SMM23, Workshop on Speech, Music and Mind 2023
18 August 2023, Trinity College Dublin, Ireland

Comparison of acoustic-to-articulatory and brain-to-articulatory mapping during speech production using ultrasound tongue imaging and EEG

Tamás Gábor Csapó[1], *Frigyes Viktor Arthur*[1], *Péter Nagy*[2,3], *Ádám Boncz*[3]

[1]Department of Telecommunications and Media Informatics,
Budapest University of Technology and Economics (BME), Budapest, Hungary
[2]Department of Measurement and Information Systems, BME, Budapest, Hungary
[3]Sound and speech perception Research Group, Institute of Cognitive Neuroscience and Psychology,
Research Centre for Natural Sciences, Budapest, Hungary

{csapot,arthur}@tmit.bme.hu, nagy.peter.ssprg@ttk.hu, adam.boncz@gmail.com

Abstract

With the investigation of speech-related biosignals we can enhance traditional speech synthesis which might be useful for future brain-computer interfaces. In a recent previous research, from the brain signal measured with EEG, we predicted directly measured articulation, i.e., ultrasound images of the tongue, with a fully connected deep neural network. The results showed that there is a weak but noticeable relationship between EEG and ultrasound tongue images, i.e., the network can differentiate articulated speech and neutral (resting state) tongue position. In the current study, we extend this with a focus on acoustic-to-articulatory inversion (AAI), and estimate articulatory movement from the speech signal. After that, we analyze the similarities between AAI-estimated articulation and EEG-estimated articulation. We compare the original articulatory data with DNN-predicted ultrasound and show that EEG input is only suitable to distinguish neutral tongue position and articulated speech, whereas melspectrogram-to-ultrasound can also predict articulatory trajectories of the tongue.

Index Terms: ultrasound, EEG, brain-computer interface

1. Introduction

Biosignals recorded during speech production might be useful to extend speech technologies: for example, trackings of articulation can be used for speech-to-articulation mapping, often called as acoustic-to-articulatory inversion (AAI). AAI has the aim of estimating articulatory movements from the acoustic speech signal [1, 2]. As another example, recordings of the brain might be useful towards brain-to-speech synthesis. In a broader sense, Brain-Computer Interfaces (BCIs) can allow computers to be controlled directly without physical activity. It is expected that in the future, the use of speech neuroprostheses may help patients with neurological or speech disorders [3].

1.1. Acoustic-to-articulatory inversion

Acoustic-to-articulatory inversion [1, 2, 4, 5, 6], a subfield of speech technology, has the goal of estimating 'direct' articulatory movements (e.g. position of the lips, jaw, tongue, velum, etc.) recorded using articulatory acquisition techniques, from the acoustic speech signal input. Learning the relationship between articulation and acoustics could improve the performance of several tasks such as speech recognition, synthesis [7, 8], or biofeedback / visualization of speech production [9]. For recording articulation, several techniques exist, which each have pros and cons. Most previous works in AAI are based on Electromagnetic Articulography (EMA) or X-ray Microbeam (XRMB) data, which can track only several points of the articulatory organs, and therefore provide limited spatial information. For example, EMA typically uses two to six sensor coils placed on the tongue, and therefore its spatial resolution is limited. Compared to EMA and XRMB, imaging methods (e.g. UTI: Ultrasound Tongue Imaging, and MRI: Magnetic Resonance Imaging) have the advantage that the tongue surface is fully visible, and ultrasound can be recorded in a non-invasive way [10, 11, 12]. During UTI recordings, usually, when the subject is speaking, the ultrasound transducer is placed below the chin, resulting in mid-sagittal images of the tongue movement. The typical result of 2D ultrasound recordings is a series of gray-scale images in which the tongue surface contour has a greater brightness than the surrounding tissue and air (for samples, see Fig. 2 top as 'raw' data, and Fig. 3 for 'wedge' orientation). Compared to EMA, XRMB and MRI, ultrasound is a technique of higher cost-benefit if we take into account equipment cost, portability, safety, spatial and temporal resolution, and visualized structures, e.g. the tongue surface is visible as a continuous line [11, 4, 6]. Therefore, for the experiments in the current paper, UTI has been used as articulatory information.

