Towards Silent Paralinguistics: Deriving Speaking Mode and Speaker ID from Electromyographic Signals

Lorenz Diener[1], *Shahin Amiriparian*[2], *Catarina Botelho*[3], *Kevin Scheck*[1], *Dennis Küster*[1], *Isabel Trancoso*[3], *Björn W. Schuller*[2,4], *Tanja Schultz*[1]

[1]Cognitive Systems Lab (CSL), University of Bremen, Germany
[2]Chair of Embedded Intelligence for Health Care and Wellbeing, Universität Augsburg, Germany
[3]INESC-ID/Instituto Superior Técnico (IST), University of Lisbon, Portugal
[4]GLAM – Group on Language, Audio, and Music, Imperial College London, UK

lorenz.diener@uni-bremen.de

Abstract

Silent Computational Paralinguistics (SCP) – the assessment of speaker states and traits from non-audibly spoken communication – has rarely been targeted in the rich body of either Computational Paralinguistics or Silent Speech Processing. Here, we provide first steps towards this challenging but potentially highly rewarding endeavour: Paralinguistics can enrich spoken language interfaces, while Silent Speech Processing enables confidential and unobtrusive spoken communication for everybody, including mute speakers. We approach SCP by using speech-related biosignals stemming from facial muscle activities captured by surface electromyography (EMG). To demonstrate the feasibility of SCP, we select one speaker trait (speaker identity) and one speaker state (speaking mode). We introduce two promising strategies for SCP: (1) deriving paralinguistic speaker information directly from EMG of silently produced speech versus (2) first converting EMG into an audible speech signal followed by conventional computational paralinguistic methods. We compare traditional feature extraction and decision making approaches to more recent deep representation and transfer learning by convolutional and recurrent neural networks, using openly available EMG data. We find that paralinguistics can be assessed not only from acoustic speech but also from silent speech captured by EMG.

Index Terms: Silent Speech, Electromyography, Paralinguistics

1. Introduction

Computational Paralinguistics have been demonstrated to reveal large amounts of information from the voice of a speaker [1]. The usage of such information, including a speaker's affective state, health status, personality, as well as manifold further states and traits bears huge potential, for example in spoken communication or voice-based user interfaces.

A silent speech interface (SSI) [2] is a system enabling spoken communication to take place even when an audible acoustic signal is unavailable from the user. By capturing sensor data from the speech production process involving the articulators, articulatory muscle activities, neural pathways and the brain itself, the resulting biosignals provide a representation of speech beyond acoustics that can be used for spoken communication [3]. Such speech-related biosignals allow to circumvent the limitations of conventional speech processing systems and are the basis of computational paralinguistics for silent speech (We define biosignals as autonomous signals produced by human activities measured in physical quantities using different sensor technologies [4]).

In particular, the human-computer interaction (HCI) community has embraced biosignals to extend the number of modalities available for developing robust and intuitive devices. Information obtained from biosignals is used to interpret not only physical states, but also affective and cognitive states, and activities of a user. Thereby, biosignals provide an inside perspective on human mental processes, intentions, and needs that complement traditional means of observing human interaction from the outside, and thus enable personalized and adaptive services [4].

Silent Computational Paralinguistics (SCP) reveal paralinguistics for situations when audible acoustic signals are not available or advisable, e. g., due to privacy concerns or disturbance of others, adverse noise conditions, or speech pathologies. While SSIs have previously addressed Automatic Speech Recognition, e. g., from video, EMG, or ultrasound [5], or examined how to synthesize silent to audible speech, e. g., for laryngectomy patients [6, 7, 8], research on privacy for paralinguistic analysis has focused mostly on whispered speech [9, 10, 11]. Some research has explored EMG for emotion recognition [12], and facial expressions to enhance human-computer interaction [13] or human-robot interaction [14]. However, the authors are not aware of any published works aimed at advancing SSIs via computational paralinguistics, or at extending computational paralinguistics to silent speech.

This paper takes the first steps towards SCP. We use EMG as our modality, exploiting physiological cues to estimate paralinguistic information. We introduce two methods: *direct EMG-based paralinguistics* which estimates speakers' traits and states directly from EMG, and subsequent *indirect EMG-to-Speech paralinguistics* which first synthesizes an audible speech signal from EMG and then applies standard acoustic paralinguistic methods. We chose one speaker trait (speaker identity) and one speaker state (speaking mode), and applied both traditional machine learning and deep learning approaches based on unsupervised and transfer learning to perform our experiments.

