Combining different modalities in classifying phonological categories

SHUNAN ZHAO[1] AND FRANK RUDZICZ[1,2]
[1]UNIVERSITY OF TORONTO
[2]TORONTO REHABILITATION INSTITUTE

UNIVERSITY OF TORONTO

UHN Toronto Rehabilitation Institute

Introduction

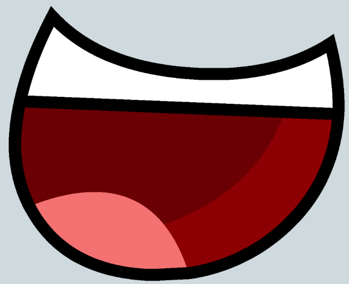
- **Imagined speech**: “hearing” one’s own voice silently to oneself, without the intentional movement of any extremities such as lips, tongue, or hands (from Wikipedia).
- Uses:
  - Clinical tool to assist those with severe paralysis.
  - “Synthetic telepathy” for the military (Bogue, 2010).
  - General purpose communication.

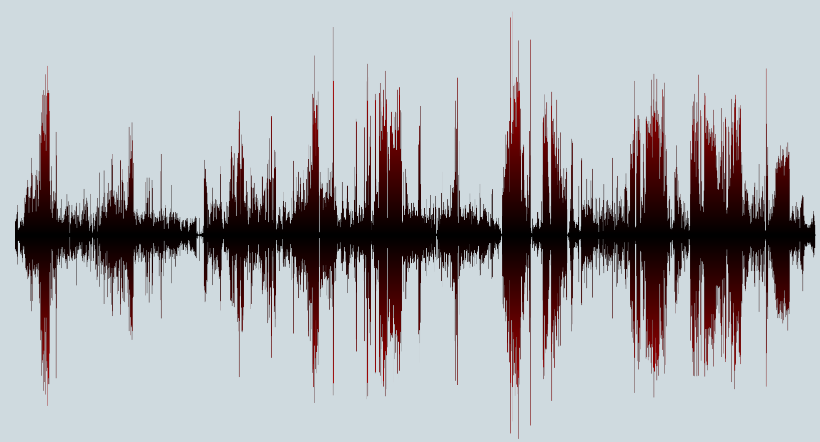
Previous Approaches

- Previous approaches at imagined speech classification
  - Invasive and partially-invasive methods (Blakely et al., 2008; Bartels et al., 2008; Kellis et al., 2010; Pasley et al., 2012).
  - EEG (Suppes et al., 1997; Brigham and Kumar, 2010; Callan et al., 2000; D'Zmura et al., 2009; DaSalla 2009)
- We are interested in discovering solutions that can be applied **more generally** and that **relate acoustics to speech production**.

Our Approach

- We collect audio, facial (from the **Kinect**) and EEG data of vocalized and imagined speech.
- This allows us to **relate** the **acoustics** with internal **speech production** and **speech articulation**.



