EMG-to-Speech: Direct Generation of Speech from Facial Electromyographic Signals

Matthias Janke, *Student Member, IEEE,* and Lorenz Diener, *Student Member, IEEE*

***Abstract*—Silent Speech Interfaces** **are systems that enable speech communication even when an acoustic signal is unavailable. Over the last years, public interest in such interfaces has intensified. They provide solutions for some of the challenges faced by today's speech-driven technologies, such as robustness to noise and usability for people with speech impediments. In this paper, we provide an overview over our Silent Speech Interface. It is based on *facial surface electromyography*, which we use to record the electrical signals that control muscle contraction during speech production. These signals are then converted *directly* to an audible speech waveform, retaining important paralinguistic speech cues for information such as speaker identity and mood. This paper gives an overview over our state-of-the-art direct EMG-to-speech transformation system. The paper describes the characteristics of the speech EMG signal, introduces techniques for extracting relevant features, presents different EMG-to-speech mapping methods and finally presents an evaluation of the different methods for real-time capability and conversion quality.**

***Index Terms*—Silent Speech Interface, Electromyography, SSI, Biosignal**

I. Introduction

WHETHER it is secret agents communicating silently while sneaking around, like in numerous cold war novels, or inhabitants of a space station silently talking to an artificial intelligence, such as in Orson Scott Card's "Ender's Game" – the ability use speech as a means of communication without speaking words out aloud has been a mainstay of fiction for decades. The approaches imagined range from sensing inaudibly quiet sounds by recording articulator movement to reading imagined speech directly from the brain.

Human speech communication is the most sophisticated form of interaction. Due to its efficiency, ease of use and information richness, it has been the focus of a lot of attention in human-computer interaction research. The majority of all widely used speech recognition based systems rely on speech transmitted over air – acoustic data. However, several scenarios exist where non-audio representations of speech might be helpful:

- Adverse Noise Conditions: Though techniques for noise-robust ASR exist [1], loud ambient background noise can make exploiting an acoustic speech signal challenging or even impossible.
- Complementing Acoustic Speech: Adding a second modality for speech representation can improve a speech processing systems performance.
- Silent operation: Audible speech can disturb bystanders and may also be overheard by eavesdroppers.
- Speech Rehabilitation: Approximately 7.5 million people in the United States have trouble using their voices according to a survey by the National Institute on Deafness and Other Communication Disorders [2]. Silent speech based systems might offer new alternatives for rehabilitation [3].

While fiction is, as usual, ahead of reality by some amount, *Silent Speech Interfaces* – speech interfaces that do not rely on the presence of an audible acoustic signal – do exist, with various alternative sensor technologies being actively investigated by research groups:

- Surface electromyography (EMG) [4]–[8]: The activation potentials of facial articulatory muscles are recorded with surface electrodes, providing information about articulatory muscle movement during speech production.
- Brain computer interfaces based on electroencephalography (EEG) [9], [10], near infrared sensors (fNIRS) [11], [12] or implants in the speech-motor cortex (ECOG) [13], [14]: Electrical (EEG, ECOG) or the hemodynamic correlate of brain activity is recorded to try to gain information about the speech production process.
- Video camera based lip reading [15], [16] – a video camera captures the movement of the mouth, and spoken words are inferred using image processing techniques.
- Permanent Magnetic or Electromagnetic Articulography (PMA or EMA) [17], [18]: The movement of magnets attached to the articulators is captured by measuring magnetic field changes using sensors around the mouth.
- Ultrasound / optical imaging of tongue and lips [19]–[22]
- Non-audible murmur (NAM) microphones [23]–[27]: Nearly-inaudible body-conducted low-amplitude acoustic waves are measured with a type of stethoscopic microphone.
- Glottal activity detection based on electroglottography (EGG) [28], [29] or vibrometry [30], [31]: Electrical activity or vibration in the larynx area is measured to infer glottal activity.

For a general overview of biosignal-based spoken interaction, refer to the survey paper in this special issue [32]. We additionally differentiate into four speech modalities:

1) Audible speech
2) NAM and Whisper: Recording of nearly silent signals that don't necessarily propagate through air, recorded using bone-conduction or stethoscopic microphones. Although these approaches still require an acoustic signal, the signal does not need to be audible to humans.
3) Silent speech: Capturing information from the vocal tract or articulatory configurations, using e.g. EMG or PMA. These approaches do not rely on an audible signal

All authors are with the Cognitive Systems Lab, University of Bremen, Germany

only silent articulation, i.e. movement of the articulators without sound production, is required.

4) Imagined speech: Direct interpretation of the process of imagined speech production by capturing brain signals.

In this paper we use EMG recorded during audible speech and perform a *direct* conversion to speech. Note that, while the speech we use is audible, our system uses *only* the EMG signal for conversion (The audible signal is required for training). It is, therefore, possible to build an EMG-based system in a way that puts it into category three – a truly *Silent Speech* system. Previous work (e.g. [33], [34]) already proposed automatic speech recognition (ASR) on EMG-based input resulting in text output, that can be synthesized using text-to-speech systems. However this recognition-followed-by-synthesis approach has some limitations: First, there is an output delay based on the additional computation time in the ASR step. Second, the ASR-based approach is restricted to a given language and vocabulary. Third, features like speaking rate and paralinguistic information – e.g. speaker identity, mood, etc. – which is crucial for a natural communication, are not transmitted.

We expect our direct feature transformation technique to have the advantages of retaining paralinguistic information and operating without the latency and vocabulary limitations imposed by an ASR step. Furthermore, the proposed direct speech synthesis needs no language information. The labeling that provides the phonetic information required by ASR-based speech transformation approaches introduces an additional potential error source – this work establishes a label-free approach.

Today, several research groups are promoting EMG-based speech processing [8], [35]–[40]. Some investigate a particular topic, e.g. Portuguese language specific factors [41] or focus on peculiarities of disordered speakers [42]. Additionally, there are multimodal approaches, i.e. the combination of acoustic and EMG signals [42]–[44].

Thus far, there has only been little work on direct EMG-to-speech conversion [45], [46]. Toth et al. [47] have reported promising results with a GMM-based technique without a speech recognition step. They use 5 EMG channels with electrodes positioned on muscles of the articulatory apparatus, recording 380 utterances (about 48 minutes) of parallel audible speech and EMG data. Additionally, a speech recognition experiment is performed using the synthesized speech output. Restricting the testing vocabulary to 108 words, 84.3 % of the words were recognized correctly. First results with speech recognition on EMG recorded during silent articulation are also presented – however, this reduces the word recognition accuracy to 20.2 %.

Denby et al. use ultrasound images from the tongue to directly generate a speaker's vocal tract parameters. [19]. Hueber et al. [48] add video image information to ultrasound data and use a combination of hidden Markov modeling and Unit Selection to synthesize speech. The authors state that the speech output is of decent quality for correctly predicted sequences; however, the error number is still too high to generate a truly usable output signal. The same authors [49] additionally present a Gaussian Mixture Model (GMM) based approach to convert similarly captured lip and tongue motions.

Toda et al. [50], [51] use electromagnetic articulography (EMA) input data for a GMM-based mapping technique, which obtains good output quality.

Although the quality of silent speech interfaces has gradually improved and the used recording devices has become considerably more affordable, many caveats still remain. Some approaches require a laboratory environment (e.g. EMA) or have a heavily restricted set of output units (e.g. Lam at al. [45] train on 8 phonemes only).

Denby et al. [52] compare different silent speech modalities, examining different mapping approaches. They value EMG for its high potential in terms of non-invasiveness, cost, silent-usage and other factors.

Finally, Wand et al. [35] introduced an electrode array grid, which is also used in this paper.

The following sections present the state-of-the-art in EMG-to-speech conversion systems. They describe the setup used for recording EMG and speech in parallel, the corpus used in this paper for evaluation and the feature processing used to extract feature sequences from EMG and audio waveforms. Several mapping techniques are described: EMG-to-speech feature transformation based on GMMs, Neural Networks and Unit Selection. The quality of their output is evaluated using objective criteria as well as subjective listening tests.

