INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Toward Silent Paralinguistics: Speech-to-EMG – Retrieving Articulatory Muscle Activity from Speech

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

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

catarina.t.botelho@tecnico.ulisboa.pt

Abstract

Electromyographic (EMG) signals recorded during speech production encode information on articulatory muscle activity and also on the facial expression of emotion, thus representing a speech-related biosignal with strong potential for paralinguistic applications. In this work, we estimate the electrical activity of the muscles responsible for speech articulation directly from the speech signal. To this end, we first perform a neural conversion of speech features into electromyographic time domain features, and then attempt to retrieve the original EMG signal from the time domain features. We propose a feed forward neural network to address the first step of the problem (speech features to EMG features) and a neural network composed of a convolutional block and a bidirectional long short-term memory block to address the second problem (true EMG features to EMG signal). We observe that four out of the five originally proposed time domain features can be estimated reasonably well from the speech signal. Further, the five time domain features are able to predict the original speech-related EMG signal with a concordance correlation coefficient of 0.663. We further compare our results with the ones achieved on the inverse problem of generating acoustic speech features from EMG features.

Index Terms: Speech, Electromyography (EMG), Silent Computational Paralinguistics, Acoustic-to-Articulatory Inversion.

1. Introduction

Speech production is a complex process that starts in the brain. Following the formulation of the intention of speech, manifested by electrical potentials in the cortex, the signal is conducted through the nervous system to the muscles involved in the speech kinematics – and finally, speech is emitted from the mouth as sound waves. At all these production levels, biosignals can be captured and studied to draw conclusions about linguistic and paralinguistic information of spoken communication [1]. Many researchers have taken advantage of biosignals, proposing systems to generate speech features from Electrocorticography (ECoG) [2] [3] [4], Electroencephalography (EEG) [5] [6], Electromyography (EMG) [7] [8] [9], ultrasound [10] [11], and video recordings of speech articulation [12] [13].

The inverse problem of transforming acoustic speech signals into the underlying biosignals involved in speech production has likewise sparked recent interest. In particular, this concerns the issue of acoustic-to-articulatory inversion (AAI), i. e., the estimation of articulatory movements from the acoustic speech signal. Most AAI works are based on Electromagnetic Articulography (EMA) data (e. g. [14] [15] [16]), and ultrasound imaging

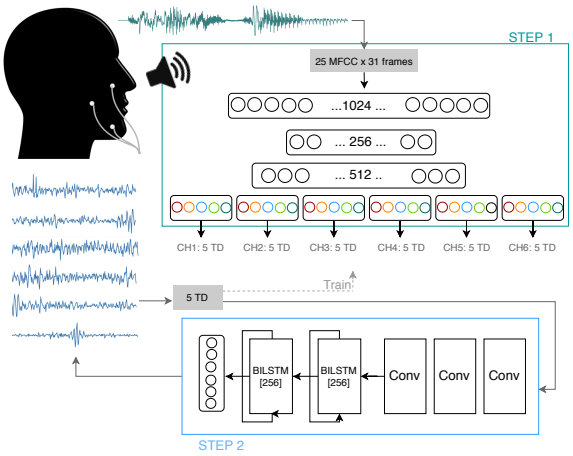


Figure 1: *Two-step Speech-to-EMG system: Acoustic$_{MFCC}$-to-EMG$_{TD}$ (step 1) followed by EMG$_{TD}$-to-EMG$_{Orig}$ (step 2).*

[17]. Krishna et al. [18] converted acoustic features to EEG features, whereas [4] estimates articulatory dynamics from audio recordings, which are included as an explicit intermediate representation in the decoding of speech from EcoG signals.

