ĐÁNH GIÁ CÁC HỆ THỐNG NHẬN DẠNG GIỌNG NÓI TIẾNG VIỆT (VAIS, VIETTEL, ZALO, FPT VÀ GOOGLE) TRONG BẢN TIN

EVALUATION OF VIETNAMESE SPEECH RECOGNITION PLATFORMS (VAIS, VIETTEL, ZALO, FPT AND GOOGLE) IN NEWS

Nguyễn Thị Mỹ Thanh, Phan Xuân Dũng,
Nguyễn Ngọc Hay, Lê Ngọc Bích, Đào Xuân Quy
Trường Đại học Quốc tế Miền Đông, Việt Nam

Ngày toà soạn nhận bài 20/10/2020, ngày phản biện đánh giá 12/11/2020, ngày chấp nhận đăng 5/2/2021

TÓM TẮT

Bài báo này giới thiệu kết quả đánh giá các hệ thống nhận dạng giọng nói tiếng Việt (VASP-Vietnamese Automatic Speech Recognition) trong bản tin từ các công ty hàng đầu của Việt Nam như Vais (Vietnam AI System), Viettel, Zalo, Fpt và công ty hàng đầu thế giới Google. Để đánh giá các hệ thống nhận dạng giọng nói, chúng tôi sử dụng hệ số Word Error Rate (WER) với đầu vào là văn bản thu được từ các hệ thống Vais VASP, Viettel VASP, Zalo VASP, Fpt VASP và Google VASP. Ở đây, chúng tôi sử dụng tập tin âm thanh là các bản tin và API từ các hệ thống Vais VASP, Viettel VASP, Zalo VASP, Fpt VASP và Google VASP để đưa ra văn bản được nhận dạng tương ứng. Kết quả so sánh WER từ Vais, Viettel, Zalo, Fpt và Google cho thấy hệ thống nhận dạng tiếng nói tiếng Việt trong các bản tin từ Viettel, Zalo, Fpt và Google đều có kết quả tốt, trong đó Vais cho kết quả vượt trội hơn.

Từ khóa: *Xử lý ngôn ngữ tự nhiên; Nhận dạng tiếng nói; WER; tin tức; Api.*

ABSTRACT

This article introduces an evaluation of Vietnamese Automatic Speech Recognition (VASR) in the news domain from top Vietnamese speech recognition companies such as Vais, Viettel, Zalo, Fpt and top world company such as Google. To evaluate speech recognition systems, Word Error Rate (WER) coefficient with recognized text inputs from Vais VASP, Viettel VASP, Zalo VASR, Fpt VASP and Google VASP platforms were utilized. The recognized texts were acquired by using audio files in the news domain and APIs from Vais VASP, Viettel VASP, Zalo VASR, Fpt VASP and Google VASP platforms to convert from speech to text. The evaluation results obtained from WER which was applied for Vais, Viettel, Zalo, Fpt and Google, show that VASP from Viettel, Zalo, FPT and Google are adequate in which Vais is superior.

Keywords: *Natural language processing; Speech recognition; WER; News; Api.*

1. GIỚI THIỆU

Trong cuộc sống của con người, tin tức và đặc biệt bản tin truyền hình là một trong những lĩnh vực quan trọng. Tuy nhiên, các bản tin được thực hiện với nhiều ngôn ngữ khác nhau, việc này gây khó khăn cho người xem. Vì vậy, một trong những giải pháp cho vấn đề này là tự động chuyển ngôn ngữ (machine translation) các bản tin thành ngôn ngữ mà người xem mong muốn. Trong những ứng dụng thực tế, Youtube cho phép hiển thị phụ đề (subtitle) trong các videos được đăng tải bên trong nó bằng ngôn ngữ khác. Để thực hiện điều này, chúng ta phải tải phụ đề lên videos được đăng tải hoặc Youtube tự động nhận dạng giọng nói (automatic speech recognition) và chuyển thành phụ đề tương ứng với ngôn ngữ khác. Dễ dàng nhận thấy là chất lượng của phụ đề sẽ phụ thuộc vào hệ thống nhận dạng giọng nói vì nó là đầu vào. Câu hỏi đặt ra là hệ thống nào là hệ thống nhận dạng giọng nói tiếng Việt tốt nhất trong các bản tin hiện tại. Câu trả lời này sẽ được đưa ra trong bài báo này.

Trí tuệ nhân tạo ngày càng đóng vai trò quan trọng trong cuộc sống của con người bởi

