CLASSIFYING PHONOLOGICAL CATEGORIES IN IMAGINED AND ARTICULATED SPEECH

Shunan Zhao[1] *and Frank Rudzicz*[1,2]

[1] Department of Computer Science, University of Toronto; [2] Toronto Rehabilitation Institute-UHN; Toronto, Canada

ABSTRACT

This paper presents a new dataset combining 3 modalities (EEG, facial, and audio) during imagined and vocalized phonemic and single-word prompts. We pre-process the EEG data, compute features for all 3 modalities, and perform binary classification of phonological categories using a combination of these modalities. For example, a deep-belief network obtains accuracies over 90% on identifying consonants, which is significantly more accurate than two baseline support vector machines. We also classify between the different states (resting, stimuli, active thinking) of the recording, achieving accuracies of 95%. These data may be used to learn multimodal relationships, and to develop silent-speech and brain-computer interfaces.

Index Terms— Phonological categories, electroencephalography, speech articulation, deep-belief networks

1. INTRODUCTION

Brain-computer interfaces (BCIs) often involve imagining gross motor movements to move a pointer on-screen. However, some research has attempted to access language centres directly. This has involved using ECoG [1, 2] and neurotrophic electrodes beneath the skull [3] to recreate words or auditory spectra [4] 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. We are interested in discovering solutions that can be applied more generally.

Suppes *et al.* [5] performed whole-word recognition using electroencephalographic (EEG) and MEG data, where participants either silently pronounced words or thought about their meaning. Porbadnigk *et al.* [6] used an HMM to classify between EEG signals associated with the imagined speech of five words with limited accuracy. The order in which the words were presented significantly affected the results, which were above chance for only one of four modes. Previous attempts to classify EEG signals associated with the imagined pronunciation of phonemes often focussed on vowels [7, 8, 9, 10], building on work by Fujimaki *et al.* [11], who identified event-related potentials during the imagined pronunciation of $/a/$. While relevant, these studies did not relate EEG signals to either articulation or acoustics during actual speech production.

2. DATA

2.1. Data Collection

Four female and eight male participants (mean age $= 27.4$, $\sigma = 5$, range $= 14$) were recruited from the University of Toronto campus. All participants were right-handed, had at least some post-secondary education, had no visual, hearing, or motor impairments, and had no history of neurological conditions or drug abuse. Furthermore, 10 of the 12 participants identified North American English as their first language and the remaining 2 spoke North American English at a fluent level, having learned the language at a mean age of 6.

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 $\mathbb{R}[-1..1]$: *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 [12]. 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 kΩ.

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*) [13]. 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.

Naturally, given the impact of movement on EEG, we expect excessive noise in the **speaking** state EEG. Once the participant has finished speaking, one of the investigators would proceed to the next trial. Each prompt was presented 12 times for a total of 132 trials. The phonemic/syllabic prompts were first presented followed by the 4 'Kent' words, and the trials were randomly permuted within each of those two sections. After every 40 trials, the participant was given the opportunity to rest. Data from 4 of the 12 participants were discarded due to unattached ground wires and two participants falling asleep during recording. Ethical approval was obtained from both the University of Toronto and the University Health Network, of which Toronto Rehab is a member.

2.2. Pre-processing

EEG was pre-processed with EEGLAB [14], including removal of ocular artifacts using blind source separation [15]. The data were band-pass filtered between 1 Hz and 50 Hz, and the mean values were subtracted from each channel. We also applied a small Laplacian filter to the data, using the neighbourhood of adjacent channels. The EEG data were segmented into different trials, and each trial was further segmented into the 4 states described above. We discarded 16 trials that did not contain facial features from the Kinect.

2.3. Feature extraction and selection

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 $\pm$ minimum, sum, spectral entropy, energy, and:

$$\text{skewness} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^{3/2}}$$

$$\text{kurtosis} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2} - 3.$$

Here, x_i is a sample of the window, $\bar{x}$ is the mean, and n is the number of samples in the window. We also compute the mean, maximum, minimum, and the sum and difference of the maximum and minimum for the absolute value of the windowed signal. Furthermore, we compute the first and second derivates of the above features. This results in 1197 features for each channel of the segment, for a total of 65,835 features across the 62 channels.

