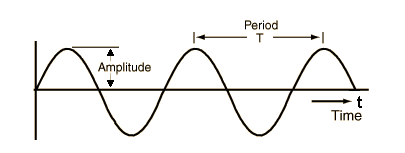
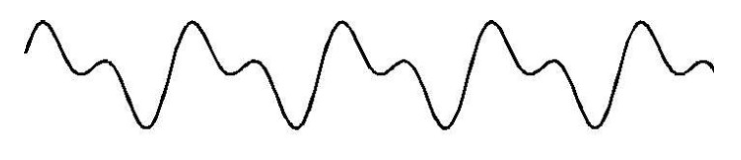
1. **Tìm hiểu chung về Sound Signal**

* Sound là một dạng tín hiệu được sinh ra từ sự thay đổi áp suất không khí, bắt nguồn từ một dao động cơ học nào đó,
* Cường độ của sự thay đổi áp suất này có thể đo được, nó là biên độ (Applitude)
* Sound Signal thường lặp đi lặp lại theo một kỳ T, đồ thị của nó dưới dạng sóng:



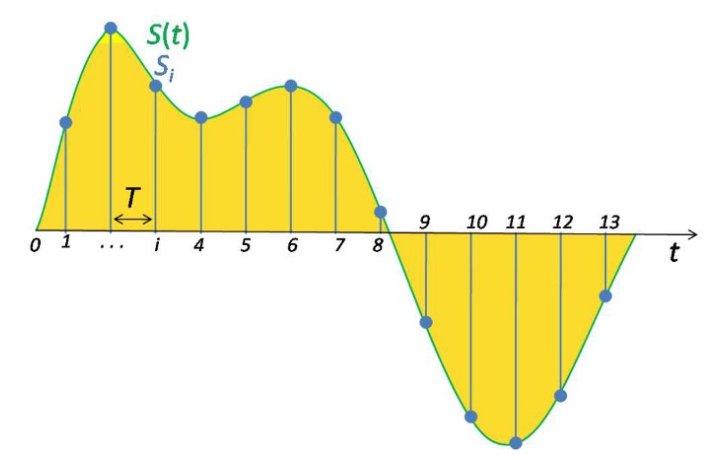
* Giá trị nghịch đảo của chu kỳ T, ký hiệu là f, gọi là tần số của Sounf Signal: nó thể hiện mức độ dao động của Signal trong thời gian 1s (bằng số đỉnh của Signal trong 1s). Đơn vị của f là Hertz.
* Trong thực tế, đồ thị của Sound Signal thường không đơn giản dạng Sin như vậy, mà phức tạp hơn rất nhiều. Tuy nhiên, chúng vẫn có dạng sóng và có chu kỳ.
* Ví dụ, đồ thị của một dụng cụ âm nhạc như dưới đây:



* Nhìều Sound Signal có thể được tổng hợp thành một Sound Signal duy nhất.
* Về mặt cảm thụ sinh học, mỗi Sound Signal có một đặc trưng riêng, gọi là âm sắc (timbre). Tai người có thể phân biệt được các Sound khác nhau dựa vào âm sắc của các Sound đó.

1. **Số hóa Sound Signal**

* Sound Signal là một dạng tín hiệu tương tự (Analog Signal). Tuy nhiên để thuận lợi trong việc lưu trữ, xử lý và truyền tải, Sound Signal được chuyển sang dạng Số (Digital Signal).
* Việc chuyển đổi này phải đảm bảo không làm mất mát quá nhiều thông tin so với tín hiệu gốc, và từ tín hiệu đã chuyển đổi có thể dễ dàng khôi khục lại gần như nguyên vẹn tín hiệu ban đầu.
* Số hóa Sound Signal được thực hiện bằng cách lấy giá trị biên độ của nó tại các vị trí cách đều nhau trong mỗi chu kỳ.



* Mỗi vị trí như vậy được gọi là một mẫu (Sample). Ta có khái niệm Tần số lấy mẫu (Sample Rate) là số lượng mẫu trong 1s.
* Một câu hỏi đặt ra là giá trị của Sample Rate là bao nhiêu là hợp lý.
  + Hai nhà khoa học Nyquist và Shannon đã đồng thời, độc lập đưa ra một định lý, gọi là Định lý lấy mẫu Nyquist–Shannon, về việc xác đinh giá trị của Sample Rate.
  + Định lý phát biểu như sau:

Một hàm số tín hiệu x(t) không chứa bất kỳ thành phần tần số nào lớn hơn hoặc bằng một giá trị fm có thể biểu diễn chính xác bằng tập các giá trị của nó với chu kỳ lấy mẫu T = 1/(2fm).