By examining how speech is converted into other biosignals involved in speech production, work on AAI may help us understand the interconnections between different biosignals and their role in communication. Among these, facial EMG is arguably one of the most important signals involved in speech production, e. g., due to its role for signalling facial expressions of emotion [19] as well as social intentions such as politeness [20]. Furthermore, the synthesis of EMG from acoustic speech shows rich potential – for example in medicine, as a means to increase articulatory awareness in speech therapy, or in computer animation, as a means to visualize realistic muscle movements [21]. AAI research further complements work on Silent Computational Paralinguistics (SCP), i. e., the assessment of speaker states and traits from non-audibly spoken communication [22], by generating large amounts of synthetic EMG data from audio. Thus, future work on EMG-based speech models may require smaller amounts of costly laboratory recordings once important features can be validated against synthetic EMG obtained from AAI.

To the best of our knowledge, this is the first work that proposes a conversion of acoustic speech to EMG signals. Ours is a two-step approach (see Figure 1), parallel to standard speech synthesis methodologies: First, we generate EMG time domain (TD) features, and then derive the EMG signal from those features.

http://dx.doi.org/10.21437/Interspeech.2020-2926

Table 1: *EMG-UKA Corpus. (*) indicates the number of sessions included in the trial corpus.*

Session Type	Number of sessions	Duration [h:m:s]	
		Average	Total
Small	61 (12)	0:03:08	3:11:34
Large	2 (1)	0:27:02	0:54:04

In this initial study, we consider these two steps as independent tasks, and consequently, when synthesizing the EMG signal from the TD EMG features, the TD EMG features considered are the ones obtained from the true EMG signal rather than the ones synthesized in step one. Thus, this second step is designed as a proof-of-concept to validate whether the EMG TD features encode sufficient information to retrieve the original EMG signal. With these two separate steps, we intend to present the basis for a single pipeline that allow the retrieval of the original EMG signal from speech, and also to validate the use of TD EMG features as intermediary representations in future silent paralinguistics systems. For step one, we propose an hourglass-shaped feed forward neural network, while for the second step, we propose a convolutional block followed by a bidirectional long short-term memory (BLSTM) block. For both steps, we use the Concordance Correlation Coefficient (CCC) [23] as the evaluation metric.

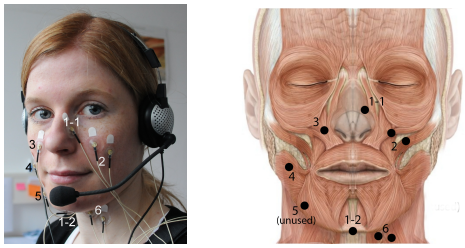
The rest of this paper is organized as follows: Section 2 describes the EMG-UKA corpus used for the experiments. Section 3 presents the methods, including the features extracted from the acoustic and audible EMG signals and the neural network architectures used at steps 1 and 2. The results are presented and discussed in section 4. Section 5 summarizes our main conclusions and directions for future work.

2. Corpora

All experiments in this initial study were performed with the EMG-UKA parallel EMG-Speech corpus [24], [25]. The corpus includes 63 small and large sessions from 8 speakers, in 3 speaking modes (audible, silent speech, and whispered speech). A subset of the sessions is freely available as a trial corpus [24], the full corpus is available from ELRA [25]. In this work, we use the EMG and speech recordings that correspond to the audible speech. Information on the number and duration of the sessions can be found in Table 1. Each speaker has a varying number of sessions: two speakers (speaker 2 and 8) recorded a larger number of sessions (32 and 19, respectively) while the other speakers recorded up to 3 sessions. Further information on the sessions can be found in [24].

The acoustic data was recorded at a sampling rate of 16 kHz with a standard close-talking microphone, whereas the speech-related EMG signals were recorded using a Becker Meditec Varioport amplifier with 6 EMG channels, operating at 600 Hz. The two signals were synchronized via an hardware marker that marks the same point in time, and assuming an electromechanical delay between muscle activation and speech production of 50 ms. Figure 2 shows the positioning of the electrodes, capturing the EMG signal of six articulatory muscles [24]: Zygomaticus major and levator anguli oris (both 2, 3), platysma (4, 5), depressor anguli oris (5), the anterior belly of the digastric (1-2), the tongue (1-2, 6), and a reference channel on the nose (1-1).

