Word Identification using Phonetic Features: towards a method to support multivariate fMRI speech decoding

Tijl Grootswagers[*1], *Karen Dijkstra*[*1], *Louis ten Bosch*[23],
Alex Brandmeyer[4], *Makiko Sadakata*[14]

[1] Artificial Intelligence Department, Radboud University Nijmegen, The Netherlands
[2] Max Planck Institute for Psycholinguistics Nijmegen, The Netherlands
[3]CLS/CLST, Radboud University Nijmegen,The Netherlands
[4]Radboud University Nijmegen, Donders Institute for Brain, Cognition and Behaviour, The Netherlands

{t.grootswagers, karendijkstra}@student.ru.nl,
l.tenbosch@let.ru.nl, {a.brandmeyer, m.sadakata}@donders.ru.nl

Abstract

Nowadays, using state of the art multivariate machine learning approaches, researchers are able to classify brain states from brain data. One of the applications of this technique is decoding phonemes that are being produced from brain data in order to decode produced words. However, this approach has been only moderately successful. Instead, decoding articulatory features from brain data may be more feasible. As a first step towards this approach, we propose a word decoding method that is based on the detection of articulatory features (words are identified from a sequence of articulatory class labels). In essence, we investigated how the lexical ambiguity is reduced as a function of confusion between articulatory features, and a function of the confusion between phonemes after feature decoding. We created a number of models based on different combinations of articulatory features and tested word identification on an English corpus with approximately 70,000 words. The most promising model used only 11 classes and identified 71% of words correctly. The results confirmed that it is possible to decode words based on articulatory features, and this offers the opportunity for multivariate fMRI speech decoding.

Index Terms: fMRI-based speech decoding, articulatory features, lexical access

1. Introduction

Recently, computational power and functional Magnetic Resonance Imaging (fMRI) precision and quality have increased to a level that allows for treating fMRI voxel spaces as pattern recognition problems. Using state of the art multivariate machine learning approaches, researchers are able to classify brain states from fMRI distributed patterns. The mapping from blood oxygen level dependent (BOLD) responses, measured by fMRI[1], to brain states can be learned using multivariate classification algorithms like support-vector machines or neural networks [1]. These patterns contain information corresponding to different cognitive states of the brain, and can be used in conjunction with an appropriately trained classifier to decode unseen BOLD responses into the stimuli that induced these responses [2, 3, 4, 5, 6].

One of the applications that has been explored is decoding phonemes that are being perceived [7] or produced [8, 9] on the basis of single-trial fMRI or electrocorticographic (ECoG) data. While these decoding approaches can be used to learn more about the functional organization of language-related processes in the brain, the application of decoding produced phonemes becomes especially interesting in the case of paralysis: patients with locked-in syndrome are unable to move their muscles, and are thus unable to produce speech. If attempted phoneme production could be decoded, this would offer a potential means of communication through speech even without muscle control.

Classification results obtained by classifying intended phoneme production from signals recorded using ECoG look promising: Some phonemes could be classified with over 60% accuracy. However, performance for many phonemes is still around chance level, thus this approach is still far from an actual application [9].

Instead of trying to find a mapping between brain patterns and phonemes, we investigated the use of acoustic-articulatory features to classify the phoneme from those features. This approach, if successful, would offer two major contributions for a word based brain reading application. First, decoding articulatory features instead of single phonemes greatly reduces the number of classes that need to be distinguished; for instance, the classification results mentioned in the previous section were based on 38 phonemes. The use of fewer classes reduces the probability of misclassification and will generally lead to higher performance in decoding [10]. Representation of the speech signal in terms of articulatory features may also lead to a better and more flexible way of describing the speech in terms of articulatory overlapping events [11, 12, 13, 14]. Second, articulatory features are likely easier to decode than single phonemes because articulatory features cause activation in the (pre)motor cortex for both speech production and speech perception [15, 16, 17, 18, 19].

The question addressed in this paper is to what extent word recognition degrades as a function of degradation of the representation of phonemes as articulatory features to investigate the effect of noisy brain reading output. In our method, we assume

[1]Magnetoencephalography or electroencephalography are commonly used as well for higher temporal resolution but lower spatial resolution than fMRI.

[*]These authors contributed equally

that a lexicon is available in which words (i.e. orthographic forms) are represented by a sequence of phonemes.

