The EMG-UKA Corpus for Electromyographic Speech Processing

Michael Wand, Matthias Janke, Tanja Schultz

Karlsruhe Institute of Technology, Karlsruhe, Germany
tanja.schultz@kit.edu

Abstract

This article gives an overview of the *EMG-UKA* corpus, a corpus of electromyographic (*EMG*) recordings of articulatory activity enabling speech processing (in particular speech recognition and synthesis) based on EMG signals, with the purpose of building *Silent Speech interfaces*. Data is available in multiple *speaking modes*, namely audibly spoken, whispered, and silently articulated speech. Besides the EMG data, synchronous acoustic data was additionally recorded to serve as a reference. The corpus comprises 63 recorded sessions from 8 speakers, the total amount of data is 7:32 hours. A trial subset, consisting of 1:52 hours of data, is freely available for download.

Index Terms: Electromyography, Speech Data Corpus, Silent Speech Interfaces

1. Introduction

This paper presents the *EMG-UKA* corpus of surface electromyographic (EMG) and acoustic recordings of speech, for the purpose of investigating EMG-based speech processing, in particular speech recognition and synthesis. The corpus contains 7:32 hours of synchronous EMG and acoustic data, coming from 63 recording sessions of 8 speakers, in English language. Data was recorded in three different *speaking modes*, namely audibly spoken, whispered, and silently articulated speech (see section 2.1 for a definition).

This database was created as part of our efforts to develop an EMG-based *Silent Speech interface*, i.e. a system which uses the EMG recordings of articulatory muscle activity to process speech even when no sound is heard or created [1]. Such systems are applicable in situations where nondisturbing, confidential communication is desired, as well as for speech-disabled persons, e.g. laryngectomees.

The electromyographic signal stems from muscle activity: Whenever a muscle contracts, a small electric potential is generated. This signal is captured by electrodes; in our case, we used small, nonintrusive surface electrodes to capture EMG signals from a user's face while speaking. Since this can be done even when no speech sound is heard or created, a Silent Speech interface can be realized based on the EMG signal. Among current Silent Speech technologies, this approach has been judged favorably in terms of usability, non-invasiveness, potential, and cost [1]. The EMG recording is fully *passive*, i.e. an existing biophysiological activity is measured. We never apply any kind of electric current to the human body; thus EMG can be used on an everyday basis, even by unexperienced persons, and outside a laboratory environment.

Based on the EMG-UKA data corpus, our group investigated several key challenges in EMG-based speech recognition and processing, including optimal modeling of articulatory activity [2], intonation generation [3], compensation for variations between recording sessions [4] and discrepancies between *speaking modes*, in particular, between audibly spoken and silently articulated speech [5, 6, 7, 8]. In section 5, we report reference Word Error Rates for several of the recognition scenarios.

A trial subset of the EMG-UKA corpus, consisting of 112 minutes of data, is available to everybody free of charge from http://csl.anthropomatik.kit.edu/EMG-UKA-Corpus.php. By the publication of this dataset, we hope to foster further improvements of the EMG-based Silent Speech interface, as well as to provide a reference database for benchmarking different technical approaches. For the full version of the data, please contact the third author (Tanja Schultz).

This paper is organized as follows: In section 2, we describe the contents of the EMG-UKA corpus, the recording setup is detailed in the subsequent section 3. Section 4 describes the data contained in the corpus distribution, this includes the EMG and acoustic data as well as supplementary information. Section 5 reports baseline Word Error Rates for different EMG recognition tasks, and section 6 concludes the paper.

2. Corpus overview

2.1. Content of the recordings

The EMG-UKA corpus consists of surface electromyographic and acoustic recordings of read speech in English language, from the Broadcast News domain. Data was recorded in three different *speaking modes*, namely audibly spoken, whispered, and silently articulated speech. The recording setup has been in use since 2005 and is the result of extensive studies [9, 10]; it is described in detail in section 3. All data is segmented at the utterance level. We additionally computed word-level and phone-level alignments, which are available together with the EMG and acoustic data.