II. Characteristics of the facial EMG signal during speech production

EMG is a biosignal that consists of electrical currents emitted from muscles during their contraction, representing neuromuscular activity [53]. Recorded as a surface signal, it is the summation of the activity of many different motor units, attenuated while crossing different tissue layers and is in practice overlaid by ambient electromagnetic noise and artifacts. The investigation into the relationship between EMG and speech has been ongoing for several decades [54]–[56].

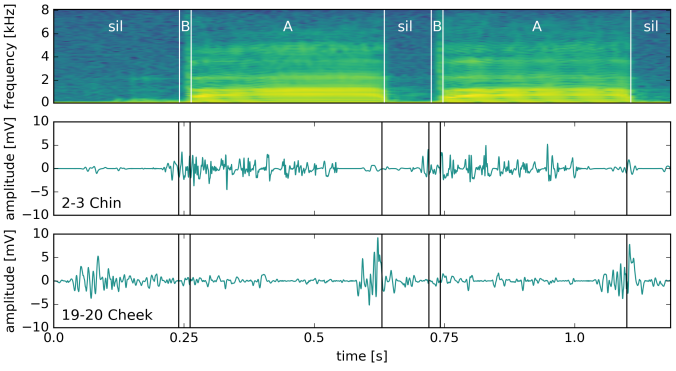

Fig. 1: Spectrogram of the acoustic signal (top) and time-series plot of the EMG signals from two channels for part of the phrase "babababa". Noise in EMG data was filtered out using a *spectral subtraction* technique [57] for visualization purposes.

As a first investigation we compare a simultaneously recorded audio signal with the EMG signal of simple spoken consonant-vowel phrases. Fig. 1 gives an example for a part of the phrase "babababa", showing the acoustic spectrogram and EMG time series – one channel from the chin array and one channel from the cheek array (Compare Fig. 4 for electrode positions). A couple of observations can be made:

- The EMG signals are noticeably different from each other, especially in onset and breakoff.
- The acoustic signal is delayed with regards to the EMG signal, due to an effect known as Electromechanical Delay (EMD) [58]. We compensate for this signal delay with a constant delay of 50 ms [59] and contextual feature stacking (see Section III-A).

III. EMG-to-Speech Transformation

The general framework of the proposed EMG-to-speech approach is shown in Fig. 2. It consists, broadly, of two stages:

1) a training stage (green arrows),
2) a conversion stage on unseen data (blue arrows).

For training, we use simultaneously recorded EMG and acoustic data (see section IV). The data consist of EMG feature vectors as source data and audio feature vectors as target data. See Sec. III-A for details on EMG and acoustic signal processing. The direct transformation from EMG features into acoustic representations is realized either by Gaussian Mapping (Sec.III-C), Unit Selection (III-D) or Deep Neural Networks (III-E). Vocoding creates the final speech wave files from the generated acoustic representations. This step is done using Mel Log Spectrum Approximation (MLSA) [60]. While this is a relatively simple vocoding scheme, our evaluation of other schemes has not yielded significant improvements – though it may be prudent to re-evaluate this in the future after improvements to mapping methods have increased quality somewhat.

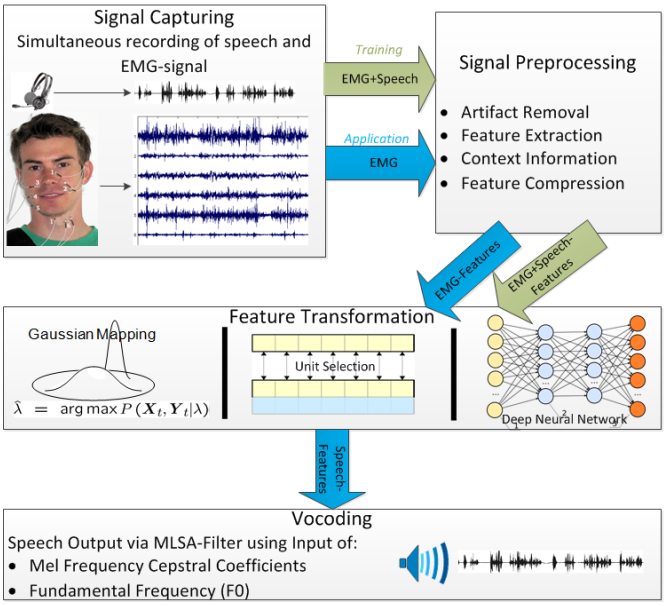

Fig. 2: *Processing steps of EMG-to-speech transformation. Green arrows represent training data flow, blue arrows represent data flow during EMG-to-speech application.*

A. Signal Preprocessing

We use a set of time-domain based features first introduced by Jou et al. [59] (*TD* features). For calculating TD-features, the signal is first split into a high-frequency and low-frequency part with a triangular filter with a cutoff frequency of 134 Hz implemented using a double moving average. The signal is then windowed rectangularly with a frame size of 27 ms and frame shift of 10 ms, respectively. The TD features, calculated for every frame, are defined as (With ZCR being the zero-crossing rate and x_{low} and x_{high} being the low and high frequency parts):

$$\mathbf{TD}(x_{\text{low}}, x_{\text{high}}) = \Bigg[\frac{1}{n}\sum_{i=1}^{n}(x_{\text{low}}[i])^2, \quad \frac{1}{n}\sum_{i=1}^{n}x_{\text{low}}[i], \quad \frac{1}{n}\sum_{i=1}^{n}(x_{\text{high}}[i])^2, \quad \frac{1}{n}\sum_{i=1}^{n}|x_{\text{high}}[i]|, \quad \text{ZCR}(x_{\text{high}})\Bigg]$$

The TD feature vectors are stacked with adjacent vectors to introduce contextual information of the immediate past and future into the final feature. The stacking is performed for 15 frames into the past and 15 frames into the future, resulting in stacked *TD15* feature vectors 31 frames in length. This relatively high amount of contextual information achieved good results in ASR experiments and partially compensates for the electromechanical delay effect.

Together, these parameters result in a feature dimensionality of 5425 (35 channels – compare section IV – times 5 features times 31 because of stacking). To reduce the dimensionality before training an EMG-to-speech conversion system, linear discriminant analysis (LDA) is applied. An LDA matrix that maximizes discriminability of phone sub-states (the beginning, middle and end of each phone) is calculated based on labels force-aligned to the EMG signal using the simultaneously recorded audio signal, taking EMD into account as a 50ms time shift. The signal is transformed by this matrix and then truncated to the 32 highest-discriminability dimensions.

For the acoustic signal, 25 Mel-Frequency Cepstral Coefficients (MFCCs) [61] and the speech fundamental frequency (F_0) are used. They are extracted as filter parameters of a MLSA filter and F_0 estimates in 32 ms Blackman-windowed frames with a 10 ms shift (resulting in frames time-aligned with the EMG feature frames).

B. Vocoding

For methods where the output of the mapping is a sequence of MFCCs and F_0s, it is necessary to convert those features back to an audio waveform. In our vocoding step, this is achieved using the MLSA filter method [60]. This is possible since the MFCCs and F_0s were extracted as MLSA filter parameters. The vocoding step is the same for all mapping methods in which it is used (i.e. all but Unit Selection with direct concatenative synthesis).

C. Feature Transformation using Gaussian Mixture Models

GMMs are a commonly used technique in voice conversion [62]. The variant used in this work is based on the GMM-based articulatory-to-acoustic mapping introduced by Toda et al. [51].

To train the feature transformation, parallel source (EMG) and target (MFCC/F0) vectors are stacked to create joint feature vectors. A GMM is fitted to these joint vectors. The joint GMM probability density can then be used during conversion by finding the MFCC/F0 feature vector that maximizes the combined likelihood given the EMG feature vector and joint density, i.e. the MFCC/F0 feature vector that, stacked with the EMG feature vector, is assigned the highest likelihood by

the GMM. This is done by calculating the expected value for the MFCC/F0 feature vector given the source vector and GMM. [63], [64]:

We have thoroughly investigated the use of GMMs in EMG-to-Speech conversion, most recently their performance depending on the number of mixtures used, in a session-dependent as well as a session-independent setting [65]. It therefore serves as a baseline to which we compare other transformation methods.

