Towards Real-life Application of EMG-based Speech Recognition by using Unsupervised Adaptation

Michael Wand, Tanja Schultz

Karlsruhe Institute of Technology, Karlsruhe, Germany

michael.wand@kit.edu, tanja.schultz@kit.edu

Abstract

This paper deals with a *Silent Speech Interface* based on Surface Electromyography (EMG), where electrodes capture the electric activity generated by the articulatory muscles from a user's face in order to decode the underlying speech, allowing speech to be recognized even when no sound is heard or created. So far, most EMG-based speech recognizers described in literature do not allow electrode reattachment between system training and usage, which we consider unsuitable for practical applications. In this study we report on our research on *unsupervised session adaptation*: A system is pre-trained with data from multiple recording sessions and then adapted towards the current recording session using data accruable during normal use, without requiring a time-consuming specific enrollment phase. We show that considerable accuracy improvements can be achieved with this method, paving the way towards real-life applications of the technology.

Index Terms: Silent Speech Interfaces, EMG, EMG-based Speech Recognition, Unsupervised Adaptation

1. Introduction

Speech is a key capability of human beings and thus of tremendous importance in our daily lives. However, voice-based communication and interaction exhibit several challenges which originate from the fact that speech must be loudly spoken and cannot be easily shielded. The main issues identified by the highly recommended overview article [1] are interference with the environment (bystanders are disturbed, and private communication is impossible) and exclusion of speech-disabled people (for example laryngectomees).

Silent Speech interfaces (SSI) [1] are systems enabling speech communication to take place when an audible acoustic signal is undesired or unavailable. This addresses both situations described above, namely covert communication in public places and communication for speech-disabled patients. Additionally, Silent Speech technology may be used to augment conventional speech recognition, for example when ambient noise masks the acoustic signal.

In this paper we report on a Silent Speech recognizer based on Surface Electromyography (EMG), where the electric activity of the user's facial muscles is captured by surface electrodes. This allows speech to be recognized even when the user only moves his or her mouth without generating audible sound. While several such systems have been described, they frequently suffer from problems in practical situations since they only apply to *session-dependent* scenarios, i.e. no repositioning of electrodes between the recording of training and evaluation data is possible. We consider this setup unsatisfatory, since in practice, a user will almost always desire to immediately apply the system when the need for communication arises, without requiring an enrollment phase before use. Yet, it is known from prior work (e.g. [2, 3]) that a recognizer severely degrades when training and test data stem from different recording sessions, i.e. reattachment of the electrodes took place.

In this paper we present results on a session-independent recognizer using *unsupervised adaptation*. Our system is pretrained on a set of recording sessions which does not contain the *target* session on which the system is to be tested. Data from the target session is used to *adapt* the recognizer towards this session, but without making assumptions about this adaptation data: in particular, the user is not required to record a set of specific, predetermined sentences as for normal training (hence, *unsupervised*). This means that in a real-world application, the adaptation data can be accrued during normal usage of the system, substantially extending and improving upon our previous study [3], where initial results on session adaptation using a *supervised* approach are reported: That method requires the user to speak a set of specific enrollment sentences whenever the recording electrodes are reattached.

2. Related Work

EMG-based speech recognition started in the 80's with the studies of Sugie and Tsunoda [4, 5], and Jorgensen and colleagues remarked that the method works even for non-audible speech [6]. The first system to recognize *continuous* speech using only EMG signals was presented by our research group in 2006 [7]. It was based on context-independent phone models; our *Bundled Phonetic Feature* models [8] substantially improved this setup, yielding Word Error Rate (WER) improvements of more than 33% relative.

Current research topics in EMG-based speech recognition include optimized signal processing [9] and acquisition [10], studying the discrepancy between audibly spoken and silently mouthed speech [11, 12, 13, 14], language-specific challenges (e.g. nasality detection [15]), direct synthesis of speech from EMG signals [16, 17, 18], and supervised session adaptation, which we build upon [3]. Other forms of non-acoustic processing of speech signals are researched as well, an overview can be found in [1].

