The Idiap Speech Synthesis System for the Blizzard Challenge 2023

Haolin Chen[1,2]*, Mutian He*[1,2]*, Louise Coppieters de Gibson*[1,2]*, Philip N. Garner*[1]

[1]Idiap Research Institute, Martigny, Switzerland
[2]École Polytechnique Fédérale de Lausanne, Switzerland
{hchen, mhe, lcoppieters, pgarner}@idiap.ch

Abstract

This paper presents the text-to-speech (TTS) system submitted by Idiap Research Institute to the Blizzard Challenge 2023. Our system follows the conventional pipeline of text analysis, acoustic modeling (AM) and vocoding. For text analysis, open-source pretrained part-of-speech (POS) taggers and lemmatizers are utilized to provide more accurate grapheme-to-phoneme (G2P) conversion on top of the eSpeak backend. The rest of the system incorporates a fully diffusion-based approach which comprises a diffusion transformer-based acoustic model and FastDiff as the vocoder, both of which are trained only on the provided data to ensure high-quality synthesis. Our entry provides a baseline for the cascading diffusion AM-vocoder architecture since no extra design is adopted to enhance the naturalness of speech. Evaluation results have demonstrated high synthesis quality of our system and the effectiveness of the proposed phonemization pipeline.

Index Terms: speech synthesis, Blizzard Challenge, French TTS, diffusion transformer

1. Introduction

The hub task of the Blizzard Challenge 2023 is to build a voice from the provided French data, which consists of around 51 hours of audiobook recordings read by a female French speaker. The spoke task focuses on speaker adaptation and aims to build a voice from around 2 hours of audiobook recordings read by another female French speaker. The Idiap system was submitted to both the hub task and the spoke task.

The top priority of the text-to-speech (TTS) task is to generate high-quality, natural, and intelligible speech. Since neural networks were first introduced to TTS [1, 2], the quality of the synthesized speech has been improved dramatically over the intervening years. In recent years, deep generative model (DGM) based TTS systems [3, 4, 5] have demonstrated their superiority in high-quality and fast synthesis over previous sequence-to-sequence modeling counterparts [6, 7, 8]. In particular, the more recent diffusion-based acoustic models [9, 10, 5] and vocoders [11, 12, 13] have dominated in terms of quality and naturalness. Since 2023, emerging large-scale pretrained language models [14, 15] and DGMs [16, 17] have revolutionized speech synthesis research in generating human-level natural speech and adapting to the target speaker, speaking style or language with very few data. However, these models are neither open to the research community nor can be trained on normal hardware.

Given the provided data are of sufficient quality and quantity, the challenges mainly lie in how to process liaisons and heterophonic homographs in the language which takes place during the text analysis. In French, the liaison refers to the act of pronouncing a linking consonant between two words in a suitable phonetic and syntactic context, which usually gives information about the grammatical structure of a noun phrase. The relatively rare heterophonic homographs refer to words that are spelled the same but pronounced differently, and almost always occur between words of different grammatical categories. These special properties require extra efforts to deliver accurate grapheme-to-phoneme (G2P) conversion in a neural TTS system that uses phoneme input. Available open-source non-neural French phonemizers include the Montreal Forced Aligner [18], Gruut [1] and eSpeak (also eSpeak-ng) [2]. Among them, the first two only perform G2P on word level and handle neither liaison nor homographs. While the eSpeak is a rule-based phonemizer and handles liaison in many cases, it is unable to distinguish heterophonic homographs at the grammatical level since it does not consider part-of-speech. There are also open-source neural G2P models [19] available for the French language, however these models are normally trained on open-source lexicons that do not usually include liaisons and homographs; this limits their performance in real-life scenarios. For systems that support character input [6, 8, 3, 4], the problem can be solved to some extent by the neural network itself given the corpus covers a wide range of the special cases. However, the use of characters as textual input will largely induce higher computation cost and decelerate training and inference due to longer input length compared to using phonemes.

