Modeling Coarticulation in EMG-based Continuous Speech Recognition

Tanja Schultz and Michael Wand

Abstract

This paper discusses the use of surface electromyography for automatic speech recognition. Electromyographic signals captured at the facial muscles record the activity of the human articulatory apparatus and thus allow to trace back a speech signal even if it is spoken silently. Since speech is captured before it gets airborne, the resulting signal is not masked by ambient noise. The resulting Silent Speech Interface has the potential to overcome major limitations of conventional speech-driven interfaces: it is not prone to any environmental noise, allows to silently transmit confidential information, and does not disturb bystanders.

We describe our new approach of phonetic feature bundling for modeling coarticulation in EMG-based speech recognition and report results on the EMG-PIT corpus, a multiple speaker large vocabulary database of silent and audible EMG speech recordings, which we recently collected. Our results on speaker-dependent and speaker-independent setups show that modeling the interdependence of phonetic features reduces the word error rate of the baseline system by over 33% relative. Our final system achieves 10% word error rate for the best-recognized speaker on a 101-word vocabulary task, bringing EMG-based speech recognition within a useful range for the application of silent speech interfaces.

Key words: EMG-based Speech Recognition, Silent Speech Interfaces, Phonetic Features

1. Introduction

In the past decade, the performance of automatic speech processing systems, including speech recognition, spoken language translation, and speech synthesis, has improved dramatically. This has resulted in an increasingly

widespread use of speech and language technologies in a large variety of applications, such as commercial information retrieval systems, call center services, voice-operated cell phones, car navigation systems, personal dictation and translation assistance, as well as applications in military and security domains. However, speech-driven interfaces based on conventional acoustic speech signals still suffer from several limitations.

Firstly, acoustic speech signals are transmitted through air and are thus prone to ambient noise. Despite tremendous efforts there are still no robust speech processing systems in sight, which provide reasonably good results in crowded restaurants, airports, or any noisy places. To overcome this problem we propose to capture and process the speech signal before it gets airborne and thus avoid to get affected by adverse noise conditions.

Secondly, conventional speech interfaces rely on audibly uttered speech, which has two major drawbacks: it jeopardizes confidential communication in public and it disturbs any bystanders. Services which require the access, retrieval, and transmission of private or confidential information, such as PINs, passwords, and security or safety information are particularly vulnerable. The proposed Silent Speech Interface (SSI) allows to utter speech silently and thus overcomes both limitations: confidential information can be submitted securely and silent speech does not disturb or interfere with the surroundings.

Finally, Silent Speech Interfaces might give hope to people with certain speech disabilities as the technologies allow the building of virtual prostheses for patients without vocal folds (Denby et al., 2009). Also, elderly and weak people may benefit since silent articulation can be produced with less effort than audible speech.

Our approach to capture speech before it gets airborne relies on surface ElectroMyoGraphy (EMG). This is the process of recording electrical muscle activity using surface electrodes. When a muscle fiber is activated by the central nervous system, small electrical currents in form of ion flows are generated. These electrical currents move through the body tissue, encountering a resistance which creates an electrical field. The resulting potential differences can be measured between certain regions on the body surface, i.e. at the skin. The amplified electrical signal obtained from measuring these voltages over time can be fed directly into electronic devices for further processing. Since speech is produced by the activity of the human articulatory muscles, the resulting myoelectric signal patterns allow to trace back the corresponding speech. These signals are not corrupted or masked by en-

vironmental noise transmitted through air. Furthermore, since EMG relies on muscle activity only, speech can even be recognized when it is produced silently, i.e. mouthed without any vocal effort.

We envision several application areas for Silent Speech Interfaces: (1) robust, private, non-distracting speech recognition for human-machine interfaces, such as silently speaking text messages, (2) recognition plus speech synthesis (at the remote side) for quietly accessing remote applications, such as speech or text-based information systems, (3) transmitting articulation parameters followed by articulatory synthesis for silent human-human communication, (4) speech prostheses, and (5) recognition of silent speech followed by text translation into another language followed by speech synthesis, which appears like speaking in a foreign tongue. In 2006 we successfully demonstrated a prototype of this last application at Interspeech (Jou et al., 2006). A video file of our latest system showcasing some of the above mentioned applications is available from our webpage[1].

2. Toward Large Vocabulary EMG-based Speech Recognition

The use of EMG for speech recognition dates back to the mid 1980s, when Sugie and Tsunoda in Japan, and Morse with colleagues in the United States published almost simultaneously their first studies. Sugie and Tsunoda (1985) used three surface electrodes to discriminate Japanese vowels, and demonstrated a pilot system which performed this task in realtime. Morse and O'Brien (1986) examined speech information from neck and head muscle activity to discriminate two spoken words, and in the following years, extended their approach to the recognition of ten words spoken in isolation (Morse et al., 1989, 1991). Although initial results were promising, with accuracy rates of 70% on a ten word vocabulary, performance decreased dramatically for slightly larger vocabularies, achieving only 35% for 17 words, and thus did not compare favorably with conventional speech recognition standards.

More competitive performance was first reported by Chan et al. (2001), who achieved an average word accuracy of 93% on a 10-word vocabulary of the English digits. A good performance could be achieved even when words were spoken non-audibly, i.e. when no acoustic signal was produced (Jorgensen et al., 2003), suggesting this technology could be used to communicate

[1] see `http://csl.ira.uka.de/index.php?id=146`

silently. Recent work (Jou et al., 2006; Walliczek et al., 2006) successfully demonstrated that phonemes can be used as modeling units for EMG-based speech recognition by carefully designing the signal preprocessing front-end, paving the way for large vocabulary speech recognition. For a more detailed review on Silent Speech Interfaces based on EMG, please refer to (Denby et al., 2009) in this journal issue.

While a lot of progress was made over the last years, there are still major limitations which need to be overcome for the application of EMG to large vocabulary speech recognition. In particular, we see three major challenges. First, the impact of speaker dependencies, such as speaking style, speaking rate, and pronunciation idiosyncrasies needs to be investigated. Second, the EMG signal is affected by changes in electrode positioning, environmental conditions (temperature and humidity), and tissue properties (Leveau and Andersson, 1992). These factors clearly favor the development of speaker dependent and often session dependent systems, i.e. systems in which training and testing is performed on data collected within the same recording session. Consequently, results known from the literature focus on a very small number of subjects. Third, little is known yet about the qualitative and quantitative articulation differences between silent and audible speech. Our experimental results in (Maier-Hein et al., 2005) and more recently in (Wand et al., 2009) suggest that EMG signals do significantly differ between silent and audible speaking mode. We assume that this is mainly due to the lack of biofeedback when speaking silently but further investigation will be necessary to continuously improve our silent speech interface. Nevertheless, we recently demonstrated the first speech recognition system that handles seamless switches between both speaking modes.

