Multilingual Text-to-Speech with Grad-TTS

Presented by:
Albert Millert, Shalini Priya, Soklay Heng, Wenjun Sun

Université de Lorraine
IDMC - Institut des sciences du digital

MSc in Natural Language Processing
UE 805 – Software Project

February 11, 2022

Table of Content

1. Project Presentation
2. Development Process
3. System Architecture
4. Data Processing
5. Models
6. Web Interface
7. Evaluation

1 Project Presentation

2 Development Process

3 System Architecture

4 Data Processing

5 Models

6 Web Interface

7 Evaluation

Introduction

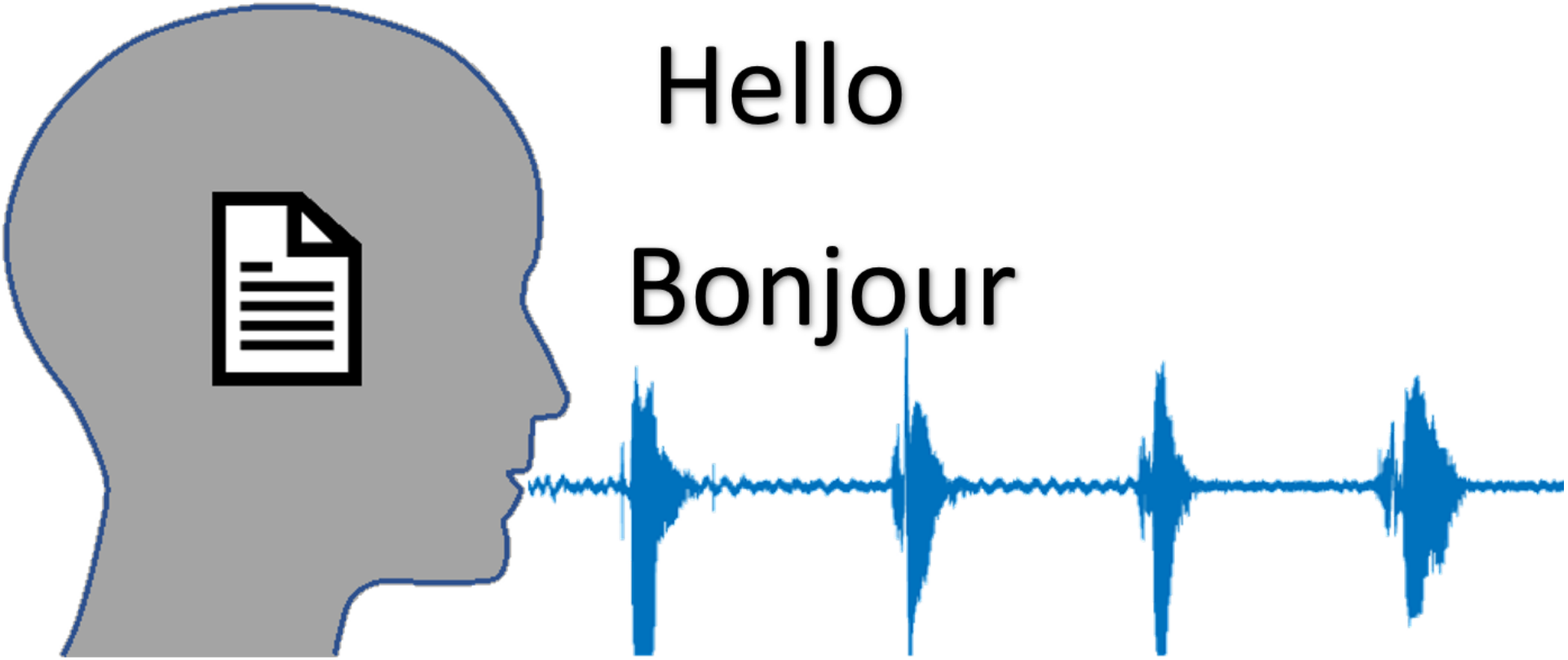


Figure 1: Speech Production

Project Objective

- Text-to-speech service on website for English and French
- Shareable & extendable tool

1 Project Presentation

2 Development Process

3 System Architecture

4 Data Processing

5 Models

6 Web Interface

7 Evaluation

Development Timeline

1. Frontend design
2. English, French corpora
3. French data preprocessing - phonetization
4. Language classification
5. Flask server
6. Embed Grad-TTS as a submodule - working on English data
7. Containerization through Docker + networking
8. Docker-compose + reverse proxy
9. French Grad-TTS model
10. Audio file generation + playing in the browser
11. Intelligibility & naturalness evaluation + analyses
12. GitHub release + Docker Hub image publish

