Vowel Tuner

Soklong HIM Nora LINDVALL Maxime MÉLOUX
Jorge VASQUEZ-MERCADO

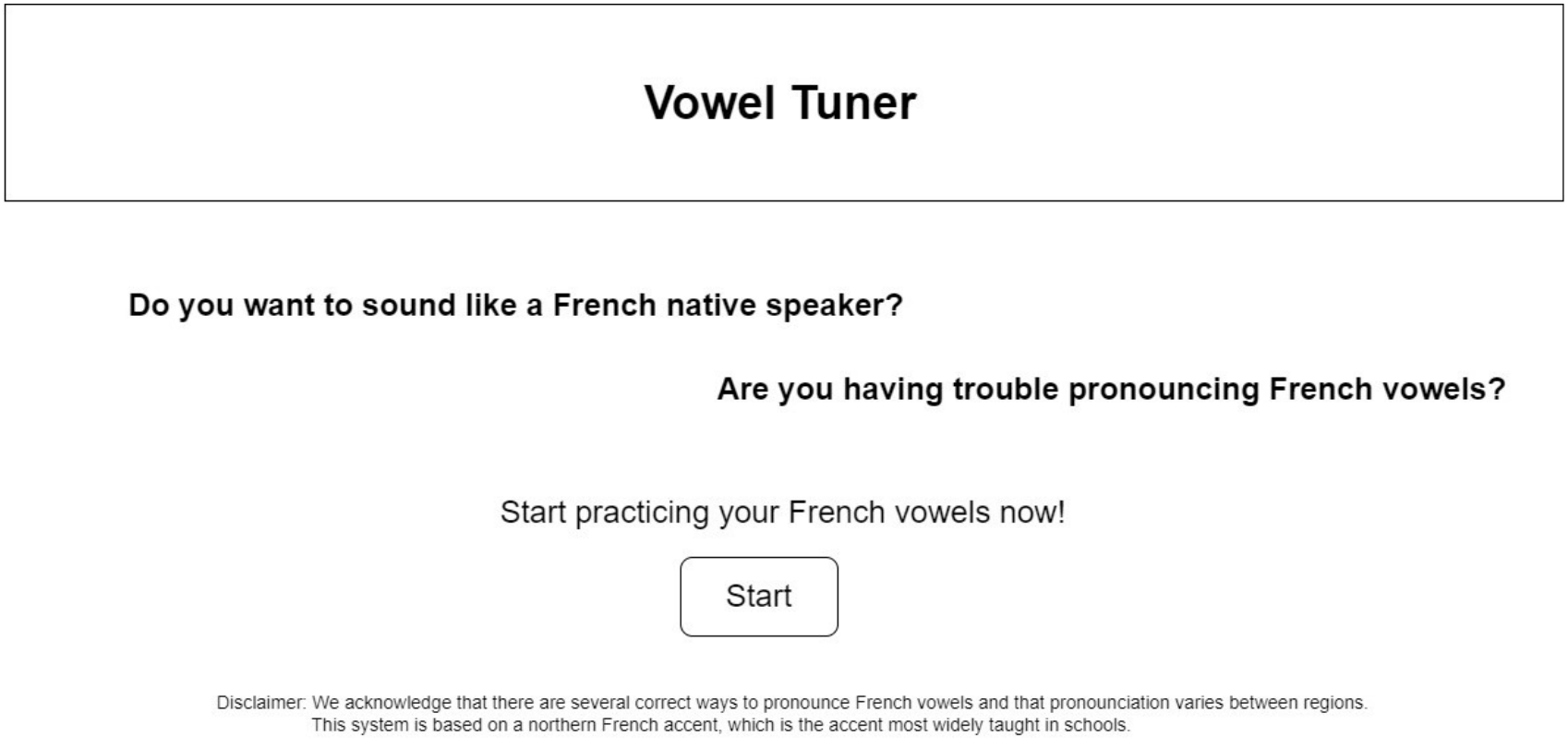
NLP M2

Software Project
Jan. 13, 2023

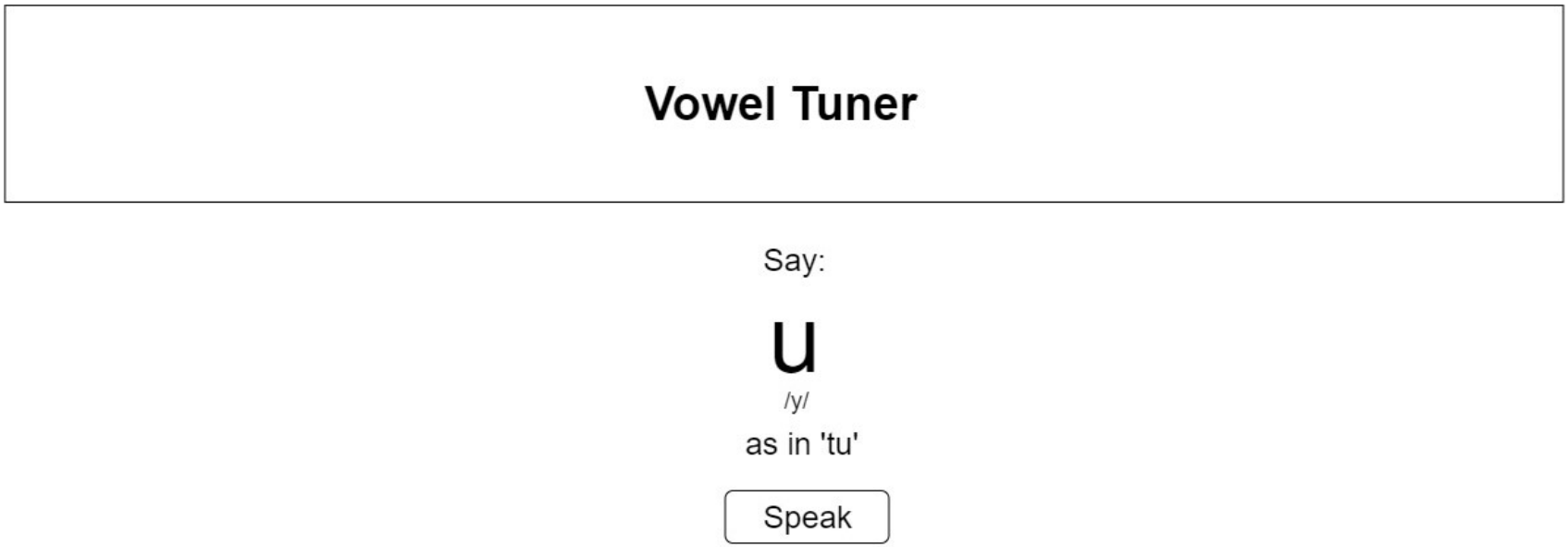
Outline

Plan

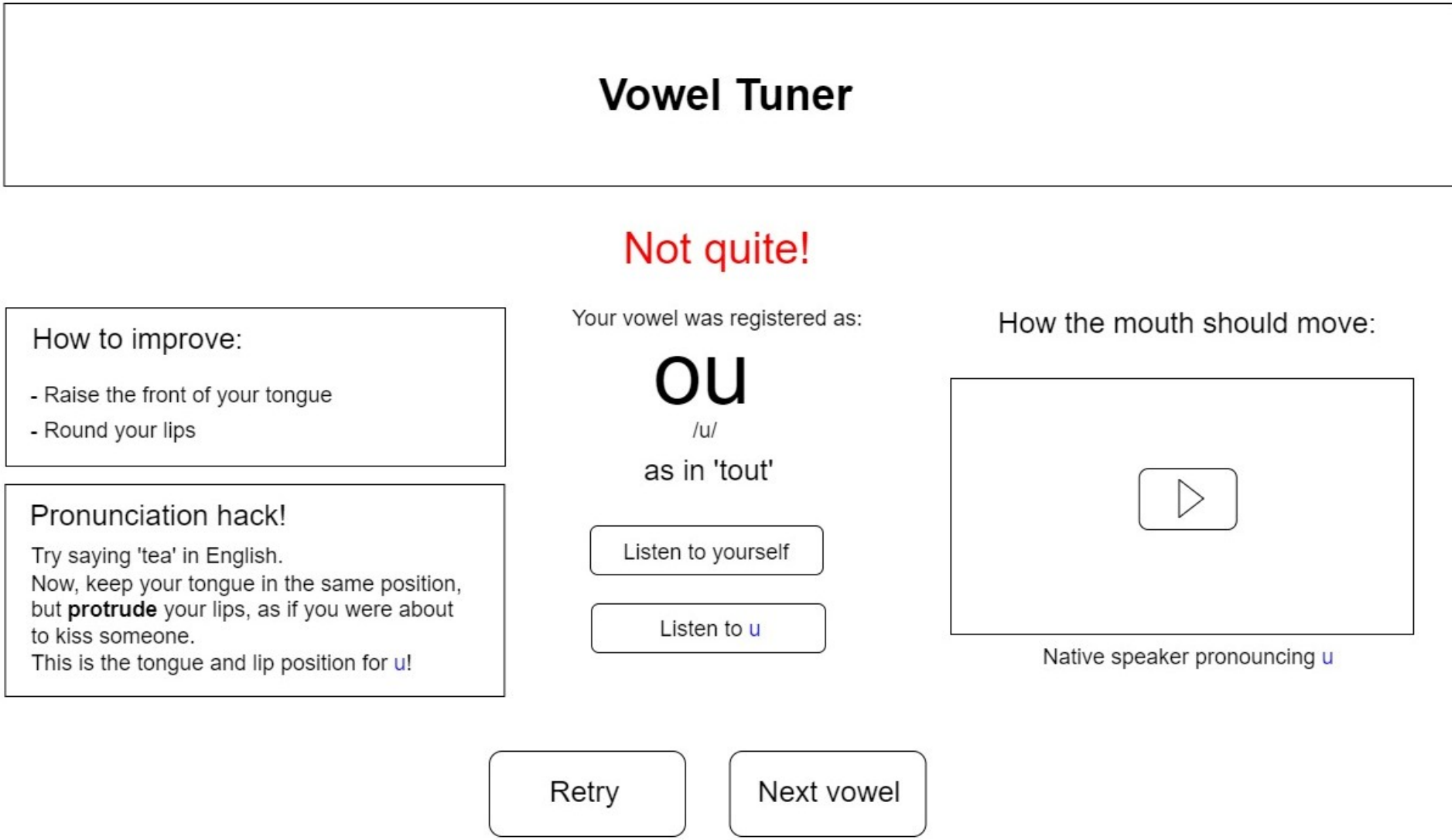
Interface


Vowel Tuner
Do you want to sound like a French native speaker?
Are you having trouble pronouncing French vowels?
Start practicing your French vowels now!
Start
Disclaimer: We acknowledge that there are several correct ways to pronounce French vowels and that pronounciation varies between regions.
This system is based on a northern French accent, which is the accent most widely taught in schools.

