Alena YAKAVETS
Cécile MACAIRE
Chanoudom PRACH
Ludivine ROBERT

UNIVERSITÉ DE LORRAINE

Institut des sciences du Digital Management & Cognition

UE 905 EC1: Software Project

MULTILINGUAL MULTISPEAKER EXPRESSIVE TEXT-TO-SPEECH SYSTEM

M2 NLP / 2020-2021
University of Lorraine, IDMC

January 26, 2021

Our Team

Alena YAKAVETS

Linguistics

Cécile MACAIRE

Computer Science

Chanoudom PRACH

Computer Science

Ludivine ROBERT

Linguistics

Outline

The project

Model Architecture

Data

Experimentation

Results

Discussion

The Project

Project Recap

WHAT

- ➔ Multilingual Multispeaker Text-to-Speech system
  - ◆ Based on Tacotron 2 [1] (end-to-end generative TTS model, sequence-to-sequence with attention paradigm).
  - ◆ Multispeaker: training models based on several speakers.
  - ◆ Multilingual: able to generate speech based on different language models.
  - ◆ Different encoders: GST, VAE, GMVAE, X-VECTORS
  - ◆ Emotional contour transfer

MAIN GOALS

- ➔ Update ERISHA[1] library to:
  - ◆ Add multilinguality module: English, French
  - ◆ Train with two additional encoders
- ➔ Speech generation trained on sparse data.

[1] https://github.com/ajinkyakulkarni14/ERISHA

Multilingual TTS

Multilingual: Possible Interpretations

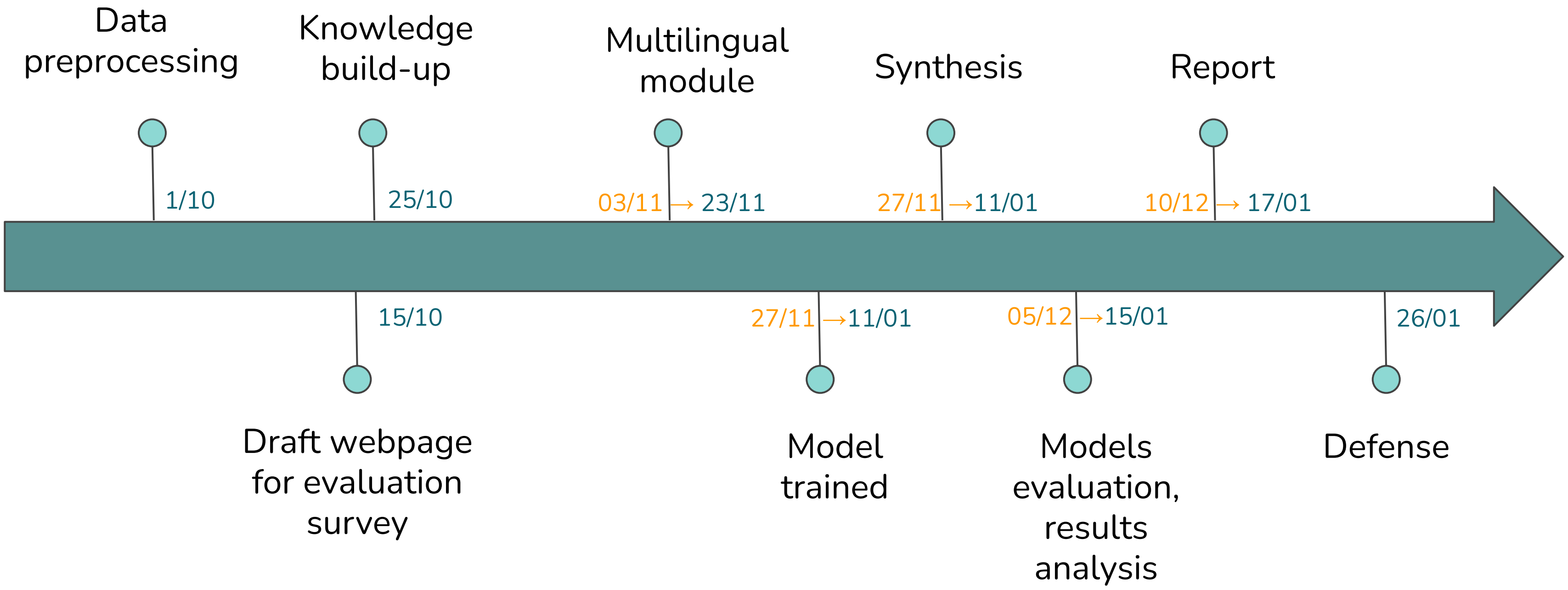
- Train speech synthesis in multiple languages.
- Select the language model to be applied to the whole document/text, regardless of the source language of the text.

- Train speech synthesis in multiple languages.
- Detect language for document or parts of it.
- Produce speech using a dedicated language model for each detected language part accordingly.

Previous work

- Waveform-based statistical speech synthesis system (WaveNet) [2]
- Tacotron [3]
- Text-Predicted Global Style Token (TP-GST) [4]

Timeline


Data preprocessing
1/10
Knowledge build-up
25/10
Multilingual module
03/11 → 23/11
Synthesis
27/11 →11/01
Report
10/12 → 17/01
15/10
Draft webpage for evaluation survey
27/11 →11/01
Model trained
05/12 →15/01
Models evaluation, results analysis
26/01
Defense

Model Architecture

Model Architecture

Baseline: Tacotron 2 [1]
➔ Attention-based sequence-to-sequence model.

Input: text sequence.

Output: sequence of log-mel spectrogram.

New: expressivity, speaker, language encoders.

