IberSPEECH 2022
14-16 November 2022, Granada, Spain

ReSSInt project: Voice Restoration using Silent Speech Interfaces

Inma Hernáez Rioja[1], *Jose A. Gonzalez-Lopez*[2], *Eva Navas*[1], *Jose L. Pérez Córdoba*[2], *Ibon Saratxaga*[1], *Gonzalo Olivares*[3], *Jon Sanchez*[1], *Alberto Galdón*[3], *Victor García Romillo*[1], *Jesús del Castillo Cabrera*[2], *Inge Salomons*[1] *Eder del Blanco Sierra*[1]

HiTZ Center - Aholab, University of the Basque Country UPV/EHU, Spain
[2]Dept. of Signal Theory, Telematicas and Communications, University of Granada, Spain
[3]Hospital Universitario Virgen de las Nieves, Granada, Spain

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

Abstract

ReSSInt is a project funded by the Spanish Ministry of Science and Innovation aiming at investigating the use of Silent speech interfaces (SSIs) for restoring communication to individuals who have been deprived of the ability to speak. These interfaces capture non-acoustic biosignals generated during the speech production process and use them to predict the intended message. In the project two different biosignals are being investigated: electromyography (EMG) signals representing electrical activity driving the facial muscles and intracraneal 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. In this paper we describe the current status of the project as well as the problems and difficulties encountered in its development.

Index Terms: Silent speech interfaces, Brain-computer interfaces, EMG to speech, EEG, speech synthesis, voice conversion, deep neural networks.

1. Introduction

Speech is one of the most unique skills of human beings and our primary means of communication. This is the reason why speech and language disorders have a major impact on the quality of life of people who suffer them, affecting not only their daily communication, but having also important economic consequences for them.

In an attempt to help these people, this coordinated project aims to develop revolutionary assistive technology to restore them the ability to communicate again. In particular, we will develop novel signal processing algorithms to decode speech from non-audible, speech-related biosignals generated by the human body during speech production [1, 2]. To this end, we will record the following biosignals from two target groups:

- **Subproject 1 (SP1)**, lead by the University of the Basque Country, will focus on total-laryngectomy patients, who have lost their voice after having the larynx removed as a treatment for throat cancer, but still retain the control over most of their upper vocal track, including jaw, lips and tongue. For this group, the electrical activity driving the facial muscles will be captured using EMG [3].
- **Subproject 2 (SP2)**, lead by the University of Granada, will focus on patients with neurodegenerative diseases or brain damage that have their speech affected. For many of these individuals, the only means of communication is through limited eye movements and blinking; however, for those with complete paralysis but intact cognitive skills, even this type of communication may not even be possible. For this patient group, we will record brain activity using either non-invasive electroencephalography (EEG) or iEEG in order to explore the feasibility to decode speech from brain activity.

With these target groups in mind, the objectives of this coordinated project are as follows:

1. Firstly, we will record a database with time-synchronous recordings of EMG and speech (SP1) or iEEG and speech (SP2).
2. Using the datasets resulting from the previous objective, we will develop high-quality baseline systems for the synthesis of speech from either EMG (SP1) or iEEG (SP2) using state-of-the-art generative deep techniques.
3. While the previous objective deals with the training of deep neural networks (DNNs) when parallel data is available, this third objective will deal with the problem of training such techniques when there is no parallel data. For instance, when it is not possible to record the audio from the participants because they suffer an speech impediment. To address this problem, we will explore different alternative solutions, including the use of temporal alignment algorithms [4, 5] or state of the art non-parallel training algorithms for DNNs such as [6].
4. Finally, SP1 will also explore the viability of real-time EMG-to-speech conversion. If this is possible, it would clear the path for a radically new solution for speaking after a laryngectomy.

In the rest of the paper, we report the progress so far in this project, focusing on the acquired data but also including the obstacles we have found and the risk mitigation mechanism we have adopted.

