Auditory Stimulus Discrimination from MEG Data

Tuomas Lukka
Harvard Society of Fellows
Harvard University
78 Mount Auburn St.
Cambridge, MA 02138
`lukka@fas.harvard.edu`

Bernd Schoner
Physics and Media Group
MIT Media Laboratory
20 Ames Street
Cambridge, MA 02139
`schoner@media.mit.edu`

Alec Marantz
Department of Linguistics and Philosophy
MIT
Cambridge, MA 02139
`marantz@mit.edu`

Abstract

We consider Magnetoencephalographic (MEG) data in a signal detection framework. Our data set consists of responses evoked by the voiced syllables /bÒ/ and /dÒ/ and the corresponding voiceless syllables /pÒ/ and /tÒ/. The data yield well to principal component analysis (PCA), with a reasonable subspace in the order of three components out of 37 channels. To discriminate between responses to the voiced and voiceless versions of a consonant we form a feature vector by either matched Æltering or wavelet packet decomposition and use a mixture-of-experts model to classify the stimuli. Both choices of a feature vector lead to a signiÆcant detection accuracy. Furthermore, we show how to estimate the onset time of a stimulus from a continuous data stream.

1 INTRODUCTION

Magnetoencephalography (MEG) uses SQUID technology to measure the small magnetic Æelds induced by electrical activity in the brain. Sensitive to roughly the same neural activity as EEG/ERP, MEG offers some advantages in data analysis and source localization. Although multi-sensor MEG systems recording magnetic Øux at kilohertz sampling rates provide an incredibly rich source of data about brain activity, most current analysis techniques make use of only a fraction of the data collected (see, e.g., Aulanko et al. 1993, Fujimaki et al. 1995). The most common approach to the analysis of stimulus evoked responses with MEG is to record 100 or more time-locked responses to the same stimulus, average these responses, and then perform single dipole source analysis on the averaged waves. While averaging serves to reduce noise and to remove ™background∫ activity unrelated to the stimulus, dipole modeling loses the statistics of the averaging and proves a data-wasteful method of reducing the dimensionality of MEG data.

In this paper, we introduce a new way of looking at MEG data from a signal processing and discrimination perspective. We show that it is possible to build a classiÆer system to discriminate between different stimuli from the unaveraged data. Principal component analysis is used to reduce the dimensionality of the data without loss of signiÆcant information.

