Classification and clustering of ultrasound tongue images in vowel production

語言所碩二 盧妍蓁
語言所博三 翁益寧

Background

- Ultrasound is a popular tool in many areas of research

Background

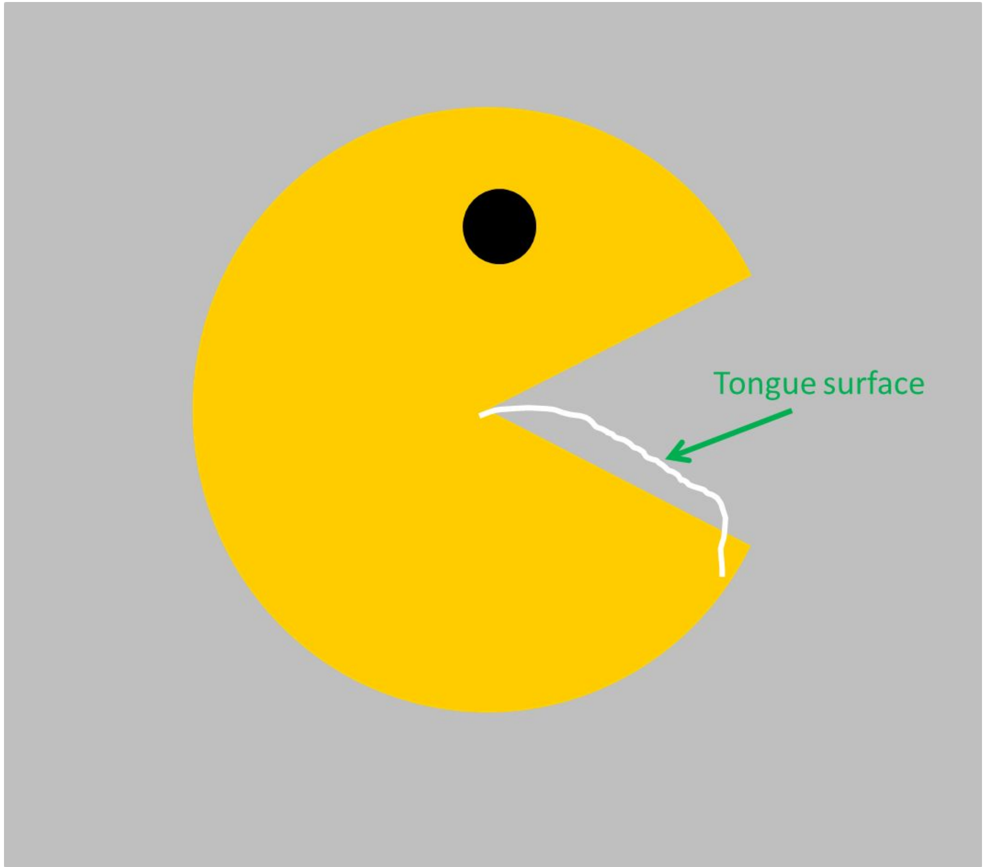
- Ultrasound is a popular tool in many areas of research
    - Tongue, larynx (vocal cords)
    - Articulatory phonetics and laboratory phonology
    - Speech pathology and therapy
    - Second language acquisition

Background

- Ultrasound is a popular tool in many areas of research
  - Tongue, larynx (vocal cords)
  - Articulatory phonetics and laboratory phonology
  - Speech pathology and therapy
  - Second language acquisition
- Pros of ultrasound
  - Non-invasive, safe, easy to set-up, accessible
  - Visible in real-time

