The GIPSA-Lab Text-To-Speech System for the Blizzard Challenge 2025

Gérard Bailly, Olivier Perrotin

The Blizzard Challenge 2025
27 August 2025, Gröningen, The Netherlands

The GIPSA-Lab Text-To-Speech System for the Blizzard Challenge 2025

Gérard Bailly, Olivier Perrotin

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

Abstract

This paper describes the GIPSA-Lab submission to the Blizzard Challenge 2025. The Text-To-Speech system trained for this challenge is an autoregressive encoder-decoder architecture based on Tacotron2. Updates of the Tacotron2 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 but with no phonetic alignments with audio. 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 TTS [1, 2, 3, 4], combined with neural vocoders [5, 6, 7], generate synthetic voices that closely mimic natural speech. However, the training is mostly conducted in favourable environments with large datasets and high-quality recordings. Thus, the good performances shown by neural TTS models may be overestimated compared to real-life applications [8].

The Blizzard Challenge 2025 aims at evaluating latest neural TTS systems in more challenging environments, in particular with minimal unaligned training data recordings from a unique speaker and on Bildts, an under-resourced Dutch language variety. More specifically, the BH1 task of this challenge includes the evaluation of utterances in context. Only readings of 138 short paragraphs were provided together with a small pronunciation dictionary.

Our approach to this challenge is to propose a TTS system very close to the original Tacotron2 model [2] but with the addition of: the training on mixed input, a phonetic prediction subtask and a stateful training of the text encoder. This extended version is named *TC2* in the following.

Tacotron2 has an autoregressive audio decoder. In this paper, we show how the letter-to-sound (L2S) alignment proposed by Lenglet et al. [9] for both French and English can be used 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 and exploit the Bildts pronunciation dictionary provided by the challenge organizers. Results indirectly show that our model benefited from this phonetic predictor and is perceived more Dutch-like than other systems.

This paper is organized as follows: Section 2 describes our proposed model and the L2S mapping used to train our *TC2* 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 4.1. Finally, results of the Blizzard evaluation are discussed in section 4.3.

2. Model: Tacotron2 with mixed inputs

This section describes the Tacotron2 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 Tacotron2 implementations [4]. The encoder, attention mechanism, prenet, decoder and postnet are kept unchanged. Note that the decoder is trained to predict two frames at the same time: the training and inference speed is reduced and convergence is more robust.

Speaker and style control is achieved through the addition of trainable speaker/style embeddings at the output of the text encoder. The model is trained on both orthographic and phonetic input sequences, following the mixed-inputs training procedure [10].

Following [9], an additional phonetic prediction layer is added at the output of the text encoder, composed of a fully-connected layer with softmax. This layer predicts a one-to-one mapping between orthographic inputs and phonetic outputs. This one-to-one L2S mapping is further described in Section 2.2. The goals of this layer are twofold: first, it helps disambiguating homographs if any as shown in [11]. Second, it enables to train the text encoder 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. The same lexicon as the Blizzard organizers was used.

The vocoder used is HiFi-GAN [7]. The original architecture remains unchanged[2]. The technical specificities and performances of our system are summed up in Table 1.

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

The one-to-one L2S mapping is used to generate phonetic sequences from the orthographic sequences of the training data,

[1] https://github.com/gerard-bailly/tacotron2
[2] https://github.com/jik876/hifi-gan

