Speech synthesis for Walloon, an under-resourced minority language

Jose Felipe Espinosa Orjuela, Philippe Boula de Mareüil, Marc Evrard

Human Language Science and Technology, Université Paris-Saclay, CNRS, LISN, France

jose.espinosa-orjuela@universite-paris-saclay.fr, mareuil@lisn.fr, evrard@lisn.fr

Abstract

This paper describes a text-to-speech synthesis system for Walloon, a Gallo-Romance language spoken in Belgium and part of France (in the Ardennes department). The system uses recordings of a translation of *The Little Prince*, read entirely by a male speaker (156 minutes) and, for the first chapters, a female speaker (18 minutes). The corpus was segmented into sentences and transcribed into phonemes by a rule-based grapheme-to-phoneme converter. The synthesis system is based on the Variational Inference with Adversarial Learning for End-to-End Text-to-Speech (VITS) architecture, and several models were trained in different conditions: with or without grapheme-to-phoneme conversion, using or not a fine-tuned model pre-trained on a French corpus. A perceptual evaluation campaign was conducted with Walloon speakers. Results suggest that the models resorting to French data are only preferred in the training condition with the 18-minute reduced corpus.

Index Terms: speech synthesis, Walloon, under-resourced languages, endangered language

1. Introduction

Europe is rich in its linguistic diversity, even though this wealth is masked by the dominant role of official languages and threatened by the interruption of intergenerational transmission. Providing minority languages with automatic processing tools may not be enough to revitalise them, but it is now essential to their revalorisation. Among these languages, Walloon has been officially recognised as an endogenous language of Belgium since 1990. It is also recognised as one of the languages of France, as catalogued by the French Ministry of Culture since 1999. It is spoken in the Pointe de Givet, a small territory north of the Ardennes department. It is a *langue d'oïl*, like French. Some digital resources are available (e.g., Walloon Wikipedia, online dictionaries), but no text-to-speech (TTS) synthesis system. The development of such a system represents a significant societal issue. It is challenging primarily due to the limited availability of speech data. Nevertheless, previous work has suggested leveraging data from related languages — such as French in the case of Walloon — as a viable strategy to address this limitation [1, 2].

Speech synthesis, now present on most smartphones and in public transport, has new educational applications for the general public when oral transmission is interrupted. TTS speech synthesis systems have recently been developed for minority languages such as Occitan [3] and Breton [4]. A recent architecture, the Variational Inference with Adversarial Learning for End-to-End Text-to-Speech (VITS) [5], has enabled high-quality speech synthesis across a very wide range of languages (Meta's Voicebox [6]). Systems such as Microsoft's VALL-E [7], or XTTS [8] include under-resourced languages. Nevertheless, Walloon remains notably absent from these developments.

Recent advances in TTS architectures have enabled the development of high-quality TTS for unseen speakers using as little as one minute of speech, an approach commonly referred to as zero-shot TTS (ZS-TTS) [5, 9, 7, 10]. Building multilingual ZS-TTS systems remains more challenging, often requiring complex model architectures [8]. Training a system using two to three hours of recorded speech is feasible with current mainstream models [11, 5, 12]. Two to three hours is approximately the time it takes to read *The Little Prince*, the second most translated book after the Bible. We had a rewritten version read in standardised Walloon by a native male speaker, and the first chapters read by a native female speaker. The corpus read according to a "neutral", supradialectal or transregional pronunciation served as a basis for the development of an orthographic-phonetic conversion system, intended to prove robust to other writing systems, inspired by Feller [13]. A grapheme-to-phoneme (G2P) conversion system has been written in the form of rules that can be parameterised to adapt to different regions. Let us specify that the G2P task, although more straightforward than for a language like English, requires hundreds of rewrite rules. It was inspired by a rule-based G2P converter for French [14, 15, 16], which was also used in the experiments we will describe.

Since the early 1990s, Walloon has undergone a standardisation process known as *rifondou walon* [17, 18], which is now widely disseminated, including through platforms such as Wikipedia. Its orthography is inspired by the more phonetic system proposed by Feller [13], which incorporates French orthographic conventions and retains most morphological markers of number, gender and person. Similar to the unified spellings of Breton and Occitan, *rifondou walon* is designed to be read according to the speaker's regional phonological norms. For instance, in the word *bijhe* 'North wind', the diasystemic digraph <jh> may be realised as [ʒ] in Liège, [χ] in Verviers, and either [h] or [ʃ] in other parts of Wallonia. The system is intended to encompass several possible pronunciations, including diphthongs, affricates and phonemes that do not exist in French [19, 20]: /ɛ̃/ <én> (alongside /ɛ̃/ <in> and /œ̃/ <un>), long vowels like /ɔː/ <å> or consonants like /h/ <h> or /x/ <xh>.

