IberSPEECH 2021
24-25 March 2021, Valladolid, Spain

Voice Restoration with Silent Speech Interfaces (ReSSInt)

Inma Hernáez Rioja[1], Jose A. Gonzalez-Lopez[2], Eva Navas[1], Jose Luis Pérez Córdoba[2], Ibon Saratxaga[1], Gonzalo Olivares [3], Jon Sanchez[1], Alberto Galdón[3], Victor García Romillo[1], Míriam González-Atienza[2], Tanja Schultz[4], Phil D. Green[5], Michael Wand[6], Ricard Marxer[7], Lorenz Diener[4]

HiTZ Center - Aholab, University of the Basque Country UPV/EHU, Spain
[2]SigMAT group, University of Granada, Spain
[3]Hospital Universitario Virgen de las Nieves, Granada, Spain
[4]Cognitive Systems Lab, University of Bremen, Bremen, Germany
[5]Speech and Hearing Group, University of Sheffield, UK
[6]Swiss AI Lab, ISDIA, Manno, Switzerland
[7]University of Toulon, France

inma.hernaez@ehu.eus, joseangl@ugr.es, eva.navas@ehu.eus

Abstract

ReSSInt aims at investigating the use of silent speech interfaces (SSIs) for restoring communication to individuals who have been deprived of the ability to speak. SSIs are devices which capture non-acoustic biosignals generated during the speech production process and use them to predict the intended message. Two are the biosignals that will be investigated in this project: electromyography (EMG) signals representing electrical activity driving the facial muscles and invasive electroencephalography (iEEG) neural signals captured by means of invasive electrodes implanted on the brain. From the whole spectrum of speech disorders which may affect a person's voice, ReSSInt will address two particular conditions: (i) voice loss after total laryngectomy and (ii) neurodegenerative diseases and other traumatic injuries which may leave an individual paralyzed and, eventually, unable to speak. To make this technology truly beneficial for these persons, this project aims at generating intelligible speech of reasonable quality. This will be tackled by recording large databases and the use of state-of-the-art generative deep learning techniques. Finally, different voice rehabilitation scenarios are foreseen within the project, which will lead to innovative research solutions for SSIs and a real impact on society by improving the life of people with speech impediments.

Index Terms: Silent speech interfaces, brain to speech conversion, EMG to speech, speech synthesis, voice conversion, deep neural networks.

1. Introduction

Speech is the first and foremost means of human communication. Unfortunately, many people are not able to speak, in particular those who have lost this ability through illness or disability. There are no many studies providing specific data about the prevalence of this disability. In [1] the authors conclude that 0.4% of the European population suffer from a speech impediment. In a later survey from 2011 [2], it is reported that 0.5% of people in Europe present 'difficulties' with communication. Focusing on Spain (data from the Spanish National Institute for Statistics (INE) published in 2008) there are more than 410,000 people with a disability to produce spoken messages [3]. For instance, laryngectomy patients (∼1200 total laryngectomies are performed every year in Spain [4]), whose voice box has been completely removed to treat larynx cancer, can no longer speak in a conventional way after the operation. Speech is also affected after brain damage, spinal cord injuries or neurodegenerative diseases such as amyotrophic lateral sclerosis (ALS), a disease which is expected to increase worldwide by 69% between 2015 and 2040 [5] due to an aging population and improved public healthcare. As this disease progresses, individuals can no longer communicate verbally and assistive devices that rely on nonverbal signals are needed for communication.

Voice loss is not only a problem for efficient communication, but also deprives the speaker of a central personal characteristic (namely, his/her own voice) which can in turn lead to occupational disability, personal isolation, and clinical depression [6]. In the absence of clinical procedures for repairing the damage caused by the above disorders, several methods are used to restore communication. However, all traditional methods are, in general, far from ideal. For laryngectomized patients, the 'gold-standard' method for voice restoration, the tracheoesophageal valve, requires frequent replacement every 3-4 months due to biofilm growth [7] and produces a masculine voice disliked by female patients. The electrolarynx, on the other hand, despite being relatively easy to use and safe, produces a very robotic and monotone voice. Esophageal speech, a method of speech production that involves oscillation of the esophagus, sounds gruff and masculine and is difficult to master. Additionally, esophageal speech is less intelligible both by humans and machines ([8,9]), which makes voice interaction with computer very difficult. Although voice conversion strategies have been investigated to improve the quality and intelligibility of these voices, there is still margin for improvement [10, 11].

