ABSTRACT
Classification of EEG-signals is error-prone, due to the small differences in the measurements and the inherent presence of continuing brain dynamics. Often these dynamics are denoted as noise, and in this view, classification is difficult, due to the small signal to noise ratio.

We investigate how multiple trials of our EEG data can be used to increase the classification rate. Two schemes are used to combine the measurements: i) combine probabilities of individual classifications; ii) average measurements before classification. The number of trials that was used was either fixed or flexible. Flexible means that as many trials as needed are used to get a certain reliability of the classification. It is found that combining probabilities works best for large variances and a flexible number of samples, and averaging a flexible number of measurements works best for small variance.

The validation of the method is tested on an EEG data set in which a subject listened to two different rhythms. On a single trial, the classification rate was 80%, a classification rate of about 90% was achieved using the average of 2 trials, and a classification rate of approximately 95% was found for 3 trials. This coincided well with the predictions.

1. INTRODUCTION
For humans, keeping track of time and temporal structure is an essential skill [1]. It allows an economic focus of perceptual attention and is crucial for well-coordinated and synchronised movements. Rhythm arises if this temporal structure is marked by discrete events. Due to its importance, evolution has equipped us with sophisticated mechanisms for processing rhythmic structures [2]. The simultaneous activation of large neuronal populations in such auditory processing can be measured as Event-Related Potentials (ERP’s), as they are time-locked to the eliciting stimulus [3].

The traces invoked by the rhythm may be extracted from the electroencephalogram (EEG) signals, and these signals allow for a direct detection of perceived temporal patterns [4]. Voluntarily imagined rhythms can be used for direct Brain-Computer Interfacing (BCI), using rhythms as coding to output classifiable signals. This BCI systems may be applied as communication prosthesis, e.g. enhancing the quality of life for ALS and locked-in patients [5].

EEG traces are different between trials due to the inherently present brain dynamics. This makes the classification based on a single trial error prone. In critical situations such as in control, it is of the utmost importance to have a reliable classification. A fixed number of trials can be combined to get a more reliable classification, but multiple samples might not be needed if a clear distinction can be made in one of the first trials. Therefore, we will investigate the effect of using a variable number of trials on the classification rate for our EEG recordings.

In this paper we will start in section 2 by investigating different methods of combining samples of computer generated Gaussian distributions. The goal of this section is to investigate how different combining methods react in different situations and what the idiosyncrasies of these methods are. In section 3 the predictive value of the methods is tested on real data. Finally, a conclusion is drawn in section 4.

2. APPROACH
The distributions we use to test the combination techniques are assumed to be Gaussian. Two classes are considered that give rise to the conditional probabilities \( P(x|z_1) \) and \( P(x|z_2) \) in which \( x \) denotes some measurement and \( z_i \) the class. The parameters of the Gaussian are \( \mu_1, \sigma_1 \) and \( \mu_2, \sigma_2 \) respectively. If the distributions \( P(x|z_1) \) and \( P(x|z_2) \) are not well separated, samples can be combined in several ways to decrease the classification error. The different schemes investigated are:

- first combine measurements, then classify; and first classify, then combine classification probabilities.
2.1. Fixed number of samples

In order to better classify samples, we can take the average of several measurements. In the EEG community this can be seen as to first make an ERP, and classify based on this result. If \( n \) samples of a Gaussian distribution are averaged, the new distribution is again a Gaussian, but the variance changes from \( \sigma^2 \) to \( n^{-1}\sigma^2 \) [6]. Due to this decrease of variance, the overlap between the two distributions is decreased, and as a consequence the error lessend (if the means are different). The overlap between two Gaussians is illustrated in figure 1. In this figure, the classes’ original conditional distributions are given by the solid lines. The overlap between these two distributions is depicted as the light gray area. The conditional distributions that result after the averaging of two measurements are given by the dashed lines. The new overlap is the dark gray area. It can be seen that the new overlap, and therefore the classification error, is decreased. The symbol used to refer to this method in figures is \( \sum n \). \( n \) is substituted for the number of samples over which the average is calculated.

