The GIPSA-Lab Text-To-Speech System for the Blizzard Challenge 2023

Martin Lenglet, Olivier Perrotin, Gérard Bailly

18th Blizzard Challenge Workshop
29 August 2023, Grenoble, France

The GIPSA-Lab Text-To-Speech System for the Blizzard Challenge 2023

Martin Lenglet, Olivier Perrotin, Gérard Bailly

Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-Lab, France
{martin.lenglet,olivier.perrotin,gerard.bailly}@grenoble-inp.fr

Abstract

This paper describes the GIPSA-Lab submission to the Blizzard Challenge 2023. The Text-To-Speech system trained for this challenge is a Transformer-based non-autoregressive encoder-decoder architecture based on FastSpeech2. Updates of the FastSpeech2 framework were provided to specifically train the model on orthographic inputs, which is our main focus for this edition of the challenge. This model was trained with both orthographic and phonetic transcriptions of the same dataset. An additional phonetic prediction layer was added to the model. This additional layer enables to train the text encoder on phonetic prediction alone, without the need for audio recordings.

Index Terms: speech synthesis, mixed-inputs TTS, phonetic prediction

1. Introduction

Latest neural networks breakthroughs have largely improved performances of various automatic speech processing tasks, including Text-To-Speech (TTS). Latest neural TTS [1, 2, 3, 4], combined with neural vocoders [5, 6, 7] generate synthetic voices that closely mimic natural speech. However, the evaluation of synthetic speech naturalness is mostly conducted in favorable environments, on test stimuli which are very close to the training corpus. Thus, the good performances shown by neural TTS models may be overestimated compared to real-life applications.

The Blizzard Challenge 2023 aims at evaluating latest neural TTS systems in more challenging environments. More specifically, the Hub-Task of this challenge includes the evaluation of intelligibility of semantically unpredictable sentences and heterophonic homographs. The Spoke-Task on the other hand is a speaker adaptation task on a smaller dataset shared by the Blizzard organizers. Only orthographic sequences can be used as inputs in the submitted systems.

Our approach to this challenge is to propose a TTS system very close to the state-of-the-art model FastSpeech2 [4] but with the addition of phonetic prediction sub-task. FastSpeech2 is a fully parallel Transformer-based [8] architecture which implements 3 secondary tasks on top of the spectrogram prediction: pitch, energy and duration prediction. The duration prediction is the key factor of this parallel architecture, since it is necessary to realize the phone-to-frames alignment at the interface between the text encoder and the audio-decoder. However, this duration predictor is also the limiting factor to train FastSpeech2 on orthographic inputs, since the time-segmentation of the training set, necessary to train this predictor, is unclear when processing orthographic sequences. In this paper, we show how the letter-to-sound alignment proposed by Lenglet et al. [9] can be used to assign duration to the orthographic sequences in order to train a FastSpeech2 model on orthographic inputs. Moreover, we show that the addition of a phonetic prediction task from the output of the FastSpeech2 text encoder allows to train the model on <orthography|phonetic> pairs without the need for audio recordings. This setup helps learning phonetic transcriptions for words and contexts that are otherwise rarely found in classical audiobooks training corpora. We show through Blizzard results that this addition helped modelling heterophonic homographs. Results also show that our model is perceived as more natural than the FastSpeech2 baseline.

This paper is organized as follows: Section 2 describes our proposed model and the letter-to-sound mapping used to train our FastSpeech2 on orthographic sequences. Section 3 describes the extended dataset we used to train our model, and the training procedure. Prior to the Blizzard Challenge results, we evaluated the accuracy of the proposed phonetic prediction layer in section 3.3. Finally, results of the Blizzard evaluation are discussed in section 4.

2. Model: FastSpeech2 with mixed inputs

This section describes FastSpeech2 baseline architecture enhanced with the proposed phonetic prediction layer. The overall architecture of the proposed model is shown in Fig.1. The implementation is available online[1].

