EMG-UKA Trial Corpus

Michael Wand, Matthias Janke, Tanja Schultz

tanja.schultz@kit.edu
http://csl.anthropomatik.kit.edu

1. Introduction

The *EMG-UKA* corpus comprises surface electromyographic (EMG) and acoustic recordings of speech, for the purpose of performing speech processing (in particular speech recognition and synthesis) based on EMG signals. Data was recorded in the *audible*, *whispered*, and *silent* speaking modes, see section 3 for details.

This distribution, the *EMG-UKA Trial corpus*, contains a subset of around 1:52 hours of data from 4 speakers. The full corpus comprises 7:40 hours of data, subdivided into 63 sessions (including two "large" ones with more than 25 minutes of data) by 8 speakers.

Table 1 gives an overview of these figures. For detailed information about the trial subset please refer to section 4. For inquiries regarding the full version of the data, please contact Tanja Schultz (`tanja.schultz@kit.edu`).

	Full EMG-UKA Corpus				**EMG-UKA Trial Corpus**			
Subset	# Spk	# Ses	Avg	Total	# Spk	# Ses	Avg	Total
Audible (Small)	8	61	3:08	3:11:34	4	12	3:19	39:47
Whispered (Small)	8	32	3:22	1:47:42	4	6	3:38	21:47
Silent (Small)	8	32	3:19	1:46:20	4	6	3:44	22:21
Audible (Large)	2	2	27:02	54:04	1	1	28:29	28:29
All data	**7:39:40**				**1:52:24**			

Table 1 – 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.

2. Setup of the Recordings

EMG data was recorded with a six-channel electrode setup, developed by L. Maier-Hein [1]. We used standard gelled Ag/AgCl surface electrodes with a circular recording area having a diameter of 4 mm. Figure 1 shows how the electrodes were positioned, capturing the EMG signal of six articular 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, 7) [2] (compare [3] 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 7, we follow [4] in removing channel 5, which tends to yield unstable and artifact-prone signals. The EMG-UKA Trial corpus distribution does, however, contain all six EMG channels.

The recordings were performed with the *Varioport* biosignal recorder (Becker Meditec, Germany). This recorder is battery-powered and includes an electrical insulation device to separate the electric currents of the amplifier and the controlling computer. 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.

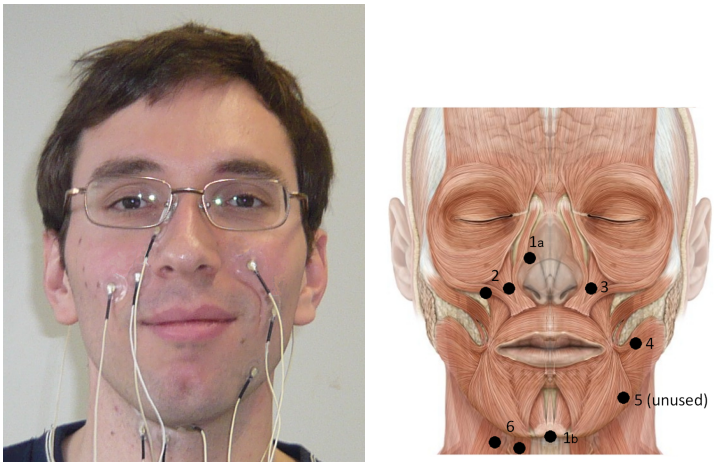

Figure 1 – Electrode positioning for the EMG-UKA corpus (muscle chart adapted from [5]). Channels are enumerated, the reference electrodes for the unipolar channels 3, 4, and 5 (behind the ears) are not shown.

Acoustic data was recorded with a standard close-talking microphone connected to a stereo USB soundcard, with a 16 kHz sampling rate. Note that the acoustic data was recorded in stereo format, where 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 with a hardware *marker* signal which is stored 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. In both cases, the first peak of the marker is to be used for synchronization; it marks the same point of time in both signals. For easier usage, we precomputed the location of the synchronization signals in terms of samples, this information it is found in the `offset` directory of the corpus. See section 6 for details.

