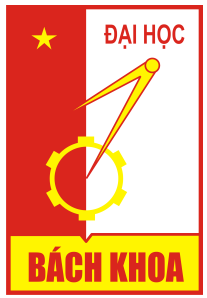
Thử nghiệm nhận dạng tiếng nói tiếng Việt cho người khuyết tật giọng nói bằng phương pháp học sâu



Phan Xuan Phuc
Mentor: TS. Nguyen Hong Quang

Hanoi University of Science and Technology

Ha Noi, January 14, 2020

Nội dung

Source code: https://gitlab.com/fakerphan/masr2
Documents: https://gitlab.com/fakerphan/masr2/tree/master/reports

Nội dung

Source code: https://gitlab.com/fakerphan/masr2
Documents: https://gitlab.com/fakerphan/masr2/tree/master/reports

Giới thiệu bài toán

Với những người bị khuyết tật giọng nói, làm sao để mọi người có thể hiểu những gì họ nói?

Giới thiệu bài toán

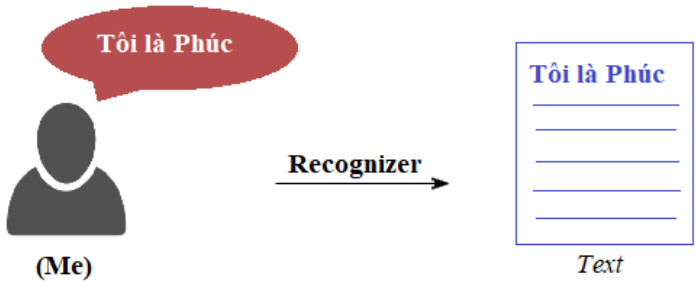
Với những người bị khuyết tật giọng nói, làm sao để mọi người có thể hiểu những gì họ nói?

- Nhận dạng tiếng nói của người bị khuyết tật giọng nói
  $\rightarrow$ "Speech Recognition"

Giới thiệu bài toán

Với những người bị khuyết tật giọng nói, làm sao để mọi người có thể hiểu những gì họ nói?

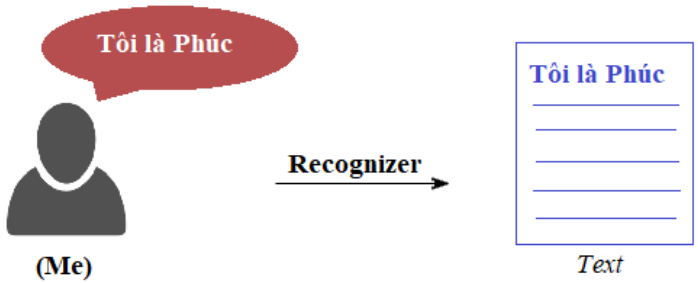
- Nhận dạng tiếng nói của người bị khuyết tật giọng nói
  → "Speech Recognition"
  - Biến đổi tiếng nói (Speech) → văn bản (Text)



Giới thiệu bài toán

Với những người bị khuyết tật giọng nói, làm sao để mọi người có thể hiểu những gì họ nói?

- Nhận dạng tiếng nói của người bị khuyết tật giọng nói
  → "Speech Recognition"
  - Biến đổi tiếng nói (Speech) → văn bản (Text)

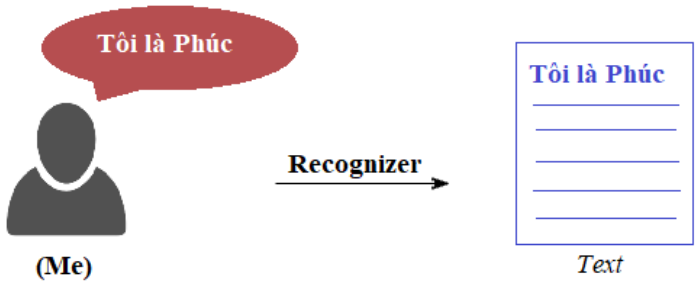


- Tập trung vào tiếng nói của người bị khuyết tật giọng nói

Giới thiệu bài toán

Với những người bị khuyết tật giọng nói, làm sao để mọi người có thể hiểu những gì họ nói?

- Nhận dạng tiếng nói của người bị khuyết tật giọng nói
  → "Speech Recognition"
  - Biến đổi tiếng nói (Speech) → văn bản (Text)

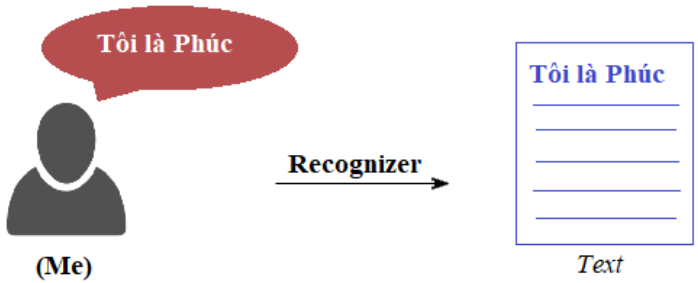


- Tập trung vào tiếng nói của người bị khuyết tật giọng nói
- Tập trung với ngôn ngữ Tiếng Việt

Giới thiệu bài toán

Với những người bị khuyết tật giọng nói, làm sao để mọi người có thể hiểu những gì họ nói?

- Nhận dạng tiếng nói của người bị khuyết tật giọng nói
  → "Speech Recognition"
  - Biến đổi tiếng nói (Speech) → văn bản (Text)



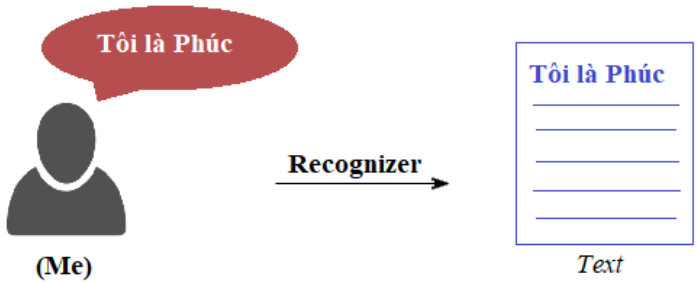
- Tập trung vào tiếng nói của người bị khuyết tật giọng nói
- Tập trung với ngôn ngữ Tiếng Việt

Input: Một âm thanh tiếng nói (Speech) của người khuyết tật giọng nói

Giới thiệu bài toán

Với những người bị khuyết tật giọng nói, làm sao để mọi người có thể hiểu những gì họ nói?

- Nhận dạng tiếng nói của người bị khuyết tật giọng nói
  → "Speech Recognition"
  - Biến đổi tiếng nói (Speech) → văn bản (Text)



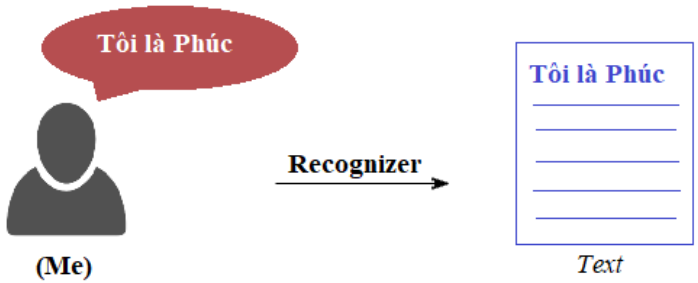
- Tập trung vào tiếng nói của người bị khuyết tật giọng nói
- Tập trung với ngôn ngữ Tiếng Việt

Input: Một âm thanh tiếng nói (Speech) của người khuyết tật giọng nói
Output: Văn bản đầu ra (Text) tương ứng của âm thanh tiếng nói đó

Giới thiệu bài toán

Với những người bị khuyết tật giọng nói, làm sao để mọi người có thể hiểu những gì họ nói?

- Nhận dạng tiếng nói của người bị khuyết tật giọng nói
  → "Speech Recognition"
  - Biến đổi tiếng nói (Speech) → văn bản (Text)



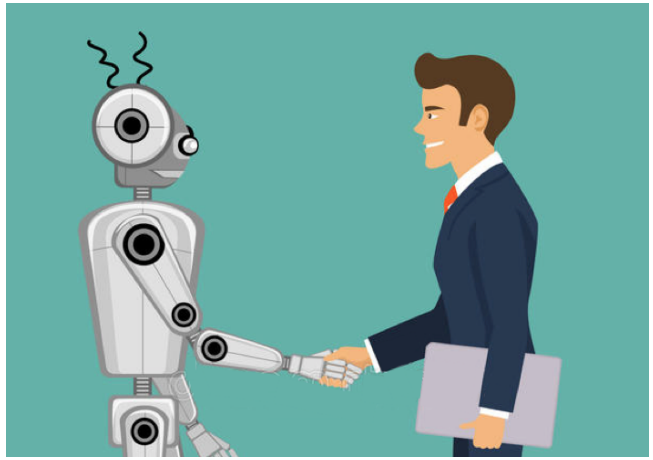
- Tập trung vào tiếng nói của người bị khuyết tật giọng nói
- Tập trung với ngôn ngữ Tiếng Việt

Input: Một âm thanh tiếng nói (Speech) của người khuyết tật giọng nói
Output: Văn bản đầu ra (Text) tương ứng của âm thanh tiếng nói đó

Motivation

Motivation

Motivation



Nội dung

Source code: https://gitlab.com/fakerphan/masr2
Documents: https://gitlab.com/fakerphan/masr2/tree/master/reports

Dataset

Ngữ cảnh:

[1]Wiki, https://www.ezglot.com/most-frequently-used-words.php?l=vie&s=wp-freq.

Dataset

Ngữ cảnh:

- Chỉ có khoảng 3 tháng để thực hiện

[1]Wiki, https://www.ezglot.com/most-frequently-used-words.php?l=vie&s=wp-freq.

Dataset

Ngữ cảnh:

- Chỉ có khoảng 3 tháng để thực hiện
- Không có dữ liệu có sẵn của người khuyết tật giọng nói

[1]Wiki, https://www.ezglot.com/most-frequently-used-words.php?l=vie&s=wp-freq.

Dataset

Ngữ cảnh:

- Chỉ có khoảng 3 tháng để thực hiện
- Không có dữ liệu có sẵn của người khuyết tật giọng nói

Giải pháp:

[1]Wiki, https://www.ezglot.com/most-frequently-used-words.php?l=vie&s=wp-freq.

Dataset

Ngữ cảnh:

- Chỉ có khoảng 3 tháng để thực hiện
- Không có dữ liệu có sẵn của người khuyết tật giọng nói

Giải pháp:

- Thu thập dữ liệu bằng cách ghi âm 1600 từ vựng thông dụng nhất trong Tiếng Việt [1] sao cho bao phủ toàn bộ âm vị và chữ cái trong Tiếng Việt,

[1] Wiki, https://www.ezglot.com/most-frequently-used-words.php?l=vie&s=wp-freq.

