Phoneme Discrimination from MEG Data

Tuomas Lukka
Harvard Society of Fellows
Harvard University
78 Mount Auburn St.
Cambridge, MA 02138
lukka@iki.fi

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
Massachusetts Institute of Technology
Cambridge, MA 02139
marantz@mit.edu

Abstract

We treat Magnetoencephalographic (MEG) data in a signal detection framework to discriminate between different phonemes heard by a test subject. Our data set consists of responses evoked by the voiced syllables /bæ/ and /dæ/ and the corresponding voiceless syllables /pæ/ and /tæ/. 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 filtering or wavelet packet decomposition and use a mixture-of-experts model to classify the stimuli. Both choices of a feature vector lead to a significant 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 fields 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 flux 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. [1, 6]). 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. This kind of analysis is interesting from a clinical point of view, when locating a particular function in the brain is important. However, 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 classifier system to discriminate between different stimuli from the un-averaged data. Principal component analysis is used to reduce the dimensionality of the data without loss of significant information and some different detection algorithms are used to discriminate between responses in the subject caused by different phonemes.
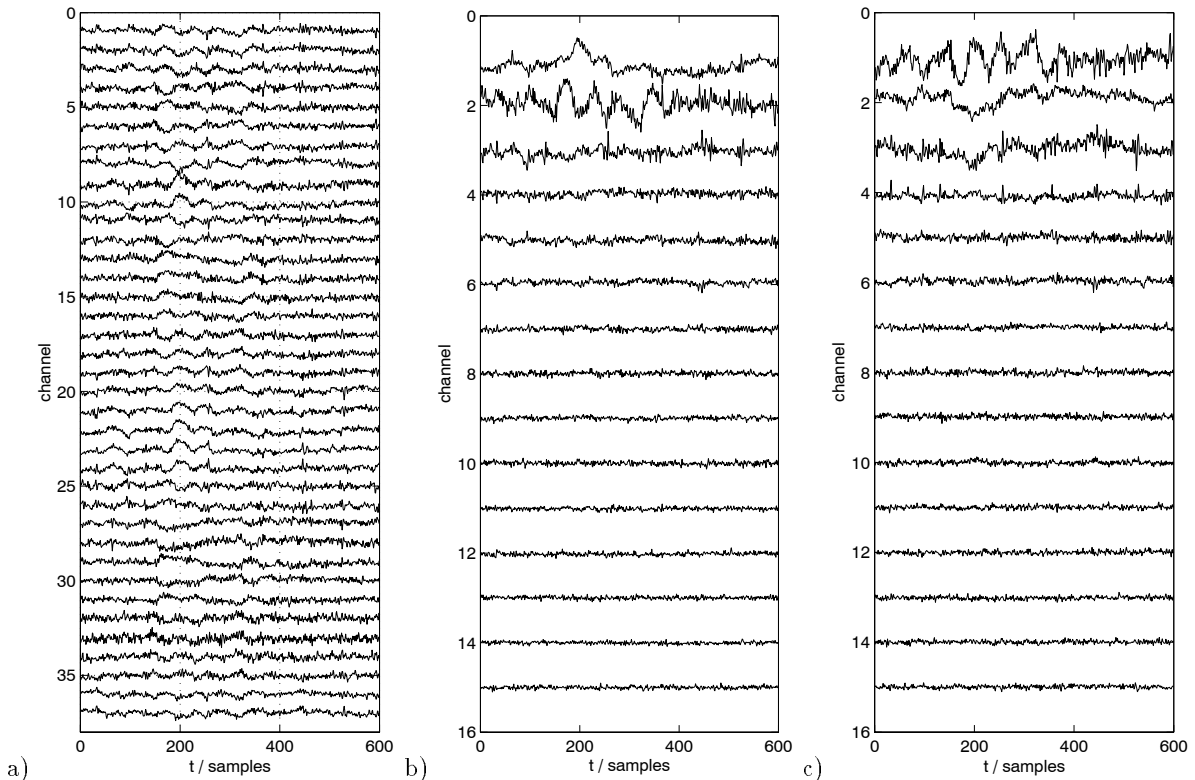
Figure 1: MEG data. a) All channels of one raw epoch. b) average-response-defined PCA and c) single-epoch-defined PCA of the same data (15 principle components).

