EEG-BASED SPEECH RECOGNITION
Impact of Temporal Effects

Anne Porbadnigk, Marek Wester, Jan-P. Calliess and Tanja Schultz
Cognitive Systems Lab, University of Karlsruhe, Am Fasanengarten 5, 76131 Karlsruhe, Germany

Keywords: Electroencephalography, Speech recognition, Unspoken speech.

Abstract: In this paper, we investigate the use of electroencephalograhic signals for the purpose of recognizing unspoken speech. The term unspoken speech refers to the process in which a subject imagines speaking a given word without moving any articulatory muscle or producing any audible sound. Early work by Wester (Wester, 2006) presented results which were initially interpreted to be related to brain activity patterns due to the imagination of pronouncing words. However, subsequent investigations lead to the hypothesis that the good recognition performance might instead have resulted from temporal correlated artifacts in the brainwaves since the words were presented in blocks. In order to further investigate this hypothesis, we run a study with 21 subjects, recording 16 EEG channels using a 128 cap montage. The vocabulary consists of 5 words, each of which is repeated 20 times during a recording session in order to train our HMM-based classifier. The words are presented in blockwise, sequential, and random order. We show that the block mode yields an average recognition rate of 45.50%, but it drops to chance level for all other modes. Our experiments suggest that temporal correlated artifacts were recognized instead of words in block recordings and back the above-mentioned hypothesis.

1 INTRODUCTION

1.1 Motivation

Electroencephalography (EEG) has proven to be useful for a multitude of new methods of communication besides the well-known clinical applications. In recent years, speech recognition has facilitated our lives by providing a new, more natural means of communication with machines. Previous research has proven that it is feasible to link these two ideas, that is to use EEG for the recognition of normal speech (Wester, 2006). This has been taken one step further by investigations of whether it is feasible to recognize unspoken speech based on the EEG signals recorded ((Wester, 2006) and (Calliess, 2006)). Unspoken speech means that a subject thinks a given word without the use of the articulatory muscles and without uttering any audible sound.

Up to now, speech recognition usually depends on audible spoken utterances. However, we can think of two scenarios in which unspoken speech is preferable. First, there are situations where using spoken speech is undesirable or even unfeasable, for instance in quiet settings or environments where uttering speech is impossible. Second, there are people who are not able to utter speech due to a physical disability. For instance, locked-in patients have extremely limited possibilities for communicating with their environment.

1.2 Related Work

Investigating EEG-based brain computer interfaces (BCIs) has evolved into an increasingly active strand of research. Good overviews can be found in (Dornhege et al., 2007) and (Wolpaw et al., 2002), while (Lotte et al., 2007) provides a review of classification algorithms. Prominent examples of BCIs include the Thought Translation Device (Birbaumer, 2000) and the Berlin Brain Computer Interface (Blankertz et al., 2006). The aim of BCIs is to translate the thoughts or intentions of a given subject into a control signal for operating devices such as computers, wheelchairs or prostheses. Using a BCI usually requires the user to explicitly manipulate his/her brain activity which is then used as a control signal for the device (Nijholt et al., 2008). This usually necessitates a learning process that may last several months as described in (Neuper et al., 2003). By contrast, we focus on the direct recognition of mentally uttered speech with the

Porbadnigk A., Wester M., Calliess J. and Schultz T. (2009).
EEG-BASED SPEECH RECOGNITION - Impact of Temporal Effects.
In *Proceedings of the International Conference on Bio-inspired Systems and Signal Processing*, pages 376-381
DOI: 10.5220/0001554303760381

aim of developping a more intuitive interface.

Several systems have been developped which couple a spelling application to a BCI for mental text entry ((Birbaumer et al., 1999), (Scherer et al., 2004), (Wolpaw et al., 2003)). None of these systems attempts to recognize words directly, though.

Another approach has been taken by developing systems that recognize silent speech based on electromyographic (EMG) data (Maier-Hein, 2005). However, this technique is based on the movement of facial muscles and thus cannot be used by locked-in patients or patients with diseases that prevent articulatory muscle movements.

In (Suppes et al., 1997), it has been shown that isolated words can be recognized based on EEG and MEG recordings. In one of their experimental conditions called internal speech, the subjects were shown one out of 12 words on a screen and asked to utter this word 'silently' without using any articulatory muscles.