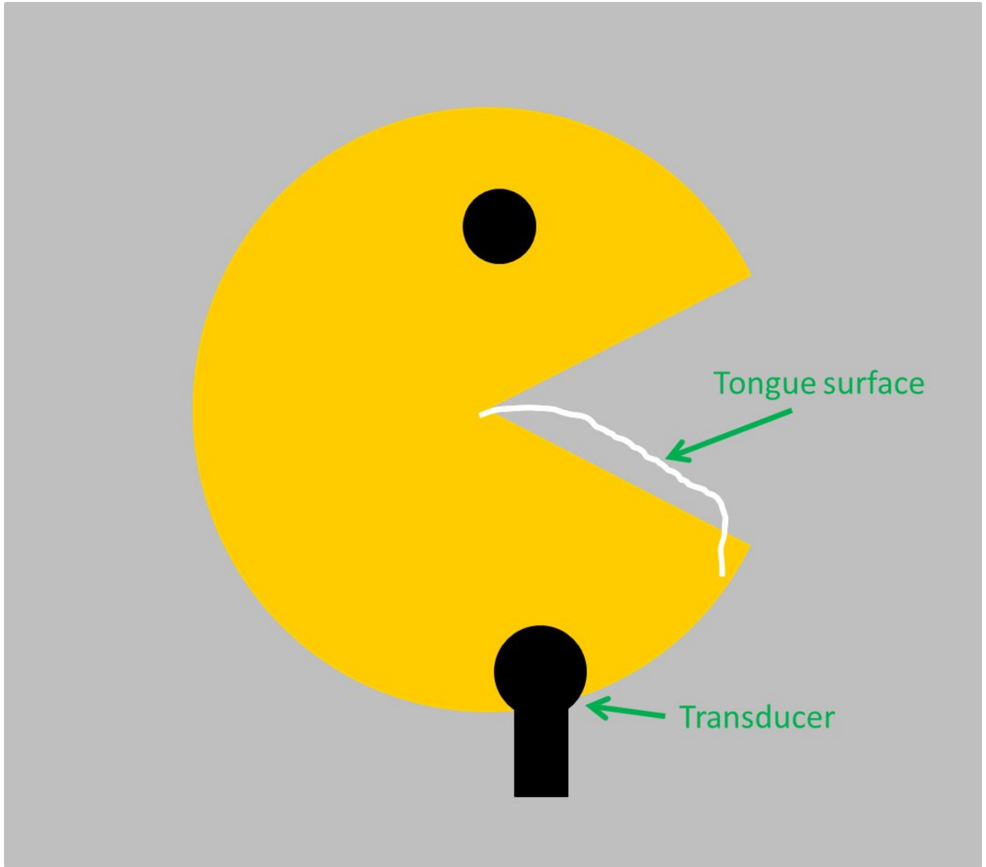
Background

- How ultrasound works



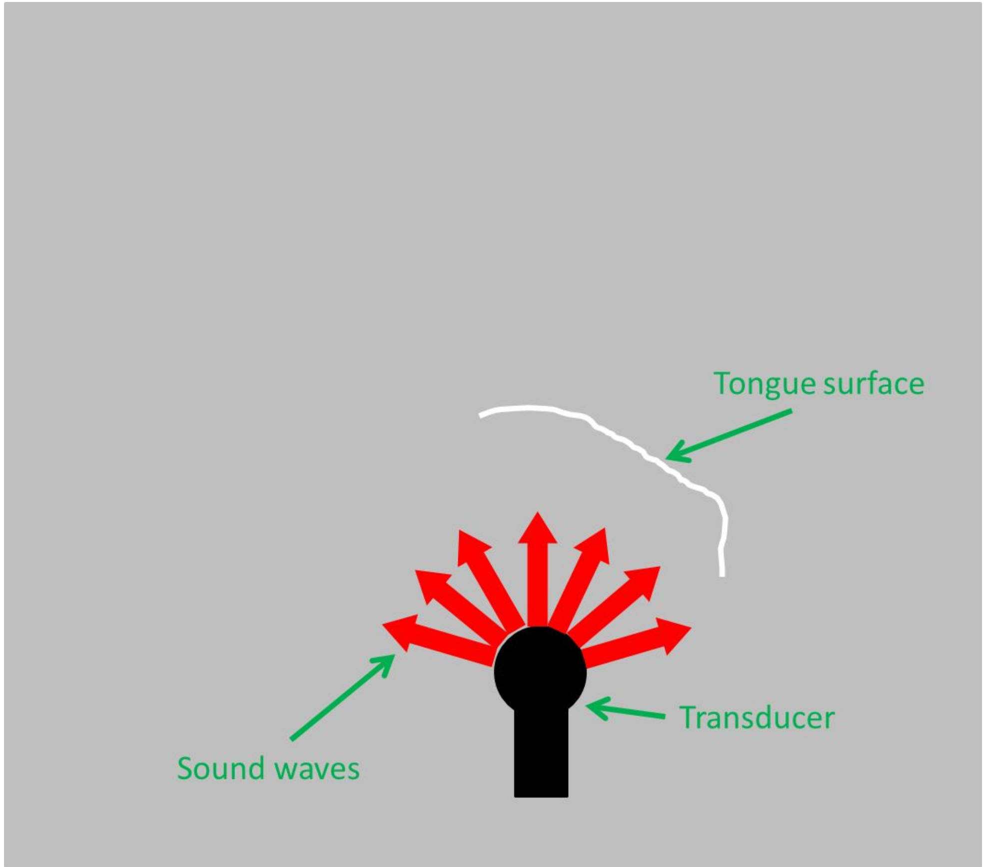
Background

- How ultrasound works



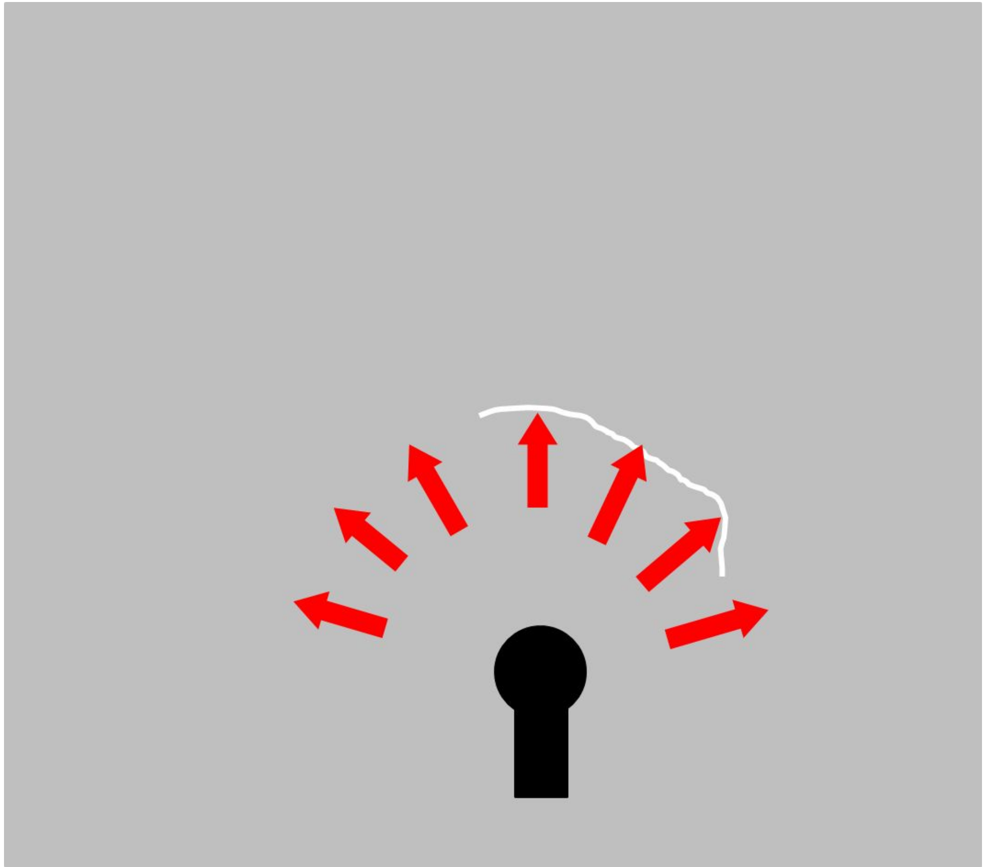
Background

- How ultrasound works



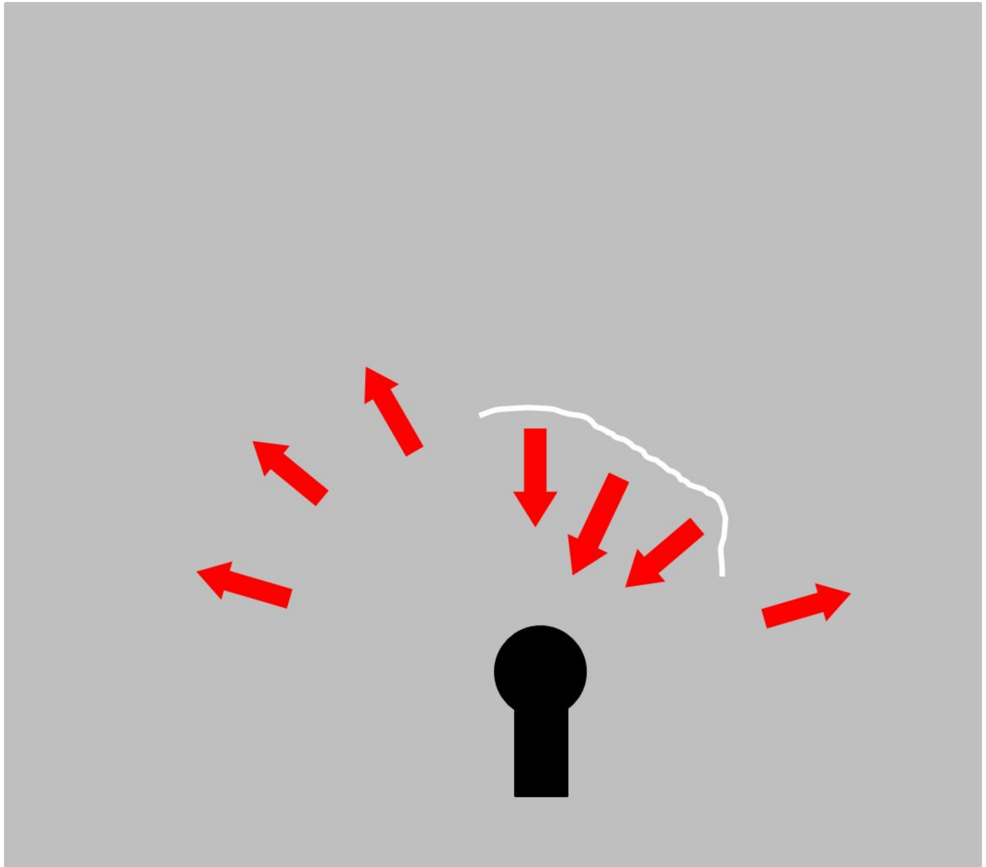
Background

- How ultrasound works



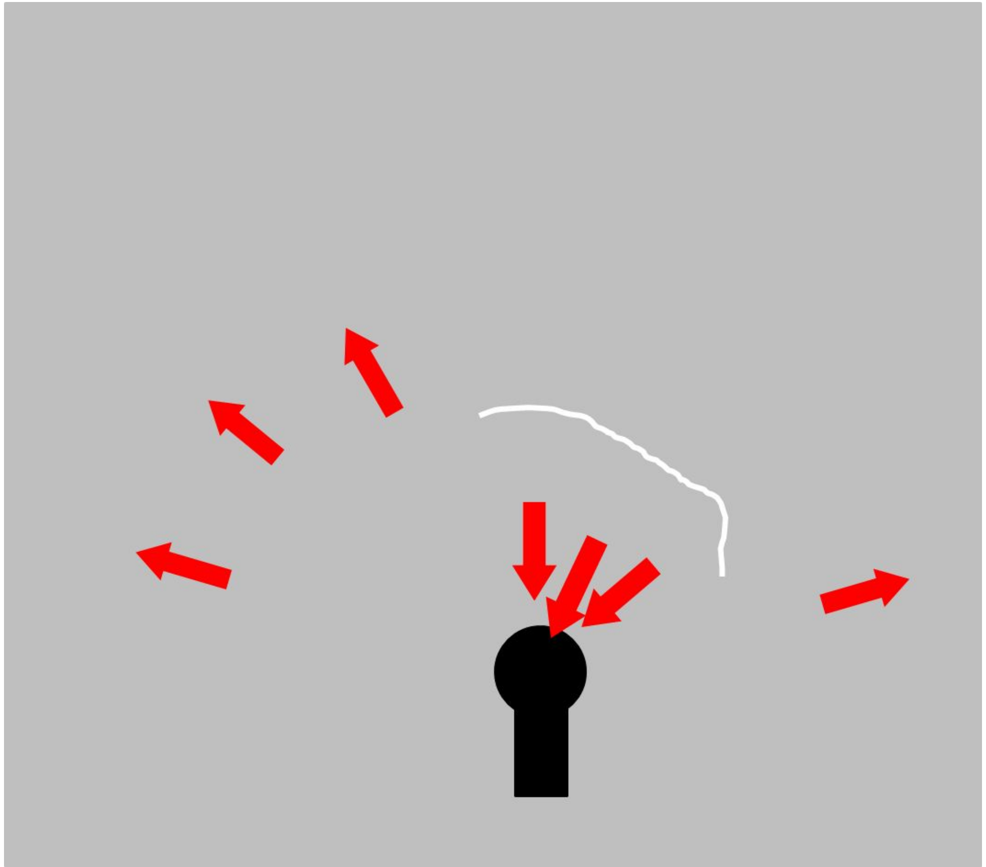
Background

- How ultrasound works



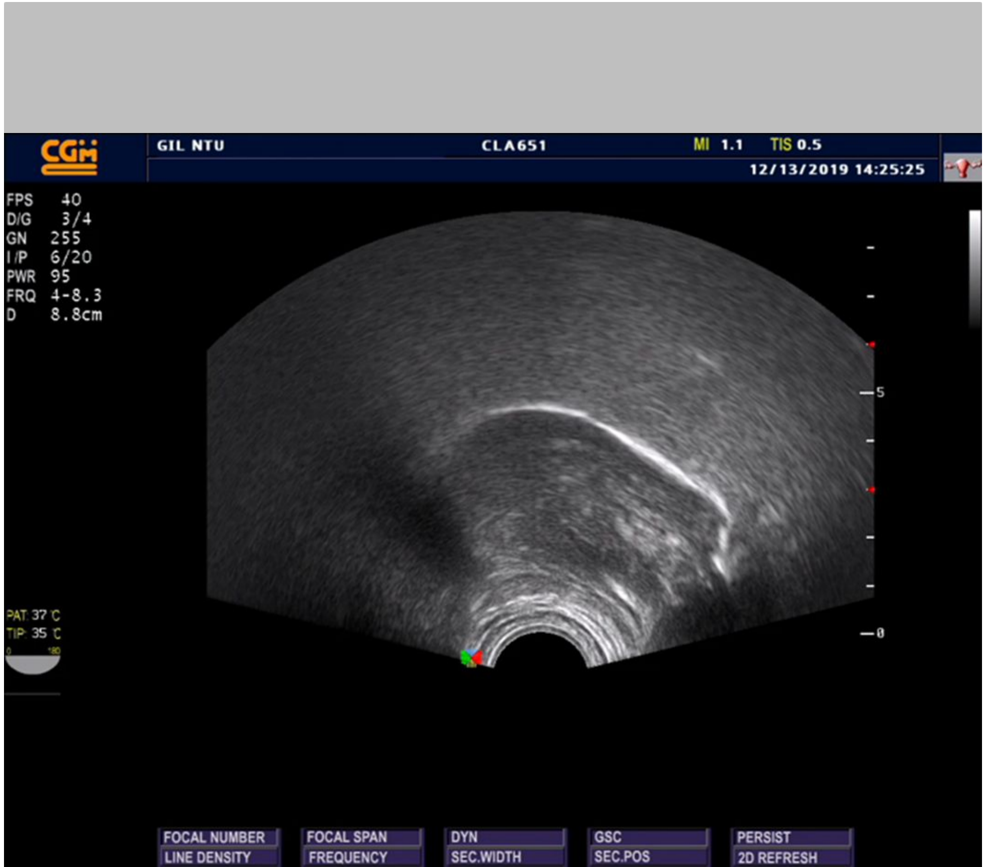
Background

- How ultrasound works



Background

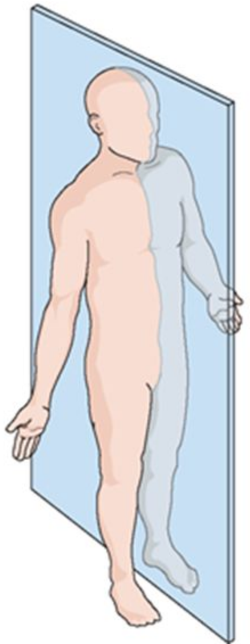
- How ultrasound works

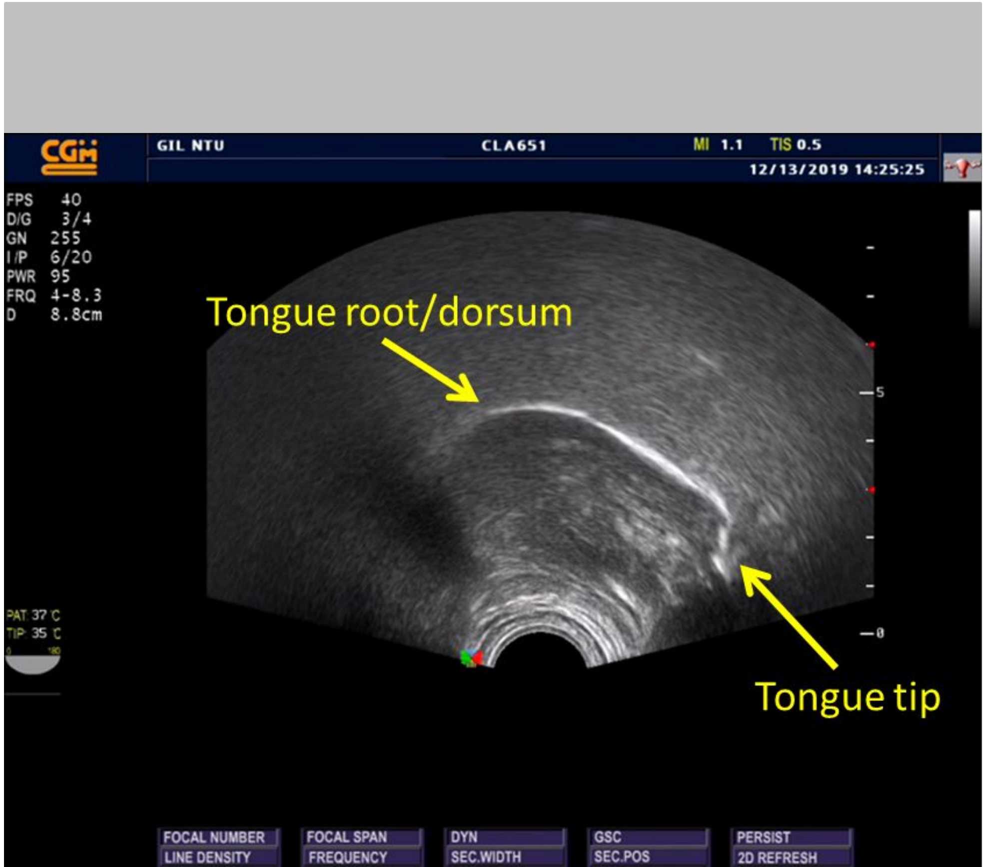


Background

- How ultrasound works

Sagittal plane



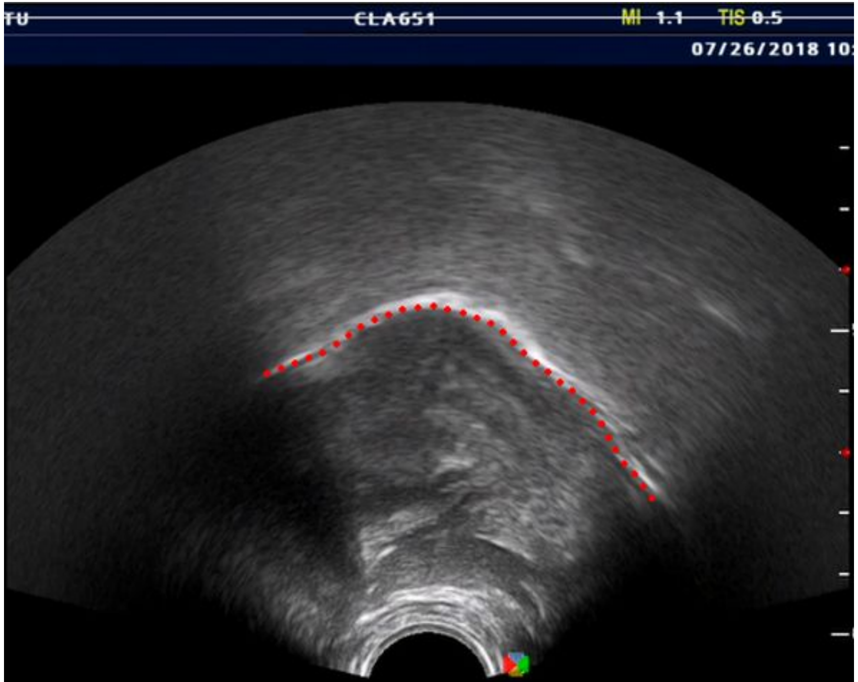


Background

- Traditional method of analyzing ultrasound data is slow and laborious
- Quantifying image data into coordinates (contour tracing)
    - Tongue contour → line (a series of points)
        - Each point described by coordinates (X, Y)

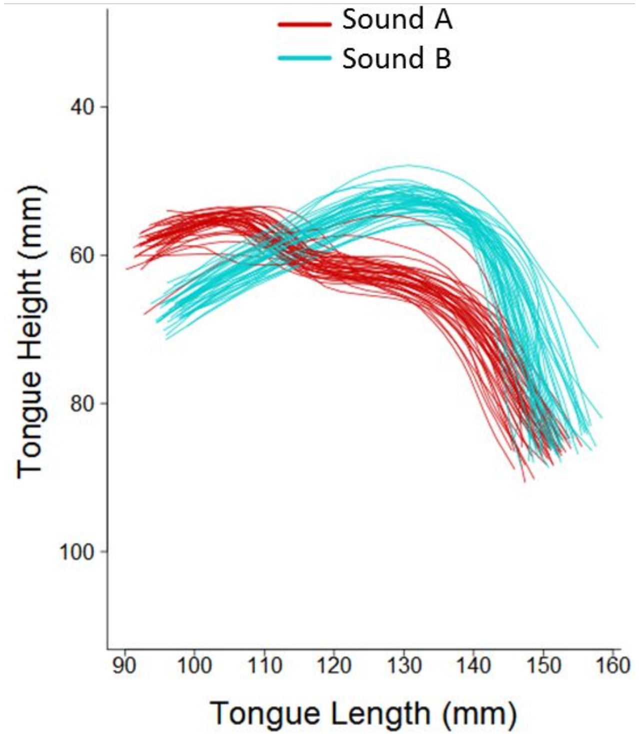
Background

- Traditional method of analyzing ultrasound data is slow and laborious
- Quantifying image data into coordinates (contour tracing)
    - Tongue contour → line (a series of points)
        - Each point described by coordinates (X, Y)



