INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

A comparative study of different EMG features for acoustics-to-EMG mapping

Manthan Sharma[1], *Navaneetha Gaddam*[1], *Tejas Umesh*[1], *Aditya Murthy* [2], *Prasanta Kumar Ghosh* [1]

[1] Electrical Engineering, Indian Institute of Science, Bengaluru, 560012, India
[2]Centre For Neuroscience, Indian Institute of Science, Bengaluru, 560012, India

manthans@iisc.ac.in, snnavaneetha95@gmail.com, tejastf3@gmail.com, adi@iisc.ac.in, prasantg@iisc.ac.in

Abstract

Electromyography (EMG) signals have been extensively used to capture facial muscle movements while speaking since they are one of the most closely related bio-signals generated during speech production. In this work, we focus on speech acoustics to EMG prediction. We present a comparative study of ten different EMG signal-based features including Time Domain (TD) features existing in the literature to examine their effectiveness in speech acoustics to EMG inverse (AEI) mapping. We propose a novel feature based on the Hilbert envelope of the filtered EMG signal. The raw EMG signal is reconstructed from these features as well. For the AEI mapping, we use a bi-directional long short-term memory (BLSTM) network in a session-dependent manner. To estimate the raw EMG signal from the EMG features, we use a CNN-BLSTM model comprising of a convolution neural network (CNN) followed by BLSTM layers. AEI mapping performance using the BLSTM network reveals that the Hilbert envelope based feature is predicted from speech with the highest accuracy, among all the features. Therefore, it could be the most representative feature of the underlying muscle activation during speech production. The proposed Hilbert envelope feature, when used together with the existing TD features, improves the raw EMG signal reconstruction performance compared to using the TD features alone.
Index Terms: silent speech, acoustics-to-EMG, BLSTM

1. Introduction

Speech production is a complex process which involves the generation of a wide range of physiological signals, apart from acoustics, including articulatory movements and facial muscle activities. Activity of the muscles can be captured using multiple electrode surface electromyography (EMG), targeting prominent muscles involved in speech production. The resultant potential differences are measured and this amplified signal obtained over time is referred to as the EMG signal.

EMG signals have been introduced in the context of speech in past works [1, 2, 3]. However, in this work, we focus on speech acoustics to EMG inverse (AEI) mapping which can be useful in many areas of research. It can be used to improve previous efforts in visualisation of facial muscle activation from speech signals. Moreover, a large number of muscles are activated while speaking; knowing the particular muscles which are activated for the enunciation of a specific speech sound could be useful in improving speech synthesis and recognition. It could also help us determine the key muscles activated for each phoneme. Synthetic EMG features obtained from AEI mapping can help in improving the EMG to Speech mapping which is helpful for people suffering from speech disorders like Laryngectomy. It can also help in solving the problem of low resource EMG-acoustic data, due to the expensive experimental setup involved in such data collection. To the best of our knowledge, there has not been any research on speech to EMG prediction, apart from the work by Botelho et al. [4], in which authors extract time domain (TD) features from the EMG signals, as introduced by Jou et al. [1], which have been popularly used for EMG to Speech mappings.

The motivation for this work is to examine features of EMG signals that can help improve the AEI mapping, as well as in the reconstruction of the raw EMG signal from those features. In this study on speech to EMG prediction, we choose five TD features [1, 5] existing in the literature for comparison, namely, Low Frequency Mean (LFM), Low Frequency Power (LFP), High Frequency Power (HFP), High Frequency Zero Crossing Rate (HFZCR), and HF Rectified Mean (HFRM). We further implement another set of four non-speech temporal (NST) features that have been successful, in the literature, for representing EMG signals in the context of non-speech motor tasks. They are Mean Absolute Value (MAV) [6, 7], Root Mean Square (RMS) [8, 9], DAV [10], and Low Frequency Band-pass (LFB). As speech production involves motor planning and execution, apart from various cognitive activities, we experiment with NST features for the AEI mapping. The TD features capture specific details of the EMG signal including high frequency/low frequency power. However, the NST features, represent the raw EMG signals typically through an envelope of the time domain waveform, and maybe better suited to reflect muscle activity associated with speech production. Apart from the above mentioned TD and NST features, we also introduce a novel feature based on the Hilbert transform. The Hilbert transform introduced by Huang et al. [11] is an efficient way to analyze non-stationary and non-linear signals like EMG. The Hilbert transform helps in estimating instantaneous frequencies as a function of time, which is an energy-frequency-time distribution, known as the Hilbert spectrum. We, in this work, use the Hilbert transform to get a low frequency envelope of the raw EMG signal that explicitly captures spectro-temporal attributes of the EMG signal unlike other feature extraction methods.