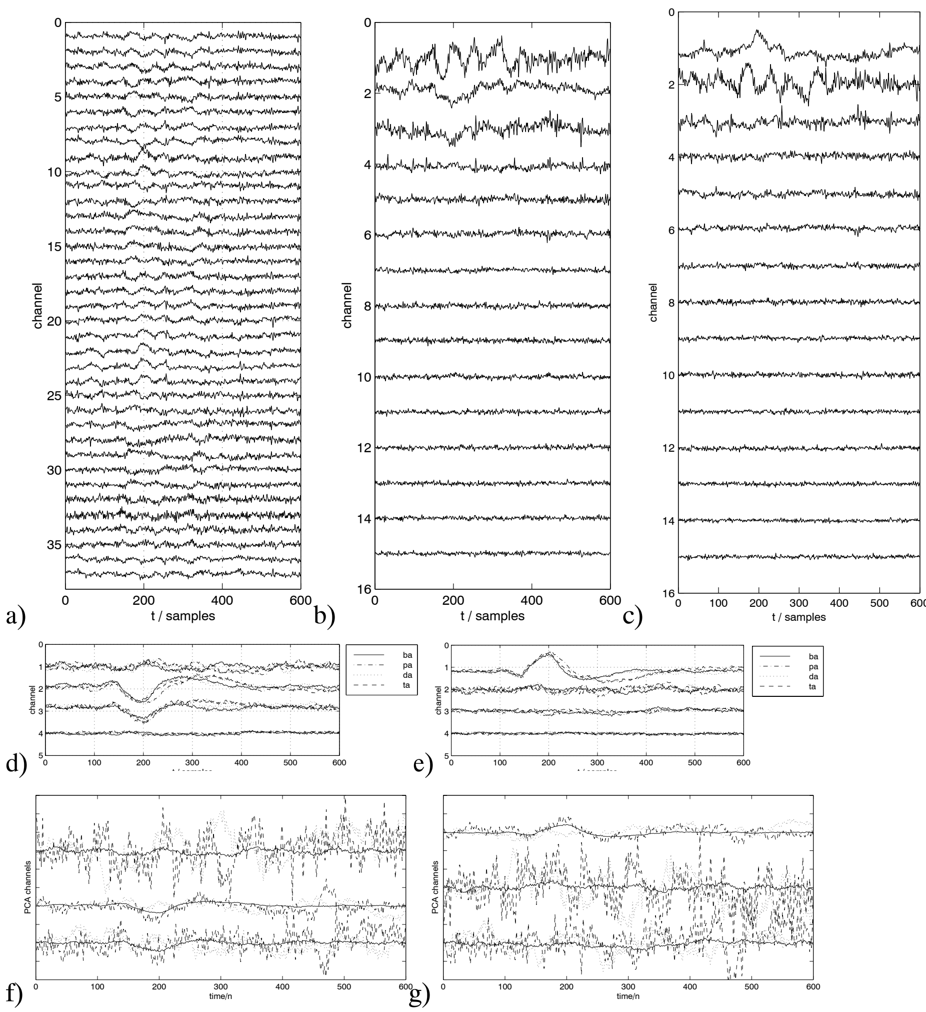

Figure 1: MEG data. a) All channels of one raw epoch. b) single-epoch-deÆned PCA and c) average-response-deÆned PCA of the same data. Average responses to the four different stimuli after d) single-epoch-deÆned PCA and e) average-response-deÆned PCA. A single epoch and the average superimposed, in f) single-epoch-deÆned PCA and g) average-response-deÆned PCA.

2 DATA

The data were collected as part of the experiment reported in Poeppel et al. 1996, where detailed description of the stimuli and data collection techniques may be found. BrieØy, the stimuli were 4 synthesized 300ms syllables, /bÒ/, /pÒ/, /dÒ/, and /tÒ/. The voiced-voiceless pairs /bÒ/-/pÒ/ and /dÒ/-/tÒ/ differ acoustically only in ™voicing onset timeʃ with the Ærst member of each pair containing 20ms of ™aspirationʃ prior to the onset of the (voiced) vocalic portion of the syllable and the second member containing 80ms of aspiration.

MEG recordings were taken in a magnetically shielded room using a 37-channel system with SQUID-based Ærst-order gradiometer sensors. The sensor array was centered over the left auditory cortex and the 4 stimuli were presented to the right ear 100 times each, in pseudorandom order at a variable ISI of 1 to 1.5 seconds. 400 epochs of 600ms were recorded, time-locked to stimulus onset, with a 100ms pre-stimulus interval. The sampling rate was 1041.7 Hz with a bandwidth of 400 Hz.

3 ALGORITHMS

Our analysis of the MEG data proceeds in three steps. In the Ærst we reduce the dimensionality of the data from 37 to the order of three by principal component analysis (PCA) (see Oja 1983). The second step is concerned with analyzing the reduced data in a time-dependent way with either matched Æltering or wavelet packet analysis. From this step we obtain a low-dimensional feature vector which we use in step three to do the actual discrimination with a local experts type model.

3.1 PCA

From Fig. 1 a) it is clear that the incoming signals are not independent. The PCA transformation reduces this redundancy by Ænding the best orthogonal linear subspace[1]. This is useful for compact visualization (Fig. 1 b) and c)) as well as for reduction of computational effort in the subsequent manipulation of the data. The effect of noise on many algorithms is also reduced as the signal is concentrated to the Ærst channels.

The transformation is deÆned by the eigenvectors of the covariance matrix of the data (see Oja 1983). With the MEG data, we can deÆne the covariance matrix either by the usual covariance over single epochs or by the covariance of the averaged responses to the stimuli.

The difference between the two deÆnitions is illustrated by Fig. 1 b)±e): in the data transformed by the PCA deÆned by the single epochs, the response is split between channels 2 and 3 whereas the average-deÆned PCA reduces the amount of noise by concentrating the response in the Ærst channels, and therefore seems preferable. However, if the response varies from epoch to epoch (e.g. if the response to /dÒ/ were to depend on some other variable such as the phase of the background brain waves), the covariance matrix of the single epochs should be used as otherwise information might be lost.

