

An Open Source Tool for the Training of the Pronunciation of Vowel Systems

Wilbert Heeringa & Hans Van de Velde

FRYSKE  AKADEMY

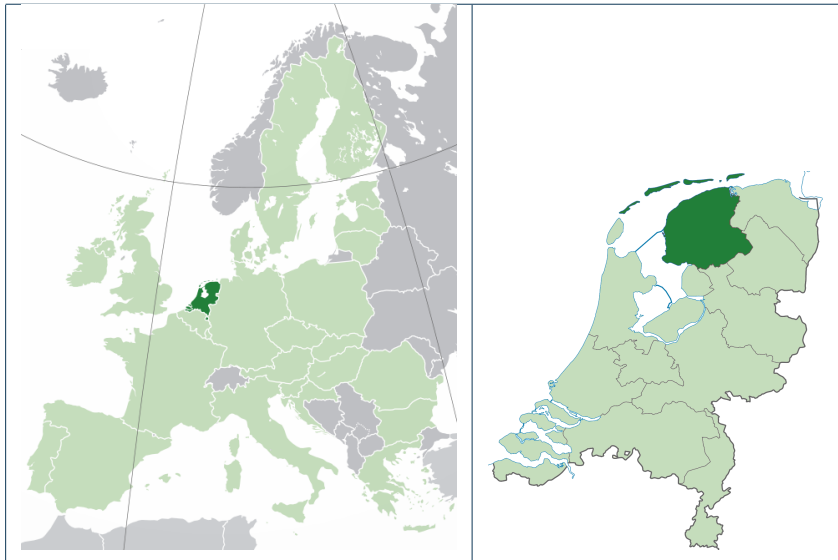
COLING workshop

Documentation for Education of Lesser Used Languages

Rēzekne, May 1-4, 2023

About the Fryske Akademy

- Founded in 1938.
- Core task is to conduct high quality scientific research on the Frisian Case, with impact.
- Fundamental scientific research in the fields of
 - linguistics, sociolinguistics, sociology, lexicography,
 - multilingualism and language learning, minority languages,
 - Old Frisian and the history of the Middle Ages, the early modern era and the latest era.
- ICT and the development of digital tools and collections.



Maps taken from Wikipedia and adapted.



Picture: Erik and Petra Hesmerg

Introduction

Goal

- Development of an online vowel pronunciation trainer
- Device independent.
- Generic: can be used for any language when the relevant speech models and samples are provided.
- Especially suitable for low-resource languages and commercially less interesting languages.

Goal

- Open-source.
- Becomes available in the public domain.
- Focus on the pronunciation of vowels.

The tool is useful for...

- learning the pronunciation of a language,
- learning or correcting a specific accent,
- speech therapy.

Scientific tool

- The app can also track the learning process and progress of users.
- It may give us insight in which vowels are easy to learn and which are not.

Outline

- Existing vowel pronunciation programs
 - Coco
 - Apps for Android/iOS
- The open-source pronunciation trainer
 - Vowels
 - Whole words
 - Interface

Existing vowel pronunciation programs

Existing vowel pronunciation programs

Coco

Coco

- Computer program for the acquisition of the pronunciation of the vowels of Belgian standard Dutch (spoken in Flanders).
- Developed in 2015–2016 for the Thomas More University College in Antwerp, department of Speech Language Therapy.
- Inspired by
 - Klinkermikken ‘vowel shooting’
 - Lorre

