fit@hcmus

Nhận dạng tiếng nói

Nguyễn Đức Hoàng Hạ

2022

Nội dung

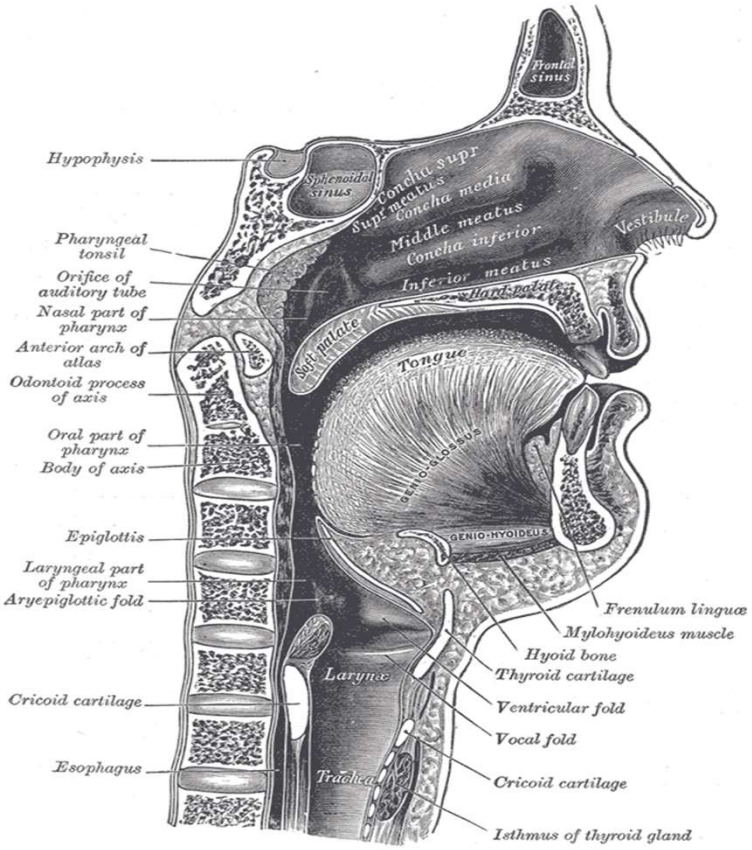
- Tín hiệu tiếng nói
- Bài toán nhận dạng tiếng nói (Automatic Speech Recognition)
- Một số bài toán nhận dạng liên quan đến tiếng nói
- Đồ án môn học

Khảo sát lớp học

https://www.menti.com/

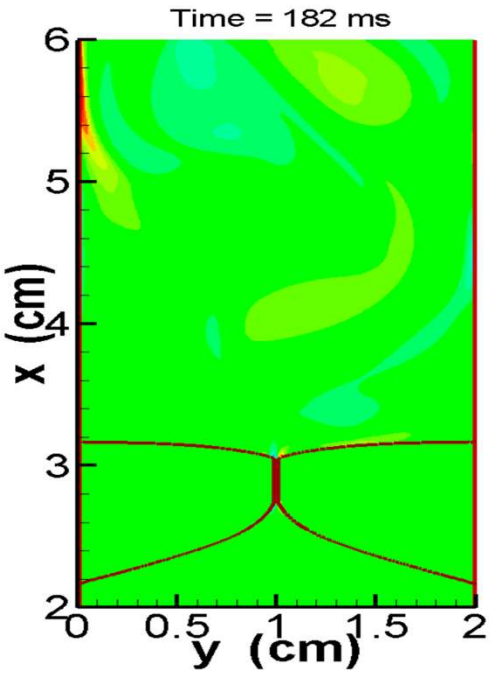
Tiếng nói được phát ra như thế nào?

Bộ phận phát âm ở người



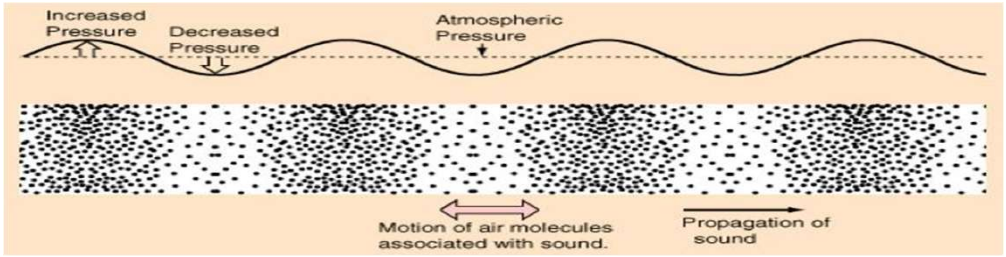
Tiếng nói được phát ra như thế nào?

Hơi ở phổi tạo dao động ở thanh quản

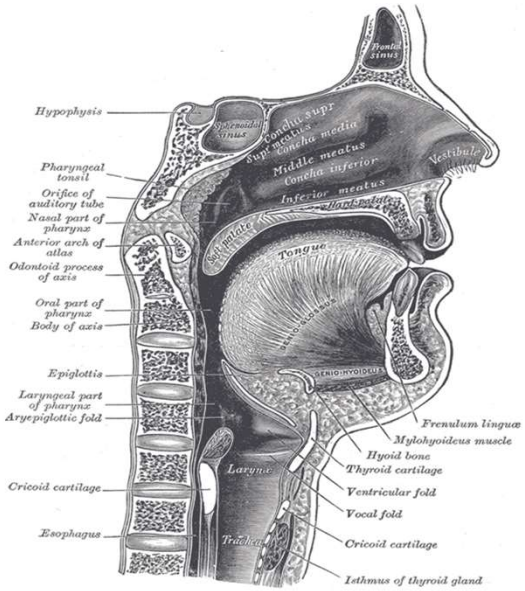


Tiếng nói được phát ra như thế nào?

Dao động của thanh quản tạo ra lan truyền dao động
áp suất trong không khí



Tiếng nói được phát ra như thế nào?



Mô hình bộ phận phát âm đơn giản

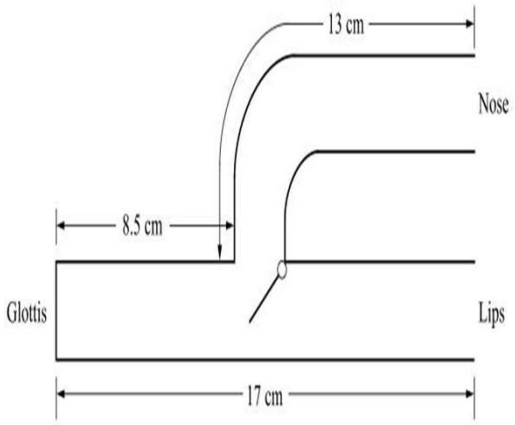
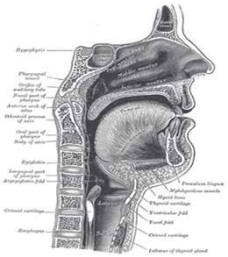


Figure 2.5: Simplified physical model of the vocal tract

Tiếng nói được phát ra như thế nào?