Doi: https://doi.org/10.54644/jte.63.2021.46

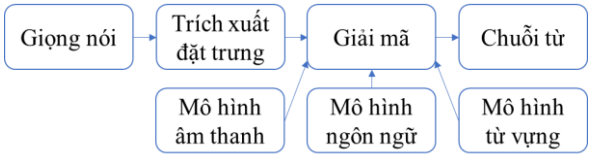
vì những tiềm năng và ứng dụng của nó. Đặc biệt những kết quả gần đây trong lĩnh vực trí tuệ nhân tạo đã làm thay đổi cách con người giao tiếp với nhau và cách con người giao tiếp với máy móc. Những hệ thống trí tuệ nhân tạo tốt nhất trên thế giới như Siri (Apple), Google Assistant (Google), Cortana (Microsoft), Alexa (Amazon) và Waston (IBM) là những trợ lý ảo có khả năng giao tiếp với con người bằng văn bản (text) hoặc giọng nói (voice). Ngoài ra còn những hệ thống như xe tự hành (self-driving-cars), tổng đài trợ lý ảo (virtual switchboard), báo nói (talking news), dịch tự động (machine translation). Các hệ thống này đều được xây dựng trên công nghệ nhận dạng giọng nói (speech recognition). Vì vậy, nhận dạng giọng nói đóng vai trò quan trọng quyết định chất lượng của các hệ thống. Câu hỏi quan trọng là hệ thống nhận dạng giọng nói nào có kết quả tốt nhất. Đối với tiếng Anh, một số nghiên cứu trong [1], [2] đưa ra kết quả so sánh giữa các hệ thống Google, IBM, Microsoft, CMU Sphinx và kết quả thể hiện Google vượt trội hơn các hệ thống khác. Đối với tiếng Việt, trong hội thảo VLSP 2019 [3] kết quả so sánh giữa Vasi, Zalo và Viettel được đưa ra; kết quả thể hiện Vais tốt hơn Zalo và Viettel. Do đó, chúng ta thấy rằng cần có một đánh giá giữa Google và các hệ thống Vais, Viettel, Zalo cho nhận dạng giọng nói tiếng Việt.

Trong bài báo này, chúng tôi giới thiệu kết quả so sánh các hệ thống nhận dạng giọng nói tiếng Việt trong lĩnh vực tin tức thông qua Api từ Vais, Viettel, Zalo, Fpt và Google. Để thực hiện, chúng tôi phát triển một công cụ để tính hệ số WER với đầu vào là các văn bản được nhận dạng từ các hệ thống Vais VASP, Viettel VASP, Zalo VASR, Fpt VASP và Google VASP. Ngoài ra, khác với kết quả trong [3], chúng tôi sử dụng các tập tin âm thanh có độ dài khoảng 20 giây với số lượng từ và số lượng câu lớn và kết quả đánh giá dựa trên API thực hiện trong tháng 9/2020 cho Vais, Viettel, Fpt, Google và tháng 12/2020 cho Zalo (demo version).

Phần còn lại của bài báo được trình bày như sau: Mục 2 mô tả năm hệ thống nhận dạng giọng nói tiếng Việt gồm Vais, Viettel, Zalo, Fpt và Google; Mục 3 mô tả phương pháp tính hệ số WER; Mục 4 trình bày kết quả từ một số bản tin; và cuối cùng Mục 5 là kết luận.

2. NHẬN DẠNG GIỌNG NÓI TIẾNG VIỆT

Thành phần của hệ thống VASR như trong **Hình 1** với cấu trúc gồm các phần: trích xuất đặc trưng (feature extraction), mô hình âm học (acoustic model), mô hình ngôn ngữ (language model), mô hình từ vựng (lexicon model) và bộ giải mã (decoder).



Hình 1. *Cấu trúc hệ thống nhận dạng giọng nói tiếng Việt*

Công nghệ lõi của các hệ thống Vais, Zalo và Viettel tại cuộc thi VASR trong Hội thảo VLSP 2019 được đưa ra trong **Bảng 1**. Trong đó, Vais, Zalo và Viettel đều sử dụng các công nghệ lõi tương tự nhau. Tuy nhiên, bài báo này đánh giá các hệ thống VASR dựa trên Api tức là các hệ thống đã triển khai thực tế nên công nghệ được triển khai thực tế có thể sẽ có sự khác biệt so với công nghệ trong **Bảng 1** vì các công ty không đề cập đến công nghệ chi tiết trong sản phẩm thương mại.

Bảng 1. *Công nghệ nhận dạng giọng nói tiếng Việt [3]*

Đặc điểm	Vais	Zalo	Viettel
Đặc trưng đầu vào	MFCC+Pitch	MFCC+Pitch	MFCC+Pitch
Tăng cường dữ liệu	Noise+RIR	Noise+RIR	Noise+RIR
Mô hình âm học	TDNN	TDNN+LSTM	TDNN+BLSTM
Ngôn ngữ	News+Conv	News+YouTube	News
Từ vựng	16k words	17k words	11k words
WER	13.7%	14.36%	27.11%

Tiếp theo chúng tôi đánh giá năm hệ thống Vais, Viettel, Zalo, Fpt và Google từ APIs:

https://vaisapis.vais.vn/analytic/v1/digitalization/audio-upsert-execute

https://viettelgroup.ai/voice/api/asr/v1/rest/decode_file

https://zalo.ai/experiments/automation-speech-recognition

https://api.fpt.ai/hmi/asr/general

https://speech.googleapis.com/v1p1beta1/speech:recognize

3. MÔ TẢ HỆ THỐNG TÍNH WER

Công thức tính Word Error Rate của một câu được nhận dạng là [4]

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

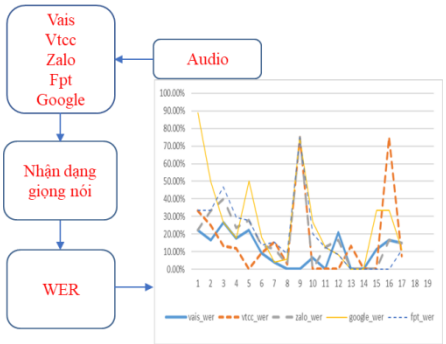
trong đó S là số từ thay thế, D là số từ bỏ đi, I là số từ chèn vào và N là tổng số từ tham khảo. Ở đây chúng ta có thể xem tổng S + D + I là số từ sai khác so với số từ tham khảo N. Cụ thể hơn, tính toán WER được đưa ra trong ví dụ:

- Thực tế (ground truth): "Hệ thống nào là hệ thống tốt nhất";
- Giả thuyết (hypothesis): "Hệ thống nào là hệ thống".

