CSL-EMG_Array: An Open Access Corpus for EMG-to-Speech Conversion

Lorenz Diener, Mehrdad Roustay Vishkasougheh, Tanja Schultz

Cognitive Systems Lab, University of Bremen

lorenz.diener@uni-bremen.de

Abstract

We present a new open access corpus for the training and evaluation of EMG-to-Speech conversion systems based on array electromyographic recordings. The corpus is recorded with a recording paradigm closely mirroring realistic EMG-to-Speech usage scenarios, and includes evaluation data recorded from both audible as well as silent speech. The corpus consists of 9.5 hours of data, split into 12 sessions recorded from 8 speakers. Based on this corpus, we present initial benchmark results with a realistic online EMG-to-Speech conversion use case, both for the audible and silent speech subsets. We also present a method for drastically improving EMG-to-Speech system stability and performance in the presence of time-related artifacts.

Index Terms: EMG, Synthesis, EMG-to-Speech, Silent Speech Interfaces

1. Introduction

Speech is the most important form of communication. It is efficient and natural and lets us easily communicate with each other. With recent advances in human-computer interaction, it also allows us to interact with increasingly complex speech-based user interfaces. However, speech interfaces are not without conceptual issues, since they typically rely on an audible acoustic speech signal. Sometimes, audible speech is simply not an option either due to situational circumstances (e. g. in situations where bystanders would be disturbed or where confidential information needs to be communicated while other people are listening in), or due to the user not being able to produce a clean audible speech signal (e. g. in the case of medical issues impairing speech production, such as laryngectomy). Additionally, when this signal is distorted (e. g. in a large crowd or on a factory floor), performance degrades. A possible approach to address these problems is to build interfaces using speech-related biosignals other than acoustics recorded with a microphone [1].

In recent years, research interest in such *silent speech interfaces* (SSIs) – speech interfaces that continue to function even when an audible acoustic signal is not present [2] – has grown substantially within the speech and general signal processing communities. There is an increasing number of works using signals from ultrasound [3, 4, 5], permanent-magnetic articulography [6, 7], microwave radar [8], *surface electromyography* (sEMG, with muscle movement [9] or sub-vocal [10]), non-audible murmur recorded with a throat microphone [11] or even electrocorticography [12].

EMG-to-Speech conversion is one type of SSI: It refers to the direct conversion of facial electrophysiological muscle activity, measured using surface electromyography, to audible speech. Such a direct conversion approach is well suited to speech prosthesis and silent telephony applications and could be used as a pre-processing step to regular acoustic speech interfaces. To enable conversational use, direct EMG-to-Speech conversion must work in real-time and with low latency. Additionally, it should work on silently recorded signals – i. e. with a user simply mouthing words. Due to data availability, however, EMG-to-Speech systems are mainly evaluated on EMG measured during audible speech – performance on truly silent speech signals is generally not addressed. Finally, an online EMG-to-Speech system – one where output is produced directly as a user speaks – should not require large amounts of time and data for user or session enrolment. Due to signal differences between sessions and speaking modes, these challenges are substantial. Moreover, to the best of our knowledge, there is no publicly available data corpus for EMG-to-Speech conversion research which can be used to investigate solutions to all of these issues.

Available corpora [13, 14], such as the EMG-UKA corpus, were designed with speech recognition in mind, and are therefore not well-suited for training and evaluating EMG-to-Speech conversion systems, especially not in an online context. Firstly, for training EMG-to-Speech conversion systems, sessions with large amounts of utterances are preferable (the EMG-UKA corpus contains only two large sessions, with most sessions containing 50 utterances). Secondly, array electrodes are a vastly more practical alternative to single-electrode setups, cutting the expertise and setup time required to use a system – and unlike with a system using a low number of electrodes, electrode attachment problems can potentially be compensated for. There is currently no publicly available array EMG speech corpus. Finally, existing corpora were recorded with training and testing utterances presented in a fully randomized order. While this is acceptable for testing offline EMG-to-Speech conversion, it is unsuitable for online scenarios. Here, the testing necessarily follows after the training because training data needs to be available before a system can be trained. To get realistic estimates of the performance of an online EMG-to-Speech conversion system, it is necessary to record test data *after* training data, which allows us to account for time-correlated changes in the signal – which an online system will have to compensate for. In addition to collecting such test data, within-session adaptation data to evaluate strategies for such compensation would also be useful.

