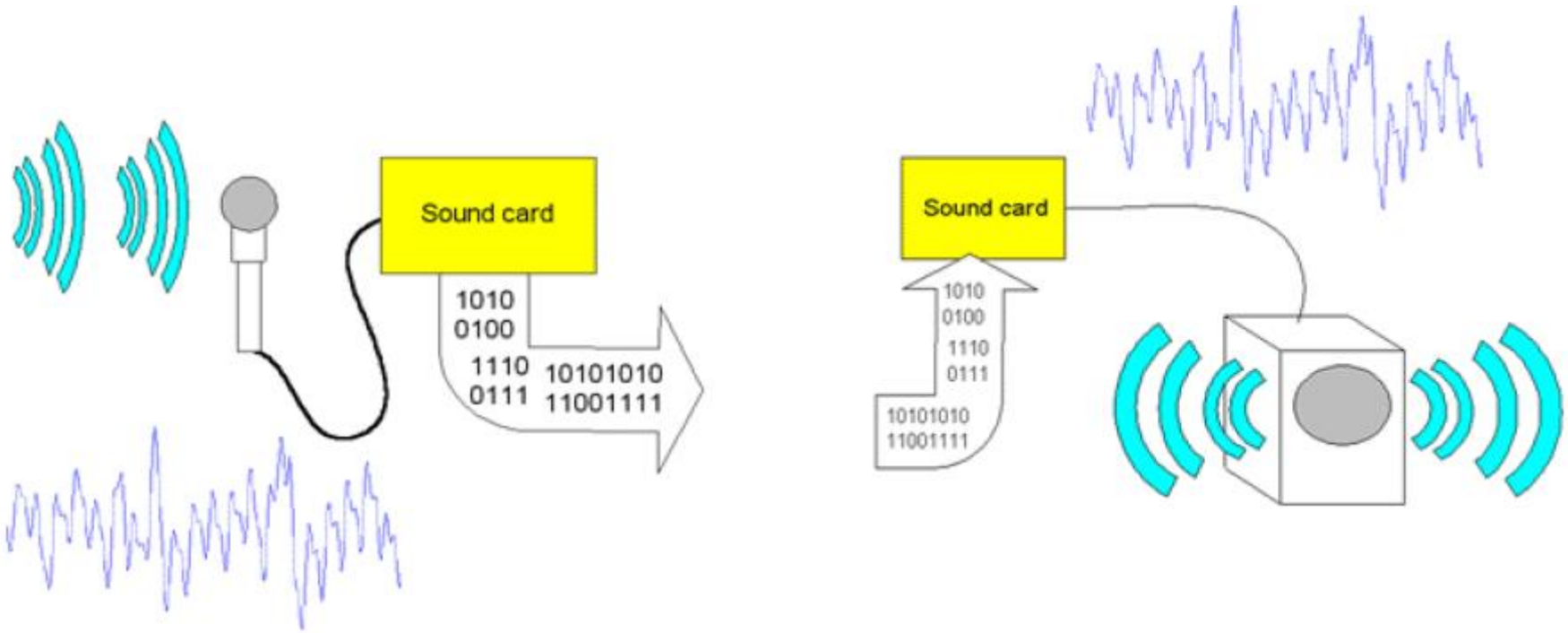
Các kỹ thuật chỉ số hóa và truy vấn dữ liệu âm thanh

Nguyễn Đình Hóa
hoand@ptit.edu.vn 0942807711

Dữ liệu âm thanh

- Âm thanh được lưu trữ theo chuỗi các mẫu, thường dưới dạng nén



Chỉ số hóa và truy vấn âm thanh

- Một số tình huống cần giải quyết:
  - Nếu ta có một tệp âm thanh về buổi biểu diễn của một diễn viên nào đó, làm thế nào để ta biết được khi nào thì diễn viên đó hát, khi nào thì diễn viên đó nói chuyện với khán giả?
  - Nếu ta có một tệp thu âm của một hội nghị khoa học, làm thế nào để ta biết chính xác khi nào thì mọi người bàn đến một vấn đề cụ thể XYZ nào đó?
  - Nếu ta có nhiều tệp âm nhạc, làm thế nào tìm được tệp nhạc cần tìm nếu như ta không nhớ tên bài hát mà chỉ nhớ điệu nhạc của nó?
  - Nếu ta cần tìm các cảnh kinh dị trong một bộ phim, làm sao để ta nhanh chóng tìm thấy chúng dựa trên âm thanh trong phim?

