ReSSint Database Description

Aholab Signal Processing Laboratory
Dec 2023

Grant PID2019-108040RB-C21/AEI/10.13039/501100011033

Content

Introduction

The database described in this report has been generated in the framework of the ReSSInt project [1][2] that aims to investigate the use of silent speech interfaces to restore communication in people who have been deprived of the ability to speak. Silent speech interfaces (SSIs) are devices designed to capture non-acoustic biological signals produced during the speech production process and utilize them to predict the intended message. While SSIs have traditionally been explored primarily within the realm of speech recognition, specifically Silent-Speech-to-Text applications, the ReSSint project takes a distinctive approach by concentrating on direct speech synthesis techniques. This involves the direct generation of the speech waveform corresponding to the captured biosignals.

The non-acoustic biosignals that are used in this work are EMG signals or, more specifically, surface (i.e., non-invasive) EMG [3]. Electromyography is a technique used to measure and record the electrical activity of muscles. When a muscle is active, it produces an electrical signal, called an action potential that can be detected by an electrode placed on the skin over the muscle. Since for this project we are interested in speech, we target muscles in the face and the neck.

Background

In order to develop an EMG-to-speech SSI, a large database of EMG and speech data is required. The main idea is to obtain a model trained on large amounts of EMG and speech data. To ensure the generalization capabilities of the models, it is important to use a diverse and representative dataset for training.

Most of the works and databases related to SSIs [4][5][6] have been developed for English, and some for other languages [7][8][9][10][11]. However, none of these works focuses on Spanish, and therefore this project intends to narrow that gap.

Two prominent challenges in the development of these interfaces are the dependency of the trained models on the session (session dependency) and on the speaker (speaker dependency). Session dependency arises from the variations observed in the obtained EMG signals when electrodes are positioned differently on the subject's face. Speaker dependency is due to differences in the way of speaking from person to person. Additionally, an important issue arises from inadequate adhesion of the electrodes to the skin, leading to the detachment of electrodes over time and the generation of noisy signals. As a consequence, long sessions are difficult to carry out thus limiting the amount of data available per session.

Database Overview

The database has been designed to contribute to the development of SSIs in Spanish, with a particular focus on laryngectomees. Standard speakers with no oral pathology and people who have undergone a laryngectomy will take part in the database recording.

We have acquired recordings of 8 surface EMG sensors while individuals pronounce words and sentences in Spanish. Each individual has recorded several sessions spanning several weeks. During certain recording sessions, speech has been articulated and recorded, while in others, speech only involved mouth movements without actual sound production. Additionally, the sessions have been captured via video using a 3D camera, ensuring synchronized images of the recording process are available.

In each recording session, three different kinds of items are recorded, namely: nonsense words including vowel-consonant-vowel structures, isolated words, and sentences. The sentences are taken from the Sharvard Corpus [12] and from a text corpus called AhoSyn that was developed to record TTS databases [13].

Data Collection

In this section, we describe the methods used for EMG data collection, as well as the recording setup and procedure.

Recording setup

Each session has been recorded in a sound-proof room using a silent computer, in an attempt to reduce interference with the audio and EMG signals as much as possible. The selected EMG sensors are bipolar CDE-C[1] with concentric connectors. The EMG signals have been recorded with a Quattrocento bio-electrical amplifier at a sampling frequency of 2048 Hz, and the voice has been captured with a Neumann TLM103 (diaphragm) microphone with a sampling frequency of 16 kHz.

For the acquisition and synchronization of the audio and EMG signals, we have used publicly available software[2] developed by the Cognitive Systems Lab at the University of Bremen. This software also incorporates a user interface that facilitates the process of recording. Additionally, synchronized video of the facial movements has been captured with an Intel RealSense D415 RGB-depth camera. These video signals are meant to provide supplementary data and allow multi-modal experiments, such as automatic lip reading.

[1] https://otbioelettronica.it/en/product/cde-c/

[2] https://github.com/cognitive-systems-lab/EMG-GUI

To ensure that the EMG signals and the audio are well aligned, a synchronization signal is shared between the Quattrocento bio-electrical amplifier and the sound device. The synchronization signal is raised by the speaker through the recording interface when they start the recording of each utterance, and it is descended when they finish the recording. The Quattrocento amplifier creates the synchronization signal and it is saved together with the EMG signals as an additional channel. At the same time, it is outputted through an analog auxiliary output, which is introduced in one of the channels of the sound interface. The stereo audio signals contain the speech signal in the left channel and the synchronization signal in the right channel.
Figure 1 shows a picture of the complete acquisition setup.