Hình thể trong miệng khác nhau sẽ tạo cộng hưởng khác nhau

==> âm vị khác nhau



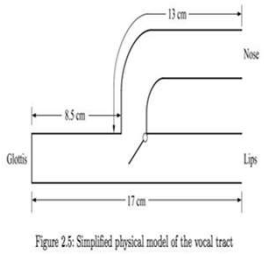
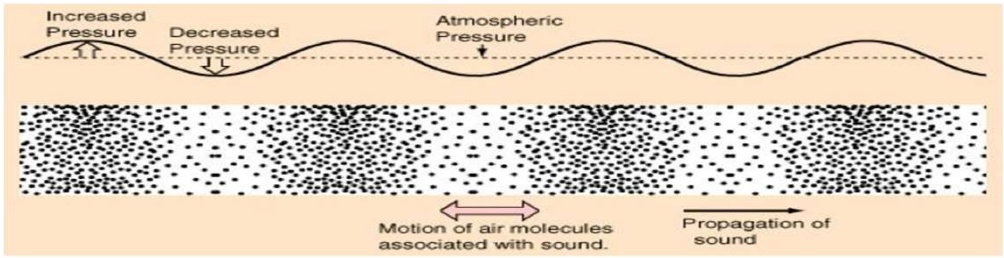
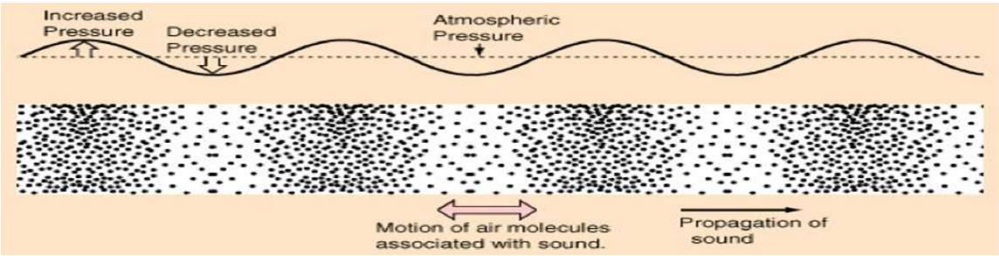


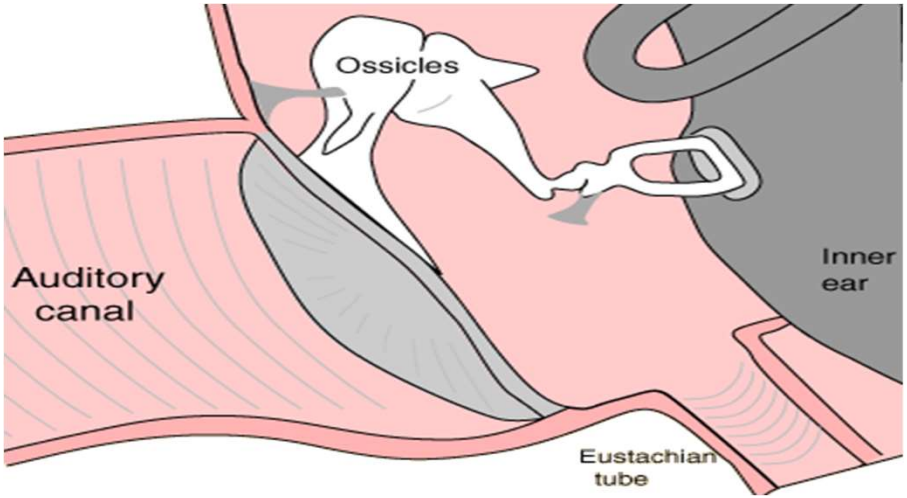
Figure 2.5: Simplified physical model of the vocal tract



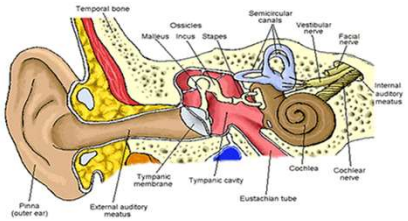
Âm thanh được cảm nhận như thế nào?

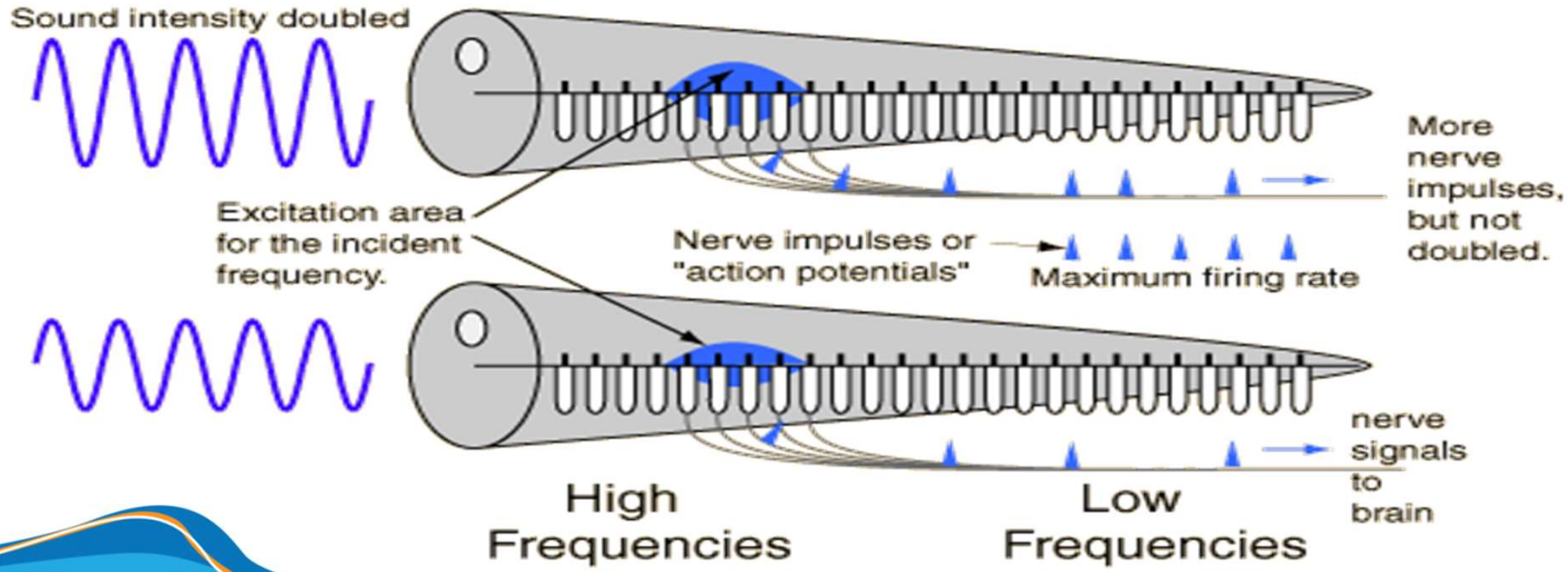
Cách cảm nhận âm thanh ở người





Âm thanh được cảm nhận như thế nào?



