Alena YAKAVETS
Cécile MACAIRE
Chanoudom PRACH
Ludivine ROBERT

Software Project

MULTILINGUAL TEXT-TO-SPEECH SYSTEM

NLP 2nd year 2020-2021

October 6, 2020

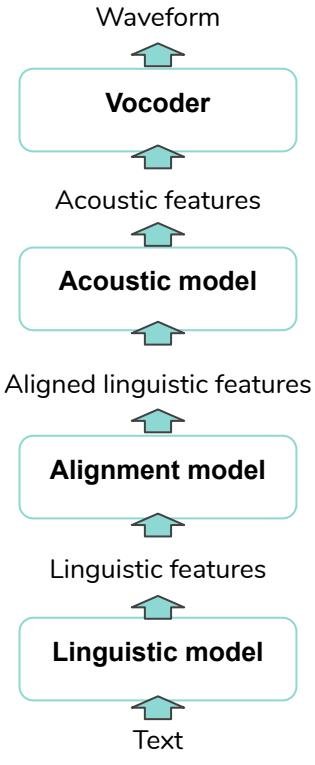
Presentation Overview

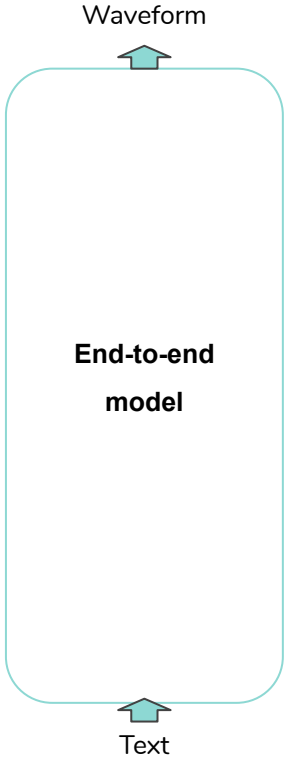
- TTS and multilingual TTS
- Tacotron
- Our approach
- Corpus
- Model Evaluation
- Timeline

Text-to-Speech Systems

Typical pipeline architecture for statistical parametric speech synthesis ⇒

Task-specific models





End-to-end system:

- directly transforms text to waveform
- doesn't require immediate feature extraction
- internal blocks are jointly optimized
- errors from different components don't accumulate

Multilingual TTS

Multilingual: Possible Interpretations

- Detect language for each document or part of document
- Produce speech using a language model for each detected part

- Define language model to be used for each document
- Apply selected language model for the whole document

Tacotron: an end-to-end speech synthesis system by Google

2017 - Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc Le, Yannis Agiomyrgiannakis, Rob Clark, Rif A. Saurous.

TTS systems are complex:

- same text can correspond to different pronunciations or speaking styles,
- output sequences are usually much longer than those of the input (prediction errors can accumulate quickly).

What is new with Tacotron?

Tacotron: an end-to-end speech synthesis system by Google

2017 - Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc Le, Yannis Agiomyrgiannakis, Rob Clark, Rif A. Saurous.

Idea: end-to-end generative TTS model based on the sequence-to-sequence with attention paradigm.

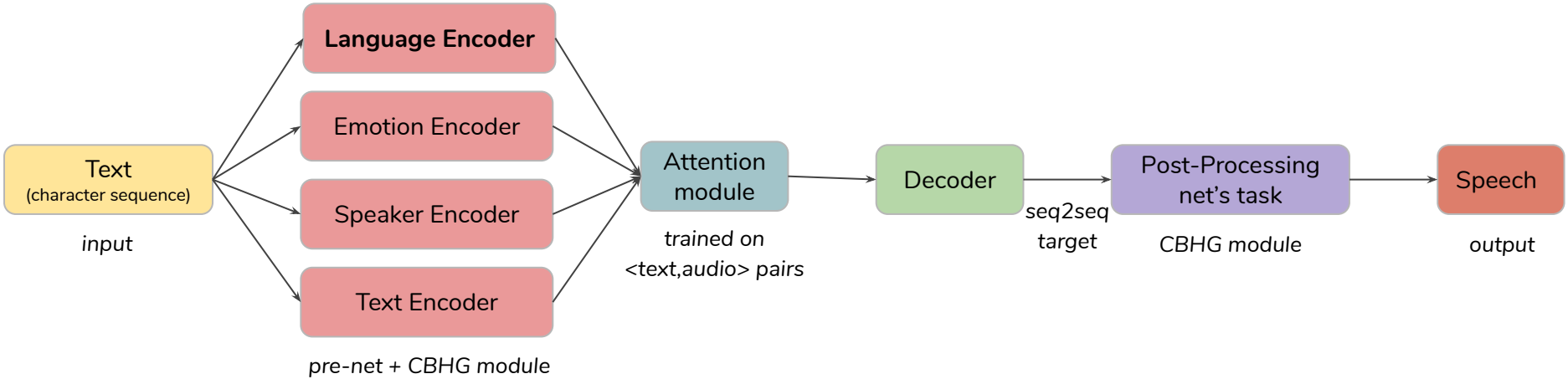


Figure 1: Tacotron Architecture

Tacotron: an end-to-end speech synthesis system by Google

2017 - Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc Le, Yannis Agiomyrgiannakis, Rob Clark, Rif A. Saurous.

