CONTACT US

Computational Linguistics
University of Toronto, Department of Computer Science

Home | People | Courses | Research | Publications | Downloads | Meetings

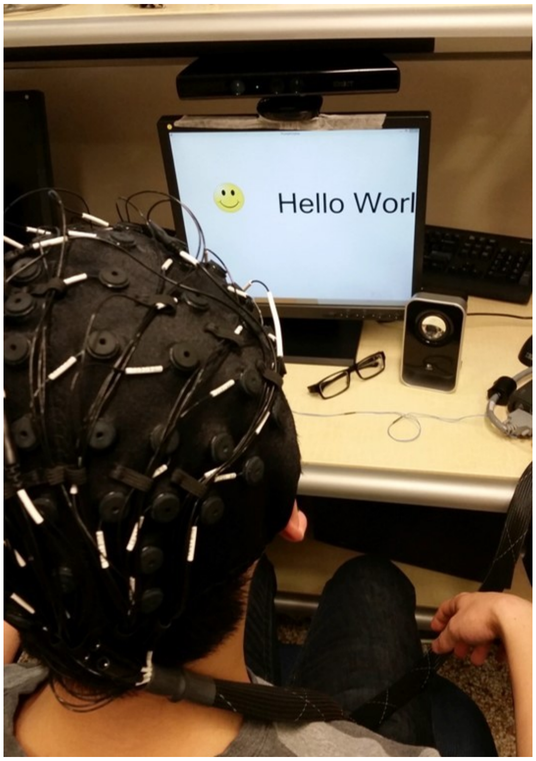
The KARA ONE Database: Phonological Categories in imagined and articulated speech

Brain-computer interfaces (BCIs) often involve imagining gross motor movements to move a pointer on-screen. This research attempts to access language centres directly. While invasive methods have high signal-to-noise ratios, they are only used in severe cases, due to the complex nature of the surgery. In practice, solutions must be applicable more generally.

Here, we present a new dataset, called Kara One, combining 3 modalities (**EEG**, **face tracking**, and **audio**) during **imagined and vocalized phonemic and single-word prompts**. This accesses the **language** and **speech production** centres of the brain. In the **associated paper**, we show how to accurately classify imagined phonological categories solely from EEG data.

These data may be used to learn multimodal relationships, and to develop silent-speech and brain-computer interfaces.

Instrumentation and stimuli



Kara One setup

Kinect face tracking

Each study was conducted in an office environment at the **Toronto Rehabilitation Institute**. Each participant was seated in a chair before a computer monitor. A Microsoft Kinect (v.1.8) camera was placed next to the screen to record facial information and the participant's speech. For each frame of video, the Kinect extracted six 'animation units' (AUs), all on R[-1..1] upper lip raiser, jaw lowerer, (lateral) lip stretcher, brow lowerer, lip corner depressor, outer brow raiser. A research assistant placed an appropriately-sized EEG cap on the participant's head and injected a small amount of gel to improve electrical conductance. We used a 64-channel Neuroscan Quick-cap, where the electrode placement follows the 10-20 system. To control for artifacts arising from eye movement, we used 4 electrodes placed above and below the left eye and to the lateral side of each eye. All EEG data were recorded using the SynAmps RT amplifier and sampled at 1 kHz. Impedance levels were usually maintained below 10 k.

After EEG setup, the participant was instructed to look at the computer monitor and to move as little as possible. Over the course of 30 to 40 minutes, individual prompts appeared on the screen one-at-a-time. We used 7 phonemic/syllabic prompts (*iy*, *uw*, *piy*, *tiy*, *diy*, *m*, *n*) and 4 words derived from Kent's list of phonetically-similar pairs (i.e., *pat, pot, knew, and gnaw*). These prompts were chosen to maintain a relatively even number of nasals, plosives, and vowels, as well as voiced and unvoiced phonemes.