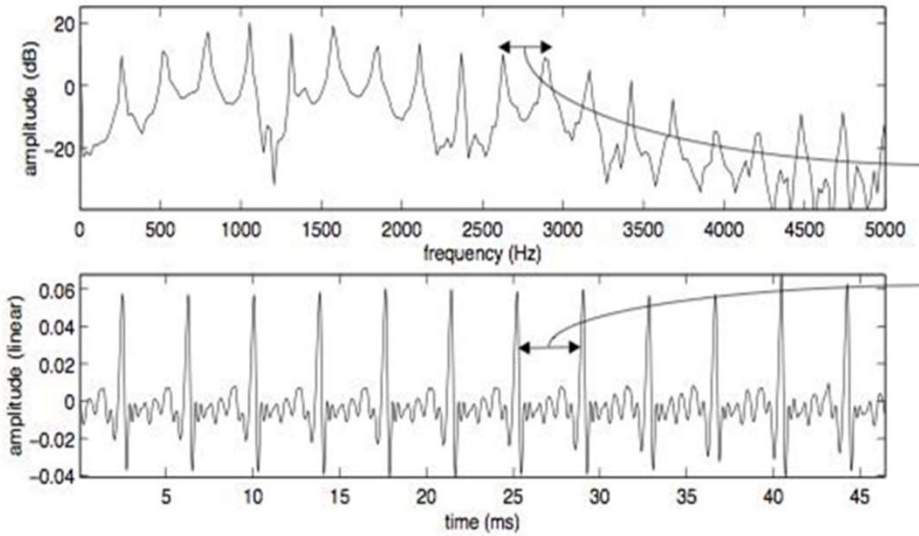


Tần số chuẩn F0 và các formants F1, F2 ...

For sine waves:

- **F0 = frequency**
- **pitch ~ frequency**

Complex harmonic sounds



Trumpet sound

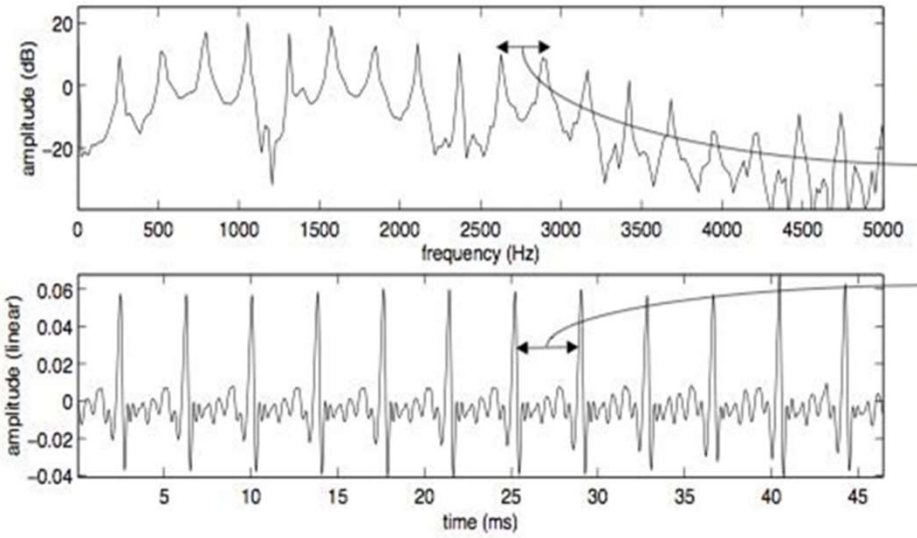
- **F0:**
  ***F* = 262 Hz**
- **wavelength**
  1/*F* = 3.8 ms

Tần số chuẩn F0 và các formants F1, F2 ...

For sine waves:

