the frequency domain. This means that a certain number of samples in the time domain will be analysed in the frequency domain. Using this technique, a heat map can be generated, with frequency as the dependent variable, time along the x-axis, and the intensity of the frequency represented by variations in color contrast.

3. Feature Extraction

The extraction of features involves gathering all the meaningful information that can be obtained from the signal. As described in the physical nature of sound, our auditory system captures and categorizes information, with the recognition process beginning in our auditory system. Even though our eardrums are sensitive only to variations in sound pressure, they can detect pressure changes as small as a few micro-pascals. This sensitivity forms the basis of the hypothesis that by replicating the information capture process of our ears, which operates as an energy detection method, we can simulate the process to recognize, classify, or localize events, instruments, or emotions.

3.1. Feature Construction

The implementation of a transformation on the input can result in a better representation of the signal, efficiently capturing the specifics of the acoustic wave for a particular task. These transformations may include basic operations, such as applying simple functions to subsets of variables. This is why audio signals offer a vast array of features to work with, as modeling our ear system alone is insufficient. Consequently, attempts have been made to model the psychoacoustic process, recognizing that hearing is not solely a mechanical phenomenon of wave propagation but also a sensory and perceptual event.

As an example of feature construction, and to illustrate its interpretability and transition from simple to complex transformations, consider the following:

3.1.1. Zero Crossing Rate

The zero crossing rate, as its name suggests, counts the number of times a signal crosses zero per frame, transitioning from negative to positive and vice versa. While this

transformation may seem simple, it can provide substantial information about the nature of the sound. For instance, a noisy sound tends to have more changes of sign, making this feature relevant for distinguishing between pitched and noisy sounds, or between vowels and consonants in speech.

Moreover, the zero crossing rate can serve as one of the features for more complex tasks such as music genre classification or emotion recognition. Changes in speaking speed have been identified as one of the features of certain emotions, with an increase in speaking rate observed during anger, in contrast to happiness, which does not exhibit a clear speaking rate[7].

3.1.2. Power

Loudness, the perception of the power of a signal, is calculated as the sum of the squared samples. Since the samples measure the energy of the signal and are related to the waveform, the power is also associated with variations in the time of sound pressure. This feature is particularly useful in distinguishing between happiness and anger, as both emotions have higher energy levels compared to others. However, anger is consistently at least 20dB louder than happiness.

3.1.3. MFCC

One of the most commonly used feature constructions in audio processing is the Mel-Frequency Cepstral Coefficients (MFCCs). This feature is characterized by modifying the signal to create a representation that models the ear's nonlinear response to sound. The process of obtaining MFCC features can be summarized in four steps:

1. Signal Framing and Windowing
2. Power Spectrum
3. Mel Filter Bank
4. Discrete Cosine Transforms (DCT)

The first step involves framing the signal to create quasi-stationary segments. It is crucial to select an appropriate window size, especially for voice signals, where a frame size of around 30ms is relevant due to the approximately 20ms interval between two glottal closures. This helps avoid artifacts such as discontinuities or frequency bleeding over. In the next step, a window is applied to create a sense of periodicity in the frame.

However, injecting a frequency into our signal can introduce features that were not present in the original information. The result depends on the properties of the window, which may introduce ripple artifacts. The Hanning window is commonly used due to its smooth decay to zero at the borders, and overlapping eliminates the addition of any other signals, while increasing spectral magnitudes of borders.

After framing and windowing the signal, the calculation of the power spectrum can be carried out. This involves applying the Fourier transform. It's important to note that from a computer science perspective, the complexity of the algorithms must be taken into consideration. Computing the discrete Fourier transform is expensive, as it requires $O(N^2)$multiplications and additions. However, the fast Fourier transform (FFT) reduces this complexity to O(N log N) by leveraging the symmetry of the transformation itself,

particularly with even sizes. After obtaining the FFT of the signal, the power is calculated to derive the power spectrum.

As mentioned, the objective of MFCC is to model the nonlinear response of the human ear, aiming to assign equal importance to frequencies as our ear does. The Mel Filter bank consists of a set of filters that map the power spectrum to the Mel-scale. It can be computed as follows:

$$m = 2595 \log_{10}(1 + \frac{f}{700})$$

After creating the logarithmic perception of pitch, the discrete cosine transform (DCT) is applied to retain the most relevant coefficients. This feature construction has become so popular that it is sometimes used as a sound representation.