Recordings 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. All recordings were supervised by CSL researchers and qualified recording assistants; all subjects were recruited from the Karlsruhe student population. Therefore, the subjects were not native speakers of English, however we made sure that they pronounced the recorded sentences correctly. The 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 was anonymized, i.e. proper names are replaced by neutral IDs, and no information will be made available that links the recordings to individuals.

3. Speech Data and Speaking Modes

The EMG-UKA data comprises three *speaking modes*: Audible, whispered, and silent speech. Here *silent speech* is defined as silent articulation: The user should perform the normal articulatory movements while suppressing the glottal airstream.

In all three speaking modes, we recorded read speech in a push-to-talk scenario: The subject was shown a text prompt on a computer screen, and started/stopped the recording by mouse click. The text corpus was taken from the broadcast news (BN) domain. Note that only a subset of the 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 of parallel utterances facilitates the comparison of speaking modes.

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", while suppressing the air flow through the glottis. In order to avoid unnatural speech across the modes, more specific instructions were not given. We observed 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 spoken utterance was not understandable; this clearly distinguishes silent and whispered speech.

4. Corpus Overview and Statistics

The EMG-UKA Trial distribution contains data from 13 sessions by 4 speakers, aged between 24 and 30 years at the time of recording. The data was chosen to allow the reproduction of a variety of our published experiments. In particular, a subset of six session contains data from all three speaking modes, so that experiments on speaking mode differences as in [6, 7, 8] are possible, and we included a set of eight sessions by one speaker (8) and a large session (2-101) with more than 27 minutes of training data to allow for multi-session and single-session experiments on extended amounts of training data [9].

Each session is divided into training and test data. This data subdivision forms the basis for all our experiments, since the recognition accuracy depends on the amount of words in the recognition vocabulary, it is recommended to keep this subdivision unchanged. The content of the training data (partially) varies between sessions, whereas the test data always consists of the same 10 sentences.

All sessions except 2-101 contain 40 training sentences and 10 test sentences, all of which are unique. Session 2-101 contains 500 training sentences and 20 test sentences, where the test set consists of the same 10 sentences as for the smaller sessions, repeated twice. Note that the session ID 101 is reserved for the large session, whereas all other sessions are enumerated consecutively in order of recording, starting from 1. Training and test sentences were always recorded in randomized order.

Table 2 shows a breakdown of session durations.

Session	Audible		Whispered		Silent	
	Train	Test	Train	Test	Train	Test
2-1	178	46	172	45	179	48
2-3	174	44	177	46	183	47
2-101	1630	79	0	0	0	0
4-1	172	46	173	44	181	46
6-1	188	51	189	52	189	49
8-1	168	45	0	0	0	0
8-2	158	41	164	44	170	44
8-3	154	42	157	44	159	46
8-4	143	40	0	0	0	0
8-5	136	39	0	0	0	0
8-6	135	39	0	0	0	0
8-7	134	38	0	0	0	0
8-8	137	39	0	0	0	0
Sum	3507	589	1032	275	1061	280

Table 2 – Breakdown of session durations in the EMG-UKA Trial corpus, in seconds. The session ID 101 designates a "large" session.

5. Directory Structure and List of Files

For each utterance of the corpus, we define a speaker ID (e.g. 002), a session id (e.g. 003), and an utterance ID (e.g. 0100). These IDs appear in the file names and supplementary data of the corpus distribution, all three IDs taken together uniquely identify an utterance.

The EMG-UKA distribution is structured as follows. In the main folder, there are the following directories.

- `audio` – Audio data files
- `emg` – EMG data files
- `offset` – Offset files with precomputed alignments for synchronization

- `Subsets` – Lists of data for training and testing in different speaking modes
- `Alignments` – Phone-level and word-level alignments of all the data
- `Transcripts` – Data transcriptions
- `Supplementary` – Additional information.

