# THÔNG TIN CHUNG CỦA BÁO CÁO

- Link YouTube video của báo cáo:
  *https://youtu.be/-SOdWQcY0yg*
- Link slides:
  https://github.com/akaihachu/CS2205.APR2023/slide.pdf
  https://github.com/akaihachu/CS2205.APR2023/proposal.pdf
  https://github.com/akaihachu/CS2205.APR2023/brochure.pdf
- *Mỗi thành viên của nhóm điền thông tin vào một dòng theo mẫu bên dưới*
- *Sau đó điền vào Đề cương nghiên cứu (tối đa 5 trang), rồi chọn Turn in*

| | |
|---|---|
| • Họ và Tên: Nguyễn Hoa<br>• MSSV: 220101044<br>  | • Lớp: CH1702-KHMT<br>• Tự đánh giá (điểm tổng kết môn): 9.0/10<br>• Số buổi vắng: 2<br>• Số câu hỏi QT cá nhân: 6<br>• Số câu hỏi QT của cả nhóm: 5<br>• Link Github: https://github.com/akaihachu/CS2205.APR2023<br>• Mô tả công việc và đóng góp của cá nhân cho kết quả của nhóm: Thực hiện cá nhân toàn bộ |

# ĐỀ CƯƠNG NGHIÊN CỨU

**TÊN ĐỀ TÀI**

HƯỚNG ĐẾN TÓM LƯỢC NỘI DUNG NÓI THÀNH VĂN BẢN TỪ ĐẦU ĐẾN CUỐI

**TÊN ĐỀ TÀI TIẾNG ANH**

TOWARDS END-TO-END SPEECH-TO-TEXT SUMMARIZATION

**TÓM TẮT**

*Tóm lược nội dung nói thành văn bản* (tóm lược nội dung S2T) là một nhiệm vụ nhằm tạo ra các bản tóm tắt ngắn gọn và súc tích các tài liệu nói, như podcast, bài giảng, hoặc phát thanh tin tức. Tóm lược nội dung S2T có thể giúp người dùng truy cập và tiếp nhận lượng lớn thông tin âm thanh một cách hiệu quả về thời gian. Hầu hết các phương pháp tóm lược nội dung S2T hiện có đều dựa vào việc ghép tầng một *bộ nhận dạng tiếng nói* (ASR) với một *mô hình tóm lược nội dung văn bản* (T2T). Tuy nhiên, cách làm này thể hiện một số lỗi và cho hiệu quả chưa cao do sự chưa khớp nhau giữa dữ liệu tiếng nói và dữ liệu văn bản, và dẫn đến sự thiếu tối ưu hóa của hai hệ thống riêng lẻ này.

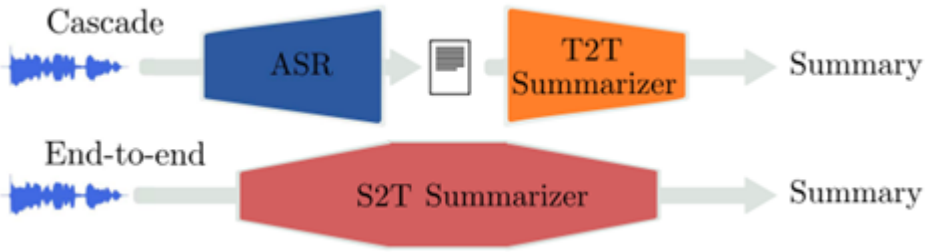
Mô hình tóm lược nội dung văn bản *từ đầu đến cuối* (E2E) của là một phương pháp thay thế hứa hẹn có thể trực tiếp tạo ra các bản tóm lược nội dung từ đầu vào tiếng nói, mà không cần phải tạo ra lời thoại trung gian, do đó chúng có thể khai thác các đặc trưng âm thanh và ngữ điệu từ tiếng nói, cũng như học được các biểu diễn tiềm ẩn giàu có mà nắm bắt các khía cạnh ngữ nghĩa và ngữ cảnh của nội dung nói để tạo ra các bản tóm lược nội dung có tính trung thực và mạch lạc hơn đối với đầu vào tiếng nói. Tuy nhiên, chúng cũng đặt ra nhiều thách thức, như sự khan hiếm của dữ liệu tiếng nói được gán nhãn tóm tắt, sự phức tạp và đa dạng của ngôn ngữ nói cùng với đòi hỏi về chi phí tính toán do phải xử lý các chuỗi âm thanh dài.