Bằng cách sử dụng công thức tính (1), chúng ta đơn giản tính được WER là 0.25 hoặc 25%.

$$WER = \frac{S + D + I}{N} = \frac{2 + 0 + 0}{8} = 0.25$$

Dựa trên so sánh WER, hệ thống đánh giá VASR được thiết kế như trong **Hình 2.** Trong đó, đầu vào của hệ thống là tập tin âm thanh (audio) chứa giọng nói cần được nhận dạng, đầu ra của hệ thống là tập tin excel chứa kết quả nhận dạng giọng nói thành văn bản. Kết quả hệ số WER được vẽ thành biểu đồ để so sánh trong năm hệ thống.



Hình 2. *So sánh hệ số WER*

Bởi vì văn bản đầu ra từ các hệ thống Vais VASP, Viettel VASP, Zalo VASR, Fpt VASP và Google VASP không đồng nhất. Để đánh giá các hệ thống VASR, chúng tôi sử dụng tập tin âm thanh có chiều dài khoảng 20 giây, đồng thời có nhiều người nói nên văn bản được nhận dạng có số lượng câu lớn hơn hai. Vì vậy, văn bản từ Vais, Zalo, Fpt và Google có dấu phẩy (,), dấu chấm (.), dấu phần trăm (%) và dấu hỏi (?) trong khi Viettel không có. Ngoài ra, văn bản từ Viettel chỉ là chữ (words), số và ngày tháng năm đều được chuyển thành chữ. Do đó, để đồng nhất kết quả và có thể so sánh được với văn bản thực tế (grounth truth), chúng tôi chuẩn hóa văn bản từ các hệ thống bằng cách loại bỏ các ký tự đặc biệt và chuyển số cũng như ngày tháng năm theo định dạng như trong ví dụ tiếp theo:

- Văn bản chưa chuẩn hóa: "Ngày *hai* tháng *mười* năm *hai nghìn không trăm hai mươi*, giá vàng giảm *hai mươi phần trăm*";
- Văn bản được chuẩn hóa: "Ngày *2* tháng *10* năm *2020* giá vàng giảm *20%*".

Chúng tôi phát triển một công cụ dựa trên kết quả từ [5] để thực hiện tính toán và so sánh WER với đầu vào là các tập tin âm thanh và văn bản tham khảo. Thuật toán của công cụ này được trình bày như sau:

```
for audio in sources do
        text = asr_api(audio)
        for i in sentences do
            sys_wer = wer(ground truth, text)
        end
end
```

Để đánh giá các hệ thống VASR trong các bản tin, chúng tôi sử dụng các bản tin trên Youtube. Sau đó chúng tôi cắt bản tin thành những tập tin âm thanh khoảng 20 giây, và thực hiện nhận dạng giọng nói trên những tập tin âm thanh đó. Vì văn bản được nhận dạng có chứa nhiều câu, mỗi câu có WER được tính dựa theo phương trình (2) trong đó "i" là câu thứ "i" trong văn bản được nhận dạng.

$$WER_i = \frac{S_i + D_i + I_i}{N_i} \quad (2)$$

Tính WER trung bình cho cả bản tin được đưa ra bởi phương trình (3) với NS là số câu trong văn bản được nhận dạng.

$$WER_M = \frac{\sum_{i=1}^{NS} WER_i}{NS} \quad (3)$$

4. KẾT QUẢ

Các bản tin được lựa chọn để đánh giá bốn hệ thống VASR được lựa chọn như sau:

- NS1 [6]: Bản tin về thể thao "*Giải vô địch quốc gia trở lại với những trận đấu đầy sôi động – VTV24*", bản tin trong phòng thu và cả ngoài sân cỏ với độ ồn thấp và có ba giọng nói khác nhau;
- NS2 [7]: Bản tin về văn hóa "*Ngôi làng của những đầu sư tử thổi lửa – VTV24*", bản tin có độ ồn cao, chứa nhiều tạp âm từ môi trường xung quanh và có bốn giọng nói khác nhau;
- NS3 [8]: Bản tin về thời tiết "*Thiệt hại ban đầu do bão số 5 tại Huế - VTV Go*", bản tin có độ ồn cao với ba giọng nói địa phương;
- NS4 [9]: Bản tin quốc tế "*Phản ứng của Quốc tế trước thông tin Tổng thống Mỹ mắc covid-19 – HTV tin tức*", độ ồn thấp, có nhiều từ và tên riêng quốc tế. Chúng tôi chia NS4 thành hai phần, NS41 là phần chỉ gồm từ và tên riêng quốc tế và NS42 là phần còn lại không chứa từ và tên riêng quốc tế.
- NS5 [10]: Bản tin với nội dung là cuộc gọi từ điện thoại "*Ông Trump mắc covid-19- Chiến dịch tranh cử Tổng thống Mỹ có thể vỡ trận – VTC Now*", độ ồn cao và có một giọng nói.

