

Âm thanh và Dạng sóng

Âm thanh là gì?

1. Được tạo ra bởi sự rung động của một vật thể.
2. Sự rung động đó khiến các phần tử trong không khí dao động.
3. Sự thay đổi áp suất không khí tạo ra sóng cơ.



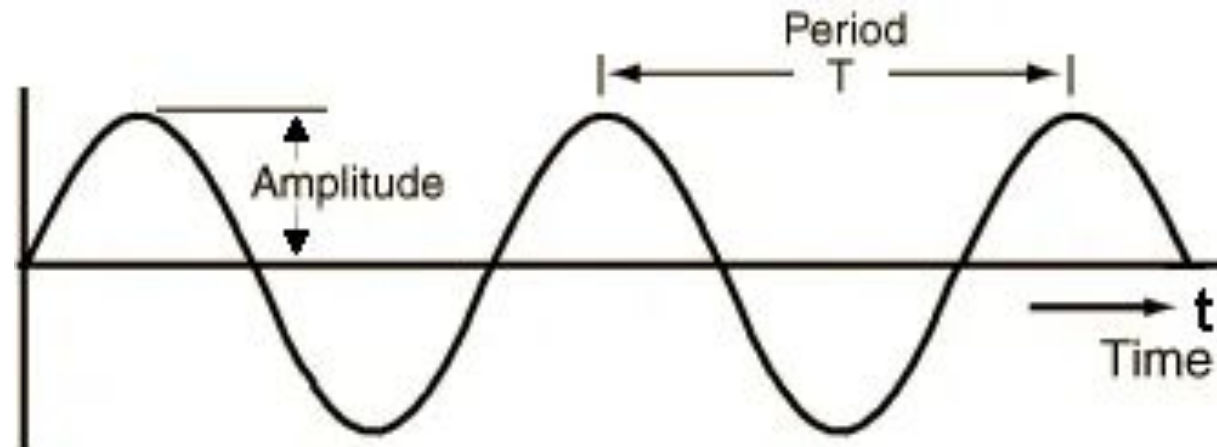
1. Dao động lan truyền trong không gian.
2. Năng lượng truyền từ điểm dao động ban đầu này đến điểm khác trong môi trường.
3. Môi trường bị biến dạng (Không khí, nước,...).

Sóng âm

1. Là một phần của sóng cơ, có tần số nằm trong khoảng mà tai người có thể cảm nhận được (khoảng 20Hz đến 20.000Hz).
2. Khi truyền đến tai người, sóng âm sẽ làm màng nhĩ dao động và tạo ra cảm giác nghe.

Âm thanh có 1 số đặc tính cơ bản sau:

1. Biên độ (Amplitude): Đây là chiều cao của sóng âm, biểu thị cường độ của âm thanh. Biên độ lớn hơn tương ứng với âm thanh lớn hơn.
2. Chu kỳ (Period): Là khoảng thời gian để tín hiệu hoàn thành một chu kỳ sóng đầy đủ.
3. Tần số (Frequency): Là số chu kỳ sóng hoàn thành trong một giây. Tần số được đo bằng đơn vị Hertz (Hz) và là nghịch đảo của chu kỳ. Ví dụ, nếu một sóng hoàn thành một chu kỳ trong 1/100 giây, tần số của nó sẽ là 100 Hz.



1. Tỷ lệ năng lượng được truyền đi
2. Năng lượng trên 1 đơn vị thời gian từ nguồn âm đẳng hướng
3. Đơn vị đo W (Watt)

Cường độ âm thanh

1. Là công suất âm thanh trên một đơn vị diện tích
2. Đơn vị đo W/m^2

Ngưỡng nghe của con người - Threshold of hearing

1. Là mức cường độ âm thanh nhỏ nhất con người có thể nghe được
2. Ngưỡng nghe của người: $TOH = 10^{-12} \text{ W/m}^2$

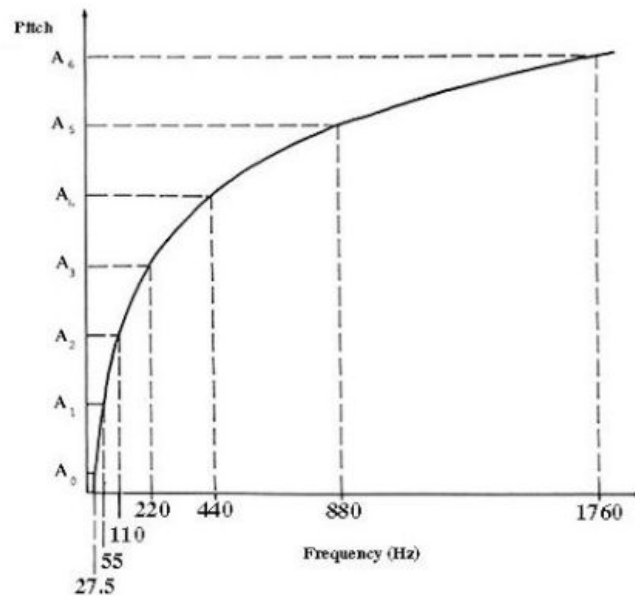
Ngưỡng đau của con người - Threshold of pain

1. Là mức cường độ âm thanh lớn nhất gây ra cảm giác đau tai
2. Ngưỡng đau của người: $TOP = 10 \text{ W/m}^2$

1. Thang logarit
2. Đơn vị đo dB
3. Tỷ số giữa 2 giá trị cường độ
4. Sử dụng ngưỡng nghe - threshold of hearing làm tham chiếu

$$dB(I) = 10 \cdot \log_{10}\left(\frac{I}{I_{TOH}}\right)$$

1. Pitch được cảm nhận theo thang logarit chứ không phải tuyến tính. VD: tăng gấp đôi tần số (100 - 200Hz). Tai người cảm nhận theo thang logarit. Để tăng gấp đôi độ cao âm cần tăng tần số theo 1 tỷ lệ khác.
2. Tỷ lệ 2:1: Hai tần số được cảm nhận tương tự nhau nếu chúng khác nhau theo lũy thừa cơ số 2. VD: sự khác biệt giữa 200Hz và 400Hz sẽ tương đương với sự khác biệt giữa 400Hz và 800Hz.
3. Độ cao âm - Tần số: dựa vào công thức và biểu đồ sau ta có thể tính tần số chuẩn của một nốt nhạc.



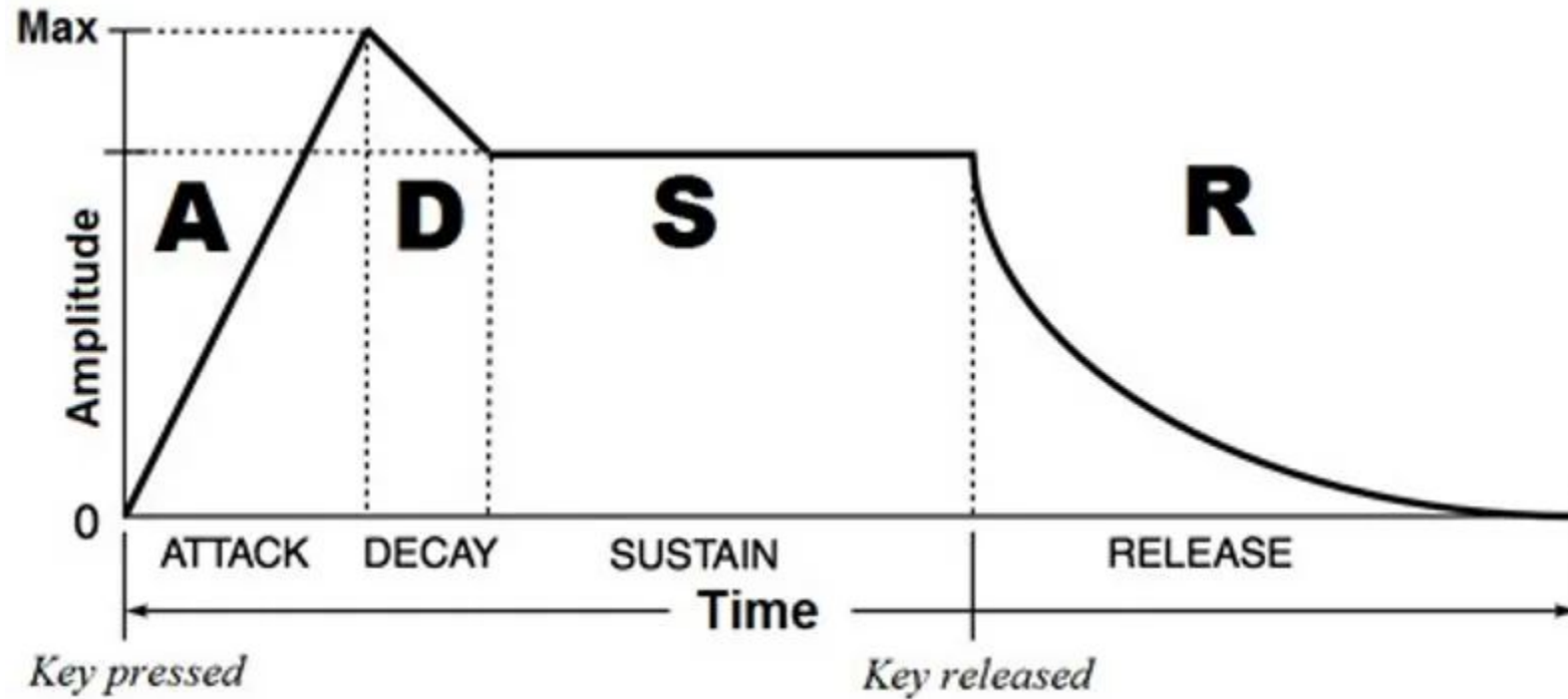
$$F(p) = 2^{\frac{p-69}{12}} \cdot 440$$

1. Sự khác biệt giữa 2 âm thành cùng cường độ, tần số, thời gian,...
2. Được miêu tả bằng các từ ngữ: sáng, tối, ấm...

Features của âm sắc

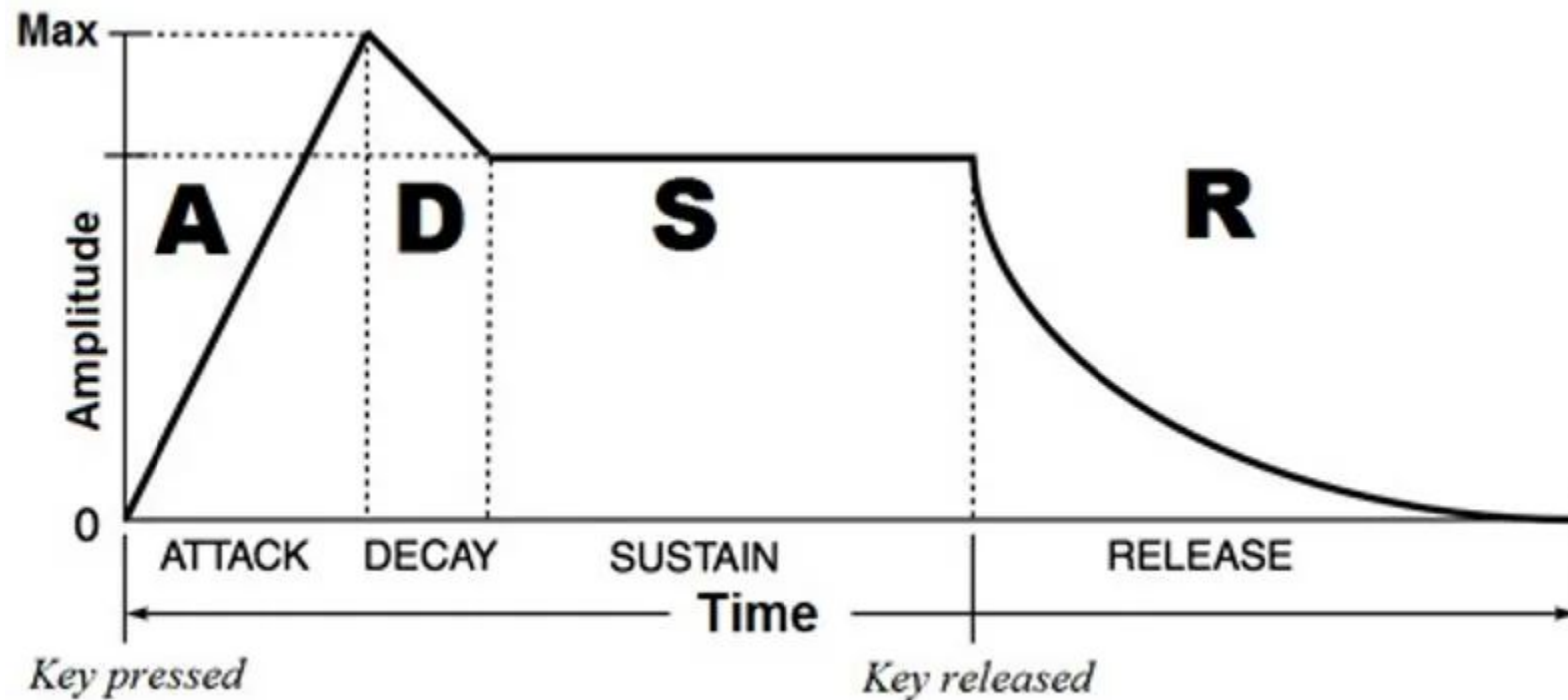
1. Đa chiều
2. Sound Envelope
3. Harmonic content
4. Amplitude/Frequency modulation