- **F0 = frequency**
- **pitch ~ frequency**

Complex harmonic sounds

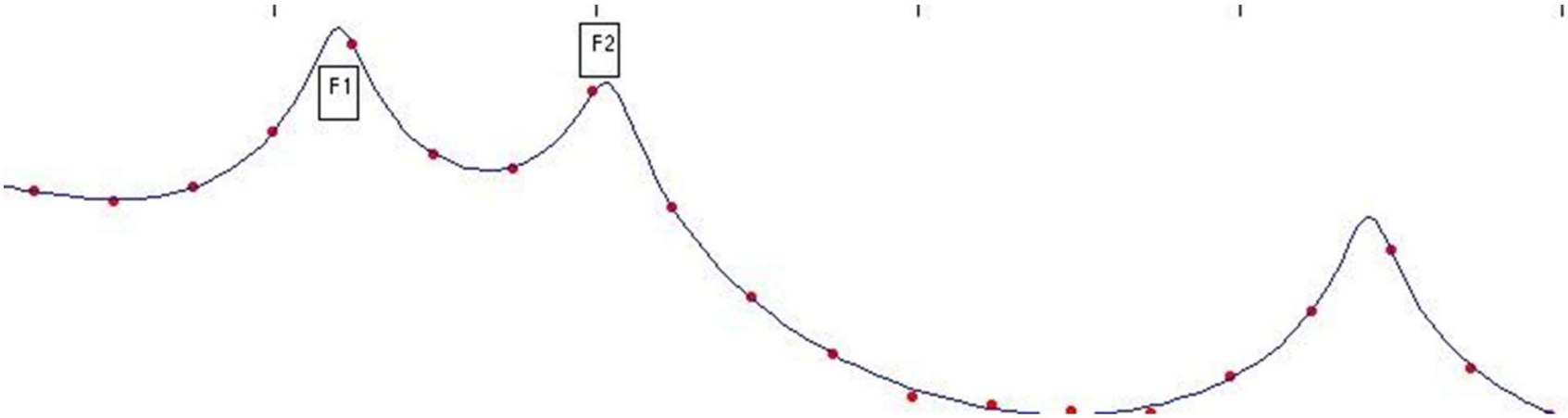


Trumpet sound

- **F0:**
  ***F* = 262 Hz**
- **wavelength**
  1/*F* = 3.8 ms

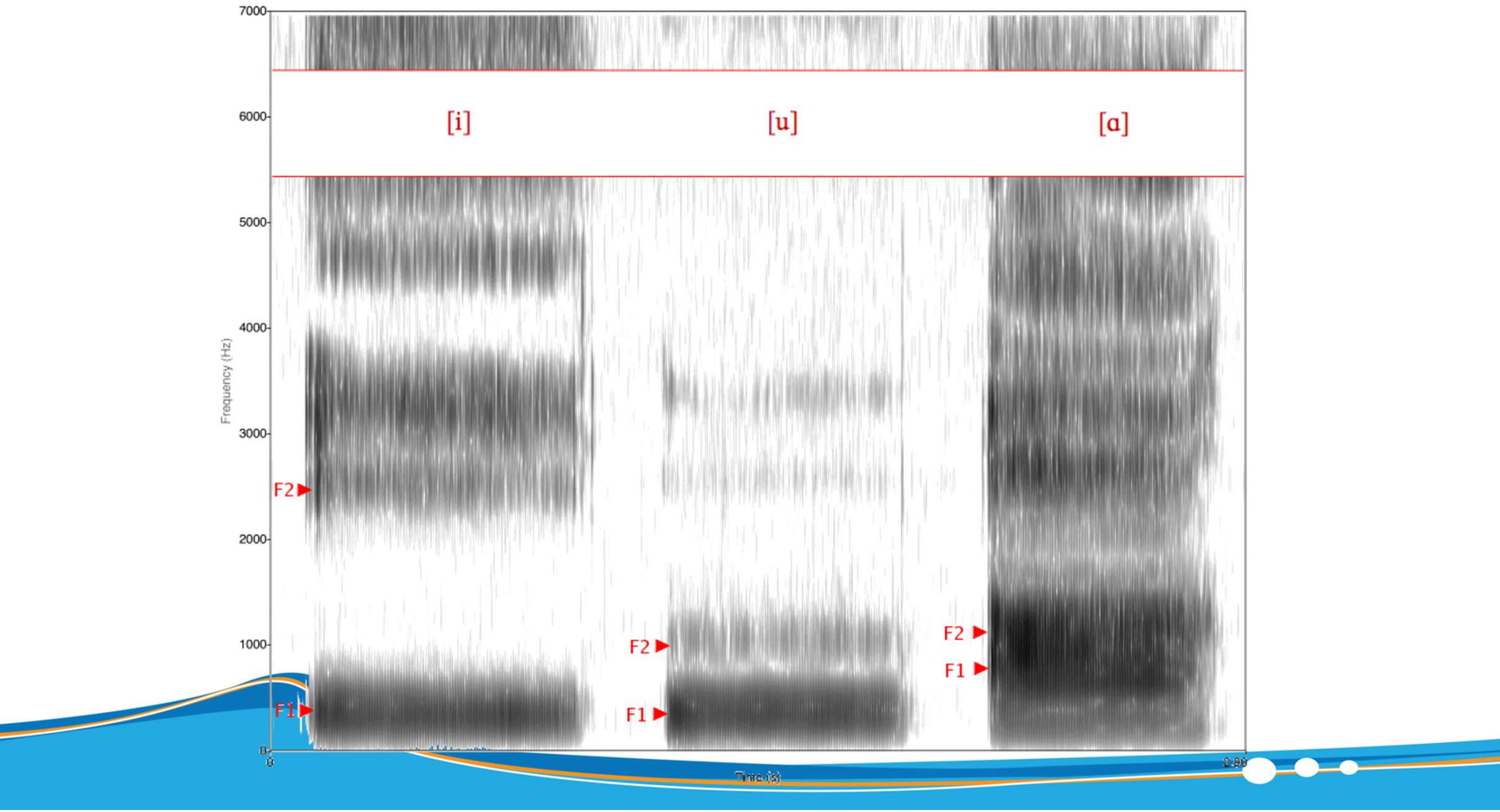
Hỏi: wavelength là bao nhiêu mét?

Tần số chuẩn F0 và các formants F1, F2 ...



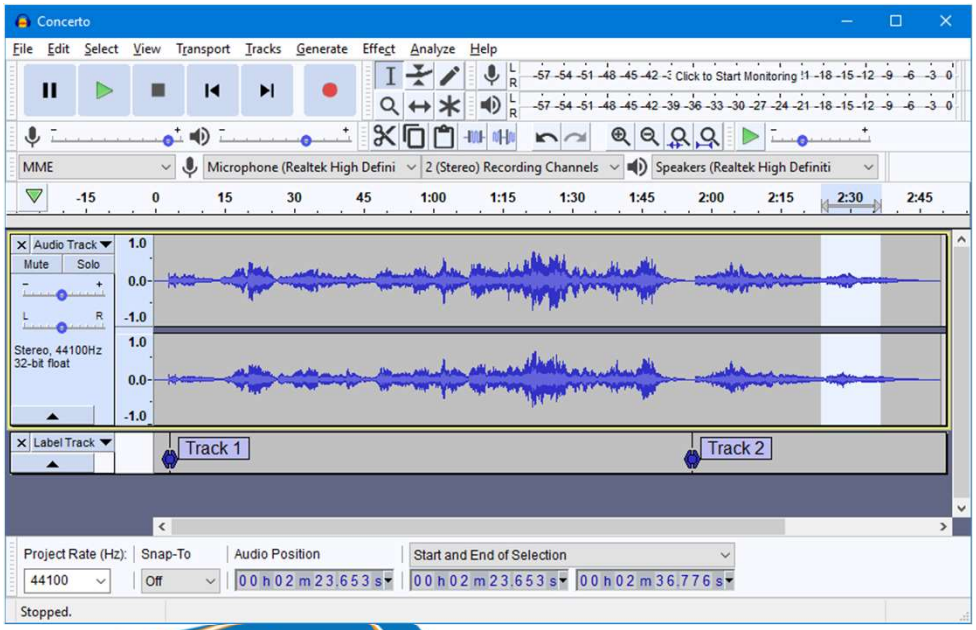
F0 liên quan đến cấu tạo dây thanh quản, thể hiện nguồn tạo năng lượng ban đầu cho âm thanh.
F1, F2, ... liên quan đến vòm họng, miệng, hốc mũi, ... thể hiện quá khứ của âm thanh đã trải qua

Tham khảo thêm: https://home.cc.umanitoba.ca/~krussll/phonetics/acoustic/formants.html


7000
6000
5000
4000
3000
2000
1000
0
Frequency (Hz)
[i]
[u]
[ɑ]
F2
F1
F2
F1
F2
F1
0
Time (s)

Ứng dụng hỗ trợ phân tích âm thanh

Audacity: https://www.audacityteam.org/



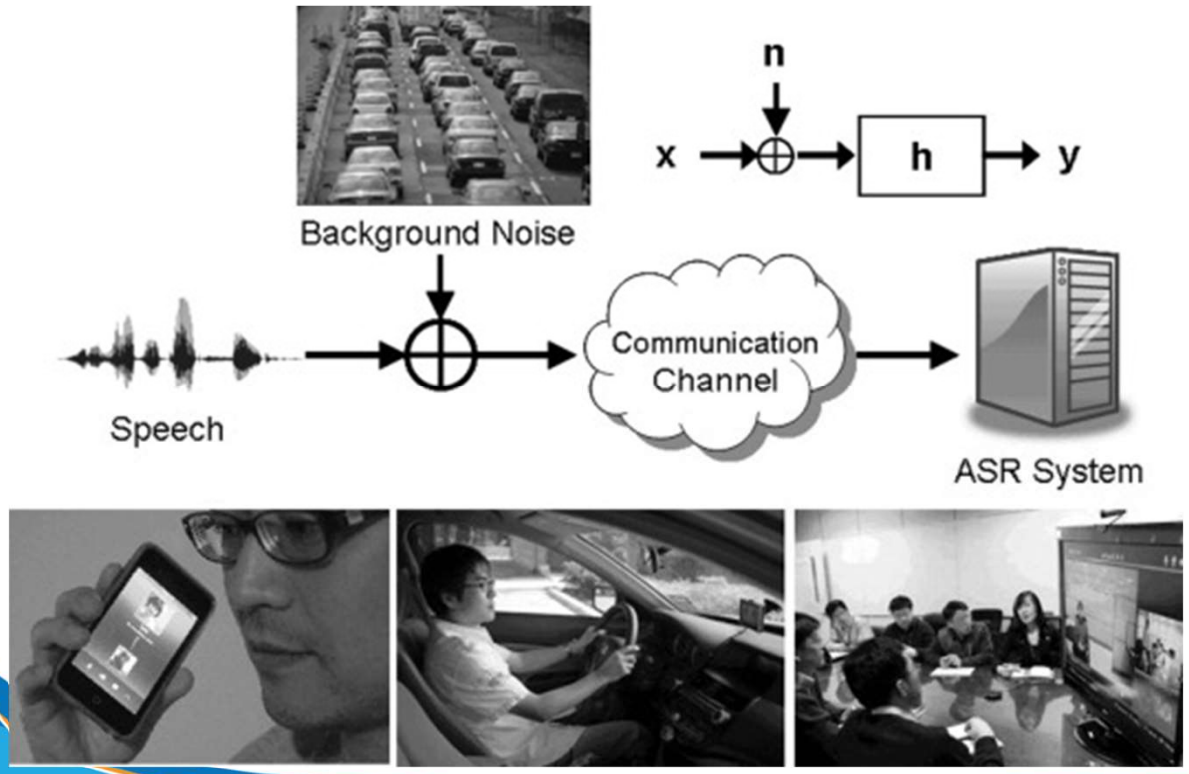
Praat: https://www.fon.hum.uva.nl/praat/

