LIUM-TTS entry for Blizzard 2023

Félix Saget, Thibault Gaudier, Meysam Shamsi, Marie Tahon

LIUM-TTS entry for Blizzard 2023

Félix Saget[1], *Thibault Gaudier*[1,2], *Meysam Shamsi*[1], *Marie Tahon*[1]

[1]LIUM, France
[2]LIA, France

{felix.saget,thibault.gaudier,meysam.shamsi,marie.tahon}@univ-lemans.fr

Abstract

This paper presents the LIUM-TTS entry for Blizzard 2023. It is the first participation of the LIUM in the Blizzard Challenge. The Blizzard Challenge 2023 focused on French language in two tasks. The Hub task was provided with 50 audio hours (with partially aligned annotation), and the Spoke task only 2 hours. The proposed TTS for the Hub task consists of a Transformer-based grapheme-to-phoneme, a FastSpeech 2-based acoustic model, and a fine-tuned Waveglow vocoder. The output of this system has been fed through a voice conversion module from Hub to Spoke voice. The perceptual evaluation of our system in comparison with other Blizzard participants shows its weaknesses and highlights future working axes to deal with upcoming challenges.

Index Terms: Blizzard Challenge 2023, Text-to-speech, voice conversion,

1. Introduction

The annual Blizzard Challenge intends to compare speech synthesis approaches on a given task. While most Text-to-Speech (TTS) systems are developed for English, the recent editions of Blizzard foster different languages such as Mandarin (2020), Spanish (2021) and French (2023). While the Spanish edition focused on the synthesis of English words within a Spanish sentence, the Mandarin and the French editions tackle the issue of small amounts of data to train a synthesizer. To do so, French audiobook data has been released to participants and two tasks are proposed:

- Hub task: almost 50 hours from a female speaker (NEB)
- Spoke task : almost 2 hours of another female speaker (AD)

With the recent progress in speech synthesis, the quality and intelligibility of the synthetic speech signals are becoming very close to natural speech, thus allowing an extensive use of such technology in various domains. However, while most efforts have concentrated on English audiobooks data, the specificity of other languages or other domains have not been extensively investigated.

One of the main challenges in TTS is to find the harmony between the written text given in input and the realization of prosodic and pronunciation characteristics in the output speech. Previous works investigated the impact of pronunciation in the perception of quality and expressivity of the synthetic speech [1]. The adaptation of the phoneme sequence obtained from the text with a pronunciation dictionary (*canonical*) to the voice corpus has shown a clear preference in terms of quality [2]. In French, four types of phoneme confusions can occur between the *canonical* sequence and the realized pronunciation : 1) related to the speaker itself (for example /o/ $\leftrightarrow$ /ɔ/), 2) elision of final liquids (/l, ʁ/), 3) voice assimilation (devoicing /ʒ/ $\rightarrow$ /ʃ/ or voicing) and 4) deletion of shwa /ə/ [2]. To these phenomenon, one should add the liaison, and the mismatch between the output of the phonetizer and the phonemes as labeled in the voice corpus. For these reasons, a special care has been made in the LIUM-TTS entry regarding pronunciation modeling.

To cope the pronunciation issue, the end-to-end neural TTS systems were first designed to take textual inputs such as the mono speaker Tacotron 2 [3] or the multi-speaker DeepVoice3 [4]. However, some recent approaches still use phoneme sequences in input, as for FastSpeech [5]. This option implies to use an external grapheme to phoneme (G2P) convertor. Data driven G2P converts input graphemes into phoneme sequences which are adapted to the voice corpus. In our submission, we decided to use this last option.

Speech synthesis is theoretically based on the source-filter model, where the source generates voiced/unvoiced signal (pitch, rhythm and energy) and the filter adapts the spectral content to the target phonemes. While Tacotron 2 enables to model the duration with an attention mechanism, the non-autoregressive FastSpeech also incorporates a variance adapter to augment latent information of duration, pitch and energy. A Tacotron 2 architecture with an additional loss between predicted and reference alignments to prevent duration modeling issues has shown nice improvements [6].

