INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

STE-GAN: Speech-to-Electromyography Signal Conversion using Generative Adversarial Networks

Kevin Scheck, Tanja Schultz

Cognitive Systems Lab, University of Bremen, Bremen, Germany
scheck@uni-bremen.de, tanja.schultz@uni-bremen.de

Abstract

With Speech-to-Electromyography Generative Adversarial Network (STE-GAN), we propose a model which can synthesize Electromyography (EMG) signals from acoustic speech. We condition the generator network on representations of the spoken content obtained from a voice conversion model. Given these representations, the generator outputs an EMG signal corresponding to the articulated content of the acoustic speech in the setting of a specific EMG recording session. In comparison to previous work, STE-GAN directly generates EMG signals from acoustic speech. As it uses more speaker-independent content representations as input, it can synthesize EMG signals from speech of speakers who were unseen during training.

Index Terms: electromyography, acoustic to articulatory inversion, silent speech interfaces, generative adversarial networks

1. Introduction

Surface Electromyography (EMG) signals of articulatory muscles reflect the speech production process [1]. As such, they are a biosignal of interest for *Silent Speech Interfaces* (SSIs) [2], which aim to enable speech communication without depending on acoustic speech. The driving force of EMG-based SSIs are EMG-to-Speech (ETS) models, which convert EMG signals to the corresponding acoustic speech signal [3, 4, 5, 6]. Recent work [7, 8] started to examine the inverse problem: Speech-to-EMG (STE) i.e. predicting EMG signals from acoustic speech. STE is related to acoustic-to-articulatory inversion, in which acoustic speech is used to predict articulator trajectories, captured by, e.g., Electromagnetic Articulography (EMA) [9], X-ray microbeam [10], or ultrasound imaging [11]. Potential applications of STE lie in speech therapy, as the activity of articulatory muscles could be predicted without EMG equipment. STE could also be explored to generate new, artificial EMG signals to improve ETS model training. These downstream tasks require that STE models can synthesize EMG signals from speech and speakers which were unseen during training. The first studies on STE [7, 8] evaluated the prediction of EMG features from acoustics, and the generation of EMG signals from their Ground-Truth (GT) EMG features as independent problems. As such, they did not investigate the direct prediction of EMG signals from speech.

To address this challenge, we propose Speech-to-EMG Generative Adversarial Network (STE-GAN). It directly converts acoustic speech to corresponding EMG signals (see Fig. 1). We base our system on neural vocoders [12, 13] and adjust the architecture and losses for EMG generation. We condition the EMG generator on speech content representations extracted by Voice Conversion (VC) models [14]. In comparison to acoustic features such as mel-spectrograms, VC content representations are designed to discard speaker information. As such, STE-GAN could be more applicable to generate EMG signals from speech of unseen speakers for downstream tasks. For STE-GAN, we use the soft Speech Units (SUs) of van Niekerk et al. [14], because soft SUs can be predicted by ETS models [6]. As EMG signals vary between recording sessions, we additionally condition the generator network on learned session embeddings. To generate realistic EMG signals, STE-GAN minimizes Time-Domain (TD) feature [15] differences between real and fake signals. Lastly, we use the gradients of a pre-trained EMG encoder during training [16]. We evaluate our model on two EMG data sets and generate EMG signals from speech inputs of unseen speakers using the LibriTTS data set [17]. We compare STE-GAN with systems based on previous work [8] which first predict EMG features from acoustic speech and then predict EMG signals.