Dataset

Ngữ cảnh:

- Chỉ có khoảng 3 tháng để thực hiện
- Không có dữ liệu có sẵn của người khuyết tật giọng nói

Giải pháp:

- Thu thập dữ liệu bằng cách ghi âm 1600 từ vựng thông dụng nhất trong Tiếng Việt [1] sao cho bao phủ toàn bộ âm vị và chữ cái trong Tiếng Việt,
  - Bao gồm: 134 âm vị, 89 chữ cái (bao gồm cả thanh điệu),

[1] Wiki, https://www.ezglot.com/most-frequently-used-words.php?l=vie&s=wp-freq.

Dataset

Ngữ cảnh:

- Chỉ có khoảng 3 tháng để thực hiện
- Không có dữ liệu có sẵn của người khuyết tật giọng nói

Giải pháp:

- Thu thập dữ liệu bằng cách ghi âm 1600 từ vựng thông dụng nhất trong Tiếng Việt [1] sao cho bao phủ toàn bộ âm vị và chữ cái trong Tiếng Việt,
  - Bao gồm: 134 âm vị, 89 chữ cái (bao gồm cả thanh điệu),
  - Bao gồm: 1546 từ đơn, 54 từ ghép,

[1] Wiki, https://www.ezglot.com/most-frequently-used-words.php?l=vie&s=wp-freq.

Dataset

Ngữ cảnh:

- Chỉ có khoảng 3 tháng để thực hiện
- Không có dữ liệu có sẵn của người khuyết tật giọng nói

Giải pháp:

- Thu thập dữ liệu bằng cách ghi âm 1600 từ vựng thông dụng nhất trong Tiếng Việt [1] sao cho bao phủ toàn bộ âm vị và chữ cái trong Tiếng Việt,
  - Bao gồm: 134 âm vị, 89 chữ cái (bao gồm cả thanh điệu),
  - Bao gồm: 1546 từ đơn, 54 từ ghép,
- Có 3 người nói (speaker):

[1] Wiki, https://www.ezglot.com/most-frequently-used-words.php?l=vie&s=wp-freq.

Dataset

Ngữ cảnh:

- Chỉ có khoảng 3 tháng để thực hiện
- Không có dữ liệu có sẵn của người khuyết tật giọng nói

Giải pháp:

- Thu thập dữ liệu bằng cách ghi âm 1600 từ vựng thông dụng nhất trong Tiếng Việt [1] sao cho bao phủ toàn bộ âm vị và chữ cái trong Tiếng Việt,
  - Bao gồm: 134 âm vị, 89 chữ cái (bao gồm cả thanh điệu),
  - Bao gồm: 1546 từ đơn, 54 từ ghép,
- Có 3 người nói (speaker):
  - 3 Man (Bố, Anh Trai và Tôi),
  - 0 Woman,

[1] Wiki, https://www.ezglot.com/most-frequently-used-words.php?l=vie&s=wp-freq.

Dataset

Ngữ cảnh:

- Chỉ có khoảng 3 tháng để thực hiện
- Không có dữ liệu có sẵn của người khuyết tật giọng nói

Giải pháp:

- Thu thập dữ liệu bằng cách ghi âm 1600 từ vựng thông dụng nhất trong Tiếng Việt [1] sao cho bao phủ toàn bộ âm vị và chữ cái trong Tiếng Việt,
  - Bao gồm: 134 âm vị, 89 chữ cái (bao gồm cả thanh điệu),
  - Bao gồm: 1546 từ đơn, 54 từ ghép,
- Có 3 người nói (speaker):
  - 3 Man (Bố, Anh Trai và Tôi),
  - 0 Woman,
- Mỗi từ vựng được ghi âm bởi 3 người và ở 4 thời điểm nói khác nhau,

[1] Wiki, https://www.ezglot.com/most-frequently-used-words.php?l=vie&s=wp-freq.

Dataset

Ngữ cảnh:

- Chỉ có khoảng 3 tháng để thực hiện
- Không có dữ liệu có sẵn của người khuyết tật giọng nói

Giải pháp:

- Thu thập dữ liệu bằng cách ghi âm 1600 từ vựng thông dụng nhất trong Tiếng Việt [1] sao cho bao phủ toàn bộ âm vị và chữ cái trong Tiếng Việt,
  - Bao gồm: 134 âm vị, 89 chữ cái (bao gồm cả thanh điệu),
  - Bao gồm: 1546 từ đơn, 54 từ ghép,
- Có 3 người nói (speaker):
  - 3 Man (Bố, Anh Trai và Tôi),
  - 0 Woman,
- Mỗi từ vựng được ghi âm bởi 3 người và ở 4 thời điểm nói khác nhau,
- Training: 3*(1,600 ghi âm/ 1 người) ở 3 thời điểm nói,

[1] Wiki, https://www.ezglot.com/most-frequently-used-words.php?l=vie&s=wp-freq.

Dataset

Ngữ cảnh:

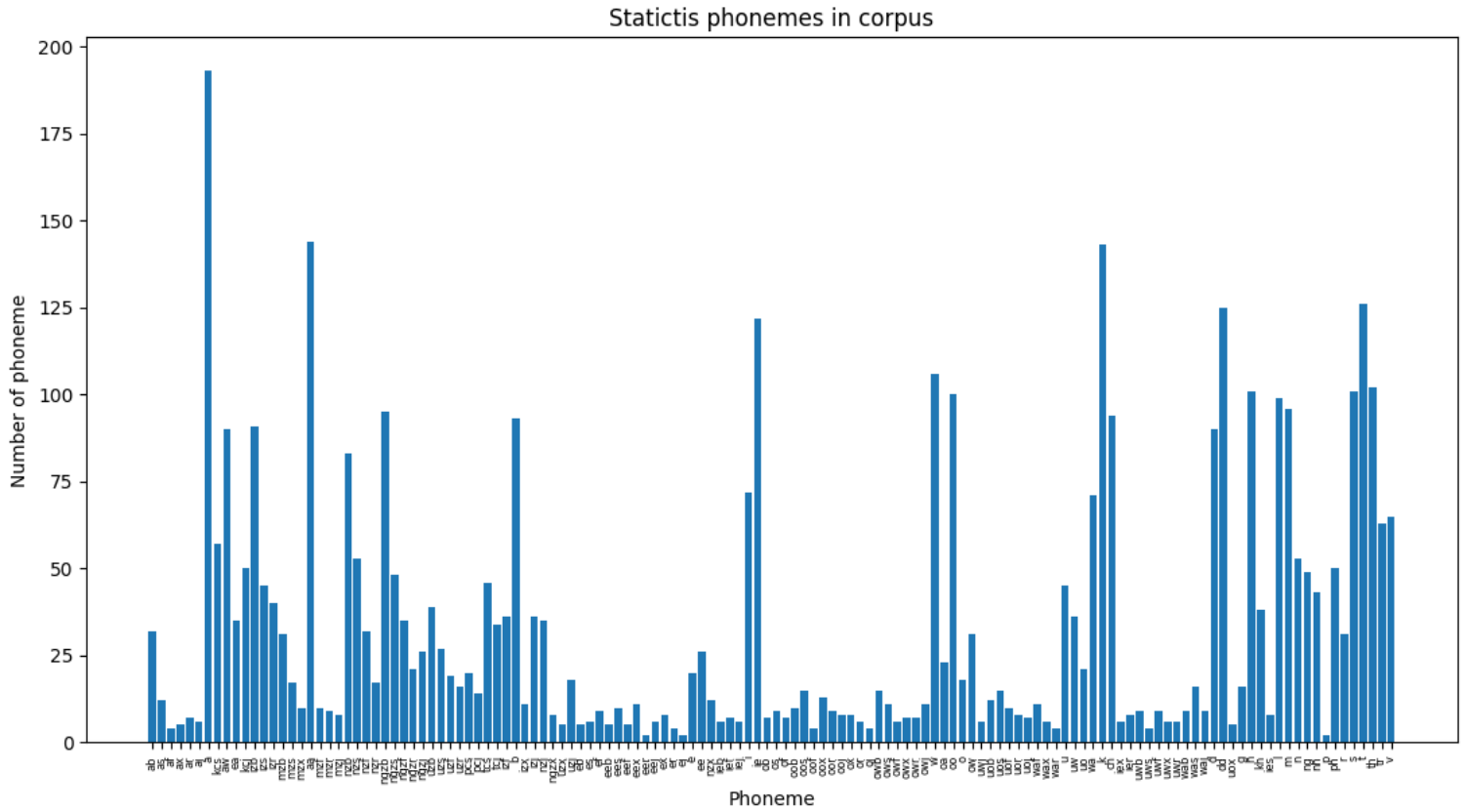
- Chỉ có khoảng 3 tháng để thực hiện
- Không có dữ liệu có sẵn của người khuyết tật giọng nói

Giải pháp:

- Thu thập dữ liệu bằng cách ghi âm 1600 từ vựng thông dụng nhất trong Tiếng Việt [1] sao cho bao phủ toàn bộ âm vị và chữ cái trong Tiếng Việt,
  - Bao gồm: 134 âm vị, 89 chữ cái (bao gồm cả thanh điệu),
  - Bao gồm: 1546 từ đơn, 54 từ ghép,
- Có 3 người nói (speaker):
  - 3 Man (Bố, Anh Trai và Tôi),
  - 0 Woman,
- Mỗi từ vựng được ghi âm bởi 3 người và ở 4 thời điểm nói khác nhau,
- Training: 3*(1,600 ghi âm/ 1 người) ở 3 thời điểm nói,
- Evaluation: 3*(1,600 ghi âm/ 1 người) ở thời điểm nói còn lại.

[1] Wiki, https://www.ezglot.com/most-frequently-used-words.php?l=vie&s=wp-freq.

