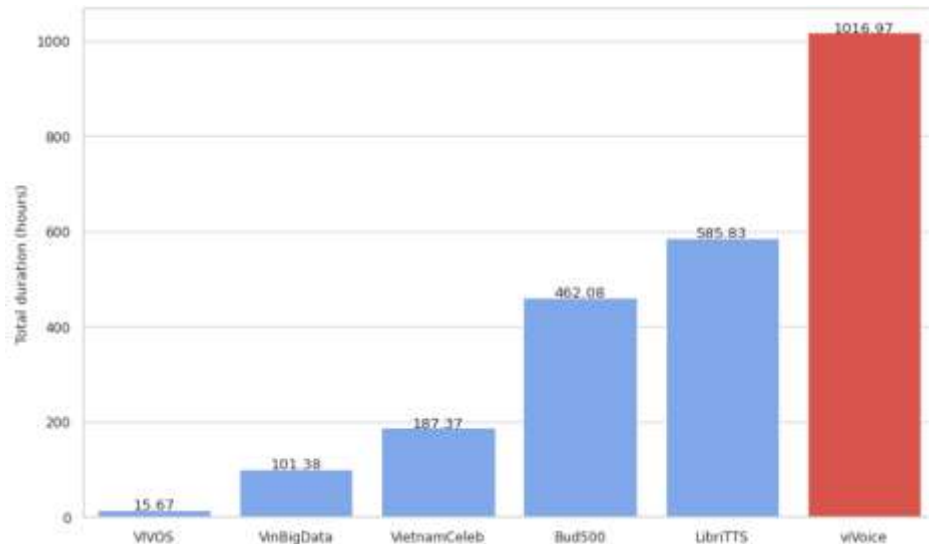


# viVoice

## Hành trình tạo ra bộ dữ liệu giọng nói tiếng Việt 1000 giờ với chi phí 0 đồng



Minh họa: kích thước của viVoice so với các bộ dữ liệu khác  
Cập nhật lần cuối tháng 6 năm 2024



# Giới thiệu viVoice

- viVoice là bộ dữ liệu 1000 giờ giọng nói tiếng Việt được chia sẻ cho cộng đồng nghiên cứu AI vào tháng 6 năm 2024.
- Tác vụ: dùng để huấn luyện các mô hình chuyển đổi văn bản thành giọng nói tiếng Việt với khả năng sao chép giọng.
  - Ứng dụng: review phim, video top top, tin tức, pod cast, sách nói v.v
- Ví dụ đầu ra của mô hình viXTTS huấn luyện bằng viVoice:



# Bối cảnh lịch sử

- Tháng 12 năm 2023: làm đề án “Chatbot **2 ngôn ngữ Anh - Việt** với **giọng nói tự nhiên**” -> cần phần chuyển văn bản thành giọng nói (TTS) **hoạt động tốt trên 2 ngôn ngữ và phải tự nhiên**.
- Những mô hình TTS như thế này với ngôn ngữ **giàu tài nguyên** như tiếng Anh thì đã tồn tại một khoảng thời gian và chúng rất tốt (Tortoise, XTTS, StyleTTS, HierSpeech...)

-> Mình nghĩ có thể *dễ dàng* làm cho tiếng Việt... **nhưng đó là một nước đi sai...**



# Vì tiếng Việt không có dữ liệu public phù hợp

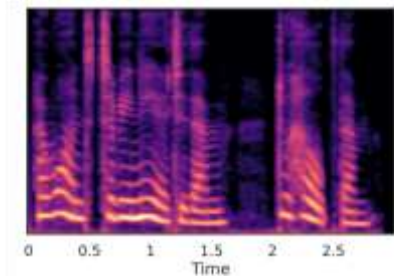
Dataset	Year	Task	Speakers	Samples	Hrs	SR
VIVOS [14]	2016	ASR	65	12420	16	16000
Common Voice [66]	2019	ASR	276	16175	18	48000
FOSD [15]	2018	ASR	N/A	25921	30	48000
VinBigdata100h [16]	2020	ASR	N/A	56427	101	16000
VietTTS [20]	2022	TTS	1	22884	36	22050
VietBibleVox [19]	2023	TTS	1	29185	67	48000
Vietnam-Celeb [68]	2023	SR/SV	1000	87140	187	16000
BUD500 [18]	2024	ASR	N/A	649158	462	16000

Bảng: Thống kê các bộ dữ liệu giọng nói tiếng Việt (không phải tất cả)

