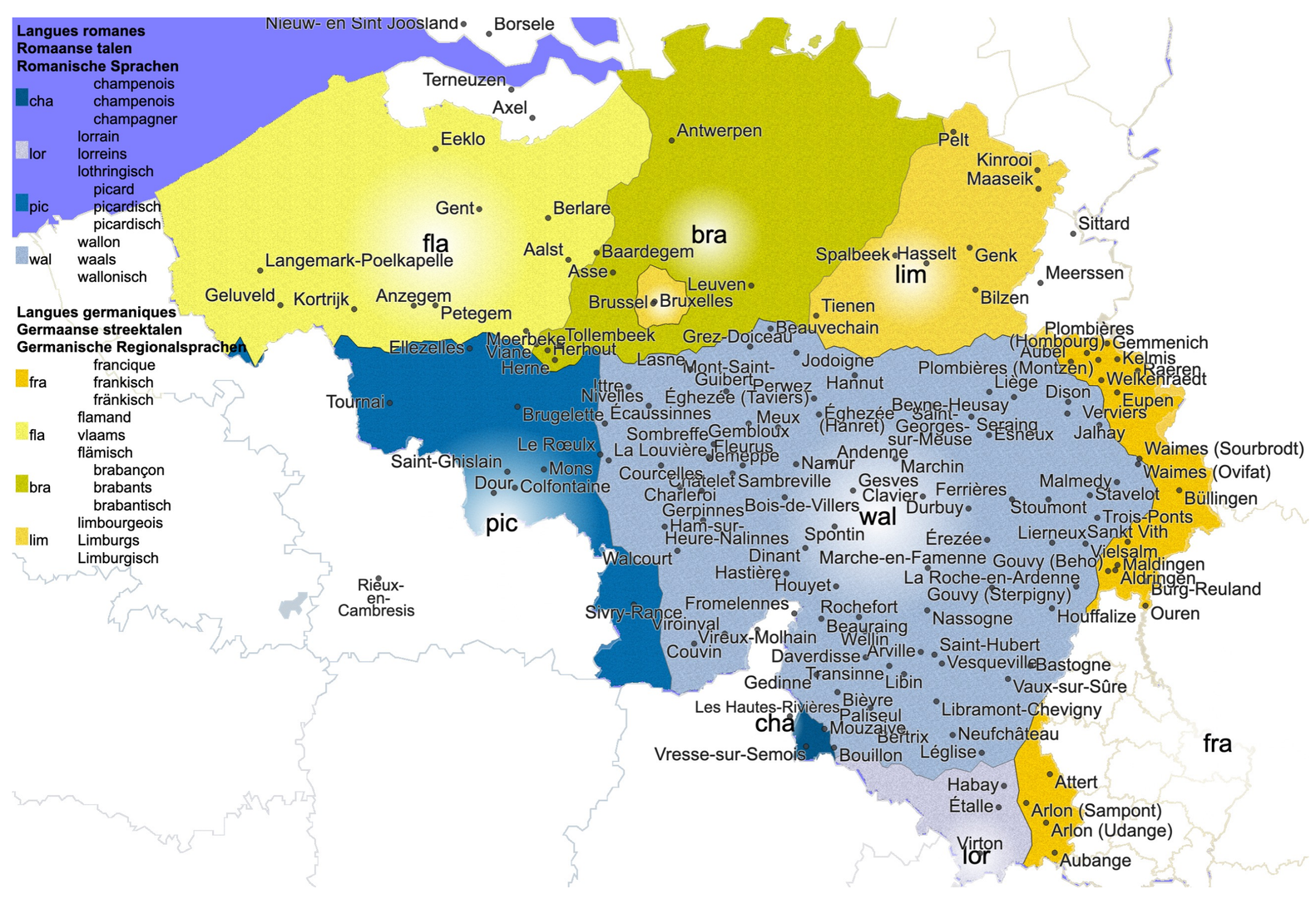
Text-to-speech synthesis in Walloon, an endogenous language of Belgium

