JTE
HCMUTE
ISSN: 1859-1272

JOURNAL OF TECHNICAL EDUCATION SCIENCE
Ho Chi Minh City University of Technology and Education
Website: https://jte.hcmute.edu.vn/index.php/jte/index
Email: jte@hcmute.edu.vn

Create the Video Subtitles Based on Voice Recognition Technology: Test for Some Programs at VTV

Phong Nguyen-Huu[1*], Vo Nguyen Quoc Bao[2], Tran Minh Trung[1]
[1]Vietnam Television, Vietnam
[2] Posts and Telecommunications Institute of Technology, Vietnam
[*] Corresponding author. *Email: phongnguyen@vtv.vn*

ARTICLE INFO

Received: 19/1/2022
Revised: 12/4/2022
Accepted: 17/8/2022
Published: 30/8/2022

KEYWORDS

STT;
WER;
VOD;
OTT;
CC.

ABSTRACT

This paper presents the trial results of Speech-To-Text (STT) recognition tool for VOD (Video On Demand) contents of the VTVgo system at Vietnam Television. In order to evaluate the accuracy of the STT tool, the word error rate (WER) was used to measuring the performance of the automatic speech recognition, the machine translation system. Test results of 10 different types of TV show with 1065 video hours were analyzed. The WER had achieved low level from 2.8% to 4.3% with some genres of news, 19h, weather forecasts, where the majority of speakers, presenters (MC) read standard voices in the Studio. The dialogue from a speaker, less interference from outside noise. Besides, to illustrating the video subtitle application, we had conducted the test on the VTVgo system, integrated the optional subtitle display tool into the VTVgo app. The test Android platform was Smart TV and SmartPhone, to demonstrating the ability to apply video subtitles on the OTT (Over The Top) - the digital content distribution platform.

Tạo Phụ Đề Video Dựa Trên Kỹ Thuật Nhận Dạng Giọng Nói: Thử Nghiệm Cho Một Số Chương Trình Tại VTV

Nguyễn Hữu Phong[1*], Võ Nguyễn Quốc Bảo[2], Trần Minh Trung[1]
[1]Đài Truyền hình Việt Nam, Việt Nam
[2]Học viện Công nghệ Bưu chính Viễn Thông Cơ sở tại TP.HCM, Việt Nam
** Tác giả liên hệ. Email:phongnguyen@vtv.vn*

THÔNG TIN BÀI BÁO

Ngày nhận bài: 19/1/2022
Ngày hoàn thiện: 12/4/2022
Ngày chấp nhận đăng: 17/8/2022
Ngày đăng: 30/8/2022

TỪ KHÓA

Nhận dạng giọng nói;
Tỉ lệ lỗi từ;
Video theo yêu cầu;
Dịch vụ OTT;
Phụ đề chi tiết.

TÓM TẮT

Bài báo này trình bày kết quả thử nghiệm công cụ nhận dạng giọng nói Speech-To-Text (STT) cho các nội dung VOD (Video On Demand) trên hệ thống VTVgo của Đài THVN. Để đánh giá độ chính xác của công cụ STT, tỷ lệ lỗi từ (WER: Word Error Rate) được sử dụng để đo hiệu suất của hệ thống nhận dạng giọng nói tự động, dịch máy. Kết quả thử nghiệm thực hiện 10 thể loại chương trình truyền hình khác nhau với 1065 giờ video. Tỉ lệ WER thấp nhất là 2.8% đến 4.3% đạt được với một số thể loại chương trình thời sự và tin tức, dự báo thời tiết, ở đó phần lớn người nói, người dẫn chương trình (MC) đọc giọng chuẩn trong Studio và lời thoại từ một người nói, ít bị nhiễu bởi tạp âm bên ngoài. Bên cạnh đó, để minh họa ứng dụng phụ đề video, chúng tôi tiến hành thử nghiệm trên hệ thống VTVgo, tích hợp công cụ hiển thị phụ đề tùy chọn vào ứng dụng VTVgo app. Nền tảng thử nghiệm là SmartTV và SmartPhone Android, nhằm minh họa khả năng ứng dụng phụ đề video trên nền tảng phân phối nội dung số OTT (Over The Top).

Doi: https://doi.org/10.54644/jte.71B.2022.1128

JTE
HCMUTE
ISSN: 1859-1272

JOURNAL OF TECHNICAL EDUCATION SCIENCE
Ho Chi Minh City University of Technology and Education
Website: https://jte.hcmute.edu.vn/index.php/jte/index
Email: jte@hcmute.edu.vn

1. Giới thiệu

Phụ đề chi tiết CC [1] (*closed captions: một số tài liệu còn gọi là phụ đề đóng, nghĩa là chỉ hiển thị text khi người xem nhấn vào nút hiển thị phụ đề tùy chọn*) cho video trở thành một phương tiện quan trọng để cung cấp thông tin cho người cao tuổi hoặc người khiếm thính gặp khó khăn khi nghe âm thanh của các chương trình truyền hình. Ngày nay, người xem có thể chọn hiển thị phụ đề theo sở thích cá nhân. Một số chương trình truyền hình của CNN headline news, ABC world news, BBC [2] chứa phụ đề kèm video khi phân phối.

Giá trị của phụ đề trong việc xem nội dung truyền hình cho những người khiếm thính của các chương trình truyền hình từ lâu đã được công nhận và được phản ánh trong luật pháp của Nhật, Mỹ và Châu Âu [3]. Các Đài TH (truyền hình) trên thế giới ví dụ như BBC, NHK [4], đã nghiên cứu và thử nghiệm phụ đề từ những năm thập niên 2000. Giai đoạn 2010-2012, BBC bắt đầu cung cấp dịch vụ phụ đề cho các chương trình phát sóng thương mại trên các nền tảng: Broadcast, OTT (Over-The-Top) và Internet. Năm 2010, NHK nghiên cứu mô hình nhận dạng giọng nói để tạo phụ đề cho các ứng dụng video trực tuyến [5]. Đến năm 2015-2016, BBC tiếp tục phát triển hệ thống nhận diện giọng nói để tạo phụ đề tự động cho nhiều thể loại chương trình, bao gồm các video clip được phân phối trên web [6].

Phụ đề cho video thời gian thực là một thách thức trong phát sóng truyền hình quảng bá do yêu cầu về tính chính xác cao của công cụ STT, đồng bộ thời gian thực. Để giải quyết vấn đề này, các trung tâm nghiên cứu TH và các trường đại học đã tham gia nghiên cứu, thử nghiệm tạo phụ đề video cho phát sóng TH từ 2015 [7] trong một dự án chung. Giải pháp phụ đề đa ngôn ngữ: Live Caption; Close Caption; Sub-titles, ứng dụng AI tạo phụ đề tự động đã được thực hiện [8].

Gần đây, kỹ thuật tạo phụ đề tự động cho ứng dụng trên các màn hình thứ hai (Smart-Phone) được nghiên cứu [9]. Tạo phụ đề cho các video trên mạng xã hội và đa màn hình đã được thực hiện trong [10], xem xét những thách thức và định dạng liên quan đến việc xuất bản nội dung có phụ đề và phụ đề trên mạng xã hội.