From the practical point of view, the limited computational resources available to us and the short time frame of the challenge are pertinent. Here at Idiap, the servers are mostly equipped with consumer GPU cards and are not optimized for multi-GPU training, leaving us a limited selection of model architectures. In addition, despite Idiap's being situated in a French speaking region, no dedicated toolboxes or dictionaries have been developed for French TTS in recent years. This requires us to utilize publicly available resources as much as possible to cope with the aforementioned particularities of the French language.

Based on the analyses above, we aim to build a TTS system that 1) employs accessible model architectures that offer high-quality and natural synthesis, 2) properly handles the special properties of the French language, and 3) can be trained efficiently on our infrastructure which allows fast verification and iteration. Specifically, for text analysis, we leverage publicly available part-of-speech (POS) taggers and lemmatizers to achieve more accurate G2P conversion on top of the eSpeak backend. For neural architectures, our system adopts a conventional cascading architecture consisting of a diffusion transformer-based acoustic model and FastD-

[1]https://github.com/rhasspy/gruut
[2]https://github.com/espeak-ng/espeak-ng

iff [13], a diffusion-based vocoder. The acoustic model employs a standard non-autoregressive encoder-decoder design that purely relies on the generative modeling power of the diffusion, which makes our system a baseline of the diffusion-based AM-vocoder architecture. Evaluation results have shown a high quality synthesis achieved by our system and the effectiveness of the text analysis pipeline.

2. Text Analysis

2.1. Liaisons

The liaison in the French language refers to the phonetic linking or connection between words in spoken language. It involves the pronunciation of a consonant sound at the end of a word when the following word begins with a vowel sound. Liaison is a characteristic feature of French pronunciation and helps maintain the smooth flow of speech. In most cases, it is limited to word sequences that have a logical connection in meaning, such as an article followed by a noun, an adjective followed by a noun, a personal pronoun followed by a verb, and similar patterns.

The presence of specific liaison patterns in French makes rule-based phonemization a highly suitable technique, which is exactly the one built into eSpeak. Other types of phonemizers also exist, such as the lexicon-based Gruut. In a lexicon-based phonemizer, words are either looked up in a pre-existing lexicon or their pronunciations are predicted using a pretrained G2P model. However, the word-by-word nature of lexicon-based phonemization necessitates additional rules to handle liaisons between words, which are often unavailable in such systems. Recent advancements in G2P solutions, such as sequence-to-sequence neural networks utilized in [20, 19], directly predict phonemes from the input text. Nevertheless, the effectiveness of these models heavily relies on the coverage of the training text corpus, limiting their practicality due to the scarcity of high-quality datasets.

2.2. Heterophonic homographs

In general, heterophonic homographs in French are words that are spelled the same but pronounced differently and have different meanings. Fortunately, their existence is relatively rare, and the phenomenon almost always occurs between words of different grammatical categories, which makes it possible to disambiguate by inferring from the grammatical context.

The first step is to understand in what grammatical categories the common homographs exist. Among publicly available resources online, Wiktionary [3] provides a comprehensive list of 813 heterophonic homographs that exist in the French language. In one blog [4] and [21], the most common scenarios are summarized and the corresponding examples are given. In summary, these scenarios include 1) indicative imperfect first person plural of a verb vs. plural of a noun that end with "-tions", 2) indicative present third person plural of a verb vs. adjective or noun that end with "-ent", 3) infinitive of a first group verb vs. nouns that end with "-er", and 4) miscellaneous cases.

Intuitively, for most cases where words in a pair fall in different grammatical categories, the disambiguation can be done by identifying the part-of-speech of words. For other cases where the two words belong to the same category, such as "convient" and "present", this can be solved by inferring the original form of the word from the context, i.e., lemmatization, to determine their pronunciations.

