* Nội dung bài báo cáo của nhóm chúng em bao gồm các nội dung: giới thiệu paper nhóm chọn để tìm hiểu và báo cáo; các vấn đề đang gặp phải và hướng giải quyết của model đối với các vấn đề đã đặt ra; cuối cùng là phần demo model của tác giả.
* Nhóm em tham khảo bài báo của nhóm tác giả này, đường link được bọn em đính kèm trong phần mô tả của video trên youtube. Nội dung chính của paper là giải quyết bài toán text to speed (TTS) - chuyển văn bản thành giọng nói với chất giọng cho trước và chạy trong thời gian thực
* Nhân bản giọng nói (voice cloning) là việc tạo ra một mô phỏng nhân tạo giọng nói của một người. Các phần mềm AI nhân bản giọng nói ngày nay có khả năng tạo ra giọng nói tổng hợp gần giống với giọng nói của con người được nhắm mục tiêu.   
  Theo các phương pháp truyền thống, để có thể nhân bản giọng nói của một đối tượng, ta phải cần tập dữ liệu huấn luyện lên đến hàng giờ ghi lại giọng nói của đối tượng đó.
* Các mô hình truyền thống, khi xử lý tác vụ tương tự đều cần rất nhiều dữ liệu của giọng nói nguồn (hàng trăm giờ) để huấn luyện.
* Khi muốn nhân bản giọng nói khác, phải lặp lại quá trình bên trên.

🡪 Tốn một chi phí lớn cho việc huấn luyện

* Cần một mô hình máy học vận hành với thời gian thực và không cần cập nhật trọng số.
* Input: Một đoạn ghi âm giọng nói ngắn (khoảng 5s) và đoạn văn bản cần chuyển thành giọng nói.
* Output: Giọng nói giống với đầu vào.
* Nhóm tác giả đã xây dựng một model tên là Multispeaker speech synthesis” bao gồm 3 mạng neural được train riêng biệt là…..”
* Mạng học sâu **Speaker Encoder**

bộ mã hóa tiếng nói (speaker encoder) là mạng học sâu đã được huấn luyện với hàng ngàn đối tượng. Bộ mã hóa sẽ tạo ra vector nhúng (embedding), đây là biểu diễn tiếng nói một cách dễ hiểu đối với máy tính (Có thể xem như một bộ trích xuất đặc trưng). Nói cách khác, bộ mã hóa này cố gắng học được bản chất tiếng nói của con người qua dữ liệu giọng nói của hàng ngàn đối tượng.

* + Mục tiêu: Trích xuất Speaker Embedding (một vector biểu diễn âm thanh để máy tính có thể hiểu được) của ngôn ngữ nói
  + Input: Dữ liệu giọng nói
  + Output: Speaker Embedding
* Output của mạng neural này được dùng làm input cho mạng **Synthesizer**
* Mạng **Synthesizer**
* Bộ tổng hợp (Synthesizer) nhận các đặc trưng mà bộ mã hóa tổng hợp được cùng với đoạn văn bản mà ta muốn tổng họp từ đó tạo ra biểu đồ Mel Spectrogram (biểu diễn ngắn gọn của giọng nói và ngữ điệu của một người nào đó). Việc triển khai mô đun này dựa trên kĩ thuật Tacotron của DeepMind.
  + Mục tiêu: Tổng hợp âm thanh từ đoạn văn bản dựa trên speaker embedding của giọng nói nguồn.
  + Input: Speaker embedding vector từ mạng Speaker Encoder và đoạn văn bản cần chuyển thành giọng nói
  + Output: Biểu đồ Mel Spectrogram – một dạng biểu diễn âm thanh
  + Kỹ thuật được sử dụng: Tacotron 2 của DeepMind
* Output của mạng này được dùng làm input cho mạng **Vocoder**
* Mạng **Vocoder**

Bộ tổng hợp tiếng nói (vocoder) tổng hợp ra âm thanh ở dạng sóng từ 1 biểu đồ Mel Spectrogram - đây là cách biểu diễn âm thanh tốt nhất để áp dụng trong học máy. Chiều x là thời gian đoạn ghi âm, chiều y là tần số âm thanh. Vocoder được triển khai dựa trên kĩ thuật WaveNet của DeepMind

* + Mục đích: Chuyển biểu đồ Mel Spectogram thành dạng âm thanh nghe được
  + Input: Biểu đồ Mel Spectogram
  + Output: Sóng âm thanh nghe được
* Kỹ thuật sử dụng: WaveNet của DeepMind
* Tổng hợp lại các quá trình trên, ta có sơ đồ tóm tắt sau:

+ Đầu tiên, input là đoạn ghi âm được đưa vào Speaker Encoder để trích xuất đặc trưng giọng nói và theo như trong paper thì các đặc trưng đó được lưu thành một embedding

+ Sau đó embedding được truyền sang mạng Synthesizer. Ở mạng Synthesizer còn phải truyền thêm input là đoạn văn bản mình muốn phần mềm đọc ra. Output của mạng này là một biểu đồ Spectogram và cuối cùng chuyển vào Vocoder để xuất ra âm thanh nghe được

* Để đánh giá đươc độ hiểu quả của hệ thống trên, theo nguyên tắc ta cần đáng giá độ tương đồng (similarity) của giọng nói nguốn với giọng nói được tổng hợp nhưng nói 2 câu nói hoàn toàn khác nhau. Ngoài ra ta còn phải đánh giá được độ tự nhiên (naturalness) của giọng nói đươc tổng hợp. Việc đánh giá này là 1 vấn đề khó khăn. Với các tập dữ liệu huấn luyện khác nhau, kết quả đánh giá có thể sẽ khác nhau.
* Để giải quyết vần đề này, tác giả sử dụng MOS (Mean Opion Score), sẽ có nhiều người đánh giá nghe đoạn âm thanh được tổng hợp và đánh giá độ tương đồng và độ tự nhiên trên thang đo từ 1 đến 5. Từ đó mà đánh giá mẫu âm thanh được chuyển thành giọng nói của con người tốt như thế nào.