COMPARISON OF STT FPT.AI VS VIETTEL.AI

Luc Duong Hai,
Software Engineering FPT, FPT University
Hanoi, VietNam
Email:haildhe141223@fpt.edu.vn,

Nguyen Long Phuong
Software Engineering FPT, FPT University
Hanoi, VietNam
Email:phuongnlhe141219@fpt.edu.vn

Ha Duc Hanh
Software Engineering FPT, FPT University
Hanoi, VietNam
Email:hanhhdhe141144@fpt.edu.vn,

Nguyen Huu Trung
Software Engineering FPT, FPT University
Hanoi, VietNam
Email:trungnhhe141239@fpt.edu.vn

***Abstract*—Currently, Artificial Intelligence (AI) is very much concerned and applied to many different sectors, industries and life, especially natural language processing. In this article, our team will show how to process an audio file into a continuous audio clip by cutting out unnecessary silence, and pauses. After that use fpt.ai to convert audio to text. Thereby creating a program to help improve the quality of sound files up to 80% and music files up to 20%..**

We used Fpt.ai API to detect audio and music after that compared with Viettelgroup.ai. The conclusion was found which one better included simple audio and music files.

Hopefully this article will help you have more ideas for developing natural language processing projects.

***Keywords*—librosa, voice, silence, split, fpt.ai, viettelgroup.ai**

I. Introduction

In this project we will create a program that can download audio. Detect play, stop and analyzing the lyrics of audio file and music using commands on the linux operating system and python. After using our program, it will detect extractly silence, reduce almost noise, and also it can download song by original link.

Ability audio processing in Fpt.ai better than Viettelgroup.ai. In the other hand, Viettelgroup.ai has music processing better than Fpt.ai.

II. Methodology and Testing

A. Requirements

In order to make a multiple continuous-speaking files without any silences, we need to prepare the following:

- Open Google Colaboratory [1].
- Import necessary libraries includes librosa [2], IPython, noisereduce, ffmpeg, scipy, AudioSegment [3] …
- Using Fpt.ai [4] to detect a **.mp3** file then store the results to data texts .
- Comparison between (Fpt.ai) and (Viettelgroup.ai [5])

B. Main Idea

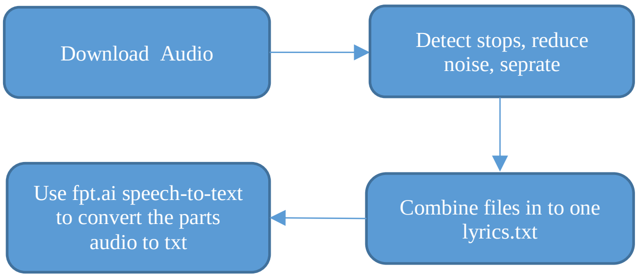

Download Audio
Detect stops, reduce noise, seprate
Combine files in to one lyrics.txt
Use fpt.ai speech-to-text to convert the parts audio to txt

Fig. 1. Flow chart

- Firstly, use urllib.request to download audio.
- Secondly, using InaSpeechSegmenter [6] to identify type of audio voice, and detect silences.
- Thirdly, reduce audio noise by using reducenoise [7] library.
- Fourthly, using API to send requests to fpt.ai and convert audio to texts.
- Finally, combine texts into one lyrics.txt.

C. Detail Process / algorithm

Step 1: Connect Colab with Google Drive

```
from google.colab import drive
drive.mount('/content/drive')
```

Step 2: Install necessary libraries

```
#Install necessary libraries
!pip install librosa
!pip install ffmpeg-python
!pip install noisereduce
!apt -qq install -y sox
!pip install pydub
!pip install sox
!sudo apt-get install sox libsox-fmt-mp3
!apt-get install libsox-fmt-
all sox libchromaprint-dev
!pip install inaSpeechSegmenter
```

Step 3: Import libraries

```
import ffmpeg
import urllib.request
import librosa
import matplotlib.pyplot as plt
import librosa.display
import IPython.display as ipd
import noisereduce as nr
import requests
from inaSpeechSegmenter import Segmenter
from pydub import AudioSegment
```

TABLE I. The Functions of Libraries

Libraries	Functions
noisereduce	Noise reduction in python using spectral gating (speech, bioacoustics, time-domain signals)
librosa	A python package for music and audio analysis, create music information retrieval systems.
inaSpeechSegmenter	Split the audio signal into homogeneous zones of speech, music, and noise. Then detects speaker gender.
pydub	A library to manipulate audio data with a simple high-level interface.
ffmpeg	A very fast video and audio converter that can also grab from a live audio/video source. It can also convert between arbitrary sample rates and resize video on the fly with a high-quality polyphase filter.