One important aspect of the 2023 Blizzard edition is to generate synthetic samples with only 2 hours of speech from the target voice. To do so, multi-speaker TTS can be conditioned on one-hot vectors [4] or speaker embeddings [7, 8]. The last option allows to generate new voices without retraining the TTS system. Another option is to train a mono-speaker TTS, then convert the synthetic speech to the target voice.

Our contribution for Blizzard 2023 is made of different components :

- The front-end is made of a G2P convertor which manages the specificity of written French
- A mono speaker FastSpeech based TTS system
- A voice conversion system based on an adapted Tacotron2 model
- A Waveglow based speech vocoder

2. Data

Each task corresponds to a target voice, either Nadine Eckert-Boulet (NEB) or Aurélie Derbier (AD). NEB has read numerous books whose recordings are available on Librivox. In the SynPaFlex project, more than 87 hours of this voice were extracted and annotated according to various expressive aspects in order

to build a corpus dedicated to French expressive TTS [9]. Indeed, the speaker is able to change her prosody and modify her voice in order to personify some characters with a distinct style from the indirect speech [10].

In the challenge version, a subset of 5 books read by NEB has been selected [11]. The available audio data is summarized in Table 1. Audio data was provided in the form of 16-bit PCM WAV files, each sampled at 22,050Hz. Each file is a complete book chapter containing book sentences read aloud, as well as potential ambient noise and background conversation. Annotation files are in CSV format, with each line corresponding to a segmented speech excerpt, the time in milliseconds at which the excerpt starts and ends within the corresponding chapter file, and a text transcript. We refer to these speech excerpts as "utterances" for the rest of this article.

A statistical examination of the two voices pitch contours on the utterance level shows the pitch distribution of AD is more spread out (average = 164.3Hz, standard deviation = 55.3) than NEB (avg. = 194.9, std = 41.4) who has a more stable voice. We also noticed that NEB speaker has a slight foreign accent when speaking French.

Among the available data, some books have been phonetically aligned, so that, we can retrieve phoneme sequences for a subset of the voices. The phonetic alphabet is derived from SAMPA [12] and includes additional silent characters, combined phones (for instance `k&s` for /k s/) and punctuation marks. This last information is crucial for speech synthesis as the prosodic content adapts to the construction of the utterance. For example, a question mark implies generally a raise in the pitch contour.

As shown in Table 1 an automatic phoneme alignment is provided for all utterances of the Spoke task. In total, 68% of utterances for the Hub task were provided with alignment information.

Task	Speaker	Utterance source	# utt	Length (h)	Alignment
Hub (FH1)	NEB	*Madame Bovary* (G. Flaubert)	14417	11.10	✓
		Les mystères de Paris (E. Sue)	28333	21.54	✓
		Les tribulations d'un chinois en Chine (J. Verne)	1279	1.10	✓
		Les tribulations d'un chinois en Chine (J. Verne)	4876	4.18	-
		La fille du pirate (H.-E. Chevalier)	6040	5.00	-
		Le vampire (P. Féval)	8998	8.61	-
		TOTAL	63943	- 51.53	
Spoke (SH1)	AD	Various audiobooks	1608	1.43	✓
		Parliament transcripts	907	0.65	✓
		TOTAL	2515	2.08	✓

Table 1: *Description of the audio data for the Blizzard Challenge 2023.*

We decided to split each dataset (only the aligned part) as follows:

- NEB: NEB-train (80% of the aligned data), NEB-valid (20% of the aligned data) and NEB-test (unaligned data).
- AD: AD-train (80% of the aligned data), AD-valid (20% of the aligned data).

We investigated the quality of the data, especially, we used an overlapping speech detection on all the data [13]. It is found that in some files from AD dataset, two speakers (a male and AD) were talking but these files were discarded from the training set in the released data.

