IberSPEECH 2022
14-16 November 2022, Granada, Spain

Phone classification using electromyographic signals

Eder del Blanco, Inge Salomons, Eva Navas, Inma Hernáez

HiTZ Basque Center for Language Technologies - Aholab,
University of the Basque Country UPV/EHU, Bilbao, Spain
eder.delblanco, inge.salomons, eva.navas, inma.hernaez@ehu.eus

Abstract

Silent speech interfaces aim at generating speech from biosignals obtained from the human speech production system. In order to provide resources for the development of these interfaces, language-specific databases are required. Several silent speech electromyography (EMG) databases for English exist. However, a database for the Spanish language had yet to be developed. The aim of this research is to validate the experimental design of the first silent speech EMG database for Spanish, namely the new ReSSInt-EMG database. The EMG signals in this database are obtained using eight surface EMG bipolar electrode pairs located in the face and neck and are recorded in parallel with either audible or silent speech. Phone classification experiments are performed, using a set of time-domain features typically used in related works. As a validation reference, the EMG-UKA Trial Corpus is used, which is the most commonly used silent speech EMG database for English. The results show an average test accuracy of 40.85% for ReSSInt-EMG, suggesting that the data acquisition procedure for the new database is valid.

Index Terms: silent speech interfaces, human-computer interaction, EMG signals, speech processing, phone classification

1. Introduction

Silent Speech Interfaces (SSI) are devices aimed at generating a speech signal using some of the biosignals generated by the human organs while speaking silently [1]. Among several existing proposals, electromyographic (EMG) signals obtained from the muscles of the face have been considered appropriate for the task [2, 3]. As we talk (silently or not), a number of muscles of the face and neck move in a coordinated way. The movements are induced by small electrical voltages which can be captured by the appropriate sensors carefully located on the skin.

A clear application area for SSI is as a communication aid that provides a voice to people with speech disabilities [4]. These interfaces can also be used in situations where private communications are needed [5, 6] or in noisy environments where audible speech would be masked by the environmental noise [7].

EMG signals obtained from the muscles of the face and neck have been used to perform phone classification [8, 9], syllable identification [10], word recognition [11, 12] and even continuous speech recognition [13]. Attempts have also been made to apply them in speaker recognition tasks [14, 15]. In addition, efforts for direct speech generation from this kind of signals have also been made [16, 17, 18].

Most of the works and databases for SSI have been developed for English (see [19, 20, 21] among others) although there are also databases for other languages [22, 23, 10, 24, 25]. The authors of this work are involved in a project with the aim of developing an EMG based SSI infrastructure to allow the development of SSI for Spanish, as described in [26]. In the current paper we describe the first efforts made in the development of such an infrastructure, namely the first phone recognition results that validate the database design decisions mainly related to the acquisition equipment and electrode setup as well as the recording corpus. The developed classification experiments have been performed also using data from an existing database (in English), as a way to validate the results. For that, the freely available trial subset of the EMG-UKA parallel EMG-Speech corpus [19] has been chosen. As described in section 6, the obtained classification accuracy is promising and allows us to continue confidently with the database recordings and to tackle in the near future more complex problems such as direct EMG-to-Speech conversion.

This paper is organized as follows. Section 2 describes the recording setup. Section 3 and Section 4 describe the data used in the classification experiments, which are detailed in section 5. The obtained results are presented and discussed in section 6. Finally, we conclude the paper with a summary of the accomplished achievements together with present and future works.

2. Recording Setup

2.1. Recording procedure

The recordings are made in a sound-proof room, using a silent computer to reduce the interference of the EMG-signals. The EMG-signals are captured by a Quattrocento bio-electrical amplifier at a sampling frequency of 2 048 Hz and the speech signals are captured by a Neumann TLM103 (diaphragm) microphone with a sampling frequency of 16 kHz. An audio interface secures a synchronization signal between the EMG and the audio signals. Additionally, video is captured in order to have supplementary data and allow multi-modal experiments in the future, but this modality has not been used in the experiments described in this paper. For the acquisition and synchronization of the audio and EMG signals, we are using the recording system from Cognitive Systems Lab (University of Bremen)[1].

In order to reduce inter-session variability in audio and video, the positions of the microphone and video camera are kept constant for all sessions. Additionally, a 3D mask with the positions of the electrodes is made during the first session of each speaker, and used in later sessions to also keep constant the electrode locations. Also, speakers were instructed to speak with a bit exaggerated articulation.