Dataset



Hình 1: Phân bố 134 âm vị tiếng Việt trong bộ dữ liệu thu thập gồm 1,600 từ vựng → *Mất cân bằng (Imbalacing)*

Nội dung

Source code: https://gitlab.com/fakerphan/masr2
Documents: https://gitlab.com/fakerphan/masr2/tree/master/reports

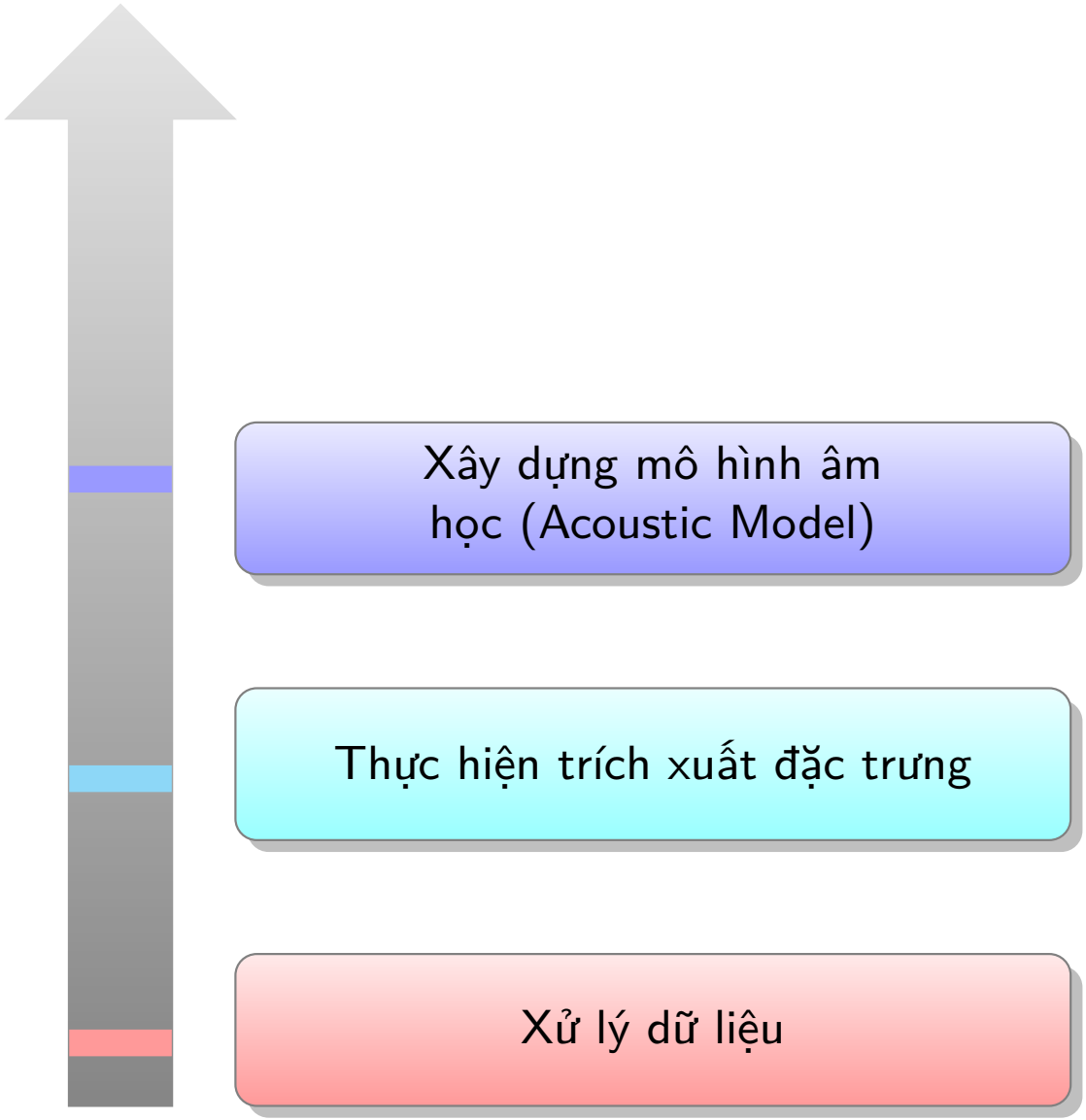
Định hướng giải pháp

Định hướng giải pháp

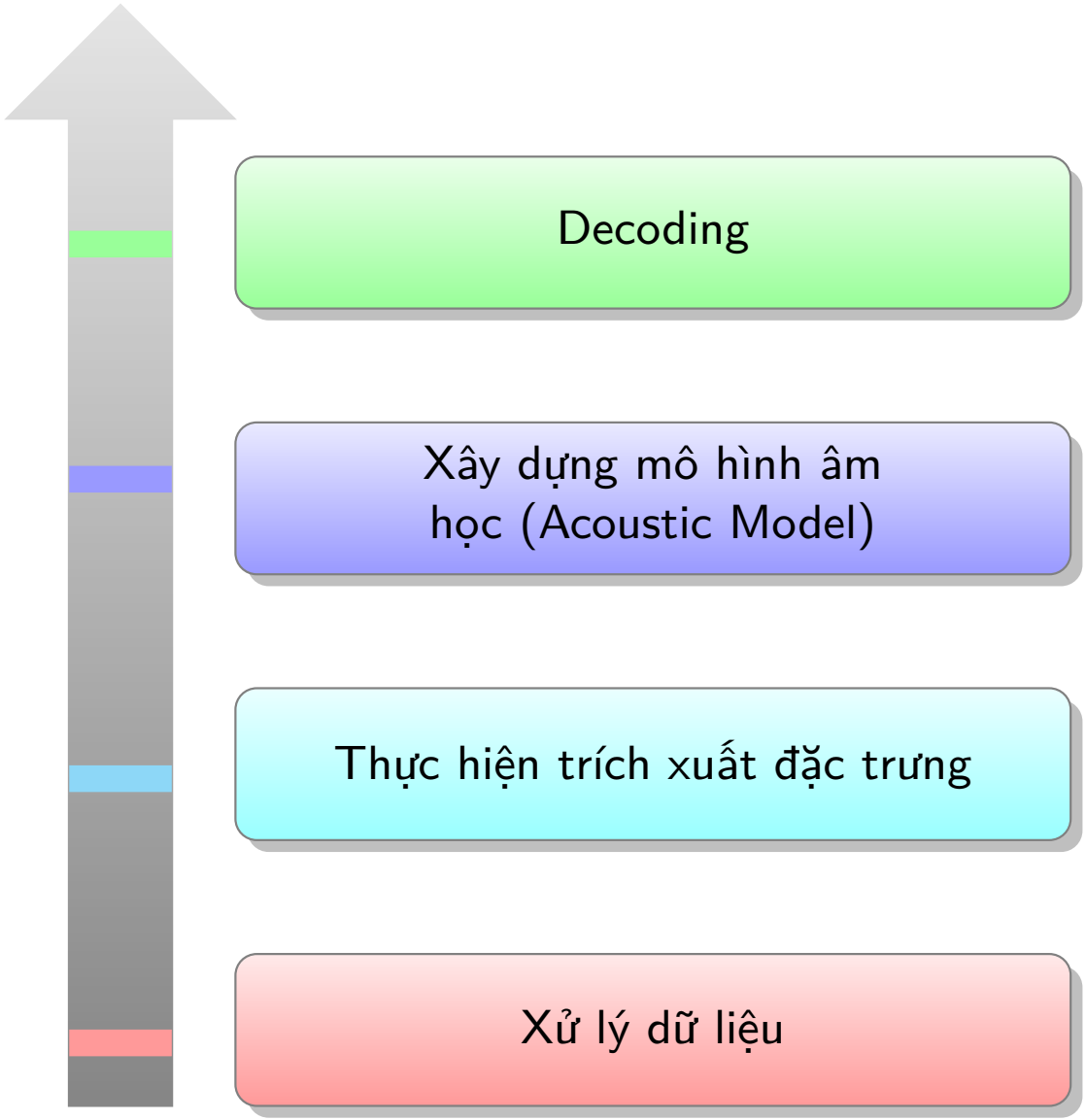
Xử lý dữ liệu

Định hướng giải pháp

Định hướng giải pháp



Định hướng giải pháp



Nội dung

Source code: https://gitlab.com/fakerphan/masr2
Documents: https://gitlab.com/fakerphan/masr2/tree/master/reports

Xử lý dữ liệu

- Dữ liệu thu âm được lưu trữ dưới định dạng .mp3 hoặc .m4a

[2] https://librosa.github.io/librosa/

Xử lý dữ liệu

- Dữ liệu thu âm được lưu trữ dưới định dạng .mp3 hoặc .m4a
- Chuyển đổi dữ liệu âm thanh thành định dạng .wav, mono-chanel, sampling rate=16kHz:

[2] https://librosa.github.io/librosa/

Xử lý dữ liệu

- Dữ liệu thu âm được lưu trữ dưới định dạng .mp3 hoặc .m4a
- Chuyển đổi dữ liệu âm thanh thành định dạng .wav, mono-chanel, sampling rate=16kHz:

Command line (Linux):

```
for f in *.m4a; do ffmpeg -i "$f" -acodec pcm_s16le -ac 1 -ar 16000
"$f/%m4a/wav"; done
```

[2] https://librosa.github.io/librosa/

Xử lý dữ liệu

- Dữ liệu thu âm được lưu trữ dưới định dạng .mp3 hoặc .m4a
- Chuyển đổi dữ liệu âm thanh thành định dạng .wav, mono-chanel, sampling rate=16kHz:

Command line (Linux):

```
for f in *.m4a; do ffmpeg -i "$f" -acodec pcm_s16le -ac 1 -ar 16000
"$f/%m4a/wav"; done
```

- Sử dụng thư viện Librosa [2] để thực hiện đọc file âm thanh từ file .wav

[2] https://librosa.github.io/librosa/

Xử lý dữ liệu

Cân bằng dữ liệu tập huấn luyện

Xử lý dữ liệu

Cân bằng dữ liệu tập huấn luyện

- Ta xét một ngưỡng cân bằng N (N=100)

Xử lý dữ liệu

Cân bằng dữ liệu tập huấn luyện

- Ta xét một ngưỡng cân bằng N (N=100)
- Với những âm vị có số lần xuất hiện $<$ N $\rightarrow$ Cluster A, ngược lại $\rightarrow$ Cluster B

Xử lý dữ liệu

Cân bằng dữ liệu tập huấn luyện

- Ta xét một ngưỡng cân bằng N (N=100)
- Với những âm vị có số lần xuất hiện $<$ N $\rightarrow$ Cluster A,
  ngược lại $\rightarrow$ Cluster B
- Với những âm vị trong Cluster A, ta thực hiện việc tăng cường số lần xuất hiện của nó bằng cách:

Xử lý dữ liệu

Cân bằng dữ liệu tập huấn luyện

- Ta xét một ngưỡng cân bằng N (N=100)
- Với những âm vị có số lần xuất hiện $<$ N $\rightarrow$ Cluster A,
  ngược lại $\rightarrow$ Cluster B
- Với những âm vị trong Cluster A, ta thực hiện việc tăng cường số lần xuất hiện của nó bằng cách:
  - Tìm kiếm từ (word) tương ứng có chứa âm vị đó, sao cho: Các âm vị được phân tích bởi từ đó có nhiều nhất có thể trong Cluster A và ít tồn tại nhất có thể trong Cluster B

Xử lý dữ liệu

Cân bằng dữ liệu tập huấn luyện

- Ta xét một ngưỡng cân bằng N (N=100)
- Với những âm vị có số lần xuất hiện $<$ N $\rightarrow$ Cluster A,
  ngược lại $\rightarrow$ Cluster B
- Với những âm vị trong Cluster A, ta thực hiện việc tăng cường số lần xuất hiện của nó bằng cách:
  - Tìm kiếm từ (word) tương ứng có chứa âm vị đó, sao cho: Các âm vị được phân tích bởi từ đó có nhiều nhất có thể trong Cluster A và ít tồn tại nhất có thể trong Cluster B
- Thuật toán dừng lại cho đến khi tất cả các âm vị vượt qua ngưỡng cân bằng $\rightarrow$ Bộ dữ liệu huấn luyện sau đó cân bằng âm vị