2.1. Model Architecture

The proposed model is very close to one of the open source FastSpeech2 implementation [4]. The encoder, variance adaptor and decoder are kept unchanged[2]. Following early implementations of FastSpeech2, the pitch predictor is trained on raw pitch values in semitones, instead of continuous wavelet transforms [10] in latter works. Pitch and energy values are extracted using WORLD pre-processing toolbox [11], and are averaged by phonemes, and normalized. The same multi-speaker model is used for the Hub-task and the Spoke-Task of this Blizzard Challenge. Speaker control is achieved through the addition of a trainable speaker embedding at the output of the text encoder. The model is trained on both orthographic and phonetic input sequences, following the mixed-inputs training procedure [12].

Following [9], an additional phonetic prediction layer is added at the output of the text encoder. This layer predicts a one-to-one mapping between orthographic inputs and phonetic outputs. This one-to-one letter-to-sound mapping (L2S) is further described in Section 2.2. The goals of this layer are twofold: first, it helps disambiguating homographs as shown in [13]. Second, it enables to train the text encoder

[1] https://github.com/MartinLenglet/Blizzard2023_TTS

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

10.21437/Blizzard.2023-3

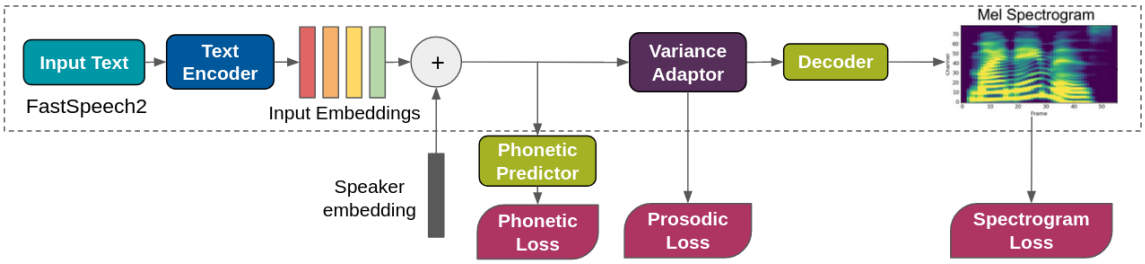


Figure 1: *Model Architecture of the multi-speaker FastSpeech2 baseline with the phonetic prediction layer. This phonetic prediction layer is trained on the output of the text encoder.*

Table 1: *Technical specificities and performances of the proposed FastSpeech2 with mixed inputs and vocoder Waveglow. Inference speed is reported as the Real-Time Factor (RTF). The loading time is the duration needed to load the model before starting the inference. This duration is not considered to compute the inference speed. Performances are computed on a single GPU Quadro RTX 8000.*

Model	# Parameters	Memory Footprint (Mbytes)	Loading Time (s)	Inference Speed (RTF)
FastSpeech2	35 630 466	1 600	4.5	1.58×10^{-2}
Waveglow	87 879 272	2 400	3	5.31×10^{-2}
Total	**123 509 738**	**4 000**	**7.5**	6.89×10^{-2}

on <orthography|phonetic> pairs without the need for corresponding audio. This eases the training of models out of audiobooks corpora, e.g. through the use of dictionaries. The cross-entropy phonetic loss trains the model on a categorization task. This loss is added to already existing MAE spectrogram-loss and MSE pitch, duration and energy-losses. The same lexicon as the Blizzard organizers was used.