3. Systems

Speech audio generation is typically divided in three separate steps: 1) a front end which processes grapheme sequences into machine readable representations, 2) an acoustic model that generates an intermediate time frequency representation of signal, and 3) a neural vocoder which produces the synthesized speech waveform.

3.1. Front-end

3.1.1. G2P architecture

A grapheme-to-phoneme (G2P) block is employed to convert the orthographic representations of words into their corresponding phonetic representations. The architecture of G2P, is based on the work from Yolchuyeva et al. [14]. It is a 6-layer forward transformer encoding followed by a dense layer trained using the Connectionist Temporal Classification (CTC) loss function at the word level. We used an available implementation of the architecture[1] with default parameters. The output of the G2P model includes special characters and punctuation marks, preserving the original textual characteristics during the grapheme-to-phoneme conversion process.

3.1.2. Training Data

For the training set of the G2P, we use two types of data. 1) the data provided by the challenge, NEB-train (and AD-train for second task) 2) an additional dataset called Lexique383 [15] which is a French lexical database covering approximately 140,000 French words along with their orthographic and phonetic representations. The words come from a corpus of books and film subtitles. By including Lexique383 in our training data, we enhance the diversity of the dataset. This increased diversity allows the model to better cover various name identities and handle particular or rare words.

3.1.3. Training Steps

Our G2P is trained from scratch using both Lexique383 and NEB-train (and AD-train for second task) in two phases. In the initial training phase, the G2P trained on word level using only Lexique383. This initial phase allows the model to learn the basic mappings from grapheme to phonemes. In the next phase, the G2P is trained on the NEB-train (and AD-train for second task) dataset at the utterance level. In this phase, the G2P would be able to account for contextual factors such as "liaisons" in French pronunciation. This fine-tuning process not only adapts the model to the specific characteristics of the NEB dataset, but also incorporates the speaker style. For example, it can handle cases where different speakers have distinct pronunciations for the same words ("magnifique" is annotated as `/m a n~ i f i k/` or `/m a n j i f i k/` in NEB, when it is annotated as `/m a n i f i k q/` in AD).

3.1.4. Evaluation

To evaluate the performance of the G2P model, we measure the Phoneme Error Rate (PER) on words without special characters. The results are as follows:

[1] `https://github.com/as-ideas/DeepPhonemizer`

- Pretraining on the Lexique383 then fine-tuning on NEB-train: testing on NEB-val yields a PER of 0.9%. This model is used for the first task.
- Pretraining on the Lexique383 then fine-tuning on NEB-train + AD-train: testing on AD-val yields a PER of 1.6%. This model is used for the second task.

These results demonstrate the importance of fine-tuning the model on dataset-specific utterances and adapting to different speaker styles. Evaluation of the first model with the data of the second model shows higher PER (PER of AD-val with first model is 3.4%).

3.1.5. Limitations of the G2P

While one of main objective of using additional dataset has been the augmentation of diversity of words, but our G2P is not still completely compatible to predict the correct phoneme sequence of some name identity. For example, "Nicaragua" would be transcribed as `/n i k a r a g a/` instead of `/n i k a r a g w a/`. Another challenge that a French G2P is facing is the liaison. While our model is able to predict most of the cases but still there are places that it makes an error such the liaison between "premier" and "entretien" (the model prediction is `/p r xˆm j e a~ t r xˆt j e~/` instead of `/p r xˆm j e r a~ t r xˆt j e~/`).

3.2. Mono speaker acoustic model

3.2.1. Model

The acoustic model must be able to predict the frequency content for a given grapheme or phoneme sequence, but also, must infer the duration of each input symbol (duration model). In Tacotron2, the duration model is implicitly tackled by the attention module, but the loss function only includes spectrogram similarity. However, alignment issues, as well as phoneme duplication and skipping problems still occur. To cope with this issue, one option is to modulate the attention block with a phoneme alignment loss which can improve prosody [16].