Tạo phụ đề cho video trực tuyến của Netflix được thực hiện từ 2015 [11]. Năm 2016, Facebook hỗ trợ tạo phụ đề cho các video trực tuyến [12]. YouTube cũng đã tích hợp công cụ làm phụ đề tiếng Anh trực tuyến và một số ngôn ngữ khác [13]. Qua khảo sát có thể thấy tạo phụ đề cho video phát sóng TH và phân phối trên các nền tảng internet, mạng xã hội (Facebook, YouTube, Netflix) đã được thực hiện từ khá sớm và đã được thương mại hóa. Sử dụng kỹ thuật STT với độ chính xác lên đến 98% cho các nội dung trực tiếp [14].

Độ chính xác của các hệ thống ASR được đo thông qua tham số WER. Kết quả thử nghiệm vào tháng 10/2021 được công bố trong tài liệu [15], với 84 file audio mẫu được ghi âm. WER trung bình cho các mẫu thử nghiệm lần lượt là: Google Standard (26.79%), IBM Waston (19.56%), Google Enhanced (11.70%), VoiceGain (11.03%), Amazone (11%), Microsoft (10.41%). Với ngôn ngữ Tiếng Việt, kết quả thử nghiệm công bố tại VLSP 2019 [16], với các mẫu thử nghiệm khoảng 16 ngàn câu gồm tin tức và bài nói chuyện. WER lần lượt là: ZALO (14.36%), Viettel (27.11%), VAIS (13.7%). Một nghiên cứu công bố trong tháng 04/2021 [17], với các mẫu thử nghiệm là 05 bản tin chứa tổng số 1834 từ. Công cụ ASR được lấy từ API của các công ty cung cấp. WER cho nhận dạng Tiếng Việt bản tin NS1 (chứa 272 từ) lần lượt là: VAIS (11.09%), Viettel (16.56%), ZALO (18.26%), FPT (19.71%), Google (27.13%).

Trong nước, Công ty FPT đã đầu tư nghiên cứu phần mềm nhận dạng giọng nói để điều khiển các thiết bị IoT [18]. Một số công cụ như voicebot chứa modul nhận dạng STT của FPT.AI đạt độ chính xác 90.51% thử nghiệm với ngôn ngữ tiếng Việt [19]. Năm 2018, nhóm nghiên cứu (Viettel Cyberspace Center: VTCC) từ Công ty Viettel [20] đã phát triển phần mềm nhận dạng giọng nói dựa trên công nghệ AI, với độ chính xác theo công bố đạt 82%. Phần mềm nhận dạng giọng nói tiếng Việt của công ty VAIS, công bố 2018 cho kết quả nhận dạng âm thanh chính xác với khoảng cách xa microphone tới 7m, hoạt động trong môi trường có nhiễu và tiếng nhạc. Giải thuật được công ty VAIS phát triển trong nhiều năm [21], nhận dạng gần thời gian thực, với độ trễ nhỏ hơn 0.5s.

Tại Việt Nam, các nghiên cứu tập trung vào nhận dạng giọng nói Tiếng Việt (dịch máy), chatbox, tổng đài tự động, điều khiển và tìm kiếm bằng giọng nói. Phụ đề chủ yếu dành cho phim phát trên

JTE
HCMUTE
ISSN: 1859-1272

JOURNAL OF TECHNICAL EDUCATION SCIENCE
Ho Chi Minh City University of Technology and Education
Website: https://jte.hcmute.edu.vn/index.php/jte/index
Email: jte@hcmute.edu.vn

internet, YouTube, chưa có nghiên cứu và đánh giá ứng dụng STT để tạo phụ đề cho video trong phát sóng TH. Tại Đài THVN, qua khảo sát sơ bộ cho thấy làm phụ đề thủ công chủ yếu cho phim nước ngoài (VTVcab, SCTV), một số bản tin Tiếng Anh (kèm phụ đề Tiếng Việt – kênh VTV4). Các đơn vị khác chưa làm phụ đề hoặc làm thủ công một số chương trình. VTVgo - nền tảng phân phối nội dung VOD phong phú, chưa tích hợp cung cấp dịch vụ phụ đề video cho người xem. Việc ứng dụng kỹ thuật STT vào làm phụ đề cho các nội dung video sẽ rút ngắn thời gian triển khai, tiết kiệm chi phí và cung cấp trải nghiệm tốt hơn cho người xem.

Trong bài báo này, chúng tôi sử dụng công cụ STT do VAIS phát triển để thử nghiệm và đánh giá độ chính xác của giải thuật nhận dạng giọng nói trên các mẫu video thực nghiệm, nhằm tiết kiệm thời gian triển khai và chi phí đầu tư nghiên cứu ban đầu. Giải thuật của VAIS đã thể hiện khả năng vượt trội đạt kết quả cao nhất trong cuộc thi nhận dạng tiếng nói cả 3 miền Bắc, Trung, Nam do VLSP tổ chức năm 2018. Ngoài ra, để minh họa tính năng tạo phụ đề tự động video cho người xem TH trên nền tảng OTT, chúng tôi phát triển thêm các Tool, phần mềm và các apk để tích hợp vào hệ thống VTVgo, cung cấp thử nghiệm phụ đề tùy chọn cho các nội dung VOD cho người xem.

