Universität Zürich^UZH

voxplorer
–
The Voice Explorer

Alessandro De Luca

LiRI – CL UZH

Schedule

1. Motivation
2. What is *voxplorer* and where to find it
3. How to install *voxplorer*
4. DEMO 1: Visualizer mode (Saarbrücken data)
5. DEMO 2: Speaker embeddings mode (CANDOR)
6. Feature extraction mode: parameter settings
7. DEMO 3: Feature extraction mode (CANDOR)
8. DEMO 4: Filtering mode
9. Future plans, goals, and Q&A

Motivation

- With the recent and fast improvements in computational power and storage capacity, driven by the need of more and better descriptive data that covers a large study population, speech corpora have started to become increasingly large and complex.
- E.g. CANDOR:
  - Multi-modal data: archived (.zip) = 1TB
  - Only full-length WAV files ≅ 440 GB
  - Processed (VAD + filtering and snipping) 28480 files [942 speakers]

Motivation

Better access to computational resources for all!

What is *voxplorer*?

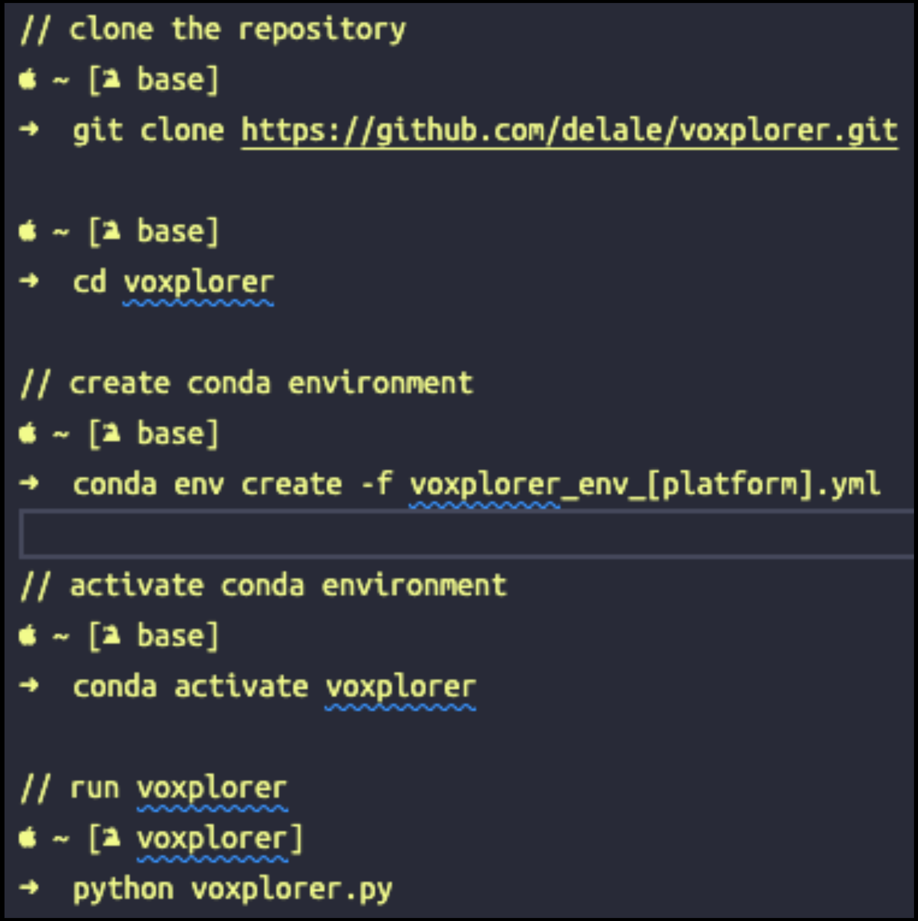
- *voxplorer* started as an interactive visualization tool, BUT…
- An interactive tool for exploration and subsetting of large corpora
- A powerful automatic feature extraction tool

Where to find *voxplorer*:

https://github.com/delale/voxplorer

Installation demo…

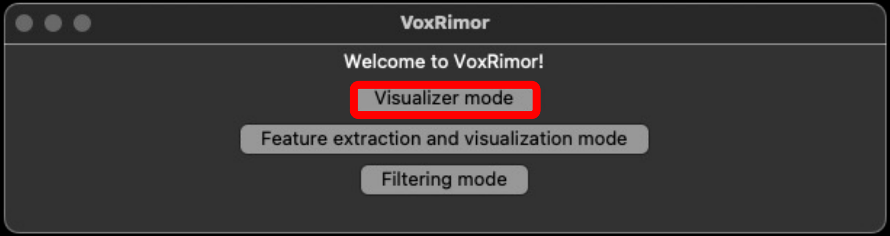
- Clone the repository
- Install miniconda or Anaconda
- Create the *voxplorer* conda environment
- Activate the environment
- Have fun!



```
// clone the repository
 ~ [👤 base]
➜  git clone https://github.com/delale/voxplorer.git

 ~ [👤 base]
➜  cd voxplorer

// create conda environment
 ~ [👤 base]
➜  conda env create -f voxplorer_env_[platform].yml

// activate conda environment
 ~ [👤 base]
➜  conda activate voxplorer

// run voxplorer
 ~ [👤 voxplorer]
➜  python voxplorer.py
```

DEMO 1: Visualizer mode

```
                filename vowel gender  age health     F0mean   F0median  ...    mfcc14    mfcc15    mfcc16    mfcc17    mfcc18    mfcc19    mfcc20
0         1-a_n_20f-hlth.wav     a      f   20   hlth 274.431673 275.862069  ...  0.356511  0.556554 -0.127219 -1.107133 -1.910601 -0.961069  0.149626
1        10-a_n_22f-hlth.wav     a      f   22   hlth 199.024390 200.000000  ...  0.353346 -0.046021  0.196454 -0.045817  0.052015  0.299132  0.111662
2      1006-a_n_32f-hlth.wav     a      f   32   hlth 205.128205 205.128205  ...  0.328761  0.124032  0.082613  0.151899  0.153524  0.160330  0.650411
3       102-a_n_39f-hlth.wav     a      f   39   hlth 198.482385 200.000000  ...  0.102856  0.202720  0.078172  0.003700  0.161462 -0.053153  0.284400
4      1022-a_n_45f-hlth.wav     a      f   45   hlth 216.750083 216.216216  ...  0.198137  0.186112 -0.140532  0.402129  0.362395  0.864775  0.845723
...                      ...   ...    ...  ...    ...        ...        ...  ...       ...       ...       ...       ...       ...       ...       ...
4126     95-u_n_51f-hlth.wav     u      f   51   hlth 222.222222 222.222222  ...  0.252393  0.344417  0.298158  0.500797  0.736506  0.685966  0.417393
4127    962-u_n_30f-hlth.wav     u      f   30   hlth 223.774250 222.222222  ...  0.316520  0.208523  0.332267  0.532395  0.701455  0.793471  0.230223
4128     97-u_n_21f-hlth.wav     u      f   21   hlth 242.760943 242.424242  ...  0.367020  0.428709  0.891417  0.819015  0.402244 -0.548269 -1.293228
4129     99-u_n_84f-hlth.wav     u      f   84   hlth 239.889087 242.424242  ...  0.257677  0.647518  0.500471  0.782175  0.399972 -0.234390 -0.587947
4130    990-u_n_29f-hlth.wav     u      f   29   hlth 233.949580 235.294118  ...  0.421929  0.605091  0.535736  0.667520  0.510483  0.192911 -0.264604
```



Welcome to VoxRimor!