The rest of the paper is organized as follows: Section 2 gives a brief overview of the EMG-UKA data corpus used in this paper and related terminology. Section 3 introduces EMG-to-Speech conversion as the foundation for later experiments on EMG-to-Speech output. Section 4 presents several methods for performing speaker identification directly on EMG data. Section 5 presents our results of performing speaker identification with deep autoencoders. Section 6 then presents our results of performing speaking mode recognition, and section 7 finally gives a summary of the work performed and potential future work.

Figure 1: *EMG-UKA electrodes with bipolar derivation (white) and derivation against a neutral reference (black numbers).*

2. Data overview

We use the EMG-UKA parallel EMG-Speech corpus [15, 16]. This corpus contains acoustic and EMG speech signals recorded in parallel, including a marker channel to compensate for different delays in the signal recording paths. The audio data was recorded at a sampling rate of 16 kHz, with a standard close-talking microphone, whereas the EMG signals were recorded using a Becker Meditec Varioport device with 6 EMG channels, operating at 600 Hz. An overview of the electrode positioning can be seen in Figure 1. The corpus includes a total of 63 sessions recorded from 8 speakers, featuring 3 different speaking modes (audible aka modal speech, silent aka mouthed speech, whispered speech) as part of 32 multi-mode sessions. While a modal speech signal is not available for all utterances, the EMG signal is always recorded – we call the EMG data for different modes "audible EMG", "whisper EMG" and "silent EMG", respectively.

A subset of the sessions is freely available as a trial corpus [15], the full corpus is available from ELRA [16]. A breakdown of these sessions by speaker can be found in Table 1, while Table 2 provides information about the size of the different mode subsets. To analyze the performance of acoustic paralinguistic analysis methods based on the EMG-to-Speech system output, the EMG data of all utterances was converted to audible speech as described below.

3. EMG-to-Speech conversion

To convert EMG signals to acoustic (audible) speech, we use a deep neural network system, as described in our previous work [17]. The bottleneck-shape 3-hidden-layer neural network systems (first layer width of 1024 neurons) were trained in a session-dependent manner, using TD-15 features as input. The networks were trained to output mel-frequency cepstral coefficients (MFCCs) and fundamental frequency values for use with a mel-log spectrum approximation vocoder [18], which can then be used to produce an acoustic speech waveform. For silent test data, conversion was performed using a model trained on the audible data of the same session.

Table 1: *EMG-UKA Corpus: Speaker Breakdown. (*) indicates session is part of the trial corpus, numbers in brackets indicate number of sessions / utterances that are part of the trial corpus.*

	#sessions			
Speaker	Total	Large	Multi-Mode	#utterances
1	3	0	3	450 (0)
2 (*)	33 (3)	1 (1)	15 (2)	3720 (820)
3 (*)	1 (1)	0	1 (1)	150 (0)
4	2	0	2	300 (150)
5	1	0	1	150 (0)
6 (*)	1 (1)	0	1 (1)	150 (150)
7	2	0	2	300 (0)
8 (*)	20 (8)	1	7 (2)	2159 (600)
Total	63 (13)	2 (1)	32 (6)	7379 (1720)

Table 2: *EMG-UKA Corpus: Subset Breakdown*

			Duration ([h:]mm:ss)	
Subset	#Spk	#Sess	Average	Total
Audible (Small)	8	61	03:08	3:11:34
Whispered (Small)	8	32	03:22	1:47:42
Silent (Small)	8	32	03:19	1:46:20
Audible (Large)	2	2	27:02	54:04
Whole Corpus	8	63		7:32:00

On the audible data (EMG signals from audible speech), we obtained a mean mel-cepstral distortion (MCD) score of 6.48 (*SD* 0.79; MIN 4.5; MAX 9.06). On the silent data (EMG signals from silent speech), using audible data to which the output was aligned using dynamic time warping (DTW) as the reference, the mean DTW-MCD score was 6.38 (*SD* 0.57; MIN 4.86; MAX 8.86). We further calculate the concordance correlation coefficient (CCC) for each of the MFCCs, as estimated from the EMG data versus from the reference audio (Figure 2), suggesting that lower coefficients (i. e., large spectral changes) were predicted relatively well, whereas the prediction quality for the higher coefficients was degraded. As demonstrated by our complementary work on retrieving articulatory muscle activity from EMG [19], the CCC for this inverse SCP problem is around 0.62. I. e., results for Speech-to-EMG CCCs have been comparable to the prediction of the first two MFCCs in figure 2.