For each audio recording, we measure the same set of features as above. For the facial data, we measure a subset of the above features for each AU, including the mean, maximum, minimum, median, skewness, and kurtosis. We further compute the first and second derivatives for each AU and measure the same set of features.

Due to the high dimensionality of the feature space, particularly for the EEG features, we rank features by their Pearson correlations with the given classes for each task independently and we select the N features with the highest correlation coefficients, where $N \in [5..100]$. Given the multiple tasks and our cross-validation scheme (see section 3), we perform feature selection on every training set independently.

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 Table 1. 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 [16], 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

Table 1. Top 10 highest mean correlations, r, between EEG channels and resulting acoustics.

3. EXPERIMENTS

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 ($K_{quad}(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i^\top \mathbf{x}_j + C)^2$) and **SVM-rbf** uses the radial basis function ($K_{rbf}(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2), \gamma > 0$), given input vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ and optimized parameters C and γ. For both SVMs, we allow 90% of data to violate the Karush-Kuhn-Tucker conditions, if necessary.

In the DBN, weights w_{ij} between nodes i and j, in different layers, are adjusted at iteration $t+1$ with gradient descent given weights at time t according to

$$\Delta w_{ij}(t+1) = w_{ij}(t) + \eta \frac{\delta \log(P(\mathbf{x}))}{\delta w_{ij}}, \tag{1}$$

for empirical learning rate η, where $P(\mathbf{x}) = \frac{\sum_h \exp(-E(\mathbf{x},h))}{Z}$ is the alternate formulation of the probability of $\mathbf{x}$ and $E(v,h)$ is the Gaussian-Bernoulli function

$$E(\mathbf{x}, h) = \sum_{i=1}^{\mathcal{X}} \frac{(\mathbf{x}_i - b_i)^2}{2} - \sum_{i=1}^{\mathcal{X}} \sum_{j=1}^{\mathcal{H}} w_{ij} \mathbf{x}_i h_j - \sum_{j=1}^{\mathcal{H}} a_j h_j, \tag{2}$$

where $\mathbf{x}_i$ is the activation at the i^{th} of $\mathcal{X}$ visible units, h_j is the activation at the j^{th} of $\mathcal{H}$ hidden units, and b_i and a_j are their bias terms. After unsupervised training, we set a linear mapping of the output and 'fine tune' the network in a supervised fashion using class predictions. In all cases, we use one hidden layer whose (bottleneck) size is empirically 25% of the size of the input. We use up to 10 iterations (to avoid overfitting) in the pretraining cycle with a batchsize of $N/4$ (given N observation vectors), a learning rate $\eta = 0.1$, a drop-out rate [17] of 0.5 and the 'cross entropy' objective function $C = -\sum_j d_j \log(p_j)$, empirically chosen, where d_j is the target probability for output j and p_j is the actual probability output of j.

3.1. Classification of phonological categories

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.

Figure 1 shows the average accuracy (with std. error $\sigma/\sqrt{n}$) of classifying $\pm/uw/$ 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. Although the SVM-quad classifier obtains significantly better-than-chance accuracy on $\pm/uw/$, the SVM classifiers, in general, obtain significantly lower accuracy than the DBNs. As suggested by the high $\sigma/\sqrt{n}$ for the SVM classifiers, this may be largely due to the interaction of the classification tasks and the modalities of the data used. Indeed, Table 2 shows that the average accuracies of the SVM-quad classifier varies greatly across these two dimensions. This is further confirmed by an analysis of variance (ANOVA) in Table 3 which not only shows significant linear effects of each of the classifier, test subject, task, and modality on the accuracy of phonological category classification, but also significant interactions between the task and both of the classifier used and the modality of the data.