Each session is divided in train and test data. The small sessions contain 40 train utterances and 10 test utterances. The large session of speaker 2 contains 500 train utterances and 20 test utterances, and the large session of speaker 8 contain 496 train utterances and 13 test utterances. While the training data partially varies across sessions, the 10 test utterances are unique and the same in all sessions. In the larger sessions, the test set includes repetitions of the 10 test utterances [25].



(a) (b)

Figure 2: *EMG electrode positioning in the EMG-UKA corpus. Electrodes numbered in black in (a) are measured against a reference electrode behind the ear, whereas white numbers indicate bipolar derivation. Muscle chart in (b) adapted from [26].*

3. Methods

We address the conversion of acoustic speech to EMG as a two-step problem. In the first step, called "$\text{Acoustic}_{\text{MFCC}} \rightarrow \text{EMG}_{\text{TD}}$", we convert speech represented by 25 Mel Frequency Cepstral Coefficients (MFCCs) [27] into 5 time domain (TD) EMG features, establishing an approach parallel to the previous work on the generation of speech from EMG signals [7]. In the second step, called "$\text{EMG}_{\text{TD}} \rightarrow \text{EMG}_{\text{Orig}}$", we assess the possibility of retrieving the original EMG signal from these 5 TD features.

EMG signals are speaker dependent, due to varying tissue and skin properties across speakers, as well as due to different muscle and fat proportions, and session dependent e. g., due to small shifts in the electrode positioning. For these reasons, we expect that cross-session experiments perform worse than single-session, as suggested by previous work [9]. Single-session experiments have, on the one hand, less data variability, but, on the other hand, provide a smaller amount of data for model training. Furthermore, single-session models are lacking in generalizability. Our two-step approach enables the use of simpler models in step one (session-dependent problem) and deeper models for step two (session- and channel-independent problem). The train and test partitions described in section 2 are maintained in all experiments. The development set was defined in each experiment, with the same dimension as the test set, as a random subset of the pool of training instances.

3.1. Feature extraction

Both acoustic and EMG signals were windowed using a 32 ms Blackman filter, shifted by 10 ms per step (i. e., an overlap of 22 ms).

We extract 25 MFCCs [27] per audio frame. To introduce some context, we stack this feature vector with a stacking height of 15 frames into the past and future, thus representing each audio frame by a vector of dimension $25 \times 31 = 775$.

The EMG is represented as a series of 5 TD features per EMG frame: low frequency (up to 134 Hz) power, low frequency (up to 134 Hz) mean, high frequency (above 134 Hz) power (HF power), high frequency zero-crossing rate (HF ZCR), and high frequency rectified mean (HF rectified mean).

This TD feature set was originally proposed by [28], [29]

and has been used to convert speech-related EMG to acoustic speech in several works, such as [9], [30], and [7].

3.2. First step: $\text{Acoustic}_{\text{MFCC}} \rightarrow \text{EMG}_{\text{TD}}$

The 775-dimensional vectors representing each audio frame with context were fed into an hourglass-shaped feed forward neural network with three hidden layers [1024, 256, 512], regularized with dropout with p= 0.5 and batch normalization, yielding six output layers (channels) for 5 dimensions (TD features) – see Figure 1. This hourglass shape has been employed in previous works addressing the inverse problem of converting TD EMG features to MFCCs [30]. The network uses rectified linear units (ReLU) as the activation function for the first three layers. The loss function is based on the CCC. I. e., in line with [23], the loss function is computed as follows:

$$loss = \frac{1}{C \times F} \sum_{c=1}^{C} (F - \sum_{f=1}^{F} CCC(y_{cf}, \hat{y}_{cf})), \qquad (1)$$

where F is the number of features, C the number of channels, y and $\hat{y}$ are the target and predicted value. The learning rate resembles 0.002, the batch size 32, and the model was trained for 50 epochs with an Adam optimizer. As the EMG signal is session and speaker dependent, we defined three sets of experiments:

1. **Single session.** The model is trained and tested with data from the same session. We perform two single session experiments, with the two large sessions available in the corpus, which belong to speakers 2 and 8.
2. **Multi-session.** Training and testing in leave-one-session-out cross-validation setting: for each speaker, the model is trained with all training utterances of all sessions but one, and tested with the test utterances of the left-out session. We perform two multi-session experiments, with speakers 2 and 8 (speakers with a larger number of sessions).
3. **Multi-speaker.** Training and testing in leave-one-speaker-out cross-validation setting, i. e., models are trained on all training utterances of all speakers but one, and tested on all test utterances of the left-out speaker. We repeat this for the 8 speakers. In these experiments, there is no session nor speaker overlap between train and test folds.

3.3. Second step: $\text{EMG}_{\text{TD}} \rightarrow \text{EMG}_{\text{Orig}}$

In this initial study, the second step consists of synthesizing the original EMG signal from the 5 EMG TD features directly extracted from the EMG signal. Being able to retrieve the EMG signal from the TD EMG features justifies their usage as intermediary representations in future silent paralinguistics works. Furthermore, while we consider the reliable generation of EMG TD features an important proof-of-concept, the generation of the original EMG signal opens up new avenues for research and understanding of EMG signals related to spoken communication, and for silent speech, which cannot be captured acoustically.

A schematic representation of the network used can be found at the bottom of Figure 1. It consists of a convolutional-BLSTM neural network, similar to what has been proposed for other paralinguistic tasks, such as detection of emotions [31] and breathing patterns from speech [32]. The neural model includes one convolutional block, one BLSTM block, and one output linear layer. The convolutional block includes three 1D convolutional layers [128, 256, 512], kernel size [5, 3, 3], stride 1, padding to keep time dimension constant, no pooling, $tanh$ as activation function, and batch normalization. The BLSTM block includes two BLSTM layers with hidden layer size 256 and dropout with a probability of 0.4, followed by batch normalization. Finally, the output linear layer has a dimension of 6 to match the sampling frequency of the EMG signal (600 Hz) and a frame shift of 10 ms used to compute the features. Each utterance is fed into the network as a tensor of dimension $f \times t$, where f is the number of TD features, and t is the number of frames in the signal. The f dimension is fed as channels for the first 1D convolutional layer, and the convolution occurs across the time dimension. The convolutional and BLSTM blocks perform a feature mapping in the sense that they keep the time dimension of the tensor constant. At the linear output layer, the time dimension is mapped to match the EMG sampling frequency. The learning rate resembles 0.001, the batch size 5, and the learning rate decay 0.1 with a period of 20 epochs. We train the model up to 50 epochs with early stopping depending on the performance on the development set.

The extraction of TD EMG features from the raw EMG signal is independent of the speaker, the session and the electrode positions (channels), thus, when retrieving the raw EMG from TD features using a neural network, it is not relevant to distinguish between speakers, sessions nor electrode positions. Therefore, the data used to train and evaluate the model includes all channels from all speakers and all sessions. This results in a training set, a development set, and a test set composed of 2793, 643, and 643 utterances.

4. Results

This section describes and discusses the results generated from the experiments described above. All results were evaluated using CCC. CCC assumes values in $[-1, 1]$, where a coefficient of 0 reflects no correlation between the true and the predicted values, a coefficient of 1 reflects perfect agreement and a coefficient of -1 reflects perfect reversed agreement. In addition, CCC also reflects the absolute correctness rather than only a relative one. No other metrics are reported for the sake of space and conciseness. We chose CCC over other standard metrics, such as mean squared error, because CCC assumes values in a bounded interval, easier to interpret when no previous baselines are available.

4.1. $\text{Acoustic}_{\text{MFCC}} \rightarrow \text{EMG}_{\text{TD}}$: session and speaker dependencies

The $\text{Acoustic}_{\text{MFCC}} \rightarrow \text{EMG}_{\text{TD}}$ results at the single-session (speaker 8), multi-session (speaker 8), and multi-speaker experiments are detailed in Figure 3. The single-session results (Figure 3, on the left) appear promising. We achieve a CCC value of 0.54 for speaker 2, and 0.63 for speaker 8, when averaging all feature scores across all channels. These results increase to 0.64 and 0.75 if the HF ZCR feature is excluded.