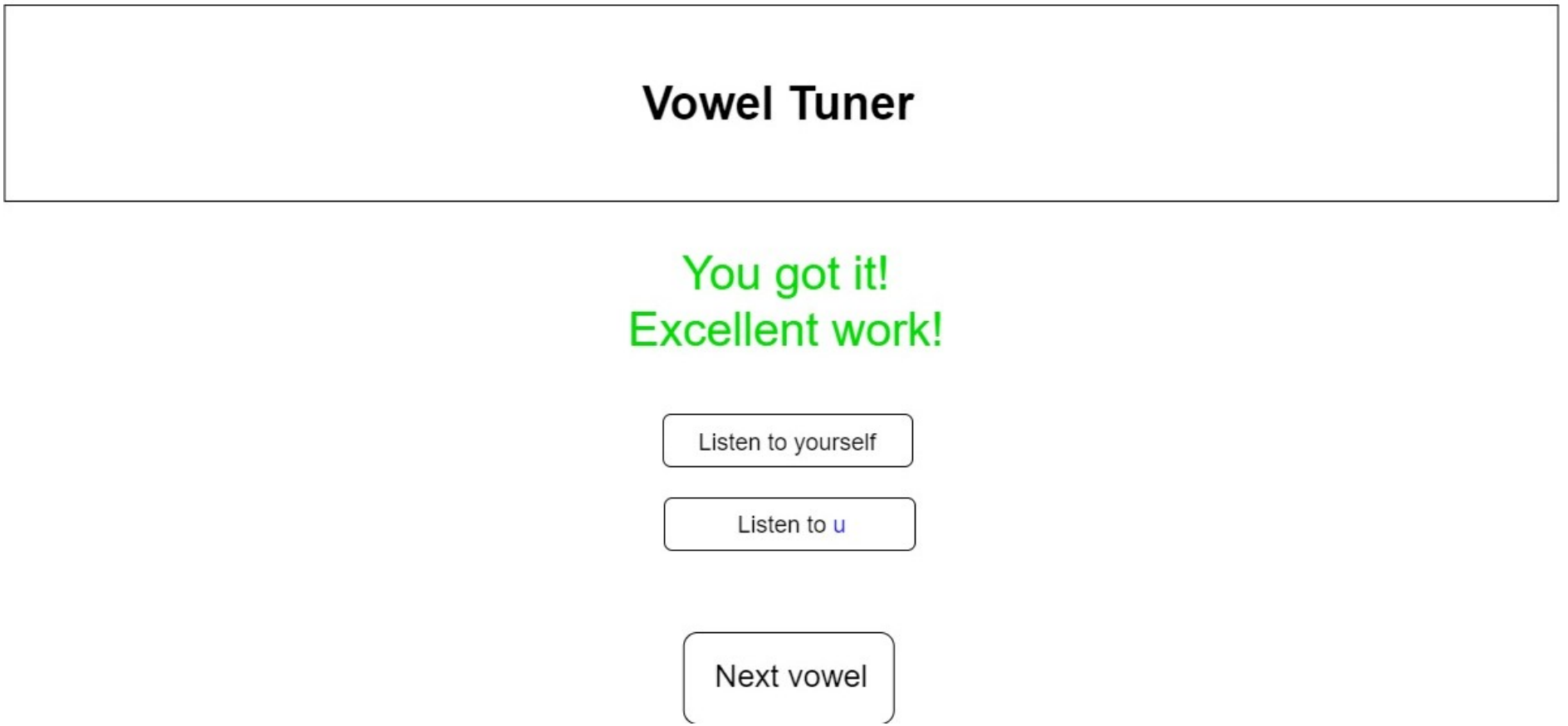
Interface



Interface

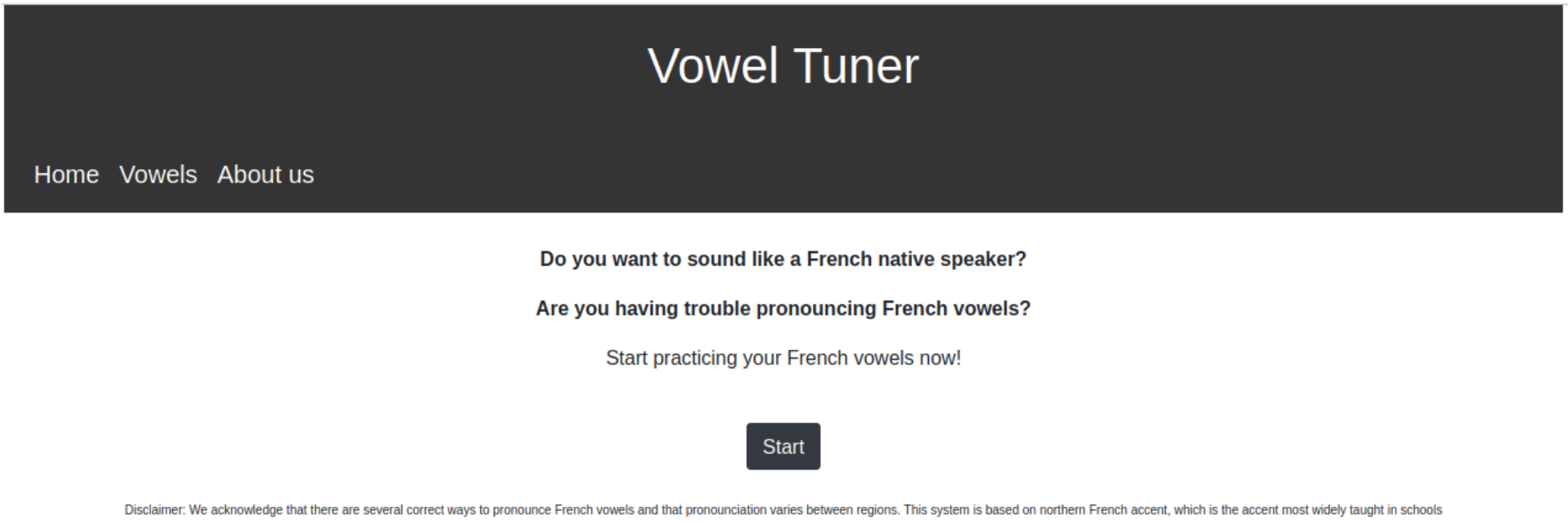

Vowel Tuner
Not quite!
How to improve:
- Raise the front of your tongue
- Round your lips
Pronunciation hack!
Try saying 'tea' in English.
Now, keep your tongue in the same position, but protrude your lips, as if you were about to kiss someone.
This is the tongue and lip position for u!
Your vowel was registered as:
ou
/u/
as in 'tout'
Listen to yourself
Listen to u
How the mouth should move:
Native speaker pronouncing u
Retry
Next vowel

Interface


Vowel Tuner
You got it!
Excellent work!
Listen to yourself
Listen to u
Next vowel

Web App Development

Welcome Interface



Web App Development

Recording Interface

Web App Development

Web App Development - now

- Welcome interface
- Get the audio record from the browser using the user's microphone
- Obtain the .wav file by clicking on the link

Web App Development - to do

- Storing the .wav file automatically in a specific location
- Create the full interface presented before

Plan

The corpus

- 17 men, 21 women
- 1754 (automatically?) annotated vowels
- ≈1750 non-annotated vowels

→ Too few samples for a large neural network?

