Vowel Tuner

Soklong HIM Nora LINDVALL Maxime MÉLOUX
Jorge VASQUEZ-MERCADO

NLP M2

Software Project
Jan. 23, 2023

Outline

Plan

So far...

The corpus has some issues...

- Badly annotated words:
  - sœur $\rightarrow$ /syʁ/ instead of /sœʁ/)
  - ti1 $\rightarrow$ tiOne $\rightarrow$ /tjɔn/
  - tant2 $\rightarrow$ tantTwo $\rightarrow$ /tɑ̃nto/
- Over half of the corpus (and speakers) had no annotations
- The existing annotations were fully automatic (Astali)

Dataset

Re-annotated the entire corpus using forced alignment (Astali)

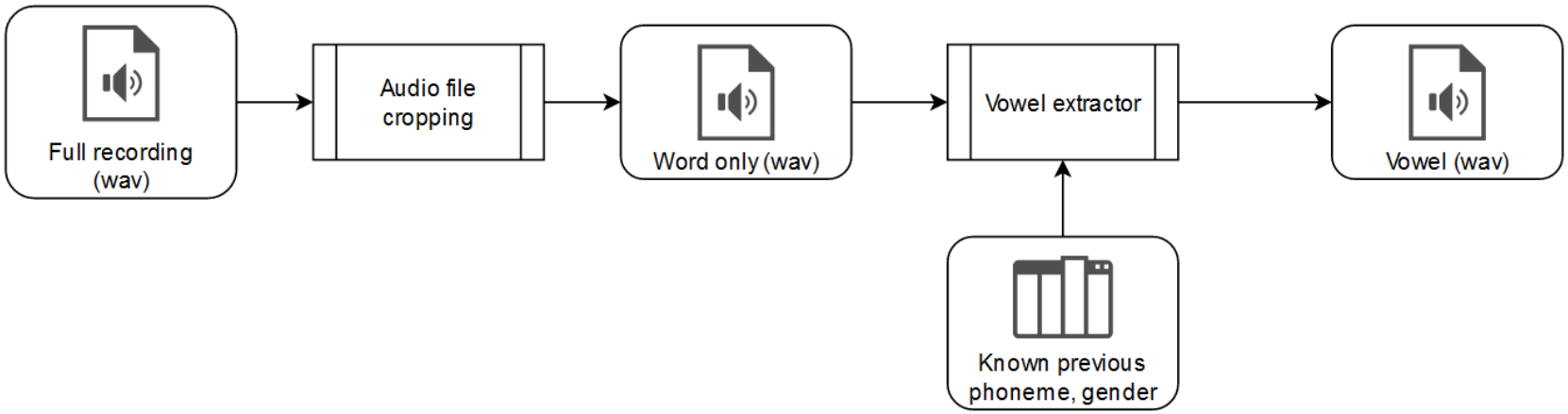
1755 vowels $\rightarrow$ **5775** vowels 😱

Plan

To summarize

It's better if the user has to pronounce an entire word than just a vowel.
But: Our classifier performs a lot better if fed with a single vowel.

→ How to go from a recording to a single vowel?



Solutions

- Full recording $\rightarrow$ word: leading/trailing silence detection (using a volume threshold)
- Word $\rightarrow$ vowel: Regression model
  - Intuition: If a neural network can recognize a vowel from a spectrum (not easy), it can identify consonant-vowel transitions (very visible)

Solutions

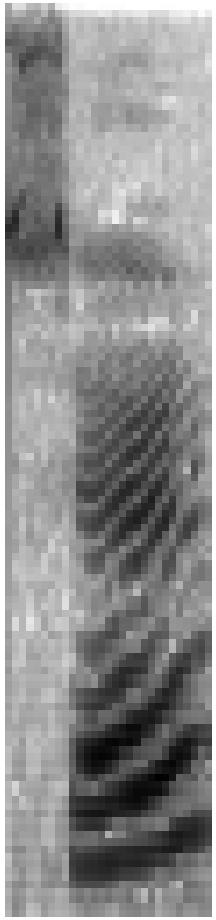


Figure: Melspectrogram of "sœur" (/sœʁ/). Can you see the transitions?