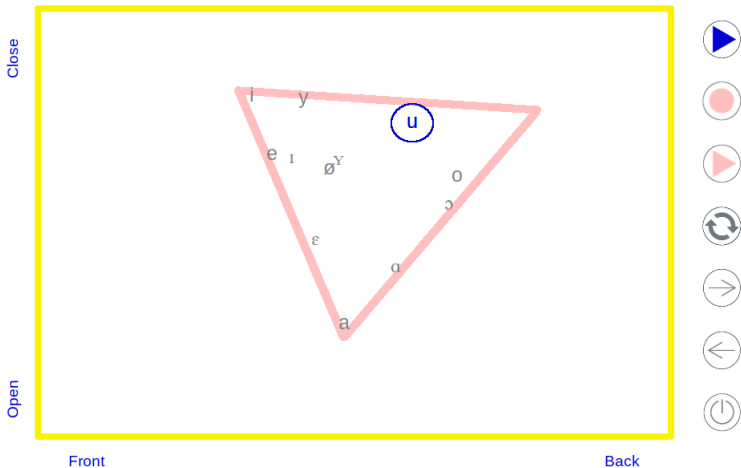
Coco

- The two programs were implemented by Ing. Jos Pacilly (Leiden University), supervised by Prof. Dr. Vincent van Heuven.
- The idea of 'Klinkermikken' originates from Povel & Wansink (1986): darting with your voice.
- Coco is implemented in Praat script.

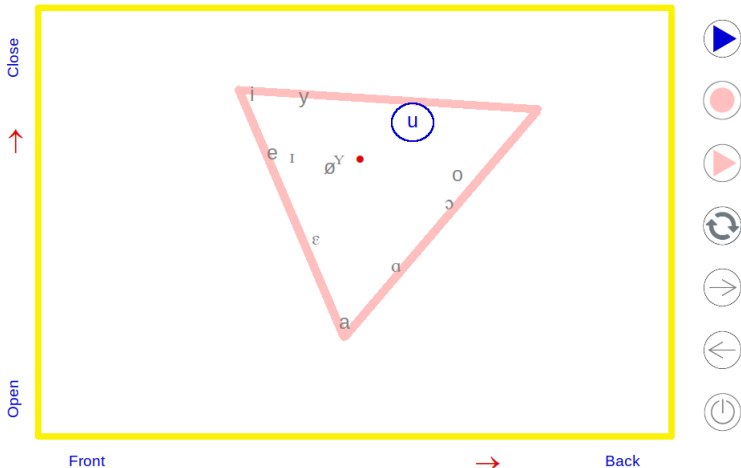
Vowels ...

- should be monophthongs;
- presented in the context of monosyllabic words, where the consonants in those words are obstruents.

soep



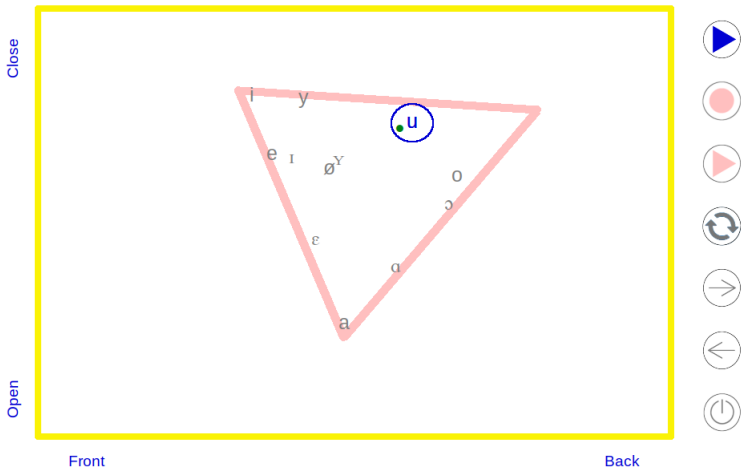
soep



Reduce the jaw opening and raise the back part of the tongue higher towards the soft palate.

Reduce the resonance space by making the constriction between tongue and palate more at the back of the mouth and/or make the lips more rounded.

soep



Existing vowel pronunciation programs

Existing vowel pronunciation programs

Apps for Android/iOS

Apps for Android/iOS

Usually a categorization in:

- short vowels
- long vowels
- diphthongs
- voiced consonants
- voiceless consonants






Apps for Android/iOS

- Within each category IPA symbols are shown.
- Per IPA sound a list of words (or phrases or sentences) is shown that contains that sound.
- The user's pronunciation is rated on a scale of five stars.