3. Data Acquisition and Corpus

Our corpus follows our prior study [3], but has been substantially enlarged since then. It is a subset of the *EMG-UKA* corpus which has been recorded at Karlsruhe Institute of Technology since 2009 [19], namely, we use data by the two speakers who each recorded a large number of sessions.

The EMG-UKA electrode configuration [2, 21] is depicted in figure 1. The setup consists of 6 EMG channels captur-

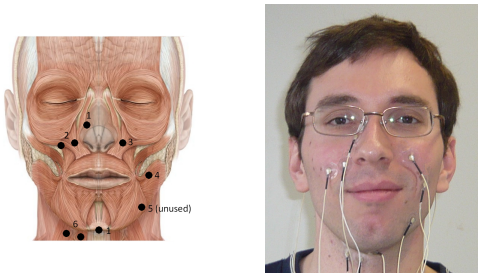

Figure 1: Overview of electrode positioning and captured facial muscles. Muscle chart adapted from [20]

ing data from major facial muscles; channel 5 yields unstable signals and is not used. Recordings were performed with the portable *Varioport* EMG data acquisition system (Becker-Meditec, Germany). EMG data was sampled at 600 Hz and filtered with an analog high-pass filter with a cut-off frequency at 60 Hz. In addition to the EMG recordings, the EMG-UKA corpus contains parallel recordings of the acoustic speech signal, performed with a standard headset microphone, as well as the reference (textual content) for each utterance. Note that in this study, we are constrained to using EMG recordings of normally spoken speech since the subset of the EMG-UKA corpus which is used in this study does not contain a full set of recordings of silently mouthed speech. We refer to our series of studies [11, 12, 13, 14] for the proof that it is possible to deal with EMG data of silently mouthed speech in a satisfactory manner.

Each session comprises 50 English sentences, consisting of a *BASE* subset of 10 sentences and a *SPEC* subset of 40 sentences. The BASE set is identical across all sessions, the SPEC set may vary between sessions. We perform all evaluation on the BASE sets, the SPEC data is used for training the session-independent *background* systems and for adaptation. The full data corpus is summarized in the following table.

Speaker	# of sessions	Average data length per session, in seconds		
		Train	Test	Total
1	32	143	41	184
2	16	140	40	180
Total amount of data: 2:26h				

4. Recognizer Setup

In this section we describe the building blocks of our recognition system, namely feature extraction, training, and decoding.

Our **Features** are based on time-domain properties of the multi-channel EMG signal, like the frame-based power or mean, see [7] for details. Initial frame-based features are stacked to allow for context modeling, finally LDA is applied to reduce the number of features to 12 for our session-dependent systems and 24 for our session-independent and session-adaptive systems; these values were found to be optimal in a series of initial baseline experiments. Note that we consider it legitimate to use different LDA dimensionalities for different systems since our systems vary vastly in training data size, and the optimal feature dimensionality should rise when the training data amount increases: In this case, a "one-size-fits-all" philosophy is clearly wrong and would lead to an unfair comparison between the different systems.

The **recognizer setup** follows a standard pattern: We use three-state left-to-right fully continuous Hidden-Markov-Models, where the emission probabilities are modeled with bundled phonetic features (BDPFs) [8]. Since we only use EMG data of audibly spoken speech, we can create phone-level time alignments from the parallely recorded acoustic data, as in [7]. Note that these alignments are only used for bootstrapping the system, the speech recognition (and, of course, the unsupervised adaptation) is only performed on EMG data.

Training is performed according to the recipe in [8], yielding a *myoelectric model* which statistically describes our EMG features. The **adaptation** modifies the myoelectric model to better match the data of the target session. Details about the unsupervised adaptation algorithm are found below.