For the AEI mapping using TD features, an hour glass shaped Deep Neural Network (DNN) model was used in step 1 of the work by Botelho et al. [4]. In this work, we instead use a BLSTM model for AEI mapping. Unlike DNNs, where temporal context needs to be provided explicitly, LSTMs are well known for sequence-to-sequence mappings because they inherently account for temporal dependencies. The prediction performance is measured using a correlation coefficient (CC) between the original and predicted feature trajectories. A CC of 0.76 is obtained using Hilbert envelope, which is found to be better than using all the other considered features. All the TD features (excluding HFZCR) perform better than the NST features except DAV which performs at par with the TD features. Following evaluation of the AEI mapping, we propose a CNN-

http://dx.doi.org/10.21437/Interspeech.2021-2240

BLSTM network based on the step 2 of the work by Botelho et al. [4] to reconstruct the raw EMG signal from these features. Since, the proposed Hilbert feature performs the best for AEI mapping, we use it together with the existing TD features for reconstructing the raw EMG signal more accurately compared to using just the TD features, which was done in the step 2 of the work by Botelho et al. [4].

2. Different EMG feature extraction methods

In the past, the EMG signals have not only been used in the context of speech production, but also for various other muscle activation related to non-speech tasks including finger, hand movement for picking objects, as well as leg motion, such as walking. This section provides a survey of different feature extraction methods that have consistently been proven effective for EMG signals, in the context of speech as well as non-speech tasks. In this study, we use all these feature extraction methods. We also introduce a novel Hilbert envelope-based feature for EMG to compare its effectiveness on AEI mapping and reconstruction of the raw EMG signal with the existing features. Implementation of these methods is made available on GitHub. [1]

2.1. EMG features used in context of speech task

Most recently, time domain (TD) features, namely, LFM, LFP, HFP, HFRM, HFZCR, introduced by Jou et al. [1, 5] have been commonly used for EMG to Speech mapping [12, 13, 3]. These features have also been used for the first work on speech to EMG mapping [4]. We extract the TD Features using a window of length 25ms and a shift of 10ms using a Blackman filter. These five TD features are concatenated to obtain a feature vector for a frame, TD_{frame}= [LFM, HFRM, LFP, HFP, HFZCR]. More details about these features can be found in the work by Jou et al. [1].

2.2. EMG features used in context of Non-speech task

In this sub-section, we describe in detail about the NST feature extraction for EMG signals used in the context of non-speech tasks. For extracting these features, the EMG signal is resampled to 1000 Hz and then band-pass filtered between 3 Hz to 300 Hz. This is done because the EMG signals have very less power outside this frequency range. We consider the range from 3 Hz to remove any DC offset present in the signal. After computing the features at the sample level, the NST features are obtained by downsampling the feature sequence to 100Hz.

Let the original EMG signal of N samples, that has been band-pass filtered from 3 Hz to 300 Hz, be represented as, $x[n], 1 \leq n \leq N$. Let the square of the signal be represented as, $s[n] = x^2[n], 1 \leq n \leq N$. As the signal is sampled at 1000 Hz, we represent the 10ms window duration as $W = 10$.

Mean Absolute Value (**MAV**) is a popular feature that detects the muscle contraction levels. It has been commonly used for EMG signals from wrist and hand movements [6], as well as for arm movements [7]. To calculate the MAV feature from the signal, we take the absolute values of the signal and run a moving average with a 10ms window. This is represented with the signal $m[k]$ as, $m[k] = \frac{1}{W}\sum_{n=k}^{k+W-1}|x[n]|, 1 \leq k \leq N$.