Xử lý dữ liệu

Cân bằng dữ liệu tập huấn luyện

- Ta xét một ngưỡng cân bằng N (N=100)
- Với những âm vị có số lần xuất hiện $<$ N $\rightarrow$ Cluster A, ngược lại $\rightarrow$ Cluster B
- Với những âm vị trong Cluster A, ta thực hiện việc tăng cường số lần xuất hiện của nó bằng cách:
  - Tìm kiếm từ (word) tương ứng có chứa âm vị đó, sao cho: Các âm vị được phân tích bởi từ đó có nhiều nhất có thể trong Cluster A và ít tồn tại nhất có thể trong Cluster B
- Thuật toán dừng lại cho đến khi tất cả các âm vị vượt qua ngưỡng cân bằng $\rightarrow$ Bộ dữ liệu huấn luyện sau đó cân bằng âm vị

Input: File dữ liệu huấn luyện và file chuyển đổi âm vị các từ trong tiếng Việt

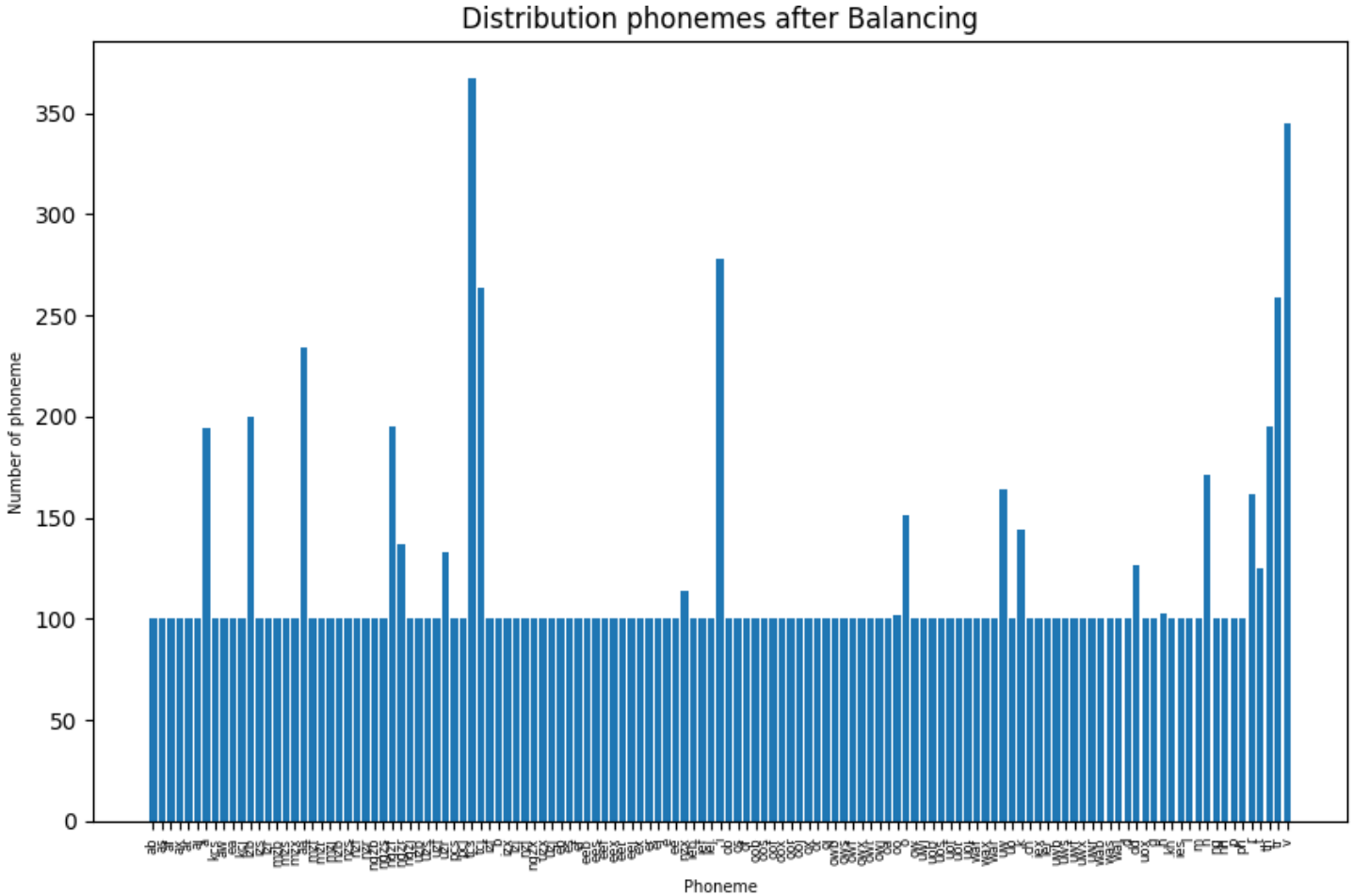
Xử lý dữ liệu

Cân bằng dữ liệu tập huấn luyện

- Ta xét một ngưỡng cân bằng N (N=100)
- Với những âm vị có số lần xuất hiện $<$ N $\rightarrow$ Cluster A, ngược lại $\rightarrow$ Cluster B
- Với những âm vị trong Cluster A, ta thực hiện việc tăng cường số lần xuất hiện của nó bằng cách:
  - Tìm kiếm từ (word) tương ứng có chứa âm vị đó, sao cho: Các âm vị được phân tích bởi từ đó có nhiều nhất có thể trong Cluster A và ít tồn tại nhất có thể trong Cluster B
- Thuật toán dừng lại cho đến khi tất cả các âm vị vượt qua ngưỡng cân bằng $\rightarrow$ Bộ dữ liệu huấn luyện sau đó cân bằng âm vị

Input: File dữ liệu huấn luyện và file chuyển đổi âm vị các từ trong tiếng Việt
Output: File dữ liệu huấn luyện sau khi cân bằng âm vị

Xử lý dữ liệu



Hình 2: Phân bố 134 âm vị tiếng Việt trong bộ dữ liệu sau khi được cân bằng

Trích xuất đặc trưng

Trích xuất đặc trưng

- Sử dụng phương pháp trích xuất đặc trưng MFCCs (Mel Frequency Cepstral Coefficients).

Trích xuất đặc trưng

- Sử dụng phương pháp trích xuất đặc trưng MFCCs (Mel Frequency Cepstral Coefficients).
  - Mỗi frame có độ dài là 25ms, khoảng cách giữa 2 frame liên tiếp là 10ms.

Trích xuất đặc trưng

- Sử dụng phương pháp trích xuất đặc trưng MFCCs (Mel Frequency Cepstral Coefficients).
  - Mỗi frame có độ dài là 25ms, khoảng cách giữa 2 frame liên tiếp là 10ms.
- Mỗi frame trích xuất được một vector đặc trưng-13 chiều tương ứng.

Trích xuất đặc trưng

- Sử dụng phương pháp trích xuất đặc trưng MFCCs (Mel Frequency Cepstral Coefficients).
  - Mỗi frame có độ dài là 25ms, khoảng cách giữa 2 frame liên tiếp là 10ms.
- Mỗi frame trích xuất được một vector đặc trưng-13 chiều tương ứng.
- "Rescale" lại đặc trưng thu được bằng Standardization:

$$x^{'} = \frac{x - \mu}{\sigma + \epsilon}$$

Trích xuất đặc trưng

- Sử dụng phương pháp trích xuất đặc trưng MFCCs (Mel Frequency Cepstral Coefficients).
  - Mỗi frame có độ dài là 25ms, khoảng cách giữa 2 frame liên tiếp là 10ms.
- Mỗi frame trích xuất được một vector đặc trưng-13 chiều tương ứng.
- "Rescale" lại đặc trưng thu được bằng Standardization:

$$x^{'} = \frac{x - \mu}{\sigma + \epsilon}$$

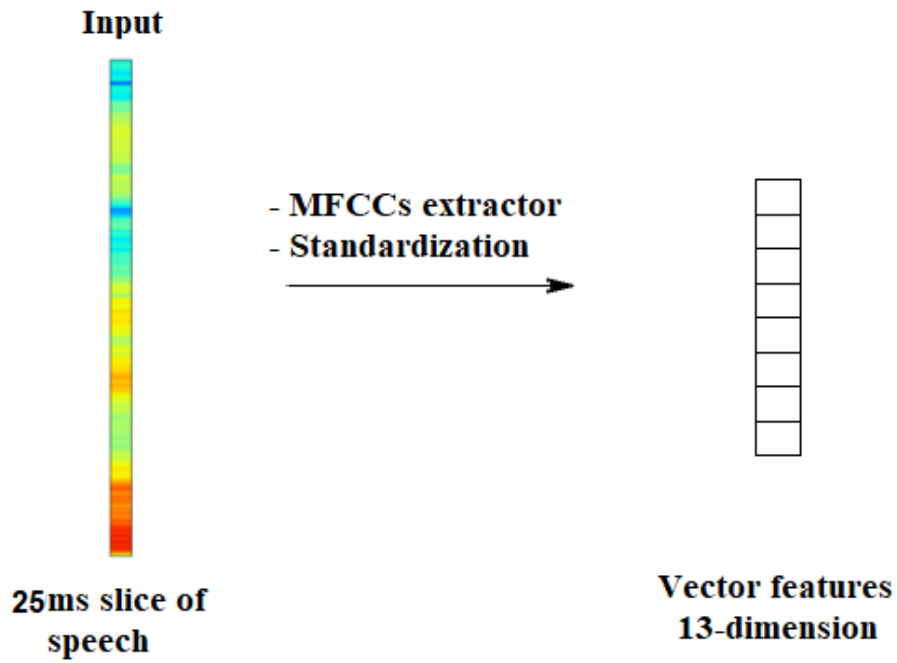
  - μ và σ được tính dựa trên K (K=200) mẫu trong tập dữ liệu huấn luyện,
  - $\epsilon = 1e - 14$

Trích xuất đặc trưng

- Sử dụng phương pháp trích xuất đặc trưng MFCCs (Mel Frequency Cepstral Coefficients).
  - Mỗi frame có độ dài là 25ms, khoảng cách giữa 2 frame liên tiếp là 10ms.
- Mỗi frame trích xuất được một vector đặc trưng-13 chiều tương ứng.
- "Rescale" lại đặc trưng thu được bằng Standardization:

$$x^{'} = \frac{x - \mu}{\sigma + \epsilon}$$

  - μ và σ được tính dựa trên K (K=200) mẫu trong tập dữ liệu huấn luyện,
  - $\epsilon = 1e - 14$