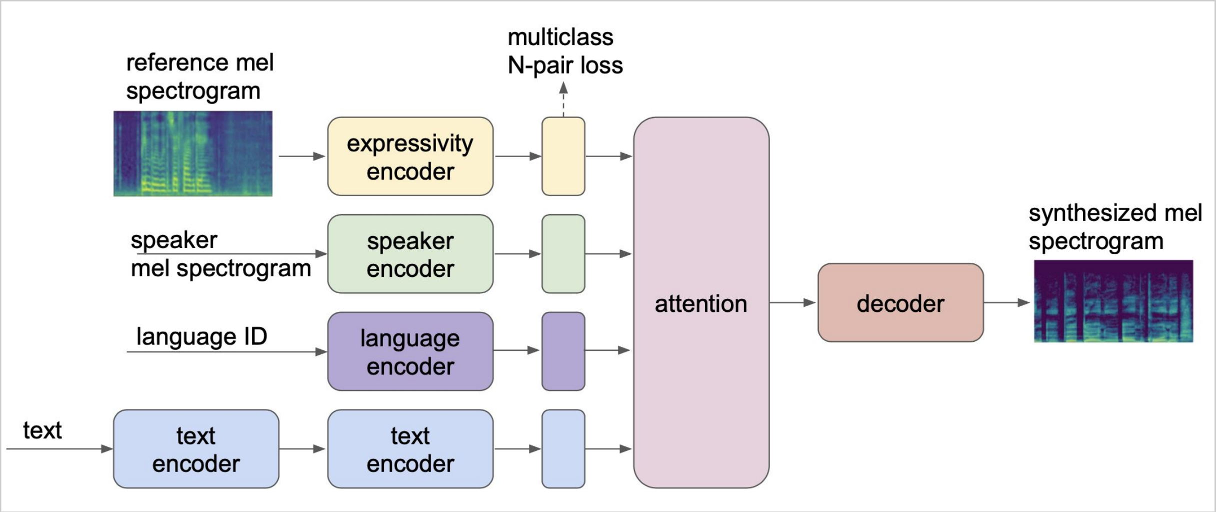


Figure 1: End-to-end Multilingual Multispeaker Expressive Text-to-Speech system.

Model Architecture

Text encoder

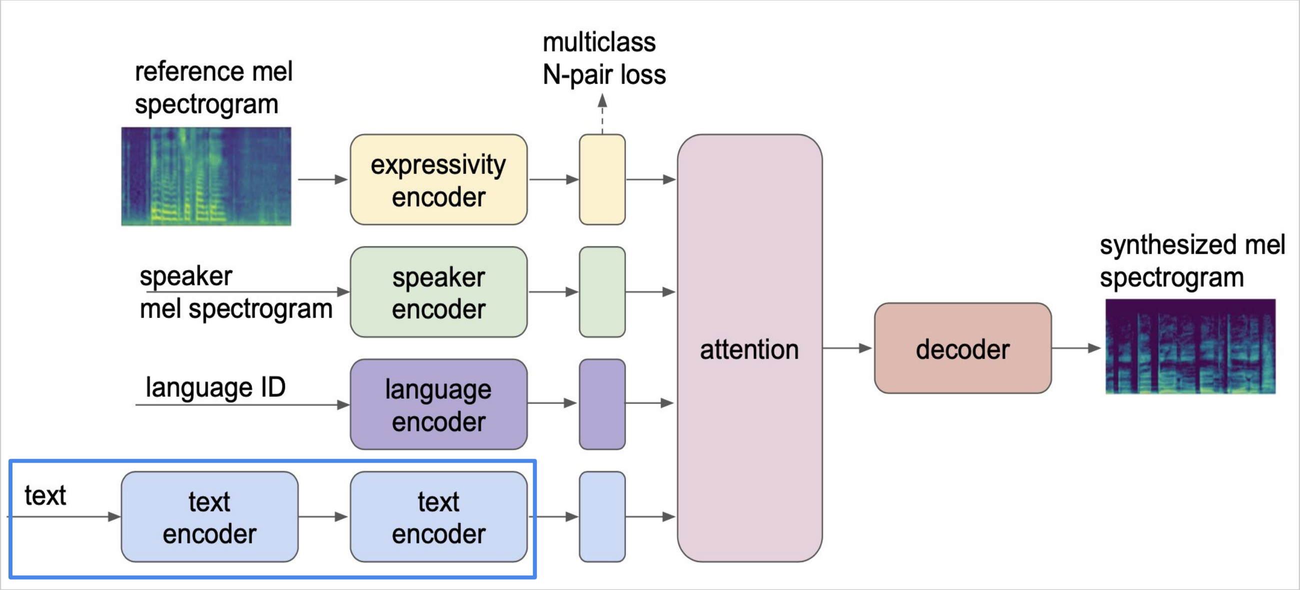


Figure 2: End-to-end Multilingual Multispeaker Expressive Text-to-Speech system.

Goal: produce a latent representation of the input text.

Steps:

1. Conversion into character embeddings,
2. Passed through 3 Conv Layers,
   ↳ N-grams
3. Latent representation z_t generated from the last layer output by a BLSTM RNN layer.

Model Architecture
Expressivity encoder

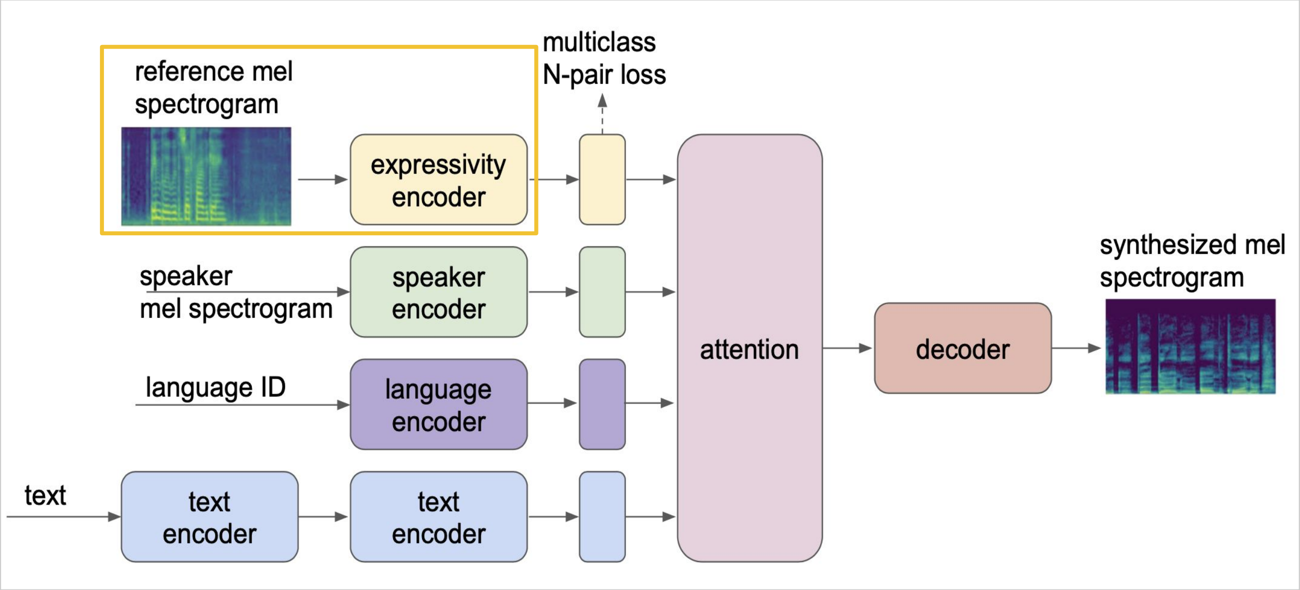


Figure 3: End-to-end Multilingual Multispeaker Expressive Text-to-Speech system.