10.21437/Blizzard.2025-5

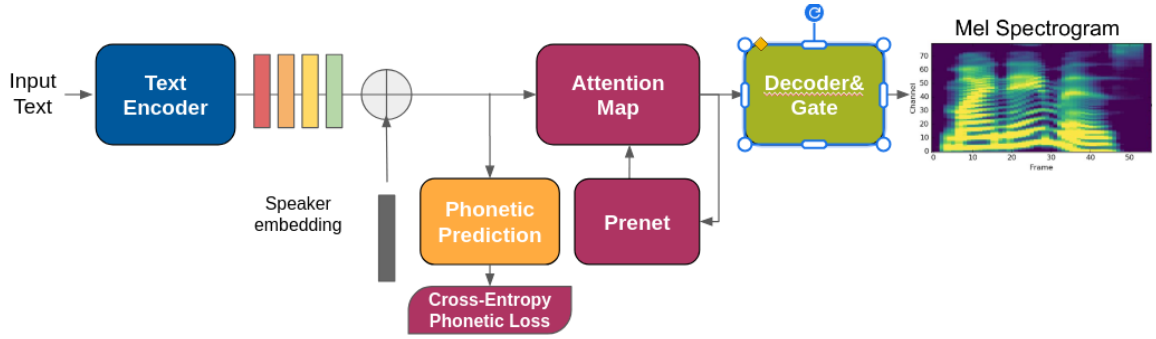

Figure 1: *Model Architecture of the Tacotron2 baseline with the phonetic prediction layer. This phonetic prediction layer is plugged to the output of the text encoder.*

Table 1: *Technical specificities and performances of the proposed TC2 with mixed inputs and vocoder HiFi-GAN. 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)
TC2	28 433 978	108.50	4.5	0.0333
HiFi-GAN	13 926 017	53.12	3.0	0.0081
Total	**42 359 995**	**161.62**	**7.5**	**0.0414**

in order to train the text encoder. This method was proved efficient on the training on French TTS [12], provided with a mapping derived from the exploration of the attention map of a fully trained Tacotron2 TTS model by Lenglet et al. [9].

In this Blizzard Challenge, the one-to-one L2S mapping is performed using the iterative method proposed by Black et al. [13] and further refined by Jiampojamarn et al. [14] seeking for the best alignment between input letters and phones. Letters not associated with phones are called mute and get aligned with a mute output symbol. Multiple phones paired with a unique letter are grouped in so-called di- or tri-phones.

Output phones includes the 47 phones used in the dictionary provided by the Blizzard organizers. 11 additional symbols were added:

1. The symbol /_ / is assigned as output of this one-to-one mapping for muted characters
2. The symbol /__ / is assigned to silences (paired with spaces and punctuations)
3. 9 di-phones for schwa insertions (often associated with vowels and /m/ and /l/ endings) and English diphthongs (e.g. `hype|h A&i p _` or `hyperlink|h A&i p q r l I ng k`)

3. Training and Early Evaluation

3.1. Dataset

Audio. We used all provided text and audio materials as training material for the text and audio decoder. We however hand-corrected most of the provided segmentation: from the readings of 138 short paragraphs split into 7177 parts of speech, we retrieved 6452 utterances separated by a minimum of 350 ms of silence with or without breath noise. The training set is further described in Table 2.

We expanded and gave in phonetic forms all numbers and abbreviations. This procedure was used 77 times in the training corpus. For example:

```
Boereweer|156811|158624|
Skriifster Annie M.G. Schmidt|
.Skriifster Annie {e^ m}{G e} Schmidt,
```

Note that a short text was also entirely phonetized, so that to help the phonetizer to operate on whole sentences:

```
Peerd fan Sundreklaas|146620|147864|
Maar dat mâg niet!|
.Maar dat mâg niet!|
__ m a _ r _ d A t _ m A R _ n i q t __
```

Also, every text associated with input utterances is appended with a punctuation mark: (a) existing introductory dialogue or parenthetical marks «, —, [, (if any; (b) else, start of paragraph (§) if any[3]; (c) else, the last punctuation mark of the preceding utterance if any; (d) then, a comma by default. This procedure complements the stateful training of the text decoder for restoring the prosodic coherence of adjacent utterances. We then systematically add 130 ms of ambient silence at the beginning and end of training utterances to ease the *TC2* attention map aligning these start and end marks with silences of fixed durations. Note that an additional 100 ms of ambient silence is padded at the end of every utterance to improve end-of-sequence prediction over several frames by the output gate.