FastSpeech [5] incorporates a dedicated module for phoneme duration prediction. FastSpeech2 [17] further improves speech quality by learning additional variation information in the form of pitch and energy. Therefore, such system has the advantage of greatly minimizing alignment issues. Furthermore, FastSpeech 2 is trained almost twice as fast as Tacotron 2, and inference time is reduced greatly because of the model's non-autoregressive nature.

Preliminary results have shown that alignment improvements with the modified Tacotron2 did not lead to satisfying results in comparison to the one obtained with FastSpeech2. So we decided to use only the FastSpeech2 system, building upon the public implementation available on GitHub[2].

3.2.2. Training data

To train the mono-speaker acoustic model for the Hub task, we use the aligned subset of NEB-train. In light of the speech quality of our initial results, and because of time constraints, we decided not to use forced alignment to obtain the remaining 32%. We also trained a mono-speaker acoustic model for the Spoke task with the AD-train set, which is fully aligned.

FastSpeech2 models are trained for 1000k iterations each with a batch size of 16, amounting to 408 epochs for the Hub task and 7160 for the Spoke task. The models are trained from scratch with the Adam optimizer, a learning rate of 10^{-3} and an eps value of 10^{-9}. Mean Absolute Error is minimized against target mel-spectrograms derived from the ground-truth audio; pitch (using the WORLD vocoder), energy and duration predictors minimize Mean Square Error against extracted ground-truth values, as described in [17].

[2] `https://github.com/ming024/FastSpeech2`

3.3. Voice conversion system

To build a synthetic AD voice, two options are available: 1) train a mono-speaker acoustic model on AD data (as described in the previous section), 2) convert the synthetic NEB voice into AD voice by using a voice conversion system. We selected the second approach, which is described in sections 3.3.1 and 3.3.2, and we discuss the reasons and limitations relative to this choice in the results section.

3.3.1. General idea

This part describes a Voice conversion system, which is used for the second task (Spoke task). The general idea of how we will achieve this task is shown in Figure 1. We make use of the acoustic model trained on Hub task data to get a first audio sample with another voice (here NEB), and this audio sample is used to extract WavLM features. These features are then fed to a modified Tacotron2, which is trained to reconstruct speech with AD voice from WavLM features coming from Spoke task dataset.

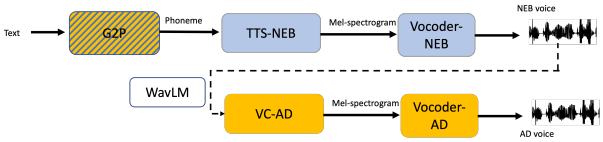

Figure 1: *Global overview of the submitted systems. Blue elements are trained on NEB data, yellow ones are on AD data.*

3.3.2. Model description and training

The model we use is an adaptation of Tacotron2 [3], in which we use WavLM features as an input instead of one-hot embeddings of graphemes or phonemes. In terms of architecture, we replaced the first embedding layer by a linear layer of same size, since the input of this layer will no longer be one-hot encoded. This model is supposed to be trained on single-speaker data of the target voice. At inference, the model is fed with WavLM features extracted from source speech, uttered by source speaker, and since the model only saw target speaker during training, the generated mel-spectrogram represents the source sentence uttered with target voice.

Prior to training, we extracted $T \times 768$-sized WavLM representations from each train sample from AD dataset using the WavLM-Base checkpoint provided by Microsoft [3], where T denotes the time dimension. These features will be used as an input to our model, to predict the 80-band mel-spectrogram. We use Tacotron2 implementation by Nvidia [4] and used a batch size of 32 and a learning rate of 1e-3. Other hyperparameters were left to default values.

[3] `https://github.com/microsoft/unilm/tree/master/wavlm`

[4] `https://github.com/NVIDIA/tacotron2`

3.4. Vocoder

3.4.1. Model

As our neural vocoder, we chose to use Waveglow [18], a non-autoregressive, flow-based generative network. This choice is based on unsatisfactory results with MelGAN and HiFiGAN, which came packaged with FastSpeech 2 in the implementation we used.