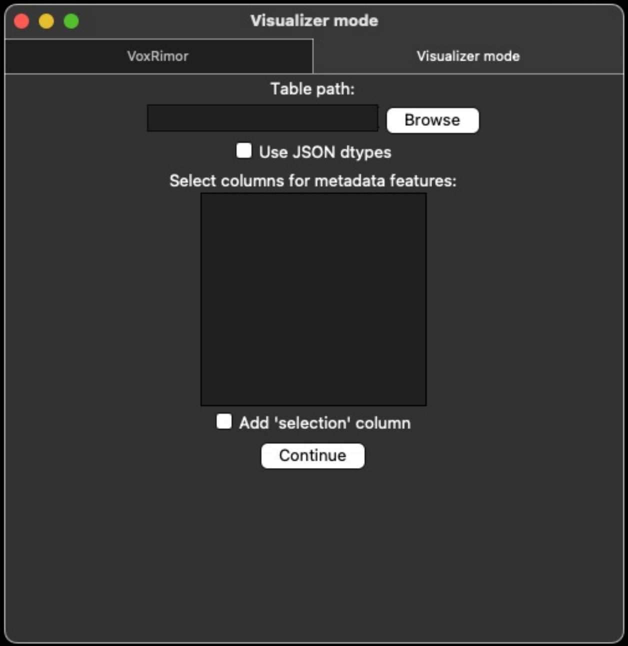


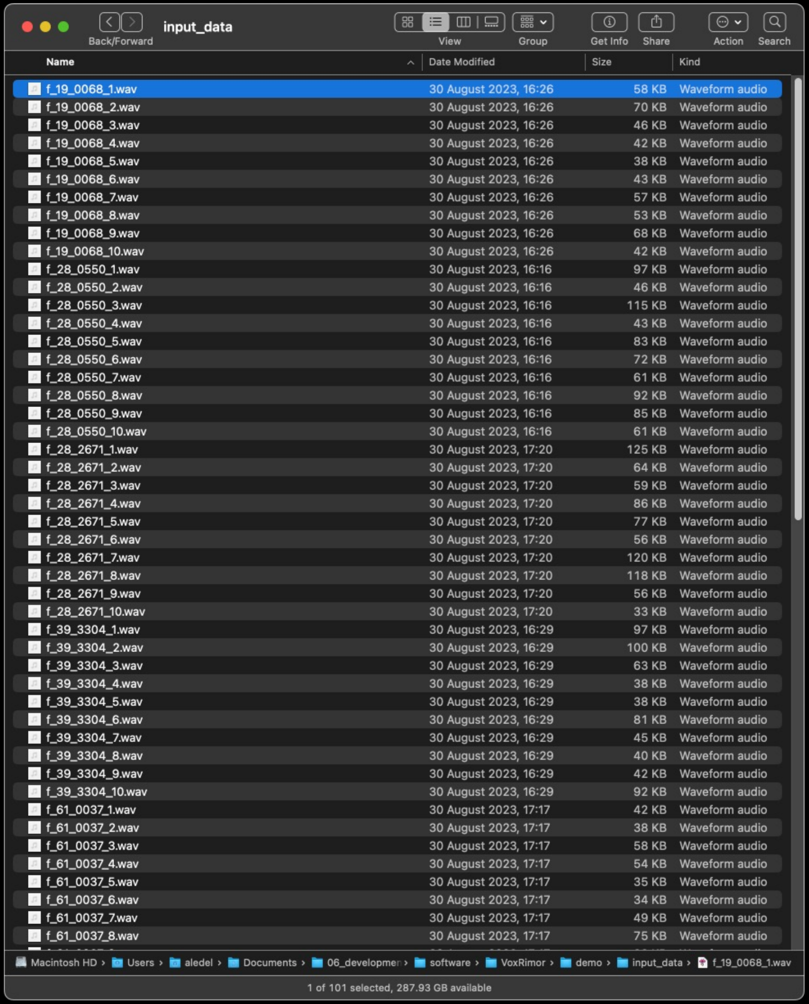
Table path:

☐ Use JSON dtypes

Select columns for metadata features:

☐ Add 'selection' column

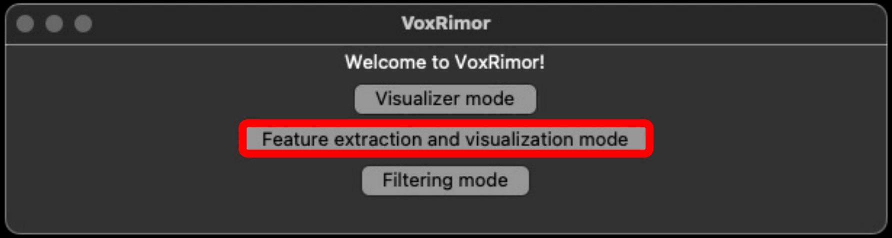
DEMO 2: Speaker embeddings mode



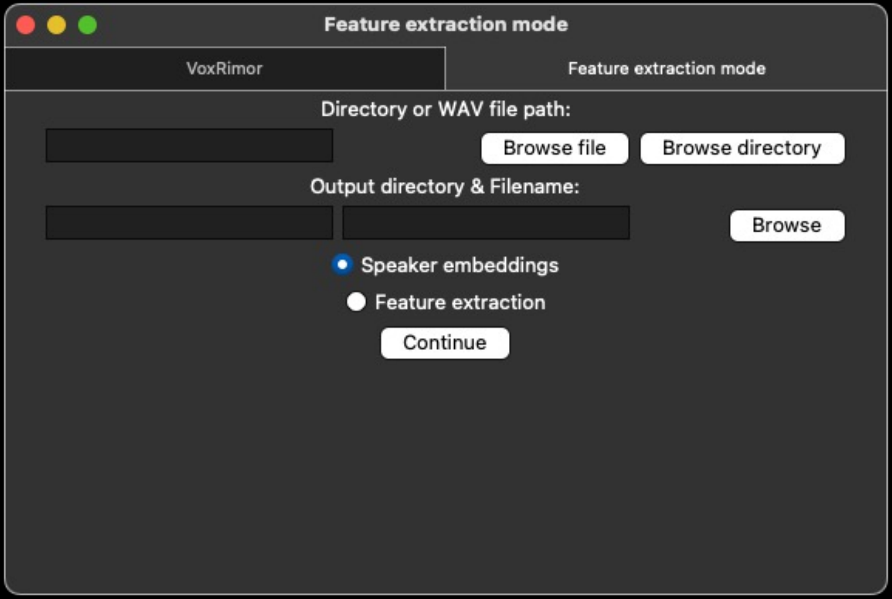
Name	Date Modified	Size	Kind
f_19_0068_1.wav	30 August 2023, 16:26	58 KB	Waveform audio
f_19_0068_2.wav	30 August 2023, 16:26	70 KB	Waveform audio
f_19_0068_3.wav	30 August 2023, 16:26	46 KB	Waveform audio
f_19_0068_4.wav	30 August 2023, 16:26	42 KB	Waveform audio
f_19_0068_5.wav	30 August 2023, 16:26	38 KB	Waveform audio
f_19_0068_6.wav	30 August 2023, 16:26	43 KB	Waveform audio
f_19_0068_7.wav	30 August 2023, 16:26	57 KB	Waveform audio
f_19_0068_8.wav	30 August 2023, 16:26	53 KB	Waveform audio
f_19_0068_9.wav	30 August 2023, 16:26	68 KB	Waveform audio
f_19_0068_10.wav	30 August 2023, 16:26	42 KB	Waveform audio
f_28_0550_1.wav	30 August 2023, 16:16	97 KB	Waveform audio
f_28_0550_2.wav	30 August 2023, 16:16	46 KB	Waveform audio
f_28_0550_3.wav	30 August 2023, 16:16	115 KB	Waveform audio
f_28_0550_4.wav	30 August 2023, 16:16	43 KB	Waveform audio
f_28_0550_5.wav	30 August 2023, 16:16	83 KB	Waveform audio
f_28_0550_6.wav	30 August 2023, 16:16	72 KB	Waveform audio
f_28_0550_7.wav	30 August 2023, 16:16	61 KB	Waveform audio
f_28_0550_8.wav	30 August 2023, 16:16	92 KB	Waveform audio
f_28_0550_9.wav	30 August 2023, 16:16	85 KB	Waveform audio
f_28_0550_10.wav	30 August 2023, 16:16	61 KB	Waveform audio
f_28_2671_1.wav	30 August 2023, 17:20	125 KB	Waveform audio
f_28_2671_2.wav	30 August 2023, 17:20	64 KB	Waveform audio
f_28_2671_3.wav	30 August 2023, 17:20	59 KB	Waveform audio
f_28_2671_4.wav	30 August 2023, 17:20	86 KB	Waveform audio
f_28_2671_5.wav	30 August 2023, 17:20	77 KB	Waveform audio
f_28_2671_6.wav	30 August 2023, 17:20	56 KB	Waveform audio
f_28_2671_7.wav	30 August 2023, 17:20	120 KB	Waveform audio
f_28_2671_8.wav	30 August 2023, 17:20	118 KB	Waveform audio
f_28_2671_9.wav	30 August 2023, 17:20	56 KB	Waveform audio
f_28_2671_10.wav	30 August 2023, 17:20	33 KB	Waveform audio
f_39_3304_1.wav	30 August 2023, 16:29	97 KB	Waveform audio
f_39_3304_2.wav	30 August 2023, 16:29	100 KB	Waveform audio
f_39_3304_3.wav	30 August 2023, 16:29	63 KB	Waveform audio
f_39_3304_4.wav	30 August 2023, 16:29	38 KB	Waveform audio
f_39_3304_5.wav	30 August 2023, 16:29	38 KB	Waveform audio
f_39_3304_6.wav	30 August 2023, 16:29	81 KB	Waveform audio
f_39_3304_7.wav	30 August 2023, 16:29	45 KB	Waveform audio
f_39_3304_8.wav	30 August 2023, 16:29	40 KB	Waveform audio
f_39_3304_9.wav	30 August 2023, 16:29	42 KB	Waveform audio
f_39_3304_10.wav	30 August 2023, 16:29	92 KB	Waveform audio
f_61_0037_1.wav	30 August 2023, 17:17	42 KB	Waveform audio
f_61_0037_2.wav	30 August 2023, 17:17	38 KB	Waveform audio
f_61_0037_3.wav	30 August 2023, 17:17	58 KB	Waveform audio
f_61_0037_4.wav	30 August 2023, 17:17	54 KB	Waveform audio
f_61_0037_5.wav	30 August 2023, 17:17	35 KB	Waveform audio
f_61_0037_6.wav	30 August 2023, 17:17	34 KB	Waveform audio
f_61_0037_7.wav	30 August 2023, 17:17	49 KB	Waveform audio
f_61_0037_8.wav	30 August 2023, 17:17	75 KB	Waveform audio

Macintosh HD › Users › aledel › Documents › 06_developmen › software › VoxRimor › demo › input_data › f_19_0068_1.wav

1 of 101 selected, 287.93 GB available



Welcome to VoxRimor!



Directory or WAV file path:

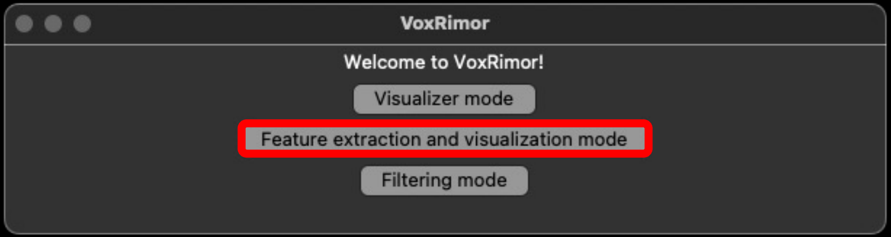
Output directory & Filename:

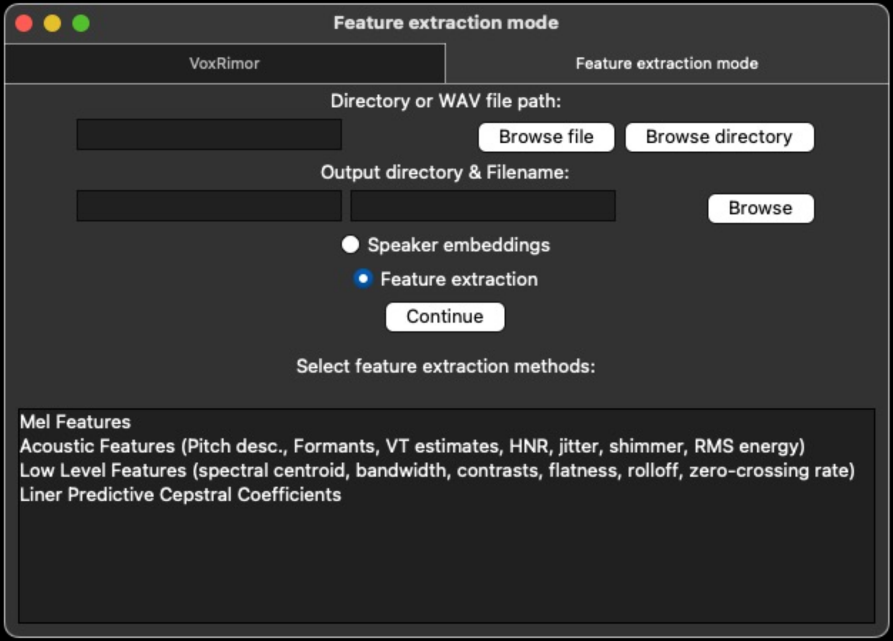
● Speaker embeddings

○ Feature extraction

Feature extraction mode

- Mel-features
- Acoustic features
- Low-level features
- Linear predictive cepstral coefficients