The audio output of the audio decoder are two frames of 80-bands Mel-spectrogram computed on the 22 050 Hz audio signal with a hop-size of 256 (which is equivalent to a spectrogram sampling rate of ≈ 86 Hz).

Dictionary. We also used the Letter-to-Sound dictionary (7221 out the 7754 words provided by the Blizzard organizers) to complement the training of the phonetic predictor. We discarded entries with stars, that required reconstruction from already transcribed morphemes. It was aligned with the procedure described in section 2.2.

[3]No paragraph breaks were provided in the training material whereas they were explicitly checked in test material

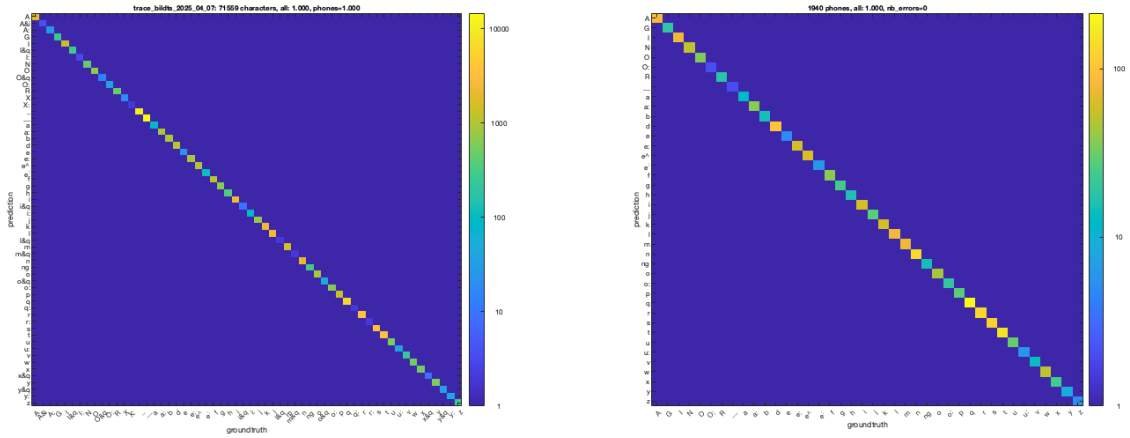

Figure 2: *Confusion Matrices of the phonetic prediction layer, for orthographic inputs (left) and phonetic inputs (right).*

Table 2: *Datasets. Durations are given in hh:mm:ss.ms*

Speaker	Metadata		Audio	
	Dataset	Gender	Duration	# Utt
BH1	Training	Male	5:56:38.272	6452
BH1	Test	Male	-	9732
Dictionary	Isolated words	-	-	7221

3.2. Training Procedure

The non-audio inputs (dictionary) 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.

The model was first trained without the stateful option for 100 epochs, using both orthographic and phonetic transcriptions. The batch size is set to 32. As mentioned above, 22 utterances from one short text (Peerd_fan_Sundreklaas) are presented twice by epoch: once with their orthographic input and once with their phonetic input. Batches are randomly selected among the whole training corpus, resulting in a mixture of input types in each batch. This mixture is not supervised. The learning rate was fixed to 10^{-3}.

Then stateful training for 50 additional epochs was performed for acoustic utterances: batches were organized such as each item in a batch was following the utterance preceding the item in the previous batch. The final activations (h_n,c_n) of hidden/context units of the forward LSTM layer of the previous utterances of the text encoder were copied to the initial activations (h_0,c_0) of hidden/context units of the forward LSTM layer of the current corresponding utterances, except at onsets of paragraphs (§) where they were set to zeros. Since no paragraph information was provided within the training short texts, only 138 onsets were reset, corresponding to the beginning of each short text.

The vocoder HiFi-GAN [7] was fine-tuned from the pretrained model shared with the GitHub implementation. The fine-tuning was performed, first for 50 epochs on the Ground-Truth spectrograms, and then for 50 additional epochs on spectrograms predicted by the final *TC2* model.