3.1 Bản tin NS1

Trước hết, chúng tôi đánh giá khả năng nhận dạng tên riêng tiếng Việt dựa trên bản tin NS1. Tên riêng tiếng Việt trong NS1 được đưa ra trong **Bảng 2**, kết quả thể hiện các công ty Việt Nam làm tốt hơn Google trong việc nhận dạng tên riêng tiếng Việt.

Bảng 2. *Nhận dạng tên riêng tiếng Việt*

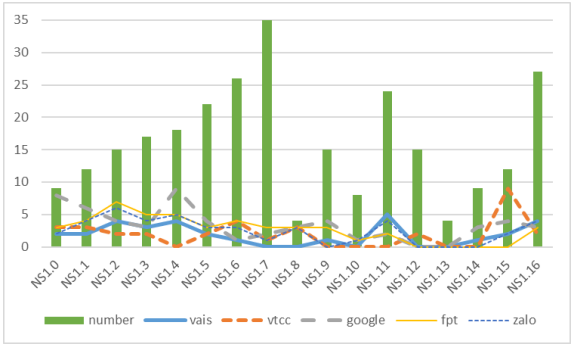
Tên riêng	Vais	Viettel	Zalo	Fpt	Google
Văn Khánh	**Văn Khánh**	**văn khánh**	**Văn Khánh**	**Văn Khánh**	**Văn Khánh**
Phúc Tịnh	**Phúc Tịnh**	**phúc tịnh**	**Phúc Tịnh**	Phước tích	Phước Thịnh
Hàng Đẫy	**Hàng Đẫy**	hàng đấy	hàng đấy	hàng đi	hàng đấy
Xuân Việt	Xuân Viện	**xuân việt**	Xuân Việt	**Xuân Việt**	công việc
Liêm Điều	lên Điều	niêm điều	liêm điền	**Liêm điều**	Liên đều

Đối với tên riêng, Vais, Zalo, Fpt và Google đều viết hoa trong khi Viettel thì không. Như đã trình bày ở phần trước, văn bản được nhận dạng của Viettel là đoạn văn bản chữ thường, không có chữ viết hoa, không có ký tự đặc biệt và số, và không có dấu tách câu.

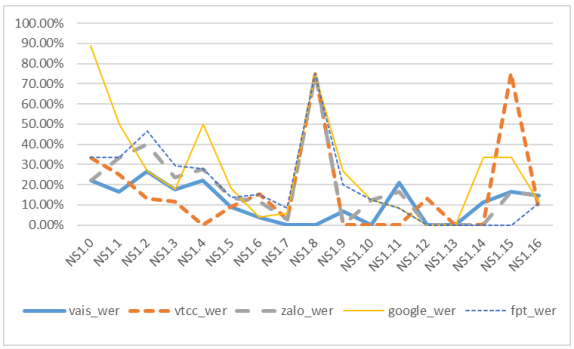
Số từ sai khác trong NS1 tương ứng với câu (NS1.0, NS1.1,…) được trình bày trong **Hình 3** với number là số từ trong câu được so sánh, và Vais, Viettel, Zalo, Fpt và Google là số từ sai khác của Vais, Viettel, Zalo, Fpt và Google tương ứng với số từ trong câu. Tổng số từ trong NS1 là 272, số từ sai khác của Vais, Viettel, Zalo, Fpt và Google tương ứng lần lượt là 31, 33, 42, 46 và 57. Vais và Viettel khác nhau tương đối nhỏ (31 và 33). Kết quả Zalo và Fpt cũng tương đối gần nhau (42 và 46). Dựa trên so sánh số từ sai khác, thứ tự sắp xếp là Vais, Viettel, Zalo, Fpt và Google.

WER của các câu trong NS1 được thể hiện trong **Hình 4**. Tại NS1.8, câu tham khảo là "trên sân *Hàng Đẫy*", kết quả tương ứng từ Vais, Viettel, Zalo, Fpt và Google là "Trên sân *Hàng Đẫy*", "chiết xuất *hàng đấy*", "chiết xuất *hàng đấy*", "Chiếc xe *hàng đi*" và "chiếc xe *hàng đấy*". Ở câu này, có tên riêng "*Hàng Đẫy*", và chỉ Vais nhận dạng chính xác nội dung câu, các hệ thống còn lại chỉ nhận được từ "*hàng*", Viettel, Zalo và Google nhận dạng giọng nói từ "*đẫy*" thành "*đấy*" và Fpt là "*đi*". Viettel và Zalo có kết quả giống nhau. Tiếp