Chỉ số hóa và truy vấn âm thanh

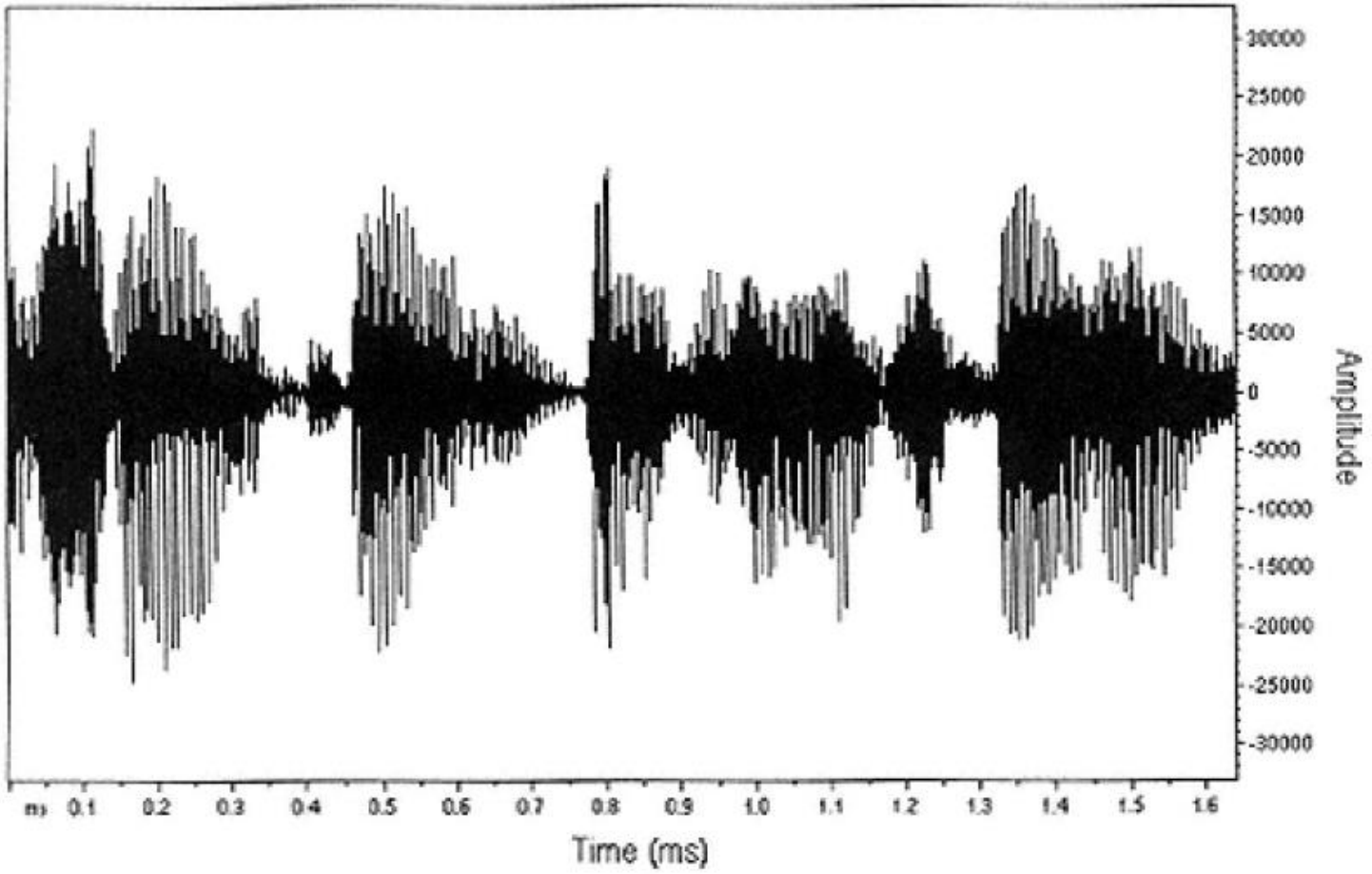
- Các phương pháp chính
  - Âm thanh được phân loại thành một trong 3 dạng chính: lời thoại, âm nhạc, nhiễu
  - Mỗi loại âm thanh được xử lý, lưu trữ và truy vấn theo các cách khác nhau
    - VD: lời thoại được phân tích / nhận diện thành các từ, và các từ này được lưu trữ.
  - Các loại âm thanh khác nhau có các ý nghĩa khác nhau theo các ứng dụng cụ thể
  - Việc phân loại âm thanh giúp cho việc truy vấn dữ liệu hiệu quả hơn.
  - Âm thanh được lưu trữ và truy vấn dựa trên sự so sánh tương tự trên các thuộc tính của chúng.

Các thuộc tính chính của âm thanh

- Các thuộc tính miền thời gian
  - Năng lượng trung bình
  - Tốc độ đổi dấu của tín hiệu
  - Phần trăm của khoảng lặng
- Các thuộc tính miền tần số
  - Phổ âm thanh
  - Băng thông
  - Phân bố năng lượng âm thanh
  - Độ điều hòa âm
  - Độ cao thấp của âm thanh
- Ảnh phổ

Các thuộc tính miền thời gian

- Biến thiên của cường độ âm thanh theo thời gian



Các thuộc tính miền thời gian

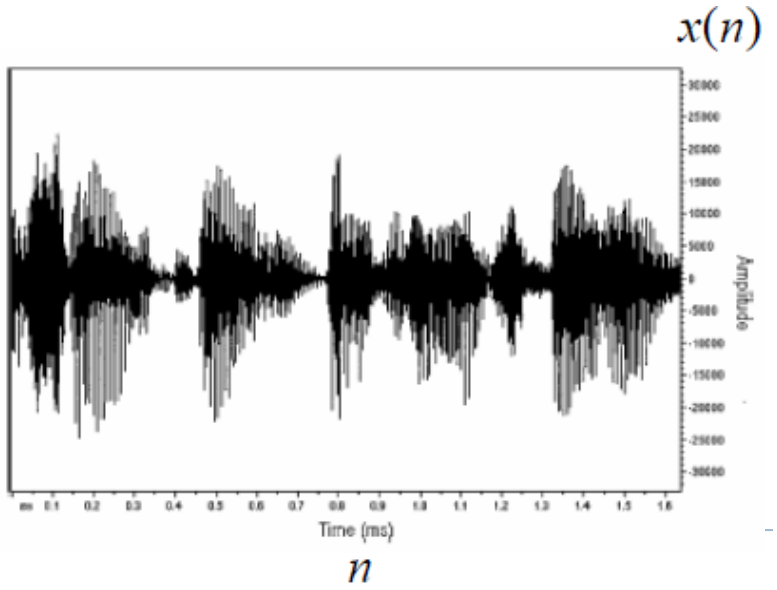
- Năng lượng trung bình: thể hiện độ to nhỏ của âm