For the work described in this paper we used a similar task. In an initial study, Marek Wester implemented a system that seemed to be capable of recognizing unspoken speech based on EEG signals at a high recognition rate (Wester, 2006). However, the words were presented in blocks. In a subsequent study, Jan-P. Calliess showed that blockwise presentation of words produces far better results than any other presentation mode he experimented with. Consequently, he formulated the question whether the good results using block mode were due to temporal correlated brain activity artifacts (Calliess, 2006). In both studies, the recognition rates were estimated since they were calculated offline with the method described in section 2.2. The two contradicting positions can be formulated as follows:

Hypothesis A. *Unspoken speech can be recognized based on EEG signals employing the method proposed in (Wester, 2006).*

Hypothesis B. *The recognition results reported in (Wester, 2006) were overestimated due to temporal correlated artifacts in the EEG signals that were recognized instead of words.*

In this paper, we investigate which one of these hypotheses is correct. For this, it was of importance to produce data which was recorded in the same recording session (that is without removing the EEG cap) but with varying presentation modes such that the data could be compared directly. A more detailed analysis of the data can be found in (Porbadnigk, 2008).

2 EXPERIMENTAL SETUP

For this study, 21 subjects (6 female, 15 male) were recorded in a total of 68 sessions. The average age was 24.5 years and all of the subjects were fluent in German.

2.1 Recording Hardware

The *VarioPort*^{TM} was used as amplifier/ADC and recording device which has a resolution of 0.22 V / bit and 16 input channels (for more details, refer to (Becker, 2004)). All recordings were conducted at a sampling rate of 300Hz. We used an elastic EEG cap by Electro-Cap International, Inc., equipped with 128 Ag/AgCl electrodes. Out of these, only 16 could be recorded simultaneously due to the limitations of our amplifier. The selection of the electrodes was based on the experience gained in (Calliess, 2006), following the layout shown in Figure 1.

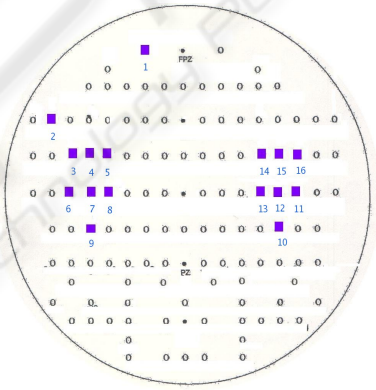

Figure 1: Layout of the high-density EEG cap used. The subset of electrode positions actually recorded are marked by squares. Position 8 and 13 correspond to C3 and C4 (according to the 10-20 system), respectively.

The main focus of recording was the area around the orofacial motor cortex. The only exception were two electrodes. One was placed above the left eyebrow for eye blink detection (electrode 1 in Figure 1). The other electrode was located as far away from the motor strip as possible but still picking up a large part of signals from Broca's area of the left cortical hemisphere (electrode 2 in Figure 1) which is reported to be dominant for speech production.

2.2 Recognition Software

The recognition of the recorded data was performed with the Janus Recognition Toolkit (JRTk), a state-of-the-art speech recognition system (Waibel et al., 2001). Word segmentation was based on eye blinks

with which the subjects marked the beginning and the end of the thinking process for each word. These were detected by an automatic eye blink detector (refer to (Calliess, 2006) for more details).

In JRTk, every word is modeled by a left-to-right Hidden Markov Model (HMM). The standard HMM used for training had five states and one Gaussian mixture per state. We also performed experiments with different numbers of HMM states (3,4,5,6,7) and Gaussians per state (4,8,16,32,64). It turned out that the ideal parameters depend on the word order but no clear trend could be detected.