2.3. Method

Having known the above particularities in the French language, we construct the text analysis module as follows. First, the text input is phonemized by the eSpeak G2P backend. Since eSpeak is able to process liaisons, we only need to refine its corresponding output of homographs considering the grammatical context. To achieve this, we first create a look-up table where different pronunciations of each homographs and the corresponding part-of-speech categories or original forms can be queried, mainly referring to the last two sources mentioned above. During inference, if any homograph in the look-up table exists in the text, we utilize publicly available pretrained POS taggers [5] and lemmatizers [6] to recognize the part-of-speech or the original form of the homograph. Using the inferred information, we refer to the look-up table to obtain the actual phonemes of each homographs. Finally, we compare the phonemes generated by eSpeak with the queried phonemes and rectify the incorrect output.

3. Neural Architectures

To balance synthesis quality and training efficiency, we employ a cascading diffusion-based architecture consisting of a diffusion transformer acoustic model and the FastDiff vocoder.

3.1. Acoustic model

The acoustic model [22] comprises 1) the transformer-based text encoder that encodes phoneme embeddings into hidden representations, 2) the variance adapter that predicts the pitch, energy, and duration of each phoneme and expands the hidden representations to the length of the mel-spectrogram, and 3) the diffusion transformer decoder which generates the mel-spectrogram through a diffusion process. The diffusion transformer is an faster alternative to the most commonly used non-causal WaveNet that offers equivalent synthesis quality.

The architecture of the acoustic model is rather standard: there are no extra components or designs that particularly enhance the naturalness or the speaking style, thus it purely relies on the generative modeling power of the diffusion to render natural speech. We take the chance to see how the standard diffusion architecture performs compared to other more advanced competitors, especially when trained on a highly expressive corpus.

3.2. Vocoder

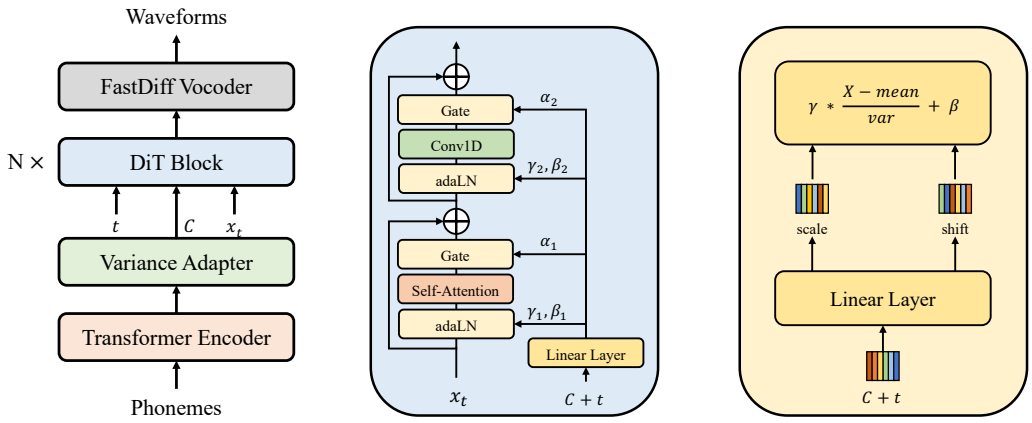
FastDiff is a conditional diffusion-based vocoder for high-quality waveform synthesis. The denoiser network employs a stack of time-aware location-variable convolutions with diverse receptive field patterns to model long-term time dependencies. Originally, a noise predictor was further adopted to derive tighter schedules to accelerate inference without distinct quality degradation. However, we found this algorithm is difficult to implement and the derived schedule must be optimized for every dataset, which makes it less favorable for the adapta-

[3] https://fr.wiktionary.org/wiki/Cat%C3%A9gorie:Homographes_non_homophones_en_fran%C3%A7ais

[4] https://a3nm.net/blog/french_non_homophonous_homographs.html

[5] https://huggingface.co/qanastek/pos-french-camembert-flair

[6] https://github.com/explosion/spacy-models/releases/tag/fr_dep_news_trf-3.5.0



(a) *Overview* (b) *DiT block* (c) *Adaptive layer norm (adaLN)*

Figure 1: *The architecture of the TTS system. The acoustic model and the vocoder are trained separately.*

tion task. Therefore, we use the linear schedule instead of the fast schedule. We also found that FastDiff can be trained more efficiently compared to its GAN-based counterparts, which usually require days of training and multiple GPUs.