1. Form của sóng âm.
2. Sử dụng ADSR Model để đánh giá.

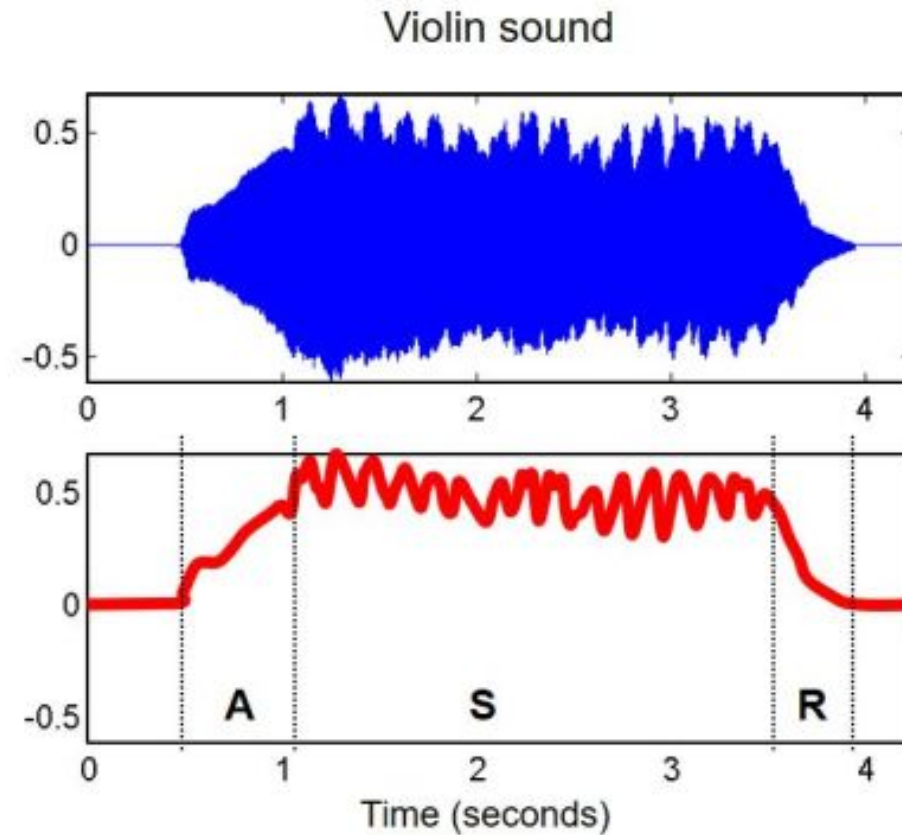
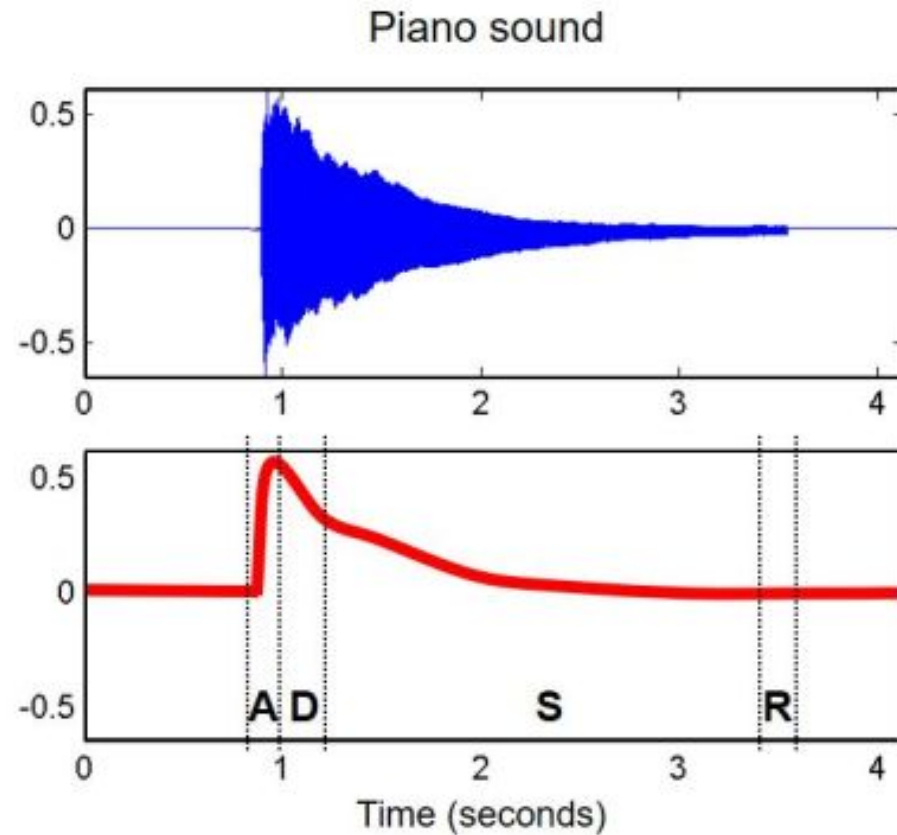


Attack - Decay - Sustain - Release model

1. Attack: Khoảng thời gian âm thanh bắt đầu phát ra đến khi đạt mức cường độ âm lớn nhất. Đây là giai đoạn âm thanh to lên nhanh chóng.
2. Decay: Là khoảng thời gian ngay sau khi đạt đến cường độ cực đại, âm thanh giảm dần xuống một mức cường độ ổn định.
3. Sustain: Là giai đoạn âm thanh duy trì ở mức cường độ ổn định.
4. Release: Khoảng thời gian giảm dần về 0 khi bạn ngừng tác động vào nguồn âm.



Attack - Decay - Sustain - Release model



1. Tổng hợp của nhiều sóng sin khác nhau.
2. Partial: Một sóng sin dùng để mô tả âm thanh.
3. Partial thấp nhất được gọi là Fundamental frequency - Tần số cơ bản.
4. Harmonic partial là một tần số gấp bội của Fundamental frequency.

VD: $f_1 = 440\text{Hz}$, $f_2 = 2 \times 440 = 880\text{Hz}$, $f_3 = 3 \times 440 = 1320\text{Hz}$,...

Tần số cơ bản, Tần số hài bậc 2, Tần số hài bậc 3,...

5. Inharmonic: Tính phi điều hòa.

Harmonic: Âm thanh điều hòa, tần số hài là bội nguyên của tần số cơ bản.

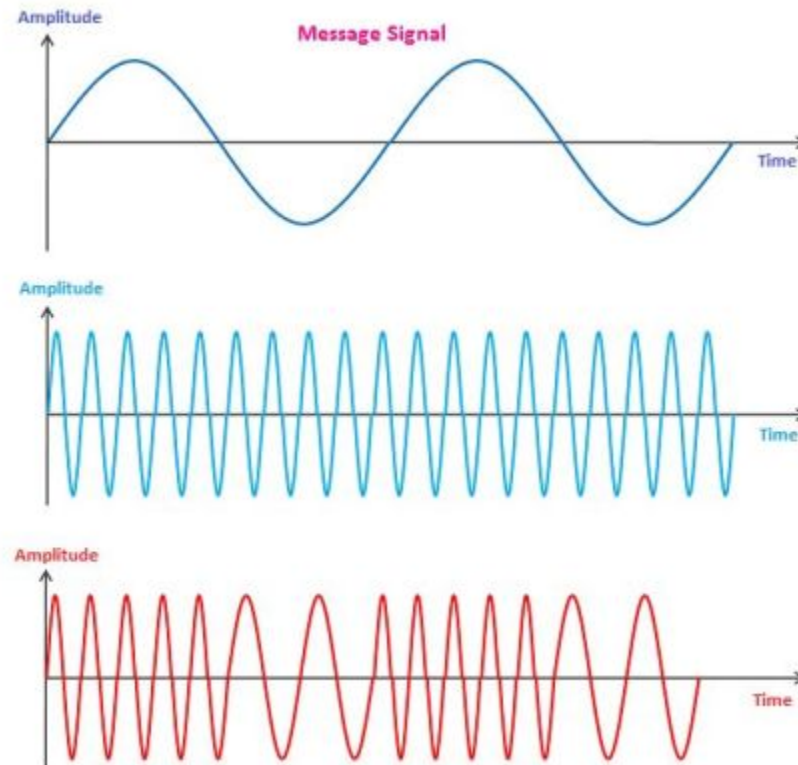
Inharmonic: Âm thanh phi điều hòa, tần số của các thành phần hài lệch so với tần số cơ bản.

VD: tiếng đàn violin, nhạc cụ - Harmonic sound

tiếng xe cộ đi lại - Inharmonic sound

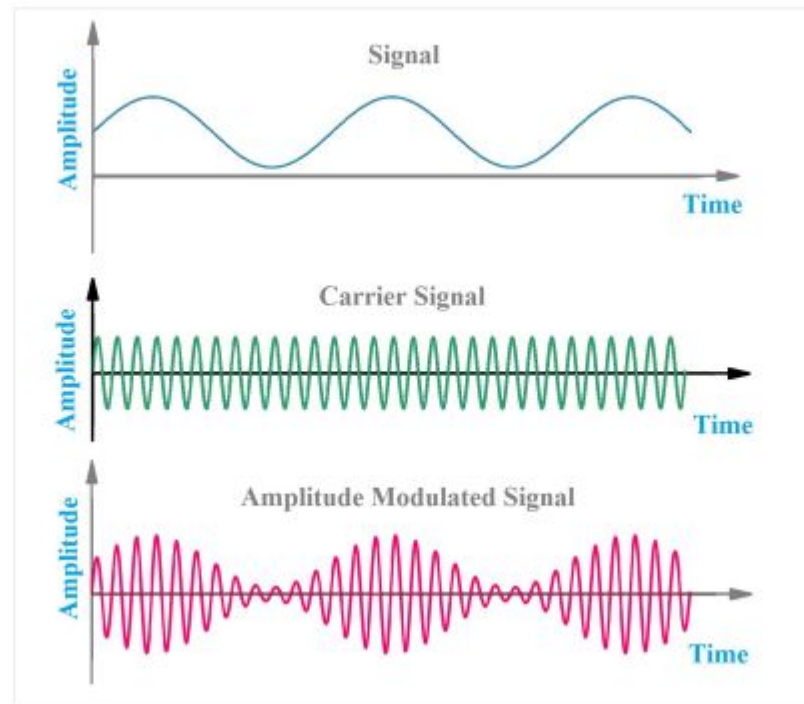
Điều chế tần số - Frequency modulation

1. Còn được gọi là vibrato
2. Sự biến đổi tần số theo chu kỳ
3. Thường được dùng để biểu thị các sắc thái khác nhau trong âm nhạc



Điều chế biên độ - Amplitude modulation

1. Còn được gọi là tremolo
2. Sự biến đổi biên độ theo chu kỳ
3. Thường được dùng để biểu thị các sắc thái khác nhau trong âm nhạc



1. Đa dạng các đặc trưng của âm
2. Sự biến đổi của biên độ (ADSR model)
3. Phân bố năng lượng trên các tần số hài
4. Điều chế tín hiệu (tần số, biên độ)

1. Là 1 sóng
2. Có các đặc trưng về cơ bản: Tần số, cường độ, âm sắc.
3. Chuyển đổi thành các đặc trưng gần với cảm nhận của con người hơn: Cao độ - Pitch, độ to - loudness, âm sắc - timbre.

Xử lý tín hiệu âm thanh cho Machine Learning

1. Biểu diễn âm thanh.
2. Mã hóa tất cả thông tin chúng ta cần để tái tạo âm thanh đó.

Analog signal

1. Giá trị thời gian liên tục.
2. Giá trị biên độ liên tục.

Digital signal

1. Chuỗi các giá trị rời rạc: Được biểu diễn bằng một chuỗi các giá trị rời rạc, không liên tục như analog.
2. Giá trị hữu hạn: Mỗi điểm dữ liệu trong tín hiệu digital chỉ có thể nhận một trong một số giá trị nhất định.

1. Sampling (lấy mẫu)
2. Quantization (lượng tử hóa)

Pulse - code modulation (PCM): Điều chế mã xung

Quá trình PCM gồm 3 bước chính:

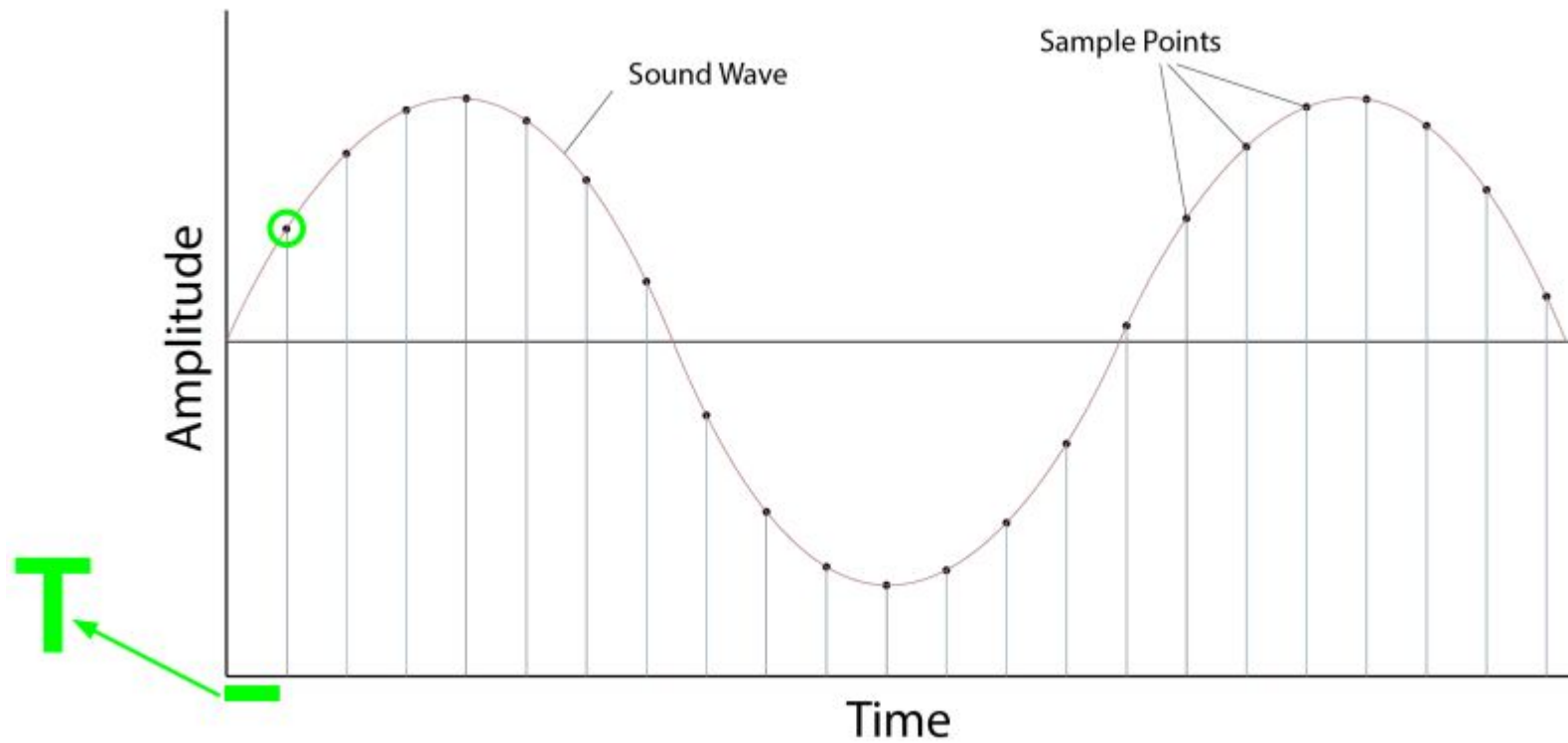
B1: Lấy mẫu Tín hiệu tương tự được lấy mẫu định kỳ để tạo thành một chuỗi các giá trị rời rạc. Tần số lấy mẫu phải đủ cao để đáp ứng định lý Nyquist - Shannon, đảm bảo tín hiệu gốc có thể được tạo lại chính xác từ các mẫu.