D. Feature Transformation using Unit Selection

Unit Selection was first introduced in the 1980s [66] and has since become a popular approach for speech synthesis [67]. To perform speech synthesis or conversion with Unit Selection, short segments of audio data are selected from a database (called a *codebook*) and then concatenated, sometimes with overlap, to create an output audio sequence. Recently, we have introduced EMG-based Unit Selection [68] for direct EMG-to-speech conversion.

The unit database (the *codebook*) is created by extracting segments of l frames length – the *unit width* – from a set of training utterances. The segments are extracted synchronously for the EMG *source features* and the acoustic *target features*. Each such pair of parallel source- and target feature segments is called a *codebook unit*. To get a large variety of units, this is done not with one unit being extracted starting after the previous one, but instead with a *unit shift* of one frame (i.e. with an overlap of $l - 1$ frames between units). Together, the extracted units make up the codebook.

To convert an EMG signal to audible speech, a sequence of *test units* is created similarly to how codebook units were extracted – however, there is only a source (EMG) feature sequence and thus, no audio data in the units. Here, the unit shift is not 1 but is instead optimized on a development set. The result of this is an *output unit sequence* made from codebook units, with one codebook unit chosen for each test unit.

For each of the test units, a codebook unit is selected according to the combination of two cost functions, a *target cost* and a *concatenation cost*. The target cost function measures how well a codebook unit fits the given test unit. It is calculated between the test- and codebook units' EMG segments.

The concatenation cost function is calculated between codebook units audio segments. It measures how well the units' acoustic segments fit to each other when they are directly adjacent in the output unit sequence, which enables a smooth transition between the audio segments. Using a weighted sum (with empirically determined weights) of these two costs as the selection criterion, the unit selection process is a search for the sequence of codebook units that minimizes the total cost given the test unit sequence. In our previous work [68] [69], we have evaluated different functions for target and concatenation cost in Unit Selection EMG-to-speech conversion, achieving the best results with the mean cosine similarity.

Fig. 3 illustrates the complete unit search process. If the concatenation cost weight is non-zero, the search for the ideal unit sequence has to consider the entire sequence at the same time. This can be done efficiently using the Viterbi algorithm [70] (as a Viterbi search on a fully connected graph with the units as nodes and the weighted costs as edge weights). Due to the large number of choices in each step, a full Viterbi search is still computationally expensive – in practice, it is better to just consider a limited number of active paths in each step. Taking this to the extreme of considering only one active path, it is possible to use a greedy algorithm to always select the best next unit given the already selected units. When the concatenation cost *is* zero (i.e. the concatenation cost is ignored), the restricted searches are equivalent to a full Viterbi search, otherwise, they trade computation time for correctness.

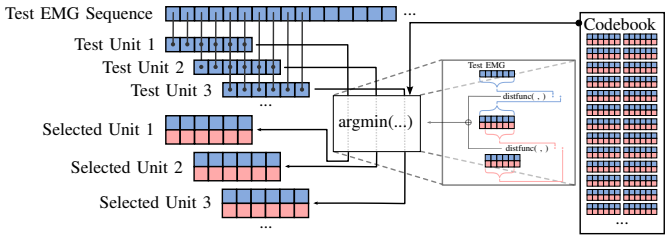

Fig. 3: Illustration of the search for the optimal unit sequence. The EMG sequence is split into test units, on which the sum of unit costs (sum of units' EMG target and audio concatenation costs) is minimized.

After determining the optimal unit sequence, the overlapping audio feature segments of the output unit sequence are used to create the final output MFCC/F0 sequence. This is done by taking the mean of all selected units' frames that correspond to one input frame. This output MFCC and F0 sequences can then be passed to the MLSA vocoder for speech generation.

We recently advanced this Unit Selection technique using a clustering approach [71] that substantially reduces the number of units in the codebook and thus the computation time, while improving the output quality. The main idea behind this clustering is the reduction of audio artifacts by creating units that are more representative of a single relation between EMG and audio. This has the benefit of reducing sensitivity to outlier units, a single one of which can already greatly reduce intelligibility. Additionally, it eliminates redundancies in the codebook, which reduces computation time requirements for the conversion process.

The base codebook units (calculated as above) are clustered in two stages, using the k-means algorithm. First, units are clustered according to the audio feature vectors of all audio frames covered by each unit. Second, the units assigned to each audio cluster are clustered (separately for each cluster) according to the EMG feature vectors covered by each of these units. Given these cluster assignments, a set of *cluster units* is created by calculating the mean audio and EMG feature frames over all units assigned to a cluster. These units are then used as the new codebook in the Unit Selection conversion process described above.

For our evaluations, we used a cluster-based system with 6000 clusters based only on audio features, with weight parameters optimized on a development set.

E. Feature Transformation using Neural Networks

Artificial Neural Networks are models whose power lays in the interconnection of many simple units ("neurons") that,

together, can perform complex calculations. This section describes EMG-to-Speech conversion with two different kinds of neural network models: Feedforward *deep neural networks* (DNNs), and *Long-Short-Term Memory* (LSTM) networks.

EMG-based neural network approaches have been introduced for phone classification [46] and, more recently, for direct EMG-to-speech feature transformation [72]. A similar articulatory-to-acoustic mapping approach based on Deep Neural Networks was introduced by Bocquelet et al. [73]. They trained on electromagnetic articulography (EMA) data which was recorded synchronously with the audible speech sounds.

1) Feedforward DNNs: Feedforward deep neural networks are the simplest form of the DNN architecture. They are used for frame-based EMG-to-speech conversion in our previous work [72]. The basic component of the networks is the *rectified-linear* neuron. Many such neurons are arranged into layers, which are then connected with all units from one layer feeding into each unit of the next layer. We construct a 3-hidden-layer feedforward neural network, separately for MFCCs and F0s.

The size of the input- and output layers is determined by the the input- and output feature dimension. The dimension of the hidden layers is a free parameter and has to be optimized. In our previous work [72], we started with a general bottleneck shape, following earlier experience with processing whispered speech [74]. Several sets of layer sizes (keeping the same overall network structure) were evaluated on a development set. Layer sizes of 2500 neurons for the first, 512 neurons for the second and 1024 neurons for the third hidden layer gave the best performance. These five layer neural networks are trained on parallel EMG and audio feature vectors using stochastic gradient descent with 1024 sample minibatches, a momentum of 0.9 and a learning rate of 0.001 for the first three epochs and then 0.01 for all following epochs. The sum-squared-error is used as the loss function. Training is stopped after the loss on a validation set (held out from the training set) stops decreasing. To avoid bias towards numerically larger EMG- or audio features, the signal is normalized to zero mean and unit variance. Dropout [75] is used to reduce overfitting.

After the training process has converged, we get a set of weight and bias matrices which fully define a mapping function from input EMG features to target acoustic speech features.

2) Long-Short-Term-Memory Networks: We also evaluate the direct EMG-to-speech approach with LSTM networks, which are state-of-the-art for several of problems, e.g. recognition of speech [76], [77] or hand-writing [78]. LSTMs, introduced in 1997 by Hochreiter et al. [79], are *recurrent neural networks* and enable a long-range temporal context by using memory cell units that store information over a longer period of time, together with non-linear gating units that regulate the data flow into and out of the cell. The usage of LSTMs in EMG-to-speech conversion is motivated by their ability to cope with temporal dependencies directly as part of the model instead of requiring the stacking of feature vectors.

The LSTMs used in this work are bidirectional LSTMs, as described by Graves et al [80]. We chose to use a training momentum of 0.9 in all our networks, as LSTM research [81] and our own preliminary experiments have shown that this parameter has only minor influence on LSTM performance. To determine the number of memory blocks per layer and learning rate, we varied numbers from literature and optimized them for our task and our own input data. The speech feature enhancement by Wolmer et al. [82] used three hidden layers, consisting of 78, 128 and 78 memory blocks. We use this as our basis, varying the number of hidden layers (from 1 to 4) and memory blocks per layer (60, 80 and 100), and evaluate several learning rates for each set of parameters. We obtained the best results using two hidden layers, consisting of 100 and 80 memory blocks, with a learning rate of $3 \cdot 10^{-7}$. To avoid overfitting, we stopped training after 20 epochs without improvement of the validation sum of squared errors. Input and output data was normalized to zero mean and unit variance.