2.2. Electrode Setup

The final electrodes locations used were defined after performing a set of pilot experiments, consisting of a simple phone classification task for 100 consonant-vowel (CV) combinations. As a result, the setup differs partially from that used in the EMG-

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

10.21437/IberSPEECH.2022-7

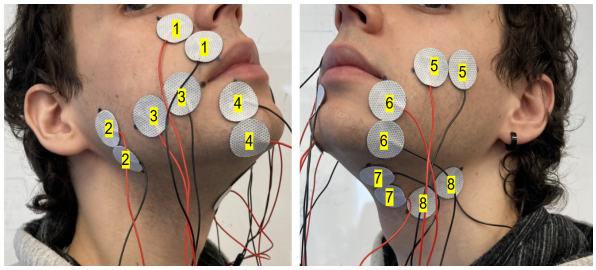

Figure 1: *Electrode locations for the ReSSInt-EMG database*

UKA corpus. The pilot experiments consisted of a series of different setups varying in the total number of electrodes and electrode locations, based on various previous approaches in related research. In summary, all facial and neck muscles were targeted, and the resulting setup contains those channels that were contributing most in the phone classification task.

The muscles targeted in the EMG-UKA corpus, using six channels, are (see [19]): Levator anguli oris (channels 2 and 3), Zygomaticus major (channels 2 and 3), Platysma (channel 4), Anterior belly of the digastric (channel 1), Tongue (channels 1 and 6) and Depressor anguli oris (channel 5). It must be noted that channel 5 is not used in most of the phone recognition experiments performed with this database [27, 13, 28, 29]. However, it has been used in [30] for word recognition and also for speaker identification [14] and speech to EMG conversion [31, 32]. Therefore, this channel has been included in our own phone recognition experiments.

For the ReSSInt-EMG database, a partially different setup is used (see figure 1). First of all, there are eight channels, each targeting a different muscle. The following three targeted muscles are similar: Anterior belly of the digastric (channel 7), Zygomaticus major (channel 5) and Depressor anguli oris (channel 6). Then there are two that are close or underneath the muscle targeted in the EMG-UKA setup: Levator labii superioris (channel 1), Depressor labii inferioris (channel 4). The remaining three are additional: Risorius (channel 3), Masseter (channel 2) and Stylohoid (channel 8).

For the EMG-UKA corpus, both monopolar and bipolar channels were used, whereas the ReSSInt-EMG database only contains signals from bipolar channels. Furthermore, the electrodes are of a different kind.

3. Materials

In this section we describe the two databases used in the experiments. The first database is the EMG-UKA Trial Corpus for English [19], which has been used as reference. The second one is the new database we are currently developing for Spanish, the ReSSInt-EMG database.

3.1. The EMG-UKA Trial Corpus

The EMG-UKA Trial Corpus is a subset of the full EMG-UKA Corpus [33] and includes four speakers and 13 recording sessions (see Table 1 for details). There are three types of utterances: audible, whispered and silent. For the experiments described in this paper, only the EMG recordings with parallel audible sentences were used. Each session includes 50 audible sentences, with the exception of session 101 from speaker 002, which contains 520 audible sentences. The sessions with 50 sentences are divided into a training set of 40 sentences and a test set of 10 sentences. Session 101 from speaker 002 is divided into 500 sentences for training and 20 sentences for testing.

Table 1: *Session information for EMG-UKA and ReSSInt-EMG databases*

Database	Speaker	Session	Gender	Duration
EMG-UKA	002	001	M	3:29
	002	003	M	3:24
	002	101	M	26:04
	004	001	F	3:23
	006	001	M	3:45
	008	001	M	3:19
	008	002	M	3:06
	008	003	M	3:02
	008	004	M	2:50
	008	005	M	2:41
	008	006	M	2:40
	008	007	M	2:38
	008	008	M	2:42
ReSSInt-EMG	001	101	M	16:51
	001	102	M	17:31
	001	103	M	17:00
	001	104	M	19:24
	002	101	F	25:25
	002	102	F	26:50
	002	103	F	25:55
	002	104	F	27:06
	003	101	M	24:38
	004	101	F	26:05

3.1.1. Database labeling

The phonetic transcription accompanying the EMG-UKA Trial Corpus has been found partially incorrect, thus we have obtained new phonetic transcriptions using the Librispeech lexicon together with some additional transcriptions for non-existing words. Also, a new phonetic segmentation has been obtained using Montreal Forced Aligner [34]. A set of 40 phones is considered in our experiments. Initial and final silences have been removed, but short pauses inside the sentences have been included in the phone classification experiments.