Things are even worse for people who have suffered a brain stroke or neurodegenerative disease. These patients normally find themselves struggling with communication and need to use alternative methods, such as augmentative and alternative communication (AAC) devices, to communicate with their family and caregivers. These devices, usually with rates lower than 15 words per minute, are only suitable for short conversations.

In recent years, SSIs [12–14] have emerged as a promising alternative to restore oral communication by decoding speech from non-acoustic (silent) speech-related biosignals generated during speech production. Lip reading is the best-known form of silent speech communication. A variety of sensing modalities have been investigated to capture those biosignals, such as

10.21437/IberSPEECH.2021-28

vocal tract imaging [15], electromagnetic articulography (magnetic tracing of articulator movements) [16, 17], EMG [18–20], which captures facial muscle activity using surface electrodes, and iEEG [21–23], which captures the electrical activity in the brain. Since SSIs allow to capture speech without requiring any acoustic signal at all, they offer a fundamentally new solution to restore communication capabilities to speech-disabled persons.

The project described in this paper will investigate SSIs-based communication systems for restoring vocal communication to individuals who have been deprived of the ability to speak. In particular, ReSSInt will address two user groups: total-laryngectomy patients and individuals affected by brain damages. Each speech impairment will be addressed by a specific interface: EMG and iEEG. In the following sections we present the main characteristics and relevant previous works using these interfaces. We also describe the main objectives of the project and how it has been structured to achieve the objectives.

2. Silent Speech Interfaces

Two approaches have been proposed to decode speech from silent, speech-related biosignals [14]: silent speech-to-text and direct speech synthesis. In the first approach, automatic speech recognition (ASR) algorithms trained on silent speech data are used to decode speech from the input data. Text-to-speech (TTS) software can then be used to synthesize speech from the decoded text if required. Several works have shown promising results on silent speech-to-text for a variety of methods. For instance, in [24], a EMG-based silent speech recognizer was proposed. The system was evaluated on a corpus of phonetically-rich sentences recorded by $n = 8$ healthy persons, achieving a word error rate (WER) of 16.8%. In [25], EMG-based silent speech recognition was evaluated as assistive communication device for $n = 8$ laryngectomized patients, achieving an average WER of 10.3%.

Although still in a more preliminary stage, speech decoding from recordings of neural activity in anatomical regions involved in continuous speech production has also been shown to be feasible. Due to the potential advantages of the brain-to-text approach, breakthrough advances have been achieved in recent years. Thus, [26] demonstrated that phonetic features, such as place and manner of articulation, and voicing status, can be decoded during continuous speech production from eletrocorticography. Mugler *et al*. [27] was the first study reporting on decoding the full set of phonemes for American English, obtaining up to 36% accuracy when classifying phonemes within word productions and up to 63% accuracy for a single phoneme. The first study to address the task of decoding continuous speech from iEEG recordings was [21]. Seven patients undergoing surgery for epilepsy treatment where implanted with electrocorticography (ECoG) sensors (a type of intracranial EEG technique where electrode strips are places directly over the exposed surface of the brain) and, later, speech and ECoG signals were simultaneously recorded while the subjects read texts aloud. Acquired brain signals were used to train speech recognizers for each subject. Up to 75% word accuracy was reported when the vocabulary consisted on 10 possible words, and up to 40% when the user could choose between 100 words.

2.1. Direct speech synthesis from silent speech data

Though appealing, the silent speech-to-text approach lacks the real-time capabilities (i.e. low latency) that a SSI system for natural human speech communication would require. In this regard, previous studies have provided estimates on the maximum latency for an ideal SSI system. In oral communication, 100 to 300 ms of propagation delay causes slight hesitation on a partner's response and beyond 300 ms causes users to begin to back off to avoid interruption [28]. Studies on delayed auditory feedback, in which subjects received delayed feedback of her/his voice, found disruptive effects on speech production in subjects with delays starting at 50 ms and maxing out around 200 ms [29]. Altogether, these results suggest a $\sim$50 ms latency for an ideal SSI system, though latencies up to $\sim$100 ms may still be reasonable. These low latencies can only be achieved through the second SSII approach, direct speech synthesis, in which audible speech is directly generated from silent speech data by mapping the input silent data into a suitable speech representation (e.g. MFCCs) and then generating a waveform from the estimated speech parameters. Most commonly, deep neural networks (DNNs) [30] trained on time-aligned speech and silent data recordings (i.e. parallel data) are used to model the silent speech-to-speech mapping.