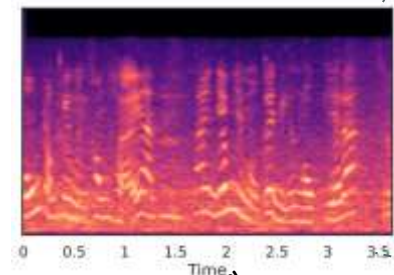
Cập nhật lần cuối: tháng 04/2024

# Ví dụ về chất lượng âm thanh

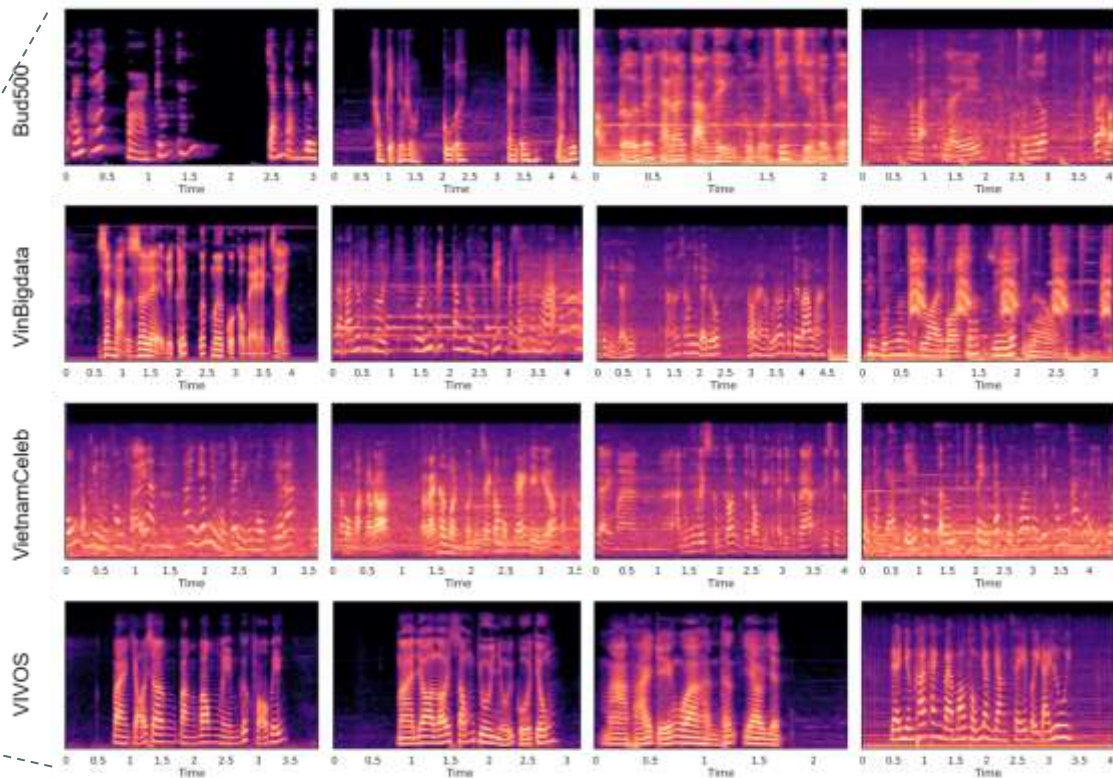
Spectrogram Visualization



Sạch



Ồn



Dữ liệu nhận diện giọng nói (STT) thường có tiếng ồn, phân giải thấp -> không phù hợp