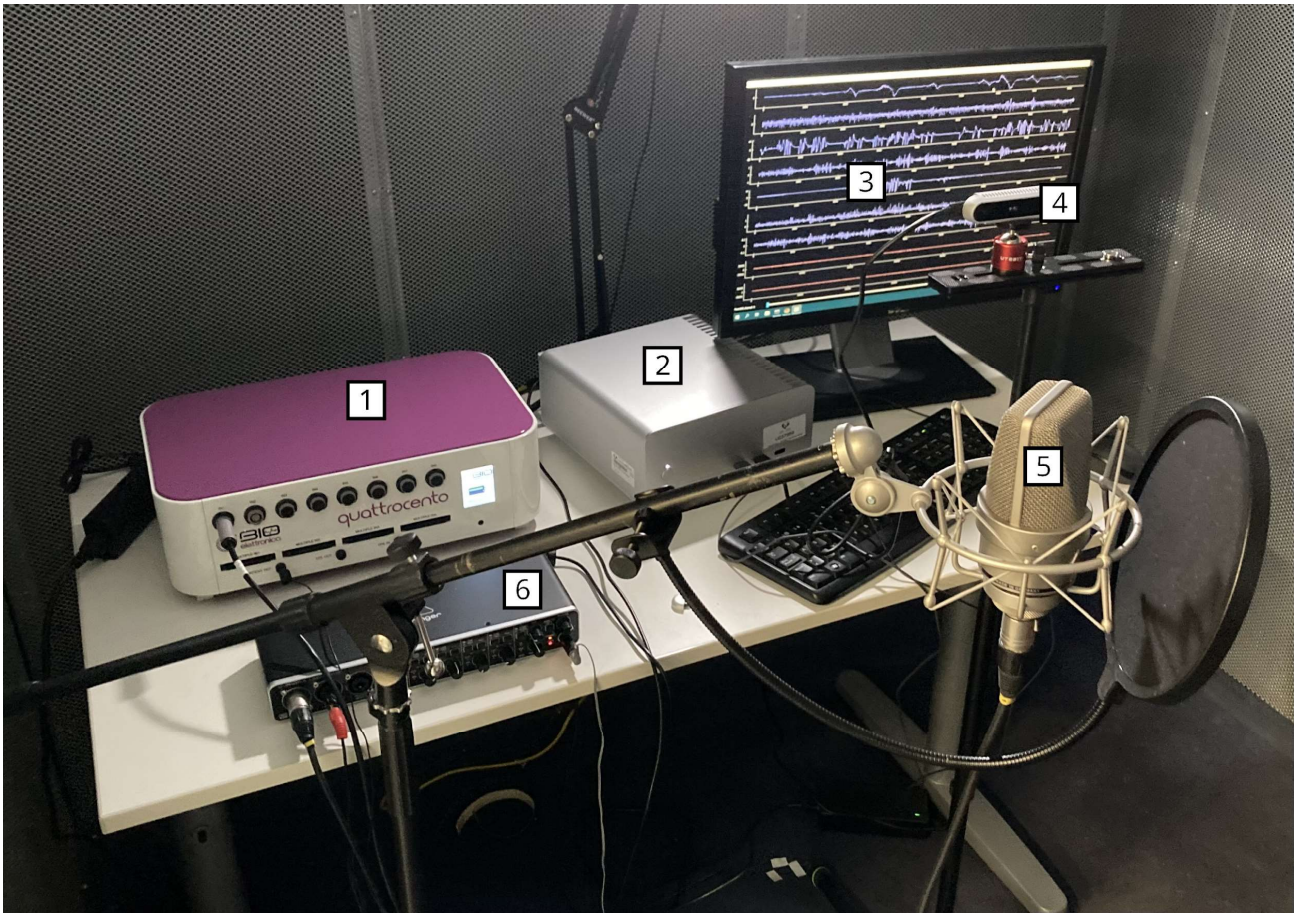

Figure 1 - Acquisition setup: (1) EMG amplifier; (2) silent computer; (3) computer screen with acquisition software; (4) camera; (5) microphone; (6) audio interface.

Recording procedure

Before commencing each recording session, a technician positioned the EMG sensors in the appropriate locations after cleaning the subject's face and throat. The recordings were conducted in a semi-professional, soundproof recording cabin. Inside the cabin, the subject was situated facing both the microphone and the 3D camera.

A dedicated technician was present in the cabin throughout the recording sessions, offering continuous assistance. While conducting the recordings, the technician stationed in the recording booth will continually monitor the generated signals to prevent errors caused by electrode detachment, interferences, mispronunciations or any other potential issues. The database contains both audible speech, where subjects read aloud the provided prompt, and silent recordings, where subjects mouthed the words and sentences displayed on a computer screen.

Before the initial session with each speaker, a trained technician determined the precise positions of the electrodes using facial landmarks and a measuring tape. For instance, in locating the risorius or laughing muscle, we positioned the first electrode adjacent to the corner of the mouth and placed the second electrode in the direction of the earlobe on the same side of the face. Three points were marked: one on each outer side of both electrodes and one in the middle. This process was repeated for all eight electrode pairs, resulting in a total of 24 reference marks.

To minimize inter-session variability in audio and video, we maintained consistent positions for the subject, microphone, and video camera across all sessions. Furthermore, a personalized 3D mask (see Figure 2) was built for each participant with the data acquired using a 3D scanner in their first recording. This approach ensured the constancy of electrode locations throughout all sessions. A 3D printing professional generated the 3D scan of the face and produced a mask with holes corresponding to the reference marks. Afterwards, during subsequent sessions, we replicated the marks on the subject's face using the holes and positioned the electrodes accordingly.

Before each recording session, speakers are given instructions to articulate their speech slightly more than they would in normal conversation.

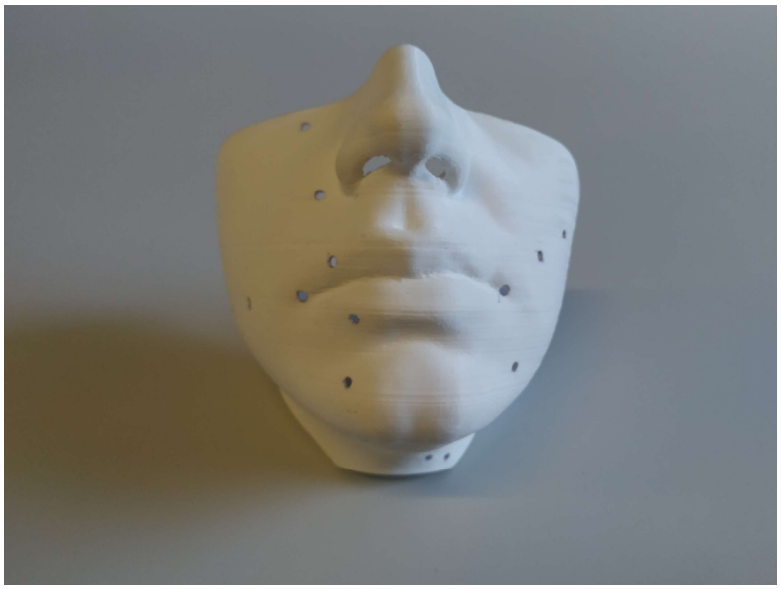

Figure 2 - A personalized 3D mask. The holes are used as reference marks to find the positions of the electrodes in the subject's face.

