Position Paper: Brain Signal-based Dialogue Systems

Odette Scharenborg[1] and Mark Hasegawa-Johnson[2]

[1]Multimedia Computing Group, Delft University of Technology, Delft, The Netherlands

[2]University of Illinois, Urbana-Champaign, IL, USA

Abstract This position paper focuses on the problem of building dialogue systems for people who have lost the ability to communicate via speech, e.g., patients of locked-in syndrome or severely disabled people. In order for such people to communicate to other people and computers, dialogue systems that are based on brain responses to (imagined) speech are needed. A speech-based dialogue system typically consists of an automatic speech recognition module and a speech synthesis module. In order to build a dialogue system that is able to work on the basis of brain signals, a system needs to be developed that is able to recognize speech imagined by a person and can synthesize speech from imagined speech. This paper proposes combining new and emerging technology on neural speech recognition and auditory stimulus construction from brain signals to build brain signal-based dialogue systems. Such systems have a potentially large impact on society.

Introduction: A speech-based dialogue system typically takes in a spoken utterance by the user on the basis of which an action from the dialogue system follows. Communication from the dialogue system to the user occurs either in text or using synthesized speech. People who have lost the ability to communicate via speech or sign language (e.g., severely paralyzed people or patients of locked-in syndrome [1]) cannot use existing dialogue systems, nor are they able to communicate with other people. In order for them to communicate, brain-computer interfaces are needed [2]. These brain-computer interfaces should be able to convert the intended message from the neural activity in the brain areas related to speech processing and production into an action carried out by the dialogue system or into text or synthesized speech in case of communication with another person [3].

Two approaches investigating the decoding of speech from neural signals can be distinguished: in the *overt* condition, listeners' neural signals when listening to speech are recorded and decoded; in the *covert* condition, listeners' neural signals are recorded while they imagine to speak and subsequently decoded. The former case is an important step to understand the relationship between the acoustic signal and the neural signal; the latter case is the situation that allows patients to communicate their thoughts and wishes and is the ultimate dream.

Neural signals: The most often used type of neural signal is electrocorticography (ECoG), which is an invasive methodology in which electrode arrays are placed directly onto the surface of the brain in patients. Electroencephalography (EEG) is less invasive as it ‘only’ requires wearing a cap with electrodes, making it a technique that is more user-friendly and cheaper than ECoG. A downside to using EEG signals compared to ECoG is that because EEG caps are placed on the outside of the skull, brain signals obtained with EEG are noisier and have a less good spatial resolution than ECoG signals. EEG signals however have good time resolution which is import in speech processing.

Overt condition: [4] presented a proof-of-concept neural speech recognition system, which used brain responses to continuous speech produced by two speakers obtained using ECoG from three patients receiving surgery related to epilepsy. Data from these three individuals were used to train three listener-dependent systems and a listener-independent system. The obtained phone error rates ranged from 70% to over 80% for the listener-dependent systems. A review of automatic speech recognition of different types of neural signals found that ECoG provided the best recognition results [5]. Although recognition is poor, these systems show that listener-independent linguistic information can be obtained from the ECoG signals.

Covert condition: The neural signals that give the best results in brain-computer interfaces are obtained using ECoG [6][7][8]. EEG signals have, however, with limited success been used to decode imagined articulation of two English vowels [9], three Dutch vowels [10], two Japanese vowels [11], and “yes” and “no” [12] with above chance accuracy. Although more research is needed before this technique can be fully used, for a dialogue system, “yes” and “no” are highly important words.

Auditory stimulus reconstruction: Auditory stimulus reconstruction is an inverse mapping technique which attempts to create an auditory signal from the neural signals [3][13][14][15]. This technique can be used to convert the neural signals from a person listening to speech (overt condition) or the articulation of words imagined by a person (covert condition) into a temporal and spectral representation. Typically, in speech-based brain-computer interfaces, the neural activity to (imagined) speech is decoded into linguistic units such as phonemes or words or acoustic units such as the speech envelope or the magnitude spectrogram (see [14] for references), which can be synthesized as speech. Recently, [14] proposed to train a DNN to directly predict the parameters of the synthesizer from ECoG signals to covert speech rather than go via intermediate representations, so combining neural speech recognition and synthesis. This approach significantly outperformed a system which used an audiospectrogram as an intermediate unit.