Why is it better than other TTS systems?

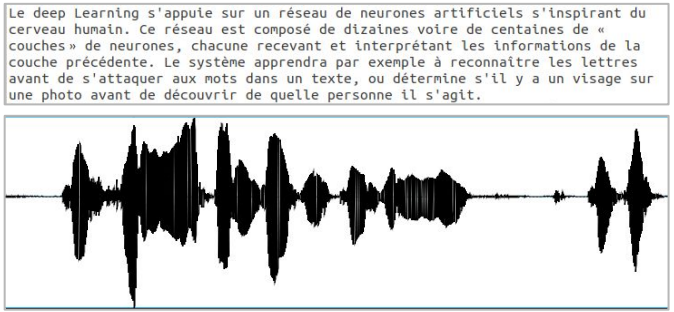
- Not needed hand-engineered linguistic features or complex components.
- Can be **trained from scratch with random initialization**.
- Use a sequence-to-sequence model:
  - capture pronunciation of words,
  - variation of human speech including volume, speed and intonation, sentiment, etc.
- Easier adaptation to new data.
- Robustness of a single model.

Tacotron: an end-to-end speech synthesis system by Google

2017 - *Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc Le, Yannis Agiomyrgiannakis, Rob Clark, Rif A. Saurous.*

Multilingual TTS system?

Multilingual support



Huge corpus of high quality data

Corpus

English: EmoV-DB

Emotional Voices Database

- Emotions: amusement, anger, sleepiness, disgust and neutral
- Speakers: native; males and females
- Reading sentences from books

French: SIWIS

French Speech Synthesis Database

- Emotion: neutral
- Speaker: native; female
- 9750 utterances from various sources: parliament debates and novels
- >10h of speech data

Preprocessing

Phonemes extraction:

- Converts text into context labels with elite web service
- Extract central phonemes with our script

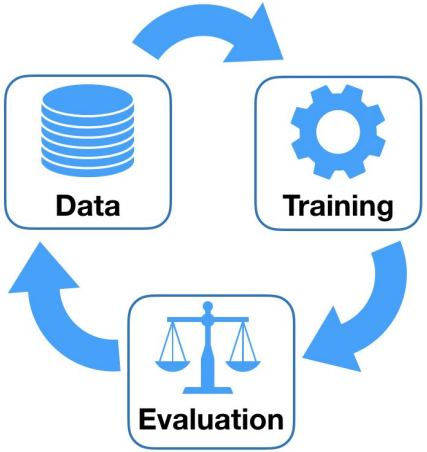
```
x^x-e+k=w@1_1/A:0_0_/B:1-1-1@1-1&1-4#0-1$0-1!0-1;0-3|e/C:1+0+3/D:x_0/
E:CONJCOOR+1@1+4&0+3#0+1/F:ADV-1/G:0_0/H:4=4@0=13|NONE/I:3_3/J:35+31-14
x^e-k+w=a@1_3/A:1_1_1/B:1-0-3@1-1&2-3#1-0$1-1!1-3;1-2|a/C:0+0+1/D:CONJCOOR_1/
E:ADV+1@2+3&0+2#0+1/F:SYMBOL-1/G:0_0/H:4=4@0=13|NONE/I:3_3/J:35+31-14
e^k-w+a=_@2_2/A:1_1_1/B:1-0-3@1-1&2-3#1-0$1-1!1-3;1-2|a/C:0+0+1/D:CONJCOOR_1/
E:ADV+1@2+3&0+2#0+1/F:SYMBOL-1/G:0_0/H:4=4@0=13|NONE/I:3_3/J:35+31-14
k^w-a+_=_@3_1/A:1_1_1/B:1-0-3@1-1&2-3#1-0$1-1!1-3;1-2|a/C:0+0+1/D:CONJCOOR_1/
E:ADV+1@2+3&0+2#0+1/F:SYMBOL-1/G:0_0/H:4=4@0=13|NONE/I:3_3/J:35+31-14
w^a-_+_=i@1_1/A:1_0_3/B:0-0-1@1-1&3-2#2-0$1-1!1-2;2-1|_/C:0+1+1/D:ADV_1/
E:SYMBOL+1@3+2&1+1#1+1/F:SYMBOL-1/G:0_0/H:4=4@0=13|NONE/I:3_3/J:35+31-14
a^_-_+i=I@1_1/A:0_0_1/B:0-1-1@1-1&4-1#2-0$1-0!2-1;3-1|_/C:1+1+2/D:SYMBOL_1/
E:SYMBOL+1@4+1&2+0#1+3/F:PRONPERSJ-1/G:0_0/H:4=4@0=13|NONE/I:3_3/J:35+31-14
_^_-i+I=I@1_2/A:0_1_1/B:1-1-2@1-1&1-3#0-2$0-1!3-1;1-2|i/C:1+0+2/D:SYMBOL_1/
E:PRONPERSJ+1@1+3&0+1#1+2/F:PRONPERCD-1/G:4_4/H:3=3@1=12|NONE/I:2_2/J:35+31-14
_^i-I+I=@@2_1/A:0_1_1/B:1-1-2@1-1&1-3#0-2$0-1!3-1;1-2|i/C:1+0+2/D:SYMBOL_1/
E:PRONPERSJ+1@1+3&0+1#1+2/F:PRONPERCD-1/G:4_4/H:3=3@1=12|NONE/I:2_2/J:35+31-14
```

```
['e', 'k', 'w', 'a', '_', '_', 'i', 'I', 'I', '@', 'v', '9', 'j', 'u', 'n', '_', 'n', 'u', 'R', '@', 'v', 'j', 'd', 'R', 'v', 'E', 'R', 'I', '@', 'n', 'O',
'R', '_', 's', 'a', '_', '_', 'E', 's', 't', 'a', '_', 'd', 'i', 'R', 'o', 'p', 'e', 'i', 'd', 'e', 'z', 'O', 'n', 'a', '_', 't', 'e', 'Z', '_', 'a']
```