IV. Recording Setup and Data Corpus

To capture EMG signals, we used the OT Bioelettronica (http://www.otbioelettronica.it) EMG-USB2 multi-channel EMG amplifier. Data acquisition using electrode-arrays [35] is a step towards practical usage, as they are easier to attach than the single electrodes conventionally used [8], [47]. and less time consuming. In addition, a higher number of EMG channels is available due to a large number of electrodes.

After some initial experiments, we chose a 2,048 Hz sampling frequency, 3 Hz cutoff frequency high-pass filter and a low-pass filter with a cutoff frequency of 900 Hz. The amplifiers Driven Right Leg (DRL) noise reduction circuit [83] was used to reduce common mode (e.g. line noise) interference. Electrode gel was applied in order to reduce electrode/skin impedance. Double-sided adhesives were used for attachment.

Following our initial experiments, we decided on capturing signals with two arrays: A cheek array consisting of 4 rows of 8 electrodes with 10 mm inter-electrode distance (IED), and a chin array with a single 8 electrodes row with 5 mm IED (The numbering of the final electrode channels is shown Fig. 4).

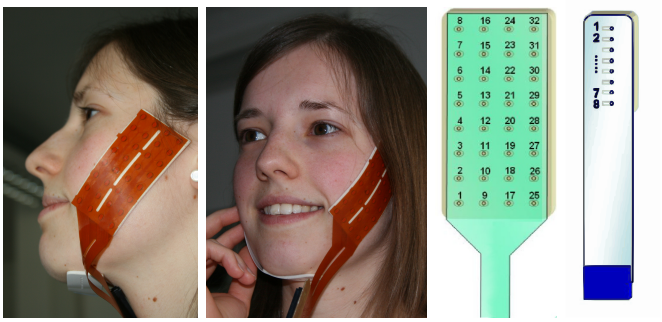

Fig. 4: Left: Electrode array positioning; 8 × 4 electrode grid on the cheek, small linear array under the chin. Right: Channel numbering of the 4 by 8 cheek electrode array, and the 1 by 8 chin array.

Bipolar derivation is used, where the potential differences between adjacent channels in each column are calculated. This results in 35 bipolar EMG channels – 28 from the 8 × 4 cheek electrodes and 7 from the 8 chin electrodes. To alleviate problems with detached electrodes, each sessions' recordings were inspected manually and visibly faulty channels were excluded for that session.

The chosen electrode positioning was the best compromise between getting rich information about the muscle movements

and allowing for minimal interference with the subjects articulation process. According to extensive experiments on electrode positioning [84], EMG-based speech processing requires, at the very least, signals from the cheek area and the throat (to capture tongue activity).

Recordings were done on a laptop with the in-house developed recording software BiosignalsStudio [85]. Each recording consists of a set of read phonetically balanced English sentences from the broadcast news domain.. In order to assure consistent pronunciation of words and to detect errors in the setup (e.g. detached electrodes), all recordings were supervised by a member of the research team.

All sentences were recorded with an EMG-recording system and a standard close-talking headset in parallel. An additional channel contains an analogue marker signal marking the begin and end of each utterance. Each of the recorded sentences was displayed on a screen (in random order) and the speakers were asked to hold the recording button, then read the sentence in normal, audible speech and release the button. The resulting utterance was considered valid if the pronunciation of each word in the sentence had been articulated properly in English. In case of mispronunciations or disfluencies, the utterance was repeatedly recorded until the pronunciation of the complete sentence was valid. Furthermore, the subjects were allowed to practice the pronunciation.

TABLE I: EMG-Audible Data Corpus Information

Speaker-Session	Sex	Length [mm:ss]			# of utterances		
		Train	Dev	Eval	Train	Dev	Eval
S1-Single	m	24:23	02:47	01:19	450	50	20
S1-Array	m	28:01	03:00	00:47	450	50	10
S1-Array-Lrg	m	68:56	07:41	00:48	984	109	10
S2-Single	m	24:12	02:42	00:49	447	49	13
S2-Array	m	22:14	02:25	01:10	450	50	20
S3-Array-Lrg	f	110:46	11:53	00:46	1,771	196	10
Total		278:32	30:28	05:39	4,552	504	83

The corpus consists of six sessions total, from three speakers, with varying amounts of data, each split into a *train*(ing), *dev*(elopment) and *eval*(uation) set. Four sessions incorporate around 500 phonetically balanced English utterances that are based on a corpus introduced in our previous work [8]. Two larger sessions exist (tagged with the suffix "Lrg"). These incorporate utterances from the Arctic [86] and TIMIT [87] corpora, resulting in a total of 1,103 utterances for the smaller and 1,977 utterances for the bigger of the large sessions. Table I lists the durations of the recorded sessions and the number of utterances per session. We additionally used two sessions with the single-electrodes setup from our previous work [8] (tagged with "Single" instead of "Array"), which allows us to compare array with single-electrodes setups.

V. Evaluation of transformation methods

This section presents different subjective and objective evaluations [88] for the presented EMG-to-speech approaches.

In our experiments, we used both the *Computational Network Toolkit* [89] and the *brainstorm* [90] neural network implementations. For the evaluation of LSTM networks, we used the *CURRENNT* implementation [91].

A. Run-Time Evaluation

Since we hypothesized that one of the advantages of the direct synthesis approach is its fast processing time, we evaluate the real-time capabilities of the different EMG-to-speech techniques using the conversion time of the evaluation set. We only state pure conversion time for mapping EMG features to MFCC features, as MLSA vocoding and file-I/O are assumed to be constant between the different feature transformation approaches and are therefore omitted.

The input EMG feature dimensionality depends on the mapping approach. We use two different input EMG feature sets: High-dimensional TD15 features and those features reduced to 32 dimensions using LDA. The TD15 features dimensionality depends on the amount of usable channels in a session, ranging from 5,425 dimensions on the 35-channel array setup to 930 dimensions on the single electrodes setup. The Gaussian mapping uses 64 Gaussian mixtures. The number of Gaussians and conversion time are linearly related(e.g. using 32 Gaussians approximately cuts the conversion time in half).

All measurements were obtained on an Intel Core i7-2700 CPU running at 3.5 GHz. The results of the evaluation can be found in Table II. The real-time (RT) factor is the ratio of feature transformation time to the duration of the converted utterances – a RT factor of one means that each second of input requires 1 second pure feature transformation time. Note that all results were obtained by calculating the time taken to convert the entire evaluation set as one batch.

The Unit Selection needs to compare with every codebook unit. Therefore, the time Unit Selection takes for feature transformation, depends on the amount of codebooks (i.e. training data). This means that mapping requires considerably higher feature transformation times which are less useful for a real-time setup (RT-factor > 20) and the results are not presented in detail. Even with the proposed unit clustering approach, which considerably reduces computation time, the real-time factor of our implementation clearly stays above 1.

TABLE II: Run-time comparison of transformation methods

Session	Time taken for feature transformation in [sec] (RT-Factor)			
	DNN (TD15)	LSTM (TD15)	LSTM (LDA)	GMM (LDA)
S1-Single	2.9 (0.02)	26.6 (0.16)	5.6 (0.03)	42.7 (0.26)
S2-Single	2.9 (0.02)	23.4 (0.14)	5.4 (0.03)	41.4 (0.26)
S1-Array	12.3 (0.07)	161.4 (0.90)	6.1 (0.03)	45.7 (0.25)
S2-Array	10.2 (0.07)	144.9 (1.00)	4.7 (0.03)	35.5 (0.24)
S1-Arr-Lrg	14.6 (0.03)	139.9 (0.30)	15.5 (0.03)	118.7 (0.26)
S3-Arr-Lrg	16.2 (0.02)	155.4 (0.22)	23.4 (0.03)	181.5 (0.25)

The model-based neural network and Gaussian Mapping approaches conversion times do not depend on the amount of training data, as they only need to load the previously trained models. However, input dimensionality influences them heavily. The neural network is faster than all other approaches. Even with high-dimensional feature input (S1-Array and S2-Array having a input feature vector size of 5,425 dimensions), the mapping is faster than 0.1 times real-time, making this method the preferred choice for an online EMG-to-speech system. Using the reduced 32-dimensional feature input, Gaussian mapping still achieves 0.25 times real time, fast enough for real-time use. While this study compares the different

EMG mapping approaches, details of an implementation with optimized real-time speech output can be found in [92].