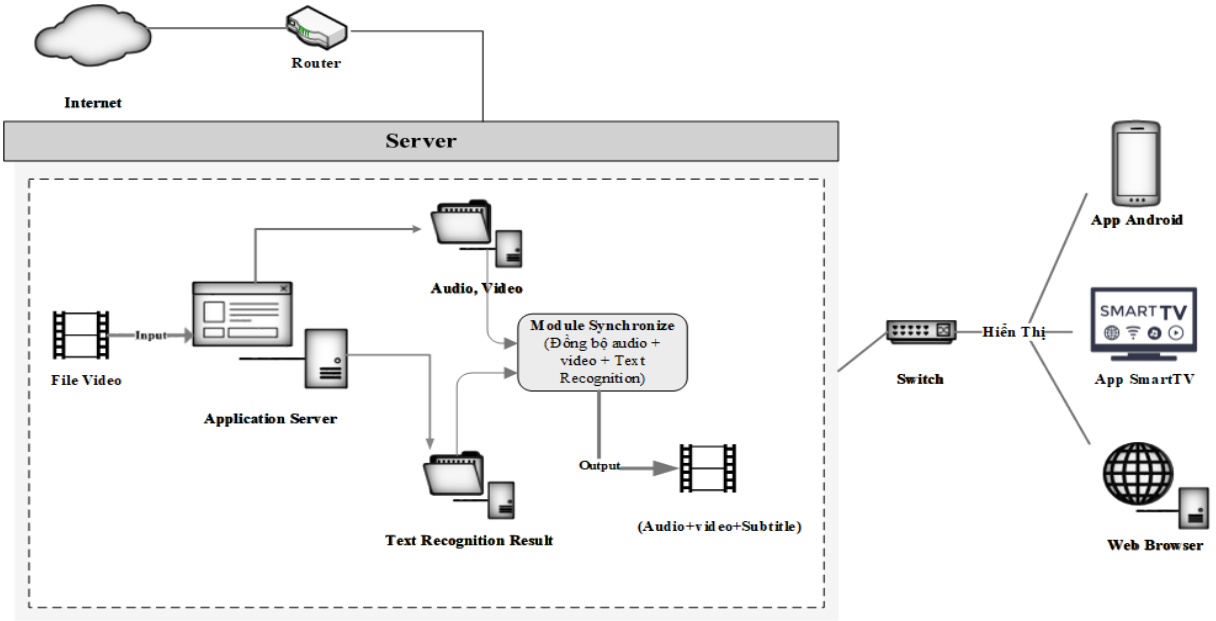
2. Mô hình thử nghiệm

Hệ thống thử nghiệm được mô tả trong **Hình 1** gồm hai phần chính. Phần máy chủ (Server) ở phía phát sẽ xử lý đầu vào các file video lấy mẫu. Đầu ra sẽ gồm video kèm phụ đề (Subtitle) được đồng bộ hóa theo mã thời gian (timecode) và hiển thị video kèm phụ đề tùy chọn để xem trước. Sơ đồ các khối chức năng như sau:

Application Server: gồm Module STT có chức năng nhận dạng giọng nói, được cài đặt trên Workstation có nhiệm vụ xử lý, nhận dạng audio ra văn bản text và xuất ra file phụ đề (định dạng *.srt).

Module tách Audio, Video (mp3 extractor): Nhiệm vụ trích xuất và tạo ra file audio (định dạng mp3) từ file video được chọn. Create Subtitles: Tạo file phụ đề (định dạng .srt), sử dụng liên kết với công cụ STT để tạo phụ đề từ file mp3 ra file sub txt. Kết quả sẽ lưu vào thư mục Text Recognition Result trong **Hình 1**. Module Synchronize: Có nhiệm vụ đồng bộ dữ liệu audio, video, file text phụ đề dựa trên timecode. Sau đó hiển thị video kèm phụ đề để kiểm tra trước khi phát sóng.

Phía client: Hiển thị video kèm phụ đề đến người dùng trên trình duyệt Web, Smartphone, SmartTV. Bao gồm các ứng dụng trên Mobile (hệ điều hành Android). Ứng dụng trên SmartTV (hệ điều hành Android). Tích hợp tính năng hiển thị phụ đề trên các trình duyệt Web (IE, Chrome, Firefox, Safari,…). Tất cả các module được tích hợp vào giao diện GUI cho thử nghiệm.



Hình 1. *Mô hình thử nghiệm đánh giá tỉ lệ lỗi từ WER*

JTE
HCMUTE
ISSN: 1859-1272

JOURNAL OF TECHNICAL EDUCATION SCIENCE
Ho Chi Minh City University of Technology and Education
Website: https://jte.hcmute.edu.vn/index.php/jte/index
Email: jte@hcmute.edu.vn

3. Phương pháp lấy mẫu và đánh giá

Kết hợp nghiên cứu lý thuyết và thực nghiệm. Xây dựng các công cụ phần mềm để đánh giá, thử nghiệm kiểm chứng, phương pháp toán học, và mô hình đánh giá độ chính xác của giải thuật nhận dạng giọng nói qua tham số: WER (tỉ lệ lỗi từ), WRR (tỉ lệ nhận dạng từ đúng), đồng bộ video/audio/text dựa trên timecode.

Bước 1: Đánh giá trong mạng nội bộ (LAN) để kiểm chứng độ chính xác giải thuật, WER, WRR, đồng bộ video/audio/text, tương thích phụ đề với các trình duyệt web (Chrome, Safari, Firefox,…). Mô hình xây dựng với các bước thực hiện và thư viện Player tương tự như hệ thống VOD.

Bước 2: Thử nghiệm và đánh giá trên hệ thống thực tế (VTVgo), ứng dụng tùy chọn hiển thị phụ đề cho người xem chạy trên Smart TV và Smart Phone Android, mô hình thử nghiệm này tương tự như cung cấp dịch vụ VOD kèm phụ đề.

Các mẫu video thử nghiệm được lấy từ một số thể loại chương trình của VTV. Hệ thống cung cấp tạo phụ đề tự động CC (Close captions) sử dụng kỹ thuật STT để chuyển giọng nói, lời thoại, âm thanh thành văn bản (text), sau đó đồng bộ với video và hiển thị lên màn hình Tivi (TV). Cấu trúc mẫu như sau:

- Lấy mẫu theo thể loại chương trình: Giải trí, Tin tức, Thời tiết, du lịch, Thời sự, Gameshow, Talkshow, Phim truyện để đảm bảo tính đa dạng của phương ngữ vùng miền, giọng Bắc – Trung – Nam trong việc đánh giá tỉ lệ lỗi từ.

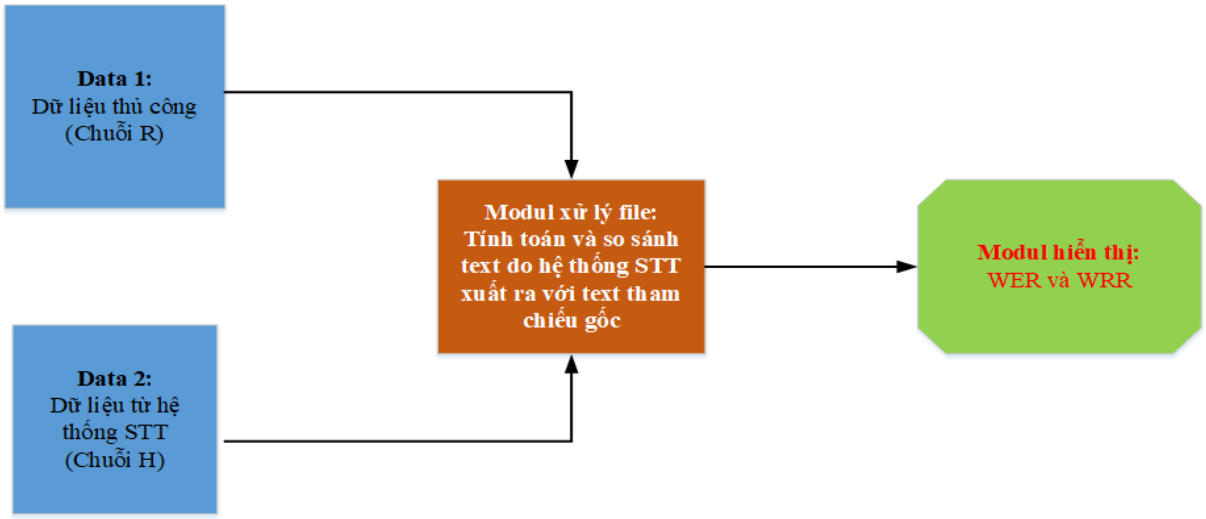
- Mỗi một chương trình lấy 3–5 mẫu, thời lượng mỗi mẫu khác nhau: 5 phút, 15 phút, 30 phút. Với các chương trình dự báo thời tiết, điểm tin thời lượng 5p đã chiếm 90% lời thoại của chương trình. Với các chương trình khác 15p-30p chiếm khoảng 50% - 80% lời thoại chương trình.