Electrode setup

Previous studies have employed various approaches in determining the optimal electrode setup, such as targeting muscles specifically [14][15][16][17][18], analyzing anatomical regions [19], and identifying patterns in a high-density electrode setup [20]. Knowing that an activation potential travels along the muscle as a wave, the most appropriate way to use bipolar acquisition is to place the two electrodes longitudinally over the muscle. In our approach, we opted to target muscles individually and conducted a pilot study, which involved addressing all relevant superficial muscles in the face and neck to identify the most suitable muscles for the task. As a result of this study, the final setup (see Figure 3) slightly differs from those used in the previously mentioned studies. These are the targeted muscles (using one channel each):

- Levator labii superioris (channel 1)
- Masseter (channel 2)
- Risorius (channel 3)
- Depressor labii inferioris (channel 4)
- Zygomaticus major (channel 5)
- Depressor anguli oris (channel 6)
- Anterior belly of the digastric (channel 7)
- Stylohyoid (channel 8)

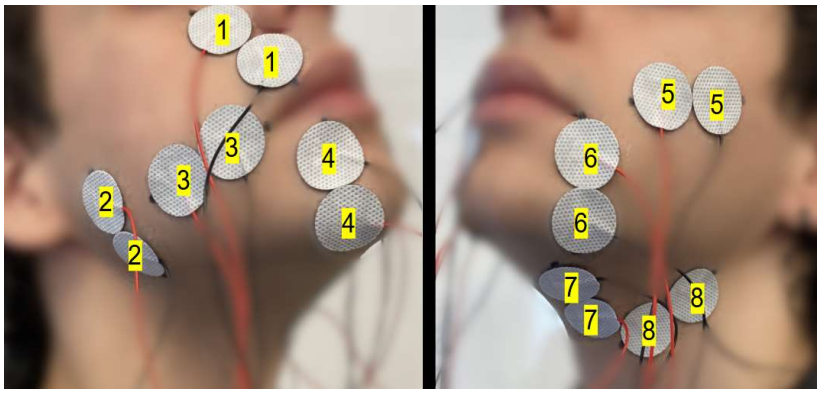

Figure 3 - Electrode setup for the ReSSInt-EMG database, showing the eight bipolar electrode pairs (eight channels), each targeting a different muscle.

To ensure consistency across all speakers, sensors must target the same muscles for all of them. The trained technicians were responsible for securely affixing the electrodes before each recording session. An anticipated challenge in EMG signal recording relates to potential sensor detachment. The sensors must be securely attached to the face and neck during speech or mouthing, involving rapid and continuous movement. This, combined with the possibility of perspiration during recordings, may lead to sensors losing proper contact or detaching entirely. To mitigate this issue, a technician continuously monitored the recorded signals in real-time. If any issues were detected, recordings were paused until sensors were repositioned correctly.

Synchronization between audio, EMG signals and video frames

During recordings, the acquisition of audio, EMG signals and video may contain inaccuracies regarding the precise timing of their start and end. The video corresponding to an utterance begins recording as soon as the preceding utterance has been recorded and the current one is initiated. The speaker is required to press a button when they start speaking and release it when they finish the utterance. Consequently, the recorded audio and EMG signals include some content before the button is pressed and after it is released. This additional duration may not be the same for both the audio and EMG signals.

As previously mentioned, to address the timing discrepancy between the audio and EMG signal, a synchronization signal is shared by the EMG amplifier and the audio interface. When the speaker presses the button to start speaking, the recording interface software sends an instruction to the EMG amplifier to raise a square signal. When the button is released, another instruction is sent to lower the square signal. The synchronization signal is incorporated as the ninth channel alongside the other eight EMG signal channels. The EMG amplifier's analog output generates a signal synchronized with the square signal, which is transmitted to the right channel of the audio interface through a cable. When

the synchronization signal raises, the analog output produces a sudden positive potential, and when it falls, a sudden negative potential is generated. Figure 4 shows an example of the synchronization signals included in the audio and EMG. The audio and the EMG signals are expected to be truncated at the array positions corresponding to the edges of their respective synchronization signals.