In this article we focus on the issue of achieving reliable and robust models for large vocabulary speech recognition systems based on EMG and show that these models significantly improve the recognition performance, even when impacting factors such as speaker and session variabilities are present.

2.1. The "EMG-Pittsburgh (EMG-PIT)" Multiple Speaker Database

Over the last two years we collected a large database of EMG signals from 78 speakers, where the speakers produced audible and silent, i.e. mouthed, speech. This collection was done in a joint effort with colleagues from the Department of Communication Science and Disorders at University of Pittsburgh (Dietrich, 2008).

The collection was carried out 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).

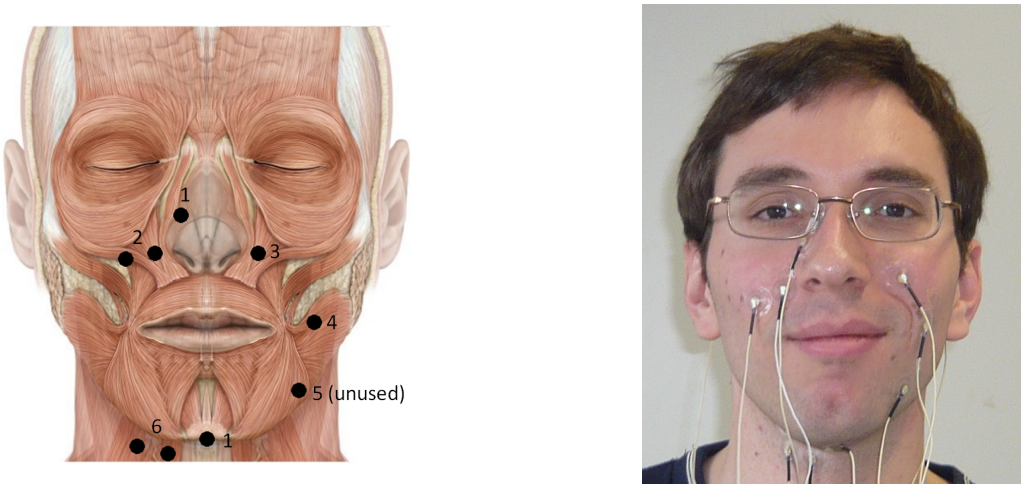

Figure 1: Overview of electrode positioning and captured facial muscles (muscle chart adapted from (Schünke et al., 2006)). See text for description.

To study similarities and differences of audible and silent speaking mode, the database covers both speaking modes with parallel utterances. The audible utterances were simultaneously recorded with a conventional air-transmission microphone. For EMG recording we used a computer-controlled 8-channel EMG data acquisition system (Varioport, Becker-Meditec, Germany). 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. Following our previous studies on the optimal positioning and number of recording electrodes (Maier-Hein et al., 2005), we adopted the electrode positioning which yielded maximal recognition results. This also ensures backward compatibility of our experiments. The electrode setting is shown in figure 2.1. It uses five channels, numbered 1, 2, 3, 4,

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

Table 1: Statistics of the EMG-PIT Multiple Speakers Database

and 6. Channel five serves for experiments with different electrode positionings, however we did not use it for the experiments described in this paper. Channels 1, 2, and 6 use bipolar derivation, whereas channels 3, 4, and 5 were derived unipolarly, with two reference electrodes placed on the mastoid portion of the temporal bone. The electrodes capture signals from the levator angulis oris (channels 2 and 3), the zygomaticus major (channels 2 and 3), the platysma (channel 4), the anterior belly of the digastric (channel 1) and the tongue (channels 1 and 6). However, due to the fact that the EMG is captured at the surface, some signals may consist of a superposition of active muscle fibers in the proximity of the recording electrode. The acoustic data were 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.

In order to get good phone coverage, and to avoid transcription work, the subjects read phonetically balanced sentences in a controlled setting rather than recording conversational, unplanned speech. To cover large amounts of linguistic 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 consisted of 101 words. All sentences from both batches were selected to be phonetically balanced. Each recording session consisted of two parts, one audible and one silent speech part. In each part we recorded one BASE set and one SPEC set. The total of 50 utterances were recorded in random order. For the pilot study subjects who recorded two sessions, the order of the audible and silent parts was reversed after the first session to control effects from utterance repetitions between the parts. Table 1 shows the statistics from the resulting EMG-PIT corpus.

2.2. Baseline EMG-based Speech Recognition System

In (Wand and Schultz, 2009b) we reported first EMG recognition results based on 26 recording sessions with 13 speakers of the audible part of the EMG-PIT pilot study subset. For each speaker, the audible part of the SPEC set was used for training, and the BASE set for testing. This EMG-based recognizer is described below and serves as a baseline for the experiments presented in this paper.

The baseline EMG-based recognition system uses 45 context independent phoneme models and a silence model. Each phoneme is modeled using a 3-state left-to-right Hidden Markov Model (beginning, middle, and end of the phoneme), silence is modeled by a single state. This results in 136 models, each of which applies Gaussian Mixtures as emission probabilities. This modeling scheme follows the traditional setup for context-independent acoustic-based speech recognition.

The amount of Gaussians is determined by a merge-and-split algorithm (Ueda et al., 2000) on the training data, resulting in roughly 2 Gaussians per model on average. In total the baseline system consists of 290 Gaussians. This small number is due to the very limited amount of training data. For the same reason our systems applies Gaussians with diagonal covariance matrices.

For feature extraction, we found that time-domain features gave optimal results. This feature extraction method is defined in the following way (Jou et al., 2006): For any feature f, $\bar{f}$ is its frame-based time-domain 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 $r[n] = |x[n] - w[n]|$. Then the TD15 feature is defined as

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

Note that (Jou et al., 2006) and (Wand and Schultz, 2009b) only used a stacking width of 5 frames. On the EMG-PIT corpus, the stacking width of

15 frames gives significantly better results (Wand and Schultz, 2009a). In these computations, we used a frame size of 27 ms and a frame shift of 10 ms. These values are reported as giving optimal results in our earlier work, therefore we adopted the same frame size and shift in the TD15 feature extraction.

The TD15 feature is computed for each of the five electrode channels, then the final feature vector per frame is built by stacking the frame-based features of the five channels. After this procedure, Linear Discriminant Analysis is applied to reduce the dimensionality of the final feature vector from 775 to 32 coefficients per frame.