The TTS system we designed, the first one for the Walloon language to our knowledge, is available online[1]. The main corpus, along with additional test recordings, and the

[1] https://github.com/lisn-speech-synthesis/Walloon-Synthesis-VITS

methodology employed are described in more detail in the following section (Section 2). Several models were trained in different conditions: with or without G2P conversion, using or not a fine-tuned model pre-trained on a French corpus. The computational experiments we conducted are reported in Section 3, alongside results from objective metrics. A perceptual evaluation campaign was conducted with Walloon speakers, the results of which are provided in Section 4.

2. Material and method

2.1. Dataset

The dataset comprises audio files in Wave format (recorded in a quiet room, in stereo and sampled at 44.1 kHz), each corresponding to a chapter of *The Little Prince* and additional data, recorded by two native Walloon speakers (one male, one female)[2]. The male speaker read the entire book, while the female speaker only read the first chapters, totalling 156 and 18 minutes of speech, respectively. In addition, translations of Aesop's fable "The North Wind and the Sun" into central, southern, western and eastern varieties of Walloon were used, coming from a speaking atlas of Belgium [21]. A few dozen sentences of this atlas were reread by the two speakers and kept aside as unseen data for a perceptual evaluation, totalling approximately 4 minutes of speech per speaker.

2.2. Model architecture

The TTS synthesis system we have developed is based on the `VITS` model, which uses a conditional variational autoencoder (CVAE) framework augmented with adversarial learning, enabling end-to-end learning from text input to speech output. This model integrates the conditional generative capabilities of CVAEs with the robustness of generative adversarial networks (GANs) [22] to produce natural-sounding speech. Figure 1 presents a simplified training and inference workflow. Given a text, CVAE models the conditional distribution of speech. This process is achieved using a dual-encoder setup: (1) a posterior encoder maps the input speech to a latent space using variational inference; (2) complementing this encoder, a prior encoder uses text input to generate a prior distribution over the latent space, facilitating the generation of speech that aligns with the given text [5]. To increase the expressive power of the model, normalising flows are applied to the distribution. Adversarial training incorporates a discriminator that critiques the output of the generative model, increasing the quality of the generated voices. Also, phoneme duration is predicted stochastically, allowing the synthesised speech to exhibit realistic temporal variations.

During training, the posterior and prior encoders collaborate to map the speech and text inputs into a shared latent space, which is then used to reconstruct the target speech spectrogram. The normalising flows enhance the mapping precision by transforming a simple prior distribution into a complex one, thereby capturing the intricate characteristics of speech. The role of the discriminator in this setup is to ensure that the generated speech not only sounds natural but also closely matches the target spectrograms in terms of rhythm and intonation.

[2]The audio corpora of The Little Prince are available upon request from the authors of this article. They are subject to special jurisdiction until 2032, because the author of the book, Alexandre de Saint-Exupéry, died for France. The audio corpus of the male speaker is also available at `https://wa.wikisource.org/wiki/Li_Ptit_Prince_(Hendschel-Mahin,_2023)`

2.3. Preprocessing and data preparation

Preprocessing and data preparation for the `VITS` model are critical to ensure the quality and consistency of the speech synthesis output. These processes are divided into data formatting, text preprocessing and audio preprocessing.

The audio files were segmented into sentences. The sampling rate was reduced to 22 kHz, and the files were converted to mono to conform to the model requirements. Classical encoding issues, typical of natural language processing, were fixed. For example, input characters, transformed into lower case, have been normalised so that they are consistent with the *rifondou walon* orthography (e.g., <à> replaced by <a>). Moreover, acronyms were expanded, and numbers were converted into letters.

As illustrated in Figure 1, for the experiments using phonemes as input, a G2P converter was employed to map the text to its phonemic representation in the International Phonetic Alphabet (IPA). This phonetic encoding is particularly advantageous for languages like Walloon, whose orthographic conventions exhibit some irregularities. A 300-rule converter was thus written: particular attention was paid to vowel lengthening, word-final consonant devoicing (e.g., the digraph <jh> /ʒ/ devoiced into [ç]), gemination, liaison and assimilation phenomena. However, the `VITS` model is designed to work effectively without explicit phonemisation. It is, therefore, a question of interest whether a phoneme-based approach or a grapheme-based approach performs better.