$$E = \frac{\sum_{n=0}^{N-1} x(n)^2}{N}$$

- Tốc độ đổi dấu của tín hiệu: đại diện tần số trung bình của tín hiệu

$$ZC = \frac{\sum_{n=1}^{N} |sgn[x(n)] - sgn[x(n-1)]|}{2N}$$

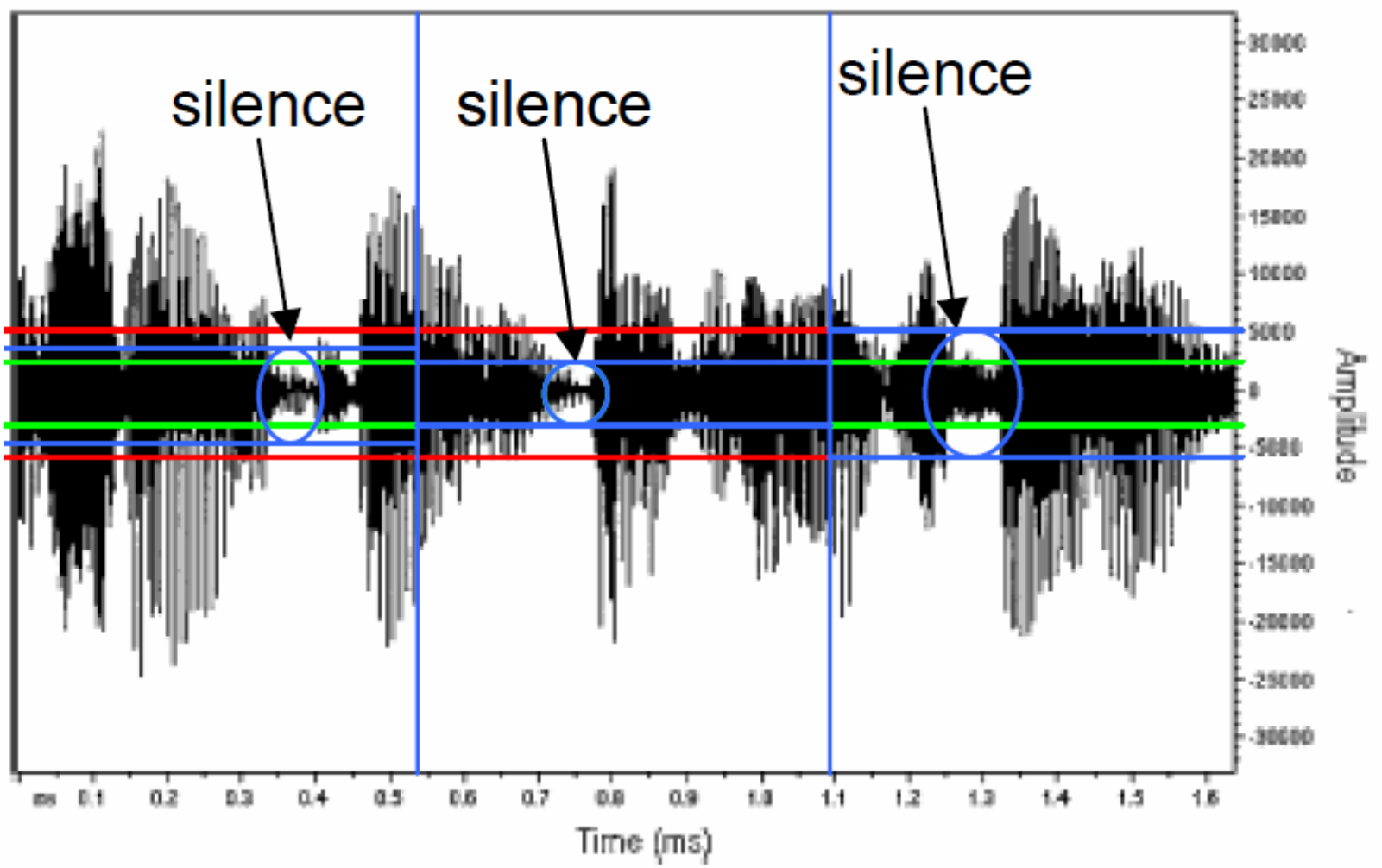
$$sgn(a) = \begin{cases} 1 & a > 0 \\ 0 & a = 0 \\ -1 & a < 0 \end{cases}$$



Các thuộc tính miền thời gian

- Tỉ lệ của khoảng lặng trong âm
  - Khoảng lặng là khoảng thời gian mà giá trị tuyệt đối của biên độ sóng âm nhỏ hơn một ngưỡng cho trước
  - Phần trăm khoảng lặng là tỉ lệ giữa tổng thời gian không có tiếng chia cho toàn bộ độ dài của tệp âm thanh
  - Có 2 giá trị ngưỡng cần xác định: ngưỡng biên độ và ngưỡng thời gian.
  - Các cách để xác định ngưỡng biên độ:
    - Sử dụng một giá trị ngưỡng cố định cho trước
    - Sử dụng ngưỡng tương đối, dựa trên một giá trị có sẵn
    - Sử dụng các giá trị ngưỡng thích ứng.

Các thuộc tính miền thời gian



Các thuộc tính miền tần số

- Phổ âm: sử dụng biến đổi Fourier hoặc các phương pháp biến đổi khác
  - Biến đổi Fourier rời rạc (DFT)