5:32 PM Basic






Short Vowel Sound

 /ɪ/	 /ʊ/	 /ʌ/	 /ʊ/	 /ʌ/
--	--	--	--	--






Long Vowel Sound

 /i:/	 /e:/	 /ɜ:/	 /ɔ:/	 /ɔ:/
---	---	---	---	---






Diphthong

 /ɪə/	 /aʊ/	 /aʊ/	 /aʊ/	 /aʊ/
---	---	---	---	---

Voiced Consonant

 /b/	 /d/	 /tʃ/	 /d/	 /d/
--	--	---	--	--

Unvoiced Consonant

 /p/	 /t/	 /k/	 /k/	 /t/
--	--	--	--	--



The screenshot shows a mobile application interface for English pronunciation. At the top, there is a status bar with the time 5:31 PM and various icons. Below that is an orange header with a back arrow and the word "Speaking". There are two green buttons: "Hide" on the left and "Video" on the right. In the center is a diagram of a human mouth in profile, showing the tongue touching the roof of the mouth. Below the diagram is a text box containing the instruction: "/ɪ/ is a short vowel sound. The short /ɪ/ sound is created with the tongue rounded upward. You'll probably be able to feel your top side teeth with the side of your tongue." Below this text are three tabs: "Words", "Phrases", and "Sentences", with "Phrases" being the active tab. The main content area is divided into four rows, each representing a phrase. Each row has a text label at the top, a set of five stars in the middle, a microphone icon below the stars, and two "Listen" buttons (US and UK) at the bottom. The first row is for "drink milk" with five green stars. The second row is for "litter bins" with one green star and four grey stars. The third row is for "biggest building" with five green stars. The fourth row is for "this guitar" with one grey star and four grey stars. At the bottom right of the screen is a red circular button with a white square icon.

The open-source pronunciation trainer

Properties

- Web app
- Device-independent, focus on smartphone and tablet.

Evaluation of ...

- vowel pronunciation
 - short vowels, long vowels, diphthongs
 - detailed feedback is provided
- pronunciation of a word as a whole
 - simple feedback
 - rating on a scale of five stars.

Generic

- A procedure for adding speech samples of new reference speakers will be included.
- A procedure will be included for adding language-specific speech models (forced alignment models, wav2vec models)
- These models make it possible to compare the pronunciation of a user with the pronunciation of the reference speakers.

The open-source pronunciation trainer

The open-source pronunciation trainer

Vowels

We need ...

- recordings of reference speakers, e.g. standard variety, regional accents or dialects, different genders, ages, types of voices;
- a procedure for segmentizing and transcribing the speech of the users and the speech of the reference speakers;
- a distance measure that reflects the difference between the pronunciation of the user and the reference speaker, either for vowel segments in words, or for whole words;

Recording of reference speakers

- If available, existing recordings are used, otherwise, new recordings will be made.
- Frisian: 21 monophthongs, 25 diphthongs and 22 sequences of three or four vowels, which are made up of a glide and a long vowel or diphthong.
- Latgalian: 10 monophthongs and 3 diphthongs.
- Each sound should be represented by at least four words that represent different phonological contexts.

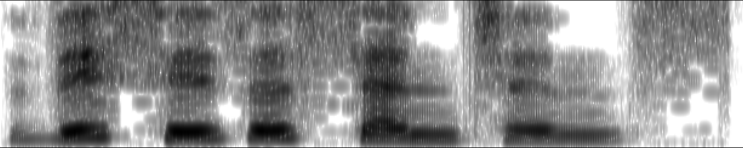
Segmentation of pronunciation recordings

- A crucial part of automatic processing and evaluation a speaker's pronunciation is the segmentation of a speaker's pronunciation in phonetic segments.
- Thus the location of the vowel in the acoustic signal can be found.
- Segmentation can be done by using a forced alignment model.
- We will train a forced alignment model with the Montreal Forced Aligner (MFA).