Mô hình âm học

Mô hình âm học

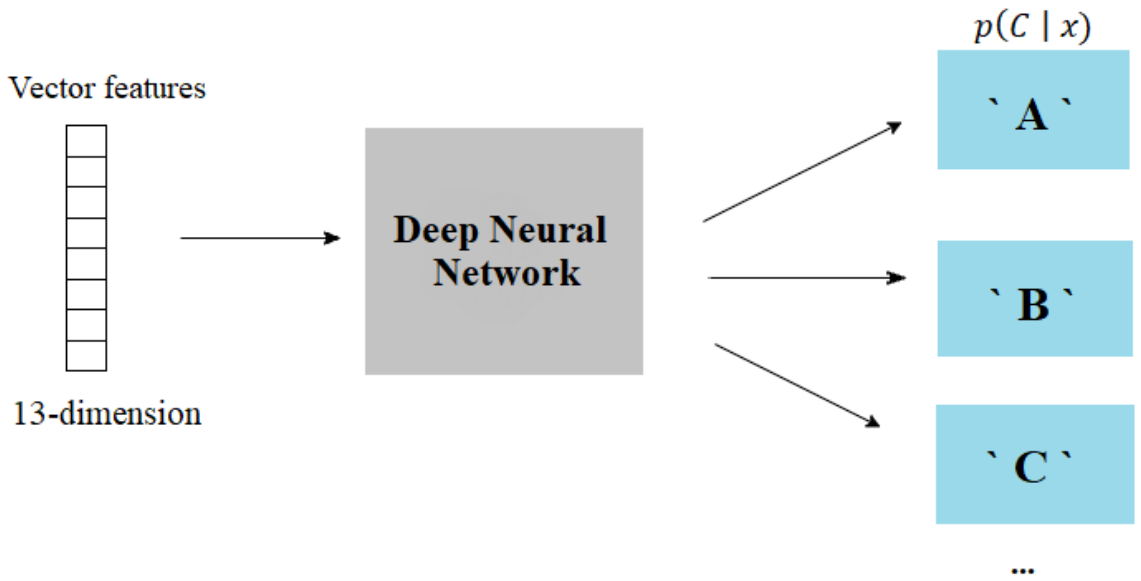
- Mỗi vector đặc trưng thu được sau đó đưa qua mô hình âm học. Đầu ra của nó là một phân phối xác suất trên các class là bộ ký tự C.

Mô hình âm học

- Mỗi vector đặc trưng thu được sau đó đưa qua mô hình âm học. Đầu ra của nó là một phân phối xác suất trên các class là bộ ký tự C.
- Bộ ký tự C bao gồm 89 chữ cái (gồm cả thanh điệu) trong Tiếng Việt, 1 ký tự *"Space"*. Để tổng quát hơn, xét cả 4 ký tự trong tiếng Anh $\{f, j, w, z\}$ và ký tự *"blank"* $\rightarrow$ Output $=$ 95 units

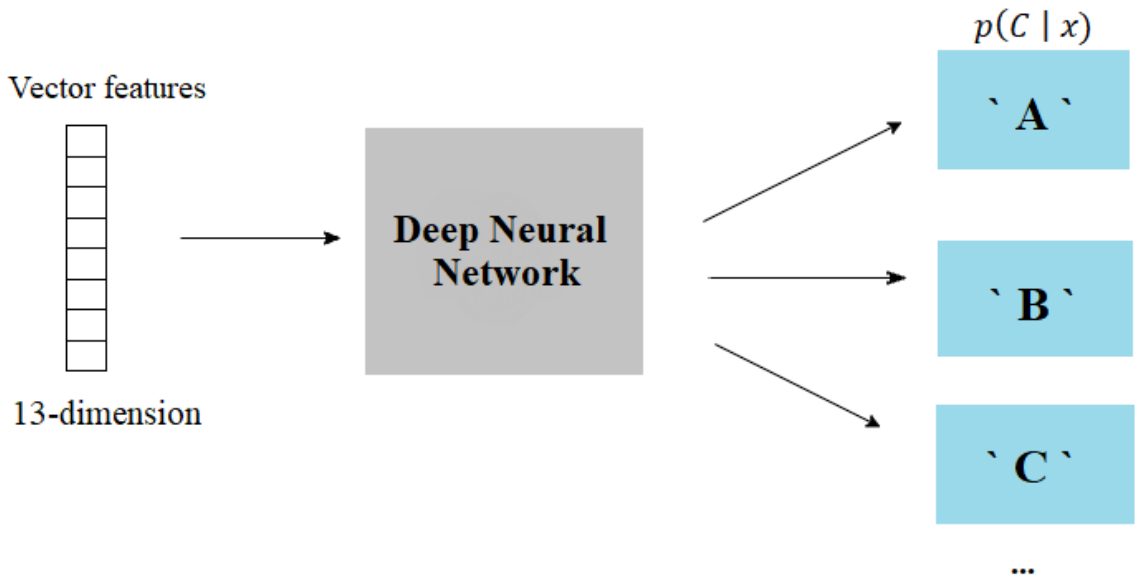
Mô hình âm học

- Mỗi vector đặc trưng thu được sau đó đưa qua mô hình âm học. Đầu ra của nó là một phân phối xác suất trên các class là bộ ký tự C.
- Bộ ký tự C bao gồm 89 chữ cái (gồm cả thanh điệu) trong Tiếng Việt, 1 ký tự *"Space"*. Để tổng quát hơn, xét cả 4 ký tự trong tiếng Anh $\{f, j, w, z\}$ và ký tự *"blank"* $\rightarrow$ Output $=$ 95 units



Mô hình âm học

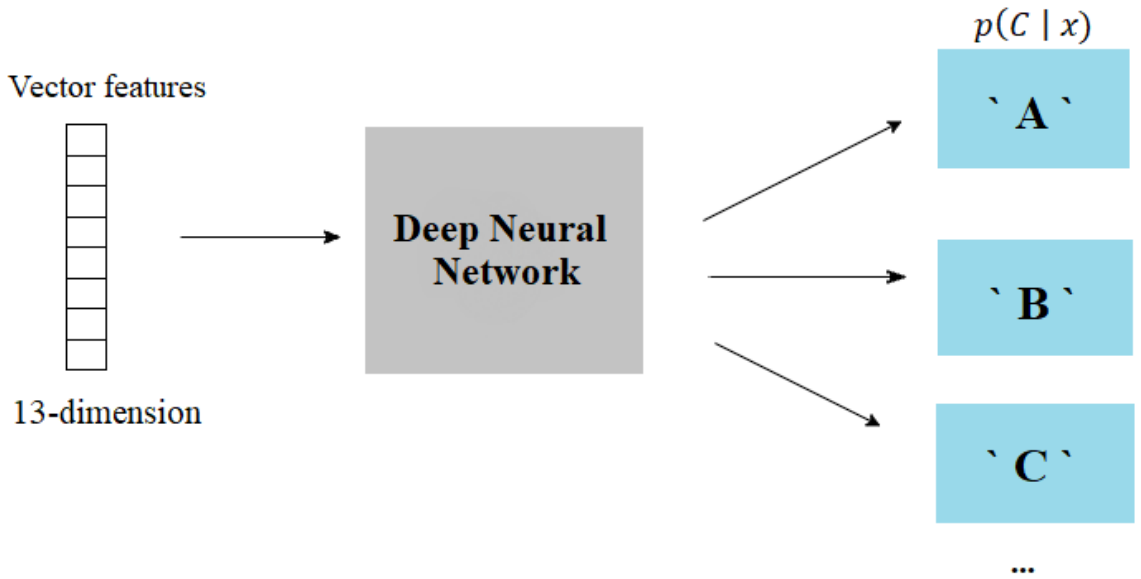
- Mỗi vector đặc trưng thu được sau đó đưa qua mô hình âm học. Đầu ra của nó là một phân phối xác suất trên các class là bộ ký tự C.
- Bộ ký tự C bao gồm 89 chữ cái (gồm cả thanh điệu) trong Tiếng Việt, 1 ký tự *"Space"*. Để tổng quát hơn, xét cả 4 ký tự trong tiếng Anh $\{f, j, w, z\}$ và ký tự *"blank"* $\rightarrow$ Output $=$ 95 units



Input: Một vector đặc trưng biểu diễn cho một frame của tín hiệu tiếng nói

Mô hình âm học

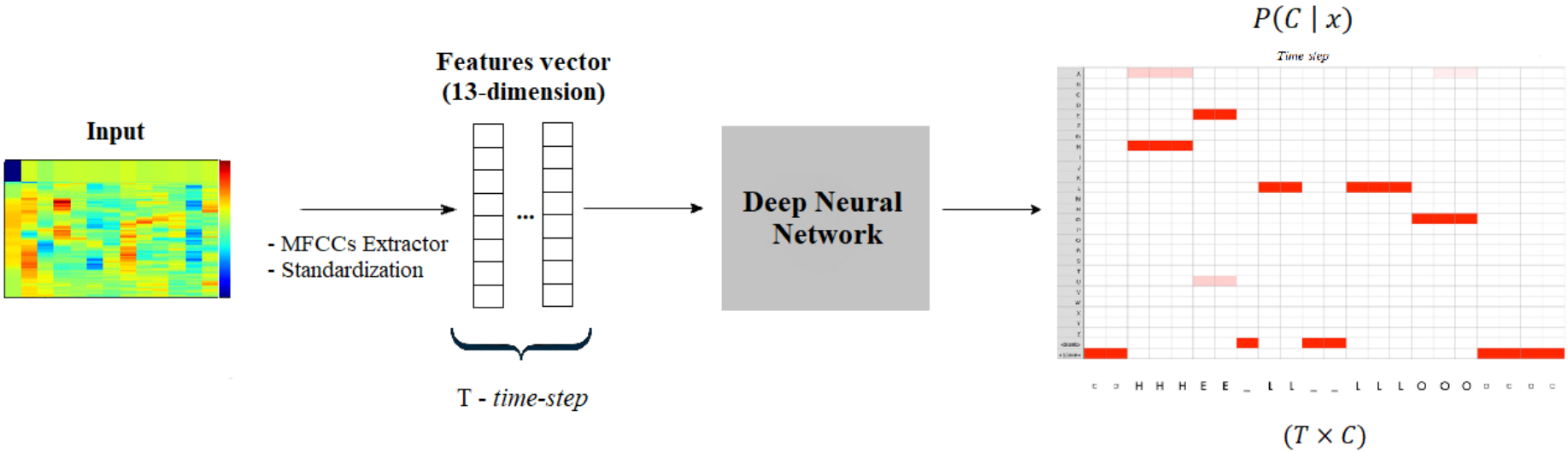
- Mỗi vector đặc trưng thu được sau đó đưa qua mô hình âm học. Đầu ra của nó là một phân phối xác suất trên các class là bộ ký tự C.
- Bộ ký tự C bao gồm 89 chữ cái (gồm cả thanh điệu) trong Tiếng Việt, 1 ký tự *"Space"*. Để tổng quát hơn, xét cả 4 ký tự trong tiếng Anh $\{f, j, w, z\}$ và ký tự *"blank"* $\rightarrow$ Output $=$ 95 units