Trong đề tài này đề xuất giải quyết yêu cầu này bằng cách phát triển các mô hình E2E đề xuất cho tóm lược nội dung S2T có thể xử lý hiệu quả dữ liệu đầu vào tiếng nói dài và đa dạng, và khai thác các kho văn bản lớn để cải thiện hiệu suất của chúng để so sánh hiệu quả với các mô hình ASR/T2T thông dụng.

**GIỚI THIỆU**

Tin tức phát sóng, podcast, bài giảng, … được phát hành với số lượng lớn trên các đa phương tiện nghe nhìn nhưng việc tìm kiếm thông tin về chúng mất khá nhiều thời gian. Hệ thống tóm lược nội dung S2T trợ giúp bằng cách xác định các nội dung quan trọng trong lời nói của tin tức và tạo ra một văn bản vắn tắt tùy theo yêu cầu là một giải pháp thường được

sử dụng. Trọng tâm nghiên cứu gần đây đang hướng đến các hệ thống tóm lược nội dung tiếng nói, trong đó các dữ liệu nói sẽ được chuyển thể thành các bản tóm lược nội dung S2T tinh gọn tương tự như một bản tóm tắt do một chuyên gia viết lại.




*Hình 1: Minh họa sơ đồ ghép tầng 2 giai đoạn và sơ đồ E2E để tóm lược nội dung S2T*

Các hệ thống tóm lược nội dung S2T sử dụng phương pháp *ghép tầng ASR/T2T* (xem Hình 1), trong đó sử dụng mô hình nhận dạng tiếng nói tự động (ASR) để tạo lời thoại trung giàn, nối tiếp với một mô hình tóm lược nội dung văn bản thành văn bản (T2T) để tạo ra bản tóm lược nội dung tiếng nói [1]. Chúng được đào tạo trên 2 bộ dữ liệu tách biệt để tóm lược nội dung đối thoại [2]. Tuy nhiên, thách thức đặt ra là các bản ghi do mô hình ASR tạo ra vẫn còn chứa một số lỗi không thể khắc phục, do không thể tận dụng các thông tin ngoài lời thoại (chẳng hạn độ lớn âm lượng, biểu cảm của tiếng nói) cho việc tóm lược nội dung [3].

Nhiều mô hình ngôn ngữ khác đã được đề xuất để giải quyết thách thức trên [4,5], trong đó mô hình từ đầu đến cuối E2E (xem **Hình 1**) được cho rằng có thể quyết vấn đề này trong hai bài viết khác nhau [6,7] do chúng không cần tạo ra lời thoại trung gian mà xử lý bằng một mô hình âm thanh và ngôn ngữ duy nhất từ đầu đến cuối. Tuy nhiên, câu hỏi được đặt ra là *liệu rằng các mô hình E2E có hiệu quả đến đâu và có đủ khả năng thay thế các mô hình ASR/T2T thông dụng để tóm lược nội dung S2T hay không* trong điều kiện lượng dữ liệu âm thanh gán nhãn tóm lược nội dung đang khan hiếm, sự phức tạp và đa dạng của ngôn ngữ nói cùng với chi phí cao hơn xử lý các chuỗi âm thanh dài.

Để trả lời câu hỏi trên, đề tài đề xuất một mô hình ghép tầng ASR/T2T thông dụng và một vài mô hình E2E được phát triển để tóm lược nội dung S2T với dữ liệu tin tức phát sóng. Trong đó, mô hình ghép tầng ASR/T2T được fine-tuning trên bộ dữ liệu tin tức phát sóng còn các mô hình E2E thì sử dụng mô hình transformer có dùng sử dụng một kiến trúc mã hóa – giải mã (encoder-decoder) và *sự tự chú ý hạn chế* (restricted self-intention) và các mô hình *học chuyển giao* (transfer learning) tận dụng các đặc trưng tiếng nói được trích xuất sử dụng mô hình biểu diễn tiếng nói đã được học trước [8]. Cả hai kiểu mô hình trên được so sánh với nhau và so sánh với các mô hình SOTA cơ sở sử dụng điểm số ROUGE để đánh giá cùng với sự đánh giá của con người.