cho TTS

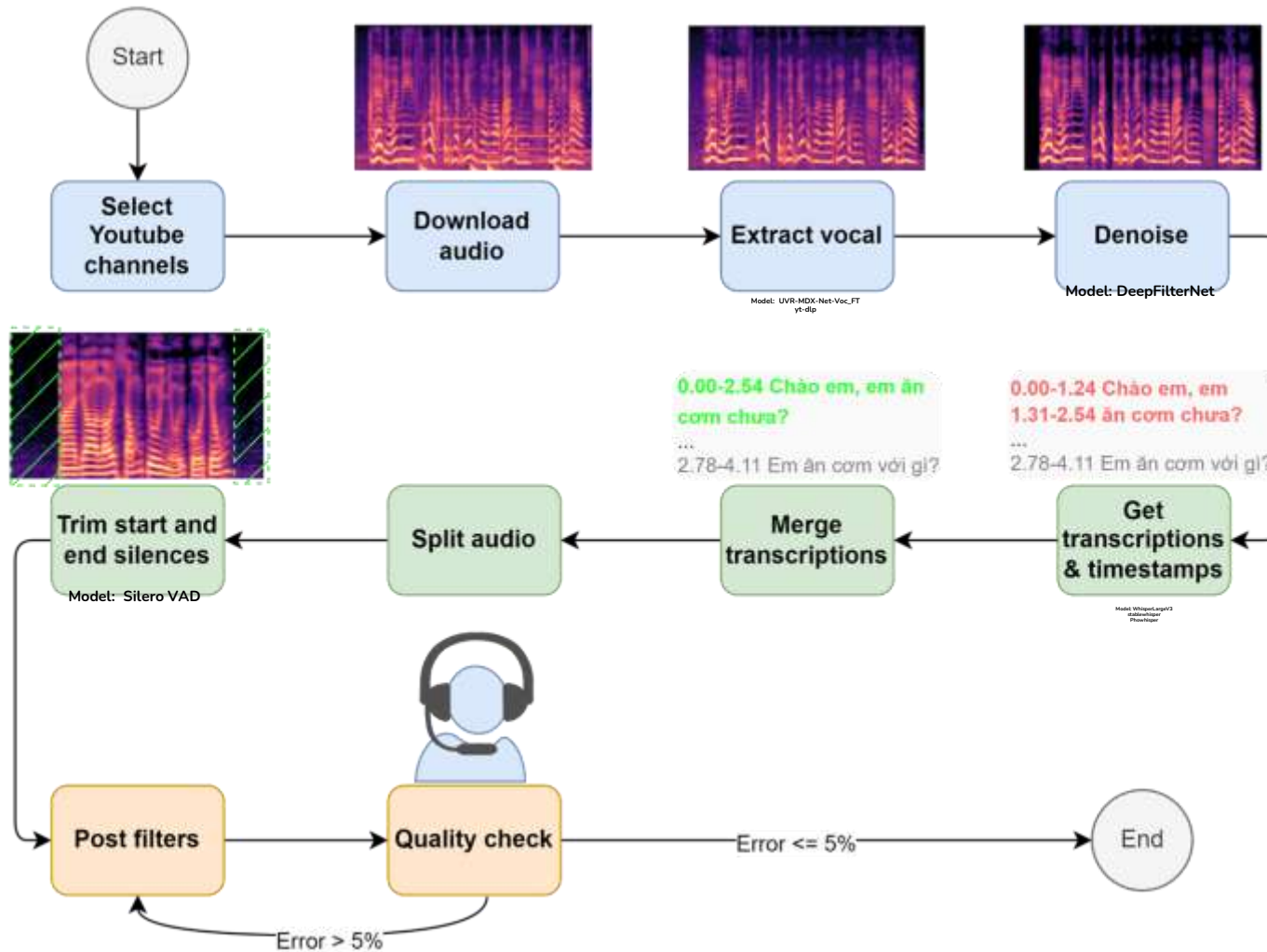
# Bắt đầu hành trình 0 đồng

- Vừa không có data vừa không có mô hình luôn -> thế đường cùng, bắt buộc phải làm đại đi hoặc cả team rút đồ án :v
- **Tự xây pipeline xử lý dữ liệu trên chiếc máy tính có sẵn** (phần cứng phổ thông)
  - GPU: 1x 3060Ti 12GB
  - CPU: I5 13500
  - SSD: 1TB





# Pipeline



**Giai đoạn 1:**  
Làm sạch âm thanh

**Giai đoạn 2:** tạo chữ  
tiếng Việt và cắt âm  
thanh

**Giai đoạn 3:** Kiểm  
tra chất lượng thủ  
công

# viVoice Pipeline – Hậu trường

## Nguồn dữ liệu (chọn hoàn toàn thủ công)

- 187 kênh youtube với chất lượng giọng nói tốt, không có nhiều người nói cùng lúc trong 1 video.
- Chọn kênh từ trí nhớ và hỏi bạn bè
- Kiểm tra chất lượng kênh thủ công