2 Data

The data were collected as part of the experiment reported in [12], where detailed description of the stimuli and data collection techniques may be found. Briefly, the stimuli were 4 synthesized 300ms syllables, /bæ/, /pæ/, /dæ/, and /tæ/. The voiced-voiceless pairs /bæ/-/pæ/ and /dæ/-/tæ/ differ acoustically only in "voicing onset time," with the first member of each pair containing 20ms of "aspiration" 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 first-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 pseudo-random 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. The example of the data in Fig. 1 is representative — obviously quite noisy, with several channels peaking slightly after the stimulus onset.

3 Algorithms

Our analysis of the MEG data proceeds in three steps. In the first we reduce the dimensionality of the data from 37 to the order of three by principal component analysis (PCA) (see [11]). The second step
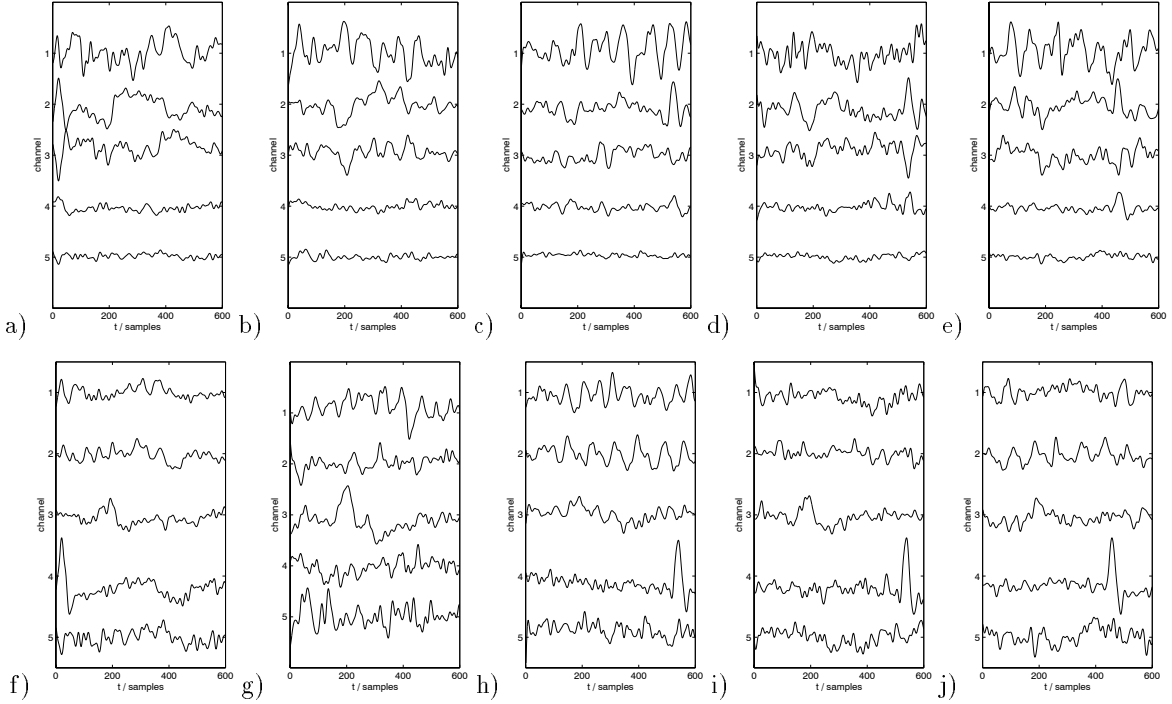
Figure 2: Recorded MEG epochs stimulated by /dæ/,/pæ/,/tæ/,/bæ/ and /pæ/. a)-e) PCA transformed (single epoch defined)responses. f)-j) Same epochs ICA transformed as suggested in Makeig et al. 1996 and 1997. Some events come out clearly, such as the heart beat in channel 4 and the stimulus response in channel 3; however, the whitening required by the algorithm has increased the noise levels.

is concerned with analyzing the reduced data in a time-dependent way with either matched filtering 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 same shape of peak is seen in many channels. The PCA transformation reduces this redundancy by finding the best orthogonal linear subspace. 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 by leaving out the redundant low-energy channels.