The research team of ReSSInt, as well as our international collaborators, have made significant contributions on the direct speech synthesis approach. Thus, in [16], we proposed a SSI system based on direct speech synthesis and electromagnetic articulography. The vocabulary consisted on digit sequences and consonant-vowel pairs. Our results showed that intelligible speech could be generated from articulatory data, although some phones were more mistakable than others (i.e. phones differing in their manner of articulation or voicing were hard to distinguish from the silent speech data). Building upon this work, in [17], we addressed the task of synthesizing continuous speech for a large vocabulary using recurrent neural networks. On average, the resulting speech was $\sim$75% intelligible, but for some subjects speech intelligibility reached up to $\sim$92%.

Although direct speech synthesis from EMG signals has experienced considerable advances in recent years, this technology still presents many challenges which keep it from becoming a product. The first limitation is the strong dependency of the results on the training session. Although array EMG sensors have provided greater signal stability and robustness in this sense (thus the relative position of the sensors is kept constant), there are still differences in data between different sessions [31, 32]. In [33] the authors show that even the training material (style, isolated words, syllables) can influence the quality of the results. More importantly, all the mentioned experiments are carried out on a speaker dependent fashion and speaker independence remains unsolved. Finally, for a real application of this kind of systems, real time performance must be achieved. Although there have been some recent attempts [34], this is still an open research issue.

In parallel with these findings, direct synthesis from neural signals is gaining growing attention due to the promises of restoring speech function in individuals unable to speak. In comparison with other sensing techniques, recording silent speech data directly from the brain has the advantage that speech is captured in an earlier form, which means that audio could be synthesized from the neural signals with lower latency. To date, however, only a few studies have addressed the task of generating speech directly from neural activity. The first study to report on this was [35] in which neural activity recorded from a completely-paralyzed individual was used to decode formant frequencies during imagined speech. As the user became engaged in more training sessions, he quickly improved with practice, learning to control the system to produce

better acoustics. Martin *et al.* [36] investigated the prediction of continuous speech from ECoG during imagined speech. In this study, $n = 7$ subjects were asked to read aloud (overt speech) and imagined reading (covert speech) short paragraphs of text. Later, models were trained for the overt condition to predict speech parameters from ECoG. These models were also applied to predict the speech parameters during the covert condition. Despite not being fully intelligible, human listeners were able to identify the reconstructed speech when they have to choose from a list of sentences. More recently, [22] proposed a two-stage approach in which they first transformed ECoG signals into anatomical representations of the vocal-tract articulators, and then transformed such intermediate representations into speech using bidirectional recurrent neural networks. Reconstructed speech waveforms for $n = 5$ volunteers were deemed quite intelligible by human listeners. Our collaborators from the University of Bremen have also made important contributions in this field. For instance, in [23, 37], they investigated the synthesis of speech from ECoG signals. Two approaches were used to map ECoG into speech: 3-dimensional convolutional neural networks (CNNs) and a concatenative, unit-selection approach. In general, despite not being fully intelligible, it was found that synthesized speech sounded natural and included features such as prosody and accentuation. Moreover, it was found that speech motor cortex provided more information for the reconstruction process than the other cortical areas.

3. Objectives of the project

Despite the promising results and advances achieved so far, SSI devices have not made it to the mass market. In our opinion, a major reason for this is the lack of focus on real-life use cases. In particular, the problem of inter-session and inter-speaker variability is not yet solved and requires intensive further investigation. Also, most works have not taken full advantage of recent advances in generative DNNs. For example, most systems have focused on predicting the spectral envelope while nowadays neural vocoders can generate prosodic information as well. Furthermore, most existing studies have been performed with able-bodied subjects, often relying on parallel recorded silent speech-and-acoustic signals. This excludes the important group of speech-disabled persons who have already lost their voice or have it severely impaired. Finally, many studies have used offline data, which disregards the fact that a user will expect acoustic feedback during the process of speaking silently. This feedback will allow the user to improve/adapt his/her own speaking patterns (we speak of coadaptation of the user and the device). Additionally, care needs to be taken to make the system flexible and easy-to-use, which implies lightweight and portable devices, fast enrollment, and graceful degradation in the case of processing errors.