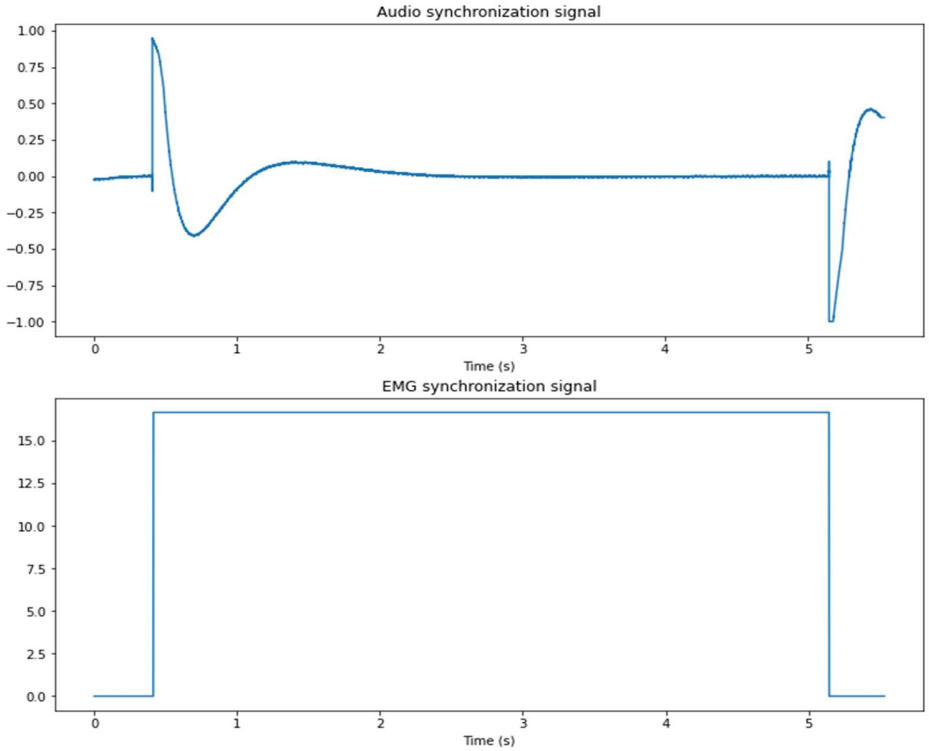

Figure 4 - Synchronization signals included in the audio signal (top) and EMG ninth channel (bottom).

Regarding the videos, there is no need to apply synchronization to them, as it was already applied during the extraction of video frames. This was necessary because certain original videos were excessively lengthy relative to the relevant content, potentially encompassing irrelevant moments in the recording process, such as periods when the speakers are resting. To synchronize the video, the frame counters of both the color and depth video streams are tracked by the recording interface. The frame number recorded when the speaker presses and releases the recording button is registered. After extracting image frames from the video files, frames recorded before the button is pressed and after it is released are deleted. Consequently, the video frames provided with the database are synchronized with the synchronization signals.

Relation between the color and the depth video frames

To utilize the color and depth video frames effectively, it is essential to acknowledge a limitation of the employed camera (Intel RealSense D415): it lacks a built-in synchronization mechanism between the color and depth frames. These frames are sourced from two separate streams within the recording interface. Both streams operate at a frame rate of 30 frames per second. To ensure temporal alignment, frames from each stream are selected based on their respective frame counters, guaranteeing they correspond to the same time interval when the synchronization signal is activated. However, it is crucial to understand that the color and depth frames are not inherently paired or synchronized, given the asynchronous nature of the video streams.

Data Structure

The data is organized in different folders following this structure:
data type > speaker > corpus > session > file(s) per utterance

- The data type indicates the type of information contained in the file and can take the following values:
    - audio: raw audio signals both in .wav and numpy format. Both formats contain stereo audio signals and store speech samples in the left channel and the synchronization signal in the right channel.
    - corpora: text files with the texts pronounced in each utterance of each corpus. The line number matches the utterance number in the file name.
    - emg: for the raw EMG signals that are provided in .adc and numpy format.
    - video: folder to store the video information in png format both for color and depth data.
    - corrected_transcripts: folder that stores text files with the target prompt corrected to match mispronunciations. The corrected transcription should prevail the default transcription for the utterances inside a specific corpus/session that have a file in this folder.
- The speaker is a 3 digit ID identifying the speaker in the recordings from 001 to 009.
- The corpus is a 3 digit ID identifying the corpus that has been recorded from 001 to 016.
- The session is a 3 digit ID identifying the session of the recordings. In this ID the first digit is 1 for the sessions that include audio and 2 or 3 for the sessions that have been recorded when mouthing the sentences.

Inside these folders, the files have been named according to the following naming convention:
data type_speaker_corpus_session_utterance.extension

- utterance: it is a 4 digit correlative ID identifying the sentence recorded in the file. The content of the utterance can be retrieved from the corpora text files, as the utterance number matches the line in which the transcription of the utterance is located in it.
- extension: is a 3 character identifier that relates to the content of the files and indicates the format or type of the file. It can take the following values:
    - wav for audio
    - wav.npy for audio in numpy format
    - adc for raw EMG data
    - adc.npy for EMG data in numpy format
    - lab for the orthographic transcriptions

In the case of the video frames, within the session folder there is a subfolder for each utterance, adhering to the name convention *speaker_corpus_session_utterance*. These folders contain two subfolders: *color* and *depth*, each containing image files alongside their corresponding metadata files. Image frames follow the naming convention *stream type_time stamp.png*, while metadata files adhere to *stream type_metadata_time stamp.txt*. Notably, both the PNG files and their associated metadata files share identical timestamps in their names. It is important to note that the color and depth frames originate from distinct asynchronous streams, rendering them independent and unable to be paired based on timestamp or any other parameter

Content and Annotations

Besides the EMG, video and audio data, the database includes the following metadata:

- Short description: stating the purpose and general features of the database
- Time and place: describing the span of the recordings and the place where they were performed
- Responsible of the recordings: names of the persons in charge of performing the recordings
- Folder structure: describing the organization of the data
- Data types: listing the different data associated with a recording
    - audio: raw audio signals (in .wav and .npy format)
    - emg: raw emg signals (in .adc and .npy format)
    - video: images corresponding to video frames both for color and depth (in .png format) along with the respective metadata (in .txt format)
    - transcripts: text files with the target prompts (in .txt format)
    - corrected_transcripts: text files with the target prompt corrected to match mispronunciations, only available for the cases where there was an error in the reading of the prompt
- Speakers: ID corresponding to each participant

- Corpus: ID corresponding to each of the text corpus recorded in each session. The database includes three types of textual material: vowel-consonant-vowel combinations, a list of 100 most useful words and isolated sentences. Vowel-consonant-vowel combinations were created pairing each of the 22 Spanish consonants once with each of the five vowels in Spanish, resulting in 110 combinations. Context was added to each combination, in the format at[V[[C][V]ta, to control for co-articulation. For the compilation of the list of the 100 most useful Spanish words, we chose several words from each category sourced from a website dedicated to enhancing daily communication for individuals with limited means of expression[3]. We carefully examined the phoneme balance of the selected list to ensure the representation of all phonemes in the Spanish language. Finally the isolated sentences were taken from the Sharvard [12] and AhoSyn [13] corpora. Table 2Table 1 shows a detailed list of the corpora recorded in each session of the database.