Root Mean Square (**RMS**) is another reliable feature that has been used for EMG signals while recording wrist and forearm motions [8], as well as leg motions during cycling [9]. It is popular because of low computational cost and great performance [14]. To calculate the RMS feature from the signal, we first calculate $s[n]$ and run a moving average with a 10ms window, and then take the square root of the values. This is represented with the signal $r[k]$ as, $r[k] = \sqrt{\frac{1}{W}\sum_{n=k}^{k+W-1} s[n]}, 1 \leq k \leq N$.

The pre-processing method introduced by d'Avella et al. [10] (named as '**DAV**' in this paper) has been quite frequently used for leg movements in animals. While this feature extraction method has not been used for EMG signals in the context of speech, a visualisation of this feature shows that it captures the EMG signal envelope very well. To calculate the DAV of an EMG signal, we first calculate $s[n]$, followed by a low pass filtering with a 20 Hz cut-off. Let us represent that with $\tilde{u}[n]$. We then take the absolute of the signal and run a moving average with a 10ms window. This is represented with the signal $d[k]$ as, $d[k] = \frac{1}{W}\sum_{n=k}^{k+W-1}|\tilde{u}[n]|, 1 \leq k \leq N$.

A feature extraction method (named '**LFB**': Low Frequency Bandpass) is also used by making slight modification to the DAV method of preprocessing EMG signals. To calculate the LFB we band-pass filter the raw EMG signal in the frequency range of 5-15 Hz. This is represented as $\hat{u}[n]$. We then take the absolute of these values. This is represented with the signal $\bar{l}[n]$] as, $\bar{l}[n] = |\hat{u}[n]|, 1 \leq n \leq N$.

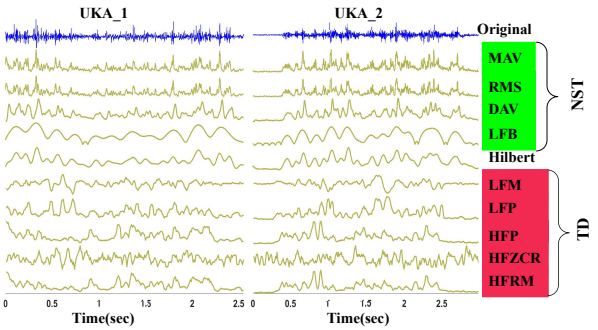

Figure 1: *Plots of different features of the EMG signals for 2 subjects (UKA_1 & UKA_2), together with the (time-aligned) original EMG signal*

2.3. Hilbert features

For the analysis of time-varying and non-linear signals, the **Hilbert Transform** has been proved to be an efficient method [11]. Therefore, Hilbert has been used as a popular tool for feature extraction for EMG signals from hand and finger movements [15]. Different versions including the Iterative Hilbert Transform (IHT) have been effective for upper limb movement [16]. Similarly, the Hilbert-Huang transform for EMG signals has been used in the context of speech recognition [17, 18]. We, in this work use the Hilbert transform to capture the envelope of the EMG signals. The method used has been described below in detail.

Given an EMG signal $x[n]$, the Hilbert transform returns a signal $x_h[n]$, which together with $x[n]$ results in a complex analytic signal as follows: $x_a[n] = x[n] + jx_h[n], 1 \leq n \leq N$, where the real part $x[n]$ contains the original data and imaginary part, $x_h[n]$, contains the Hilbert transform of the signal. The analytic signal can be represented in magnitude phase form as follows: $x_a[n] = A[n]e^{j\phi[n]}, 1 \leq n \leq N$

We then take the absolute of the analytic signal to get $A[n]$. We finally apply a low pass filter with cut-off frequency of 20

[1] https://github.com/snnavaneetha95/AcousticstoEMGmapping

Hz [2] to obtain the Hilbert feature, denoted by $h[n]$ which is then downsampled to 100Hz like the NST features. Fig. 1 illustrates ten different features used in this work for two example utterances from UKA_1 and UKA_2 subjects.