1.2. Brain-to-speech synthesis

For recording the brain signal, several technologies are available: e.g., electroencephalography (EEG) [13, 14], stereotactic deep electrodes (sEEG) [15], intracranial electrocorticography (ECoG) [16], magnetoencephalography (MEG) [17], Local Field Potential (LFP) [16]. Among these brain signal recording methods, EEG may be the most suitable for BCI, as it is affordable, involves significantly less risk than invasive methods, and can be portable [18]. Initial research has already been carried out to develop EEG and speech-based BCI [19, 14, 15, 20], but this has not yet resulted in clearly intelligible speech. For example, Sharon and Murthy show that multi-phasal correlation can enhance imagined speech recognition from EEG, but the prediction is not fully accurate yet [14]. The reason is that EEG only measures the brain signal on the scalp; therefore, it is less accurate than invasive technologies. Using invasive methods, it has already been possible to create speech-like synthesized speech based on brain signals, e.g. ECoG [21, 22] and sEEG [23, 15, 24], but due to the above disadvantage (primarily the invasive nature), the latter are not expected to be widespread.

1.3. Brain-to-articulation mapping

Articulatory movements have only sporadically been studied in parallel with brain signals during speech production. Lesaja et al. investigates neural correlates of lip movements, and predicts

10.21437/SMM.2023-4

lip-landmark position from ECoG input [25]. Besides, Csapó et al. [26] recorded non-invasive EEG as the brain signal and used ultrasound tongue imaging as articulatory representation, all recorded in parallel during speech production. The aim of the study was to predict ultrasound images from EEG input; and the initial results indicate that there is a weak but noticeable relationship between EEG and ultrasound tongue images, i.e. a simple DNN can differentiate articulated speech and neutral (resting state) tongue position [26].

Besides, the other related studies all use estimated articulatory data, i.e. they take into account the articulatory information inferred from the speech signal or from textual contents (e.g., [27, 22, 28]. Several studies appeared this year which have the aim to predict articulatory-related information from the brain signal. Amigó-Vega et al. [29] and Wairagkar et al. [30] both use invasive EEG for brain representation. The former has VocalTractLab parameters as the target [29], whereas the latter aims to predict EMA representation resulting from speaker-independent AAI with pre-trained models [30].

1.4. Goal of the current study

The conclusion of the above studies is that for patients whose cortical / neural processing of articulation is still intact, a speech-based BCI decoder using articulatory information can be more intuitive or more natural, and easier to learn to use. However, according to the overview above, there are only a few methods that examine the brain-related information, articulation, and speech together. In the current paper, we contribute to this field using EEG and ultrasound tongue imaging, and extend [26] with acoustic-to-articulatory inversion experiments. The motivation here was to highlight articulatory movement prediction patterns that EEG struggles with.

2. Methods

2.1. Recordings: EEG, ultrasound and speech

Our recordings were made in an electromagnetically shielded quiet room of the ELKH Research Centre for Natural Science, Budapest, Hungary, for a previous study [26]. The EEG signal was recorded with a 64-channel Brain Products actiCHamp type amplifier, using actiCAP active electrodes. Four channels were used to track horizontal and vertical eye movements. The electrodes were placed according to the international 10-20 arrangement. The impedance of the electrodes was kept below 15 kOhm. During the recording, the FCz electrode played the role of the reference electrode. The signal was sampled at a frequency of 1000 Hz.

The midsagittal movement of the tongue was recorded using the "Micro" system (AAA v220.02 software, Articulate Instruments Ltd.) with a 2–4 MHz (penetration depth), 64-element, 20 mm radius convex ultrasound probe at 81.67 fps, and we also used a headset for probe fixing. The metal headset was placed above the EEG sensors so that the devices did not interfere with each other. Recording arrangement is shown in Fig. 1 of [26]. The speech was recorded with a Beyerdynamic TG H56c tan omnidirectional condenser microphone and digitized with an M-Audio M-Track 2x2 / FocusRite Scarlett 2i2 USB external sound card at 44,100 Hz.