In order to initialize the EMG phoneme models, we require time alignments for the audible EMG training utterances. We obtain these time alignments by using the acoustic data which has been simultaneously recorded. These acoustic data are forced-aligned with a Broadcast News (BN) speech recognizer trained with the Janus Recognition Toolkit (JRTk). This HMM-based recognizer uses quintphones with 6000 distributions sharing 2000 codebooks. The baseline performance of this acoustic speech recognizer is 10.2% Word Error Rate (WER) on the clean speech condition (F0) of the official BN test set (Yu and Waibel, 2000).

After the initial EMG phoneme models have been obtained, four iterations of Viterbi training are performed.

For decoding, we apply a trigram language model trained on Broadcast News data. The testing process consists of a Viterbi decoding followed by a lattice rescoring based on a matrix of word penalty and language model weighting parameters in order to obtain optimal recognition results. We use the batch of speaker-specific audible SPEC utterances as training set, and the audible BASE utterances as testing set. Therefore, in total we have a test set of 28 sessions of 14 speakers, with 10 utterances per speaker with a vocabulary of 101 words. On the test set, the trigram-perplexity of the language model is 24.24. The average Word Error Rate obtained with this baseline speaker-dependent EMG-based recognition system is 47.15%.

3. Speaker-dependent EMG-based Recognition System

In this section, we report results of *speaker-dependent* EMG-based speech recognition on the audible sentences of the 14 speakers of the *pilot study* of the EMG-PIT corpus as described in section 2.1. The basic recognizer setup is the same as described in section 2.2.

3.1. Modeling Phonetic Features

In (Wand and Schultz, 2009b) we considered speaker-dependent and speaker-independent *phoneme-based* EMG recognizers. This means that 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 (Kirchhoff, 1999) that an acoustic speech recognizer can benefit from additionally modeling *phonetic features (PFs)*, which represent properties of a given phoneme, such as the place or the manner of articulation.

Note that in previous works, i.e. (Kirchhoff, 1999; Metze, 2005), Phonetic Features are also called "Articulatory Features". Since the phonetic feature 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.

It is empirically shown (Kirchhoff, 1999) that a speech recognizer which combines phoneme models and PF models performs better under adverse conditions, like poor signal quality or background noise. While EMG-based speech recognition does not suffer from ambient noise, we face the challenge of other noise artifacts, such as the impact of temperature and humidity on the electrodes, or superposition of muscle activity. Therefore, we investigate in this study the effect of PFs on EMG-based speech recognition. Also, when only a small data set is available, PF models get a more robust parameter estimation: Since a phonetic feature is generally shared by multiple phonemes, we can use the combined training data of these phonemes in order to train a phonetic feature model more reliably than a single phone model.

The remainder of this section deals with the effect of modeling phonetic features for EMG-based speech recognition. We use PFs which 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 are directly derived from the phonemes and correspond to the IPA phonological features (International Phonetic Association, 1999). The PFs do intentionally not form an orthogonal set because we want the PFs to benefit from redundant information.

Figure 2 shows PF classification F-scores [2] for different phonetic features,

[2]With C_{tp} = true positive count, C_{fp} = false positive count, C_{fn} = false negative count, precision $P = C_{tp}/(C_{tp} + C_{fp})$, and recall $R = C_{tp}/(C_{tp} + C_{fn})$, the (balanced) F-Score is defined as F-Score $= 2PR/(P + R)$.

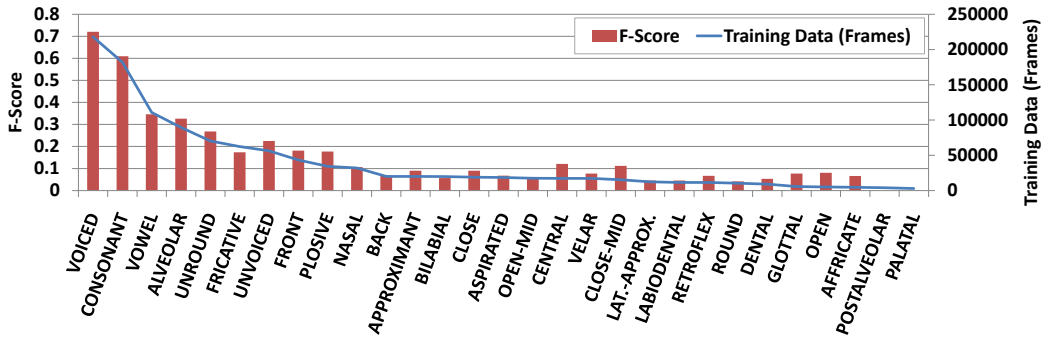

Figure 2: Phonetic Feature Classification Accuracies (F-scores) and Training Data Amounts (Frames). Note that only the training data amount for the *present* PF models is charted.

where the features are sorted according to the amount of data which is available to train the *present* models of these phonetic features. We did not consider the training data amount for the *absent* models, since in the vast majority of cases, the *absent* model receives much more training data than the *present* model. It can be seen that the classification accuracy for EMG measured in F-Score roughly corresponds to the amount of training data and that only a small number of phonetic features receives sufficient training data to yield good classification rates. To ensure reliable estimates for the PFs in our experiments, we limited ourselves to the nine phonetic features in the database which had more than 50000 frames of training data. This leads to the list of the following PFs: {Voiced, Consonant, Vowel, Alveolar, Unround, Fricative, Unvoiced, Front, Plosive}.

3.2. Phonetic Features as Additional Knowledge Source

The architecture we employ for the PF-based EMG decoding system is a *multi-stream* architecture (Metze and Waibel, 2002; Jou et al., 2007), see figure 3. This means that the models draw their *emission probabilities* not from one single source (or stream), but from various sources. The additional sources correspond to phonetic features, like "Vowel" or "Fricative". The conventional EMG phoneme-based recognizer contributes as well. The emission probabilities are always modeled with Gaussian Mixtures.