4. Direct EMG-based speaker recognition

The speech EMG signal is known to be session-dependent, i. e. recording sessions may produce markedly different patterns of EMG data, even when the same utterance is spoken. These signal differences cannot be explained by simple hyper- or hypoarticulation but instead are more complex. Between-session differences in EMG signals are caused not only by shifts in electrode positions and impedances, but also by differences in tissue, padding, facial hair, skin and muscle conditions, makeup ("daily form") as well as speaker idiosyncrasies in speech production. Therefore, it should be possible to perform *speaker recogni-*

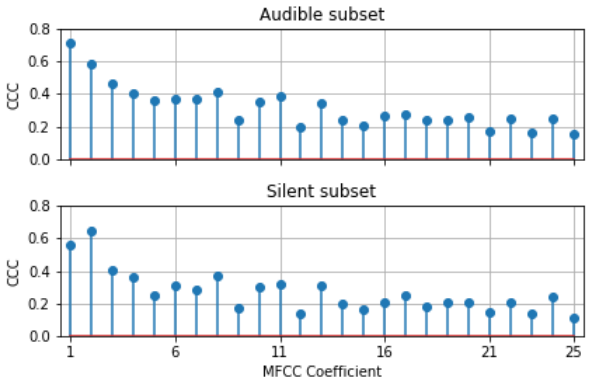

Figure 2: *CCC between MFCCs estimated from EMG data and MFCCs calculated from reference audio, for the audible subset (above) and the silent subset (aligned via DTW, below).*

tion from EMG signals. From the three problems of speaker recognition, i. e., speaker identification (SID), verification, and diarization, we chose the first one for the present experiments.

4.1. Experimental setup

To evaluate the SID rate from the speech-related EMG signal, we picked those speakers from the EMG-UKA corpus for whom more than one session is available (speakers 1, 2, 4, 7 and 8; 60 sessions in total). We trained each system on all but one session and evaluated on the hold-out session to ensure that results are based on speaker characteristics rather than session-specific aspects. From the EMG signal we derived a set of time-domain (mean absolute value, root mean square, sum absolute values, variance, simple square integral, waveform length, average amplitude change, zero crossing rate, slope sign change) and frequency-domain (median frequency, weighted mean frequency) features. Using these features, we trained and evaluated a Linear Discriminant Analysis (LDA) and a Random Forest classifier with 290 trees and a maximum depth of 120 for utterance-wise speaker identification.

4.2. Results and discussion

Table 3 shows the accuracy of the LDA classifier, suggesting that results were impacted by the imbalanced data – with higher performance for the speakers with a larger amount of training data (e. g., accuracy for speaker 8 is nearly perfect). However, the LDA and Random Forest classifiers achieved 83 % and 84 % accuracy, respectively, and were well above chance level (chance level for the accuracy being the prevalence of the most common class, 55 %). The worst performing speaker is speaker 7, for which, likely due to imbalance, not a single utterance is assigned.

To compensate for the imbalanced data, we calculated the unweighted average recall (UAR). The LDA classifier obtained a UAR of 73%, and clearly outperformed the Random Forest, which favored the frequently seen speakers and achieved a UAR of 64 %. Chance level for UAR would be 12.5 %. Still, it can be seen that for both classifiers speakers who were under-represented in training had a higher chance for mis-classification.

A potential caveat is possible sequence effects between sessions due to electrode condition and speaker form if the same speaker is recorded for multiple sessions back to back. Because the UKA corpus does not have information on recording dates and times available, we cannot investigate this with just the UKA data – further recordings may be necessary to fully exclude it.