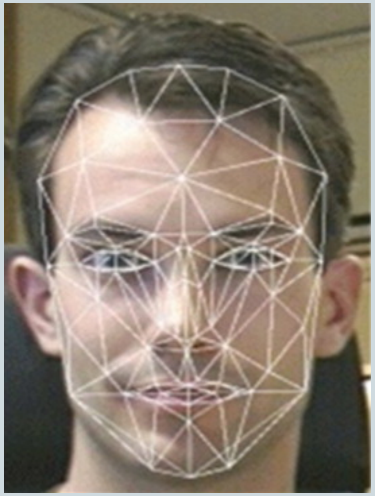


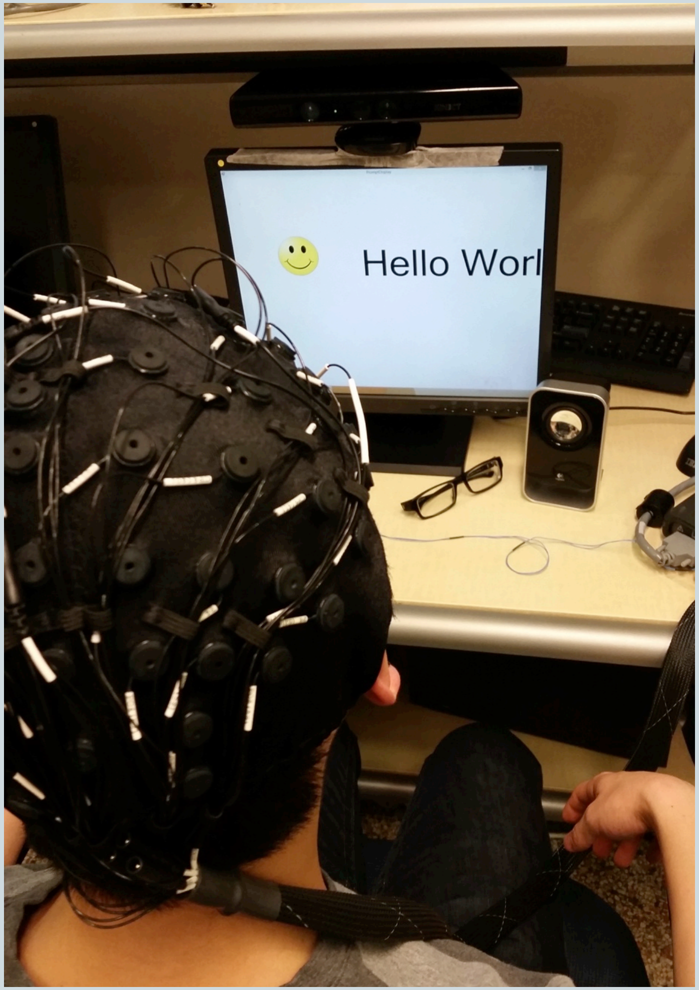
Participants

- **12 participants** (mean age = 27.4, σ = 5, range = 14) were recruited from the University of Toronto campus.
- All participants were **right-handed**, had **some post-secondary education**, and had **no history of neurological conditions or substance abuse**.
- 10 participants identified **NA English** as their native language and 2 spoke NA English at a fluent level.

Recording

- A **Microsoft Kinect** camera was used to record **facial information** (6 animation units) and **audio**, while EEG was recorded using a 64-channel cap.