Conclusion: The question whether it is possible to build dialogue systems for people who have lost the ability to communicate via speech using their brain responses to speech cannot yet be answered in the affirmative. However, initial building blocks are in place to build such systems, especially for the construction of dialogue systems which require “yes”/”no” answers. The ultimate goal is to make it possible for the patient to communicate his or her thoughts by imagining speech which subsequently can be synthesized, ideally including emotional and speaker-dependent characteristics. Because of user-friendliness, EEG-based technology is preferred over the invasive ECoG-based approach. More research is needed to improve the independent modules and integrate them into working dialogue systems.

References

[1] Laureys, S. et al. “The locked-in syndrome: what is it like to be conscious but paralyzed and voiceless?” *Prog. Brain Res.*, 150, 495–611, 2005.

[2] Sellers, E. W., Ryan, D. B., and Hauser, C. K. “Noninvasive brain-computer interface enables communication after brainstem stroke.” *Sci. Transl. Med.*, 6(257), 257re7, 2014.

[3] Iljina, O. et al. “Neurolinguistic and machine-learning perspectives on direct speech BCIs for restoration of naturalistic communication.” *Brain-Computer Interfaces*, 4, 186–199, 2017.

[4] D.A. Moses, N. Mesgarani, M. K., Leonard, and E. F. Chang. “Neural speech recognition: Continuous phoneme decoding using spatiotemporal representations of human cortical activity.” *J. Neural Eng.*, 13, 19 pages, 2016.

[5] Herff, C. and Schultz, T. “Automatic speech recognition from neural signals: a focused review.” *Front. Neurosci.*, 10, 429, 2016.

[6] Martin, S., et al. “Word pair classification during imagined speech using direct brain recordings.” *Sci. Rep.*, 6, 25803, 2016.

[7] Leuthardt, E. C., et al. “Using the electrocorticographic speech network to control a brain–computer interface in humans.” *J. Neural Eng.*, 8, 2011.

[8] Pei, X., Barbour, D. L., Leuthardt, E. C., and Schalk, G. “Decoding vowels and consonants in spoken and imagined words using electrocorticographic signals in humans.” *J. Neural Eng.*, 8, 46028, 2011.

[9] Dasalla, C. S., Kambara, H., Sato, M., and Koike, Y. “Single-trial classification of vowel speech imagery using common spatial patterns.” *Neural Netw.*, 22, 1334–1339. doi: 10.1016/j.neunet.2009.05.008, 2009.

[10] Hausfeld, L., De Martino, F., Bonte, M., and Formisano, E. “Pattern analysis of EEG responses to speech and voice: influence of feature grouping.” *Neuroimage*, 59, 3641–3651. doi: 10.1016/j.neuroimage.2011.11.056, 2012.

[11] Natsue, Y., Atsushi, N., Nasreddine, B. A., Duk, S., Hiroyuki, K., Takashi, H., and Yasuharu, K. “Decoding of Covert Vowel Articulation Using Electroencephalography Cortical Currents.” *Frontiers in Neuroscience*, 10, 175. https://www.frontiersin.org/article/10.3389/fnins.2016.00175 DOI=10.3389/fnins.2016.00175, 2016.

[12] Lopez-Gordo, M. A., Fernandez, E., Romero, S., Pelayo, F., and Prieto, A. “An auditory brain-computer interface evoked by natural speech.” *J. Neural Eng.*, 9, 036013. doi: 10.1088/1741-2560/9/3/036013, 2012.

[13] Pasley, B. N. B. N., et al. “Reconstructing speech from human auditory cortex.” *PLoS Biol.*, 10, 2012.

[14] Akbari, A. H., Khalighinejad, B., Herrero, J. L., Mehta, A. D., and Mesgarani, N. “Towards reconstructing intelligible speech from the human auditory cortex.” *bioRxiv, the preprint server for biology*, doi: https://doi.org/10.1101/350124, 2018.

[15] Chakrabarti, S., Sandberg, H. M., Brumberg, J. S., and Krusienski, D. J. “Progress in speech decoding from the electrocorticogram.” *Biomed. Eng. Lett.*, 5, 10–21, 2015.