The data in the `audio`, `emg`, `offset`, `Alignments`, and `Transcripts` directories is always subdivided into subdirectories according to speaker and session, see section 4. Each audio file follows the naming scheme `a_<speaker>_<session>_<utterance>.wav`, e.g. `a_002_003_0100.wav`. Similarly, each EMG file is named `e07_<speaker>_<session>_<utterance>.adc`, e.g. `e07_002_003_0100.adc` (the `07` indicates the number of channels). Each offset file is named `offset_<speaker>_<session>_<utterance>.txt`, e.g. `offset_002_003_0100.txt`.

The alignment files are named

- `phones_<speaker>_<session>_<utterance>.txt` (for phone-level alignments)
- `words_<speaker>_<session>_<utterance>.txt` (for word-level alignments)

Similarly, the transcript files are named `transcript_<speaker>_<session>_<utterance>.txt`.

The `Subsets` directory contains six files `{train|test}.{audible|whispered|silent}` which give a standard division of the data set into training and test data.

The data formats of all files in the directories mentioned so far are described below in section 6.

The `Supplementary` directory contains the additional files

- `dictionary.full|test` – a pronunciation dictionary covering all words appearing in the corpus, and a dictionary of all words appearing in the test set.
- `phoneList` – a list of all phones used in the dictionaries.

6. Data Formats

We use the following data formats:

- **Audio Data** is saved in RIFF Wave format with 16kHz sampling rate and two channels (i.e. stereo), where the first channel contains the actual signal, and the second channel contains the hardware marker signal.
- **EMG Data** is saved in raw, uncompressed, headerless, little endian short integer (16 bit) format. We use the file suffix `adc` (analog/digital converted). *All EMG data files comprise seven channels*, which includes the marker signal. Samples are saved consecutively, i.e. bytes 1 to 14 of any file contain the first sample, bytes 15 to 28 contain the second sample, and so on. As an example, ADC data could be read using MATLAB as follows:

  ```
  fid = fopen(EMGFile.adc,'r','n');
  channels = 7;
  ADC = fread(fid,[channels,inf], '*int16');
  fclose(fid)
  ```

  Now each *row* of the variable `ADC` contains an EMG channel.
- **Offset** files contain two lines with two numbers each: The first line contains the beginning and end offsets for acoustics, the second line contains the beginning and end offsets for EMG, both in samples. Note that these numbers are typically vastly different due to the difference in sampling rates, see below for an example. Also note that the beginning offsets are shifted by 0.2 seconds for both acoustics and EMG in order to cut out the marker signal, which sometimes causes an artifact in the first audio channel.
- **Subsets** are given in text format, with one line per session: Each line gives the session ID and the corresponding utterances, separated by a colon.

