Speaker-Adaptive Speech Recognition Based on Surface Electromyography

Michael Wand and Tanja Schultz

Universität Karlsruhe (TH), Germany,
mwand@ira.uka.de, tanja@ira.uka.de,
WWW home page: http://csl.ira.uka.de

Abstract. We present our recent advances in *silent speech* interfaces using electromyographic signals that capture the movements of the human articulatory muscles at the skin surface for recognizing continuously spoken speech. Previous systems were limited to speaker- and session-dependent recognition tasks on small amounts of training and test data. In this article we present speaker-independent and speaker-adaptive training methods which allow us to use a large corpus of data from many speakers to train acoustic models more reliably. We use the speaker-dependent system as baseline, carefully tuning the data preprocessing and acoustic modeling. Then on our corpus we compare the performance of speaker-dependent and speaker-independent acoustic models and carry out model adaptation experiments.

1 INTRODUCTION

Automatic Speech Recognition (ASR) has now matured to a point where it is successfully deployed in a wide variety of every-day life applications, including telephone-based services and speech-driven applications on all sorts of mobile personal digital devices.

Despite this success, speech-driven technologies still face two major challenges: first, recognition performance degrades significantly in the presence of noise. Second, confidential and private communication in public places is difficult due to the clearly audible speech.

In the past years, several alternative techniques were proposed to tackle these obstacles, among them the recognition of whispered speech with a throat microphone [1] or non-audible murmur with a special stethoscopic microphone [2]. Other approaches include using optical or ultrasound images of the articulatory apparatus, i.e. [3], or subvocal speech recognition [4].

In this article, we present our most recent investigations in electromyographic (EMG) speech recognition, where the activation potentials of the articulatory muscles are directly recorded from the subject's face via surface electrodes[1].

In contrast to many other technologies, the major advantage of EMG is that it allows to recognize *non-audible*, i.e. *silent* speech. This makes it an interesting technology not

[1] Strictly spoken, the technology is called *surface electromyography*, however we use the abbreviation EMG for simplicity.

only for mobile communication in public environments, where speech communication may be both a confidentiality hazard and an annoying disturbance, but also for people with speech pathologies.

Research in the area of EMG-based speech recognition has only a short history. In 2002, [5] showed that myoelectric signals can be used to discriminate a small number of words. In 2006, [6] showed that speaker dependent recognition of continuous speech via EMG is possible. The recognition accuracy in this task could be improved by a careful design of acoustic features and signal preprocessing [7], and advances in acoustic modeling using phonetic features in combination with phone models [8]. However, the described experiments were based on relatively small amounts of data, and consequently were limited to speaker-dependent modeling schemes. In [9], first results on EMG recognition across recording sessions were reported, however these experiments were run on a small vocabulary of only 10 isolated words.

This article reports EMG-based recognition results on continuously spoken speech comparing speaker-dependent, speaker-adaptive, and speaker-independent acoustic models. We use recognition results from the speaker-dependent system as baseline and show that the accuracy of this system improves by appropiately tuning the data preprocessing and adapting the acoustic modeling. For the speaker-independent and speaker-adaptive experiments we first develop generic speaker independent acoustic models based on a large amount of training data from many speakers and then adapt these models based on a small amount of speaker specific data. The baseline performance of the speaker-dependent EMG recognizer is 52.08% Word Error Rate on a testing vocabulary of 108 words.

The article is organized as follows: In section 2, we describe the data acquisition and the resulting data corpus *EMG-PIT*. In section 3, we explain the setup of the EMG recognizer, the feature extraction methods, as well as the different training and adaptation variants. In section 4, we present the recognition results of the different methods and section 5 concludes the article.

2 THE EMG-PIT DATA CORPUS

During the years 2007 - 2008 we collected a large database of EMG signals from 78 native speakers of American English. This collection was done in a joint effort with colleagues from the Department of Communication Science and Disorders at University of Pittsburgh [10]. The resulting data corpus bears the name *EMG-PIT*; to the best of our knowledge it is the largest corpus of EMG recordings of speech so far.

The collection was done in two phases, a pilot study with 14 speakers, and the final collection of 64 speakers. The 14 pilot study subjects participated in two recording sessions, the other speakers participated in one recording session. All participants were female adults between 18 and 35 years of age with normal vocal qualities. The subjects were recruited primarily from the student population of Pittsburgh (University of Pittsburgh and Carnegie Mellon University).