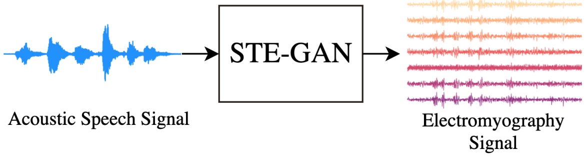

Figure 1: *STE-GAN converts acoustic speech to corresponding multi-channel electromyography signals.*

2. Related Work

GANs [18] have been utilized to generate various biosignals, such as EMG [19, 20, 21], Electrencephalography (EEG) [22, 23], Electrocardiography (ECG) [24, 25], or EMA [16]. The SynSigGAN model of Hazra and Byun [19] is able to generate a variety of biosignals, such EEG, ECG or EMG. Chen et al. [20] train a Deep Convolutional GAN (DCGAN) to synthesize EMG features of hand gestures. Zanini and Colombini [21] use a modified DCGAN model to generate EMG signals from Parkison's disease patients. To the best of our knowledge, no prior work has used GANs for STE. An early study on STE is the work of Botelho et al. [7]. The authors first train a Neural Network (NN) to predict EMG TD features from Mel-Frequency Cepstral Coefficients (MFCCs). Then, they train a Convolutional Neural Network (CNN) - Long Short-Term Memory (LSTM) network to predict EMG signals from GT TD features. Sharma et al. [8] evaluate additional features and models for this task. The authors use a 5-layer Bidirectional LSTM (BLSTM) for mapping MFCCs to EMG features. For predicting EMG signals from GT EMG features, they find that adding Hilbert envelope features increases the generation quality.

10.21437/Interspeech.2023-174

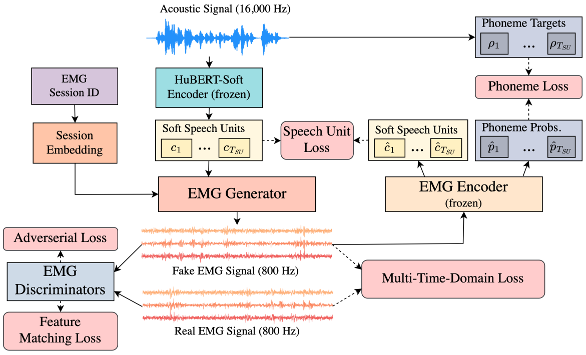

Figure 2: *Training of the STE-GAN model.*

3. Speech-to-Electromyography GAN

3.1. Network architectures

We base STE-GAN on neural vocoders, in particular HiFi-GAN [12] and CARGAN [13]. We modify the network architectures for processing multi-channel EMG signals at a sample rate of 800 Hz (see Fig. 2). The network architectures are available in the code repository of STE-GAN[1].

EMG Generator The generator G outputs the C-channel EMG signal given a sequence of speech features and the session identifier of the EMG signal (see Fig. 2). For speech features, we use the "HuBERT-Soft" encoder [14] to extract soft SUs at a frame rate of 50 Hz. We combine the SUs with a learnable, 64-dimensional session embedding by feature vector concatenation. The resulting input feature sequence is processed by a modified GAN-TTS architecture [26]. It comprises eight *Gblocks*, of which the four middle blocks upsample the feature sequence by a factor of 2 respectively. The output layer is an 1D convolution layer with a *tanh* activation function and outputs a C-channel EMG signal at 800 Hz.

EMG Discriminators The discriminator networks process the C-channel EMG signals at 800 Hz. We use modified Multi-Scale Discriminator (MSD) and Multi-Period Discriminator (MPD) architectures of HiFi-GAN [12]. We lower the number of layers and kernel sizes to approximate the original receptive field of the discriminators at the lower EMG sample rate. We use 3 MSDs, processing EMG signals at different temporal resolutions, and 5 MPDs with periods 2, 3, 5, 7, and 11.

EMG Encoder Similar to the ArticulationGAN [16], we incorporate a model which was trained on the same modality as the generator output. We use a pre-trained EMG encoder E which predicts soft SUs from the EMG signal. As such it performs the inverse prediction as the EMG generator G. Additionally, E predicts framewise phoneme classes. We use the network architecture of Gaddy and Klein [4] which was modified for soft SU prediction [6]. Convolutional layers first downsample the EMG signal to features with a 50 Hz frame rate. A transformer encoder then processes the resulting feature sequence. The model has two output layers which predict framewise soft SUs and phonemes respectively.