- *INPUT:* Một bản tin phát sóng audio.
- *OUTPUT:* Một đoạn văn bản tóm lược nội dung lại tạo ra để tóm lược nội dung lại

input.

## MỤC TIÊU

- Nghiên cứu và đánh giá các phương pháp hiện có về việc tóm lược nội dung S2T cho dữ liệu tin tức phát sóng theo phương pháp ghép tầng ASR/T2T thông dụng và chỉ ra những hạn chế, tồn tại cần giải quyết.
- Đề xuất các phương pháp mới cho S2T tổng hợp theo phương pháp E2E đề xuất (cả transformer và transfer learning) và so sánh hiệu suất với các mô hình cơ sở.
- Thực nghiệm và đánh giá các mô hình E2E đề xuất trên bộ dữ liệu có nhãn có số lượng khan hiếm và đánh giá khả năng mở rộng dữ liệu tin tức phát sóng trên các ngôn ngữ khác nhau, cũng như các loại dữ liệu loại khác như video, podcast,....

## NỘI DUNG VÀ PHƯƠNG PHÁP

**Nội dung**

- **Giới thiệu:** Trình bày bối cảnh, đặt vấn đề, mục tiêu, phương pháp và đóng góp của luận án.
- **Cơ sở lý thuyết:** Trình bày các khái niệm cơ bản, các mô hình và kỹ thuật liên quan đến S2T tổng hợp theo phương pháp E2E.
- **Các công trình liên quan:** Khảo sát và phân tích các công trình đã công bố về S2T tổng hợp, cả kiểu ghép tầng ASR/T2T và kiểu đầu cuối E2E, đồng thời chỉ ra những hạn chế và thách thức cần giải quyết.
- **Phương pháp đề xuất:** Trình bày chi tiết các phương pháp đề xuất cho S2T tổng hợp theo phương pháp E2E, bao gồm mô hình dựa trên transformer và mô hình dựa trên transfer learning.
- **Thực nghiệm và đánh giá:** Trình bày các thiết lập mô hình thực nghiệm, các chỉ số đánh giá, các kết quả thực nghiệm các mô hình E2E đề xuất, cả transformer và transfer learning, và một mô hình ghép tầng ASR/T2T thông dụng được train cùng dữ liệu giọng nói đầu vào. Đề tài sẽ tiến hành các thử nghiệm rộng rãi trên các lĩnh vực với các ngôn ngữ khác nhau của tài liệu nói, chẳng hạn như video hướng dẫn, chương trình phát sóng tin tức và podcast. Đề tài sẽ sử dụng bộ dữ liệu có sẵn công khai cho các các ngôn ngữ khác nhau, chẳng hạn như How-2 (tiếng Anh), AMI (tiếng Anh), ESTER (tiếng Pháp) hoặc Giga Speech (tiếng Trung). Đề tài sẽ so sánh các mô hình đề xuất với các mô hình cơ sở thông dụng bằng cách sử dụng điểm số ROUGE cũng như dựa vào đánh giá của con người.
- **Kết luận và hướng phát triển:** Tổng kết lại các kết quả đạt được, nhận xét về ưu

nhược điểm của các mô hình đề xuất và đưa ra các hướng phát triển trong tương lai. Đánh giá tiềm năng của mô hình E2E trong việc thay thế các mô hình ASR/T2T thông dụng.