Nguồn? Trust me bro

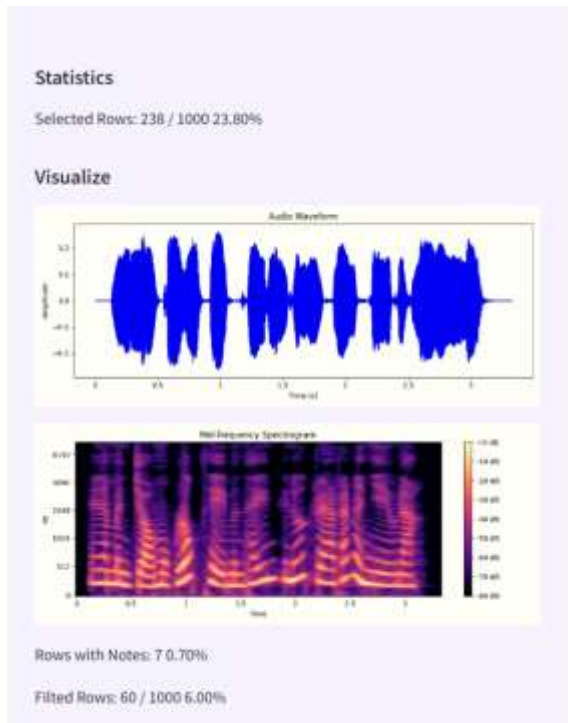


# viVoice Pipeline – Hậu trường

- Bắt đầu từ những bài báo về data như LibriTTS, LibriTTS-R,...
- Mỗi bước trong pipeline đều được chọn lựa kỹ từ các mô hình có sẵn, là quá trình thử/sai nhiều lần
- Đặc biệt quan trọng: output của mỗi bước trong pipeline đều được lưu lại để dễ dàng chạy lại khi tìm được model tốt hơn
- Kiểm tra chất lượng âm thanh thủ công và viết các quy luật lọc dữ liệu là bước quan trọng nhất và cũng tốn thời gian nhất

# Đảm bảo chất lượng thủ công

Làm nhanh một ứng dụng Streamlit để đảm bảo chất lượng



DATA DIR

/home/thinhpg/data/youtube-01\_vad\_dfn/

Select a file

/home/thinhpg/data/youtube-01\_vad\_dfn/20240301\_metadata\_inspect\_1000\_01.csv