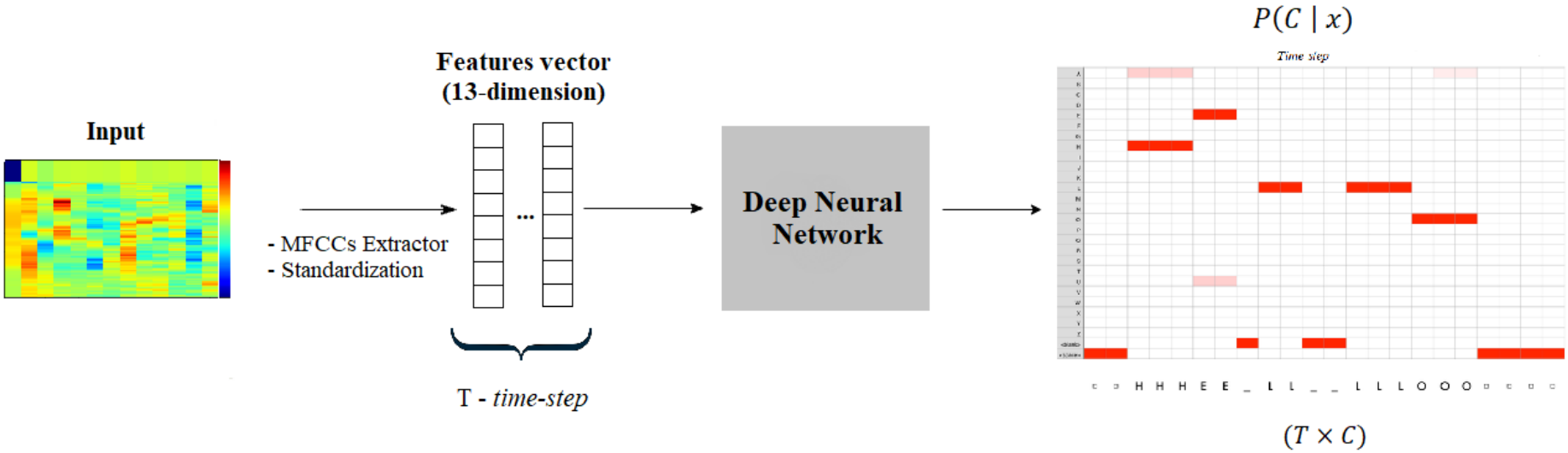
Input: Một vector đặc trưng biểu diễn cho một frame của tín hiệu tiếng nói
Output: Một phân phối xác suất trên bộ ký tự nhận dạng

Mô hình âm học

Mô hình âm học

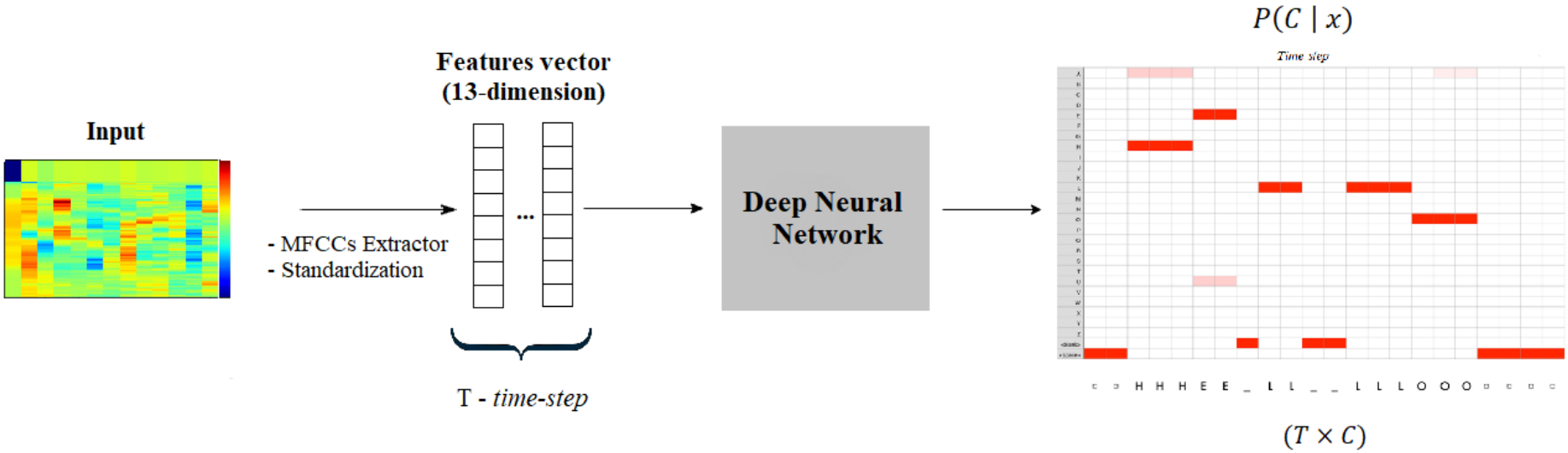


Mô hình âm học



Input: Một tín hiệu tiếng nói (speech) được phát ra bởi người khuyết tật giọng nói

Mô hình âm học



Input: Một tín hiệu tiếng nói (speech) được phát ra bởi người khuyết tật giọng nói
Output: Một ma trận phân phối xác suất trên bộ ký tự C gồm 95 ký tự.

Mô hình âm học - Kiến trúc

[3] Awni Hannun, "Deep Speech: Scaling up end-to-end speech recognition". *In: arXiv:1412.5567 (2014).*

[4] Dario Amodei, "Deep Speech 2: End-to-End Speech Recognition in English and Mandarin". *In: arXiv:1512.02595 (2015).*

Mô hình âm học - Kiến trúc

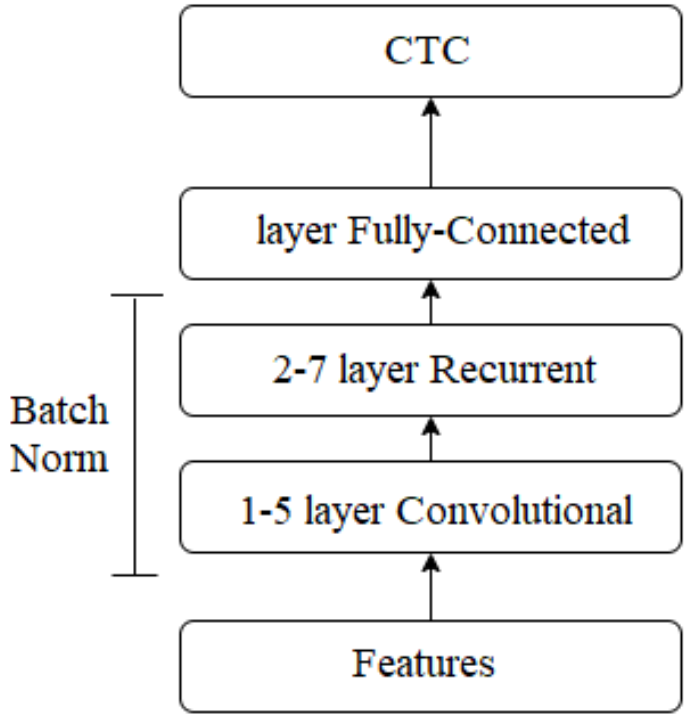
- Tham khảo kiến trúc mô hình của Deep Speech 1 [3] và Deep Speech 2 [4]

[3] Awni Hannun, "Deep Speech: Scaling up end-to-end speech recognition". *In: arXiv:1412.5567 (2014).*

[4] Dario Amodei, "Deep Speech 2: End-to-End Speech Recognition in English and Mandarin". *In: arXiv:1512.02595 (2015).*

Mô hình âm học - Kiến trúc

- Tham khảo kiến trúc mô hình của Deep Speech 1 [3] và Deep Speech 2 [4]
- Kiến trúc chung cho các thử nghiệm mô hình âm học

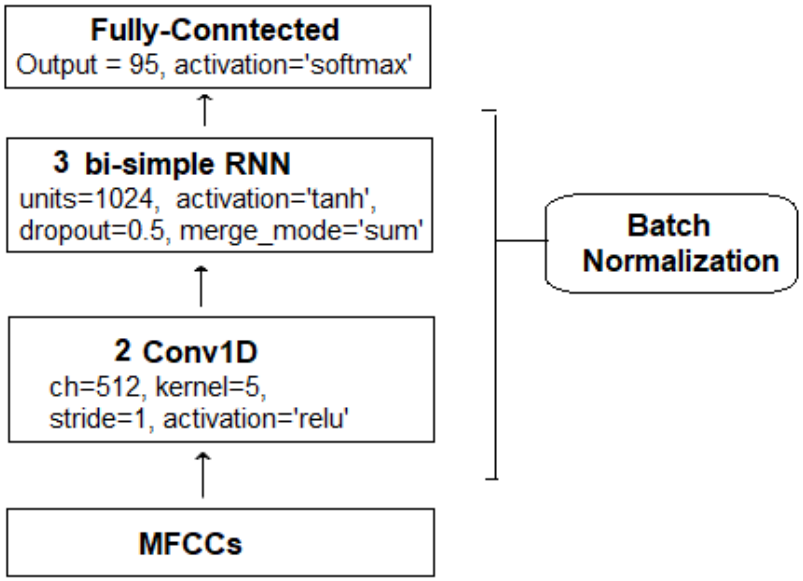


[3] Awni Hannun, "Deep Speech: Scaling up end-to-end speech recognition". *In: arXiv:1412.5567 (2014).*

[4] Dario Amodei, "Deep Speech 2: End-to-End Speech Recognition in English and Mandarin". *In: arXiv:1512.02595 (2015).*