3.4.2. Training

A universal checkpoint, pre-trained on the English LJSpeech dataset [19], was made available by the original publishers on the official Waveglow repository[5]. We fine-tuned two versions of this universal checkpoint with the Hub and Spoke task data, respectively, incorporating the unaligned Hub data.

We found out both the universal vocoder (referred to as *Universal Waveglow* in the following) and the fine-tuned models (*FT Waveglow*) to be very close in audio quality. Because the Universal checkpoint benefits from a lower computational cost, we considered it as a valid candidate for our system. The final decision between universal and fine-tuned was taken through an internal perceptual test presented in 4.2.

On inference, we remark that generated audio contains an audibly noticeable amount of high-pitched noise. The official code repository for Waveglow packages the model with a denoiser module, which removes some of the vocoder fingerprint by substracting a noise signal from the original output. The noise is created by traversing the Waveglow layers with a zero-filled tensor; the resulting waveform is then substracted from the original audio. A strength parameter of the denoiser allows for some control over the scaling factor λ of the noise tensor. In our experiments, we found $\lambda = 0.1$ and 0.2 were producing the better sounding results.

4. Results

4.1. Internal objective evaluation

In order to enhance the cost-effectiveness of perceptual tests, an objective evaluation approach is adopted to automatically estimate the quality of synthetic signals. This approach considers various indicators : the loss values of the acoustic model, which encompasses alignment, pitch, and energy errors, in addition to Perceptual Evaluation of Speech Quality (PESQ) [20] and Mean Opinion Score (MOS) prediction.

PESQ was initially developed in the telecommunication domain, and asses the quality of a speech signal by comparing it to a reference. It has been demonstrated to be significantly correlated with perceptual evaluations in the context of synthetic speech [21]. In order to calculate PESQ score[6], the synthetic and original waveforms are downsampled to 16kHz. It ranges approximately from 1 (bad) to 5 (excellent). The MOS prediction model consists of a 2-layer MLP head fed with a wav2vec representation of the signal, referred to as WV-MOS [21]. We use the trained model[7] by the authors on the Voice Conversion Challenge 2018 dataset [22]. MOS scores range from 0 (bad) to 5 (excellent). The Table 2 compares the PESQ and MOS score (the higher, the better) of different configurations of TTS on NEB-val. The confidence intervals are computed on the mean score of NEB test files, assuming a Gaussian distribution, with a confidence level of 0.95.

It is important to acknowledge certain limitations of the evaluation methods employed. While PESQ can provide insights, its suitability for assessing the quality of synthetic speech may be limited, specially when a particular aspect of quality is in question. The PESQ is very powerful to estimate acoustic quality. However, a mistaken pronunciation would probably leads to a good acoustic quality but bad perceptive quality. Moreover, the WV-MOS model used for evaluation was not specifically trained on French speech data, which could affect its accuracy when evaluating French synthetic speech.

System	Universal Waveglow			FT Waveglow		
Metric	no DN	$\lambda = 0.1$	$\lambda = 0.2$	no DN	$\lambda = 0.1$	$\lambda = 0.2$
PESQ	1.24±0.02	1.23±0.02	1.22±0.02	1.27±0.02	1.26±0.02	1.26±0.02
WV-MOS	4.21±0.04	3.63±0.04	3.51±0.04	3.88±0.04	3.82±0.04	3.72±0.04

Table 2: *Objective evaluation of two vocoders with three level of denoising on NEB-val. no DN means that no denoiser is applied while λ corresponds to the strength of the denoiser*

An important observation made during the evaluation is that by increasing the power of the denoiser (comparing $\lambda = 0.1$ and $\lambda = 0.2$ with no denoiser), the fine-tuned vocoder demonstrates a significantly higher score than the universal vocoder. By listening to generated samples, we are convinced that the use of denoiser is necessary, so the fine-tuned Waveglow with denoiser is selected as the vocoder.