Script: https://lennes.github.io/spect/#forced-alignment

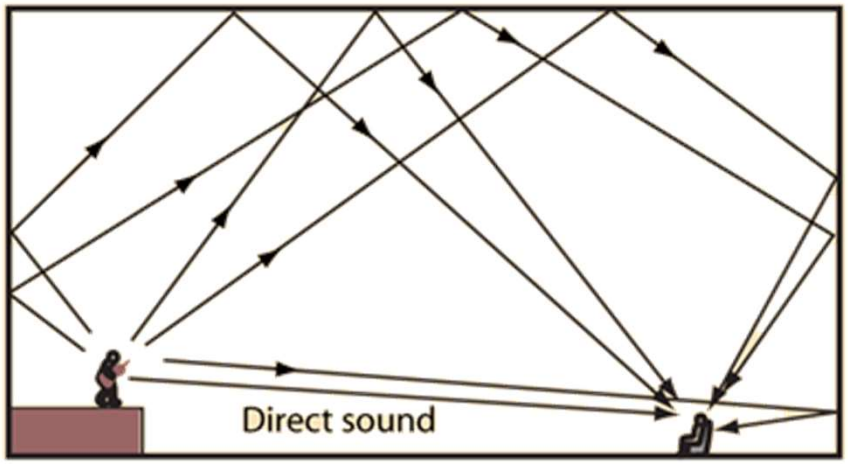
Môi trường truyền âm

- Tiếng nói không đến ngay lập tức từ miệng người nói đến tai người nghe (hay microphone) mà được truyền trong không khí
- Hỏi: Các loại nhiễu trong môi trường truyền âm?