2. CSL-EMG_Array corpus overview

In this work, we present a corpus of parallel EMG and audio data which can be used to build and evaluate EMG-to-Speech conversion systems in a realistic usage scenario, with speech recorded in two speaking modes – audibly and silently (i. e. merely mouthing words without producing audible acoustic speech) produced.

2.1. Design

The CSL-EMG_Array corpus consists of sessions recorded in a block-wise manner, with a total of 7 blocks recorded in a fixed sequence in numerical order (i. e. first block 1, then block 2, then block 3, etc.) and utterances within a block presented in randomized order (as opposed to previous corpora recording all utterances in a randomized manner with no time structure, as one single block). This closely mirrors the real online EMG-

Table 1: *Amount of utterances for different recording blocks (amounts in parentheses include additional sentences only present for silent testing mode sessions).*

Subset	train	dev	eval
(Block0_Align)	-	(50)	(40)
Block1_Initial	250	50	40
Block2_Adapt1	20	20	20
Block3_Eval1	-	30	20
Block4_Adapt2	20	-	-
Block5_Eval2	-	30	20
Block6_Adapt3	20	-	-
Block7_Eval3	-	30	20
Total	310	160 (210)	120 (160)

to-Speech conversion scenario and therefore (unlike corpora where training and testing data do not have any temporal structure) allows for the development and testing of online EMG-to-Speech conversion systems with realistic estimates of online performance. The utterances are English sentences from the broadcast news domain, and are split into training, development and evaluation subsets.

Block 1 includes recordings of the entire set of sentences available in the corpus (the full training, development and evaluation sets). It can be used to train and optimize EMG-to-Speech systems and to create a baseline for evaluation in a manner that is comparable to offline EMG-to-Speech conversion.

Block 2, 4 and 6 each contain 20 training sentences (identical in each case). Block 2 additionally contains 20 utterances each for development and evaluation. These can be used to implement strategies for session enrolment and within-session adaptation.

Finally, Block 3, 5 and 7 contain 30 development and 20 evaluation utterances to evaluate these strategies on data not recorded concurrently with the data that the system is being trained on. This matches the evaluation scenario of a real online EMG-to-Speech conversion system, where compensation for time-related artifacts such as electrode detachment or impedance drift is required. Table 1 presents an overview of the utterance counts in each block and subset.

There are two types of sessions in the corpus: Audible-testing-mode sessions, and silent-testing-mode sessions. For the audible-testing-mode sessions, subjects were prompted to simply read out the utterances as they normally would, and parallel EMG- and Audio signals are included for each utterance. For silent-testing-mode sessions, subjects were asked to silently mouth all sentences that are part of the development or evaluation subset in blocks 1 through 7 (i. e. mouthing without producing sound while reading along) – for these, only an EMG signal is included, as a reference audio signal is not produced. Note that this means that for these sessions, it is not possible to directly compare the systems output with a reference signal since an acoustic signal does not exist when people speak silently.

The lack of audible acoustic reference data in silent-testing-mode sessions is a problem when trying to evaluate EMG-to-Speech systems built for this mode: Common measures such as the mel-cepstral distortion score rely on such a reference signal and cannot be computed when it is not available. To still allow for objective evaluation, silent sessions include an additional block 0 that contains an audible recording of the development and evaluation utterances (both EMG and Audio). This data can be used to evaluate EMG-to-Speech conversion output using *dynamic time warping* (DTW) alignment or similar techniques.