3. Model architectures

In this section, we present the architecture of the neural networks used for the AEI mapping and reconstruction of raw EMG signal. Codes for all the models are made available on our GitHub repository. [3]

3.1. BLSTM: Speech Acoustics to EMG Inversion

The work by Botelho et al. [4] is the first and only work, where authors used a DNN architecture for the AEI mappings. We propose to use a BLSTM model for the speech acoustics to EMG inversion task because of the inherent property of the LSTM networks to account for temporal dependencies. The network has 5 layers with 128 neurons in each layer. The 'linear' activation function is used for the output layer, which has 6 dimensions (one for every channel). We use 'tanh' as the activation function for all the layers with a Dropout [19] of 0.2. Mean Squared Error (MSE) and Concordance Correlation Coefficient (CCC) Loss introduced in [4] [20] are separately used as the objective function. The 'Adam' optimizer [21] is used to update model parameters. The MSE objective function is found to perform better than CCC loss in terms of the CC value between the original and predicted features. Hence, we report all results using MSE objective function in this study. We also embed a masking layer for the BLSTM network. A masking layer is generally embedded to use a mask value of zero, to skip some time frames during weight updation. This is especially useful for neural networks where the inputs have been padded to obtain inputs of identical length. Hence, masking is used to avoid the network from learning from the padded values and adding to the loss unnecessarily. The neural network is trained for 50 epochs with an early stopping criterion to avoid overfitting.

3.2. CNN-BLSTM: Raw EMG reconstruction

Botelho et al.[4] used 3 layered 1D CNN followed by 2 BLSTM layers and a fully connected layer to map TD features to raw EMG signals. In this work, we propose to use a combination of two features, i.e, TD and Hilbert instead of just TD features. The network architecture is shown in Fig. 2. The three 1D CNN layers are identical to those described in step 2 of [4]. Each BLSTM layer has 128 units with a 'tanh' activation and a dropout probability of 0.2. The output of the network is 6 dimensional time distributed fully connected layer. The model is trained on MSE loss with an 'Adam' optimizer. A masking technique, similar to that used for AEI mapping, is used and the model is trained for 50 epochs with an early stopping criterion. We have also experimented with a different setup where information from Hilbert features were extracted using CNN kernels in a manner similar to that of the TD features but the raw EMG reconstruction performance did not improve.

[2] We have experimented with various cut-off frequencies ranging from 5-40 Hz. It was found that the performance did not alter significantly when the cut-off frequency was chosen in the range 15-40 Hz.

[3] https://github.com/snnavaneetha95/AcousticstoEMGmapping

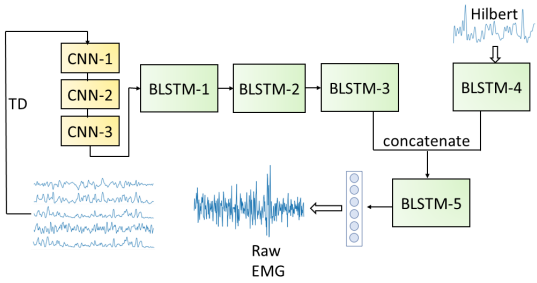

Figure 2: *CNN-BLSTM Architecture for Raw EMG reconstruction*

4. Experimental setup

In this work, the AEI experiments are performed on 2 subjects. The 2 large sessions from 2 different subjects, i.e., UKA_1 and UKA_2 from the EMG-UKA corpus, recorded at the Karlsruhe Institute of Technology, are used for the AEI mapping. Subjects UKA_1 and UKA_2 spoke 520 and 509 audible sentences in one session, respectively. For the reconstruction of raw EMG, we use the entire EMG-UKA corpus, i.e., all the audible sentences from all sessions. The six EMG electrodes capture the signals from 6 prominent muscles involved in speech production. The EMG signals and speech are recorded synchronously at 600 Hz and 16 kHz sampling rates, respectively. For details, please refer to the corpus paper by Wand et al. [22].

