Towards Continuous Speech Recognition for BCI

Christian Herff, Adriana de Pesters, Dominic Heger, Peter Brunner, Gerwin Schalk and Tanja Schultz

Abstract For the last two decades, brain-computer interface (BCI) research has worked towards practical and useful applications for communication and control. Yet, many BCI communication approaches suffer from unnatural interaction or time-consuming user training. As continuous speech provides a very natural communication approach, it has been a long standing question whether it is possible to develop BCIs that perform speech recognition from cortical activity. Imagined speech as a BCI paradigm for locked-in patients would mean a large improvement in communication speed and usability without the need for cumbersome spelling using individual letters. We showed for the first time that automatic speech recognition from neural signals is possible. Here, we evaluate the feasibility of speech recognition from neural signals using only temporal offsets associated with speech production and omitting information from speech perception. This analysis provides first insights into the potential usage of imagined speech processes for speech recognition, for which no perceptive activity is present.

Keywords Speech · BCI · Automatic speech recognition · ASR · Brain-computer interface

C. Herff (✉) · D. Heger · T. Schultz
Cognitive Systems Lab, University of Bremen (formerly at Karlsruhe Institute of Technology), Enrique-Schmidt-Str. 5, 28359 Bremen, Germany
e-mail: christian.herff@uni-bremen.de
URL: http://www.csl.uni-bremen.de

A. de Pesters · P. Brunner · G. Schalk
New York State Department of Health, National Resource Center for Adaptive Neurotechnologies, Wadsworth Center, Albany, USA

P. Brunner · G. Schalk
Department of Neurology, Albany Medical College, Albany, USA

C. Guger et al. (eds.), *Brain-Computer Interface Research*,
SpringerBriefs in Electrical and Computer Engineering,
DOI 10.1007/978-3-319-57132-4_3

1 Introduction

Previous neuroscientific studies provided evidence for neural representations of speech, such as phones and phonetic features during speech perception [3, 9, 12]. Other studies classified [1, 8, 10] or investigated the production [4, 18] of limited sets of phones, syllables, and words. A complete set of manually labeled phones was classified in single word production in [13]. However, it was unclear whether the brain encodes a complete repertoire of phonetic representations during the production of continuous speech that allows the decoding of words and phrases.

In a study with 7 participants [6], we showed for the first time that continuously spoken speech is represented in the brain as a sequence of phones. These phones can be decoded from electrocorticographic (ECoG) recordings and allow the composition of the spoken words, which we call *Brain-to-Text*. All participants were undergoing surgery for intractable epilepsy and agreed to participate in our experiment. Electrode locations were determined based solely on clinical needs of the patients. We used electrode grids (Ad-Tech Medical Corp., Racine, WI; PMT Corporation, Chanhassen, MN) with inter-electrode distances of 0.6–1 cm. BCI2000 [16] was used to record ECoG signals from eight 16-channel g.USBamp biosignal amplifiers (g.tec, Graz, Austria).

In our experiment, we recorded ECoG activity and the acoustic waveform simultaneously, while participants read aloud different texts consisting of childrens' literature, fan fiction or political speeches. We time-aligned the neural data to a phone labeling obtained from the acoustic data using our in-house speech recognition toolkit BioKIT [17]. This allowed us to identify the neural activity corresponding to the production of each phone. See Fig. 1 for data recording in our experiment and aligning of ECoG and acoustic data. We segmented the texts into phrases and used the recorded ECoG data of all but one phrase for feature selection and training, then