3.2. Training losses

For training the STE-GAN, we use a subset of the losses of HiFi-GAN [12]. In particular, we minimize the Least-Squares GAN [27] adversarial losses $\mathcal{L}_{Adv}$. We further use the feature matching loss $\mathcal{L}_{FM}$, in which the L1 loss of the discriminator activations between real and fake inputs is minimized. We refer to Kong et al. [12] for further details on $\mathcal{L}_{Adv}$ and $\mathcal{L}_{FM}$. STE-GAN uses the following additional losses (see Fig. 2).

Speech Unit Loss We minimize the Euclidian distance between soft SUs predicted by the EMG encoder using the generator output, and the GT soft SUs from the speech input [6]:

$$\mathcal{L}_{SU}(G) = \frac{1}{T_{SU}} \sum_{t=1}^{T_{SU}} ||\mathbf{c}_t - E_c(G(\mathbf{c}, \mathbf{s}))_t||_2 \tag{1}$$

where $\mathbf{c}$ denotes the GT soft SU target sequence and $\mathbf{s}$ is the session embedding of the EMG signal. $E_c(G(\mathbf{c}, \mathbf{s}))$ denotes the soft SU output of E using the EMG synthesis of G given $\mathbf{c}$ and $\mathbf{s}$. During training, E is frozen and only G is updated.

Phoneme Loss We minimize the cross-entropy loss between phonemes predicted by E, given the fake EMG signal, and the GT phoneme sequence ρ as phoneme loss [4, 6].

$$\mathcal{L}_P(G) = -\frac{1}{T_{SU}} \sum_{t=1}^{T_{SU}} \sum_{i=1}^{|\mathcal{P}|} \mathbf{b}_{t,i} \cdot log\ E_p(G(\mathbf{c}, \mathbf{s}))_{t,i} \tag{2}$$

where $\mathcal{P}$ is the set of phonemes and $\mathbf{b}_{t,i}$ indicates whether phoneme i at frame index t is the target phoneme ρ_t. $E_p(G(\mathbf{c}, \mathbf{s}))_{t,i}$ denotes the probability for phoneme i at frame t predicted by the EMG encoder given the generator output.

Multi-Time-Domain Loss TD features of the EMG signal have been used in various work on ETS [3, 28, 29] and were used as model inputs for EMG signal generation in previous work [7, 8]. We therefore minimize the difference between TD features of original and generated EMG signals as loss function. We use the TD feature implementation of Jou et al. [15] and apply a 9-point double-averaging filter on the EMG signal $\mathbf{x}$ to obtain a low-frequency signal $\mathbf{w}$. We then compute a rectified, high-frequency residual signal $\mathbf{r} = |\mathbf{x} - \mathbf{w}|$. The used TD feature sets then comprise the mean and power values of windowed frames of $\mathbf{w}$ and $\mathbf{r}$ respectively. For the Multi-Time-Domain (MTD) loss, we use K_{TD} different TD feature implementations with varying window sizes and shifts to evaluate the EMG signal at multiple resolutions. We minimize the sum of mean L1 distances between TD vectors for K_{TD} TD sets:

$$\mathcal{L}_{TD}(G) = \sum_{k=1}^{K_{TD}} \frac{1}{T_{TD}^{(k)}} \sum_{j=1}^{T_{TD}^{(k)}} ||TD_j^{(k)}(\mathbf{x}) - TD_j^{(k)}(G(\mathbf{c}, \mathbf{s}))||_1 \tag{3}$$

where $\mathbf{x}$ denotes the real EMG signal. $T_{TD}^{(k)}$ is the number of windows of the EMG signals using the kth TD feature implementation. $TD_j^{(k)}$ denotes the TD feature vector of the jth signal window using the kth TD implementation. For our experiments, we use three TD feature sets with window sizes / shifts of 25 / 10, 64 / 16, and 100 / 25 milliseconds respectively.