Training FastSpeech2 on orthographic inputs is usually tricky, since the training of the explicit duration predictor relies on the time-segmentation of characters in the training dataset. When using phonetic sequences, every input character is attributed a unambiguous duration, either through expert analysis of the audio signal, or with automatic tools like Montreal-Forced Aligner [14]. On the other hand, in the case of opaque languages like French which require a wider visual attention span to achieve the grapheme-to-phoneme (G2P) transcription [15], it is unclear how to distribute the duration between the multiple orthographic characters involved in one phoneme, called complex phoneme in the following. Thanks to this one-to-one L2S mapping, we are able to attribute the duration to the character of interest in case of complex phonemes, and a null duration to the other characters involved. This procedure enables to train FastSpeech2 with orthographic inputs, without relying on a front-end phonetic transcription. As a result, the raw orthographic sequence is used as is during inference.

The vocoder used is Waveglow [6]. The original architecture remains unchanged[3]. The technical specificities and performances of our system are summed up in Table 1.

2.2. One-to-one Letter-to-Sound Mapping

Following the exploration of the attention map of a fully trained Tacotron2 TTS model [2], a one-to-one L2S mapping was proposed by Lenglet et al. [9]. The main results of this study are reported in this section. This mapping is deduced from the number of frames which focus on a particular grapheme in case of complex phonemes. Examples of most commonly seen patterns are given in Fig.2. Empirical rules were deduced from these observations, summed up in Table 2. The symbol /_ / is assigned

[3] https://github.com/NVIDIA/waveglow

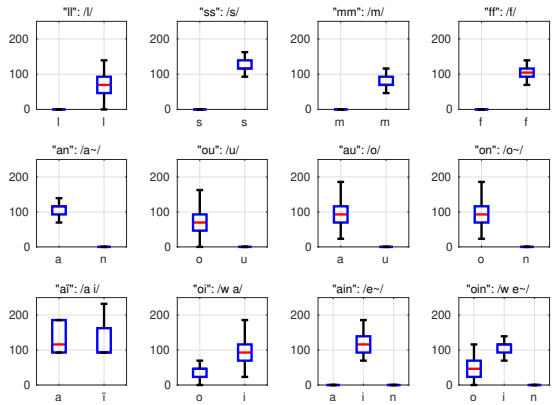


Figure 2: *Distributions of durations of activation (ms) of common character sequences in complex phonemes.*

Table 2: *Activation rules on grapheme recurrent schemes. C and V stand for consonant and vowel respectively. _ stands for muted character.*

Schemes	Activation	Examples
C C	_ C	“nn”, “ll”, “ss”
V V	V _	“an”, “ou”, “au”
V V V	_ V _	“eau”, “ain”

as output of this one-to-one mapping for muted characters.

This L2S mapping is used twice to train the model. First, characters durations when using orthographic inputs are attributed following the rules given in Table 2. This enables to train FastSpeech2 directly on orthographic sequence, and use raw orthographic sequences at inference. Thus, it enables the model to handle French liaisons on its own, which can otherwise be an issue with G2P front-end [16]. Similarly, the FastSpeech2 encoder is also able to learn how to disambiguate homographs by relying on the contextualisation provided by its successive Transformer layers [8].

Second, the phonetic prediction layer uses this L2S mapping as targets to be predicted from the input sequence in case of orthographic inputs. This phonetic prediction layer further helps the disambiguation of homographs at the output of the

encoder. In case of phonetic inputs, the input phoneme is set as the target of the prediction layer (except for punctuation marks and spaces, which are given two possible outputs: /_ / in case of null duration, or /__ / otherwise).

3. Training and Early Evaluation

3.1. Dataset

The same multi-speaker model described in section 2 was trained for both the Hub-task and the Spoke-task. To strictly follow the mixed-inputs training, utterances without phonetic alignment were excluded from the training set (19 986 out of 64 015 utterances for speaker NEB). 3 200 utterances (5% of the NEB corpus) were randomly picked in this excluded portion of the corpus as the validation set. In order to maximize the multi-speaker performances of our model, 2 additional speakers were added to the Blizzard dataset. The whole aligned corpus was included in the training set. Following Blizzard rules for the challenge, the two additional speakers are taken from open-access online databases, specified in Table 3. A part from this extended training dataset, our model is not specifically designed to achieve few-shot speaker adaptation. Nonetheless, we took part in the Spoke-Task to evaluate the benefits of our mixed representations FastSpeech2 in this context.