3.2. Feature Reduction

Feature reduction, or more commonly used dimensionality reduction, is an essential technique in many applications dealing with high-dimensional datasets. The challenge with high-dimensional datasets lies not only in their computationally expensive processing but also in the quality of the features used as input. These features may include redundancies or irrelevancies[8], which introduce noise into the dataset, making it challenging to analyze using other tools.

3.2.1. PCA

Principal Component Analysis (PCA) aims to identify the K eigenvectors with the highest eigenvalues, mapping the data to a new plane where the representation is more interpretable for the given task. Initially, the principal components across all dimensions are determined, followed by selecting a number of components based on the explained variance. It has demonstrated to be effective in high dimensional datasets[9].

3.2.2. U-map

Uniform Manifold Approximation and Projection (UMAP), unlike PCA, is a nonlinear technique for dimension reduction. UMAP is constructed from a theoretical framework based on Riemannian geometry and algebraic topology. Due to its emphasis on neighboring points, it preserves more of the global structure when determining the new data space compared to other methods such as tSNE [10].

3.2.3. Embeddings

Embeddings are features constructed by an autoencoder, a downsampling architecture that retains essential information for signal reconstruction. The mid-level representation of MFCC is often generated for use as an input. In the field of emotion recognition, embeddings have been employed to demonstrate how emotions modulate language and content through the distribution of reconstruction errors.

4. Feature Selection

The quality of a feature space is crucial to mitigate the most significant issues in modern architectures, particularly impacting classifier performance in terms of accuracy, complexity, speed, and interpretability.

The list of features of audio signal is extensive and should be selected based on the task objective. As mentioned, audio signals can be represented in both the frequency and time domains, each offering a set of features. The domain is the primary method for categorizing features, distinguishing between continuous or spectral representations, and can also include qualitative or Teager energy operator-based (TEO-based) features.[11]

5. References

[1] Brian, "Psychoacoustics," Springer eBooks, pp. 459–501, Jan. 2007

[2] S. Salazar, A. Kapur, G. Wang, and P. Cook, Programming for Musicians and Digital Artists. Simon and Schuster, 2014.

[3] V. Despotovic, M. Ismael, M. Cornil, R. M. Call, and G. Fagherazzi, "Detection of COVID-19 from voice, cough and breathing patterns: Dataset and preliminary results," Computers in Biology and Medicine, vol. 138, p. 104944, Nov. 2021, doi: https://doi.org/10.1016/j.compbiomed.2021.104944.

[4] A. Maria and A. S. Jeyaseelan, "Development of Optimal Feature Selection and Deep Learning Toward Hungry Stomach Detection Using Audio Signals," Journal of Control, Automation and Electrical Systems, vol. 32, no. 4, pp. 853–874, Apr. 2021, doi: https://doi.org/10.1007/s40313-021-00727-8.

[5] M. Müller, T. Prätzlich, B. Bohl and J. Veit, "Freischutz digital: A multimodal scenario for informed music processing,"2013 14th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS), Paris, France, 2013, pp. 1-4, doi: 10.1109/WIAMIS.2013.6616168.

[6] M. Müller, Fundamentals of Music Processing. Cham: Springer International Publishing, 2015. doi: https://doi.org/10.1007/978-3-319-21945-5.

[7] D. Morrison, R. Wang, and L. C. De Silva, "Ensemble methods for spoken emotion recognition in call-centres," Speech Communication, vol. 49, no. 2, pp. 98–112, Feb. 2007, doi: https://doi.org/10.1016/j.specom.2006.11.004.

[8] J. A. Richards, "Feature Reduction," Springer eBooks, pp. 403–446, Jan. 2022, doi: https://doi.org/10.1007/978-3-030-82327-$6_1 0$.

[9] H. Xu, C. Caramanis and S. Mannor, .ºutlier-Robust PCA: The High-Dimensional Case,ïn IEEE Transactions on Information Theory, vol. 59, no. 1, pp. 546-572, Jan. 2013, doi: 10.1109/TIT.2012.2212415.

[10] "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction — umap 0.5 documentation," umap-learn.readthedocs.io. https://umap-learn.readthedocs.io/en/latest/

[11] R. Jahangir, Y. W. Teh, F. Hanif, and G. Mujtaba, "Deep learning approaches for speech emotion recognition: state of the art and research challenges," Multimedia Tools and Applications, Jan. 2021, doi: https://doi.org/10.1007/s11042-020-09874-7.