Releases

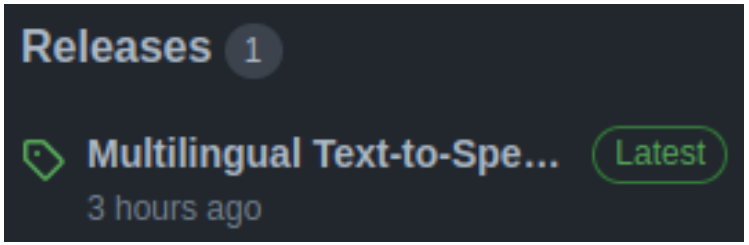


Figure 2: GitHub release

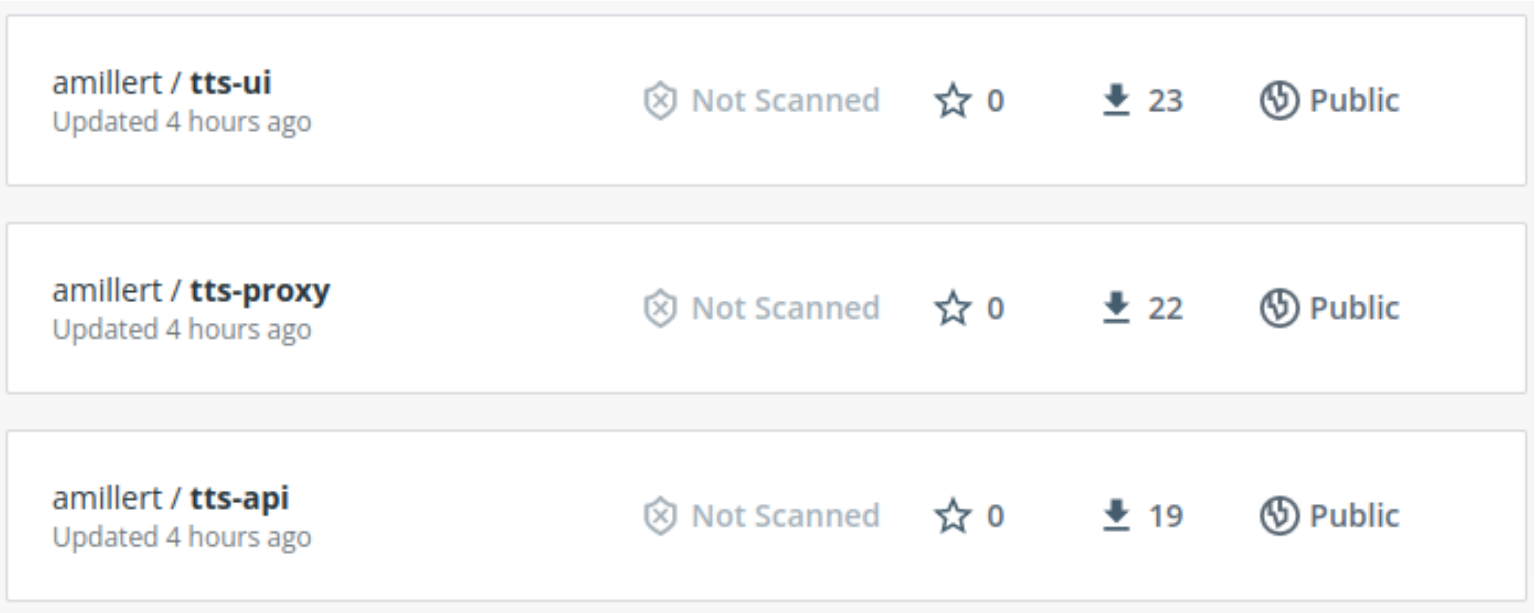


Figure 3: Docker Hub images

1 Project Presentation

2 Development Process

3 System Architecture

4 Data Processing

5 Models

6 Web Interface

7 Evaluation

System Architecture

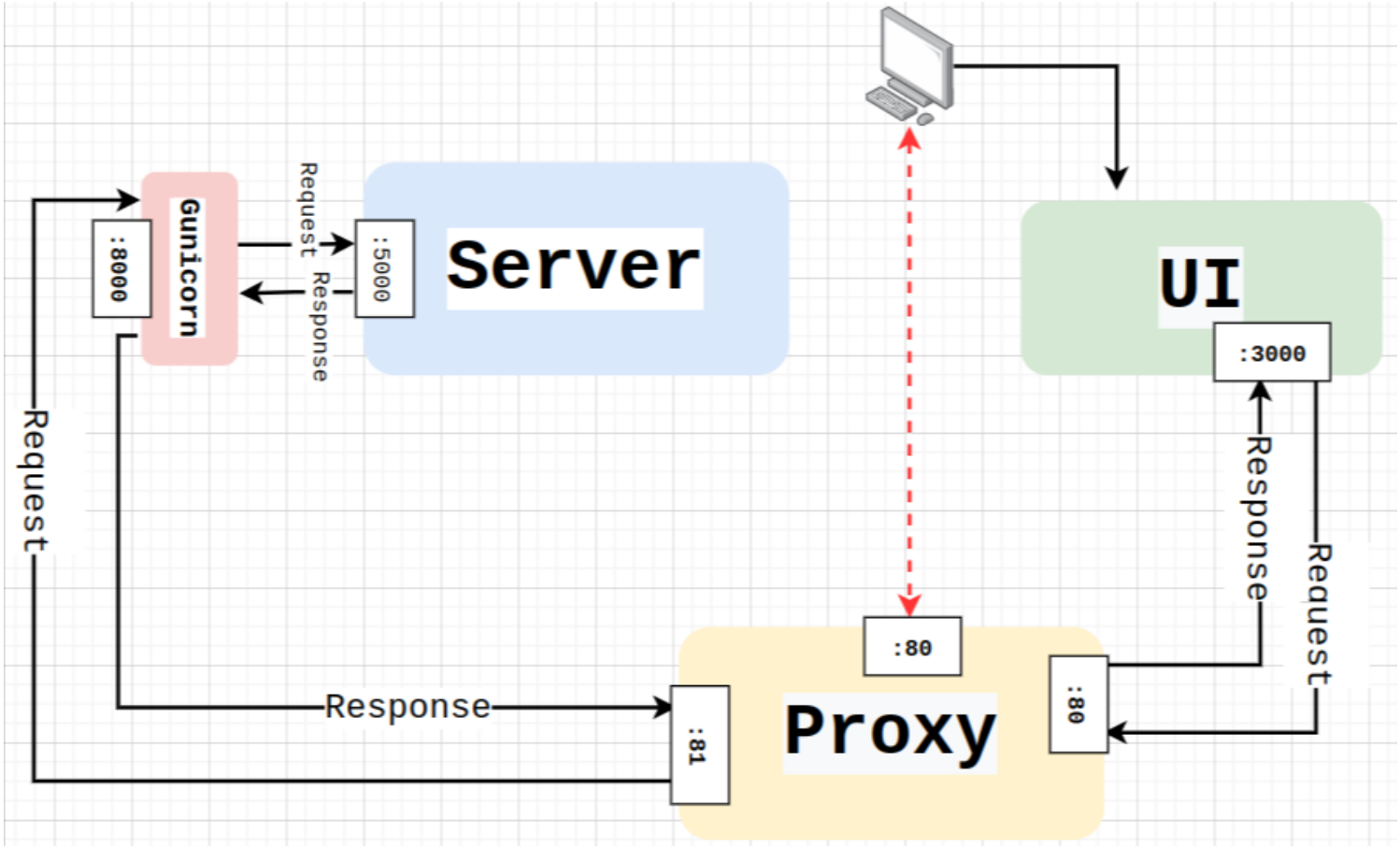


Figure 4: System contenairized architecture with request travel visualization

1 Project Presentation

2 Development Process

3 System Architecture

4 Data Processing

5 Models

6 Web Interface

7 Evaluation

Dataset

- English: LJS - female voice (already preprocessed)
  https://keithito.com/LJ-Speech-Dataset/
  13100 utterances (total length 24h)
- French: SIWIS - female voice (required preprocessing)
  https://datashare.ed.ac.uk/handle/10283/2353
  9750 utterances (10+h)