Audible speech is speech with normal voicing and intonation. Since we always recorded read speech in a quiet environment, it is expected to be free of overarticulation or strong prosodic variation.

Whispered speech is produced when the vocal cords do not vibrate, but are adducted to produce a narrow constriction at the glottis. This results in an excitation of the vocal tract where the normal fundamental frequency is replaced with a "hissing" sound. Note that whispered speech is, of course, also audible, yet in the EMG-UKA corpus we consider audible and whispered speech as clearly distinct speaking modes.

Silent speech means that the speaker performs normal articulatory movements while suppressing the glottal airstream.

The subjects were instructed to produce audible and whispered speech as they felt most natural. Similarly, we asked the subjects to articulate silent speech "as normally as possible", just with suppressed glottal airflow. In order to avoid unnatural speech, more specific instructions were not given. We observed

Full EMG-UKA Corpus			
Speaker	Number of Sessions		
	Total	Large	Multi-Mode
1	3	0	3
2	33	1	15
3	1	0	1
4	2	0	2
5	1	0	1
6	1	0	1
7	2	0	2
8	20	1	7
Total	63	2	32

EMG-UKA Trial Corpus			
Speaker	Number of Sessions		
	Total	Large	Multi-Mode
2	3	1	2
3	1	0	1
6	1	0	1
8	8	0	2
Total	13	1	6

Table 1: *Breakdown of sessions in the full EMG-UKA corpus and the trial subset. Each session contains at least 50 utterances recorded in audible speaking mode. In* multi-mode *sessions, these 50 utterances were recorded three times in all three speaking modes, the two* large *sessions contain more than 500 utterances of audible speech.*

that some phones tended to be slightly audible even in the silent speaking mode (for example, plosives). We did not correct such articulation as long as the content of the silently spoken utterance was not understandable; so silent and whispered speech are clearly distinguished. Our speakers did not speak English natively, however during the recordings we made sure that English words were pronounced correctly.

Note that only a subset of the recorded sessions contain data from all three speaking modes. Whenever multiple speaking modes are present in a session, the recorded utterances use the same text corpus across those speaking modes: We refer to this by the term “parallel utterances”. This scheme facilitates the comparison of speaking modes.

2.2. Recorded speakers and sessions

The EMG-UKA corpus comprises data from eight speakers, all of whom were recruited from the Karlsruhe student population [11]. The speakers were between 24 and 30 years old, seven of them were male, one female. The number of sessions recorded by each of the speakers varies between 1 and 33; in particular, two of the speakers recorded a large number of sessions, as well as one large session with more than 500 utterances each[1]. Out of these eight speakers, four are comprised in the trial subset as well, with a reduced number of sessions. All subjects were informed about the nature of the project and agreed by signing a consent form that their data can be used for further research and distribution. To protect privacy, all data is anonymized, i.e. proper names are replaced by neutral IDs and no information will be made available that links the recordings to individuals.

Table 1 contains a complete list of sessions in the EMG-UKA corpus. All sessions comprise at least a set of 50 utterances in audible speaking mode (*small* sessions). The *multi-mode* sessions contain three repetitions of the same 50 sentences in all three speaking modes (parallel utterances), the *large* sessions consist of 510 resp. 520 utterances of audible speech. The total amount of data in these setups is given in table 2.

Full EMG-UKA Corpus				
Subset	# Spk	# Ses	Avg	Total
Audible (Small)	8	61	3:08	3:11:34
Whispered (Small)	8	32	3:22	1:47:42
Silent (Small)	8	32	3:19	1:46:20
Audible (Large)	2	2	27:02	54:04
Total amount of data: 7:32h				

EMG-UKA Trial Corpus				
Subset	# Spk	# Ses	Avg	Total
Audible (Small)	4	12	3:19	39:47
Whispered (Small)	4	6	3:38	21:47
Silent (Small)	4	6	3:44	22:21
Audible (Large)	1	1	28:29	28:29
Total amount of data: 1:52h				