4. Evaluation

4.1. Phonetic Prediction Evaluation

The phonetic prediction was computed on the training set, and confusion matrices are reported in Fig. 2, distinguishing between orthographic vs. phonetic inputs. No errors are detected among the 71 559 orthographic characters nor 1940 phones.

4.2. Effect of full-state training

The stateful training positively impacts MAE spectrogram-loss: it almost halved the global loss over the last 50 training periods. Our procedure is however rather naive: the gradient obtained for (h_0,c_0) for the current batch is not back-propagated to the final state (h_n, c_n) of the previous one. This should be revisited with a within-batch approach.

4.3. Blizzard listening test results

Seven teams participated in the Blizzard Challenge, and our system was given the letter A. This year, the Blizzard Challenge evaluated speech produced by TTS on multiple dimensions either using subjective rating of utterances in isolation or in ground-truth context. The evaluation was conducted online. There were three separate audience (listener) groups: Bildts speakers (audience a1), Dutch/Frisian speakers (a2), and international speakers (a3).

For the first set of dimensions, the listener only listen to audio (with no textual content) and answers the following questions:

a3-overall-quality "How do you rate the overall quality of what you have just heard?"

a3-listening-effort "How would you describe the effort your were required to make in order to understand the message?"

a3-voice-pleasantness "How would you describe the voice?"

a3-human-likeness "How human-like does the previous sample sound?"

For a second set of dimensions, the text is displayed, the listener listen to audio and answers the following questions:

a1-bildts-likeness "How close to Bildts is the synthesized sentence?"

a1-naturalness "Choose a score for how appropriate the sentence sounded considering the provided context (pay attention to the sentence intonation, word stress, pronunciation,