We focused on offline recognition in this work. For training, four iterations of Expectation Maximization were applied to improve the HMM models. After the Viterbi path had been computed for each word, the one with the best score was chosen as recognizer hypothesis. For training and testing, a round robin scheme was used for as many times as each single word was recorded (20 times in this work). In each round i ($i \in \{1,...,20\}$), one sample of each word was left out of the training procedure and used for testing, resulting in a test set of 95 samples and a training set of 5 samples. Each round resulted in a percentage c_i of how many of these 5 samples in the test set were recognized correctly. The final recognition rate R was computed as follows:
R (%) = $\frac{\sum_i c_i}{20}$ with $i \in \{1,...,20\}$,
$c_i \in \{0\%, 20\%, 40\%, 60\%, 80\%, 100\%\}$

Thus, the recognition rate R is the average likelihood that the system could recognize a word correctly from the given EEG data, using the leave-one-out method described above. The recognition rates which we obtained are in fact estimates since we did not have an online system for testing. The recognition rates reported in (Wester, 2006) and (Calliess, 2006) were calculated in the same way.

Based on previous findings (Wand, 2007), the Double-Tree Complex Wavelet Transform (DTCWT) with decomposition level 3 was chosen for preprocessing. Also, a Linear Discriminant Analysis (LDA) was applied to the feature vectors.

2.3 Database Collection

As vocabulary domain, we chose the first five words of the international radiotelephony spelling alphabet (alpha, bravo, charlie, delta, echo) each of which was repeated 20 times. The total number of recordings is provided in Table 1.

A standard recording for a subject consisted of three sessions. A session is defined as a tuple $S_i = (W_i, O_i)$ with $i \in \{1,2,3\}$, where W_i is the given word list and O_i is the word order in which this word list is presented to the subject. Each of these sessions had the same word list W_i of length 100, but the word order was varied between *blocks*, *blocksReordered*, *shortBlocks*, *sequential* and *randomized*.

If the words were presented in *blocks*, they were presented to the subjects in blocks of 20 words:
((alpha)^{20},(bravo)^{20},(charlie)^{20},(delta)^{20},(echo)^{20})

The order of the blocks was randomized for some subjects which is referred to as *blocksReordered*. *ShortBlocks* means that the words were presented in short blocks of five repetitions for each word:
(((alpha)^5,(bravo)^5,(charlie)^5,(delta)^5,(echo)^5)^4)

If words were presented in *sequential* word order, the quintuple (alpha, bravo, charlie, delta, echo) was repeated 20 times. *Randomized* means that the words were shown to the subject in random order, subject to the constraint that each word occurred 20 times.

Table 1: Overview over the number of recordings in the database and the average recognition rates R (%).

Word Order	Subjects	Sessions	R
Blocks	7	11	45.95
BlocksReordered	11	11	45.05
ShortBlocks	10	10	22.10
Randomized	16	20	19.48
Sequential	9	15	18.09

2.4 Data Acquisition

The subject was seated at a desk in front of a wall, facing a CRT display which was conncected to a laptop and showed the instruction. The supervisor was sitting in front of this laptop, out of sight of the subject, to control the recordings. The laptop was used for both the online control of the experiments and the actual data recording.

Each of the recording steps consisted of four phases and had the following structure: In phase 1, the word w_i was shown to the subject for two seconds. Subsequently, the screen turned blue without showing any word (phase 2). When the screen turned white again after 2 seconds, the recording phase started (phase 3). The subject had the instruction to do the following in this phase: First, s/he had to blink with both eyes and then imagine speaking the word that had been shown in phase 1 without moving any articulatory muscles. Then, s/he had to signal the end of the thought with a second blink.

3 EXPERIMENTS AND RESULTS

In our experiments, we investigated the impact of word order and conducted cross session experiments.

3.1 Impact of Word Order

The first part of the experiments was based on variations in the presentation mode during recordings. We recorded three sessions per subject without removing the EEG cap, with varying word order across these sessions. Thus, the sessions of one subject could be compared directly in order to investigate the influence of the word order. We observed a clear correlation between the word order in which the words were presented and the recognition rate. A breakdown of the average recognition recognition rate per word order is given in Table 1. The overall picture was as follows (as can be seen in Figure 2):

$$R_{blocks} > R_{shortBlocks} > R_{randomized} > R_{sequential}$$

where the latter three rates were basically at chance level. R_{blocks} includes both alphabetical and reordered blocks and had a value of 45.50%.