From the acoustics, we extract Mel Frequency Cepstral Coefficients (MFCC) [23] features which capture the spectral characteristics of the speech. For the BLSTM model used for AEI mapping, we compute a 25 dimensional MFCC with a 25 ms window length and a shift of 10 ms. The TD features as well as the NST features are computed from the EMG signal at 100Hz. These frame wise features are directly fed to the BLSTM model unlike the DNN model in [4], in which 15 frames from the past and future are stacked to account for time dependency. For NST, TD and Hilbert features the EMG feature matrices are of the form $T \times C$ where T is the number of frames in an utterance (with padding as required) and C is the number of EMG channels.

For the AEI mapping, EMG feature matrices are the output of the model and the input is the corresponding sequence of 25 dimensional MFCC vectors. It is known that the EMG signals are strongly session-dependent due to differences in electrode placement, as well as channel impedance and skin conditions that may vary [13]. Therefore, in this work, we report the AEI results on a session-based training i.e., each subject is trained and tested separately.

Unlike AEI, the reconstruction of raw EMG signal from the EMG features is a session independent mapping and, hence, does not depend on the position of electrodes or variation in subjects [4]. Thus, we use the entire audible UKA corpus and consider every channel individually as a data sample for the raw EMG signal reconstruction task. When reconstructing raw EMG signal using the CNN-BLSTM model, the inputs to the model are two EMG feature arrays–TD features to the CNN layers, and Hilbert feature array to the BLSTM-4 as shown in Fig. 2. To compare with the baseline work [4], two different sets of features are considered in place of the TD features: 1) LFM, LFP, HFP, HFZCR, HFRM and 2) LFM, LFP, HFP, HFRM. The output of the model is 6 dimensional vectors formed by non-overlapping window of 10ms on the raw EMG signal. Here, the input for TD features is $T \times F$ per channel where T is the number of frames in an utterance and F is the number of TD

Table 1: *The mean Correlation Coefficients (with standard deviation), across all EMG channels, for each feature and for both subjects. Red and Green represent the TD features and NST features, respectively*

Feature	UKA_1	UKA_2	Avg	Feature	UKA_1	UKA_2	Avg
LFM	0.66 (0.02)	0.66 (0.02)	0.66	MAV	0.48 (0.08)	0.63 (0.07)	0.55
HFM	0.66 (0.03)	0.67 (0.01)	0.66	LFB	0.58 (0.12)	0.51 (0.12)	0.54
HFP	0.68 (0.01)	0.73 (0.03)	0.70	DAV	0.65 (0.07)	0.71 (0.10)	0.68
HFZCR	0.20 (0.01)	0.27 (0.01)	0.23	RMS	0.58 (0.08)	0.63 (0.08)	0.60
HFRM	0.68 (0.01)	0.77 (0.02)	0.72	Hilbert	0.70 (0.07)	0.76 (0.08)	0.73

features used. The Hilbert input matrix is of dimension $T \times 1$ and the output of the network is $T \times 6$, where 6 corresponds to the vector obtained from 10ms window.

For both the tasks, a 10 fold cross validation is done to get the final average Pearson's CC [24] [25] across all test samples in all folds. 10% of the training data is used as the validation set.

5. Results

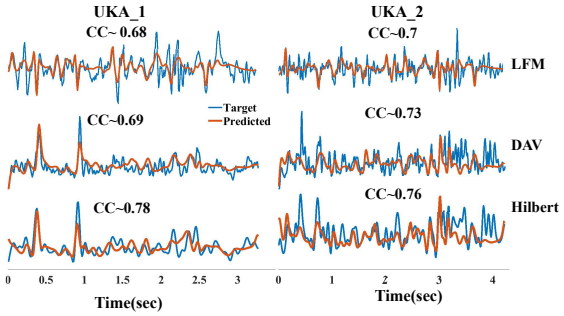

Figure 3: *Illustration of original and predicted LFM, DAV and Hilbert features from the AEI mapping separately for UKA_1 and UKA_2*

In this section, we present the results obtained from the AEI mapping and the raw EMG signal reconstruction. For both the tasks, we cross-validate our results and report the CC averaged across all folds.