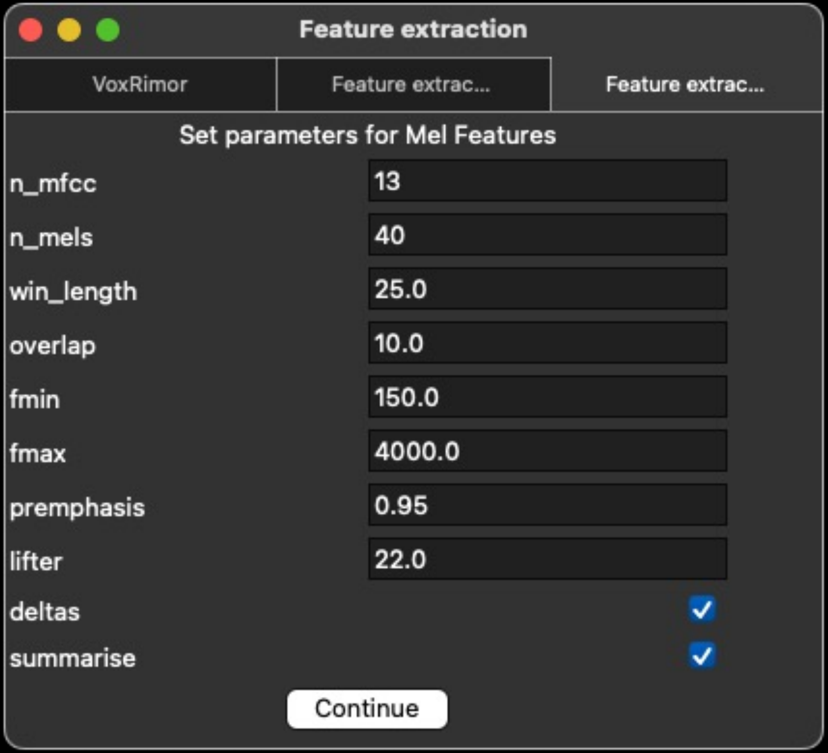


○ Speaker embeddings
● Feature extraction

Mel Features
Acoustic Features (Pitch desc., Formants, VT estimates, HNR, jitter, shimmer, RMS energy)
Low Level Features (spectral centroid, bandwidth, contrasts, flatness, rolloff, zero-crossing rate)
Liner Predictive Cepstral Coefficients

Feature extraction mode: Mel-features

- MFCCs: coefficients representing the short-term power spectrum of an audio signal computed using Mel-frequency scaling and often used in automatic speech processing tasks.
- Delta features:
  - delta: first derivative of the MFCCs
  - delta-delta: second derivative of the MFCCs

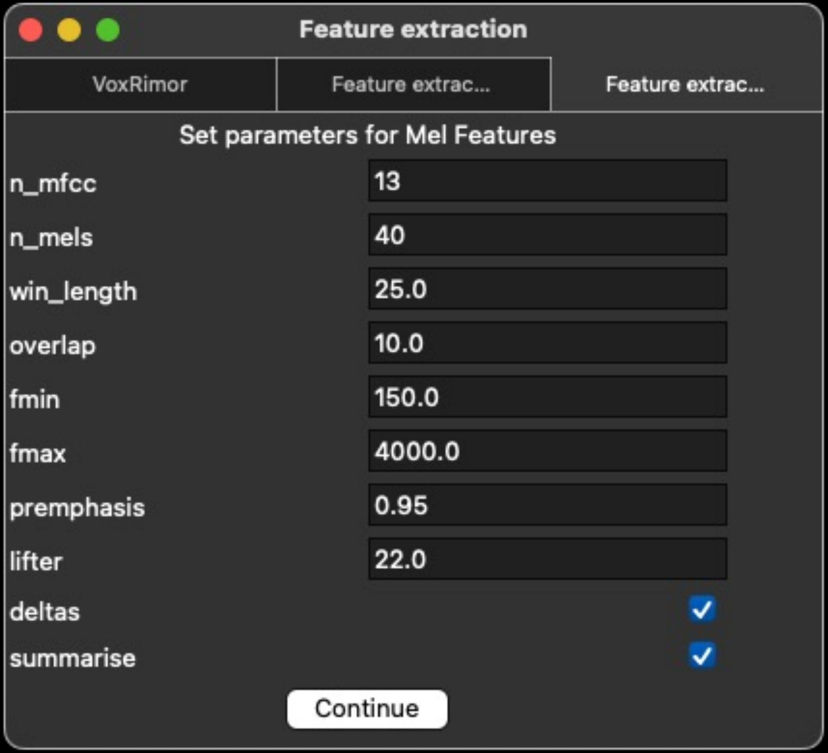


Set parameters for Mel Features

n_mfcc	13
n_mels	40
win_length	25.0
overlap	10.0
fmin	150.0
fmax	4000.0
premphasis	0.95
lifter	22.0
deltas	☑
summarise	☑

Feature extraction mode: Mel-features

- n_mfcc: number of coefficients
- n_mels: number of Mel bands
- win_length: length of analysis frame (ms)
- overlap: length of overlap between successive frames (ms)
- fmin: lowest frequency (Hz)
- fmax: highest frequency (Hz)
- premphasis: pre-emphasis coefficient
- lifter: cepstral filtering coefficient
- deltas: compute also delta and delta-delta features?
- summarise: summarise (μ and σ) of each feature at the utterance (file) level

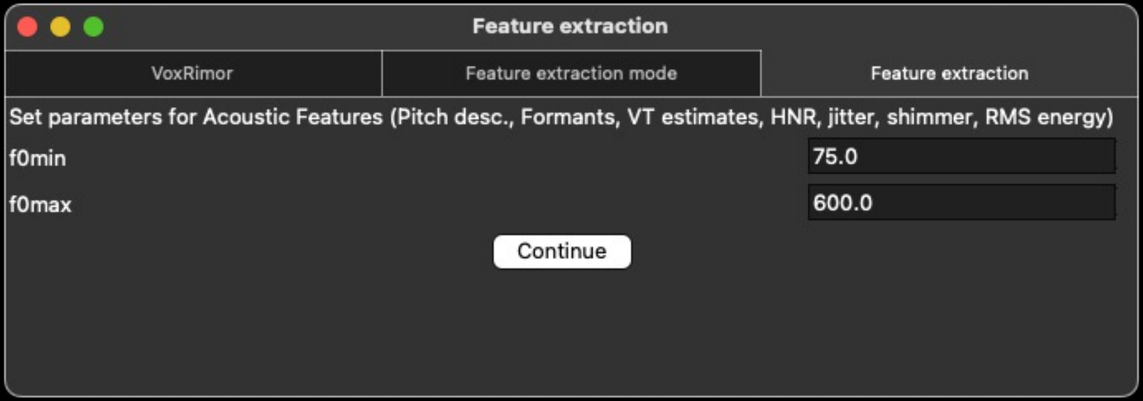


Set parameters for Mel Features

n_mfcc	13
n_mels	40
win_length	25.0
overlap	10.0
fmin	150.0
fmax	4000.0
premphasis	0.95
lifter	22.0
deltas	☑
summarise	☑