Corpus ID	Corpus name	Session ID														
		X01	X02	X03	X04	X05	X06	X07	X08	X09	X10	X11	X12	X13	X14	X15
001	110 VCV combinations	a	a	a	a	a+s	s			s						
002	100 isolated words	a	a	a	a	a+s	s	a+s	a+s	s						
003	Sharvard sentences 1-100	a	a	a	a	a+s	s	a+s	a+s	s						
004	Sharvard sentences 101-400	a					s									
005	Sharvard sentences 401-700		a							s						
006	Ahosyn sentences 1-150			a												
007	Ahosyn sentences 151-300				a											
008	Ahosyn sentences 301-400							a								
009	Ahosyn sentences 401-500								a							
010	Ahosyn sentences 501-505										a+2s	a+2s	a+2s	a+2s	a+2s	a+2s
011	Ahosyn sentences 506-570										a+2s					
012	Ahosyn sentences 571-635											a+2s				
013	Ahosyn sentences 636-700												a+2s			
014	Ahosyn sentences 701-765													a+2s		
015	Ahosyn sentences 766-830														a+2s	
016	Ahosyn sentences 896-960															a+2s

Table 1 – Corpora recorded in each session. Codes used in this table are explained in Table 2

- Session: ID identifying the number of session. Sessions starting by 1 contain audible recordings, while sessions starting by 2 or 3 contain silent speech. Code 3 is used when the same content is recorded twice in silent mode during a session. Table 1 presents the correspondence between the recording sessions and the corpora recorded during each session. If two session IDs share the same last two digits, it means that they form one recording session, in which some corpora were recorded audibly and some silently. In one session, we designate utterances recorded in

[3] https://arasaac.org/pictograms/search

both modalities as 'parallel utterances' and those recorded in only one modality as 'non-parallel utterances'.

Code	Description
a	audible (X=1 in session code)
s	silent (X=2 in session code)
a+s	audible (X=1 in session code) and silent (X=2 in session code)
a+2s	audible (X=1 in session code) and 2 repeated silent (X=2 and X=3 in session code)

Table 2 – Codes used to describe session contents

The final database comprises recordings from 9 speakers, including 6 standard speakers (ID 001 to 006) and 3 laryngectomized individuals (ID 007 to 009). The number of sessions recorded by speaker vary from one session (for speaker 008) to 15 sessions (for speaker 001). Table 3 presents speaker details and the corresponding number of recorded sessions in the database.

Speaker ID	Gender	Age	Number of recorded sessions
001	Male	29	15
002	Female	29	8
003	Male	51	4
004	Female	46	4
005	Male	45	8
006	Female	61	4
007	Female	61	2
008	Male	77	1
009	Male	64	2

Table 3 – Information about speakers in the database

The database contains a total of 22.5 hours of recordings, that distribute among the different content type and parallel/non parallel categories as specified in Table 4.

	Non-parallel duration	Parallel audible duration	Parallel silent duration	Total duration
VCV	1:22:01	0:19:02	0:20:02	2:01:06
Words	1:11:16	0:30:12	0:32:39	2:14:05
Session-common sentences	3:07:53	1:19:49	1:29:31	5:57:12
Session-specific sentences	10:25:24	0:36:08	1:18:52	12:20:23
Total duration	16:06:36	2:45:09	3:41:04	22:32:48

Table 4 - Total duration of the database per recorded content type, expressed in the format of hh:mm:ss.

Table 5 provides a comprehensive breakdown of the content recorded by each speaker, accompanied by partial summaries per recorded content type and speaker.

Speaker ID	Session	Corpus	Non-parallel duration	Parallel audible duration	Parallel silent duration	Total duration
001	001	001	2:36	--:--	--:--	19:25
		002	--:--	--:--	--:--	
		003	4:27	--:--	--:--	
		004	12:21	--:--	--:--	
	002	001	2:24	--:--	--:--	21:50
		002	1:57	--:--	--:--	
		003	4:31	--:--	--:--	
		005	12:58	--:--	--:--	
	003	001	2:15	--:--	--:--	21:03
		002	1:50	--:--	--:--	
		003	4:42	--:--	--:--	
		006	12:15	--:--	--:--	

Speaker ID	Session	Corpus	Non-parallel duration	Parallel audible duration	Parallel silent duration	Total duration
	004	001	2:56	--:--	--:--	24:42
		002	2:27	--:--	--:--	
		003	5:21	--:--	--:--	
		007	13:59	--:--	--:--	
	005	001	--:--	2:54	2:57	21:02
		002	--:--	2:10	2:11	
		003	--:--	5:19	5:30	
	006 (silent)	001	3:09	--:--	--:--	28:56
		002	2:12	--:--	--:--	
		003	5:37	--:--	--:--	
		004	17:58	--:--	--:--	
	007	002	--:--	2:03	2:02	25:00
		003	--:--	5:08	5:45	
		008	10:03	--:--	--:--	
	008	002	--:--	2:27	2:16	25:43
		003	--:--	5:20	6:03	
		009	9:37	--:--	--:--	
	010	010	--:--	0:31	1:12	22:53
		011	--:--	6:31	14:39	
	011	010	--:--	0:30	1:07	22:52
		012	--:--	6:33	14:42	
	012	010	--:--	0:27	0:58	20:21
		013	--:--	6:02	12:53	
	013	010	--:--	0:29	1:00	20:31
		014	--:--	6:01	13:01	