4.2. Internal perceptive evaluation

Two AB listening tests were set up in order to select our final submission.

The first test was designed to estimate the impact of fine-tuning the vocoder on NEB data for the Hub task (universal vs. fine-tuned), and the strength of the denoiser ($\lambda = 0.1, 0.2$). 40 sentences from our test set were generated with the mono-speaker TTS model and the 4 vocoders. 16 participants answered to 40 AB pairs. The results show that a denoiser strength of $\lambda = 0.2$ (25.0% with fine-tuned vocoder, resp. 23.4% with universal) is preferred over $\lambda = 0.1$ (16.9%, resp. 18.2%). These results contradict the objective measures (Table 2), highlighting the limitations of current objective measures, such as WV-MOS. One possible explanation is that objective measures tend to focus on specific aspects of quality, which may not be as important in perceptual evaluation. This observation confirms that objective metrics are not entirely reliable.

We also observed that with a denoiser strength of $\lambda = 0.2$, the fine-tuned vocoder was preferred (49.3%) over the universal version (37.7%).

The second test aims at finding the best approach for the Spoke task. To do so, we compare the mono-speaker TTS-AD trained on AD data only with TTS-NEB + VC-AD. We also included re-synthesized samples to have an upper bound. 13 participants answered to 40 AB pairs. From this test we conclude that the conversion pipeline is preferred (48.4%) to the mono-speaker system (39.9%).

To conclude on this evaluation experiment, we choose the two approaches illustrated in Figure 1 for submission:

- Hub Task: mono-speaker TTS-NEB + vocoder fine-tuned on NEB data with a denoiser strength of $\lambda = 0.2$
- Spoke Task: mono-speaker TTS-NEB + Vocoder-NEB ($\lambda = 0.2$) + voice conversion VC-AD + vocoder-AD ($\lambda = 0.2$)

From our work, we can infer some limitations. First, the G2P is not able to generate correct pronunciations in some

[5] https://github.com/NVIDIA/waveglow
[6] https://github.com/ludlows/PESQ
[7] https://github.com/AndreevP/wvmos

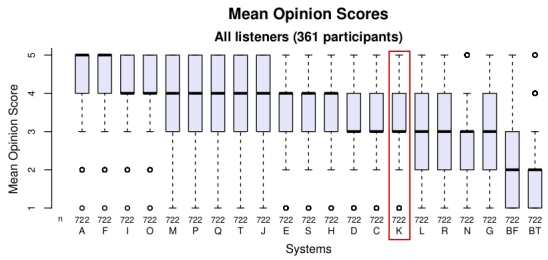

Figure 2: *Boxplot of quality scores of each submitted system for all listeners in the Hub task (NEB).*

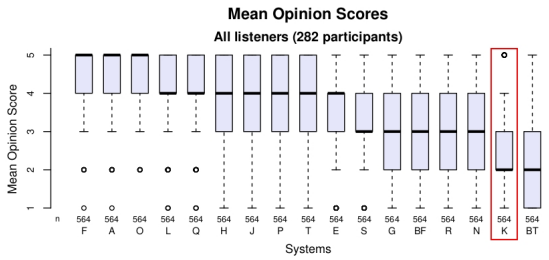

Figure 3: *Boxplot of quality scores of each submitted system for all listeners in the Spoke task (AD).*

cases. Second, the denoiser introduces an artefact that is not stable over the different utterances. Moreover, it slightly modifies the timbre of the voice, what could explain the bad results we obtained regarding speaker similarity.

4.3. Blizzard results

Beside natural voice (A), the generated signals of 18 participant teams and 2 baseline system (BF and BT) have been evaluated for the FH1 task. There were 14 participants for the FS1 task. Our system was affected the letter "K".