Table 2: *Data amount in the EMG-UKA corpus ([h:]mm:ss). Each* small *session contains 50 utterances per speaking mode, the two* large *sessions comprise 510 resp. 520 utterances.*

3. Recording setup

EMG data was recorded with a six-channel electrode setup [10]. We used standard gelled Ag/AgCl surface electrodes with a circular recording area having a diameter of 4 mm. Figure 1 shows the positioning of the electrodes, capturing the EMG signal of six articulatory muscles: the levator anguli oris (channels 2, 3), the zygomaticus major (channels 2, 3), the platysma (channels 4, 5), the depressor anguli oris (channel 5), the anterior belly of the digastric (channel 1) and the tongue (channels 1, 6) [9] (also compare [12] for further details).

EMG channels 2 and 6 were derived bipolarly, the other channels used unipolar derivation, with a reference electrode on the nose (channel 1) respectively two connected reference electrodes behind the ears (channels 3, 4, 5). An additional ground electrode was placed on the subject’s wrist. Note that in our experiments, including the ones reported in section 5, we follow [13] in removing channel 5, which tends to yield unstable and artifact-prone signals. The EMG-UKA corpus distribution does, however, contain all six EMG channels.

The recordings were performed with the portable *Varioport* biosignal recorder (Becker Meditec, Germany). Technical specifications include an amplification factor of 1170, 16 bits A/D conversion, a resolution of 0.033 microvolts per bit, and a frequency range of 0.9-295 Hz. EMG signals were sampled with a 600 Hz sampling rate. Recordings were performed in a push-to-talk setting and were controlled with the inhouse *UKA EEG-EMG Studio* software [14], they were performed in quiet rooms, but without electrical shielding: We expect this to be closer to real-life usage than using a specialized recording room.

Acoustic data was recorded with a standard close-talking microphone connected to a stereo USB soundcard, with a 16 kHz sampling rate. The acoustic data was recorded in stereo format, the first channel contains the acoustic signal, and the second channel contains a *marker* signal for synchronization.

Synchronization of EMG and acoustic data was performed

[1] Having a large number of sessions by one and the same speaker is a requirement for doing experiments on session independent systems (see section 5).

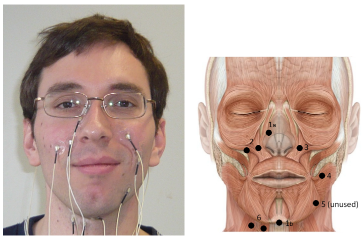

Figure 1: *Electrode positioning for the EMG-UKA corpus (muscle chart adapted from [15])*

with a hardware marker signal which is saved as the seventh channel with the EMG data and as the second (stereo) channel with the acoustic data, respectively. The marker signal appears as a binary signal in the EMG data, and as an analog signal in the audio data, see figure 2. The first peak of the marker is to be used for synchronization. For easier usage, we precomputed the location of the synchronization signals in terms of samples, this data is included in the corpus distribution (see section 4).

4. Data formats and supplementary files

Audio and EMG data are saved in raw, uncompressed format. Additionally, the corpus distribution comprises diverse supplementary data which may be used as the basis for future experiments:

- **Offset** files with precomputed synchronization points, as described above.
- Standard **subsets** for training and testing, as used in our experiments (see section 5). The 50-sentence sessions are subdivided into a training data set of 40 utterances and a test set of 10 utterances, where the textual content of training and test data is always different, and the test set is identical across sessions. The large sessions comprise 13 resp. 20 test utterances.
- **Transcriptions** of all recorded utterances.
- Phone-level **alignments** of all data. These alignments were computed from the synchronous acoustic data in the case of audible and whispered speech [13], and from the EMG data in the case of silent speech, according to the *cross-modal labeling* approach from [5]. All alignments are provided on an "AS IS" basis. Due to their different creation methods, it is assumed that the alignments of the silent EMG data are less accurate than the ones on audible and whispered recordings.
- A pronunciation **dictionary** of the corpus, and a list of used phones.