To further study the similarities and differences of audible and silent speaking mode [9], the database covers both speaking modes with parallel utterances, i.e. each speaker read the same sentences in both silent and audible speaking mode. The audible utter-

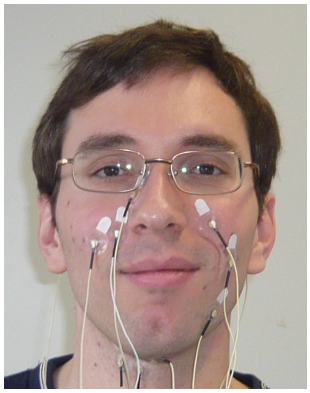

Fig. 1. Electrode Positioning

ances were simultaneously recorded with a conventional air-transmission microphone. For EMG recording we used a computer-controlled 8-channel EMG recorder (Varioport, Becker-Meditec, Germany), together with a self-developed recording tool. Technical specifications of the Varioport system include an amplification factor of 1170, 16 bits A/D conversion, a step size (resolution) of 0.033 microvolts per bit, and a frequency range of 0.9-295 Hz. All EMG signals were sampled at 600 Hz. To allow for backward compatibility with our former experiments in [9], we adopted the electrode positioning which yielded optimal results (see figure 1). This electrode setting uses five channels and captures signals from the *levator angulis oris*, the *zygomaticus major*, the *platysma*, the *anterior belly* of the *digastric* and the *tongue*. To also be able to experiment with new electrode positions, we applied one unused channel to collect new electrode positions. The acoustic data was recorded at 16kHz, 16bit resolution and stored in PCM encoding. All subjects were recorded with a close-up video Camcorder while producing audible and silent speech.

To get good phone coverage and to avoid transcription work, the subjects read phonetically balanced English sentences in a controlled setting rather than to record conversational, unplanned speech. These sentences were taken from the Broadcast News domain. To cover large amounts of context but at the same time allow for mode and variability comparisons, the speaker read one batch of 10 BASE utterances, which are the same for each speaker, and one batch of 40 speaker specific SPEC utterances, only read by one speaker. The vocabulary of the BASE sentences consists of 108 words. Each recording session consisted of two parts, one part audible and one part silent speech. In each part we recorded one BASE set and one SPEC set. The total of 50 BASE and SPEC utterances in each part were recorded in random order. For the pilot study, subjects recorded two sessions, where the order of the audible and silent parts was reversed after the first session to control effects from utterance repetitions between the parts. In the main study, each subject recorded first the audible part and then the silent part. The following table shows the statistics from the EMG-PIT corpus.

Phase	Speakers	Sessions	Utterances		Duration [min]	
			Audible	Silent	Audible	Silent
Pilot	14	28	1400	1400	108	110
Main	64	64	3200	3200	287	251
Total	78	92	4600	4600	395	361

This article reports results on the *audible utterances* of the *pilot study* only, leaving the remaining data as verification set for future studies. Thus the corpus of utterances which was used for this study has the following properties:

Speakers	14 females speakers
Sessions	2 sessions per speaker
Average Length (total) (training set) (test set)	 231 seconds per session 180 seconds 51 seconds
Decoding vocabulary	83 words (108 words including pronunciation variants)

3 EMG-BASED SPEECH RECOGNIZER

The initial EMG recognizer was taken from [7], which in turn was set up according to [6]. It used an HMM-based acoustic modeling, which was based on fully continuous Gaussian Mixture Models. For the initial context-independent phoneme recognizer there were 136 codebooks (three per phoneme, modeling the beginning, middle and end of a phoneme, and one silence codebook). It should be noted that due to the small amount of training data, most speaker dependent codebooks ended up with about one to four Gaussians after the initial automatic merge-and-split codebook generation.

The training concept worked as follows: The time-aligned training data (see section 3.1) was used either for a full training run (see section 3.3), or we applied MLLR adaptation on models which were pre-trained on a large set of speakers to adapt them to the current speaker and session (see section 3.4). The latter is especially important in practical applications since it allows setting up a recognizer with a very small amount of individual training data: in section 4.2 we describe how the recognition results change when the size of the set of speaker-specific training data is reduced.

During the decoding, we used the trained acoustic model together with a trigram language model trained on Broadcast News data. The testing process consisted of an initial testing run followed by a lattice rescoring in order to obtain optimal results. See section 3.6 for details.

In section 3.5 we present our investigations on using bundled phonetic feature models for the EMG recognizer.

3.1 Initial Time Alignment

In order to find a time alignment for the training sentences, the *audio data* which had been simultaneously recorded was used. The audio data was forced-aligned with an

English Broadcast News (BN) speech recognizer trained with the Janus Recognition Toolkit (JRTk). This recognizer is HMM-based, and makes use of quintphones with 6000 distributions sharing 2000 codebooks. The baseline performance of this system is 10.2% WER on the official BN test set (Hub4e98 set 1), F0 condition [11].

The resulting time-alignment can not be mapped directly to the EMG data since the EMG signal precedes the audio signal by about 30ms - 60ms [8]. Accordingly, we modeled this effect by delaying the EMG signal for an amount of 0 ms to 90 ms (in steps of 10 ms). Additionally, in this article we demonstrate that considering a large frame context during acoustic modeling makes the acoustic models more robust with respect to the time delay and yields better recognition results. The effect of the EMG signal delay and of considering a large frame context is charted in section 4.1.

3.2 Feature Extraction

We compare two methods for feature extraction, which are both based on *time-domain (TD) features*. Their only difference is the amount of context which is considered for the final features.