- Module chạy trên Python/MatLab để tính toán tỷ lệ lỗi từ (WER) và nhận dạng lỗi từ (WRR) trong văn bản. Quá trình xử lý dữ liệu: So sánh tỷ lệ giữa 2 đoạn văn bản từ 1 nguồn dữ liệu (video, audio). Định đạng đầu vào: File TXT chứa nội dung văn bản của Audio/Video.

Quy trình so sánh hai file text dữ liệu:

- Data 1 (Chuỗi R: text tham chiếu chuẩn): Dữ liệu thủ công được lấy từ việc nghe trực tiếp và nguồn dữ liệu biên tập có sẵn (kịch bản chương trình đã biên tập) tỉ lệ chính xác 100%.

- Data 2 (Chuỗi H: text cần đánh giá WER): Dữ liệu từ hệ thống là nguồn audio được chuyển đổi sang định dạng file txt do hệ thống nhận dạng giọng nói xuất ra.



Hình 2. *So sánh tính toán WER với file text tham chiếu*

- Module xử lý file: so khớp 2 file để đánh giá tỉ lệ lỗi từ (WER) và nhận dạng từ (WRR), từ đó hiển thị ra kết quả đánh giá.

- Module hiển thị: Hiển thị kết quả tỷ lệ lỗi từ (WER) và nhận dạng số từ (WRR), chạy trên Python/Matlab.

Tính toán WER:

Tỷ lệ lỗi từ (WER) [22] là một thước đo phổ biến về hiệu suất của hệ thống nhận dạng giọng nói và dịch máy. WER là một thông số có giá trị để so sánh các hệ thống khác nhau cũng như để đánh giá giải thuật trong một hệ thống nhận dạng giọng nói. Tỷ lệ lỗi từ (WER) là cách tiếp cận tiêu chuẩn để đánh giá hiệu suất của hệ thống nhận dạng giọng nói liên tục từ vựng lớn. Trình tự từ do hệ thống STT giả thuyết được căn chỉnh với phiên âm tham chiếu và số lỗi được tính bằng tổng các lần thay thế (S), chèn (I), và xóa (D).

Nếu có tổng số N từ trong bản phiên âm tham chiếu, thì WER được tính như sau [23]:

$$WER = \frac{S + D + I}{N} = \frac{S + D + I}{S + D + C} \quad (1)$$

trong đó, S là số lần thay thế, D là số lần xóa, I là số lần chèn, C là số từ đúng và N là tổng số từ trong văn bản tham chiếu (N=S+D+C).

Để có được một ước tính đáng tin cậy của WER, cần ít nhất 200 giờ dữ liệu thử nghiệm đối với một hệ thống STT điển hình. Trong thử nghiệm, chúng tôi sử dụng khoảng 1065h video.

Để tính toán WER, chúng tôi sử dụng chuỗi từ H (giả thuyết) và chuỗi từ R (tham chiếu), với chuỗi H là do hệ thống nhận dạng giọng nói xuất ra và chuỗi R là chuỗi text tham chiếu chuẩn dùng so sánh với giữ nguyên định dạng và timecode giống chuỗi H.

Khi đánh giá hiệu suất của hệ thống nhận dạng giọng nói, đôi khi độ chính xác của từ (WAcc) được sử dụng để thay thế:

$$WAcc = 1 - WER = \frac{N - S - D - I}{N} = \frac{C - I}{N} \quad (2)$$

WAcc là một tham số được sử dụng để đánh giá hệ thống nhận dạng giọng nói. Độ chính xác (%) phần trăm từ được định nghĩa: % WAcc = 100 - % WER. Cần lưu ý rằng độ chính xác của từ có thể bị âm. WER là thông số được sử dụng phổ biến hơn so với WAcc.

Ngoài ra, tỉ lệ nhận dạng từ WRR (Word Recognition Rate) liên quan đến số từ hệ thống ASR nhận dạng được và đưa vào tham chiếu so khớp để tính toán WER. Tỉ số WRR là số từ nhận dạng được chia cho tổng số từ tham chiếu.

$$WRR = \frac{C}{N} \quad (3)$$

Tool đánh giá lỗi từ được thực hiện bằng ngôn ngữ Python từ dự án mã nguồn mở Github [24], đây là công cụ phổ biến được khuyến cáo sử dụng cho nhận dạng lỗi từ WER không phân biệt ngôn ngữ. Nhóm nghiên cứu cũng sử dụng ngôn ngữ MatLab để kiểm tra lại WER cho kết quả tương tự với các file tham chiếu.

4. Kết quả thử nghiệm WER

Bảng 1 trình bày kết quả đánh giá tỉ lệ WER với 10 thể loại chương trình khác nhau. Mỗi chương trình được lựa chọn 5 mẫu với WER tương ứng. Ngoài ra, để thống kê độ chính xác cho từng thể loại chương trình thử nghiệm, chúng tôi sử dụng thêm tỉ lệ WER trung bình, là trung bình cộng WER của 5 mẫu thử nghiệm. Tất cả các mẫu thử nghiệm đều lấy từ hệ thống VTVgo, các file video sẽ được tách audio sau đó đưa vào hệ thống nhận dạng giọng nói. Công cụ STT sẽ trích xuất audio thành văn bản text. Văn bản từ hệ thống STT xuất ra được chuẩn hóa và giữ nguyên định dạng, timecode, sau đó so sánh với file text tham chiếu chuẩn để tính toán thông số WER.

1. **Chào buổi sáng** (thể loại Thời sự/Tin tức): Độ dài 15 phút, tỷ lệ WER trung bình là 4.3214%. Các từ tiếng Anh nhận diện sai. Một số từ mới hệ thống chưa hiểu có thể xuất ra nhiều từ khác nhau (ví dụ, Covid-19 thành Covit 19, Cô biết 19, VIC. 19, Viên 19, có viết Mười chín, cúm H 19, Huynh 19). Các

JTE
HCMUTE
ISSN: 1859-1272

JOURNAL OF TECHNICAL EDUCATION SCIENCE
Ho Chi Minh City University of Technology and Education
Website: https://jte.hcmute.edu.vn/index.php/jte/index
Email: jte@hcmute.edu.vn

tên nước ngoài đa phần hệ thống nhận diện sai như "Brazin" thành "bờ ra. Din" hoặc "Newzilan" thành "Niu Di Lân".