5. Reference recognition results

This section summarizes baseline recognition results for the corpus. We first give a very brief description of our recognition system, more details can be found in the referenced literature. We then report results on *session dependent* speech recognition on audible, whispered, and silent speech EMG data, and on *session independent* speech recognition on audible speech EMG data. A session dependent system is characterized by using training and test data from a single recording session; session independent systems [4] are trained on data from several recording sessions and tested on the test data of an *unseen* session.

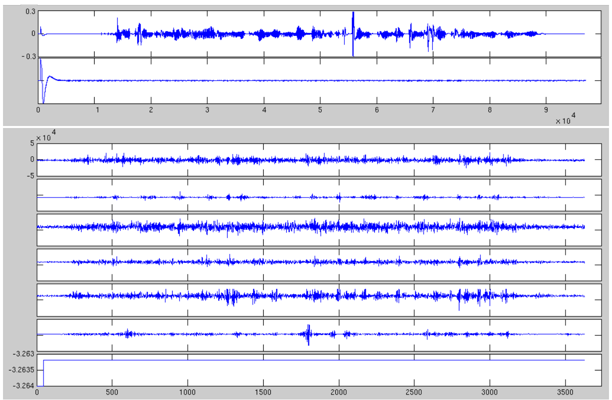

Figure 2: *Example audio signal (above) and EMG signal (below) for speaker 2, session 1, utterance 0100, with marker signal (second audio/7th EMG channel). Note the differing sampling rates: Acoustic data was sampled with 16kHz, EMG data was sampled with 600Hz.*

5.1. EMG features

Our standard EMG features are *time-domain features* [13]: For a time series $\mathbf{x}$, $\bar{\mathbf{x}}$ is its frame-based time-domain mean, $\mathbf{P_x}$ is its frame-based power, and $\mathbf{z_x}$ is its frame-based zero-crossing rate. For a frame-based feature $\mathbf{f}$, $S(\mathbf{f}, n)$ is the stacking of $2n + 1$ ($-n$ to n) adjacent frames.

For an EMG signal with normalized mean $x[n]$, the nine-point double-averaged signal $w[n]$ is defined as

$$w[n] = \frac{1}{9} \sum_{k=-4}^{4} v[n+k], \quad \text{where} \quad v[n] = \frac{1}{9} \sum_{k=-4}^{4} x[n+k].$$

The high-frequency signal is $p[n] = x[n] - w[n]$, and the rectified high-frequency signal is $r[n] = |p[n]|$. The final time-domain feature $\mathbf{TD}10$ is defined as follows:

$$\mathbf{TD}10 = S(\mathbf{TD}0, n)$$

where $\mathbf{TD}0 = [\bar{\mathbf{w}}, \mathbf{P_w}, \mathbf{P_r}, \mathbf{z_p}, \bar{\mathbf{r}}]$.

The $\mathbf{TD}10$ feature is computed for each channel, and all channel-wise feature vectors are combined. Frame size and frame shift are set to 27ms respective 10ms, note that when time alignments are to be taken from the acoustic data, the frame shifts applied to the acoustic data and the EMG data must match. In particular, the alignments which are comprised in the EMG-UKA Trial corpus distribution are based on a 10ms frameshift. For the transfer of the alignments from acoustic data to EMG data, we used the hardware synchronization signal contained in the recordings (see section 3), and we delayed the EMG signal by 50ms according to [13].

We always apply Linear Discriminant Analysis (LDA) on the $\mathbf{TD}10$ feature. The LDA matrix is computed by dividing the training data into 136 classes corresponding to the begin, middle, and end parts of the 45 English phones, plus one silence phone. From the 135 dimensions after LDA, we retain 12 dimensions for the session dependent systems, and 32 dimensions for the session independent systems.

5.2. Recognizer setup

Our EMG-based speech recognizer performs continuous speech recognition based on tristate Hidden Markov Models (HMM).

Session Type	WER on Audible EMG
Small	22.82%
Large	12.00%

Table 3: *Average session dependent Word Error Rates (WER) on audible speech EMG data, for small and large sessions.*

Each word is composed from its phones, which are taken from the pronunciation dictionary, each phone has three substates (begin, middle, end).