We use the following definitions [6]: For any feature f, $\bar{f}$ is its frame-based time-domain (amplitude) mean, P_f is its frame-based power, and z_f is its frame-based zero-crossing rate. $S(f,n)$ is the stacking of adjacent frames of feature f 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 $w[k]$ is defined as

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

The rectified high-frequency signal is $r[n] = |x[n] - w[n]|$. In [6], the best WER was obtained with the following feature:

$$\mathbf{TD5} = S(\mathbf{f2}, 5), \text{where } \mathbf{f2} = [\bar{\mathbf{w}}, \mathbf{P_w}, \mathbf{P_r}, \mathbf{z_r}, \bar{\mathbf{r}}].$$

We use this feature as baseline in this article and call it TD5, where the number 5 stands for the stacking width. We found that we got optimal results by increasing the context width to 15 frames, yielding a total of 31 frames to be stacked. The resulting feature is called TD15 and is defined as

$$\mathbf{TD15} = S(\mathbf{f2}, 15), \text{where } \mathbf{f2} = [\bar{\mathbf{w}}, \mathbf{P_w}, \mathbf{P_r}, \mathbf{z_r}, \bar{\mathbf{r}}].$$

In section 4.1, we compare the features TD5 and TD15.

In these computations, we used a frame size of 27 ms and a frame shift of 10 ms since we found earlier that these values give optimal results [12]. In both cases, the features from the five EMG channels are stacked to create a final "joint" feature consisting of the synchronized data from all channels, On the resulting joint feature vector, Linear Discriminant Analysis (LDA) is applied to reduce the dimensionality of the final feature vectors to 32 according to [6].

We compare the performance of features TD5 and TD15 in section 4.1 and demonstrate that TD15 performs better than the original TD5 feature. Therefore in the later sections we only use the TD15 feature.

3.3 Training Process

A full training run consisted of the following steps: First, an LDA transformation matrix for feature dimensionality reduction was calculated based on the time-aligned data. Initial codebooks were created by a merge-and-split algorithm in order to adapt to the small amount of training data and to compensate for differences in the available number of samples per phoneme. After this, four iterations of Viterbi EM training were performed to improve the initial models.

3.4 Across-Speaker Experiments and Adaptation

Speaker independent acoustic models were obtained by initially training acoustic models based on the training data of all speakers *but* the two sessions of the test speaker. On the trained models, we tested with the test set of the respective test speaker ("cross-speaker training"). In the adaptation experiments, we performed MLLR-based speaker adaptation of the models prior to the test ("speaker-adaptive training"). The results of these experiments are charted in section 4.2.

3.5 Bundling of Phonetic Features

In the first batch of experiments, we consider (speaker-dependent and speaker-independent) *phoneme models* of the EMG signal, i.e. we regard each frame of the EMG signal as the representation of the beginning, middle, or end state of a phoneme. However, it has been shown in acoustic speech recognition that the recognizer may benefit from additionally modeling *phonetic features (PFs)*, which represent properties of a given phoneme, such as the place of articulation or the manner of articulation [13].

Note that in some previous works, i.e. [13, 14] these models are titled "Articulatory Features". Since this modeling approach does *not* reflect the movements of the articulators, but rather represents phonetic properties of phonemes, we use the term "Phonetic Features" (PFs) in our work.

We derive the PFs from phonemes as described in [15], i.e. we use the IPA phonological features for PF derivation. In this work, we use PFs that have binary values. For example, each of the articulation places Glottal, Palatal and Labiodental is a PF that has a value either present or absent. These PFs do intentionally not form an orthogonal set because we want the PFs to benefit from redundant information. In the experiments reported in this article, we use nine different PFs, namely the set $\{$ Consonant, Vowel, Alveolar, Voiced, Fricative, Affricate, Glottal, Labiodental, Palatal $\}$, since on the relatively small vocabulary of the speaker-dependent systems these PFs are found to receive sufficient training data to allow for good classification (compare [16], figure 2).

The architecture we employ for the PF-based EMG decoding system is a *multi-stream* architecture [15, 17], see figure 2. This essentially means that the models draw their *acoustic probabilities* not from one single source (or stream) but from a weighted sum of various sources. These additional sources correspond to acoustic models representing substates of PFs, like "middle of a vowel" or "end of a non-fricative". The conventional EMG phoneme-based recognizer contributes as well.