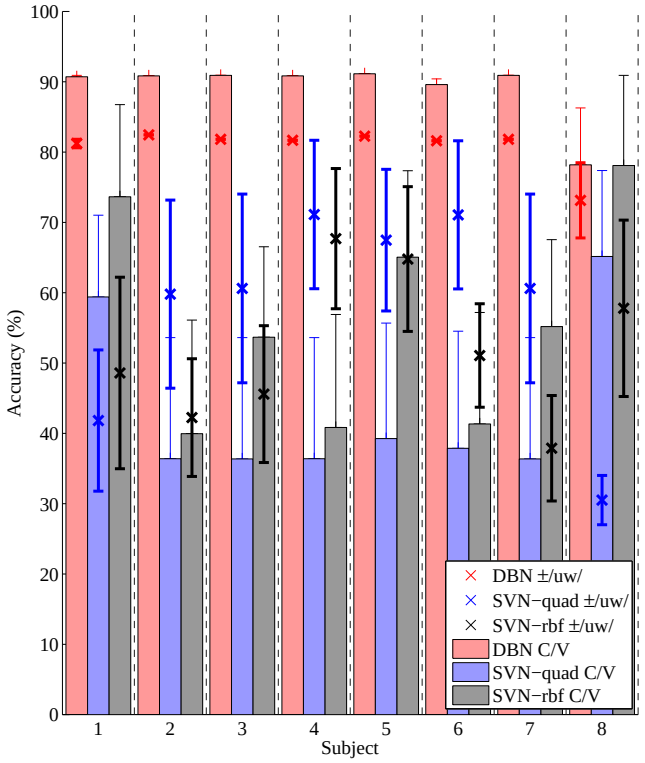

Fig. 1. Average accuracies across models for DBN, SVM-quad, and SVM-rbf classifiers for the $\pm/uw/$ and C/V tasks, across subjects. Error bars are $\sigma/\sqrt{n}$.

	Task				
	C/V	± Nasal	± Bilab.	$\pm/iy/$	$\pm/uw/$
EEG	18.08	63.50	56.64	59.60	79.16
FAC	62.54	48.10	63.73	40.25	20.68
AUD	81.05	40.48	39.98	37.63	18.33
EEG+FAC	72.17	48.41	63.73	56.03	19.60
EEG+AUD	61.13	62.72	39.99	49.15	83.75
ALL	75.72	51.87	63.73	46.01	20.20

Table 2. Average accuracies (%) across modalities and classes given the SVM-quad classifier.

Source	Sum Sq.	F-statistic	p
Classifier	85.44	$F_2 = 3591.28$	< 0.001
Subject	3.34	$F_6 = 47.53$	< 0.001
Task	46.47	$F_4 = 975.58$	< 0.001
Modality	5.77	$F_5 = 97.00$	< 0.001
Classifier×Task	55.38	$F_8 = 581.92$	< 0.01
Task×Modality	7.94	$F_{20} = 33.38$	< 0.01

Table 3. ANOVA of classification accuracies according to main and select interaction effects.

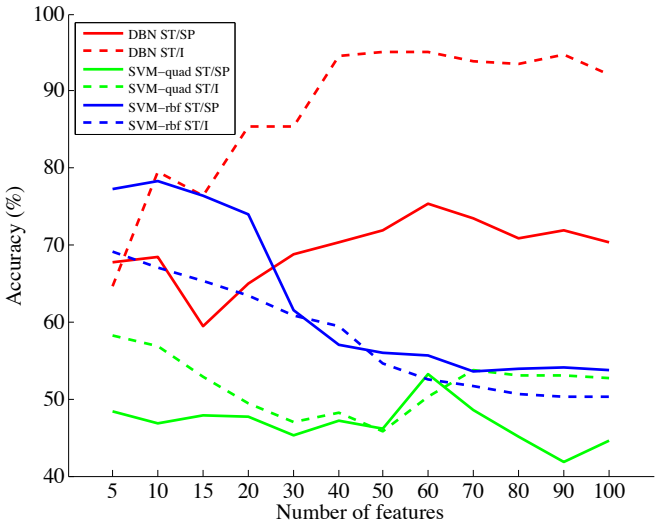

Fig. 2. Average accuracies (%) across number of features for DBN, SVM-quad, and SVM-rbf in the ST/SP and ST/I tasks.