The output of the sound card (which contains the synchronizing signal of the „Micro" ultrasound, i.e., 'frame sync', and the speech signal from the microphone) was connected to the AUX channel of the EEG – so the brain and articulation signals were recorded on separate computers, but after the session, we synchronized the data, as described in [26]. This made sure that all biosignals are in full synchrony.

As for initial data, we recorded approximately 15 minutes of EEG, ultrasound, and speech from a single native Hungarian male speaker (the first author). The Hungarian sentences were selected from the PPBA database [10].

2.2. Preprocessing the EEG, ultrasound and speech data

The EEG signal was pre-processed based on [15], similarly to [26]. We calculated the Hilbert envelope for each channel of the EEG signal (except EEG AUX) in four frequency bands: 1–50 Hz, 51–100 Hz, 101–150 Hz, and 151–200Hz. Notch filters were used to filter out the 50 Hz line noise and its harmonics. The envelope was averaged every 50 ms and offset by 12 ms to be consistent with the ultrasound tongue images (which were recorded at 81.67 fps).

Ultrasound tongue images were recorded as 8-bit grayscale pixels, in the form of raw ultrasound of the „Micro" system. The originally 64x842 pixel images were resized to 64x128 pixels (Fig. 3, left), as this does not cause significant information loss [31], but the amount of data to be processed for the target of machine learning is less.

For the analysis of speech, we extracted 80-dimensional mel-spectrograms using the 'librosa' library, with 12 ms frame shift, to be in synchrony with the EEG and ultrasound data.

2.3. DNN training and predicting articulatory information from speech input

The goal of the first experiment was to obtain direct ultrasound tongue images (being the target, 64x128 pixels) from the spectral features (being the input of the neural network, 80 dimensions), similarly to [4, 6].

We used a neural network structure with 5 hidden layers, each layer containing 1000 neurons, with ReLU activations, and a linear output layer (similar to earlier ultrasound-based studies [6, 10, 26], i.e. a fully connected deep 'rectifier' neural network, FC-DNN). The input spectral values and the output ultrasound pixels were normalized to 0–1 before training. We trained until up to 100 epochs, but applied early stopping, with a patience of 3. For training, MSE error was used.

During the prediction step, the corresponding articulatory trajectories are predicted from the trained network, and the final result is a raw ultrasound image sequence, in synchrony with the input speech signal. The raw data is converted back to wedge orientation for visualization.

2.4. Predicting articulatory information from EEG input

In the next experiment, similarly to [26], we trained a FC-DNN, during which we predicted the ultrasound tongue images (64x128 image pixels), from the Hilbert-transformed EEG input that was scaled to the range of [0–1] (being the input of the DNN, 62 channels x 4 frequency ranges, altogether 248 dimensions). The DNN structure and the training details were the same as in Sec. 2.3, with the only difference of the input.

3. Experiments and results

We performed training from the 155 sentences, using 80% of the data for training the network, 10% for validation, and the remaining 10% for testing (31 000 / 3900 / 3900 sample points). For both melspectrogram-to-ultrasound and EEG-to-ultrasound, the same train-validation-test split was used.

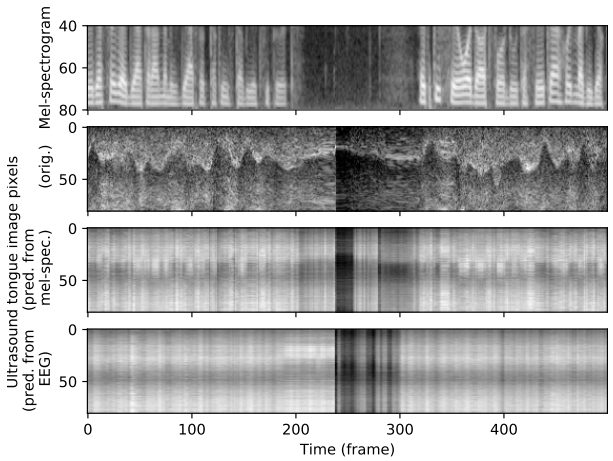

Figure 1: *Demonstration sample: a) Lower half of 80-dimensional mel-spectrogram of the original speech sample, b) original ultrasound kymogram, c) DNN-predicted ultrasound kymogram from mel-spectrogram input, d) DNN-predicted ultrasound kymogram from EEG input.*