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{-\frac{j2\pi nk}{N}}$$

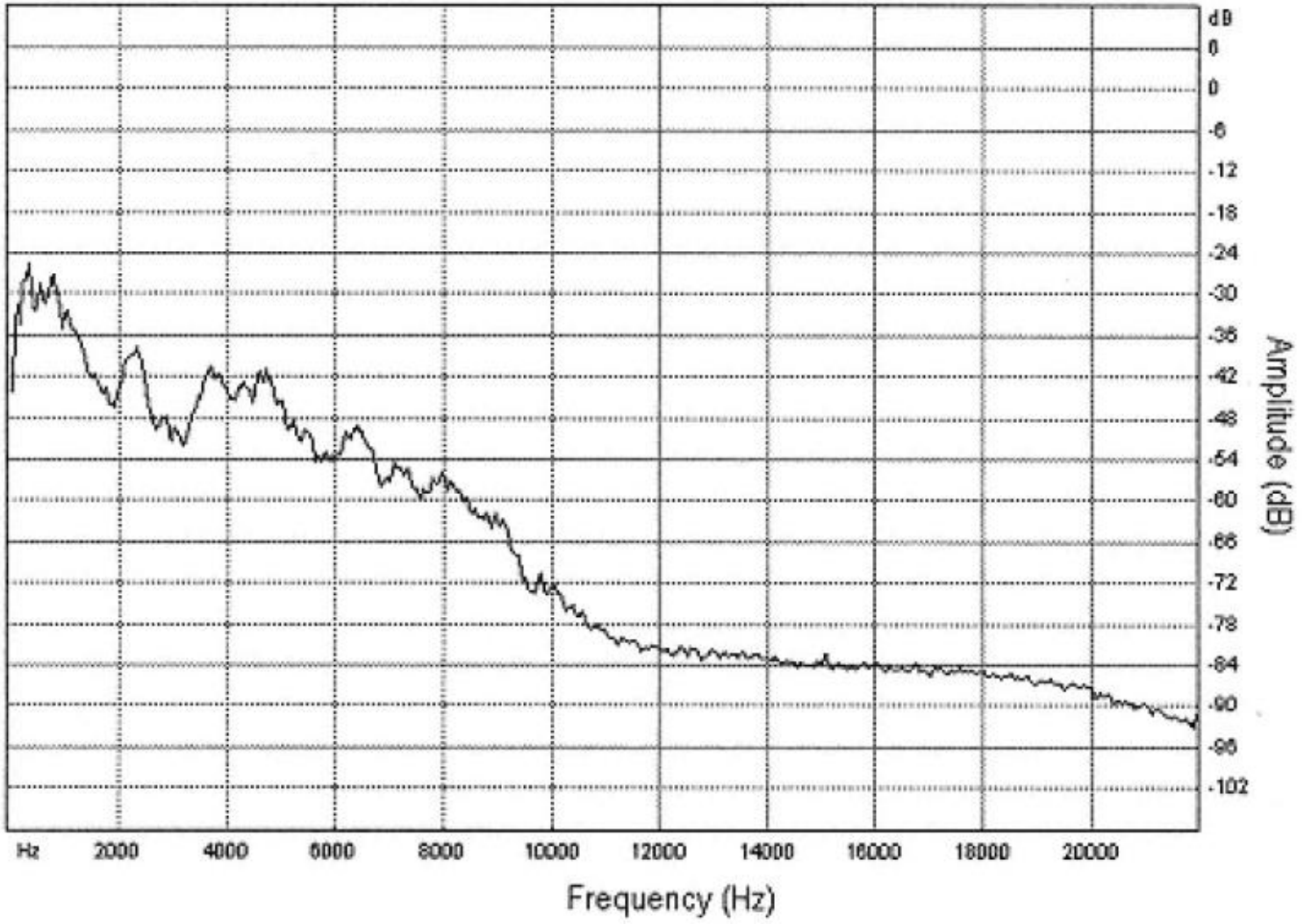
$X(k)$ tuần hoàn với tấn số $f_k = f_s \frac{\omega_k}{N} = f_s \frac{k}{N}$

  - Biến đổi ngược Fourier rời rạc (IDFT)

$$x(n) = \frac{1}{N} \sum_{k=0}^{N-1} X(k)\, e^{\frac{j2\pi nk}{N}}$$

Nếu N lớn, tín hiệu được chia thành các đoạn nhỏ (khung), DFT được thực hiện trên các đoạn này – gọi là chuyển đổi Fourier trên thời gian ngắn (short time FT).

Ví dụ phổ của một tín hiệu âm thanh



Các thuộc tính miền tần số

- Băng thông
  - Là dải tần số của âm thanh
  - Được tính bằng hiệu số giữa tần số cao nhất và tần số thấp nhất của các thành phần phổ dương.
  - “dương” được xác định khoảng 3dB ở trên mức câm.
- Phân bố năng lượng
  - Là sự phân bố tín hiệu dọc theo các thành phần tần số
  - Thông thường, tín hiệu nhạc có tần số cao hơn tín hiệu thoại. Tần số của tín hiệu thoại thường không quá 7kHz.
  - Dựa trên sự phân bố năng lượng âm, ta có thể xác định được trọng tâm (centroid) của âm, là điểm trung bình của năng lượng âm. Trọng tâm của âm còn được gọi là độ sáng (brightness).

