**Thành viên nhóm:**

17021311 Cao Minh Nhật

17021313 Đinh Công Phan

**Sản phẩm:**

Xây dựng một web application với chức năng:

* Ghi âm thanh
* Ghi âm thanh ra dạng text
* Cho phép người đọc download file text nhận dạng

**Mục tiêu và kết quả đạt được:**

*Mục tiêu:*

Xây dựng được web app speech to text có thể dự đoán được một đoạn audio dài.

*Kết quả đạt được:*

Xây dựng được web app có thể dự đoán được đoạn câu ngắn, với một số lượng từ hạn chế.

**Công nghệ và phương hướng tiếp cận:**

*Công nghệ:*

-Sử dụng flask api để hỗ trợ xây dựng ứng dụng web application.

-Đi kèm với flask là trang html với giao diện css và backend javascript.

-Về mặt model xử lý tiếng nói, nhóm dùng thư viện librosa, sklearn, hmmlearn để thực hiện công việc.

*Tiếp cận:*

- Ban đầu nhóm cân nhắc sử dụng cái cấu trúc deep learning dành cho nhận diện chuỗi âm thanh như RNN, LSTM hay trực tiếp luôn text to speech engines Deep Speech để có thể đạt được hiệu quả cao. Nhưng vì nhiều lý do và khó khăn, nhóm không thể ứng dụng được các phương pháp trên.

- Nhóm nhắm đến dùng hướng sử dựng môn hình học máy HMM để giải quyết vấn đề, cụ thể như sau:

* Nhóm sử dụng dữ liệu tiếng việt của lớp và file srt để tách các audio wave tương ứng với các từ và đặt vào các folder khác nhau, gọi là các classes.

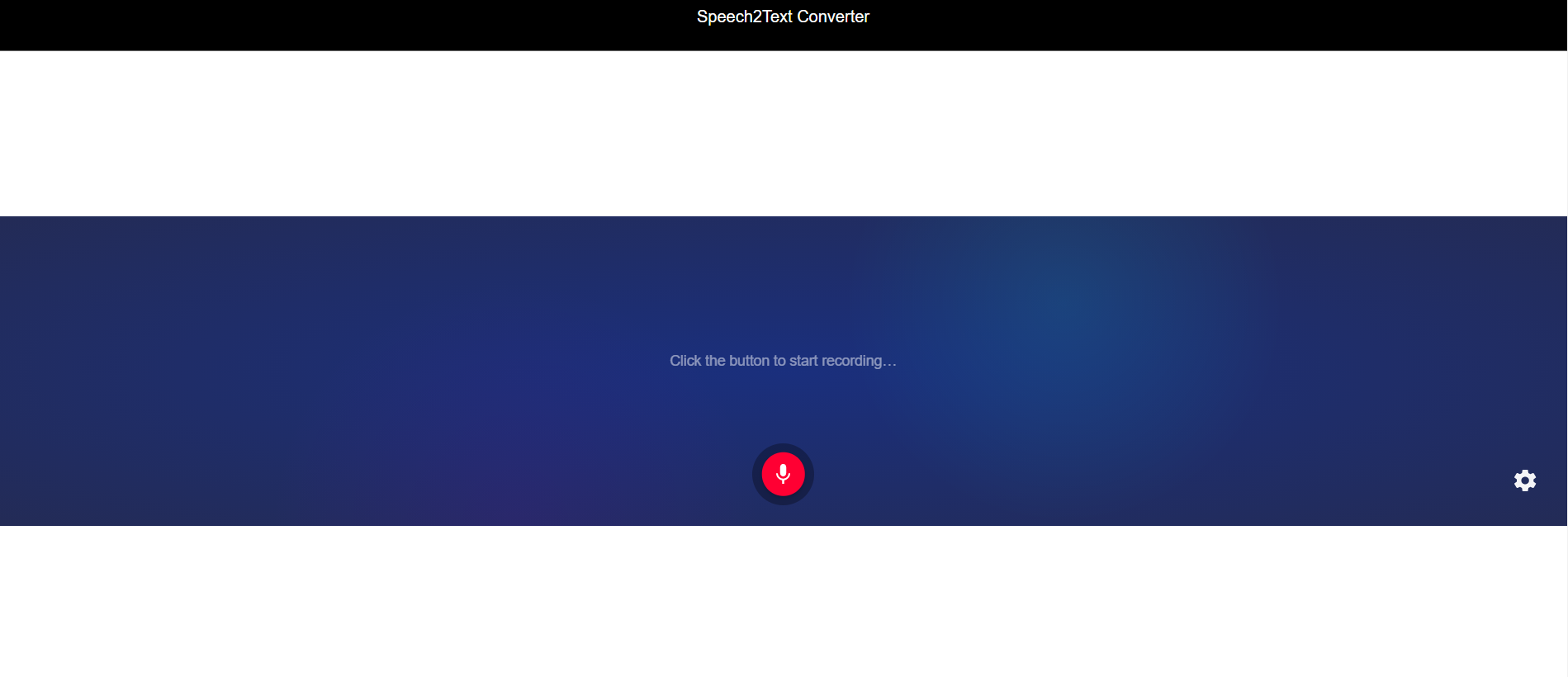
Ex: “Hà” và “Nội” sẽ bị coi là hai từ khác nhau đặt vào 2 mục khác nhau.