Speaker ID	Session	Corpus	Non-parallel duration	Parallel audible duration	Parallel silent duration	Total duration
	014	010	--:--	0:26	0:58	18:59
		015	--:--	5:36	11:59	
	015	010	--:--	0:26	0:56	18:25
		016	--:--	5:25	11:38	
	Total	VCV	13:20	2:54	2:57	19:11
		Words	8:26	6:41	6:29	21:36
		Session-common sentences	24:38	18:36	23:29	1:06:42
		Session-specific sentences	1:29:10	36:08	1:18:52	3:24:09
		Total duration	2:15:34	1:04:19	1:51:47	5:11:39
002	001	001	2:19	--:--	--:--	29:09
		002	1:29	--:--	--:--	
		003	5:50	--:--	--:--	
		004	19:31	--:--	--:--	
	002	001	--:--	--:--	--:--	33:24
		002	2:55	--:--	--:--	
		003	7:22	--:--	--:--	
		005	23:08	--:--	--:--	
	003	001	3:22	--:--	--:--	28:28
		002	2:34	--:--	--:--	
		003	6:39	--:--	--:--	
		006	15:53	--:--	--:--	

Speaker ID	Session	Corpus	Non-parallel duration	Parallel audible duration	Parallel silent duration	Total duration
	004	001	4:39	--:--	--:--	37:00
		002	3:41	--:--	--:--	
		003	8:13	--:--	--:--	
		007	20:27	--:--	--:--	
	005	001	--:--	4:29	4:32	31:52
		002	--:--	3:00	3:07	
		003	--:--	7:28	9:16	
	006 (silent)	001	4:22	--:--	--:--	42:50
		002	3:11	--:--	--:--	
		003	7:57	--:--	--:--	
		004	27:20	--:--	--:--	
	007	002	--:--	3:53	3:36	40:04
		003	--:--	8:54	8:54	
		008	14:47	--:--	--:--	
	008	002	--:--	3:42	3:04	36:51
		003	--:--	8:33	8:10	
		009	13:22	--:--	--:--	
	Total	VCV	14:43	4:29	4:32	23:45
		Words	13:51	10:35	9:48	34:13
		Session-common sentences	36:00	24:54	26:20	1:27:14
		Session-specific sentences	2:14:27	--:--	--:--	2:14:27

Speaker ID	Session	Corpus	Non-parallel duration	Parallel audible duration	Parallel silent duration	Total duration
		Total duration	3:19:01	39:58	40:40	4:39:39
003	001	001	4:28	--:--	--:--	32:05
		002	3:03	--:--	--:--	
		003	5:55	--:--	--:--	
		004	18:39	--:--	--:--	
	002	001	2:55	--:--	--:--	25:01
		002	2:51	--:--	--:--	
		003	4:36	--:--	--:--	
		005	14:39	--:--	--:--	
	005	001	--:--	3:01	3:12	22:05
		002	--:--	2:18	2:50	
		003	--:--	5:08	5:37	
	006 (silent)	001	2:52	--:--	--:--	27:44
		002	2:27	--:--	--:--	
		003	5:18	--:--	--:--	
		004	17:08	--:--	--:--	
	Total	VCV	10:15	3:01	3:12	16:28
		Words	8:21	2:18	2:50	13:29
		Session-common sentences	15:48	5:08	5:37	26:33
		Session-specific sentences	50:25	--:--	--:--	50:25

Speaker ID	Session	Corpus	Non-parallel duration	Parallel audible duration	Parallel silent duration	Total duration
		Total duration	1:24:49	10:26	11:39	1:46:54
004	001	001	--:--	--:--	--:--	28:23
		002	2:23	--:--	--:--	
		003	6:16	--:--	--:--	
		004	19:44	--:--	--:--	
	002	001	2:44	--:--	--:--	29:28
		002	2:39	--:--	--:--	
		003	5:43	--:--	--:--	
		005	18:22	--:--	--:--	
	005	001	--:--	2:42	2:32	21:18
		002	--:--	2:22	2:13	
		003	--:--	5:28	6:00	
	006 (silent)	001	3:08	--:--	--:--	31:55
		002	2:18	--:--	--:--	
		003	5:58	--:--	--:--	
		004	20:31	--:--	--:--	
	Total	VCV	5:52	2:42	2:32	11:06
		Words	7:20	2:22	2:13	11:55
		Session-common sentences	17:56	5:28	6:00	29:25
		Session-specific sentences	58:38	--:--	--:--	58:38

Speaker ID	Session	Corpus	Non-parallel duration	Parallel audible duration	Parallel silent duration	Total duration
		Total duration	1:29:47	10:32	10:46	1:51:05
005	001	001	3:01	--:--	--:--	29:03
		002	2:33	--:--	--:--	
		003	5:51	--:--	--:--	
		004	17:39	--:--	--:--	
	002	001	3:31	--:--	--:--	28:13
		002	2:14	--:--	--:--	
		003	5:38	--:--	--:--	
		005	16:50	--:--	--:--	
	003	001	3:41	--:--	--:--	23:27
		002	2:17	--:--	--:--	
		003	5:14	--:--	--:--	
		006	12:15	--:--	--:--	
	004	001	2:57	--:--	--:--	25:19
		002	2:46	--:--	--:--	
		003	6:10	--:--	--:--	
		007	13:27	--:--	--:--	
	005	001	--:--	2:52	3:13	20:11
		002	--:--	--:--	2:32	
		003	--:--	5:26	6:09	
	006 (silent)	001	2:51	--:--	--:--	31:49
		002	2:20	--:--	--:--	
		003	6:25	--:--	--:--	
		004	20:13	--:--	--:--	