For **decoding**, we use the (adapted) myoelectric model together with a trigram Broadcast News language model. On the evaluation set, the perplexity is 24.24. Decoding is also performed on the adaptation data, as a step in the unsupervised adaptation process. Therefore we use three different vocabularies for decoding:

- Our classical experiments on the EMG-UKA corpus [8, 3] use a decoding vocabulary consisting of the 108 words appearing in the BASE test data: The *Base* vocabulary.
- The BASE vocabulary is, of course, unsuitable for decoding the SPEC adaptation data, from which we take the *Spec* vocabulary. It is different for each session: The number of words vocabulary varies between 259 and 311, with an average of 299.
- Finally, we also run experiments where we extend the decoding vocabulary (in both the adaptation and evaluation phases) to 2102 words. This is called the *Full* vocabulary, since it covers the entire EMG-UKA corpus.

The **Session-Independent *Background* Systems** are trained with 7 or 15 training sessions, as follows:

- The 32 sessions of speaker 2 are divided into four blocks of eight sessions, and the 16 sessions of speaker 8 are divided into two blocks of eight sessions.
- We train eight systems on each block by a leave-one-out method, i.e. each system is trained on seven of the sessions, the remaining utterance is the designated *target* session for the adaptation experiments. For training we always use the SPEC data subsets of the respective sessions. We obtain 48 systems, each of which is trained on 280 utterances of 7 sessions.
- We additionally train systems on 15 sessions. For this experiment, blocks 1 and 2 of each speaker are combined to contain 16 sessions, and so are blocks 3 and 4 (for speaker 2). Then leave-one-out training is performed as above, with 15 training sessions and one target session. We again obtain 48 systems, each trained on 600 utterances of 15 sessions.

Note that all systems remain speaker-dependent.

5. The Unsupervised MLLR Algorithm

MLLR [22] is a *model* adaptation method, which means that the myoelectric model is transformed to better match the adaptation data. In this section we describe unsupervised MLLR as it is implemented in our system, for details about MLLR we refer to the classical article [22] by Gales and Woodland.

Assume for now that time alignments of the adaptation data are available. The first step of the adaptation process is the collection of statistics about the training samples which are assigned to each BDPF model. Second, the Gaussian parameters

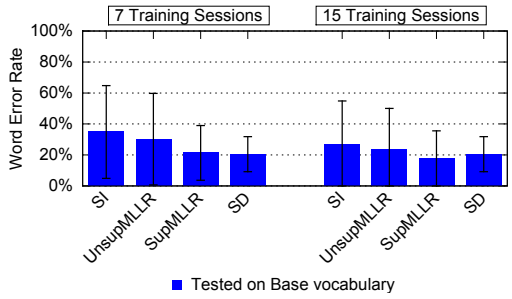

Figure 2: Word Error Rates for the initial adaptation experiment and competing methods, see text for details. Here the unsupervised adaptation (UnsupMLLR) uses the *Spec* vocabulary for decoding the adaptation data and forming hypotheses for adaptation, and the *Base* vocabulary for the final decoding and evaluation step. Results are averaged over all sessions, bars indicate standard deviation.

of each model are updated according to a maximum-likelihood criterion [22]. Gaussians are grouped using a binary regression class tree [23] and jointly transformed, which allows the adaptation of a large number of Gaussians with a small amount of adaptation data and is a key feature of the MLLR algorithm.

When the EMG-based SSI is used in practice, EMG data is produced, but the content of this data is unknown. In order to use this data for adaptation, BDPF-level time alignments must be obtained. Here we apply decoding to determine the most likely sequence of words and, consequently, the most likely BDPF-level alignment—however, since the session-independent recognizer is not perfect, this alignment will usually be partly wrong. If it is used for adaptation without precaution, a deterioration of the recognition accuracy is to be expected.