In ReSSInt, we intend to overcome the limitations of both traditional voice rehabilitation methods and previous SSI studies by investigating SSI-based communication systems for restoring communication to individuals who have been deprived of the ability to speak. From the whole spectrum of speech disorders which may affect a person's voice, ReSSInt will address two conditions, each being the objective of a particular subproject:

- **Subproject 1**: total-laryngectomy patients. These persons still retain the control over their speech articulators and, hence, silent speech data reflecting the movements of the articulators can be easily captured using EMG.
- **Subproject 2**: neurodegenerative diseases and other traumatic injuries which may leave an individual paralyzed and, eventually, unable to speak. For many of these individuals, the only means of communication is through limited eye movements and blinking; however, for those with complete paralysis, even this type of communication may not even be possible. An SSI-based communication system could provide a more effective and efficient way to communicate. Such a technology could dramatically improve these people's lives and, arguably, its potential benefits would outweigh the risks of brain surgery for implanting iEEG electrodes.

For an SSI system to be truly beneficial for these persons, it must satisfy the following criteria, which have guided us in the definition of the main goals of the project:

- It must be able to generate intelligible speech with a reasonable quality and naturalness.
- The SSI system needs to be robust to intra- and inter-speaker differences.
- The system must be flexible enough to deal with a variety of rehabilitation scenarios, in particular:
    1. Patients who are able to record synchronous silent speech and acoustic data before losing their voice,
    2. Recordings of a patient's original voice may be available, but silent speech biosignals is only recorded after s/he has completely lost her/his voice,
    3. No recordings of the original voice are available, so a substitute voice (e.g. a voice donor, perhaps a close relative) needs to be used instead. The third scenario is particularly relevant to SP2 given the difficulty of recording speech for paralyzed patients.
- Finally, a practical SSI must be able to generate audio from silent speech data in close to real-time (latency $<$100 ms), so its user can receive synchronous acoustic feedback while speaking and can adapt her/his articulation style to improve the output.

We will accomplish these goals by taking advantage of the background work on speech synthesis and SSIs of both groups and by recording large datasets, which in turn will foster the use of cutting edge deep learning techniques to improve the performance beyond the state-of-the-art. The real-time system will play a central role during the evaluation phase to assess the performance of the SSI in terms of speech intelligibility, quality, and naturalness. This system will also pave the way for studies of user-in-the-loop strategies, where both the user and the system co-adapt themselves to optimize the output.

Summarizing, the specific objectives of the coordinated project are:

1. To explore the paths and advances in the application of state-of-the-art deep generative neural network architectures to improve the present quality and intelligibility of current SSIs using EMG and ECoG.
2. To develop corpus, databases, protocols and best practices for research on SSI in Spanish language.
3. To establish a new research line and, consequently, a research infrastructure for SSI in Spain.
4. To strengthen the links between two of the most consolidated research groups on speech technologies at the national level: Aholab at UPV/EHU and SiGMAT at UGR.