In (Jou et al., 2007), the authors presented Word Error Rate results on EMG-based speech recognition with phonetic features and showed that using the PFs yields a relative WER improvement of up to 10% over the

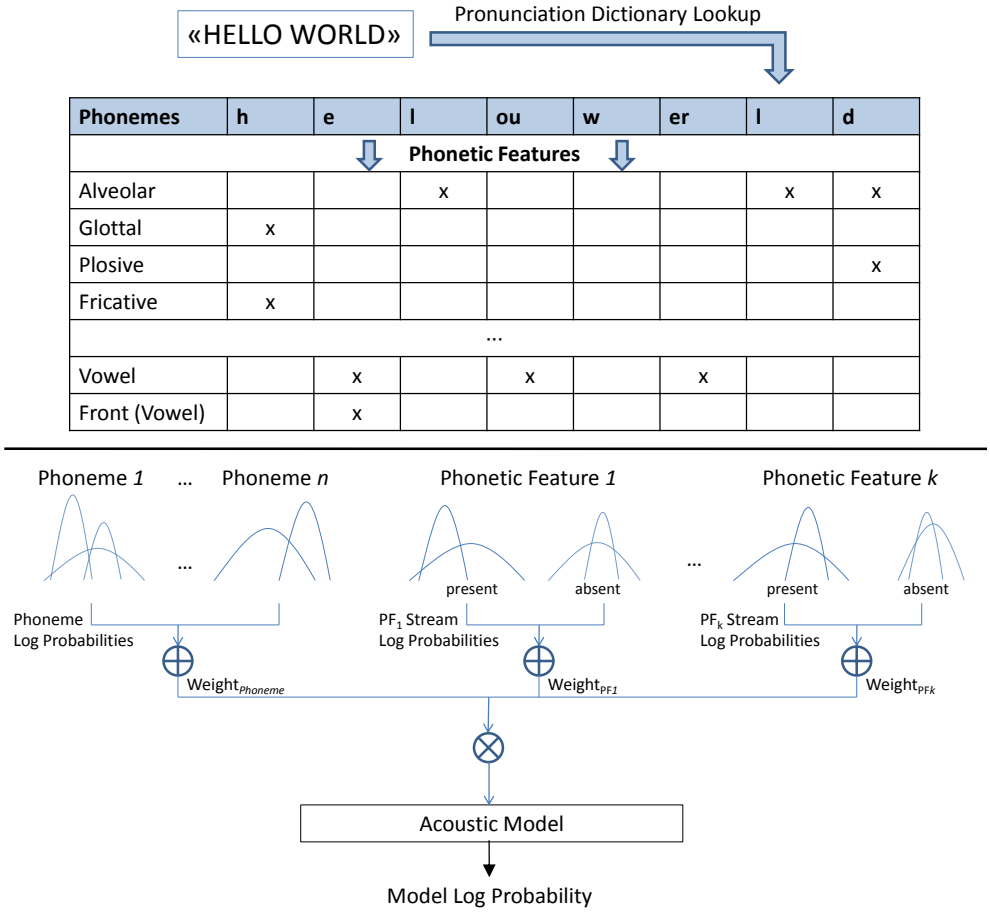

Figure 3: The Multi-Stream Phonetic Features Decoding Architecture. The upper part shows how the PFs are obtained from the phonetic information, the lower part shows the weighting of the various information sources.

conventional phoneme-only model. This system used the *middle* frames of phonemes to train the phonetic feature classifier, since these were assumed to be more stable than beginning and end frames.

We extended the PF recognition system to model PFs for the beginning, middle, and end states of phonemes. We therefore have in each stream six PF models, modeling the *beginning*, the *middle* and the *end* of a *present* or *absent* feature. In addition, each stream has one single model for silence. Since we currently handle planned speech, we refrained from using additional noise models.

For example, the end of "H" in the word "hello" would be modeled using

- the model "H-e" (end of phoneme "H") in the phoneme stream,
- the model "Alveolar-absent" in the "Alveolar" stream,
- the model "Glottal-present" in the "Glottal" stream,
- the model "Plosive-absent" in the "Plosive" stream,
- the model "Fricative-present" in the "Fricative" stream,
- the model "Vowel-absent" in the "Vowel" stream,
- the model "Front-absent" in the "Front" stream,
- etc.

The final score (i.e. the negative log-likelihood $-\log \mathrm{p}(x|\text{model H-e}) =: \text{Score(model H-e)}$) of an observation x for the model "H-e" is then computed by the formula

$$\begin{aligned}\text{score(H-e)} = &\ \text{Weight}_{\text{Phoneme}} \cdot \text{Score(phoneme H-e)} \\ &+ \text{Weight}_{\text{alveolar}} \cdot \text{score(alveolar-absent-e)} \\ &+ \text{Weight}_{\text{glottal}} \cdot \text{score(glottal-present-e)} \\ &+ \text{Weight}_{\text{plosive}} \cdot \text{score(plosive-absent-e)} \\ &+ \text{Weight}_{\text{fricative}} \cdot \text{score(fricative-present-e)} \\ &+ \text{Weight}_{\text{vowel}} \cdot \text{score(vowel-absent-e)} \\ &+ \text{Weight}_{\text{front}} \cdot \text{score(front-absent-e)} \\ &+ \text{further PF scores,}\end{aligned}$$

where the weight constants $\text{Weight}_{\text{stream}}$ may be chosen according to some optimization criterion or be experimentally determined (see next paragraph). Note that the streams are *synchronized*: Only one Hidden Markov Model is constructed, and the streams transit from one state into the next state at the same time frame.

We refer to this multi-stream architecture as *Context-Independent (CI) PF system* and apply it to our corpus. The phoneme-based baseline system, which is described in section 2.2 achieves a word error rate of 47.15% averaged over the 14 speakers. With the context-independent PF system we obtain

45.50% WER. In these experiments, the optimal PF stream weighting was experimentally determined on the test set to be 0.04 for each stream, which leaves a weight of 0.64 for the phoneme stream. The results, including a breakdown for speakers, are given in figure 5.

The decrease from 47.15% to 45.50% WER corresponds to an absolute system improvement of 1.65% WER and a relative improvement of 3.5% WER. In the remainder of this article we will refer to relative improvements in order to compare the impact of our different modeling schemes in relation to the performance level, i.e. independent of the absolute word error rates. Furthermore, we statistically testified the significance of our improvements by applying the Student's t-test for paired measures. In case of the comparison between the phoneme-based baseline system with the context-independent PF system, the significance level α is at 0.7% (0.007). In other words, the chance that the improvement happened by coincidence is only seven in a thousand.

3.3. Data-Driven Bundling of Phonetic Features

So far, we used phonetic feature classifiers as secondary sources of knowledge by augmenting the conventional phoneme-based model. While this yields slight improvements over the phoneme-only classifier, our experimental data indicates that the PF classification is not powerful enough to make the phoneme classification obsolete. In particular, increasing the weight of the PF streams beyond 0.04 shows a decreasing performance, which clearly indicates that the PF streams, even when taken together, are not as accurate as the phoneme models.

It was suggested by (Frankel et al., 2004) that one major shortcoming of the Context-Independent PF recognition system is that features are modeled as statistically independent. The independence assumption is not correct since physiologically every phonetic feature is generated by the interplay of various articulators, i.e. the interdependent activity of several facial muscles.