In order to estimate the quality of parts of a hypothesis, a *confidence measure* is applied. Our confidence computation method is taken from [24]. It is based on *lattices*, which are compact representations of different possible recognition hypotheses in the form of directed graphs: The graph nodes represent words, and edges represent possible successors and predecessors in the set of hypotheses. The words in the lattice are saved together with their (log-)probabilities coming from the myoelectric model (and from the language model), so the total probability of any word at a given timeframe can be computed from the lattice. These probabilities are now used as a confidence measure: When statistics are collected, each training data sample is *weighted* with the local confidence, ranging between 0 (do not use this sample) and 1 (give this sample full weight). The remainder of the MLLR algorithm is performed as usual.

6. Experiments and Results

As a first experiment, we trained session-independent systems as described above, then four setups were investigated:

SI The session-independent systems were left unchanged. No adaptation was performed.

UnsupMLLR Unsupervised adaptation was performed, using the 40 SPEC sentences of the target session. They were decoded with the *Spec* vocabulary, and confidences were computed accordingly.

SupMLLR Supervised adaptation was performed by using the reference text of the 40 adaptation sentences instead of the decoding hypotheses. Confidences are not applicable in this approach.

SD We trained a *session-dependent* system on the 40 SPEC sentences of the target session. This means that the background system is not used at all.

THE 1.00	FEDERAL 1.00	AVIATION 1.00	ADMINISTRATION 1.00						
IS 1.00	FIERCELY 1.00	DEFENDING 1.00	ITS 1.00	OPERATIONS 1.00					
	IN 0.96	TESTIMONY 0.72	*MORE* 0.32	*CARD* 0.28	*IS* 0.47				

Figure 3: Example hypothesis during unsupervised adaptation, with confidences. The reference was "The federal aviation administration is fiercely defending its operations in testimony before congress", so the last three words were wrongly recognized and, consequently, exhibit lower confidence levels than the correctly recognized part of the utterance.

Finally, we decoded the BASE data of the target session, using the *Base* vocabulary.

Figure 2 depicts the results of these experiments. First of all, we observe that unsupervised adaptation brings a considerable improvement compared to the session-independent system, where no form of adaptation is applied: With the 7-session background system, the WER drops from 34.9% to 30.2%, which is a relative improvement of 13.5%. With the 15-session background system, the WER is reduced from 27.0% to 23.3%, which is a very similar improvement of 13.7% relative. These results are highly significant (one-tailed t-test with paired samples, $p = 5 \times 10^{-5}$ and $p = 1 \times 10^{-4}$, respectively).

This is a very satisfactory result, since it proves that on average, both systems allow improvements even without additional supervised training. From prior results in [3], we expect that further adaptation data would yield even better results. Due to the limited size of the sessions, we cannot perform this experiment, however when the EMG-based speech recognizer will be applied in a real-life setting, extra unsupervised adaptation data *is permanently generated just by using the system* and thus readily available.

In order to gain understanding about the inner workings of the system, we inspected the hypotheses which are generated during the decoding of the adaptation data. On the adaptation data, the average WER is 44.4% on the 7-session background system and 40.2% on the 15-session background system. Figure 3 shows an example of a typical hypothesis of the first session of speaker 2, whose WER on the BASE data reduces from 34.30% to 26.30% by unsupervised MLLR. Here the reference is "The federal aviation administration is fiercely defending its operations in testimony before congress", the hypothesis on the unadapted system is "The federal aviation administration is fiercely defending its operations in testimony *more card is*", so the last three words are wrongly decoded. Since we have two word substitutions and one word insertion between the reference and the hypothesis, the WER of this particular utterance is $\frac{3}{13} \approx 23\%$. We see from figure 3 that indeed, the last three words receive far lower confidence probabilities than the first part of the utterance, so the MLLR algorithm will "do the right thing".