We investigated the predictive value of a number of articulatory features by creating classification models based on various combinations of features. These models were compared on the basis of their predictive performance on a 70,000 word English corpus.

We hypothesized to find a clear trade-off between the complexity and the performance of the model, where more complex models (i.e. those containing more articulatory features) would lead to higher predictive performances. Whether or not the performance of a given model is sufficient would depend on the application for which it would be intended. For a multivariate fMRI-decoding application using a model containing 15 classes, the performance should be on the order of 80% words correctly classified. This is used as a benchmark for evaluating the present results.

2. Methods

2.1. Models

Twelve models were created in which phonemes were allocated into different classes based on articulatory features. For consonants, place/manner of articulation and voicing were used and for vowels backness, height and roundness. For each of these features a model was created, and the remaining models were constructed using combinations of features. As diphthongs (also known as gliding vowels) essentially consist of two vowels (they 'glide' from one vowel towards another vowel), for clarity and simplicity's sake, we assigned these to the features of the vowel that they glide towards [2].

Table 1 shows the models that were tested on performance, ordered by complexity. For example, models containing more articulatory features are defined as more complex. The complexity is relevant with respect to the multivariate fMRI-decoding problem, with more complex models requiring a larger number of classes.

[2]How diphthongs should be treated with regard to phonetics is still subject to debate [20, 21].

	consonant features			vowel features			
name	voiced	place	manner	round	height	backns	classes
RV	x			x			4
MB			x			x	11
MH			x		x		12
PB		x				x	13
PH		x			x		14
MBRV	x		x	x		x	18
PBRV	x	x		x		x	21
MHRV	x		x	x	x		21
PHRV	x	x		x	x		24
PBMH		x	x		x	x	27
PBMH	x	x	x	x	x	x	41

Table 1: Specification of the models used in this study, showing which features they combine, and the number of classes that combination yields.

min phonemes	min frequency	# words
0	0	71404
0	1	53766
0	10	28079
0	100	8336
4	0	69050
4	1	51501
4	10	26257
4	100	7242

Table 2: The size of the dictionary when applying filters for frequency and size.

2.2. Database

To test the performance of the models, a dictionary was constructed using information from the Celex lexical database [22]. Specifically, we used the English phonology wordforms database [3]. The Celex database contains word forms, lemma id, relative frequency information and the phonetic transcription (along with additional information beyond the scope of this study). Celex entries containing spaces (e.g. 'accounted[ ]for' and 'cut[ ]off'), were removed, and the remaining 71.404 words served as the dataset. Using this list, we also created datasets containing only words with higher frequencies (1, 10 or 100) and/or words consisting of at least 4 phonemes to investigate the effect of minimum frequency and minimum size of included words on performance. Table 2 shows an overview of the effect of these filters on the number of words in the dictionary.

2.3. Procedure

The procedure used to calculate the performance of a model (a combination of features) can be roughly divided into these three steps:

1. The Celex database was read into a dictionary and all transcriptions were parsed according to the features.
2. Using the unique phoneme features, a new dictionary was created, the model. When an entry (i.e. a sequence of feature labels, representing the phonemes) was inserted into the model, it returned the word that translates into these features with the highest frequency.
3. The percentage of correctly classified words was then calculated by feeding every entry in the dataset into the model. A word was correctly classified if the lemma id of the output equals the lemma id of the input(e.g., "busier" as an input and "busy" as an output would be a correct classification because both words share their lemma id).

Table 3 shows an example of these steps.

[3]A copy can be obtained from `celex.mpi.nl`.

Step 1

Word	Lemma ID	Frequency	IPA	Place of articulation or tongue backness
busy	5794	1012	bɪzɪ	bilabial.near-front.alveolar.near-front
green	19636	2648	gɾin	velar.alveolar.front.alveolar
pity	34026	512	pɪtɪ	bilabial.near-front.alveolar.near-front
ready	37434	2241	ɾɛdɪ	alveolar.front.alveolar.near-front

Step 2

Model input	Model ouput
bilabial.near-front.alveolar.near-front	busy (5794, 1012)
velar.alveolar.front.alveolar	green (19636, 2648)
alveolar.front.alveolar.near-front	ready (37434, 2241)

Step 3

Word	Model input	Model Prediction	Correct?
busy	bilabial.near-front.alveolar.near-front	busy	yes
green	velar.alveolar.front.alveolar	green	yes
pity	bilabial.near-front.alveolar.near-front	busy	no
ready	alveolar.front.alveolar.near-front	ready	yes

Table 3: The procedure to calculate performance using place/backness as features. Note that 'busy' is predicted instead of 'pity'. They have the same features, therefore the prediction of the model is 'busy' as it has a higher frequency. In this example the performance of the model would be 75%.