2. **Tài Chính Kinh Doanh** (thể loại Thời sự/Tin tức): Độ dài 15 phút, tỷ lệ WER thấp, 2.811%. Nhận diện giọng đọc của MC khá tốt. Các từ tiếng Anh, từ mới đa phần hệ thống chưa nhận diện được, ví dụ: "189.000 tỷ đồng" thành "18 mươi chín 9.000 tỷ đồng", "2014" thành "2000, mười bốn", "2015" thành "2.000 mười năm". Các từ viết tắt chưa nhận diện được, ví dụ: "SHB" thành "anh hát bê". Các âm lặp lại hệ thống chỉ nhận diện 1 âm ví dụ: "PPC" thành "PC".

Bảng 1. *Số liệu thống kê tỉ lệ WER các mẫu video thử nghiệm*

STT	Chương trình	Kênh	Thời lượng (phút)	Mẫu	WER - Mẫu 1	WER- Mẫu 2	WER- Mẫu 3	WER- Mẫu 4	WER- Mẫu 5	WERtb %
1	Chào buổi sáng	VTV1	15	CT01	3.680	5.436	6.925	3.700	1.866	4.3214
2	Tài chính kinh doanh	VTV1	15	CT02	2.396	4.452	1.970	2.223	3.014	2.811
3	Thời sự 19h	VTV1	15	CT03	1.909	4.619	2.855	2.505	3.178	3.013
4	Nhịp đập 360 độ thể thao	VTV6	30	CT04	5.288	8.322	12.151	4.744	4.814	7.0636
5	Người Việt Bốn Phương	VTV4	15	CT05	6.255	4.765	6.417	4.757	3.631	5.165
6	Dự báo thời tiết	VTV1	5	CT06	3.967	2.301	4.76	1.047	2.331	2.8812
7	Du lịch: Khám Phá Việt Nam	VTV1	15	CT07	6.684	3.941	6.044	3.414	2.787	4.574
8	Phim Truyện(người phản xử)	VTV Giải Trí	15	CT08	13.668	14.62	18.131	16.377	10.572	14.673
9	Gameshow: ai là triệu phú	VTV3	15	CT09a	15.167	14.17	14.250	11.402	7.345	12.466
	Gameshow: Chúng tôi là chiến sĩ	VTV3	30	CT09b	10.014	7.947	11.977	2.579	4.475	7.398
10	Chinh phục kỳ thi THPTQG môn GDCD	VTV7	30	CT10a	9.189	6.921	7.045	2.572	5.726	6.2906
	Chinh phục kỳ thi THPTQG môn Ngữ Văn	VTV7	30	CT10b	5.939	3.111	4.748	8.919	8.016	6.146

3. **Thời sự 19h**: Độ dài 15 phút, tỷ lệ WER thấp 3.013%. Nhận diện giọng đọc của MC khá tốt. Ngôn ngữ phỏng vấn với phương ngữ vùng miền nhận diện chưa tốt. Lỗi nhận diện số với từ ví dụ: "Một không gian", hệ thống nhận diện thành "10 gian". Các từ tiếng Anh nhận diện sai như: "Bluezone" thành "Zalo John".

4. **Nhịp đập 360 độ thể thao**: Độ dài 30 phút, tỷ lệ WER cao, 7.0636%. Các tên cầu thủ thể thao tiếng Anh hệ thống không nhận diện được hoặc nhận diện sai. Tốc độ MC nói nhanh dẫn đến tình trạng nhiều từ hệ thống không nhận diện được. Tạp âm nhiều: "Nhạc nền, phỏng vấn các cầu thủ,…" dẫn đến tỉ lệ WER cao.

5. **Người Việt bốn phương**: Độ dài 30 phút, WER=5.165%. Nhận diện giọng đọc của MC tốt. Các từ tiếng Anh nhân diện sai. MC phỏng vấn có các giọng vùng miền dẫn đến tỷ lệ lỗi từ cao. Tỷ lệ lỗi cao chủ yếu do các từ mới, các từ ngữ tiếng Anh hệ thống chưa được huấn luyện và các đoạn phỏng vấn dẫn đến tỷ lệ lỗi từ cao.

6. **Dự Báo Thời Tiết**: Độ dài 5 phút, tỷ lệ WER thấp, 2.8812%. Hệ thống nhận diện giọng đọc MC rất tốt. Tỷ lệ WER xảy ra chủ yếu ở các câu đoạn đọc nhỏ, chèn nhạc nền, hoặc các tạp âm xen lẫn.

7. **Khám Phá Việt Nam** (thể loại du lịch): Độ dài 15 phút, tỷ lệ WER=4.574%. Nhận diện giọng đọc của MC tốt. Ngôn ngữ vùng miền xuất hiện nhiều dẫn đến tỷ lệ lỗi từ cao. Chương trình chứa nhiều tạp âm. Chương trình nhiều MC, chen lẫn các giọng phỏng vấn vùng miền khác nhau, dẫn đến tỷ lệ WER cao.

JTE
HCMUTE
ISSN: 1859-1272

JOURNAL OF TECHNICAL EDUCATION SCIENCE
Ho Chi Minh City University of Technology and Education
Website: https://jte.hcmute.edu.vn/index.php/jte/index
Email: jte@hcmute.edu.vn

8. **Phim Truyện**: "Người phán xử". Độ dài 15 phút, tỷ lệ WER rất cao 14.673%. Lời thoại với nhiều từ địa phương khác nhau. Tạp âm quá nhiều (nhiều hơn 1 lời thoại, nhạc nền trong đoạn hội thoại). Tốc độ nói của lời thoại nhanh hệ thống không nhận diện được.