Each trial consisted of 4 successive states:

1. A 5-second rest state, where the participant was instructed to relax and clear their mind of any thoughts.

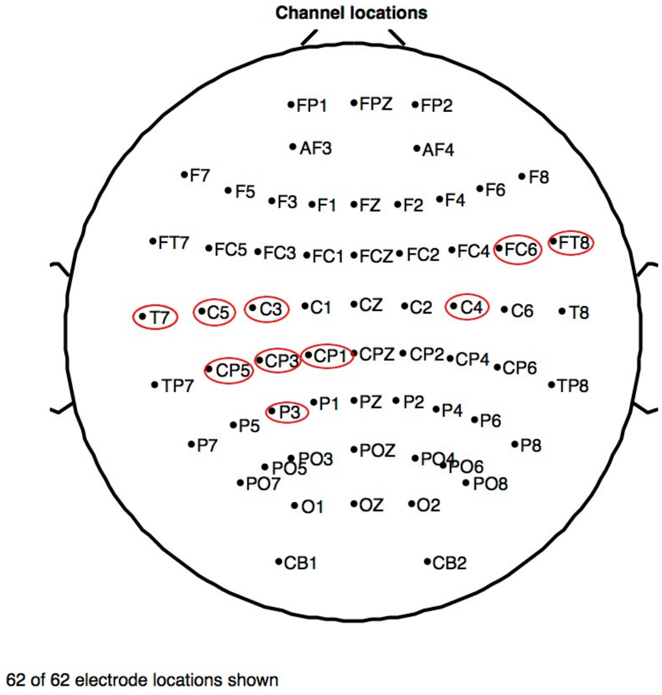
2. A stimulus state, where the prompt text would appear on the screen and its associated auditory utterance was played over the computer speakers. This was followed by a 2-second period in which the participant moved their articulators into position to begin pronouncing the prompt.
3. A 5-second imagined speech state, in which the participant imagined speaking the prompt without moving.
4. A speaking state, in which the participant spoke the prompt aloud. The Kinect sensor recorded both the audio and facial features during this stage.

Preprocessing and feature extraction

Pre-processing for the EEG data was done using **EEGLAB** and ocular artifacts were removed using blind source separation. 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.

For each EEG segment and each non-ocular channel, we window the data to approximately 10% of the segment, with a 50% overlap between consecutive windows. We then compute various features over each window, including the mean, median, standard deviation, variance, maximum, minimum, maximum minimum, sum, spectral entropy, energy, kurtosis, and skewness.

As an aside, we also compute the Pearson correlations, *r*, between all 1197 features in the audio and in each of the 62 EEG channels over all imagined speech segments in our dataset. This provides an estimate of how well each EEG channel predicts the resulting audio. The top 10 highest absolute correlations (which all turned out to be moderately positive) are shown in the table below. Interestingly, these features are dominated by central locations, with only two temporal locations (one left, T7, and one right, FT8), generally around the auditory cortex (CP3, CP5), superior to the lateral fissure. That these features are also dominated laterally on the left (C5, CP3, P3, T7, CP5, C3, CP1) appears to confirm the involvement of these regions during the planning of speech articulation, which is being investigated.



Sensor	FC6	FT8	C5	CP3	P3
Mean *r*	0.3781	0.3758	0.3728	0.3720	0.3696
Sensor	T7	CP5	C3	CP1	C4
Mean *r*	0.3686	0.3685	0.3659	0.3626	0.3623

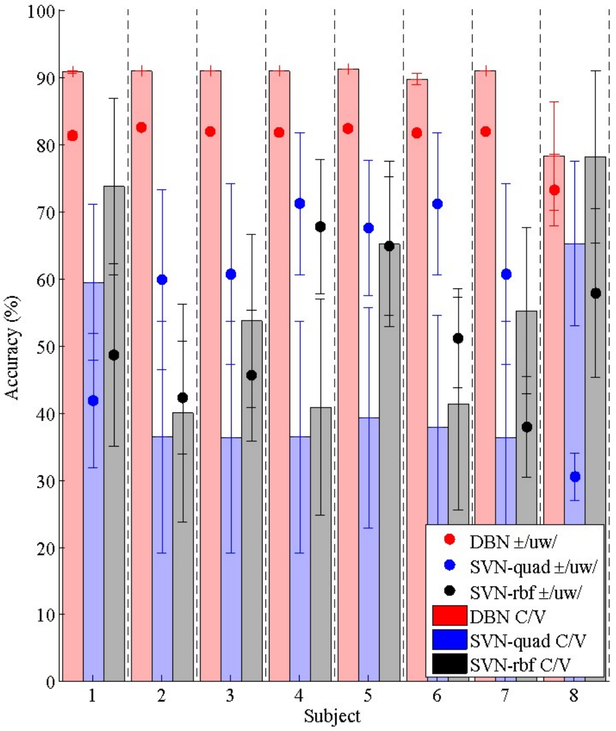
Mean Pearson correlations between acoustic features and most-correlated EEG sensor locations.