B. Objective Evaluation using Mel-Cepstral Distortion

To *objectively* evaluate our results, we employ the *Mel-Cepstral Distortion* (MCD) score [93], defined as a scaled Euclidean distance between MFCC vectors excluding the first coefficient, computed between a synthesized utterance and the reference utterance. Since MCD represents a distance measure, lower numbers imply better results.

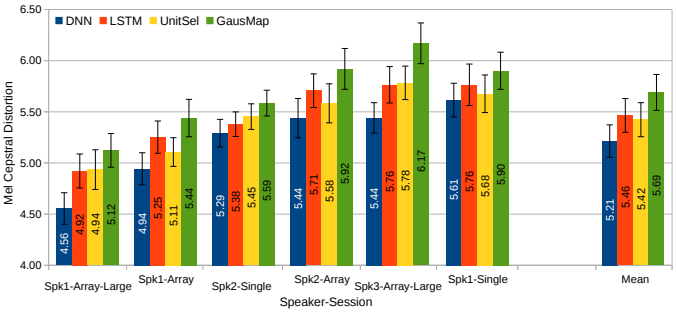

Fig. 5: MCD comparison on the evaluation set of all investigated EMG-to-speech mapping approaches: Deep Neural Networks (DNN), Long-Short-Term Memory Networks (LSTM), Unit Selection (UnitSel) and Gaussian Mapping (GausMap). Error bars show standard deviation.

Fig. 5 shows that while there is only a small difference between LSTM and Unit Selection (mean MCD of 5.46 vs 5.42), DNNs give the best results by a significant margin ($p < 0.01$) with a mean MCD of 5.21. Gaussian Mapping obtains the highest (worst) average MCD with 5.69. The best session-dependent result is achieved on the Spk1-Array-Large session with an MCD of 4.56.

A comparison of single-electrode versus array-based setups shows no clear tendency in terms of MCD. While the first EMG-to-speech approach presented by Toth et al. [47] introduced an MCD of 6.37, our current results improve the best performance to 4.51.

C. Objective Intelligibility Evaluation using ASR

In addition to the MCD evaluation, we decode the synthesized speech output using an automatic speech recognition (ASR) engine [8] to evaluate the intelligibility of the synthesized speech more directly compared to using spectral similarity measures. The speech decoder we use is based on three-state left-to-right fully continuous Hidden-Markov-Models using bundled phonetic features (BDPFs), an advanced variant of articulatory features [94]. Details about the recognition system can be found in our previous work [8].

The acoustic speech recognizer is trained on the acoustic output that was generated from the EMG training set input and finally tested on the synthesized evaluation set. This is done on each of the proposed mapping techniques. The recognizer uses a trigram language model. To enable comparability to EMG-based speech recognition on the same data [35], we restrict the decoding vocabulary to 3 different sizes: 108, 905 and 2,111 words, including variants. The Word Error Rate (WER) is used to measure the performance of the ASR.

The results of the ASR evaluation are presented in Fig. 6.

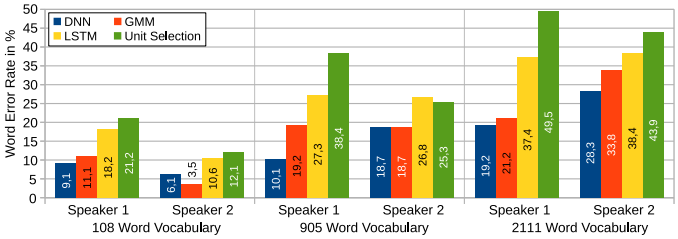

Fig. 6: Acoustic speech recognizer based comparison of the EMG-to-speech approaches using three different decoding vocabulary sizes: 108, 905 and 2,111 words, evaluated for two different speakers.

The DNN approach consistently obtains the best word error rates, confirming the results that have been achieved in the MCD evaluation. The Gaussian Mapping output gets second best results, outperforming LSTM and Unit Selection. As the Gaussian Mappings GMMs are trained on MFCCs, while the ASR acoustic model GMMs are trained on phone classes, a systematic bias of the evaluation towards GMMs is unlikely.

Comparing the obtained WER to our previous work that uses a comparable setup with an EMG-based ASR system [35], these results are encouraging. Wand et al. [35] achieve an average WER of 10.9 % using a vocabulary of 108 words with 160 training utterances, while we report a mean WER of 7.3 % using our synthesized speech approach. Toth et al. [47] state that "WER results for EMG-to-speech are actually better than results from training the ASR system directly on the EMG data". The increased amount of training data may result in the fact that we get even better results. While a direct comparison is difficult, we have shown that good performance can be achieved when the synthesized output is used on an ASR system.

D. Naturalness Evaluation using Listening Tests

To evaluate how an end-user judges the naturalness of the output of the EMG-to-speech conversion, we evaluate the F0 generation, which is the essential component for prosody generation. Is the generated F0 signal stemming from EMG better than an F0 signal that is generated without knowledge of the input or even omitted? To investigate this question we conduct a preliminary experiment: We use the target MFCCs from the reference audio file (allowing us to factor out intelligibility), and generate three different excitations:

1) the mapped *EMG-to-F0* output,
2) white Gaussian noise, resulting in unvoiced – whisper-like – speech signal, entitled *0 F0*
3) a constant *flat F0* contour, resulting in voiced, robot-like acoustic output.

We also add two variations of the reference speech recording: the original unaltered (reference), plus the re-synthesized reference recording (resynth). The latter one is generated by using extracted speech features (MFCCs + F0) from the target audio with the MLSA filter to produce a "re-synthesized" reference. This contains the quality degradation from the acoustic preprocessing and MLSA vocoding steps and thus represents the best output we can achieve with our EMG-to-speech setup. A listening test is conducted to evaluate naturalness. The participant is asked to answer the question "How natural does the presented speech recording sound?, please rate between very unnatural and very natural." This follows the

naturalness test from the Blizzard challenge [95]. A continuous slider, which is internally scaled on the interval between 0 and 100, is given and the mentioned five output variations are presented in randomized order. We randomly selected three different utterances from four different speakers, resulting in 12 utterances – each of the 12 utterances synthesized in 5 variations. This results in a total of 60 played to each participant. The following is the result of this evaluation performed with a total of 20 participants. The resulting scores are shown in Table III.

TABLE III: Naturalness evaluation of transformation methods

	Reference	Resynthesized	EMG-to-F0	Zero F0	Flat F0
Score	81	41	28	22	7

The highest drop in naturalness can be found between reference and resynthesized output. EMG-to-F0 mapping significantly ($p < 0.01$) gives the most natural output among the three F0 generation methods. 0 F0 obtained the second best naturalness score, while flat F0 was perceived worst. This implies that the whisper-like output is regarded more natural than the robotic-like F0 generation. This implies the superiority of the EMG-to-F0 mapping over simple artificial F0 generation.

E. Intelligibility Evaluation using Listening Tests

Since the MCD score does not give any insights to the generated naturalness or prosodic information, we conduct a set of subjective listening tests, where participants were asked to, as above, compare the outputs of the proposed EMG-to-speech systems to each other and to the given reference and resynthesized reference, this time asking listeners to rate intelligibility. This comparative approach is used since overall intelligibility is low – only some utterances can actually be understood, making a human-transcription approach to intelligibility evaluation hard to use.

Ten different evaluation-set utterances were randomly selected and each of them is synthesized in 5 variations. Thus, a total of 50 utterances are played to the listening test participant. 10 listeners participated in the test. The results can be seen in Fig. 7. On average, the participants preferred the DNN-based

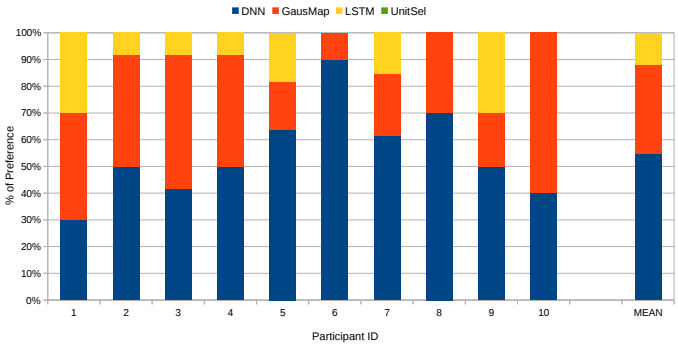

Fig. 7: Listening test preference comparison of all investigated EMG-to-speech mapping approaches.

output in more than 50 % of the cases, while Unit Selection was never preferred in a single utterance.