Data Processing

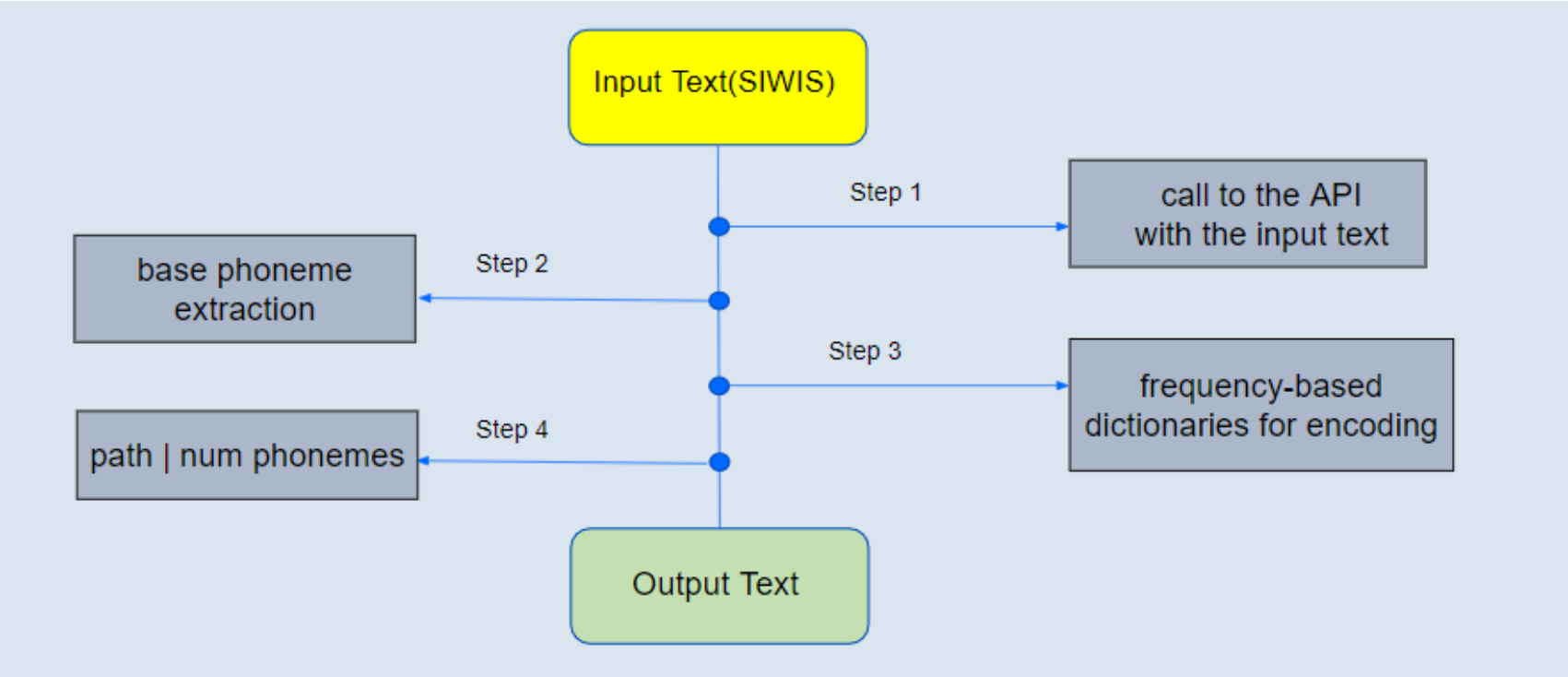


Figure 5: Phonemes Generation

Obtained data

```
./s04_0275 | 11 10 22 7 11 8 3 12 3 2 0 2 10 1 4 0 1 1 25 11 9 25 15 5 13 10 7 4 14 10 20
./s03_0245 | 5 4 7 26 13 5 3 2 11 12 16 20 9 28 17 10 17 14 6 23 26 6 19 3 5 15 5 2 0 1 1
./s03_0473 | 4 19 3 2 23 16 11 8 17 12 20 6 30 13 7 3 14 0 1 3 19 17 0 1 1 12 28 3 5 27 0
./s05_0455 | 3 5 12 4 0 14 6 24 5 3 5 1 12 4 0 14 7 6 0 9 29 6 1
./s02_0363 | 27 8 15 18 13 28 15 5 2 0 3 7 2 20 0 2 0 9 1
./s05_0480 | 5 4 16 12 3 2 5 2 10 25 21 9 30 33 22 6 28 4 0 18 5 4 22 25 11 6 7 2 0 14 34
./s04_0529 | 4 0 22 29 9 11 8 6 18 2 1 15 23 2 6 8 22 29 9 11 8 6 18 2 1 3 0 1 1 19 4 8 6
./s03_0567 | 27 8 15 18 4 8 16 12 3 7 9 1 3 31 0 9 19 6 7 13 23 17 0 20 4 1
./s01_0296 | 17 18 3 0 15 12 2 18 16 8 3 19 13 7 26 13 5 27 24 3 0 20 5 2 4 8 21 12 11 32
./s01_0287 | 2 6 31 0 32 24 2 9 19 21 23 20 5 4 5 13 4 8 26 18 15 7 6 17 8 21 24 1
```