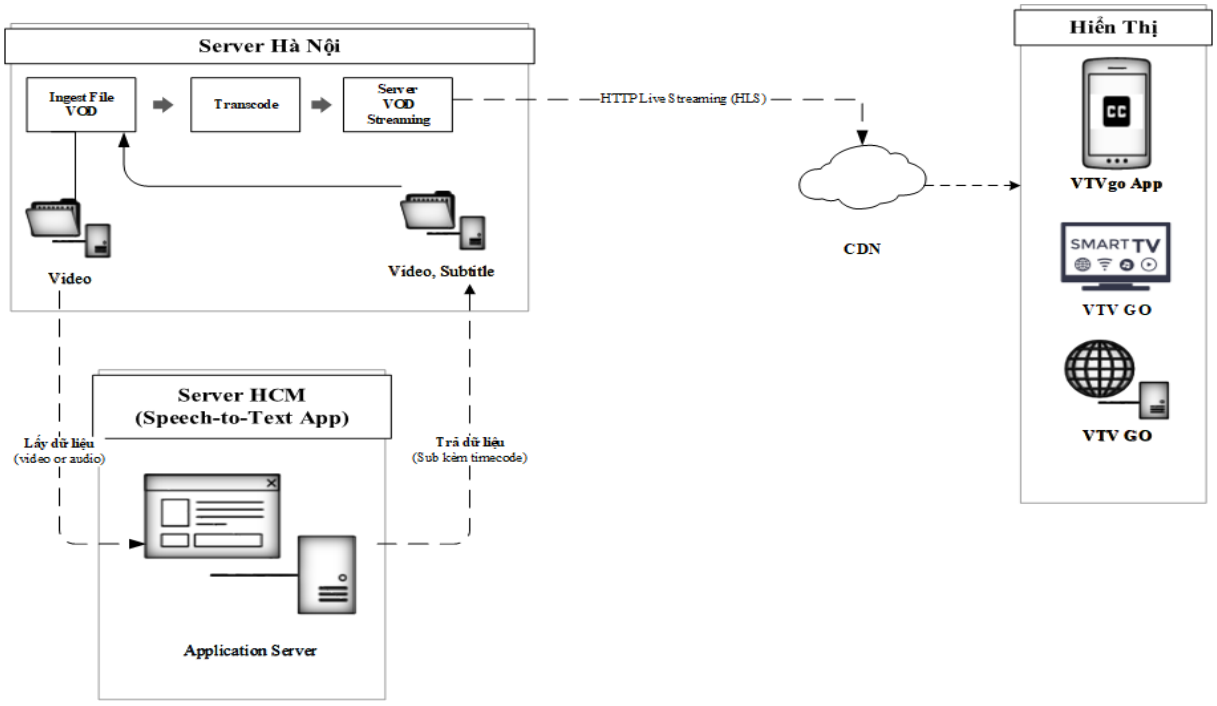
9. **Gameshow**: "Ai là triệu phú". Độ dài 15 phút, tỷ lệ WER rất cao, 12.466%. Nhận diện giọng đọc MC khá tốt. Tạp âm nhiều (âm thanh nền và tiếng vỗ tay khán giả, giọng MC, người chơi bị xen lẫn nhau). Không nhận diện được các đáp án: A, B, C, D trong chương trình. Chúng tôi là chiến sĩ: Độ dài 30 phút, tỷ lệ WER cao, 7.398%. Tạp âm quá nhiều (Nhạc nền, tiếng cười, nhiều âm thanh xen lẫn nhau). Chương trình gồm nhiều giọng vùng miền khác nhau, dẫn đến tỷ lệ lỗi từ cao. WER cao chủ yếu do tạp âm: nhiều hơn 2 giọng đọc, các biểu cảm cảm xúc người chơi dẫn đến tỷ lệ lỗi từ cao.

10. **Chinh phục kỳ thi THPTQG** môn GDCD: Độ dài 30 phút, tỷ lệ WER cao, 6.2906%. Nhận diện giọng đọc của MC tốt. Tỷ lệ lỗi từ cao chủ yếu ở các câu đáp án. Lỗi cú pháp câu sai nhiều. Nhận diện sai số với từ: "2 thành hai", "3 thành ba", "câu 15 thành câu mười 5". Lỗi các từ ngắt nghỉ trên đáp án: "A, B, C, D". Một số câu đọc nhanh, hệ thống không nhận diện kịp bị loại bỏ. Chinh phục kỳ thi THPTQG môn Ngữ Văn: Độ dài 30 phút, WER=6.146%. Nhận diện giọng đọc của MC tốt. Không nhận diện được số điểm lẻ: 0.5; 0.7; 0.1. Tốc độ giọng đọc MC nhanh. Số đọc liên tiếp không nhận diện được khoảng cách, ví dụ "0.5 0.5" thành "0.50.51".

Đánh giá tổng quan: Hệ thống nhận diện giọng chuẩn tốt, nhất là giọng MC đọc trong Studio, giọng 1 người nói tốc độ vừa phải, không có nhiều tạp âm. Hệ thống nhận diện số chưa chính xác: "ví dụ, lúc từ lúc số", một số từ Tiếng Anh, từ viết tắt chưa nhận dạng được. Tỷ lệ lỗi từ WER cao nhất trong danh sách chương trình thử nghiệm là: Phim truyện, Gameshow, chúng tôi là chiến sĩ, chinh phục kỳ thi môn THPTQG môn GDCD, người Việt bốn phương, WER từ 5%-14.6%. Chương trình tài chính kinh doanh, thời sự 19h, dự báo thời tiết tỷ lệ WER thấp từ 2.8%-4.3%. Để cải thiện độ chính xác, cần training dữ liệu cho MC/chương trình cụ thể, từ mới, từ tiếng Anh,…để hệ thống nhận dạng giọng nói học dần.

5. Thử nghiệm trên hệ thống VTVgo

Phần này trình bày kết quả thử nghiệm trên hệ thống phân phối nội dung số VTVgo của Đài THVN. Đây là một bước quan trọng nhằm đánh giá tính năng phụ đề, thử nghiệm các ứng dụng app (VTVgo Smart-Phone, VTVgo Smart TV) để minh họa ứng dụng phụ đề kèm video trên nền tảng phân phối OTT. Mô hình thử nghiệm trên VTVgo được mô tả trong **Hình 3**.



Hình 3. *Mô hình thử nghiệm trên hệ thống VTVgo*

Xử lý ở phần phát:

Bước 1: Video gốc sẽ được gửi lên 1 FTP server.

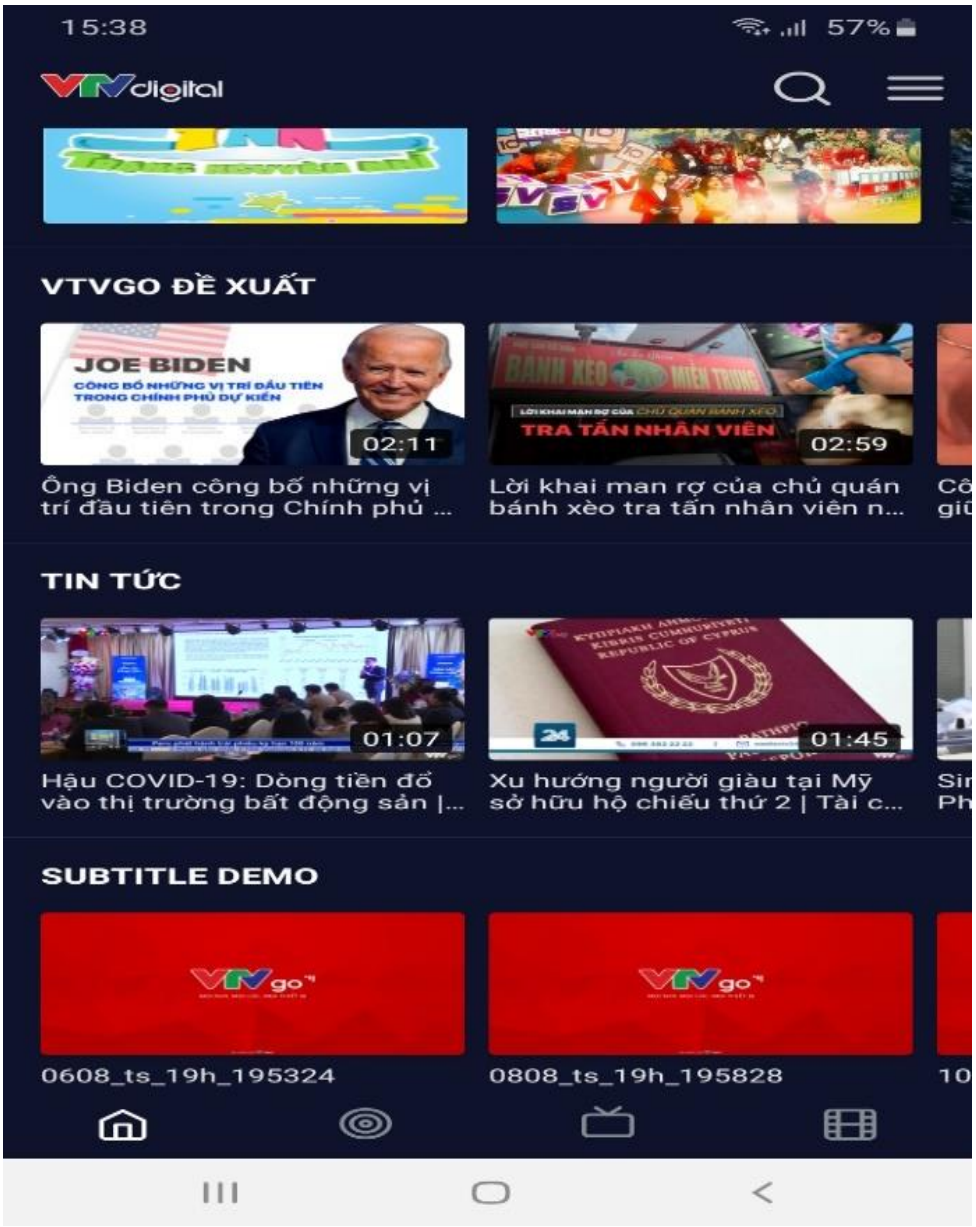
Bước 2: Video sẽ đưa vào GUI tách video->audio.