Therefore, modeling the interdependence of phonetic features should help in creating more accurate PF models and thus might improve the recognition performance. However, for EMG signals it is not clear from the start which features depend on each other, so the choice of a good algorithm to find dependencies between features is crucial. Also, we consider it to be important to find those dependencies in a data-driven fashion, i.e. by an algorithmic process, instead of relying on any kind of rules or educated guesses. We call the process of pooling dependent features together *feature bundling*, since

eventually we will end up with a set of PF models which represent *bundles* of PFs, like "voiced fricative" or "rounded front vowel". Accordingly, we call these models *Bundled Phonetic Features (BDPF)*.

As the data-driven algorithm we opted for a standard decision-tree based clustering approach (Bahl et al., 1991), as it is successfully used in traditional acoustic-based speech recognition to cluster phoneme contexts for context-dependent modeling. This algorithm works by creating a *context decision tree* that assigns classes of context similarities by asking linguistic questions. In our experiments, the predefined set of questions contained questions about phonetic features of the current phoneme. Examples of these categorical questions are: *Is the current phone voiced?* or *Is the current phone a fricative?*. Note that the set of PFs which may occur in these questions consists of about 90 different PFs, i.e. it is not limited to the PFs which actually occur as streams in the multi-stream model.

The context decision tree is created separately for each PF stream, from top to bottom. This means that the initial set of models, such as for the stream "FRICATIVE" consists of six models: namely the beginning, middle and end of a "FRICATIVE" as well as the beginning, middle, and end of "NON-FRICATIVE". Each context question splits one model into two new models. As splitting criterion we used the maximization of the loss of entropy caused by the respective split, calculated over the Gaussian mixture weights. 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 models, is generated for each PF stream. This number was experimentally found to yield optimal recognition results. Note that due to the small number of training utterances, we optimized the parameters for the PF bundling on the test set.

We call the decision tree algorithm described above *PF Bundling Algorithm*. Figure 4 graphically shows an excerpt of an example tree which may have been generated for the VOICED stream.

The full training process consists of three steps, as follows:

1. A common context-independent EMG recognizer, i.e. the baseline recognizer described in section 2.2, is trained on the given training data.

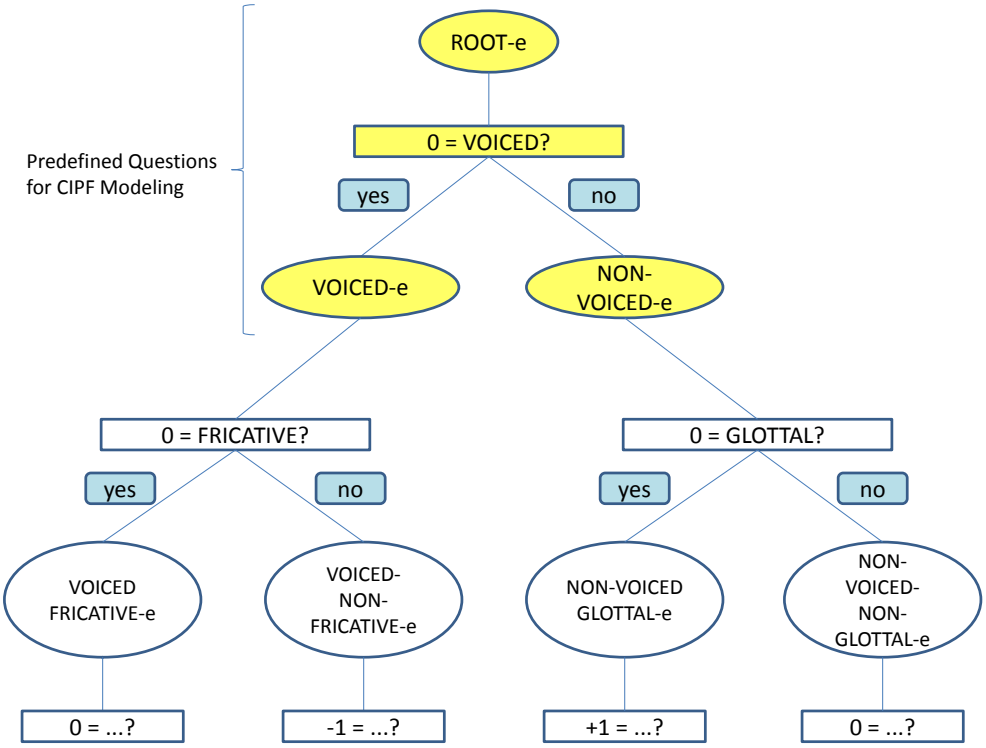

Figure 4: Example of a BDPF tree for the VOICED stream. Note that these models only apply to *end* states of phonemes (*begin* and *middle* states have their own BDPF trees, similar to this one). The upper nodes with yellow background are predefined and are also present when context-independent unbundled PFs are used; when BDPF models are used, the BDPF tree is generated from this basis.

This recognizer uses both phoneme and PF models, but *no* PF bundling yet.

2. The PF bundling algorithm is performed *for each stream*, so that a set of bundled phonetic features (BDPFs) is generated for each stream.
3. Finally, the BDPF EMG recognizer is trained using the models defined in the previous step.

The bundling process is performed on the nine most frequent PFs (see Figure 2), i.e. {Voiced, Consonant, Vowel, Alveolar, Unround, Fricative, Unvoiced, Front, Plosive}. We decided to give the PF streams identical weights and found that under this condition, the optimal weighting of the

PF streams was 0.11 for each of the nine features, while the phoneme stream was factored in by a weight of only 0.01. In other words, recognition accuracy was achieved almost exclusively by the PF classifiers. Further optimization of the stream weights will be investigated in the future, applying automated methods for stream weight training as presented for example in Beyerlein (2000) and Metze (2005).