Plan

Linguistic model

- Nothing new here, just more data.
- Input: Formants F1-F4, speaker gender, previous phoneme

Classifier	Jan 13	Jan 23	Delta
*Bagging	73.96%	78.86%	+4.90%
Decision trees	60.42%	74.25%	**+13.83%**
***Extra trees**	**79.79%**	**82.93%**	+3.14%
K neighbors	67.71%	79.40%	+11.69%
Logistic regression	66.67%	77.51%	+10.84%
Multilayer perceptron	75.00%	81.30%	+6.30%
*Random forests	75.00%	82.93%	+7.93%
*Stacking	71.88%	80.22%	+8.34%

Table: Test set accuracy of various classifiers. Stars denote ensemble methods.

Linguistic model

Results:

- Large improvements across the board
- Explainable models reaching 80% accuracy
- Chosen model: extra trees again (highest accuracy, 82.93%)
- But large (400 estimators, 200 MB)

Neural models

Regression model: Given a sound file, predict the vowel boundaries

- Output: Two values between 0 (beginning of file) and 1 (end of file)

Classification model: Given a sound file, predict the vowel

- Output: One of the 10 vowels

In practice, we take the melspectrogram as the input.

- Now using 128 mel bands instead of 1
- Values retrieved properly
- Images are now resized instead of padded