Machine learning and results

Our experiments use a subject-independent approach with leave-one-out cross-validation in which each subject's data are tested in turn using models trained with all other data combined. The results therefore may provide more generalizable conclusions than subject-specific models which depend on individual, non-transferable models. Our experiments use two types of classifier: a **deep-belief network (DBN)** and **support vector machine (SVM)** baselines. Two variants of the latter are tested, with different kernels; SVM-quad uses a quadratic kernel and SVM-rbf uses the radial basis function. For both SVMs, we allow 90% of data to violate the Karush-Kuhn-Tucker conditions, if necessary

We first classify between various phonemic and phonological classes given different modalities of data. Specifically, we consider five binary classification tasks: vowel-only vs. consonant (C/V), presence of nasal (±Nasal), presence of bilabial (±Bilab.), presence of high-front vowel (±*iy*), and presence of high-back vowel (±*uw*) using six modalities: EEG-only, facial features (FAC)-only, audio (AUD)-only, EEG and facial features (EEG+FAC), EEG and audio features (EEG+AUD), and all modalities.

The figure below shows the average accuracy (with std. error) of classifying ±*uw* and C/V, across the three classifiers and for each test subject (given subject-independent models trained on all other data) given *N = 5* input features. For both tasks, the DBN classifiers obtain **between 80% and 91% accuracy**.



Average accuracies across models for DBN, SVMquad, and SVM-rbf classifiers. Error bars represent standard error.

The paper

Use of this database is free for academic (non-profit) purposes. If you use these data in any scientific publication, you must reference the following paper:

- Shunan Zhao and Frank Rudzicz (2015) Classifying phonological categories in imagined and articulated speech. *In Proceedings of ICASSP 2015*, Brisbane Australia. **PDF**

Much of the same information can be found in these slides: **PDF**.

The data

We are making the following **14** sets of data, one per participant, available for free, **subject to the requirements above**, totalling **24GB**. These data are provided as-is -- neither the researchers nor the University of Toronto are responsible for any challenges you may encounter as a result of downloading or using these data.

- **MM05.tar.bz2**
- **MM08.tar.bz2**
- **MM09.tar.bz2**
- **MM10.tar.bz2**
- **MM11.tar.bz2**
- **MM12.tar.bz2**
- **MM14.tar.bz2**
- **MM15.tar.bz2**
- **MM16.tar.bz2**
- **MM18.tar.bz2**
- **MM19.tar.bz2**
- **MM20.tar.bz2**
- **MM21.tar.bz2**
- **P02.tar.bz2**

Some helper scripts:

- **split_data.m**
- **get_all_features.m**

A few notes:

1. **The .cnt file.** This contains the continuous EEG recordings. The EMG channel contains the colour sensors. A few of the other

channels aren't usually useful: M1, M2, EKG, and Trigger.

2. **epoch_inds.mat.** This contains the indices for the different trials. There should 132 for most participants. One of them should be 131 and two of the participants should have more trials because the study at the time was a bit longer. These indices do not include trials that were lost because the Kinect sensor wasn't working or because of some mishap.
3. **The kinect_data subdirectory.** You should include all the wav and AnimU files. You can ignore the .3D files in most cases. The AnimU files contain the animation units. Also, include the ID.txt and ID_p.txt files, which contains the order in which the prompts were supposed to be presented and the order in which they were presented in practice (which contains repeats), respectively. You can also usually ignore labels.txt and ID_ind.txt.

If you have any questions, please contact Frank Rudzicz at **frank@cs.toronto.[EDUCATION_SUFFIX]**.