Analysis on Audio Generalized Key Features and its Elementary Extraction Process

Chandini M.S.[1],Dept.of AIML,BITM,Ballari,Karnataka
Dr.Usha B.A[2]Dept of CSE,BMSIT,Bengaluru,Kanataka

Abstract

In the rapidly growing technological era, Virtual assistants such as Alexa, Siri and Google Home are the major Applications of machine intelligent models.These are largely built atop models that can perform artificial cognition from audio data.To train any statistical or ML model, we need to first extract useful features from an audio signal. Audio feature extraction is a necessary step in audio signal processing, which is a subfield of signal processing. It deals with the processing or manipulation of audio signals. It removes unwanted noise and balances the time-frequency ranges by converting digital and analog signals. It focuses on computational methods for altering the sounds.This paper introduces most commonly used audio features that are used as inputs to models and the process of extracting it from the audio input.

Keywords: Audio signal processing, Machine learning.

I. Introduction

An audio signal is a representation of sound. According to physics, sound is a travelling vibration, i.e. a wave that moves through a medium such as the air. The sound wave is transferring energy from particle to particle until it is finally "received" by our ears and perceived by our brains. The two basic attributes of sound are **amplitude** (loudness) and **frequency** (a measure of the wave's vibrations per time unit).It encodes all the necessary information required to reproduce sound. Sound is an analog signal that has to be transformed to a digital signal, in order to be stored in computers and analyzed by software. This analog to digital conversion includes two processes: **sampling** and **quantization**.

Sampling is used to convert the time-varying continuous signal x(t) to a discrete sequence of real numbers x(n). The interval between two successive discrete samples is the sampling period (Ts). We use the sampling frequency (fs = 1/Ts) as the attribute that describes the sampling process.**Quantization** is the process of replacing each real number, x(n), of the sequence of samples with an **approximation** from a finite set of discrete values as shown in Fig 1. In other words, quantization is the process of reducing the infinite number precision of an audio sample to a finite precision as defined by a particular number of bits.

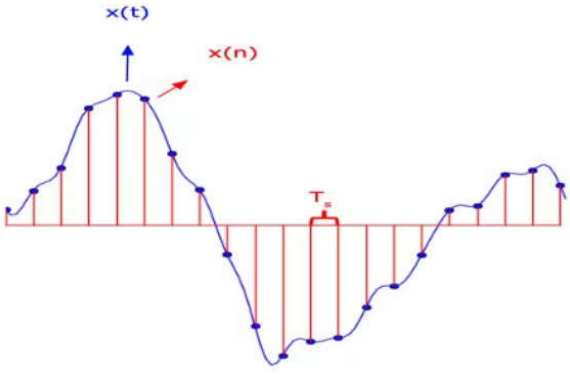


Fig 1: Sample Sound segment

II. Key Audio Features:

Audio features are description of sound or an audio signal that can basically be fed into statistical or ML models to build intelligent audio systems. Audio applications that use such features include audio classification, speech recognition, automatic music tagging, audio segmentation and source separation, audio fingerprinting, audio de-noising, music information retrieval, and more.Generally audio features are categorised with regards to the following aspects:

- **Level of Abstraction**: High-level, mid-level and low-level features of musical signals.
- **Temporal Scope**: Time-domain features that could be instantaneous, segment-level and global.
- **Musical Aspect**: Acoustic properties that include beat, rhythm, timbre (colour of sound), pitch, harmony, melody, etc.
- **Signal Domain**: Features in the time domain, frequency domain or both.
- **ML Approach**: Hand-picked features for traditional ML modeling or automatic feature extraction for deep learning modelling

a. Audio features in perspective ofSignal domain

1.Time domain: These are extracted from waveforms of the raw audio. Zero crossing rate, amplitude envelope, and RMS energy are examples.

Some time-domain audio features:

- **Amplitude Envelope** of a signal consists of the maximum amplitudes value among all samples in each frame. This feature gives a rough idea of loudness. It is however, sensitive to outliers. This feature has been extensively used for onset detection and music genre classification.