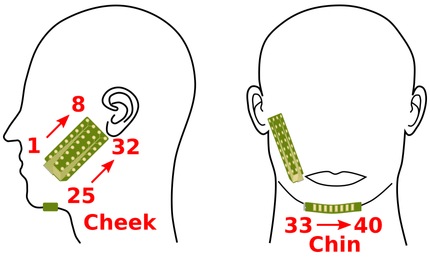

Figure 1: *EMG array electrode positions and numbering for the recordings in the CSL-EMG_Array corpus. Derivation is chained-differential, i. e. channel 1 is between electrode 1 and 2, channel 2 is between electrode 2 and 3, etc.*

In addition to the EMG- and audio data, metadata about the recordings (including transcripts) are also included.

2.2. Recording setup and data formats

Recordings were performed in a recording chamber shielded against acoustic and electromagnetic interference. The audio signals included in the corpus were recorded using a RØDE NT-1 condenser microphone and a Behringer Xenyx 302 audio interface. The EMG signals were recorded using an OT Bioelettronica Quattrocento EMG amplifier. Two array electrodes were used: One 4x8 electrode array on the left cheek, and one 8 electrode strip below the chin. The electrodes were used in a chained differential derivation configuration (see Figure 1 for positioning). Cross-row channels were not excluded and instead provided as-is. Finally, one channel was added to both the EMG- and audio signal, containing a marker that is pulled high by the EMG amplifier at the start of each utterance, allowing for easy synchronization of the EMG and audio signals by alignment of the markers. Audio data was sampled at 16000 Hz. The EMG signal was sampled at 2048 Hz with a 0.3 Hz DC offset removal and a 500 Hz anti-aliasing filter applied, and re-scaled to millivolt range (i. e. an EMG signal value of 1 for a channel means 1 mV of measured voltage difference).

The audio and EMG data is provided as one file per utterance, with files for each block in a separate folder and utterance files in separate sub-folders inside the block folder. Acoustic and EMG signal data are provided in the numpy version 3 format [15]. Additionally, for convenience, the acoustic data is also provided as 16-bit unsigned PCM wav files. The metadata is provided in a text-based json format, with one json file per block containing information about the block itself (sampling rate and filter parameters, ids of training, development and evaluation sets, order of recording) as well as about each utterance (utterance text and recording timestamp as unix epoch).

2.3. Recorded speakers and sessions

The corpus contains twelve sessions from a total of eight speakers. Four speakers (speakers 2, 4, 5 and 7) recorded audible sessions only, the other 4 (speakers 1, 3, 6 and 8) recorded both an audible and a silent session. The recorded speakers read English sentences but, as they were recruited from the general student population at a German university, are not native English speakers. They were allowed to re-attempt recording as often as desired if they felt they needed to correct their pronunciation. Three of the speakers were female, and five speakers were male. Speakers ages ranged between 19 and 32 years old. A detailed breakdown of the sessions can be found in Table 2. In total, 9.5 hours of data are available. Informed written consent of all recorded speakers was acquired prior to the collection of data.