Nhiễu do kênh thu âm và âm nền



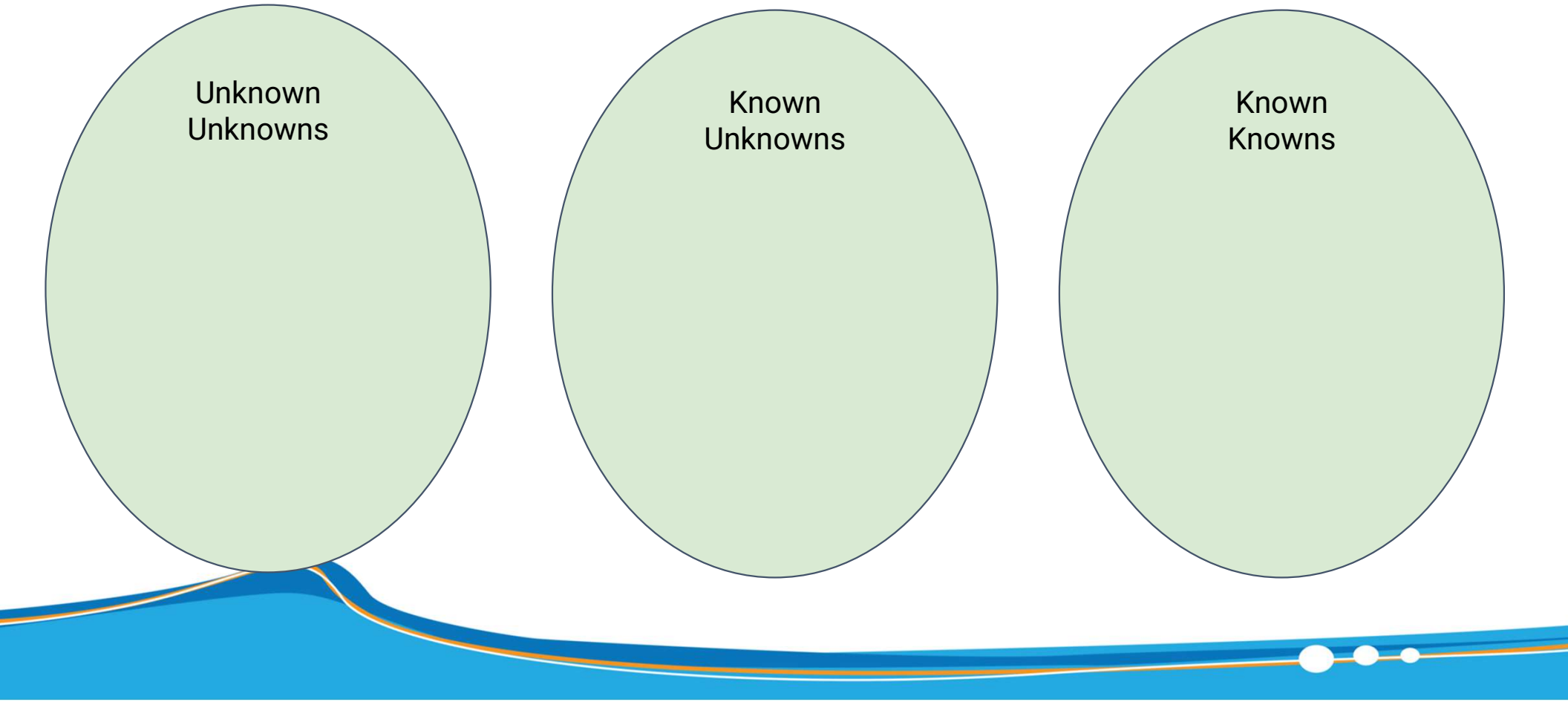
Nhiễu do tiếng vọng trong phòng



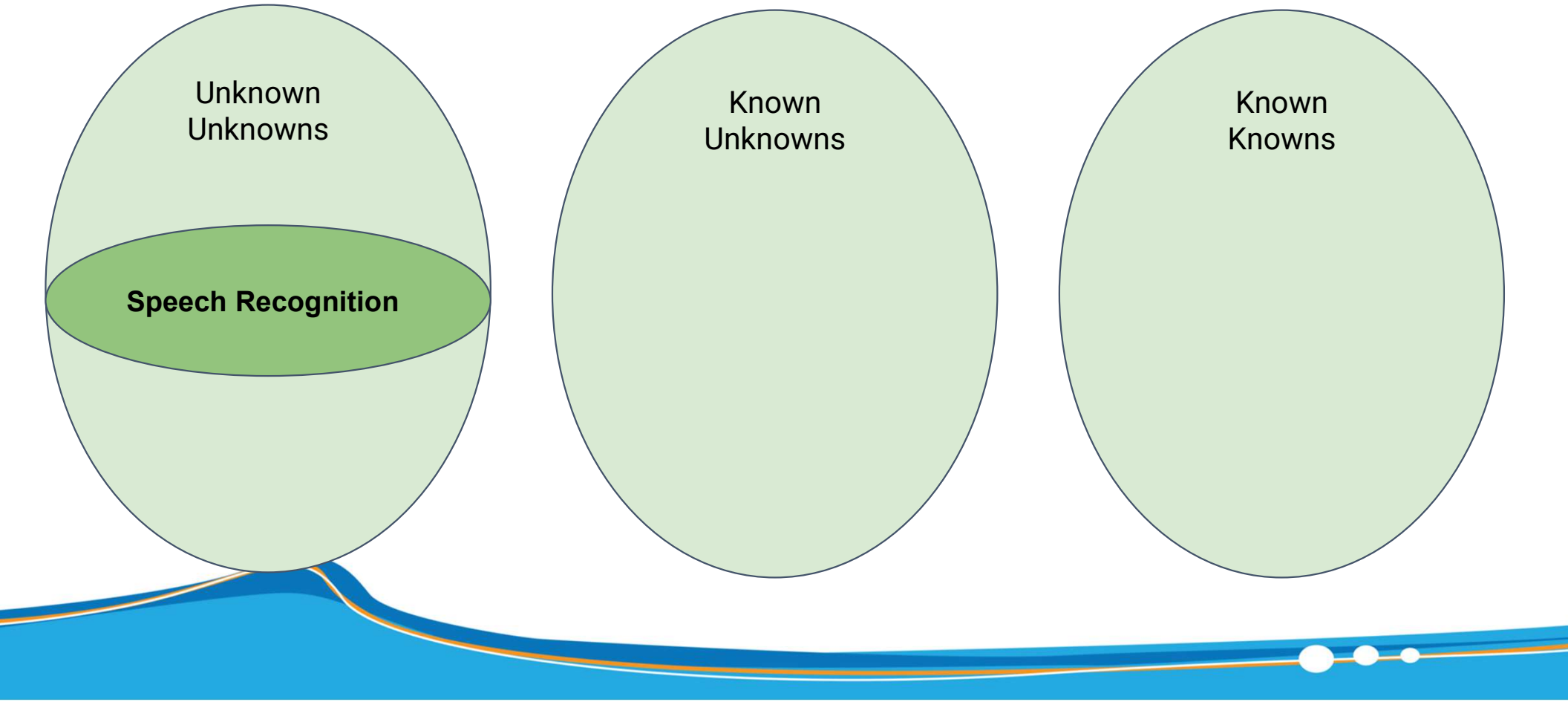
Bài toán nhận dạng tiếng nói

- Chuyển đổi từ tiếng nói thành văn bản
  - Cho 1 người cụ thể
  - Cho nhiều giọng
  - Cho nhiều ngôn ngữ
- Mô hình ngữ âm và mô hình ngôn ngữ
  - Mô hình ngữ âm
  - Mô hình ngôn ngữ

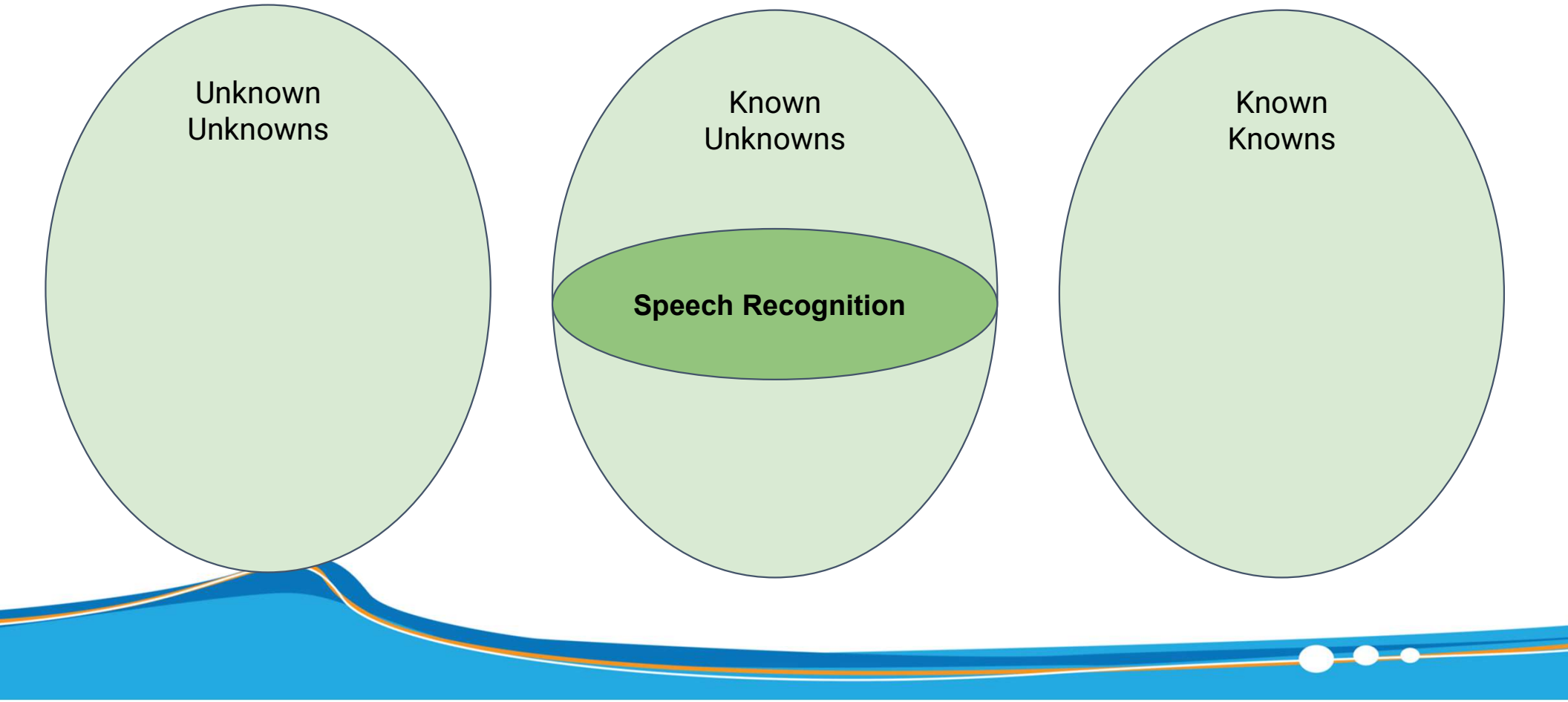
“There are known knowns. There are things we know that we know. There are known unknowns. That is to say, there are things that we now know we don't know. But there are also unknown unknowns. There are things we do not know we don't know.” Donald Rumsfeld