Moreover, since the phonetic prediction layer enables the training of the text encoder without audio recordings, we also added to the training set phonetic transcriptions from the ROBERT French-dictionary, as well as common in-context homographs. These homographs are taken from various online open-access articles, similar to [13]. The training set is further described in Table 3.

The audio output is a 80-bands Mel-spectrogram computed on the 22 050Hz audio signal with an hop-size of 256 (which is equivalent to a spectrogram sampling rate of $\approx$ 86Hz).

3.2. Training Procedure

The non-audio inputs (dictionary and homographs) are used at every stage of the training process. They are mixed with audio-inputs in each batch, with a ratio of 2/3 for audio inputs and 1/3 for non-audio inputs. While the training on non-audio inputs helps learning phonetic representations for rare words not seen in the audio corpus, this ratio minimizes the risks of degradation of the prosodic predictions and audio quality due to the absence of spectrogram-loss on the non-audio part of the corpus.

Our model was first trained following a single-speaker setup on NEB. We believe that this step helps the text encoder and decoder to focus on their primary goal which is the modulation of acoustic and prosodic local patterns according to the sequence to utter. The addition of the speaker embedding latter in the process is seen as an offset manipulation of these mean features, which is supposedly easier to learn by the model.

Table 3: *Multi-Speaker Training Dataset. Durations are given in hh:mm:ss.*

Speaker	Metadata		Audio	
	Dataset	**Gender**	**Duration**	**# Utt**
NEB	Blizzard	Female	33:33:41	44 029
AD	Blizzard	Female	2:04:53	2 515
DG	LibriVox [17]	Male	6:17:22	7 539
RO	SIWIS [18]	Female	0:35:21	586
Dictionary	Robert	-	-	95 879
Homographs	Various [13]	-	-	17 285
Total	**-**	**-**	**42:31:17**	**167 833**

The model was trained for 100 epochs on NEB only, using both orthographic and phonetic transcriptions. The batch size is set to 32. All utterances are presented twice by epoch: once with the orthographic input and once with the phonetic input. Batches are randomly selected among the whole training corpus, resulting in a mixture of speakers and input types in each batch. This mixture is not supervised.

The learning rate was fixed to 10^{-3} during this first step. Following the 2/3 - 1/3 ratio, this training includes about 50 epochs on the non-audio corpus. Following this initialization step, all other speakers were added to the training set. The learning rate exponentially decreased from this step, to reach 10^{-4} after 170 epochs. The training continued with 50 epochs on the multi-speaker corpus. When training with the multi-speaker setup, dictionary inputs are duplicated for each speaker, in order to train the phonetic predictor's dependency to the speaker. Finally, the model was trained on an evenly distributed corpus among speakers for another 50 epochs. Utterances were randomly selected to match the number of utterances in the AD corpus, when enough utterances were available. All utterances from RO were kept for this final training step. We empirically found that this final step helps modelling rarest speakers behaviors instead of copying the behavior of the most seen speaker.

The vocoder Waveglow [6] was fine-tuned from the pre-trained model shared with the GitHub implementation. The fine-tuning was performed on the NEB corpus, first for 50 epochs on the Ground-Truth spectrograms, and then for 50 additional epochs on spectrograms predicted by the FastSpeech2 model.