Các thuộc tính miền tần số

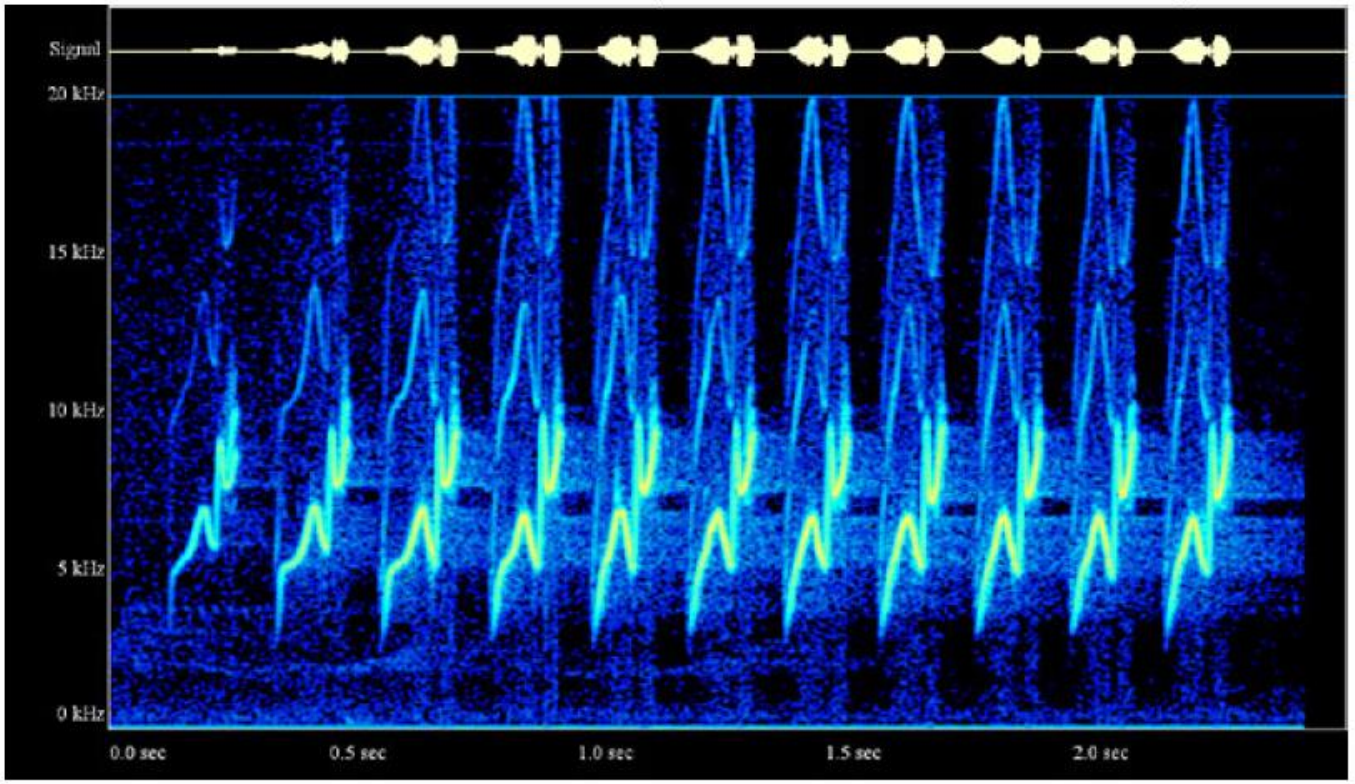
- Độ điều hòa âm (harmonicity)
  - Trong các âm thanh có giai điệu, các thành phần phổ hầu hết bao gồm các hệ số của tần số thấp nhất (thường là to nhất) của âm.
  - Tần số thấp nhất của âm được gọi là tần số cơ bản
  - Âm nhạc thường có độ điều hòa âm lớn hơn các âm thanh khác.
  - Ví dụ: âm thanh của sáo thường có tần số của các đỉnh phổ âm là 400Hz, 800Hz, 1200Hz, 1600Hz, … tương ứng với f, 2f, 3f, 4f,… trong đó f=400Hz là tấn số cơ bản của âm sáo.

Các thuộc tính miền tần số

- Độ cao của âm (pitch)
  - Là đặc tính riêng về chất lượng âm thanh, phụ thuộc vào tần số của nguồn âm
  - Chỉ có các âm thanh có tính tuần hoàn, hay các âm thanh do nhạc cụ phát ra là có thể có âm cao
  - Tần số cơ bản thường được dùng để ước lượng độ cao của âm

Ảnh phổ (spectrogram)

- Là hình thức diễn tả đồng thời các thành phần ở miên tần số và miên thời gian của âm



frequency

time

Intensity: Power of a frequency component at a particular time interval

So sánh giữa âm thoai và âm nhạc

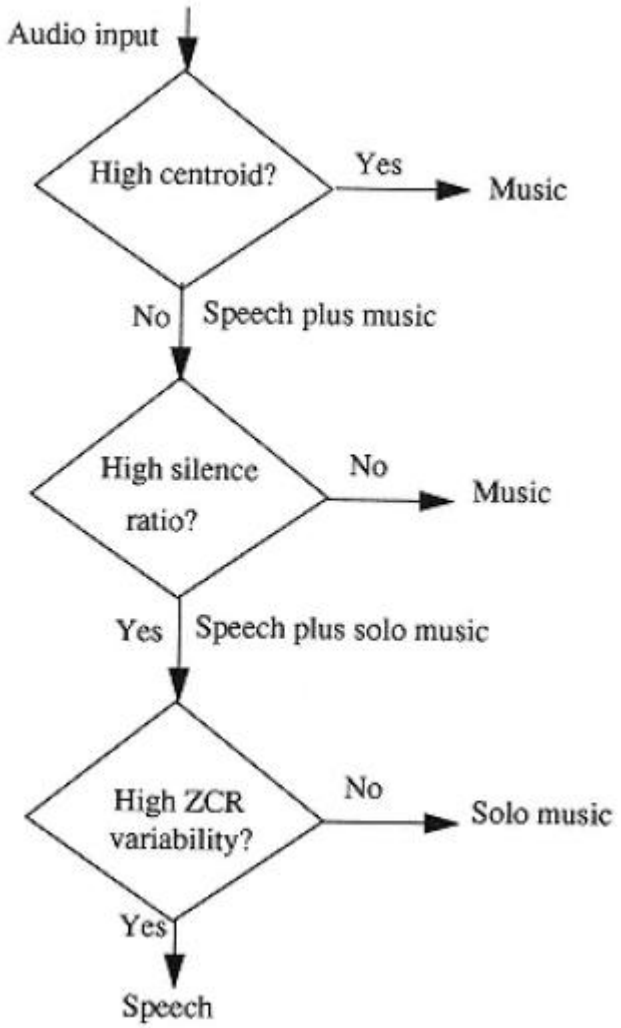
Features	Speech	Music
Bandwidth	0–7 kHz	0–20 kHz
Spectral centroid	low	high
Silence ratio	high	low
Zero-crossing rate	more variable	less variable
Regular beat	no existing	often existing

Phân loại âm thanh

- Phân loại theo từng bước
  - Tại mỗi bước, từng thuộc tính của âm thanh được sử dụng riêng rẽ để phân loại âm thanh
  - Mỗi thuộc tính đóng vai trò như một bộ lọc
  - Thứ tự lựa chọn các thuộc tính để phân loại rất quan trọng, phụ thuộc vào độ phức tạp cũng như khả năng nhận biết các âm thanh khác nhau của từng thuộc tính
  - Những thuộc tính có độ phức tạp thấp và khả năng nhận diện loại âm thanh cao sẽ được xếp trước.
  - Qúa trình phân loại có thể trải qua nhiều bước với nhiều thuộc tính được sử dụng, hoặc cũng có thể chỉ sử dụng một thuộc tính đặc trưng nào đó.