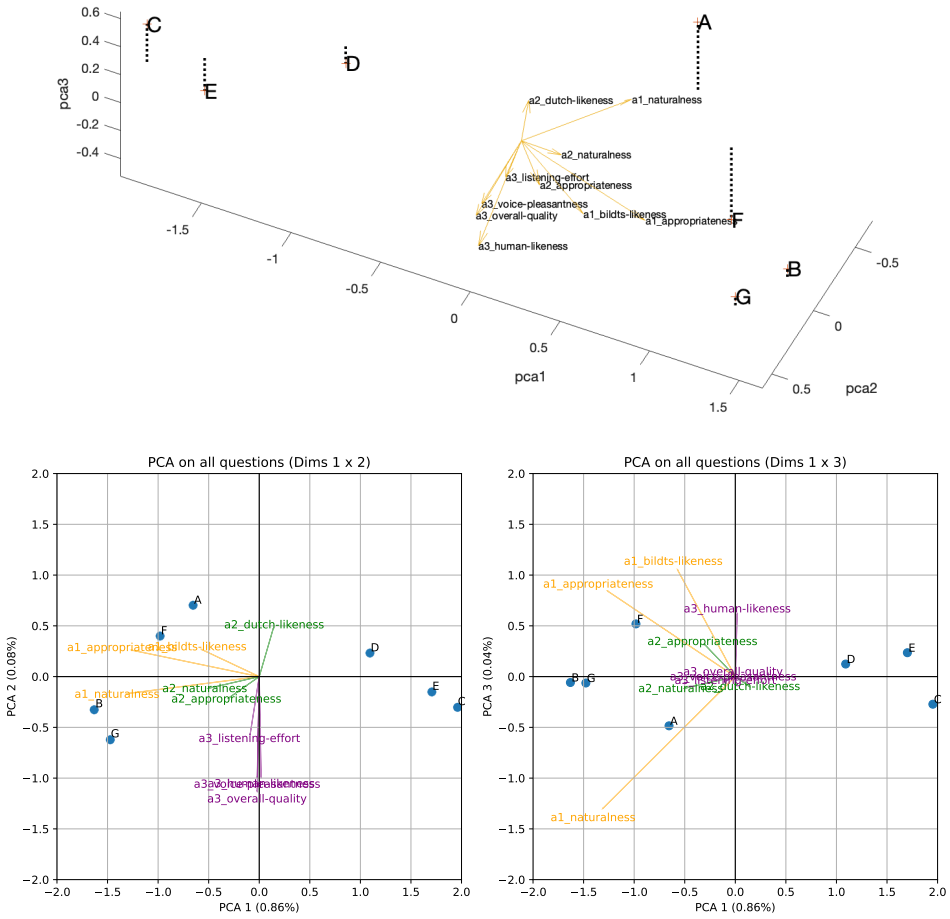

Figure 3: *Projection of the 7 systems participating in BH1 on the first 3 principal planes obtained by principal component analysis of the 10 mean MOS rated dimensions. Our system A clearly lies inbetween two groups (C, D, E) vs. (B, F, G). Original MOS directions are superimposed.*

. . . and not the content)"

a2-dutch-likeness "How close to (standard) Dutch is the synthesized sentence?"

a2-naturalness "Now choose a score for how natural or unnatural the sentence sounded."

For the third set of dimensions, the text is displayed together with the sentence(s) preceding and following that sentence. The listener listens to audio (only the synthesized central part but aware of textual context if foreseen by the model) and answers the following questions:

a1-appropriateness "Choose a score for how appropriate the sentence sounded considering the provided context (pay attention to the sentence intonation, word stress, pronunciation, . . . and not the content)"

a2-appropriateness same as previous

Beyond the individual Mean Opinion Scores (MOS) reported by the organizers, we performed a principal component analysis between the 10 averaged MOS per question obtained by the 7 systems. The 3 first principal components (PC) explain more than 98 % of the variance, (86, 8 and 4 %, respectively). The projection of the 7 systems and the 10 MOS directions are displayed in Figure 3.

Questions: We first note a high correlation between the questions asked to participants. The first PC is collinear with all questions asked to a1 and a2 audiences, except for Dutch likeliness, which is orthogonal. The second PC is collinear with all questions asked to a3. Thus, this second PC is more related to the evaluation of the signal, i.e., without understanding of what is being said, compared to the first PC which is more global. The third PC distinguishes between questions addressed to a1, i.e., Bildts speaker, although it explains little variance in the answers.

Groups of systems: The first PC, which explains most of the MOS variance evidences two groups of systems: (C, D, E) vs. (A, B, G, F). (A, B, G, F) are given the best score on all questions included in the first PC (all a1 and a2 except for Dutch likeliness), with a subgroup (B, G) getting higher

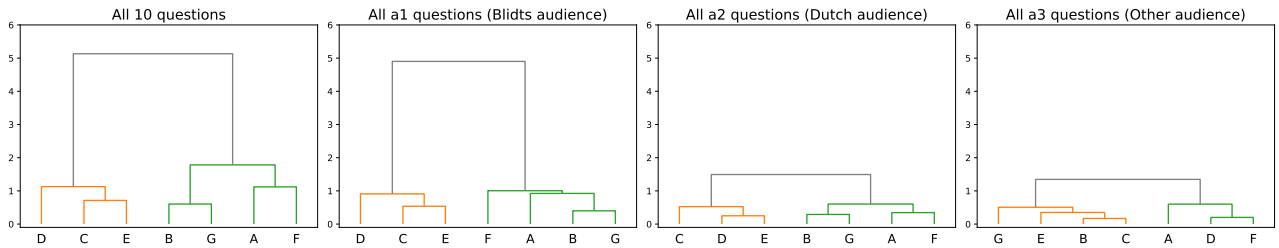

Figure 4: *Dendrograms using mean squared distances (Ward method) between MOS performances of systems according to sets of questions (top) vs. audience (bottom). It clearly shows that: (1) two groups of systems are identified; (b) judgments of Bildts native speakers are sharper than native or non native Dutch listeners.*

scores than (A, F). This is confirmed by the hierarchical clusterings displayed in the second and third panels Fig. 4, when considering only a1 and only a2, respectively. The two groups along with the (B, G) subgroup are well visible. The second PC, which is related to a3 audience shows two different groups, with higher scores for (B, C, D, E, G) than (A, D, F). This is confirmed by the right panel of Fig. 4. The left panel of Fig. 4 displays a global clustering when considering all questions. We find again the two groups (C, D, E) vs. (A, B, G, F) observed along the first PC, as the latter explains most of the MOS variance. Finally, we note that judgments of Bildts native speakers (a1) are sharper than native (a2) or non native Dutch listeners (a3), i.e., by putting higher distance between systems.