The emission probabilities for the HMM are based on multi-stream *Bundled Phonetic Features* (BDPF) [2]. A Phonetic Feature (PF) stream is a knowledge source corresponding to a phonetic (or articulatory) feature [16], which is a binary-valued property of a phone, like the place or manner of articulation: For example, each of the places of articulation *Glottal*, *Palatal*, …, *Bilabial* is a phonetic feature which may or may not be present. The key feature of our BDPF approach, detailed in [2], is the modeling of dependencies between PFs using a decision-tree approach, hence, we obtain *Bundled* Phonetic Features. Several BDPF knowledge sources are combined to yield the final emission log probability, this structure is called a *multi-stream* model [17].

The experiments reported in this section are based on eight BDPF streams, chosen to correspond to the most frequent phonetic features in the corpus. Each BDPF stream uses a decision tree with a fixed number of 120 leaves. Phonetic context questions about the left and right neighboring phones (i.e. up to a context width of 1) are allowed. All BDPF streams receive equal weights of $1/8$, no phone stream is used.

5.3. Training and decoding

For bootstrapping the recognizer (including the LDA computation), we use the phone-level time alignments from the corpus (see section 4). Training comprises intializing context-independent, unbundled phone and phonetic feature models, performing phonetic feature bundling, and retraining with the newly created models, according to the recipe in [2, section 3.3].

Decoding uses the trained *myoelectric model* together with a trigram BN language model. The test set perplexity is 24.24. Lattice rescoring is *not* applied. The recognition vocabulary is restricted to the 108 words appearing in the test set; this is our standard procedure for the small session dependent systems, where only a few minutes of training data is available. Larger (session dependent and session independent) systems also enable larger vocabularies, see [4].

All systems use a fixed training/test data subdivision (see section 4). The amount of training data is 40 utterances for all sessions except for the two "large" ones by speakers 2 and 8, where around 500 utterances are used.

5.4. Session-dependent recognition of audible speech

Table 3 shows average (test set) Word Error Rates of the 63 session dependent recognizers on audible EMG data. It can be seen that the large sessions, where more training data is available, perform substantially better than the 40-sentence sessions.

5.5. Session-dependent recognition of audible, whispered, and silent speech

Table 4 lists the Word Error Rates of our recognizer on the different speaking modes, for *session dependent* recognition. All systems are *mode dependent*: Training and testing was always performed on data from only one speaking mode. Therefore, each system was trained on 40 training utterances. It is observed that silent speech is recognized with higher error rate than whispered or audible speech, which is in particular due to a few speakers who were observed to find silent articulation especially difficult.

Speaker	WER by Speaking Mode		
	Audible	Whispered	Silent
1	43.43%	38.37%	92.23%
2	22.23%	23.24%	32.80%
3	64.60%	61.60%	99.00%
4	27.75%	30.30%	61.10%
5	37.40%	30.30%	80.80%
6	11.10%	16.20%	61.60%
7	37.85%	52.00%	76.25%
8	21.81%	23.67%	37.23%
Total	26.90%	28.19%	48.29%

Table 4: *Session-dependent Word Error Rates (WER) for all three speaking modes, averaged over sessions per speaker. For the numbers on audible EMG, only the sessions which also contain the other speaking modes are considered. Also note that the total WERs are computed on a per-session basis, not on a per-speaker basis.*

5.6. Session-independent recognition

Block	Session-independent WER
Speaker 2, Block 1	15.61%
Speaker 2, Block 2	49.30%
Speaker 8	16.16%
Average	27.02%

Table 5: *Average Word Error Rates (WER) in the session independent setup. See text for details.*

For the session independent system, only audible EMG data from the small sessions was used for training and testing. The sessions were subdivided as follows: We defined three blocks of 16 consecutive sessions, namely two blocks for speaker 2 (sessions 1 …16 and 17 …32), and one block for speaker 8, taking the first 16 sessions. For each block we trained 16 systems, where one session was designated as *target* session, and training was performed on the training data of the remaining sessions. Thus we get 16 different setups per block.