Total Loss The total loss of the EMG generators and discriminators is:

$$\begin{aligned} \mathcal{L}_G &= \sum_{k=1}^{K} [\mathcal{L}_{Adv}(G; D_k) + \lambda_{FM} \mathcal{L}_{FM}(G; D_k)] \\ &\quad + \lambda_{SU} \mathcal{L}_{SU}(G) + \lambda_P \mathcal{L}_P(G) + \lambda_{TD} \mathcal{L}_{TD}(G) \end{aligned} \tag{4}$$

$$\mathcal{L}_D = \sum_{k=1}^{K} \mathcal{L}_{Adv}(D_k; G) \tag{5}$$

where K denotes the number of discriminators. We scale the feature matching, speech unit, phoneme, and MTD losses with scalar weights respectively.

[1] https://github.com/scheck-k/ste-gan

4. Experiment Setup

4.1. EMG data sets

We evaluate STE-GAN with two data sets. First, we use the corpus of Gaddy and Klein [29] in the open vocabulary condition. One subject reads English sentences in audible and silent articulation respectively. 8-channel EMG signals are recorded at a 1000 Hz sample rate. We ignore utterances of silent articulation, since no parallel GT audio is available. We use the same EMG filtering steps as in the authors' implementation[2]. We downsample the EMG signals to 800 Hz and apply the *tanh* function, such that the amplitude range is -1 to 1. We use the pre-defined validation and testing splits, but use utterances with EMG signals of audible articulation. We ignore samples which transcriptions contain no alphanumerical characters. The train, validation, and test splits contain 6755, 199, and 98 utterances.

We further evaluate our model on the ”EMG-ArraySingle-A-500+” data set [30], which contains EMG signals recorded during normal articulation of English sentences. Subjects wore multi-array electrodes on their cheek and chin. We focus on the largest session ”S3-Array-Lrg”, for which we use a pre-defined channel set of 15 channels. We first process the EMG signals with a notch filter at 50 Hz and a bandpass filter with cut-offs at 10 Hz and 400 Hz. We downsample the EMG signals to 800 Hz and z-normalize each channel with utterance-level statistics. Lastly, we apply amplitude downscaling by a factor of 10 and the *tanh* function. We use the pre-defined train, validation, and test splits, which contain 1771, 196, and 40 utterances.

4.2. Training settings and implementation

We base our implementation on the CARGAN [13] repository[3]. We use the ”HuBERT-Soft” model of van Niekerk et al. [14] to extract soft SUs from acoustic speech[4]. The EMG encoder is pre-trained on the respective training sets as outlined by Scheck and Schultz [6]. We train the STE-GAN with the AdamW optimizer [31] with a learning rate of 2e$-$4 and a batch size of 32. We set the loss weights, based on values of CARGAN, to $\lambda_{FM} = 7.0$, $\lambda_{TD} = 15.0$, $\lambda_{SU} = 1.0$, and $\lambda_{P} = 1.0$. We found that these values give an adequate performance. As such, instead of a hyper-parameter search, we focus on ablation studies to investigate the impact of each individual loss by setting loss weights to 0.0. During training, we randomly slice 2048 EMG samples and ignore training utterances with less samples. We initialize the EMG generator and discriminators from scratch and train the models for $25k$ steps. When no losses of the EMG encoder are optimized, we train up to $250k$ steps, as validation metrics require more steps to improve. We use an Intel(R) Xeon(R) Gold 5118 CPU and an NVIDIA GeForce RTX 2080 Ti GPU. The time for $25k$ steps is approx. 4 hours.