2. EMG reference systems

One of the goals of this project is to develop a baseline EMG to Speech system using SoA technology which could serve as reference for future developments. With this goal in mind, a set of conversion experiments using a established methodology were performed using the available EMG databases, namely the freely available trial subset of the EMG-UKA parallel EMG-Speech corpus [7] and the CSL-EMG-Array [8]. However our attempts to obtain intelligible speech from those databases were

10.21437/IberSPEECH.2022-46

not successful. In fact, the literature usually reports on objective distortion and intelligibility measures such as mel cepstral distortion (MCD) or STOI [9, 10] and rarely reports on the results of human transcription tasks. One of this rare case is the work of D. Gaddy who obtained a 32.3% WER for speech generated from 19 hours of EMG data from a single speaker [11].

On the other hand, the use of an array of sensors such as the one used to obtain the CSL-EMG-Array is discouraged in [12]. This kind of sensors was also tested in our laboratory and we were not able to obtain good quality EMG signals. As a consequence of that, we turned our investigations to the use of individual sensors as in [7] and many others [13, 14]. Additionally, concentric bipolar electrodes were also tested, but although they are more compact, use less space and are easier to manage, they also provided lower quality EMG signals and performed worst in our classification experiments.

Finally, the experiments oriented towards performing recognition (continuous recognition or isolated commands recognition tasks) seem to be more successful in the literature (see for example [15, 16, 17]).

All these facts led us to the following design decisions:

- Beside sentences, we decided to include phones (VCV combinations) in our corpora as well as isolated words to widen the range of possible future experiments.
- We designed and performed a set of phone classification experiments, aimed at validating our database recording setup and methodology. The main results of these experiments are described in [18].
- Speech conversion experiments are being conducted presently using our own data and using modern Deep network architectures.
- Continuous speech recognition experiments are also going to be performed.
- Due to the low expected performance of the EMG-based conversion system, we added video recordings of the face of the participants. We think that adding this modality can help in the future to improve the results.

3. Recording setup

The acquisition of data was a key Work package (WP) of the project and has been the main developed task during the first half of the project. The first step was to define and establish the recording protocols in order to obtain the positive reports from the Ethics Committees. In this section we describe these recording protocols and procedures.

3.1. Setting up EMG acquisition

In SP1, we are recording EMG signals using a set of 8 channels, each targeting a different muscle. This setup was chosen after performing a set of phone classification experiments varying the total number of electrodes and electrode locations. EMG-signals are sampled at 2 048 Hz and captured by a Quattrocento bio-electrical amplifier. Speech signals are recorded with a Neumann TLM103 (diaphragm) microphone using a sampling frequency of 16 kHz.

Simultaneously, video recordings are being captured using an Intel Realsense D415 depth camera with a framerate of 33fps. This camera records color images that have been acquired with a 1280 x 720 pixel resolution using 24 bits per pixel and depth images that have been obtained with a 640 x 480 pixel resolution using 16 bits per pixel.

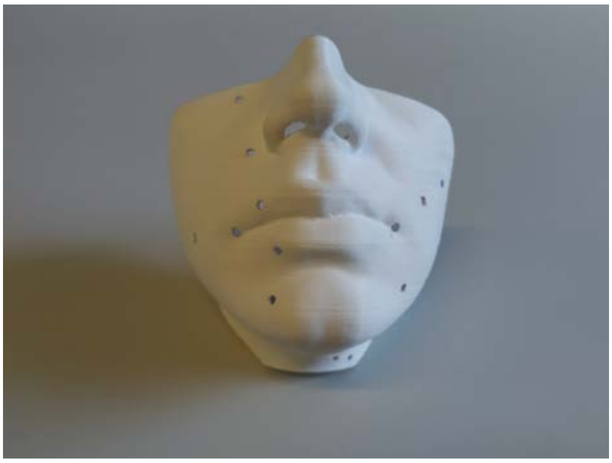

Figure 1: *3D mask with electrode locations from one participant*

Each recording session lasts around two hours including all the preparation needed. In order to control and reduce inter-session variability a 3D mask is made containing the exact positions of the electrodes. To do this, the participant's face is scanned with a mark on the exact position of the electrodes. Then a 3D mask is printed with holes at those positions (see Fig. 1). In the following sessions, the mask is used to precisely position the electrodes.