Bước 3: File audio sẽ đưa vào Tool Speech-to-text để nhận dạng và xuất ra phụ đề *.srt (đặt ở server tại TP.HCM).

Bước 4: Trả lại phụ đề *.srt về hệ thống VTVgo tại Hà Nội.

Bước 5: Thực hiện đóng gói video/sub, đồng bộ phụ đề, sử dụng giao thức HLS để phát sóng thử nghiệm.

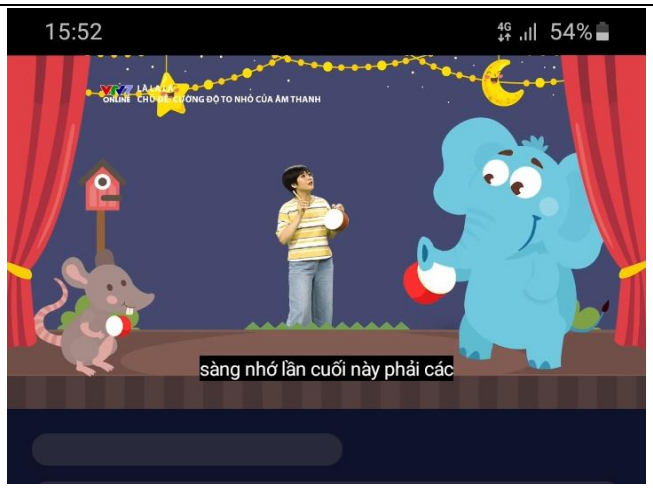
Phía user (Smart TV và Smart Phone): Nhiệm vụ dưới app client là load luồng stream HLS và thêm link subtitle vào Player, sau đó xử lý hiển thị tùy chọn (tắt/mở phụ đề) trên Player của VTVgo app. Tích hợp nút hiển thị phụ đề tùy chọn 'CC' trên App.



Hình 4. *Thư mục phụ đề video cho thử nghiệm (Subtitle Demo)*

JTE
HCMUTE
ISSN: 1859-1272

JOURNAL OF TECHNICAL EDUCATION SCIENCE
Ho Chi Minh City University of Technology and Education
Website: https://jte.hcmute.edu.vn/index.php/jte/index
Email: jte@hcmute.edu.vn



Hình 5. *Ứng dụng VTVgo app trên Smart Phone (nút CC tắt mở phụ đề tùy chọn)*

Hình 6. *Minh họa hiển thị phụ đề tùy chọn trên Smart TV (tắt mở phụ đề trên màn hình dùng Remote Tivi hoặc click vào nút CC)*

Ứng dụng VTVgo được cài đặt trên Smart TV, tích hợp nút hiển thị phụ đề "cc" tùy chọn cho người xem. Tính năng này có thể chọn on/off hiển thị phụ đề tùy chọn cho người dùng bằng cách sử dụng remote Tivi hoặc bằng click chuột vào nút "cc" trên màn hình TV hoặc Smart Phone. Các app do nhóm nghiên cứu VTV Digital xây dựng để thử nghiệm. Phần này đã trình bày kết quả thử nghiệm phụ đề cho video trên hệ thống VTVgo. Minh họa trực quan về hoạt động và ứng dụng phụ đề trên nền tảng phân phối nội dung số OTT. Các ứng dụng được viết trên Smart Phone Android và Smart TV Android có tích hợp nút hiển thị phụ đề tùy chọn cho người xem. Hạ tầng hiện nay của VTVgo đã tích hợp sẵn các hệ thống phần cứng và phần mềm để có thể triển khai ứng dụng phụ đề cho video. Tuy nhiên, độ chính xác phụ đề cần thêm thời gian cải tiến, trước mắt với nội dung VOD, cần phải có bước thủ công biên tập lại trước khi phát sóng. Các ứng dụng này có thể được kế thừa và cải tiến để triển khai về sau.

6. Kết luận

Bài báo đã trình bày kết quả thử nghiệm và kiểm chứng độ chính xác của giải thuật nhận dạng giọng nói trên các mẫu video từ hệ thống VTVgo cho một số thể loại chương trình tiêu biểu được lựa chọn. Để thử nghiệm trên hệ thống VTVgo, các ứng dụng tích hợp hiển thị phụ đề tùy chọn cho người xem trên Smart Phone và Smart TV đã được xây dựng. Tạo phụ đề tự động sử dụng các công cụ STT giảm thời gian và chi phí so với cách làm thủ công. Với thời lượng file video dài, mất nhiều giờ/ngày để làm phụ đề cho video, trong khi cùng nội dung tương tự chỉ mất vài giây hoặc vài phút để tạo phụ đề tự động bằng công cụ nhận dạng giọng nói, tiết kiệm hơn 80% thời gian và nhân sự để biên tập. Tuy vậy với độ chính xác WRR khoảng 97%−98% cho một số thể loại (thời sự, tin tức) cần thêm một số bước chỉnh sửa phụ đề thủ công trước khi phát sóng. Giải thuật cần cải thiện cho một số phương ngữ vùng miền, cần thêm dữ liệu huấn luyện để cải thiện độ chính xác. Hiển thị phụ đề tùy chọn cho người xem giúp gia tăng trải nghiệm xem tốt hơn, đặc biệt là khu vực công cộng. Từ đó có thể xem xét chọn lọc một số chương trình có độ chính xác cao để cung cấp dịch vụ tạo phụ đề tự động.