4.3. Baselines for comparison

We first compare the STE-GAN with the two-step system of Sharma et al. [8]. It consists of two models: First, a BLSTM predicts EMG features of all channels from MFCCs. Subsequently, a CNN-BLSTM predicts the EMG signal of a single channel from its GT EMG features. We additionally condition the first model on the target session with a 64-dimensional session embedding by feature concatenation. We set the output size of the CNN-BLSTM to 8 to produce 800 Hz signals. Lastly, we increase the dimensionality of the first BLSTM model to 384 such that the number of parameters of both models approximates the size of the EMG generator (see Table 1). As Sharma et al. [8], we train the models individually using the Mean Squared Error (MSE) loss. To compare the system with STE-GAN, we use the predictions of the first model as inputs for the second model. We also evaluate STE-GAN variants which use similar approaches. *STE-GAN (MFCCs+MSE)* uses MFCCs as input and is solely trained with the MSE loss applied on EMG signals. *STE-GAN (MSE)* is also trained with the MSE loss, but takes soft SUs as input. *STE-GAN (MFCCs)* uses MFCCs, but is trained with the same losses as the proposed *STE-GAN (SU)*.

Table 1: *Evaluated models and their parameter count in million.*

Model	Param. (m)
EMG Generator	23.50
EMG Discriminators (total)	11.90
EMG Encoder	58.50
Two-Step: BLSTM (MFCC → EMG Feat.)	15.70
Two-Step: CNN-LSTM (EMG Feat. → EMG Signal)	8.60

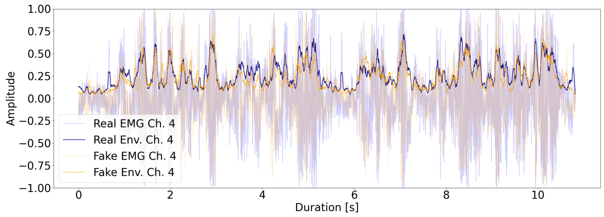

Figure 3: *The real and fake EMG signal, generated by STE-GAN, of an utterance of the Gaddy and Klein corpus. The navy and dotted, orange lines are the envelopes (Env. CC* = 0.80*).*

4.4. Evaluation metrics

For measuring the similarity between real and fake EMG signals, we compare their envelopes, as the amplitude of raw EMG signals is quasi-random [32]. As Zanini et al. [21], we compute the envelopes by rectifying the signals and applying an average filter with a window size of 50 ms (see Fig. 3). We then measure the mean Envelope Correlation Coefficient (Env. CC). We furthermore use the pre-trained EMG encoder to assess whether the generator synthesizes EMG signals which reflect the speech input. As Speech Unit Distance (SU Dist.), we compute the mean Euclidian distance between GT soft SUs of the audio input and the soft SUs predicted by the EMG encoder from generated EMG signals. We compute averaged utterance-level, framewise phoneme accuracies, excluding silence, using EMG encoder phoneme predictions on fake EMG signals (Phon. Acc.). We furthermore convert soft SUs predicted from fake EMG back to acoustic speech using the acoustic model and vocoder of van Niekerk et al. [14][4]. We transcribe the speech synthesis with the Whisper [33] ”medium.en” model and calculate the Word Error Rate (WER). To evaluate whether models can generate EMG signals from speech of unseen speakers, we generate EMG signals from 200 utterances of the LibriTTS [17] ”test-clean” split, uniformly distributed between its 39 speakers.

[2] `https://github.com/dgaddy/silent_speech` (Commit `f12495e`)

[3] `https://github.com/descriptinc/cargan` (Commit: `61051fa`)

[4] `https://github.com/bshall/soft-vc` (Commit `eb21417`)

Table 2: *Results obtained by STE models using audio utterances of the EMG test sets and 200 utterances of LibriTTS.*