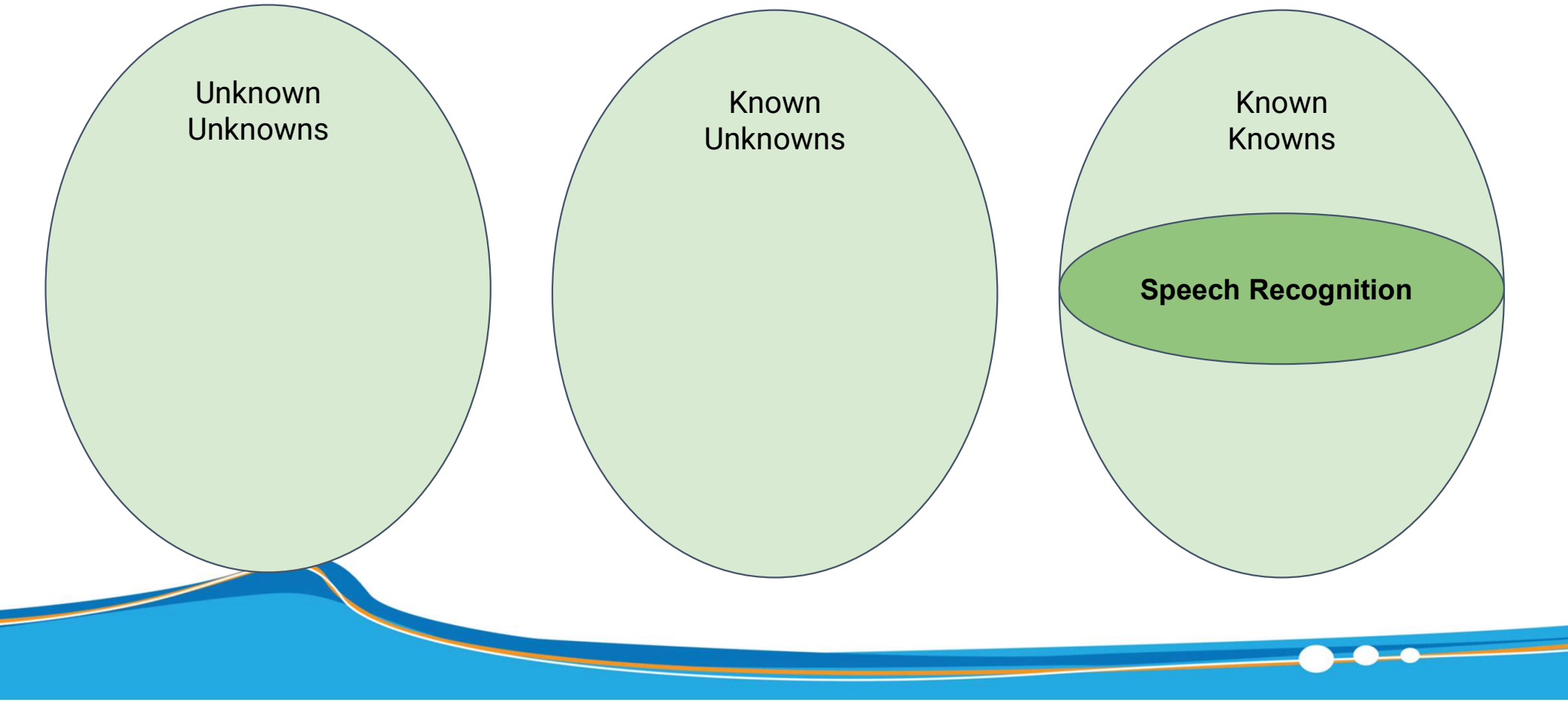
Hiện tại


Unknown
Unknowns
Speech Recognition
Known
Unknowns
Known
Knowns

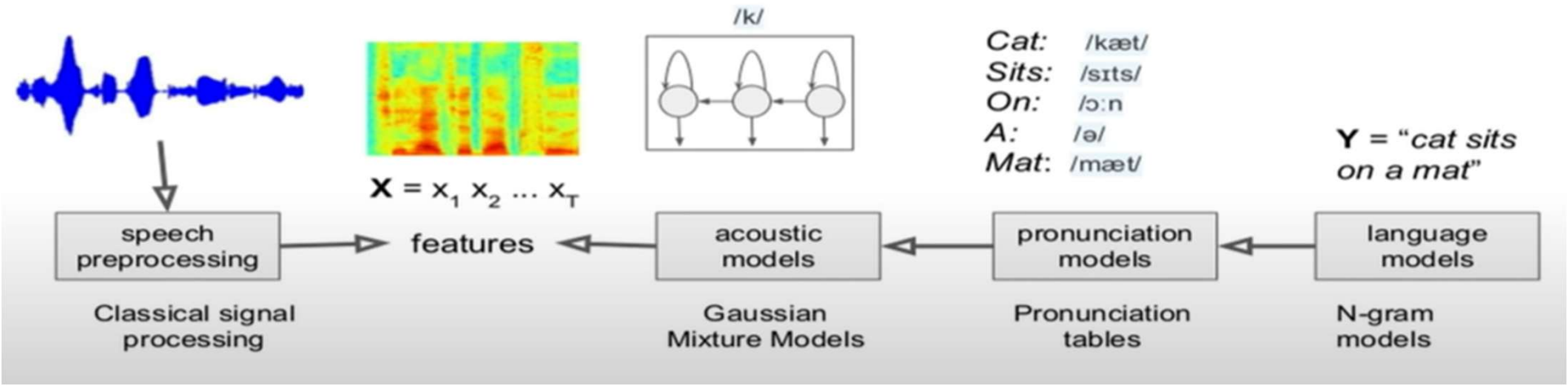
Học xong khóa học này



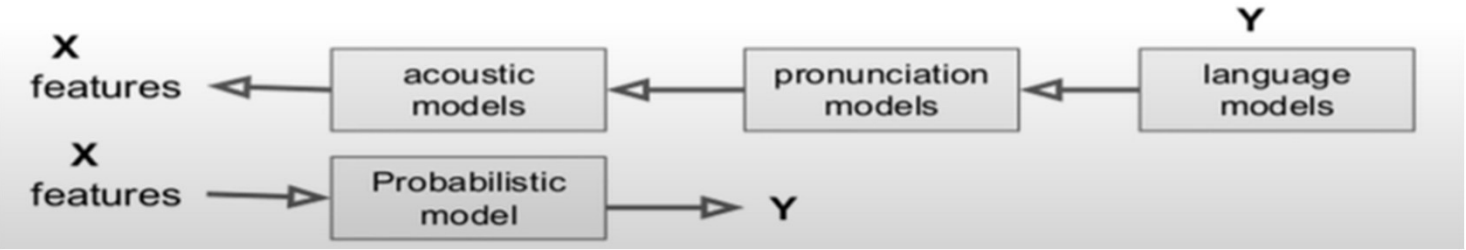
Nghiên cứu sâu về ASR sau khóa học



Bài toán nhận dạng tiếng nói (Automatic Speech Recognition)