Speaker ID	Session	Corpus	Non-parallel duration	Parallel audible duration	Parallel silent duration	Total duration
	007	002	--:--	3:13	3:27	28:43
		003	--:--	6:05	7:06	
		008	8:52	--:--	--:--	
	008	002	--:--	2:50	2:52	27:02
		003	--:--	5:43	6:34	
		009	9:03	--:--	--:--	
	Total	VCV	16:00	2:52	3:13	22:05
		Words	12:10	6:03	8:51	27:03
		Session-common sentences	29:17	17:14	19:49	1:06:19
		Session-specific sentences	1:38:19	--:--	--:--	1:38:19
		Total duration	2:35:46	26:08	31:52	3:33:46
006	001	001	2:30	--:--	--:--	31:16
		002	2:03	--:--	--:--	
		003	6:19	--:--	--:--	
		004	20:23	--:--	--:--	
	002	001	2:45	--:--	--:--	33:54
		002	2:13	--:--	--:--	
		003	7:14	--:--	--:--	
		005	21:42	--:--	--:--	
	005	001	--:--	3:04	3:36	28:06
		002	--:--	2:13	2:28	

Speaker ID	Session	Corpus	Non-parallel duration	Parallel audible duration	Parallel silent duration	Total duration
		003	--:--	8:29	8:16	
	006 (silent)	001	3:42	--:--	--:--	34:31
		002	3:08	--:--	--:--	
		003	7:08	--:--	--:--	
		004	20:34	--:--	--:--	
	Total	VCV	8:57	3:04	3:36	15:37
		Words	7:24	2:13	2:28	12:05
		Session-common sentences	20:41	8:29	8:16	37:26
		Session-specific sentences	1:02:39	--:--	--:--	1:02:39
		Total duration	1:39:42	13:46	14:20	2:07:48
007	006 (silent)	001	--:--	--:--	--:--	46:44
		002	4:14	--:--	--:--	
		003	12:07	--:--	--:--	
		004	30:23	--:--	--:--	
	009 (silent)	001	2:48	--:--	--:--	48:48
		002	2:36	--:--	--:--	
		003	10:11	--:--	--:--	
		005	33:12	--:--	--:--	
	Total	VCV	2:48	--:--	--:--	2:48
		Words	6:50	--:--	--:--	6:50

Speaker ID	Session	Corpus	Non-parallel duration	Parallel audible duration	Parallel silent duration	Total duration
		Session-common sentences	22:18	--:--	--:--	22:18
		Session-specific sentences	1:03:35	--:--	--:--	1:03:35
		Total duration	1:35:31	--:--	--:--	1:35:31
008	006 (silent)	001	2:44	--:--	--:--	32:47
		002	2:08	--:--	--:--	
		003	6:29	--:--	--:--	
		004	21:26	--:--	--:--	
	Total	VCV	2:44	--:--	--:--	2:44
		Words	2:08	--:--	--:--	2:08
		Session-common sentences	6:29	--:--	--:--	6:29
		Session-specific sentences	21:26	--:--	--:--	21:26
		Total duration	32:47	--:--	--:--	32:47
009	006 (silent)	001	3:40	--:--	--:--	36:48
		002	2:47	--:--	--:--	
		003	7:20	--:--	--:--	
		004	23:01	--:--	--:--	
		001	3:42	--:--	--:--	36:51

Speaker ID	Session	Corpus	Non-parallel duration	Parallel audible duration	Parallel silent duration	Total duration
	009 (silent)	002	2:00	--:--	--:--	
		003	7:26	--:--	--:--	
		005	23:43	--:--	--:--	
	Total	VCV	7:22	--:--	--:--	7:22
		Words	4:46	--:--	--:--	4:46
		Session-common sentences	14:46	--:--	--:--	14:46
		Session-specific sentences	46:45	--:--	--:--	46:45
		Total duration	1:13:39	--:--	--:--	1:13:39

Table 5 – Detailed speaker and session information for the ReSSInt-EMG database. The duration is expressed in the format of hh:mm:ss.

Challenges and Limitations

While meticulous efforts were invested to ensure the overall quality and reliability of all EMG signals within the database, it is important to acknowledge the inherent complexities of data acquisition. Despite exercising utmost care to prevent the inclusion of signals associated with detached electrodes, it is recognized that, in isolated instances, the database may inadvertently contain signals that deviate from the intended standard. This acknowledgment underscores our commitment to transparency and continual improvement, encouraging a collaborative approach to refine and enhance the database over time.

Another issue with the database derives from the fact that the sensors' polarity was not verified during the electrode attachment process at the beginning of each session, resulting in a lack of uniformity in the recorded EMG signals' polarity within the database. Despite these challenges, we remain committed to addressing any potential inconsistencies and continually refining the dataset to meet the highest standards of accuracy and reliability.

Conclusion

The ReSSint database is a novel initiative, representing a significant advancement in the field. Notably, it marks the first database to offer comprehensive, simultaneous data encompassing surface EMG, speech, and video recordings in the Spanish language. The inclusion of data from laryngectomees within the ReSSint database further enhances its significance and impact. By incorporating recordings from laryngectomized speakers, this database addresses a crucial gap in research by providing valuable insights into speech production and muscle activity in this specific population. This innovative resource not only expands the scope of available datasets but also opens avenues for diverse research applications, especially in the domain of laryngectomy rehabilitation and speech therapy. The integration of multiple modalities within the ReSSint database broadens the landscape of research possibilities, allowing multimodal studies and facilitating a deeper understanding of the complex interactions between speech and muscle activity.