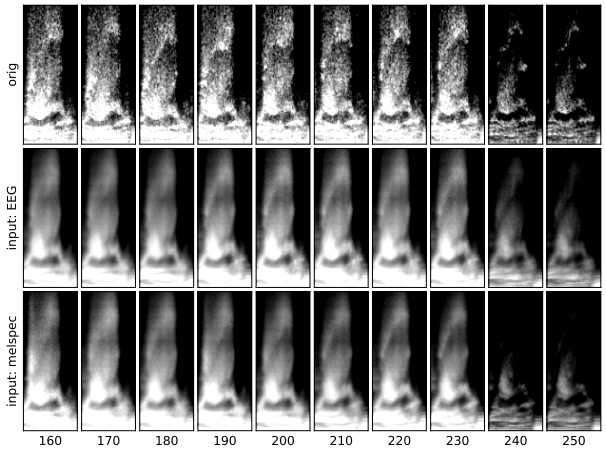

Figure 2: *Original (top), EEG-predicted (middle) and melspec-predicted (bottom) ultrasound tongue images, in 'raw' format.*

3.1. Demonstration samples for melspectrogram-to-ultrasound and EEG-to-ultrasound

After the DNN training, ultrasound tongue image prediction was performed separately from melspectrogram and EEG input, on the test set.

First, we show the results as a 'kymogram' representation: we cut out the middle vertical line from each ultrasound tongue image and plotted the change of this line over time – this kind of visualization has been useful for previous ultrasound-related studies. 'Kymogram' [32, Fig. 8], is 'articulatory signal over time': the middle slices (midline) of the ultrasound tongue images are cut (approximately corresponding to the middle of the tongue) and plotted as a function of time, similarly to a spectrogram; thus the tongue movement is roughly visible together with the speech spectrogram.

Fig. 1 shows the result of this: at the top (a) spectrogram belonging to speech, lower 40 dimensions (to emphasize the region of the first and second formants, where the tongue articulation has most effect). Next, (b) is the ultrasound image center line sequence as a function of time (belonging to the same utterance). After that is the ultrasound tongue center line predicted by the DNN from mel-spectrogram input (c), and from EEG input (d). The similarity between (a) mel-spectrogram and (b) articulatory movement is clearly noticeable: the formant movements in speech and the vertical movement of the tongue can be roughly observed in the figures. In case of ultrasound-based AAI (c), we can see that the estimated tongue movement pattern follows a similar trend as the original data, but is somewhat blurred. On the other hand, in (d) DNN-predicted tongue ultrasound, tongue movement is not visible on the midline, i.e., the FC-DNN could not learn well the relation between EEG and ultrasound tongue images. At the same time, some information can still be seen in the DNN-predicated images: at the end of the 240th frame, one sentence ends, and the next begins, which can be clearly seen in the original ultrasound (b) and also in the estimated ultrasound (c and d). Overall, we can say that according to this visualization, there is a clear relation ship between mel-spectrogram of speech and ultrasound tongue images, whereas we also found a weak but noticeable relationship between EEG and UTI, i.e. the network can differentiate articulated speech vs. neutral tongue position.

Fig. 2 shows some original and estimated ultrasound images from the test data of the speaker, in the 'raw' representation of the ultrasound machine. The contour of the tongue ultrasound is not always visible even in the original images – this is due to the dependence of the ultrasound tongue images on the speaker – it seems that this subject has a tongue that is difficult to acquire.