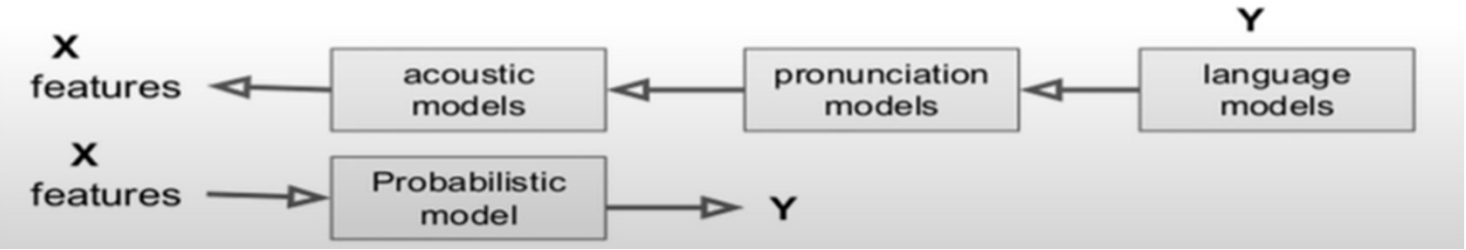


Bài toán nhận dạng tiếng nói (End-to-end Speech Recognition)



$$Y^* = \arg\max_{Y} p(Y|X)$$

Bài toán nhận dạng tiếng nói (End-to-end Speech Recognition)



$$Y^* = \arg\max_{Y} p(Y|X)$$

Given audio $X = x_1 x_2 \ldots x_T$ and corresponding output text $Y = y_1 y_2 \ldots y_L$ where $y \in \{a, b, c, d, \ldots z, ?, !, \ldots\}$

Y is just a text sequence (transcript), **X** is the audio / processed spectrogram

Perform speech recognition, by learning a probabilistic model $p(Y|X)$

Stanford University School of Engineering, Lecture 12: End-to-End Models for Speech Processing
https://www.youtube.com/watch?v=3MjIkWxXigM

Một số thư viện mở xây dựng hệ nhận dạng tiếng nói

KALDI

Homepage: http://kaldi-asr.org/;

Code: https://github.com/kaldi-asr/kaldi

Tutorial: http://kaldi-asr.org/doc/tutorial.html

HTK Homepage: http://htk.eng.cam.ac.uk/

CMU Sphinx Homepage: http://www.speech.cs.cmu.edu/sphinx/doc/Sphinx.html

Một số thư viện mở xây dựng hệ nhận dạng tiếng nói

ESPNet an end-to-end speech processing toolkit:
https://github.com/espnet/espnet

wav2letter:
https://github.com/flashlight/flashlight/tree/master/flashlight/app/asr

NVIDIA NeMo:
https://github.com/NVIDIA/NeMo/tree/main/tutorials/asr

Một số bài toán nhận dạng liên quan đến tiếng nói

- Bài toán nhận dạng người nói (Speaker Recognition)
- Bài toán xác nhận người nói (Speaker Verification)
- Phát hiện từ khóa (Keyword Detection)
- Phân đoạn người nói (Speaker Diarization)

Bài toán nhận dạng người nói (Speaker Recognition)

- Xác định ai là người nói trong danh sách những người biết trước

Bài toán xác nhận người nói (Speaker Verification)

- Kiểm tra có đúng đây là giọng của một người biết trước không.
- Có ai giả giọng nói này không? (Speaker Recognition Anti-Spoofing)

Phát hiện từ khóa (Keyword Detection)

Tương tự như tính năng "ok google" hay "hey siri" trên thiết bị Android phone và iphone.

Phân đoạn người nói (Speaker Diarization)

- Đầu vào là 1 đoạn âm thanh nhiều người nói
- Đầu ra là thông tin: ai nói và nói lúc nào trong đoạn âm thanh đó

Đồ án môn học

Học viên chọn 1 trong 2 hướng làm đề tài sau để làm đồ án môn học:

- Hướng nghiên cứu: Chọn một chủ đề trong xử lý/nhận dạng/tổng hợp tiếng nói; tìm hiểu, khảo sát và viết báo cáo về đề tài đó
- Hướng ứng dụng: Xây dựng 1 ứng dụng, có sử dụng các công nghệ về xử lý/nhận dạng/tổng hợp tiếng nói

Đồ án môn học - Tiêu chí đánh giá

- Hướng nghiên cứu
  - Có tham khảo tối thiểu 5 bài báo về chủ đề được chọn (50%)
  - Có tổng hợp, so sánh, phần tích chi tiết tối thiểu 2 phương pháp (30%)
  - Trình bày, vấn đáp (20%)
  - Có chạy thử nghiệm (+10%)
  - Có thu dữ liệu mới để kiểm tra thử nghiệm (10%)
  - Có cái tiến, kết hợp, đề xuất mới (+30%)

Đồ án môn học - Tiêu chí đánh giá

- Ví dụ về hướng nghiên cứu:
  - Học viên đọc 5 bài báo về chủ đề tự chọn, có liên quan đến công nghệ xử lý tiếng nói. Học viên cần tổng hợp nội dung chính các bài báo này (Tác giả giải bài toán gì? Hướng tiếp cận giải quyết là gì? Mô hình đề xuất? Dữ liệu sử dụng để đánh giá? Kết quả như thế nào? ...).
  - Khi đã hiểu cơ bản ý tưởng các bài báo trên, Học viên chọn 2 phương pháp có tính tương đồng để so sánh, đánh giá chi tiết sự giống nhau và khác nhau của 2 phương pháp được chọn.
  - Học viên có thể cài đặt và chạy thử nghiệm các đề tài. **Cho phép sử dụng lại các thư viện mở.**
  - Nếu có thu thập lại dữ liệu để kiểm chứng thuật toán thì sẽ có điểm cộng
  - Nếu có đề xuất giải pháp mới và có thử nghiệm thì sẽ có điểm cộng
  - Điểm trình bày khi chấm vấn đáp

Danh mục bài báo tham khảo

- Interspeech Sep-2022: https://www.interspeech2022.org/program/
- ICASSP Oct-2022: https://2022.ieeeicassp.org/technical_program.php?T=R
- VLSP Nov-2022: https://vlsp.org.vn/vlsp2022

Đồ án môn học - Tiêu chí đánh giá

- Hướng ứng dụng
  - Học viên cần có bản đặc tả yêu cầu, thiết kế dữ liệu và **thiết kế hệ thống** để xây dựng ứng dụng đã chọn.
  - Học viên cần viết báo cáo về công nghệ xử lý tiếng nói đã chọn
  - Điểm demo hoạt động thực tế của hệ thống
  - Điểm cộng cho việc huấn luyện lại hô hình (không dùng mô hình đã có trên mạng)
  - Điểm cộng cho việc thu thập thêm dữ liệu để cải tiến mô hình huấn luyện
  - Điểm trình bày báo cáo và vấn đáp

Hình thức thi cuối kỳ

- Học viên **ghi hình lại bài trình bày đồ án môn học** và nộp trong tuần thứ 11 và 12
- Thời điểm thi vấn đáp sẽ diễn ra sau tuần nộp đồ án. Giáo viên sẽ hỏi các kiến thức liên quan đến đề tài và trong môn học.