The transformation is defined by the eigenvectors of the covariance matrix of the data ([11]). With the MEG data, we can define the covariance matrix either over single epochs

$$v_{\text{sing, } c_1 c_2} = \sum_s \sum_{e \in E_s} \sum_t x_{ec_1t} x_{ec_2t} \tag{1}$$

or over averaged responses to the stimuli

$$v_{\text{avg, } c_1 c_2} = \sum_s \sum_t (\sum_{e \in E_s} x_{ec_1t})(\sum_{e \in E_s} x_{ec_2t}), \tag{2}$$
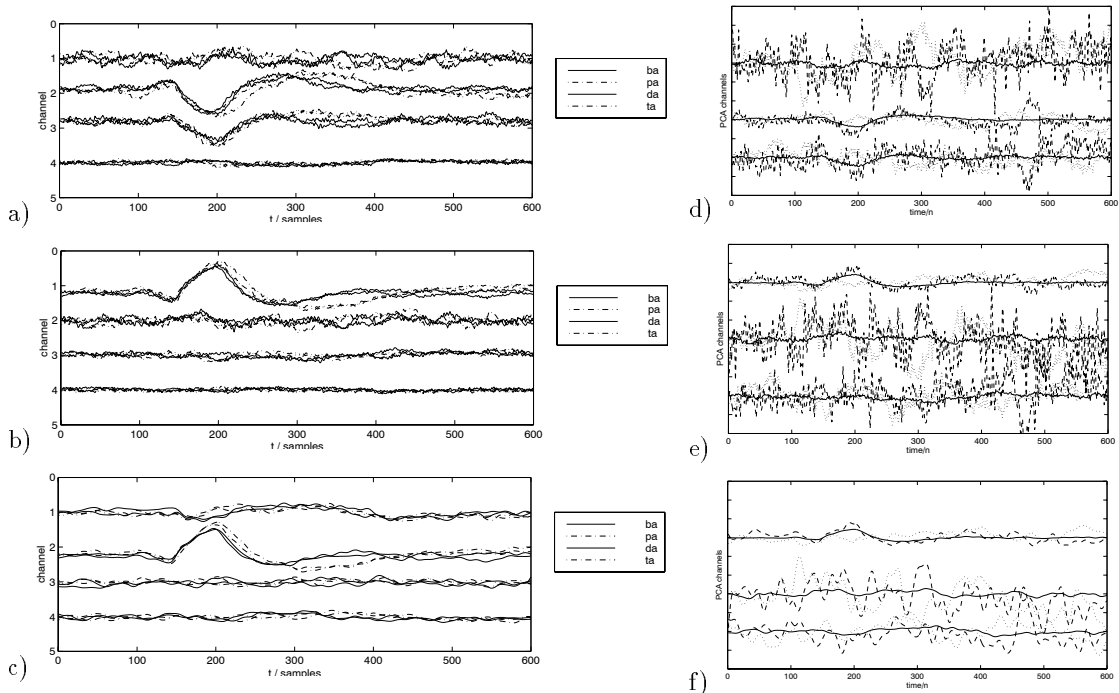
Figure 3: Average responses to the four different stimuli after a) single-epoch-defined PCA, b) average-response-defined PCA and c) ICA transform (low-pass filtered at 60 Hz). A single epoch and the average superimposed, in d) single-epoch-defined PCA, e) average-response-defined PCA and f) ICA transformed data.

where x_{ect} is the zero-average data, e is the epoch, c is the channel, t is the time step, s is a type of stimulus, E_s is the set of epochs with stimulus s and N_s is the cardinality of E_s (we leave out a constant factor from both formulas as it has no effect on the PCA transform).