theo, chúng ta xem xét tại NS1.15, câu tham khảo là "và sau đó là thủ môn vào thay thế *Trần Liêm Điều*". Kết quả tương ứng từ Vais, Viettel, Zalo, Fpt và Google là "và sau đó là thủ môn vào thay thế *chân lên Điều*", "thay thế *trận niêm điều*", "và sau đó là thủ môn và thay thế *trần liêm điền*", "và sau đó là thủ môn vào thay thế *Trần Liêm điều*" và "và sau đó là thủ môn và thay thế *trận Liên đều*". Ở câu này có tên riêng "*Trần Liêm Điều*", và chỉ Fpt nhận dạng chính xác nội dung của câu, Vais, Viettel, Zalo và Google không nhận dạng chính xác nội dung câu. Trong kết quả của Viettel, từ "*liêm*" được nhận dạng thành "*niêm*", đây là một trong những vấn đề khó của tiếng Việt, cách đọc gần như giống nhau nên máy rất dễ nhầm lẫn trong nhận dạng giọng nói. WER trung bình của Vais, Viettel, Zalo, Fpt và Google lần lượt tương ứng là 11.09%, 16.56%, 18.26%, 19.71% và 27.13%. Chúng ta tìm lại được kết quả sắp xếp là Vais, Viettel, Zalo, Fpt và Google theo WER trung bình.



Hình 3. *Số từ sai khác trong NS1 (trục x: câu tương ứng, trục y: số từ sai khác trong câu)*

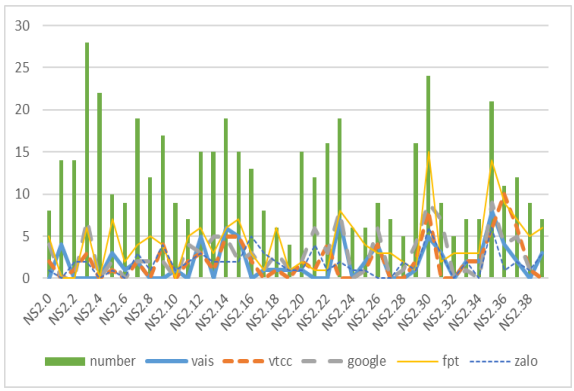


Hình 4. *WER của các câu trong NS1 (trục x: câu tương ứng, trục y: WER)*

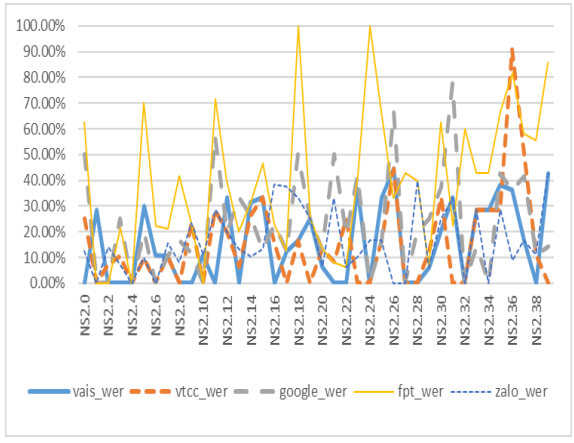
3.2 Bản tin NS2

Bản tin NS2 có độ ồn cao, nhiều tiếng tạp âm trong phóng sự. Số từ sai khác trong NS2 được mô tả trong **Hình 5**. Tổng số từ trong NS2 là 487, số từ sai khác của Vais, Viettel, Zalo, Fpt và Google lần lượt tương ứng là 74, 81, 74, 169 và 114. Số từ sai khác của Fpt trong trường hợp này cao gấp đôi các hệ thống còn lại. Kết quả này thể hiện Fpt không ổn định khi độ ồn của tạp âm lớn. Kết quả đưa ra vị trí sắp xếp là Vais, Zalo, Viettel, Google và Fpt.

WER của các câu trong NS2 được đưa ra trong **Hình 6**. WER trung bình của Vais, Viettel, Zalo, Fpt và Google lần lượt tương ứng là 15.41%, 15.63%, 16.36%, 39.25% và 23.08%. Kết quả này thể hiện tính ổn định với độ ồn từ tạp âm bên ngoài của các hệ thống Vais, Viettel và Vais. Tương tự như kết quả trong NS1, sự khác nhau giữa Vais, Viettel và Zalo không lớn. Dựa trên WER trung bình thì chúng ta có sắp xếp là Vais, Viettel, Zalo, Google và Fpt. Chú ý kết quả này khác với kết quả sử dụng số từ sai khác.