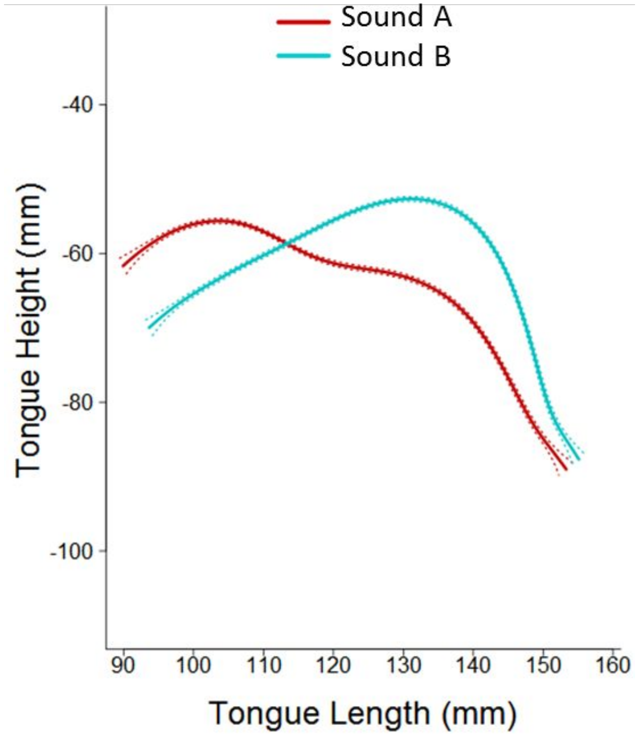
Background

- Traditional method of analyzing ultrasound data is slow and laborious
- Quantifying image data into coordinates (contour tracing)
  - Tongue contour → line (a series of points)
    - Each point described by coordinates (X, Y)
  - Statistical analysis
    - Smoothing Spline ANOVA
    - Generalized Additive Mixed Models



Background

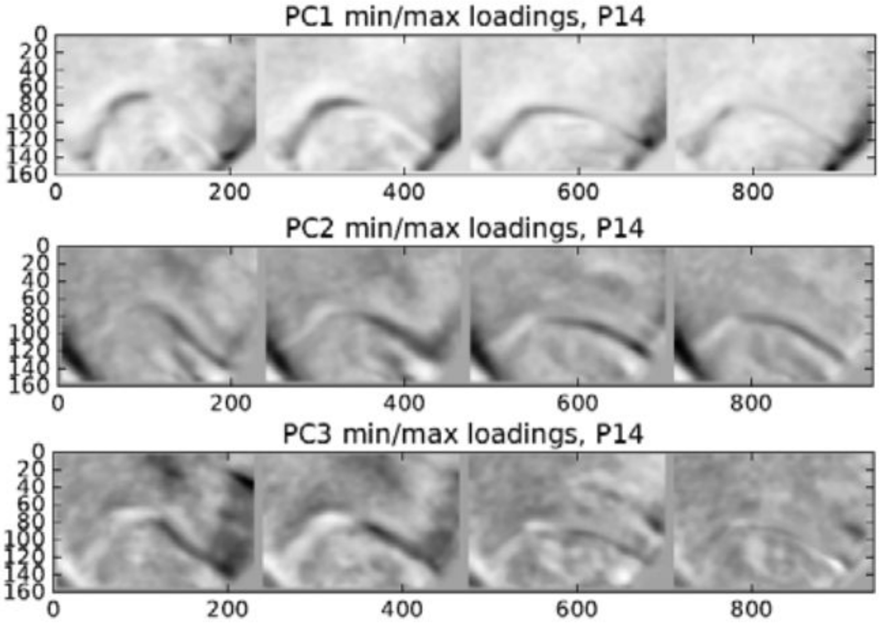
- Traditional method of analyzing ultrasound data is slow and laborious
- Quantifying image data into coordinates (contour tracing)
  - Tongue contour → line (a series of points)
    - Each point described by coordinates (X, Y)
  - Statistical analysis
    - Smoothing Spline ANOVA
    - Generalized Additive Mixed Models



Background

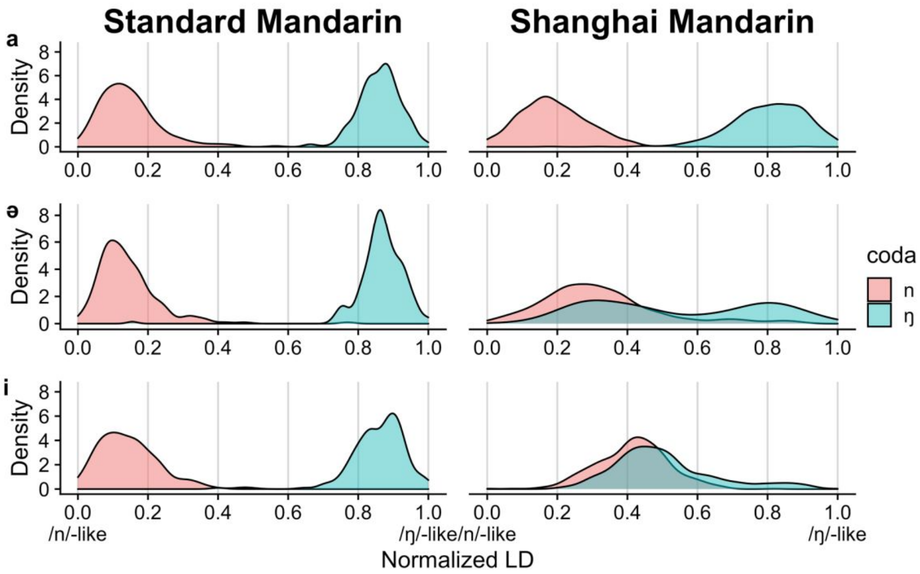
- Recent studies started to use raw-image-based analysis methods

Principal Component Analysis



Kochetov et al. (2019) Manner differences in the Punjabi dentalretroflex contrast: An ultrasound study of time-series data

Linear Discriminant Analysis



Faytak et al. (2020) Nasal coda neutralization in Shanghai Mandarin: Articulatory and perceptual evidence

Background

- Our goal:

  Try out raw-image-based methods for

  - Classification of vowels
    - Convolutional Autoencoder

  - Visualization of vowel clusters
    - PCA (Principal Component Analysis)
    - t-SNE (t-distributed Stochastic Neighbor embedding) (library install failed)
    - UMAP (Uniform Manifold Approximation and Projection)

Data collection

- Subjects
  - 2 native Mandarin speakers (1M, 1F)

Data collection

- Subjects
  - 2 native Mandarin speakers (1M, 1F)

- Materials
  - 6 vowels in Mandarin **/a i u e o ə/**
  - Each vowel pronounced 100 times in isolation
  - 6 vowels * 100 repetitions * 2 subjects = 1200 trials in total

Data collection

- Subjects
  - 2 native Mandarin speakers (1M, 1F)

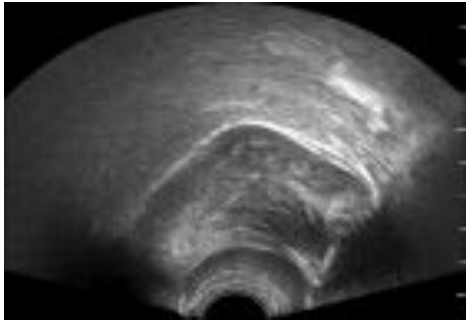
- Materials
  - 6 vowels in Mandarin **/a i u e o ə/**
  - Each vowel pronounced 100 times in isolation
  - 6 vowels * 100 repetitions * 2 subjects = 1200 trials in total