In the AEI mapping, we train the model individually for all the features. As reported in Table 1, Hilbert based feature performs better than all the other features for AEI mapping when trained using each feature separately. Amongst the TD features, the CC for HFZCR is found to be the lowest, similar to that reported in [4]. The DAV performs the best amongst the NST features. However, its performance is lower than best of the TD features. In general, the subject UKA_2 performs better separately for every feature which shows the session-dependent characteristics of the EMG signals. Fig. 3 illustrates original and predicted LFM feature (the best among the TD features), DAV feature (the best among the NST features) and the Hilbert feature for an utterance from UKA_1 and UKA_2 separately. It can be seen that, for an utterance considered, Hilbert based feature is predicted with the highest CC values.

The goal of the next set of experiments is to reconstruct raw EMG signal as accurately as possible. For this we do a 10 fold cross validation for TD features (including and excluding HFZCR) both with and without Hilbert features. Here, we consider only TD features and not the NST features for reconstruction task because all the TD features except HFZCR perform, on average, better than the NST features for the AEI mapping. We

Table 2: *The mean (standard deviation) correlation coefficient (CC) values for the raw EMG signal reconstruction across folds with different feature combinations*

Raw EMG reconstruction	
Features	**CC**
LFM+LFP+HFP+HFZCR+HFRM	0.65 (0.17)
LFM+LFP+HFP+HFRM	0.61 (0.11)
LFM+LFP+HFP+HFZCR+HFRM+Hilbert	0.73 (0.02)
LFM+LFP+HFP+HFRM+Hilbert	0.66 (0.02)

observe that jointly using Hilbert with the TD features improves raw EMG signal reconstruction performance yielding better CC values for both blue (including HFZCR) and orange (excluding HFZCR) cases as reported in Table 2. We get an average CC value of 0.73 when the raw EMG signal is reconstructed using all TD features and Hilbert together. To compare the performance with [4], which uses only TD features to reconstruct, we compute average CCC values for the reconstructed raw EMG signal and it is found to be 0.69 for TD+Hilbert features and 0.64 for TD features alone. Fig. 4 shows an example of raw EMG signal reconstruction along with CC values using TD and TD+Hilbert features separately. Thus, the Hilbert feature helps the model to capture the finer details of the raw EMG waveform.

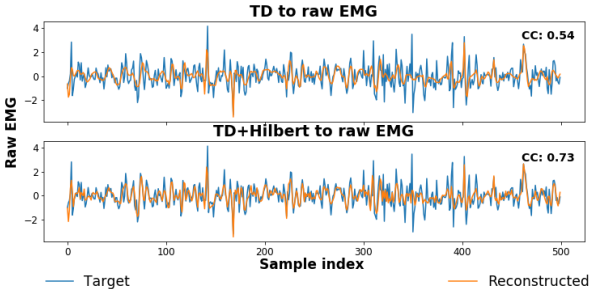

Figure 4: *Target and reconstructed raw EMG signal, using TD and TD+Hilbert features*

6. Conclusion

In this work, we perform a comparative study of ten different feature extraction techniques for AEI mapping and raw EMG signal reconstruction. We aim to determine a feature extraction technique that would best represent the EMG signals, which, in turn, would lead to an accurate raw EMG signal prediction from speech acoustics. We introduce a novel Hilbert feature that performs the best, on average, across two subjects for the AEI mapping, used in this work. The proposed Hilbert feature also helps in reconstructing the raw EMG signal more efficiently compared to existing TD features only. In future, we would like to reconstruct raw EMG signal using the predicted features from the AEI models instead of original EMG features. Methods to obtain intermediate learnable representations of the EMG signals for the AEI mapping can be investigated. Exploring different objective functions to train the models is also an interesting area to explore in future. In this work we use the full-band (i.e., 3 Hz to 300Hz) features, but the sub-band features can be potentially more useful to capture intricacies of the complex EMG signals compared to the full-band features.