Goal: produce a latent representation of the emotion.

Input: mel spectrogram.

Output: expressive embedding z_e

Model Architecture

Speaker encoder

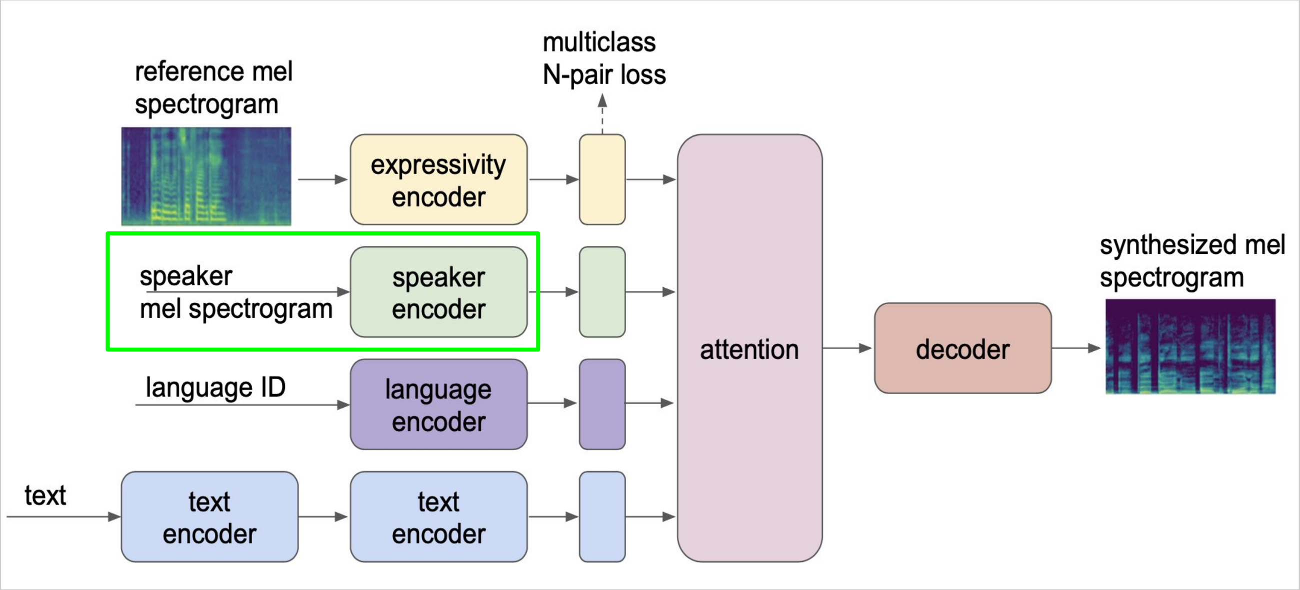


Figure 4: End-to-end Multilingual Multispeaker Expressive Text-to-Speech system.

Goal: capture the properties of different speakers.

Input: reference speech signal.

Output: non linear fixed-dimensional embedding vector z_s

Model Architecture

Language encoder

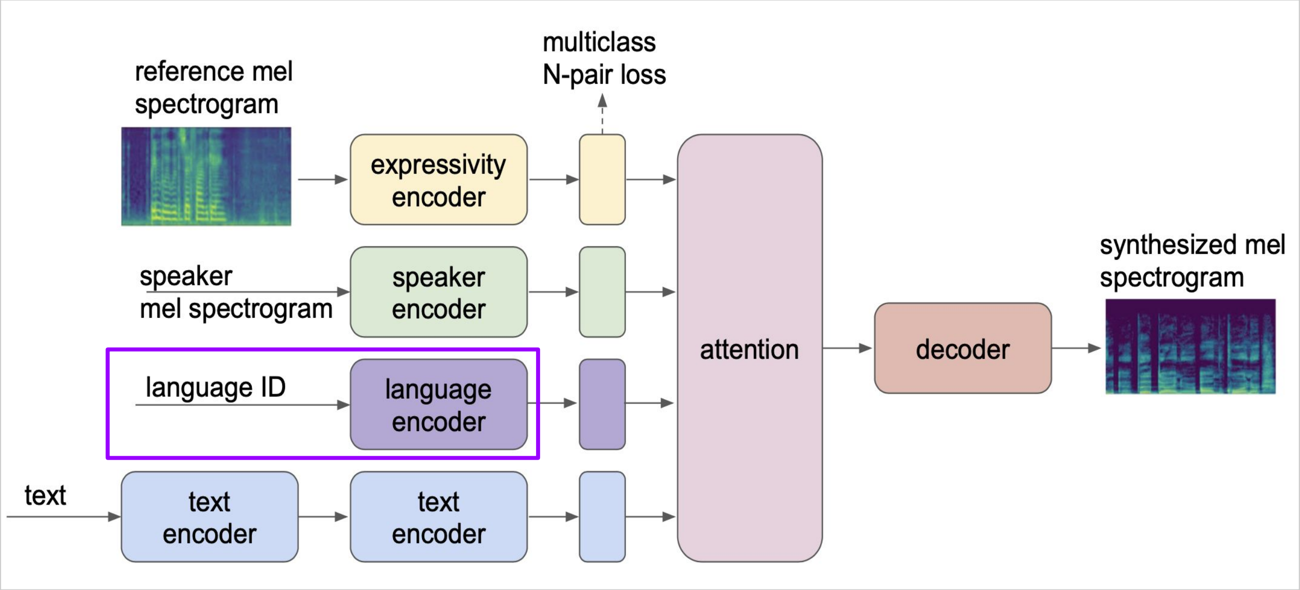


Figure 5: End-to-end Multilingual Multispeaker Expressive Text-to-Speech system.

Goal: encode the languages (multilinguality).

Input: language ID
- ↳ 0: English,
- ↳ 1: French.