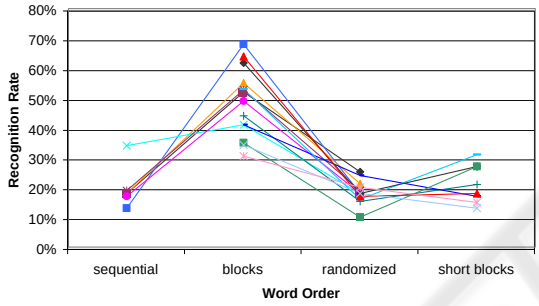

Figure 2: Recognition rate depending on word order for 13 subjects. Data points of the same subject have the same color and are connected.

First, the word order O_i was varied between *blocks*, *randomized* and *sequential*. If hypothesis B were right, *randomized* and *sequential* would yield worse results than *blocks*. This was the case. Only *blocks* yield results above chance level and this word order is vulnerable to temporal artifacts.

Result 1. *Only block recordings yield recognition rates significantly above chance level (average over all block recordings: 45.50%).*

However, the feedback from the subjects suggested an alternative explanation: The subjects reported that they could concentrate better when the words were presented in blocks. So hypothesis A can be amended with the following:

Hypothesis A_1**.** *Block recordings facilitate thinking the words in a consistent way.*

In order to evaluate this new hypothesis, we experimented with *shortBlocks* which shares one property of *blocks*: The same word is presented to the subject *n* times in a row (*n*=20 for *blocks*, n=5 for *shortBlocks*). If A_1 were right, *shortBlocks* should yield much better results than *randomized*, but worse than *blocks*. This was more or less the case as can be derived from Figure 2. However, the recognition results were very close to chance level (average recognition rate: 22.10%).

Result 2. *Although shortBlocks share an important property with blocks, the average recognition rate (22.10%) is much lower than for blocks (45.50%).*

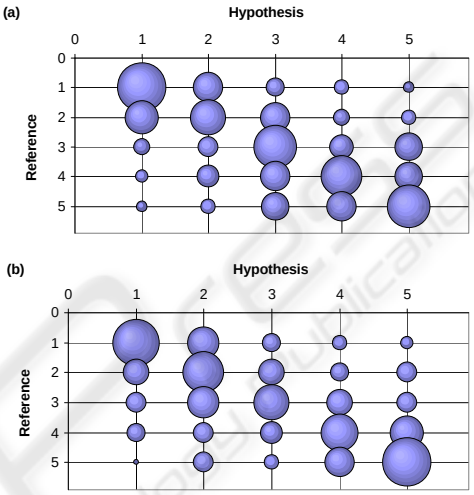

Figure 3: Impact of temporal closeness on recognition rate for alphabetical *blocks* (a) and *blocksReordered* (b). The size of a bubble indicates the number of times the system recognizes a given reference word at time y as the hypothesis at time x (x=y if the the word is recognized correctly).

The next step was to use *reordered* blocks in the recordings. *Reordered* means that the blocks still consisted of blocks of 20 words but those blocks were arranged in a different order. We then calculated two separate confusion matrices, one over all the reordered blocks and one over all the alphabetical blocks. In both cases, we ordered the matrix such that reference 1 was the first reference timewise and reference 5 was the last one. These two matrices showed the same characteristic pattern (compare diagram a) and b) of Figure 3). Furthermore, it can be seen that the more distant in time a block B is from reference block A, the less likely it is that a word from block B gets confused with a word from block A.

Thus, our experiments suggest that temporal artifacts indeed superimpose the signal of interest in *block* recordings. The results indicate that the second part of hypothesis B seems to be correct which claims that the recognition results of *block* recordings were overestimated due to temporal correlated artifacts. In fact, we assume that it is these temporal correlated artifacts which are actually recognized by JRTk.

Result 3. *For blocks recordings, it seems that tempo-*

ral correlated artifacts are recognized instead of the signal of interest.

However, this does not yet answer the fundamental question of whether it is possible to recognize unspoken speech based on EEG signals since it does not necessarily mean that there is no speech-related signal to be identified.

3.2 Cross Session Experiments

The purpose of the cross session experiments was to examine an alternative explanation for why *blocks* yield better results.

If hypothesis A_1 were right, that is if the words were indeed thought in a more consistent way during *block* recordings, these recordings would deliver more reliable data. Consequently, the *blocks* model would be trained more accurately, resulting in higher recognition rates. Thus, we ammended hypothesis A:

Hypothesis A_2. *Block recordings lead to more reliable data containing less noise and showing less variance in the length of the utterances.*

In order to examine this, we trained an HMM with *blocks* data and tested it on data recorded from the same subject, but with word orders different from *blocks*. If hypothesis A were right, the recognition rate should improve. The results show that they remained at chance level instead.