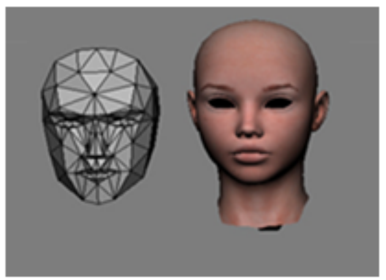


Task

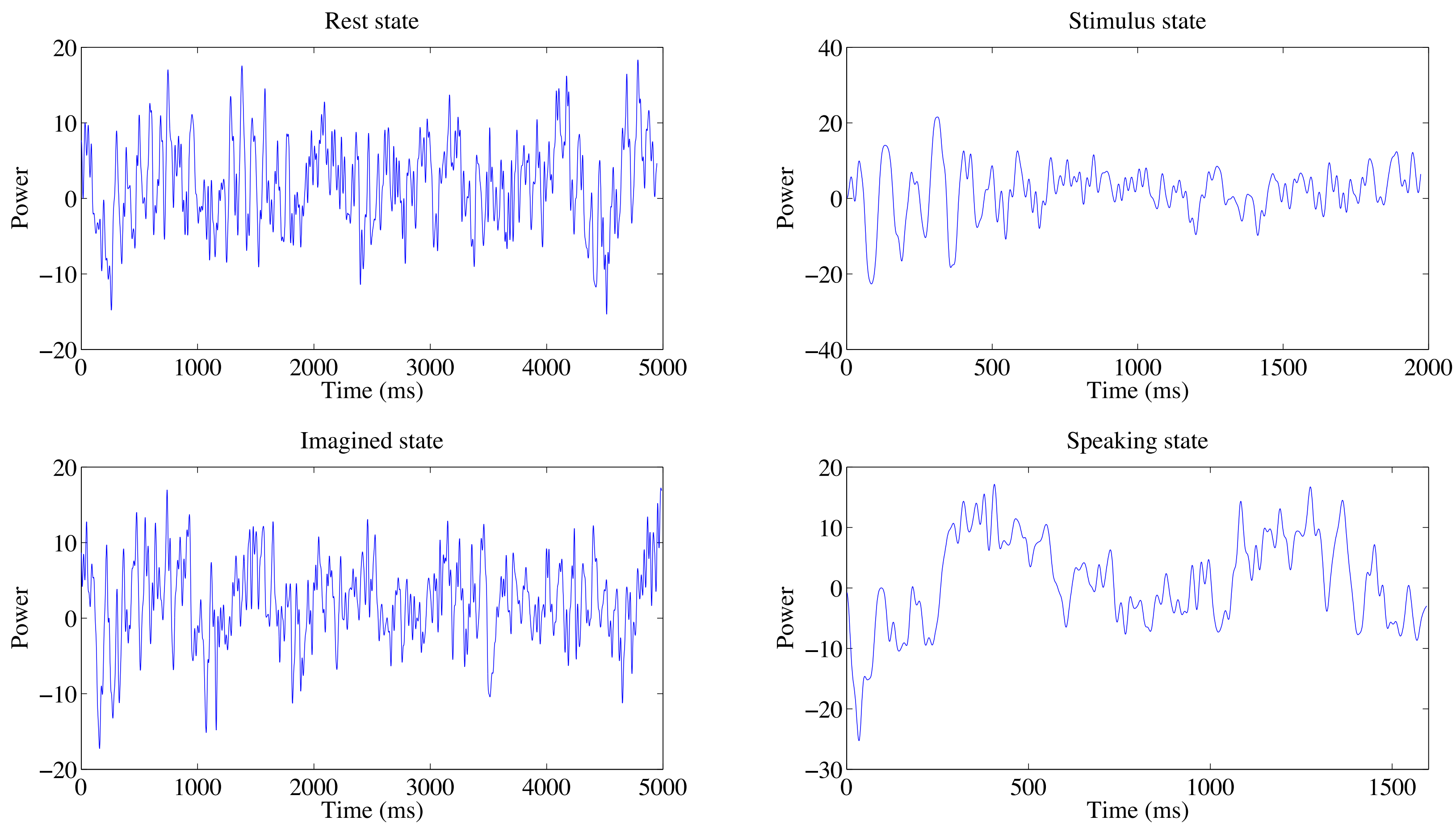
- Participants performed the following task:
  1. **Rest state:** (5 sec.) Participants were instructed to clear their mind.
  2. **Stimulus state:** A prompt appeared on the screen and was played over the computer's speakers. Participants were instructed to move their articulators into position to begin pronouncing the prompt.
  3. **Imagined state:** (5 sec.) Participants imagined speaking the prompt without moving.
  4. **Speaking state:** Participants spoke the prompt aloud.

Animation Units

- Upper Lip Raiser
- Jaw Lowerer
- Lip Stretcher
- Brow Lowerer
- Lip Corner Depressor
- Outer Brow Raiser



Different States

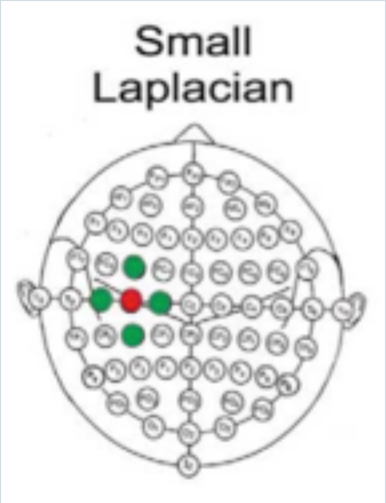


Prompts

- We used **7 phonemic/syllabic prompts**.
  - */iy/, /uw/, /piy/, /tiy/, /diy/, /m/, /n/*
- And, **4 words** from Kent's list of phonetically-similar pairs (Kent et al., 1989)
  - *pat, pot, knew, gnaw*
- Each prompt was presented **12 times**, for a total of **132 trials** per person.
- The phonemic prompts were first presented, followed by the 4 "Kent" words. Within each section, the trials were randomly permuted.