Model / EMG Data	**EMG Test Audio (1 Speaker)**				**LibriTTS Audio (39 Speakers)**		
	Env. CC	**Phon. Acc.**	**SU Dist.**	**WER**	**Phon. Acc.**	**SU Dist.**	**WER**
Gaddy and Klein [29]							
Two-Step: BLSTM → CNN-BLSTM	0.46	0.00	9.37	106.37	0.00	10.41	108.09
STE-GAN (MFCCs + MSE)	0.55	2.66	7.96	99.94	3.85	8.52	98.42
STE-GAN (MSE)	0.60	1.67	8.27	100.49	3.74	8.58	99.71
STE-GAN (MFCCs)	0.62	77.24	2.28	21.38	37.99	5.13	87.91
STE-GAN (SU)	**0.66**	**82.65**	**1.87**	**10.29**	**76.86**	**2.63**	**12.28**
S3-Array-Lrg [30]							
Two-Step: BLSTM → CNN-BLSTM	0.38	0.33	7.65	125.13	0.00	7.82	125.22
STE-GAN (MFCCs + MSE)	0.56	36.23	5.16	103.77	7.88	7.56	104.50
STE-GAN (MSE)	0.61	45.01	4.64	98.74	33.28	5.78	104.82
STE-GAN (MFCCs)	0.63	79.25	2.17	8.04	20.77	6.26	99.95
STE-GAN (SU)	**0.66**	**81.73**	**1.88**	**6.53**	**65.01**	**3.12**	**21.71**

5. Results

5.1. EMG signal reconstruction quality

Table 2 lists the obtained results of the evaluated models on the respective data test splits. The *Two-Step* model achieves worse evaluation results compared to all STE-GAN models. A possible reason for this could be that some TD features, such as the zero-crossing rate, are challenging to predict from acoustic speech, but provide important information for EMG signal reconstruction [7]. The most similar model *STE-GAN (MFCCs + MSE)* achieves better results, which could indicate that STE models benefit from end-to-end training. STE-GAN models trained with the proposed losses, *STE-GAN (MFCCs)* and *STE-GAN (SUs)*, outperform models trained with the MSE loss on all metrics. The proposed model *STE-GAN (SUs)* achieves the overall best results. In particular, it achieves the highest Env. CC of 0.66 for both data sets. It further obtains a Phon. Acc. of over 80% for generated EMG using audio of the EMG test splits. This indicates that the EMG generator can synthesize EMG signals which reflect the articulated content of the speech input.

5.2. EMG generation from speech of unseen speakers

While *STE-GAN (MFCCs)* and *STE-GAN (SU)* both perform well for the EMG test splits when compared to other models, *STE-GAN (SU)* outperforms *STE-GAN (MFCCs)* for LibriTTS in the multi-speaker setup. Using MFCCs instead of SUs leads to worse results in this setting. For instance, the WER of *STE-GAN (MFCCs)*, trained on S3-Array-Lrg, increases from approx. 8% to 100%. In comparison, the WER increase for *STE-GAN (SU)* is from 7% to 22% for S3-Array-Lrg. For *STE-GAN (SU)* trained on the Gaddy and Klein corpus, the WER increases by only 2%. As such, using soft SUs as input enables the EMG generator to generalize to speech of unseen speakers.

5.3. Impact of loss functions

Table 3 lists the results obtained by STE-GAN models which were trained with subsets of the proposed losses. Removing the MTD loss leads to the largest decrease in Env. CC, however improves most other metrics. Removing the phoneme loss does not lead to a consistently lower performance when comparing to the full STE-GAN model. However, removing both phoneme and SU loss leads to a worsening in all metrics. In this setting, the WER also increases from approx. 10% to 19% for the Gaddy and Klein data and from 7% to 81% for S3-Array-Lrg. A possible reason for the difference in WER increase between data sets could be their number of training utterances. For the smaller data set S3-Array-Lrg, incorporating at least one loss involving the EMG encoder is required for adequate performance. For the larger Gaddy and Klein corpus, the EMG generator learns to predict EMG signals reflecting the speech properties without using the EMG encoder during training.