4.3.1. Results of Hub task (NEB)

The submitted systems have been evaluated in three main aspects: general quality, intelligibility and similarity with original voice of the NEB speaker. The intelligibility scores of systems has been reported by two metrics. First the homographs (HOMOS) pronunciation accuracy, which evaluates the ability to choose the right phonemes for words that share the same grapheme representation as another word but have a different pronunciation. Second, the intelligibility in the context of different semantically unpredictable sentences (SUS).

The boxplot of mean opinion scores (MOS) in terms of systems' general quality from all listeners in the Hub task (NEB) is showed in the Figure 2. (K) achieved a score of 3.2 (with SD=1.07) in average. When our system is ranked 14 by taking into account all listeners, based on only *Non native and Non speech experts* listeners, it is ranked 12 and obtain an average score of 4.1 (with SD=0.83). This reveals the different quality perspectives in the different target community.

Our system has performed with a homographs pronunciation accuracy of 0.57 (SD=0.50) and a word error rate in SUS context of 0.22 (SD=0.26) in terms of intelligibility. Its ranking stands at 16, which can be concluded that our system faced a bigger difficulty in SUS and HOMOS intelligibility compared. The Blizzard result reveals a similarity MOS score of 2.5 (with SD=1.22) for our system, although a large SD (also a high p-value in pair comparisons) of systems has made the ranking open to question. As an example, two participant systems archived higher similarity MOS (in average) than natural voice.

4.3.2. Results of Spoke task (AD)

For the second task, the generated signals of systems have been evaluated on the aspect of general quality and the similarity with original voice of the AD speaker.

Our system does not perform very well in the Spoke task and has been obtained a quality MOS score of 2.5 (SD=1.05). Its performance in terms of similarity to original speaker is the same as for the overall quality, with a MOS of 2.3 (SD=1.18).

5. Conclusion

The LIUM TTS composed of three main modules adapted for the French language and Blizzard challenge. A Transformer based G2P is trained from scratch to deal with the specificity of read French. For the Hub Task, a FastSpeech based model is employed as the acoustic model and a Waveglow vocoder. For the Spoke Task with limited volume of data from another speaker, a voice conversion approach is followed. In this approach, we developed a WavLM based feature extractor followed by a voice conversion and adapted vocoder. In comparison to an approach where a full system is trained for a second voice, our approach has the advantage of reducing the computational cost by using the former pipeline. The proposed system for the Hub Task has 136M trainable parameters (115M for the Spoke Task). The training runtime for the complete pipeline of Hub Task was 144 hours, while the Spoke Task took 60 hours on a GTX 1080 GPU.

While this is our first LIUM TTS, we have reached acceptable perception results for the Hub task. However, several aspects need to be improved. The performance of the TTS in terms of intelligibility convinced us that not only the G2P should be trained on more data on utterance level, but also a dictionary of proper and foreign names can help a better performance. The results we obtained in terms of speaker similarity, indicates that we raised some signal quality issues with our fine-tuned Waveglow vocoder. We believe that the vocoder fine-tuning and the choice of denoiser can be more optimized.

Even though each module has an objective function that has been optimized during training, we observed that checkpoints corresponding to the lowest error, does not necessarily lead to an improvement of synthetic quality. One of the main difficulties that have been faced is the automatic evaluation of the synthetic voice, and more precisely the objective evaluation of the different modules. We conclude that there is room for improvement in the quality estimation during the development process.

6. Acknowledgements

This study has been realized under the project PULSAR from Region of Pays de la Loire (grant agreement No 2022-09747) and the European Horizon 2020 Research and Innovation Action project SELMA (grant agreement No 957017).

7. References

[1] S. Brognaux, B. Picart, and T. Drugman, "Speech synthesis in various communicative situations: Impact of pronunciation variations," in *Interspeech*, Sep. 2014, pp. 1524–1528.