Furthermore, we observe from figure 2 that supervised MLLR brings even more improvement, with Word Error Rates of 21.3% and 17.5% using the 7-session and 15-session back-

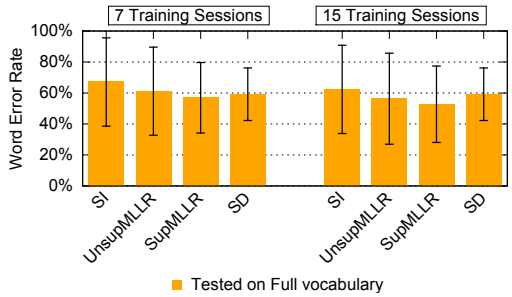

Figure 4: Word Error Rates for adaptation experiment and competing methods, see text for details. Here the unsupervised adaptation (UnsupMLLR) uses the *Spec* vocabulary for decoding the adaptation data and forming hypotheses for adaptation, and the *Full* vocabulary for the final decoding and evaluation step. Results are averaged over all sessions, bars indicate standard deviation.

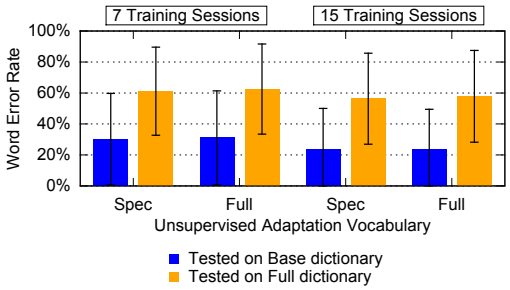

Figure 5: Word Error Rates for unsupervised adaptation using the session-dependent *Spec* vocabulary and the vastly larger *Full* vocabulary for adaptation. The results only deteriorate very slightly.

ground system, respectively. This is clear since supervised MLLR does not suffer from recognition errors on the adaptation data which may deteriorate the result, but there is also an additional reason for its better performance: Of course, the impact of applying adaptation depends on the amount of training samples, i.e. feature frames which are used for training. In the case of unsupervised MLLR, each training data sample is weighted with its confidence, so that formally, each sample is *partially* used for training. This affects the number of used samples: On average, during *supervised* MLLR 10700 frames per BDPF stream are accumulated, excluding "silence" frames. This corresponds to 107 seconds of data. *Unsupervised* MLLR on the 7-session/15-session background systems uses 8155 resp. 8244 frames on average, which is around 25% less and clearly makes a lot of difference [3].

The session-dependent systems, trained on the 40 SPEC sentences of the target session, achieve 20.5% WER on average, so our 15-session system with unsupervised adaptation and 23.3% WER comes quite close to this result without using any transcribed data from the target session at all. With supervised MLLR, it even outperforms the session-dependent system.

So far we performed decoding on the 108-word *Base* vocabulary. Of course, this means that the user is constrained to using only these 108 words, which is not judged sufficient for communication. Therefore in our next experiment, we decode the evaluation data on the *Full* vocabulary, which comes a good deal closer to allowing fluent speech. Note that the adaptation itself remains unaffected.

The resulting WERs are charted in figure 4. We see that the recognizer accuracy declines, now the 7-session and 15-session SI systems achieve WERs of 67.1% and 62.3%, respectively, which are improved by using unsupervised MLLR to 61.1% resp. 56.3%. These are still significant relative improvements of 8.9% resp. 9.6% ($p = 4 \times 10^{-4}$ resp. $p = 3 \times 10^{-5}$); similar results are observed for the other two setups, in line with [3]. While the increasing WER appears detrimental at first, we observed that in many cases, the errors stem from very short words (e.g. "and", "the"), which are often function words. Longer content-bearing words are less affected. Also note that even the 15-session background training data only amounts to around 35 minutes. Since we have seen that more training data incurs better results, we expect that a dissatisfied user can improve the system by simply adding a few more minutes of training or adaptation data. High-quality recognition is certainly possible: Several session combinations achieve WERs of 5% to 7% on the 2102-word vocabulary even without adaptation.