4.3. Speaker recognition using acoustic speaker embeddings and transfer learning

Next, we evaluate whether the speaker identification can be improved by using speaker embeddings obtained from acoustic data. We first use the softmax variant of the Generalized End-to-End (GE2E) loss [20] to train a long short-term memory (LSTM) recurrent neural network (RNN) that maps sequences of 25 MFCCs and the fundamental frequency to a single 64-dimensional speaker embedding. Its architecture consists of three stacked, unidirectional LSTM layers with a hidden size of 192. The final state of the last LSTM layer is downsized by a linear layer to the embedding dimensionality and is lastly L2-normalized. For training, we use slices of 32 consecutive acoustic features of non-silent utterances of all sessions. Subsequently, we trained an EMG speaker encoder with the same architecture to produce equivalent embedding vectors given slices of EMG TD-0 features. To train the EMG encoder, we minimize the L2 loss between speaker embeddings of parallel acoustic and EMG features of all but the hold-out session. For predicting the speaker ID of EMG data of a hold-out session, we use an LDA model fit on utterance-level embeddings of the training sessions. Compared to the LDA approach, this approach results in slightly lower performance: The UAR is 45.23 % (Accuracy 72.03 %). A detailed breakdown of per-speaker accuracy can be found in Table 3. While, unlike for the LDA and Random Forest models, some utterances were assigned to speaker 7, the lower UAR seems to indicate a strong bias towards the prior.

5. Direct and indirect deep EMG-based speaker recognition with autoencoders

In addition to our speaker identification approach introduced in Section 3, we apply recurrent autoencoders to obtain a new set of features. The evaluation mode is the same as the setup described in Section 4.1.

We obtain our deep features through unsupervised representation learning with recurrent sequence to sequence autoencoders, using the AUDEEP toolkit[1] [21, 22]. Such representation learning avoids manual feature engineering. Furthermore, learned feature sets have repeatedly been shown to be superior to hand-crafted feature sets for a variety of tasks [23, 21].

5.1. Results on EMG

In the AUDEEP approach, 128-band Mel-scale spectrograms were first extracted from the channel mean of the raw EMG measurements. Here, we use 40 ms FFT windows with a hop size of 20 ms. To eliminate some noise from the input signal, power levels are clipped below four different given thresholds (-30, -45, -60 and -75dB), resulting in four separate sets of spectrograms per data set. A distinct recurrent sequence to sequence autoencoder (2 hidden layers, 256 gated recurrent units/layer, unidirectional en- and bidirectional decoder) was trained on each of these sets of spectrograms in an unsupervised way for 64 epochs in batches of 256 samples with a learning rate of 0.001 and a dropout rate of 20 %. The learnt representations were extracted as feature vectors for the corresponding instance, and concatenated to obtain the final feature vector. A linear support vector machine (SVM) algorithm with the complexity value $C = 0.001$ was employed for the classification of the AUDEEP features. Using the introduced hyperparameters and configurations, we achieve 51.7 % UAR (86.6 % accuracy), 55.2 % UAR (82.1 % accuracy), and

[1] https://github.com/auDeep/auDeep

Table 3: *Session-wise MIN, MAX, and mean (+/- SD) of the per utterance SID accuracy from EMG using three different methods.*

	LDA			Random Forest			Embedding Transfer		
Spk#	Worst	Best	Mean	Worst	Best	Mean	Worst	Best	Mean
1	0.97	0.99	0.98±0.01	0.58	0.99	0.85±0.19	0.25	0.33	0.29±0.03
2	0.34	1.0	0.95±0.13	0.96	1.0	0.99±0.01	0.1	0.98	0.81±0.2
4	0.0	1.0	0.5±0.5	0.01	0.68	0.35±0.33	0.29	0.49	0.39±0.1
7	0.0	0.0	0.0±0.0	0.0	0.0	0.0±0.0	0.01	0.04	0.02±0.01
8	0.99	1.0	1.0±0.0	0.83	1.0	0.99±0.04	0.43	0.9	0.75±0.14
All	0.0	1.0	0.92±0.24	0.0	1.0	0.93±0.22	0.01	0.98	0.72±0.25

55.4 % UAR (82.0 % accuracy) for audible EMG, whisper EMG, and silent EMG, respectively. Compared to the results provided in Section 3, we observe that AUDEEP results are below the LDA models. We assume, the reason for this difference is twofold: First, our recurrent autoencoders cannot generalise well on the small dataset, and second: The recurrent models are still more affected by class imbalance than the less complex LDA model.