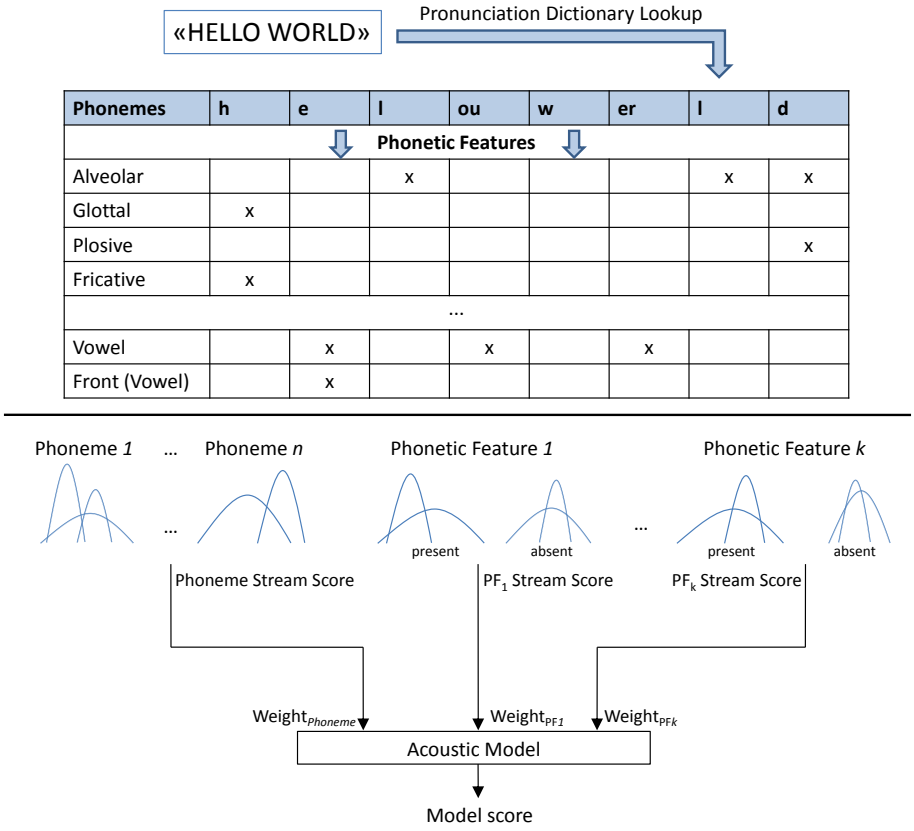

Fig. 2. The Multi-Stream Phonetic Features Decoding Architecture. The upper part shows how the PFs are obtained from the phonetic information, the lower part shows the weighting of the various information sources.

It was suggested by [18] that one major shortcoming of previous PF recognition systems was that features were modeled as statistically independent. The independence assumption is not correct since physiologically every phonetic feature describes the interplay of various articulators, i.e. the interdependent activity of several facial muscles.

We described a data-driven algorithm for finding dependencies between phonetic features in [19]. We call the process of pooling dependent features together "feature bundling", since eventually we will end up with a set of PF acoustic models which represent *bundles* of PFs, like "voiced fricative" or "rounded front vowel". We additionally allow these bundled phonetic features (BDPFs) to depend not only on properties of the current phoneme, but also on the right and left context phonemes. In [19] we reported that an EMG recognizer based on bundled phonetic features outperforms a recognizer based on context-independent phonemes only by more than 30%.

The algorithm which performs this pooling is a standard decision-tree based clustering approach [20], as it is successfully used in large vocabulary acoustic speech recognition to determine phoneme context clusters. This clustering works by creating a *context decision tree*, which classifies phonemes by asking linguistic questions about the current phoneme and its left and right context. The set of all possible questions is predefined, examples of these categorical questions are: *Is the current phone voiced?* or *Is the right-context phone a fricative?*.

The context tree is created separately for each PF stream, from top to bottom. This means that the initial set of acoustic models e.g. for the stream "FRICATIVE" consists of six models: namely the beginning, middle and end of a "FRICATIVE" or "NON-FRICATIVE". Each context question splits one acoustic model into two new models. The splitting criterion is maximizing the loss of entropy caused by the respective split. Note that both the models representing the *presence* and *absence* of a phonetic feature take part in the splitting process. The process ends when a pre-determined termination condition is met. This condition must be chosen based on the properties of the available data to create a good balance between the accuracy and the trainability of the context-dependent models.

Our termination criterion is that a fixed number of 70 tree leaves for each phonetic feature, corresponding to 70 independent acoustic models, is generated for each PF stream, since this number was experimentally found to yield optimal results. Then the general training process is as follows:

- First, an ordinary context-independent EMG recognizer is trained on the given training data. This recognizer uses both phoneme and PF models, but *no* PF bundling yet.
- In a second step, the context decision tree is grown as described above, and a set of bundled phonetic features (BDPFs) is generated.
- Finally, the BDPF EMG recognizer is trained using the acoustic models defined in the previous step.

With the BDPF recognizer, we perform the same set of cross-speaker and adaptation experiments as with the phoneme-based recognizer, see section 3.4. The results are charted in section 4.2.

3.6 Testing

For decoding, we use the trained acoustic model together with a trigram BN language model. We restricted the decoding vocabulary to the words appearing in the test set. This resulted in a test set of 10 sentences per speaker with a vocabulary of 108 words. On the test sentences, the trigram-perplexity of the language model is 24.24.

The testing process used lattice rescoring in order to determine the optimal weighting of the language model compared to the acoustic model.