The recording time in each session never exceeds one hour, to ensure a good attaching of the electrodes. Even so, it is very common for the electrodes to come off being the cause of invalidation of some acquired data. The sessions are time spaced, leaving at least one week between them. The participants will receive a small gratification after finishing the recordings.

3.2. Setting up iEEG acquisition

In SP2, we are recording iEEG and audio signals from patients with implanted EEG electrodes while they perform a series of language production and understanding tasks, as described in more detail in [19]. Patients are recruited from those admitted to the Epilepsy Surgery Unit of the "Hospital Virgen de las Nieves" in Granada, Spain. These are patients suffering drug-resistant epilepsy who have intracranial EEG electrodes implanted to determine the epileptogenic zone, that is, the area of the brain generating the epileptic seizures. Thus, the electrode placement is based solely on clinical needs for diagnostic purpose.

Either electrocorticography (ECoG) or stereotactic electroencephalography (sEEG) is used to capture brain activity from the patients. The choose between one technology or the other is based on clinical needs only. ECoG [20] uses electrode strips placed directly on the exposed surface of the brain to record activity in the cortical surface. SEEG [21], on the other hand, utilizes localized deep electrodes with the advantage that brain activity from inner regions can be monitored. Although scalp EEG was also recorded using 64 shielded Ag/AgCl electrodes, we did not process these data further because of the poor signal quality resulting from the placement of postoperative bandages on the scalp.

The recording setup used in this subproject to record simultaneously audio and iEEG signals is shown in Fig. 2. Neural data was recorded using a Natus Quantum amplifier with 256 channels (Natus Medical Inc., Middleton, USA) and streamed

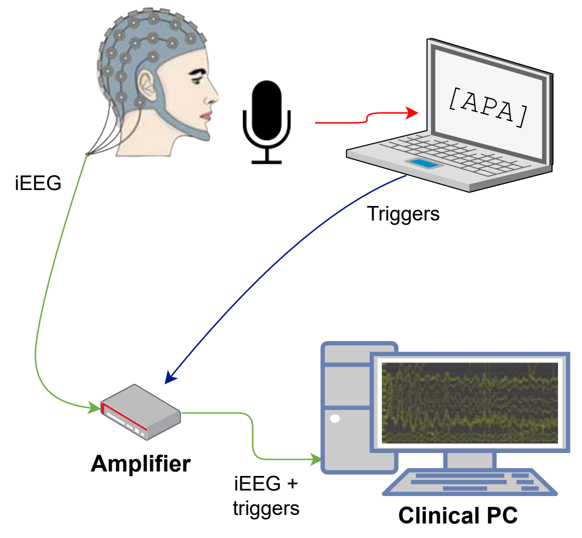

Figure 2: *Setup used in SP2 for the recording of time-synchronous iEEG and audio signals while participants performed language production tasks.*

via TCP/IP to a clinical PC for storage and clinical inspection. Recordings were continuously digitized at a sampling rate of 1024 Hz. Speech from the patients was recorded using a cardioid microphone at 44.1 kHz and stored on a Dell Inspiron 15” laptop where participants performed the experimental tasks. Both streams were synchronized by synchronization triggers send from the laptop to the Natus amplifier. Lab streaming layer (LSL) [22] was used in the laptop to synchronize audio and experimental markers.

4. Databases

4.1. ReSSint dabatase: EMG and Video

The designed text corpus being used in the recordings contains different recording tasks:

VCV pseudowords: nonsense words including VCV structures (a total of 110 combinations).

Isolated words: 100 isolated words, intended for limited vocabulary experiments.