3.2. Classification of mental state

We also classify between the different states of each trial, specifically in three binary tasks: stimulus vs. speaking (**ST/SP**), rest vs. imagined (**R/I**), and stimulus vs. imagined (**ST/I**). We again use DBN and SVM systems, as in section 3.1, with the same hyper-parameters. Initial results yielded average accuracies between 50% to 60%. To improve performance, we concatenate the band-pass filtered data from 6 of the 8 participants and perform independent component analysis (ICA). Due to practical considerations, we do not perform ICA on the data from all participants. Given observed multivariate data S, ICA assumes observations in S are linear mixtures of unknown, statistically independent sources X and computes W in $S = WX$, yielding 64 components, because we include ocular channels.

Using the same feature selection method as in section 2.3, the DBN obtains accuracies of 69%, 56%, 88% for the ST/SP, R/SP, and ST/I tasks, respectively, averaged over all subjects and feature sizes. Figure 2 shows results of two tasks. The high accuracy obtained from classifying between the imagined and speaking states is not surprising, as artifacts related to speech production are present in the EEG data. The DBN clearly outperforms the SVM baselines and scales better with the number of features, up to 60 to 70 features.

4. DISCUSSION

This paper presents the first classification of phonological categories combining acoustic, facial, and EEG data. Usually such multimodality is only possible with expensive magnetoencephalography. Instead, we use an affordable (and portable) Kinect sensor and 64-channel EEG cap, which is a much more viable setup for BCIs. Furthermore, all our reported experiments use leave-one-out cross-validation, so our models are subject-independent and generalizable.

We are continuing to record additional subjects and plan to release the data publicly. Future work includes methods to reconstruct acoustic features from EEG, after Pasley *et al.*'s work with invasive methods [4], potentially towards mapping imagined speech to synthetic speech. There is also recent evidence that the types of linguistic statistics derivable from text corpora are highly indicative of brain activity. For instance, Mitchell *et al.* [18] showed that one can predict fMRI patterns for previously unseen word stimuli given semantic information about those words (e.g., whether they refer to animate objects), demonstrating the relation between neural patterns and the distributional semantics of words. Murphy *et al.* [19] also found that EEG activation patterns encode enough information to discriminate broad conceptual categories. Anderson *et al.* [20] showed strong correlations between image and text-based distributional semantic models and fMRI recordings. Our future work will therefore not only encode correlations among EEG, articulatory, and acoustic features and their phonological categories, but among semantics as well.

5. ACKNOWLEDGEMENTS

This research is funded by the Toronto Rehabilitation Institute, the Natural Sciences and Engineering Research Council of Canada (RGPIN 435874), and a grant from the Nuance Foundation. Data collection was assisted by Selvana Morcos, Aaron Marquis, Chaim Katz, and César Márquez-Chin.

6. REFERENCES

[1] T. Blakely, K.J. Miller, R. P N Rao, Mark D. Holmes, and J.G. Ojemann, "Localization and classification of phonemes using high spatial resolution electrocorticography (ECoG) grids," in *Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE*, Aug 2008, pp. 4964–4967.

[2] Spencer Kellis, Kai Miller, Kyle Thomson, Richard Brown, Paul House, and Bradley Greger, "Decoding spoken words using local field potentials recorded from

the cortical surface," *Journal of Neural Engineering*, vol. 7, no. 5, pp. 1–10, 2010.

[3] Jess Bartels, Dinal Andreasen, Princewill Ehirim, Hui Mao, Steven Seibert, E. Joe Wright, and Philip Kennedy, "Neurotrophic electrode: Method of assembly and implantation into human motor speech cortex," *Journal of Neuroscience Methods*, vol. 174, no. 2, pp. 168–176, 2008.

[4] Brian N. Pasley, Stephen V. David, Nima Mesgarani, Adeen Flinker, Shihab A. Shamma, Nathan E. Crone, Robert T. Knight, and Edward F. Chang, "Reconstructing speech from human auditory cortex," *PLoS ONE*, vol. 10, no. 1, pp. 1–13, 2012.

[5] Patrick Suppes, Zhong-Lin Lu, and Bing Han, "Brain wave recognition of words," *Proceedings of the National Academy of Sciences*, vol. 94, no. 26, pp. 14965–14969, 1997.