B2: Lượng tử hóa: Mỗi giá trị mẫu được làm tròn thành một giá trị gần nhất trong tập hợp các giá trị rời rạc. Quá trình này giới hạn độ chính xác của tín hiệu số, nhưng giúp giảm kích thước dữ liệu.

B3: Mã hóa: mỗi giá trị lượng tử hóa được encoding thành một chuỗi bit.



Sampling period



t_n : thời điểm lấy mẫu thứ n

T : khoảng thời gian giữa 2 lần lấy mẫu liên tiếp

$$t_n = n \cdot T$$

1. Giả sử có một âm thanh có tần số cao nhất là 20kHz. Để tái tạo chính xác tín hiệu này tần số lấy mẫu phải $> 2 \times 20\text{kHz} = 40\text{kHz}$
2. Do liên quan tới một số tiêu chuẩn về truyền hình nên đĩa CD thường sử dụng tần số lấy mẫu 44.1kHz.

Ngưỡng nghe của người: 20Hz - 20kHz

$$s_r = \frac{1}{T}$$

$$f_N = \frac{s_r}{2}$$

Sr: Sampling rate - tần số lấy mẫu

fN: Giới hạn Nyquist

1. Resolution: tổng số lượng bit ghi lại trong 1s.
2. Bit depth: số lượng mức âm lượng có thể biểu diễn.

Mỗi âm thanh trên đĩa CD được mã hóa = 16 bit

16 bit biểu diễn được 2^{16} số (tín hiệu rời rạc) cung cấp độ chính xác tương đối cao.

Dynamic Range

1. Khoảng cách giữa tín hiệu lớn nhất và nhỏ nhất mà 1 hệ thống có thể thu được.



resolution



dynamic range

1. Mối quan hệ giữa cường độ tín hiệu tối đa và lỗi lượng tử hóa.
2. Tương quan với Dynamic range

$$SQNR \approx 6.02 \cdot Q$$

$$SQNR(16) \approx 96dB$$

96dB tức là tín hiệu mạnh hơn 96 lần so với nhiễu

Khi độ sâu bit tăng thêm 1 thì giá trị SQNR tăng thêm xấp xỉ 6.02 dB

Time Domain

1. Amplitude Envelope
2. Root-mean square Energy
3. Zero-crossing rate

Frequency Domain

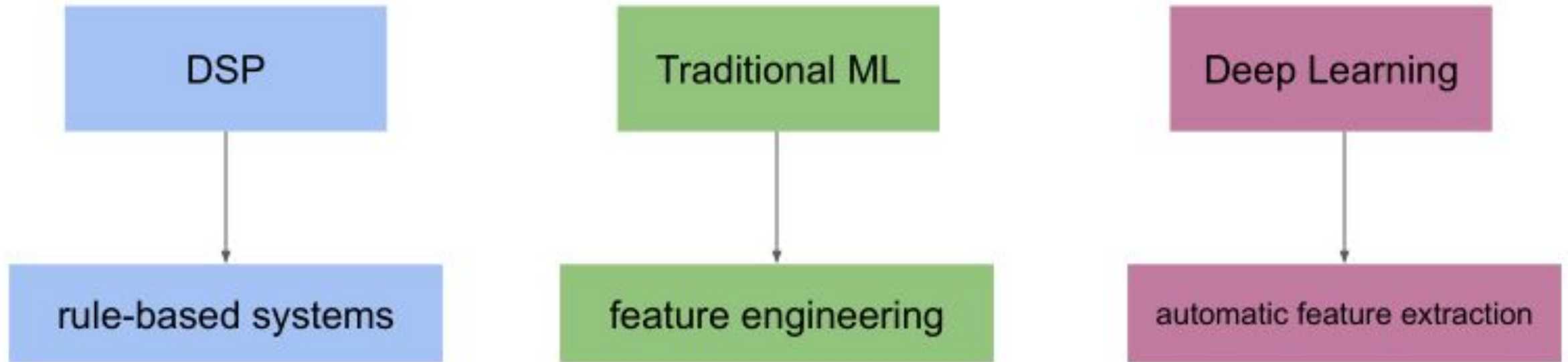
1. Band energy ratio: tỷ lệ năng lượng băng tần
2. Spectral Centroid: trọng tâm phổ
3. Spectral Flux: thông lượng phổ

Time-Frequency Representation

1. Spectrogram
2. Mel-Spectrogram
3. Mel Frequency Cepstral Coefficients



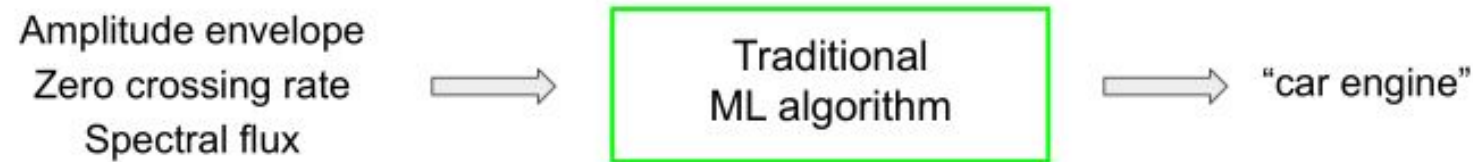
Types of intelligent audio system



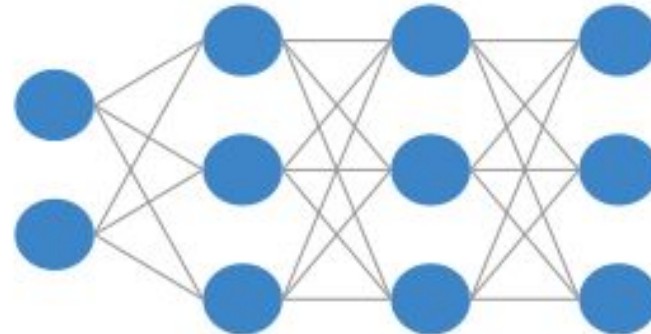
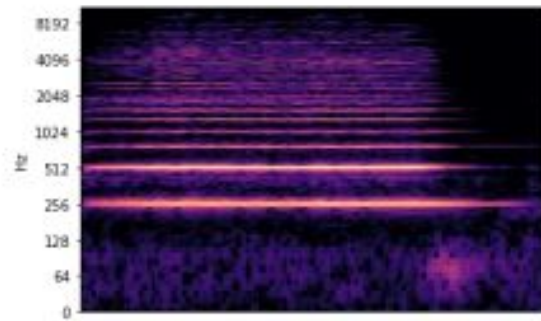
DSP

1. Xử lý tín hiệu số
2. Trích xuất đặc trưng: Xử lý tín hiệu số như biến đổi Fourier, lọc, để trích xuất thủ công từ tín hiệu dựa trên các quy tắc và thuật toán.

1. Sử dụng các thuật toán Machine Learning như Logistic Regression, Support Vector Machine, random forrest,... Yêu cầu Feature Engineer phải xác định và trích xuất thủ công từ dữ liệu

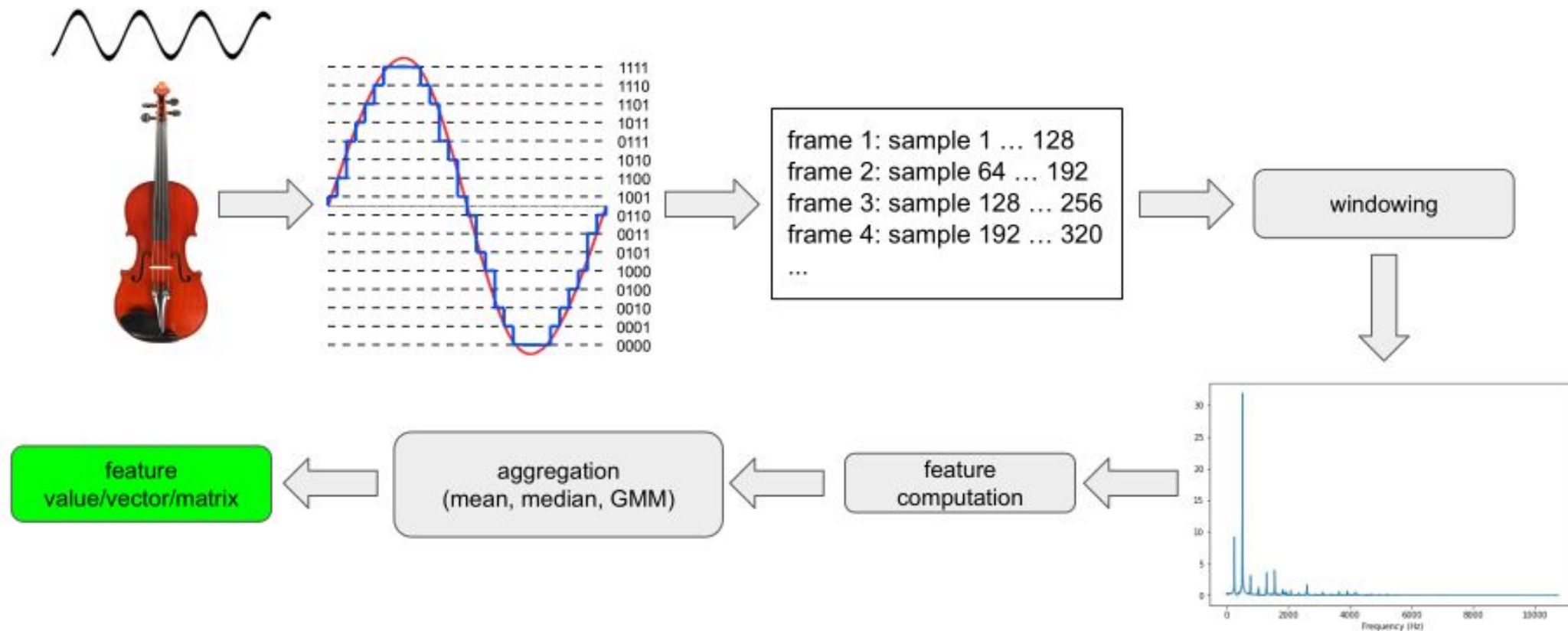


1. Sử dụng các mạng neural network để học trực tiếp từ dữ liệu. Tự động học các đặc trưng từ dữ liệu không cần con người



“car engine”

Time-domain feature pipeline



1. Những đoạn âm thanh mà tai người có thể cảm nhận được.
2. Chia nhỏ các mẫu giúp chúng ta có cái nhìn chi tiết hơn về âm thanh.
3. Phân tích tần số: mẫu càng ngắn, độ chi tiết càng cao
4. Số lượng mẫu là lũy thừa của 2. Số lượng mẫu trong 1 frame thường trong khoảng từ 256 - 8192 samples
5. Công thức tính độ dài một frame

$$d_f = \frac{1}{s_r} \cdot K$$

s_r : tần số lấy mẫu

K : số lượng mẫu trong 1 frame

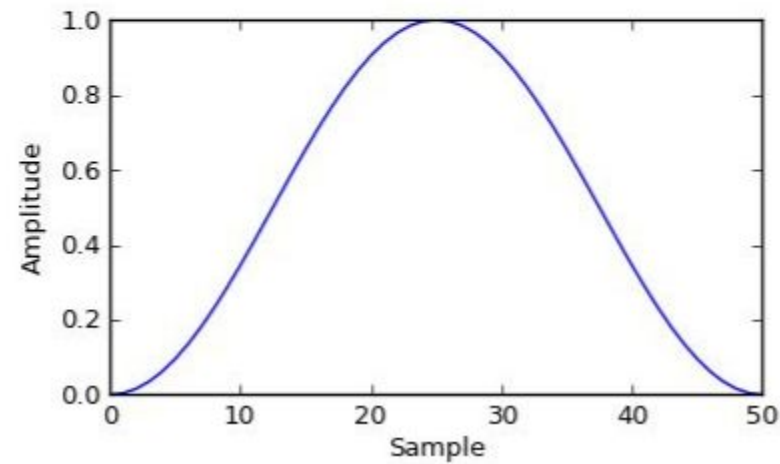
Hiệu ứng rò rỉ phổ (Spectral Leakage)

1. Tín hiệu đã xử lý không có chu kỳ là 1 số nguyên.
2. Điểm cuối của tín hiệu không liên tục.
3. Sự gián đoạn xuất hiện dưới dạng các thành phần tần số cao không có trong tín hiệu gốc.