3. Results

Figure 1 presents the performance of the different models. Detailed information about the models can be found in Table 1. The results show that the lowest performance was obtained using the simplest feature, while the highest performance was obtained using all the features.

This is in line with our hypothesis that more complex models (using more features) would lead to higher predictive performance. One surprising result is that models using manner of articulation as a feature had higher overall performance than those using place of articulation, even though such models had relatively fewer classes to distinguish. These differences can be explained in part by the more even distribution of phonemes over different manners of articulation.

Table 4 presents the effect of filtering the dictionary using word frequency and phoneme count on the overall performance. It shows that the only filter that had a noticeable effect was restricting the corpus to words with at least four phonemes and a frequency of 1000. Note, however, that when using this filter, the size of the dictionary gets reduced from 71,404 to 7,274 words. Therefore, the total number of words that are correctly classified is still larger for the unfiltered dictionary (50,786 words).

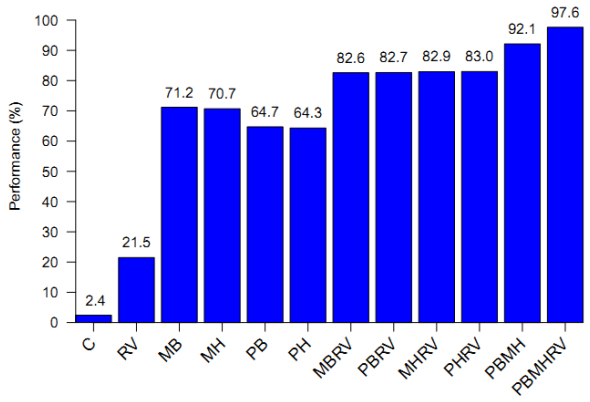

Figure 1: The performance for the models. The labels in this plot refer to the labels used in Table 1, with exception of *C*, in which only a consonant/vowel distinction is made.

min phon.	min freq.	correct	total	performance
0	0	50786	71404	71.18
0	1	37743	53766	70.25
0	10	19788	28079	70.53
0	1000	6009	8336	72.08
4	0	50416	69050	73.08
4	1	37382	51501	72.64
4	10	19460	26257	74.18
4	1000	5751	7274	79.41

Table 4: The number of correctly identified lemma's, total number of words and the performance for each of the different filters, using a model based on the *manner(consonant)* and *backness(vowel)* features.

4. Discussion

This study investigated whether methods based on the decoding of articulatory features instead of individual phonemes would be sufficient for word identification. We predicted to find a clear trade-off between the complexity and the performance of different models. Such a trade-off is indeed evident in Figure 1, though the effect diminishes with a higher number of classes. 'Manner of articulation' performed better as a feature than 'place of articulation' even though it has fewer classes within the feature.

We set a benchmark for performance at 80% correct for a model containing 15 classes. The model that performed closest to this benchmark uses the feature 'manner of articulation' for consonants and tongue position ('backness') for vowels. This model has 12 classes and achieved a performance of 71.2%. While the performance is notably lower than that of the bench-

mark, it also uses fewer classes. Ultimately the choice of features for a decoding application depends not only on the performance of the model on word identification but also on the decoding performance of the features on brain data. This can be investigated in a future study using fMRI to decode phoneme features from phoneme production in subjects. For now these findings suggest that classifying on articulatory features rather that individual phonemes yields a high word classification performance, while severely reducing the number of classes. This indicates that such an fMRI study would be worthwhile. Such results also support the idea that the use of phonetic features as opposed to single phonemes would be more effective in multivariate fMRI studies that focus on word decoding.

We also investigated the effect of filtering the word list on the basis of phoneme count and/or frequency. While filtering the word list increases performance in some cases, it also reduced the total number of words that could be identified. For example, when restricting the corpus to words with at least four phonemes and a frequency of at least 100, the performance increases to 79% compared to 71% using no filter. At the same time this filter reduces the size of the word list from around seventy thousand to seven thousand words. These results lead us to the conclusion that it is not beneficial to filter the data in this manner.