Sentences: 1300 sentences, 700 from the Sharvard corpus [23] and 600 from Ahosyn [24]. These sentences are organized in one subset with 100 sentences, and 4 subsets of 300 sentences each.

Most sessions include recordings from the nonsense words set, the 100 isolated words set and the subset of 100 sentences. The rest of the session varies mainly in the used subset of sentences.

All items in the corpus can be produced in two main modes: audible, where the audio signal is produced and recorded; and silent, where the participant articulates the presented item, but no audio is produced. In silent mode, the participants are previously trained and asked to articulate clearly.

A total of 8 speakers are planned for recording, two of them laryngectomees. The laryngectomees will only record in silent speech mode but the rest of the speakers will provide recordings in both modes. Not all speakers will be recording the same amount of sessions: 2 speakers will be recording a total of 8 recording sessions, 4 speakers will record a total of 4 sessions, and finally the laryngectomees are planned to record a total of two sessions. Table 1 shows the present status of the performed recordings.

Table 1: *Basic information about the status of the EMG recordings.*

Id.	Gender	Age	Recorded Sessions	Recorded Audio Duration
S01	M	29	5	1:39:58
S02	F	29	8	3:08:47
S03	M	51	2	1:00:02
S04	F	46	3	1:09:26
S05	M	45	1	0:29:40
S06	F	61	6	1:25:13

A full detailed description of the contents and modes of each session for each speaker is out of the scope of this paper and will be done once all the recordings are finished.

4.2. UGR database: iEEG and Audio

So far, in SP2, data from 10 epilepsy patients have been recorded (mean age 37.8 ± 12.1 years; 6 female, 1 male), as shown Table 2. For each participant, the table shows some basic demographic information, the language production tasks performed, and the type of invasive EEG technology used to capture brain activity during the language tasks. In total, approximately 9 hours of data have been recorded so far for all patients.

Table 2: *Basic demographic information, details of the tasks performed by each patient and the type of invasive EEG electrodes used to capture brain activity.*

Id.	Gender	Age	iEEG electrodes	Tasks
S01	F	27	ECoG	VCV
S02	M	24	sEEG	Sentences
S03	M	30	sEEG	Sentences
S04	M	20	sEEG	Sentences
S05	F	48	sEEG	VCV, Picture naming
S06	M	54	sEEG	VCV, Picture naming
S07	F	36	sEEG	VCV, Picture naming
S08	F	41	sEEG	VCV, Picture naming
S09	F	52	sEEG	VCV, Picture naming
S10	F	46	sEEG	VCV, Picture naming

As described in more detail in [19], the following tasks were designed to evaluate the cognitive processes underlying language production at the neural level:

VCV pseudowords: Participants have to read aloud or imagine speaking, but without actually speaking, a list of 50 vowel-consonant-vowel pseudwords, such as APA, UJU, EME, etc. The list 0f 50 VCVs was created by combining the 5 vowels in Castillian Spanish with 10 consonants.

Sentences: In this task, the patients have to read aloud a set of 100 sentences extracted from the AhoSLABI corpus [25]. This corpus contains 100 phonetically balanced sentences encompassing all the phonemes in Castilian Spanish.

Picture naming: Participants have to name aloud or imagine naming a set of 30 images extracted from the phase II of the bank of standardized stimuli (BOSS) image dataset [26]. Five images from 6 semantic categories (body parts, animals, food, clothes, kitchen items and utensil and musical instruments) were used for this task.