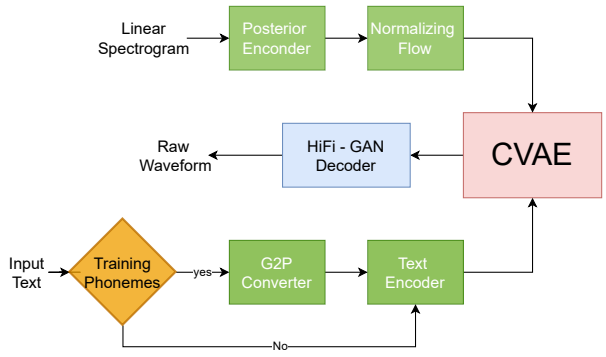
The model also incorporates an alignment estimation strategy that does not rely on external aligners. Instead, it uses an internally developed Monotonic Alignment Search to optimise the alignment between the text and the audio. Audio processing consists of converting the raw signal into mel-spectrograms, which serve as the target for the reconstruction loss during model training.

To stick to a "neutral" Walloon pronunciation, the male speaker, especially, read *The Little Prince* with a slow speech rate, resulting in long pauses. Therefore, a voice activity detector was used [23] (with -60 dB and 50 ms thresholds), and the utterance-internal pause duration was reduced by 50% in the original recordings of this speaker.

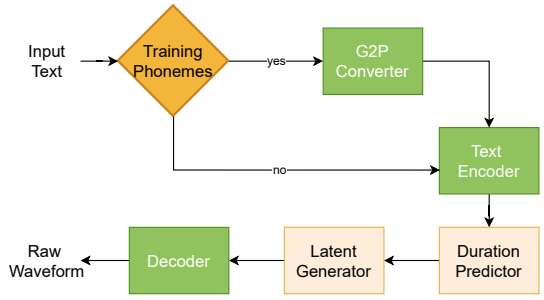
2.4. Training process

The model is trained on a dataset consisting of paired text and audio files. Each audio file is converted into a mel-spectrogram, which serves as the target output for the model. Training the `VITS` model involves three loss functions: (1) the *reconstruction loss* is the primary loss function used to minimise the difference between the generated mel-spectrogram and the target spectrogram from the training data; (2) Kullback–Leibler divergence (D_{KL}), used in the variational auto-encoder component of the model, encourages the encoded latent variables to approximate a prior distribution, which helps regularise the model and avoid overfitting; (3) the *adversarial loss*, employed as part of the GAN framework, involves training a discriminator alongside the generator (the main model) to distinguish generated audio from real audio.

The model uses the Adam optimiser with parameters $\beta_1 = 0.8$ and $\beta_2 = 0.99$, a learning rate of 2×10^{-4} with a decay of 0.999875 was applied throughout the training process. The chosen batch size was 32. Throughout the training



(a) *The block diagram illustrates the steps for training the TTS system.*



(b) *The block diagram illustrates the steps for synthesising speech.*

Figure 1: *The two block diagrams illustrate the steps for training and synthesising speech with the TTS system.*

process, model performance was monitored on a validation set comprising 10% of the data from *The Little Prince*, randomly selected per speaker. An additional 10% was held out as a test set to compute objective evaluation metrics and compare the generated audio against the original recordings across different model configurations. The loss monitored during training corresponds to the total loss, which is the sum of the reconstruction, D_{KL}, duration, adversarial and feature-matching components.

3. Computational experiments

3.1. Experimental setup

Several experiments were conducted, using distinct approaches to input representation and training strategy. In the character-based setting, plain text is used directly as input without the need for G2P conversion: the model was trained for 6,000 epochs. In the phoneme-based setting, G2P conversion was applied before training (an approach that allows for adaptations to different dialects more elegantly than spelling tricks). The model was also trained for 6,000 epochs.