Philippe Boula de Mareüil, Felipe Espinosa, Marc Evrard
Université Paris-Saclay & CNRS, LISN, Orsay, France

1. Introduction

- Linguistic diversity of Europe hidden and threatened
  → equip minority languages with automatic processing tools: e.g., Walloon (spoken in parts of Belgium and France)
  ↓
- Northern Gallo-Romance language with digital resources and the possibility to use French data, but no text-to-speech (TTS) synthesis system
  ↓
  didactic applications



Map of Belgium

- State-of-the-art systems developed, e.g., for Breton and Occitan
  → 10 or 20 hours of speech required
  → 2 or 3 hours of recordings using advances in the field of neural networks (a thousand languages in Meta's system)
  ↓
- Reading time of *The Little Prince*,
  translated and recorded, in standardised Walloon* by a native male speaker
  (and the first chapters by a female native speaker)

* *Rifondou walon* = a diasystemic system which can be read according to the speaker's own habits
  - In the word *bijhe* 'North wind', the diasystemic digraph <jh> may be interpreted as [ç] in Liège, [χ] in Verviers, [h] or [ʃ] in the rest of Wallonia.
  - The system is intended to encompass several possible pronunciations, including phonemes that do not exist in French: /ẽ/ <én> (alongside /ɛ̃/ <in> and /œ̃/), long vowels like /ɔː/ <å> or consonants like /h/ <h> or /x/ <xh>.

2. Corpus and Method

- Version of *The Little Prince* recorded using *rifondou walon* spelling, read using a supradialectal, neutral pronunciation,
  - 156 minutes for the male voice
  - 18 minutes for the female voice
  + 4 minutes for testing, from translations of "The North Wind and the Sun" based on a grapheme-to-phoneme (G2P) conversion system **robust to other writing systems** inspired by Feller (1900)

- G2P conversion system developed at LISN
  - written in the form of rules which can be parameterised to adapt to different regions
  - paying particular attention to vowel lengthening, word-final consonant devoicing, gemination, liaison and assimilation phenomena
  - to segment the audio corpus into sentences and phonemes

3. Experiments

- Preprocessing (classical encoding problems in Natural Language Processing)

- VITS architecture (used in Meta's system)
  → components dominated by deep learning
  - Conditional Variational Autoencoder (CVAE)
  - adversarial training
  - stochastic duration prediction