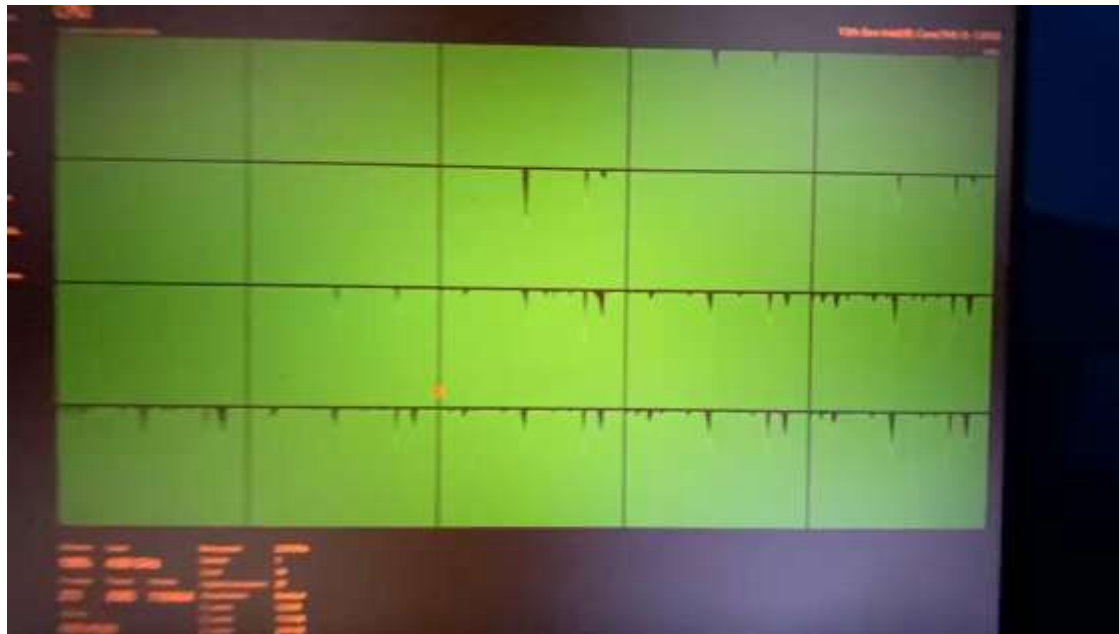
Select columns to display

Select × text × Notes × filter ×

	Select	text	Notes	filter
974	<input type="checkbox"/>	Thì ghi chú làm gì cho mất thời gian?		<input checked="" type="checkbox"/>
945	<input type="checkbox"/>	Theo các bạn, liệu rằng trong tương lai những nguyên tắc củ		<input checked="" type="checkbox"/>
908	<input checked="" type="checkbox"/>	Không.		<input checked="" type="checkbox"/>
896	<input checked="" type="checkbox"/>	Bên cạnh đó, tài liệu cũng gợi ý rằng trong kịch bản này True	Thiếu chữ cuối	<input checked="" type="checkbox"/>
853	<input checked="" type="checkbox"/>	Nên là mình trộn 1 lúc thì mình thấy cái nước ra rất là nhiều		<input checked="" type="checkbox"/>
832	<input type="checkbox"/>	Ừ mà thật ra thì ý đồ của hai đồng này là như thế thật mà.		<input checked="" type="checkbox"/>
800	<input checked="" type="checkbox"/>	D&G.		<input checked="" type="checkbox"/>
795	<input type="checkbox"/>	Không anh phải viết một chữ ra cơ.		<input checked="" type="checkbox"/>
786	<input checked="" type="checkbox"/>	Nà, gan sẵn nà.		<input checked="" type="checkbox"/>
765	<input type="checkbox"/>	Brian vẫn làm như những gì mà cậu ta được bảo và đưa cho		<input checked="" type="checkbox"/>

Activate Windows  
Go to Settings to activate Windows.

# Kinh nghiệm 101



**Bài học 1: Hãy trân trọng giấc ngủ của bạn**



**Bài học 2: Ổ cứng càng nhiều dung lượng càng tốt**

# Mô hình viXTTS

- Mô hình gốc: XTTSv2 (model tốt nhất trên BXH lúc đó có thể finetune được)
- Dữ liệu: viVoice
- Là mô hình TTS sao chép giọng nói **public** đầu tiên cho tiếng Việt.

## 🏆 Leaderboard

Vote to help the community determine the best text-to-speech (TTS) models.

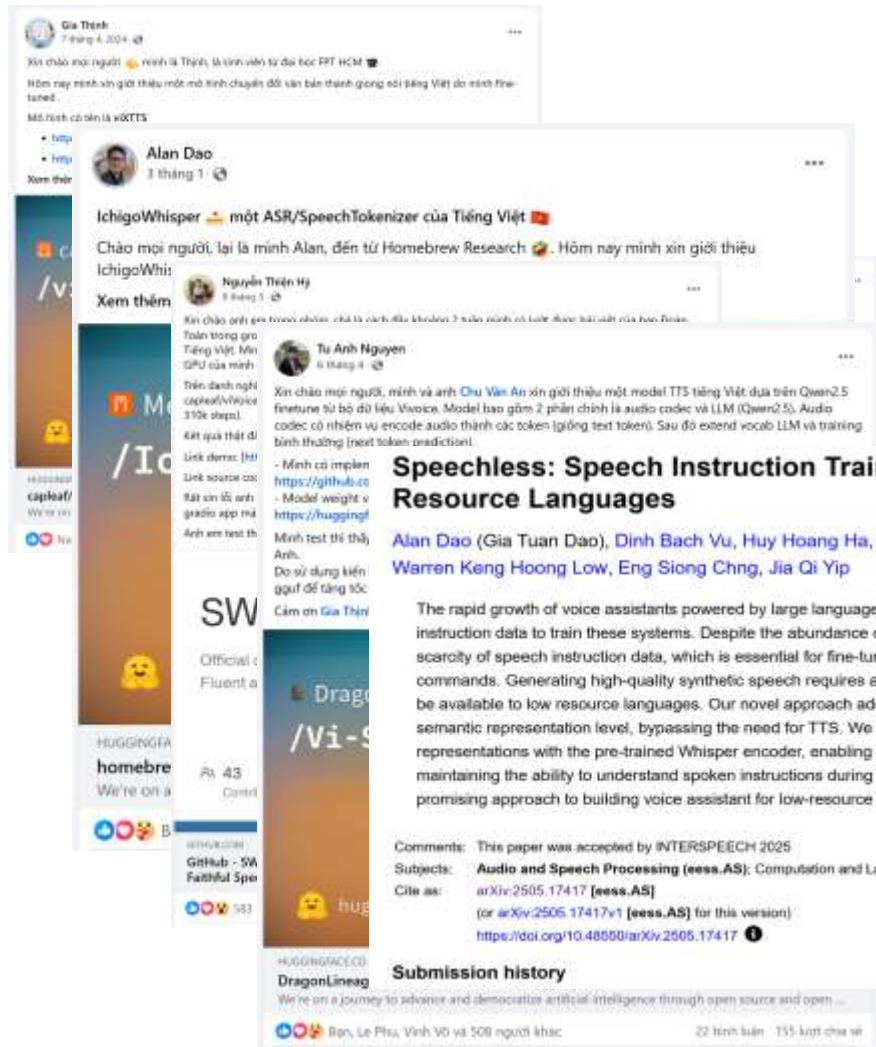
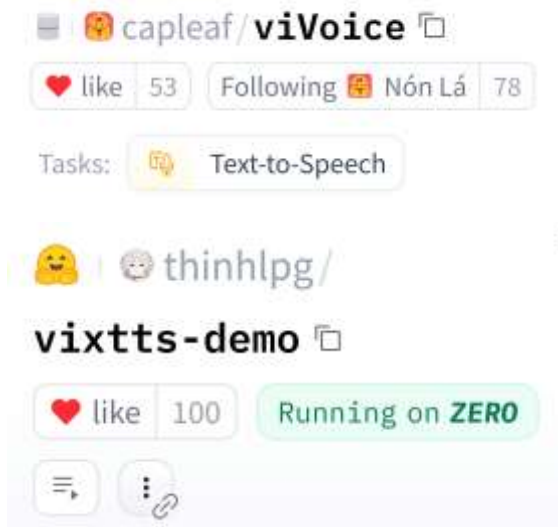
The leaderboard displays models in descending order of how natural they sound (based on votes cast by the community).