In the fine-tuned configuration, a model pre-trained on French — a language closely related to Walloon — was employed. This pre-trained model was subsequently fine-tuned using our Walloon dataset. A custom model was built by combining two datasets for training: SIWIS [24] and CSS10 [25], with care taken to match voice characteristics. Specifically, the first three sections of the SIWIS dataset were used, segmented for the female voice, while the entire CSS10 dataset was used for the male voice. Graphemes served as the input representation during both training and inference in this configuration. The model was initially trained on French data for 3,000 epochs and subsequently fine-tuned on Walloon data for 2,000 epochs. Similarly, in a fourth configuration, a model was trained using French data converted into phonemes via a rule-based G2P converter [15]. This training also followed a two-stage process: 3,000 epochs on French data, followed by 2,000 epochs on Walloon data.

The models were trained on a GPU NVIDIA Tesla V100 with 32 GiB of RAM, and each training session lasted approximately one day to complete 800 to 1,000 epochs. Across all experiments, model weights were saved every 100 epochs. The quality of the generated speech improved significantly as the training progressed, with notable enhancements observed after approximately 3,000 epochs, including fewer phonetic errors and enhanced fluency in the synthesised voices.

3.2. Objective evaluation

Three objective metrics were used to assess the quality of the model: Mel Cepstral Distortion (MCD) [26], Perceptual Evaluation of Speech Quality (PESQ) [27] and Speaker Encoder Cosine Similarity (SECS) [28]. Moreover, we resorted to an automated mean opinion score (MOS) prediction system, the `UTMOSv2` [29]. The results of this system will be commented on in Section 4, where we will compare it to the perceptual MOS experiment. For the three metrics, dynamic time warping was applied to handle the different sizes of the original and the generated audio. These metrics provided a quantitative assessment of the model's performance, complemented by pause-based measurements. They were computed on the 10% test set (with reduced pauses for the male voice). The values are shown in Figure 2, for the male and female voices, comparing how similar the generated audio is to the original. For the PESQ and SECS metrics, the higher the bars, the better the evaluation, and conversely for the MCD metric. Confidence intervals of 95% were generated using bootstrapping. The extrema of the error bars on Figures 2 and 3 correspond to the 0.025 and 0.975 quantiles on 1,000 bootstrap samples.

Despite a considerable margin of error, particularly for the female voice, PESQ values are consistently higher for the male voice than for the female voice. The phoneme-based approach yields results comparable in quality to the grapheme-based setup. In the French fine-tuned configuration, PESQ scores for the male voice are similar to those of the Walloon-only approach. In contrast, the female voice exhibits a marked improvement, suggesting enhanced perceptual quality. A similar trend is observed with the SECS metric, although this measure is primarily intended to quantify voice differences between distinct speakers. The MCD metric aligns relatively well with the PESQ and SECS results, revealing a notable disparity in reported distortion between the male and female voices. The highest reported distortion was observed in the Walloon-only phoneme-based female model. However, for the female voice, the type of linguistic input — phoneme versus grapheme — appears to exert a greater influence on the results than the inclusion of French data during training. The lowest reported distortion was the French-pretrained phoneme male voice model, which is consistent with the PESQ results.

As stated above, the pauses resulting from focusing on reading in a "generic" Walloon could be problematic. Utterance-internal pauses and speech intervals were measured using the `pyannote` voice activity detector [23] (with the same thresholds as stated above) and analysed using a Python