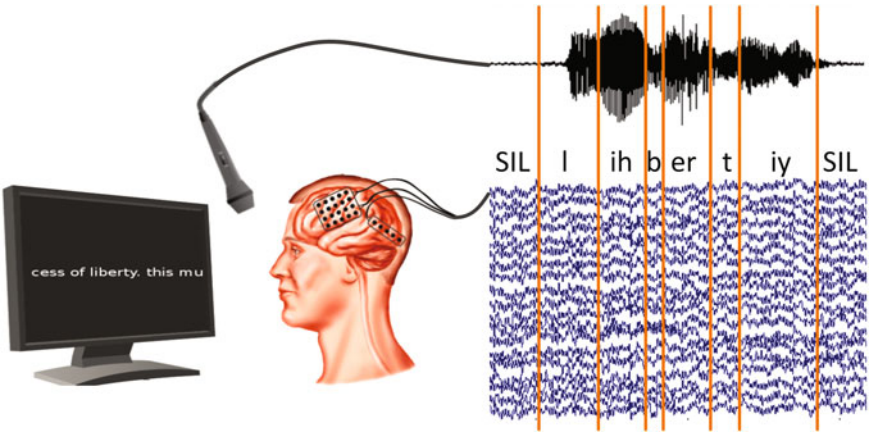

Fig. 1 Synchronized data recording of ECoG data and the acoustic stream

evaluated our approach on the ECoG data of the remaining phrase in a round-robin manner (leave-one-phrase-out validation). We compared the results from temporal offsets associated with speech production to productive and perceptive temporal offsets to analyze the feasibility of continuous speech recognition from imagined speech processes, as perceptive activity is only present when participants hear their own voice.

2 Phone Modeling in ECoG

To model phones in ECoG data, we extracted broadband-gamma (70–170 Hz) activity in 50 ms windows for each channel. The temporal dynamics of speech production were captured by including the features of the four neighboring windows before and after each window in the feature vector, i.e. representing a context of 450 ms length. We modeled each phone with a multivariate Gaussian distribution representing the mean broadband-gamma activity and the corresponding variance for all locations and time lags. We analyzed the discriminability between the different phone models by employing their Kullback-Leibler divergences (KL-div) for every electrode position and time lag. The spatio-temporal distributions of KL-div results give interesting insights into the spatio-temporal dynamics of cortical activity during continuously spoken speech. Figure 2 illustrates discriminability between phones for cortical locations and time offsets on a combined electrode montage of all participants. Phone discriminability can be observed 200 ms prior to phone production in prefrontal areas associated with speech planning (Broca's area). 100 ms prior to phone production, discriminability increases in motor areas and auditory cortex and vanishes in previously observed regions. At phone onset, discriminability peaks in motor cortex, while discriminability is largest in auditory cortex 100 ms after phone production. 200 ms after phone production, phone models can be discriminated in auditory cortex. The activations after the actual phone production are presumably triggered by the participants' perception of their own voice.

We also use the KL-div values to automatically select the best ECoG features for our *Brain-To-Text* system.

To evaluate the feasibility of our system for realistic brain-computer interfaces based on imagined speech production, we performed an analysis that focuses on activity prior to phone onset. By only keeping the temporal offsets between −200 and 0 ms (see Fig. 2), no perceptive activity from hearing one's own voice should remain in the data. This restriction to productive temporal offsets is a first simulation of imagined speech, in which no perceptive activity is present, as participants do not hear their own voice. We refer to these results as *Production* and compare them to those obtained with all temporal offsets, refereed to as *Production and Perception*. This analysis therefore provides a first insight into the feasibility of our system for imagined speech.