The second scheme used to increase performance is to first classify, and combine the probabilities of the individual classifications. For two independently and identically distributed samples it holds:

\[
P(z_i|x_1, x_2) = \frac{p(x_1, x_2|z_i)P(z_i)}{p(x_1, x_2)}
= \frac{p(x_1, x_2|z_i)P(z_i)}{p(x_1, x_2|z_1)P(z_1) + p(x_1, x_2|z_2)P(z_2)}
= \frac{p(x_1|x_i)p(x_2|z_i)P(z_i)}{p(x_1|z_1)p(x_2|z_1)P(z_1) + p(x_1|z_2)p(x_2|z_2)P(z_2)},
\]

which can be easily extended to the combination of more probabilities. The symbol used to refer to this method in figures is \( \Pi n \). \( n \) is substituted for the number of samples over which the probability is combined.

2.1.1. Threshold

In situations where the difference which class is more probable is small, we refrain from classifying, and leave the result undetermined. For this goal, we use a threshold, and if both probabilities are below it, we reject the classification. The probability of a class, given a measurement can be obtained by Bayes’ rule, and in our two classes problem becomes [7]:

\[
P(z_i|x) = \frac{p(x|z_i)P(z_i)}{p(x|z_1)P(z_1) + p(x|z_2)P(z_2)}.
\]
and therefore, the expected classifications per sample equals the acceptance rate.

The influence of the threshold on the classification rate and the expected classifications per sample is determined to illustrate some effects that can be encountered in practice. The scheme that averages over measurements before classification is exemplified first. To calculate the rates mentioned, the initial step is to find the locations in which one of the two posterior probabilities is equal to the threshold, $t$. These locations can be found by solving:

$$-\ln(\sigma_1) - \frac{(x - \mu_1)^2}{2\sigma_1^2} = \pm \ln\left(\frac{t}{1-t}\right) -\ln(\sigma_2) - \frac{(x - \mu_2)^2}{2\sigma_2^2},$$

which can be solved by the cubic formula. The $\pm$ is in this equation so that all four solutions are found. Two of these solutions are given as dots in figure 2, the other solutions occur at negative values of $x$. When these location are known, the overlap and the amount of classified data can be calculated from the cumulative distribution function.

Figure 3 shows results for $\mu_1 = 0, \mu_2 = 1, \sigma_1 = 0.5$ and $\sigma_2 = 0.5$. In the left part of the figure, both the classification rate and the expected classifications per sample are plotted as function of the threshold. If the threshold goes to one, the classification rate also goes to one, while the expected classifications per sample goes to zero: all classification are correct, but no classifications are made. In the right part of the figure, the expected classifications per sample is plotted as function of the classification rate.

The overlap for these two distributions is calculated to be 16%. This is consistent with the intersection of the classification rate and the left $y$-axis at 84%. At this threshold of 0.5 all samples are classified, and it can be observed that averaging over more samples indeed gives a better classification rate. Furthermore, it can be seen that an equal performance can be gotten if one uses one sample and a threshold. More interesting is to note that for a given error rate, it is better to classify only when one exceeds a threshold than to use several samples, as on average the number of classified samples is larger, except for classification rates near one. This is clear from the right part of the figure.

The results of figure 3 are for equal variances. However, it will often be the case that the variances are different. The same plots are made for standard deviations 6 and 7 respectively. These large values are chosen, as they better reflect the situation in which two classes are difficult to distinguish, common in EEG research. The results are given in figure 4. This figure is different in general shape from the previous figure.

The classification rate that is found for a threshold of 0.5, is of course much smaller than what was found previously. Although averaging over samples improves the classification rate, the improvement is marginal because the new overlap is still large. Interesting to observe, is that for a threshold larger than 0.52, the classification rate is better when classified by one sample than when using an average over several samples! This sounds counterintuitive, but due to the slightly smaller variance, a region is classified that was previously ignored, and this results in a large error.

A further obvious difference is the kink in the classification rate from a significance of approximately 0.55. With the help of figure 2 this can be understood. In this figure it can be seen that for a measurement between zero and one, the posterior probability does not reach one, but remains at some smaller value. If the threshold is increased, there will be a point when the region on the left in the figure is no longer classified at all. This happens at the kink in the curve.