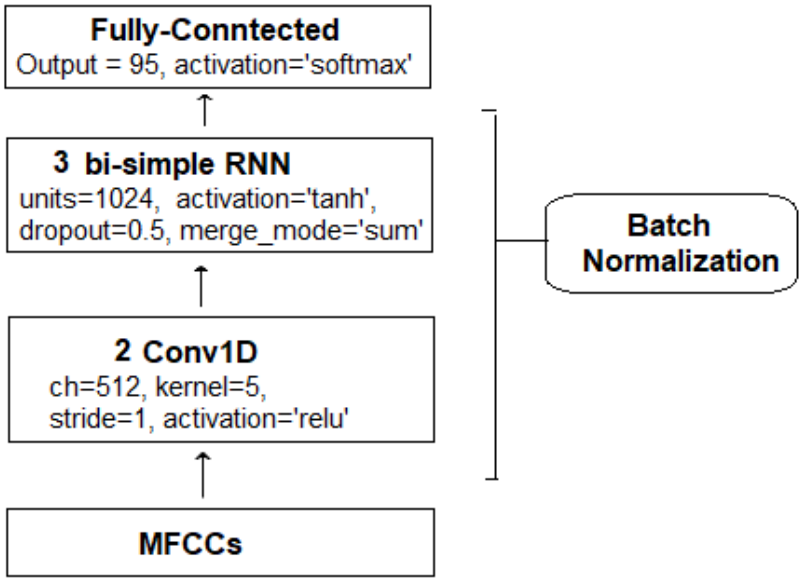
Mô hình âm học - Kiến trúc

Mô hình âm học - Kiến trúc

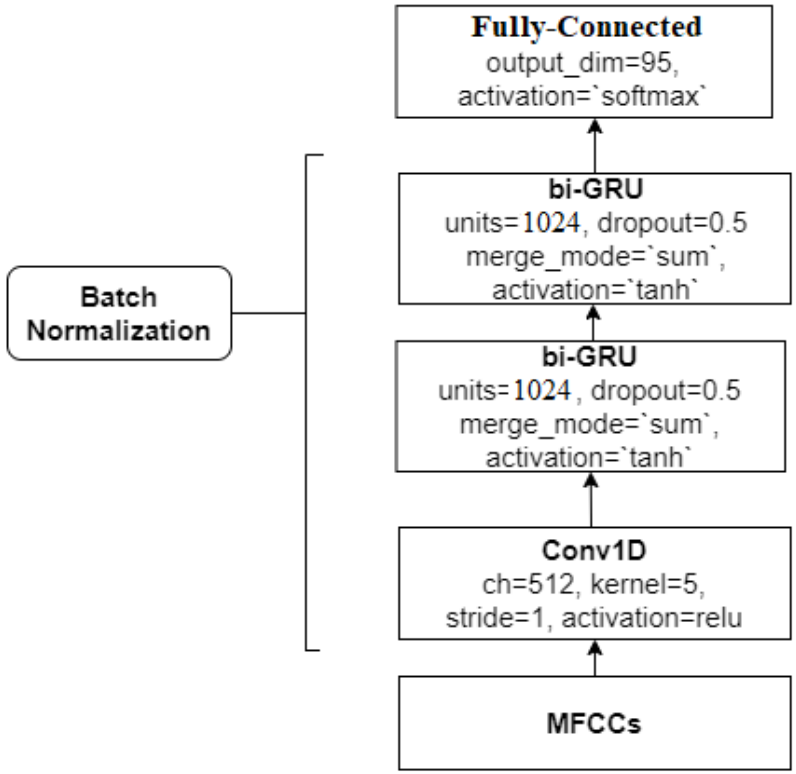


Hình 3: Model 1

Mô hình âm học - Kiến trúc



Hình 3: Model 1



Hình 4: Model 2

Mô hình âm học

[5] Alex Graves, "Towards End-to-End Speech Recognition with Recurrent Neural Networks". *In ICML, 2014.*

Mô hình âm học

- Training với hàm CTC loss (Connectionist Temporal Classification) [5]

[5] Alex Graves, "Towards End-to-End Speech Recognition with Recurrent Neural Networks". *In ICML, 2014.*

Mô hình âm học

- Training với hàm CTC loss (Connectionist Temporal Classification) [5]
- Sử dụng phương pháp tối ưu SGD (Stochastic Gradient Descent) với:

[5] Alex Graves, "Towards End-to-End Speech Recognition with Recurrent Neural Networks". *In ICML, 2014.*

Mô hình âm học

- Training với hàm CTC loss (Connectionist Temporal Classification) [5]
- Sử dụng phương pháp tối ưu SGD (Stochastic Gradient Descent) với:
  - learning rate = 0.01
  - momentum = 0.9
  - nesterov = True
  - L2 weight decay = 1e-6
  - clipnorm = 5

[5] Alex Graves, "Towards End-to-End Speech Recognition with Recurrent Neural Networks". *In ICML, 2014.*

Decoding

Decoding

- Để lấy được văn bản đầu ra cho tín hiệu tiếng nói đó, chúng ta cần phải thực hiện decoding (giải mã) ma trận output của mô hình âm học.

Decoding

- Để lấy được văn bản đầu ra cho tín hiệu tiếng nói đó, chúng ta cần phải thực hiện decoding (giải mã) ma trận output của mô hình âm học.
- Giải thuật để decoding:

Decoding

- Để lấy được văn bản đầu ra cho tín hiệu tiếng nói đó, chúng ta cần phải thực hiện decoding (giải mã) ma trận output của mô hình âm học.
- Giải thuật để decoding:
  - Greedy Search (Max decoding)

Decoding

- Để lấy được văn bản đầu ra cho tín hiệu tiếng nói đó, chúng ta cần phải thực hiện decoding (giải mã) ma trận output của mô hình âm học.
- Giải thuật để decoding:
  - Greedy Search (Max decoding)
  - Beam Search

Decoding

- Để lấy được văn bản đầu ra cho tín hiệu tiếng nói đó, chúng ta cần phải thực hiện decoding (giải mã) ma trận output của mô hình âm học.
- Giải thuật để decoding:
  - Greedy Search (Max decoding)
  - Beam Search

Input: Một ma trận phân phối xác suất trên các ký tự

Decoding

- Để lấy được văn bản đầu ra cho tín hiệu tiếng nói đó, chúng ta cần phải thực hiện decoding (giải mã) ma trận output của mô hình âm học.
- Giải thuật để decoding:
  - Greedy Search (Max decoding)
  - Beam Search

Input: Một ma trận phân phối xác suất trên các ký tự
Output: Một văn bản (text) đầu ra được giải mã (decoded output).

Nội dung

Source code: https://gitlab.com/fakerphan/masr2
Documents: https://gitlab.com/fakerphan/masr2/tree/master/reports

Đánh giá mô hình

- Mô hình được đánh giá dựa trên ba độ đo (metrics):

Đánh giá mô hình

- Mô hình được đánh giá dựa trên ba độ đo (metrics):
  - CER (Char Error Rate): *Tỷ lệ lỗi của ký tự*,

Đánh giá mô hình

- Mô hình được đánh giá dựa trên ba độ đo (metrics):
  - CER (Char Error Rate): *Tỷ lệ lỗi của ký tự*,
  - WER (Word Error Rate): *Tỷ lệ lỗi của từ*,

Đánh giá mô hình

- Mô hình được đánh giá dựa trên ba độ đo (metrics):
  - CER (Char Error Rate): *Tỷ lệ lỗi của ký tự*,
  - WER (Word Error Rate): *Tỷ lệ lỗi của từ*,
  - SER (Sentence Error Rate): *Tỷ lệ lỗi của câu*.

Kết quả

	CER	WER	SER
Model 1	17.03	41.70	42.44
Model 2	7.73	19.67	20.16

Bảng 1: So sánh kết quả hai mô hình dựa trên các độ đo CER, WER và SER (%).

- Nhận xét:

Kết quả

	CER	WER	SER
Model 1	17.03	41.70	42.44
Model 2	7.73	19.67	20.16

Bảng 1: So sánh kết quả hai mô hình dựa trên các độ đo CER, WER và SER (%).

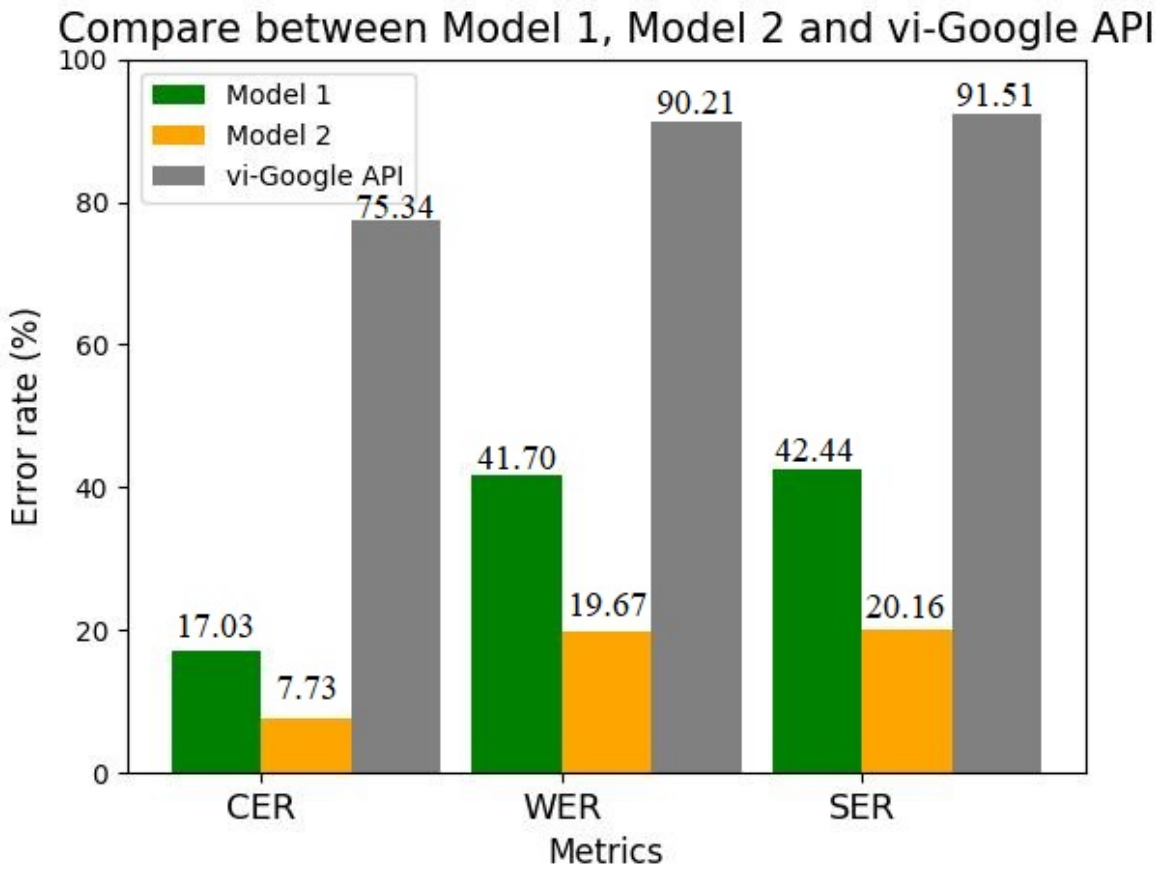
- Nhận xét:
  - Mô hình sử dụng mạng GRU 2 chiều mang lại hiệu suất vượt trội so với mạng RNN thuần 2 chiều.