The results obtained at the multi-session and multi-speaker (Figure 3 – center and right) experiments appear worse when compared to the single-session experiment. The average CCC for all the features and channels for the multi-session experiments is 0.50 and 0.57, respectively, for speakers 2 and 8, while for the multi-speaker experiment, the CCC is 0.46. The average CCCs improves to 0.59, 0.66 and 0.55 when excluding HF ZCR.

The multi-session and multi-speaker experiments were evaluated in a cross validation setting. The aforementioned figure shows the mean and standard deviation of CCC for each feature at each channel, obtained for all of the folds. We find that there is some variation of the results for the different folds. These results

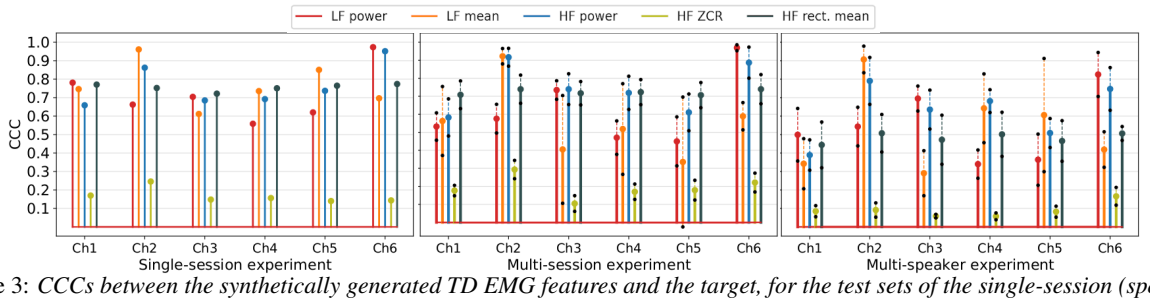
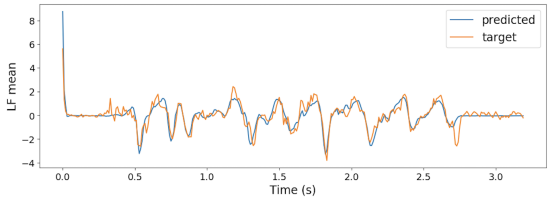
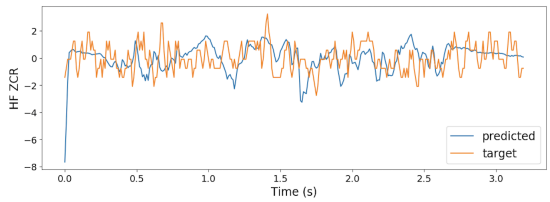


Figure 3: *CCCs between the synthetically generated TD EMG features and the target, for the test sets of the single-session (speaker 8), multi-session (speaker 8), and multi-speaker experiments. The black dots represent the mean ± standard deviation obtained for the different sessions in the cross validation experiments.*



(a) *LF mean. CCC* = 0.870



(b) *HF ZCR. CCC* = 0.182

Figure 4: *Examples of the target and predicted features LF mean and HF ZCR (single-session, speaker 8)*

support the notion that EMG signals recorded in this type of setting may be strongly session dependent, likely due to small shifts in electrode positioning. The multi-speaker results appear worse than the multi-session results, although it is not evident whether this is caused by physiological differences between speakers (e. g., fat, muscles, skin), or a result of an increasing variability in electrode position (due to an increased number of sessions, and different number of sessions per speaker) which may hinder the learning ability of the system.

4.2. $\text{Acoustic}_{\text{MFCC}} \rightarrow \text{EMG}_{\text{TD}}$: feature analysis

Figure 3 suggests that HF ZCR is much harder to predict than the rest of the TD features in all the experiments. Figure 4 presents an example of the target and the predicted LF mean and HF ZCR of channel 0, obtained in the single-session experiment with speaker 8, to illustrate the meaning of the different CCCs.