5.2. Results on EMG-to-Speech audio

In addition to performing AUDEEP evaluation on the EMG signal directly, we also evaluate a system trained to work on the output on an EMG-to-Speech system. In terms of UAR, the model operating on EMG converted to audible speech obtains an UAR of 56.32% and an accuracy of 80.63%. Overall, speaker identification based on speech obtained from EMG-to-Speech systems seems feasible.

6. Direct EMG-based recognition of speaking mode

In addition to differences between sessions and speakers, the speech EMG signal also changes depending on speaking *mode* – i. e., the signal changes depending on whether a speaker is producing modal (audible) speech, whispered speech, or is speaking silently (i. e., performing articulatory gestures without producing sound), with the differences going beyond simple hyper- or hypoarticulation. Classifying in which mode an utterance is spoken would be useful for assisting silent speech interface research. E. g., such classifications could be used to dynamically select models for different modes, and the EMG-based parameters of a model of a speaking mode may help to characterize important signal differences. Furthermore, mode recognition provides a powerful and complementary validation test for SCP because most of the between-subjects variance (e. g., differences in skin, facial hair) is held constant.

6.1. Experimental setup

To evaluate EMG-based mode classification, we use the multi-mode sessions from the EMG-UKA corpus, a total of 32 sessions. Each of these sessions contains 50 utterances spoken in each mode, split into a development and an evaluation set. We trained an LDA classifier on all sessions' training sets and evaluated it on each sessions' test set. As features, we used the mean, *SD*, and the 0th, 25th, 75th and 100th percentile values of the TD features that were also used as the input for EMG-to-Speech conversion [18].

6.2. Results and discussion

As shown by Figure 3, the mode classifier performs well for silent audio but less well for audible and whispered speech, which the classifier often confused. This matches our expectation that audible and whispered speech should be broadly similar in production, whereas the speech EMG signal for fully silent speech (where no audible feedback is available to the speaker whatsoever) is very different.

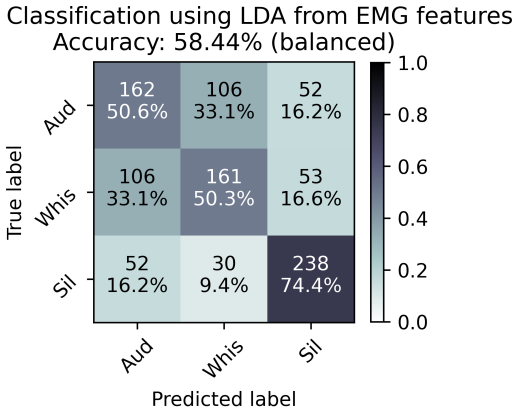

Figure 3: *Confusion matrix of performing mode classification on EMG data using an LDA model.*

7. Conclusion

In this first step towards Silent Computational Paralinguistics (SCP), we introduced two methods, *direct EMG-based paralinguistics* and subsequent *EMG-to-Speech paralinguistics* to estimate speaker ID and speaking mode on the basis of EMG biosignals. The experimental results suggest that SCP are feasible, and that both methods achieve results well above chance level. Our results for SCP mode recognition further suggest that EMG data may be sufficient to allow detection of, at least, broad differences in speech production. However, so far, direct SCP appears to outperform the indirect approach of first synthesizing speech from EMG. Thus, as current results suffer from small and imbalanced training sets, we hope to collect and share more data in the near future.

From our point of view, Silent Computational Paralinguistics offers an exciting new direction in which a lot of research can still be done to further our understanding of speech beyond acoustics.

8. Acknowledgements

This work was partially supported by national funds through FCT, Fundação para a Ciência e a Tecnologia, under project UIDB/50021/2020 and grant number SFRH/BD/149126/2019.

9. References

[1] B. Schuller and A. Batliner, *Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing*. Wiley, November 2013.

[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. Schultz, M. Wand, T. Hueber, K. D. J., 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.

[4] T. Schultz, C. Amma, D. Heger, F. Putze, and M. Wand, "Biosignale-basierte mensch-maschine-schnittstellen," *at - Automatisierungstechnik, 2013*, vol. 61, no. 11, pp. 760 – 769, 2013.