Phân loại âm thanh theo từng bước

- Lu và Hankinson 1998

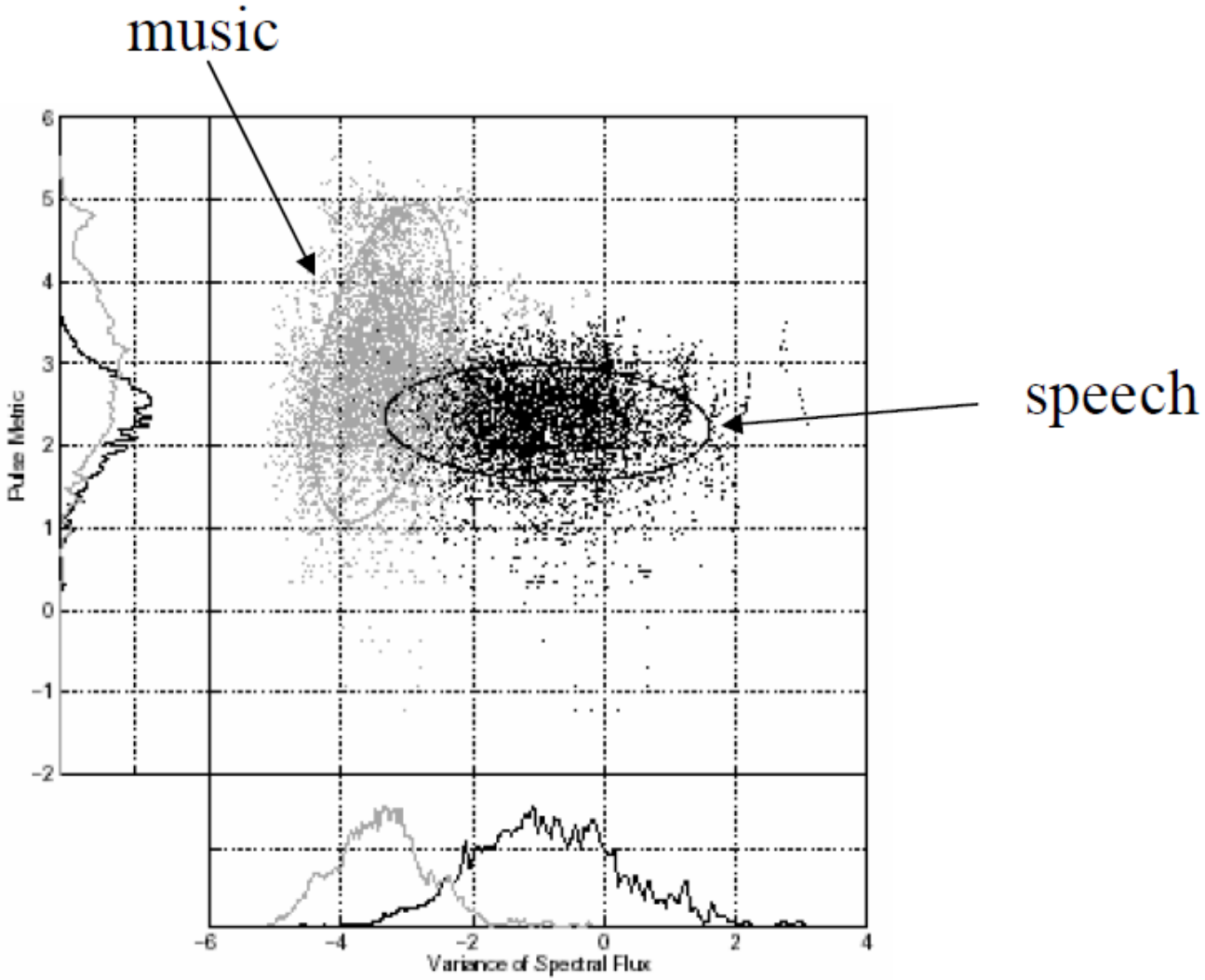


Phân loại âm thanh

- Phân loại dựa trên vector thuộc tính
  - Một tập thuộc tính được ghép với nhau thành một vector
  - Yêu cầu độ tính toán nhiều hơn, nhưng mạng lại kết quả tốt hơn việc dùng các thuộc tính đơn lẻ
  - Truy vấn không gian vector có thể được sử dụng để đánh giá sự tương đồng giữa các bản ghi âm thanh
  - Âm thanh cũng có thể được phân loại dựa trên quá trình phân cụm các vector thuộc tính.

Phân loại âm thanh dựa trên vector thuộc tính

- Scheirer và Stanley 1997



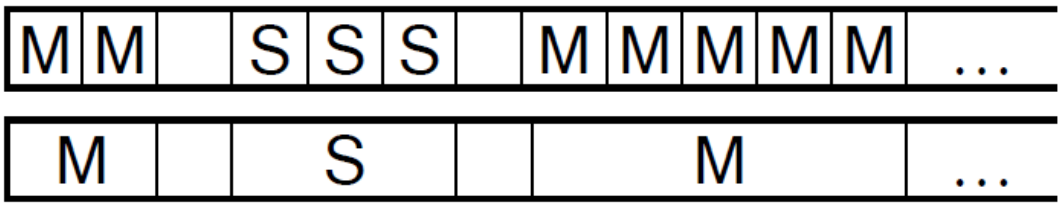
Ví dụ các loại âm thanh

- Liu và Wan 2001

Class name	Number of files	Class name	Number of files
1.Speech	131	Mammals	174
Female	66	3.2Effect	344
Male	54	Alarm	17
Mixture	11	Beat	42
2.Music	458	Bells	50
2.1Instruments	365	Car	12
Brass	44	Clock	15
Keyboard	6	Explosion	44
Percussion	125	Gun	25
String	139	Machines	110
Synthesis	14	Telephone	29
Woodwind	37	3.3Nature	41
2.2Melody	36	Storm	11
Melody	36	Thunder	5
2.3Song	57	Water	20
Cartoon Song	7	Wind	5
Normal Song	50	3.4People	44
3:Sound	649	Applause	8
3.1Animal	220	Laughter	36
Birds	46	Total	1238