4 EXPERIMENTAL RESULTS

4.1 Preprocessing for the Speaker-Dependent System

Figure 3 compares the word error rates (WER) of the speaker-dependent recognition systems of the feature preprocessings TD5 and TD15 and phoneme-based and BDPF modeling. The results were obtained on speaker-dependent systems, i.e. by training on the training data of *one* session and tested on test data from *the same* session. Note that we give the averages over all 28 sessions.

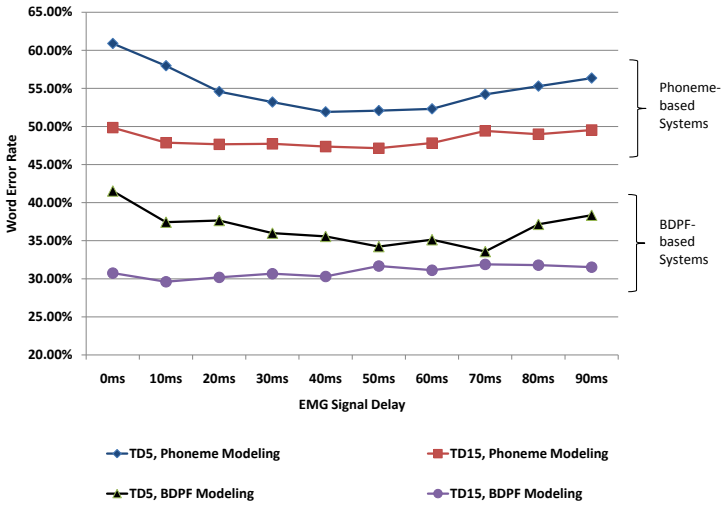

Fig. 3. Average Speaker-dependent Word Error Rate for Different Time Delays

The average WER of the baseline recognizer, equipped with the TD5 feature, is 54.89%, where the WER ranges from 51.92% to 60.90%, with a noticeable minimum between 30 ms and 60 ms. Using the TD15 feature, we get an average WER of 48.34%, where for EMG signal delays between 0 ms and 90 ms, the WER ranges from 47.15% to 49.85%. Thus one can see that the wider context yields not only a performance improvement of about 12% relative, but also a much higher robustness when the delay between audio and EMG signal is varied. This can be important in settings where the exact synchronization between EMG and audio signal may not be exactly determined during the training run.

Similarly, the average word error rate of the BDPF recognizer with the TD5 feature is 36.67%, which by using the TD15 feature is reduced by about 11.5% relative to 30.95%. Again, one can see that the graph becomes “flatter” by using a higher context width, i.e. the performance becomes less dependent on the EMG signal delay.

Experiments showed that increasing the context width consistently increases the recognition performance until a critical width of about 15 frames to each side, which

corresponds to a total context window of about 300 ms. Beyond that value no more significant improvement occurs. One can conclude from this that speech-relevant observable patterns in the EMG signal may have a length of up to 300 ms and that a purely frame-based preprocessing, which only considers a window of about 27 ms, does not fully capture the discriminating properties of the EMG signal. Also note that the optimal delay for the TD5 preprocessing still lies around 50ms, which is consistent with the results in [6].

In accordance with the results given above, for all further experiments we used the following recognizer setup:

- Feature Preprocessing: Time-domain feature TD15
- Modeling: Bundled Phonetic Features (BDPFs)
- EMG signal delay: 50ms.

4.2 Cross-Speaker and Adaptive Experiments

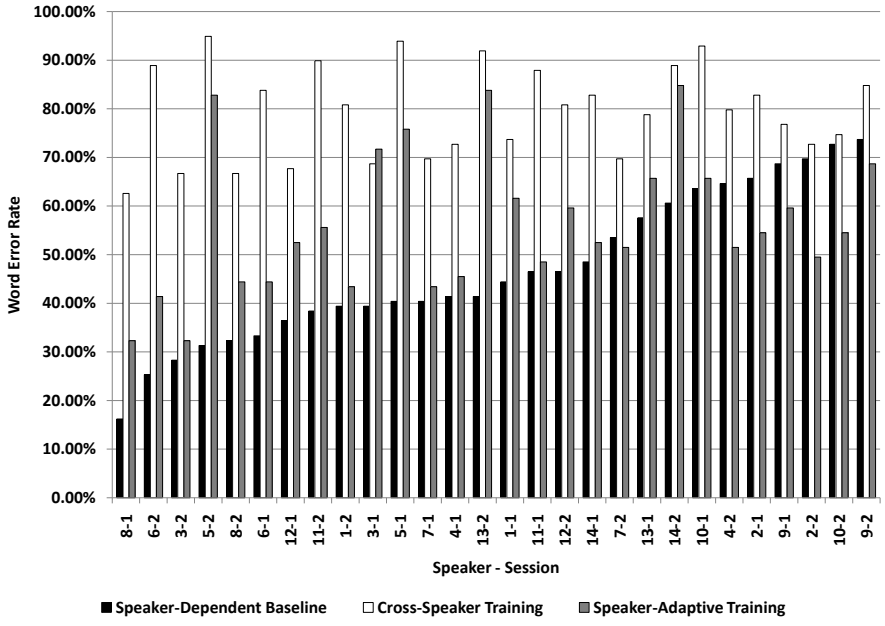

Fig. 4. Comparison of Word Error Rates for the Phoneme-based Recognizer

In the following experiments we compare three training scenarios:

- Speaker-Dependent Training: As above, the system is trained and tested with data from one speaker and one session only.