Figure 6: File-phonemes format

1 Project Presentation

2 Development Process

3 System Architecture

4 Data Processing

5 Models

6 Web Interface

7 Evaluation

Language Classifier

- English-French Corpus

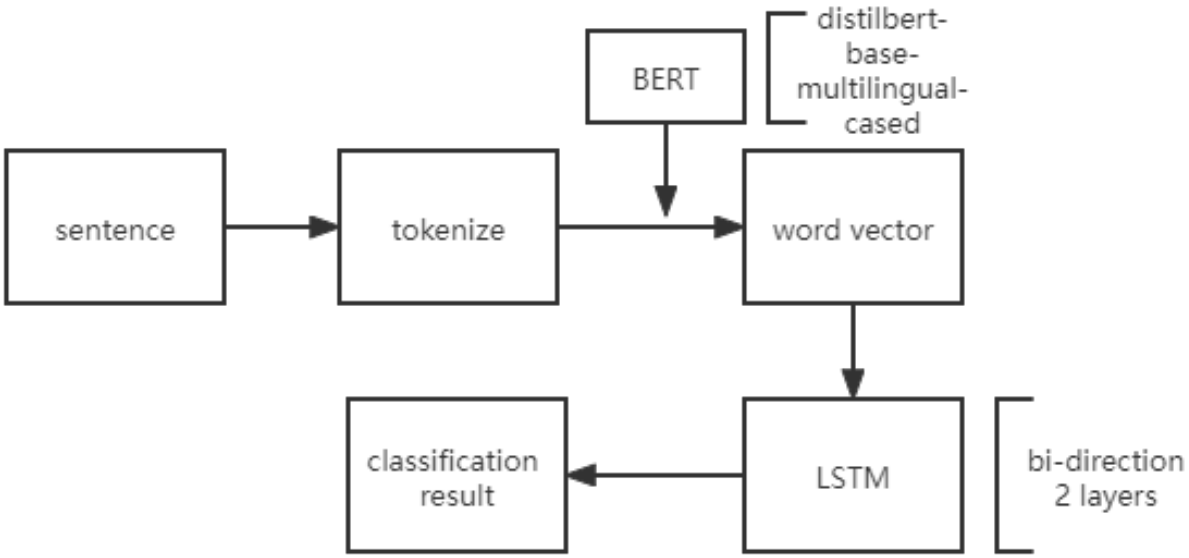


Figure 7: Language classifier

Grad TTS

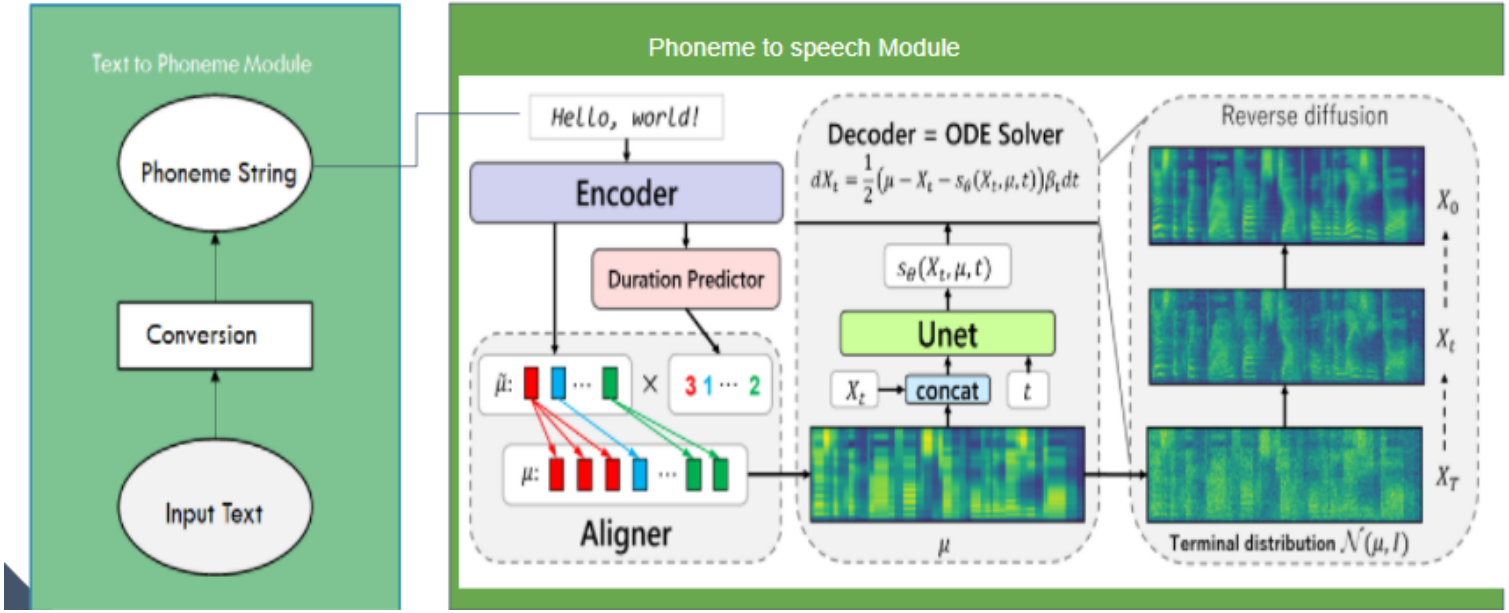


Figure 8: Grad-TTS inference scheme

1 Project Presentation

2 Development Process

3 System Architecture

4 Data Processing

5 Models

6 Web Interface

7 Evaluation

English Interface

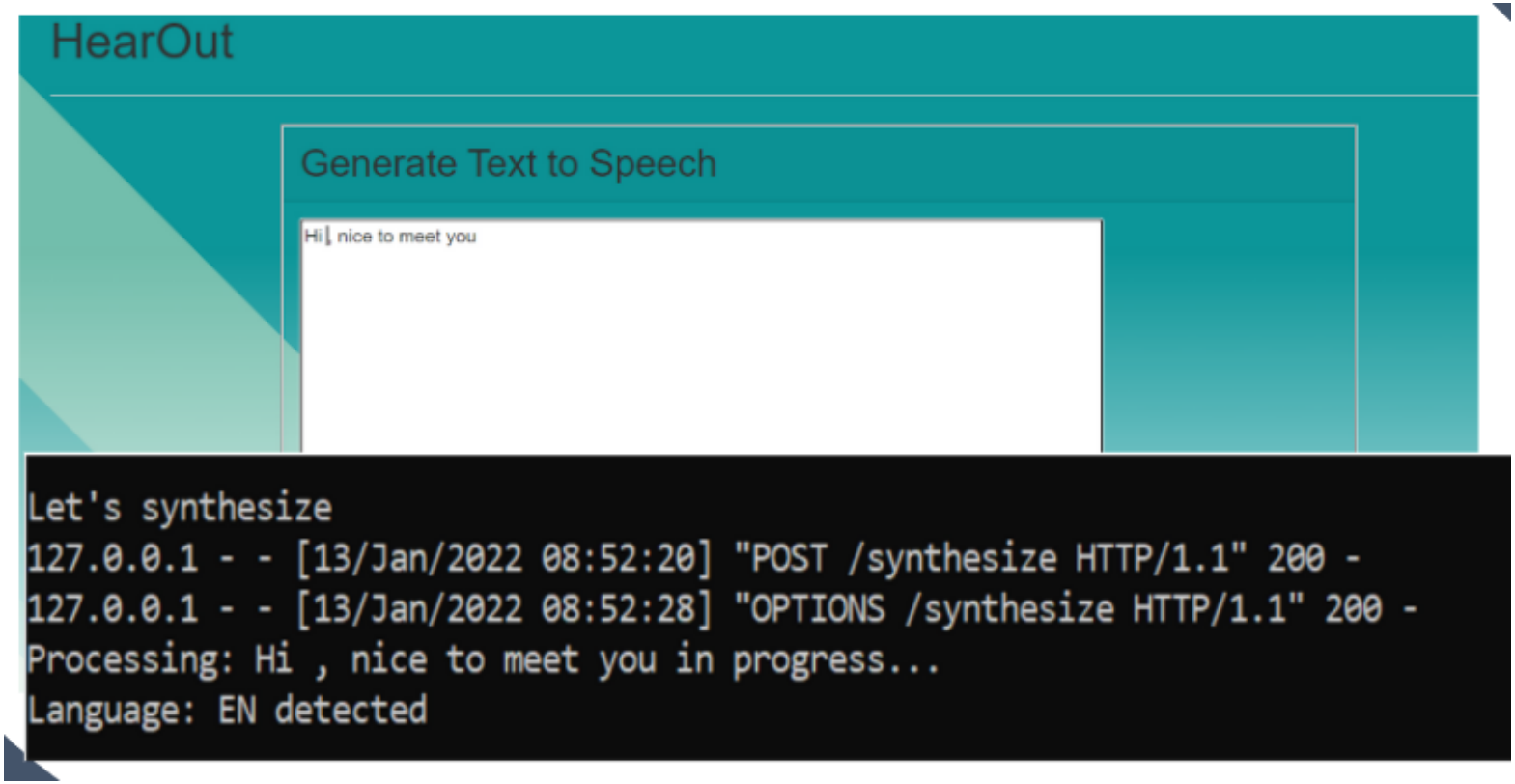


Figure 9: Detected language: English

French Interface

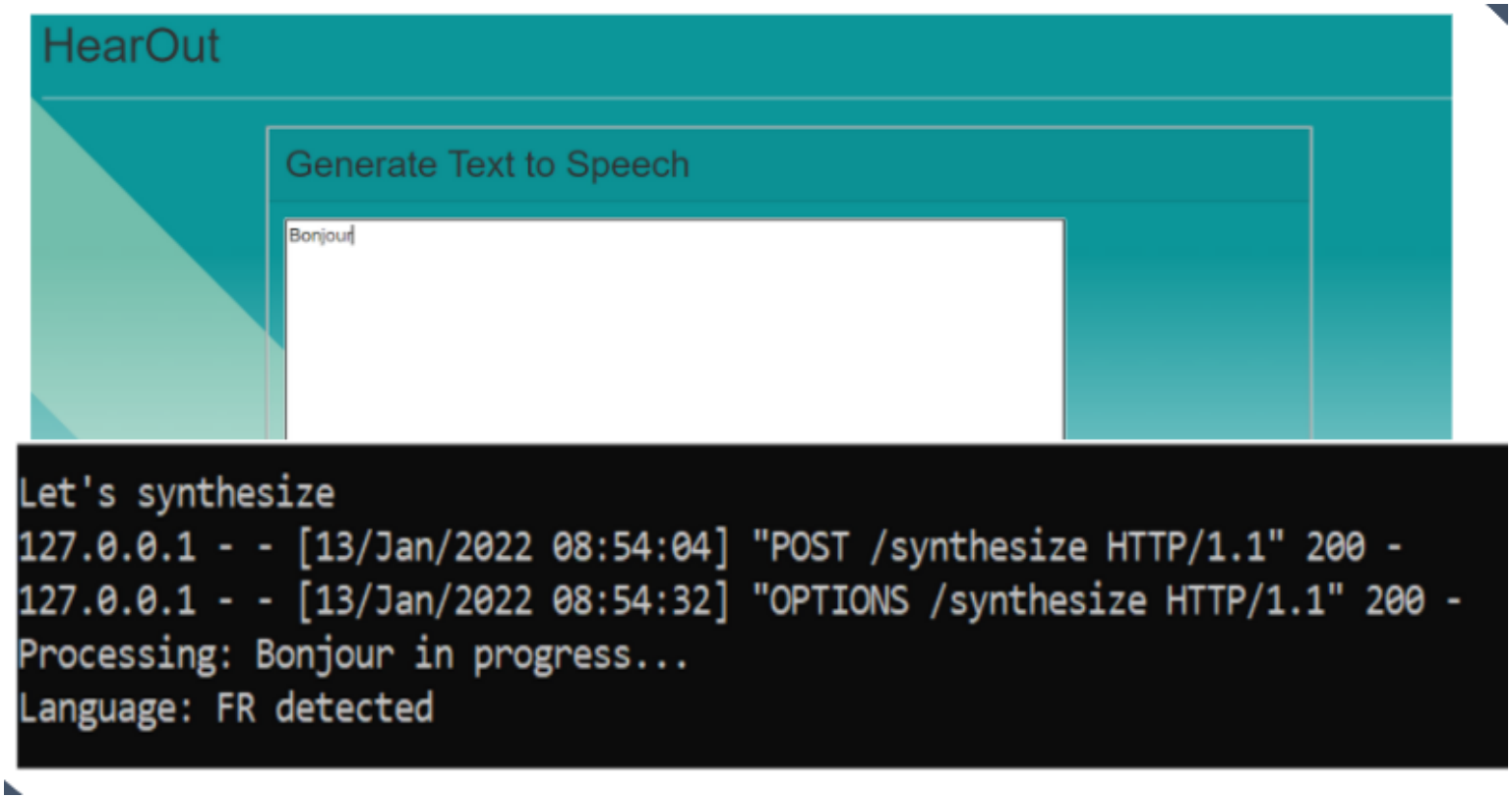


Figure 10: Detected language: French

1 Project Presentation

2 Development Process

3 System Architecture

4 Data Processing