References

[1] Hernáez Rioja, I., Gonzalez-Lopez, J.A., Navas, E., Córdoba, J.L.P., Saratxaga, I., Olivares, G., Sanchez, J., Galdón, A. and Romillo, V.G., 2022, ReSSInt project: Voice Restoration using Silent Speech Interfaces. Proc. IberSPEECH 2022, pp. 226-230

[2] Hernáez Rioja, I., González López, J.A., Navas, E., Pérez-Córdoba, J.L., Saratxaga, I., Olivares, G., Sanchez, J., Galdón, A., García Romillo, V., Gónzalez Atienza, M. and Schultz, T., 2021, January. Voice restoration with silent speech interfaces (ReSSInt), Proc. IberSPEECH 2021, pp. 130-134

[3] De Luca, C.J., 2002. Surface electromyography: Detection and recording. Technical Report, DelSys Incorporated, 10(2), pp.1-10.

[4] Wand, M.; Janke, M.; Schultz, T. The EMG-UKA corpus for electromyographic speech processing. In Proceedings of the Interspeech, 2014, pp. 1593–1597.

[5] Gaddy, D.; Klein, D. Digital voicing of silent speech. arXiv preprint arXiv:2010.02960 2020.

[6] Diener, L.; Roustay Vishkasougheh, M.; Schultz, T. CSL-EMG_Array: An Open Access Corpus for EMG-to-Speech Conversion. In Proceedings of the INTERSPEECH, 2020.

[7] Safie, S.I.; Yusof, M.I.; Rahim, R.; Taib, A. EMG database for silent speech Ruqyah recitation. In Proceedings of the 2016 IEEE EMBS Conference on Biomedical Engineering and Sciences (IECBES), 2016, pp. 712–715.

[8] Freitas, J.; Teixeira, A.; Dias, J. Multimodal corpora for silent speech interaction. Multimodal corpora for silent speech interaction 2014, pp. 4507–4511.

[9] Lopez-Larraz, E.; Mozos, O.M.; Antelis, J.M.; Minguez, J. Syllable-based speech recognition using EMG. In Proceedings of the 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology, 2010, pp. 4699–4702.

[10] Lee, K.S. EMG-based speech recognition using hidden Markov models with global control variables. IEEE Transactions on biomedical engineering 2008, 55, 930–940.

[11] Ma, S.; Jin, D.; Zhang, M.; Zhang, B.; Wang, Y.; Li, G.; Yang, M. Silent Speech Recognition Based on Surface Electromyography. In Proceedings of the 2019 Chinese Automation Congress (CAC); IEEE: Hangzhou, China, 2019; pp. 4497–4501.

[12] Aubanel, V.; Lecumberri, M.L.G.; Cooke, M. The Sharvard Corpus: A phonemically-balanced Spanish sentence resource for audiology. International Journal of Audiology 2014, 53, pp. 633–638.

[13] Sainz, I.; Erro, D.; Navas, E.; Hernáez, I.; Sanchez, J.; Saratxaga, I.; Odriozola, I. Versatile Speech Databases for High Quality 472 Synthesis for Basque. In Proceedings of the Proceedings of the Eighth International Conference on Language Resources and 473 Evaluation (LREC'12), 2012.

[14] Chan, A.D.C.; Englehart, K.; Hudgins, B.; Lovely, D.F. Myo-Electric Signals to Augment Speech Recognition. Medical & Biological Engineering & Computing 2001, 39, pp. 500–504.

[15] Maier-Hein, L.; Metze, F.; Schultz, T.; Waibel, A. Session independent non-audible speech recognition using surface electromyography. In Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding, 2005.; IEEE: San Juan, Puerto Rico, 2005; pp. 331–336.

[16] Jou, S.C.; Schultz, T.; Walliczek, M.; Kraft, F.; Waibel, A. Towards Continuous Speech Recognition Using Surface Electromyography 2006. p. 4.

[17] Schultz, T.; Wand, M. Modeling coarticulation in EMG-based continuous speech recognition. Speech Communication 2010, 52, pp. 341–353.

[18] Diener, L.; Janke, M.; Schultz, T. Direct conversion from facial myoelectric signals to speech using Deep Neural Networks. In Proceedings of the 2015 International Joint Conference on Neural Networks (IJCNN); IEEE: Killarney, Ireland, 2015; pp. 1–7.

[19] Meltzner, G.S.; Heaton, J.T.; Deng, Y.; De Luca, G.; Roy, S.H.; Kline, J.C. Silent Speech Recognition as an Alternative Communication Device for Persons with Laryngectomy. IEEE/ACM Transactions on Audio, Speech, and Language Processing 2017, 25, pp. 2386–2398.

[20] Zhu, M.; Zhang, H.; Wang, X.; Wang, X.; Yang, Z.; Wang, C.; Samuel, O.W.; Chen, S.; Li, G. Towards Optimizing Electrode Configurations for Silent Speech Recognition Based on High-Density Surface Electromyography. Journal of Neural Engineering 2021, 18, 016005.

Contact information

Database authors: AhoLab Signal Processing Laboratory
https://aholab.ehu.eus/
contact email: aholab@aholab.ehu.eus

Funding project: Voice Restoration with Silent EMG Speech Interfaces (ReSSint)
Grant number PID2019-108040RB-C21/AEI/10.13039/501100011033
https://aholab.ehu.eus/ressint/