Output: embedding z_l

Model Architecture

Multiclass N-pair loss

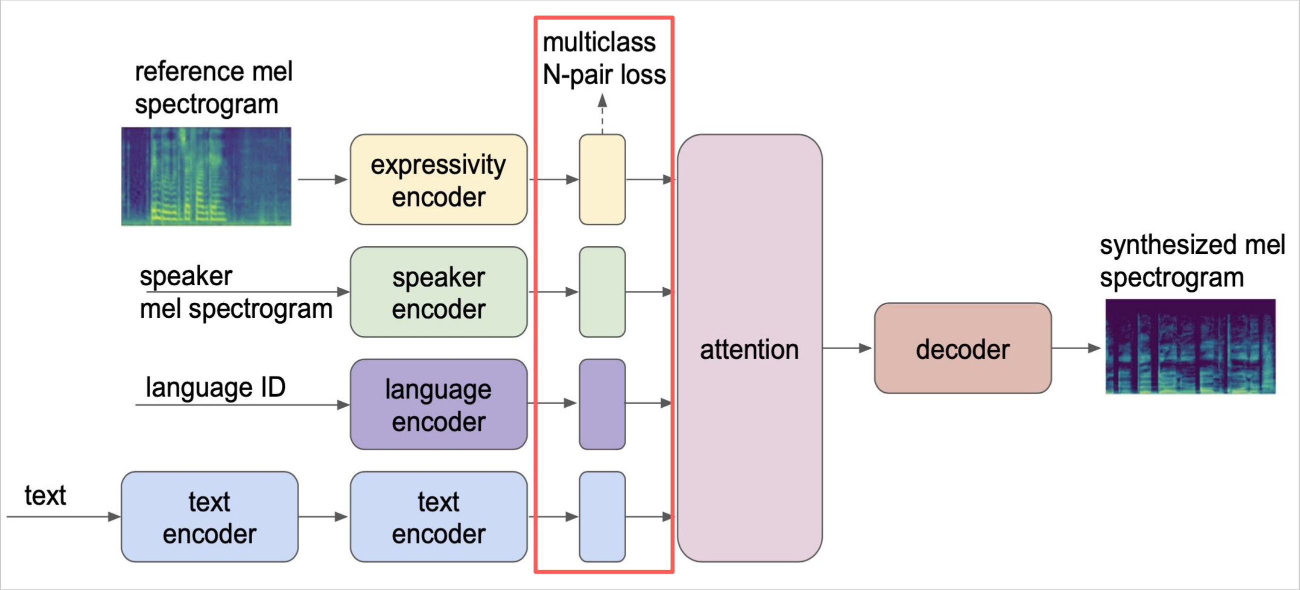


Figure 6: End-to-end Multilingual Multispeaker Expressive Text-to-Speech system.

Idea: enhance expressivity representation to make sure its transfer is correct.

Solution: learning framework with multiclass N-pair loss.

Principle:

- calculates the distance of a baseline input with positive examples (latent variables from the same emotion class) and multiple negative examples,

- reduces the distance between latent variables of the same emotion class.

Model Architecture
Attention module & Decoder

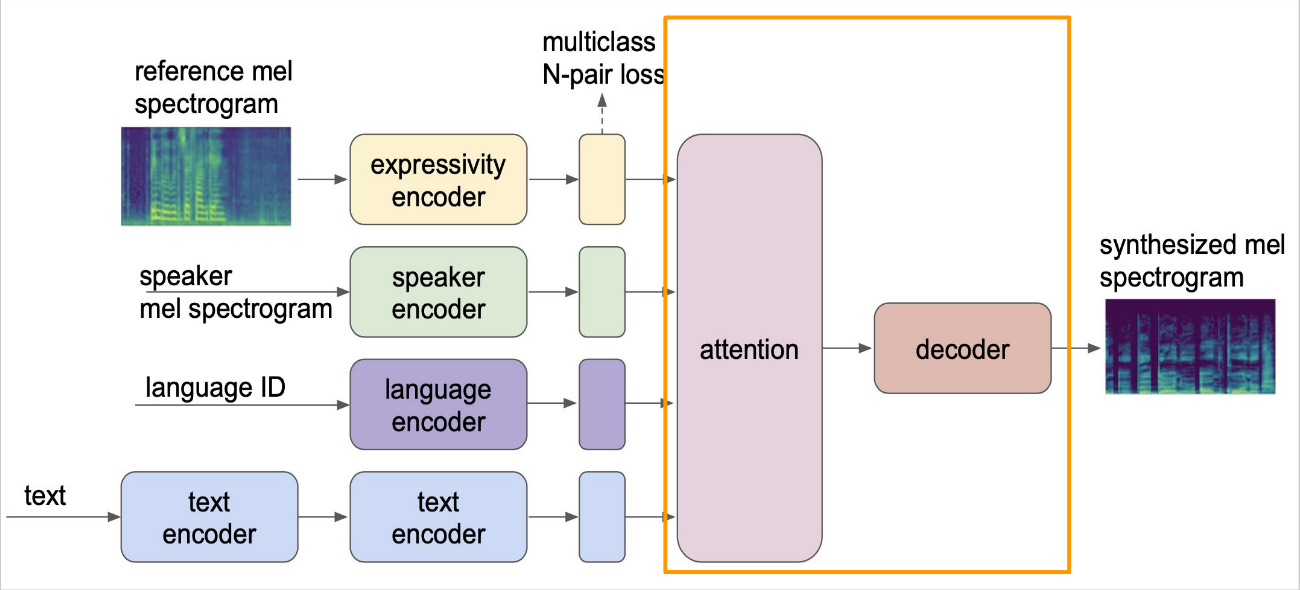


Figure 7: End-to-end Multilingual Multispeaker Expressive Text-to-Speech system.

Attention module

Goal: learn the alignment between the sequence of phonemes and desired mel spectrogram.

Steps:
1. concatenation of z_t, z_e, z_s and z_l into a fixed-length context vector,
2. 128-dimensional hidden representations are generated to compute the attention probabilities.

Decoder

Autoregressive RNN including:
1. pre-net,
2. BLSTM,
3. convolutional layers based post-net.

Neural Vocoder: WaveGlow

Generates speech waveform from mel-spectrograms.

Flow-based network.

Architecture:

1. Squeeze operation – speech samples as vectors,
2. Steps of flows – invertible 1 × 1 convolution and affine coupling layer,
3. Concatenation of the final vectors with the previous output channels – output z.

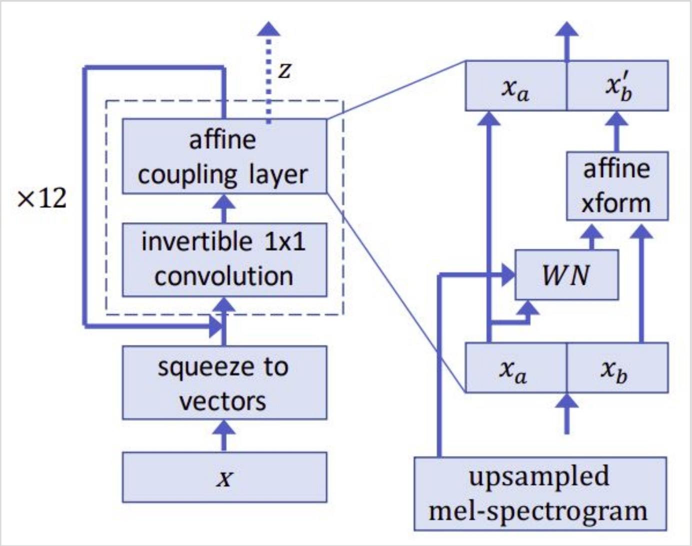


Figure 8: WaveGlow architecture [5].