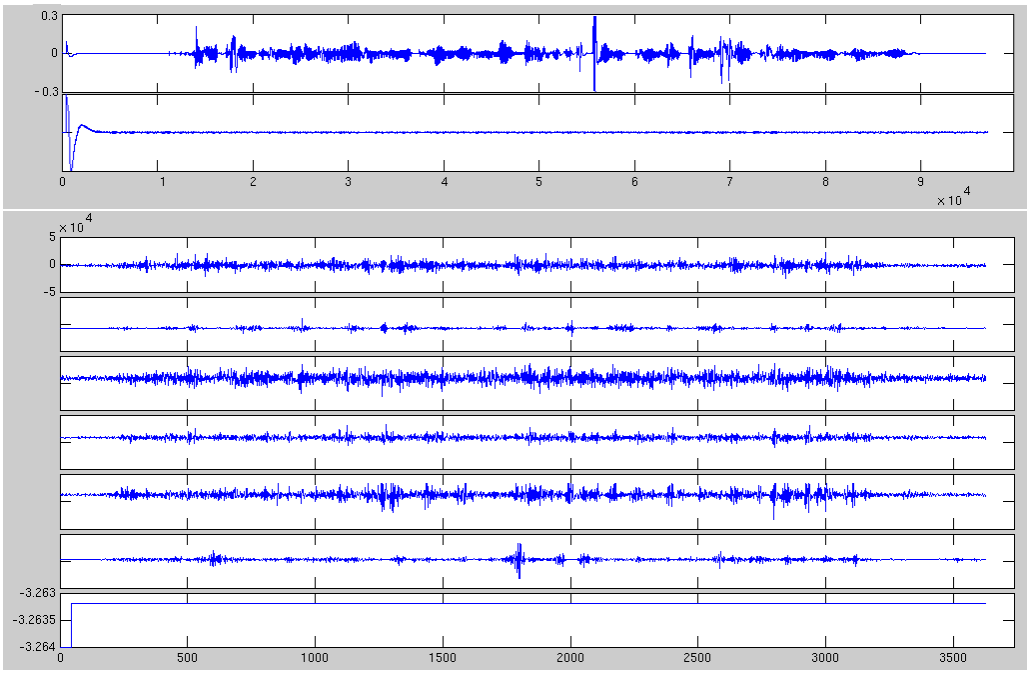

Figure 2 – Example audio signal (above) and EMG signal (below) for ID 002-001-0100. Note the differing sampling rates: Acoustic data was sampled with 16kHz, EMG data was sampled with 600Hz.

- **Alignments** are rather self-explanatory, they essentially use the TIMIT format; see below for an example. However note that in these files, we do *not* use samples as the basic unit of time: Instead, we apply our standard preprocessing (as one would do for a subsequent recognition step anyway), which comprises *cutting* the EMG signal as given by the offset file, and applying a *framing* with a 10 ms frameshift. The preprocessing is defined in detail in section 7.

As an example, figure 2 shows the audio and EMG data for utterance 002-001-0100. Note the different sampling rates: The EMG signal contains 3628 samples, the acoustic signal contains 97024 samples. Given a sampling rate of 600 Hz vs 16 kHz, this amounts to around 6.05 seconds. Also note that the marker position differs, which can be seen from the file `offset_002_001_0100.txt`: the boundaries are 3669 ... 96014 for acoustics, and 165 ... 3627 for EMG. While the starting points are computed from the marker (and shifted by 0.2 seconds so that the marker signal itself is cut out), the end points are computed so that the EMG and audio signal parts have the same length: In the example, there remain 3462 samples for EMG and 92345 samples for audio, which is 5.77 seconds.

Finally, consider the alignment files for this utterance. The word-level alignment is given in the file `Alignments/002/001/words_002_001_0100.txt`. It reads as follows:

```
0 49 $
50 79 THIS
80 123 COUNTRY
124 143 HAS
144 176 RELIED
177 186 ON
187 245 IMMIGRANTS
246 256 AND
257 271 IS
272 312 FOUNDED
313 341 UPON
342 346 A
347 396 PRINCIPLE
397 414 OF
```

```
415 467 WELCOMING
468 538 IMMIGRANTS
539 575 $
```

As described above, we see that the boundaries are given in *frames*, where only the signal part given by the offset file is considered, and framing was performed with a 10ms frameshift: hence, there are 575 frames, corresponding to 5.77 seconds with a slight rounding error. The word `$` is the silent word, it contains the single phone `SIL` (refer to the dictionary files). The phone-level alignment for this utterance begins with

```
0 49 SIL
50 64 DH
65 70 IH
71 79 S
80 88 K
89 97 AH
98 103 N
104 110 T
111 117 R
118 123 IY
...
```

The correct phonetic pronunciations are taken from the dictionary: For example, the word "this" is pronounced `DH IH S`.