  - Images from the midpoint of each trial were extracted for analysis

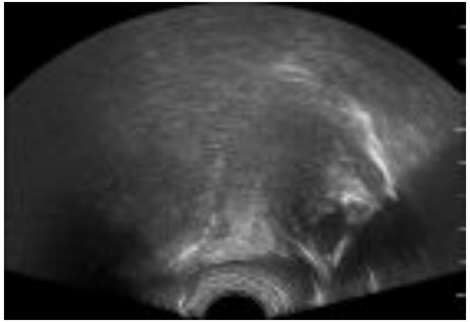
Data collection

- Speaker 1

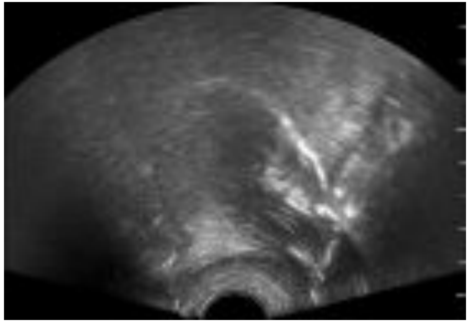
/a/



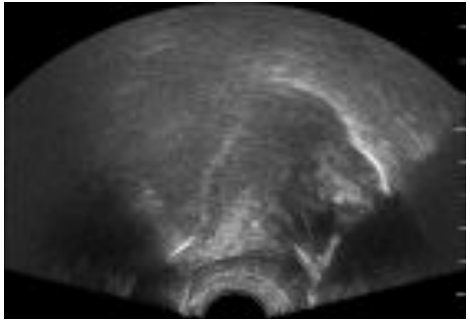
/i/



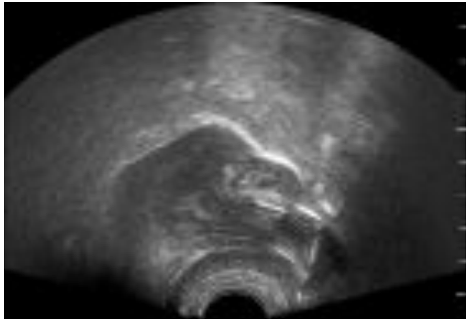
/u/



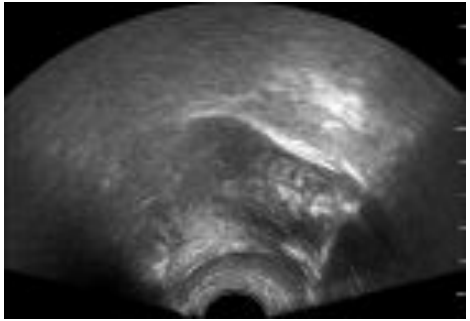
/e/



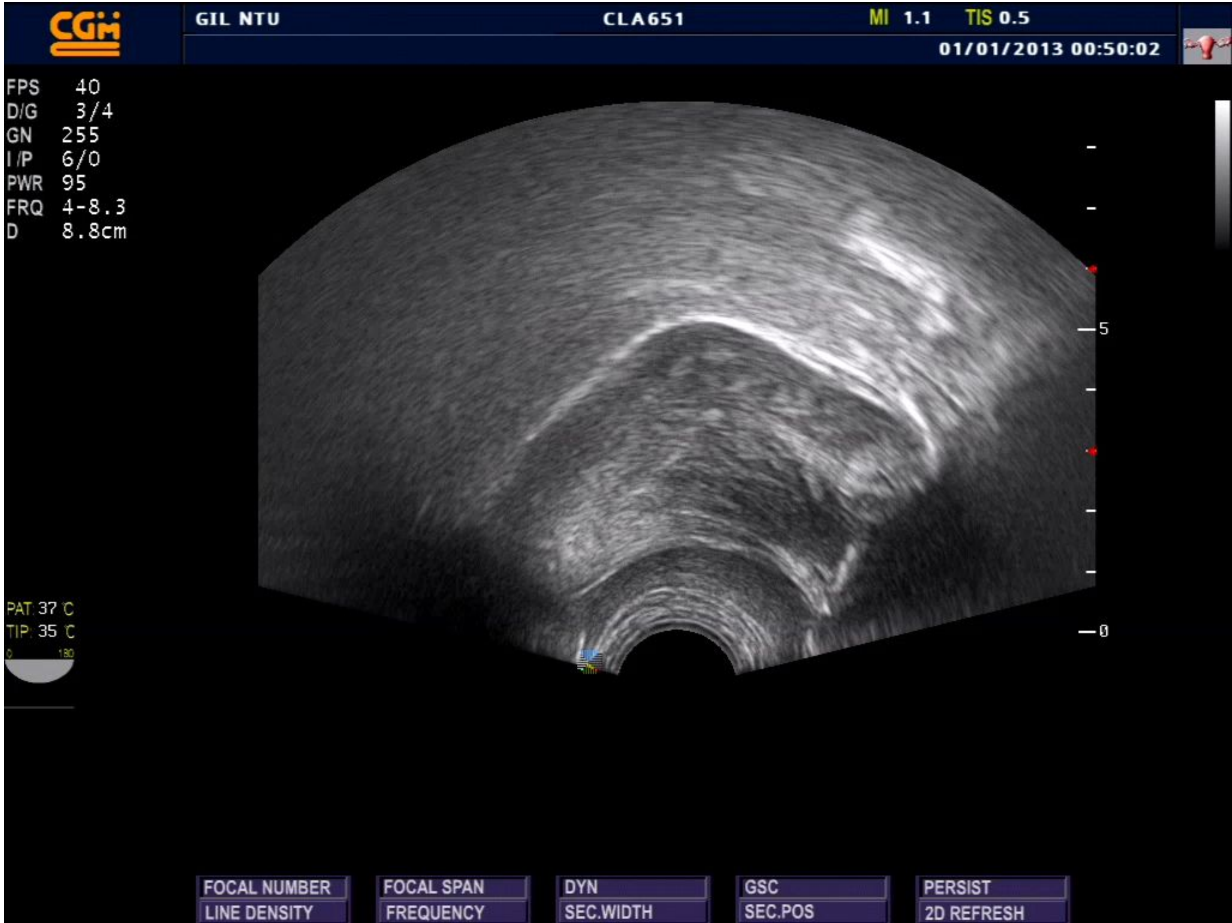
/o/



/ə/

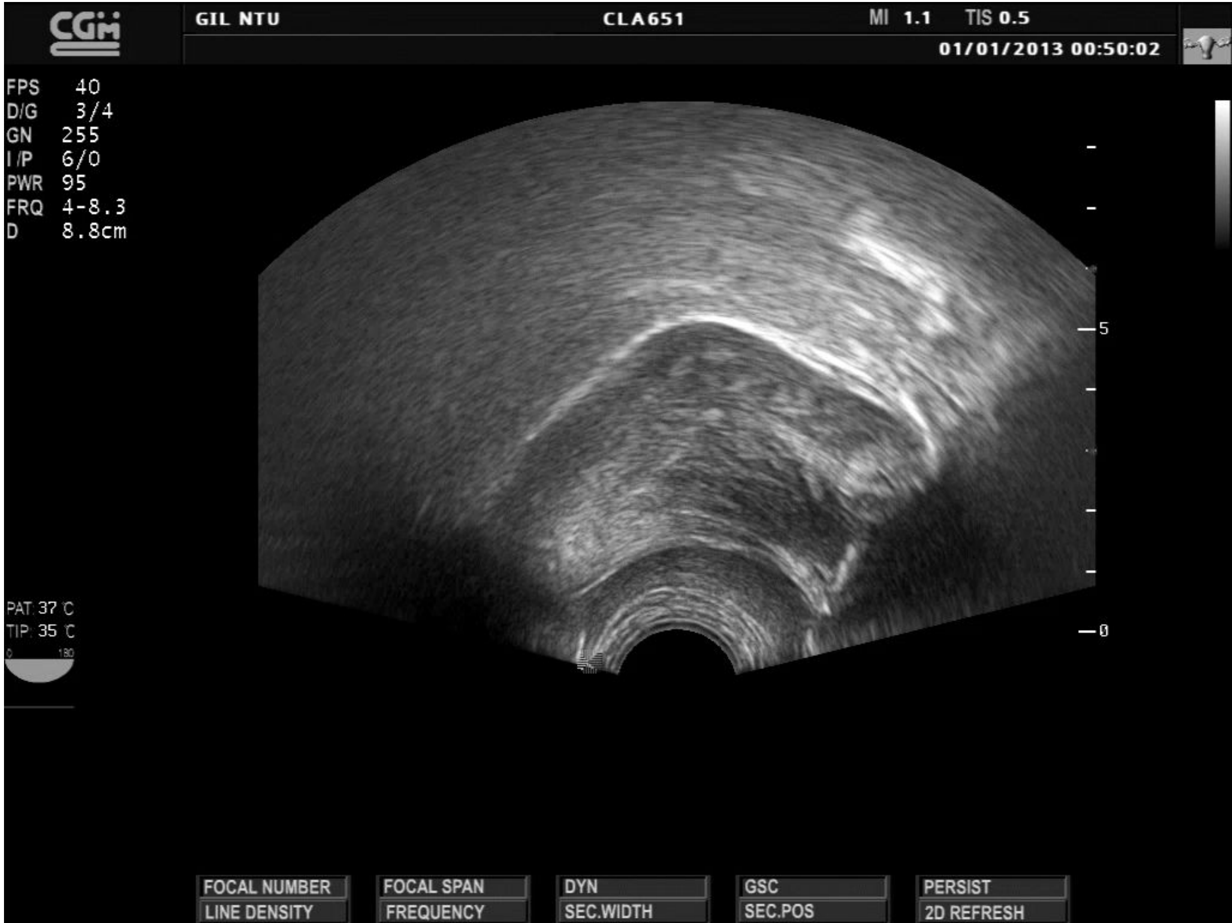


Data wrangling



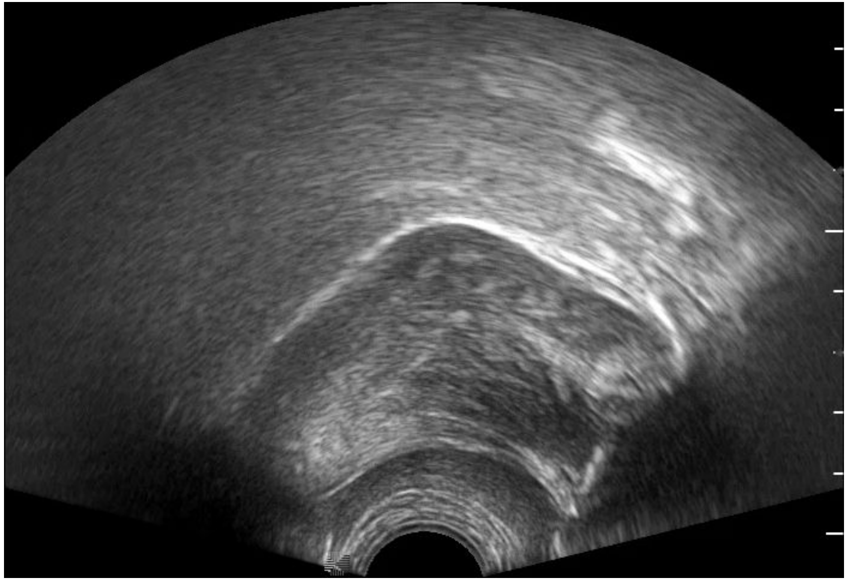
Data wrangling

- Convert to greyscale



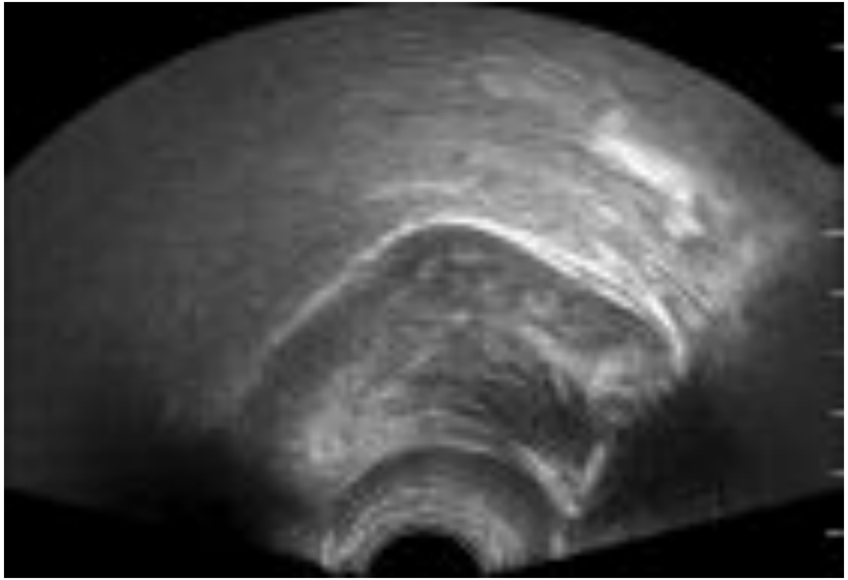
Data wrangling

- Convert to greyscale
- Cropping to meaningful area



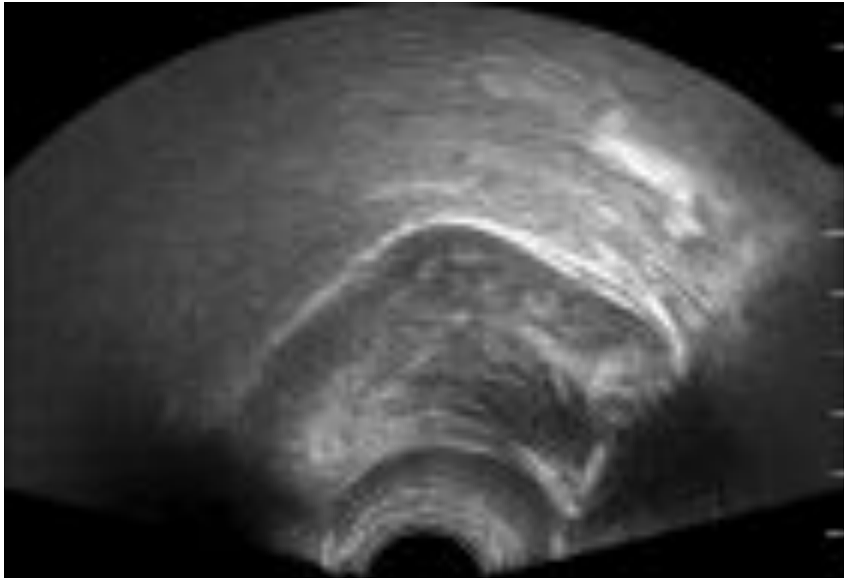
Data wrangling

- Convert to greyscale
- Cropping to meaningful area
- Downscaling to 96 (h) * 140 (w)



Data wrangling

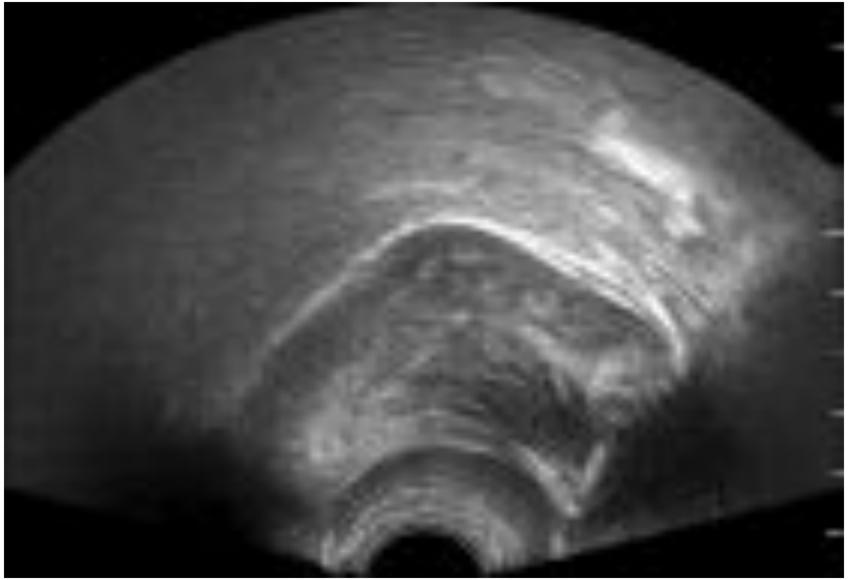
- Convert to greyscale
- Cropping to meaningful area
- Downscaling to 96 (h) * 140 (w)
- Flattening to (1 * 13440) vector
    - Only necessary for some analysis



	Pixel								
Image	P1	P2	P3	P4	P5	…	P13438	P13439	P13440
a001	0	0	34	157	255	…	27	0	0
a002	0	0	0	58	169	…	84	2	0

Data wrangling

- Convert to greyscale
- Cropping to meaningful area
- Downscaling to 96 (h) * 140 (w)
- Flattening to (1 * 13440) vector
    - Only necessary for some analysis
- Normalizing to [0, 1]



	Pixel								
Image	P1	P2	P3	P4	P5	…	P13438	P13439	P13440
a001	0.000	0.000	0.133	0.616	1.000	…	0.106	0.000	0.000
a002	0.000	0.000	0.000	0.227	0.663	…	0.329	0.008	0.000