Feature extraction mode: Acoustic features

- Pitch: mean, median, minimum, maximum, std. dev.
- Formants: F1, F2, F3, F4
- VT estimates: formant dispersion, avg. formant, formants’ geometric mean, Fitch VTL, VTL Δf
- HNR
- Jitter & Shimmer
- RMS energy



Set parameters for Acoustic Features (Pitch desc., Formants, VT estimates, HNR, jitter, shimmer, RMS energy)

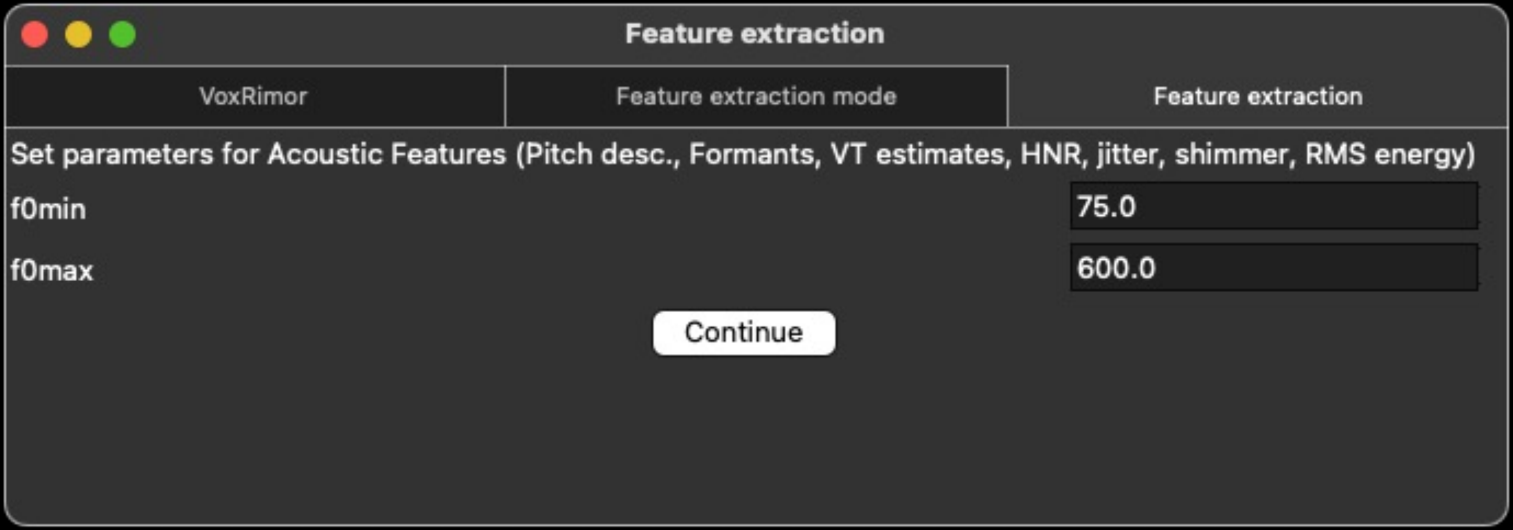
f0min 75.0

f0max 600.0

Feature extraction mode: Acoustic features

- f0min: pitch floor (Hz)
- f0max: pitch ceiling (Hz)

Uses Praat (Sound: To Pitch (cc)); Soon more parameters to come…



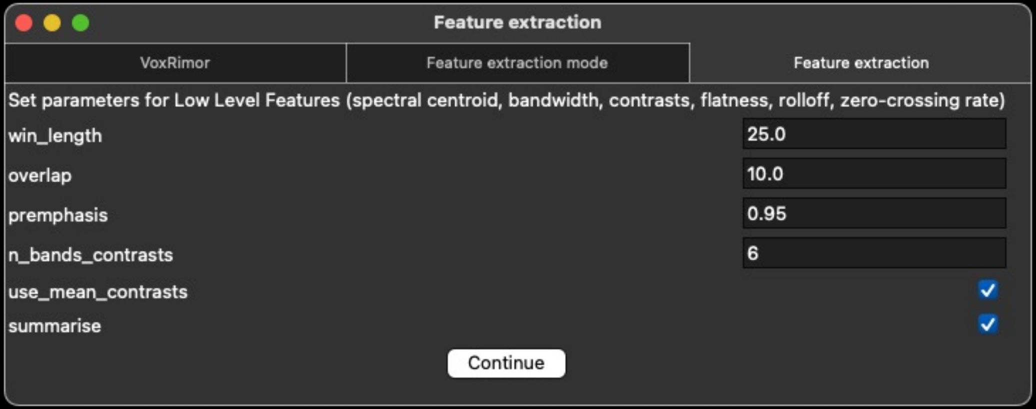
Set parameters for Acoustic Features (Pitch desc., Formants, VT estimates, HNR, jitter, shimmer, RMS energy)

f0min 75.0

f0max 600.0

Feature extraction mode: Low-level features

- spectral centroid: “centrer of mass” of avg. frequency in frame
- spectral bandwidth
- spectral contrast: mean energy contrast between peaks and valleys in each frequency sub-band
- spectral flatness: how “noise-like” vs. “tone-like” is the sound?
- spectral roll-off: roll-off frequency (freq. below which 85% of the energy is contained)
- zero-crossing rate

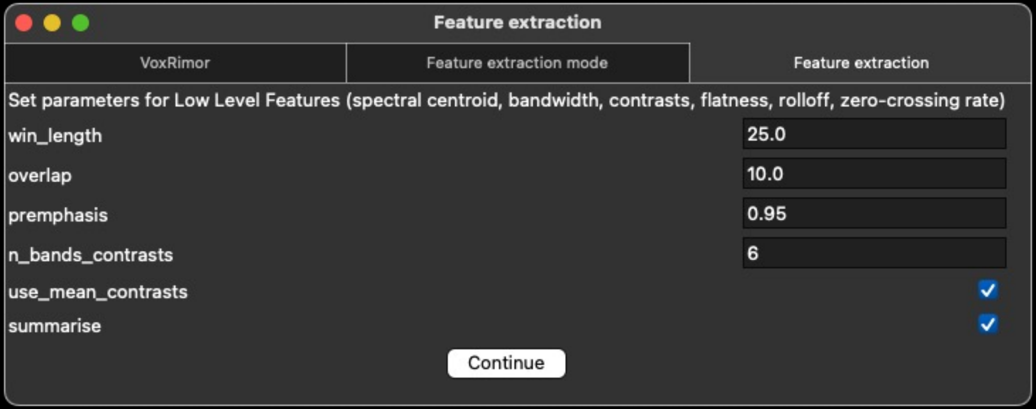


Set parameters for Low Level Features (spectral centroid, bandwidth, contrasts, flatness, rolloff, zero-crossing rate)

win_length	25.0
overlap	10.0
premphasis	0.95
n_bands_contrasts	6
use_mean_contrasts	☑
summarise	☑

Feature extraction mode: Low-level features

- win_length: length of analysis frame (ms)
- overlap: length of overlap between successive frames (ms)
- premphasis: pre-emphasis coefficient
- n_bands_contrasts: number of frequency bands by which to divide the spectrum to calculate contrasts (num. contrasts = n_bands+1)
- use_mean_contrasts: calculates the average contrast
- summarise: summarise (μ and σ) of each feature at the utterance (file) level