The CCCs achieved at the single session experiments for LF mean, LF power, HF rectified mean, and HF power are above 0.5 for all channels. These results are at a comparable level with the prediction of the first three MFCCs when converting audible EMG to acoustic speech [22]. For the remaining MFCCs, the results in the speech-to-EMG direction are much better than the results in the direction EMG-to-speech.

4.3. $\text{EMG}_{\text{TD}} \rightarrow \text{EMG}_{\text{Orig}}$

We use the five true TD features to generate the original EMG signal. Figure 5 shows an example of the predicted and target signals, suggesting a reasonable match.

As HF ZCR appeared to be harder to predict than the remaining features, we also generate the speech-related EMG signal based on the remaining four TD features. Table 2 shows that although the CCCs obtained with both feature sets are very satisfactory, the results are better in the presence of ZCR. Thus, we conclude that ZCR contains relevant information for the generation of EMG_{Orig}.

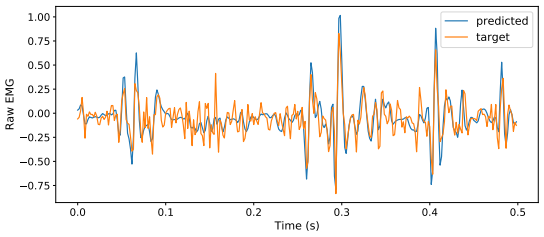


Figure 5: *EMG signal generated from TD features.*

Table 2: *CCC between the target and the predicted EMG signal, using either four or five TD features.*

Input Features	CCC
LF mean, LF power, HF rectified mean, HF power, HF ZCR	0.663
LF mean, LF power, HF rectified mean, HF power	0.602

5. Conclusions

We presented initial results on a novel two-step approach to generate speech-related EMG signals from acoustic speech. In the first step, we successfully converted MFCCs into TD EMG features. Thus, we established the foundation of the EMG-to-speech approach - i. e., the inverse of prior works that have aimed to generate speech from EMG data. The CCCs achieved in the single session experiments for the prediction of LF mean, LF power, HF rectified mean, and HF power were comparable to the first three MFCCs generated from audible EMG in our recent work on EMG-based SCP [22]. Multi-session and multi-speaker experiments, although performing worse than the single-session experiments, still achieved satisfactory results. We expect that this may be improved with deeper and more complex models when more data is available. In the second step, we generated a signal that follows reasonably the true EMG signal, using the TD features. As future work, we plan to integrate both speech-to-EMG and EMG-to-speech in one system, as well as retrieve paralinguistic information from the generated EMG signals.

6. Acknowledgements

This work was 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.

7. References

[1] 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, 2017. [Online]. Available: https://www.csl.uni-bremen.de/cms/images/documents/publications/TASLP-2017-biosignal-based-spoken.pdf

[2] C. Herff, G. Johnson, L. Diener, J. Shih, D. Krusienski, and T. Schultz, "Towards direct speech synthesis from ECoG: A pilot study," in *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 1540–1543.

[3] M. Angrick, C. Herff, E. Mugler, M. C. Tate, M. W. Slutzky, D. J. Krusienski, and T. Schultz, "Speech synthesis from ECoG using densely connected 3D convolutional neural networks," *Journal of neural engineering*, vol. 16, no. 3, p. 036019, 2019.

[4] G. K. Anumanchipalli, J. Chartier, and E. F. Chang, "Speech synthesis from neural decoding of spoken sentences," *Nature*, vol. 568, no. 7753, pp. 493–498, 2019.

[5] G. Krishna, Y. Han, C. Tran, M. Carnahan, and A. H. Tewfik, "State-of-the-art speech recognition using EEG and towards decoding of speech spectrum from EEG," *arXiv preprint arXiv:1908.05743*, 2019.