- **Root Mean Square Energy** is based on all samples in a frame. It acts as an indicator of loudness, since higher the energy, louder the sound. It is however less sensitive to outliers as compared to the Amplitude Envelope. This feature has been useful in audio segmentation and music genre classification tasks.

- **Zero-Crossing Rate** is simply the number of times a waveform crosses the horizontal time axis. This feature has been primarily used in recognition of percussive vs pitched sounds, monophonic pitch estimation, voice/unvoiced decision for speech signals, etc.

2. **Frequency domain**: These focus on the frequency components of the audio signal. Signals are generally converted from the time domain to the frequency domain using the *Fourier Transform*. Band energy ratio, spectral centroid, and spectral flux are examples.**Time-frequency representation**combine both the time and frequency components of the audio signal. The time-frequency representation is obtained by applying the Short-Time Fourier Transform (STFT) on the time domain waveform. Spectrogram, mel-spectrogram, and constant-Q transform are examples.

b. **Audio features under the** ML **approach**

1.Traditional Machine Learning approach considers all or most of the features from both time and frequency domain as inputs into the model. Features need to be hand-picked based on its effect on model performance. Some widely used features include Amplitude Envelope, Zero-Crossing Rate (ZCR), Root Mean Square (RMS) Energy, Spectral Centroid, Band Energy Ratio, and Spectral Bandwidth.

2.Deep Learning approach considers unstructured audio representations such as the spectrogram or MFCCs. It extracts the patterns on its own. By late 2010s, this became the preferred approach since feature extraction is automatic. It's also supported by the abundance of data and computation power.

3. Features input into neural network architectures

Commonly used features or representations that are directly fed into neural network architectures are spectrograms, mel-spectrograms, and Mel-Frequency Cepstral Coefficients (MFCCs).

- **Spectrograms:** A spectrogram is a visual depiction of the spectrum of frequencies of an audio signal as it varies with time. Hence it includes both time and frequency aspects of the signal. It is obtained by applying the Short-Time Fourier Transform (STFT) on the signal. In the simplest of terms, the STFT of a signal is calculated by applying the Fast Fourier Transform (FFT) locally on small time segments of the signal.

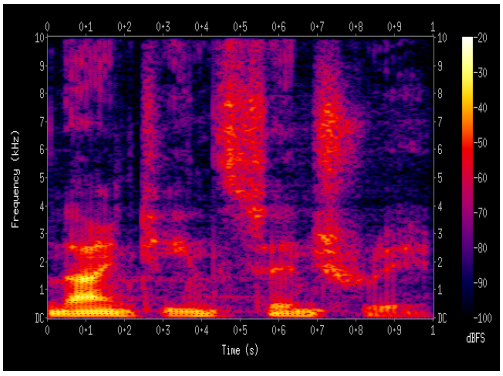
- **mel-spectrogram:** Apparently, we humans perceive sound logarithmically. We are better at detecting differences in lower frequencies than higher frequencies. For example, we can easily tell the difference between 500 and 1000 Hz, but we will hardly be able to tell a difference between 10,000 and 10,500 Hz, even though the distance between the two pairs is the same. Hence, the **mel scale** was introduced. It is a logarithmic scale based on the principle that equal distances on the scale have the same *perceptual* distance.

Conversion from frequency (f) to mel scale (m) is given by

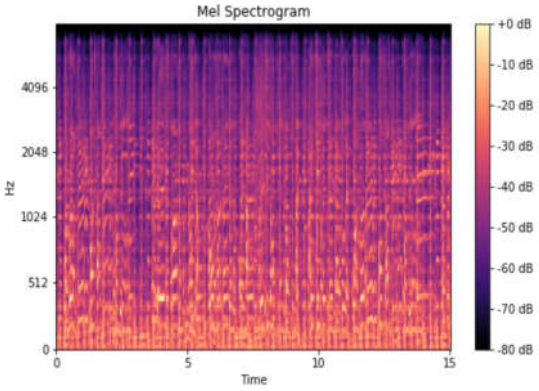
$$m=2595\cdot\log(1+f500)$$

A mel-spectrogram is a therefore a spectrogram where the frequencies are converted to the mel scale.