The difference between the two definitions is illustrated in Fig. 1 and Fig. 3: in the data transformed by the PCA defined by the single epochs, the response is split between channels 2 and 3 whereas the average-defined PCA reduces the amount of noise by concentrating the response in the first channels, and therefore seems preferable. However, if the response varies from epoch to epoch (e.g. if the response to /dæ/ 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 when the number of channels is cut after the PCA.

Independent component analysis (ICA) [9, 10, 8] has recently gained popularity in the signal processing community. It works basically by taking out the orthogonality restriction of PCA and using higher-order statistics to obtain potentially more meaningful components. Figure 2 f)–j) show the results of an ICA transformation on an selection of epochs. Some events come out clearer than after the PCA transform, for example the heart beat in one of the channels. However, for noisy data such as ours, ICA can also increase the effect of noise and make classification of signals more difficult. In a limited number of preliminary trials with ICA we did not observe any improvement over other methods.

3.2   Matched filtering

It is well known that time-correlating noisy signals with the original signal leads to efficient estimators and detectors of linear time series (matched filtering, see e.g. [2]). In a similar approach we estimate a
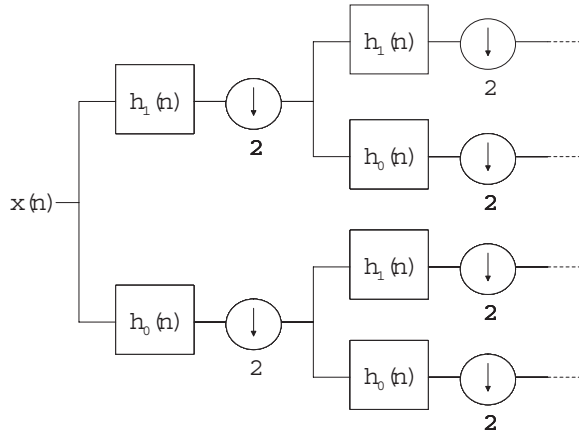
Figure 4: Filtering steps in the discrete wavelet packet transform. $h_0(n)$ and $h_1(n)$ are the half band low-pass and high-pass filters, $2 \downarrow$ stands for down-sampling by a factor 2.

noise free response by averaging over the training epochs and correlate incoming signals

with these 'true responses':

$$C_s(t) = \sum_{\tau=1}^{T} \overline{s}_\tau \cdot s_{t+\tau} \tag{3}$$

where T is the detectable length of the response and $\overline{s}$ is the averaged response. The signal $C_s(t)$ peaks when a stimulus is applied and hence the onset time can be detected. Moreover, the convolution with average responses of different stimuli at known onset times can be used to discriminate between different stimuli.

If the noisy signals are not constructed from the exact same 'true response' every time but there is a distribution of true responses, it can be that simply choosing the stimulus type whose average response correlates best with the sample may not yield the best detection results. In this kind of situation, applying a non-linear detector can improve the results. A non-linear detector may also use non-linear effects between channels.

Because matched filtering is linear, it should perform equally well with both the raw and the PCA transformed data. In practice the data set is large and performing the computation only on the largest principal components improves the efficiency markedly — discarding the low-energy channels has only a minimal effect on the results.

3.3   Wavelet packets

The windowed training signals are expanded in an orthonormal wavelet packet basis that assigns coefficients in a time-frequency grid (see e.g. [3]). The transform is based on the repeated application of a quadrature mirror filter (Daubechies 6 was used in this work) followed by a down-sampling step as illustrated in Fig. 4.