Finally, note that these alignments have *not* been hand-created. On the audible and whispered data, they were created by forced-aligning the audio recordings of the corpus with a standard Broadcast News acoustic speech recognizer [4]. On the silent data, we computed alignments using session-dependent EMG-based recognizers trained on the corresponding audible (training) data, according to the *cross-modal labeling* approach from [6]. All alignments are provided on an "AS IS" basis. Due to their different creation methods, it is assumed that the alignments of the silent data are less accurate than the ones on audible and whispered recordings.

7. Baseline Results

This section summarizes baseline recognition results for the corpus. We first give a very brief description of our recognition system, the reader is suggested to consider the referenced literature for details. We then report results on the following experiments: Session-dependent speech recognition on audible, whispered, and silent EMG data, and session-independent recognition using the eight sessions of speaker 8 in a leave-one-out setup. A *session-independent* system needs to cope with variations in electrode positioning, as well as with changing recording conditions (e.g. environmental artifacts, varying skin properties, etc.). This is a challenge which is not present in acoustic speech recognition, it has been shown that it negatively affects the recognition accuracy [9]. A *session-dependent* system is characterized by only using training and test data from a single recording session.

EMG features

Our standard EMG features are *time-domain features* [4]. They are computed as follows:

For any given 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 given frame-based feature $\mathbf{f}$, $S(\mathbf{f}, n)$ is the stacking of adjacent frames in the size of $2n + 1$ ($-n$ to n) 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 **TD**10 is defined as follows:

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

where

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

i.e. a stacking of adjacent feature vectors with context $2 \cdot 10 + 1 = 21$ is performed. This context width is shown to be optimal in [10].

This process is performed for each channel, and all channel-wise feature vectors are combined. Frame size and frame shift are set to 27 ms respective 10 ms, 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.

We always apply Linear Discriminant Analysis (LDA) on the **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 from the pronunciation dictionary, plus one silence phone. From the 135 dimensions which the LDA algorithm yields, we retain 12 dimensions for the session-dependent systems, and 32 dimensions for the session-independent systems.

Recognizer setup

The baseline EMG-based speech recognizer performs continuous speech recognition based on tristate Hidden Markov Models (HMM). 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) [11]. A Phonetic Feature (PF) stream is a knowledge source corresponding to a phonetic (or articulatory) feature [12], which is a binary-valued property of a phone, like the place or manner of articulation: For example, each of the places or articulation *Glottal*, *Palatal*, ..., *Bilabial* is a phonetic feature which may be true (present) or false (not present). The key feature of our BDPF approach, detailed in [11], is the modeling of dependencies between PFs using a decision-tree approach, hence, we obtain *Bundled* Phonetic Features. Several BDPF knowledge sources are combined by weighted summation to yield the final emission log probability, this structure is called a *multi-stream* model [13].

The experiments reported in this section are based on eight BDPF streams, chosen to correspond to the most frequent phonetic features in the corpus [10]. 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 stream receive equal weights of 1/8, no phone stream is used.

For bootstrapping the recognizer (including the LDA computation), we use the phone-level time alignments which are comprised in the EMG-UKA corpus, as described in section 6, they were created based on the acoustic data for the audible and whispered speaking modes (see [4]), and based on the EMG data for the silent speaking mode (see [6]). 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 [11, section 3.3]. The first and third training steps use split-and-merge training to initialize models [14], followed by four iteration of Viterbi training. Note that during the first step, phone models are still required in order to obtain suitable forced alignment during the Viterbi steps. We only discard the phone models at the very end of the training phase.

Decoding uses the trained *myoelectric model* together with a trigram BN language model. The test set perplexity is 24.24. 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. This vocabulary can be found in the file `Supplementary/dictionary.test`. Larger session-dependent and session-independent systems also enable larger vocabularies, see [9] for details and reference Word Error Rates.