* Với mỗi class trường ứng: từ các audio wave, nhóm chiết xuất mfcc, dùng K mean clustering cho từng class để dự đoán phân cụm đặc trưng.
* Với mỗi class: Các cụm đặc trưng sẽ được đưa vào mô hình HMM, kết quả của từng class lưu riêng.
* Khi dự đoán âm thanh, model dự đoán trên mọi class và lấy kết quả cao nhất.
* Để dự đoán được một đoạn câu ngắn, nhóm sử dụng khoảng lặng để cắt audio các từ. Dự đoán từ của các audio trên gộp lại là câu ngắn mà ứng dụng sẽ trả về.

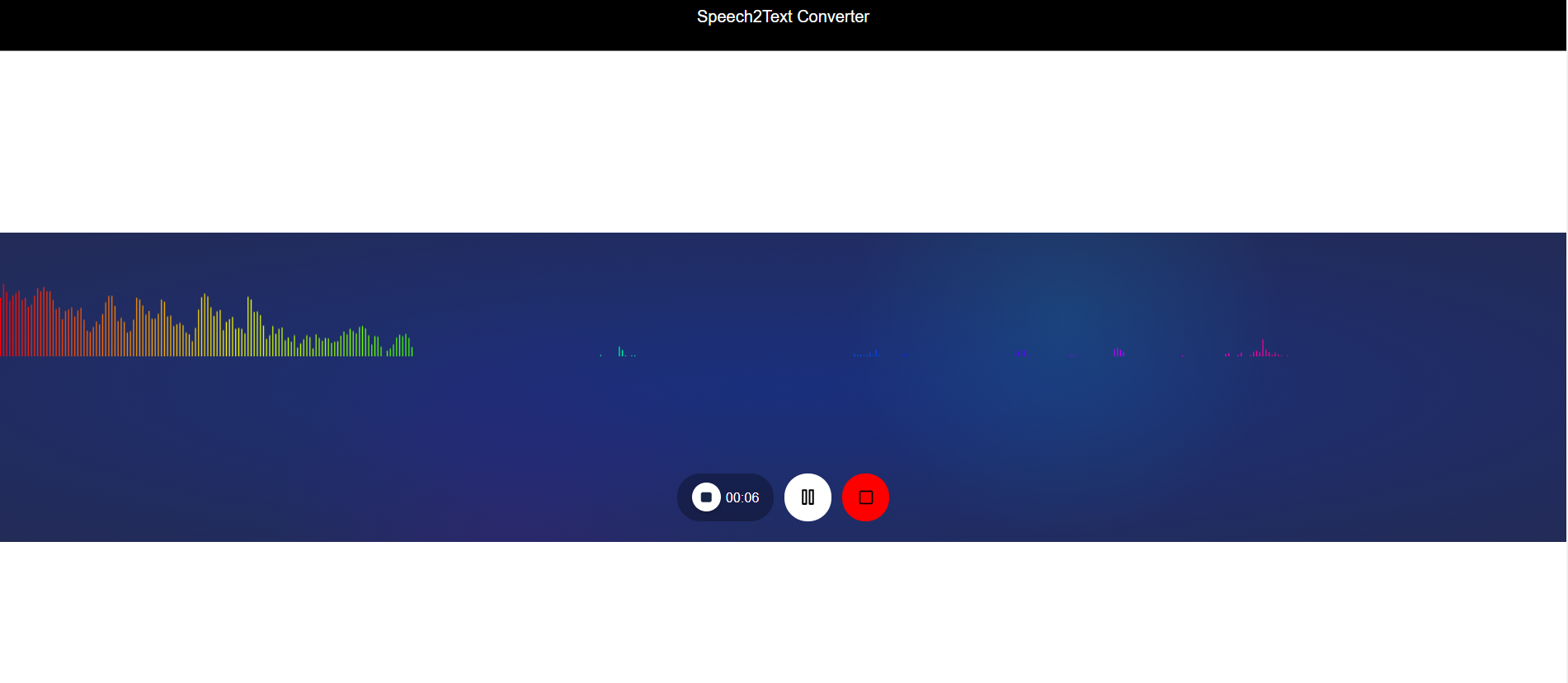
Nhược điểm: Theo nhóm hiểu, các từ đa dạng khác nhau cần các tham số đầu vào khác nhau, tuy vậy, nhóm chỉ chọn một tham số nhất định và để mô hình chạy trên mọi từ. Việc chọn tham số thực chất nhóm cũng chưa hiểu sâu thực sự bản chất, có tính ước đoán nhiều.