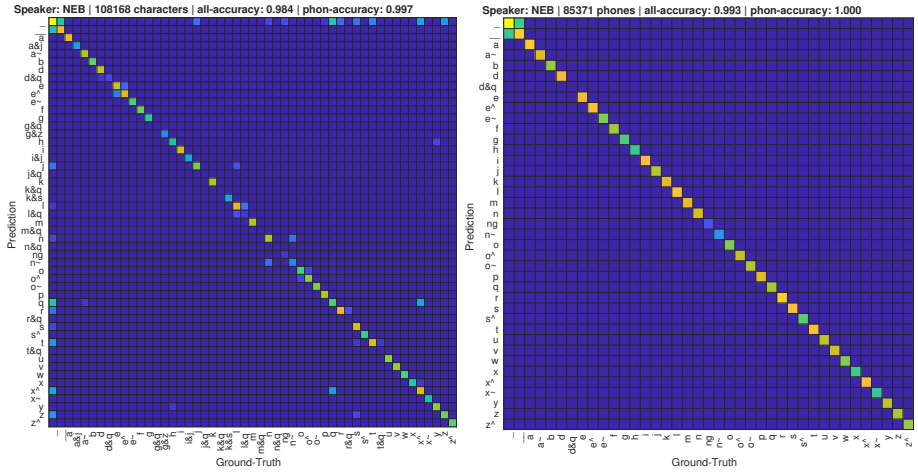
3.3. Phonetic Prediction Evaluation

On top of the evaluation performed for the Blizzard Challenge, we evaluated the performances of the phonetic prediction layer, as an indicator of the potential benefits of the proposed architecture compared to the traditional FastSpeech2 training.

As a test set, we randomly extracted 2230 additional utterances recorded by the same NEB speaker from the original M-AILABS corpus [19]. These utterances are not part of the dataset shared by Blizzard organizers, thus they have not been seen by the model during the training phase.

The phonetic prediction was computed on this test set, and confusion matrices are reported in Fig.3, using orthographic inputs (Fig.3a) and phonetic inputs (Fig.3b). Among the 108168 orthographic characters of this test set, the overall accuracy reaches 0.984 (0.997 when excluding muted characters). Interestingly, most remaining errors are confusions between close phonetic variants: mid-closed vowels VS mid-opened vowels, and full closed vowels VS semi-vowels. Most errors with muted characters are miss-predicted liaisons on ending /r/, /t/ or /z/. Note that the errors highlighted here may just reflect divergences between the Ground-Truth and the model decision on optional liaisons. On the other hand, when using phonetic inputs, this prediction is almost flawless, reaching 0.993 overall, and 1.00 when excluding spaces and punctuation marks.

While this evaluation of the phonetic accuracy of the proposed model is promising regarding the production of heterophonic homographs and the intelligibility of semantically unpredictable sentences, these tasks are specifically designed to test the model out of what has been seen during the training. Thus, results may differ on these specific tasks.



(a) *Orthographic Inputs* (b) *Phonetic Inputs*

Figure 3: *Confusion Matrices of the phonetic prediction layer, for orthographic inputs (left) and phonetic inputs (right).*

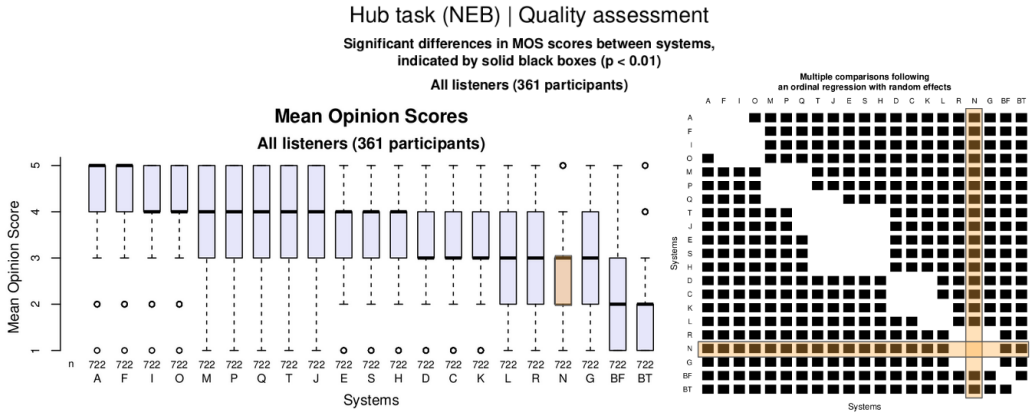


Figure 4: *Mean Opinion Scores (MOS) for the Hub Task. Our model N is highlighted in orange. The left graph shows MOS by system. In the right graph, black squares show that the difference between the two models is significant (p<0.01).*