– Cross-Speaker Training: The system is trained on all sessions from all speakers *except* the two sessions from the test speaker. The system is tested on the test data of one session.
– Speaker-Adaptive Training: We use the trained models from the cross-speaker training step, but the resulting system is then adapted toward the test speaker using MLLR adaptation [21] on the training data from one session. As above, testing is done on the test data from the same session.

Figure 4 shows a breakdown of the results of these experiments for each speaker and indicates that the speaker-dependent and adaptive systems clearly outperform the cross-speaker system. This is not very surprising as the speaker independent models have to capture speaker variabilities but at the same time suffer from slight variations in the electrode positioning across speakers. Furthermore, we see that speaker dependent model training achieves better results than MLLR adaptation for most of the speakers and sessions. However for sessions where speaker-dependent training performs badly, particularly for speakers 2 and 9 and to some extent 4 and 10, the performance of the adapted system does not degrade similarly and may outperform the speaker-dependent system.

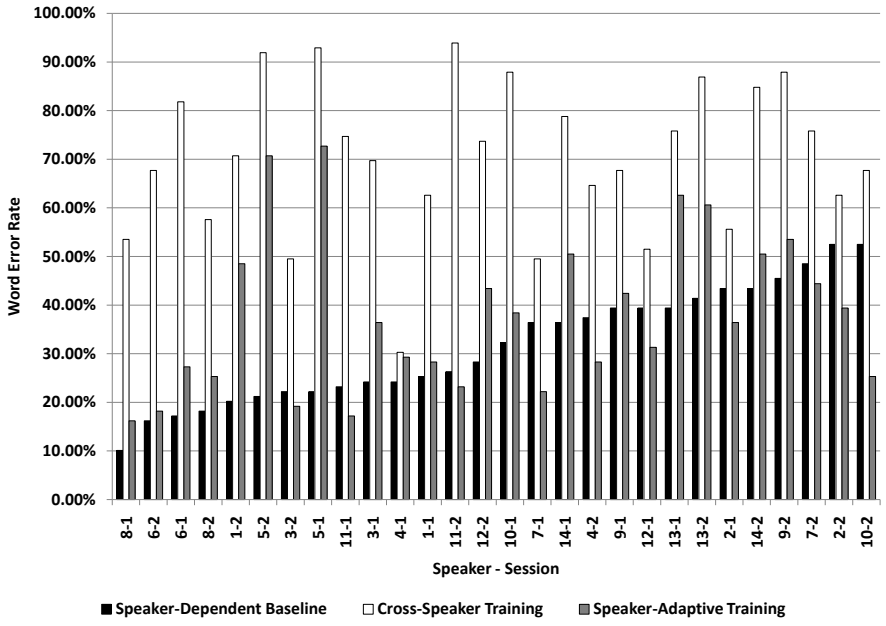

Fig. 5. Comparison of Word Error Rates for the BDPF-based Recognizer

We repeated the above experiment, comparing the training scenarios *Speaker-Dependent Training*, *Cross-Speaker Training* and *Speaker-Adaptive Training*, now using bundled

phonetic features (BDPFs) as acoustic models. The results are charted in figure 5 and indicate that BDPF modeling brings a clear performance improvement in all the training scenarios. Beyond this, the same pattern as for the phoneme-based recognizer holds: Speaker-dependent recognition generally still achieves better results than both cross-speaker and speaker-adaptive recognition, however for most speakers, with the notable exception of speaker 5, the adaptive system has a performance close to the one of the speaker-dependent system and even outperforms it for 10 out of the 28 sessions.

As a final experiment, we investigated whether MLLR adaptation is applicable for very small sets of speaker-specific training data. For this purpose we took subsets of 10, 20 and 30 sentences out of the full training sets of 40 sentences for each speaker and used each of these reduced sets to train a speaker-dependent system and to create a speaker-adaptive system by performing MLLR adaptation on the *original* cross-speaker system, trained on the full set of training data from all speakers except the one to be tested. Note that the test set remained unchanged in these experiments.

Figure 6 displays the average word error rate of these recognizers and clearly shows that while for the full set of training sentences the average WER is better for the speaker-dependent systems, the situation is very different for smaller sets of training data: With 10 adaptation sentences, the best speaker-adaptive system yields an average WER of 52.26% (with no adaptation at all, i.e. for the corresponding cross-speaker system, the average WER is 70.27%), while a speaker-dependent system yields a high average WER of 85.49%. When the training set grows, all systems quickly improve, but for up to 30 training sentences, the speaker-adaptive systems on average have a better performance than the speaker-dependent systems.