3.2 MATCHED FILTERING

It is well known that time-correlating noisy signals with the known `true signal' leads to efÆcient estimators and detectors of linear time signals (matched Æltering, see e.g. Brown and Hwang 1992). We calculate the convolutions of the data with the time-reversed average responses to the stimuli. These convolved signals peak whenever a stimulus occurs so the

[1] We considered ICA for this task but the data seem too noisy as well as low-dimensional for it to help much more than the average-deÆned PCA.

onset time of the stimulus can be estimated. Alternatively, the values of the convolved signals at a known onset time can be used as a feature vector for discriminating between different stimuli.

Because matched Æltering is linear, it should perform equally well with both the raw and the PCA transformed data. However, in practice the data set is large and performing the computation only on the largest principal components improves the efÆciency markedly.

3.3 WAVELET PACKETS

The windowed training signals are expanded in an orthonormal wavelet packet basis that assigns coefÆcients in a time-frequency grid (see e.g. Coifman and Saito 1994). The transform is based on the repeated application of a quadrature mirror Ælter (Daubechies 6 was used in this work) followed by a downsampling step so that at each transform level the coefÆcients represent the time domain behavior of a particular frequency band.

In the Ærst approach a low-dimensional orthonormal subset of coefÆcients is chosen to maximize the square distance discrimination measure D_{SD}:

$$D_{SD} = (\overline{w}_{i1} - \overline{w}_{i2})^2/(\sigma_{w_{i1}}\sigma_{w_{i2}}); \tag{1}$$

where $\overline{w}_{ic}$ denotes the averaged coefÆcient i of stimulus class c, and $\sigma_{w_{ic}}$ is the standard deviation of coefÆcients w_{ic}.

In the second approach we select a optimal complete orthonormal basis from the time frequency grid. The discriminant power of the squared and normalized coefÆcients is evaluated in terms of the symmetrized relative entropy (Kullback-Leibler distance) between either two stimuli (for discrimination) or a `stimulus' and a `non-stimulus' window (for onset detection). The algorithm for selecting the basis is described in detail in Coifman and Wickerhauser (1992). The expansion and basis selection is done for all selected PCA channels.

3.4 CLUSTER-WEIGHTED DETECTION

We use Gaussian-weighted local experts in a Cluster-Weighted Modeling framework (Gershenfeld et al 1997) to discriminate between stimulus classes based on the feature vectors obtained in the previous sections. As opposed to conventional density approximation techniques, each local expert represents a probability distribution in the joint input-output space. The likelihood of a class C_i given a particular feature vector $\overline{x}$ is

$$p(C_i|\overline{x}) = \sum_j p(C_i|E_j)p(E_j|\overline{x}) \tag{2}$$

where E_j is the expert j and

$$p(E_j|\overline{x}) = \frac{p(\overline{x}|E_j)p(E_j)}{\sum_k p(\overline{x}|E_k)p(E_k)}. \tag{3}$$

The domain of each expert is characterized by the probability distribution $p(E_j|\overline{x})$ which in this work is Gaussian. The model is trained by the Expectation Maximization algorithm (Gershenfeld et al.).

For comparison a statistical discriminator based on the Kullback-Leibler distance is tested. The complete set of normalized coefÆcients of new data is compared in probability to the averaged energy distribution of the different reference stimuli. The data is classiÆed according to the best match.