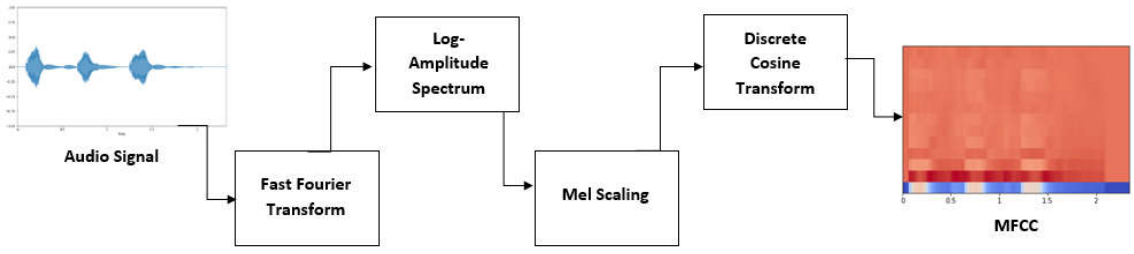
- **MFCCs :** The information of the rate of change in spectral bands of a signal is given by its cepstrum. A cepstrum is basically a spectrum of the log of the spectrum of the time signal. The resulting spectrum is neither in the frequency domain nor in the time domain and hence, it was named the **quefrency** (an anagram of the word *frequency*) domain. The Mel-Frequency Cepstral Coefficients (MFCCs) arenothing but the coefficients that make up the mel-frequency cepstrum.The cepstrum conveys the different values that construct the formants (a characteristic component of the quality of a speech sound) and timbre of a sound. MFCCs thus are useful for deep learning models.



a. Spectrogram



b. Mel Spectrogram



c. MFCCs

Fig 2: Illustrating Variations input feature fed into neural network

MFCC, have been largely employed in the speech recognition field but also in the field of audio content classification due to the fact that their computation is based on perceptual-based
frequency scale in the first stage (the human auditory model in which is inspired the frequency
Mel-scale). After obtaining the frame-based Fourier transform, outputs of a Mel-scale filter bank are logarithmized and finally they are decorrelated by means of the Discrete Cosine Transform(DCT). Only first DCT coefficients are used to gather information that represents the low frequency component of the signal's spectral envelope.

MFCC''s have been used also for music classification, singer identification, environmental sound classification, audio-based surveillance systems, being also embedded in hearing aids and even employed detect breath sound as an indicator of respiratory health and disease. Also, some particular extensions of MFCC have been introduced in the context of speech recognition and speaker verification in the aimof obtaining more robust spectral representation in the presence of noise

c. **Band Energy Ratio, Spectral Centroid and Spectral Bandwidth features**

The Band Energy Ratio (BER) provides the relation between the lower and higher frequency bands. It can be thought of as the measure of how dominant low frequencies are. This feature has been extensively used in music/speech discrimination, music classification etc.

The Spectral Centroid provides the center of gravity of the magnitude spectrum. In other words, it gives the frequency band where most of the energy is concentrated. It maps into a very prominent timbral feature called "brightness of sound" (energetic, open, dull). Mathematically, the spectral centroid is the weighted mean of the frequency bins.

The spectral bandwidth or spectral spread is derived from the spectral centroid. It is the spectral range of interest around the centroid, that is, the variance from the spectral centroid. It has a direct correlation with the perceived timbre. The bandwidth is directly proportional to the energy spread across frequency bands. Mathematically, it is the weighted mean of the distances of frequency bands from the Spectral Centroid

III. Audio Feature Extraction Methodology

Regardless of any uniquesort ofproblem given, the shape of the underlying machinemay bedefinedby way of a usual and common structurelayoutthat is depicted in Figure 3.