Windowing

1. Một kỹ thuật xử lý tín hiệu số dùng để giảm thiểu hiệu ứng rò rỉ phổ khi thực hiện biến đổi Fourier trên 1 đoạn tín hiệu
2. Nguyên lý hoạt động: Thay vì phân tích trực tiếp một đoạn tín hiệu, chúng ta sẽ nhân tín hiệu đó với một hàm toán học đặc biệt gọi là window function. Hàm này có giá trị bằng 0 ở hai đầu đoạn tín hiệu và tăng dần về giá trị một ở giữa.
3. Giảm sự gián đoạn: Việc nhân với hàm window giúp giảm bớt sự gián đoạn đột ngột ở 2 đầu đoạn tín hiệu, từ đó làm giảm các thành phần tần số cao giả gây ra hiệu ứng rò rỉ phổ.
4. Hàm window tạo ra một tín hiệu gần giống với một chu kỳ hoàn chỉnh
5. Các hàm window phổ biến: Hamming, Hanning, Blackman

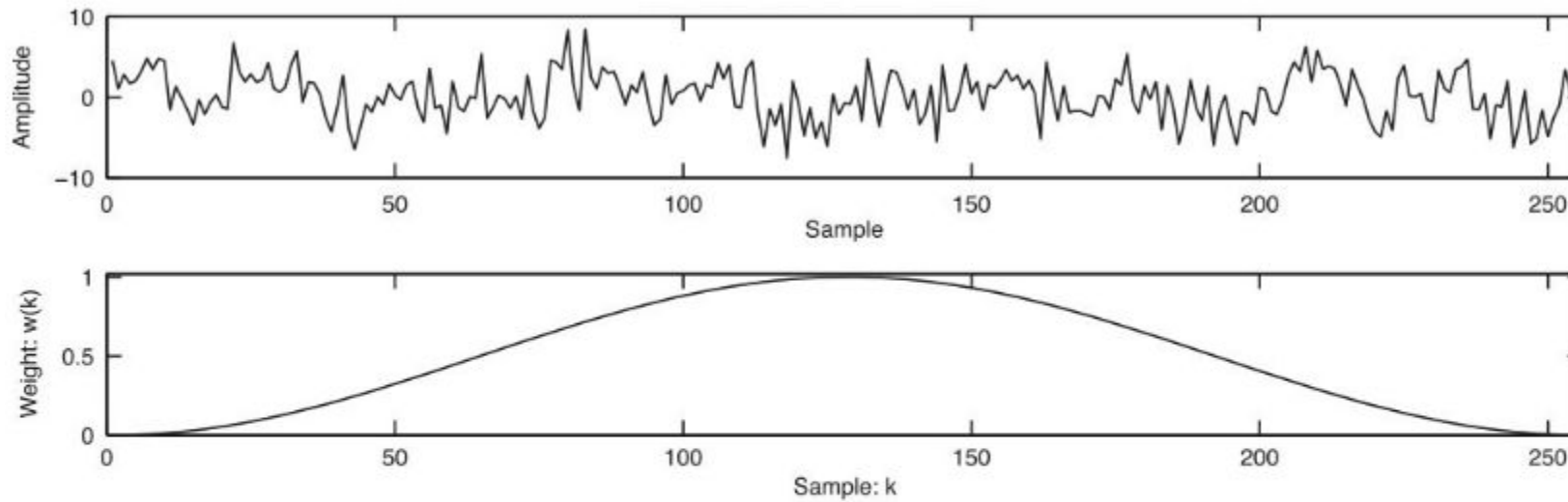
$$w(k) = 0.5 \cdot \left(1 - \cos\left(\frac{2\pi k}{K-1}\right)\right), k = 1 \dots K$$



$w(k)$: Giá trị của hàm tại mẫu thứ k

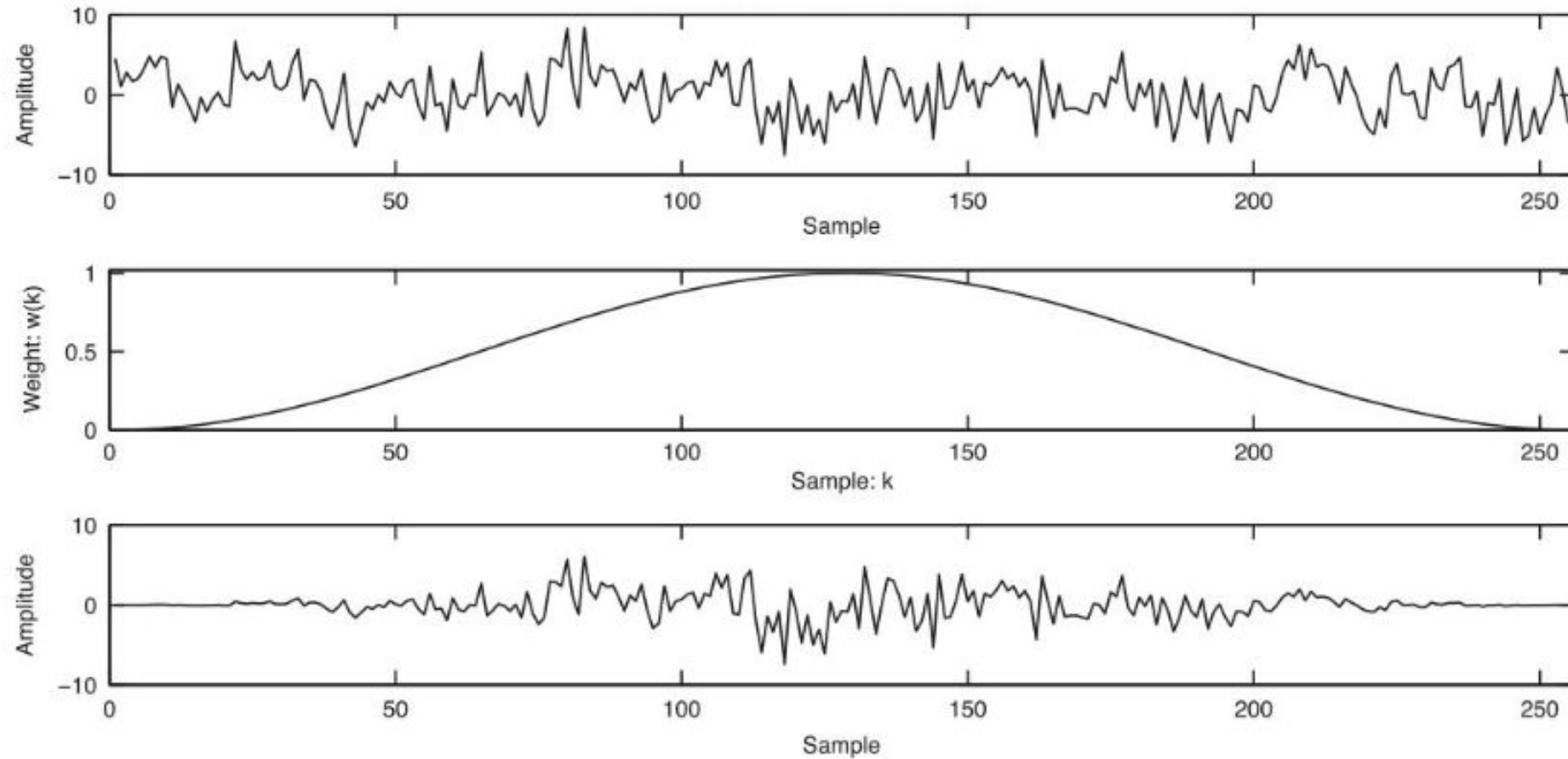
K : tổng số mẫu trong window

k : chỉ số mẫu, chạy từ $1 \dots K$



$$s_w(k) = s(k) \cdot w(k), k = 1 \dots K$$

Windowing

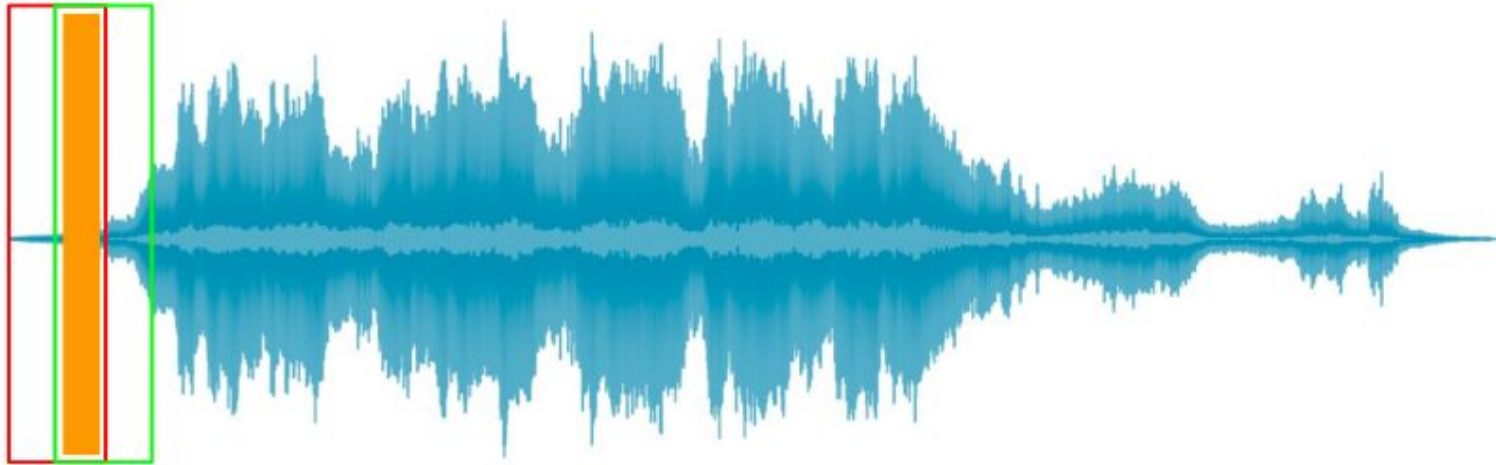


Hiệu ứng biên

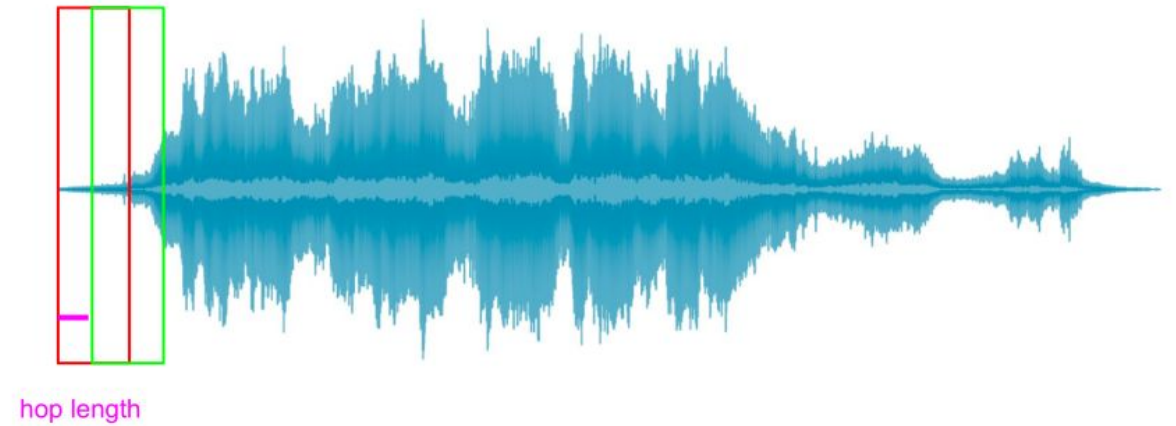
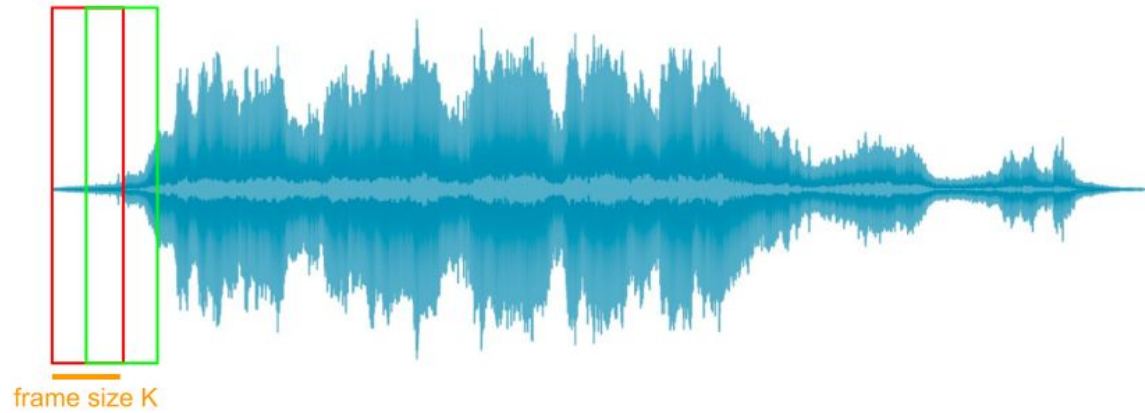
1. Các windowing thường có giá trị giảm dần về 0 ở hai đầu, điều này làm giảm biên độ của các tín hiệu ở phần cuối mỗi frame.

Overlapping frames

1. Là một kỹ thuật phân chia tín hiệu thành các đoạn ngắn (frames), trong đó các frames này có thành phần chồng lên nhau. Khác với non-overlapping frame, nơi mỗi mẫu dữ liệu chỉ thuộc về một frame duy nhất.
2. Overlapping frames giúp tăng độ chi tiết của tín hiệu cho phép ta theo dõi sự thay đổi của tín hiệu theo thời gian một cách chi tiết hơn



Overlapping frame



Frame size K: Độ dài 1 frame

Hop length: Khoảng cách giữa điểm bắt đầu của 2 frame

1. Amplitude envelop (AE)
2. Root-mean-square Energy (RMS)
3. Zero-crossing rate (ZCR)
4. ...

1. Biên độ max của tất cả các mẫu trong một frame
2. Cung cấp một cái nhìn sơ bộ về độ to của một đoạn âm thanh
3. Do chỉ lấy giá trị lớn nhất, Amplitude envelop dễ bị ảnh hưởng bởi các giá trị đột biến hoặc bất thường trong tín hiệu, có thể làm sai lệch thông tin về độ to thực sự.
4. Amplitude được sử dụng trong việc xác định thời điểm bắt đầu của âm thanh (onset) và phân loại dựa trên biên độ