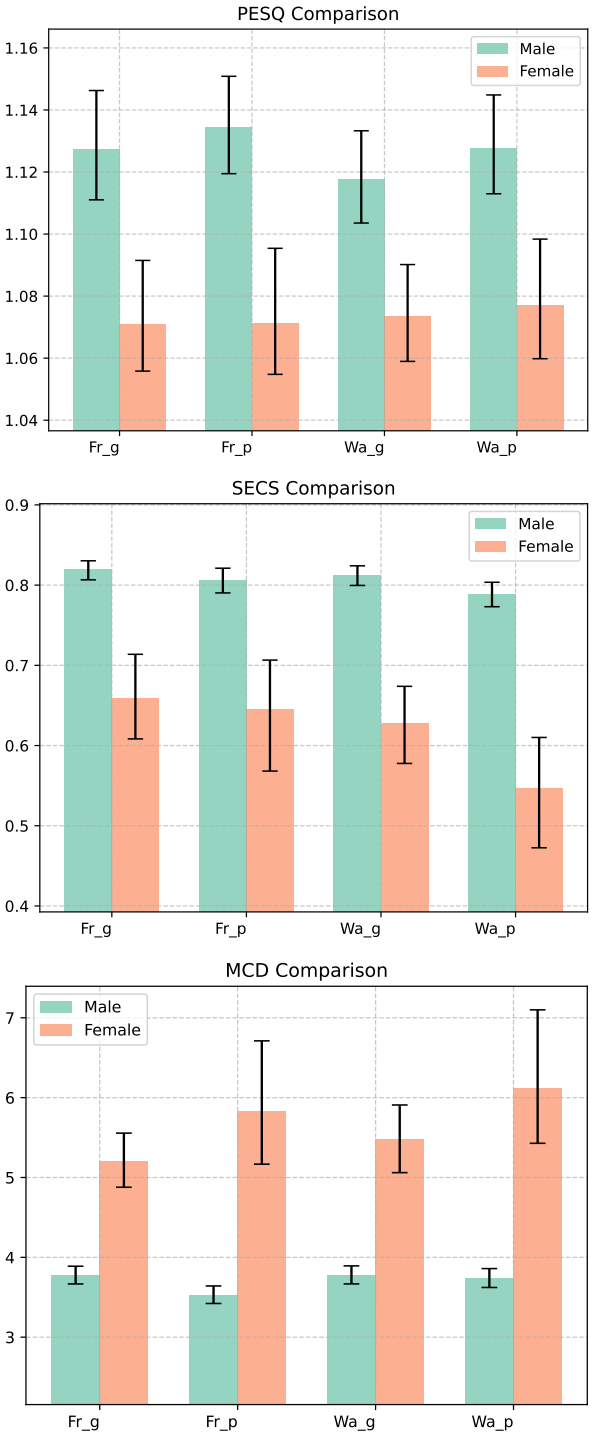


Figure 2: *PESQ, SECS and MCD values for the male voice and the female voice (`Fr` = French pretrained; `Wa` = Walloon only; `g` = grapheme-based; `p` = phoneme-based).*

script[3]. The 50 ms threshold may appear short compared to values reported in other studies [30]. However, this value was selected to account for intervocalic pauses that we perceived and judged to be erroneous. In our synthesised

[3]http://github.com/lisn-speech-synthesis/Walloon-Synthesis-Results

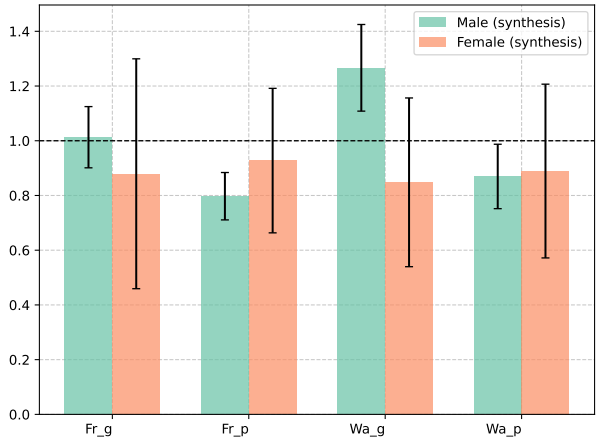


Figure 3: *Ratios between synthesised and original utterance-internal pause counts for the male voice (larger corpus) and the female voice (reduced corpus), (`Fr` = French pretrained; `Wa` = Walloon only; `g` = grapheme-based; `p` = phoneme-based).*

speech, such pauses typically ranged between 70 and 80 ms. Figure 3 presents the results as ratios between the number of pauses in the synthesised utterances and those in the original recordings. The different models appear to perform similarly in maintaining natural pausing close to the original audio, with pause ratios slightly below 1. The exception is the Walloon-only grapheme-based model, which departs from the others: it tends to make more silent pauses than the original with the male synthetic voice. Pauses were analysed as a means of obtaining an objective measure aligned with our perceptual evaluation of pause-related issues. However, a more fine-grained prosodic analysis may be necessary, as the perceptual salience of pauses is not solely determined by their number and duration. Even a few brief pauses, as short as 70 ms, can sound unnatural when occurring in atypical positions within a sentence.

The performance of G2P conversion has not been formally assessed in Walloon, unlike that of French, where the word error rate was estimated at less than 1% [14, 15]. Informal tests lead us to expect a similar percentage in Walloon. In any case, no apparent G2P conversion errors were observed in the test corpus used for the perceptual experiment described in the remainder of this paper.