Step 4: Download audio and show original audio file

```
urllib.request.urlretrieve("https://media1.vocaroo
.com/mp3/1o453JLtmsSa", "/content/drive/My Drive/A
udio Folder/audio.mp3")

#Create audio wave:
x, sr = librosa.load('/content/drive/My Drive/Audi
o Folder/audio.mp3')
%matplotlib inline
plt.figure(figsize=(12, 4))
librosa.display.waveplot(x, sr=sr)
plt.title('Audio Wave')
plt.ylabel('Amplitude')
plt.xlabel('Time (Sec)')
ipd.Audio('/content/drive/My Drive/Audio Folder/au
dio.mp3')
```

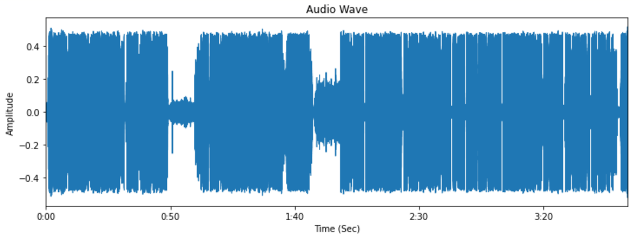


Fig. 2. Original audio

Step 5: Using InaSpeechSegment to identified type of audio file and detect silences from

```
#Using InaSpeechSegmenter:
media = '/content/drive/My Drive/Audio Folder/audi
o.mp3'
seg = Segmenter()
segmentation = seg(media)
print(segmentation)

#DETECT SILENCE:
sound_file = AudioSegment.from_mp3('/content/drive
/My Drive/Audio Folder/audio.mp3')
# milliseconds in the sound track
# ranges define the start and the end times of
each audio cut
ranges = [(0,50000),(60000,93000),(99780,107000),(
118000,133800),(134000,163000),(163500,173500),(17
4000,218700),(219000,230000)]
count = 0
for x, y in ranges:
    new_file = sound_file[x : y]
    #output file
    new_file.export("/content/drive/My Drive/Audio
 Folder/Audio/audio" + "" + str(count) +".mp3", fo
rmat="mp3")
    count+=1
```

Step 6: Reduce noise by Reduce noise libraries

```
#REDUCE NOISE:
'''REDUCE NOISE ONE BY ONE
Because once audio files had diffirent noisies par
t so we have to detected one by one to get the goo
d one'''
#Load audio file:
audio, sr = librosa.load('/content/drive/My Drive/
Audio Folder/Audio/audio0.mp3')
#Noise reduction audio0
#noise_part define the audio-frequency between 0
to 10Db
noisy_part = audio[0:10000]
reduced_noise = nr.reduce_noise(audio_clip=audio,
noise_clip=noisy_part, verbose=False)

#Check audio wave after reduce:
librosa.display.waveplot(reduced_noise, sr=sr)
plt.title('Reduced noise audio')
plt.ylabel('Amplitude')
plt.xlabel('Time (Sec)')
plt.show()
```

```
'''-----------------------------------
OUTPUT GENERATOR:
    receives a destination path, file name, audio matrix, and sample rate,
    generates a wav file based on input
-----------------------------------'''
def output_file(destination ,filename, y, sr, ext=""):
    destination = destination + filename[:-4] + ext + '.mp3'
    librosa.output.write_wav(destination, y, sr)
#Generating output file
output_file(path_name , 'audio0.mp3', reduced_noise, sr, '_reduced_noise')
ipd.Audio(reduced_noise, rate=sr)
```