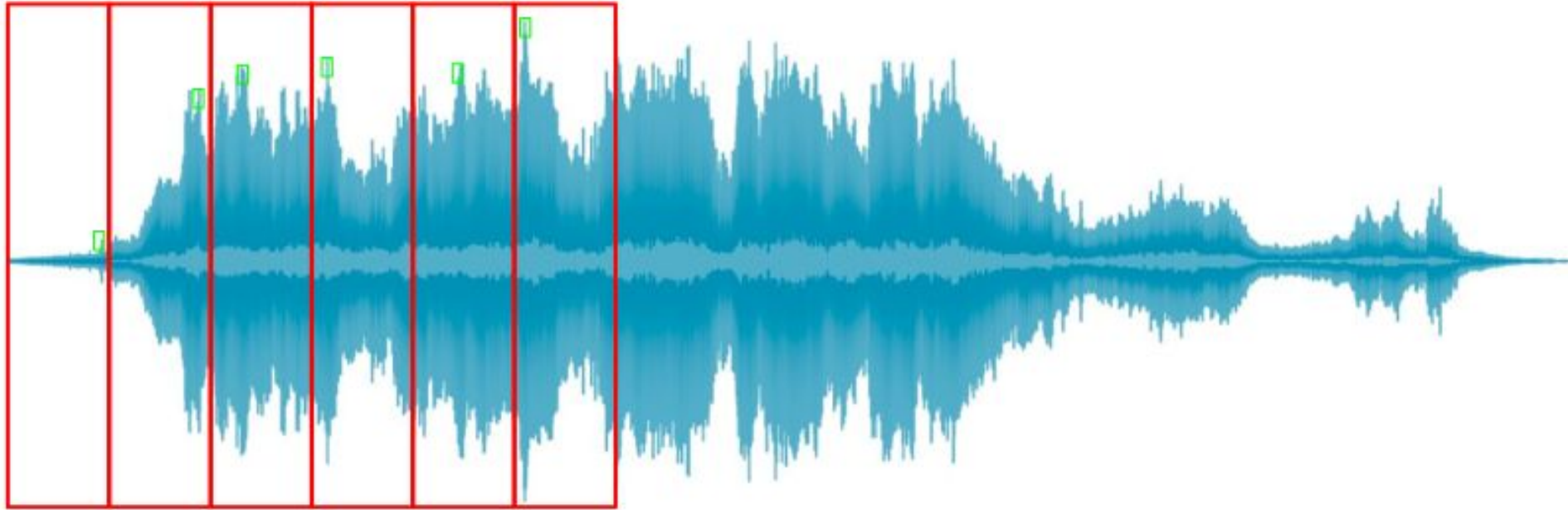
$$AE_t = \max_{k=t \cdot K}^{(t+1) \cdot K - 1} s(k)$$

AE_t : Amplitude envelope tại frame t

K : frame size

$s(k)$: Amplitude của K samples

Amplitude Envelop



1. RMS của tất cả các samples trong một frame
2. RMS cung cấp một chỉ số tốt hơn về mức độ to của tín hiệu so với chỉ sử dụng các giá trị cực đại như Amplitude Envelop
3. RMS ít bị ảnh hưởng bởi các giá trị cực đại bất thường hơn AE
4. RMS được sử dụng trong nhiều ứng dụng xử lý âm thanh, bao gồm phân đoạn âm thanh (audio segmentation) để xác định âm thanh có mức năng lượng khác nhau và phân loại nhạc dựa trên đặc trưng năng lượng của tín hiệu.
5. RMS cung cấp một thước đo cường độ trung bình của tín hiệu trong một khung thời gian. Nó ít bị ảnh hưởng bởi các giá trị cực đại bất thường so với Amplitude Envelope, do đó cung cấp một chỉ số về độ to ổn định hơn.

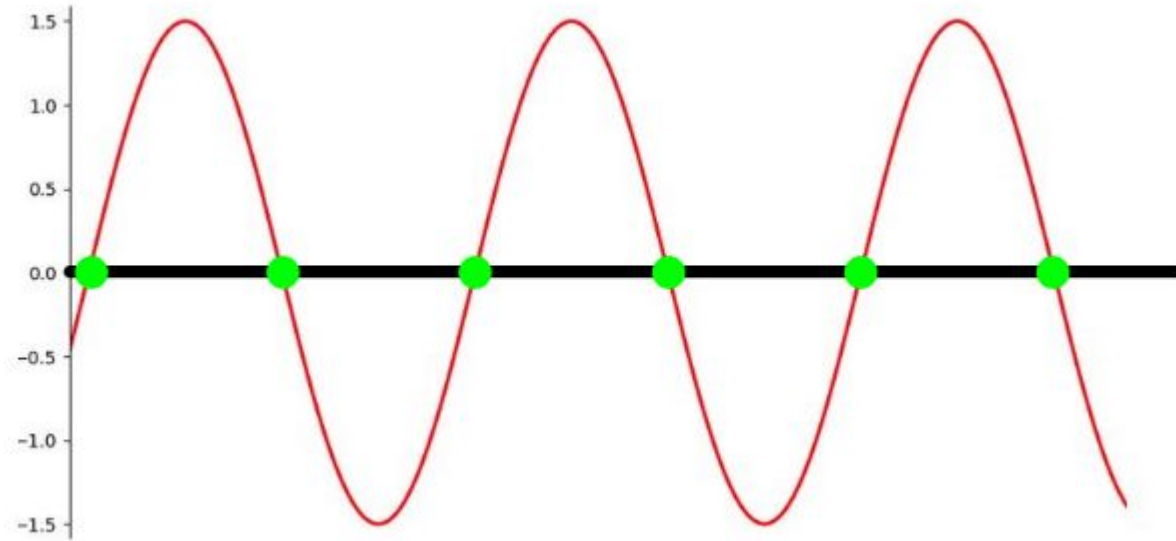
$$RMS_t = \sqrt{\frac{1}{K} \cdot \sum_{k=t \cdot K}^{(t+1) \cdot K - 1} s(k)^2}$$

$s(k)^2$: Năng lượng của sample thứ k

Tổng năng lượng của tất cả các mẫu trong frame t

Giá trị mean của tổng năng lượng

1. Số lần 1 tín hiệu đi qua trục hoành



$$ZCR_t = \frac{1}{2} \cdot \sum_{k=t \cdot K}^{(t+1) \cdot K - 1} | \boxed{\text{sgn}(s(k))} - \text{sgn}(s(k+1)) |$$

Sign function:

- $s(k) > 0 \rightarrow +1$
- $s(k) < 0 \rightarrow -1$
- $s(k) = 0 \rightarrow 0$

1. Phân biệt âm thanh gõ (percussive) và âm thanh có cao độ pitched
2. Ước lượng cao độ đơn âm
3. Phân biệt âm thanh có thanh và không có thanh trong tín hiệu giọng nói

Biến đổi Fourier

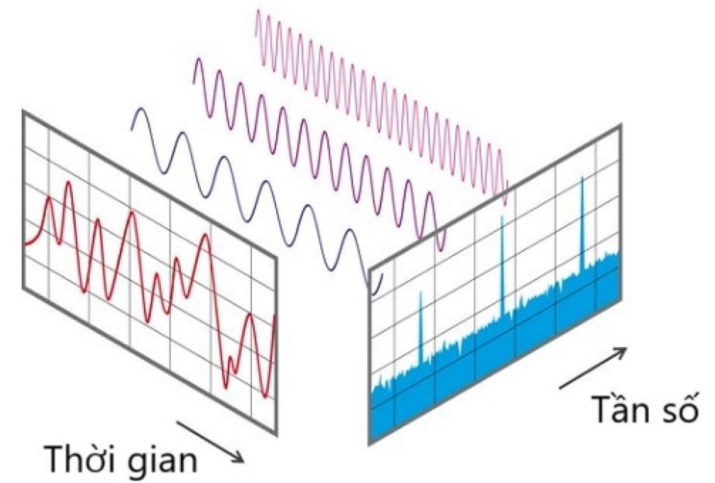
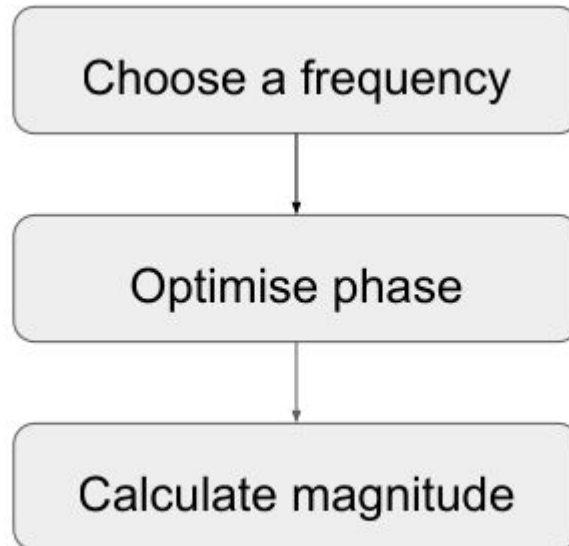
1. Biến đổi Fourier là phân tách một âm thanh phức tạp thành các thành phần tần số của nó.

Tại sao phải biến đổi từ miền thời gian qua miền tần số?

1. Chuyển đổi từ miền thời gian qua miền tần số giúp chúng ta hiểu rõ hơn về bản chất của tín hiệu và thực hiện nhiều phép xử lý phức tạp hơn.
2. Miền tần số cho thấy tín hiệu được cấu tạo từ những thành phần tần số nào và mỗi thành phần đó có cường độ như thế nào. Điều này giúp hiểu rõ bản chất tín hiệu, từ đó có thể phát hiện các nhiễu, các đặc trưng riêng biệt của tín hiệu.

Mở rộng hơn

1. Biến đổi Fourier thực chất là một phép so sánh tín hiệu với các hàm sin có tần số khác nhau
2. Đối với mỗi tần số chúng ta thu được một độ lớn (magnitude) và một pha (phase)
3. Độ lớn cao cho thấy mức độ tương đồng cao giữa tín hiệu và hàm sin
4. Tuy nhiên, hạn chế của biểu diễn miền tần số là không có thông tin về thời gian



$$\varphi_f = \operatorname{argmax}_{\varphi \in [0,1)} \left(\int s(t) \cdot \sin(2\pi \cdot (ft - \varphi)) \cdot dt \right)$$

$$d_f = \max_{\varphi \in [0,1)} \left(\int s(t) \cdot \sin(2\pi \cdot (ft - \varphi)) \cdot dt \right)$$

$$s(t) \cdot \sin(2\pi \cdot (ft - \varphi))$$

Tổng hợp của tín hiệu phức tạp và sóng sin

$$\int s(t) \cdot \sin(2\pi \cdot (ft - \varphi)) \cdot dt$$

Tính toán diện tích

$$\varphi_f = \operatorname{argmax}_{\varphi \in [0,1)}$$

Giá trị của pha để diện tích lớn nhất

$$d_f = \max_{\varphi \in [0,1)}$$

Diện tích lớn nhất

1. Tổng hợp của các sóng sin: Kết hợp nhiều sóng sin với nhau
2. Cân chỉnh theo tỷ lệ độ lớn: Điều chỉnh biên độ (cường độ) của mỗi sóng sin dựa trên độ lớn tương đối của nó trong biến đổi Fourier
3. Sử dụng pha tương đối: Xác định vị trí bắt đầu (pha) tương đối của mỗi sóng sin
4. Tín hiệu gốc và biến đổi Fourier có cùng thông tin: Nguyên tắc cơ bản của biến đổi Fourier cho rằng tín hiệu gốc và biến đổi Fourier tương ứng của nó chứa cùng một lượng thông tin
5. Biến đổi Fourier ngược: Từ miền tần số => miền thời gian

1. Ta có thể mã hóa cả độ lớn và pha trong một số phức. Giả sử ta có một tín hiệu âm thanh liên tục $g(t): \mathbb{R} \rightarrow \mathbb{R}$, biến đổi Fourier phức có thể được mô tả một cách ngắn gọn như sau:

$$\varphi_f = \operatorname{argmax}_{\varphi \in [0,1)} \left(\int s(t) \cdot \sin(2\pi \cdot (ft - \varphi)) \cdot dt \right)$$

$$d_f = \max_{\varphi \in [0,1)} \left(\int s(t) \cdot \sin(2\pi \cdot (ft - \varphi)) \cdot dt \right)$$

$$c = |c| \cdot e^{i\gamma}$$

$$c_f = \frac{d_f}{\sqrt{2}} \cdot e^{-i2\pi\varphi_f}$$

1. Biến đổi Fourier ngược (IFT): Để đưa tín hiệu từ miền tần số về miền thời gian, ta sử dụng biến đổi Fourier ngược

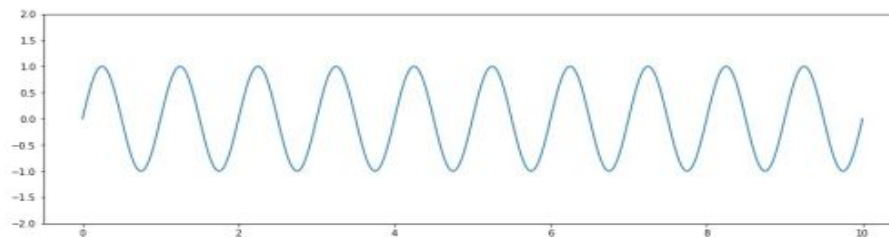
$$g(t) = \int c_f \cdot e^{i2\pi ft} df$$

Trong đó, $e^{i2\pi ft}$ là tone nguyên thủy của tần số f , $c_f \cdot e^{i2\pi ft}$ là tone nguyên thủy có trọng là độ lớn và pha biểu thị trong số phức c_f , $\int c_f \cdot e^{i2\pi ft} df$ là tổng tất cả các sóng sin có trọng.