Lời cám ơn

Nghiên cứu này được thực hiện với đề tài: "*Nghiên cứu kỹ thuật nhận dạng giọng nói tạo phụ đề tự động video ứng dụng trong phân phối nội dung số tại Đài THVN*", Quỹ Phát triển KH&CN - Đài THVN

TÀI LIỆU THAM KHẢO

[1] G. Galvez, "Closed Captioning and Subtitling for Social Media," in *SMPTE 2017 Annual Technical Conference and Exhibition*, 2017.
[2] C. J. Hughes and M. Armstrong, "Automatic retrieval of closed captions for web clips from broadcast TV content," in *National Association of Broadcasters Conference*, 2015, pp. 318-324.
[3] A. Lambourne, J. Hewitt, C. Lyon, and S. J. I. J. o. S. T. Warren, "Speech-based real-time subtitling services," vol. 7, no. 4, pp. 269-279, 2004.
[4] N. Nitta and N. Babaguchi, "Automatic Story Segmentation of Closed-Caption Text for Semantic Content Analysis of Broadcasted Sports Video," in *Multimedia information systems*, 2002, pp. 110-116.
[5] T. Imai, S. Homma, A. Kobayashi, T. Oku, and S. Sato, "Speech recognition with a seamlessly updated language model for real-time closed-captioning," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.
[6] M. J. S. M. I. J. Armstrong, "Automatic recovery and verification of subtitles for large collections of video clips," vol. 126, no. 8, pp. 1-7, 2017.
[7] P. Bell *et al.*, "The MGB challenge: Evaluating multi-genre broadcast media recognition," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2015, pp. 687-693: IEEE.
[8] IBM, "AI Closed Captioning Services for Local and State Governments," vol. 2018, pp. 1-7
[9] E. Costa-Montenegro, F. M. García-Doval, J. Juncal-Martínez, and B. J. U. A. i. t. I. S. Barragáns-Martínez, "SubTitleMe, subtitles in cinemas in mobile devices," vol. 15, no. 3, pp. 461-472, 2016.
[10] M. Montagud, F. Boronat, J. Pastor, D. J. M. T. Marfil, and Applications, "Web-based platform for a customizable and synchronized presentation of subtitles in single-and multi-screen scenarios," vol. 79, pp. 21889-21923, 2020.
[11] K. J. C. Ellis, Politics and Culture, "Netflix closed captions offer an accessible model for the streaming video industry, but what about audio description?," vol. 47, no. 3, pp. 3-20, 2015.
[12] L. N. Y. Tirumala, "Captioning Social Media Video," *Public Relations Education* vol. 7, no. 1, pp. 169-187, 2021.
[13] E. B. Marrese-Taylor, Jorge A Matsuo, Yutaka, "Mining fine-grained opinions on closed captions of YouTube videos with an attention-RNN," *arXiv:02420*, 2017.
[14] P. J. L. Romero-Fresco and Communication, "Accessing communication: The quality of live subtitles in the UK," vol. 49, pp. 56-69, 2016.
[15] J. Jarmulak, "Speech-to-Text Accuracy Benchmark: Word Error Rate for major Speech-to-Text platforms," October 31, 2021.
[16] T. D. Mai Luong, "A Report on the Speech-to-Text Shared Task in VLSP Campaign 2019," presented at the VLSP, 2019.
[17] N. T. M. D. Thanh, Phan Xuan Hay, Nguyen Ngoc Quy, Dao Xuan "Đánh giá các hệ thống nhận dạng giọng nói tiếng việt (vais, viettel, zalo, fpt và google) trong bản tin," *Journal of Technical Education Science*, no. 63, pp. 28-36, 2021.
[18] D. C. Tran, D. L. Nguyen, H. S. Ha, and M. F. Hassan, "Speech Recognizing Comparisons Between Web Speech API and FPT. AI API," in *Proceedings of the 12th National Technical Seminar on Unmanned System Technology 2020*, 2022, pp. 853-865: Springer.
[19] D. C. Tran, D. L. Nguyen, M. F. J. B. o. E. E. Hassan, and Informatics, "Development and testing of an FPT. AI-based voicebot," vol. 9, no. 6, pp. 2388-2395, 2020.
[20] Q. B. Nguyen, B. Q. Dam, and M. H. Le, "Development of a Vietnamese speech recognition system for Viettel call center," in *2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*, 2017, pp. 1-5: IEEE.
[21] Q. T. Do, "VAIS-Speech: An Overview of Automatic Speech Recognition and Text-to-speech Development at VAIS," in *VLSP 2018*, Ha Noi, Vietnam, 2018.
[22] G. Saon, B. Ramabhadran, and G. Zweig, "On the effect of word error rate on automated quality monitoring," in *2006 IEEE Spoken Language Technology Workshop*, 2006, pp. 106-109: IEEE.
[23] A. Ali and S. Renals, "Word error rate estimation for speech recognition: e-WER," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2018, pp. 20-24.
[24] Github. (2021). Available: https://github.com/belambert/asr-evaluation

Phong Nguyen-Huu received the B.E. degree in Telecommunications Engineering from University of Transport and communications–Campus 2 (UTC2), Vietnam in 2006 and Master of Telecom from HCMC Posts and Telecommunications Institute of Technology (PTIT), Vietnam in 2014. From Aug 2016, he has been working toward the PhD. degree in Faculty of Telecommunications, Ho Chi Minh city University of Technology (HCMUT). Currently, he is working for Vietnamese Television (VTV). His research interests include the areas of mobile communication network (Two-way communications, Full-Duplex transmission), energy harvesting, audio/video coding and broadcast technology.

Vo Nguyen Quoc Bao received the Ph.D. degree in electrical engineering from University of Ulsan, South Korea, in 2010. Dr. Bao is an associate professor of Wireless Communications at Posts and Telecommunications Institute of Technology (PTIT), Vietnam. He is currently serving as Director of the Wireless Communication Laboratory (WCOMM). He is senior member of IEEE. He is the Technical Editor in Chief of REV Journal on Electronics and Communications. He is also serving as an Editor of Transactions on Emerging Telecommunications Technologies (Wiley ETT), and VNU Journal of Computer Science and Communication Engineering. He served as a Technical Program co-chair for ATC (2013, 2014), NAFOSTED-NICS (2014, 2015, 2016), REV-ECIT 2015, ComManTel (2014, 2015), and SigComTel 2017. His research interests include wireless communications and information theory with current emphasis on MIMO systems, cooperative and cognitive communications, physical layer security, and energy harvesting.

Tran Minh Trung received his M.Eng. degree in Bachelor of Science at University of Natural Sciences in 1998 in Vietnam. Currently, he is working for vietnamese television station in the south region. He is interested in television technology and its application in life