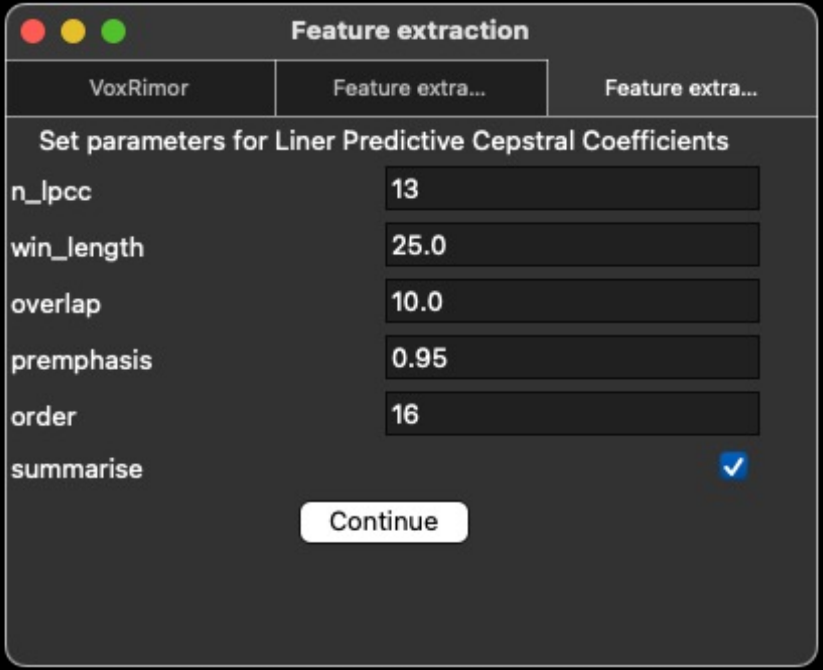


Set parameters for Low Level Features (spectral centroid, bandwidth, contrasts, flatness, rolloff, zero-crossing rate)

win_length	25.0
overlap	10.0
premphasis	0.95
n_bands_contrasts	6
use_mean_contrasts	☑
summarise	☑

Feature extraction mode: LPC features

- LPC: linear prediction coefficients (Burg); a representation of the spectral envelope after applying a linear filter
- LPCCs: linear predictive cepstral coefficients; coefficients representing the cepstrum (log-power spectrum) of the linear prediction coefficients

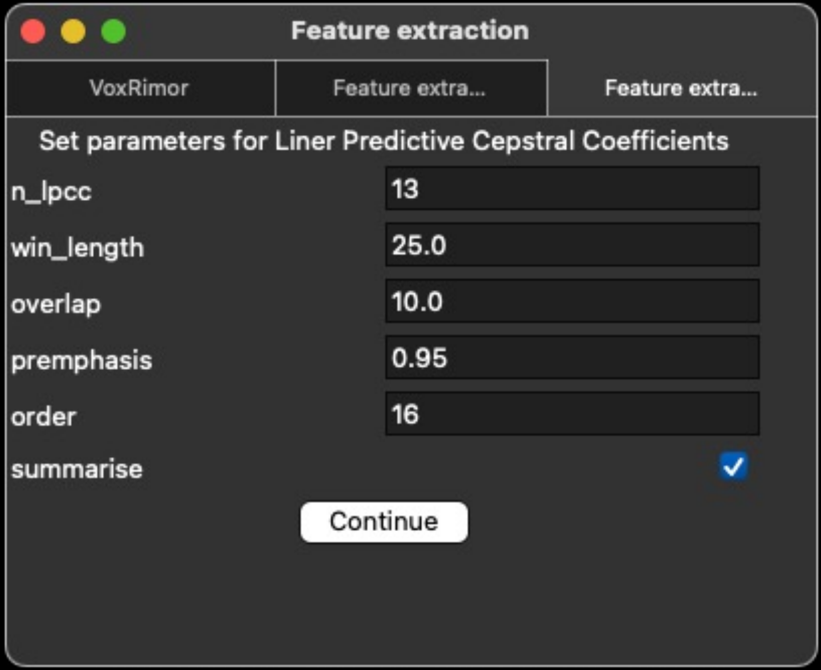


Set parameters for Liner Predictive Cepstral Coefficients

n_lpcc	13
win_length	25.0
overlap	10.0
premphasis	0.95
order	16
summarise	☑

Feature extraction mode: LPC features

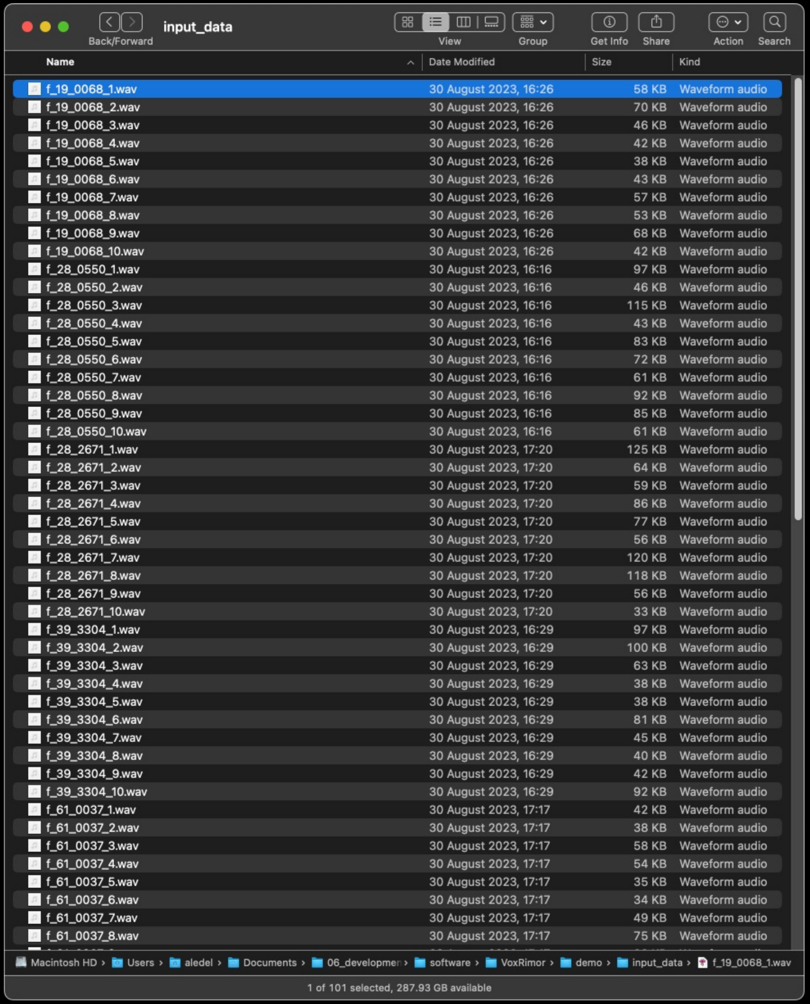
- n_lpcc: number of coefficients
- win_length: length of analysis frame (ms)
- overlap: length of overlap between successive frames (ms)
- premphasis: pre-emphasis coefficient
- order: order of the linear filter
- summarise: summarise (μ and σ) of each feature at the utterance (file) level