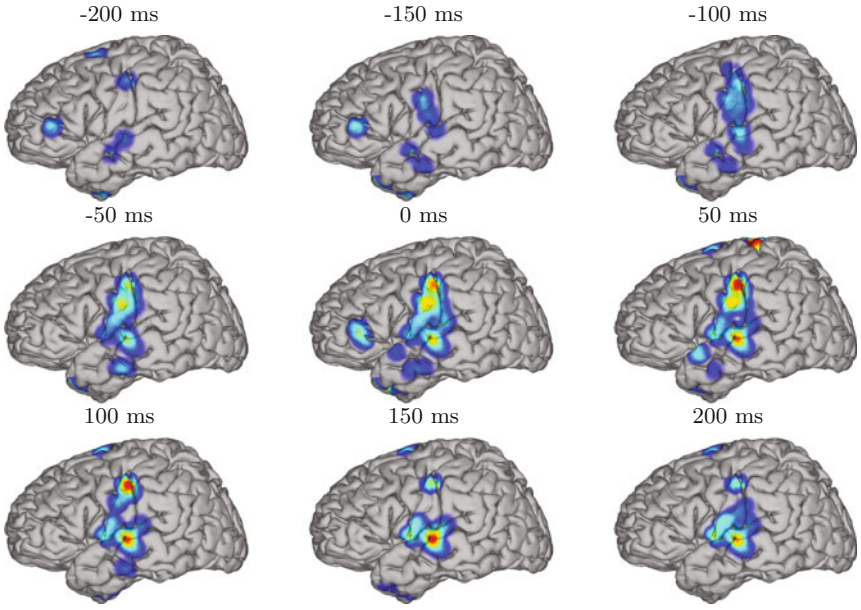

Fig. 2 Discriminability (Mean Kullback-Leibler Divergences) between phones for electrode position of all participants. Color overlays on the rendered average brain show regions of high discriminability (*red*) to lower discriminability (*blue*), all overlays are larger than random discriminability. Early differences can be observed in diverse areas up to 200 ms before phone production. Sensorimotor cortex shows high discriminability 50 ms before production, while discriminability in auditory regions of the superior temporal gyrus peaks 100 ms after production

3 Automatic Speech Recognition for BCI

We combined the phone-based speech representations of cortical activity with language information using automatic speech recognition technology to reconstruct the words in unseen spoken phrases. Language information is included into the decoding process through a language model and a pronunciation dictionary. The pronunciation dictionary contains the mapping of phone sequences to words. The language model statistically models syntactic and semantic information by predicting the next words given the preceding words [7].

Our results show that, with a limited set of words in the dictionary, *Brain-to-Text* is able to reconstruct full sentences. Figure 3 illustrates the different steps of decoding continuously spoken phrases from neural data. *ECoG signals over time* are recorded at every electrode and divided into 50 ms segments. For each 50 ms interval of recorded *broadband gamma activity*, stacked feature vectors are calculated (*Signal processing*). For each *ECoG phone model* calculated on the training data, the likelihood that this model emitted a segment of ECoG features can be calculated, resulting in *phone likelihoods over time*. Combining these Gaussian *ECoG phone models* with

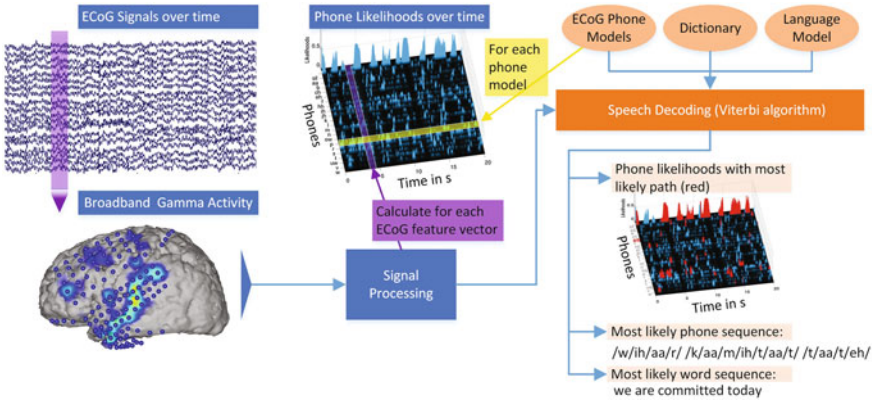

Fig. 3 Overview of the *Brain-to-Text* decoding process

language information in the form of a *dictionary* and an n-gram *language model*, the *Viterbi* algorithm calculates the *most likely word sequence* and corresponding *phone sequence*. To visualize the decoding path, the *most likely phone sequence* can be shown in the *phone likelihoods over time* (red marked areas). The system outputs the decoded word sequence. Once the ECoG phone models are trained, phrases can be decoded in real-time.