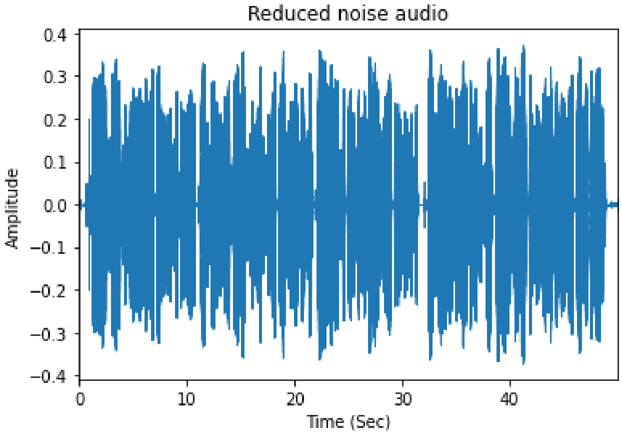


Fig. 3. Reduced Noise Audio

<u>Step 7:</u> Sent request to Fpt.ai to convert audio to texts.

```
#API FPT:
'''For somehow FPT.AI API (free API) will get limit rate to use then we have to using more codelines
'''--Once audio we will try once API
Anyway, if it's got limited. Don't worry. We prepared,
 below are some API keys we prepared, try it.
    '''-- nYvfiwnSbtuKy0OK37W4aTVYaVzM3q3Q: Key 1
        zMMA0wOD4S0dlHNtIoULGickGb0ojvt1: Key 2
            GZhmL9nML2pJ8dncoXQ8IMLXvPnDQsHN
            aDStwoJ17CqKwfF1QZ3gxA4aCo0exMnk
            iedpov8BdCqIFWxXNr4LeAIIaAjcsfEo
         EdyezNlk7VhZEuiz3N7qGHyKZ5KtRg8Y---'''
# if not work.. Connect FPT.AI:
# Sign in an account then make your own API.
#Convert File 0
import requests
url = 'https://api.fpt.ai/hmi/asr/general'
payload = open('/content/drive/My Drive/Audio Folder/Audio Reduce/audio0_reduced_noise.mp3', 'rb').read()
headers = {
    'api-key': 'OhFLAJ7gpoPXIXQESHw3FowlrBRbjGio'
}
response = requests.post(url=url, data=payload, headers=headers)
print(response.json())
#for some how you have to try again if it not work
obj = response.json()
obj2 = obj['hypotheses']
listToStr = ' '.join([str(elem) for elem in obj2])
text = listToStr.split(':')
print(text[2])
```

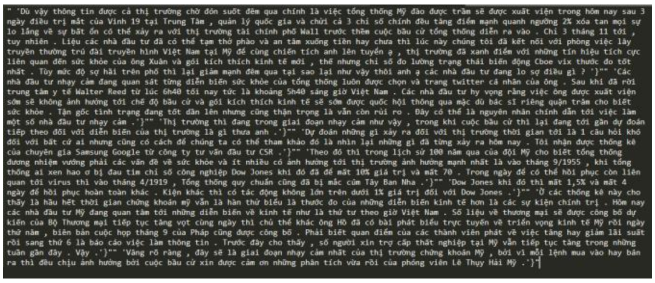
<u>Step 8:</u> Combine texts to one lyrics file, format .txt

```
#Combine all data texts above to one lyrics.txt
import io, json
with io.open('/content/drive/My Drive/Audio Folder/lyrics.txt', 'w', encoding='utf-8') as f:
  f.write(json.dumps(text[2], ensure_ascii=False))
 f.write(json.dumps(text1[2], ensure_ascii=False))
 f.write(json.dumps(text2[2], ensure_ascii=False))
 f.write(json.dumps(text3[2], ensure_ascii=False))
 f.write(json.dumps(text4[2], ensure_ascii=False))
 f.write(json.dumps(text5[2], ensure_ascii=False))
 f.write(json.dumps(text6[2], ensure_ascii=False))
 f.write(json.dumps(text7[2], ensure_ascii=False))
```

III. TESTING RESULT

It's can be applied to many systems such as voice recognition, voice control and lyrics synchronization.