F. Spectrogram Comparison

To visualize the final speech output, Fig. 8 depicts the spectrograms from the synthesized DNN-based EMG-to-speech output (on bottom) and additionally the resynthesized reference signal from an exemplary utterance taken from Spk-1-Array-Large (on top). The similar spectral shape is recognizable. However, there are also visible artifacts, the investigation of which may lead to quality improvements: Occasional minor single-frame artifacts (Marked with the $\star$ symbol) and an overall blurring of the spectrum in both the time and frequency.

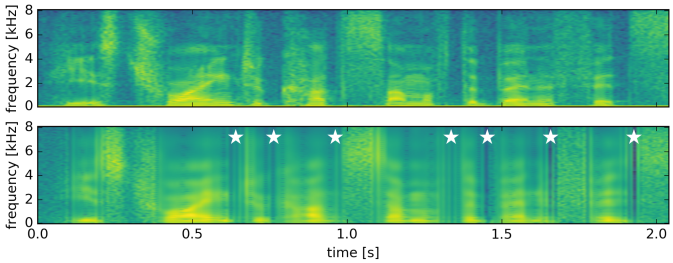

Fig. 8: Spectrograms of the utterance “He is trying to cut some of the benefits.”, reference on top, DNN-based EMG-to-speech on bottom.

VI. Conclusion

This paper introduced a speech synthesis technique to directly convert surface EMG signals of the articulatory muscles to audible speech. This approach has multiple advantages over a recognition-based conversion. First, There is no restriction to a given phone-set, vocabulary or even language. Second, paralinguistic information like speech cues for speaker mood and emotion and the characteristics of a speakers voice can be preserved, as they are implicitly modeled by the direct transformation. Third, direct mapping enables faster processing compared to EMG-to-text-to-speech and is therefore suitable for real-time use, enabling feedback to the speaker.

Four transformation approaches were presented: GMMs, DNNs, LSTMs and Unit Selection. Several sets of parameters for these methods were evaluated on a development data set. Comparative evaluations on a held-out evaluation set revealed that out of our current approaches, the Deep Neural Network based method performs best in real-time behaviour, naturalness and intelligibility. The feasibility of real-time processing was investigated and evaluated with a real-time factor lower than 0.1 using the proposed feed-forward DNNs.

In the future, we hope to further improve the real-time latency and output quality and investigate the effect of co-adaptation. We hope to improve intelligibility to the point where most converted utterances can be understood by further investigating the properties of the speech EMG signal and by incorporating prior knowledge (such as linguistic information) into the conversion process.

Acknowledgements

This research project was partially funded by the German Research Foundation (DFG), Research Grant SCHU 2452.5-1 LZV (2012-2016) entitled ”MAPS - Myoelectric Array-based Processing of Speech”. The authors thank their colleagues from CSL and Tanja Schultz for in-depth comments and corrections. Also, the authors would like to thank the anonymous reviewers for their valuable suggestions.

REFERENCES

[1] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, "An Overview of Noise-Robust Automatic Speech Recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 745–777, 2014.

[2] National Institute on Deafness and Other Communication Disorders (NIDCD), "Statistics on Voice, Speech, and Language," 2016. [Online]. Available: http://www.nidcd.nih.gov/health/statistics/pages/vsl.aspx

[3] H. Doi, T. Toda, K. Nakamura, H. Saruwatari, and K. Shikano, "Alaryngeal Speech Enhancement Based on One-To-Many Eigenvoice Conversion," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 1, pp. 172–183, 2014.

[4] N. Sugie and K. Tsunoda, "a Speech Prosthesis Employing a Speech Synthesizer – Vowel Discrimination from Perioral Muscle Activities and Vowel Production," *IEEE Transactions on Biomedical Engineering*, vol. 32, no. 7, pp. 485–490, 1985.

[5] M. S. Morse and E. M. O'Brien, "Research Summary of a Scheme to Ascertain the Availability of Speech Information in the Myoelectric Signals of Neck and Head Muscles using Surface Electrodes," *Computers in Biology and Medicine*, vol. 16, no. 6, pp. 399–410, 1986.

[6] A. D. C. Chan, K. Englehart, B. Hudgins, and D. Lovely, "Myoelectric Signals to Augment Speech Recognition," *Medical and Biological Engineering and Computing*, vol. 39, pp. 500–506, 2001.

[7] C. Jorgensen, D. Lee, and S. Agabont, "Sub Auditory Speech Recognition Based on EMG Signals," in *Proceedings of the International Joint Conference on Neural Networks, 2003*, vol. 4, Portland, Oregon, 2003, pp. 3128–3133.

[8] T. Schultz and M. Wand, "Modeling Coarticulation in EMG-Based Continuous Speech Recognition," *Speech Communication*, vol. 52, no. 4, pp. 341–353, 2010.

[9] A. Porbadnigk, M. Wester, J.-P. Calliess, and T. Schultz, "EEG-Based Speech Recognition - Impact of Temporal Effects," in *Proc. Biosignals*, 2009.

[10] S. Deng, R. Srinivasan, T. Lappas, and M. D'Zmura, "EEG Classification of Imagined Syllable Rhythm using Hilbert Spectrum Methods," *Journal of Neural Engineering*, vol. 7, no. 4, 2010.

[11] A. Villringer, J. Planck, C. Hock, L. Schleinkofer, and U. Dirnagl, "Near Infrared Spectroscopy (NIRS): A New Tool to Study Hemodynamic Changes During Activation of Brain Function in Human Adults," *Neuroscience Letters*, vol. 154, pp. 101–104, 1993.

[12] E. Formisano, F. D. Martino, M. Bonte, and R. Goebel, "Who Is Saying What? Brain-Based Decoding of Human Voice and Speech," *Science*, vol. 322, no. 5903, pp. 970–973, 2008.

[13] J. S. Brumberg, A. Nieto-Castanon, P. R. Kennedy, and F. H. Guenther, "Brain-Computer Interfaces for Speech Communication," *Speech Communication*, vol. 52, pp. 367–379, 2010.

[14] C. Herff, D. Heger, A. de Pesters, D. Telaar, P. Brunner, G. Schalk, and T. Schultz, "Brain-To-Text: Decoding Spoken Phrases from Phone Representations in the Brain," *Frontiers in Neuroscience*, vol. 9, pp. 1–11, 2015.

[15] T. Hasegawa and K. Ohtani, "Oral Image to Voice Converter-Image Input Microphone," in *Singapore ICCS/ISITA'92.'Communications on the Move*, 1992, pp. 617–620.

[16] R. Bowden, S. Cox, R. Harvey, Y. Lan, and E.-J. Ong, "Recent Developments in Automated Lip-Reading," *SPIE Security+ Defence. International Society for Optics and Photonics*, 2013.

[17] M. J. Fagan, S. R. Ell, J. M. Gilbert, E. Sarrazin, and P. M. Chapman, "Development of a (silent) Speech Recognition System for Patients Following Laryngectomy," *Medical Engineering and Physics*, vol. 30, no. 4, pp. 419–425, 2008.

[18] J. M. Gilbert, S. I. Rybchenko, R. Hofe, S. R. Ell, M. J. Fagan, R. K. Moore, and P. D. Green, "Isolated Word Recognition of Silent Speech using Magnetic Implants and Sensors," *Medical Engineering and Physics*, vol. 32, pp. 1189–1197, 2010.

[19] B. Denby and M. Stone, "Speech Synthesis from Real Time Ultrasound Images of the Tongue," in *Proc. ICASSP*, 2004, pp. 685–688.