After each filtering step the block of coefficients describes the time behavior of more and more refined frequency bands. Fig. 5 shows the first 194 bins of a Wavelet packet in time and frequency where the bins denote the average energy difference between the two stimulus classes. It can be seen how the discriminating power, originally distributed over the entire time interval, is concentrated in very few frequency bins after the transform is applied.
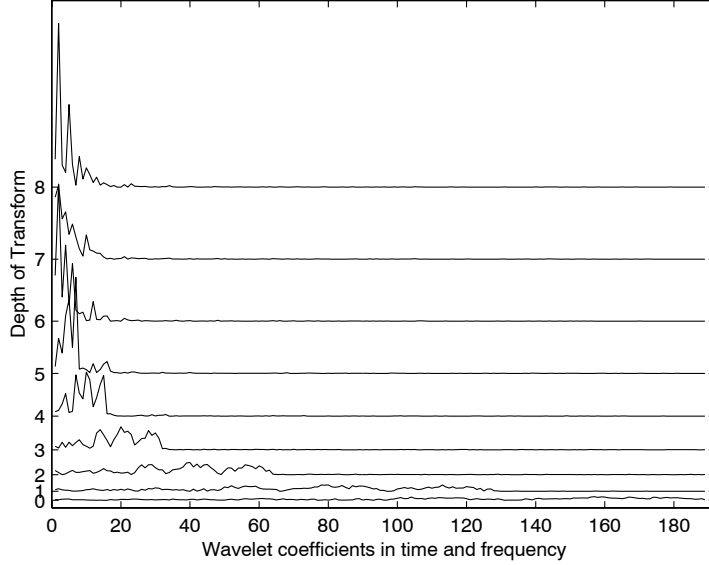
Figure 5: Wavelet packet transform (averaged discriminating energy): The y-axis refers to the depth of the transform while the x-axis represents the sub-bands, ordered from left to right. Hence the 0th-order data shows a pure time-domain picture and the 8th-order transform gives a pure frequency representation.

A key problem is the selection of a reasonable number of coefficients to form the feature vector. In a first approach an orthonormal subset of coefficients is chosen to maximize the square distance discrimination measure D_{SD}:

$$D_{SD} = (\bar{w}_{i1} - \bar{w}_{i2})^2 / (\sigma_{w_{i1}} \sigma_{w_{i2}});$$ (4)

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

In a second approach we select a optimal complete orthonormal basis from the time frequency grid. The discriminant power of the squared and normalized coefficients 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):

$$D_{KL} = \sum_i \bar{w}_{i1} \log \frac{\bar{w}_{i1}}{\bar{w}_{i2}} + \bar{w}_{i2} \log \frac{\bar{w}_{i2}}{\bar{w}_{i1}}$$ (5)

From the orthogonal coefficients $\bar{w}_i$, those are picked that maximize D_{KL}. See [4] for a detailed description of the algorithm. Expansion and basis selection are done for all PCA channels that show a significant signal to noise level.

3.4   Cluster-weighted detection

We use Gaussian-weighted local experts in a Cluster-Weighted Modeling framework [7] to discriminate between stimulus classes. In this framework each local kernel c_j represents a distribution over classes y_i, such that the likelihood of a particular output y_i given a feature vector $\mathbf{x}$ is