3.2. The ReSSInt-EMG database

The goal of the project is to obtain recordings from eight to ten speakers (including two laryngectomees) with four to eight sessions each. As we are still in the process of data acquisition, the data considered for the experiments described in this paper consist of the first sessions recorded by four speakers, with a total of ten sessions. The lower part of Table 1 shows the details of the already recorded sessions. The duration corresponds to the sum of the audio signals per session, after synchronization with the EMG signals. In this section, we describe the contents of the corpus and the details of the recorded sessions.

3.2.1. The corpus

In each recording session, three different kinds of items are recorded, namely: non-sense words including VCV structures, isolated words, and sentences. The sentences are taken from the Sharvard Corpus [35] and from a text corpus called Ahosyn that was developed to record TTS databases [36] (see Table 2). Both sets are phonetically balanced.

Table 2: *Corpus information for ReSSInt-EMG sessions*

Session	Corpus
all	110 VCV combinations
all	100 isolated words
all	Sharvard sentences 1-100
101	Sharvard sentences 101-400
102	Sharvard sentences 401-700
103	Ahosyn sentences 1-150
104	Ahosyn sentences 151-300

As can be seen in Table 2, the first 100 sentences of the Sharvard corpus are repeated during each session. For the experiments described in this paper, the last 20 of these sentences have been assigned to the test set, in order to have a similar test set size as in the longer session of EMG-UKA database (002-101). The training set consists of the remaining sentences recorded in that session, either 230 or 380, depending on the session number.

One session from one speaker (002-102) was left out of the experiments due to issues with the recordings.

3.2.2. Database labeling

Each speech utterance was aligned with the corresponding phonetic transcription using Montreal Forced Aligner. The phonetic dictionary was created using the Aholab transcriber, which uses the SAMPA phone set (30 different phones). As with EMG-UKA, initial and final silences have been removed but short pauses inside the sentences have been taken into account in the classification experiments.

4. Feature extraction

The extraction of the EMG Time Domain (TD) features has been done following the procedure proposed in [37]. Similar parameters with small variations have also been used in [9, 38]. The frame-based features used are the following: the time-domain mean of the nine-point double-averaged signal, the power of the nine-point double-averaged signal, the time-domain mean of the rectified high-frequency signal, the power of the rectified high-frequency signal, and the zero-crossing rate.

A window of 25 ms duration and 5 ms frame shift has been used to extract the EMG features. A total of 5 x N features are calculated for each frame, where N is the number of channels (N=6 for EMG-UKA and N=8 for ReSSInt-EMG). In order to add context information to the frames, the features of surrounding frames and the features of the current frame are stacked together. With this purpose, a stacking filter is used. The width of the stacking filter indicates the number of adjacent frames before and after the actual frame. Since 5 TD features are being calculated for each of the N EMG channels, the length of the parameter vector assigned to each frame can be calculated as $N \cdot 5 \cdot (2k+1)$, where $k = 15$ is the width of the stacking filter and N is the number of channels.

To reduce the dimension of the parameter vector, Linear Discriminant Analysis (LDA) [39] has been applied, as in [38] and [9]. For EMG-UKA, a reduction from 930 to 32 features is performed, while for ReSSInt-EMG the features are reduced from 1240 to 28. The difference in the number of phone classes for each database partially explains the difference in the number of features, since the maximum allowed number of features in LDA reduction is the number of classes minus 1. We decided to adopt the number 32 from previous work with the EMG-UKA database.

For the classification task using acoustic features, MFCCs have been extracted from the audio signals using a Hamming window, calculating 13 coefficients for each frame. To obtain the coefficients, a 30 filters *filterbank* has been used.

5. Experiments

The experiment consists in performing a phone classification task from EMG signals. Phone classification experiments have been previously performed with the EMG-UKA database ([9] and [38]) and the results can be used as an indicator of the database performance.

We used a bagging classifier [40], with 100 decision trees as estimators and a required minimum of 50 samples in the leaf node. We tried and compared other classifiers, such as neural networks, GMMs, and bagging classifiers with different kinds of estimators, but the chosen algorithm yielded the highest validation accuracy. The classification was done in speaker- and session-dependent modality. Cross-validation is performed using the K-Fold method, dividing the utterances in the training subset into five groups. Five classifiers are trained, using four different folds each time and testing them with the unseen fold. Then the obtained results are averaged. Finally, a new classifier is trained using all the training data, which is tested with the test subset to obtain the test accuracy.