Pre-processing

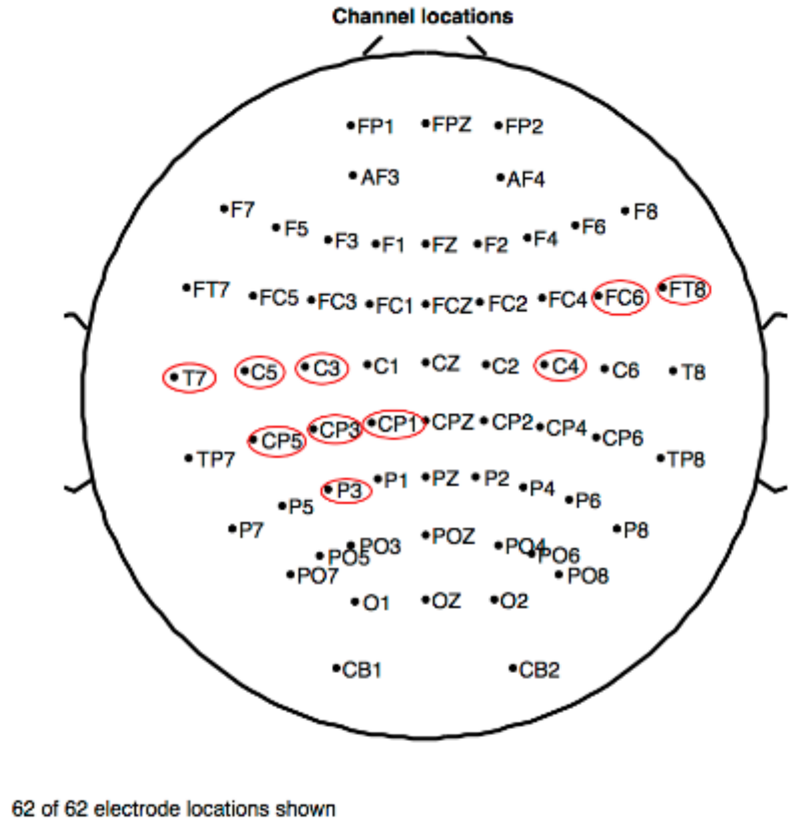
- Pre-processing for the EEG data was done using **EEGLAB** (Delorme and Makeig, 2004) and **ocular artifacts** were removed using **BSS** (Gomez-Herrero et al., 2006).
- The data was filtered between **1 and 50 Hz** and mean values were subtracted from each channel.
- We applied a small **Laplacian filter** to each channel, using the neighbourhood of adjacent channels.



Features

- For the EEG and audio data, we window the data to approximately **10% of the segment**, with a **50% overlap** between consecutive windows.
  - For each window, we compute various statistical measures, spectral entropy, energy, kurtosis, and skewness. We also compute the first and second derivative of the above features.
  - This gives us 65,835 EEG features (over 62 channels) and 1197 acoustic features.
- For the **facial** data, we compute a **subset** of the above features.
- We perform **feature selection** by ranking features by their Pearson correlations with the given classes, for each task independently.

- We computed the **Pearson correlations** between all features in the audio and each of the 62 channels.
- The **10 channels** with the **highest absolute correlations** are circled in red in the image on the right.
- This seems to confirm the involvement of the **motor cortex** in the planning of speech articulation (Pulvermuller et al., 2005)