**Phương pháp**

- **Phương pháp dựa trên transformer**: sử dụng một kiến trúc encoder-decoder, trong đó encoder nhận đầu vào là dữ liệu tiếng nói và decoder sinh ra văn bản tóm tắt. Cả encoder và decoder đều sử dụng các khối transformer sử dụng restricted self-attention, trong đó chỉ có một số lượng nhỏ các vị trí được chú ý đến.
- **Phương pháp dựa trên transfer learning:** từ mô hình BART, một mô hình ngôn ngữ được tiền huấn luyện nhiệm vụ tổng hợp văn bản trừu tượng, được hoán cải bằng cách thêm vào một lớp *tích chập 1 chiều* (1D convolution) ở đầu của encoder nhằm giảm số chiều của chuỗi âm thanh để có thể nhận input mới là tài liệu nói. Sau đó, tiếp tục huấn luyện mô hình trên bộ dữ liệu S2T đã gán nhãn của đề tài để điều chỉnh cập nhật các hệ số theo dữ liệu mới.

## KẾT QUẢ DỰ KIẾN

- Xây dựng được các mô hình S2T theo phương pháp E2E có khả năng sinh ra các bản tóm lược nội dung trừu tượng chất lượng cao từ tiếng nói.
- Cải thiện được hiệu suất của S2T tổng hợp theo phương pháp E2E tiến đến gần tương đương so với các mô hình ghép tầng ASR/T2T thông dụng nhằm tạo ra một hướng đi mới cho việc hoàn thiện mô hình tóm lược nội dung S2T và khắc phục được một số lỗi mà các mô hình phổ biến hiện nay đang gặp phải.
- Đánh giá khả năng mở rộng dữ liệu với các ngôn ngữ khác nhau, cũng như các loại dữ liệu khác như podcast, video,...

## TÀI LIỆU THAM KHẢO

*[1] Dana Rezazadegan, Shlomo Berkovsky, Juan C. Quiroz, Ahmet Baki Kocaballi, Ying Wang, Liliana Laranjo, Enrico W. Coiera: Automatic Speech Summarisation: A Scoping Review. CoRR abs/2008.11897 (2020)*

*[2] Yusen Zhang, Ansong Ni, Tao Yu, Rui Zhang, Chenguang Zhu, Budhaditya Deb, Asli Celikyilmaz, Ahmed Hassan Awadallah, Dragomir R. Radev: An Exploratory Study on Long Dialogue Summarization: What Works and What's Next. CoRR abs/2109.04609 (2021)*

*[3] Solène Evain, Ha Nguyen, Hang Le, Marcely Zanon Boito, Salima Mdhaffar, Sina Alisamir, Ziyi Tong, Natalia A. Tomashenko, Marco Dinarelli, Titouan Parcollet, Alexandre Allauzen, Yannick Estève, Benjamin Lecouteux, François Portet, Solange Rossato, Fabien Ringeval, Didier Schwab, Laurent Besacier: Task Agnostic and Task Specific Self-Supervised Learning from Speech with*

*LeBenchmark. NeurIPS Datasets and Benchmarks 2021*

*[4] Atsunori Ogawa, Tsutomu Hirao, Tomohiro Nakatani, Masaaki Nagata: ILP-based Compressive Speech Summarization with Content Word Coverage Maximization and Its Oracle Performance Analysis. ICASSP 2019: 7190-7194*

*[5] Shi-Yan Weng, Tien-Hong Lo, Berlin Chen: An Effective Contextual Language Modeling Framework for Speech Summarization with Augmented Features. CoRR abs/2006.01189 (2020)*

*[6] Sharma, R., Palaskar, S., Black, A.W., Metze, F.: End-to-end speech summarization using restricted self-attention. In: ICASSP 2022: 8072–8076.*

*[7] Kohei Matsuura, Takanori Ashihara, Takafumi Moriya, Tomohiro Tanaka, Atsunori Ogawa, Marc Delcroix, Ryo Masumura: Leveraging Large Text Corpora for End-to-End Speech Summarization. CoRR abs/2303.00978 (2023)*

*[8] Akos T¨undik, M., Kaszas, V., Szaszak, G.: Assessing the Semantic Space Bias Caused by ASR Error Propagation and its Effect on Spoken Document Summarization. In: Proc. Interspeech 2019/1333–1337 (2019).*