Hình 5. *Số từ sai khác trong NS2*

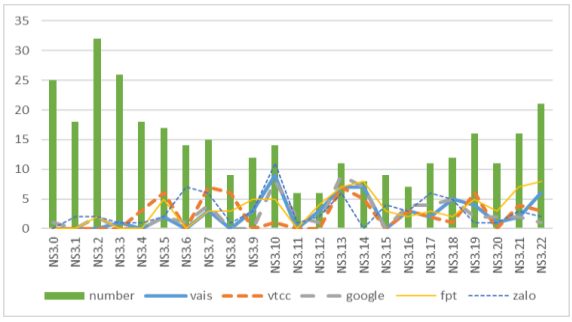


Hình 6. *WER của các câu trong NS2*

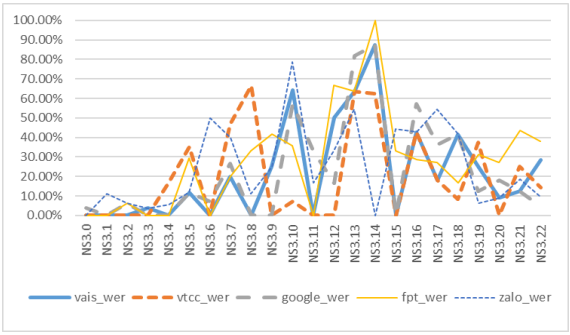
3.3 Bản tin NS3

Bản tin NS3 với giọng địa phương (giọng Huế) cùng với độ ồn cao, tốc độ giọng nói nhanh. Số từ sai khác được trình bày trong **Hình 7**. Tổng số từ trong NS3 là 334, số từ sai khác của Vais, Viettel, Zalo, Fpt và Google lần lượt tương ứng là 58, 54, 70, 75 và 57. Trong trường hợp này, không có sự sai khác lớn ở bốn hệ thống, sự sai khác giữa Vais, Viettel và Google là rất nhỏ. Kết quả sắp xếp là Viettel, Google, Vais, Zalo và Fpt.

WER của các câu trong NS3 được đưa ra trong ***Hình 8***. WER trung bình của Vais, Viettel, Zalo, Fpt và Google lần lượt tương ứng là 21.91%, 19.34%, 25.00%, 27.95% và 22.41%. Tại NS3.12-NS3.14, giọng địa phương với tốc độ nói nhanh. Tại NS3.14, câu tham khảo là "tuy nhiên *cái* mức độ chưa ở cấp độ *báo động*", kết quả đưa ra từ Vais, Viettel, Zalo, Fpt và Google lần lượt tương ứng là "tuy nhiên mức độ lực dân ở các *hội đồng*", "tuy nhiên mức độ nước dân ở *hội đồng*", "tuy nhiên *cái* mức độ lực chiến ở *khu đồng*", "tuy nhiên mức độ nổi tiếng ở các *hội đồng*", và "khi nghiên cứu mức độ nước dân ở các *hội đồng*". Chỉ có hệ thống Zalo VASR nhận dạng được giọng nói từ "*cái*". Tuy nhiên tất cả hệ thống VASR đều dạng nhầm từ "*báo động*" thành "*hội đồng*" hoặc "*khu đồng*". Mặc dù độ ồn từ tạp âm bên ngoài không bằng trong bản tin NS2, nhưng giọng địa phương không rõ làm cho các hệ thống VASR đưa ra kết quả thiếu chính xác trong nhiều trường hợp. Kết quả sắp xếp dựa trên WER trung bình là Viettel, Vais, Google, Zalo và Fpt.



Hình 7. *Số từ sai khác trong NS3*



Hình 8. *WER của các câu trong NS3*

Bảng 3. *Nhận dạng tên riêng quốc tế*

Tên riêng	Vais	Viettel	Zalo	Fpt	Google
biden	bà đinh	ba đơn	bà đinh	bà đinh	**biden**
boris johnson	**boris johnson**	boris trangxin	cris trên xin	rối rít con xin	**boris johnson**
kremlin	**kremlin**	cờ rem lin		**kremlin**	phân li
donald trump	donald trum	donald trum	rồi nào trum	đồ trâm	**donald trump**
donald trump	**donald trump**	donald trum		hội đông	**donald trump**
down jones	đầu tròn	đau tròn		**down jones**	đó cho
nasdaq	**nasdaq**	sờ tát			**nasdaq**

3.4 Bản tin NS4

Như đã trình bày trước đó, bản tin NS4 sẽ được tách thành hai phần. Phần NS41 chỉ chứa các từ và tên riêng quốc tế. **Bảng 3** trình bày kết quả nhận dạng từ và tên riêng quốc tế từ các hệ thống VASR. Kết quả này thể hiện Google và Vais có kết quả tốt hơn Viettel, Fpt và Zalo. Vais không có sự ổn định khi nhận dạng giọng nói từ "trump" thành "trum" và "trump", có sự nhầm lẫn giữa hai lần khác nhau. Nhận dạng nhầm

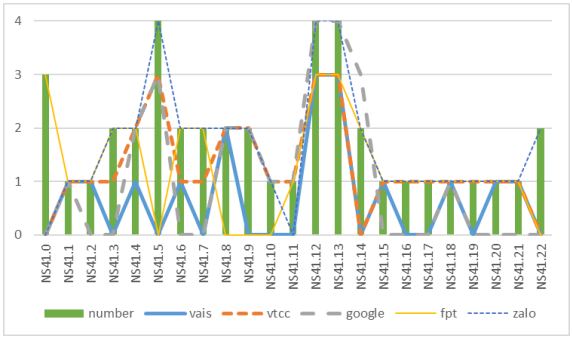
"trump" thành "trum" cũng gặp lại ở kết quả của Viettel, Zalo. Mặc dù độ chính xác nhận dạng giọng nói các từ tên riêng quốc tế phụ thuộc nhiều vào người nói với giọng chính xác hay không, nhưng nhìn chung Google vẫn có lợi thế ở lĩnh vực này vì Google hỗ trợ cả tiếng Việt và tiếng Anh cũng như có hệ thống xác định ngôn ngữ đầu vào.