4. Experiments

4.1. Data

For the hub task, the NEB corpus consists of 289 chapters of 5 audiobooks from Librivox read by a female French speaker Nadine Eckert-Boulet (NEB), totaling 51 hours and 12 minutes. Around two thirds of the utterances are annotated with texts, phonemes and phoneme durations, while the other one third have texts only. We found the phoneme annotations provided in the dataset lack the tonal and stress marks that are offered by eSpeak, and are likely to be generated by speech recognition models since minor errors can be found. Given the phonemes are unavailable during inference as part of the challenge, and the provided data are insufficient to train a dedicated G2P model, we decide to use eSpeak's phoneme set and run the phoneme-audio alignment using Montreal Forced Aligner [18] to obtain the phoneme durations. Two sets of 500 utterances are selected as the validation and test set, while the rest are used as training set. All data are preprocessed following the practice in FastSpeech 2 [7], with a sampling rate of 22,050 Hz.

For the spoke task, the AD corpus consists of 2515 utterances read by another female French speaker Aurélie Derbier (AD), totaling 2 hours and 3 minutes. We randomly select 50 utterances for the validation set and test set respectively, while the rest specifications follow the hub task.

4.2. Implementation details

The model configurations of the acoustic model follow [22], including a 4-layer transformer encoder with 256 hidden size, a variance adapter same as the one in [7], and a 4-layer diffusion transformer decoder with 256 hidden size and 2 heads. For the vocoder, we use the official implementation [7] without modification. The number of parameters of the acoustic model is around 29M, while the vocoder has around 13M parameters.

[7] https://github.com/Rongjiehuang/FastDiff

4.3. Training and inference

All experiments are conducted on a single NVIDIA RTX 3090 GPU. For the hub task, the acoustic model is trained using a batch size of 40,000 speech frames for 200k iterations, with the "rsqrt" (reciprocal of the square root) scheduler, 4,000 warm-up steps, and a learning rate factor of 2. For the diffusion process, a beta schedule of 16 steps is used for both training and inference. The vocoder is trained using a batch size of 25,600 samples for 1M iterations, with a constant learning rate of 2×10^{-4}. We use a diffusion schedule of 1000 steps for training and a faster schedule of 200 steps for inference. Both of the acoustic model and the vocoder are trained from scratch, which takes around 1 day and 2 days, respectively. The real-time factor of the entire system is 0.48, in which the acoustic model counts for 0.01 while the vocoder makes up the majority of inference time.

For the spoke task, we finetune the entire acoustic model and vocoder used for the hub task to adapt to the AD voice. Specifically, the acoustic model is finetuned for 20k steps with a learning rate of 2×10^{-4}, while the vocoder is finetuned for 10k steps with a learning rate of 1×10^{-4}.

5. Results and Analyses

Our system is identified as *T*, whereas *A* represents natural speech, and *BF* and *BT* are two reference systems.

5.1. Hub task

5.1.1. Quality

Our system is ranked the 7th among 18 participants with a mean MOS score of 3.8. Three systems achieved significantly higher synthesis quality compared to ours, while four together with our system yielded comparable results. In the detailed results broken down by the qualification of testers, we found that non-native listeners and non-speech experts tended to give higher scores compared to native listeners and speech experts. The results suggest that despite our system offering high signal quality, it might be at a disadvantage in terms of naturalness. This can be attributed to the lack of more advanced prosody modeling techniques in the acoustic model, since only the conventional variance adapter was used.