Important: In order to help keep results fair, the leaderboard hides results by default until the number of votes passes a threshold. Tick the [Reveal preliminary results](#) to show models without sufficient votes. Please note that preliminary results may be inaccurate.

order	name	score	votes
#1	ElevenLabs	1286	4499
#2	XTTSv2	1232	4184
#3	MetaVoice	1191	4348
#4	OpenVoice	1191	3999
#5	WhisperSpeech	1182	4185
#6	Phone	1116	4417

# Đóng góp nhỏ nhỏ

- Bộ dữ liệu viVoice
- Mô hình viXTTS



## Speechless: Speech Instruction Training Without Speech for Low Resource Languages

Alan Dao (Gia Tuan Dao), Dinh Bach Vu, Huy Hoang Ha, Tuan Le Duc Anh, Shreyas Gopal, Yue Heng Yeo, Warren Keng Hoong Low, Eng Siong Chng, Jia Qi Yip

The rapid growth of voice assistants powered by large language models (LLM) has highlighted a need for speech instruction data to train these systems. Despite the abundance of speech recognition data, there is a notable scarcity of speech instruction data, which is essential for fine-tuning models to understand and execute spoken commands. Generating high-quality synthetic speech requires a good text-to-speech (TTS) model, which may not be available to low resource languages. Our novel approach addresses this challenge by halting synthesis at the semantic representation level, bypassing the need for TTS. We achieve this by aligning synthetic semantic representations with the pre-trained Whisper encoder, enabling an LLM to be fine-tuned on text instructions while maintaining the ability to understand spoken instructions during inference. This simplified training process is a promising approach to building voice assistant for low-resource languages.

Comments: This paper was accepted by INTERSPEECH 2025

Subjects: **Audio and Speech Processing (eess.AS)**; Computation and Language (cs.CL); Sound (cs.SD)

Cite as: [arXiv:2505.17417 \[eess.AS\]](https://arxiv.org/abs/2505.17417)

(or [arXiv:2505.17417v1 \[eess.AS\]](https://arxiv.org/abs/2505.17417v1) for this version)

<https://doi.org/10.48550/arXiv.2505.17417>

### Submission history

We're on a journey to advance and democratize artificial intelligence through open source and open ...

Đào, Lê Phú, Vĩnh Vũ và 500 người khác

22 bình luận 155 lượt chia sẻ

# Bonus: Làm vợ ảo (waifu)?



# Lời kết

- Bạn không cần phần cứng xịn để bắt đầu. Bạn cần mục tiêu, đủ kiến thức, một ít áp lực, niềm tin, đam mê, sự kiên trì và sức mạnh tình bạn.
- Đừng ngại chia sẻ sản phẩm hay công trình của mình với cộng đồng, và bạn sẽ thấy vô vàn cơ hội bất ngờ mở ra. Những buổi AI Meetup nơi chơi tuyệt vời để bạn làm điều đó.