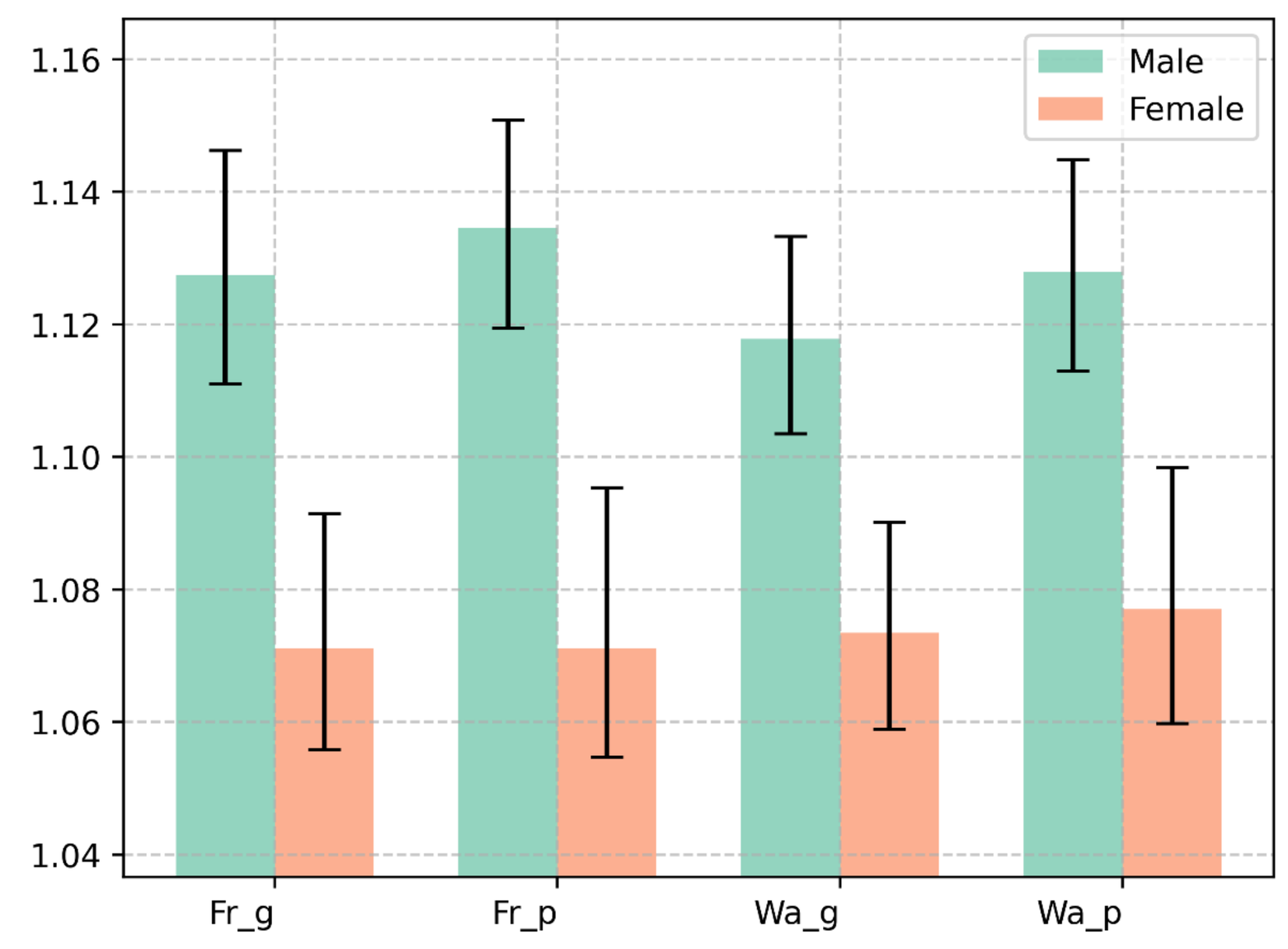
- 4 configurations for each voice (available on Hugging Face)

Male	Female
Ufsing or not G2P conversion Using or not French data	Using or not French data Using or not G2P conversion