Biến đổi Fourier ngược

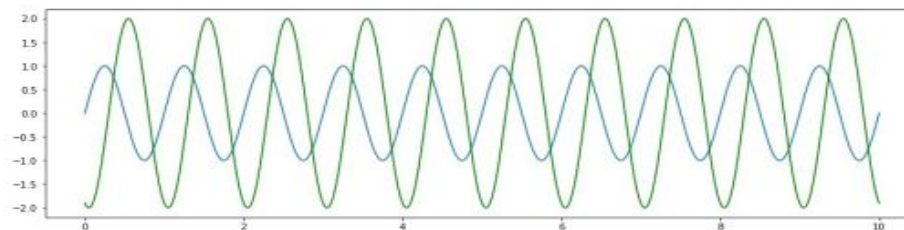
Pure tone of frequency f

$$g(t) = \int c_f \cdot e^{i2\pi ft} df$$



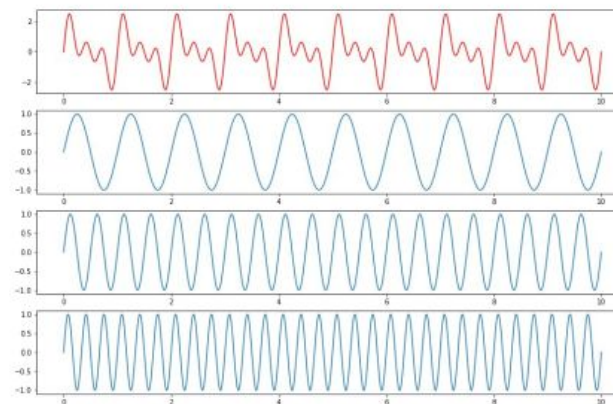
Weight pure tone with magnitude and add phase

$$g(t) = \int c_f \cdot e^{i2\pi ft} df$$



Add up all (weighted) sinusoids

$$g(t) = \int c_f \cdot e^{i2\pi ft} df$$



1. Đầu tiên, ta chuyển tín hiệu từ miền thời gian sang tần số sau đó chuyển đổi ngược trở lại từ miền tần số về thời gian (Fourier nghịch). Ta phân tích một tín hiệu thành các thành phần tần số sau khi thực hiện xử lý: loại bỏ một số thành phần tần số không cần thiết,... ta tổng hợp lại từ các thành phần đó để thu được tín hiệu gốc

$$\hat{g}(f) = \int g(t) \cdot e^{-i2\pi ft} dt$$

$$g(t) = \int c_f \cdot e^{i2\pi ft} df$$

Biến đổi Fourier rời rạc (DFT): Để số hoá một âm thanh liên tục ta thực hiện lấy mẫu, một tín hiệu kỹ thuật số $g(t)$ có thể được xấp xỉ bởi một tín hiệu rời rạc $x(n)$

$$g(t) \mapsto x(n)$$

$$t = nT$$

Với T là quãng thời gian tương ứng biến đổi Fourier $g(f)$ sẽ được xấp xỉ bởi biến đổi Fourier rời rạc $x(f)$ tương ứng.

$$\hat{g}(f) = \int g(t) \cdot e^{-i2\pi ft} dt$$

$$\hat{x}(f) = \sum_n x(n) \cdot e^{-i2\pi fn}$$

Trong 1 vòng (chu kì 2π) (tức thời gian của 1 chu kì là NT , tần số là $\frac{1}{NT} Hz$ hay $\frac{2\pi}{NT} rad/s$), chọn số lượng mẫu N là một số hữu hạn, để thuận tiện cho biến đổi Fourier ngược ta cũng chọn số lượng tần số cơ bản bằng với số lượng mẫu N ,

$$F(k) = \frac{k}{NT} = \frac{kS_r}{N}, k = [0, \dots, N-1],$$

$$\hat{x}\left(\frac{k}{NT}\right) = \sum_{n=0}^{N-1} x(n)e^{-i2\pi\frac{k}{NT}nT}.$$

Khi $k = \frac{N}{2}$ ta có $F\left(\frac{N}{2}\right) = \frac{S_r}{2}$ (tần số Nyquist), lúc này ta có biến đổi Fourier nhanh (FFT), FFT hoạt động khi số lượng mẫu N là một hàm mũ cơ số 2.

Short-time Fourier Transform: Bằng cách trượt hàm cửa sổ dọc theo tín hiệu và tính DFT cho tín hiệu ở vị trí mới ta thu được một ma trận phổ, mỗi hàng là phổ của 1 khung.

$$\hat{x}(k) = \sum_{n=0}^{N-1} x(n) \cdot e^{-i2\pi n \frac{k}{N}}$$

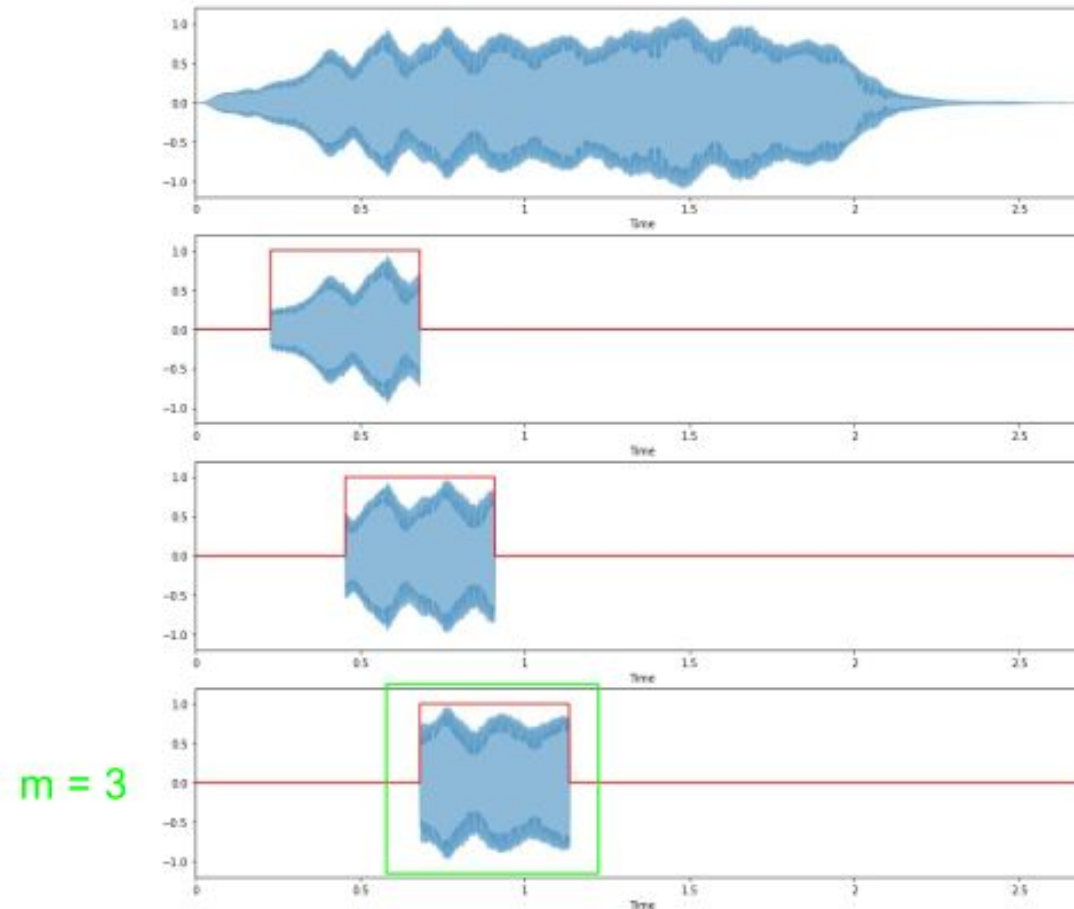
$$S(m, k) = \sum_{n=0}^{N-1} x(n + mH) \cdot w(n) \cdot e^{-i2\pi n \frac{k}{N}}$$

Starting sample of
current frame

m: chỉ số khung

H: độ dài bước nhảy giữa các khung - Hop size

Biến đổi Fourier thời gian ngắn



DFT

- Spectral vector (frequency bins)
- N complex Fourier coefficients - N số lượng các hệ số

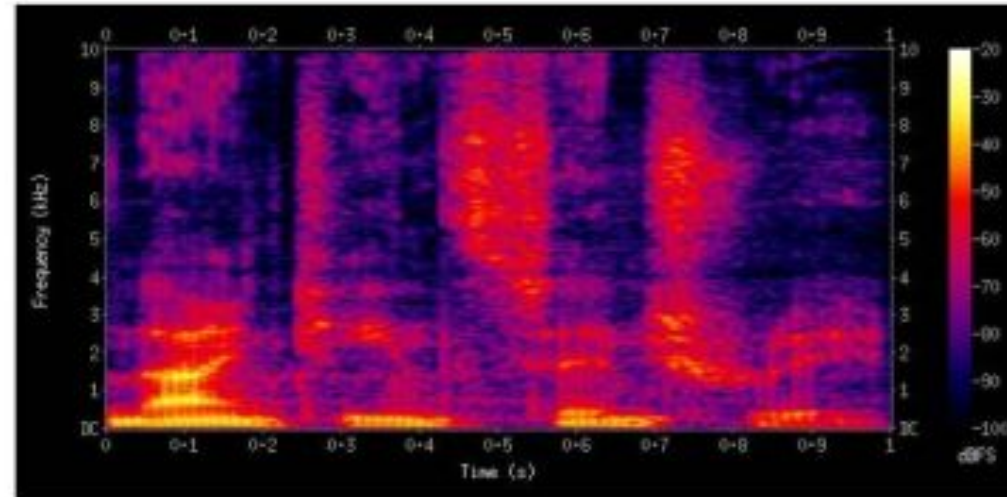
STFT

- Spectral Matrix(frequency bins, frames)
- Complex Fourier coefficients

$$\# \text{ frequency bins} = \frac{\text{framesize}}{2} + 1$$

$$\# \text{ frames} = \frac{\text{samples} - \text{framesize}}{\text{hopsize}} + 1$$

$$Y(k, m) = |S(k, m)|^2$$



$Y(m,k)$: giá trị cường độ tại tần số thứ k của khung thứ m trong Spectrogram

$S(m,k)$: giá trị phức tại tần số k của khung thứ m khi thực hiện STFT

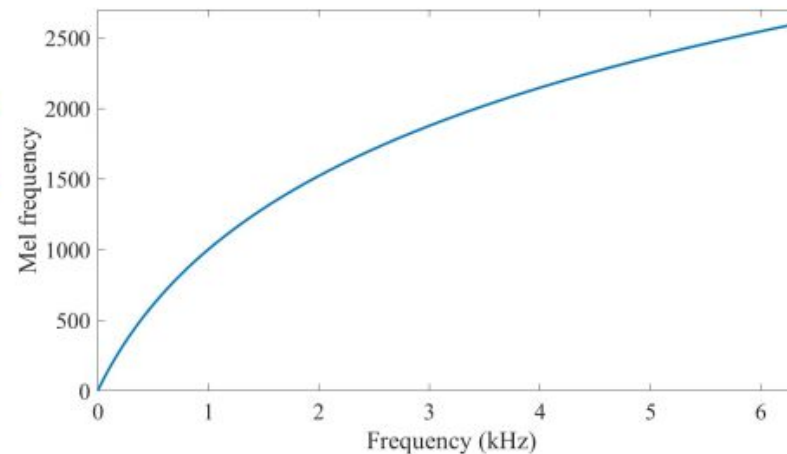
Mel-Frequency Cepstral Coefficients

Mel Scale

- Các nghiên cứu chỉ ra rằng con người không cảm nhận được tần số trên thang đo tuyến tính. Con người có thể dễ dàng nhận ra sự khác biệt giữa âm thanh 100Hz và 200Hz, nhưng lại khó nhận ra sự khác biệt giữa 2000Hz và 2100Hz mặc dù khoảng cách giữa 2 bộ âm thanh là như nhau. Đây là điều khiến Mel Scale trở thành nền tảng cơ bản trong ứng dụng Machine Learning đối với âm thanh vì nó mô phỏng cách cảm nhận âm thanh của con người
- Công thức biến đổi f Hz thành m Mels

$$m = 2595 \cdot \log\left(1 + \frac{f}{500}\right)$$

$$f = 700(10^{m/2595} - 1)$$



Để trích xuất Mel - Spectrograms ta dùng các bước sau:

1. Sử dụng STFT đối với các tín hiệu âm thanh
2. Biến đổi các biên độ thành dBs
3. Biến đổi các tần số thành thang Mel

Tại sao phải dùng thang đo dBs?

1. Thang đo Decibel là một thang đo logarit, giúp nén phạm vi hoạt động của tín hiệu. Điều này có nghĩa là các giá trị lớn và nhỏ sẽ được biểu diễn trong một phạm vi hẹp hơn, giúp dễ dàng so sánh và phân tích.
2. Thang đo Decibel cũng được xây dựng trên ngưỡng nghe của con người từ đó biểu diễn khả năng cảm nhận âm của người tốt hơn

1. Chọn số lượng mel bands
 - Hãy xác định độ chi tiết mà bạn muốn phân tích phổ tần số. Số lượng băng tần Mel càng nhiều thì độ phân giải tần số càng cao, nhưng cũng sẽ tăng tài nguyên tính toán
2. Xây dựng bộ lọc tam giác Mel
 - Tạo ra các bộ lọc có trọng số khác nhau, tập trung vào các vùng tần số quan trọng theo cách mà tai người cảm nhận
3. Áp dụng bộ lọc Mel lên Spectrogram
 - Lọc và tổng hợp năng lượng trong mỗi băng tần Mel để tạo ra Mel Spectrogram, giúp biểu diễn tần số theo cách phù hợp hơn với thính giác con người

Xây dựng bộ lọc Mel tam giác

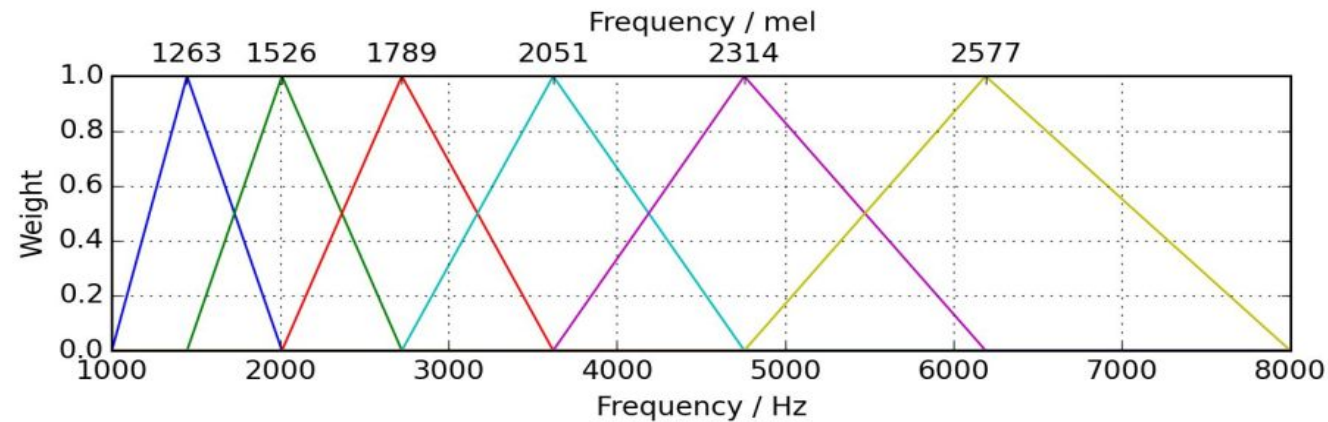
1. Biến đổi tần số thấp nhất và cao nhất của tín hiệu thành Mel.
2. Tạo các dải Mel cách đều 2 điểm