Table 3: *Ablation study results for the EMG data test split.*

Model / EMG Data	**Env. CC**	**Phon. Acc.**	**SU Dist.**	**WER**
Gaddy and Klein [29]				
STE-GAN (SU)	0.66	82.65	1.87	10.29
- MTD Loss	0.63	**84.36**	**1.62**	**7.17**
- Phoneme Loss	**0.67**	82.72	1.85	10.78
- SU Loss	0.66	81.96	2.34	11.09
- SU & Phoneme Loss	0.65	75.76	2.60	19.18
S3-Array-Lrg [30]				
STE-GAN (SU)	**0.66**	**81.73**	1.88	6.53
- MTD Loss	0.60	81.46	**1.84**	**5.03**
- Phoneme Loss	**0.66**	81.07	**1.84**	5.28
- SU Loss	**0.66**	77.92	2.47	14.32
- SU & Phoneme Loss	0.62	51.53	4.24	80.65

6. Conclusions

We have proposed Speech-to-Electromyography Generative Adversarial Network (STE-GAN), a model which generates EMG signals from acoustic speech signals. In comparison to previous work, it performs the mapping in an end-to-end fashion i.e. it converts acoustic signals to the EMG signal without the need to first predict intermediate EMG features. Furthermore, since it uses soft speech units as inputs, it can create EMG signals from speech of multiple speakers who are not included in the training set. Since STE-GAN is not yet applicable to online processing, the practical use is still limited. We therefore aim to develop a real-time capable version of the model and evaluate it on EMG data sets with multiple speakers.

7. Acknowledgements

This research is partially funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through the project ”MyVoice: Myoelectric Vocal Interaction and Communication Engine” (460884988).

8. 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, 2017.

[2] 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.

[3] 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.

[4] D. Gaddy and D. Klein, “An Improved Model for Voicing Silent Speech,” in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, 2021, pp. 175–181.

[5] H. Li, H. Lin, Y. Wang, H. Wang, M. Zhang, H. Gao, Q. Ai, Z. Luo, and G. Li, “Sequence-to-Sequence Voice Reconstruction for Silent Speech in a Tonal Language,” *Brain Sciences*, vol. 12, no. 7, 2022.

[6] K. Scheck and T. Schultz, “Multi-Speaker Speech Synthesis from Electromyographic Signals by Soft Speech Unit Prediction,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[7] 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 *Proc. Interspeech 2020*, 2020, pp. 354–358.

[8] 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 *Proc. Interspeech 2021*, 2021, pp. 616–620.

[9] A. Illa, A. Nair, and P. K. Ghosh, “The impact of cross language on acoustic-to-articulatory inversion and its influence on articulatory speech synthesis,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 8267–8271.

[10] R. Parikh, N. Seneviratne, G. Sivaraman, S. Shamma, and C. Espy-Wilson, “Acoustic To Articulatory Speech Inversion Using Multi-Resolution Spectro-Temporal Representations Of Speech Signals,” in *Proc. Interspeech 2022*, 2022, pp. 4681–4685.

[11] D. Porras, A. Sepúlveda-Sepúlveda, and T. G. Csapó, “DNN-based Acoustic-to-Articulatory Inversion using Ultrasound Tongue Imaging,” in *2019 International Joint Conference on Neural Networks (IJCNN)*, 2019, pp. 1–8.

[12] J. Kong, J. Kim, and J. Bae, “HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis,” in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 17 022–17 033.

[13] M. Morrison, R. Kumar, K. Kumar, P. Seetharaman, A. Courville, and Y. Bengio, “Chunked autoregressive GAN for conditional waveform synthesis,” in *International Conference on Learning Representations*, 2022.

[14] B. van Niekerk, M.-A. Carbonneau, J. Zaïdi, M. Baas, H. Seuté, and H. Kamper, “A comparison of discrete and soft speech units for improved voice conversion,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6562–6566.