In case of Acoustic-to-Articulatory Inversion, i.e., mel-spectrogram-to-ultrasound prediction (Fig. 2 bottom), the movement of the tongue is roughly in accordance with the original ultrasound images (Fig. 2 top). Although the UTI images are relatively blurred (because the properties of the data for this particular subject), it is clear that the AAI-estimated images follow the original articulatory patterns to some extent.

In the images estimated from EEG input (i.e., EEG-to-UTI prediction, Fig. 2 middle), the contour of the tongue is fully blurred, and the change in the position of the tongue from frame to frame is also difficult to observe – i.e., the DNN was able to learn the general shape of the tongue (the average image), but the fine details of the tongue movement cannot be seen. However, some general change of brightness is visible as a function of time: if the original images were darker, then this is also mapped on the predicted images (e.g., around frames 230–240).

The same series of images are shown in the 'wedge' representation in Fig. 3. Because the data that was used for plotting is the same, just in different visual representation, a similar trend can be noticed as in Fig. 3: the upper surface of the tongue can be roughly seen in the original images (left), and to some extend on the AAI-estimated images (right) but in the images estimated based on the EEG (middle), the ultrasound pixels are blurred, and the contour of the tongue is not visible. However, between frames 230–240, the change in light intensity can be noticed in the DNN-predicated case.

3.2. Objective measures

The mean squared error (MSE) values achieved with the above FC-DNN network are included in Table 1. The values themselves are difficult to interpret, but for example, in previous UTI-based acoustic-to-articulatory inversion experiments (during which ultrasound tongue images were predicted from the speech signal [4, 6]), the obtained NMSE validation error val-

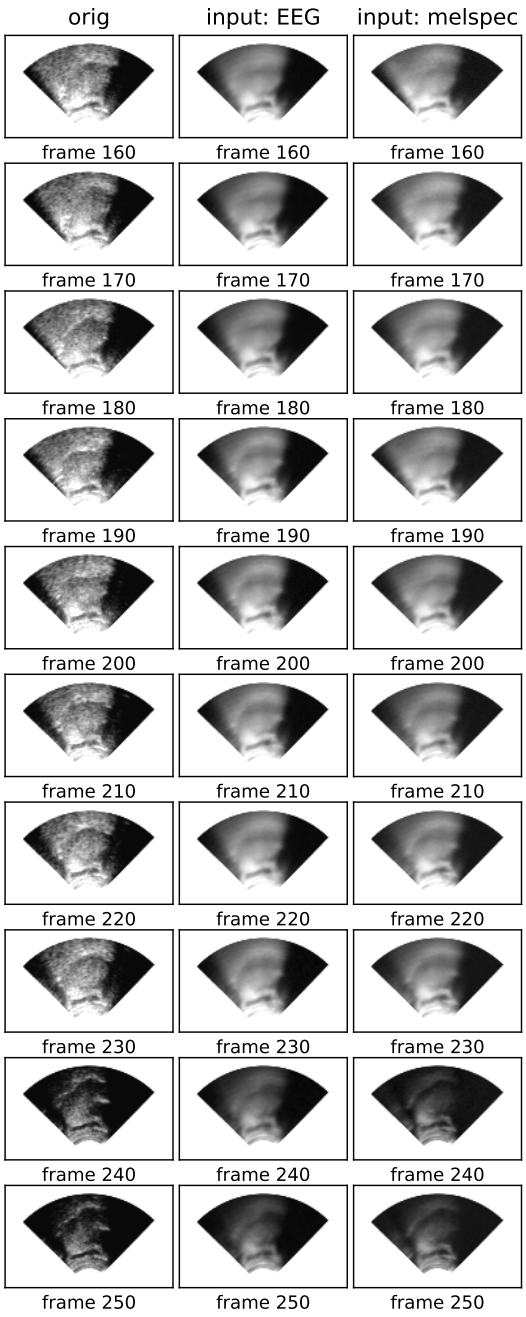

Figure 3: *Original (left), EEG-predicted (middle) and melspec-predicted (right) ultrasound tongue images, in 'wedge' format.*