[20] T. Hueber, G. Aversano, G. Chollet, B. Denby, G. Dreyfus, Y. Oussar, P. Roussel, and M. Stone, "Eigentongue Feature Extraction for an Ultrasound-Based Silent Speech Interface," in *Proc. ICASSP*, 2007, pp. 1245–1248.

[21] V.-M. Florescu, L. Crevier-Buchman, B. Denby, T. Hueber, A. Colazo-Simon, C. Pillot-Loiseau, P. Roussel, C. Gendrot, and S. Quattrocchi, "Silent Vs Vocalized Articulation for a Portable Ultrasound-Based Silent Speech Interface," in *Proc. Interspeech*, 2010, pp. 450–453.

[22] 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, pp. 288–300, 2010.

[23] Y. Nakajima, "Development and Evaluation of Soft Silicone NAM Microphone," IEICE, Tech. Rep., 2005.

[24] Y. Nakajima, H. Kashioka, K. Shikano, and N. Campbell, "Non-Audible Murmur Recognition," in *Proc. Eurospeech*, 2003.

[25] V. A. Tran, G. Bailly, H. Loevenbruck, and T. Toda, "Improvement to a NAM-Captured Whisper-To-Speech System," *Speech Communication*, vol. 52, pp. 314–326, 2010.

[26] P. Heracleous, T. Kaino, H. Saruwatari, and K. Shikano, "Unvoiced Speech Recognition using Tissue-Conductive Acoustic Sensor," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, pp. 1–11, 2007.

[27] T. Toda, M. Nakagiri, and K. Shikano, "Statistical Voice Conversion Techniques for Body-Conducted Unvoiced Speech Enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 9, pp. 2505–2517, 2012.

[28] M. Rothenberg, "a Multichannel Electroglottograph," *Journal of Voice*, vol. 6, no. 1, pp. 36–43, 1992.

[29] I. R. Titze, B. H. Story, G. C. Burnett, J. F. Holzrichter, L. C. Ng, and W. a. Lea, "Comparison Between Electroglottography and Electromagnetic Glottography," *The Journal of the Acoustical Society of America*, vol. 107, no. 1, p. 581, 2000.

[30] J. C. Bos and D. W. Tack, "Speech Input Hardware Investigation for Future Dismounted Soldier Computer Systems," DRCD Toronto CR 2005-064, 2005.

[31] S. A. Patil and J. H. L. Hansen, "the Physiological Microphone (PMIC): A Competitive Alternative for Speaker Assessment in Stress Detection and Speaker Verification," *Speech Communication*, vol. 52, pp. 327–340, 2010.

[32] T. Schultz, M. Wand, T. Hueber, D. J. Krusienski, C. Herff, and J. S. Brumberg, "Biosignal-based spoken communication: A survey," *submitted to IEEE/ACM Transactions on Audio, Speech and Language Processing*.

[33] A. D. C. Chan, K. Englehart, B. S. Hudgins, and D. Lovely, "Hidden Markov Model Classification of Myolectric Signals in Speech," *Engineering in Medicine and Biology Magazine, IEEE*, vol. 21, no. 9, pp. 143–146, 2002.

[34] S.-C. Jou, L. Maier-Hein, T. Schultz, and A. Waibel, "Articulatory Feature Classification using Surface Electromyography," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2006.

[35] M. Wand, C. Schulte, M. Janke, and T. Schultz, "Array-Based Electromyographic Silent Speech Interface," in *Proc. Biosignals*, 2013.

[36] G. S. Meltzner, J. Sroka, J. T. Heaton, L. D. Gilmore, G. Colby, S. Roy, N. Chen, and C. J. D. Luca, "Speech Recognition for Vocalized and Subvocal Modes of Production using Surface EMG Signals from the Neck and Face," in *Proc. Interspeech*, 2008.

[37] K.-S. Lee, "Prediction of Acoustic Feature Parameters using Myoelectric Signals." *IEEE Transactions On Biomedical Engineering*, vol. 57, no. 7, pp. 1587–1595, 2010.

[38] C. Jorgensen and S. Dusan, "Speech Interfaces Based upon Surface Electromyography," *Speech Communication*, vol. 52, no. 4, pp. 354–366, 2010.

[39] Y. Deng, G. Colby, J. T. Heaton, and G. S. Meltzner, "Signal Processing Advances for the MUTE SEMG-Based Silent Speech Recognition System," in *Military Communications Conference*, 2012, pp. 1–6.

[40] T. Kubo, M. Yoshida, T. Hattori, and K. Ikeda, "Towards Excluding Redundancy in Electrode Grid for Automatic Speech Recognition Based on Surface EMG," *Neurocomputing*, 2014.

[41] J. Freitas, A. Teixeira, and M. S. Dias, "Towards a Silent Speech Interface for Portuguese," in *Proc. Biosignals*, 2012, pp. 91–100.

[42] Y. Deng, R. Patel, J. Heaton, G. Colby, L. Gilmore, J. Cabrera, S. Roy, C. Luca, and G. Meltzner, "Disordered Speech Recognition using Acoustic and SEMG Signals," in *Proc. Interspeech*, 2009, pp. 644–647.

[43] E. J. Scheme, B. Hudgins, and P. A. Parker, "Myoelectric Signal Classification for Phoneme-Based Speech Recognition," *IEEE Transactions On Biomedical Engineering*, vol. 54, no. 4, pp. 694–9, 2007.

[44] Q. Zhou, N. Jiang, K. Englehart, and B. Hudgins, "Improved Phoneme-Based Myoelectric Speech Recognition," *IEEE Transactions on Biomedical Engineering*, vol. 56, p. 8, 2009.

[45] Y.-M. Lam, P. H.-W. Leong, and M.-W. Mak, "Frame-Based SEMG-To-Speech Conversion," in *IEEE International Midwest Symposium on Circuits and Systems (MWSCAS)*, 2006, pp. 240–244.

[46] T. Tsuji, N. Bu, J. Arita, and M. Ohga, "a Speech Synthesizer using Facial EMG Signals," *International Journal of Computational Intelligence and Applications*, vol. 7, no. 1, pp. 1–15, 2008.

[47] A. Toth, M. Wand, and T. Schultz, "Synthesizing Speech from Electromyography using Voice Transformation Techniques," in *Proc. Interspeech*, 2009, pp. 652–655.

[48] T. Hueber, G. Chollet, B. Denby, M. Stone, and L. Zouari, "Ouisper: Corpus Based Synthesis Driven by Articulatory Data," in *Proceedings of 16th International Congress of Phonetic Sciences*, 2007, pp. 2193–2196.

[49] T. Hueber, E.-L. Benaroya, B. Denby, and G. Chollet, "Statistical Mapping Between Articulatory and Acoustic Data for an Ultrasound-Based Silent Speech Interface," in *Proc. Interspeech*, 2011.

[50] T. Toda, A. W. Black, and K. Tokuda, "Mapping from Articulatory Movements to Vocal Tract Spectrum with Gaussian Mixture Model for Articulatory Speech Synthesis," in *Proc. of the 5th ISCa Speech Synthesis Workshop*, 2004.

[51] ——, "Statistical Mapping Between Articulatory Movements and Acoustic Spectrum using a Gaussian Mixture Model," *Speech Communication*, 2008.

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

[53] M. B. I. Reaz, M. S. Hussain, and F. Mohd-Yasin, "Techniques of EMG Signal Analysis: Detection, Processing, Classification and Applications," *Biol Proced Online*, vol. 8, pp. 11–35, 2006.

[54] K. Faaborg-Andersen and A. Edfeldt, "Electromyography of Intrinsic and Extrinsic Laryngeal Muscles During Silent Speech: Correlation with Reading Activity: Preliminary Report," *Acta oto-laryngologica*, vol. 49, no. 1, pp. 478–482, 1958.

[55] D. Kewley-Port, "EMG Signal Processing for Speech Research," *Haskins Laboratories Status Report on Speech Research*, vol. SR-50, pp. 123–146, 1977.

[56] H. Hirose, "Electromyography of the Articulatory Muscles: Current Instrumentation and Techniques," *Haskins Laboratories Status Report on Speech Research*, vol. SR-25/26, pp. 73 – 86, 1971.

[57] S. F. Boll, "Suppression of Acoustic Noise in Speech using Spectral Subtraction," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 27, no. 2, pp. 113–120, 1979.