In the right part of the figure the expected classifications per sample is plotted as function of the classification error. The curved gray lines on the right of this figure are the results of the previous distributions and are included to show the difference. Based on this figure, one can deduce that it is better to classify with only one sample and use a threshold, than to average over a fixed number of samples.

The combination of samples can also be done by the combination of their probabilities (1). The same figures are constructed as for averaging. However, since the required integrals can only be solved numerically [8], the figures are constructed completely numerically. Two set of 10000 samples were created, and the expected classifications per
sample and classification rate were evaluated on these sets. This procedure was repeated for each threshold and the results are given in figure 5. Compared to the figures 3 and 4 the differences are indistinguishable. However, inspection of the numerical values shows that for large variances the combination of the probabilities results in more classifications per sample.

### 2.2. Flexible number of samples

Instead of refraining to classify if the threshold is not exceeded, as done before, the sample can be used in combination with the next samples so that the combination of them is easier to classify. This can be implemented by averaging over more and more samples, or by keep updating the probability (1). More samples are included until the threshold is exceeded. The number of samples required for a classification is uncertain, and the results reported are on the expected classifications per sample. The same holds for the classification rate.

The legend in the figures will use a $\Sigma$ if the averaging over a flexible number of is used, a $\Pi$ for combination with probabilities and a 1 to denote the use of one sample. The last will form the baseline. Results for the situations encountered before are summarised in figure 6.

In the left hand side of this figure it can be observed that both methods that incorporate previous information in the classification instead of omitting it, classify more samples for a given classification rate. Averaging over several measurements before the classification is slightly better if the standard deviation is small.

It is striking to see, that in the right hand part of the figure, averaging over a flexible number of measurements might even result in a smaller expected number of classification per sample. This is consistent to the observation made concerning figure 4, in which we saw that it might be better to use one sample to classify for some significance, than to use several. This should kept in mind when one wants to improve the classification by using an ERP instead of the best classification from a set. The same results were encountered for equally large variances, but this is not depicted. In all our simulations the combination of probabilities was better than that of a single sample classification.

So, depending on the variance, the way how information is combined makes a difference. Similar results are reported by [9].

### 3. EXPERIMENT

#### 3.1. Data

Data from [4] was used in which a single subject, a musician, participated in five sessions of an experiment. In each session thirty trials were presented in random order for each of five short rhythmic patterns. The rhythms were selected to be perceptually and musically different [10]. Of these five rhythms, the ‘stretch’ and ‘random’ rhythm were arbitrarily selected to be used in the following classification.

The location of the electrodes measured are shown in figure 7. The data was filtered between $1-50$ Hz and captured at a 250 Hz sample rate. The horizontal and vertical EOGs were recorded to detect eye-movements. Trials in which the subject moved with his eyes, or with some other muscle, were manually removed from the data set. This resulted in a data set for the first rhythm of 132 trials for 19 electrode measurements during 3 seconds. The data set of the second rhythm contains 127 trials of the same dimension. Approximately $75\%$ of the original trials remained after artefact rejection.

The differences between the two rhythms presented is in the timing in the tones. We therefore concentrate on differences in the time response of the EEG signals, and not on differences in activation between location [11]. The feature that is used for classification, is the correlation of a new trial with the differences between the two classes’ average wave forms:

$$
    f^i = d^T_n \left( \frac{1}{N_b} \sum_{j=1}^{N_b} d_{a,j}^i - \frac{1}{N_a} \sum_{j=1}^{N_a} d_{b,j}^i \right).
$$

(4)

In this equation $d_{a,j}^i$ is time sampled data for class $a$, trial $j$ on channel $i$. There are $N_a$ trials of class $a$. The new data that is to be classified is $d_n$. All signals are sampled in time and normalised after the bandpass filtering. The reason that
the correlation is used as feature for classification is that if \( a_n \) is a measurement of class a, the feature would be large, while it would be small if it is a trial from rhythm b. In the ideal situation in which the waveforms are orthogonal and there is no noise or any other disturbance, then the feature would be 1 or -1.