4. Perceptual test

4.1. Protocol: stimuli, task and participants

A perceptual evaluation campaign was conducted with Walloon speakers. For this purpose, 20 sentences with the male voice and 16 sentences with the female voice were used from translations of the fable “The North Wind and the Sun”. The written sentences were synthesised, and one version per sentence was selected to balance the number of configurations (with or without G2P conversion, with or without French data). Consequently, the experiment included 20 synthetic stimuli with the male voice and 16 synthetic stimuli with the female voice, in addition to the 36 original stimuli.

Given the large combinatorial space, it would have been impractical for participants to listen to 10 versions of each sentence. Thus, 72 stimuli, averaging 10 seconds in length,

Table 1: *Results of the 1–5 MOS perceptual test. The mean values for the synthesised utterances are reported together with 95% confidence intervals, estimated using 1,000 bootstrap samples (smaller value pairs). Legend: `G` = Grapheme; `P` = Phoneme.*

		Using French data		Walloon data only	
Voice	**Original**	**G-based**	**P-based**	**G-based**	**P-based**
Male	4.48 $^{4.56}_{4.40}$	4.10 $^{4.23}_{3.97}$	4.01 $^{4.22}_{3.85}$	4.03 $^{4.14}_{3.93}$	3.90 $^{4.02}_{3.78}$
Female	4.55 $^{4.61}_{4.49}$	4.10 $^{4.13}_{4.07}$	3.96 $^{4.14}_{3.79}$	3.38 $^{3.63}_{3.13}$	2.70 $^{2.79}_{2.60}$

Table 2: *Results of the 1–5 automatic MOS test with `UTMOSv2`. The mean values for the synthesised utterances are reported together with 95% confidence intervals, estimated using 1,000 bootstrap samples (smaller value pairs). Legend: `G` = Grapheme; `P` = Phoneme.*

		Using French data		Walloon data only	
Voice	**Orig.**	**G-based**	**P-based**	**G-based**	**P-based**
Male	3.37 $^{3.51}_{3.24}$	2.96 $^{3.25}_{2.73}$	2.82 $^{3.13}_{2.61}$	2.89 $^{3.04}_{2.74}$	2.87 $^{3.06}_{2.68}$
Female	3.36 $^{3.57}_{3.17}$	2.10 $^{2.15}_{2.04}$	2.43 $^{2.89}_{2.18}$	1.86 $^{2.11}_{1.68}$	2.13 $^{2.34}_{1.92}$

were selected and integrated into the online `PsyToolkit` platform [31]. After providing personal information (e.g., age, gender), participants listened to three pre-test stimuli as part of a familiarisation phase (these were not included in the final analysis). They then completed the main test by listening to the stimuli presented in a randomised order — unique to each participant — and were given the opportunity to leave comments at the end of the session. The actual task consisted of a MOS test, in which subjects had to rate the quality of the sentences they heard on a scale from 1 (very poor) to 5 (very good). The test lasted 15–20 minutes.

The participants (23 in total: 17 male, 5 female and 1 who preferred not to disclose their gender) were 63 years old on average. They were drawn from various regions of Wallonia: 8 from the Centre, 9 from the South, 5 from the East, and 1 from the West.

4.2. Results

The results are reported in Table 1. As for the previous objective metric measurements, 95% confidence intervals are reported along with the mean values for the synthesised utterances. They were estimated using 1,000 bootstrap samples. Hence, the top and bottom values account for the 0.975 and 0.025 quantiles, respectively. Moreover, the results were subjected to an analysis of variance (ANOVA), carried out with the participants' responses as the dependent variable, and stimulus types as fixed factors (10 levels), using the `R` programming language [32]. The stimulus type has a significant effect. $[F(9, 1646) = 62.93; p < 0.001]$: the originals (around 4.5) are rated better than the synthesised stimuli (around 4.0 in most conditions). The synthetic female voice using only Walloon data is judged worse (around 3.4 based on graphemes, 2.7 based on phonemes), with highly significant differences according to a Tukey test $[p << 0.001]$. However, according to this post hoc test, the male synthetic voice does not show significant differences depending on whether G2P conversion and fine-tuning are used.