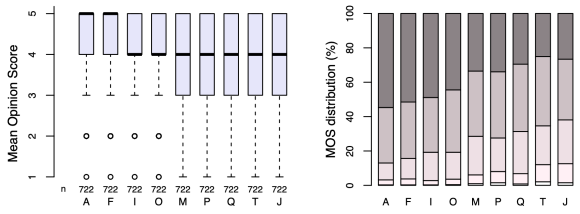


Figure 2: *MOS results of quality, hub task.*

5.1.2. Similarity

For the similarity test, the ranking is 9/18 with a mean MOS score of 3.0. Similar patterns can be found in the results breakdown as in the quality test. We also notice that the speaking style of the generated speech can sometimes be distinct from the reference, which can be attributed to the generative modeling nature of the diffusion decoder and the highly variable voice in the audio book. Additional style modeling methods should be introduced to alleviate the issue.

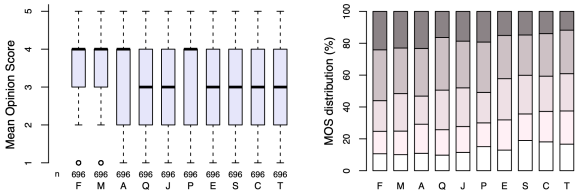
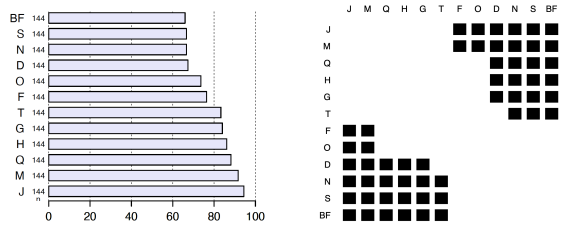


Figure 3: *MOS results of similarity, hub task.*

5.1.3. Intelligibility

In the heterophonic homograph intelligibility test, our system, ranked 6/18, achieves an accuracy of 83%, which is 17% higher than the reference system *BF* that relies solely on the eSpeak backend. The results demonstrate the effectiveness of our proposed text analysis pipeline. Since our method mainly depends on the POS taggers and lemmatizers to correct the incorrect output of eSpeak, we would expect using more accurate models can further improve the phonemization accuracy.



(a) *Pronunciation accuracy* (b) *Significant difference*

Figure 4: *Intelligibility of heterophonic homographs, hub task.*

However, in the conventional intelligibility test, the word error rate of our system is surprisingly high at 19.4%. One possible explanation for this phenomenon is that the lack of speaking style and prosody modeling techniques in the acoustic model results in the fast speaking rate commonly existing in the audio book corpus, which hampers the understanding of such semantically unpredictable sentences. It could also have been caused by the inaccurate alignment between phonemes and speech frames generated by MFA, in which case using a more advanced forced alignment tool would help mitigate the issue.

5.2. Spoke task

In the spoke task of speaker adaptation, our system, ranked in the middle, receives a quality MOS of 3.9 and a similarity MOS of 3.6. Around four systems achieved significantly higher scores than our system in both tests. The results are reasonable since we only perform finetuning on the acoustic model and the vocoder without other dedicated adaptation techniques.

6. Summary

In this paper, we described our entries to the Blizzard Challenge 2023. Our system employed a cascading pipeline of text analysis, acoustic modeling, and vocoding. For text analysis, we utilized pretrained POS taggers and lemmatizers to refine the output phonemes of eSpeak in case of heterophonic homographs. The neural architecture of our system adopted a full-diffusion approach to ensure high-quality synthesis, consisting of a diffusion transformer-based acoustic model and the FastDiff vocoder. Our system completely relied on the generative modeling power of diffusion without any extra design for enhancing naturalness, thus presented a baseline of the cascading diffusion AM-vocoder architecture. Finally, we analyzed official evaluation results and discussed potential means of improvements.

7. Acknowledgments

This project received funding under NAST: Neural Architectures for Speech Technology, Swiss National Science Foundation grant 185010.

8. References

[1] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *CoRR*, vol. abs/1609.03499, 2016.

[2] Y. Wang, R. J. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. V. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: Towards end-to-end speech synthesis," in *Interspeech 2017*. ISCA, 2017, pp. 4006–4010.