Table 1: *MSE scores after training.*

	training	validation	test
melspec-to-UTI	0.0051	0.0053	0.0051
EEG-to-UTI	0.0051	0.0052	0.0051

ues were in the order of 0.0053–0.0088; and in this case, the ultrasound tongue video generated from the speech input approximated the original articulatory movement. Note that because of the properties of the ultrasound images, this range is slightly dependent on the actual subject. It can also be seen from Table 1 that the MSE values alone are not sufficient to judge the quality of the results, and a visual inspection is necessary. In the case of previous ultrasound research, there have been experiments examining other error measures, such as Structural Similarity Index (SSIM) [33] and Complex Wavelet Structural Similarity (CW-SSIM) [34], on ultrasound tongue images [35, 6], which might be useful to check on these data.

4. Discussion and conclusions

In Sec. 1, we have shown several previous approaches for brain-to-speech synthesis with articulatory information included; but such articulation was always indirectly measured / estimated using acoustic-to-articulatory inversion, and not recorded in parallel with brain-related data [22, 27, 28, 29, 30]. As suggested by the papers above, an obvious solution for speech BCIs is the examination of articulation as an intermediate representation between the brain signal and the resulting final speech, which we dealt with in [26] and extended in this article with ultrasound-based acoustic-to-articulatory inversion. We argue that measuring and analyzing articulation with real equipment during speech production could result in further advantages and improvement in the long-term for brain-to-speech synthesis.

In the current research, we have focused on acoustic-to-articulatory inversion (AAI), and estimated articulatory movement from the speech signal. After that, we have analyzed the similarities between AAI-estimated articulation and EEG-estimated articulation. We have compared the direct articulation (resulting from recordings of ultrasound tongue images) with DNN-predicted ultrasound and have shown that EEG input is suitable to distinguish neutral (resting state) tongue position and articulated speech, i.e. the relationship between EEG and ultrasound tongue images was clearly demonstrated, because the network can differentiate articulated speech and neutral tongue position, like Voice Activity Detection. Besides, in the AAI experiments we have shown that melspectrogram-to-ultrasound can predict articulatory movements of the tongue with higher accuracy, and we have presented demonstration samples and analyzed the results visually.

Similarly to tongue movement recorded with ultrasound, lip movement recorded with a camera would be a reasonable information to use in the context of the present study. As we have also recorded lip video in the current setup, it is a reasonable next step to investigate lip articulation as well, and how this relates to the brain. Besides of the above FC-DNN, it might be useful to utilise some pre-trained convolutional neural network (CNN), and then fine-tune it on the present task. In the future, we plan to contribute to speech-based brain-computer interfaces, by adding directly recorded speech articulation information to the processing pipeline.

5. Acknowledgements

This research was funded by the National Research, Development and Innovation Office of Hungary (FK 142163 grant). T.G.Cs. was supported by the Bolyai János Research Fellowship of the Hungarian Academy of Sciences and by the ÚNKP-22-5-BME-316 New National Excellence Program of the Ministry for Culture and Innovation from the source of the NRDIF.

6. References

[1] K. Richmond, "Estimating articulatory parameters from the acoustic speech signal," PhD thesis, University of Edinburgh, 2002.

[2] A. Toutios and K. Margaritis, "A rough guide to the acoustic-to-articulatory inversion of speech," *6th Hellenic European Conference of Computer Mathematics and its Applications, HERCMA-2003*, 2003.

[3] S. L. Metzger, J. R. Liu, D. A. Moses, M. E. Dougherty, M. P. Seaton, K. T. Littlejohn, J. Chartier, G. K. Anumanchipalli, A. Tu-Chan, K. Ganguly, and E. F. Chang, "Generalizable spelling using a speech neuroprosthesis in an individual with severe limb and vocal paralysis," *Nature Communications 2022 13:1*, vol. 13, no. 1, pp. 1–15, nov 2022.

[4] D. Porras, A. Sepúlveda-Sepúlveda, and T. G. Csapó, "DNN-based Acoustic-to-Articulatory Inversion using Ultrasound Tongue Imaging," in *International Joint Conference on Neural Networks*, Budapest, Hungary, 2019, pp. N–19 221.