4 Results

To evaluate the performance of *Brain-to-Text*, we compared the decoding results of our approach to randomized models (randomization test by shifting the labels of the training data by half the session length). The randomized results illustrate the impact that the language model and dictionary have when no usable neural information is present. Figure 4 shows phone classification accuracies for all participants and sessions. Classification accuracies for combined productive and perceptive areas (purple bars) are better than accuracies achieved with randomized models (yellow bars) for all sessions of all participants. To estimate how well a hypothetical device based on imagined speech production might be, we evaluated our approach only based on productive areas, by excluding all activations from time offsets after phone onset. As the participants cannot hear their own voice prior to the onset of the phone, this ensures that no perceptive activity should be used in this evaluation. Results on productive areas only (turquoise bars) outperform the randomized models for all sessions, but are usually worse than accuracies achieved when using all neural activity.

As *Brain-to-Text* outputs word sequences, we evaluated the Word Error Rate between our predicted word sequence and the reference phrase. One of the major limitations in our study is the small amount of training data per session, with only a

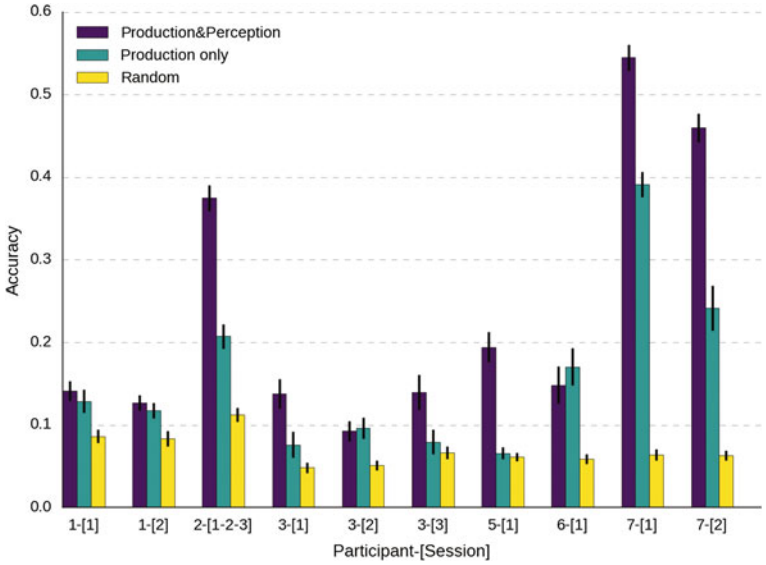

Fig. 4 Phone classification accuracies for all participants and sessions. Error bars depict standard errors. Our system shows significantly better accuracies than random models (*yellow bars*) when using all information (*purple bars*) and when only using productive temporal offsets (*turquoise bars*)

few minutes of data. For comparison, speech recognition systems based on acoustic speech are usually trained on thousands of hours of data. To account for the limited amount of training data, we restricted the amount of recognizable words in the dictionary to a range of 10–100 words. We were able to achieve Word Error Rates as low as 25% when using a dictionary of 10 words. Word Error Rates depending on dictionary size for the best performing participant are shown in Fig. 5. Word Error Rates are lowest (between 25% and just over 60%) when using perceptive and productive (purple line) time offsets. Neural activity only resulting from speech production yields slightly higher Word Error Rates (turquoise line) than perceptive and productive activity, but still outperforms randomized models (yellow line) for all dictionary sizes. Using productive activity only, more than 60% of words are recognized correctly for a dictionary of 10 words.

To ensure that word recognition is not based on the robust recognition of a small subset of phones, we also analyzed average phone true positive rates. For this analysis, we obtained the ground truth of phone timings from the audio alignment described earlier. Bars in Fig. 5 show true positive rates averaged across all phones on window-level. Again, productive and perceptive time offsets (purple) combined yield the best results, but using only productive neural activity (turquoise) still yields high average phone true positive rates above 20%. Both systems using neural activity outperform random true positive rates (yellow). Average phone true positive rates remain rather stable even when dictionary sizes increase.