Compared to the perceptual MOS, the results from the automatic MOS evaluation of `UTMOSv2` presented in Table 2 are significantly lower, with an average score of 2.68, as opposed to 3.92 for the perceptual MOS. The original recording received consistent ratings of 3.37 and 3.36 for the male and female voices, respectively, which are notably lower than the perceptual MOS scores of 4.48 and 4.55. These differences suggest that the system, primarily optimised for the English language [29], struggled to generalise effectively to the Walloon language. While a trend is still observed between the female and male synthetic voices, the gap is smaller compared to the perceptual MOS. Within the male voice category, the scores for all four models fall within the margin of error. Regarding the female voices, some differences can be observed among the synthetic voices, although these do not align with the perceptual MOS. The lowest score was attributed to the grapheme-based Walloon-only model, whereas the phoneme-based version received the lowest score in the MOS evaluation. Interestingly, the `UTMOSv2` system tends to assign lower ratings to the female grapheme-based models compared to the other female models, which contrasts with the trend observed using the MCD metric.

5. Conclusion

This study highlighted critical insights for implementing TTS synthesis systems in low-resourced languages: it explored the potential of models for Walloon using a limited dataset. Different input representations and training strategies were evaluated: grapheme-based or phoneme-based, and whether or not a French pre-trained model was used for fine-tuning. The results of a perceptual test, corroborated by objective metrics (except for the Auto-MOS test, which did not seem to generalise well on our data), showed that the grapheme-based model and the slightly more complex phoneme-based model produced comparable quality. Using French data only

proved valuable in the training condition with the 18-minute reduced corpus; the impact is negligible with a 156-minute corpus. This finding is interesting, as it may be generalised to other lesser-resourced languages, which present similar challenges for speech technologies. Additionally, different data augmentation techniques could be explored to improve performance.

Grapheme-to-phoneme conversion, while not directly improving the results in this study, presents a promising avenue for future research. It offers enhanced control over the phonetic output, which can be tailored to accommodate different dialects. This flexibility aligns with one of the core principles of *rifondou walon* ('normalised Walloon'), where certain digraphs can be pronounced differently depending on the region. Further experiments and listening tests are necessary to determine whether the parameterisation of a few G2P conversion rules can enhance the system's acceptability and better reflect the diversity of Walloon varieties.

6. Acknowledgements

We express our sincere gratitude to the speakers who provided their voices for the recordings, as well as to all the participants who took part in the perceptual experiment.

7. References

[1] C. Soria, J. Mariani, and C. Zoli, "Dwarfs sitting on the giants' shoulders–how lts for regional and minority languages can benefit from piggybacking major languages," in *Proceedings of XVII FEL Conference*, 2013, pp. 73–79.

[2] D. Bernhard, A.-L. Ligozat, M. Bras, F. Martin, M. Vergez-Couret, P. Erhart, J. Sibille, A. Todirascu, P. Boula de Mareüil, and D. Huck, "Collecting and annotating corpora for three under-resourced languages of France: Methodological issues," *Language Documentation & Conservation*, vol. 15, pp. 316–357, 2021. [Online]. Available: https://hal.science/hal-03273196

[3] A. Corral, I. Leturia, A. Séguier, M. Barret, B. Dazéas, P. B. de Mareüil, and N. Quint, "Neural text-to-speech synthesis for an under-resourced language in a diglossic environment: the case of gascon occitan," in *Proceedings of the 1st Joint Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*. Marseille: European Language Resources Association (ELRA), 2020.

[4] D. Guennec, H. Hajipoor, G. Lecorvé, P. Lintanf, D. Lolive, A. Perquin, and G. Vidal, "Breizhcorpus: a large breton language speech corpus and its use for text-to-speech synthesis," in *Odyssey Workshop 2022*. ISCA, 2022, pp. 263–270.

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

[6] 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," *Advances in neural information processing systems*, vol. 36, pp. 14 005–14 034, 2023.

[7] S. Chen, C. Wang, Y. Wu, Z. Zhang, L. Zhou, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li *et al.*, "Neural codec language models are zero-shot text to speech synthesizers," *IEEE Transactions on Audio, Speech and Language Processing*, 2025.

[8] E. Casanova, K. Davis, E. Gölge, G. Göknar, I. Gulea, L. Hart, A. Aljafari, J. Meyer, R. Morais, S. Olayemi, and J. Weber, "Xtts: a massively multilingual zero-shot text-to-speech model," in *Interspeech 2024*, 2024, pp. 4978–4982.

[9] E. Casanova, J. Weber, C. D. Shulby, A. C. Junior, E. Gölge, and M. A. Ponti, "Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone," in *International conference on machine learning*. PMLR, 2022, pp. 2709–2720.