[6] Anne Porbadnigk, Marek Wester, Jan Calliess, and Tanja Schultz, "EEG-based speech recognition - impact of temporal effects.," in *BIOSIGNALS*, Pedro Encarnao and Antnio Veloso, Eds. 2009, pp. 376–381, INSTICC Press.

[7] K. Brigham and B.V.K.V. Kumar, "Imagined speech classification with EEG signals for silent communication: A preliminary investigation into synthetic telepathy," in *Bioinformatics and Biomedical Engineering (iCBBE), 2010 4th International Conference on*, June 2010, pp. 1–4.

[8] Michael D'Zmura, Siyi Deng, Tom Lappas, Samuel Thorpe, and Ramesh Srinivasan, "Toward EEG sensing of imagined speech," in *Human-Computer Interaction. New Trends*, JulieA. Jacko, Ed., vol. 5610 of *Lecture Notes in Computer Science*, pp. 40–48. Springer Berlin Heidelberg, 2009.

[9] Daniel E Callan, Akiko M Callan, Kiyoshi Honda, and Shinobu Masaki, "Single-sweep EEG analysis of neural processes underlying perception and production of vowels," *Cognitive Brain Research*, vol. 10, no. 1-2, pp. 173–176, 2000.

[10] Charles S. DaSalla, Hiroyuki Kambara, Makoto Sato, and Yasuharu Koike, "Single-trial classification of vowel speech imagery using common spatial patterns," *Neural Networks*, vol. 22, no. 9, pp. 1334–1339, 2009, Brain-Machine Interface.

[11] N. Fujimaki, F. Takeuchi, T. Kobayashi, S. Kuriki, and S. Hasuo, "Event-related potentials in silent speech," *Brain Topography*, vol. 6, no. 4, pp. 259–267, 1994.

[12] F Sharbrough, GE Chatrian, RP Lesser, H Lüders, M Nuwer, and TW Picton, "American electroencephalographic society guidelines for standard electrode position nomenclature," *Journal of Clinical Neurophysiology*, vol. 8, no. 2, pp. 200–202, 1991.

[13] Ray D. Kent, Gary Weismer, Jane F. Kent, and John C. Rosenbek, "Toward phonetic intelligibility testing in dysarthria," *Journal of Speech and Hearing Disorders*, vol. 54, no. 4, pp. 482–499, 1989.

[14] Arnaud Delorme and Scott Makeig, "EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis," *Journal of Neuroscience Methods*, vol. 134, no. 1, pp. 9–21, 2004.

[15] G. Gomez-Herrero, W. De Clercq, H. Anwar, O. Kara, K. Egiazarian, S. Van Huffel, and W. Van Paesschen, "Automatic Removal of Ocular Artifacts in the EEG without an EOG Reference Channel," in *Signal Processing Symposium, 2006. NORSIG 2006. Proceedings of the 7th Nordic*, June 2006, pp. 130–133.

[16] Friedemann Pulvermüller, Martina Huss, Ferath Kherif, Fermin Moscoso del Prado Martin, Olaf Hauk, and Yury Shtyrov, "Motor cortex maps articulatory features of speech sounds," *Proceedings of the National Academy of Sciences (PNAS)*, vol. 103, no. 20, pp. 7865–7870, 2005.

[17] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 2014.

[18] Tom M Mitchell, Svetlana V Shinkareva, Andrew Carlson, Kai-Min Chang, Vicente L Malave, Robert A Mason, and Marcel Adam Just, "Predicting human brain activity associated with the meanings of nouns," *Science*, vol. 320, no. 5880, pp. 1191–1195, 2008.

[19] Brian Murphy, Marco Baroni, and Massimo Poesio, "EEG responds to conceptual stimuli and corpus semantics," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, Singapore, August 2009, pp. 619–627, Association for Computational Linguistics.

[20] Andrew J. Anderson, Elia Bruni, Ulisse Bordignon, Massimo Poesio, and Marco Baroni, "Of words, eyes and brains: Correlating image-based distributional semantic models with neural representations of concepts," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, Seattle, Washington, USA, October 2013, pp. 1960–1970, Association for Computational Linguistics.