[5] S. Udupa, A. Illa, and P. Ghosh, "Streaming model for Acoustic to Articulatory Inversion with transformer networks," in *Proc. Interspeech*, no. September, Incheon, Korea, 2022, pp. 625–629.

[6] T. G. Csapó and A. Sepúlveda, "Ultrasound tongue image synthesis for acoustic-to-articulatory inversion using convolutional and recurrent neural networks," *submitted to Infocommunications Journal*, 2022.

[7] Z.-H. Ling, K. Richmond, J. Yamagishi, and R.-H. Wang, "Integrating Articulatory Features Into HMM-Based Parametric Speech Synthesis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1171–1185, aug 2009.

[8] T. G. Csapó, C. Zainkó, L. Tóth, G. Gosztolya, and A. Markó, "Ultrasound-based Articulatory-to-Acoustic Mapping with WaveGlow Speech Synthesis," in *Proc. Interspeech*, 2020, pp. 2727–2731.

[9] W. Katz, T. Campbell, J. Wang, E. Farrar, J. Eubanks, A. Balasubramanian, B. Prabhakaran, and R. Rennaker, "Opti-speech: A real-time, 3D visual feedback system for speech training," in *Proc. Interspeech*, Singapore, Singapore, 2014, pp. 1174–1178.

[10] T. G. Csapó, T. Grósz, G. Gosztolya, L. Tóth, and A. Markó, "DNN-Based Ultrasound-to-Speech Conversion for a Silent Speech Interface," in *Proc. Interspeech*, Stockholm, Sweden, 2017, pp. 3672–3676.

[11] M. Stone, "A guide to analysing tongue motion from ultrasound images," *Clinical Linguistics and Phonetics*, vol. 19, no. 6-7, pp. 455–501, jan 2005.

[12] V. Ramanarayanan, S. Tilsen, M. Proctor, J. Töger, L. Goldstein, K. S. Nayak, and S. Narayanan, "Analysis of speech production real-time MRI," *Computer Speech and Language*, vol. 52, pp. 1–22, 2018.

[13] D. J. McFarland and J. R. Wolpaw, "EEG-based brain–computer interfaces," *Current Opinion in Biomedical Engineering*, vol. 4, pp. 194–200, dec 2017.

[14] R. A. Sharon and H. A. Murthy, "Correlation based Multi-phasal models for improved imagined speech EEG recognition," in *Proc. Workshop on Speech, Music and Mind (SMM 2020)*, 2020, pp. 21–25.

[15] M. Verwoert, M. C. Ottenhoff, S. Goulis, A. J. Colon, L. Wagner, S. Tousseyn, J. P. van Dijk, P. L. Kubben, and C. Herff, "Dataset of Speech Production in intracranial Electroencephalography," *Scientific Data 2022 9:1*, vol. 9, no. 1, pp. 1–9, jul 2022.

[16] G. Buzsáki, C. A. Anastassiou, and C. Koch, "The origin of extracellular fields and currents — EEG, ECoG, LFP and spikes," *Nature Reviews Neuroscience*, vol. 13, no. 6, pp. 407–420, may 2012.

[17] D. Dash, P. Ferrari, A. Babajani-Feremi, A. Borna, P. D. Schwindt, and J. Wang, "Magnetometers vs Gradiometers for Neural Speech Decoding," *IEEE EMBC*, vol. 2021, pp. 6543–6546, nov 2021.

[18] A. J. Casson, "Wearable EEG and beyond," *Biomedical engineering letters*, vol. 9, no. 1, pp. 53–71, feb 2019.

[19] G. Krishna, C. Tran, Y. Han, M. Carnahan, and A. H. Tewfik, "Speech Synthesis Using EEG," in *Proc. ICASSP*, online, 2020, pp. 1235–1238.

[20] S. Luo, Q. Rabbani, ·. Nathan, and E. Crone, "Brain-Computer Interface: Applications to Speech Decoding and Synthesis to Augment Communication," *Neurotherapeutics 2022*, vol. 1, pp. 1–11, jan 2022.