For the short sessions in EMG-UKA, the 40 sentences of the training subset are used for the cross-validation experiments while for the larger session 500 sentences are used. On the other hand, the ReSSInt-EMG sessions have a maximum of 400 sentences. From each session, 20 utterances are left out of training, to have the same number of test utterances as in the larger EMG-UKA session. Then, 5-fold cross-validation is applied to the 380 (Sharvard) / 230 (Ahosyn) remaining sentences.

6. Results and discussion

Tables 3 and 4 show the obtained results of the phone classification performed on the EMG-UKA Trial Corpus and the ReSSInt-EMG database, respectively. Since the classification is speaker-dependent and session-dependent, the results are shown for each individual session. The tables show:

- The test accuracy based on the MFCCs (acoustic signals). They are provided as practical reference for the general performance.
- The validation accuracy based on the EMG signals. Since the K-fold validation method is used, the mean and standard deviation of the accuracy results obtained with each of the 5 folds is shown.
- The test accuracy based on the EMG signals.

All the shown accuracy values are frame-based. The results for the long session 101 from speaker 002 have been underlined in Table 3. Because of the different duration of this session, the averages shown in the last line of Table 3 do not include that long session.

Although not comparable because of the several differences in the experimental framework (the new labeling and different classifiers), the results we have obtained in the EMG-UKA Trial Corpus are in the range of those presented in the literature

[9][38]. It is worth to mention that in the EMG-UKA Trial Corpus, the validation accuracy obtained for the shorter sessions is on average similar to that of the long session in spite of the differences in the amounts of training material. In any case, as expected, the highest test accuracy has been obtained for the longer session 101 from speaker 002.

The results for the ReSSInt-EMG sessions show some variability among speakers as well as among different sessions from the same speaker, in spite of the effort done to avoid session variability within speakers, by using the mask. However, this variability is also observed in the results for the acoustic data, so it cannot be directly related to a variation in the electrodes locations.

A side-by-side comparison between the results of the experiments performed on the two databases cannot be made, due to the already mentioned differences in the experimental setup. To start with, the languages are different, resulting in a difference in the number and type of phone classes (40 phones for EMG-UKA vs. only 30 for ReSSInt-EMG). Furthermore, the recording conditions and materials (number, type and locations of the electrodes) used were not the same.

However, the goal of this research was to validate the acquisition procedure of the new database. For this, we matched the characteristics of the ReSSInt-EMG sessions with the one larger EMG-UKA session in duration and used the same classifier with the same features and parameters. The idea was that if the new database obtains at least the same results as the reference database, then the new data is valid. The results show that this is the case.

Finally, we would like to remark that the experiments shown in this paper were performed using only sentences, to match the corpus from the EMG-UKA database as much as possible. However, as described in section 3.2.1, the ReSSInt-EMG sessions also contain recorded VCV combinations and isolated words. Experiments that we performed using the VCV combination data only (not included in this paper) show a promising average phone accuracy of 45.03%, with a maximum of 54.43% for one of the sessions. Note that the data for these experiments on average only include an audio duration of 58 seconds per session (split into 80% for training and 20% for testing), thus offering very good accuracy values even for very small amounts of data.

Table 3: *EMG-based phone classification validation and test accuracy obtained with EMG-UKA corpus*

Speaker	Session	MFCC acc. (%)	EMG Valid. acc. (%)	EMG Test acc. (%)
002	101	52.87	26.82±0.35	28.32
002	001	44.04	23.26±0.97	19.99
002	003	45.07	28.11±1.20	25.69
004	001	40.81	20.63±0.75	16.14
006	001	45.44	22.11±2.03	22.26
008	001	44.13	29.75±0.50	25.94
008	002	43.47	29.50±0.88	24.08
008	003	41.25	27.08±0.92	22.78
008	004	43.97	28.55±1.29	23.63
008	005	44.77	29.37±1.29	25.81
008	006	42.99	29.43±1.67	25.79
008	007	43.05	28.33±1.29	23.17
008	008	40.05	27.21±0.48	24.77
average±std		**43.25±1.65**	**26.94±3.02**	**23.33±2.76**