Encoder: GST

Overview

GST = Global Style Tokens [6]

- ➢ Bank of style embeddings, that are trained together within Tacotron.
- ➢ Trained on expressive speech data with no explicit prosodic labels.
- ➢ Learns to model acoustic expressiveness independently of text content.
- ➢ Yields interpretable embeddings that can be used to control and transfer style.

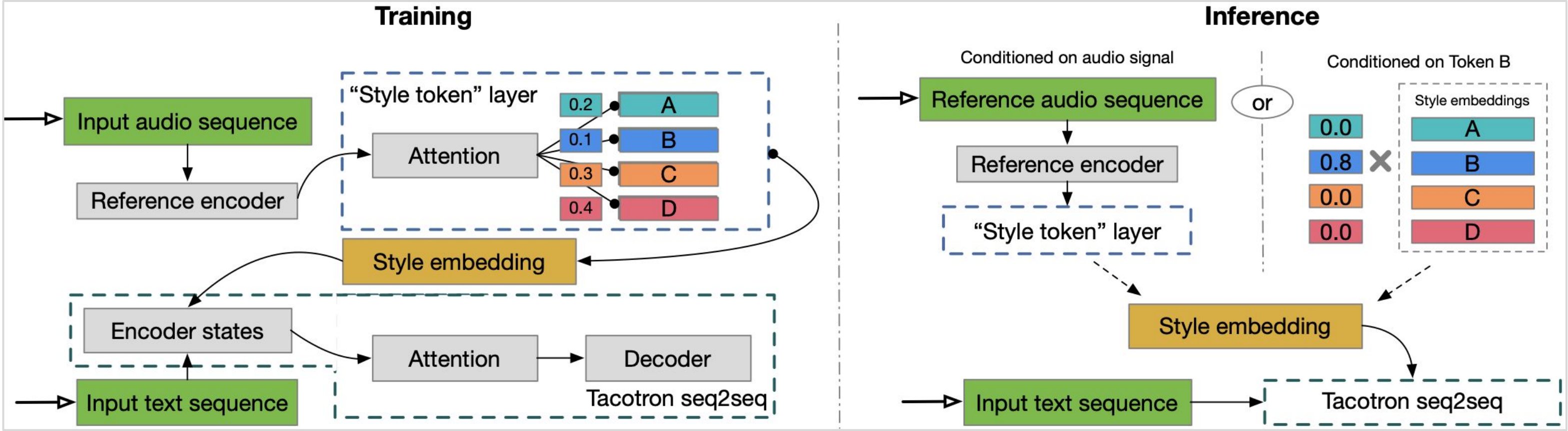


Figure 9: Global Style Tokens architecture [6].

Encoder: GST

Training

Training: train target log-mel spectrogram is submitted to the reference encoder, then processed in the style token layer. The resulting style embedding conditions text encoder states in Tacotron.

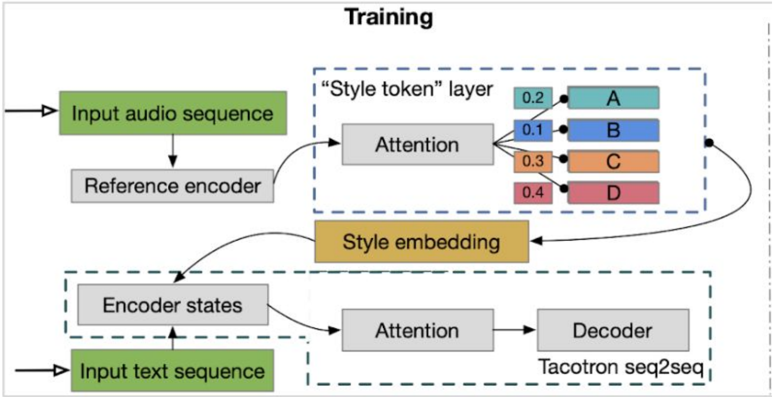


Figure 10: GST architecture - training modules [6].

1. Reference encoder [7]: encodes the prosody of the speech signal into a fixed-length vector (reference embedding).
   Input: log-mel spectrogram.
   Structure: convolutional stack with a RNN.

2. Attention module: style token layer learns similarities between reference embeddings and global style tokens (GSTs).
   Input: generated reference embedding.

3. Text encoder: processes style embedding for conditioning.
   Input: style embeddings.
   Output: set of weights for each style token.

4. Tacotron 2 [1] architecture is trained in parallel with the style token layer.

Encoder: GST

Inference

Inference: text synthesis with a designated speaking style.

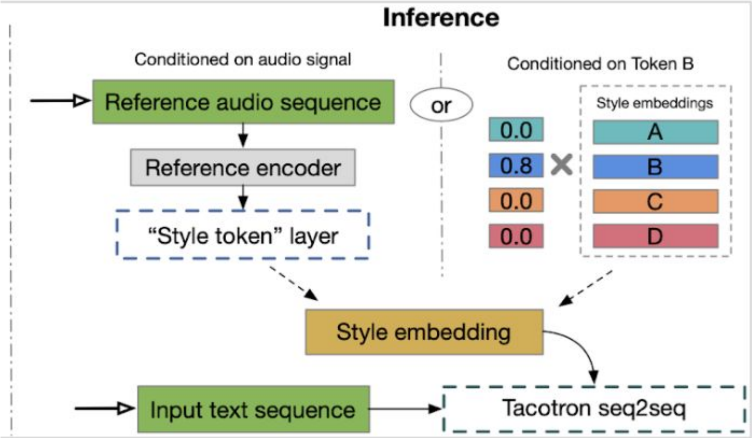


Figure 11: GST architecture - inference modules [6].

<u>Two ways</u>

1. Conditioned on audio signal:
    - reference audio sequence is used in reference encoder for expressive style transfer.
2. Conditioned on token:
    - reference encoder is skipped,
    - select certain tokens (learned interpretable token) to control the style without a reference signal.