4 RESULTS

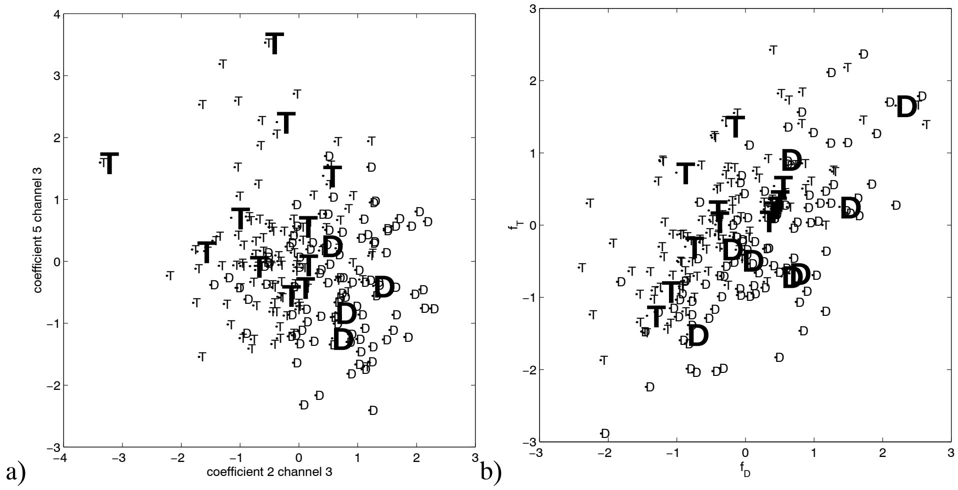

Figure 2: Two dimensions of the feature vector for the bÒ/dÒ discrimination: a) A/WP b) A/MF. The small letters refer to the actual sample points; the large letters are the centers of the local experts. The letter T refers to the voiceless and D to the voiced version of the consonant.

4.1 VOICED/VOICELESS DISCRIMINATION

We applied the above methods to the data described in section 2. Two different windows with different offsets were tested, both 256 samples long. The offset for the second window is beyond the acoustic difference between the stimuli, which ensures that we are detecting based on brain activity and not simply a MEG recording of the actual stimulus.

As seen in Table 1, it is possible to get a statistically signiÆcant detection accuracy for voiced/voiceless discrimination. The number of local experts N_e in the detector was found by cross-validation. Figure 2 shows slices of example input spaces to the mixture of experts classiÆer. We show the results for one speciÆc subject. The data taken from a second subject led to nearly identical results. There were no signiÆcant differences between matched Æltering and the wavelet packet decomposition methods, nor was there signiÆcant difference between different quadrature mirror Ælters (Haar, CoiØet and Daubechies Ælter were tested). Two coefÆcients were used to form the wavelet coefÆcient feature vector, as using more coefÆcients didn't improve performance and led to overÆtting.

Discrimination between the two voiced consonants (/bÒ/-/dÒ/) or the two voiceless consonants (/pÒ/-/tÒ/) was impossible with the available data. The results indicate that more MEG channels are needed for discrimination in this case (see Fig. 1).

4.2 ONSET DETECTION

Fig. 3 shows the results of using a matched Ælter as well as Kullback-Leibler distance estimator on some out-of-sample data. Due to the lack of an actual continuous data stream, chained single epochs were used for this experiment. From these signals, the onset times of stimuli can be estimated by some peak detection algorithm. It is clear that the Kullback-Leibler distance is much more sensitive to noise. The periodic structure of the signal between the onsets is mostly due to the periodicity of the background brain waves.

As a proof-of-principle experiment the local performance of the matched Ælter onset esti-

Table 1: Results for discriminating voiced/voiceless syllables. The last four columns are the detection results, the numbers before/after the slash are the number of correct/incorrect classiÆcations.

			Window	ClassiÆcation			
			Offset	Training		Testing	
Syllables	Method	N_e [a]	(samples)	C_1	C_2	C_1	C_2
bÒ/pÒ	A[b]/WP[c]	10	105	52/18	62/8	25/5	21/9
bÒ/pÒ	S[d]/WP	4	105	50/20	53/17	25/5	21/9
bÒ/pÒ	A/KL[e]	N/A	205	59/11	63/7	25/5	18/12
bÒ/pÒ	A/MF[f]	15	205	52/18	56/14	19/11	25/5
dÒ/tÒ	A/WP	4	205	45/25	51/19	19/11	20/10
dÒ/tÒ	A/WP	2	105	50/20	49/21	21/9	22/8
dÒ/tÒ	A/MF	15	205	57/13	65/5	21/9	25/5

[a] Number of clusters (local experts)
[b] Average-deÆned PCA
[c] Wavelet packet coefÆcient and cluster-weighted detection
[d] Single-epoch-deÆned PCA
[e] Kullback-Leibler distance discrimination
[f] Matched Æltering discrimination and cluster-weighted detection