Phân đoạn âm thanh

- Một tệp âm thanh thường bao gồm nhiều đoạn chứa các loại âm thanh khác nhau
- Tệp âm thanh có thể chia thành các đoạn chứa riêng rẽ từng loại âm thanh
- Phương pháp:
  - Chia tệp âm thanh thành các đoạn nhỏ, sau đó dùng các phương pháp phân loại để phân loại âm thanh trong các đoạn đó
  - Các đoạn âm thanh liền kề được ghép lại với nhau nếu chúng cùng loại

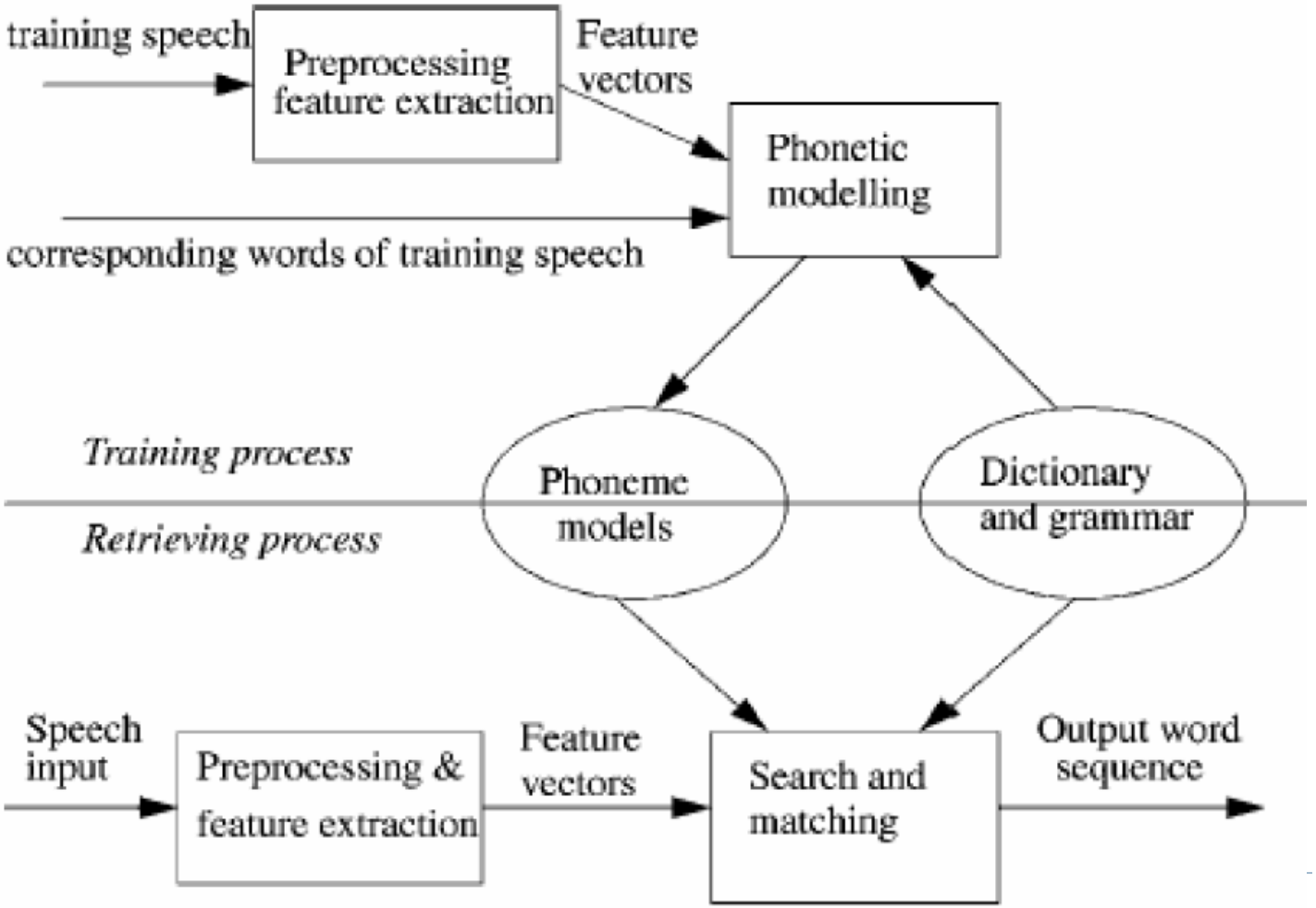


Nhận dạng và truy vấn âm thoại

- Sử dụng các kỹ thuật nhận dạng âm thoại để chuyển đổi âm thoại sang dạng văn bản, sau đó sử dụng hệ thống IR để lưu trữ và truy vấn
  - Nhận dạng âm thoại
    - Các khái niệm cơ bản về nhận dạng âm thoại tự động (ASR)
    - Các phương pháp cải tiến
    - Các kỹ thuật dựa trên mô hình Markov ẩn (HMM)
    - Dựa trên mạng nơ ron nhân tạo
  - Nhận dạng người nói