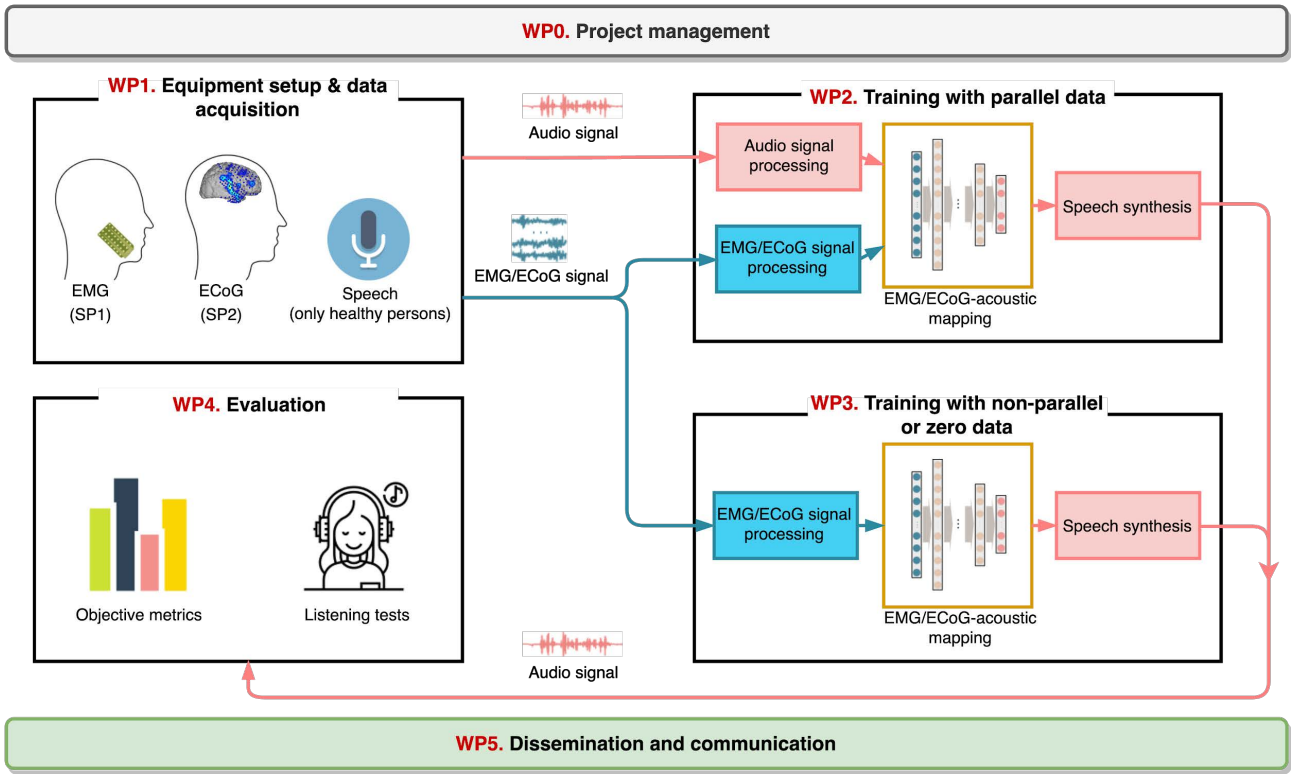

Figure 1: *Work package diagram for ReSSInt.*

4. Methodology

ReSSInt is split into 5 work packages, representing functional units which are necessary to tackle the goals of the project. The major workflow between the work packages is shown in Figure 1.

WP0 and WP5 are transverse work packages dedicated to the project management and dissemination/communication activities, respectively. WP1 is dedicated to the experimental part of the project (i.e. equipment setup, stimuli definition, participant recruitment and data acquisition). Data recorded in WP1 will be used by WP 2 and 3, where the foundational algorithms for direct speech synthesis from silent speech data will be developed. In particular, WP2 deals with training DNN architectures in a supervised fashion using parallel data. WP3, in contrast, is dedicated to training with non-parallel data. WP4 is dedicated to the evaluation of the algorithms and the user tests, which will provide continuous input, assessment, and improvement requests to the technical work packages. Our development model is iterative, with frequent interactions between work packages and partners. Most work packages run for the entire length of the project; yet research is structured by the subordinate tasks within the work packages.

5. Conclusions

In this paper we have presented the project ReSSInt, which will be executed in the period from July 2020 till June 2023. The project involves two research groups located in Spain (at the University of the Basque Country UPV/EHU and the University of Granada) in collaboration with expert researchers from other countries.

The beginning of the project has been greatly affected by COVID-19 and some of the task corresponding to the first year are suffering a delay. The acquisition of ECoG data in SP2 has not started yet, due to strict limitations on non-urgent surgery. Also, the acquisition of the equipment needed in SP1 has been delayed. However, the preparation of baseline systems is going on using external data provided by our collaborators. Our expectation is that we will recover from the initial delay and the main goals of the project remain viable.

Updated information about this project can be found at http://aholab.ehu.eus/ressint.

6. Acknowledgments

This work has been funded by the Agencia Estatal de Investigación ref.PID2019-108040RB-C21/AEI/10.13039/501100011033 and PID2019-108040RA-C22/AEI/10.13039/501100011033. Jose A. Gonzalez-Lopez holds a Juan de la Cierva-Incorporation Fellowship from the Spanish Ministry of Science, Innovation and Universities (IJCI-2017-32926).