4. Blizzard Results

This year Blizzard Challenge evaluates speech produced by TTS on multiple criteria. The Hub-Task evaluates models capacities to reproduce natural behaviors of NEB. The naturalness is evaluated with Mean Opinion Scores (MOS). Intelligibility is evaluated on heterophonic homographs disambiguation and Semantically Unpredictable Sentences (SUS). Similarly, the Spoke-task evaluates the ability of the model to produce natural voice with few examples on AD. Our system did not perform better than the FastSpeech2 baseline in this speaker adaptation task. Thus, this section is focused on the most interesting results of our proposed system: naturalness and intelligibility on the Hub-task. Results commented in this section have been computed regardless of listeners experience in the domain. Among all presented systems, A is the original recording, BT is the baseline Tacotron2, BF is the baseline FastSpeech2, and N is our proposed model. Our model N is highlighted in orange in all figures.

4.1. Naturalness on the most seen speaker

The results of the naturalness assessment on the Hub-Task are reported in Fig.4. Although not showing impressive results, our model was significantly preferred over the FastSpeech2 baseline. Training our model on orthographic sequences may have helped to produce more accurate phonetic patterns. In comparison, the FastSpeech2 Baseline BF has been trained solely on phonetic inputs. Thus, BF relies on a G2P front-end to convert the orthographic sequences of the test set before synthesis. Depending on the front-end used, it may produce errors, in particular with French liaisons which may be hard to predict.

We also believe that our overall MOS score could have benefited from simple post-treatments to reduce the produced noise. We are aware that our Waveglow vocoder produces background noise which can be detrimental to listeners judgment. However, in an attempt to avoid the use of heuristics, we decided to enter the challenge without post-processing denoising techniques.

4.2. Heterophonic Homographs Disambiguation

Intelligibility assessment on heterophonic homographs is reported in Fig.5. Our model N achieves an average score among all systems. Our model shows global improvements over the BF, which was expected thanks to the addition of mixed representations and the training of the phonetic prediction layer on the auxiliary dictionary and homographs corpus.

More specifically, our model performs very well on homographs that have been seen with enough examples in its homograph corpus. “Fils” (261 examples) pronounced /f i s/: “son (en)” VS /f i l/: plural of “fil (fr)”, “wire (en)” has a intelligibility score or 100% for both variants, whereas systems with over-