3. Biến đổi lại các điểm đó về thang Hz

$$f = 700(10^{m/2595} - 1)$$

4. Làm tròn đến gần tần (frequency bins) gần nhất
5. Tạo triangular filters



$$M = (\# \text{ bands, framesize} / 2 + 1)$$

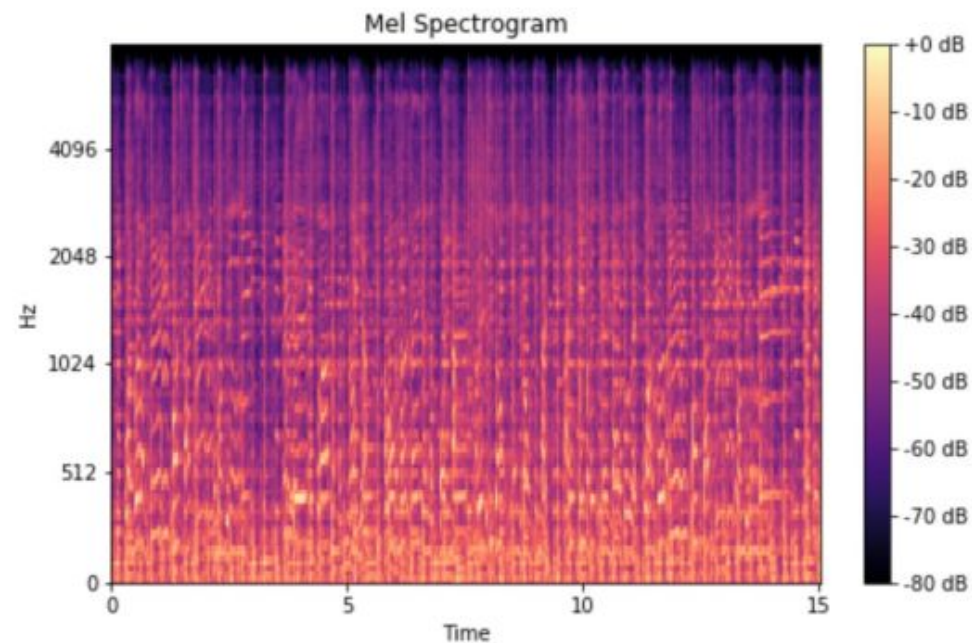
Sau bước trên ta có ma trận (kích thước Mel filter)

$M = (\text{Mel bands, frequency bins})$

$Y = (\text{frequency bins, frames})$

$MY = (\text{Mel bands, frames})$

$\text{frequency bins} = \text{frame size}/2 + 1$



Cepstrum là gì?

Cepstrum là một khái niệm trong xử lý tín hiệu, được sử dụng chủ yếu để phân tích tín hiệu âm thanh, đặc biệt là trong nhận dạng giọng nói. Cepstrum được định nghĩa là biến đổi Fourier ngược của logarit biên độ phổ của một tín hiệu. Thuật ngữ "cepstrum" thực chất là từ "spectrum" đảo chữ.

Cepstrum
↓
Spectrum

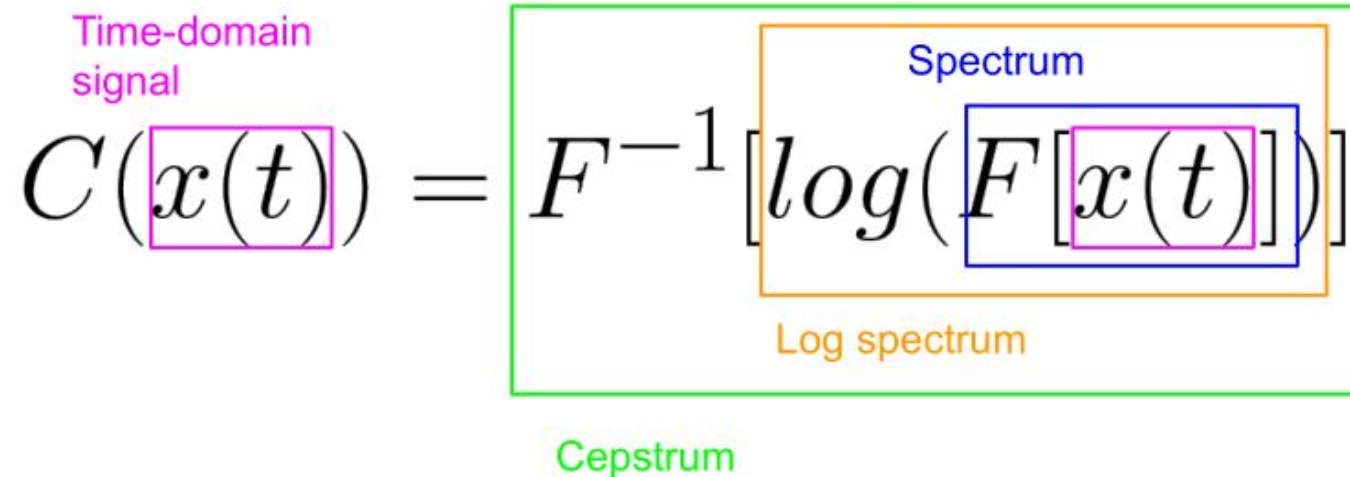
Quefrency
↓
Frequency

Liftering
↓
Filtering

Rhamonic
↓
Harmonic

1. Được phát triển trong quá trình nghiên cứu về tiếng vang trong tín hiệu địa chấn (1960s)
2. Đặc trưng âm thanh được lựa chọn cho nhận dạng giọng nói/nhận dạng người (1970s).
3. Xử lý âm nhạc (2000s).

1. Đầu tiên, tín hiệu thời gian $x(t)$ được chuyển đổi sang miền tần số bằng biến đổi Fourier $F[x(t)]$. Điều này cho ta biết thành phần tần số của tín hiệu
2. Lấy logarit: Kết quả của biến đổi Fourier được lấy logarit tự nhiên. Việc này giúp nén các giá trị lớn và làm nổi bật các cấu trúc tương đối hơn trong phổ tần số.
3. Biến đổi Fourier ngược: Cuối cùng kết quả từ bước 2 được biến đổi ngược lại miền thời gian bằng biến đổi Fourier ngược F^{-1} . Tín hiệu thu được ở bước này gọi là cepstrum

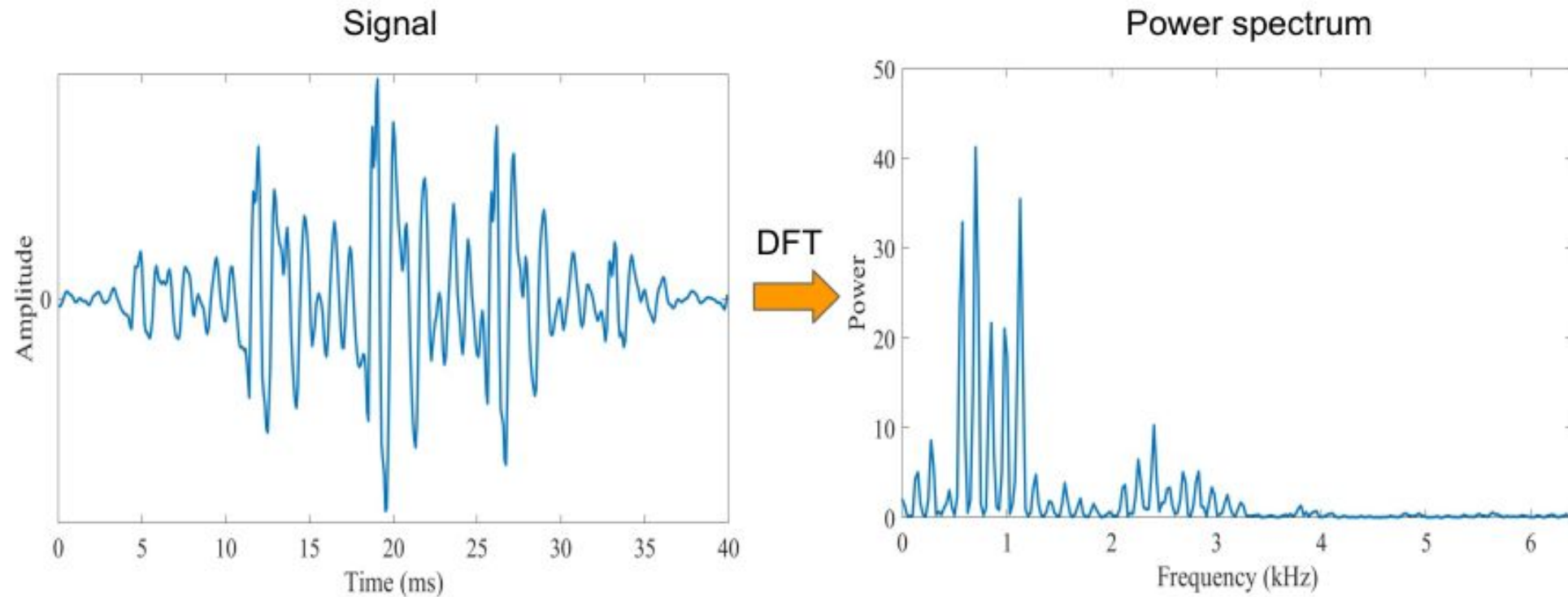


The diagram illustrates the mathematical process of calculating the cepstrum. It features the equation $C(x(t)) = F^{-1}[\log(F[x(t)])]$. The components are color-coded and boxed to show their relationship:

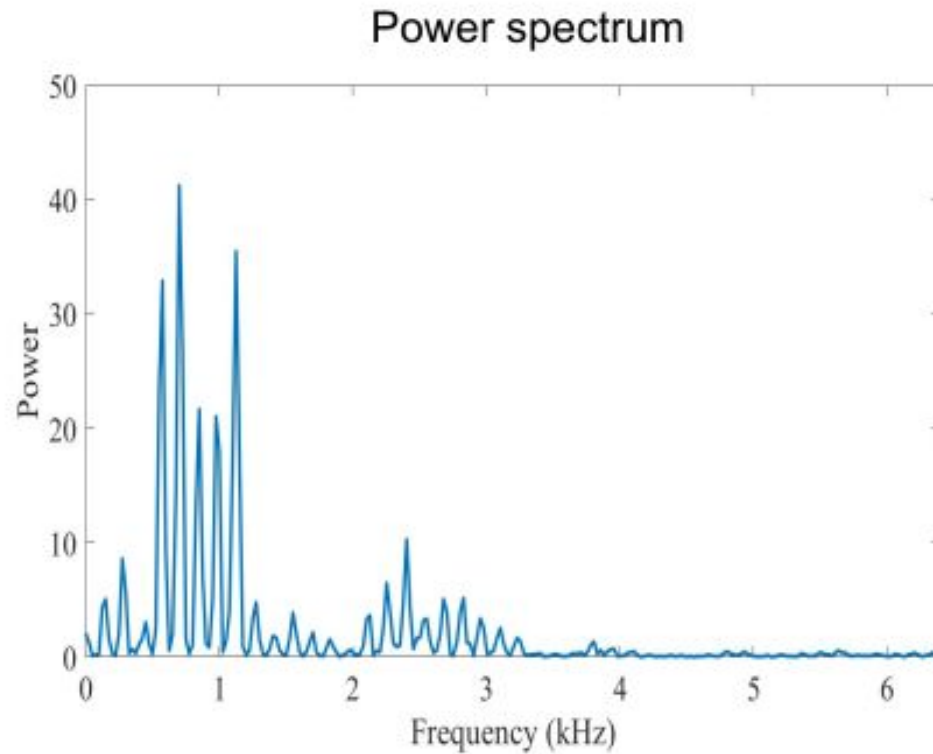
- Time-domain signal** (purple text) points to the $x(t)$ term, which is enclosed in a purple box.
- Spectrum** (blue text) points to the $F[x(t)]$ term, which is enclosed in a blue box.
- Log spectrum** (orange text) points to the $\log(F[x(t)])$ term, which is enclosed in an orange box.
- Cepstrum** (green text) points to the entire expression $C(x(t))$, which is enclosed in a green box.