Hình 9 minh họa số lượng từ sai khác và WER của các câu trong NS41. Tổng số từ quốc tế tham khảo trong NS41 là 42, số từ sai khác của Vais, Viettel, Zalo, Fpt và Google lần lượt là 16, 29, 38, 29 và 24.

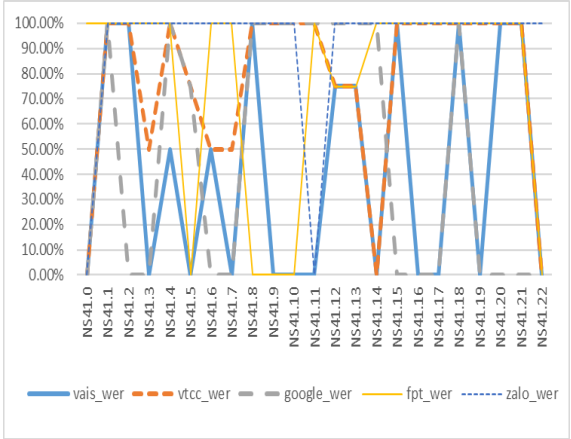
Hình 10 WER trung bình của Vais, Viettel, Zalo, Fpt và Google lần lượt tương ứng là 41.30%, 77.17%, 91.30%, 76.09% và 48.91%. Mặc dù kết quả của Vais và Google tốt hơn Viettel, Zalo và Fpt nhưng tỉ lệ nhận dạng giọng nói từ và tên riêng quốc tế với sai số quá lớn. Kết quả thể hiện là các hệ thống VASR chưa nhận dạng được giọng nói với từ và tên riêng quốc tế trong bản tin tiếng Việt.

Tách phần NS41 chứa từ và tên riêng quốc tế trong NS4, phần NS42 chỉ chứa từ tiếng Việt. Tổng số từ trong NS42 là 335, số từ sai khác của Vais, Viettel, Zalo, Fpt và Google tương ứng là 18, 12, 79, 48 và 49. WER trung bình của Vais, Viettel, Zalo, Fpt và Google tương ứng lần lượt là 5.45%, 3.6%, 31.24%, 19.36% và 22.91%. Trừ Zalo, bốn hệ thống VASR đều đạt được kết quả tốt, đặc biệt là Viettel và Vais.

Tổng hợp NS41 và NS42, tổng số từ trong NS4 là 377. Số từ sai khác của Vais, Viettel, Zalo, Fpt và Google lần lượt tương ứng là 34, 41, 117, 77 và 73. WER trung bình của Vais, Viettel, Zalo, Fpt và Google lần lượt tương ứng là 17.76%, 28.86%, 51.86%, 38.36% và 31.84%. Dựa trên số từ sai khác và WER trung bình, chúng ta có thứ tự là Vais, Viettel, Google, Fpt và Zalo.



Hình 9. *Số từ sai khác trong NS41*

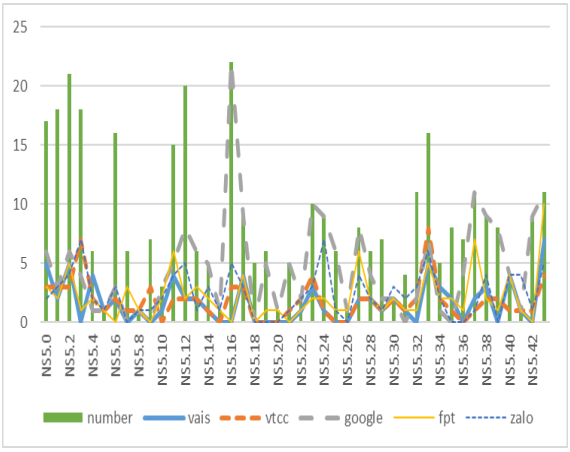


Hình 10. *WER của các câu trong NS41*

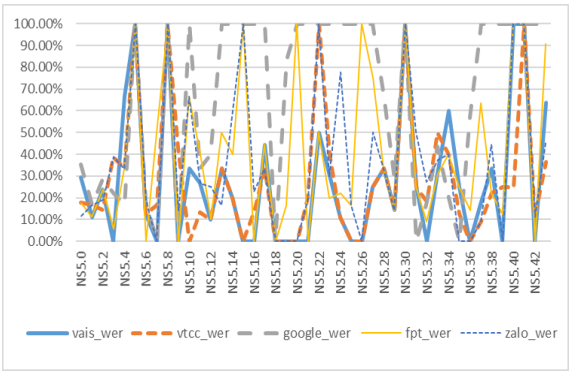
3.5 Bản tin NS5

Bản tin NS5 là một cuộc phỏng vấn thông qua điện thoại với độ ồn và nhiễu lớn. Số từ sai khác được thể hiện trong **Hình 11**. Tổng số từ trong NS5 là 364, số từ sai khác của Vais, Viettel, Zalo, Fpt và Google tương ứng lần lượt là 74, 80, 104, 95 và 206. Do ảnh hưởng bởi nhiễu và giọng bị biến dạng, Google không có kết quả tốt trong khi Vais, Viettel, Fpt và Zalo có kết quả chấp nhận được. Các hệ thống VASR được sắp xếp là Vais, Viettel, Fpt, Zalo và Google.