Result of sample output lyrics.txt:



```
" 'Dù vậy thông tin được cả thị trường chờ đón suốt đêm qua chính là việc tổng thống Mỹ đào được trăm sẽ được xuất viện trong hôm nay sau 3
ngày điều trị mắt của Vinh 19 tại Trung Tâm , quản lý quốc gia và chửi cả 3 chỉ số chính đều tăng điểm mạnh quanh ngưỡng 2% xóa tan mọi sự
lo lắng về sự bất ổn có thể xảy ra với thị trường tài chính phố Wall trước thêm cuộc bầu cử tổng thống diễn ra vào . Chỉ 3 tháng 11 tới ,
tuy nhiên . Liệu các nhà đầu tư đã có thể tạm thở phào và an tâm xuống tiền hay chưa thì lúc này chúng tôi đã kết nối với phòng việc lấy
truyền thường trú đài truyền hình Việt Nam tại Mỹ để cùng chiến tích anh lên tuyến ạ , thị trường đã xanh điểm với những tín hiệu tích cực
liên quan đến sức khỏe của ông Xuân và gói kích thích kinh tế mới , thế nhưng chỉ số đo lường trạng thái biến động Cboe vix thước đo tốt
nhất . Tùy mức độ sợ hãi trên phố thì lại giảm mạnh đêm qua tại sao lại như vậy thôi anh ạ các nhà đầu tư đang lo sợ điều gì ? '}"" 'Các
nhà đầu tư nhạy cảm đang quan sát từng diễn biến sức khỏe của tổng thống luôn được chọn và trang twitter cá nhân của ông . Sau khi đã rời
trung tâm y tế Walter Reed từ lúc 6h40 tối nay tức là khoảng 5h40 sáng giờ Việt Nam . Các nhà đầu tư hy vọng rằng việc ông được xuất viện
sớm sẽ không ảnh hưởng tới chế độ bầu cử và gói kích thích kinh tế sẽ sớm được quốc hội thông qua mặc dù bác sĩ riêng quận trăm cho biết
sức khỏe . Tận gốc tình trạng đang tốt dần lên nhưng cũng thận trọng là vẫn còn rủi ro . Đây có thể là nguyên nhân chính dẫn tới việc làm
một số nhà đầu tư nhạy cảm .'}"" 'Thị trường thì đang trong giai đoạn nhạy cảm như vậy , trong khi cuộc bầu cử thì lại đang tới gần dự đoán
tiếp theo đối với diễn biến của thị trường là gì thưa anh .'}"" 'Dự đoán những gì xảy ra đối với thị trường thời gian tới là 1 câu hỏi khó
đối với bất cứ ai nhưng cũng có cách để chúng ta có thể tham khảo đó là nhìn lại những gì đã từng xảy ra hôm nay . Tôi nhận được thống kê
của chuyên gia Samsung Google từ công ty tư vấn đầu tư CSR .'}"" 'Theo đó thì trong lịch sử 100 năm qua của đội Mỹ cho biết tổng thống
đương nhiệm vướng phải các vấn đề về sức khỏe và ít nhiều có ảnh hưởng tới thị trường ảnh hưởng mạnh nhất là vào tháng 9/1955 , khi tổng
thống ai xen hao ơ bị đau tim chỉ số công nghiệp Dow Jones khi đó đã để mất 10% giá trị và mất 70 . Trong ngày để có thể hồi phục còn liên
quan tới virus thì vào tháng 4/1919 , Tổng thống quy chuẩn cũng đã bị mắc cúm Tây Ban Nha .'}"" 'Dow Jones khi đó thì mất 1,5% và mất 4
ngày để hồi phục hoàn toàn khác . Kiện khác thì có tác động không lớn trên dưới 1% giá trị đối với Dow Jones .'}"" 'Ở các thống kê này cho
thấy là hầu hết thời gian chứng khoán mỹ vẫn là hàn thử biểu là thước đo của những diễn biến kinh tế hơn là các sự kiện chính trị . Hôm nay
các nhà đầu tư Mỹ đang quan tâm tới những diễn biến về kinh tế như là thứ tư theo giờ Việt Nam . Số liệu về thương mại sẽ được công bố dự
kiến của Bộ Thương mại tiếp tục tăng vọt cùng ngày thì chủ thể khác ông Hồ đã có bài phát biểu trực tuyến về triển vọng kinh tế Mỹ rồi ngày
thứ năm , biên bản cuộc họp tháng 9 của Pháp cũng được công bố . Phải biết quan điểm của các thành viên phát về việc tăng hay giảm lãi suất
rồi sang thứ 6 là báo cáo việc làm thông tin . Trước đây cho thấy , số người xin trợ cấp thất nghiệp tại Mỹ vẫn tiếp tục tăng trong những
tuần gần đây . Vậy .'}"" 'Vâng rõ ràng , đây sẽ là giai đoạn nhạy cảm nhất của thị trường chứng khoán Mỹ , bởi vì mỗi lệnh mua vào hay bán
ra thì đều chịu ảnh hưởng bởi cuộc bầu cử xin được cảm ơn những phân tích vừa rồi của phóng viên Lê Thụy Hải Mỹ .'}"
```