Kết quả

	CER	WER	SER
Google API	75.34	90.21	91.51

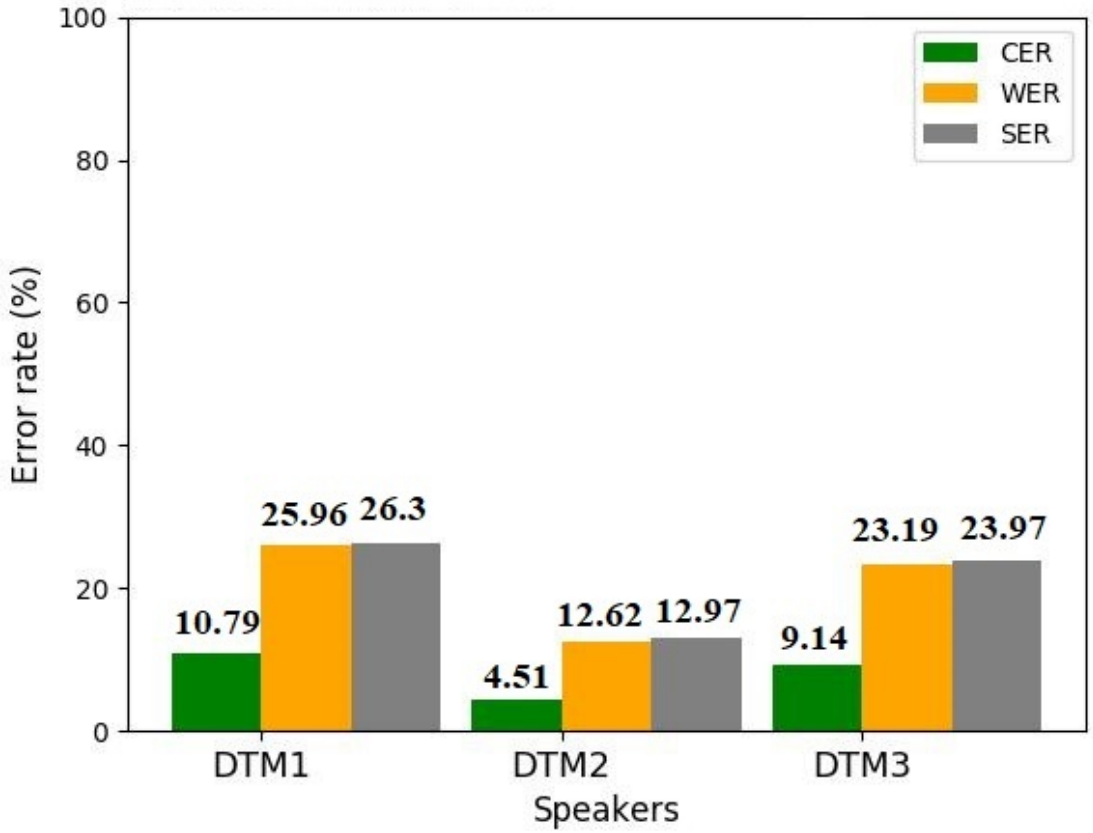
Bảng 2: Kết quả đánh giá khi sử dụng Google API nhận dạng tiếng nói Tiếng Việt của người khuyết tật giọng nói dựa trên các độ đo CER, WER và SER (%).

Kết quả



Hình 5: So sánh giữa các mô hình Model 1, Model 2 và sử dụng Google API

Kết quả



Hình 6: Kết quả đánh giá của Model 2 cho từng người nói *(DTM1, DTM2, DTM3)*

Nhận xét tổng quan

Nhận xét tổng quan

Nhận xét một số trường hợp dẫn đến nhận dạng sai dựa trên phân tích định tính đầu ra:

Nhận xét tổng quan

Nhận xét một số trường hợp dẫn đến nhận dạng sai dựa trên phân tích định tính đầu ra:

- Các chữ cái phát âm giống nhau, ví dụ như: "ia" với "ya"; "iê" với "yê"; "i" với "y"; "âu" với "ô"; "au" với "o"; ...

Nhận xét tổng quan

Nhận xét một số trường hợp dẫn đến nhận dạng sai dựa trên phân tích định tính đầu ra:

- Các chữ cái phát âm giống nhau, ví dụ như: "ia" với "ya"; "iê" với "yê"; "i" với "y"; "âu" với "ô"; "au" với "o"; ...
- Một số chữ cái mà người khuyết tật khó phát âm phân biệt, nhận dạng bị lệch lạc như: "tr" với "t"; "x" với "s"; "gi" với "d"; "nh" với "n"; "kh" với "c"; ...

Nhận xét tổng quan

Nhận xét một số trường hợp dẫn đến nhận dạng sai dựa trên phân tích định tính đầu ra:

- Các chữ cái phát âm giống nhau, ví dụ như: "ia" với "ya"; "iê" với "yê"; "i" với "y"; "âu" với "ô"; "au" với "o"; ...
- Một số chữ cái mà người khuyết tật khó phát âm phân biệt, nhận dạng bị lệch lạc như: "tr" với "t"; "x" với "s"; "gi" với "d"; "nh" với "n"; "kh" với "c"; ...
- Một số trường hợp nhận diện bị thiếu âm, chẳng hạn như "trong" thành "tong" hoặc "rong"; "dấu" thành "ấu"; ...

Nhận xét tổng quan

Nhận xét một số trường hợp dẫn đến nhận dạng sai dựa trên phân tích định tính đầu ra:

- Các chữ cái phát âm giống nhau, ví dụ như: "ia" với "ya"; "iê" với "yê"; "i" với "y"; "âu" với "ô"; "au" với "o"; ...
- Một số chữ cái mà người khuyết tật khó phát âm phân biệt, nhận dạng bị lệch lạc như: "tr" với "t"; "x" với "s"; "gi" với "d"; "nh" với "n"; "kh" với "c"; ...
- Một số trường hợp nhận diện bị thiếu âm, chẳng hạn như "trong" thành "tong" hoặc "rong"; "dấu" thành "ấu"; ...
- Ngoài ra mô hình vẫn còn nhận dạng sai khi xuất hiện nhiễu phát âm hoặc môi trường.

Nhận xét tổng quan

Nhận xét một số trường hợp dẫn đến nhận dạng sai dựa trên phân tích định tính đầu ra:

- Các chữ cái phát âm giống nhau, ví dụ như: "ia" với "ya"; "iê" với "yê"; "i" với "y"; "âu" với "ô"; "au" với "o"; ...
- Một số chữ cái mà người khuyết tật khó phát âm phân biệt, nhận dạng bị lệch lạc như: "tr" với "t"; "x" với "s"; "gi" với "d"; "nh" với "n"; "kh" với "c"; ...
- Một số trường hợp nhận diện bị thiếu âm, chẳng hạn như "trong" thành "tong" hoặc "rong"; "dấu" thành "ấu"; ...
- Ngoài ra mô hình vẫn còn nhận dạng sai khi xuất hiện nhiễu phát âm hoặc môi trường.

Nội dung

Source code: https://gitlab.com/fakerphan/masr2
Documents: https://gitlab.com/fakerphan/masr2/tree/master/reports

Kết luận

Kết quả đạt được

Kết luận

Kết quả đạt được

- Thực hiện quá trình thu thập và xử lý dữ liệu

Kết luận

Kết quả đạt được

- Thực hiện quá trình thu thập và xử lý dữ liệu
- Tìm hiểu, nghiên cứu xử lý tiếng nói và trích xuất đặc trưng cho âm thanh tiếng nói

Kết luận

Kết quả đạt được

- Thực hiện quá trình thu thập và xử lý dữ liệu
- Tìm hiểu, nghiên cứu xử lý tiếng nói và trích xuất đặc trưng cho âm thanh tiếng nói
- Nghiên cứu các mô hình học âm học và thực hiện xây dựng mô hình giải quyết bài toán.

Kết luận

Kết quả đạt được

- Thực hiện quá trình thu thập và xử lý dữ liệu
- Tìm hiểu, nghiên cứu xử lý tiếng nói và trích xuất đặc trưng cho âm thanh tiếng nói
- Nghiên cứu các mô hình học âm học và thực hiện xây dựng mô hình giải quyết bài toán.
- Đánh giá, so sánh các mô hình và sử dụng Google API với nhau.

Kết luận

Vấn đề hạn chế

Kết luận

Vấn đề hạn chế

- Vấn đề thu thập dữ liệu còn nhiều hạn chế về độ đa dạng về thời lượng, từ vựng, số lượng người nói.

Kết luận

Vấn đề hạn chế

- Vấn đề thu thập dữ liệu còn nhiều hạn chế về độ đa dạng về thời lượng, từ vựng, số lượng người nói.
- Việc trích xuất đặc trưng và decoding vẫn còn có thể được cải thiện.

Kết luận

Vấn đề hạn chế

- Vấn đề thu thập dữ liệu còn nhiều hạn chế về độ đa dạng về thời lượng, từ vựng, số lượng người nói.
- Việc trích xuất đặc trưng và decoding vẫn còn có thể được cải thiện.
- Việc huấn luyện và đánh giá mô hình còn hạn chế về thời gian và máy huấn luyện $\rightarrow$ hạn chế trong việc thử nghiệm thay đổi các tham số trong mô hình và quá trình huấn luyện.

Hướng phát triển trong tương lai

Hướng phát triển trong tương lai

Hướng phát triển trong tương lai

Hướng phát triển trong tương lai

- Giải quyết vấn đề dữ liệu: tăng cường dữ liệu; có thể thử nghiệm sử dụng các mô hình sinh hiệu quả để tăng cường dữ liệu tổng quan hơn, ...

Hướng phát triển trong tương lai

Hướng phát triển trong tương lai

- Giải quyết vấn đề dữ liệu: tăng cường dữ liệu; có thể thử nghiệm sử dụng các mô hình sinh hiệu quả để tăng cường dữ liệu tổng quan hơn, ...
- Cải thiện chất lượng việc trích xuất đặc trưng cho tiếng nói, mô hình âm học và việc decoding.

Hướng phát triển trong tương lai

Hướng phát triển trong tương lai

- Giải quyết vấn đề dữ liệu: tăng cường dữ liệu; có thể thử nghiệm sử dụng các mô hình sinh hiệu quả để tăng cường dữ liệu tổng quan hơn, ...
- Cải thiện chất lượng việc trích xuất đặc trưng cho tiếng nói, mô hình âm học và việc decoding.
- Giải quyết các vấn đề thách thức như nhiễu môi trường, tốc độ, ...

Hướng phát triển trong tương lai

Hướng phát triển trong tương lai

- Giải quyết vấn đề dữ liệu: tăng cường dữ liệu; có thể thử nghiệm sử dụng các mô hình sinh hiệu quả để tăng cường dữ liệu tổng quan hơn, ...
- Cải thiện chất lượng việc trích xuất đặc trưng cho tiếng nói, mô hình âm học và việc decoding.
- Giải quyết các vấn đề thách thức như nhiễu môi trường, tốc độ, ...
- Tìm hiểu, nghiên cứu hướng tiếp cận chuyển đổi giọng nói giữa người khuyết tật và người nói chuẩn (Voice Conversion).

Cảm ơn thầy cô và các bạn đã lắng nghe !

Q & A