[3] J. Kim, S. Kim, J. Kong, and S. Yoon, "Glow-tts: A generative flow for text-to-speech via monotonic alignment search," in *Advances in Neural Information Processing Systems 2020*, 2020.

[4] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *38th International Conference on Machine Learning, ICML*, vol. 139, 2021, pp. 5530–5540.

[5] S. Lee, H. Kim, C. Shin, X. Tan, C. Liu, Q. Meng, T. Qin, W. Chen, S. Yoon, and T. Liu, "Priorgrad: Improving conditional denoising diffusion models with data-dependent adaptive prior," in *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*, 2022.

[6] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Ryan, R. A. Saurous, Y. Agiomyrgiannakis, 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 2018, Calgary, AB, Canada, April 15-20, 2018*. IEEE, 2018, pp. 4779–4783.

[7] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T. Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," in *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, 2021.

[8] Y. Zheng, X. Li, F. Xie, and L. Lu, "Improving end-to-end speech synthesis with local recurrent neural network enhanced transformer," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020*, 2020, pp. 6734–6738.

[9] M. Jeong, H. Kim, S. J. Cheon, B. J. Choi, and N. S. Kim, "Diff-tts: A denoising diffusion model for text-to-speech," in *Interspeech 2021*, 2021, pp. 3605–3609.

[10] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, and M. A. Kudinov, "Grad-tts: A diffusion probabilistic model for text-to-speech," in *38th International Conference on Machine Learning, ICML*, vol. 139, 2021, pp. 8599–8608.

[11] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro, "Diffwave: A versatile diffusion model for audio synthesis," in *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, 2021.

[12] M. W. Y. Lam, J. Wang, D. Su, and D. Yu, "BDDM: bilateral denoising diffusion models for fast and high-quality speech synthesis," in *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*, 2022.

[13] R. Huang, M. W. Y. Lam, J. Wang, D. Su, D. Yu, Y. Ren, and Z. Zhao, "Fastdiff: A fast conditional diffusion model for high-quality speech synthesis," in *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, 2022, pp. 4157–4163.

[14] C. Wang, S. Chen, Y. Wu, Z. Zhang, L. Zhou, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li, L. He, S. Zhao, and F. Wei, "Neural codec language models are zero-shot text to speech synthesizers," *CoRR*, vol. abs/2301.02111, 2023.

[15] P. K. R. et al., "Audiopalm: A large language model that can speak and listen," *CoRR*, vol. abs/2306.12925, 2023.

[16] K. Shen, Z. Ju, X. Tan, Y. Liu, Y. Leng, L. He, T. Qin, S. Zhao, and J. Bian, "Naturalspeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers," *CoRR*, vol. abs/2304.09116, 2023.

[17] M. Le, A. Vyas, B. Shi, B. Karrer, L. Sari, R. Moritz, M. Williamson, V. Manohar, Y. Adi, J. Mahadeokar *et al.*, "Voicebox: Text-guided multilingual universal speech generation at scale," *arXiv preprint arXiv:2306.15687*, 2023.

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

[19] J. Zhu, C. Zhang, and D. Jurgens, "Byt5 model for massively multilingual grapheme-to-phoneme conversion," in *Interspeech 2022*. ISCA, 2022, pp. 446–450.

[20] K. Rao, F. Peng, H. Sak, and F. Beaufays, "Grapheme-to-phoneme conversion using long short-term memory recurrent neural networks," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2015, South Brisbane, Queensland, Australia, April 19-24, 2015*. IEEE, 2015, pp. 4225–4229.

[21] M. 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 - 24th International Conference, SPECOM 2022, Gurugram, India, November 14-16, 2022, Proceedings*, vol. 13721. Springer, 2022, pp. 265–278.

[22] H. Chen and P. N. Garner, "Diffusion transformer for adaptive text-to-speech," in *12th ISCA Speech Synthesis Workshop (SSW) 2023*, 2023.