The average Word Error Rates on the test sets of the target sessions are given in table 5. We observe that for two out of three blocks, the WER is excellent even in the session independent setup, where no training data from the test session is used at all. We showed that this system can be further improved by both supervised or unsupervised *session adaptation* [4, 18, 19].

6. Conclusion

During the past decade, Silent Speech interfaces have become a major topic of research. The electromyographic approach is of particular interest, since it is highly developed, particularly regarding continuous speech recognition. In this paper we presented the *EMG-UKA* corpus, which has been used as the basis for a large number of investigations conducted by our group. A trial subset of 1:52 hours is freely available for download, serving as a starting point for other Silent Speech researchers, and as a benchmark database.

7. References

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

[2] T. Schultz and M. Wand, “Modeling Coarticulation in Large Vocabulary EMG-based Speech Recognition,” *Speech Communication*, vol. 52, no. 4, pp. 341 – 353, 2010.

[3] K. Nakamura, M. Janke, M. Wand, and T. Schultz, “Estimation of Fundamental Frequency from Surface Electromyographic Data: EMG-to-F0,” in *Proc. ICASSP*, 2011, pp. 573 – 576.

[4] M. Wand and T. Schultz, “Session-independent EMG-based Speech Recognition,” in *Proc. Biosignals*, 2011, pp. 295 – 300.

[5] M. Janke, M. Wand, and T. Schultz, “A Spectral Mapping Method for EMG-based Recognition of Silent Speech,” in *Proc. B-INTERFACE*, 2010, pp. 22 – 31.

[6] ——, “Impact of Lack of Acoustic Feedback in EMG-based Silent Speech Recognition,” in *Proc. Interspeech*, 2010, pp. 2686 – 2689.

[7] M. Wand, M. Janke, and T. Schultz, “Investigations on Speaking Mode Discrepancies in EMG-based Speech Recognition,” in *Proc. Interspeech*, 2011, pp. 601 – 604.

[8] ——, “Decision-Tree based Analysis of Speaking Mode Discrepancies in EMG-based Speech Recognition,” in *Proc. Biosignals*, 2012, pp. 101 – 109.

[9] L. Maier-Hein, “Speech Recognition Using Surface Electromyography,” Diploma thesis, Interactive Systems Labs, University of Karlsruhe, 2005.

[10] L. Maier-Hein, F. Metze, T. Schultz, and A. Waibel, “Session Independent Non-Audible Speech Recognition Using Surface Electromyography,” in *Proc. ASRU*, 2005, pp. 331 – 336.

[11] M. Janke, “Spektrale Methoden zur EMG-basierten Erkennung lautloser Sprache,” Diploma Thesis, Cognitive Systems Lab, Karlsruhe Institute of Technology, 2010.

[12] UCLA Phonetics Laboratory, “Dissection of the Speech Production Mechanism,” Department of Linguistics, University of California, Tech. Rep., 2002, available online: http://www.linguistics.ucla.edu/people/ladefoge/manual.htm.

[13] 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.

[14] C. Mayer, “UKA EMG/EEG Studio v2.0.”

[15] M. Schünke, E. Schulte, and U. Schumacher, *Prometheus - Lernatlas der Anatomie*. Stuttgart, New York: Thieme Verlag, 2006, vol. [3]: Kopf und Neuroanatomie.

[16] K. Kirchhoff, “Robust Speech Recognition Using Articulatory Information,” Dissertation, University of Bielefeld, 1999.

[17] F. Metze, “Articulatory Features for Conversational Speech Recognition,” Dissertation, University of Karlsruhe, 2005.

[18] M. Wand and T. Schultz, “Towards Real-life Application of EMG-based Speech Recognition by using Unsupervised Adaptation,” in *Proc. Interspeech*, 2014.

[19] M. Wand, “Advancing Electromyographic Continuous Speech Recognition: Signal Preprocessing and Modeling,” Dissertation, Karlsruhe Institute of Technology, 2014.