Table 4: *EMG-based phone classification validation and test accuracy obtained with ReSSInt-EMG corpus*

Speaker	Session	MFCC acc. (%)	EMG Valid. acc. (%)	EMG Test acc. (%)
001	101	69.98	45.92±1.24	44.01
001	102	71.93	44.17±0.57	45.25
001	103	71.50	42.07±1.18	43.98
001	104	67.64	40.06±1.10	36.42
002	101	70.57	40.52±0.58	40.91
002	103	68.17	35.86±1.04	37.65
002	104	73.02	36.44±0.98	40.76
003	101	71.01	42.84±0.39	41.80
004	101	69.55	38.78±0.68	36.84
average±std		**70.37±1.64**	**40.74±3.18**	**40.85±3.09**

7. Conclusions and future works

In this paper we have presented the results of the first phone classification experiments carried out on the new ReSSInt-EMG database. For the creation of this new database we had to determine the acquisition setup, most importantly the type, number and locations of the electrodes, the acquisition equipment and the contents and duration of the recording sessions. To validate the final setup, phone classification experiments have been performed using similar experimental conditions for the new database and a well known already developed database. The obtained results provide reassurance on the established data acquisition procedure and open the path to more complex tasks such as continuous speech recognition or direct EMG-to-Speech conversion.

As the recordings are still on-going, the analysis of session- and speaker-dependency has been left for future works. For this, also other architectures such as Deep Neural Networks are being considered. Similarly, experiments will also be developed using the EMG signals obtained in the silent speech modality. Additionally, data from laryngectomees are yet to be obtained and analysed.

Finally, the database also includes video recordings of the face, to process the lips and the area around the lips. We expect that the video modality can provide additional information for those cases where EMG signals are not sufficient (such as tongue and lip movements).

The ReSSInt-EMG database will be made publicly available at the end of the project.

8. Acknowledgements

This work was supported by the Spanish State Research Agency (AEI) Grant PID 2019-108040RB-C21 /AEI/ 10.13039/ 501100011033. We are grateful to the participants for their invaluable contribution.

9. 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] A. R. Toth, M. Wand, and T. Schultz, "Synthesizing speech from electromyography using voice transformation techniques," in *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[3] M. Janke, M. Wand, K. Nakamura, and T. Schultz, "Further investigations on EMG-to-speech conversion," in *ICASSP*, 2012, pp. 365–368.

[4] 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.

[5] K. Sun, C. Yu, W. Shi, L. Liu, and Y. Shi, "Lip-interact: Improving mobile device interaction with silent speech commands," in *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*, 2018, pp. 581–593.

[6] R. Zhang, M. Chen, B. Steeper, Y. Li, Z. Yan, Y. Chen, S. Tao, T. Chen, H. Lim, and C. Zhang, "SpeeChin: A Smart Necklace for Silent Speech Recognition," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 5, no. 4, pp. 1–23, 2021.

[7] G. Krishna, C. Tran, J. Yu, and A. H. Tewfik, "Speech recognition with no speech or with noisy speech," in *ICASSP*. IEEE, 2019, pp. 1090–1094.

[8] Q. Zhou, N. Jiang, and B. Hudgins, "Improved phoneme-based myoelectric speech recognition," *IEEE Transactions on Biomedical Engineering*, vol. 56, no. 8, pp. 2016–2023, 2009.

[9] M. Wand and T. Schultz, "Analysis of phone confusion in EMG-based speech recognition," in *ICASSP*. IEEE, 2011, pp. 757–760.

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

[11] K.-S. Lee, "EMG-Based Speech Recognition Using Hidden Markov Models With Global Control Variables," *IEEE Transactions on Biomedical Engineering*, vol. 55, no. 3, pp. 930–940, 2008.

[12] M. Wand and T. Schultz, "Session-independent EMG-based Speech Recognition," in *Biosignals*, 2011, pp. 295–300.

[13] M. Wand and J. Schmidhuber, "Deep Neural Network Frontend for Continuous EMG-Based Speech Recognition," in *Interspeech*, 2016, pp. 3032–3036.

[14] L. Diener, S. Amiriparian, C. Botelho, K. Scheck, D. Küster, I. Trancoso, B. W. Schuller, and T. Schultz, "Towards Silent Paralinguistics: Deriving Speaking Mode and Speaker ID from Electromyographic Signals," in *Interspeech*, 2020, pp. 3117–3121.