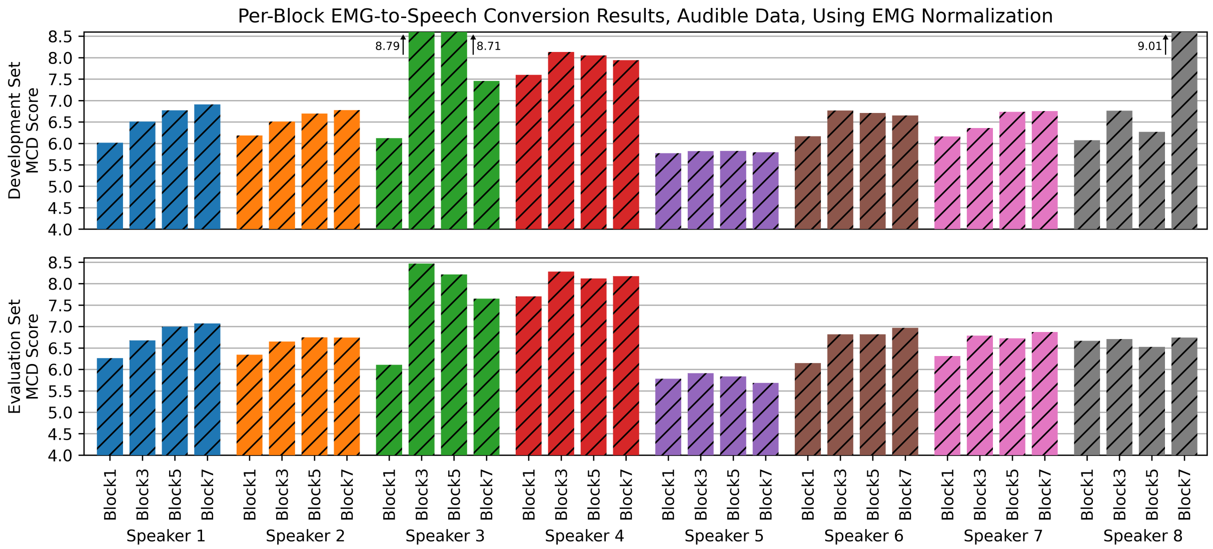

Figure 2: *Baseline MCD scores (lower is better) for real-time session-dependent EMG-to-Speech conversion with EMG normalization, training on Block 1 and evaluating on Blocks 1, 3, 5 and 7.*

Table 2: *Session durations broken down by training, development and testing set as well as speaker gender and session mode.*

			Total (mm:ss)			Mean (mm:ss)		
Session	mode	m/f	train	dev	eval	train	dev	eval
Spk1	aud	m	25:21	12:06	10:33	4.9	4.5	5.3
Spk1-Sil	sil	m	21:33	13:17	11:46	4.2	3.8	4.4
Spk2	aud	f	26:14	11:50	10:09	5.1	4.4	5.1
Spk3	aud	f	24:33	11:16	10:16	4.8	4.2	5.1
Spk3-Sil	sil	f	23:20	15:33	13:42	4.5	4.4	5.1
Spk4	aud	m	31:31	14:04	12:26	6.1	5.3	6.2
Spk5	aud	m	20:53	9:29	8:14	4.0	3.6	4.1
Spk6	aud	m	28:42	13:09	11:30	5.6	4.9	5.8
Spk6-Sil	sil	m	28:40	16:25	14:21	5.5	4.7	5.4
Spk7	aud	f	25:13	11:37	10:22	4.9	4.4	5.2
Spk8	aud	m	20:50	10:02	8:30	4.0	3.8	4.2
Spk8-Sil	sil	m	20:44	12:51	10:59	4.0	3.7	4.1
All			297:35	151:38	132:49	4.8	4.3	5.0

3. Initial EMG-to-Speech evaluation

To provide a baseline for future results and to further illustrate the usefulness of the novel recording protocol of the CSL-EMG_Array corpus, we present initial EMG-to-Speech conversion results on this corpus. Here, we provide initial results using the *mel-cepstral distortion* (MCD) score, a distance measure for comparing system output to a reference audio file [16].

3.1. Real-time C-TD15 EMG features

The original TD-15 feature set [17] has proven to be resilient and effective for EMG-to-Speech conversion. It is, however, of limited use when building a system that needs to output data with very low latency to enable natural spoken communication. In particular, it requires both explicit future context through stacking (150 ms into the future) as well as implicitly (to calculate a 9 point double average, which also limits its usefulness when using sample rates other than 600 Hz). For this reason, we introduce a new feature set: The *causal TD15* (C-TD15) features, which can be calculated with low latency and no explicit future context.