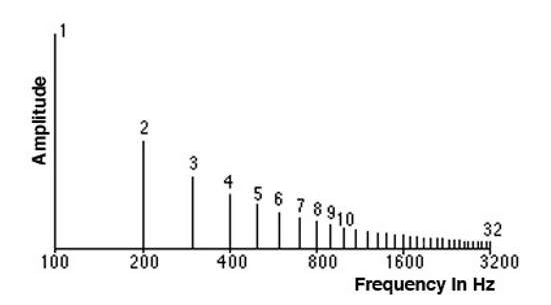
=> Như vậy, tần số lấy mẫu phải thoả mãn điều kiện fs ≥ 2fm. Tần số giới hạn fs/2 này được gọi là tần số Nyquist và khoảng (-fs/2; fs/2) gọi là khoảng Nyquist. Thực tế, tín hiệu trước khi lấy mẫu sẽ bị giới hạn bằng một bộ lọc để tần số tín hiệu nằm trong khoảng Nyquist.

* Định lý Nyquist–Shannon được áp dụng cho tín hiệu nói chung chứ không phải chỉ riêng tín hiệu âm thanh.

=> Đến dây ta phân biệt được 2 khái niệm Audio và Sound. Sound có nguồn gốc là các dao động cơ học lan truyền trong các môi trường đàn hồi (rắn, lỏng, khí), còn Audio được sinh ra từ các thiết bị điện tử thông qua quá trình lấy mẫu, ghi âm, …

1. **Chuẩn bị dữ liệu Audio cho Deep Learning model**

* Đối với sự trợ giúp của Deep Learning, đối với dữ liệu Audio, chúng ta sẽ chuyển chúng sang dạng Image và sử dụng kiến trúc CNN kinh điển để xử lý chúng.
* Để hiểu rõ từ Audio chuyển sang Image thì tiếp theo cần tìm hiểu Spectrum và Spectrogram.
* Hình dưới đây thể hiện Spectrum của một đoạn nhạc.
  + Trục tung là giá trị biên độ.
  + Trục hoành là giá trị tần số của mỗi tín hiệu thành phần.



Tần số có giá trị nhỏ nhất được gọi là tần số Cơ bản. Các tần số khác là bội số của tần số cơ bản được gọi là Sóng hài (harmonics). Ví dụ, nếu tần số cơ bản là 200Hz, thì các sóng hài của nó là 400Hz, 600Hz, …

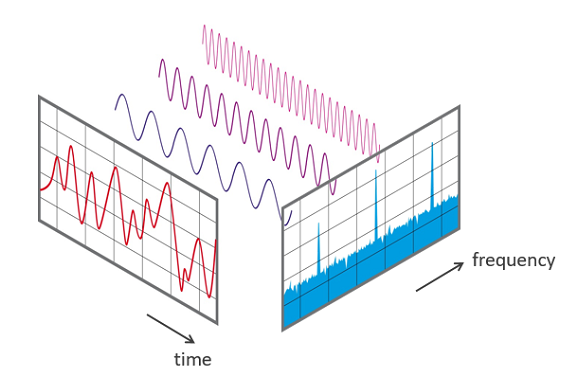
**3.1 Spectrum**

* Sound Signal trong thực tế thường là sự tổng hợp của nhiều tín hiệu thành phần khác nhau.
  + Ví dụ: tiếng nói của chúng ta bao gồm các tạp âm (Noise) xung quanh. Mỗi tín hiệu thành phần đó lại có tần số khác nhau, do đó tổng hợp các tần số thành phần ta có tần số của Sound Signal.

=> Spectrum chính là tập hợp các tần số của các tín hiệu thành phần tạo nên Sound Signal.

**3.2 Miền thời gian và miền tần số**

* Mỗi Sound Signal đều có 2 miền giá trị: Thời gian và Tần số. Trong mỗi miền đó, Sound Signal được thể hiện theo cách khác nhau.