In the last experiment, we again direct our attention to the question of practical applicability. So far, we used the small *Spec* vocabulary for decoding the adaptation data: As described above, this means that a speaker would be constrained to using this vocabulary if adaptation data is to be collected. In real-world situations, a larger vocabulary is desired, so we now use the 2102-word *Full* vocabulary not only for system evaluation, but also for the decoding step in unsupervised adaptation.

Figure 5 charts the WER on the BASE evaluation set for different systems using unsupervised MLLR with the *Spec* vocabulary and the *Full* vocabulary. (All other systems are not affected by changing the adaptation data decoding vocabulary and are therefore omitted.) We observe that the WERs hardly increase: For example, with the 7-session background system, the WER rises from 30.2% to 31.0% resp. 61.1% to 62.5% when evaluation is performed with the *Base* resp. *Full* vocabulary.

This is convincing evidence for the robustness of the confidence computation: Accumulation of adaptation statistics works even though the decoding of the adaptation data becomes much harder. The latter reflects in the WER on the adaptation data, which rises from 44.4% to 53.0% on the 7-session background system and from 40.2% to 50.6% on the 15-session background system. Indeed confidences are lower than for *Spec* vocabulary decoding, e.g. on the 7-session background systems the weighted adaptation data after decoding with the *Full* vocabulary only amounts to an average of 7100 frames, compared to 8200 frames when the *Spec* vocabulary is used. We conclude that even though the WER on the *Full* vocabulary of 2102 words is still relatively high, unsupervised adaptation remains possible, which is a key prerequisite when the method is to be used in a real-life setting.

7. Summary

In this paper we present an EMG-based Silent Speech recognizer using *unsupervised adaptation* for improving its accuracy on unseen recording sessions. The system may be pre-trained at any time prior to usage and can then be applied without requiring specific enrollment. The key contribution of this study is the ability of the system to accumulate EMG adaptation data during normal usage even with a relatively large allowable vocabulary; this capability paves the way towards practical uses of the system.

8. 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] 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.

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

[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] C. Jorgensen, D. D. Lee, and S. Agabon, "Sub Auditory Speech Recognition Based on EMG/EPG Signals," in *Proc. IJCNN*, Portland, Oregon, 2003, pp. 3128 – 3133.

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

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

[9] G. S. Meltzner, G. Colby, Y. Deng, and J. T. Heaton, "Signal Acquisition and Processing Techniques for sEMG based Silent Speech Recognition," in *Proc. EMBC*, 2011, pp. 4848 – 4851.

[10] M. Wand, C. Schulte, M. Janke, and T. Schultz, "Array-based Electromyographic Silent Speech Interface," in *Proc. Biosignals*, 2013, pp. 89 – 96.

[11] 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.

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

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

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

[15] J. Freitas, A. Teixeira, S. Silva, C. Oliveira, and M. S. Dias, "Velum Movement Detection based on Surface Electromyography for Speech Interface," in *Proc. Biosignals*, 2014, pp. 13 – 20.

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

[17] K.-S. Lee, "Prediction of Acoustic Feature Parameters using Myoelectric Signals," *IEEE Transactions on Biomedical Engineering*, vol. 57, pp. 1587 – 1595, 2010.

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

[19] M. Wand, M. Janke, and T. Schultz, "The EMG-UKA Corpus for Electromyographic Speech Processing," in *Proc. Interspeech*, 2014.

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

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

[22] M. J. F. Gales and P. C. Woodland, "Mean and Variance Adaptation within the MLLR Framework," *Computer Speech and Language*, vol. 10, pp. 249 – 264, 1996.

[23] M. J. F. Gales, "The Generation and Use of Regression Class Trees for MLLR Adaptation," Cambridge University Engineering Department, Tech. Rep., 1996.

[24] T. Kemp and T. Schaaf, "Estimating Confidence Using Word Lattices," in *Proc. Eurospeech*, 1997, pp. 827 – 830.