→ More data potentially available

Corpus processing

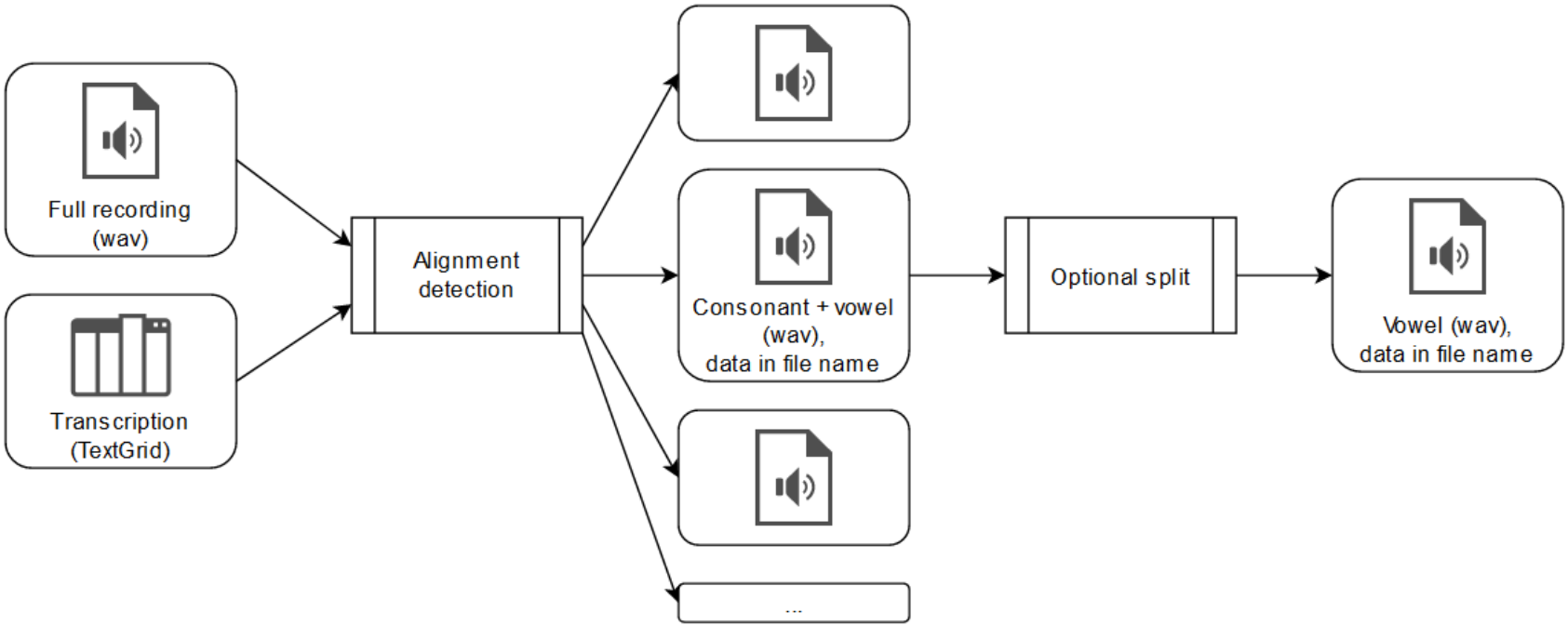


Figure: How a training example is processed

Plan

Architecture

- CNN-based architecture using the image of the spectrogram

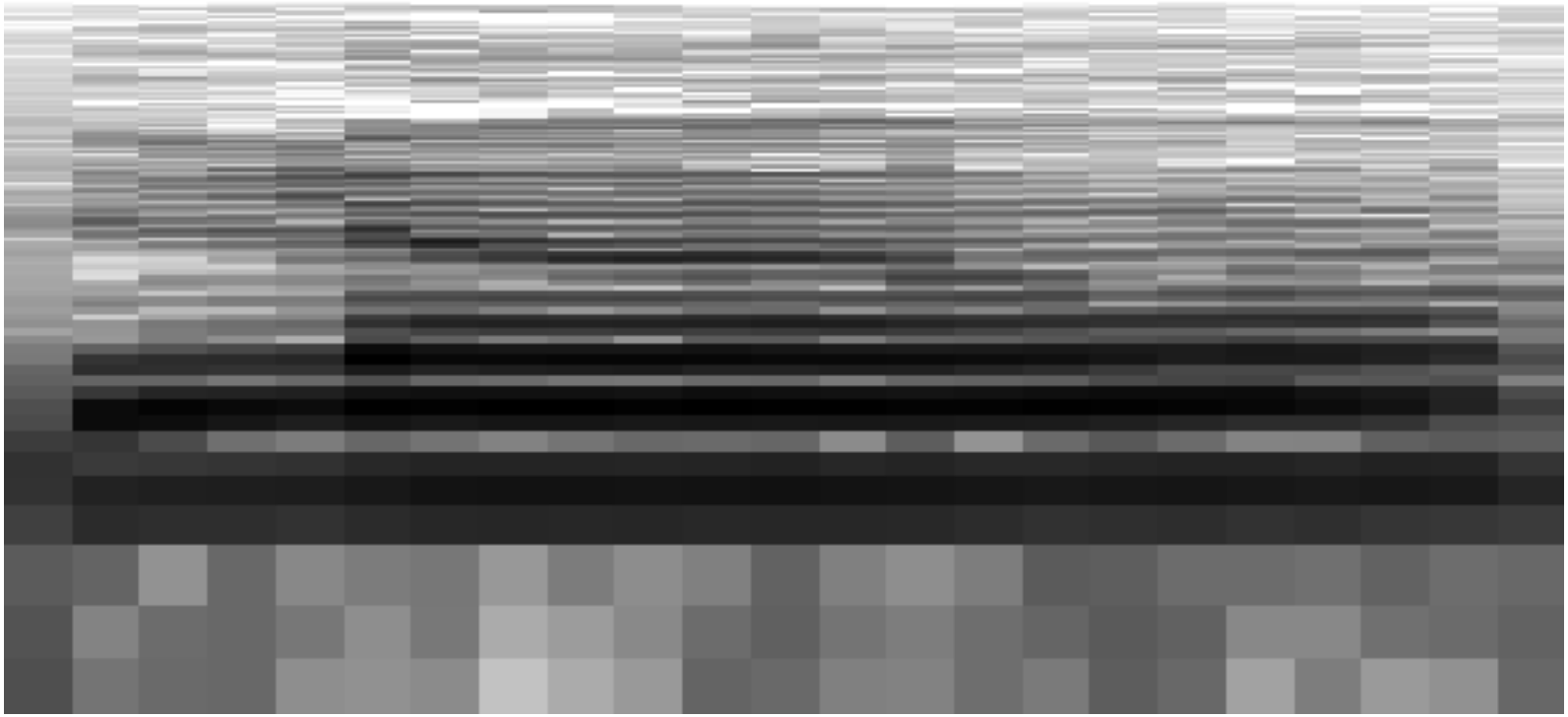


Figure: A cropped recording of the word "lors" /lɔ(ʁ)/.