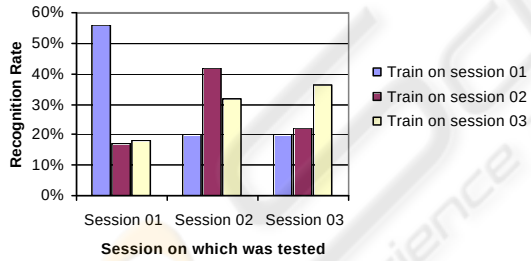

Figure 4: Cross session testing with blocks for subject 05.

This could have been due to the fact that we tested the system on a word order different from *blocks* or due to a high variance in the signals between sessions in general. Therefore, we recorded three sessions from the same subject but set the word order to *blocks* for all of them. This was done for two subjects. For each of them, the system was trained with one *blocks* session and tested on a different *blocks* session. The recognition result droped significantly to chance level (see Figure 4). This is essentially the same result as for the previous cross session experiments with different word orders. The results can be summarized as follows:

Result 4. *Cross session training with the method proposed in (Wester, 2006) does not yield recognition rates above chance level, even if the recording is done with the same word order.*

However, it has been shown in (Suppes et al., 1997) that subject-independent predictions are possible. It has to be taken into account, though, that their subject-independent model was built by averaging over half of their database which was much larger than ours.

Since cross session training did not work out, independent of whether the same or a different presentation mode was used, we cannot conclude if this contradicts hypothesis A_2.

4 DISCUSSION AND CONCLUSIONS

The main goal of the work presented here was to investigate whether it is possible to reliably recognize unspokend speech based on EEG signals using the method proposed in (Wester, 2006). While data presented in (Wester, 2006) was given a promising interpretation, (Calliess, 2006) suggested that temporal correlated artifacts in the EEG signals may have been recognized instead of words. These two hypotheses were refined during the course of our work:

Hypothesis A. *Unspoken speech can be recognized based on EEG signals employing the method proposed in (Wester, 2006). The fact that other word orders yield worse recognition rates may be explained by the following assumptions:*

- A_1: *Block recordings facilitate thinking the words in a consistent way.*
- A_2: *Block recordings lead to more reliable data containing less noise and showing less variance in the length of the utterances.*

Hypothesis B. *Unspoken speech cannot be recognized based on EEG signals employing the method proposed in (Wester, 2006). The recognition results reported in (Wester, 2006) were overestimated due to temporal correlated artifacts in the brainwave signals that were recognized instead of words.*

It could be shown that except for the *block* mode which yielded an average recognition rate of 45.50%, all other modes had recognition rates at chance level. This may be partially explained by the assumptions stated in A_1 and A_2. However, our experiments suggest that temporal correlated artifacts indeed superimpose the signal of interest in block recordings. Therefore, the promising initial results presented in (Wester,

2006) seem most likely to have been caused by an artifact in the experimental design, that is temporal correlated patterns were recognized rather than words. Thus, we conclude that hypothesis B is probably correct. Furthermore, our experiments showed that cross session training (within subjects) only yields recognition rates at chance level, even if the same word order was used for the recordings.

Of course, our analysis does not imply that it is impossible in general to correctly extract (and classify) unspoken speech from EEG data. It has to be pointed out that we do not address the general question of whether this is feasible. Instead, we focus on the method proposed in (Wester, 2006) and show that it is not well suited for the task. Furthermore, it should be taken into account that some assumptions are proposed here which we cannot prove so far.

However, the approach taken here could be changed and improved in several respects. First, using a vocabulary of words with semantic meaning might lead to improvements. Apart from this, it would prove useful to provide JRTk with more training data by recording a higher number of repetitions per word. Second, the recognizing system itself needs to be changed. Due to the high variation of the length of the utterances, normalization would most probably improve the performance of the system. Furthermore, a different word model might be more suitable than HMMs since it turned out that HMMs with just one state yield fairly good results. A one state HMM however does not model temporal data anymore.