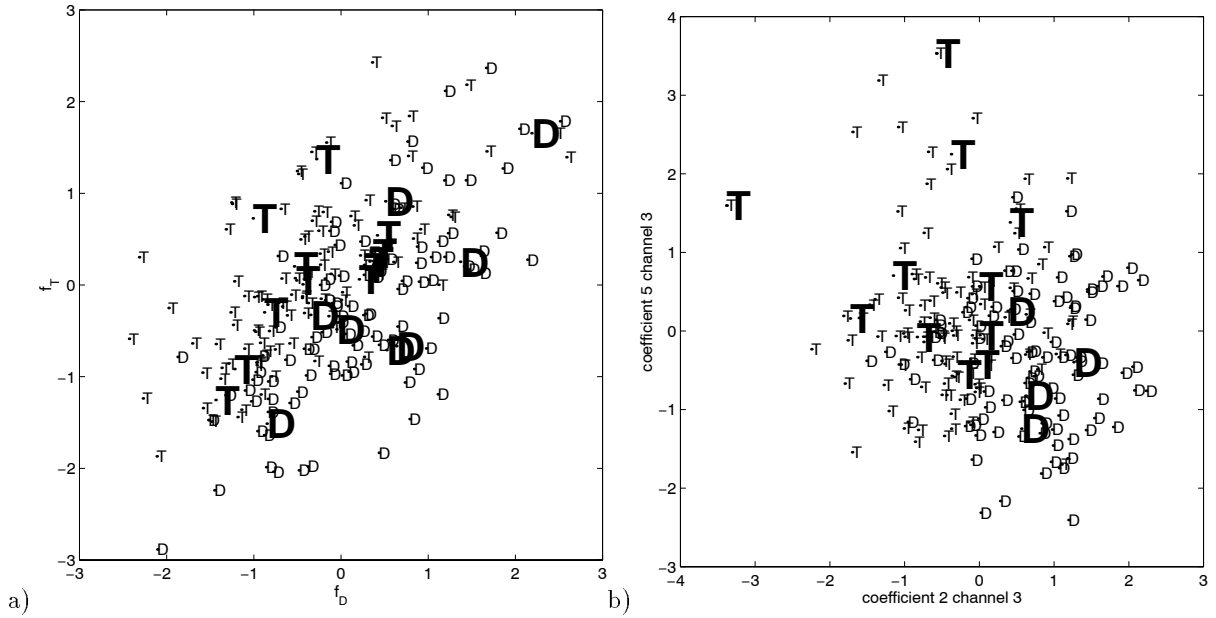$$p(y_i|\mathbf{x}) = \sum_j p(y_i|c_j)p(\mathbf{x}|c_j)p(c_j)$$ (6)

Figure 6: Two dimensions of the feature vector for the bæ/dæ discrimination: a) A/MF b) A/WP. 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.

where $p(c_j)$ is the unconditioned probability of kernel c_j, $p(\mathbf{x}|c_j)$ is taken to be a multivariate Gaussian $N(\mu_j, \mathbf{P}_j)$ that describes the domain of influence of a cluster and $p(y_i|c_j)$ is a probability table characterizing the output distribution of kernel c_j. The classification is done by choosing the y_i that maximizes the probability $p(y_i|\mathbf{x})$.

The model is trained by the Expectation Maximization algorithm which maximizes the likelihood of the data by iterating between an E-step and an M-step [5, 7]. In the **E-step** the current model is assumed correct and the data distribution is computed according to

$$ p(c_j|\mathbf{x}, y_i) = \frac{p(y_i|c_j)p(\mathbf{x}|c_j)p(c_j)}{\sum_k p(y_i|c_k)p(\mathbf{x}|c_k)p(c_k)} \quad . \tag{7} $$

In the **M-step** the data distribution is assumed correct and the data likelihood is maximized. The new model parameters become

$$
\begin{aligned}
p(c_j) &= \frac{1}{N}\sum_n p(c_j|y_n, \mathbf{x}_n) \\
\mu_j &= \frac{\sum_n \mathbf{x}_n\, p(c_j|y_n, \mathbf{x}_n)}{\sum_n p(c_j|y_n, \mathbf{x}_n)} \\
[\mathbf{P}_j]_{kl} &= \frac{\sum_n (x_{k,n} - \mu_{k,j})(x_{l,n} - \mu_{l,j})\, p(c_j|y_n, \mathbf{x}_n)}{\sum_n p(c_j|y_n, \mathbf{x}_n)} \\
p(y_i|c_j) &= \frac{\sum_{n|y_n = y_i} p(c_j|y_n, \mathbf{x}_n)}{\sum_n p(c_j|y_n, \mathbf{x}_n)} \quad .
\end{aligned}
\tag{8}
$$

See Fig. 6 for an illustration of the input-space showing labeled data points and clusters.

3.5 Kullback-Leibler distance detection

For comparison with the cluster-weighted detector a statistical discriminator based on the Kullback-Leibler distance is tested. The complete set of normalized coefficients of new data is compared in probability to the averaged energy distribution of the different reference stimuli (see Equation 5). The data is classified according to the best match.

4 Results

4.1 Voiced/voiceless discrimination

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 classifications.