Segmentation of pronunciation recordings

- Forced alignment refers to the process by which orthographic transcriptions are aligned to audio recordings to automatically generate phone level segmentation.

https://linguistics.berkeley.edu/plab/guestwiki/index.php?title=Forced_alignment



This is a sentence

This

is

a

sentence

ð

ɪ

s

ɪ

z

ɐ

s

ɛ

n

t

ə

n

s

Montreal Forced Aligner

- MFA is an open-source trainable forced aligner.
- Provides a user-friendly wrapper to the Kaldi ASR toolkit for acoustic model training and alignment.
- Kaldi is an actively maintained, open-source automatic speech recognition (ASR) toolkit.

Montreal Forced Aligner

- Many forced aligners use monophone acoustic models.
- MFA uses more complex triphone acoustic models and speaker-adapted features (Barth et al. 2020)
- Triphone models allow for context sensitivity (coarticulation).
- MFA offers the phoneme-level of granularity (Biczysko 2022), is trainable on any language and well documented.

For running MFA we need ...

- a pronunciation dictionary;
- at least 45 minutes of speech where the sentences are labeled by orthographic transcriptions



Montreal Forced Aligner

Pronunciation dictionaries

- For Frisian:
Frysk Hânwurdbboek and Foarkarswurldlist
- For Latgalian:
a pronunciation dictionary is in preparation

Speech corpora: Frisian

- Common Voice Corpus 13.0
sample sentences recorded by volunteers, the sentences are orthographically transcribed, 67 validated hours;
- FAME corpus
annotated radio broadcasts in the Frisian language, 50 years time span, 8.5 hours training data, 1 hour development set, 1 hour test set, used for ASR;
- Boarnsterhim Corpus
parallel corpus, sentences, Frisian speakers of a former municipality in the geographical center of the Frisian language area;
- Corpus council meetings
spontaneous speech of council meetings, 70 hours.

Speech corpora: Latgalian

- MuLaR: Spoken Latgalian in audio recordings and transcripts
- MuLa2022: Corpus of Contemporary Latgalian Texts 2022

Comparing to reference speakers

- Compare a user's vowel pronunciation to the vowel pronunciations of a reference speaker by comparing Mel-frequency cepstral coefficients (MFCCs).
- Compare MFCCs at multiple time points (e.g. at 25%, 38%, 50%, 62% and 75% of the vowel duration).
- MFCCs are popular due to their greater invariance to physical differences between speakers.

Comparing to reference speakers

- 39 MFCC coefficients per time frame.
- The first 12 parameters are related to the amplitude of frequencies.
- The 13th parameter is the energy in the frame.
- The last 13 parameters are the dynamic changes from the current frame to the next frame.

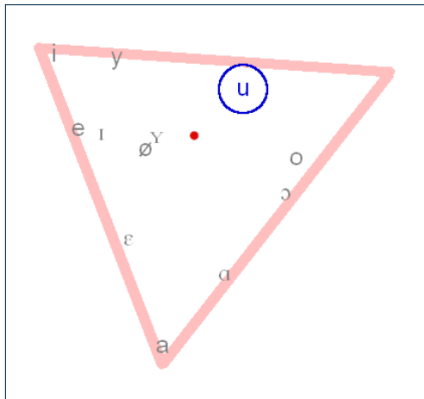
<https://jonathan-hui.medium.com/speech-recognition-feature-extraction-mfcc-plp-5455f5a69dd9>

Procedure

- Compare the user's pronunciation to all vowels pronounced by the reference speaker and find the closest vowel.
- If the closest vowel is not the target vowel, give feedback on how to improve the pronunciation and hit the target vowel.

Example

- A user is supposed to pronounce an [u].
- The closest reference speakers's vowel is [ɤ].
- The user should lower the pronunciation and pronounce the vowel more to the front.



Feedback:

Reduce the jaw opening and raise the back part of the tongue higher towards the soft palate.

Reduce the resonance space by making the constriction between tongue and palate more at the back of the mouth or make more roundness with the lips.

The open-source pronunciation trainer

The open-source pronunciation trainer

Whole words

wav2vec

- Bartelds & Wieling 2022 calculated linguistic distances between Dutch dialects by comparing their word pronunciations acoustically with a wav2vec 2.0 model.
- Discretizes an audio waveform into timesteps.
- When comparing samples, the time points of the one are aligned to the time points of the other with Dynamic Time Warping.
- The differences between aligned time points are calculated and added together → pronunciation distance.