[5] M. Wand, T. Schultz, and J. Schmidhuber, "Domain-adversarial training for session independent emg-based speech recognition." in *Interspeech*, 2018, pp. 3167–3171.

[6] L. Diener, T. Umesh, , and T. Schultz, "Improving fundamental frequency generation in emg-to-speeech conversion using a quantization approach," in *Automatic Speech Recognition and Understanding*, 2019.

[7] T. Hueber, E.-L. Benaroya, G. Chollet, B. Denby, G. Dreyfus, and M. Stone, "Development of a silent speech interface driven by ultrasound and optical images of the tongue and lips," *Speech Communication*, vol. 52, no. 4, pp. 288–300, 2010.

[8] J. A. Gonzalez, L. A. Cheah, J. M. Gilbert, J. Bai, S. R. Ell, P. D. Green, and R. K. Moore, "Direct speech generation for a silent speech interface based on permanent magnet articulography," in *Proceedings of the International Joint Conference on Biomedical Engineering Systems and Technologies*, 2016, pp. 96–105.

[9] Y. Jin, Y. Zhao, C. Huang, and L. Zhao, "Study on the emotion recognition of whispered speech," in *Proc. 2009 WRI Global Congress on Intelligent Systems*, vol. 3. IEEE, 2009, pp. 242–246.

[10] G. Chenghui, Z. Heming, Z. Wei, W. Yanlei, and W. Min, "A preliminary study on emotions of chinese whispered speech," in *Proc. 2009 International Forum on Computer Science-Technology and Applications*, vol. 2. IEEE, 2009, pp. 429–433.

[11] J. Deng, X. Xu, Z. Zhang, S. Frühholz, and B. Schuller, "Exploitation of Phase-based Features for Whispered Speech Emotion Recognition," *IEEE Access*, vol. 4, pp. 4299–4309, July 2016.

[12] B. Cheng and G. Liu, "Emotion recognition from surface emg signal using wavelet transform and neural network," in *Proc. 2nd international conference on bioinformatics and biomedical engineering (ICBBE)*, 2008, pp. 1363–1366.

[13] G. Gibert, M. Pruzinec, T. Schultz, and C. Stevens, "Enhancement of human computer interaction with facial electromyographic sensors," in *23nd conference of the computer-human interaction special interest group of Australia on Computer-human interaction: design (OZCHI 2009), Melbourne, Australia*, 2009, cHISIG.

[14] A. Jones, D. Küster, C. A. Basedow, P. Alves-Oliveira, S. Serholt, H. Hastie, L. J. Corrigan, W. Barendregt, A. Kappas, A. Paiva *et al.*, "Empathic robotic tutors for personalised learning: A multidisciplinary approach," in *International conference on social robotics*. Springer, 2015, pp. 285–295.

[15] 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. [Online]. Available: http://www.csl.uni-bremen.de/CorpusData/download.php?crps=EMG

[16] ELRA Catalogue ID ELRA-S0390, "Parallel EMG-Acoustic English GlobalPhone, ISLRN 910-309-096-5," 2014. [Online]. Available: http://www.islrn.org/resources/910-309-096-523-6/

[17] 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.

[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, nov 2017.

[19] 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 (to appear).

[20] L. Wan, Q. Wang, A. Papir, and I. Lopez Moreno, "Generalized End-to-End Loss for Speaker Verification," *arXiv e-prints*, p. arXiv:1710.10467, Oct. 2017.

[21] S. Amiriparian, M. Freitag, N. Cummins, and B. Schuller, "Sequence to Sequence Autoencoders for Unsupervised Representation Learning from Audio," in *Proc. DCASE 2017*, Munich, Germany, 2017, pp. 17–21.

[22] M. Freitag, S. Amiriparian, S. Pugachevskiy, N. Cummins, and B. Schuller, "auDeep: Unsupervised Learning of Representations from Audio with Deep Recurrent Neural Networks," *Journal of Machine Learning Research*, vol. 18, pp. 1–5, 2018.

[23] S. Amiriparian, "Deep representation learning techniques for audio signal processing," Ph.D. dissertation, Technische Universität München, 2019.