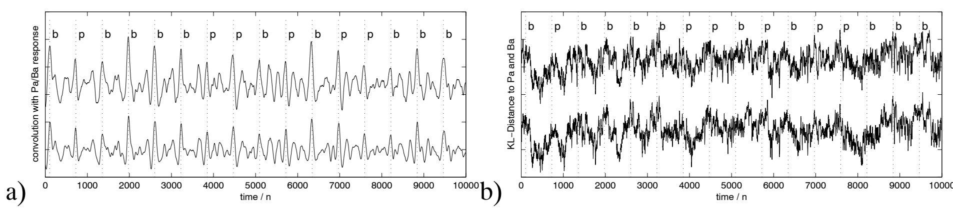

Figure 3: Two example signals from the onset detection. a) matched Æltering b) Kullback-Leibler distance

mator was estimated on 60 out-of-sample epochs (mixed /pÒ/-/bÒ/ stimuli) by taking the onset time to be the local maximum within 100 samples of the true onset in either direction. The estimator worked with an average bias of -0.6 and a standard deviation of 15.3 time samples.

5 CONCLUSIONS AND FUTURE WORK

The fact that the nonlinear wavelet packet approaches and a simple matched Ælter work equally well indicates that for the current case where the stimulus is always the same the response is essentially linear. However, it is not clear whether this would be the case e.g. if there were several different speakers for each stimulus. Also, given the relatively small number of recording channels and the apparent subtlety of the contrastive response to the test stimuli, more training samples will be required to fully test the non-linear methods. We are arranging to collect at least 400 sample responses to each of several stimuli similar to the /dÒ/-/tÒ/ stimuli already employed. Continuously recorded data including responses to these stimuli will be used to test the signal detection ability of the model derived from these samples.

We are planning to develop an event-based maximum likelihood model for interpreting the data. Such a model would be able to attribute parts of the signal to ™uninteresting events∫

based on information in the other channels. It should then be possible to obtain a much purer signal (e.g. canceling out the background brain waves and heartbeats) and thereby further improve the accuracy of the onset estimation and stimulus discrimination.

Since MEG provides an extremely rich source of data on brain function, it is important for cognitive neuroscience to develop analysis techniques for extracting signal from noise and for identifying crucial features of evoked responses. For computational neuroscience, the data provide a very good test case for a variety of neural algorithms, as they are time-dependent, multidimensional, noisy, but regular. In this paper, we have only just begun the task of mining MEG data.

Acknowledgements

The authors would like to thank Neil Gershenfeld and Josh Smith for their comments on this manuscript. T.L. would like to thank the Emil Aaltonen foundation for Ænancial support during the initial part of this work.

References

Aulanko, R., Hari, R., Lounasmaa, O.V., Näätänen, R. & Sams, M. (1993) Phonetic invariance in the human auditory cortex. *NeuroReport* **4** 1356-1358

Brown, R. G. & Hwang, P. Y. (1992) Introduction to Random Signals and Applied Kalman Filtering. New York: Wiley.

Coifman, R. R. & Saito, N. (1994) Constructions of local orthonormal bases for classiÆcation and regression. *Comptes Rendus Acad. Sci. Paris, Serie I* **319** (2) 191-196.

Coifman, R.R. & Wickerhauser, M.V. (1992) Entropy-based algorithms for best basis selection. *IEEE Transactions on Information Theory*, **38**(2): 713-718.

Fujimaki, N., Hirata, Y., Kuriki, S. & Nakajima, H. (1995) Event-related magnetic Æelds at latencies of over 400 ms in silent reading of Japanese Katakana meaningless words. *Neuroscience Research* **23** 419-422

Gershenfeld, N., Metois, E., & Schoner, B. (1997) Cluster Weighted Modeling, to be published.

Oja, E. (1983) Subspace Methods of Pattern Recognition. New York: J. Wiley, Letchworth, England: Research Studies Press.

Poeppel, D., Yellin, E., Phillips, C. , Roberts, T.P.L., Rowley, H.A., Wexler, K. & A. Marantz (1996) Task-induced asymmetry of the auditory evoked M100 neuromagnetic Æeld elicited by speech sounds, *Cognitive Brain Research* **4**: 231-242.