5. Dissemination to society

The main scientific goals of this project aim at improving the life of people with communication disabilities. Dissemination to society of the used technology, the possibilities it offers but also its present limitations is important. A well-informed society is able to take decisions about complex problems that can sometimes affect its own health and welfare. We describe here some of the dissemination activities where we have participated in order to make our knowledge accessible to the whole society. Most of the activities have been organized by the university, faculty or department:

- we took part on the activity ”The secrets of telecommunications” organized in the context of the ”European researchers night” [27]. The target audience were young students of secondary (11-15 years old students).
- we presented our project in the ”Open doors event” [28] organized by the Faculty of Engineering of Bilbao, where the target audience are students of last year of high school (16-17 years old students).
- we presented a poster of the project in the ”XXI. Semana de la Ciencia, la tecnología y la innovación” organized by the UPV/EHU (5-7 November 2022).
- we presented the project in the event ”Mas mujeres en Tecnologías de Telecomunicación” [29]. The target audience were teachers and those responsible for the guidance of secondary students.
- we gave two talks in the form of interviews in the ”Jornadas Internacionales de Comunicación Aumentativa y Alternativa asistida por Tecnología” [30, 31].
- we presented the project to ”HiTZ Basque Center for Language Technology” [32].
- we participated in the initiative ”rEgaLA tu voz” together with the company Irisbond [33].

6. Conclusions

In this paper we have described the current state of the project ReSSInt, a coordinated project funded by the Ministry of Science and Innovation, which is being be executed 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.

ReSSInt is an interdisciplinary project whose aim is to deliver real societal impact by developing assistive technology to restore communication to people who have lost the ability to speak. To achieve this challenging goal, the project makes use of cutting-edge sensors and deep learning techniques in order to decode speech from either EMG or iEEG signals captured from the potential users of this technology.

The beginning of the project was greatly affected by COVID-19 and some of the tasks corresponding to the first year were delayed, mainly those related to participant recruitment. Thus, the main results of the project until now mostly consist in the compiled resources and databases, while the algorithmic part is still under active development. With these databases in place and the know-how of both groups in speech technology, we expect the objectives of the project to start being full-filled in the near future.

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

7. 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.

8. References

[1] 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.

[2] 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.

[3] K. R. Mills, “The basics of electromyography,” *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 76, no. suppl 2, pp. ii32–ii35, 2005.

[4] J. A. Gonzalez-Lopez, M. Gonzalez-Atienza, A. Gomez-Alanis, J. L. Pérez-Córdoba, and P. D. Green, “Multi-view temporal alignment for non-parallel articulatory-to-acoustic speech synthesis,” *arXiv preprint arXiv:2012.15184*, 2020.

[5] J. A. Gonzalez-Lopez, A. Gomez-Alanis, J. L. Pérez-Córdoba, and P. D. Green, “Non-parallel articulatory-to-acoustic conversion using multiview-based time warping,” *Applied Sciences*, vol. 12, no. 3, p. 1167, 2022.

[6] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[7] M. Wand, M. Janke, and T. Schultz, “The EMG-UKA corpus for electromyographic speech processing,” in *Interspeech*, 2014, pp. 1593–1597.

[8] L. Diener, V. M. Roustay, and T. Schultz, “Csl-emg array: An open access corpus for emg-to-speech conversion.” in *INTERSPEECH*, 2020, pp. 3745–3749.

[9] 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.

[10] L. Diener and T. Schultz, “Investigating objective intelligibility in real-time emg-to-speech conversion.” in *INTERSPEECH*, 2018, pp. 3162–3166.

[11] D. Gaddy and D. Klein, “An improved model for voicing silent speech,” in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Association for Computational Linguistics, Aug. 2021, pp. 175–181.

[12] L. Diener, “The impact of audible feedback on emg-to-speech conversion,” Ph.D. dissertation, Universität Bremen, 2021.

[13] D. Gaddy and D. Klein, “Digital voicing of silent speech,” in *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

[14] 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.

[15] A. Ratnovsky, S. Malayev, S. Ratnovsky, S. Naftali, and N. Rabin, “EMG-based speech recognition using dimensionality reduction methods,” *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–11, 2021.

[16] M. Wand, T. Schultz, and J. Schmidhuber, “Domain-Adversarial Training for Session Independent EMG-based Speech Recognition,” in *Interspeech*, 2018, pp. 3167–3171.