Table 3: *Baseline MCD scores (lower is better) for an EMG-to-Speech conversion system without within-session normalization, audible testing mode sessions only.*

	Block (Dev. set)				Block (Eval. set)			
Session	1	3	5	7	1	3	5	7
Spk1	7.54	7.78	8.54	8.95	7.84	7.94	8.49	8.83
Spk2	7.55	9.98	8.22	8.46	8.12	9.57	8.21	8.41
Spk3	7.5	17.14	14.71	9.06	7.87	17.33	14.4	9.07
Spk4	8.66	9.82	9.39	9.4	8.71	9.75	9.51	9.48
Spk5	7.49	7.95	7.69	7.65	7.46	7.93	7.64	7.56
Spk6	7.41	7.96	7.94	7.88	7.64	8.01	8.0	8.08
Spk7	7.46	8.23	8.58	8.63	8.08	8.48	8.63	8.8
Spk8	7.83	8.18	8.01	8.31	7.96	8.27	8.25	8.33

To calculate C-TD15 features for a single channel, the signal is first split into a high-frequency-band and low-frequency-band part using a third-order Butterworth high- and low-pass filters with a cutoff frequency of 134 Hz (resulting in a delay of approx. 12 samples). The high- and lowband signals are then each processed into frames with a Blackman window of 32 ms length and 10 ms shift. From the resulting frames, the lower-band power, lower-band mean, higher-band power, higher-band zero-crossing rate and higher-band absolute-value mean are calculated, resulting in one C-TD1 frame. The C-TD1 frame is stacked together with the 14 preceding C-TD1 frames to obtain the final C-TD15 feature vector for that channel. To calculate the C-TD15 features for a multi-channel EMG signal, the C-TD15 features for each channel are calculated separately and then concatenated to obtain the combined multi-channel EMG feature vector.

3.2. Target acoustic speech features for audio files

Our system uses *Mel-Frequency Cepstral Coefficients* (MFCCs) together with the non-continuous *Fundamental Frequency* (F_o) to represent the target acoustic signal as audio file. The acoustic signal was first windowed with the same parameters as the EMG signal. MFCCs were then extracted for each window as the filter parameters of a Mel-Log Spectrum Approximation filter [18], allowing for efficient re-synthesis. F_o was extracted using the

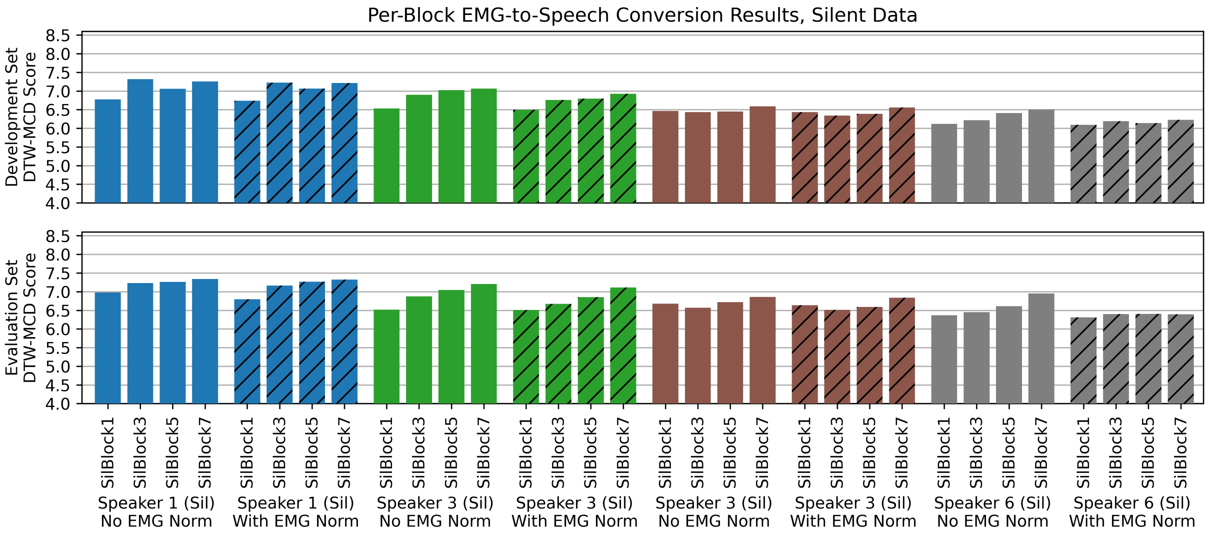

Figure 3: *Baseline MCD scores (lower is better) for real-time session-dependent EMG-to-Speech conversion, training on Block 1 and evaluating on silent data from Blocks 1, 3, 5 and 7.*