Fig. 4. Sample output lyrics

“The news conference awaited by the whole market last night was the fact that the US President who dug the bass would be discharged today after 3 days of MAX treatment of Boeing 79 at the National Management Center and cursed all 3 The main numbers all gained strongly around the 2% threshold. Eliminate all worries about the instability that might happen to Wall Street financial markets before the presidential election takes place. On November 3, however, whether investors can temporarily breathe a sigh and feel secure in money, at this time, we are connected with the permanent transmission room of Vietnam television station in the US In order to join the achievements, the market was green with positive signals related to the health of Mr. Xuan and the new economic stimulus package, but the index measures the state of volatility. My video is the best measure, depending on the level of fear, Tran Phu has dropped sharply last night, why is that? What do investors fear? "

"Sensitive investors are observing every evolution of the always-chosen president's health and his personal Twitter. But when he left the Walter Reed medical center. At 6:40 am this evening at 5:40 am Vietnam time, investors hope that his early release from hospital will not affect the electoral regime and the economic stimulus package will soon be approved by the National Assembly. though hundred district private doctor said the health. His condition is getting better, but he is also cautious that there is still a risk. This may be the main reason for making some investors sensitive ."

"The market is in such a sensitive period, while the election is nearing the next forecast for the market's movement, sir.

“Predicting what will happen to the school next time is a difficult question for anyone but there is a way for us to refer to that is to look back on what happened today. I received statistics from Samsung Google expert from consulting firm CSR”

“Accordingly, in the past 100 years of the American team, the incumbent president has encountered health problems and more or less affected the strongly influenced market, especially in September 1955 President who called me a heart attack Dow Jones Industrial Index, then lost 10% of its value and took 70 5 days to be able to recover. As for the virus, in April 1919, the Standard President also caught the Spanish flu .”

“Dow Jones, then, lost 1.5% and took 4 days to recover completely different. Another event has a small impact on less than 1% of value for the Dow Jones .”

“These statistics show that most of the time US stocks are still Korean, as a measure of economic developments rather than political events. Today US investors are interested in economic developments such as Wednesday, Vietnam time, data on trade will be announced, expected by the Ministry of Commerce to continue to skyrocket on the same day. another, Mr. Ho gave a speech online about the US economic prospects and then on Thursday the minutes of the September meeting of France were announced. Must know the views of the members about the increase or decrease of interest rates and then on Friday is the report on employment information. Previously, the number of people applying for unemployment benefits in the US continued to increase in recent weeks.”

Fig. 5. Sample output lyrics after translated by NH. Trung

❖ **Quiet environment**: The program displays excellent sound with a bit noise and silience.

- ➢ Ability processing in this environment is up to 80% with a simple audio file (not music file).
- ➢ An audio file with more than 900 words after using Fpt.ai to detected, it has 827 words. And also we tried another one, audio with 560 words and once we used Fpt.ai to detected, it has 536 words (95,71%).
- ➢ The accuracy of the first file is 795/827 words subequal 96% but it’s worked with simple audio, not music files.
- ➢ The accuracy of the second file is 509/536 words subequal 94,96% also with simple audio.
- ➢ The first audio file we trimmed to 7 files, the first trimmed file with 16 seconds and 63 words, after using Fpt.ai detected, we got 65 words (60/63 95,23%), and also Viettelgroup. ai detected too, we got 67 words (56/63 88,89%).