Most informative electrode positions

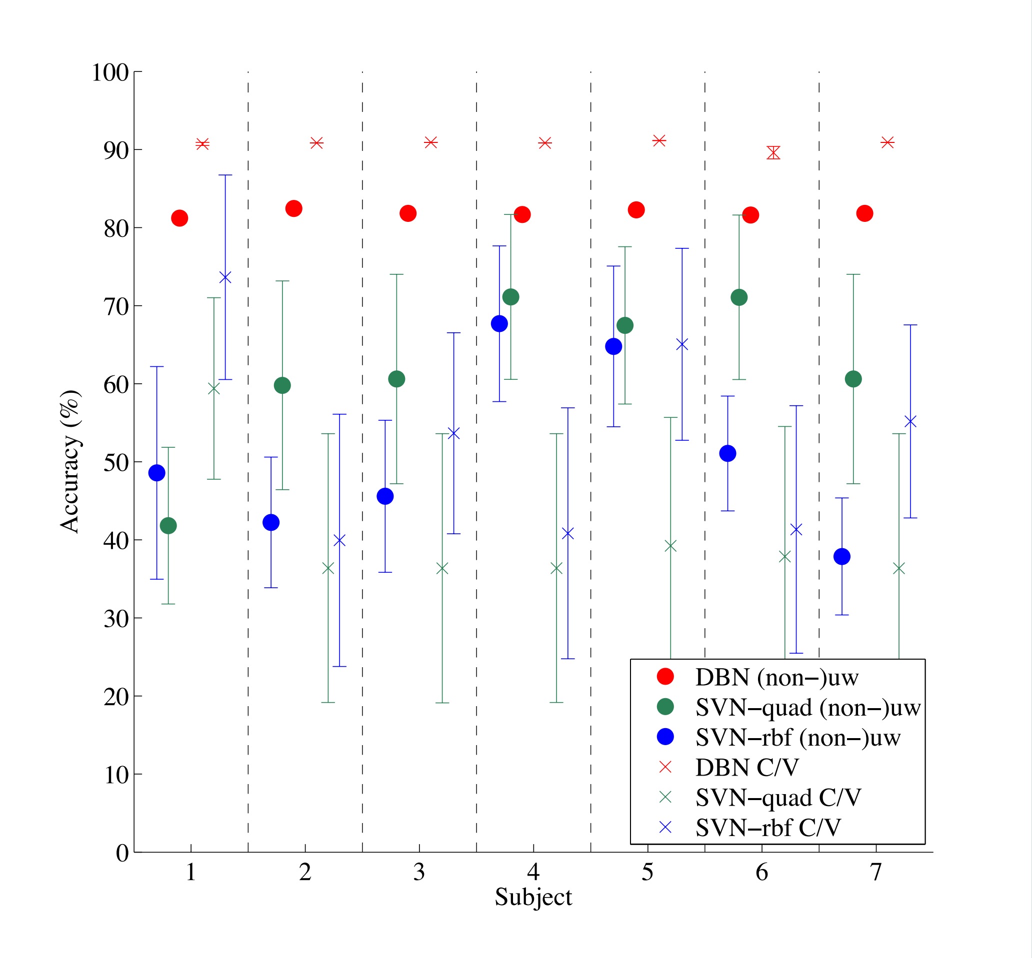
Experiments

- We use **subject-independent leave-one-out cross-validation** for our experiments.
- We use three classifiers:
  - A deep-belief network (**DBN**), with one hidden layer whose size is 25% of the input size. We also do up to 10 iterations of pre-training, a learning rate of 0.1, and a dropout rate of 0.5.
  - An SVM with a quadratic kernel (**SVM-quad**).
  - An SVM with a radial basis function kernel (**SVM-rbf**)

Classification of Phonological Categories

- We classify between various phonological categories.
- We consider the 5 binary classification tasks:
  - Vowel-only vs. consonant (**C/V**)
  - Presence of nasal (±**Nasal**)
  - Presence of bilabial (±**Bilab.**)
  - Presence of high-front vowel (±/**iy**/)
  - Presence of high-back vowel (±/**uw**/)
- We use six different feature sets: **EEG**-only, facial features (**FAC**)-only, audio (**AUD**)-only, EEG and facial features (**EEG+FAC**), EEG and audio features **(EEG+AUD)**, and **all** modalities.