XLSR-53 model

- Bartelds & Wieling 2022 obtained good results using the XLSR-53 model.
- Pre-trained on 56,000 hours of speech in 53 languages including Dutch, German and Latvian.
- They fine-tuned it on 243 hours of Dutch speech from the Spoken Dutch Corpus (Oostdijk et al. 2000).

Available models

- XLSR-53 model fined-tuned on Frisian Common Voice dataset by Wietse de Vries.

28 March 2021 <https://huggingface.co/wietsedv/wav2vec2-large-xlsr-53-frisian>

- XLSR-53 model fined-tuned on FAME corpus by 'techsword'.

23 July 2022 <https://huggingface.co/techsword/wav2vec-large-xlsr-53-frisian-fame/tree/main>

- XLSR-53 model fined-tuned on Latvian Common Voice dataset by Anton Lozhkov.

28 March 2021 <https://huggingface.co/anton-l/wav2vec2-large-xlsr-53-latvian>

Fine-tuning

- Common Voice (CV) datasets consist of separate sentences.
- Therefore, models fine-tuned on CV datasets are likely suitable for evaluation of word pronunciations.
- The CV datasets have grown since 28 March 2021, so we will fine-tune the models again.
- Instead of using the Latvian model we may fine-tune the XLSR-53 model with Latgalian speech data.

Rating

- The DTW distances are converted into a scale of five stars.

The open-source pronunciation trainer

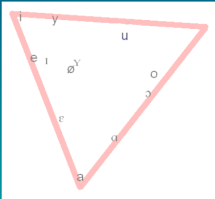
The open-source pronunciation trainer Interface

Interface

- The example in the next slides is created by adapting the layout of the 'Speak English Pro' app.

6:00 PM 1/20

sheep
/ʃi:p/



☆☆☆☆

< [Microphone icon] >

The image shows a screenshot of a mobile application interface. At the top, the status bar displays '6:00 PM' and '1/20'. Below this, a dark blue header contains a close button (X), the page number '1/20', and a square icon. The main content area has a teal background. The word 'sheep' is displayed in white, with its phonetic transcription '/ʃi:p/' below it. A white square contains a vowel chart with a red triangle highlighting the position of the vowel /i:/. The chart shows various vowel symbols: 'i' and 'y' at the top left; 'e', 'ɪ', and 'ø' on the left side; 'u' at the top right; 'o' and 'ɔ' on the right side; 'ɛ' and 'a' at the bottom left; and 'ɑ' at the bottom right. Below the chart are five white stars. At the bottom of the screen, there is a navigation bar with a back arrow, a microphone icon, and a forward arrow. A speaker icon is visible in the bottom right corner of the teal area.

6:00 PM 1/20

sheep

/ʃi:p/

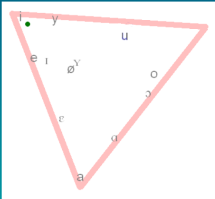
☆☆☆☆

Reduce the jaw opening and raise the back part of the tongue higher towards the soft palate.

< [Microphone icon] >

6:00 PM 1/20

sheep
/ʃi:p/



★★★★★

< [Microphone icon] >

The image shows a screenshot of a language learning application. At the top, the time is 6:00 PM and the page number is 1/20. The word 'sheep' is displayed in white text on a dark blue background, with its phonetic transcription /ʃi:p/ below it. In the center, a white square contains a vowel chart with a red triangle. The triangle's vertices are labeled with the vowels /i:/ (top-left), /e:/ (bottom-left), and /a:/ (bottom-right). Other vowels are scattered within the triangle: /y/ near /i:/, /ø/ and /y/ near /e:/, /u/ near the top, /o/ and /ɔ/ near the right side, and /ɑ/ near /a:/. Below the chart are five yellow stars. At the bottom, there are navigation arrows and a microphone icon.

Paɫdis!