[6] G. Krishna, C. Tran, Y. Han, and M. Carnahan, "Speech synthesis using EEG," *arXiv preprint arXiv:2002.12756*, 2020.

[7] 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.

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

[9] L. Diener, G. Felsch, M. Angrick, and T. Schultz, "Session-independent array-based EMG-to-speech conversion using convolutional neural networks," in *Speech Communication; 13th ITG-Symposium*. VDE, 2018, pp. 1–5.

[10] B. Denby and M. Stone, "Speech synthesis from real time ultrasound images of the tongue," in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. I–685.

[11] N. Kimura, M. Kono, and J. Rekimoto, "SottoVoce: an ultrasound imaging-based silent speech interaction using deep neural networks," in *2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–11.

[12] K. Vougioukas, P. Ma, S. Petridis, and M. Pantic, "Video-driven speech reconstruction using generative adversarial networks," *arXiv preprint arXiv:1906.06301*, 2019.

[13] D. Michelsanti, O. Slizovskaia, G. Haro, E. Gómez, Z.-H. Tan, and J. Jensen, "Vocoder-based speech synthesis from silent videos," *arXiv preprint arXiv:2004.02541*, 2020.

[14] A. Illa and P. K. Ghosh, "Low resource acoustic-to-articulatory inversion using bi-directional long short term memory." in *Interspeech*, 2018, pp. 3122–3126.

[15] A. Illa and P. Ghosh, "Representation learning using convolution neural network for acoustic-to-articulatory inversion." in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5931–5935.

[16] P. Liu, Q. Yu, Z. Wu, S. Kang, H. Meng, and L. Cai, "A deep recurrent approach for acoustic-to-articulatory inversion," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4450–4454.

[17] D. Porras, A. Sepúlveda-Sepúlveda, and T. G. Csapó, "DNN-based acoustic-to-articulatory inversion using ultrasound tongue imaging," in *2019 IEEE International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8.

[18] G. Krishna, C. Tran, M. Carnahan, Y. Han, and A. H. Tewfik, "Generating EEG features from acoustic features," *arXiv preprint arXiv:2003.00007*, 2020.

[19] A. Kappas, E. Krumhuber, and D. Küster, "Facial behavior," in *In: Hall, Judith A.; Knapp, Mark L. (Ed.), Nonverbal communication (pp. 131-166). Berlin: de Gruyter, 2013*. de Gruyter, pp. 131–166.

[20] D. Küster, "Hidden tears and scrambled joy: On the adaptive costs of unguarded nonverbal social signals," in *Social Intelligence and Nonverbal Communication*. Springer, 2020, pp. 283–304.

[21] M. Sagar and R. Scott, "System and method for tracking facial muscle and eye motion for computer graphics animation," Jun. 30 2009, uS Patent 7,554,549.

[22] 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.

[23] L. Lin, "A concordance correlation coefficient to evaluate reproducibility," *Biometrics*, pp. 255–268, 1989.

[24] M. Wand, M. Janke, and T. Schultz, "The EMG-UKA corpus for electromyographic speech processing," in *Interspeech*, 2014. [Online]. Available: trial data at http://www.csl.uni-bremen.de/CorpusData/download.php?crps=EMG

[25] 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/

[26] M. Schünke, E. Schulte, and U. Schumacher, *Prometheus-Lernatlas der Anatomie*. Stuttgart, New York: Thieme Verlag, 2006.

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

[28] 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.

[29] S.-C. S. Jou, "Automatic speech recognition on vibrocervigraphic and electromyographic signals," Ph.D. dissertation, Carnegie Mellon University, Language Technologies Institute, 2008.

[30] 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 (IJCNN)*, 2015, pp. 1–7.

[31] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network," in *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 5200–5204.

[32] B. W. Schuller, A. Batliner, C. Bergler, E.-M. Messner, A. Hamilton, S. Amiriparian, A. Baird, G. Rizos, M. Schmitt, L. Stappen *et al.*, "The Interspeech 2020 computational paralinguistics challenge: elderly emotion, breathing & masks," in *Interspeech*, 2020.