The reported performance could potentially be improved upon by various means of post processing. For example, the probability distribution of phonemes can be taken into account. Using such a distribution, which can be obtained from the same word corpus, more information is provided about most likely phoneme in a given context. Note that the fact that we used frequency information for the prediction already makes use of the probability distribution, though it can be extended from the word level to the phoneme level, thereby increasing its precision. Another way to improve word-prediction performance would be to use contextual information, such as word bigram frequency. Further study is needed to assess these post-processing methods.

Finally, a number of important limitations need to be considered with regards to the findings in this study. First, to be able decode words in multivariate fMRI studies we have assumed that words are pronounced without any form of reduction. In a restricted experimental setup, it would be possible to instruct participants to clearly pronounce words. However, in natural speech, reduction plays an important role, with around 20% of word utterances containing one or more deleted segments [23, 24].

Secondly, the spatial and temporal resolution of fMRI is possibly too low to detect the subtle differences in the features we described. For example, the difference (in brain activation) between a front and near-front tongue position (as in the vowel 'backness' feature), might be too small to pick up with fRMI [25]. Using ECoG, spatial resolution is much more precise (although very invasive), allowing different features to be more reliably distinguished [26, 9, 27]. However, noisy detection of a feature can be compensated by precise detection of another phoneme. Cue trading describes the effect that a phonetic distinction is the result of a combined effect of several different acoustic cues. These cues contribute to the perception of a phonetic distinction. In general, a trading relation among the cues can be demonstrated in an identification task as long as the speech stimuli are phonetically ambiguous [11]. Taken together, cue trading can improve the translation from utterances to articulatory features of the phonemes.

The most important limitation lies in the fact that the feature division used in this study is arguably incomplete and possibly not the most optimal choice for fMRI studies. It has been shown that phonemes can be classified using many features other than the ones used in this study (for an overview of the debate on phonetic features, see van Oostendorp et al. 2011, chapters 17,19,21,27) . Future research should therefore concentrate on which features perform well in multivariate fMRI decoding. Finally, it is worth mentioning that the method presented here can be extended to other languages, though it is very likely that the optimal feature set differs across languages or even across individuals [30].

5. References

[1] K. Norman, S. Polyn, G. Detre, and J. Haxby, "Beyond mind-reading: multi-voxel pattern analysis of fMRI data," *Trends in cognitive sciences*, vol. 10, no. 9, pp. 424–430, 2006.

[2] J. Haynes, K. Sakai, G. Rees, S. Gilbert, C. Frith, and R. Passingham, "Reading hidden intentions in the human brain," *Current Biology*, vol. 17, no. 4, pp. 323–328, 2007.

[3] Y. Miyawaki, H. Uchida, O. Yamashita, M. Sato, Y. Morito, H. Tanabe, N. Sadato, and Y. Kamitani, "Visual image reconstruction from human brain activity using a combination of multiscale local image decoders," *Neuron*, vol. 60, no. 5, pp. 915–929, 2008.

[4] M. Mur, P. Bandettini, and N. Kriegeskorte, "Revealing representational content with pattern-information fMRI–an introductory guide," *Social cognitive and affective neuroscience*, vol. 4, no. 1, p. 101, 2009.

[5] B. Murphy, M. Poesio, F. Bovolo, L. Bruzzone, M. Dalponte, and H. Lakany, "Eeg decoding of semantic category reveals distributed representations for single concepts," *Brain and language*, vol. 12, no. 1, pp. 12–22, 2011.

[6] S. Nishimoto, A. Vu, T. Naselaris, Y. Benjamini, B. Yu, and J. Gallant, "Reconstructing visual experiences from brain activity evoked by natural movies," *Current Biology*, vol. 21, no. 19, pp. 1641–1646, 2011.

[7] E. Formisano, F. De Martino, M. Bonte, and R. Goebel, ""Who" Is Saying" What"? Brain-Based Decoding of Human Voice and Speech," *Science*, vol. 322, no. 5903, p. 970, 2008.

[8] T. Blakely, K. Miller, R. Rao, M. Holmes, and J. 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*. IEEE, 2008, pp. 4964–4967.

[9] J. Brumberg, E. Wright, D. Andreasen, F. Guenther, and P. Kennedy, "Classification of intended phoneme production from chronic intracortical microelectrode recordings in speech-motor cortex," *Frontiers in Neuroscience*, vol. 5, 2011.