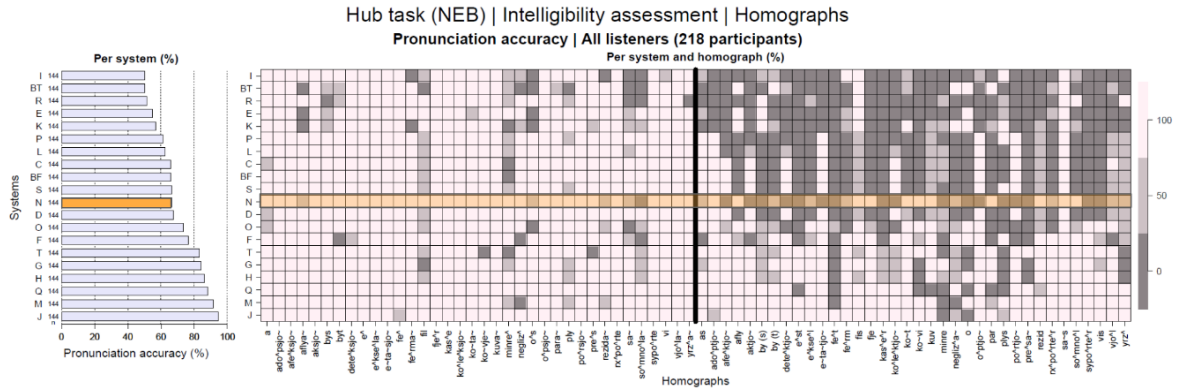


Figure 5: *Homographs intelligibility scores for the Hub Task. Our model N is highlighted in orange. The right graph shows the percentage of correct pronunciation by system. The right graph shows this intelligibility assessment by homograph.*

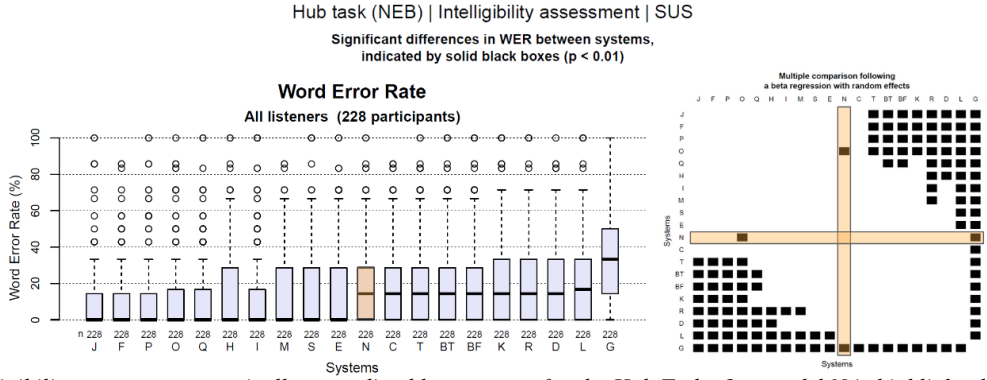


Figure 6: *Intelligibility scores on semantically unpredictable sentences for the Hub Task. Our model N is highlighted in orange. The right graph shows the percentage of correct pronunciations by system. In the right graph, black squares show that the difference between the two models is significant (p<0.01).*

all better scores do not achieve such accuracy on this specific homograph. This is also true for "convient" (181 examples) or "fier" (366 examples) (/k o~v i e~/: "suit (en)" VS /k o~v i/: "invite (en)" — /f j eˆr/: "proud (en)" VS /f j e/: "trust (en)"), with the most common forms /k o~v i e~/ and /f j eˆr/ being systemically pronounced by other TTS regardless of the context. On the contrary, "options" (117 examples), "intentions" (141 examples) and "portions" (145 examples) also appear in the homographs training corpus, but with fewer examples. The number of examples and the balance between variants impact the performances of the system. However, the proposed method helps modelling homographs if enough examples are given during training.

4.3. Semantically Unpredictable Sentences (SUS)

Intelligibility scores on SUS are reported in Fig.6 for all systems. All models but one perform similarly on SUS. Our system only statistically differs from G which shows the worst results on this task, and from O which performs better. On the other hand, BF is found to statistically differ from the top 5 performing systems. The mixed representations and phonetic prediction layer may have help to achieve this task.

5. Conclusions and Discussion

This paper has described the GIPSA-Lab system for the Blizzard Challenge 2023. This system is very similar to the original FastSpeech2 architecture, with two major additions: the training on orthographic sequences and the phonetic prediction layer. The phonetic prediction layer was evaluated before the Blizzard Challenge, and showed very promising performances. The results of the proposed system in Blizzard evaluation confirm the benefits of these additions compared to the baseline FastSpeech2 system. Our system performed better than the baseline FastSpeech2 on naturalness and intelligibility on the most seen speaker in the corpus. On the other hand, our system did not show much difference in terms of speaker adaptation.