Encoder: VAE

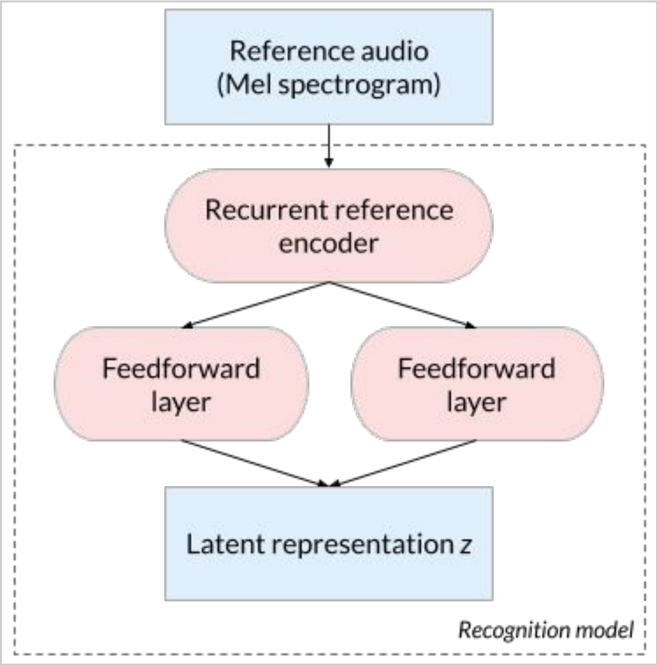


Figure 12: VAE encoder architecture [8].

Variational autoencoder (VAE) [8]: deep generative model.

Idea: map the input onto a distribution.

Architecture:
- ➢ Input: reference speech,
- ➢ Recurrent reference encoder,
- ➢ Two fully connected feedforward layers to generate mean and standard deviation of the latent variable z.

➜ Kullback Leibler (KL) annealing problem.

Encoder: GMVAE

GAUSSIAN MIXTURE VAE - GMVAE [9]: capable of controlling speaker, noise, and style

- ➢ **Comprised of three modules**: *Synthesizer, Latent Encoder* and *Observed encoder.*
- ➢ Latent attributes using a mixture of distribution.
- ➢ Learns an interpretable and disentangled latent representation.

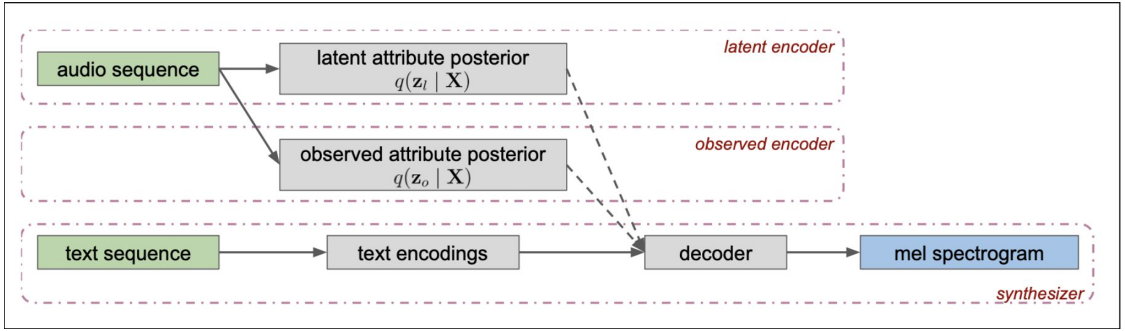


Figure 13: GMVAE architecture.

Encoder: x-vector

X-VECTORS [10] : deep neural network based embeddings.

➔ Efficient in speaker verification and recognition.

Architecture

1. Frame-level:
   - TDNN Layers.
2. Segment-level:
   - Statistical pooling layer,
   - Layers,
   - Softmax output layer.

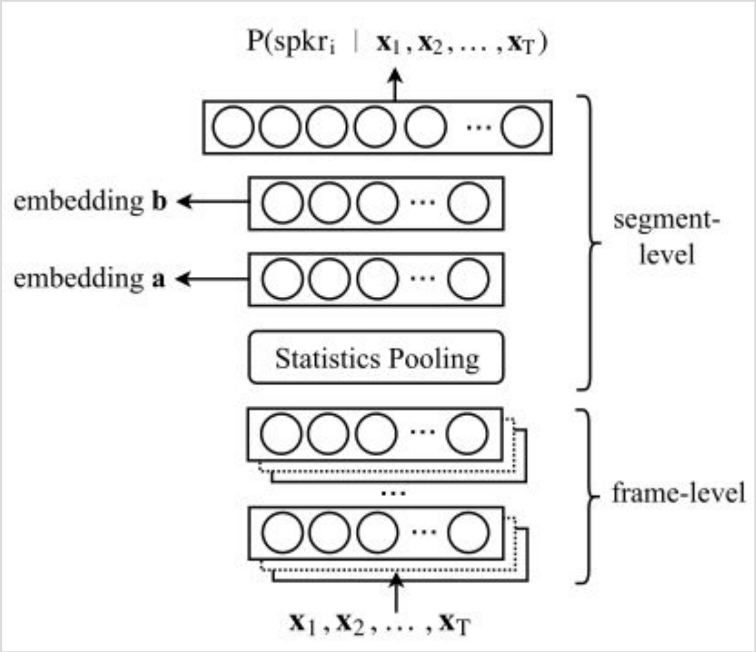


Figure 14: x-vector DNN architecture [11].

Data

Corpus

French: SIWIS[3] [12]

French Speech Synthesis Database

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

English: EmoV-DB[1] [13]

Emotional Voices Database

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

English: LJSpeech[2] [14]

- Emotion: neutral
- Speaker: native
- 13,100 audio clips
- Reading passages from 7 non-fiction books
- ≈24h

[3] https://datashare.ed.ac.uk/handle/10283/

[1] https://github.com/numediart/EmoV-DB
[2] https://keithito.com/LJ-Speech-Dataset/

Sample of the train filelist

```
/srv/storage/multispeechedu@talc-data2.nancy/software_project/corpus/EmoV-
DB/sam/Disgusted/converted/Disgust_85-112_0106.wav|The emotion which she had suppressed
burst forth now in a choking sob.|0|3|0
/srv/storage/multispeechedu@talc-data2.nancy/software_project/corpus/EmoV-
DB/bea/Angry/converted/anger_281-308_0291.wav|The weeks had gone by, and no overt acts had
been attempted.|1|2|0
/srv/storage/multispeechedu@talc-data2.nancy/software_project/corpus/LJSpeech/wavs/LJ008-
0178.wav|the bodies for identification, the wounded to hospitals, a cart-load of shoes,
hats, petticoats, and fragments of wearing apparel were picked up.|5|0|0
/srv/storage/multispeechedu@talc-
data2.nancy/software_project/corpus/SIWIS/wavs/neut_parl_s01_0346.wav|25, 9, 8, 0, 21, 7,
1, 5, 9, 10, 14, 14, 10, 1, 9, 12, 0, 15, 15, 0, 21, 8, 14, 21, 24, 30, 0, 15, 1, 0, 5, 8,
0, 15, 0, 23, 13, 4, 3, 0, 15, 15, 27, 26, 7, 1, 6, 4, 3, 31, 30, 24, 5, 27, 12, 24, 26,
8, 9, 21, 2, 11, 15|4|0|1
```

Figure 15: Extract from input filelist.