Our system: Our system A is in the first group (second subgroup) along the first PC, i.e., when considering Bildts and Dutch audience. Inversely, it was rated in the last group along the second PC, i.e., when rated by listeners that are not familiar with Bildts (a3).

As we can hypothesize that the a3 audience mostly focused on signal aspects, this suggests that our model did not provided the highest speech quality among participants. One explanation could be an insufficient fine-tuning of the HiFi-GAN vocoder, which produced identifiable artefacts, especially on consonants.

If the quality of our output signal was rated badly by a3, then we can make the hypothesis that the good scores given by a1 and a2 (on the first PC) are language-related. On the one hand, our phonetic prediction improves the pronunciation of phones, on the other hand the stateful training targeted coherence of prosody between utterances. While those listening test results do not particularly disentangled the segmental and supra-segmental aspects of the output speech, they suggest a successful contribution of one or two of our additions.

Finally, the last orthogonal dimension is Dutch likeliness, on which we got the highest score. While we do not know whether this is a good or bad result (wrong speech accent?), this is surprising as our model was only trained on the provided Bildts dataset and did not see any Dutch data.

5. Conclusions

This paper has described the GIPSA-Lab system for the Blizzard Challenge 2025. This system is very similar to the original Tacotron2 architecture, with three major additions: the training on mixed input, the phonetic prediction layer and stateful training of the text encoder. The phonetic prediction layer was evaluated before the Blizzard Challenge, and showed very promising performances. The results of the proposed system in Blizzard evaluation shows that our three additions allowed our system—even a bit outdated—to still competes with more state-of-the-art solutions, when targeting a native audience able to evaluate language-related aspects of the synthetic speech.

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. J. Skerry-Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu, "Natural TTS synthesis by conditioning wavenet on Mel-spectrogram predictions," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Calgary, AB, Canada, April 15-20 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*, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. Long Beach, California, USA: PMLR, June 9-15 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 (ICLR)*, Virtual, May 3-7 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 *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Brighton, UK, May 12-17 2019, pp. 3617–3621.

[7] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Advances in Neural Information Processing Systems*, vol. 33, Vancouver, Canada, December 6-12 2020, pp. 17 022–17 033.

[8] G. Bailly, E. André, E. Cooper, E. Klabbers, B. Cowan, J. Edlund, N. Harte, S. King, S. Le Maguer, R. K. Moore, B. Möbius, S. Möller, A. Pandey, O. Perrotin, F. Seebauer, S. Strömbergsson, D. R. Traum, C. Tånnander, P. Wagner, J. Yamagishi, and Y. Yasuda, "Hot topics in speech synthesis evaluation," in *ISCA Speech Synthesis Workshop*, Leeuwarden, The Netherlands, August 24-26 2025.

[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, pp. 788–796.

[10] 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.

[11] 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 *Speech and Computer*, S. R. M. Prasanna, A. Karpov, K. Samudravijaya, and S. S. Agrawal, Eds. Springer International Publishing, 2022, pp. 265–278.

[12] G. Bailly, M. Lenglet, O. Perrotin, and E. Klabbers, "Advocating for text input in multi-speaker text-to-speech systems," in *ISCA Speech Synthesis Workshop*, Grenoble, France, August 26-28 2023, pp. 1–7.

[13] 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.

[14] S. Jiampojamarn, G. Kondrak, and T. Sherif, "Applying many-to-many alignments and hidden markov models to letter-to-phoneme conversion," in *Human Language Technologies*. Rochester, New York: ACL, 2007, pp. 372–379.