[17] K. Proroković, M. Wand, T. Schultz, and J. Schmidhuber, "Adaptation of an EMG-Based Speech Recognizer via Meta-Learning," in *2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, 2019, pp. 1–5.

[18] E. del Blanco, I. Salomons, E. Navas, and I. Hernaez, "Phone classification using elecromyographic signals," in *Proc. IberSPEECH 2022*, 2022, p. to appear.

[19] J. A. Gonzalez-Lopez, A. Galdón, G. Olivares, S. Raman, D. Muñoz, P. Macizo, J. L. Pérez-Córdoba, A. M. Peinado, A. M. Gomez, V. Sanchez, and A. B. Chica, "Clinical applications of neuroscience: Locating language areas in epileptic patients and restoring speech in paralyzed people," in *Submitted to IberSPEECH*, 2022.

[20] E. C. Leuthardt, K. J. Miller, G. Schalk, R. P. Rao, and J. G. Ojemann, "Electrocorticography-based brain computer interface-the seattle experience," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 14, no. 2, pp. 194–198, 2006.

[21] C. Herff, D. J. Krusienski, and P. Kubben, "The potential of stereotactic-eeg for brain-computer interfaces: current progress and future directions," *Frontiers in neuroscience*, vol. 14, p. 123, 2020.

[22] C. Kothe, "Lab streaming layer (LSL)," 2014. [Online]. Available: https://github.com/sccn/labstreaminglayer

[23] V. Aubanel, M. L. G. Lecumberri, and M. Cooke, "The Sharvard Corpus: A phonemically-balanced Spanish sentence resource for audiology," *International Journal of Audiology*, vol. 53, no. 9, pp. 633–638, Sep. 2014.

[24] I. Sainz, D. Erro, E. Navas, I. Hernáez, J. Sanchez, I. Saratxaga, and I. Odriozola, "Versatile speech databases for high quality synthesis for Basque," in *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, May 2012.

[25] L. S. García, S. Raman, I. H. Rioja, E. N. Cordón, J. Sanchez, and I. Saratxaga, "A spanish multispeaker database of esophageal speech," *Computer Speech & Language*, vol. 66, p. 101168, 2021.

[26] M. B. Brodeur, K. Guérard, and M. Bouras, "Bank of standardized stimuli (BOSS) phase ii: 930 new normative photos," *PloS one*, vol. 9, no. 9, p. e106953, 2014.

[27] J. Sanchez de la Fuente, "Los secretos de las telecomunications." [Online]. Available: https://www.ikertzaileengaua-ehu.org/programacion/ciencia-circular/secretos-de-las-telecomunicaciones/

[28] "Jornadas de puertas abiertas 2022." [Online]. Available: https://www.ehu.eus/es/web/bilboko-ingeniaritza-eskola/alumnado/jornada_de_puertas_abiertas/bloques_formativos/bloque_tics

[29] "Mas mujeres en tecnologías de telecomunicación." [Online]. Available: https://www.ehu.eus/es/web/kis/jornada_eso

[30] "Jornadas internacionales de comunicación aumentativa y alternativa asistida por tecnología." [Online]. Available: https://www.upo.es/investiga/capacesdecomunicar/entrevista-inmaculada-hernaez/

[31] "Jornadas internacionales de comunicación aumentativa y alternativa asistida por tecnología." [Online]. Available: https://www.upo.es/investiga/capacesdecomunicar/entrevista-jose-andres-gonzalez/

[32] "Webinar en HiTZ: The ReSSInt project." [Online]. Available: https://aholab.ehu.eus/ressint/dissemination-webinarHITZ/

[33] "'¿Me regalas tu voz?': la historia de los bancos de voces para enfermos de ELA." [Online]. Available: https://cadenaser.com/nacional/2022/06/20/me-regalas-tu-voz-la-historia-de-los-bancos-de-voces-para-enfermos-de-ela-cadena-ser/