Các khái niệm cơ bản về ASR

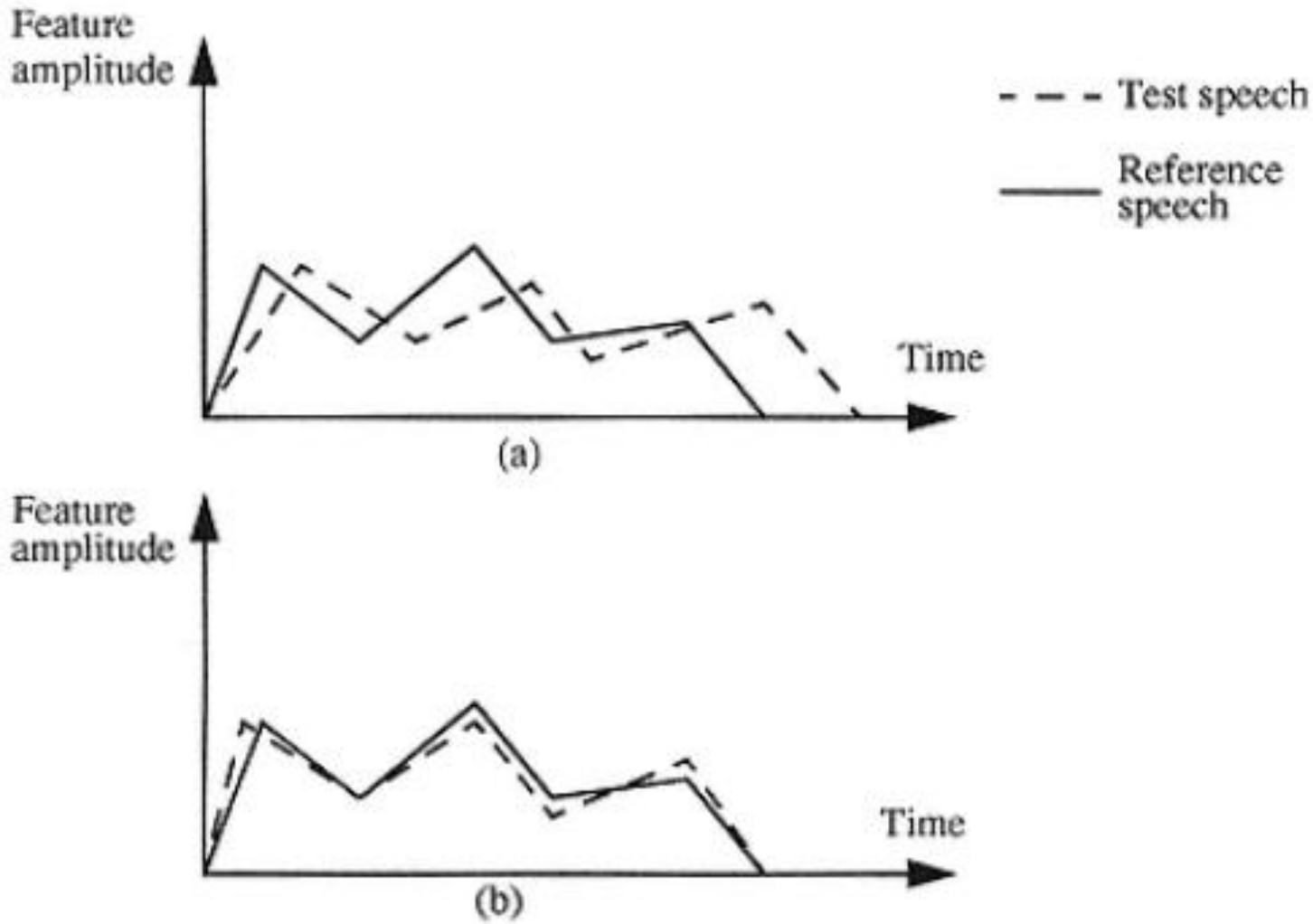
- Hệ thống ASR nói chung



Các hạn chế của ASR

- Phụ thuộc vào chủ đề
- Phụ thuộc vào thời gian
- Bị ảnh hưởng bởi nhiễu
- Sự khác nhau giữa âm của các từ đơn lẻ với âm của các từ nói trong câu liên tục
- Sự khác nhau giữa âm đọc và âm nói
- Phụ thuộc vào kích thước bộ từ vựng

Các hạn chế của ASR



Lưu trữ và truy vấn âm nhạc

- Có hai loại dữ liệu âm nhạc
  - Âm nhạc có cấu trúc và các hiệu ứng âm thanh
  - Âm nhạc dựa trên lấy mẫu
    - Truy vấn dựa vào vector thuộc tính / đặc trưng
    - Truy vấn dựa trên độ cao thấp (pitch)

Âm nhạc có cấu trúc

- Có thể được biểu diễn bởi một tập các câu lệnh hoặc thuật toán
- Ví dụ:
  - MIDI
  - MPEG-4
- Được phát triển để phục vụ việc truyền âm thanh, tái tạo âm thanh, chứ không nhằm mục đích truy vấn dữ liệu âm thanh. Tuy nhiên, cấu trúc của âm thanh lại rất thuận tiện trong việc tra cứu chúng.
- Phù hợp trong việc tìm kiếm chính xác dữ liệu

Âm nhạc dựa trên lấy mẫu

- Được truy vấn dựa trên các thuộc tính của chúng
  - Phân loại âm thanh dựa trên thuộc tính
  - Tra cứu bằng cách so sánh các thuộc tính

Feature		Mean	Variance	Importance
Duration		2.71982	0.191312	6.21826
Loudness:	Mean	–45.0014	18.9212	10.3455
	Variance	200.109	1334.99	5.47681
	Autocorrelation	0.955071	7.71106e-05	108.762
Brightness:	Mean	6.16071	0.0204748	43.0547
	Variance	0.0288125	0.000113187	2.70821
	Autocorrelation	0.715438	0.0108014	6.88386
Bandwidth:	Mean	0.363269	0.000434929	17.4188
	Variance	0.00759914	3.57604e-05	1.27076
	Autocorrelation	0.664325	0.0122108	6.01186
Pitch:	Mean	4.48992	0.39131	7.17758
	Variance	0.207667	0.0443153	0.986485
	Autocorrelation	0.562178	0.00857394	6.07133

Class Model for Laughter (Wold et al., 1996)