5 Models

6 Web Interface

7 Evaluation

Language Classifier

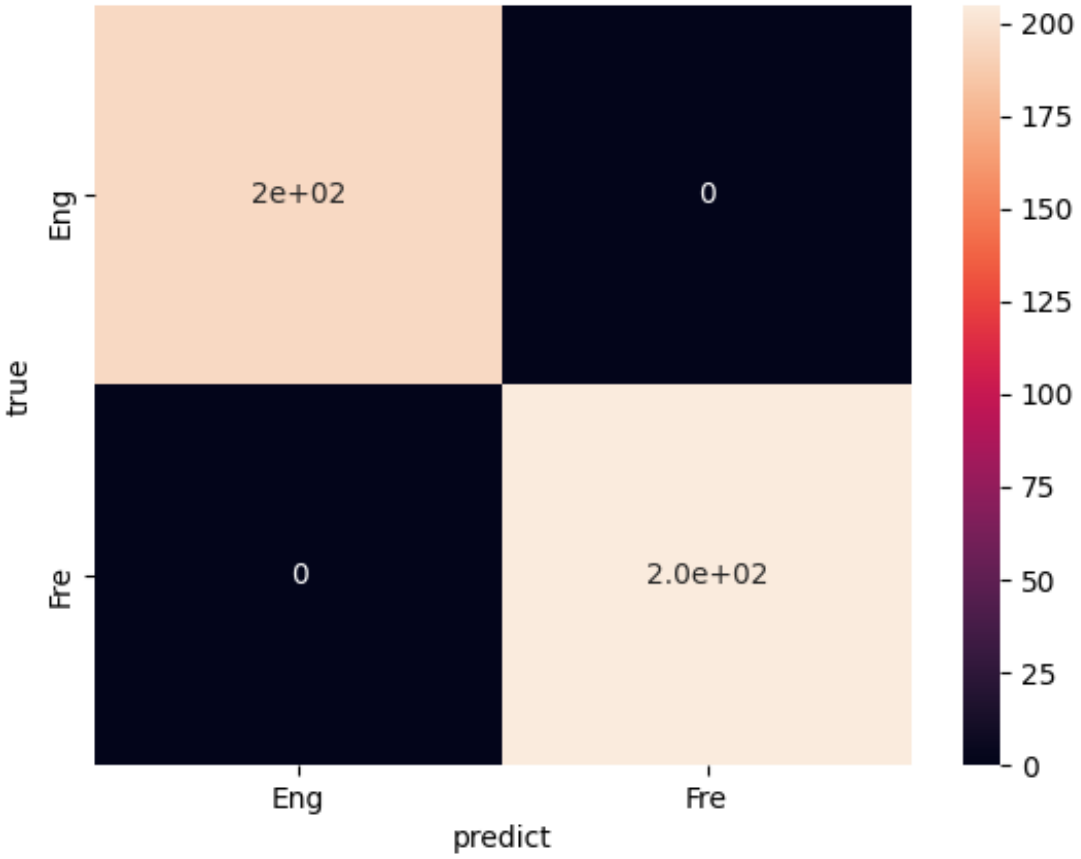


Figure 11: Result of classifier

Grad TTS

Quality of speech:

- Naturalness
- Intelligibility

Approaches:

- Objective
- Subjective

Objectivity Evaluation (Intelligibility)

- Google ASR Python library
- WER with jiwer Python library

Result

```
hypo = []
for f in file:
    transcription = transcribe_audio(f)
    hypo.append(transcription)
```

```
hypo
```

```
['community to help homeless',
 'Council chief executive fails to secure',
 'cancel to contest',
 'cancel notes to protect test heritage god',
 'cancel welcomes ambulance',
 'cancel welcomes insurance break',
 'cream tales of leadership kruphix 2',
 'start a fire expected to rise',
 'death toll continues to climb in South Korean sub',
 'Coldplay ms-80ver Iraqi concert']
```

```
ground_truth = ["community urged to help homeless youth","council chief executive fails to secure position",'councillor to contest wollor
"council welcomes insurance breakthrough","crean tells alp leadership critics to shut up","dargo fire threat expected to rise","death tol
,"dems hold plebiscite over iraqi conflict"]
```

```
#The most simple use-case is computing the edit distance between two strings:
from jiwer import wer
error = wer(ground_truth, hypo)
```

Figure 12: Google ASR and WER for Objectivity Evaluation

Language	Sentences	WER compared with Google ASR
English	10	0.56
French	10	0.30

Figure 13: Result of Objectivity Evaluation for Intelligibility

Subjectivity Evaluation (Intelligibility)