[10] C. Bishop, *Pattern recognition and machine learning*. New York: Springer, 2006.

[11] L. Ten Bosch and O. Scharenborg, "Modeling cue trading in human word recognition," in *Proceedings of the 13th Annual Conference of the International Speech Communication Association (Interspeech 2012)*, 2012.

[12] L. Ten Bosch, "Speech variation and the use of distance metrics on the articulatory feature space," in *Speech Recognition and Intrinsic Variation Workshop*, 2006.

[13] J. Frankel, M. Wester, and S. King, "Articulatory feature recognition using dynamic bayesian networks," *Computer Speech & Language*, vol. 21, no. 4, pp. 620–640, 2007.

[14] K. Livescu, A. Bezman, M. Borges, L. Yung, O. Cetin, J. Frankel, S. King, X. Xhi, and L. Lavoie, "Manual transcription of con-

versational speech at the articulatory feature level," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 4. IEEE, 2007, pp. IV–953.

[15] A. D'Ausilio, F. Pulvermüller, P. Salmas, I. Bufalari, C. Begliomini, and L. Fadiga, "The motor somatotopy of speech perception," *Current Biology*, vol. 19, no. 5, pp. 381–385, 2009.

[16] I. Meister, S. Wilson, C. Deblieck, A. Wu, and M. Iacoboni, "The essential role of premotor cortex in speech perception," *Current Biology*, vol. 17, no. 19, pp. 1692–1696, 2007.

[17] F. Pulvermüller, M. Huss, F. Kherif, F. del Prado Martin, O. Hauk, and Y. Shtyrov, "Motor cortex maps articulatory features of speech sounds," *Proceedings of the National Academy of Sciences*, vol. 103, no. 20, pp. 7865–7870, 2006.

[18] S. Wilson, A. Saygin, M. Sereno, and M. Iacoboni, "Listening to speech activates motor areas involved in speech production," *Nature neuroscience*, vol. 7, no. 7, pp. 701–702, 2004.

[19] R. Wise, J. Greene, C. Büchel, and S. Scott, "Brain regions involved in articulation," *The Lancet*, vol. 353, no. 9158, pp. 1057–1061, 1999.

[20] J. Durand, *On the phonological status of glides: the evidence from Malay*. Dordrecht, Holland: Foris Publications, 1987.

[21] J. Padgett, "Glides, vowels, and features," *Lingua*, vol. 118, no. 12, pp. 1937–1955, 2008.

[22] R. Baayen, R. Piepenbrock, and L. Gulikers, "The celex lexical database (release 2)[cd-rom]," *Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania [Distributor]*, 1995.

[23] M. Ernestus, "Voice assimilation and segment reduction in casual Dutch," *A corpus-based study of the phonology-phonetics interface. LOT, Utrecht*, 2000.

[24] K. Johnson, "Massive reduction in conversational American English," in *Spontaneous speech: Data and analysis*, K. Yoneyama and K. Maekawa, Eds. Tokyo: National Institute for Japanese Language, 2004, pp. 29–54.

[25] M. van Gerven, J. Farquhar, R. Schaefer, R. Vlek, J. Geuze, A. Nijholt, N. Ramsey, P. Haselager, L. Vuurpijl, S. Gielen *et al.*, "The brain–computer interface cycle," *Journal of Neural Engineering*, vol. 6, p. 041001, 2009.

[26] K. Bouchard, N. Mesgarani, K. Johnson, and E. Chang, "Functional organization of human sensorimotor cortex for speech articulation," *Nature*, vol. 495, pp. 327–334, 2013.

[27] B. Pasley, V. David, N. Mesgarani, A. Flinker, S. Shamma, N. Crone, R. Knight, and E. Chang, "Reconstructing speech from human auditory cortex," *PLoS biology*, vol. 10, no. 1, p. e1001251, 2012.

[28] International Phonetic Association, *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge University Press, 1999.

[29] M. van Oostendorp, C. Ewen, E. Hume, and K. Rice, *The Blackwell Companion to Phonology*. Wiley-Blackwell, 2011.

[30] A. Marchal and W. Hardcastle, "ACCOR: Instrumentation and database for the cross-language study of coarticulation," *Language and Speech*, vol. 36, no. 2-3, pp. 137–153, 1993.