Fig 3: Audio recognizing machine Architecture

In a first stage, the continuous audio stream captured by a microphone is segmented into shorter signal chunks by means of a **windowing process**. This is achieved by sliding a window function over the theoretically infinite stream of samples of the input signal, and ends up by converting it into a continuous sequence of finite blocks of samples. Thus, the system will be capable of operating on sample chunks of finite length. Moreover, depending on the length of the window function, the typically non-stationary audio signal can be assumed to be quasi-stationary within each frame, thus facilitating subsequent signal analysis. The choice of the type and length of the window function, as well as the overlap between consecutive signal frames, is intimately related to the machine hearing application at hand. It seems logical that, for instance, the length of the window function should be proportional to the minimum length of the acoustic events of interest. Therefore, window lengths between 10 and 50 milliseconds are typically employed to process speech or to detect transient noise events , while windows of several seconds are used in computational auditory scene analysis (CASA) applications.

Once the incoming audio stream has been segmented into finite length chunks, audio features are extracted from each one of them using stage-11 module. The goal of **feature extraction** is to obtain a compact representation of the most salient acoustic characteristics of the signal, converting a N samples long frame into K scalar coefficients (with K << N), thus attaining a data compaction that allows increasing the efficiency of subsequent processes . To that effect, these features may consider the physical or perceptual impact of signal contents computed in the time, frequency, etc. domains. To keep this time information, the features extracted from several subsequent signal frames can be merged into a single feature vector. It should be noted that,due to this feature merging process, the feature vectors acquire a very high dimensionality that may represent a hurdle to the subsequent audio analysis process.

In order to compact the feature vectors, feature extraction techniques are sometimes followed by a data dimensionality reduction process.

Finally, an **audio analysis** task must be conducted upon the feature vectors obtained in the previous step. Of course, audio analysis is a generic label that tries to encompass any audio processing necessary to tackle the specific machine hearing application at hand. For instance, in case that recognizing a specific type of sound was the goal of our hearing machine, this audio analysis block would consist of a supervised machine learning algorithm that should first build representative acoustic models upon multiple samples from each sound class that we want the system to recognize, to subsequently classify any incoming unknown sound signal into one of the predefined classes based on the information acquired during the algorithm's training phase.

Conclusion

Machine Intelligent systems learnt based on audio signals are significantly increasing in the digital trend. It is essential to understand for we humans the structure, characteristic feature and the behaviour for audio signal to adapt to the interesting innovation. This paper has presented an up-to-date review of the most relevant generalized audio features and its extraction techniques related to audio recognizing machine which have been developed for the analysis of speech, music and environmental sounds. Further, this can be revisited to elaborate and describefeature extraction techniques computed on the wavelet and imagedomains, obtained from multilinear or non-linear parameterizations, together with those derived from specific representations such as the machine-pursuit algorithm.

References:

1. Chu, S.; Narayanan, S.S.; Kuo, C.J. Environmental Sound Recognition With Time-Frequency Audio Features. IEEE Trans. Audio Speech Lang. Process. 2009, 17, 1142–1158.
2. Peltonen, V.; Tuomi, J.; Klapuri, A.; Huopaniemi, J.; Sorsa, T. Computational Auditory Scene Recognition. In Proceedings of the 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Orlando, FL, USA, 13–17 May 2002; Volume 2, pp. II:1941 – II:1944
3. Oppenheim, A.V.; Schafer, R.W. Discrete-Time Signal Processing; Prentice Hall: Upper Saddle River, NJ, USA, 1989
4. Gerhard, D. Audio Signal Classification: History and Current Techniques; Technical Report TR-CS 2003-07; Department of Computer Science, University of Regina: Regina, SK, Canada, 2003
5. Center Point Audio. 2021. "Understanding the difference between Analog and Digital Audio." Center Point Audio. Accessed 2021-05-23
6. Wikipedia. 2021b. "Audio Signal Processing." Wikipedia, May 7. Accessed 2021-05-23.
7. Velardo, Valerio. 2020c. "Frequency-Domain Audio Features." The Sound of AI, on YouTube, October 12. Accessed 2021-05-23.