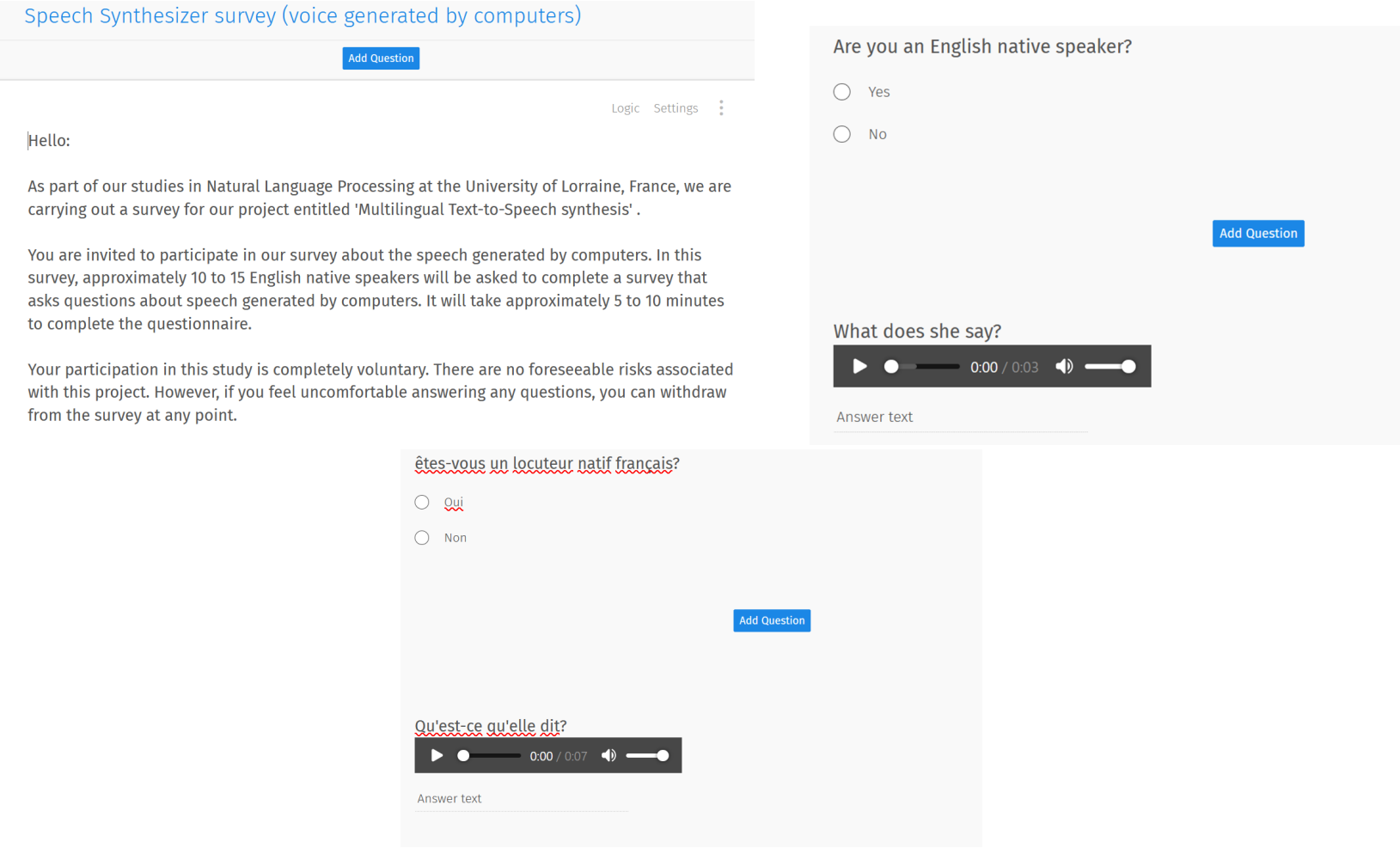
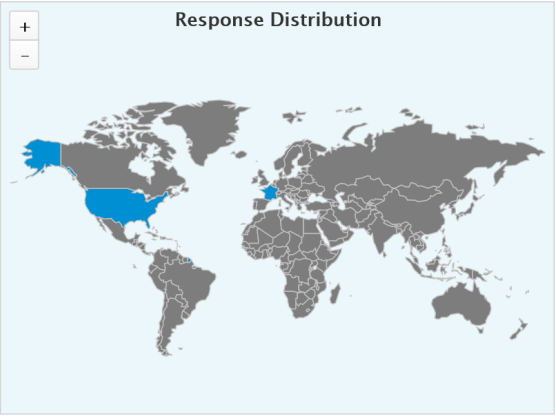
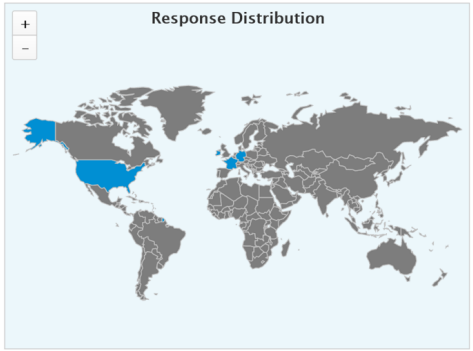


Figure 14: English and French Online Survey

Result



Countries	Responses
FR	98.15%
US	1.85%
Total	100.00%



Countries	Responses
FR	66.67%
US	22.22%
IE	7.41%
DE	3.70%
Total	100.00%

02/07/2022	65741973	Ayrton raconta alors en peu de mots ce qui s'était passé où du moins ce qu'il savait
02/07/2022	65740791	Herton raconta alors en peu de mots ce qu'il s'était passé, ou du moins ce qu'il savait.
02/07/2022	65738137	Herron raconta alors en peu de mots ce qu'il s'était passé ou du moins ce qu'il savait

02/03/2022	65519406	National gathering gets all clear after
02/02/2022	65512401	National gallery gets all clear after
01/30/2022	65330228	National gallery gets all clear after
01/30/2022	65330103	National gallery gets all clear after
01/29/2022	65280832	National gallery gets all clear after

Figure 15: English and French Online Survey

- Analysis with WER

Language	Participants (Native Speakers)	WER
English	10	0.55
French	15	0.42

Figure 16: Result of Subjectivity Evaluation for Intelligibility

Objectivity Evaluation (Naturalness) with AutoMOS

- The model has been built, but still needs to be adjusted and verified

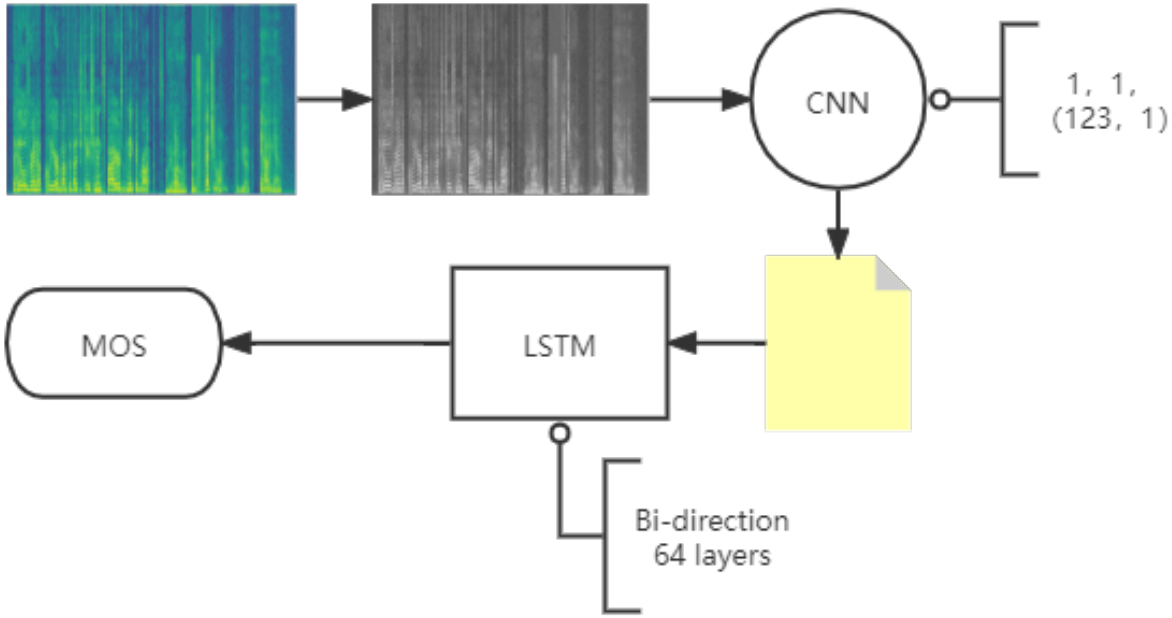


Figure 17: Pipeline of AutoMOS

Subjectivity Evaluation (Naturalness)