Architecture

- CNN-based architecture using the image of the spectrogram
- Images padded to the same maximal width (resizing?)
- Quite shallow architecture due to dataset size
- 10 classes output (only oral vowels) and cross-entropy loss
- Hyperparameters: Number of convolutional layers, number and size of fully connected layers, kernel size, max pool kernel size, dropout, activation function, padding, stride, batch normalization, optimizer, batch size, learning rate, vowel only or consonant+vowel...

Preliminary experiments

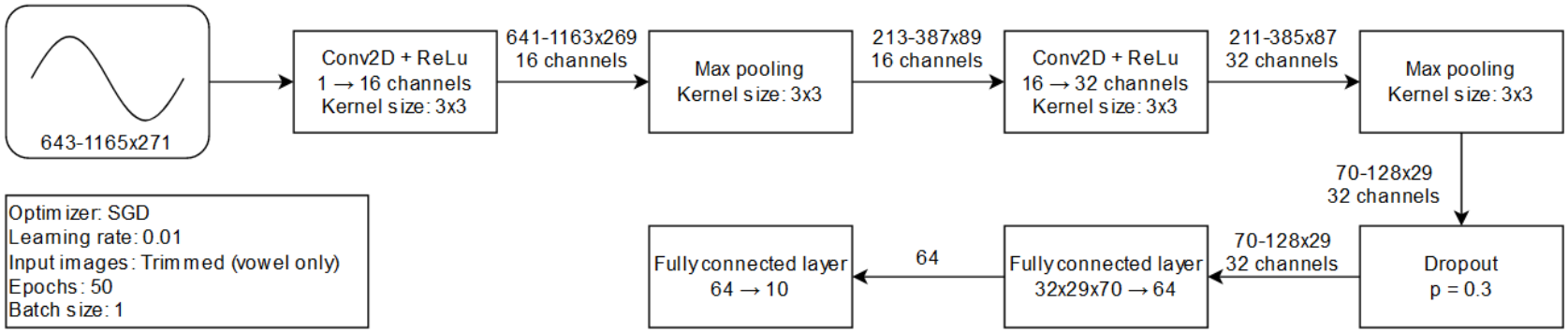


Figure: The best performing architecture so far (4,162,954 parameters)

Accuracy on test set (10% of dataset): 81.51%

Plan

Rule-based approach

Using reference formants: 33% accuracy (last presentation), can be increased to 40-45% with tweaks.

Can we go further?

Input features:

- Formants F1-F4, automatically extracted
- "Only" 960 vowels
- The gender of the speaker
- The previous consonant (m, p, l, s, t, t1)

$\rightarrow$ 4 analog dimensions and 7 binary ones

Input visualization

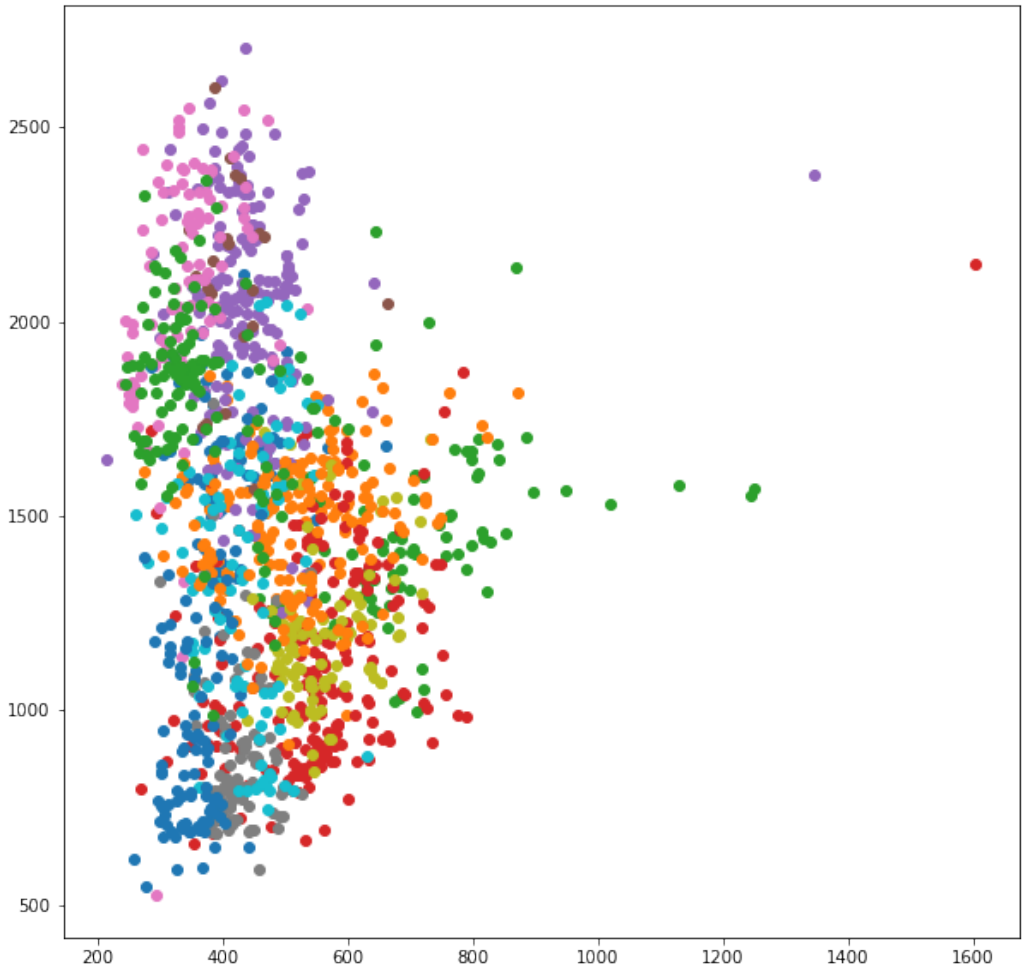


Figure: (F1, F2) for vowels in the dataset

Input visualization (better)