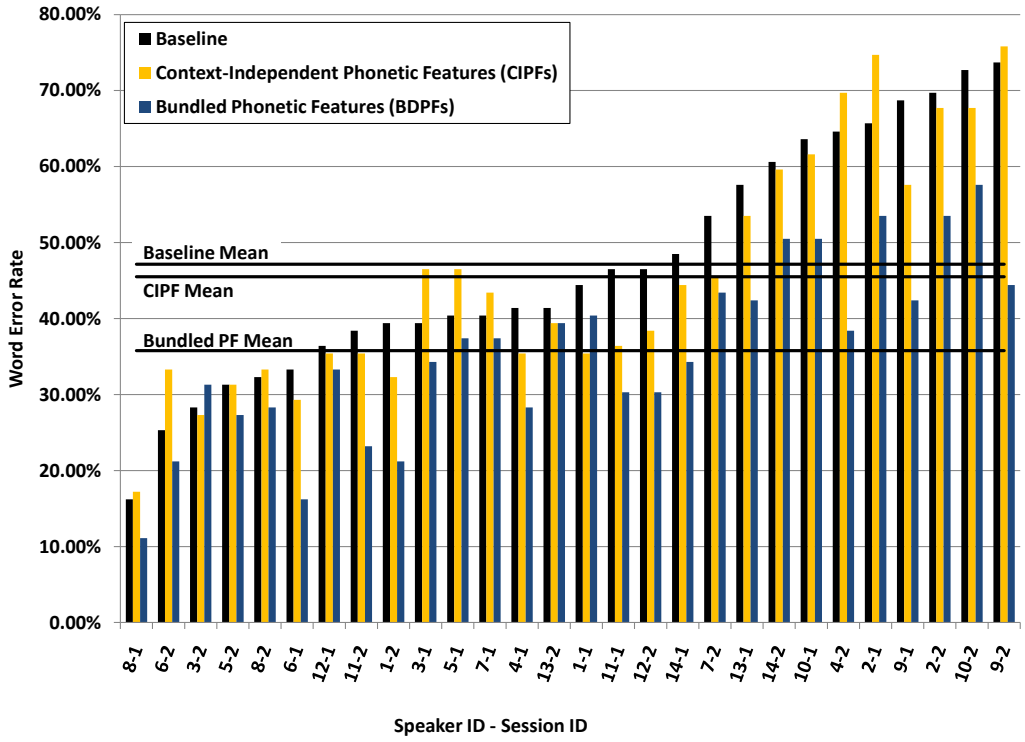

Figure 5: Phonetic Feature Bundling: Breakdown of Word Error Rates for Speaker-Dependent System

The performance of the resulting system can be seen in figure 5, which gives a breakdown of the word error rates for the baseline system, the Context-Independent PF system and the Bundled PF system. The indices on the horizontal axis have the form **Speaker ID - Session ID**. The sessions are ordered according to the baseline performance. It can be seen that while context-independent PFs only give a small improvement over the baseline of 47.15% WER to 45.50%, PF clustering drastically reduces the WER to

35.78%, which is a relative improvement of 24.1% compared to the baseline system. Again, the improvement is statistically highly significant.

3.4. Bundling of Context-Dependent Phonetic Features

The PF bundling algorithm can be adjusted in various ways. In particular, we can create *context-dependent* PFs. This means that the models for a given PF not only depend on the presence or absence of other PFs for the *current* phoneme, but also on the PF's *context*, i.e. its neighboring phonemes.

This modification is easily done by extending the set of predefined linguistic questions which is used for building the context decision tree with questions for the left and right phoneme contexts (i.e. *Is the left context phone a vowel?* or *Is the right context phone a fricative?*). We refer to this system as "Context-Dependent (CD) Bundled PF system."

From traditional acoustic speech recognition it is known that modeling context-dependent phonemes reduces word error rates by about 30%. We investigated the use of triphones for EMG in (Wand and Schultz, 2009b), which gave a relative improvement of 11.5% for a speaker-independent EMG recognizer trained on about 77 minutes of EMG data. The limited success was certainly partly due to the lack of training data, since context-dependent modeling requires a large amount of data for reliable model estimates. However, modeling context-dependent PFs instead of context-dependent phonemes provides a partial solution to this problem, since due to the overall smaller amount of PF units and thus a better data sharing, we have more training data available for PF-based models than for phoneme models.

To investigate the performance of this strategy we ran two experiments: First, we created a *Context-Dependent Bundled PF System* as described above. Second, we computed PF models which depended on their left and right contexts, but *not* on the current value of the other PFs, i.e. were *not* bundled in the sense of our definition. We consistently call this system *Context-Dependent PF System*. The context-independent *Bundled PF System* described in section 3.3 serves as the baseline.

Figure 6 depicts the results of these experiments and shows that by using context-dependent bundled PFs, we can reduce the word error rate by another 12% relative, giving the best average WER of 31.49% so far, again a statistically highly significant improvement. However, if we use context-dependent *unbundled* PFs, the WER drastically increases to about 42%.

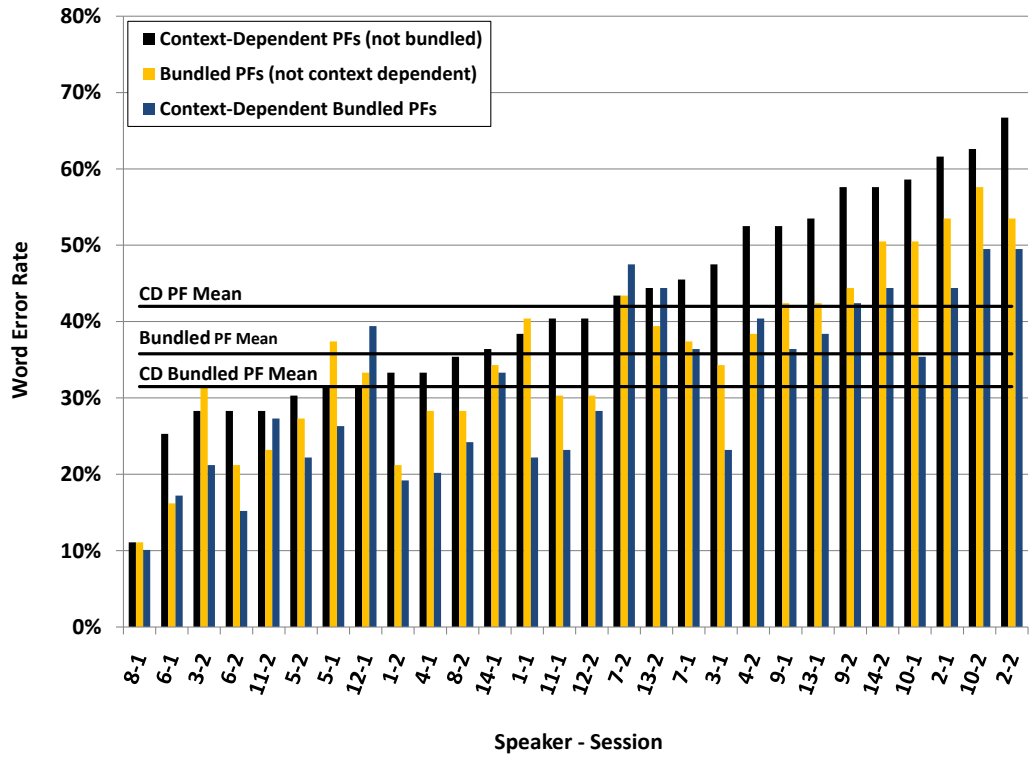

Figure 6: Different Phonetic Feature Bundling Methods: Breakdown of Word Error Rates for the Speaker-Dependent System

The first result is not very surprising, since using context-dependent bundled PFs essentially means giving the decision-tree clustering algorithm more flexibility. As long as the available data is relatively homogeneous and overfitting of the trained PF bundling is avoided, this approach should always give better results. However, the drop in accuracy when only context-dependency is used, clearly proves that PF bundling plays an important role in capturing the variability of PF representations.