Online Survey for English TTS Model

- Mean Opinion Score(MOS)
- 35 sentences
- 3 evaluator for each sentence
- average as final MOS

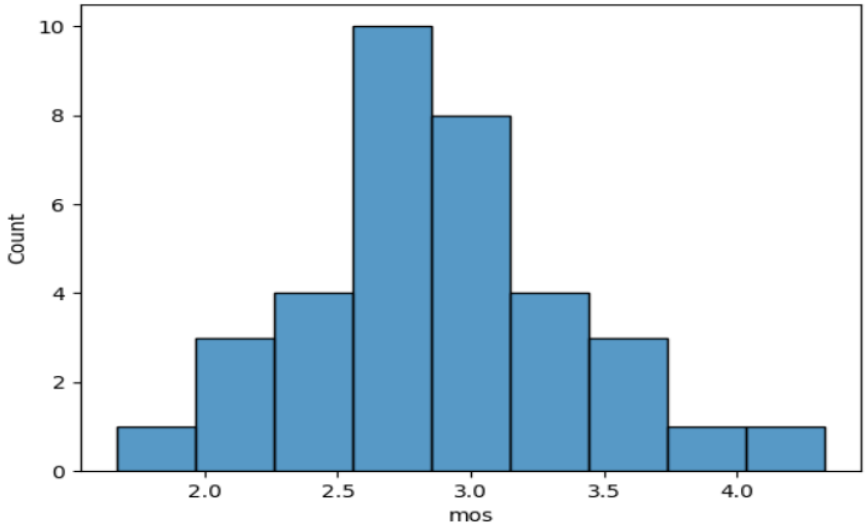


Figure 18: MOS

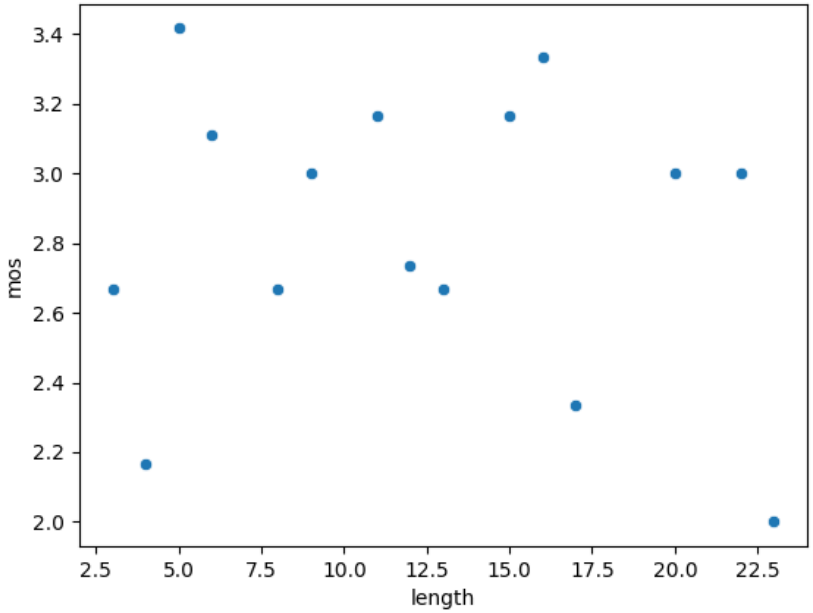
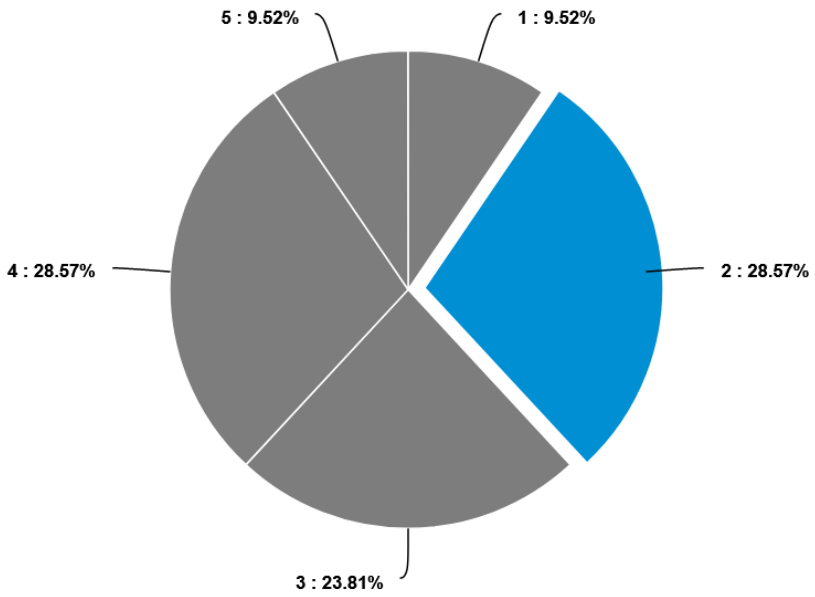


Figure 19: Length and MOS

Subjectivity Evaluation (Naturalness)

Online Survey for French TTS Model

À quel point pensez-vous que le discours est naturel ?



Question	Count	Score	1	2	3	4	5
À quel point pensez-vous que le discours est naturel ?	21	3					

Figure 20: Result of Online Survey for French TTS's Naturalness

DEMO & Conclusion

tinyurl.com/multitts-app

Thanks for your attention!!!

Q&A