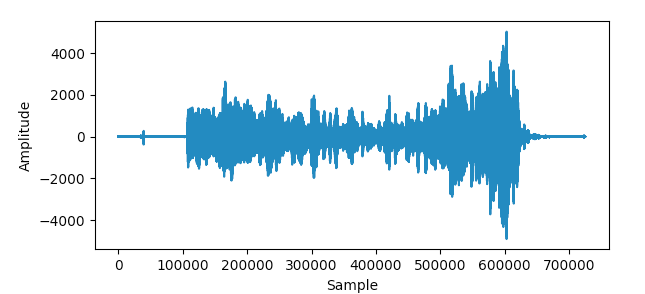
* Trong miền thời gian: Sound Signal mô tả sự thay đổi của biên độ theo thời gian. Biên độ nằm trên trục tung và thời gian nằm trên trục hoành.
* Trong miền tần số: Sound Signal mô tả sự thay đổi của biên độ theo tần số. Biên độ nằm trên trục tung và tần số nằm trên trục hoành.

**3.3 Spectrogram**

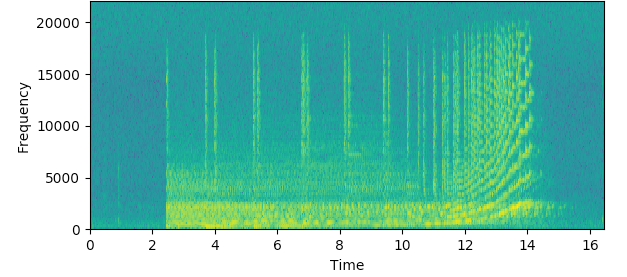
* Ở trên ta đã nói đến mối liên hệ giữa biên độ với thời gian và biên độ với tần số. Hơn thế nữa tần số và thời gian cũng có liên hẹ với nhau. Đồ thị thể hiện mối quan hệ đó gọi là Spectrogram:
  + Trục X thể hiện thời gian.
  + Trục Y thể hiện tần số.

=> Spectrogram thể hiện sự thay đổi của tần số theo thời gian. Độ lớn của biên độ cũng được Spectrogram thể hiện thông qua màu sắc. Màu sắc càng sáng thì thì biên độ càng lớn và ngược lại.

* Ví dụ:
  + Soung Signal trong miền thời gian:



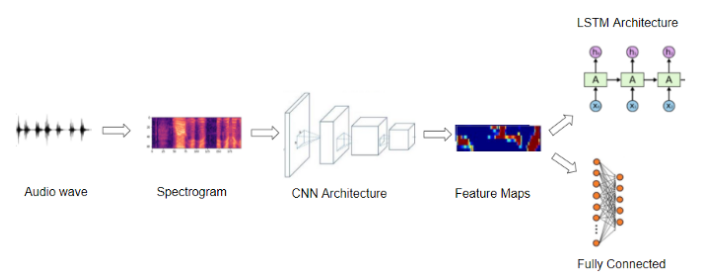
* + Spectrogram của Sound Signal trên:



=> Có thể nói, Spectrogram là cách thể hiện tốt nhất của Sound Signal, dưới dạng một hình ảnh, bởi vì nó mang đầy đủ thông tin về thời gian, tần số và biên độ. Và do vậy, Spectrogram được sử dụng làm dữ liệu đầu vào cho các Deep Learning model, như một bức ảnh thông thường.

* Spectrogram được sinh ra bằng cách áp dụng phép biến đổi Fourier lên một Signal để phân tách Signal đó thành các tần số thành phần. Python có thư viện giúp húng ta thực hiện Fourier Transform một cách dễ dàng.

1. **Audio Deep Learning models**



**B1:** Chuyển đổi Raw Audio sang Spectrogram.

**B2:** Áp dụng một số kỹ thuật tăng cường dữ liệu. Các kỹ thuật này cũng có thể áp dụng cho Raw Audio. (Không bắt buộc)

**B3:** Xây dựng CNN model và huấn luyện nó.

**B4:** Output của CNN là các Feature Maps. Tùy vào bài toán cụ thể mà chúng ta sinh ra các dạng Output khác nhau.

* **Audio Classification:** Cho Feature Maps đi qua một bộ phân lớp (FC, SVM, …) để sinh ra nhãn cho Audio.
* **Speech to Text:** Cho Feature Maps đi qua RNN để sinh ra Text tương ứng với Audio.