$$C(x(t)) = F^{-1}[\log(F[x(t)])]$$

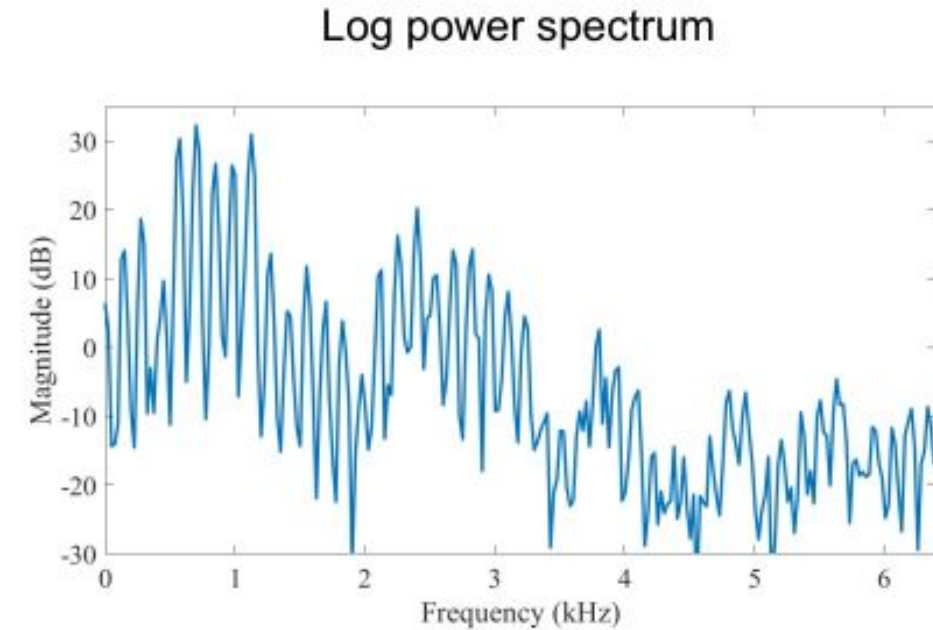
Visualize Cepstrum



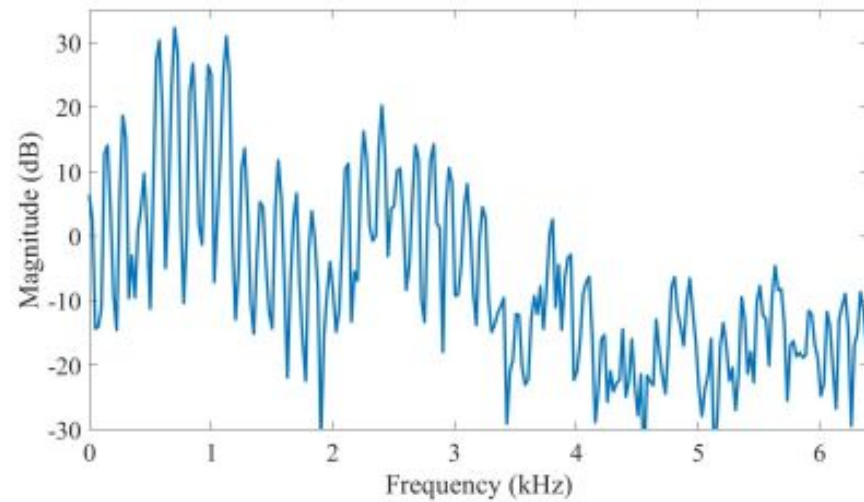
Visualize Cepstrum



log
→



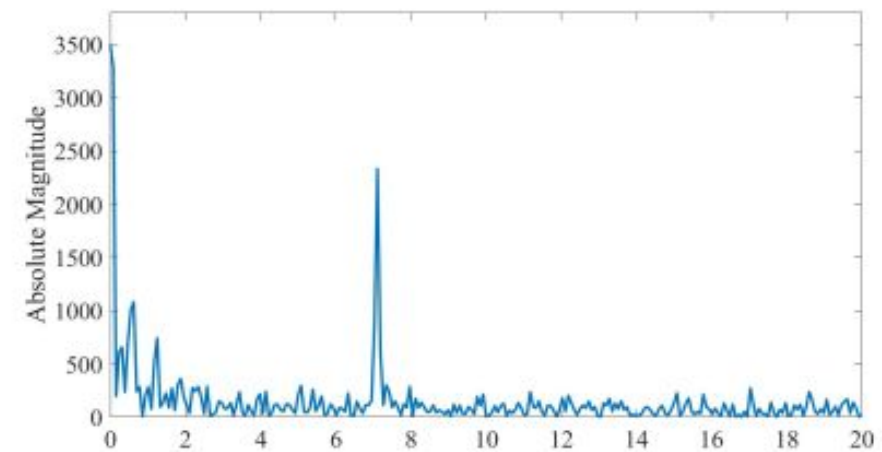
Log power spectrum



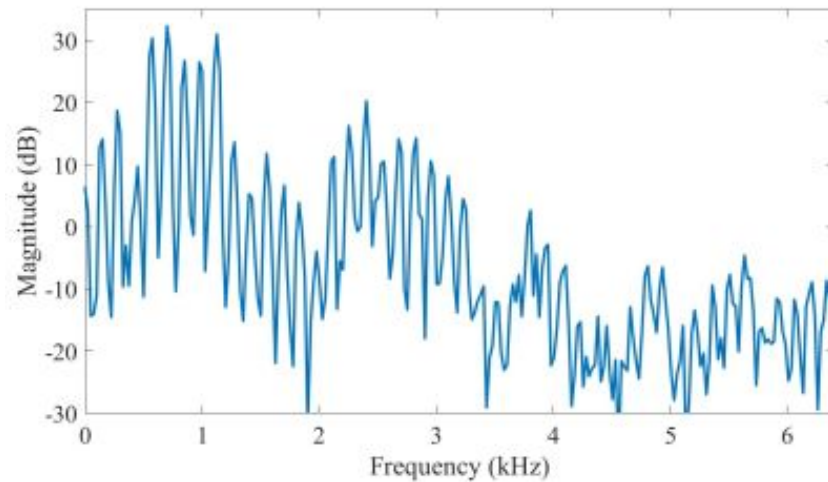
IDFT



Cepstrum



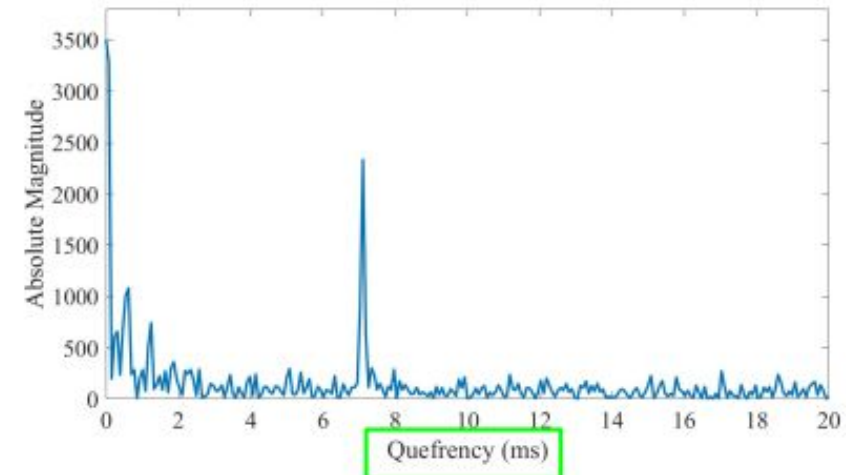
Log power spectrum

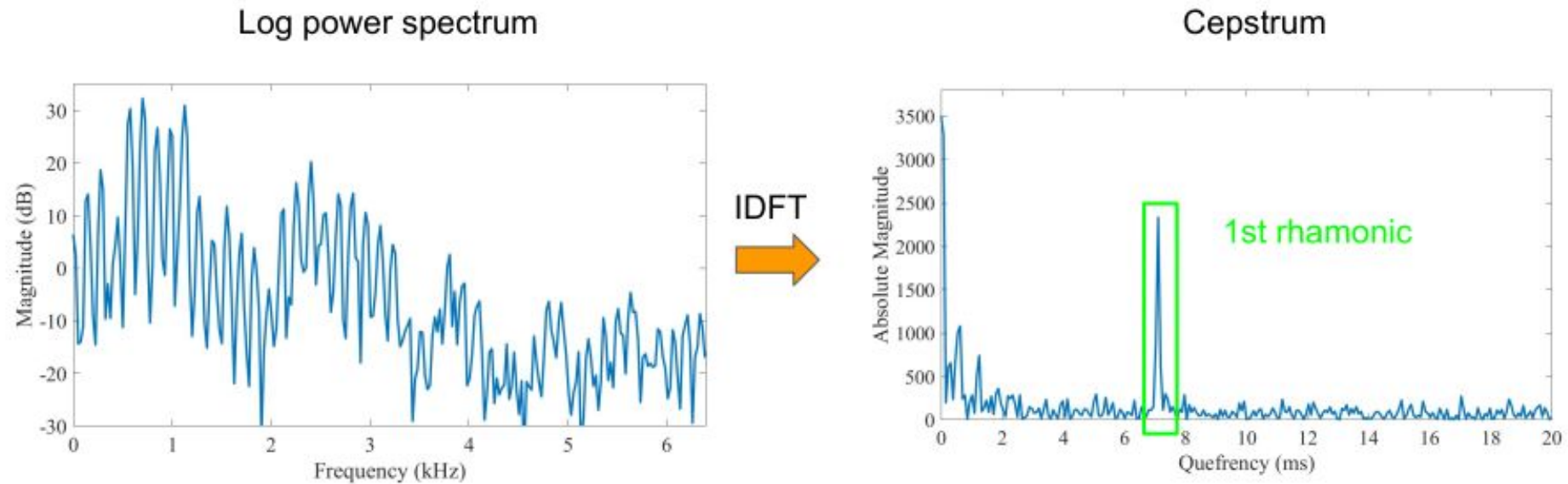


IDFT



Cepstrum





Tổng kết, các bước của phương pháp trích xuất đặc trưng tín hiệu âm thanh MFCCs như sau:

Dạng sóng \rightarrow STFT \rightarrow Log(độ lớn quang phổ) \rightarrow thang Mel \rightarrow biến đổi Cosin rời rạc \rightarrow MFCCs.

Sử dụng biến đổi cosin rời rạc (một dạng đơn giản của biến đổi Fourier) cho ta các hệ số là các giá trị thực, do đó sẽ giảm số chiều của phổ quang biểu diễn. Thông thường ta sẽ chọn 13 hệ số đầu tiên trong các hệ số của MFCCs (chứa hầu hết các thông tin như: các đỉnh biên độ, bao quang phổ...), các hệ số tiếp theo là đạo hàm và đạo hàm cấp 2 theo thời gian của MFCCs, như vậy thông thường có 39 hệ số MFCCs trong một frame.

Tại sao sử dụng MFCC thay vì Mel Spectrogram?

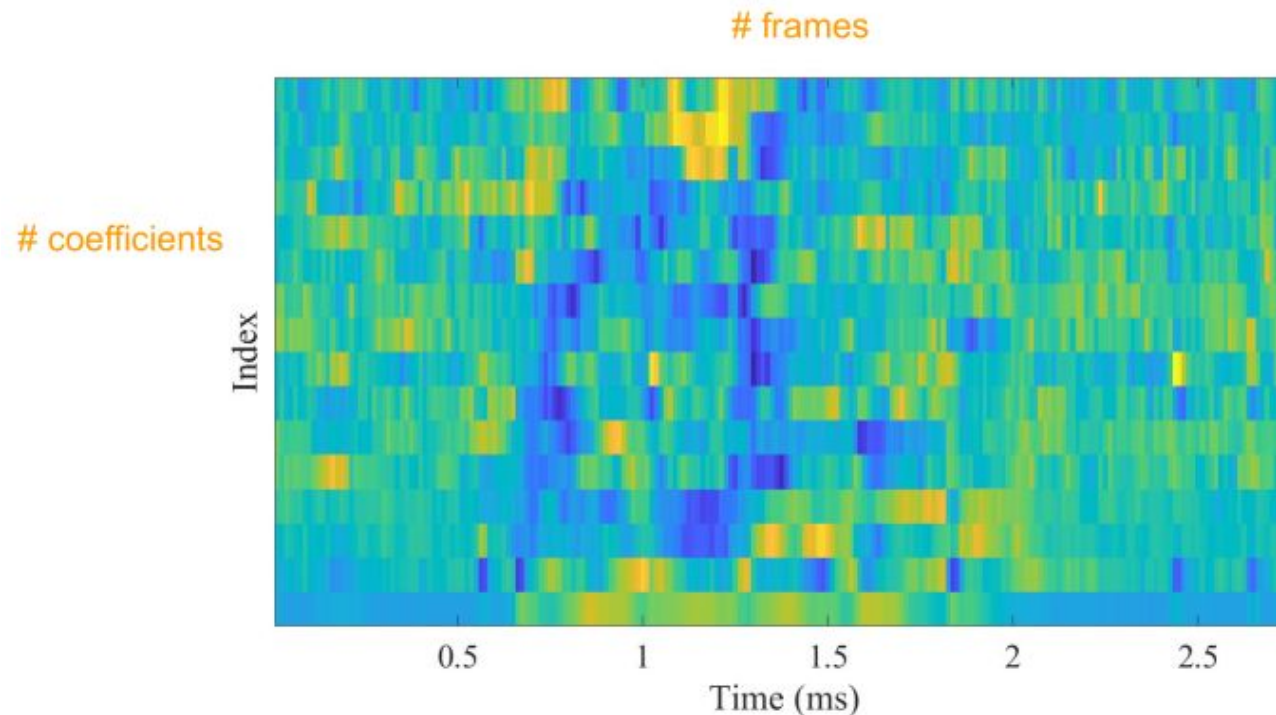
Bắt chước cách con người nghe âm thanh: MFCC loại bỏ các thông tin không cần thiết như cường độ, tập trung vào nội dung.

Giảm kích thước dữ liệu: Chỉ giữ lại vài hệ số quan trọng, giảm tải tính toán.

Phù hợp nhận diện giọng nói: MFCC tách biệt tốt nội dung âm vị, Mel Spectrogram phù hợp hơn cho âm nhạc.

Các hệ số MFCC giữ lại:

1. Nội dung phổ tần số quan trọng:
 - Các hệ số bậc thấp trong MFCC mô tả dạng tổng quát của phổ âm thanh, phản ánh cấu trúc âm học chính.
 - Chúng chứa thông tin quan trọng về giọng nói, như âm vị (phoneme) và các đặc điểm phát âm.
2. Loại bỏ nhiễu và thông tin không liên quan: Các hệ số bậc cao hơn chứa chi tiết nhỏ và thường là nhiễu hoặc đặc điểm không quan trọng (như dao động nhanh ở phổ). MFCC thường chỉ giữ lại 12 hệ số đầu.
3. Hệ số năng lượng (Energy): Ngoài các hệ số cepstral, MFCC thường giữ lại hệ số năng lượng tổng của khung âm để bổ sung thông tin về cường độ tín hiệu.



Các thành phần chính trong hình ảnh spectrogram:

- **Trục ngang (Time):** Thể hiện trục thời gian, cho biết sự thay đổi của tần số theo thời gian. Trong hình ảnh này, đơn vị thời gian là mili giây (ms).
- **Trục dọc (Coefficient):** đại diện cho một hệ số MFCC.
- **Màu sắc:** Mỗi điểm trên spectrogram đại diện cho cường độ của một thành phần tần số tại một thời điểm nhất định. Màu sắc thường được sử dụng để thể hiện cường độ này, với màu sáng hơn đại diện cho cường độ cao hơn.

XIN CẢM ƠN



Từ 13 đặc trưng ban đầu, ta tính delta để biểu diễn tốc độ thay đổi của âm giữa các frame theo công thức:

$$\Delta C(t) = \frac{\sum_{n=1}^N n(C_{t+n} - C_{t-n})}{2 \sum_{n=1}^N n^2}$$

Thông thường N được chọn là 2, chính là số frame trước và sau frame 't' hiện tại. C_t là véc tơ đặc trưng của frame hiện tại (bao gồm cả năng lượng). Tiếp theo tính thêm $\Delta\Delta C$ thể hiện gia tốc thay đổi của âm giữa các frame tín hiệu, công thức tính giống như trên chỉ có thay C_t bằng ΔC_t .

Cuối cùng ta được một véc tơ đặc trưng MFCC có 39 đặc trưng bao gồm: 12 hệ số cepstral, 12 delta của hệ số cepstral, 12 delta-delta của hệ số cepstral, 1 năng lượng của khung tín hiệu, 1 delta năng lượng và 1 delta-delta năng lượng.