Syllables	Method	$N_e{}^a$	Window Offset (samples)	Classification Training C_1	C_2	Testing C_1	C_2
bæ/pæ	A^b/WP^c	10	105	52/18	62/8	25/5	21/9
bæ/pæ	S^d/WP	4	105	50/20	53/17	25/5	21/9
bæ/pæ	A/KL^e	N/A	205	59/11	63/7	25/5	18/12
bæ/pæ	A/MF^f	15	205	52/18	56/14	19/11	25/5
dæ/tæ	A/WP	4	205	45/25	51/19	19/11	20/10
dæ/tæ	A/WP	2	105	50/20	49/21	21/9	22/8
dæ/tæ	A/MF	15	205	57/13	65/5	21/9	25/5

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

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 significant 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 classifier. We show the results for one specific subject. The data taken from a second subject led to nearly identical results. There were no significant differences between matched filtering and the wavelet packet decomposition methods, nor was there significant difference between different quadrature mirror filters (Haar, Coiflet and Daubechies filter were tested). Two coefficients were used to form the wavelet coefficient feature vector, as using more coefficients didn't improve performance and led to over-fitting.

Discrimination between the two voiced consonants (/bæ/-/dæ/) or the two voiceless consonants (/pæ/-/tæ/) was impossible with the available data. The results indicate that more MEG channels are needed for discrimination in this case (see Fig. 1) or that the subject's internal discrimination between these consonants is not reflected in the MEG data..
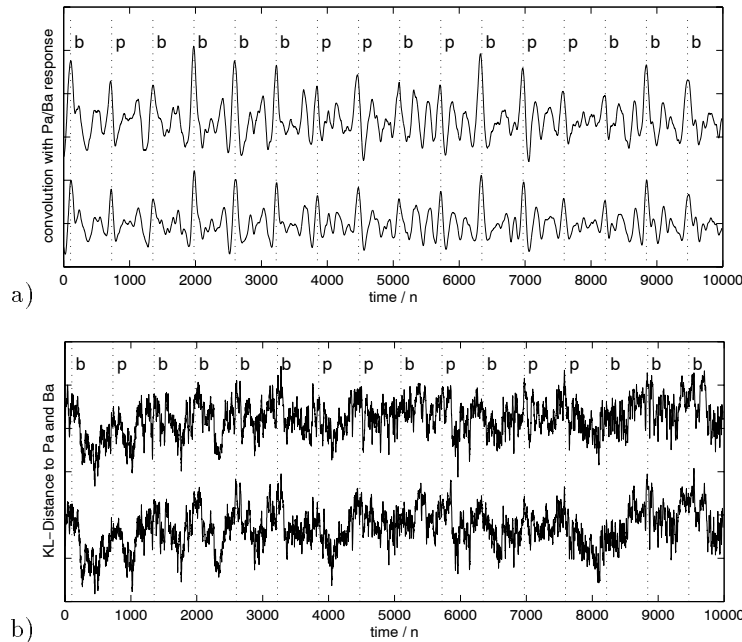
Figure 7: Two example signals from the onset detection. a) matched filtering b) Kullback-Leibler distance

We also made some tests of using ICA-transformed data to discriminate between the different cases but we were not able to get better detection results than with PCA components.

4.2 Onset detection

The average response of the processing techniques described above can be used in a slightly modified way to detect the presence and the onset of a stimulus in a continuous data stream. The convolution of signal and reference epoch peaks whenever a stimulus occurs. Similarly the onset can be estimated based on the wavelet expansion, in which case the best basis is defined with respect to the discriminating power between 'stimulus event' and 'zero event'.

Fig. 7 shows the results of using a matched filter 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 filter onset estimator was estimated on 60 out-of-sample epochs (mixed /pæ/-/bæ/ 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.

Another way of estimating stimulus onsets is to pick out the ICA channel that corresponds to the response. It is clear from Fig. 2 that this approach could work straightforwardly.

5 Conclusions and future work

The fact that the nonlinear wavelet packet approaches and a simple matched filter 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 if, for example, there were several different speakers for each stimulus. Given the relatively small number of recording channels and the apparent subtlety of the contrastive response to the test stimuli, more training samples would be required to fully test the non-linear methods.

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.