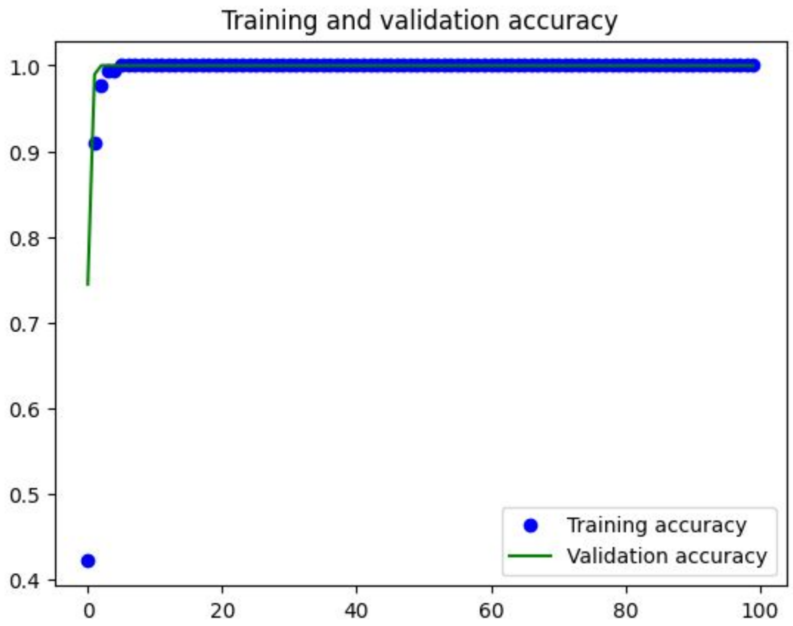
Results: classification

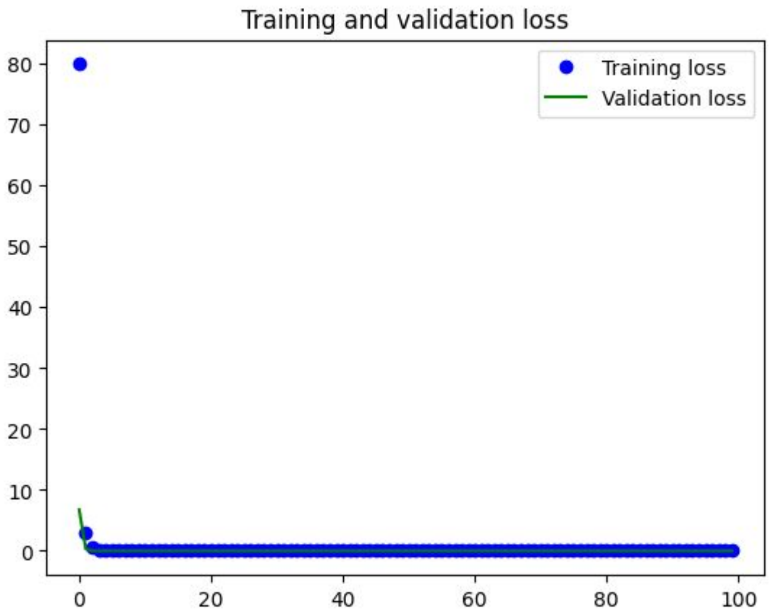
- Total 1200 images

- Training set: 960 images
  - 80% train
  - 20% validation
  - 200 total epochs

- Test set: 240 images

Results: classification



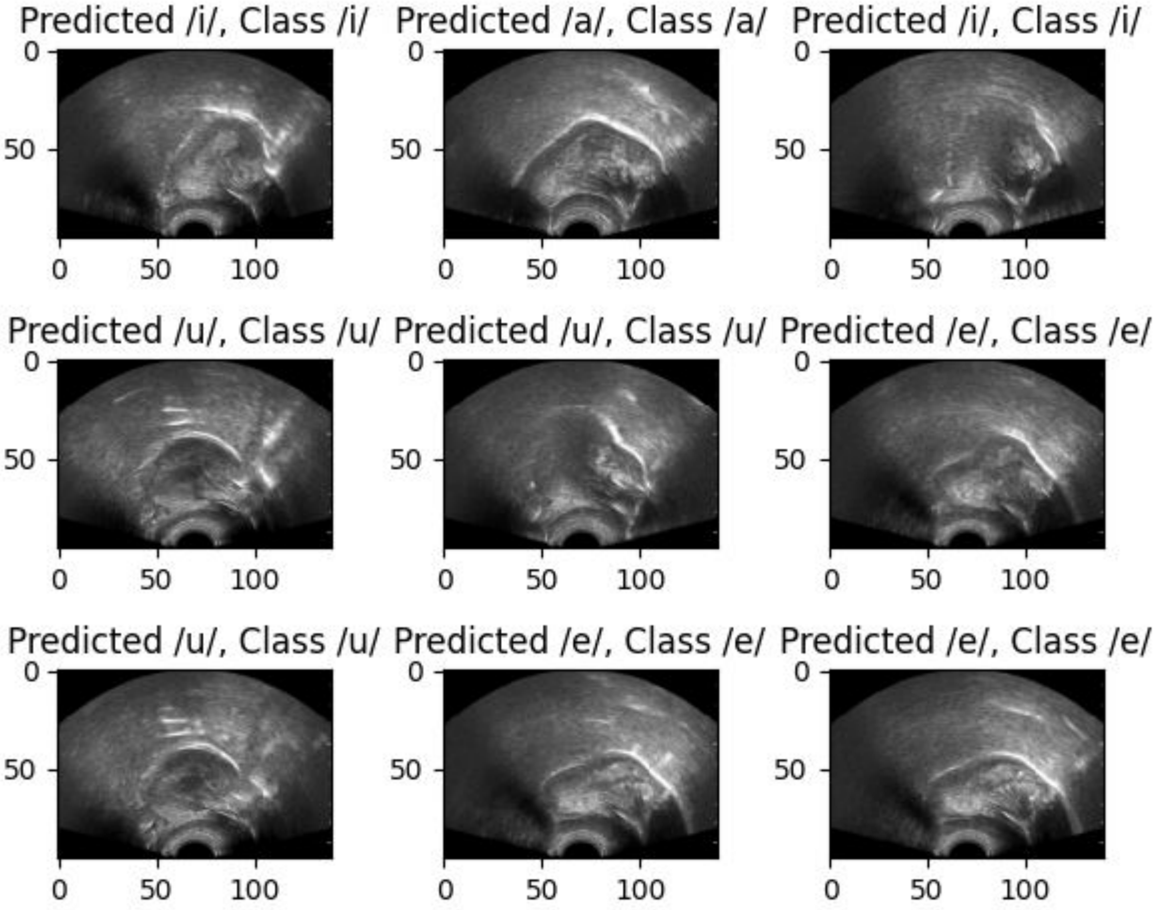


Test loss: 0.095
Test accuracy: 0.992

No overfitting
Robust performance

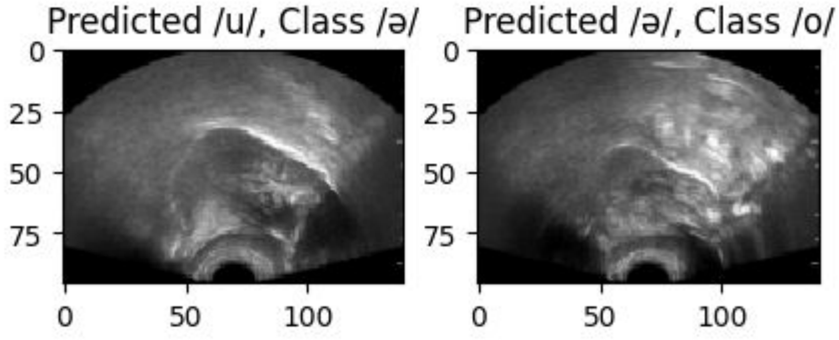
Results: classification

238 correct labels



Results: classification

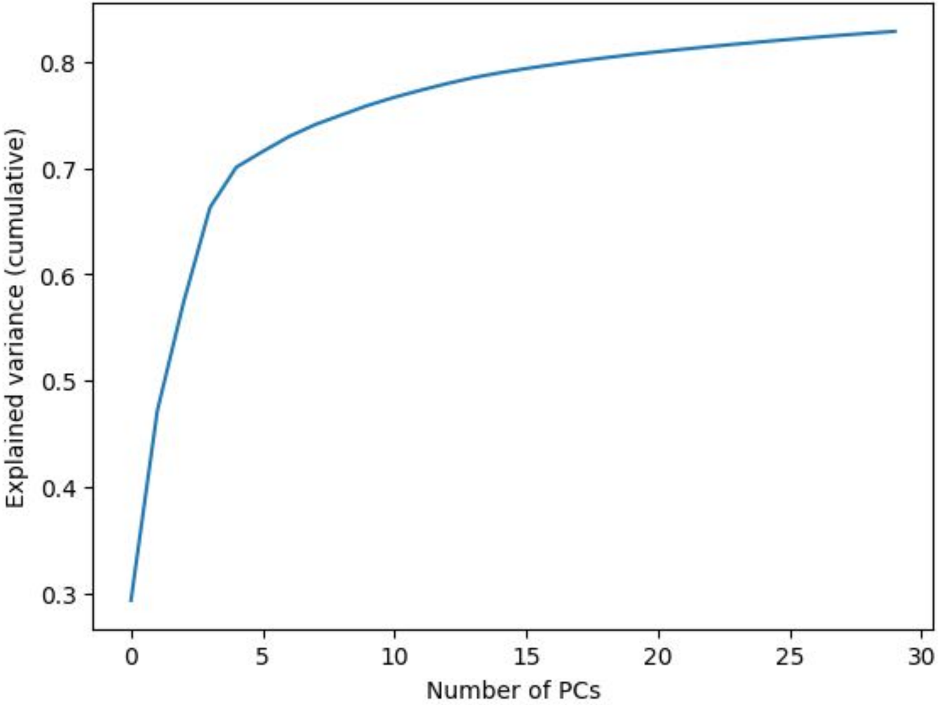
2 incorrect labels



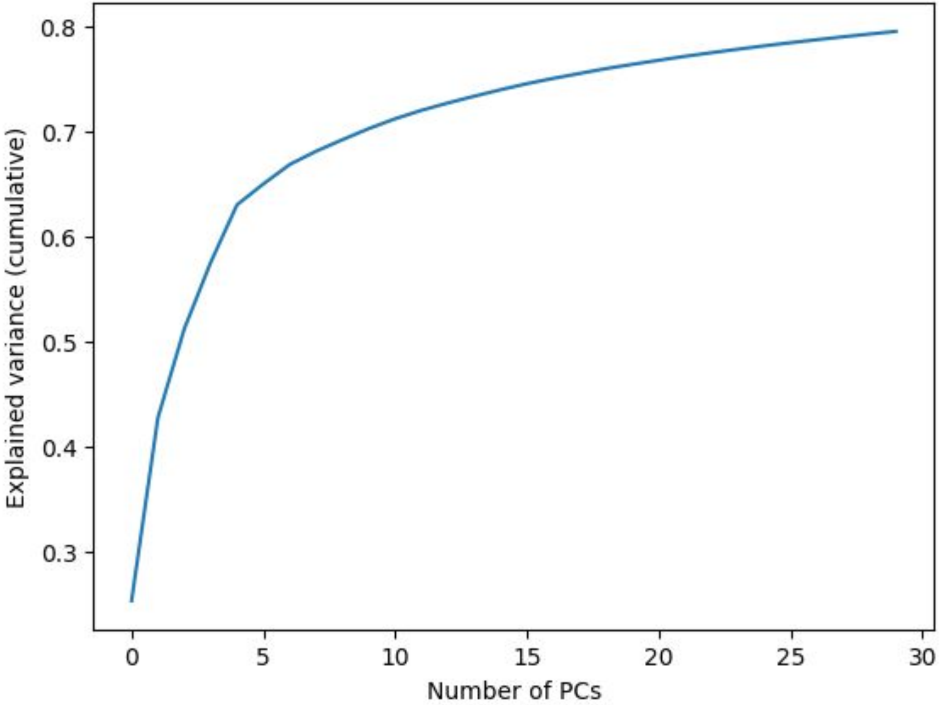
Results: clustering-PCA

- Explained variance (cumulative)

Speaker M



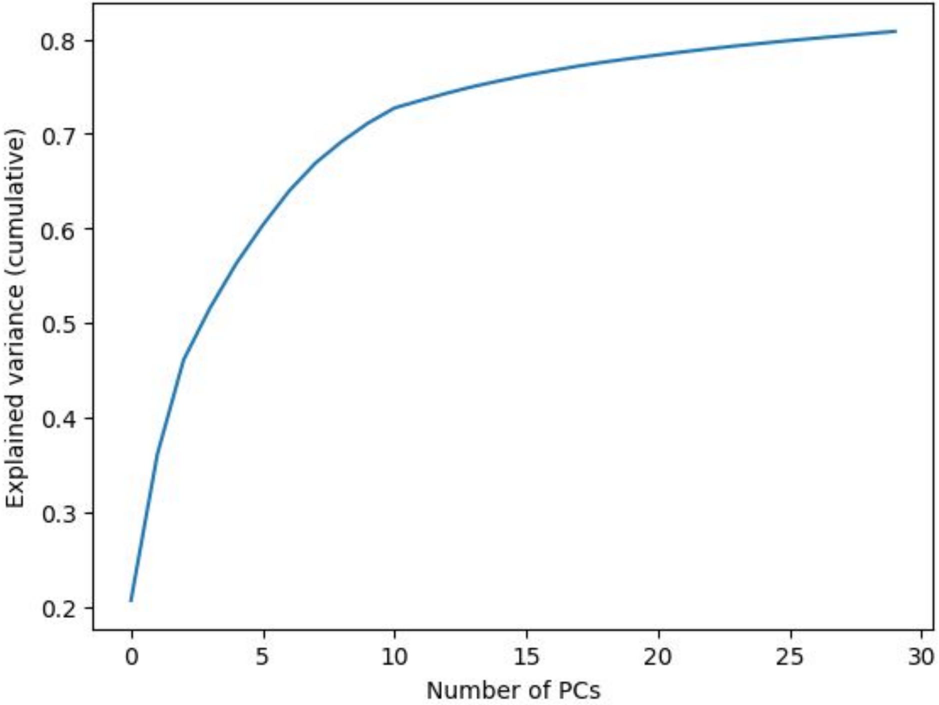
Speaker F



Results: clustering-PCA

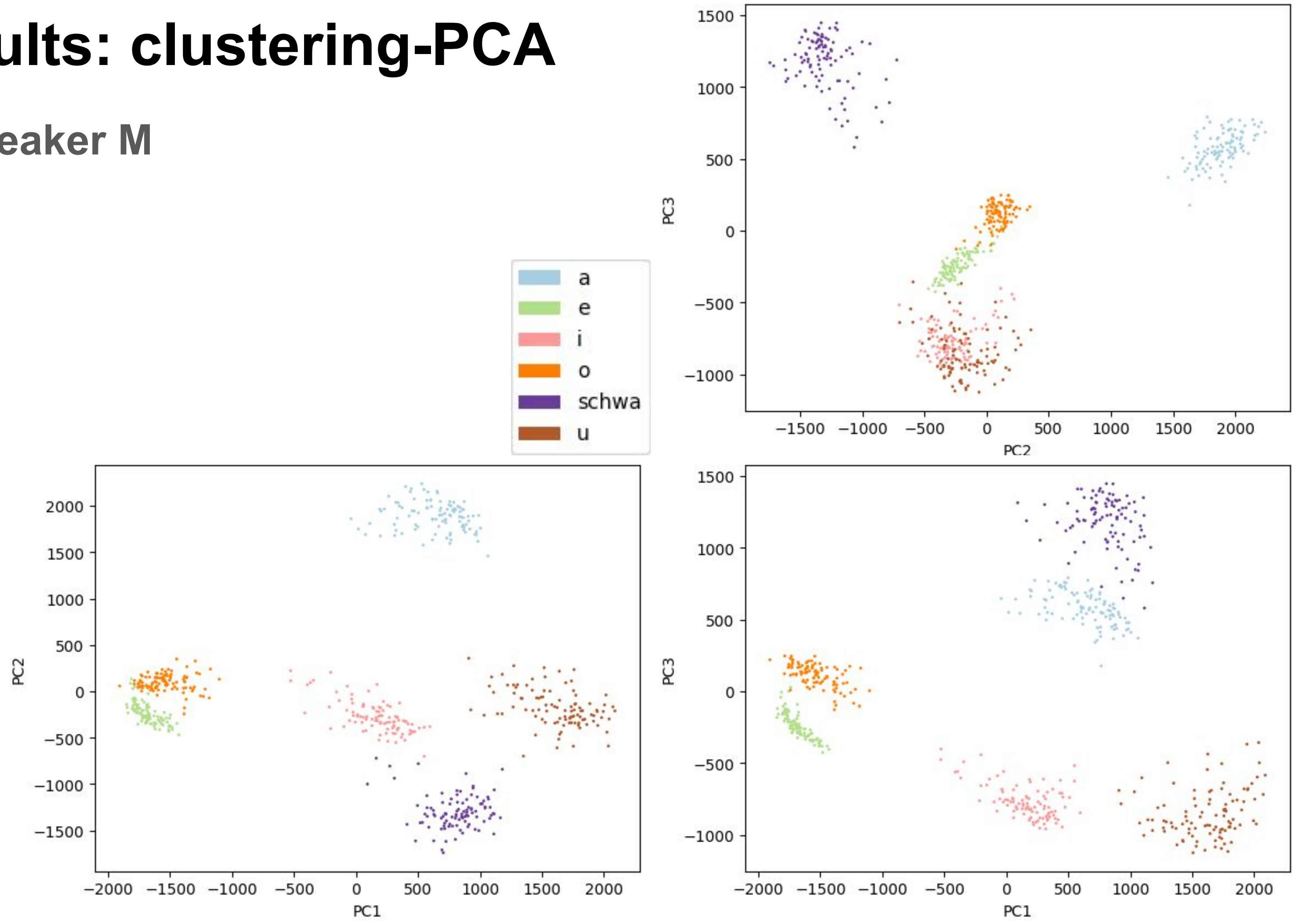
- Explained variance (cumulative)