With the context-dependent BDPF system in place, it remains to be investigated why the BDPF-based systems perform so much better than the classical phoneme-based system. In order to do so, we ran a further series of experiments, where we started from the phoneme-based system and then *incrementally added* the context-dependent BDPF streams, ordered according to the frequency of the underlying PFs. Since we found in the previous

experiments that in the optimal BDPF system, the phoneme stream practically receives a weight of zero, in this new experiment we distributed the stream weights equally among the available PF streams. This means that after adding one BDPF stream, this stream received a weight of 1, with two BDPF streams, each received a weight of 0.5, and so on.

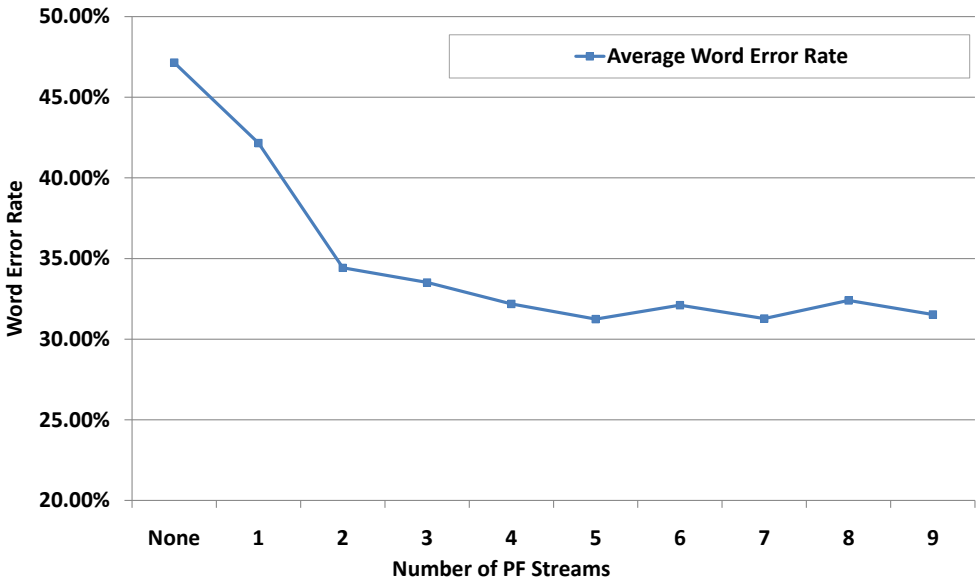

Figure 7: Word Error Rates for Different Numbers of Context-Dependent BDPF Streams. The phoneme-based system is labeled "none".

The results are charted in figure 7: Without PF streams, we have the baseline performance of 47.15% WER, with one stream, we achieve a WER of 42.16%, and with two streams, the WER drops to 34.42%. With five streams, the WER is 31.25%, which is essentially the same result as the WER of 31.49% for nine streams.

It can thus be seen that adding just a few streams already yields a large performance improvement. However we can also see that using only one BDPF stream, which is an approach similar to the one described in (Yu and Schultz, 2003), clearly is not enough: Indeed, the largest performance gain is achieved by adding the second stream, and the further gains up to five streams are still significant.

It should be emphasized that even though the first stream is based on

the VOICED/NON-VOICED pair, and the second stream is based on the CONSONANT/NON-CONSONANT pair, and so on, it would be incorrect to say that e.g. the second stream gives the system the ability to distinguish consonants and vowels, since this distinction will also appear in the first stream due to the PF bundling. Rather it appears that the first stream cannot yield the full distinctive power of a larger system since the bundling is not sufficiently fine-grained: As described in section 3.3, there are 70 codebooks in each PF stream, whereas the context-independent phoneme recognizer contains 136 codebooks.

Growing a larger tree in the first stream might partly remedy this problem if enough training data is present, however our initial experiments showed that this is not the case for our corpus, and that 70 codebooks per stream give optimal results on the EMG-PIT corpus.

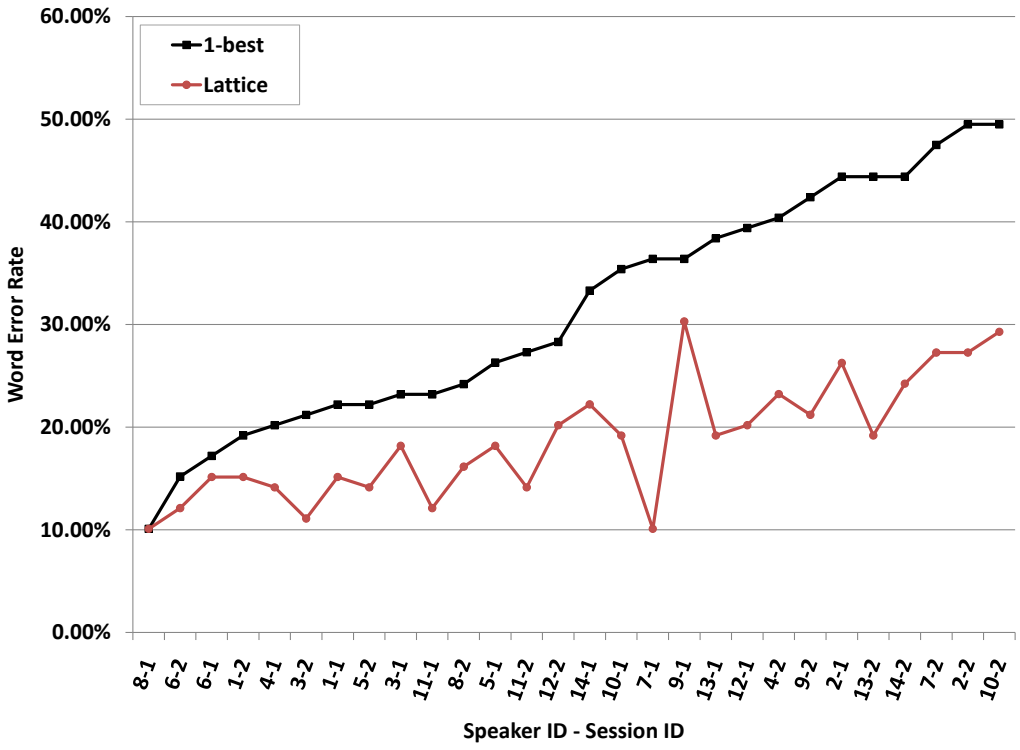

Figure 8: Lattice-based Word Error Rates (Lattice density=100)