[2] M. Tahon, G. Lecorvé, and D. Lolive, "Can we Generate

Emotional Pronunciations for Expressive Speech Synthesis?" *IEEE Transactions on Affective Computing*, vol. 11, no. 4, pp. 684–695, 2018. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01802463

[3] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, R. A. Saurous, Y. Agiomvrgiannakis, and Y. Wu, "Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2018, pp. 4779–4783.

[4] W. Ping, K. Peng, A. Gibiansky, S. O. Arik, A. Kannan, S. Narang, J. Raiman, and J. Miller, "Deep Voice 3: Scaling Text-to-Speech with Convolutional Sequence Learning," in *International Conference on Learning Representations (ICLR)*, 2018.

[5] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "FastSpeech: Fast, Robust and Controllable Text to Speech," *NeurIPS*, p. 10, 2019.

[6] V. G. Romillo, I. H. Rioja, and E. Navas, "The AHOLAB Text-to-Speech system for Blizzard Challenge 2021," in *Blizzard Challenge (satellite of Interspeech)*, 2021.

[7] A. Kulkarni, V. Colotte, and D. Jouvet, "Improving transfer of expressivity for end-to-end multispeaker text-to-speech synthesis," in *2021 29th European Signal Processing Conference (EUSIPCO)*. IEEE, 2021, pp. 31–35.

[8] E. Cooper, C.-I. Lai, Y. Yasuda, F. Fang, X. Wang, N. Chen, and J. Yamagishi, "Zero-shot multi-speaker text-to-speech with state-of-the-art neural speaker embeddings," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6184–6188.

[9] A. Sini, D. Lolive, G. Vidal, M. Tahon, and E. Delais-Roussarie, "SynPaFlex-Corpus: An Expressive French Audiobooks Corpus Dedicated to Expressive Speech Synthesis," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018.

[10] M. Tahon and D. Lolive, "Discourse phrases classification: direct vs. narrative audio speech," in *Speech Prosody*, Poznan, Poland, Jun. 2018.

[11] G. Bailly, O. Perrotin, and M. Lenglet, "Ressources for End-to-End French Text-to-Speech Blizzard challenge," Jan. 2023. [Online]. Available: https://doi.org/10.5281/zenodo.7560290

[12] J. C. Wells *et al.*, "Sampa computer readable phonetic alphabet," *Handbook of standards and resources for spoken language systems*, vol. 4, pp. 684–732, 1997.

[13] M. Lebourdais, M. Tahon, A. Laurent, and S. Meignier, "Overlapped speech and gender detection with WavLM pre-trained features," in *Interspeech 2022*. ISCA, Sep. 2022, pp. 5010–5014.

[14] S. Yolchuyeva, G. Németh, and B. Gyires-Tóth, "Transformer Based Grapheme-to-Phoneme Conversion," in *Proc. Interspeech 2019*, 2019, pp. 2095–2099.

[15] B. New, C. Pallier, M. Brysbaert, and L. Ferrand, "Lexique 2: A new french lexical database," *Behavior Research Methods, Instruments, & Computers*, vol. 36, no. 3, pp. 516–524, 2004.

[16] X. Zhu, Y. Zhang, S. Yang, L. Xue, and L. Xie, "Pre-alignment guided attention for improving training efficiency and model stability in end-to-end speech synthesis," *IEEE Access*, vol. 7, pp. 65 955–65 964, 2019.

[17] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," in *International Conference on Learning Representations*, 2021.

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

[19] K. Ito and L. Johnson, "The lj speech dataset," https://keithito.com/LJ-Speech-Dataset/, 2017.

[20] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*, vol. 2. IEEE, 2001, pp. 749–752.

[21] P. Andreev, A. Alanov, O. Ivanov, and D. Vetrov, "Hifi++: a unified framework for bandwidth extension and speech enhancement," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[22] J. Lorenzo-Trueba, J. Yamagishi, T. Toda, D. Saito, F. Villavicencio, T. Kinnunen, and Z. Ling, "The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods," in *The Speaker and Language Recognition Workshop*. ISCA, 2018, pp. 195–202.