Combined



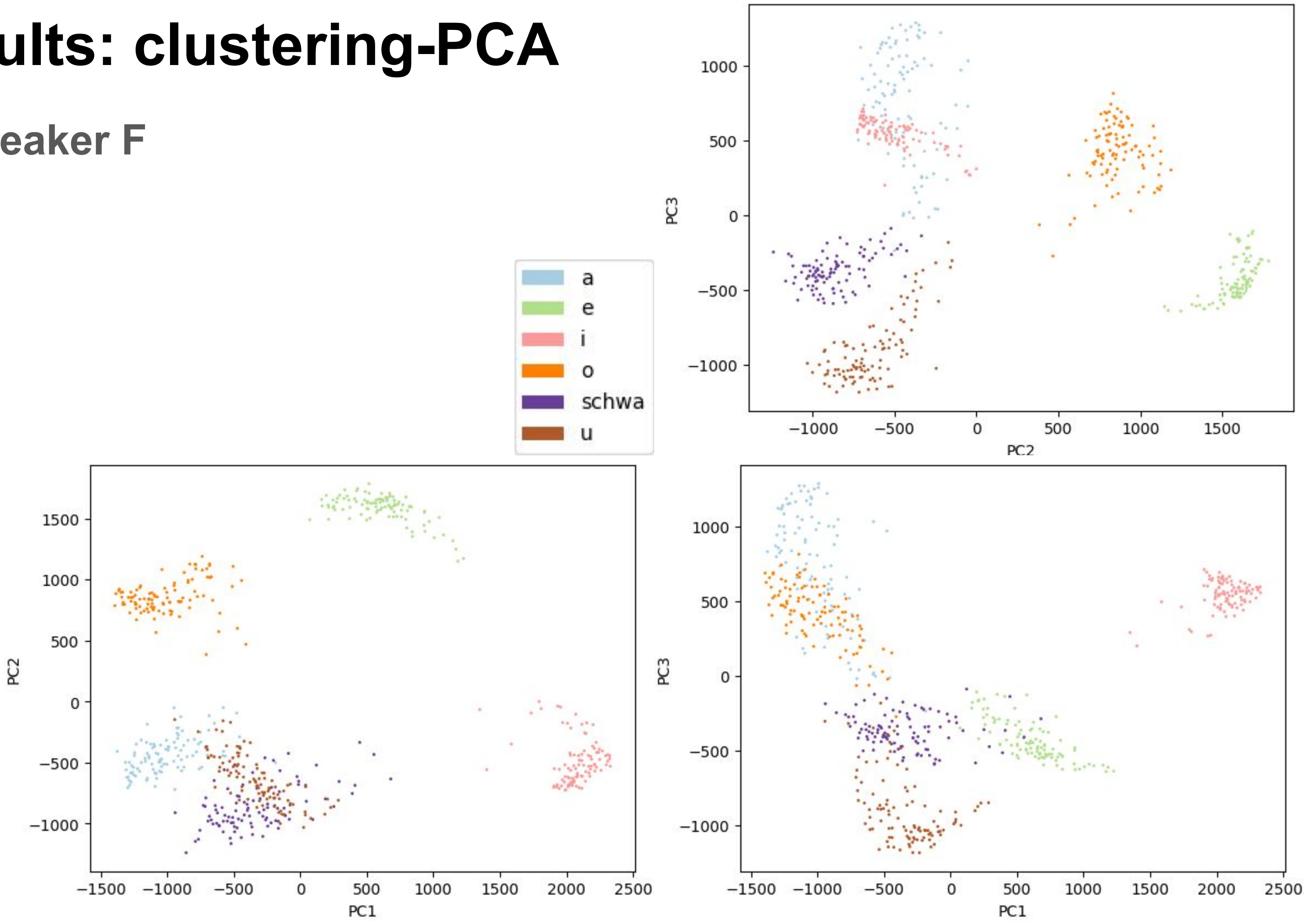
Results: clustering-PCA

- **Speaker M**



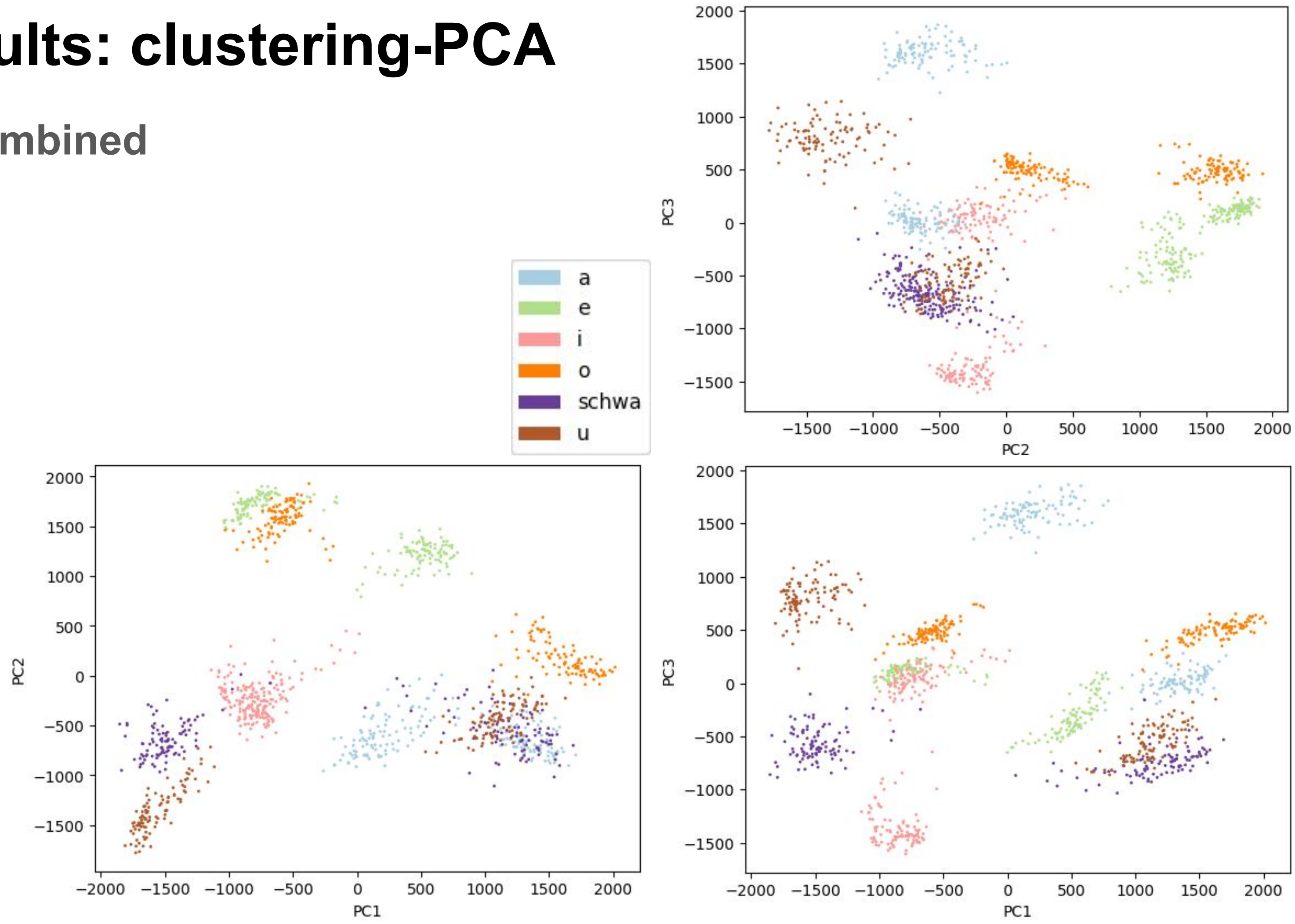
Results: clustering-PCA

- **Speaker F**



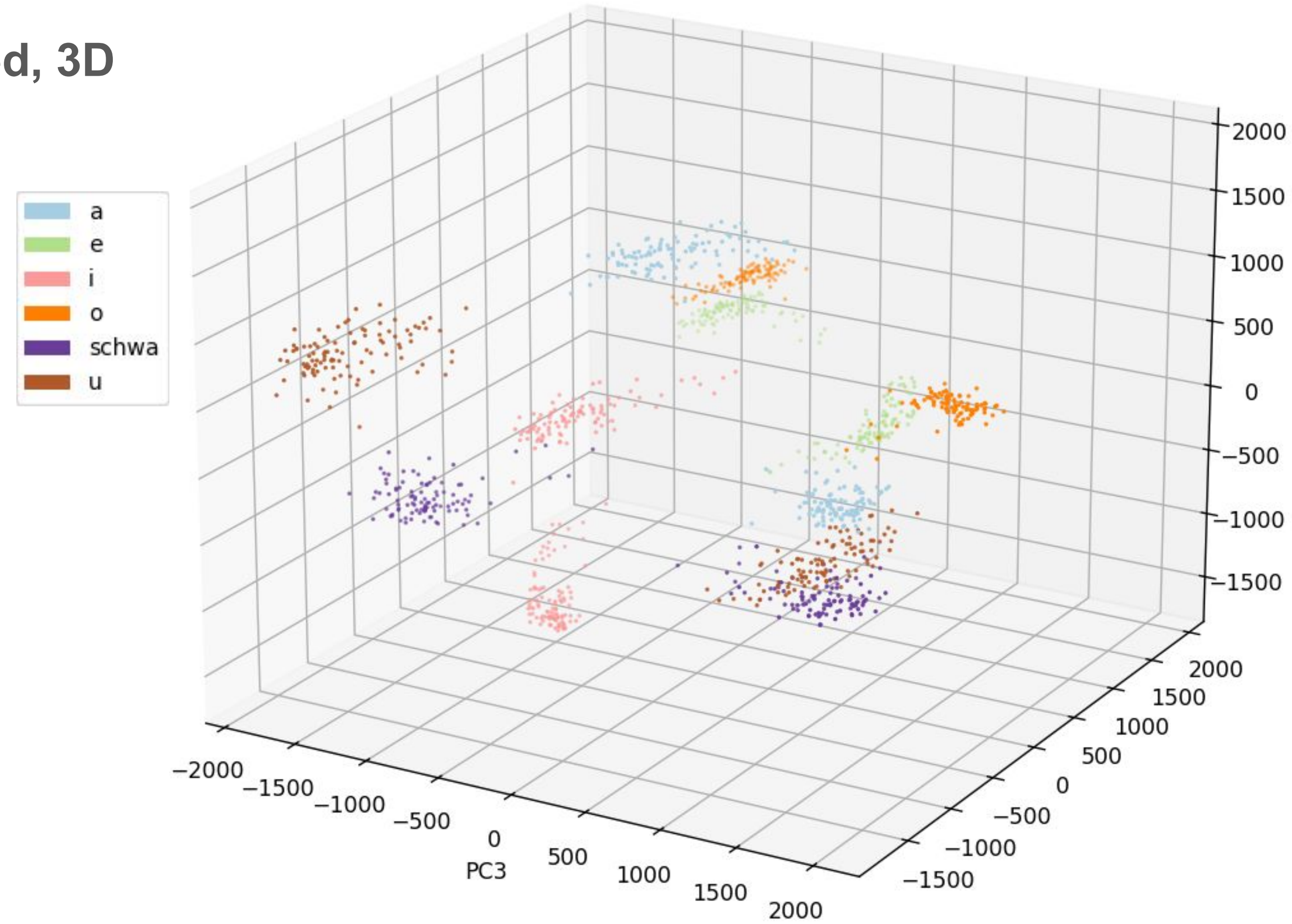
Results: clustering-PCA

- Combined



Results: clustering-PCA

- **Combined, 3D**



Results: clustering-UMAP

Speaker M

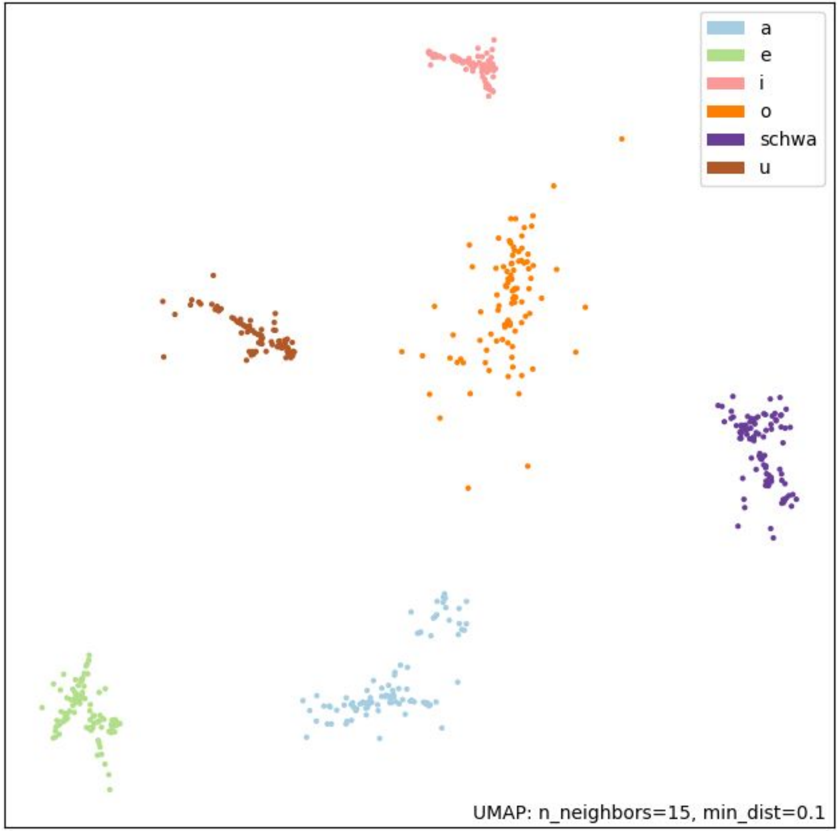
2D embedding

connectivity

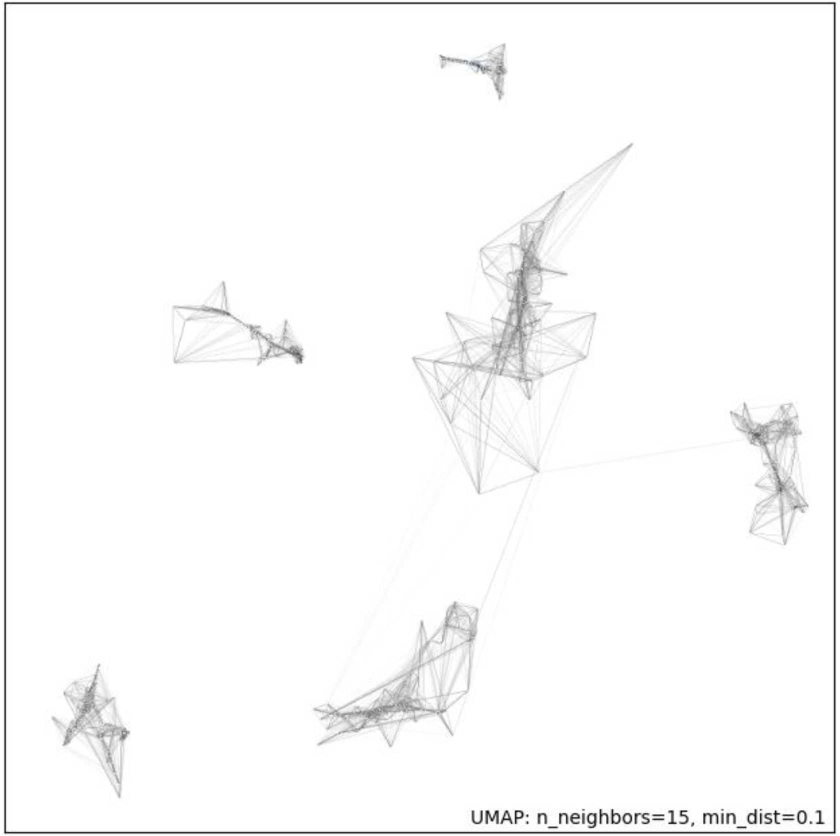
Results: clustering-UMAP

Speaker F

2D embedding



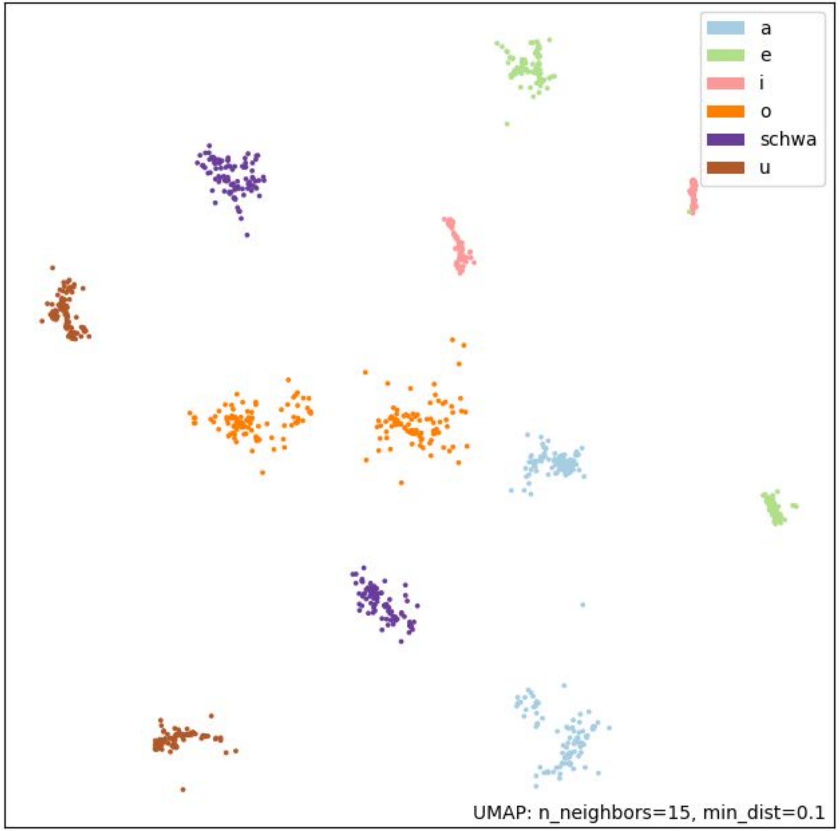
connectivity



Results: clustering-UMAP

Combined

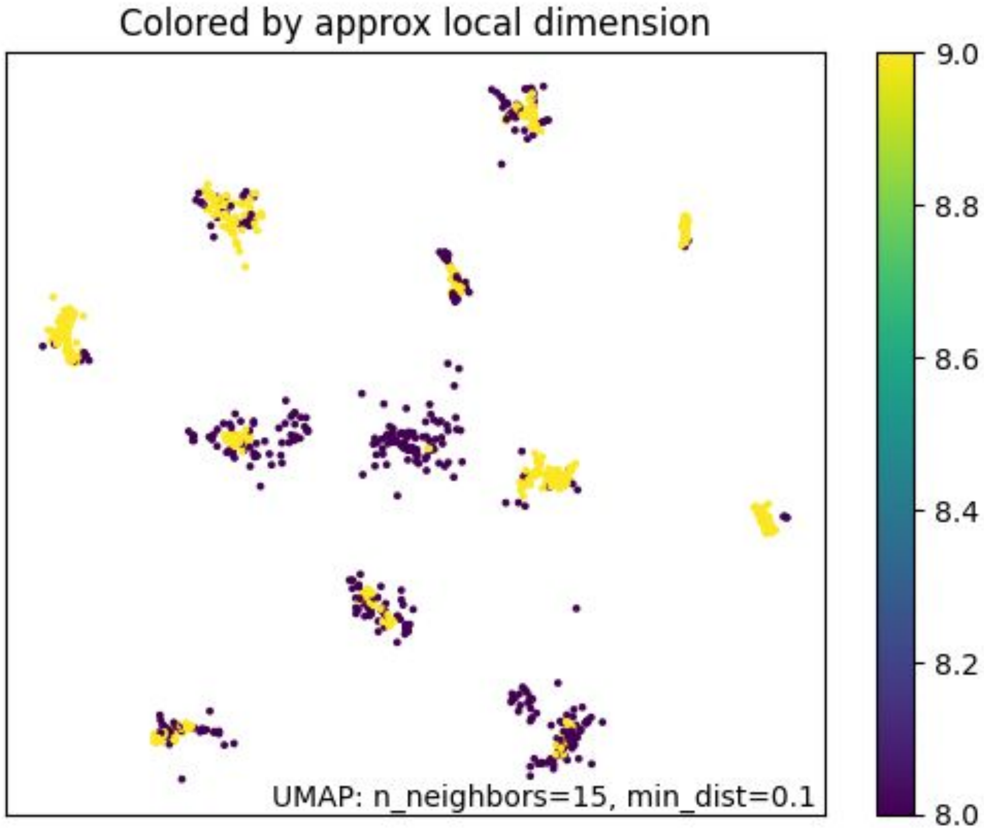
2D embedding



connectivity

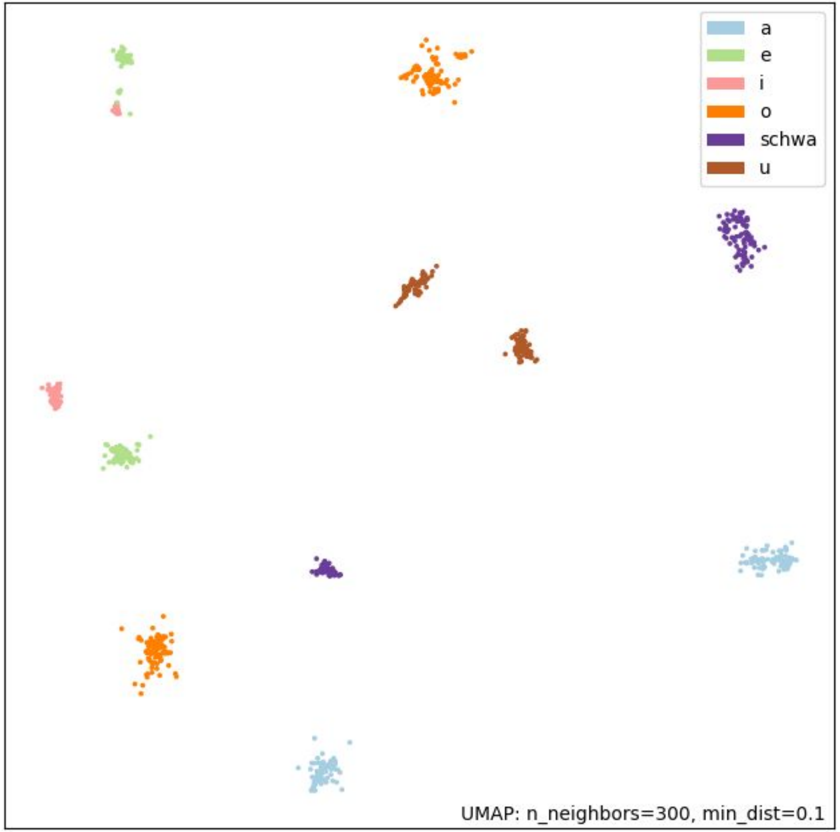
Results: clustering-UMAP

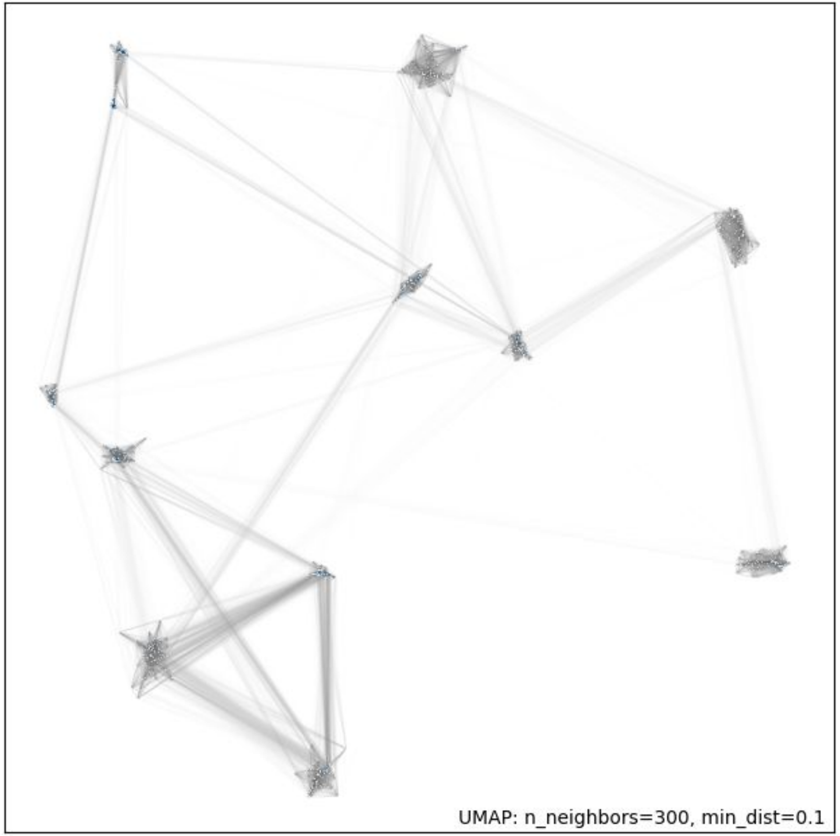
Combined local dimensions



Results: clustering-UMAP

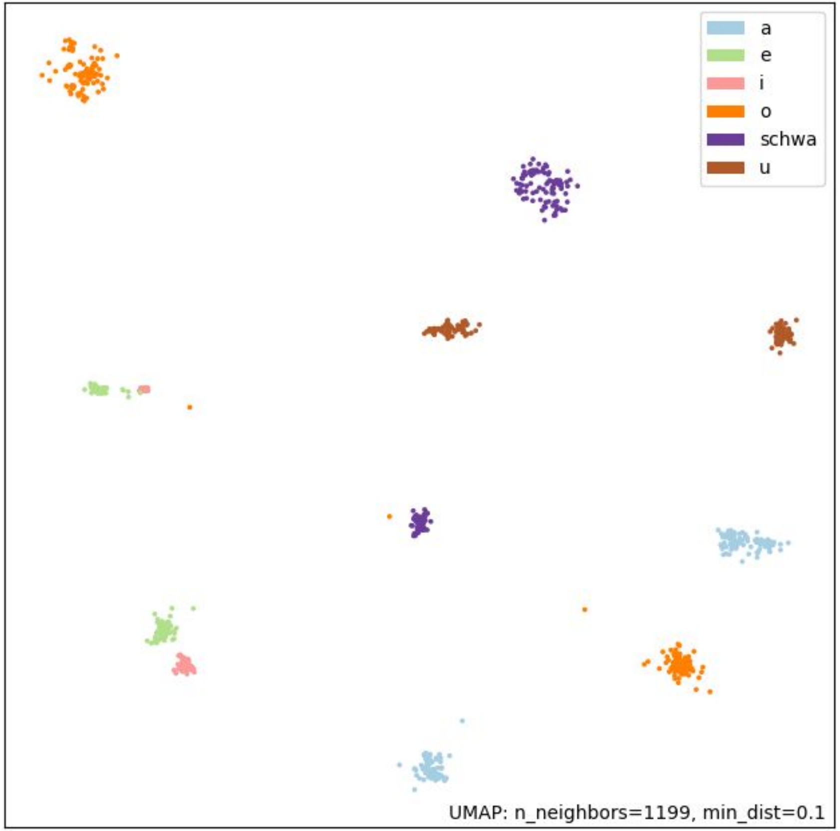
Combined, n_neighbors=300 **2D embedding** **connectivity**

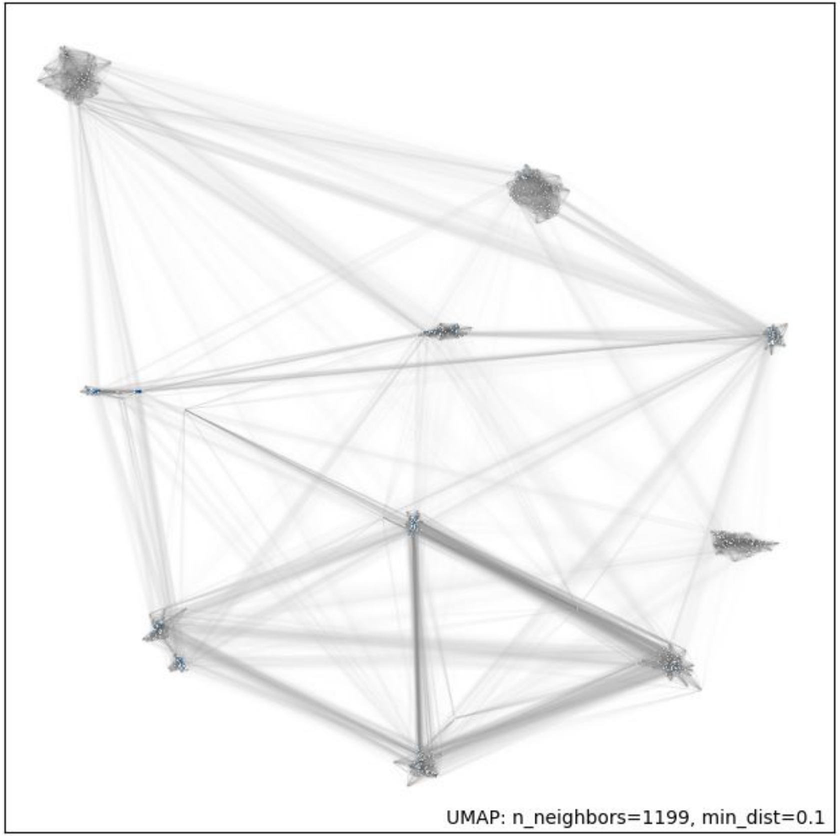




Results: clustering-UMAP