Yin algorithm [19]. Before calculating acoustic features, we normalize the audio waveform to be between -0.99 and 0.99 to ensure a consistent signal amplitude range between utterances.

3.3. EMG-to-Speech conversion

To train a basic EMG-to-Speech conversion system for one session, we first align the EMG- and acoustic signals of all training utterances from Block 1 of that session using the marker channel. We then extract parallel EMG C-TD15 and acoustic MFCC and F_o features. Using these features, we trained two neural networks: One to predict MFCC features and one to predict F_o trajectories, both from C-TD15 input. We used a bottleneck network structure that has proven to work well for this task [20] and amount of data, with hidden layer sizes of 2048, 512 and 1024, with dropout regularization after each layer. We trained the networks for 500 epochs using stochastic gradient descent with a learning rate of 0.01 and a minibatch size of 1024 (training times for these systems, on an Nvidia RTX2080Ti GPU, ranged between 20 and 25 minutes). No further tuning of model structure and hyper-parameters was performed for this evaluation.

For evaluation, we converted the EMG features of the development and evaluation data of blocks 1, 3, 5 and 7 of the same session to acoustic speech features using the trained networks and compare the resulting MFCCs to reference MFCCs, aligned using the marker. Results of this evaluation are summarized in Table 3. The systems operate as expected on the training block (block 1). However, performance gets worse or breaks down entirely for the online scenario (i. e. on those evaluation data which are recorded after training data, as evaluated using blocks 3, 5 and 7). This degradation is likely due to changing conditions over time, e. g. electrode gel drying out or speaker fatigue. Note that, while generating F_o from EMG is possible we do not consider F_o evaluation, as the results presented in this paper are merely given as a baseline for future work.

3.4. Real-Time EMG normalization

We next present an initial method for addressing these differences, enabling EMG-to-Speech conversion in a realistic real-time online scenario. We achieve this by performing running normalization of the EMG signal. We keep track of the 99th percentiles of the absolute value of EMG channels over 250 ms and normalize all samples using this 99th percentile value unless this would result in an amplification greater than 100. This keeps the signal in a range of -1 to 1, compensating for drift and short artifacts while not amplifying noise from detached electrodes. The results of applying the normalization are shown in Figure 2. Compared to the baseline (not shown), MCD scores for the evaluation blocks have improved and the normalized system is able to produce output for all blocks, enhancing practical usability.

3.5. Evaluation on silent data

Last but not least, we evaluate our systems on silently recorded speech data. The training procedure is as described in the previous sections. However, since there is no parallel reference audio data, evaluation is performed by first aligning the system output MFCCs with audible reference MFCCs from block 0 using the DTW algorithm. We then calculate the MCD score between output and aligned MFCCs as before, resulting in a DTW-MCD measure. Baseline results for this evaluation mode are plotted in Figure 3. Note that, due to the alignment, direct comparison of these scores to non-DTW MCD scores is not possible, and the DTW alignment is expected to reduce differences between systems – however, it can be seen that the silent sessions follow the same general trends described in the previous sections.

4. Summary and outlook

We have presented the CSL-EMG_Array corpus for EMG-to-Speech conversion. This corpus uses a state-of-the-art EMG electrode array setup. Due to the corpus design, it allows the easy and comparable training and testing of EMG-to-Speech systems for many paradigms, including offline and online systems, session adaptive systems and systems operating on silently produced speech. The corpus is openly available for research purposes[1]. In the future, we plan to evaluate different approaches to adaptation using this corpus, with the goal of improving performance in an online evaluation context as well as in direct EMG-to-Speech conversion of silent speech.