Create metadata file

<filepath wav>|<text>|<speakerid>|<emotions>|<languageid>

Our Model



ML Model Training Workflow

- Technology: Deep Learning - Pytorch
- Library Usage by Ajinkya and it will be completed by the end of October
- Model Training
  - Approximately 2-3 weeks for training
  - Evaluated by different language speakers
- Languages: English, French

PyTorch

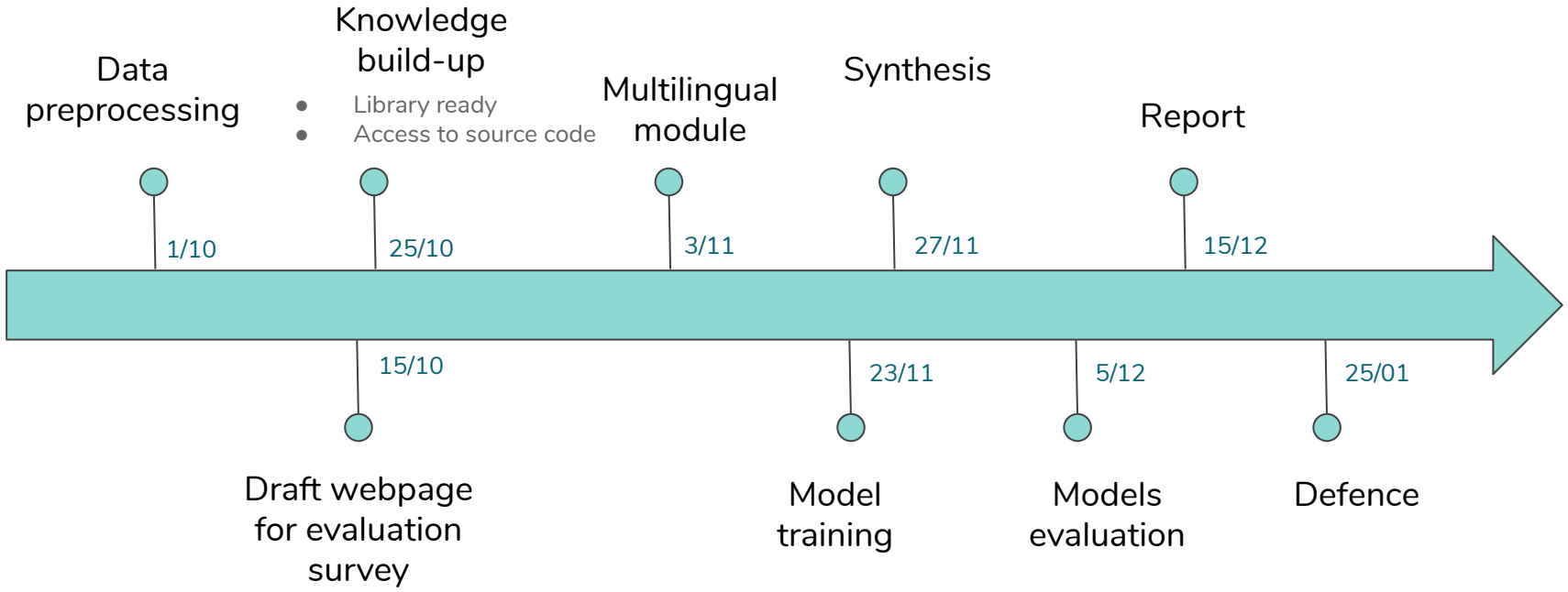
Model Evaluation

- Evaluator - Human
  - Native and non-native speaker
  - Mean Opinion Score (MOS)
- Online evaluation via website:
  - At least 12-15 inputs

Writing Part

- Interspeech 2021 Paper
- Final Report
  - Results interpretation

Timeline


Data preprocessing
1/10
Knowledge build-up
Library ready
Access to source code
25/10
Draft webpage for evaluation survey
15/10
Multilingual module
3/11
Model training
23/11
Synthesis
27/11
Models evaluation
5/12
Report
15/12
Defence
25/01

Thank you for your attention!

DO YOU HAVE ANY QUESTIONS?