Neural models

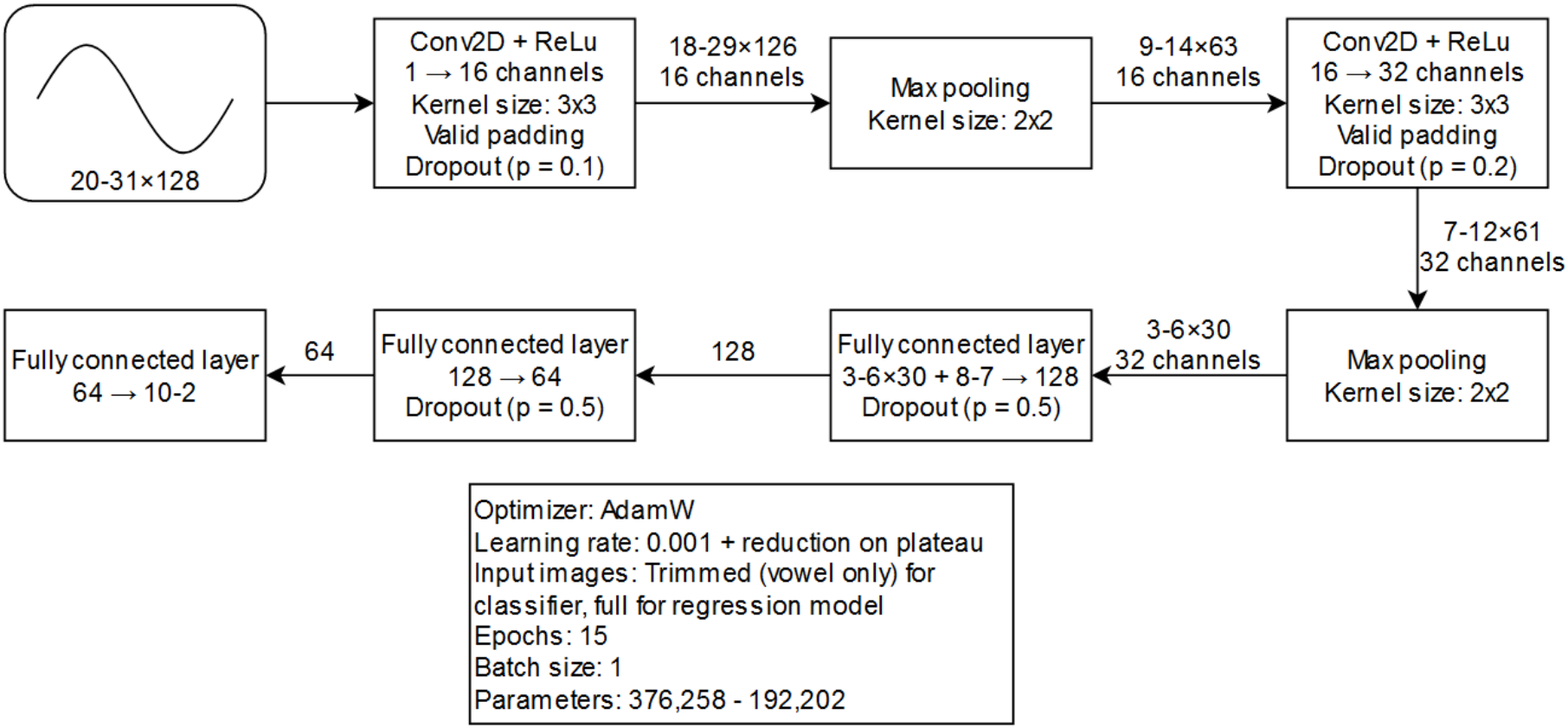


Figure: Final best architecture for both the classifier and regression model

Neural models

Regression model:

- Total mean square error of 1.1376 over the test set
- Qualitatively and quantitatively good

Classifier:

- Final test set accuracy: **94.59%**

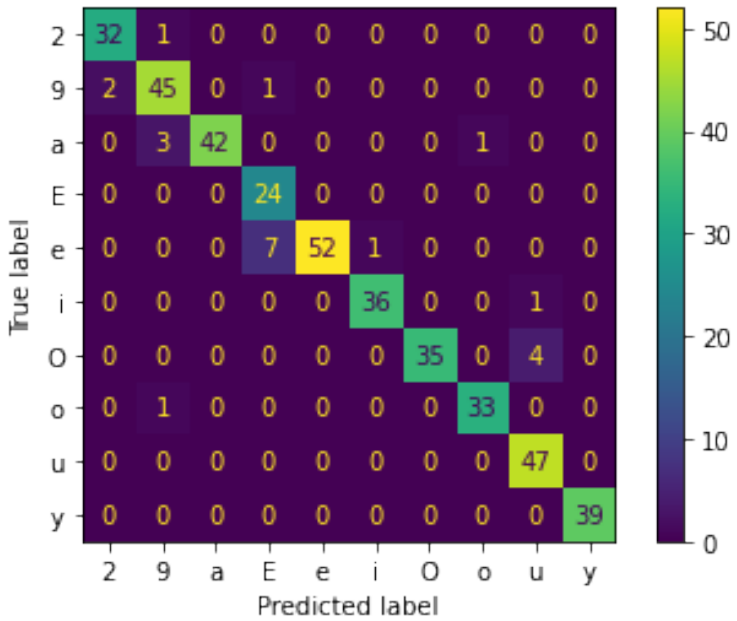


Figure: Confusion matrix for the classifier

Discussion

In theory:

- 94.59% accuracy for the neural model
- 82.93% accuracy for the extra trees model
- Good (hard to quantify) performance for the vowel extractor

But: Real world users (usually) do not have access to the LORIA recording room.

Discussion

What's the performance on real data?