[15] S.-C. Jou, T. Schultz, M. Walliczek, F. Kraft, and A. Waibel, “Towards Continuous Speech Recognition Using Surface Electromyography,” in *Proc. Interspeech*, 2006.

[16] G. Beguš, A. Zhou, P. Wu, and G. K. Anumanchipalli, “Articulation gan: Unsupervised modeling of articulatory learning,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[17] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, “LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech,” in *Proc. Interspeech*, 2019, pp. 1526–1530.

[18] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative Adversarial Nets,” in *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, Eds., vol. 27. Curran Associates, Inc., 2014.

[19] D. Hazra and Y.-C. Byun, “SynSigGAN: Generative Adversarial Networks for Synthetic Biomedical Signal Generation,” *Biology*, vol. 9, no. 12, 2020.

[20] Z. Chen, Y. Qian, Y. Wang, and Y. Fang, “Deep Convolutional Generative Adversarial Network-Based EMG Data Enhancement for Hand Motion Classification,” *Front Bioeng Biotechnol*, vol. 10, p. 909653, Jul. 2022.

[21] R. Anicet Zanini and E. Luna Colombini, “Parkinson’s Disease EMG Data Augmentation and Simulation with DCGANs and Style Transfer,” *Sensors (Basel)*, vol. 20, no. 9, May 2020.

[22] Y. Luo and B.-L. Lu, “Eeg data augmentation for emotion recognition using a conditional wasserstein gan,” in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2018, pp. 2535–2538.

[23] S. M. Abdelfattah, G. M. Abdelrahman, and M. Wang, “Augmenting the size of eeg datasets using generative adversarial networks,” in *2018 International Joint Conference on Neural Networks (IJCNN)*, 2018, pp. 1–6.

[24] J. Chen, K. Liao, K. Wei, H. Ying, D. Z. Chen, and J. Wu, “ME-GAN: Learning panoptic electrocardio representations for multi-view ECG synthesis conditioned on heart diseases,” in *Proceedings of the 39th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 162. PMLR, 17–23 Jul 2022, pp. 3360–3370.

[25] T. H. Rafi and Y. Woong Ko, “HeartNet: Self Multihead Attention Mechanism via Convolutional Network With Adversarial Data Synthesis for ECG-Based Arrhythmia Classification,” *IEEE Access*, vol. 10, pp. 100 501–100 512, 2022.

[26] M. Bińkowski, J. Donahue, S. Dieleman, A. Clark, E. Elsen, N. Casagrande, L. C. Cobo, and K. Simonyan, “High Fidelity Speech Synthesis with Adversarial Networks,” in *International Conference on Learning Representations*, 2020.

[27] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley, “Least Squares Generative Adversarial Networks,” in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[28] L. Diener, S. Bredehöft, and T. Schultz, “A comparison of EMG-to-Speech Conversion for Isolated and Continuous Speech,” in *13th ITG Conference on Speech Communication*, 2018.

[29] D. Gaddy and D. Klein, “Digital Voicing of Silent Speech,” in *Proc. Empirical Methods in Natural Language Processing*, 2020, p. 5521–5530.

[30] L. Diener, “The impact of audible feedback on emg-to-speech conversion,” Ph.D. dissertation, University of Bremen, 2021. [Online]. Available: https://www.csl.uni-bremen.de/cms/images/documents/publications/Diener2021Diss.pdf

[31] I. Loshchilov and F. Hutter, “Decoupled Weight Decay Regularization,” in *International Conference on Learning Representations (ICLR)*, 2019.

[32] R. H. Chowdhury, M. B. I. Reaz, M. A. B. M. Ali, A. A. A. Bakar, K. Chellappan, and T. G. Chang, “Surface Electromyography Signal Processing and Classification Techniques,” *Sensors*, vol. 13, no. 9, pp. 12 431–12 466, 2013.

[33] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “Robust speech recognition via large-scale weak supervision,” 2022.