[1] https://www.uni-bremen.de/en/csl/research/silent-speech-communication/csl-emg-array-corpus

5. References

[1] T. Schultz, M. Wand, T. Hueber, D. J. Krusienski, C. Herff, and J. S. Brumberg, "Biosignal-based spoken communication: A survey," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 25, no. 12, pp. 2257–2271, nov 2017. [Online]. Available: https://www.csl.uni-bremen.de/cms/images/documents/publications/TASLP-2017-biosignal-based-spoken.pdf

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

[3] T. Grósz, G. Gosztolya, L. Tóth, T. G. Csapó, and A. Markó, "F0 estimation for dnn-based ultrasound silent speech interfaces," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 291–295.

[4] G. Gosztolya, T. Grósz, L. Tóth, A. Markó, and T. G. Csapó, "Applying dnn adaptation to reduce the session dependency of ultrasound tongue imaging-based silent speech interfaces," *Acta Polytechnica Hungarica*, vol. 17, no. 7, 2020.

[5] D. Fabre, T. Hueber, L. Girin, X. Alameda-Pineda, and P. Badin, "Automatic animation of an articulatory tongue model from ultrasound images of the vocal tract," *Speech Communication*, vol. 93, pp. 63–75, 2017.

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

[7] B. Cao, N. Sebkhi, T. Mau, O. T. Inan, and J. Wang, "Permanent magnetic articulograph (pma) vs electromagnetic articulograph (ema) in articulation-to-speech synthesis for silent speech interface," in *Proceedings of the Eighth Workshop on Speech and Language Processing for Assistive Technologies*, 2019, pp. 17–23.

[8] P. Birkholz, S. Stone, K. Wolf, D. Plettemeier, K. Wolf, D. Plettemeier, S. Stone, and P. Birkholz, "Non-invasive silent phoneme recognition using microwave signals," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 26, no. 12, pp. 2404–2411, 2018.

[9] 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, nov 2017.

[10] A. Kapur, S. Kapur, and P. Maes, "Alterego: A personalized wearable silent speech interface," in *23rd International Conference on Intelligent User Interfaces*. ACM, 2018, pp. 43–53.

[11] T. Toda and K. Shikano, "Nam-to-speech conversion with gaussian mixture models," in *Proceedings of the Annual Conference of the International Speech Communication Association*, 2005, pp. 1957–1960.

[12] C. Herff, G. Johnson, L. Diener, J. Shih, D. Krusienski, and T. Schultz, "Towards direct speech synthesis from ecog: A pilot study," in *Engineering in Medicine and Biology Society (EMBC), 2016 38th Annual International Conference of the IEEE*, Aug 2016.

[13] 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)*, N. C. C. Chair), K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis, Eds. Reykjavik, Iceland: European Language Resources Association (ELRA), may 2014.

[14] M. Wand, M. Janke, and T. Schultz, "The emg-uka corpus for electromyographic speech processing," in *The 15th Annual Conference of the International Speech Communication Association, Singapore*, 2014, interspeech 2014.

[15] T. Oliphant, "NumPy: A guide to NumPy," USA: Trelgol Publishing, 2006–, [Online; accessed May 2nd, 2020]. [Online]. Available: http://www.numpy.org/

[16] R. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," in *Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing*, vol. 1, May 1993, pp. 125–128 vol.1.

[17] S.-C. S. Jou, T. Schultz, M. Walliczek, F. Kraft, and A. Waibel, "Towards continuous speech recognition using surface electromyography," in *In proceedings of the 9th ISCA International Conference on Spoken Language Processing*, 2006.

[18] S. Imai, "Cepstral analysis synthesis on the mel frequency scale," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 8, 1983, pp. 93–96.

[19] A. De Cheveigné and H. Kawahara, "Yin, a fundamental frequency estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, 2002.

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