Combined, n_neighbors=1199 **2D embedding** **connectivity**

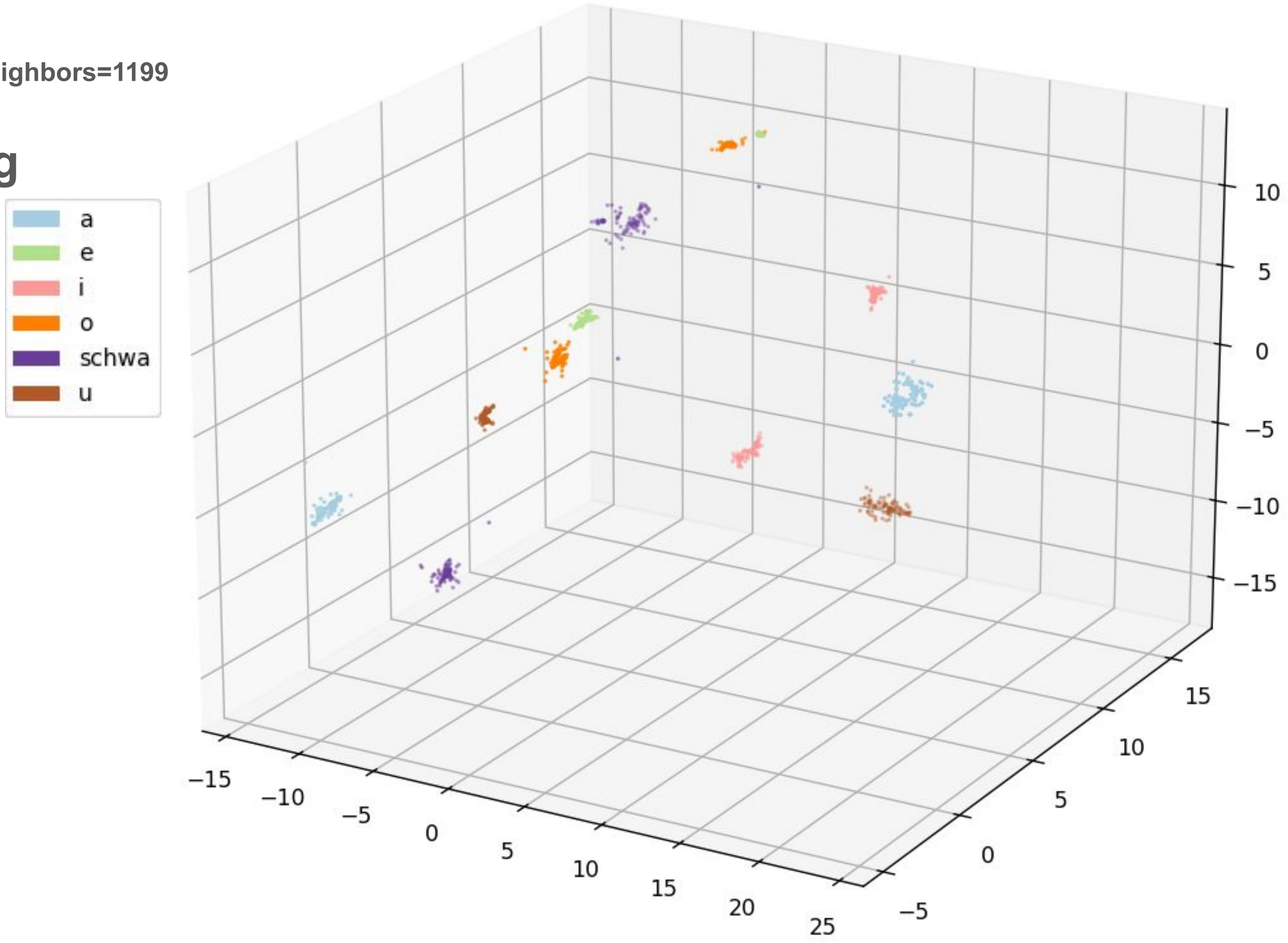




Results: clustering-UMAP

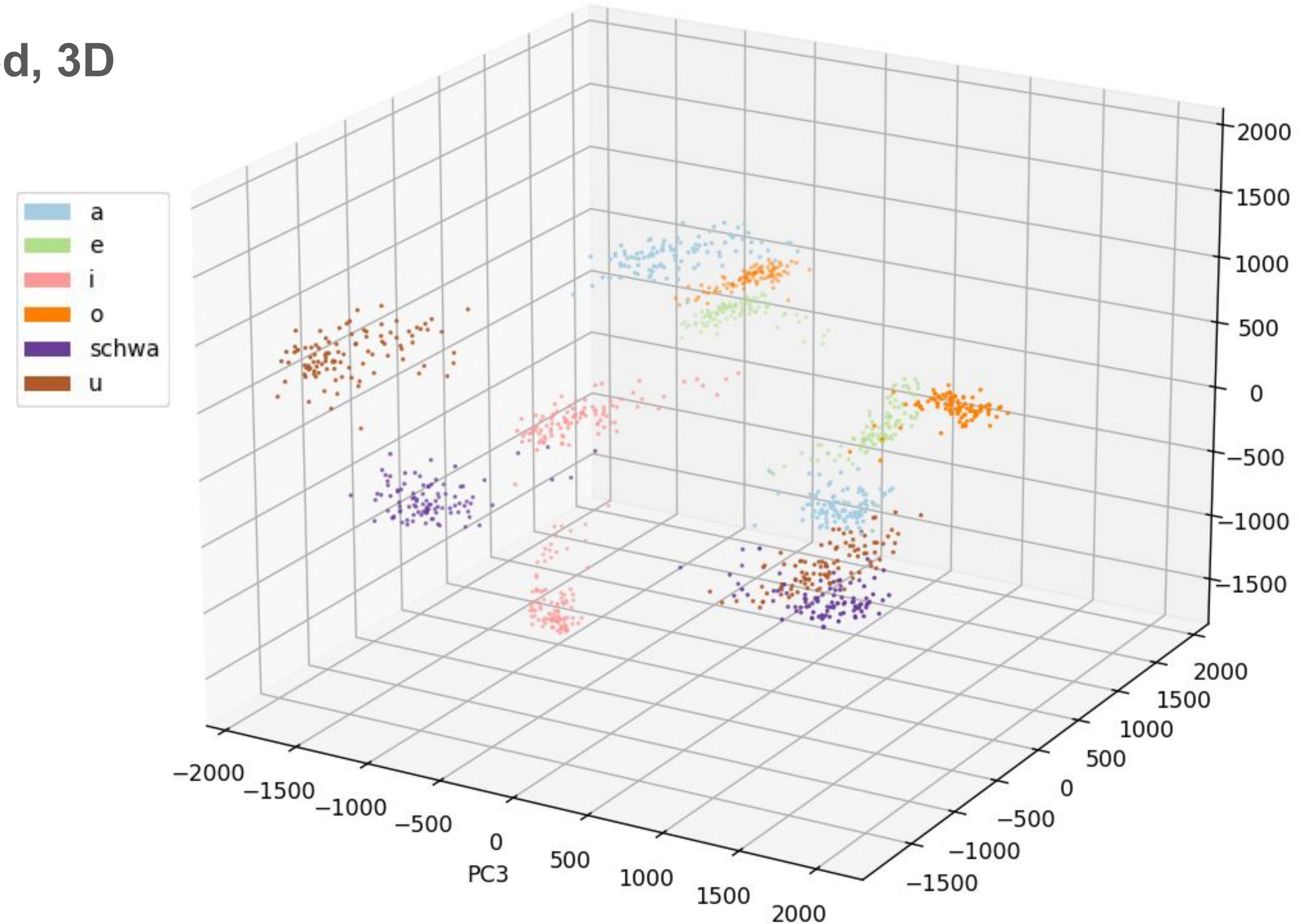
Combined, n_neighbors=1199

3D embedding



Results: clustering-PCA

- **Combined, 3D**



Discussions

- Both classification and clustering are very effective for stable vowel images

Discussions

- Both classification and clustering are very effective for stable vowel images
- Convolutional Autoencoder
  - Accuracy is surprisingly very high
  - Maybe input data is not variable enough / too simple

Discussions

- Both classification and clustering are very effective for stable vowel images
- Convolutional Autoencoder
  - Accuracy is surprisingly very high
  - Maybe input data is not variable enough / too simple
- PCA
  - Needs a lot of PCs to capture >80% total variance
  - Overlapping in combined data set

Discussions

- Both classification and clustering are very effective for stable vowel images
- Convolutional Autoencoder
  - Accuracy is surprisingly very high
  - Maybe input data is not variable enough / too simple
- PCA
  - Needs a lot of PCs to capture >80% total variance
  - Overlapping in combined data set
- UMAP
  - Vowels fall into distinct clusters while preserving local density
  - Distance between clusters is hard to interpret when graph is disconnected
    - Suggesting low variability of input data

Discussions

- Both classification and clustering are very effective for stable vowel images
- Convolutional Autoencoder
  - Accuracy is surprisingly very high
  - Maybe input data is not variable enough / too simple
- PCA
  - Needs a lot of PCs to capture >80% total variance
  - Overlapping in combined data set
- UMAP
  - Vowels fall into distinct clusters while preserving local density
  - Distance between clusters is hard to interpret when graph is disconnected
    - Suggesting low variability of input data
- The same vowels from different speakers don’t seem to cluster together
  - Image normalization

Thank you!