4.3 Summary

The following table summarizes the average Word Error Rates of the different recognizers we presented in the above sections. Note that all values are based on the TD15 preprocessing, and that we always give results for the full set of 40 speaker-specific training/adaptation sentences.

System	Word Error Rate		Rel. Gain by BDPF
	Phonemes	BDPF	
Speaker-Dependent	47.15%	31.68%	32.8%
Cross-Speaker	79.50%	70.27%	11.6%
Speaker-Adaptive	56.34%	37.92%	32.7%

The results are also charted in figure 7 and clearly show that MLLR model adaptation is applicable to the task of EMG speech recognition. Furthermore it can be seen that BDPF modeling yields a clear improvement in all training scenarios. In the speaker-dependent and speaker-adaptive cases, the improvements are both about 33%, whereas in the cross-speaker case, the gain is at about 11.5%. In particular, BDPF modeling improves the performance of the adaptation step, making this modeling approach a very interesting perspective for further investigations in speaker-adaptive EMG signal processing.

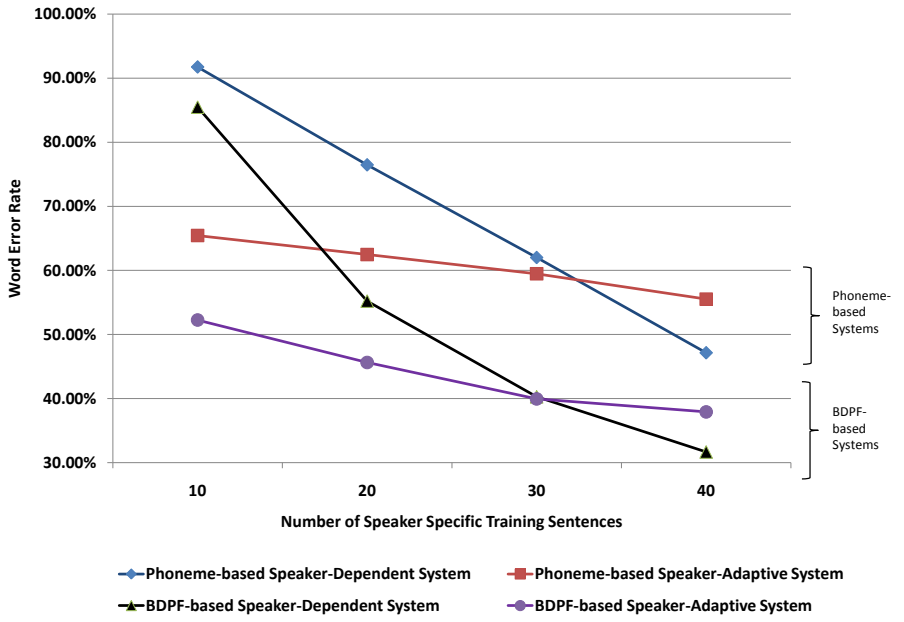

Fig. 6. Comparison of Word Error Rates for speaker-dependent and speaker-adaptive systems on different amounts of training data.

5 CONCLUSIONS

We have compared speaker-dependent, speaker-independent and speaker-adaptive systems for EMG speech recognition, reporting results on the performance of EMG speech recognition across multiple speakers and sessions of the EMG-PIT data corpus. We found that while for the full training sets of the EMG-PIT corpus the speaker- and session-dependent EMG system still performs best, for small speaker-specific training sets of up to approximately 30 utterances, on average a speaker-adaptive system outperforms a speaker- and session-dependent EMG recognizer; the speaker-adaptive system still yields acceptable recognition results with a set of only 10 adaptation sentences. This shows that the MLLR adaptation method is feasible for EMG speech recognition and that adaptation methods may be a lever for increasing the usability of EMG-based speech recognition.

We also showed that phonetic feature bundling consistently outperforms phoneme-based systems and in particular significantly increases the performance of MLLR adaptation.

ACKNOWLEDGMENTS

We would like to thank Maria Dietrich and Katherine Verdolini Abbott for the cooperation in recording the EMG data as part of a larger study which was supported in

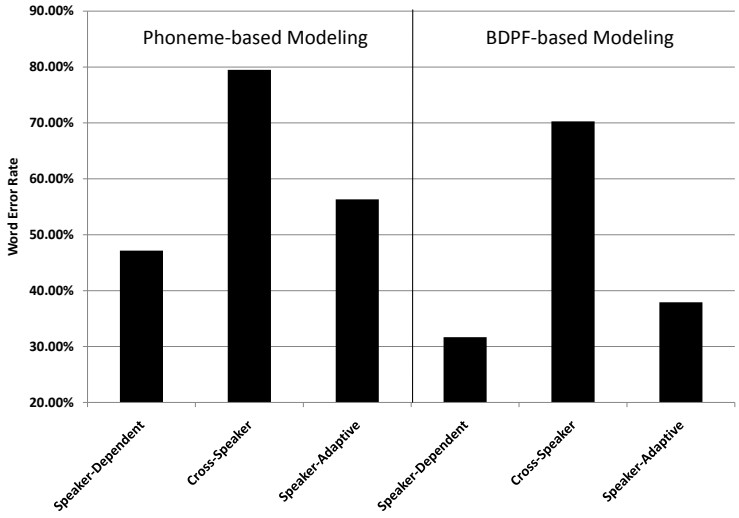

Fig. 7. Comparison of Average Word Error Rates for All Systems

part through funding received from the SHRS Research Development Fund, School of Health and Rehabilitation Sciences, University of Pittsburgh to Maria Dietrich and Katherine Verdolini Abbott.