Dự đoán những gì xảy ra đối với thị trường thời gian tới là 1 câu hỏi khó đối với bất cứ ai.
Nhưng cũng có cách để chúng ta có thể tham khảo, đó là nhìn lại những gì từng xảy ra.
Hôm nay, tôi nhận được thống kê của chuyên gia Sam Stove từ công ti tư vấn đầu tư CSRA.

Fig. 6. Original audio cut lyrics

Dự đoán những gì xảy ra đối với thị trường thời gian tới là 1 câu hỏi khó đối với bất cứ ai nhưng cũng có cách để chúng ta có thể tham khảo , đó là nhìn lại những gì đã từng xảy ra hôm nay . Theo tôi nhận được thống kê của chuyên gia Samsung Google từ Công ty tư vấn đầu tư CSRA.

Fig. 7. Fpt.ai audio detect

dự đoán những gì xảy ra đối với thị trường thời gian tới là một câu hỏi khó đối với bất cứ ai
nhưng cũng có cách để chúng ta có thể tham khảo đó là nhìn lại những gì đã từng xảy ra
hôm nay tôi nhận được thống kê của chuyên gia samsung google từ công ty tư vấn đầu tư xi ép
sẽ ra

Fig. 8. ViettelGroup audio detect

- ➢ A music file has 273 words when detected stops and pauses. It trimmed to 7 files (20-40s) each. A trimmed file with 15s and 25 words. Fpt.ai detects 11 words (Accuracy 0%) and Viettel.ai detects 13 words (Accuracy 16%).

Em, ngày em đánh rơi nụ cười vào anh
Có nghĩ sau này em sẽ chờ
Và vô tư cho đi hết những ngây thơ

Fig. 9. Original music cut lyrics

nghệ an đang sôi nổi
cứ nghĩ sơ mc
cho đi hết như

Fig. 10. Viettelgroup.ai music detect

Đen . 2 . Ư nghi sơn . Ư cho I Hate You . Ư .

Fig. 11. Fpt.ai music detect

❖ **Noise stable environment**: The program displays fair sound without silience, sometimes the output contains tiny noises but it is trivial. Ability processing in noise stable environment is about 90%.

❖ **Noise unstable environment**: The program displays sound without silience but it still has noise because in this environment, noise level changes through time so the program can not get sample noise exactly in audio file to reduce noise . Ability processing in this environment is low, about 70%.

❖ **Summary**: This program performs well in quiet and noise stable environment. In noise unstable environment, program still works but it’s result contains unnessesary noise.

IV. Conclusion

In this article, we want to mention the audio processing function to create a complete file by removing unnecessary silence and noise in the audio file. Currently, this audio processing program is very interested and focused because it is applied to many areas of life such as: Speech recognition as know as (Speech recognition and then converting them into corresponding text), speech synthesis as know as (From an automatically synthesized text into speech), or text summary which is (From a long texts summarizes into just a shorter text as desired but still contains the most essential content). In short, the above application is a very useful tool and my team hopes this program can be further improved and developed in the future.

V. References

[1] E. Bisong and E. Bisong, “Google Colaboratory,” in *Building Machine Learning and Deep Learning Models on Google Cloud Platform*, 2019, pp. 59–64.

[2] and O. N. McFee, Brian, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, “librosa: Audio and music signal analysis in python,” *https://librosa.org/doc/latest/install.html*, 2013. https://groups.google.com/g/librosa.

[3] M. Strange, “AudioSegment,” 2018. https://github.com/MaxStrange/AudioSegment.

[4] T. D. Chung, M. Drieberg, M. F. Bin Hassan, and A. Khalyasmaa, “End-to-end Conversion Speed Analysis of an FPT.AI-based Text-to-Speech Application,” in *LifeTech 2020 - 2020 IEEE 2nd Global Conference on Life Sciences and Technologies*, 2020, pp. 136–139, doi: 10.1109/LifeTech48969.2020.1570620448.

[5] V. C. Center, “Speech Recognition,” 2018. https://viettelgroup.ai/.

[6] David.Doukhan., “Speech Segmentation Program,” 2018. https://github.com/ina-foss/inaSpeechSegmenter.

[7] T. Sainburg, “Noise reduction in python using spectral gating,” 2018. https://github.com/timsainb/noisereduce.