Equation 4 results in a feature for each channel measured. For now, we only want to use the best channel for the classification. The smallest classification error if found for the channel with the smallest overlap between the two classes. This minimal overlap found was 20.4% for channel F3. In figure 7, created with EEGLAB [12] the overlap is plotted mapped to the electrodes. The smaller the overlap, the darker the colour. Although it is not the intention of this paper to relate this figure to the physiological origins of the signals, the symmetry of the overlap strengthens our believes in the channel selection, as the auditory cortex is symmetrically located.

3.2. Evaluation

The mean and the standard deviation of the correlation for both classes are determined for electrode F3, and are given in table 1. The statistics of the feature make it possible to plot the classification rate and the acceptance rate as function of the threshold. This can be used to select a combination method and a threshold setting that results in the required performance of our classifier. The acceptance rate as function of the error for these statistics is given in figure 8. The differences between combining with the average and with probabilities is minimal. The overlap is surprisingly small for EEG measurements.

The different combination techniques are tested for the EEG data measured. The data set was repeatedly divided into a training and an evaluation set and the final results were averaged over these divisions. 100 different threshold values are tested. The experimental results are given in figure 9.

If all the samples are classified, so when a threshold of 0.5 is used, then the classification error rate is 21.2%. This was of course equal for all the methods, as no combination is made yet. The classification error is close to the error predicted from its statistics, which was 20.4%.

The calculated curves for the given statistics coincide rather well with the experimentally determined curves. The general form is similar, and the values are not too far off. However, there are differences; the first that stand out is the rebound of the classification rate if a large threshold is used for classification with one sample. Few classifications are made when this rebound is starting, and outliers are starting to play an important role. Furthermore, the limited number of samples in the evaluation set might influence this behaviour. Second, the amount of accepted samples is smaller for the true data than it is for the calculated data. Again, this is likely due to the limited number of samples in our data set.

The difference between using multiple samples in contrast to one sample is significantly larger than predicted in figure 8. When using one sample for the classification, a single extremity of the data set will result in a classification, as this will give a large probability for one of the two classes. However, if more samples were used, then this extremity has less influence on the total probability. Only if this extreme measurement is found as the first measurement,
then it will have the same effect as without combination. This results in a more robust classification if a flexible number of samples is used.

As final result the information in bits per trial is plotted in figure 10. This figure shows that if one is interested in transferring as much information as possible, it is best to classify all samples and take the errors as they occur.

4. CONCLUSION

We examined several methods for improving of the classification of our EEG signals, in which we want to classify which rhythm a subject is listening to. The focus was on combining successive trials to increase the performance. EEG signals are generally difficult to classify, and combining classifications is a practical and useful way to improve the performance.

Combining measurements only if the certainty of a classification is below some threshold gives good results in comparison with the use of a fixed number of trials. A larger number of samples is classified when a flexible number of trials is used to achieve an equal error rate, because it does not require a new sample if the first sample is good enough. Which of the methods that use a flexible number of samples was best, depends on the variance of the distributions. If the variance is small, averaging over samples until the threshold is exceeded results in a good classification. For large variances, combining the probabilities of individual classifications works better.

A combination of measurements was successfully applied in the classification of EEG signals. Based on one electrode, F3, it was possible to classify which of the two rhythms a subject was listening to in 80% of the trials. This result increased to 90% if on average two trials were used, and to 95% if three trials were used. The prediction that was made on the samples required coincided rather well with the true number of samples that were required.

As our EEG signals had a strong temporal structure, the signal processing scheme that was used was straightforward: classify based on the inner product of a new trial with the difference of the average waveforms. Due to the simplicity of this method, the computational load is minimal, and thus it has great potential for use in a real-time brain computer interface. More work need to be done before rhythmic stimuli can be used in a brain computer interface. First the classification needs to be changed from rhythms perceived to rhythms imagined. Second, the classification scheme should be extended to deal with multiple rhythms.

5. REFERENCES