7. Acknowledgements

The authors thank the Pratiksha Trust for their support.

8. References

[1] 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, pp. 573–576.

[2] T. Schultz and M. Wand, "Modeling coarticulation in EMG-based continuous speech recognition," *Speech Communication*, vol. 52, no. 4, pp. 341–353, 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] 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," *Small*, vol. 61, p. 12, 2020.

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

[6] A. Phinyomark, "A novel feature extraction for robust EMG pattern recognition," *Journal Of Computing*, vol. 1, pp. 71–80, 2009.

[7] K. Kiguchi, T. Tanaka, and T. Fukuda, "Neuro-fuzzy control of a robotic exoskeleton with EMG signals," *IEEE Transactions on fuzzy systems*, vol. 12, no. 4, pp. 481–490, 2004.

[8] A. Phinyomark, F. Quaine, Y. Laurillau, S. Thongpanja, C. Limsakul, and P. Phukpattaranont, "EMG amplitude estimators based on probability distribution for muscle–computer interface," *Fluctuation and Noise Letters*, vol. 12, no. 1350016, pp. 1–18, 2013.

[9] C. A. Hautier, K. Arsac, Laurent Mauriceand Deghdegh, J. Souquet, A. Belli, and J.-R. Lacour, "Influence of fatigue on EMG/force ratio and cocontraction in cycling," *Medicine and Science in Sports and Exercise*, vol. 32, no. 4, pp. 839–843, 2000.

[10] A. d'Avella, P. Saltiel, and E. Bizzi, "Combinations of muscle synergies in the construction of a natural motor behavior," *Nature neuroscience*, vol. 6, no. 3, pp. 300–308, 2003.

[11] N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N.-C. Yen, C. C. Tung, and H. H. Liu, "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis," *Proceedings of the Royal Society of London. Series A: mathematical, physical and engineering sciences*, vol. 454, no. 1971, pp. 903–995, 1998.

[12] L. Maier-Hein, F. Metze, T. Schultz, and A. Waibel, "Session independent non-audible speech recognition using surface electromyography," in *Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2005, pp. 331–336.

[13] L. Diener, T. Umesh, and T. Schultz, "Improving fundamental frequency generation in EMG-to-Speech conversion using a quantization approach," in *Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 682–689.

[14] Y. Fang, H. Liu, G. Li, and X. Zhu, "A multichannel surface EMG system for hand motion recognition," *International Journal of Humanoid Robotics*, vol. 12, no. 02, p. 1550011, 2015.

[15] A. O. Andrade, P. Kyberd, and S. J. Nasuto, "The application of the Hilbert spectrum to the analysis of electromyographic signals," *Information Sciences*, vol. 178, no. 9, pp. 2176–2193, 2008.

[16] J. L. Dideriksen, F. Gianfelici, L. Z. P. Maneski, and D. Farina, "EMG-based characterization of pathological tremor using the iterated Hilbert transform," *IEEE transactions on biomedical engineering*, vol. 58, no. 10, pp. 2911–2921, 2011.

[17] C. Jorgensen and S. Dusan, "Speech interfaces based upon surface electromyography," *Speech Communication*, vol. 52, no. 4, pp. 354–366, 2010.

[18] C. Jorgensen, D. D. Lee, and S. Agabont, "Sub auditory speech recognition based on EMG signals," in *Proceedings of the International Joint Conference on Neural Networks*, vol. 4. IEEE, 2003, pp. 3128–3133.

[19] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[20] I. Lawrence and K. Lin, "A concordance correlation coefficient to evaluate reproducibility," *Biometrics*, pp. 255–268, 1989.

[21] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[22] M. Wand, M. Janke, and T. Schultz, "The EMG-UKA corpus for electromyographic speech processing," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014, pp. 1593–1597.

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

[24] J. Benesty, J. Chen, Y. Huang, and I. Cohen, "Pearson correlation coefficient," in *Noise Reduction in Speech Processing*. Springer, 2009, pp. 1–4.

[25] 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.