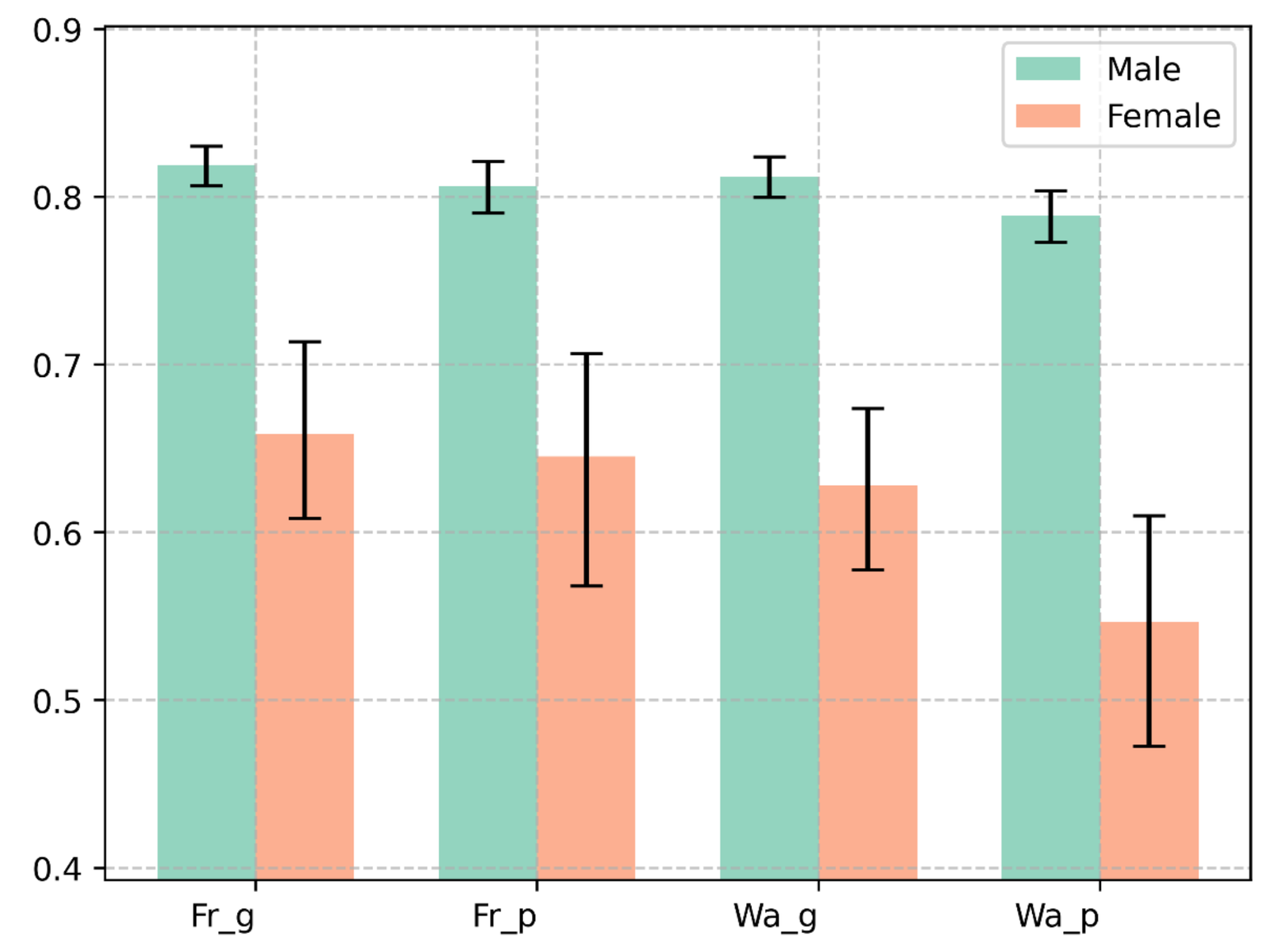
4. Results

- Objective evaluation

Perceptual Evaluation of Speech Quality (PESQ)