7. References

[1] D. Dupré and A. Karjalainen, "Employment of disabled people in europe in 2002," *Statistics in focus*, pp. 3–26, 2003.

[2] Eurostat, "Population by type of basic activity difficulty, sex and age (hlth_dp040)," https://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=hlth_dp040&lang=en, 2011, accessed: 23-10-2019.

[3] "Portal Web del Instituto Nacional de Estadística. Encuesta de Discapacidad, Autonomía Personal y Situaciones de Dependencia 2008," https://www.ine.es/jaxi/Tabla.htm?path=/t15/p418/a2008/hogares/p01/modulo1/l0/&file=01002.px&L=0, accessed: 2020-12-12.

[4] P. D. de Cerio Canduela, I. A. González, R. B. Durban, A. S. Suárez, M. T. Secall, P. L. P. Arias, C. of Head, L. R. W. Group *et al.*, "Rehabilitation of the laryngectomised patient. recommendations of the spanish society of otolaryngology and head and neck surgery," *Acta Otorrinolaringologica (English Edition)*, vol. 70, no. 3, pp. 169–174, 2019.

[5] K. C. Arthur, A. Calvo, T. R. Price, J. T. Geiger, A. Chio, and B. J. Traynor, "Projected increase in amyotrophic lateral sclerosis from 2015 to 2040," *Nature communications*, vol. 7, no. 1, pp. 1–6, 2016.

[6] H. Danker, D. Wollbrück, S. Singer, M. Fuchs, E. Brähler, and A. Meyer, "Social withdrawal after laryngectomy," *European Archives of Oto-Rhino-Laryngology*, vol. 267, no. 4, pp. 593–600, 2010.

[7] S. Ell, "Candida-the cancer of silastic," *Journal of laryngology and otology*, vol. 110, no. 3, pp. 240–242, 1996.

[8] S. Raman, I. Hernaez, E. Navas, and L. Serrano, "Listening to Laryngectomees: A study of Intelligibility and Self-reported Listening Effort of Spanish Oesophageal Speech," in *Proc. IberSPEECH 2018*, 2018, pp. 107–111. [Online]. Available: http://dx.doi.org/10.21437/IberSPEECH.2018-23

[9] S. Raman, L. Serrano, A. Winneke, E. Navas, and I. Hernaez, "Intelligibility and listening effort of spanish oesophageal speech," *Applied Sciences*, vol. 9, no. 16, p. 3233, 2019.

[10] L. Serrano, D. Tavarez, X. Sarasola, S. Raman, I. Saratxaga, E. Navas, and I. Hernaez, "Lstm based voice conversion for laryngectomees." in *IberSPEECH*, 2018, pp. 122–126.

[11] L. Serrano, S. Raman, D. Tavarez, E. Navas, and I. Hernaez, "Parallel vs. non-parallel voice conversion for esophageal speech." in *INTERSPEECH*, 2019, pp. 4549–4553.

[12] B. Denby, T. Schultz, K. Honda, T. Hueber, J. M. Gilbert, and J. S. Brumberg, "Silent speech interfaces," *Speech Communication*, vol. 52, no. 4, pp. 270–287, 2010.

[13] T. Schultz, M. Wand, T. Hueber, D. J. Krusienski, C. Herff, and J. S. Brumberg, "Biosignal-based spoken communication: A survey," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2257–2271, 2017.

[14] J. A. Gonzalez-Lopez, A. Gomez-Alanis, J. M. Martín-Doñas, J. L. Pérez-Córdoba, and A. M. Gomez, "Silent speech interfaces for speech restoration: A review," *IEEE Access*, vol. 8, pp. 177 995–178 021, 2020.

[15] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip reading sentences in the wild," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 3444–3453.

[16] J. A. Gonzalez, L. A. Cheah, J. M. Gilbert, J. Bai, S. R. Ell, P. D. Green, and R. K. Moore, "A silent speech system based on permanent magnet articulography and direct synthesis," *Computer Speech & Language*, vol. 39, pp. 67–87, 2016.

[17] J. A. Gonzalez, L. A. Cheah, A. M. Gomez, P. D. Green, J. M. Gilbert, S. R. Ell, R. K. Moore, and E. Holdsworth, "Direct speech reconstruction from articulatory sensor data by machine learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2362–2374, 2017.