Session-dependent recognition of audible speech based on EMG data

Table 3 lists the Word Error Rates of our baseline recognizer on the different speaking modes, for *session-dependent* and *mode-dependent* recognition, i.e. all systems are trained and tested on data from only one session and one speaking mode.

Session (Train/Test)	Word Error Rate by Speaking Mode		
	Audible	Whispered	Silent
2-1	21.20%	6.10%	27.30%
2-3	12.10%	17.20%	11.10%
2-101	11.60%	-	-
4-1	34.30%	27.30%	68.70%
6-1	11.10%	16.20%	61.60%
8-1	14.10%	-	-
8-2	27.30%	39.40%	43.40%
8-3	30.30%	18.20%	40.40%
8-4	26.30%	-	-
8-5	17.20%	-	-
8-6	19.20%	-	-
8-7	17.20%	-	-
8-8	22.20%	-	-

Table 3 – Test set Word Error Rates in the session-dependent setup

Training and test data was taken from the corresponding data lists in the `Subsets` directory. Decoding was performed based on the 108-word vocabulary containing exactly the words and variants appearing in the test set (see the file `dictionary.test`), together with a trigram Broadcast News language model. We emphasize that *all* experiments *ever* use only EMG data for the recognition; the audio data is only used for writing alignments, but *never* to train the recognizer. Lattice rescoring is *not* applied.

Session-independent recognition

For the session-independent system, only audible EMG data is used for training and testing. Each system is trained using the training data of seven sessions from speaker 601, and then tested on the test data of the remaining session. The system setup is as above, with the exception that we use 32 components after LDA, which yields better results on this particular set of sessions. The Word Error rates are given in table 4.

Training Sessions	Test Session	Word Error Rate
8-2, ..., 8-8	8-1	40.40%
8-1, 8-3, ..., 8-8	8-2	30.30%
8-1, 8-2, 8-4, ..., 8-8	8-3	17.20%
8-1, ..., 8-3, 8-5, ..., 8-8	8-4	7.10%
8-1, ..., 8-4, 8-6, ..., 8-8	8-5	1.00%
8-1, ..., 8-5, 8-7, 8-8	8-6	11.10%
8-1, ..., 8-6, 8-8	8-7	6.10%
8-1, ..., 8-7	8-8	8.10%

Table 4 – Test set Word Error Rates in the session-*independent* setup

References

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

[2] L. Maier-Hein, "Speech Recognition Using Surface Electromyography," diploma thesis, Interactive Systems Labs, University of Karlsruhe, 2005.

[3] UCLA Phonetics Laboratory, "Dissection of the Speech Production Mechanism," tech. rep., Department of Linguistics, University of California, 2002. Available online: http://www.linguistics.ucla.edu/people/ladefoge/manual.htm.

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

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

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

[7] M. Janke, M. Wand, and T. Schultz, "Impact of Lack of Acoustic Feedback in EMG-based Silent Speech Recognition," in *Proc. Interspeech*, pp. 2686 – 2689, 2010.

[8] M. Wand, M. Janke, and T. Schultz, "Decision-Tree based Analysis of Speaking Mode Discrepancies in EMG-based Speech Recognition," in *Proc. Biosignals*, pp. 101 – 109, 2012.

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

[10] M. Wand, *Advancing Electromyographic Continuous Speech Recognition: Signal Preprocessing and Model*. Dissertation, Karlsruhe Institute of Technology, 2014.

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

[12] K. Kirchhoff, *Robust Speech Recognition Using Articulatory Information*. Dissertation, University of Bielefeld, 1999.

[13] F. Metze, *Articulatory Features for Conversational Speech Recognition*. Dissertation, University of Karlsruhe, 2005.

[14] N. Ueda, R. Nakano, Z. Ghahramani, and G. E. Hinton, "Split and Merge EM Algorithm for Improving Gaussian Mixture Density Estimates," *Journal of VLSI Signal Processing*, vol. 26, pp. 133 – 140, 2000.