Set parameters for Liner Predictive Cepstral Coefficients

n_lpcc	13
win_length	25.0
overlap	10.0
premphasis	0.95
order	16
summarise	☑

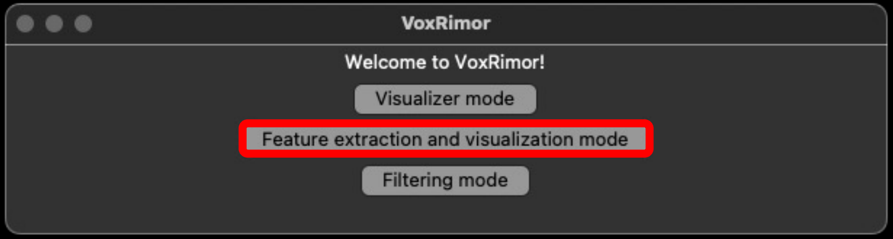
DEMO 3: Feature extraction mode



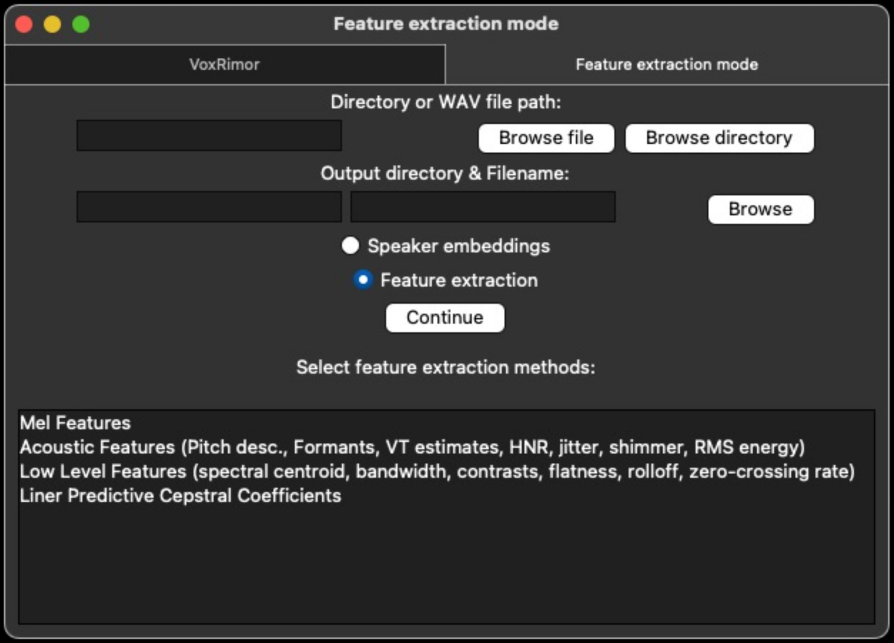
Name	Date Modified	Size	Kind
f_19_0068_1.wav	30 August 2023, 16:26	58 KB	Waveform audio
f_19_0068_2.wav	30 August 2023, 16:26	70 KB	Waveform audio
f_19_0068_3.wav	30 August 2023, 16:26	46 KB	Waveform audio
f_19_0068_4.wav	30 August 2023, 16:26	42 KB	Waveform audio
f_19_0068_5.wav	30 August 2023, 16:26	38 KB	Waveform audio
f_19_0068_6.wav	30 August 2023, 16:26	43 KB	Waveform audio
f_19_0068_7.wav	30 August 2023, 16:26	57 KB	Waveform audio
f_19_0068_8.wav	30 August 2023, 16:26	53 KB	Waveform audio
f_19_0068_9.wav	30 August 2023, 16:26	68 KB	Waveform audio
f_19_0068_10.wav	30 August 2023, 16:26	42 KB	Waveform audio
f_28_0550_1.wav	30 August 2023, 16:16	97 KB	Waveform audio
f_28_0550_2.wav	30 August 2023, 16:16	46 KB	Waveform audio
f_28_0550_3.wav	30 August 2023, 16:16	115 KB	Waveform audio
f_28_0550_4.wav	30 August 2023, 16:16	43 KB	Waveform audio
f_28_0550_5.wav	30 August 2023, 16:16	83 KB	Waveform audio
f_28_0550_6.wav	30 August 2023, 16:16	72 KB	Waveform audio
f_28_0550_7.wav	30 August 2023, 16:16	61 KB	Waveform audio
f_28_0550_8.wav	30 August 2023, 16:16	92 KB	Waveform audio
f_28_0550_9.wav	30 August 2023, 16:16	85 KB	Waveform audio
f_28_0550_10.wav	30 August 2023, 16:16	61 KB	Waveform audio
f_28_2671_1.wav	30 August 2023, 17:20	125 KB	Waveform audio
f_28_2671_2.wav	30 August 2023, 17:20	64 KB	Waveform audio
f_28_2671_3.wav	30 August 2023, 17:20	59 KB	Waveform audio
f_28_2671_4.wav	30 August 2023, 17:20	86 KB	Waveform audio
f_28_2671_5.wav	30 August 2023, 17:20	77 KB	Waveform audio
f_28_2671_6.wav	30 August 2023, 17:20	56 KB	Waveform audio
f_28_2671_7.wav	30 August 2023, 17:20	120 KB	Waveform audio
f_28_2671_8.wav	30 August 2023, 17:20	118 KB	Waveform audio
f_28_2671_9.wav	30 August 2023, 17:20	56 KB	Waveform audio
f_28_2671_10.wav	30 August 2023, 17:20	33 KB	Waveform audio
f_39_3304_1.wav	30 August 2023, 16:29	97 KB	Waveform audio
f_39_3304_2.wav	30 August 2023, 16:29	100 KB	Waveform audio
f_39_3304_3.wav	30 August 2023, 16:29	63 KB	Waveform audio
f_39_3304_4.wav	30 August 2023, 16:29	38 KB	Waveform audio
f_39_3304_5.wav	30 August 2023, 16:29	38 KB	Waveform audio
f_39_3304_6.wav	30 August 2023, 16:29	81 KB	Waveform audio
f_39_3304_7.wav	30 August 2023, 16:29	45 KB	Waveform audio
f_39_3304_8.wav	30 August 2023, 16:29	40 KB	Waveform audio
f_39_3304_9.wav	30 August 2023, 16:29	42 KB	Waveform audio
f_39_3304_10.wav	30 August 2023, 16:29	92 KB	Waveform audio
f_61_0037_1.wav	30 August 2023, 17:17	42 KB	Waveform audio
f_61_0037_2.wav	30 August 2023, 17:17	38 KB	Waveform audio
f_61_0037_3.wav	30 August 2023, 17:17	58 KB	Waveform audio
f_61_0037_4.wav	30 August 2023, 17:17	54 KB	Waveform audio
f_61_0037_5.wav	30 August 2023, 17:17	35 KB	Waveform audio
f_61_0037_6.wav	30 August 2023, 17:17	34 KB	Waveform audio
f_61_0037_7.wav	30 August 2023, 17:17	49 KB	Waveform audio
f_61_0037_8.wav	30 August 2023, 17:17	75 KB	Waveform audio

Macintosh HD › Users › aledel › Documents › 06_developmen › software › VoxRimor › demo › input_data › f_19_0068_1.wav

1 of 101 selected, 287.93 GB available



Welcome to VoxRimor!