The results of the disambiguation of heterophonic homographs shows the potential of the proposed training of the text encoder on <orthography|phonetic> pairs without the need for audio recordings. However, disambiguation was only improved for the most seen examples in the homographs training corpus. Wider corpora may help to achieve better results. The training procedure may also impact the final result. The ratio of non-audio inputs in the training batches may vary to include more phonetic training during the learning phase.

The vocoder used also contributed to the mitigated MOS evaluated during quality assessments. We experience mitigated audio quality with Waveglow, which tends to add background noise in our samples. The impact of this noise can be reduced with post-processing denoising, that we did not explore in our Blizzard submission. Other vocoders like Hifi-GAN may also help regarding this issue.

6. Acknowledgements

This research has received funding from the BPI project THERADIA and MIAI@Grenoble-Alpes (ANR-19-P3IA-0003). This work was granted access to HPC/IDRIS under the allocation 2023-AD011011542R2 made by GENCI.

7. References

[1] 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," *arXiv preprint arXiv:1710.07654*, 2017.

[2] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan *et al.*, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *ICASSP*. IEEE, 2018, pp. 4779–4783.

[3] T. Kenter, V. Wan, C.-A. Chan, R. Clark, and J. Vit, "Chive: Varying prosody in speech synthesis with a linguistically driven dynamic hierarchical conditional variational network," in *International Conference on Machine Learning*. PMLR, 2019, pp. 3331–3340.

[4] 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.

[5] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv:1609.03499*, 2016.

[6] R. Prenger, R. Valle, and B. Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," in *ICASSP*. IEEE, 2019, pp. 3617–3621.

[7] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17 022–17 033, 2020.

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017.

[9] M. Lenglet, O. Perrotin, and G. Bailly, "Modélisation de la parole avec tacotron2 : Analyse acoustique et phonétique des plongements de caractère," in *Actes des Journées d'Etudes sur la Parole (JEP)*, Noirmoutiers, France, June 13-17 2022.

[10] M. Vainio, A. Suni, and D. Aalto, "Continuous wavelet transform for analysis of speech prosody," *Tools and Resources for the Analysys of Speech Prosody (TRASP)*, 2013.

[11] M. Morise, H. Kawahara, and H. Katayose, "Fast and reliable f0 estimation method based on the period extraction of vocal fold vibration of singing voice and speech," in *Audio Engineering Society Conference: 35th International Conference: Audio for Games*. Audio Engineering Society, 2009.

[12] K. Kastner, J. F. Santos, Y. Bengio, and A. Courville, "Representation mixing for tts synthesis," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5906–5910.

[13] M.-L. Hajj, M. Lenglet, O. Perrotin, and G. Bailly, "Comparing nlp solutions for the disambiguation of french heterophonic homographs for end-to-end tts systems," in *SPECOM*. Springer, 2022, pp. 265–278.

[14] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldi." in *Interspeech*, vol. 2017, 2017, pp. 498–502.

[15] A. W. Black, K. Lenzo, and V. Pagel, "Issues in building general letter to sound rules," in *The third ESCA/COCOSDA workshop (ETRW) on speech synthesis*, Jenolan Caves House, Blue Mountains, Australia, 1998.

[16] G. Bailly, M. Lenglet, O. Perrotin, and E. Klabbers, "Advocating for text input in multi-speaker text-to-speech systems," in *12th ISCA Speech Synthesis Workshop*. ISCA, 2023, pp. 13–18.

[17] J. Kearns, "Librivox: Free public domain audiobooks," *Reference Reviews*, 2014.

[18] P.-E. Honnet, A. Lazaridis, P. N. Garner, and J. Yamagishi, "The siwis french speech synthesis database. design and recording of a high quality french database for speech synthesis," Idiap, Tech. Rep., 2017.

[19] I. Solak, "The M-AILABS speech dataset," https://www.caito.de/2019/01/the-m-ailabs-speech-dataset/, 2019.