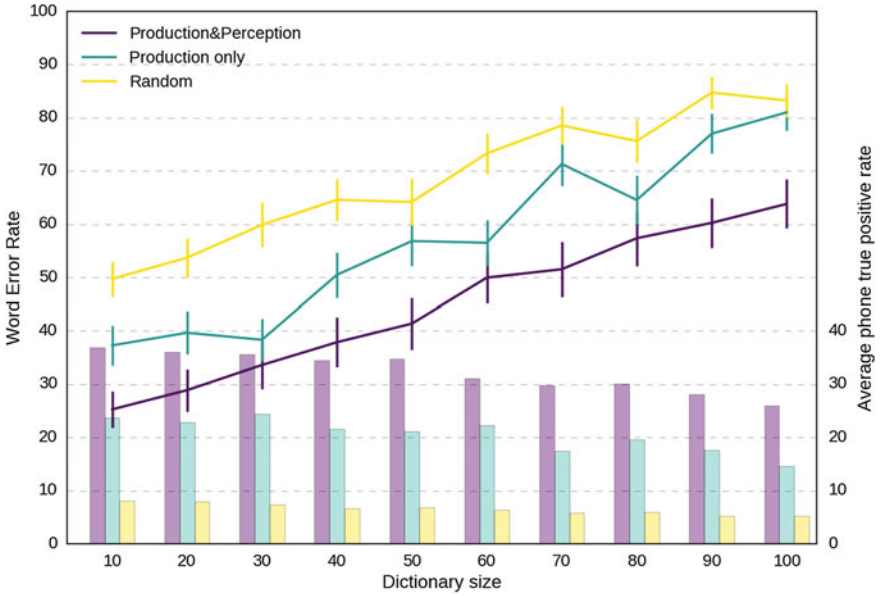

Fig. 5 Word Error Rates over dictionary size (*lines*); average true positive rates across phones depending on dictionary size (*bars*). Error bars depict standard errors. While the full set of temporal offsets performs best (*purple*), information from productive time offsets (*turquoise*) still outperforms random models (*yellow*) for all dictionary sizes both in Word Error Rates and true positive rates

Even though detailed results are only shown for the participant which gave the best recognition results, we found significantly better results than random models in Word Error Rate and single phone true positive rates for all sessions in this study.

5 Conclusion

In summary, our results support the hypothesis that *Brain-to-Text* may eventually allow people to communicate using brain signals associated with continuous spoken language, i.e. without the current limitations of a restricted set of commands or unnatural selection procedures. We showed that participants' neural activity could be used to decode continuously spoken phrases into a textual representation, even when omitting neural activity associated with the perception of their own voice. This illustrates the feasibility of speech recognition from neural activity when participants only imagine to speak. Thus, using continuous speech production for BCIs has the potential to increase naturalness and information transfer rates and the practical utility of current BCI communication approaches. Ultimately, speech processes for BCIs

might lead to information transfer rates similar to that of continuous speech while being more natural to the user.

While the generative models used in this study allow for a good illustration and fast training of phone models, we have shown that more advanced discriminative models can improve results [5].

Recent advances in the modeling of imagined phones [2], reconstruction of imagined speech spectra [11] and investigations in silent reading [14, 15], suggest that covert and overt speech share a neural substrate. Our presented results suggest that neural activity from productive temporal offsets allows reconstruction of a textual form, without the need for perceptive information. These findings highlight the potential of *Brain-to-Text* to be used on imagined continuous speech in the future.

