Alena YAKAVETS
Cécile MACAIRE
Chanoudom PRACH
Ludivine ROBERT

Software Project

MULTILINGUAL TEXT-TO-SPEECH SYSTEM

M2 NLP 2020-2021

November 24, 2020

Presentation Overview

- What’s Done
- Erisha library
- What’s In Progress
- What’s Next To Do
- Timeline

Items Done

- Website for evaluating synthesized speech
  - UI and text content is finalized
  - Website available both in English and in French
  - *Challenge!* Long-term hosting
- Access to Grid5000
  - Accounts for everyone in the team
  - Each of us will train a specific version of the TTS model
- Hands on to train the original model in the Grid5000 environment
- Library code updated
  - Include the multi-language component
  - Other modifications

How does Erisha library work?

- Model based on Tacotron 2
- Hyperparameters:
  - number of speakers
  - number of emotions
  - number of languages
  - encoder type
  - sampling rate = 22050 Hz
  - ...
- Classes:
  - TextLoader -- loads audio, text pairs, normalizes text and converts them to sequences of one-hot vectors, computes mel-spectrograms from audio files
  - TextCollate -- trains batch from normalized text and mel-spectrogram

Items In Progress

- Training the models (4 different versions):
  - GST - Global Style Tokens
  - VAE - Variational Autoencoder
  - GMVAE - Gaussian Mixture VAE
  - X-vector

- Training time per model:
  - ~2 weeks (2 x 5 days)

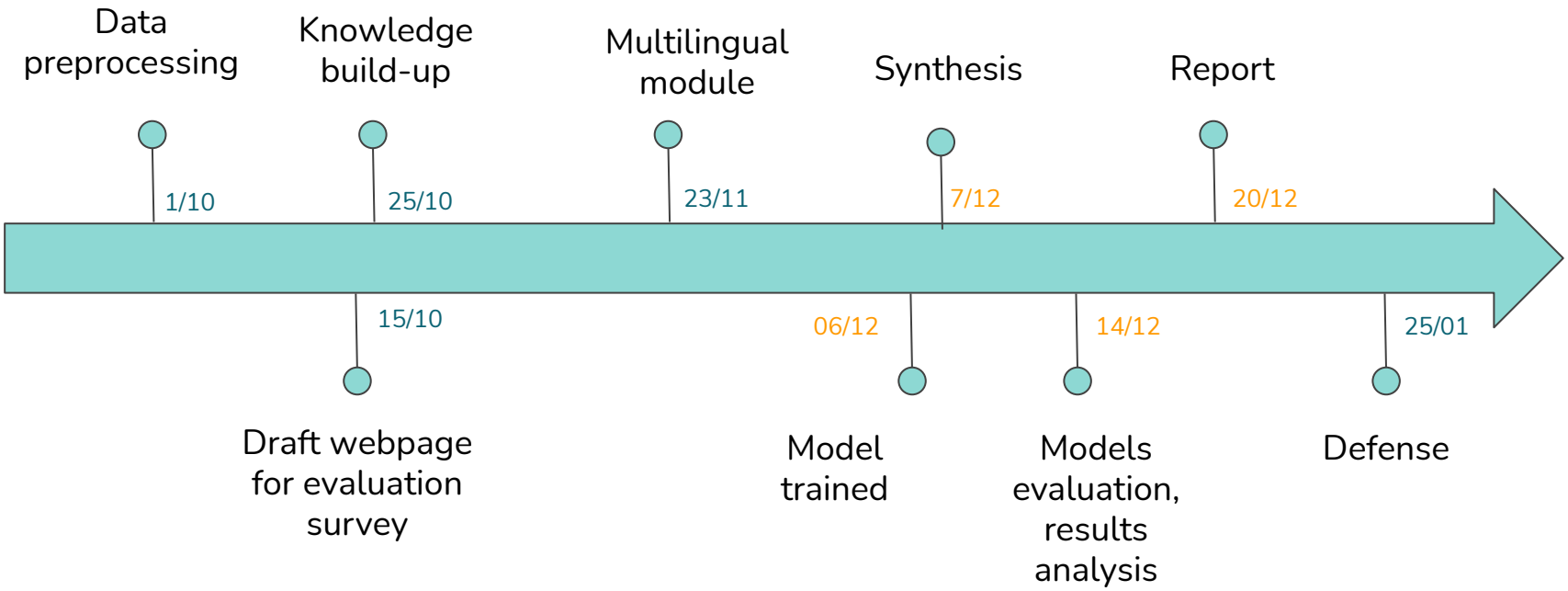
Models for training our multilingual TTS

- **GST - Global Style Tokens**
  - A bank of embeddings that are jointly trained within Tacotron. Trained with no explicit labels, but learn to model acoustic expressiveness independently of text content. Can be used for style transfer, replicating the speaking style of a single audio clip across an entire long-form text corpus.
- **VAE - Variational Autoencoder**
  - Deep generative model. VAE is an autoencoder whose encodings distribution is regularised during the training in order to ensure that its latent space has good properties allowing us to generate some new data.
- **GMVAE - Gaussian Mixture VAE**
  - A variant of the variational autoencoder model (VAE) with a Gaussian mixture as a prior distribution, with the goal of performing unsupervised clustering through deep generative models.
- **X-vector**
  - Deep NN-based embeddings trained on time-delay neural networks with a statistical pooling layer trained for the speaker recognition task. Maps the variable-length utterances to text independent fixed dimensional embeddings which are trained using a deep NN that discriminates between speakers.

Items To Do

- Evaluate synthesized speech
  - Synthesize speech samples for each trained model
  - Upload samples to the website
  - Call-to-action for participants
- Results analysis & interpretation
- Writing
  - Paper for Interspeech conference
  - Final report

Timeline


Data preprocessing
1/10
Knowledge build-up
25/10
Multilingual module
23/11
Synthesis
7/12
Report
20/12
15/10
Draft webpage for evaluation survey
06/12
Model trained
14/12
Models evaluation, results analysis
25/01
Defense

Thank you for your attention!

DO YOU HAVE ANY QUESTIONS?