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

Experimentation

Experimentation

1. Training 4 models: GST, VAE, GMVAE, x-vector
2. Data:

   - French: SIWIS [12],

   - English: EmoV-DB[1] [13], LJSpeech [14]
3. Model Training: Grid5000[4] [15]

Parameters	Value
Epoch	500
Learning rate	0.001
Weight decay	0.000001
Convolution Layer 1	Kernel Size = 3
Batch Size	1

Table 1: Hyperparameters shared by the models.

```
Validation loss 120000:  4.592939
Saving model and optimizer state at iteration 120000 to /srv/
storage/multispeechedu@talc-data2.nancy/software_project/vae/
output/checkpoint_120000
Train loss 120001 0.182466 Grad Norm 2.207243 1.26s/it
```

Figure 16: Extract from output training file.

2841259	cprach	Running production	-l "{(((type = 'default') AND (type != 'default' OR max_walltime >= 86400 OR max_walltime <= 0)) AND production = 'YES') AND exotic = 'NO'}/host=1,walltime=24:0:0"	PASSIVE	(cluster='grele' or cluster='graffiti' or cluster='grue') AND maintenance = 'NO'	None	24:0:0	2020-12-17 14:23:14	2020-12-17 14:23:15	2020-12-17 14:23:15
2841309	lrobert	Running production	-l "{(((type = 'default') AND (type != 'default' OR max_walltime >= 86400 OR max_walltime <= 0)) AND production = 'YES') AND exotic = 'NO'}/host=1,walltime=24:0:0"	PASSIVE	(cluster='grele' or cluster='graffiti' or cluster='grue') AND maintenance = 'NO'	None	24:0:0	2020-12-17 14:43:36	2020-12-17 14:43:49	2020-12-17 14:43:49

Figure 17: Sample from Grid5000 website.

[4]https://www.grid5000.fr/w/Grid5000:Hom

Results

Samples from Each Model

1. GST
   Samples in English:
2. VAE
   Sample in English:
3. GMVAE
   Sample in French:
4. x-vector
   Sample in French:

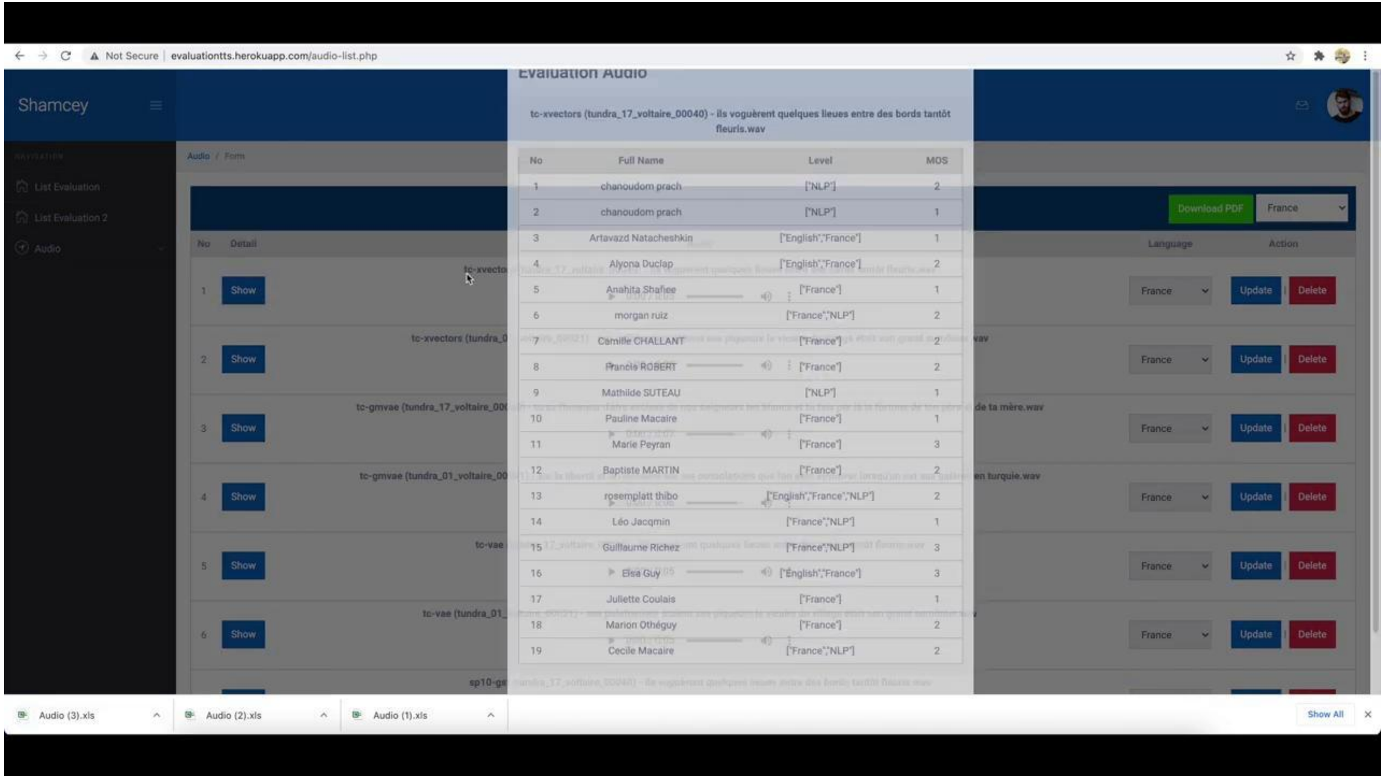
1. It was a curious coincidence.
2. Their forces were already moving into the north country.
3. /anger/ I am going to surprise father, and you will go with Pierre!
4. J’y serai même un peu en retard.
5. Ils voguèrent quelques lieues entre des bords tantôt fleuris.