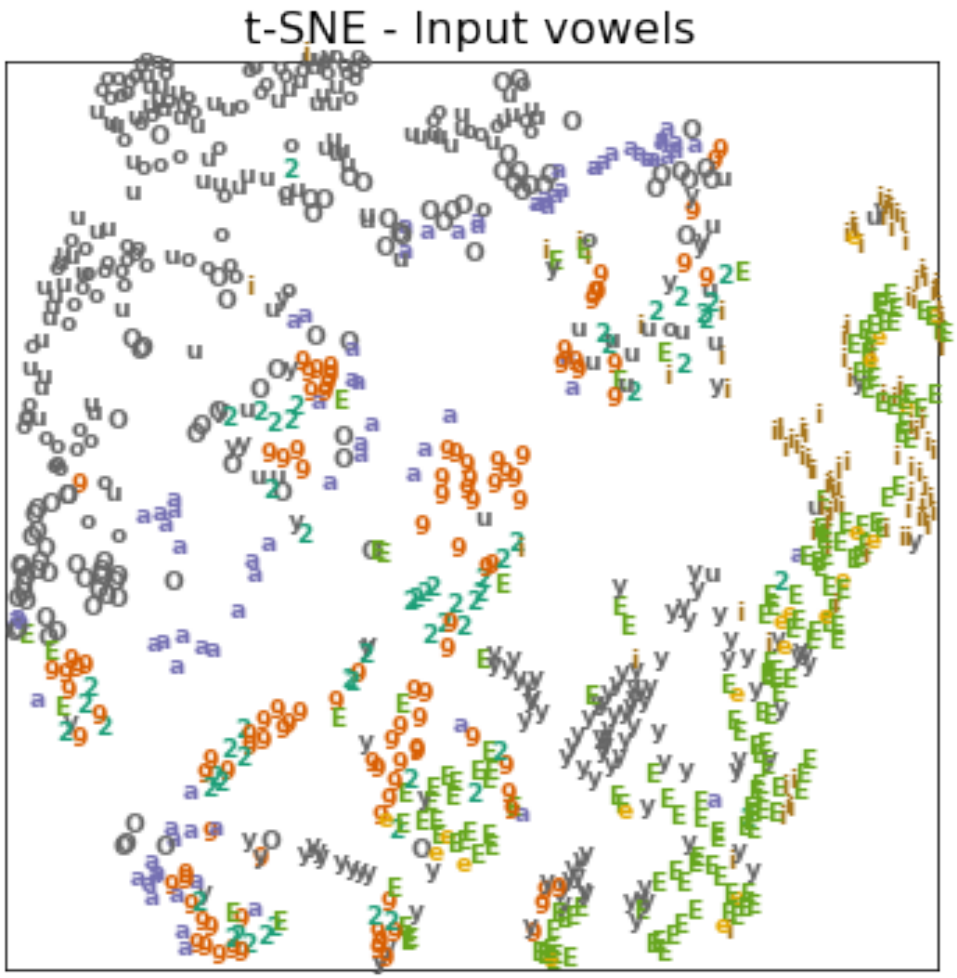


Figure: T-SNE of the 11-dimensional dataset

Some classifiers - Decision Trees

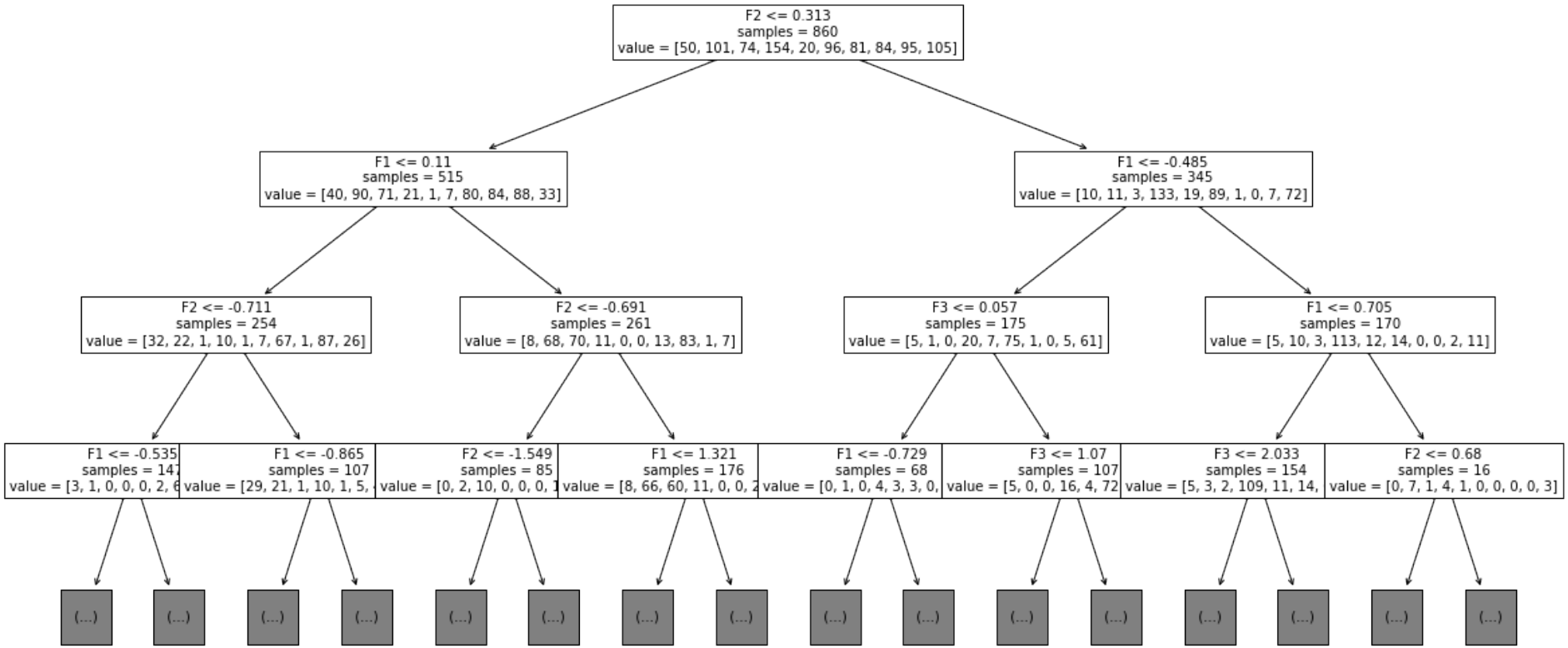


Figure: Decision trees classifier rules

Some classifiers - Multilayer Perceptron

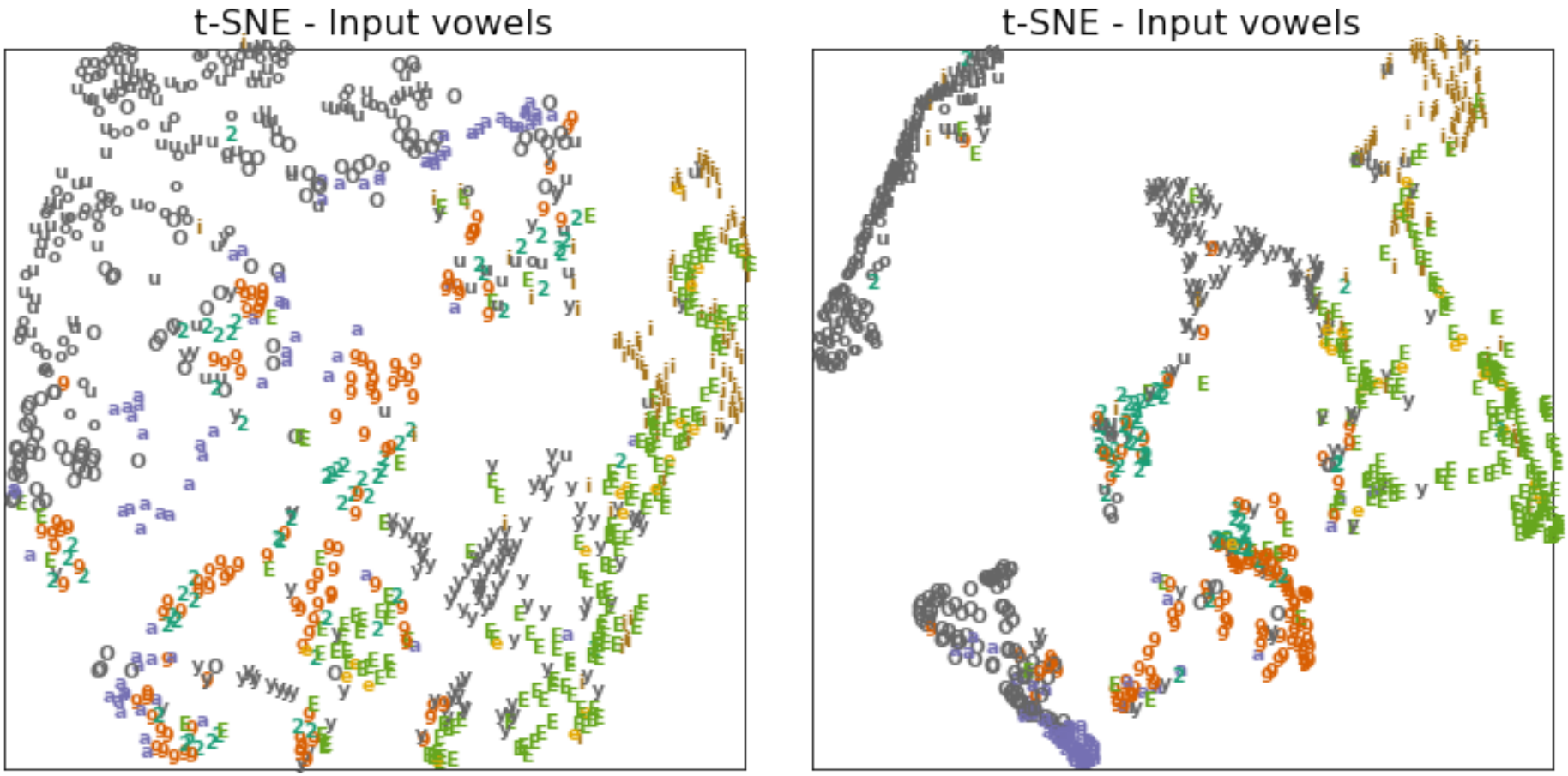


Figure: T-SNE of the input dataset vs. after the last hidden layer of the MLP

Best classifier results

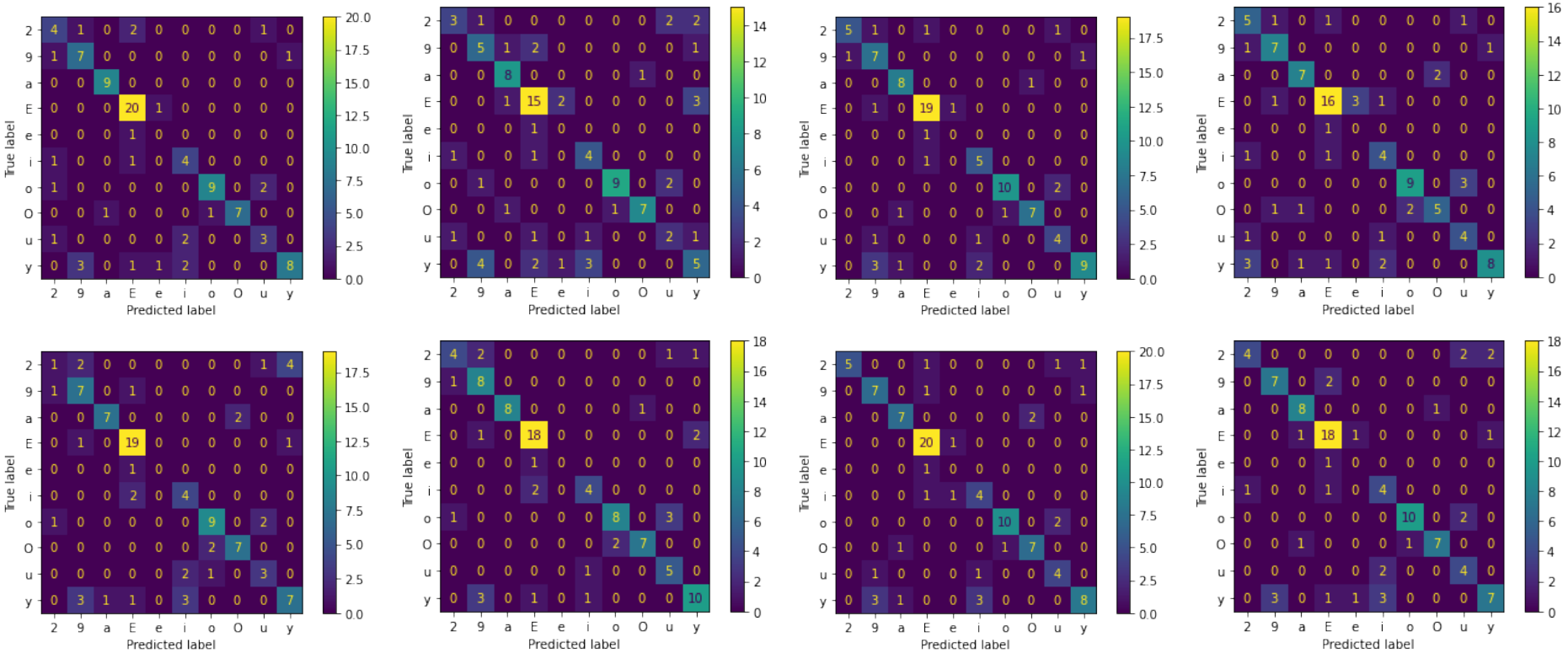