[18] T. Schultz and M. Wand, "Modeling coarticulation in emg-based continuous speech recognition," *Speech Communication*, vol. 52, no. 4, pp. 341–353, 2010.

[19] M. Wand and T. Schultz, "Session-independent emg-based speech recognition." in *Biosignals*, 2011, pp. 295–300.

[20] L. Diener, M. Janke, and T. Schultz, "Direct conversion from facial myoelectric signals to speech using deep neural networks," in *2015 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2015, pp. 1–7.

[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, 2019.

[23] M. Angrick, C. Herff, E. Mugler, M. C. Tate, M. W. Slutzky, D. J. Krusienski, and T. Schultz, "Speech synthesis from ecog using densely connected 3d convolutional neural networks," *Journal of neural engineering*, vol. 16, no. 3, 2019.

[24] M. Wand, M. Janke, and T. Schultz, "Tackling speaking mode varieties in emg-based speech recognition," *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 10, pp. 2515–2526, 2014.

[25] G. S. Meltzner, J. T. Heaton, Y. Deng, G. De Luca, S. H. Roy, and J. C. Kline, "Silent speech recognition as an alternative communication device for persons with laryngectomy," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 25, no. 12, pp. 2386–2398, 2017.

[26] F. Lotte, J. S. Brumberg, P. Brunner, A. Gunduz, A. L. Ritaccio, C. Guan, and G. Schalk, "Electrocorticographic representations of segmental features in continuous speech," *Frontiers in human neuroscience*, vol. 9, p. 97, 2015.

[27] E. M. Mugler, J. L. Patton, R. D. Flint, Z. A. Wright, S. U. Schuele, J. Rosenow, J. J. Shih, D. J. Krusienski, and M. W. Slutzky, "Direct classification of all american english phonemes using signals from functional speech motor cortex," *Journal of neural engineering*, vol. 11, no. 3, p. 035015, 2014.

[28] S. Na and S. Yoo, "Allowable propagation delay for voip calls of acceptable quality," in *International Workshop on Advanced Internet Services and Applications*. Springer, 2002, pp. 47–55.

[29] A. Stuart, J. Kalinowski, M. P. Rastatter, and K. Lynch, "Effect of delayed auditory feedback on normal speakers at two speech rates," *The Journal of the Acoustical Society of America*, vol. 111, no. 5, pp. 2237–2241, 2002.

[30] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[31] M. Janke and L. Diener, "EMG-to-speech: Direct generation of speech from facial electromyographic signals," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2375–2385, 2017.

[32] L. Diener, G. Felsch, M. Angrick, and T. Schultz, "Session-independent array-based emg-to-speech conversion using convolutional neural networks," in *Speech Communication; 13th ITG-Symposium*. VDE, 2018, pp. 1–5.

[33] L. Diener, S. Bredehoeft, and T. Schultz, "A comparison of emg-to-speech conversion for isolated and continuous speech," in *Speech Communication; 13th ITG-Symposium*. VDE, 2018, pp. 1–5.

[34] L. Diener, C. Herff, M. Janke, and T. Schultz, "An initial investigation into the real-time conversion of facial surface emg signals to audible speech," in *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2016, pp. 888–891.

[35] F. H. Guenther, J. S. Brumberg, E. J. Wright, A. Nieto-Castanon, J. A. Tourville, M. Panko, R. Law, S. A. Siebert, J. L. Bartels, D. S. Andreasen *et al.*, "A wireless brain-machine interface for real-time speech synthesis," *PloS one*, vol. 4, no. 12, p. e8218, 2009.

[36] S. Martin, P. Brunner, C. Holdgraf, H.-J. Heinze, N. E. Crone, J. Rieger, G. Schalk, R. T. Knight, and B. N. Pasley, "Decoding spectrotemporal features of overt and covert speech from the human cortex," *Frontiers in neuroengineering*, vol. 7, p. 14, 2014.

[37] C. Herff, L. Diener, M. Angrick, E. Mugler, M. C. Tate, M. A. Goldrick, D. J. Krusienski, M. W. Slutzky, and T. Schultz, "Generating natural, intelligible speech from brain activity in motor, premotor, and inferior frontal cortices," *Frontiers in neuroscience*, vol. 13, p. 1267, 2019.