Evaluation approach

- **Type:** subjective.
- **Where:** dedicated website, open to external participants.
- **Participants:** 14 for EN, 16 for FR. Majority were NLP students.
- **Metric:** Mean Opinion Score (MOS) [16] with absolute category ranking from 1 to 5 to rate naturalness, intelligibility and overall quality.
- **Text display:** yes.
- **Samples per model:** 10 for each language (~2 per speaker).
- **Estimated time for evaluation:** ~20 minutes.
- **Evaluation focus:** speech synthesis (emotional contour transfer dropped).

Evaluation Website Technologies

Evaluation Website Demo

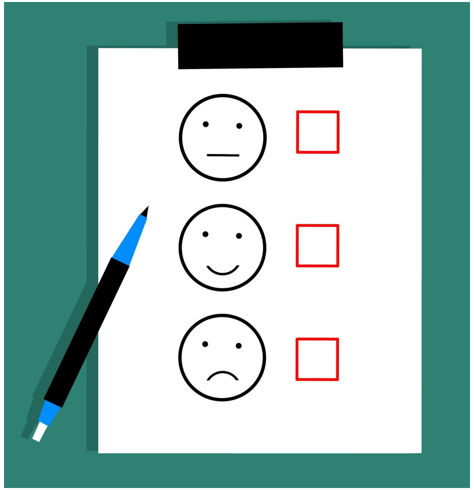


Short Video Clip of the Evaluation Website

Evaluation results

	MOS		
Model	**English**	**French**	**English+French**
x-vectors	2.43	2.59	2.51
vae	2.26	**2.73**	2.50
gst	**2.72**	2.59	**2.66**
gmvae	**2.56**	**2.83**	**2.70**

Table 2: MOS scores from each model.



Observations:

- MOS [16] for all models is below 3, both for EN and FR.
- Best performers for FR: VAE, GMVAE
  - In Erisha[5], GST was better than VAE by 0.25
  - Overall the samples were rated higher than English
- Best performers for EN: GST, GMVAE.
- For EN samples LJ Speech rated better than EmoV-DB (neutral).
- Number of speakers and their dedicated corpus subparts are not equal, which may have caused poorer results for some models.

[5] https://github.com/ajinkyakulkarni14/ERISHA

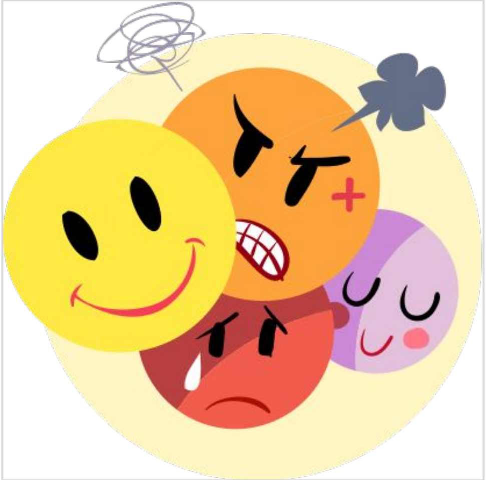
Discussion

Challenges

1. Issues with data preprocessing.
2. Restrictions with Grid5000.
3. Time consuming on ERISHA training.
4. Only performed subjective evaluation.
5. Low performance in emotional generated speech.

Perspectives

On multilinguality, ...



... on the emotion transfer, ...

Parameters	Value
Epoch	?
Learning rate	
Weight decay	
Convolution Layer 1	
Batch Size	

... and on the training parameters.

References

[1] Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., ... & Wu, Y. (2018, April). Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4779-4783). IEEE.

[2] Oord, A. V. D., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., ... & Kavukcuoglu, K. (2016). Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*.

[3] Wang, Y., Skerry-Ryan, R. J., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., ... & Saurous, R. A. (2017). Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135*.

[4] Stanton, D., Wang, Y., & Skerry-Ryan, R. J. (2018, December). Predicting expressive speaking style from text in end-to-end speech synthesis. In *2018 IEEE Spoken Language Technology Workshop (SLT)* (pp. 595-602). IEEE

[5] Prenger, R., Valle, R., & Catanzaro, B. (2019, May). Waveglow: A flow-based generative network for speech synthesis. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 3617-3621). IEEE.

[6] Wang, Y., Stanton, D., Zhang, Y., Ryan, R. S., Battenberg, E., Shor, J., ... & Saurous, R. A. (2018, July). Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. In *International Conference on Machine Learning* (pp. 5180-5189). PMLR.

[7] Skerry-Ryan, R. J., Battenberg, E., Xiao, Y., Wang, Y., Stanton, D., Shor, J., ... & Saurous, R. A. (2018, July). Towards end-to-end prosody transfer for expressive speech synthesis with tacotron. In *international conference on machine learning* (pp. 4693-4702). PMLR.

References (2)

[8] Zhang, Y. J., Pan, S., He, L., & Ling, Z. H. (2019, May). Learning latent representations for style control and transfer in end-to-end speech synthesis. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6945-6949). IEEE.

[9] Wei-Ning Hsu, Yu Zhang, Ron Weiss, Heiga Zen, Yonghui Wu, Yuxuan Wang, Yuan Cao, Ye Jia, Zhifeng Chen, Jonathan Shen, Patrick Nguyen, and Ruoming Pang. Hierarchical generative modeling for controllable speech synthesis. In International Conference on Learning Representations, 2019.

[10] Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., & Khudanpur, S. (2018, April). X-vectors: Robust dnn embeddings for speaker recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5329-5333). IEEE.

[11] Snyder, D., Garcia-Romero, D., Povey, D., & Khudanpur, S. (2017, August). Deep Neural Network Embeddings for Text-Independent Speaker Verification. In *Interspeech* (pp. 999-1003).

[12] Yamagishi, J., Honnet, P. E., Garner, P., & Lazaridis, A. (2017). The siwis french speech synthesis database.

[13] Adigwe, A., Tits, N., Haddad, K. E., Ostadabbas, S., & Dutoit, T. (2018). The emotional voices database: Towards controlling the emotion dimension in voice generation systems. *arXiv preprint arXiv:1806.09514*.

[14] Ito, K. (2017). The lj speech dataset.

[15] Balouek, D., Amarie, A. C., Charrier, G., Desprez, F., Jeannot, E., Jeanvoine, E., ... & Sarzyniec, L. (2012, April). Adding virtualization capabilities to the Grid'5000 testbed. In *International Conference on Cloud Computing and Services Science* (pp. 3-20). Springer, Cham.

References (3)

[16] Streijl, R. C., Winkler, S., & Hands, D. S. (2016). Mean opinion score (MOS) revisited: methods and applications, limitations and alternatives. *Multimedia Systems*, *22*(2), 213-227.t

Thank you for your attention!

DO YOU HAVE ANY QUESTIONS?