[10] S. Chen, S. Liu, L. Zhou, Y. Liu, X. Tan, J. Li, S. Zhao, Y. Qian, and F. Wei, "Vall-e 2: Neural codec language models are human parity zero-shot text to speech synthesizers," arXiv, June 2024.

[11] 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*, 2020.

[12] E. Strickland, A. Lacheret, M. Evrard, S. Kahane, D. Aubakirova, D. Doncenco, D. Torres, P. Quennehen, and B. Guillaume, "De nouvelles méthodes pour l'exploration de l'interface syntaxe-prosodie: un treebank intonosyntaxique et un système de synthèse pour le pidgin nigérian," in *Actes de la 31ème Conférence sur le Traitement Automatique des Langues Naturelles, volume 1: articles longs et prises de position*, 2024, pp. 376–383.

[13] J. Feller, *Essai d'orthographe Wallone*. Liège: Vaillant-Carmanne, 1900.

[14] P. Boula de Mareüil, "Étude linguistique appliquée à la synthèse de la parole à partir du texte," Ph.D. dissertation, Université Paris-Sud (11), Orsay, France, 1997.

[15] ——, "Conversion grapheme-phoneme: de la formalisation à l'évaluation," in *Ressources et évaluation en ingénierie des langues*, N. M. K. Chibout, J. Mariani and F. Néel, Eds. De Boeck/AUPELF-UREF, 2000, pp. 509–525.

[16] M. Evrard, "Synthèse de parole expressive à partir du texte : Des phonostyles au contrôle gestuel pour la synthèse paramétrique statistique," Ph.D. dissertation, Université Paris-Sud (Paris 11), France, 2015.

[17] L. Hendschel, "Quelle planification linguistique pour le wallon ?" in *Actes du Colloque international de Charleroi*, L. U. C. W. éditeur, Ed., 1996, pp. 3–21.

[18] L. Mahin, "Témoignage," in *Singuliers*, M. de la parole en Ardenne, Ed., vol. 2, 1993, pp. 13–16.

[19] L. Remacle, *La différenciation dialectale en Belgique romane avant 1600*. Geneva: Droz, 1992, vol. 256.

[20] J. Haust, M. Boutier, L. Remacle, M. Counet, E. Legros, J. Lechanteur, and E. Baiwir, *Atlas linguistique de la Wallonie*, F. d. p. e. l. Université de Liège, Ed. Liège: Vaillant-Carmanne, 1953–2011.

[21] P. Boula de Mareüil, L. Mahin, and F. Vernier, "Les parlers romans dans l'atlas sonore des langues et dialectes de belgique," *Bien dire et bien aprandre-Revue de médiévistique*, no. 35, pp. 85–108, 2020.

[22] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[23] H. Bredin, "pyannote. audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe," in *24th Interspeech Conference*. ISCA, 2023, pp. 1983–1987.

[24] J. Yamagishi, P.-E. Honnet, P. Garner, and A. Lazaridis, "The siwis french speech synthesis database? design and recording of a high quality french database for speech synthesis," University of Edinburgh. School of Informatics. The Centre for Speech Technology Research., 2017 [dataset]. [Online]. Available: https://datashare.ed.ac.uk/handle/10283/2353

[25] K. Park and T. Mulc, "Css10: A collection of single speaker speech datasets for 10 languages," in *Proc. Interspeech 2019*, 2019, pp. 1566–1570.

[26] R. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," in *Proceedings of IEEE pacific rim conference on communications computers and signal processing*, vol. 1. IEEE, 1993, pp. 125–128.

[27] 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.

[28] 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.

[29] K. Baba, W. Nakata, Y. Saito, and H. Saruwatari, "The t05 system for the voicemos challenge 2024: Transfer learning from deep image classifier to naturalness mos prediction of high-quality synthetic speech," in *2024 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2024, pp. 818–824.

[30] C. Fauth and J. Trouvain, "Détails phonétiques dans la réalisation des pauses en français: étude de parole lue en langue maternelle vs en langue étrangère," *Langages*, vol. 211, no. 3, pp. 81–95, 2018.

[31] G. Stoet, "Psytoolkit: A novel web-based method for running online questionnaires and reaction-time experiments," *Teaching of Psychology*, vol. 44, no. 1, pp. 24–31, 2017.

[32] R. Core Team, *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing, 2021.