One future possibility would be 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.

References

[1] AULANKO, R., HARI, R., LOUNASMAA, O., NÄÄTÄNEN, R., AND SAMS, M. Phonetic invariance in the human auditory cortex. *Neuroreport 4* (1993), 1356–1358.

[2] BROWN, R., AND HWANG, P. *Introduction to Random Signals and Applied Kalman Filtering.* New York, 1992.

[3] COIFMAN, R., AND SAITO, N. Constructions of local orthonormal bases for classification and regression. *Comptes Rendus Acad. Sci. Paris, Serie I 2* (1994), 191–196.

[4] COIFMAN, R., AND WICKERHAUSER, M. Entropy-based algorithms for best basis selection. *IEEE Transactions on Information Theory 38* (1992), 713–718.

[5] DEMPSTER, A., LAIRD, N., AND RUBIN, D. Maximum Likelihood From Incomplete Data via the EM Algorithm. *J. R. Statist. Soc. B 39* (1977), 1–38.

[6] FUJIMAKI, N., HIRATA, Y., KURIKI, S., AND NAKAJIMA, H. Event-related magnetic fields at latencies of over 400 ms in silent reading of japanese katakana meaningless words. *Neuroscience Research 23* (1995), 419–422.

[7] GERSHENFELD, N., SCHONER, B., AND METOIS, E. Cluster-weighted modelling for time series analysis. *Nature 397* (1999), 329–332.

[8] MAKEIG, S., AND AL. Ica matlab package for psychophysiological data, 1998. http://www.cnl.salk.edu/scott/ica-download-form.html.

[9] MAKEIG, S., BELL, A., JUNG, T.-P., AND SEJNOWSKI, T. Independent component analysis of electroencephalographic data. In *Advances in Neural Information Processing Systems*, D. Touretzky, M. Mozer, and M. Hasselmo, Eds., vol. 8. MIT Press, Cambridge, MA, 1996, pp. 145--151.

[10] MAKEIG, S., JUNG, T.-P., GHAHREMANI, D., BELL, A., AND SEJNOWSKI, T. Blind separation of auditory event-related brain responses into independent components. *Proc. Natl. Acad. Sci USA 94* (1997), 10979--10984.

[11] OJA, E. *Subspace Methods of Pattern Recognition.* Research Studies Press, New York, 1983.

[12] POEPPEL, D., YELLIN, E., PHILLIPS, C., ROBERTS, T., ROWLEY, H., WEXLER, K., AND MARANTZ, A. Task-induced asymmetry of the auditory evoked m100 neuromagnetic field elicited by speech sounds. *Cognitive Brain Research 4* (1996), 231--242.

Biographies:

Tuomas Lukka

`lukka@iki.fi`

Tuomas Lukka received his Master's and PhD from the University of Helsinki in 1995 in physical chemistry under the direction of Prof. Lauri Halonen. Since 1996 he is Junior Fellow at the Harvard Society of Fellows, working on both neural and symbolic-statistical learning systems and free software for scientific computing on the Linux operating system.

Bernd Schoner

`schoner@media.mit.edu`

Bernd Schoner received engineering diplomas (M.Sc.) in Electrical Engineering from the Rheinische Wesfälische Technische Hochschule Aachen, Germany, and in Industrial Engineering from École Centrale de Paris, France, both in 1996. Since then he has been a Ph.D. candidate at the MIT Media Laboratory working with Prof. Neil Gershenfeld on the prediction and analysis of driven dynamical systems. His main research interests include machine learning, statistical inference, and the application of these techniques to problems in computer music and musical synthesis.

Alec Marantz

`marantz@mit.edu`

Alec Marantz received the B.A. in psycholinguistics from Oberlin College in 1978 and the Ph.D. in linguistics from M.I.T. in 1981. He is currently Professor of Linguistics in the Department of Linguistics and Philosophy at M.I.T. His research interests include the syntax and morphology of natural languages, linguistic universals, and the neurobiology of language. He is currently involved in revising morphological theory within linguistics and in exploring MEG techniques to uncover how the brain processes language.