[21] C. Herff, D. Heger, A. de Pesters, D. Telaar, P. Brunner, G. Schalk, and T. Schultz, "Brain-to-text: decoding spoken phrases from phone representations in the brain," *Frontiers in Neuroscience*, vol. 9, p. 217, 2015.

[22] G. K. Anumanchipalli, J. Chartier, and E. F. Chang, "Speech synthesis from neural decoding of spoken sentences," *Nature*, vol. 568, no. 7753, pp. 493–498, apr 2019.

[23] M. Angrick, M. C. Ottenhoff, L. Diener, D. Ivucic, G. Ivucic, S. Goulis, J. Saal, A. J. Colon, L. Wagner, D. J. Krusienski, P. L. Kubben, T. Schultz, and C. Herff, "Real-time synthesis of imagined speech processes from minimally invasive recordings of neural activity," *Communications Biology*, vol. 4, no. 1, p. 1055, 2021.

[24] F. V. Arthur and T. G. Csapó, "Speech synthesis from intracranial stereotactic Electroencephalography using a neural vocoder," in *submitted*, 2022.

[25] S. Lesaja, C. Herff, G. D. Johnson, J. J. Shih, T. Schultz, and D. J. Krusienski, "Decoding Lip Movements during Continuous Speech using Electrocorticography," in *International IEEE/EMBS Conference on Neural Engineering, NER*, San Francisco, CA, USA, 2019, pp. 522–525.

[26] T. G. Csapó, F. V. Arthur, P. Nagy, and Ádám Boncz, "Towards Ultrasound Tongue Image prediction from EEG during speech production," in *Proc. Interspeech*, 2023.

[27] F. H. Guenther, J. S. Brumberg, E. Joseph Wright, A. Nieto-Castanon, J. A. Tourville, M. Panko, R. Law, S. A. Siebert, J. L. Bartels, D. S. Andreasen, P. Ehirim, H. Mao, and P. R. Kennedy, "A Wireless Brain-Machine Interface for Real-Time Speech Synthesis," *PLoS ONE*, vol. 4, no. 12, 2009.

[28] P. Favero, J. Berezutskaya, N. F. Ramsey, A. Nazarov, and Z. V. Freudenburg, "Mapping Acoustics to Articulatory Gestures in Dutch: Relating Speech Gestures, Acoustics and Neural Data," in *IEEE EMBC*, 2022, pp. 802–806.

[29] J. Amigó-Vega, M. Verwoert, M. C. Ottenhoff, P. L. Kubben, and C. Herff, "Decoding articulatory trajectories during speech production from intracranial EEG," in *BCI meeting*, 2023.

[30] M. Wairagkar, L. Hochberg, D. Brandman, and S. Stavisky, "Continuous speech synthesis and articulatory kinematics decoding from intracortical neural activity," in *BCI meeting*, 2023.

[31] T. G. Csapó, G. Gosztolya, L. Tóth, A. H. Shandiz, and A. Markó, "Optimizing the Ultrasound Tongue Image Representation for Residual Network-Based Articulatory-to-Acoustic Mapping," *Sensors*, vol. 22, 2022.

[32] S. M. Lulich, K. H. Berkson, and K. de Jong, "Acquiring and visualizing 3D/4D ultrasound recordings of tongue motion," *Journal of Phonetics*, vol. 71, pp. 410–424, 2018.

[33] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image Quality Assessment: From Error Visibility to Structural Similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, apr 2004.

[34] M. P. Sampat, Z. Wang, S. Gupta, A. C. Bovik, and M. K. Markey, "Complex Wavelet Structural Similarity: A New Image Similarity Index," *IEEE Transactions on Image Processing*, vol. 18, no. 11, pp. 2385–2401, nov 2009.

[35] K. Xu, T. G. Csapó, P. Roussel, and B. Denby, "A comparative study on the contour tracking algorithms in ultrasound tongue images with automatic re-initialization," *The Journal of the Acoustical Society of America*, vol. 139, no. 5, pp. EL154–EL160, may 2016.