[58] P. R. Cavanagh and P. V. Komi, "Electromechanical Delay in Human Skeletal Muscle Under Concentric and Eccentric Contractions," *European Journal of Applied Physiology and Occupational Physiology*, vol. 42, no. 3, pp. 159–163, 1979.

[59] S.-C. Jou, T. Schultz, M. Walliczek, F. Kraft, and A. Waibel, "Towards Continuous Speech Recognition using Surface Electromyography," in *Proc. Interspeech*, 2006, pp. 573–576.

[60] S. Imai, "Cepstral Analysis Synthesis on the Mel Frequency Scale," *Proc. ICASSP*, pp. 93–96, 1983.

[61] T. Fukada and K. Tokuda, "an Adaptive Algorithm for Mel-Cepstral Analysis of Speech," in *Proc. ICASSP*, 1992, pp. 137–140.

[62] T. Toda, A. W. Black, and K. Tokuda, "Voice Conversion Based on Maximum-Likelihood Estimation of Spectral Parameter Trajectory," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 8, pp. 2222–2235, 2007.

[63] Y. Stylianou, O. Cappe, and E. Moulines, "Continuous Probabilistic Transform for Voice Conversion," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 2, pp. 131–142, 1998.

[64] A. Kain and M. W. Macon, "Spectral Voice Conversion for Text-To-Speech Synthesis," in *Proc. ICASSP*, 1998, pp. 285–288.

[65] M. Janke, M. Wand, K. Nakamura, and T. Schultz, "Further Investigations on EMG-To-Speech Conversion," in *Proc. ICASSP*, 2012, pp. 365–368.

[66] Y. Sagisaka, "Speech Synthesis by Rule using an Optimal Selection of Non-Uniform Synthesis Units," in *Proc. ICASSP*, 1988, pp. 679 – 682.

[67] A. J. Hunt and A. W. Black, "Unit Selection in a Concatenative Speech Synthesis System using a Large Speech Database," in *Proc. ICASSP*, 1996, pp. 373–376.

[68] M. Zahner, M. Janke, M. Wand, and T. Schultz, "Conversion from Facial Myoelectric Signals to Speech: A Unit Selection Approach," in *Proc. Interspeech*, 2014, pp. 1184 – 1188.

[69] L. Diener, "Improving Unit Selection Based EMG-To-Speech Conversion," Master's thesis, Karlsruher Institut für Technologie, Germany, 2015.

[70] A. J. Viterbi, "Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm," *IEEE Transactions on Information Theory*, vol. 13, no. 2, pp. 260–269, 1967.

[71] L. Diener, M. Janke, and T. Schultz, "Codebook Clustering for Unit Selection Based EMG-To-Speech Conversion," in *Proc. Interspeech*, 2015.

[72] ——, "Direct Conversion from Facial Myoelectric Signals to Speech using Deep Neural Networks," in *International Joint Conference on Neural Networks (IJCNN)*, 2015, pp. 1–7.

[73] F. Bocquelet, T. Hueber, L. Girin, P. Badin, and B. Yvert, "Robust Articulatory Speech Synthesis using Deep Neural Networks for BCI Applications," in *Proc. Interspeech*, 2014, pp. 2288–2292.

[74] M. Janke, M. Wand, T. Heistermann, K. Prahallad, and T. Schultz, "Fundamental Frequency Generation for Whisper-To-Audible Speech Conversion," in *Proc. ICASSP*, 2014, pp. 2579–2583.

[75] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout : A Simple Way to Prevent Neural Networks from Overfitting," *Journal of Machine Learning Research (JMLR)*, vol. 15, pp. 1929–1958, 2014.

[76] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech Recognition with Deep Recurrent Neural Networks," in *Proc. ICASSP*, no. 6, 2013.

[77] H. Sak, A. Senior, and F. Beaufays, "Long Short-Term Memory Recurrent Neural Network Architectures for Large Scale Acoustic Modeling," in *Proc. Interspeech*, 2014, pp. 338–342.

[78] P. Doetsch, M. Kozielski, and H. Ney, "Fast and Robust Training of Recurrent Neural Networks for Offline Handwriting Recognition," in *Proc. International Conference on Frontiers in Handwriting Recognition*, 2014, pp. 279–284.

[79] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1–32, 1997.

[80] A. Graves and J. Schmidhuber, "Framewise Phoneme Classification with Bidirectional LSTM Networks," in *Proceedings of the International Joint Conference on Neural Networks*, vol. 4, 2005, pp. 2047–2052.

[81] K. Greff, R. K. Srivastava, J. Koutnik, B. Steunebrink, and J. Schmidhuber, "LSTM: A Search Space Odyssey," *arXiv:1503.04069*, 2015.

[82] M. Wöllmer, Z. Zhang, F. Weninger, B. Schuller, and G. Rigoll, "Feature Enhancement by Bidirectional LSTM Networks for Conversational Speech Recognition in Highly Non-Stationary Noise," in *Proc. ICASSP*, 2013, pp. 6822–6826.

[83] B. B. Winter and J. G. Webster, "Driven-Right-Leg Circuit Design," *IEEE Transactions On Biomedical Engineering*, vol. 30, no. 1, pp. 62–66, 1983.

[84] L. Maier-Hein, F. Metze, T. Schultz, and A. Waibel, "Session Independent Non-Audible Speech Recognition using Surface Electromyography," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, San Juan, Puerto Rico, 2005.

[85] D. Heger, F. Putze, C. Amma, M. Wand, I. Plotkin, T. Wielatt, and T. Schultz, "BiosignalsStudio: A Flexible Framework for Biosignal Capturing and Processing," in *KI 2010: Advances in Artificial Intelligence*, 2010, pp. 33–39.

[86] J. Kominek and A. W. Black, "the CMU Arctic Speech Databases," in *Fifth ISCA Workshop on Speech Synthesis*, 2004, pp. 223–224.

[87] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPa TIMIT Acoustic-Phonetic Continous Speech Corpus CD-ROM. NIST Speech Disc 1-1.1," *NASA STI/Recon Technical Report N*, vol. 93, p. 27403, 1993.

[88] Y. Hu and P. C. Loizou, "Evaluation of Objective Quality Measures for Speech Enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 229–238, 2008.

[89] D. Yu, A. Eversole, M. Seltzer, K. Yao, Z. Huang, B. Guenter, O. Kuchaiev, Y. Zhang, F. Seide, H. Wang, J. Droppo, G. Zweig, C. Rossbach, J. Currey, J. Gao, A. May, B. Peng, A. Stolcke, and M. Slaney, "an Introduction to Computational Networks and the Computational Network Toolkit," Microsoft Research, Tech. Rep. MSR-TR-2014-112, 2014.

[90] K. Greff and R. Srivastava, "Brainstorm," 2015. [Online]. Available: https://github.com/IDSIA/brainstorm

[91] F. Weninger and J. Bergmann, "Introducing CURRENNT - the Munich Open-Source CUDa RecurREnt Neural Network Toolkit," *Journal of Machine Learning Research*, vol. 16, pp. 547–551, 2015.

[92] L. Diener, C. Herff, M. Janke, and T. Schultz, "An Initial Investigation into the Real-Time Conversion of Facial Surface EMG Signals to Audible Speech," in *Engineering in Medicine and Biology Society (EMBC), 2016 38th Annual International Conference of the IEEE*, Aug 2016.

[93] R. F. Kubichek, "Mel-Cepstral Distance Measure for Objective Speech Quality Assessment," in *IEEE Pacific Rim Conference on Communications, Computers and Signal Processing*, 1993, pp. 125–128.

[94] K. Kirchhoff, G. A. Fink, and G. Sagerer, "Combining Acoustic and Articulatory Feature Information for Robust Speech Recognition," *Speech Communication*, vol. 37, no. 3-4, pp. 303–319, 2002.

[95] K. Prahallad, A. Vadapalli, N. Elluru, G. Mantena, B. Pulugundla, P. Bhaskararao, H. A. Murthy, S. King, V. Karaiskos, and A. W. Black, "the Blizzard Challenge 2013 - Indian Language Tasks," in *Blizzard Challenge Workshop*, no. 1, 2013.