Plan

Presentation

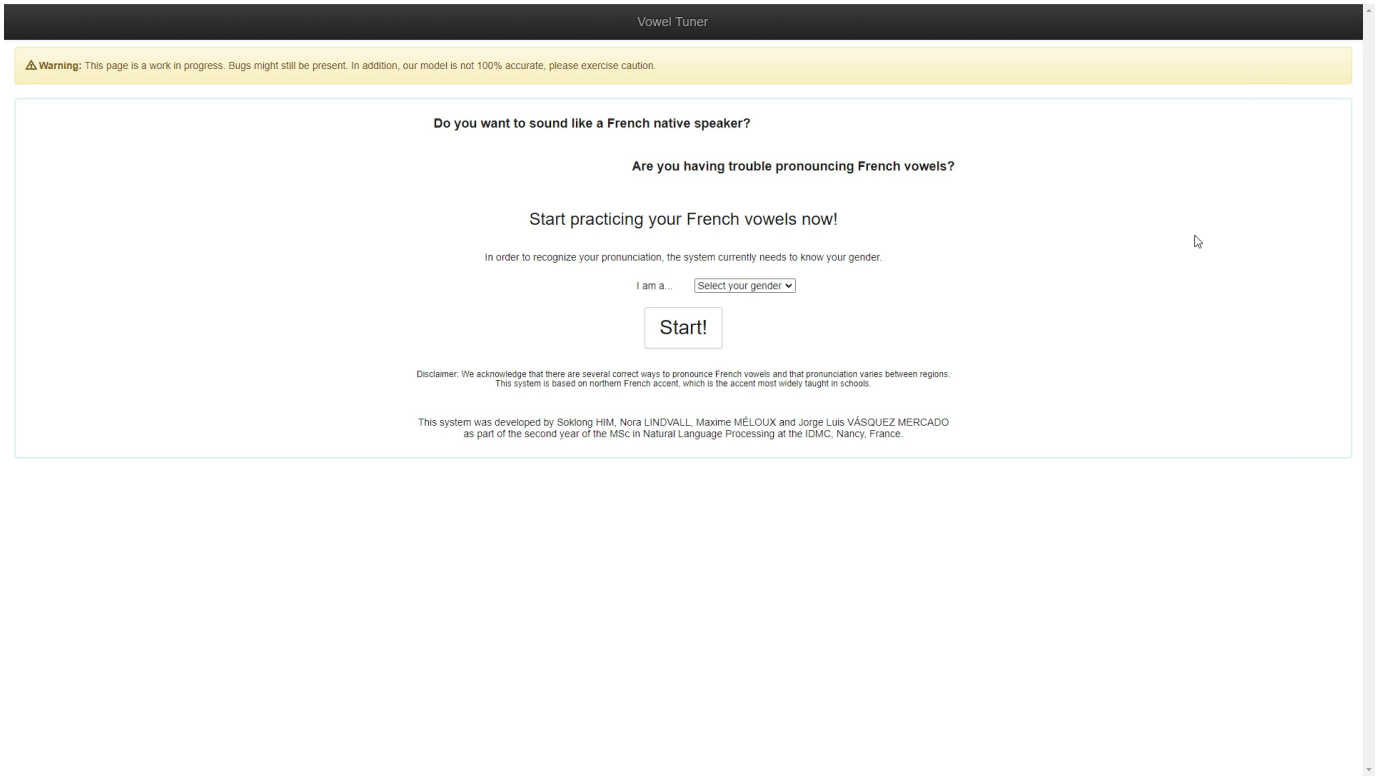
Frontend: HTML5/JavaScript using Bootstrap 3, jQuery and Recorderjs

- The user inputs their gender
- They then record themselves saying a word and click the prediction button
- RGPD-compliant website (see privacy policy)

Backend: Flask server (Python)

- Crops the sound file to a single vowel
- Feeds the vowel to the linguistic or neural classifier
- Generates user-friendly but linguistic-based custom pronunciation feedback

Demo


Vowel Tuner
⚠ Warning: This page is a work in progress. Bugs might still be present. In addition, our model is not 100% accurate, please exercise caution.
Do you want to sound like a French native speaker?
Are you having trouble pronouncing French vowels?
Start practicing your French vowels now!
In order to recognize your pronunciation, the system currently needs to know your gender.
I am a...
Select your gender
Start!
Disclaimer: We acknowledge that there are several correct ways to pronounce French vowels and that pronunciation varies between regions. This system is based on northern French accent, which is the accent most widely taught in schools.
This system was developed by Soklong HIM, Nora LINDVALL, Maxime MÉLOUX and Jorge Luis VÁSQUEZ MERCADO as part of the second year of the MSc in Natural Language Processing at the IDMC, Nancy, France.