Figure: Confusion matrices for selected classifiers (bagging, decision trees, extra trees, k-neighbors, logistic regression, MLP, random forests, stacking)

Best classifier results

Classifier	Accuracy
*Bagging	73.96%
Decision trees	60.42%
***Extra trees**	**79.79%**
K neighbors	67.71%
Logistic regression	66.67%
Multilayer perceptron	75.00%
*Random forests	75.00%
*Stacking	71.88%

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

Plan

Vowel processing

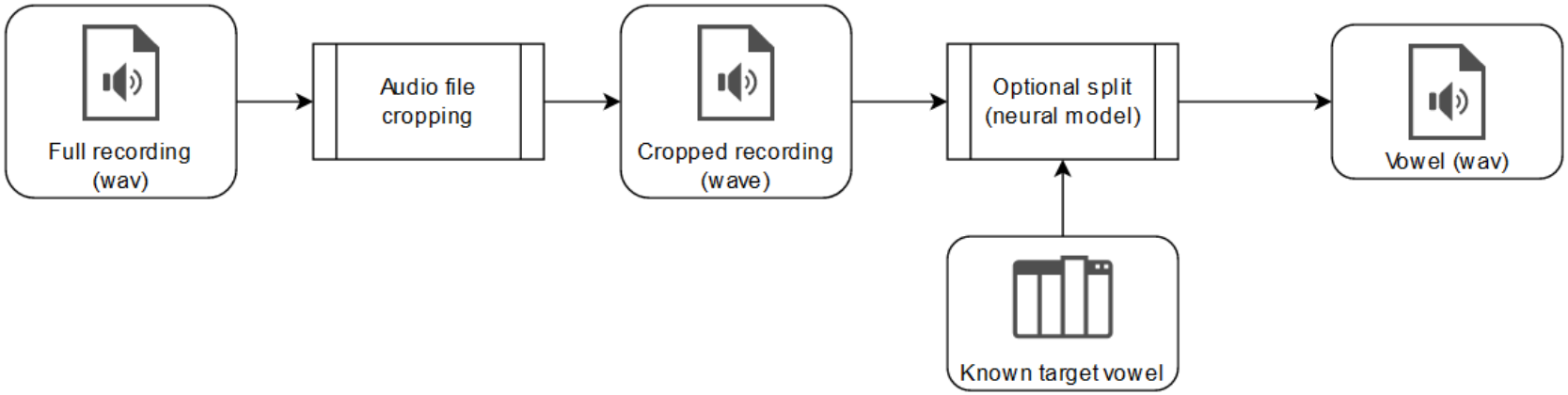


Figure: Vowel extraction from the initial recording