[15] M. U. Khan, Z. A. Choudry, S. Aziz, S. Z. H. Naqvi, A. Aymin, and M. A. Imtiaz, "Biometric Authentication based on EMG Signals of Speech," in *2020 International Conference on Electrical, Communication, and Computer Engineering (ICECCE)*, 2020, pp. 1–5.

[16] M. Zahner, M. Janke, M. Wand, and T. Schultz, "Conversion from facial myoelectric signals to speech: a unit selection approach," in *Interspeech*, 2014, pp. 1184–1188.

[17] 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)*, 2015, pp. 1–7.

[18] 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.

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

[20] D. Gaddy and D. Klein, "Digital voicing of silent speech," *arXiv preprint arXiv:2010.02960*, 2020.

[21] L. Diener, M. Roustay Vishkasougheh, and T. Schultz, "CSL-EMG_Array: An Open Access Corpus for EMG-to-Speech Conversion," in *INTERSPEECH*, 2020.

[22] J. Freitas, A. Teixeira, and M. Dias, "Multimodal corpora for silent speech interaction," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland: European Language Resources Association (ELRA), may 2014.

[23] S. I. Safie, M. I. Yusof, R. Rahim, and A. Taib, "Emg database for silent speech ruqyah recitation," in *2016 IEEE EMBS Conference on Biomedical Engineering and Sciences (IECBES)*, 2016, pp. 712–715.

[24] K.-S. Lee, "Emg-based speech recognition using hidden markov models with global control variables," *IEEE Transactions on biomedical engineering*, vol. 55, no. 3, pp. 930–940, 2008.

[25] S. Ma, D. Jin, M. Zhang, B. Zhang, Y. Wang, G. Li, and M. Yang, "Silent Speech Recognition Based on Surface Electromyography," in *2019 Chinese Automation Congress (CAC)*. Hangzhou, China: IEEE, Nov. 2019, pp. 4497–4501.

[26] I. Hernaez, J. A. González-López, E. Navas, J. L. Pérez Córdoba, I. Saratxaga, G. Olivares, J. Sánchez de la Fuente, A. Galdón, V. García Romillo, M. González-Atienza, T. Schultz, P. Green, M. Wand, R. Marxer, and L. Diener, "Voice Restoration with Silent Speech Interfaces (ReSSInt)," in *Proc. IberSPEECH 2021*, 2021, pp. 130–134.

[27] M. Wand and T. Schultz, "Towards real-life application of EMG-based speech recognition by using unsupervised adaptation," in *Interspeech*, 2014, pp. 1189–1193.

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

[29] 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.

[30] A. Abdullah and K. Chemmangat, "A computationally efficient semg based silent speech interface using channel reduction and decision tree based classification," *Procedia Computer Science*, vol. 171, pp. 120–129, 2020, third International Conference on Computing and Network Communications (CoCoNet'19).

[31] C. Botelho, L. Diener, D. Küster, K. Scheck, S. Amiriparian, B. W. Schuller, T. Schultz, A. Abad, and I. Trancoso, "Toward Silent Paralinguistics: Speech-to-EMG — Retrieving Articulatory Muscle Activity from Speech," in *Interspeech*, 2020, pp. 354–358.

[32] M. Sharma, N. Gaddam, T. Umesh, A. Murthy, and P. K. Ghosh, "A Comparative Study of Different EMG Features for Acoustics-to-EMG Mapping," in *Interspeech*, 2021, pp. 616–620.

[33] ELRA Catalog ID-S0390. (2017) Parallel EMG-Acoustic English GlobalPhone. [Online]. Available: https://catalog.elra.info/en-us/repository/browse/ELRA-S0390/ , urldate = 2022-31-07

[34] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldi." in *Interspeech*, vol. 2017, 2017, pp. 498–502.

[35] 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.

[36] 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.

[37] S.-C. Jou, T. Schultz, M. Walliczek, F. Kraft, and A. Waibel, "Towards continuous speech recognition using surface electromyography," in *Ninth International Conference on Spoken Language Processing*, 2006.

[38] M. Wand, *Advancing electromyographic continuous speech recognition: Signal preprocessing and modeling*. KIT Scientific Publishing, 2015.

[39] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of eugenics*, vol. 7, no. 2, pp. 179–188, 1936.

[40] L. Breiman, "Bagging predictors," *Machine learning*, vol. 24, no. 2, pp. 123–140, 1996.