Human evaluation

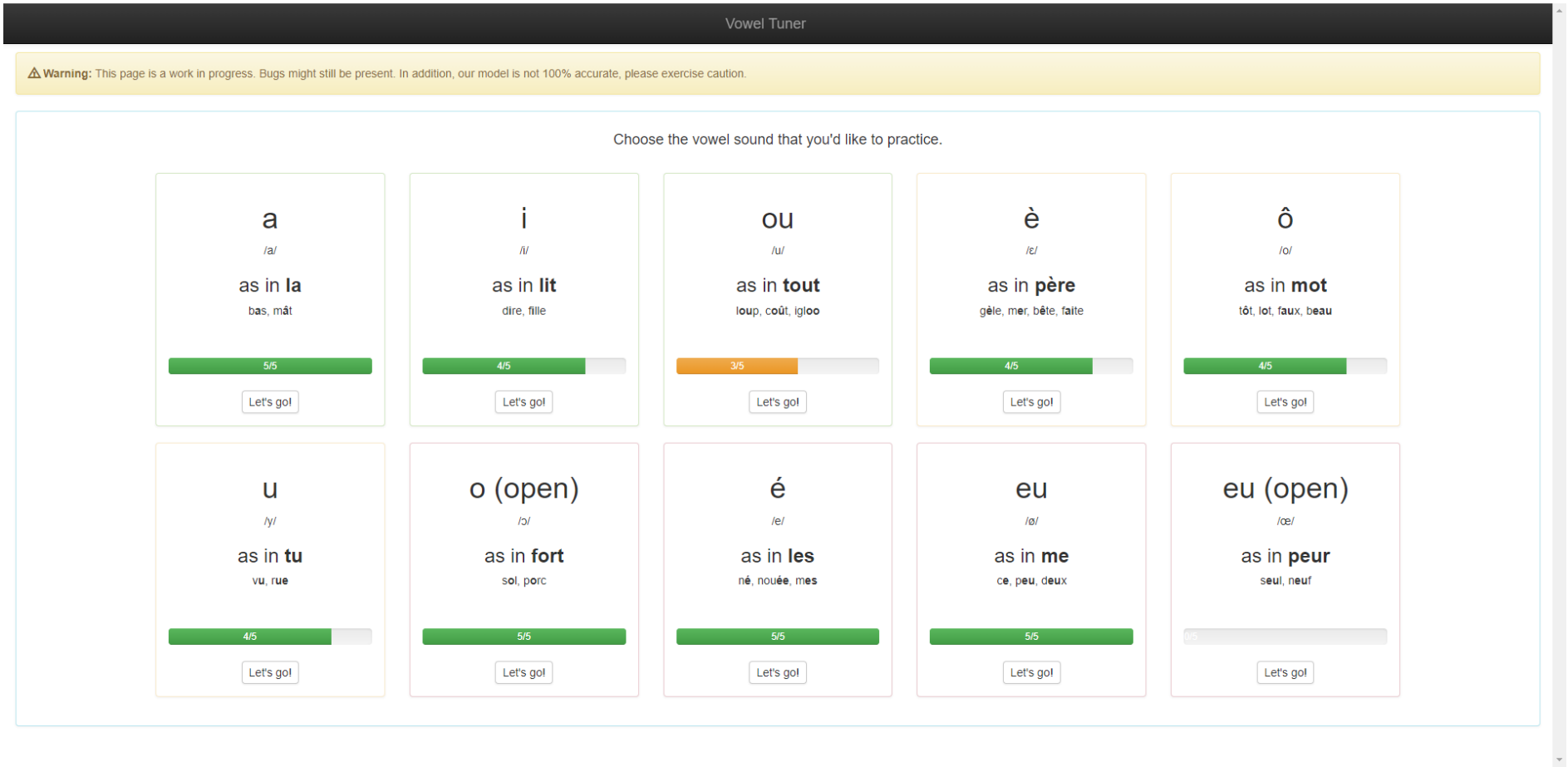


Figure: A native speaker pronouncing each vowel 5 times (linguistic model).