WER của các câu trong NS5 được đưa ra trong ***Hình 12***. WER trung bình của Vais, Viettel, Zalo, Fpt và Google lần lượt tương ứng là 27.99%, 28.06%, 44.56%, 40.33% và 66.57%. Dựa trên kết quả so sánh trung bình WER, chúng ta có kết quả sắp xếp các hệ thống VASR là Vais, Viettel, Fpt Zalo và Google. Chúng ta có kết quả tương tự như sử dụng số từ sai khác.



Hình 11. *Số từ sai khác trong NS5*



Hình 12. *WER của các câu trong NS5*

3.6 Tổng hợp

Kết quả tổng hợp so sánh các hệ thống VASR dựa trên số từ sai khác và WER trung bình được trình bày trong **Bảng 4.** Vais và Viettel có kết quả tương đối gần nhau (271 và 289). Zalo, Fpt và Google có kết quả độ chính xác thấp hơn Vais và Viettel. Tỉ lệ sai khác của Google (509) gần gấp đôi của Vais (271).

Kết quả đánh giá các hệ thống Vais, Viettel, Zalo, Fpt và Google trong việc nhận dạng giọng nói tiếng Việt trong các bản tin được đưa ra mang tính tương đối với các mẫu là các bản tin truyền hình trên Youtube. Mặc dù số lượng mẫu còn khiêm tốn nhưng cũng đủ để thấy là kết quả thể hiện Vais vượt trội hơn các hệ thống còn lại.

Kết quả bài báo là một tham khảo giúp các nhà phát triển lựa chọn hệ thống nhận dạng giọng nói tiếng Việt trong số các hệ thống Vais, Viettel, Zalo, Fpt và Google để phát triển và triển khai các ứng dụng như dịch tự động, trợ lý ảo dựa trên giọng nói, tổng đài tự động.

Bảng 4. *Đánh giá các hệ thống VASR*

	Số từ	**Vais**	**Viettel**	**Zalo**	**Fpt**	**Google**
NS1	272	**31-11.09%**	33-16.56%	42-18.26%	46- 19.71%	57-27.13%
NS2	487	**74-15.41%**	81-15.63%	74-16.36%	169-39.25%	114-23.08%
NS3	334	58-21.91%	**54-19.34%**	70-25.00%	75-27.95%	57-22.41%
NS4	377	**34-17.76%**	41-28.86%	117-51.86%	77-38.36%	73-31.84%
NS5	364	**74-27.99%**	80-28.06%	104-44.56%	95-40.33%	206–66.57%
Tổng	**1834**	**271**	289	407	462	509

5. KẾT LUẬN

Bài báo giới thiệu phương pháp đánh giá và đưa ra kết quả so sánh các hệ thống nhận dạng giọng nói tiếng Việt gồm Vais, Viettel, Zalo, Fpt và Google trong lĩnh vực bản tin truyền hình. Kết quả của Viettel, Zalo, Fpt và Google là chấp nhận được nhưng Vais có kết quả tốt hơn mặc dù không có khác biệt quá lớn giữa Vais và Viettel.

Trong nghiên cứu tiếp theo, chúng tôi sẽ đánh giá các hệ thống nhận dạng giọng nói tiếng Việt trong các lĩnh vực khác như phim ảnh, âm nhạc.

TÀI LIỆU THAM KHẢO

[1] V. Këpuska and G. Bohouta, *Comparing speech recognition systems (Microsoft API, Google API and CMU Sphinx)*, Int. J. Eng. Res. Appl, 7(03), pp. 20-24. 2017.
[2] F. Filippidou and L. Moussiades, *A Benchmarking of IBM, Google and Wit Automatic Speech Recognition Systems*, IFIP International Conference on Artificial Intelligence Applications and Innovations, pp. 73-82, 2020.
[3] L.C. Mai and D.Q. Truong, *Report on the Speech-to-Text Shared Task in VLSP Campaign 2019*, Vietnamese Language Signal Processing, 2019. (https://vlsp.org.vn/sites/default/files/2019-10/VLSP2019-ASR-summary.pdf)
[4] A. C. Morris, V. Maier and P. Green, *From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition*, Eighth International Conference on Spoken Language Processing, pp. 2786-2768, 2004.
[5] Jitsi, *JiWER: Similarity measures for automatic speech recognition evaluation*. https://github.com/jitsi/jiwer
[6] *Giải vô địch quốc gia trở lại với những trận đấu đầy sôi động – VTV24* https://youtu.be/N2FfBEWO84A
[7] *Ngôi làng của những đầu sư tử thổi nữa – VTV24* https://youtu.be/YZc5TiXi_DE
[8] *Thiệt hại ban đầu do bão số 5 tại Huế - VTV Go* https://youtu.be/kqnmPdwk62A
[9] *Phản ứng của Quốc tế trước thông tin Tổng thống Mỹ mắc covid-19 – HTV tin tức* https://youtu.be/k6OTsmpKtbc
[10] *Ông Trump mắc covid-19-Chiến dịch tranh cử Tổng thống Mỹ có thể vỡ trận – VTC Now* https://youtu.be/QehJIcATgH8

Tác giả chịu trách nhiệm bài viết:
TS. Đào Xuân Quy
Trường Đại học Quốc tế Miền Đông
Email: quy.dao@eiu.edu.vn