Directory or WAV file path:

Output directory & Filename:

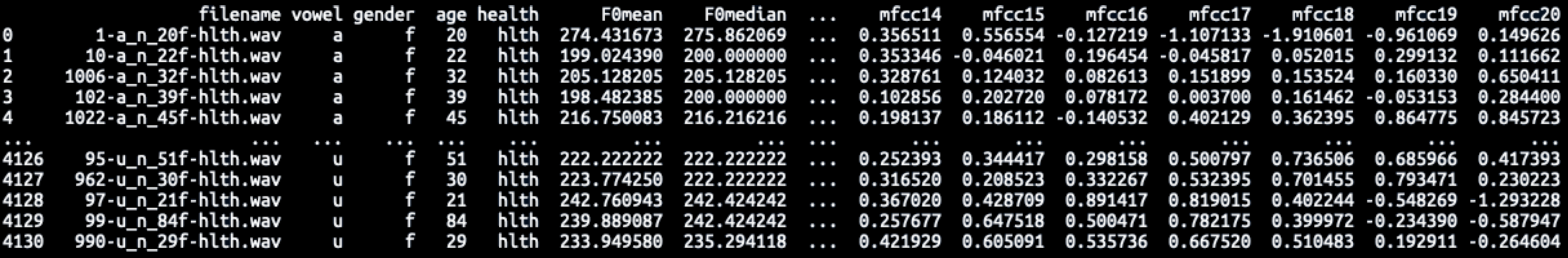
○ Speaker embeddings

● Feature extraction

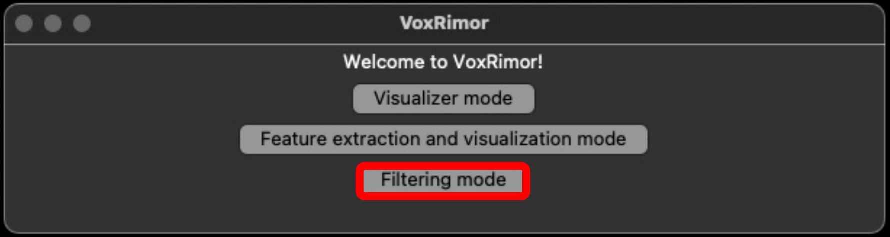
Select feature extraction methods:

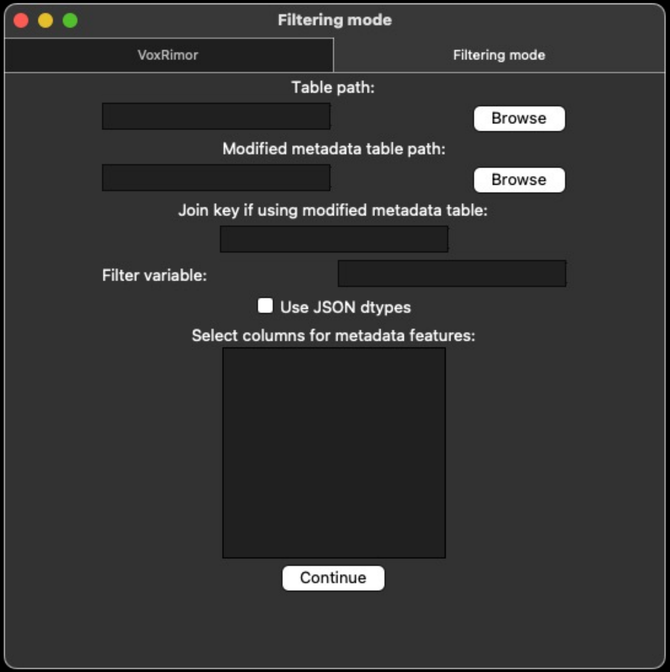
Mel Features
Acoustic Features (Pitch desc., Formants, VT estimates, HNR, jitter, shimmer, RMS energy)
Low Level Features (spectral centroid, bandwidth, contrasts, flatness, rolloff, zero-crossing rate)
Liner Predictive Cepstral Coefficients

DEMO 4: Filtering mode



```
                  filename vowel gender  age health     F0mean   F0median  ...    mfcc14    mfcc15    mfcc16    mfcc17    mfcc18    mfcc19    mfcc20
0       1-a_n_20f-hlth.wav     a      f   20   hlth 274.431673 275.862069  ...  0.356511  0.556554 -0.127219 -1.107133 -1.910601 -0.961069  0.149626
1      10-a_n_22f-hlth.wav     a      f   22   hlth 199.024390 200.000000  ...  0.353346 -0.046021  0.196454 -0.045817  0.052015  0.299132  0.111662
2    1006-a_n_32f-hlth.wav     a      f   32   hlth 205.128205 205.128205  ...  0.328761  0.124032  0.082613  0.151899  0.153524  0.160330  0.650411
3     102-a_n_39f-hlth.wav     a      f   39   hlth 198.482385 200.000000  ...  0.102856  0.202720  0.078172  0.003700  0.161462 -0.053153  0.284400
4    1022-a_n_45f-hlth.wav     a      f   45   hlth 216.750083 216.216216  ...  0.198137  0.186112 -0.140532  0.402129  0.362395  0.864775  0.845723
...                    ...   ...    ...  ...    ...        ...        ...  ...       ...       ...       ...       ...       ...       ...       ...
4126   95-u_n_51f-hlth.wav     u      f   51   hlth 222.222222 222.222222  ...  0.252393  0.344417  0.298158  0.500797  0.736506  0.685966  0.417393
4127  962-u_n_30f-hlth.wav     u      f   30   hlth 223.774250 222.222222  ...  0.316520  0.208523  0.332267  0.532395  0.701455  0.793471  0.230223
4128   97-u_n_21f-hlth.wav     u      f   21   hlth 242.760943 242.424242  ...  0.367020  0.428709  0.891417  0.819015  0.402244 -0.548269 -1.293228
4129   99-u_n_84f-hlth.wav     u      f   84   hlth 239.889087 242.424242  ...  0.257677  0.647518  0.500471  0.782175  0.399972 -0.234390 -0.587947
4130  990-u_n_29f-hlth.wav     u      f   29   hlth 233.949580 235.294118  ...  0.421929  0.605091  0.535736  0.667520  0.510483  0.192911 -0.264604
```





Filtering mode
VoxRimor
Filtering mode
Table path:
Browse
Modified metadata table path:
Browse
Join key if using modified metadata table:
Filter variable:
□ Use JSON dtypes
Select columns for metadata features:
Continue

Future plans, goals, and Q&A

- Praat TextGrid support.
- More parameter settings → better feature extraction.
- More features: e.g. VoiceSauce
- Integrated speaker verification system and classification tools (ECAPA-TDNN VoxCeleb2).

Fully-integrated web application and UI:

- *voxplorer* as an open-source online LiRI service.
- Direct selection & downloading of data.
- Downloadable reduced dimensions versions of data.
- Statistical tools: distances, box-plots, distributions.

voxplorer Thanks You!

Any questions?