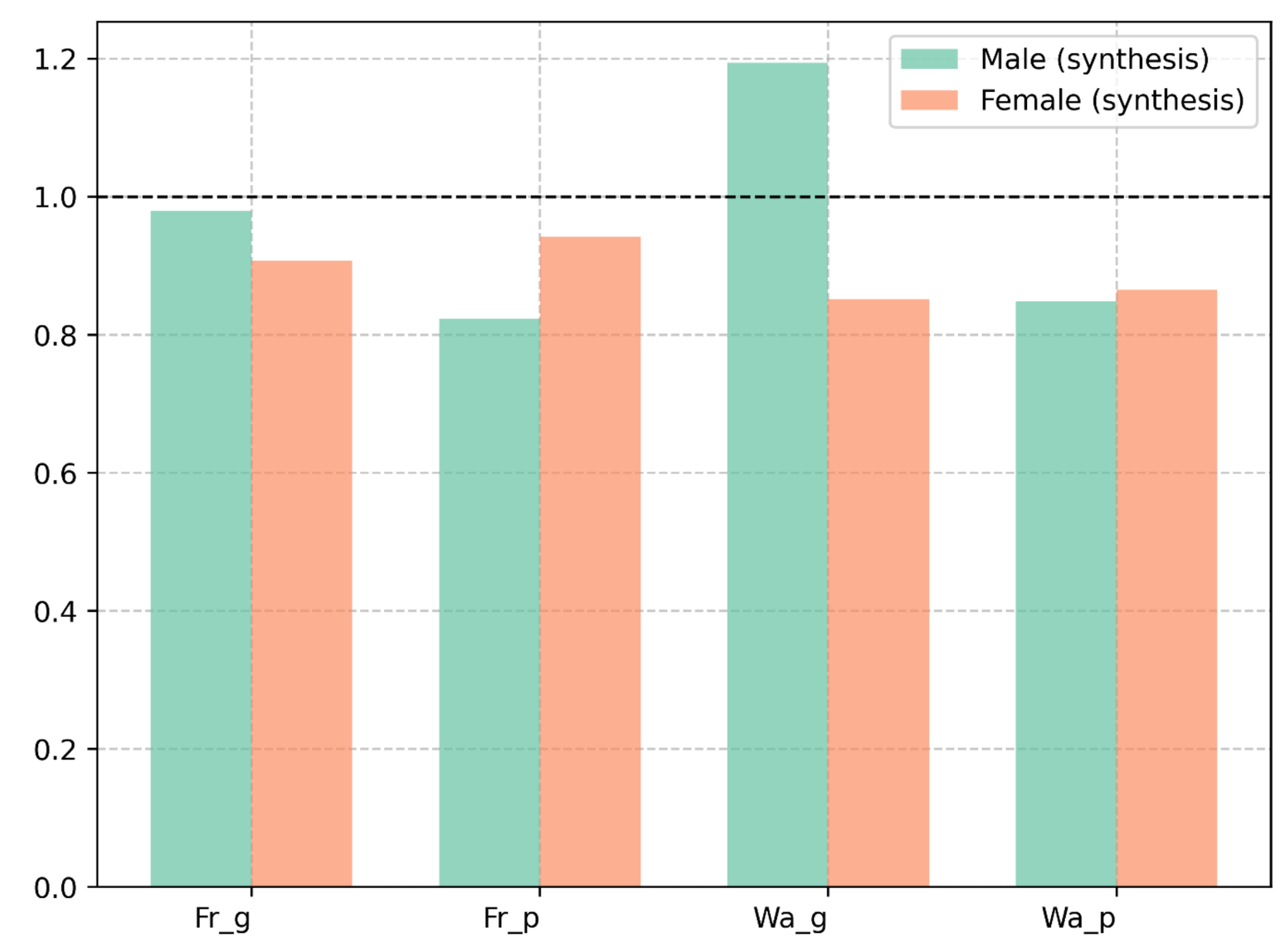


Speaker Encoder Cosine Similarity (SECS)



  - values consistently higher (= better) for the male voice than for the female voice, despite a sizeable margin of error
  - results of the phoneme-based approach similar to that of the grapheme-based setup
  - better scores, for the female voice, with the French fine-tuned approach, compared to the Walloon-only approach

- Pause-related measures



  - Walloon-only grapheme-based model producing more pauses than the original audio, for the male voice
  - other results difficult to interpret

- Perceptual test: Mean Opinion Score (MOS)
  - rating from 1 (very poor) to 5 (very good)
  - 23 Walloon listeners (63 years old on average) on unseen data
  - 72 sentences (from translations of "The North Wind and the Sun")

		Using French data		Using only Walloon data	
Voice	Original	G-based	P-based	G-based	P-based
Male	4.48	4.10	4.01	4.03	3.90
Female	4.52	4.10	3.96	3.38	2.70

  - synthesised stimuli rated around 4.0 in most conditions (around 4.5 for the originals)
  - synthetic female voice using only Walloon data judged worse (around 3.0) → highly significant differences

5. Conclusion and Future Work

- ≠ input representations and training strategies, using a limited dataset
  → good quality produced by grapheme-based and phoneme-based models

- Models using French data only preferred in the training condition with the (18-minute) reduced corpus
  → possible generalisation to other lesser-resourced languages

→ Disclosure to Walloon networks
→ Adaptation to different regional varieties

6. Acknowledgements

- Thanks to the two speakers and all the listeners