**Chương trình cụ thể:**

*Giao diện đầu vào và giao diện ghi âm:*



Hình 1: Giao diện đầu vào của ứng dụng

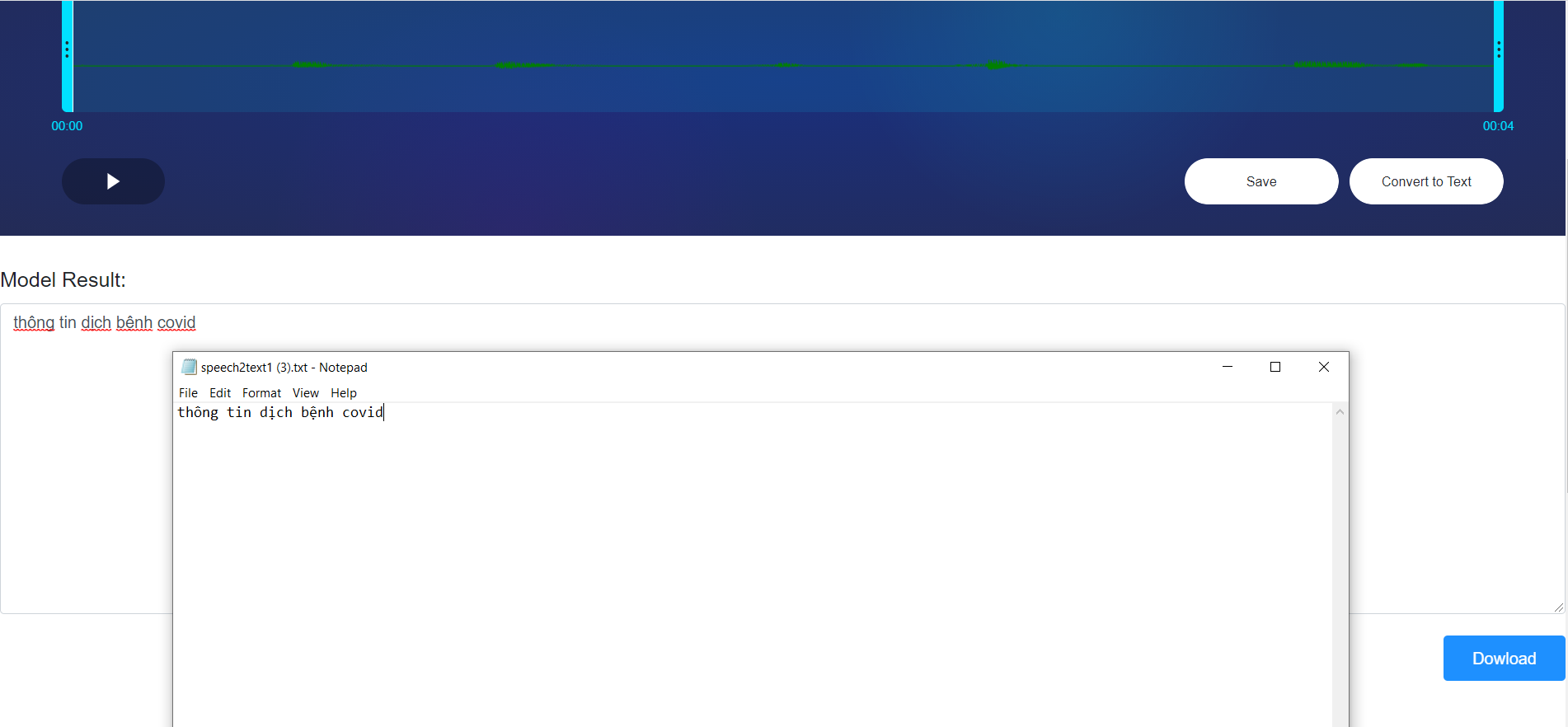


Hình 2: Giao diện khi ghi âm của ứng dụng

*Biến đổi audio thành text và lưu kết quả:*



Hình 3: Minh họa chuyển đổi audio của công cụ



Hình 4: Minh họa kết quả chuyển đổi audio thành văn bản