References

1. Szu-Chen Jou, Tanja Schultz, and Alex Waibel. Whispery Speech Recognition Using Adapted Articulatory Features. In *Proc. ICASSP*, 2005.
2. Yoshitaka Nakajima, Hideki Kashioka, Kiyohiro Shikano, and Nick Campbell. Non-Audible Murmur Recognition. In *Proc. Eurospeech*, 2003.
3. Thomas Hueber, Gérard Chollet, Bruce Denby, Gérard Dreyfus, and Maureen Stone. Continuous-Speech Phone Recognition from Ultrasound and Optical Images of the Tongue and Lips. In *Proc. Interspeech*, pages 658–661, 2007.
4. Chuck Jorgensen and Kim Binsted. Web Browser Control Using EMG Based Sub Vocal Speech Recognition. In *Proceedings of the 38th Hawaii International Conference on System Sciences*, 2005.
5. A. Chan, K. Englehart, B. Hudgins, and D. Lovely. Hidden Markov Model Classification of Myolectric Signals in Speech. *Engineering in Medicine and Biology Magazine, IEEE*, 21(9):143–146, 2002.
6. Szu-Chen Jou, Tanja Schultz, Matthias Walliczek, Florian Kraft, and Alex Waibel. Towards Continuous Speech Recognition using Surface Electromyography. In *Proc. Interspeech*, Pittsburgh, PA, Sep 2006.
7. Michael Wand, Szu-Chen Stan Jou, and Tanja Schultz. Wavelet-based Front-End for Electromyographic Speech Recognition. In *Proc. Interspeech*, 2007.
8. Szu-Chen Jou, Lena Maier-Hein, Tanja Schultz, and Alex Waibel. Articulatory Feature Classification Using Surface Electromyography. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-2006), Toulouse, France, May 15-19, 2006*, 2006.

9. Lena Maier-Hein, Florian Metze, Tanja Schultz, and Alex Waibel. Session Independent Non-Audible Speech Recognition Using Surface Electromyography. In *Proc. ASRU*, 2005.
10. Maria Dietrich. *The Effects of Stress Reactivity on Extralaryngeal Muscle Tension in Vocally Normal Participants as a Function of Personality*. PhD thesis, University of Pittsburgh, 2008.
11. Hua Yu and Alex Waibel. Streamlining the Front End of a Speech Recognizer. In *Proc. ICSLP*, 2000.
12. Matthias Walliczek, Florian Kraft, Szu-Chen Jou, Tanja Schultz, and Alex Waibel. Sub-Word Unit Based Non-Audible Speech Recognition Using Surface Electromyography. In *Proc. Interspeech*, Pittsburgh, PA, Sep 2006.
13. Katrin Kirchhoff. *Robust Speech Recognition Using Articulatory Information*. PhD thesis, University of Bielefeld, 1999.
14. Florian Metze. *Articulatory Features for Conversational Speech Recognition*. PhD thesis, University of Karlsruhe, 2005.
15. Florian Metze and Alex Waibel. A Flexible Stream Architecture for ASR Using Articulatory Features. In *Proc. ICSLP*, Sep 2002.
16. Szu-Chen Stan Jou and Tanja Schultz. *BIOSTEC - BIOSIGNALS 2008 best papers*, chapter Automatic Speech Recognition based on Electromyographic Biosignals, page accepted for publication. Communications in Computer and Information Science (CCIS) series by Springer, Heidelberg, 2009.
17. Szu-Chen Stan Jou, Tanja Schultz, and Alex Waibel. Continuous Electromyographic Speech Recognition with a Multi-Stream Decoding Architecture. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-2007), Honolulu, Hawaii, US, April 15-20, 2007*, 2007.
18. Joe Frankel, Mirjam Wester, and Simon King. Articulatory Feature Recognition Using Dynamic Bayesian Networks. In *Proc. ICSLP*, 2004.
19. Tanja Schultz and Michael Wand. Modeling Coarticulation in Large Vocabulary EMG-based Speech Recognition. Speech Communication Journal, 2009. To Appear.
20. L. R. Bahl, P. V. de Souza, P. S. Gopalakrishnan, D. Nahmoo, and M. A. Picheny. Decision Trees for Phonological Rules in Continuous Speech. In *Proc. ICASSP*, 1991.
21. C. J. Leggetter and P. C. Woodland. Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models. *Computer Speech and Language*, 9:171–185, 1995.