Results

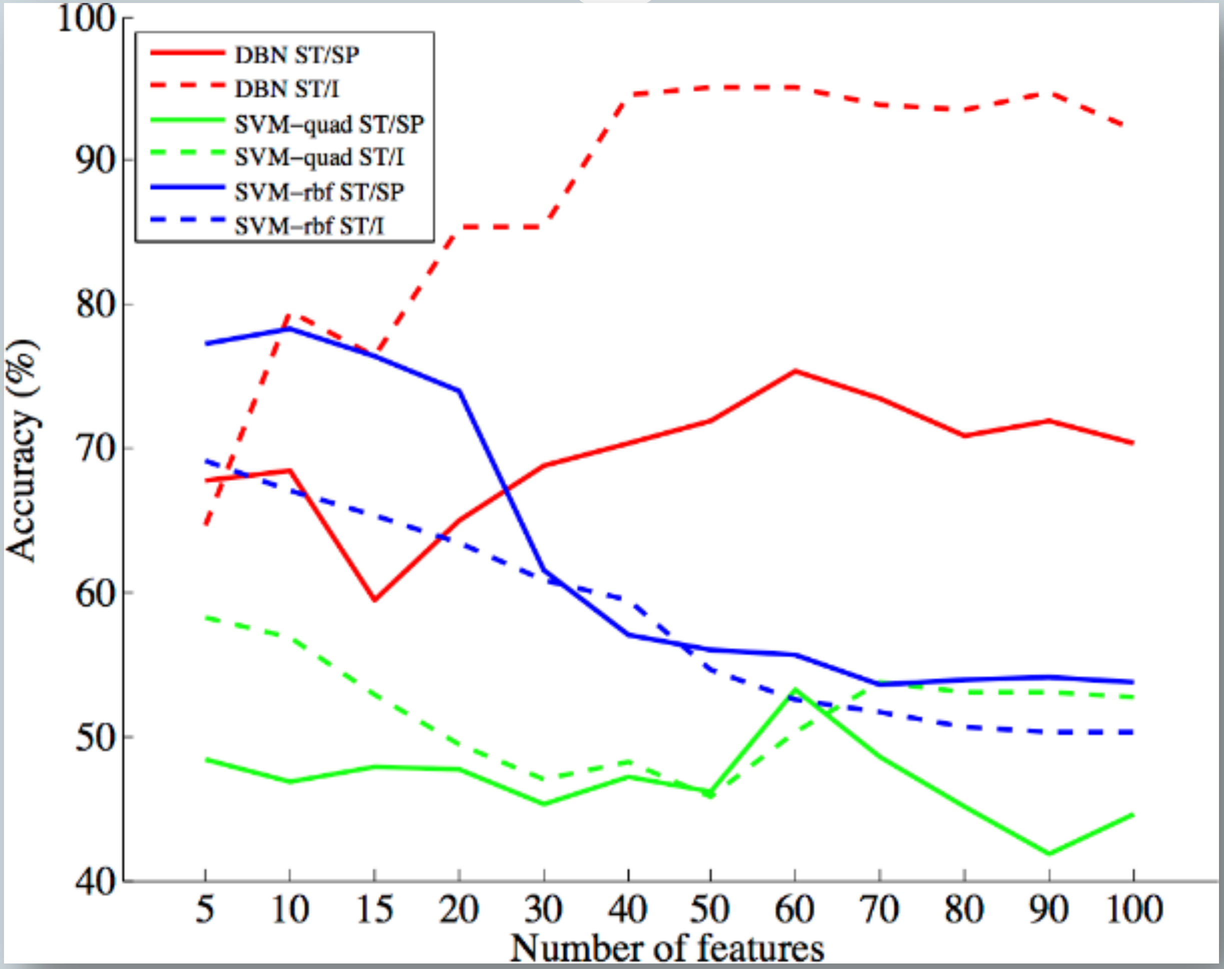


100
90
80
70
60
50
40
30
20
10
0
Accuracy (%)
1
2
3
4
5
6
7
Subject
DBN (non–)uw
SVN–quad (non–)uw
SVN–rbf (non–)uw
DBN C/V
SVN–quad C/V
SVN–rbf C/V

Classification of Mental State

- As a second experiment, we classify the different states of each trial in three binary tasks:
  - Stimulus vs. speaking (**ST/SP**)
  - Rest vs. imagined (**R/I**)
  - Stimulus vs. imagined (**ST/I**)
- We use the same classifiers as before with the same hyper-parameters.
- To improve performance, we concatenate the band-pass filtered data from 6/8 participants and perform **ICA**.

Classification Results



Conclusions and Future Work

- We present the **first** classification of **phonological** categories combining **acoustic**, **facial**, and **EEG** data, using relatively inexpensive equipment.
- We plan on making the data publicly available in the near future.
- Future work will involve methods to reconstruct acoustic features from the EEG.