Third, the subject could be provided with feedback on whether a given word was recognized correctly. It has been shown in (Birbaumer, 2000) that subjects can indeed be trained to modify their brain waves for using an EEG-based BCI. Thus, we would expect that the subject could adapt his/her brain waves such that they are recognized more easily .

REFERENCES

Becker, K. (2004). Gebrauchsanweisung für *VarioPort™*.

Birbaumer, N. (2000). The thought translation device (ttd) for completely paralyzed patients. *IEEE Trans Rehabil Eng.*, 8(2):190–3.

Birbaumer, N., Ghanayim, N., Hinterberger, T., Iversen, I., Kotchoubey, B., Kübler, A., Perelmouter, J., Taub, E., and Flor, H. (1999). A spelling device for the paralysed. *Nature*, 398:2978.

Blankertz, B., Dornhege, G., Krauledat, M., Müller, K.-R., Kunzmann, V., Losch, F., and Curio, G. (2006). The berlin brain-computer interface: Eeg-based communication without subject training. *IEEE Trans. Neural Sys. Rehab. Eng.*, 14(2):147–152.

Calliess, J.-P. (2006). Further investigations on unspoken speech. Institut für Theoretische Informatik Universität Karlsruhe (TH), Karlsruhe, Germany.

Dornhege, G., del R. Millan, J., Hinterberger, T., McFarland, D., and Müller, K.-R., editors (2007). *Towards Brain-Computer Interfacing*. MIT Press.

Lotte, F., Congedo, M., Lecuyer, A., Lamarche, F., and Arnaldi, B. (2007). A review of classification algorithms for eeg-based braincomputer interfaces. *J. Neural Eng.*, 4:R1–R13.

Maier-Hein, L. (2005). Speech recognition using surface electromyography. Master's thesis, Institut für Theoretische Informatik Universität Karlsruhe (TH), Karlsruhe, Germany.

Neuper, C., Müller, G. R., Kübler, A., Birbaumer, N., and Pfurtscheller, G. (2003). Clinical application of an eeg-based braincomputer interface: a case study in a patient with severe motor impairment. *Clin. Neurophysiol.*, 114:399–409.

Nijholt, A., Tan, D., Pfurtscheller, G., Brunner, C., Millan, J., Allison, B., Graimann, B., Popescu, F., Blankertz, B., and Müller, K.-R. (2008). Brain-computer interfacing for intelligent systems. *IEEE Intell. Syst.*, 23:7279.

Porbadnigk, A. (2008). Eeg-based speech recognition: Impact of experimental design on performance. Institut für Algorithmen und Kognitive Systeme, Universität Karlsruhe (TH), Karlsruhe, Germany.

Scherer, R., Müller, G., Neuper, C., Graiman, B., and Pfurtscheller, G. (2004). An synchronously controlled eeg-based virtual keyboard: Improvement of the spelling rate. *IEEE Trans. Neural Syst. Rehabil. Eng.*, 51(6):979984.

Suppes, P., Lu, Z.-L., and Han, B. (1997). Brain wave recognition of words. *Proc. Natl. Acad. Sci. USA*, 94:14965–14969.

Waibel, A., Bett, M., Metze, F., Ries, K., Schaaf, T., Schultz, T., Soltau, H., Yu, H., and Zechner, K. (2001). Advances in automatic meeting record creation and access. In *Proc. ICASSP '01*, volume 1, pages 597–600.

Wand, M. (2007). Wavelet-based preprocessing of eeg and emg signals for speech recognition. Institut für Theoretische Informatik Universität Karlsruhe (TH), Karlsruhe, Germany.

Wester, M. (2006). Unspoken speech - speech recognition based on electroencephalography. Master's thesis, Institut für Theoretische Informatik Universität Karlsruhe (TH), Karlsruhe, Germany.

Wolpaw, J. R., Birbaumer, N., McFarland, D., Pfurtscheller, G., and Vaughan, T. (2002). Brain-computer interfaces for communication and control. *Clin. Neurophysiol.*, 113(6):767791.

Wolpaw, J. R., McFarland, D. J., Vaughan, T. M., and Schalk, G. (2003). The wadsworth center brain-computer interface (bci) research and development program. *IEEE Trans. Neural Syst. Rehabil. Eng.*, 11(2):207–207.