Vowel classification and feedback

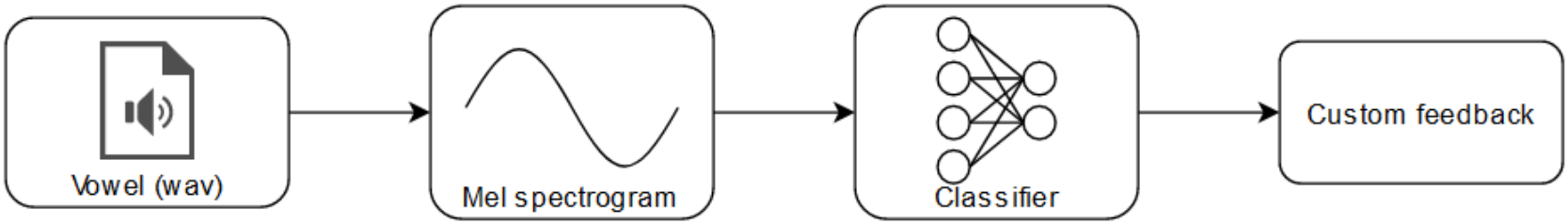


Figure: The deep-learning pipeline

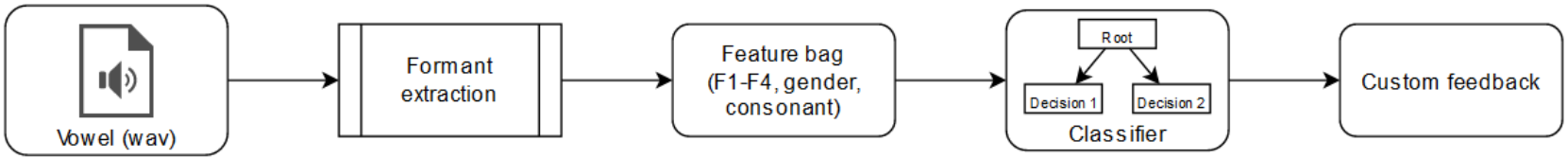


Figure: The rule-based pipeline

Idea: combine both to have a stronger model?

Plan

Plan

What's next?

- Create content (text prompts)
- Perform more experiments (neural network)
- Record visuals
- Complete interface
- (Annotate more .wav files)
- Write report

What about the corpus collection??

-Organizational issues
-No time

Thank you!

Questions? Feedback?