Human evaluation

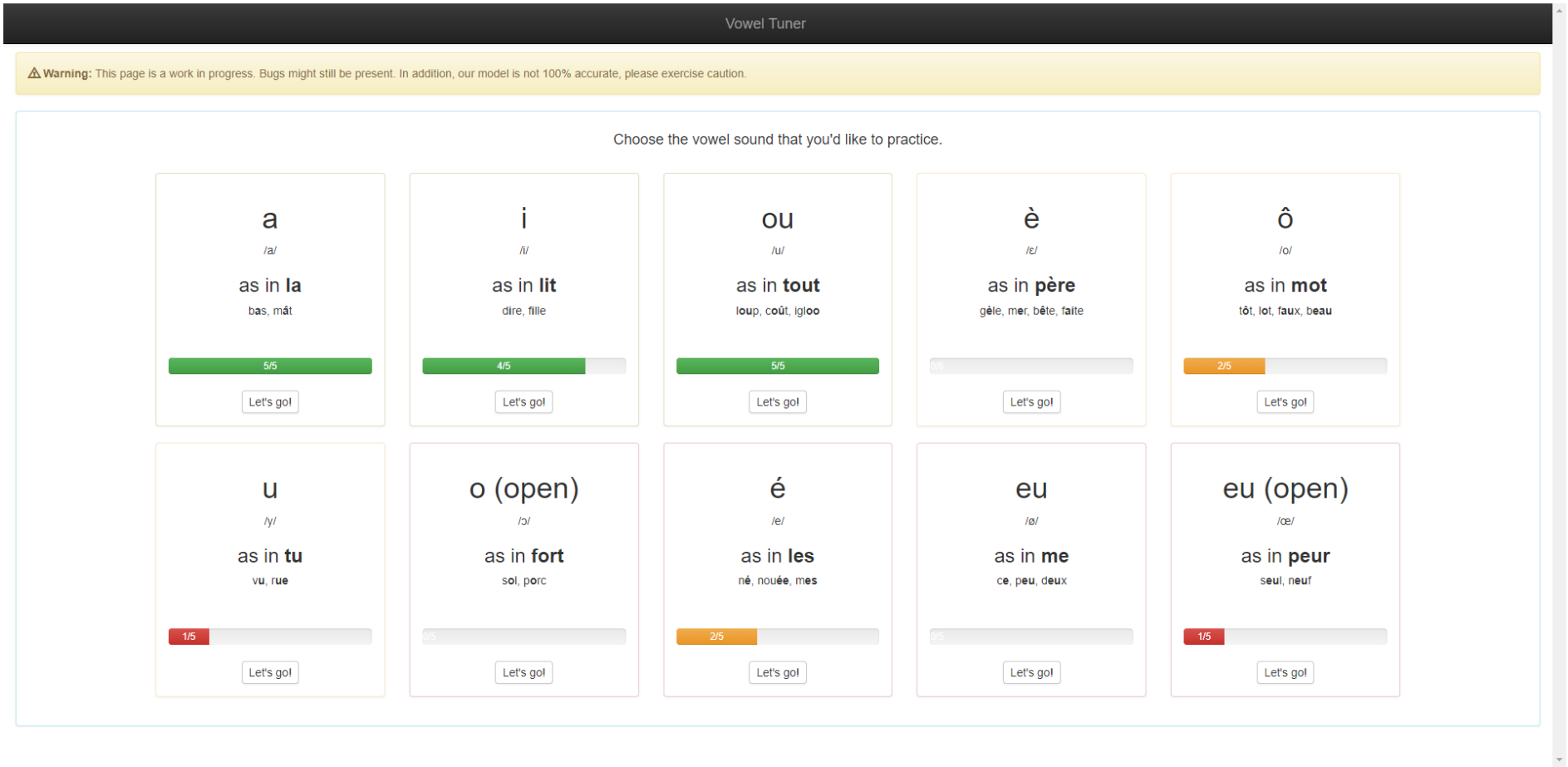


Figure: A native speaker pronouncing each vowel 5 times (neural model).

Results and evaluation

The **linguistic model** is currently selected.

- A lot more robust
- Real performance completely overshadows the NN model
- Remaining issues with /œ/ and /u/ (on our sample)

The **neural model** is currently only able to reliably detect cardinal vowels.

Plan

Plan

What's next?

- Record reference audio
- Record reference visuals
- Get human evaluation from native speakers
- Write the report

Plan

And later?

- Better sound level detection
- Train on more data (from app users?)
- Improve NN robustness (data augmentation)
- Remove gender input if possible
- Nasal vowels (easy)
- Custom feedback based on the user's native language (long)
- Consonants (hard, especially for feedback)
- More varieties of French (hard)

Thank you!

Questions? Feedback?