References

1. T. Blakely, K.J. Miller, R.P.N. Rao, M.D. Holmes, J.G. Ojemann, Localization and classification of phonemes using high spatial resolution electrocorticography (ECoG) grids, in *30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2008. EMBS 2008* (IEEE, 2008), pp. 4964–4967
2. S.J. Brumberg, E.J. Wright, D.S. Andreasen, F.H. Guenther, P.R. Kennedy, Classification of intended phoneme production from chronic intracortical microelectrode recordings in speech-motor cortex. Front. Neurosci. **5** (2011)
3. F. Edward, Chang, J.W. Rieger, K. Johnson, M.S. Berger, N.M. Barbaro, R.T. Knight, Categorical speech representation in human superior temporal gyrus. Nat. Neurosci. **13**(11), 1428–1432 (2010)
4. M. Fukuda, R. Rothermel, C. Juhász, M. Nishida, S. Sood, E. Asano, Cortical gamma-oscillations modulated by listening and overt repetition of phonemes. Neuroimage **49**(3), 2735–2745 (2010)
5. D. Heger, C. Herff, A. de Pesters, D. Telaar, P. Brunner, G. Schalk, T. Schultz, Continuous speech recognition from ECoG, in *Sixteenth Annual Conference of the International Speech Communication Association* (2015)
6. C. Herff, D. Heger, A. de Pesters, D. Telaar, P. Brunner, G. Schalk, T. Schultz, Brain-to-text: decoding spoken phrases from phone representations in the brain, Front. Neurosci. **9**(217) (2015)
7. F. Jelinek, *Statistical Methods for Speech Recognition* (MIT Press, 1997)
8. S. Kellis, K. Miller, K. Thomson, R. Brown, P. House, B. Greger, Decoding spoken words using local field potentials recorded from the cortical surface. J. Neural Eng. **7**(5), 056007 (2010)
9. J. Kubanek, P. Brunner, A. Gunduz, D. Poeppel, G. Schalk, The tracking of speech envelope in the human cortex. PLoS ONE **8**(1), e53398 (2013)
10. C.E. Leuthardt, C. Gaona, M. Sharma, N. Szrama, J. Roland, Z. Freudenberg, J. Solis, J. Breshears, G. Schalk, Using the electrocorticographic speech network to control a brain-computer interface in humans. J. Neural Eng. **8**(3), 036004 (2011)
11. S. Martin, P. Brunner, C. Holdgraf, H.-J. Heinze, N.E. Crone, J. Rieger, G. Schalk, R.T. Knight, B. Pasley, Decoding spectrotemporal features of overt and covert speech from the human cortex. Front. Neuroeng. **7**(14) (2014)
12. N. Mesgarani, C. Cheung, K. Johnson, E.F. Chang, Phonetic feature encoding in human superior temporal gyrus. Science 1245994 (2014)
13. M.E. Mugler, J.L. Patton, R.D. Flint, Z.A. Wright, S.U. Schuele, J. Rosenow, J.J. Shih, D.J. Krusienski, M.W. Slutzky, Direct classification of all American English phonemes using signals from functional speech motor cortex. J. Neural Eng. **11**(3), 035015 (2014)

14. M. Perrone-Bertolotti, J. Kujala, J.R. Vidal, C.M. Hamame, T. Ossandon, O. Bertrand, L. Minotti, P. Kahane, K. Jerbi, J.-P. Lachaux, How silent is silent reading? intracerebral evidence for top-down activation of temporal voice areas during reading. J. Neurosci. **32**(49), 17554–17562 (2012)
15. I.C. Petkov, P. Belin, Silent reading: does the brain hear both speech and voices? Curr. Biol. **23**(4), R155–R156 (2013)
16. G. Schalk, D.J. McFarland, T. Hinterberger, N. Birbaumer, J.R. Wolpaw, Bci2000: a general-purpose brain-computer interface (BCI) system. IEEE Trans. Biomed. Eng. **51**(6), 1034–1043 (2004)
17. D. Telaar, M. Wand, D. Gehrig, F. Putze, C. Amma, D. Heger, N.T. Vu, M. Erhardt, T. Schlippe, M. Janke et al., BioKIT—real-time decoder for biosignal processing, in *The 15th Annual Conference of the International Speech Communication Association (Interspeech 2014)* (2014)
18. L.V. Towle, H.-A. Yoon, M. Castelle, J.C. Edgar, N.M. Biassou, D.M. Frim, J.-P. Spire, M.H. Kohrman, ECoG gamma activity during a language task: differentiating expressive and receptive speech areas. Brain **131**(8), 2013–2027 (2008)