Besides the structure of the phonetic models, the systems described above

System	# of Gaussians (total)	# of Gaussians per Stream
Phoneme-based system	290	290
Context-independent PF system	1690	169
Context-dependent BDPF system	2489	248

Table 2: Number of Gaussians for different systems averaged over Speakers

differ in the number of parameters (i.e. Gaussian distributions) which have to be trained. Therefore we compared the total number of Gaussian distributions in the different recognition systems. Note that in the PF-based systems, we must get a much higher number of Gaussians than in the phoneme-based system, since the number is determined automatically during the training process based on the available training data, and in the PF-based systems, each training data sample is indeed used ten times, once for the phoneme stream and once for each of the nine PF streams. The numbers of Gaussians are charted in table 2.

Comparing the phoneme-based system and the BDPF-based system, we can see that if we use *only one* BDPF stream, both systems have got a comparable number of Gaussians. However, as shown in figure 7, the one-stream BDPF system is significantly better than the phoneme-based system, with a performance difference of 10.6% relative.

These experiments suggest that the performance gain achieved by BDPFs is indeed due to the BDPF modeling scheme, and also that with the given constraints in training data size, the multi-stream structure, which has the great advantage of re-using training data for each of the streams, is indeed crucial for success.

As a final experiment in this section, we investigate the word lattice generated by our best recognition system, i.e. the context-dependent bundled PFs system. A word lattice is the common output format of speech recognition systems to provide a memory-efficient representation of a large number of alternative hypotheses. By calculating the lattice-based word error rate of our best system, we get an estimate of how much information is available in our representation of the EMG signal. We investigate a lattice with a density of 100 in our experiments. This means that for an utterance where the reference text contains n words, the lattice pruning retains $100 \cdot n$ nodes. Each node corresponds to a word of the search vocabulary at a specific position

System	PF	Bundled	Context	WER	rel. Gain
Baseline	no	-	-	47.15	-
Context-independent PFs	yes	no	no	45.50	3.5*
Context-dependent PFs	yes	no	yes	41.99	7.7*
Bundled PFs	yes	yes	no	35.78	14.8*
Context-dependent bundled PFs	yes	yes	yes	31.49	12.0*
Lattice (density of 100)	yes	yes	yes	18.76	-

Table 3: Summary of Single Speaker System Performances (averaged over Speakers)

within the utterance.

Figure 8 shows the lattice-based word error rates. We achieve a lattice WER of 18.76% compared to the 31.49% first-best error rate of our currently best system. Table 3 summarizes the results of our experiments on the speaker-dependent setup. We started from a baseline of 47.15% word error rate. This is the averaged performance of the speaker dependent systems trained on each of the 14 subjects from the pilot subset of the EMG-PIT database. The context-independent PF system gave a 3.5% relative gain over the baseline. By using the context and by bundling the PFs we achieved a drastic improvement of 7.7% and 14.8% respectively, and the context-dependent bundled PFs further improved the system by another 12% relative. Our currently best speaker dependent EMG-based speech recognizer gives 31.49% word error rate, with about 10% for the best performing speaker and 50% for the worst performing speaker. The relative performance gains are all statistically significant as indicated by '*' in table 3, with the significance level $\alpha \leq 0.07\%$ for all tests.

4. Multi-Speaker Recognition System

Having successfully introduced PF clustering, in this section we report the results of PF clustering for a multi-speaker scenario. All systems described in this section are trained with the combined training data from the 14 speakers of the EMG-PIT pilot study and then tested for each speaker and each session on the respective test set. The baseline system, as before, is the context-independent phoneme-based recognizer described in section 2.2, which for the multi-speaker scenario yields a WER of 62.15%, averaged over all 28 sessions of the 14 speakers.

In (Wand and Schultz, 2009b), we applied context-dependent (CD) phoneme modeling to the EMG-based speech recognition task for the first time. The context-dependent phoneme recognizer was based on generalized triphones sharing 600 codebooks, where the triphone clustering was performed with the standard decision tree approach (Bahl et al., 1991) which we also use in a modified form for BDPF clustering (see section 3.3). We applied this recognizer, equipped with the TD15 preprocessing, to the multi-speaker scenario. Consequently, the performance improved to 56.55% word error rate.

Given the context-dependent phoneme models, we were able to conduct two experiments to investigate in which aspects the context-dependent phonemes and the context-dependent bundled PFs differ. In one experiment, we used the optimal context-dependent BDPF recognizer described in section 3.4 to train multi-speaker models. Recall that this recognizer consists of a context-independent phoneme-based recognizer augmented by nine additional acoustic knowledge streams of bundled, context-dependent PFs, as shown in figure 3. In a second experiment, we used the same recognizer structure, but used a phoneme stream with *context-dependent* phoneme models instead of the original context-independent phoneme stream. This context-dependent phoneme stream was modeled according to the context-dependent phoneme-based EMG speech recognizer from (Wand and Schultz, 2009b) (see last paragraph).

The goal of these two experiments is to establish that context-dependent BDPFs capture at least as much coarticulation information as traditional context-dependent phonemes. This is a logical assumption: Comparing the way the decision tree clustering algorithm (Bahl et al., 1991) is applied to phoneme models and to phonetic features, we see that the context-dependent BDPF clustering essentially adds significant flexibility to phoneme context clustering, while retaining all the power of the original algorithm, a fact which has also been described in (Yu and Schultz, 2003). The results of these experiments do indeed support this claim. Note that on our corpus, it is necessary to employ a multi-speaker scenario for these experiments, since the amount of training data for any single speaker is too small to allow the training of context-dependent phoneme models for single speakers.

In both multi-speaker experiments, we used the same set of nine PFs which was used in the speaker-dependent experiments. However, due to the larger amount of training data, the BDPF clustering trees were grown to an experimentally determined optimal number of 220 leaf nodes, instead of 70 leaf nodes in the speaker-dependent case.

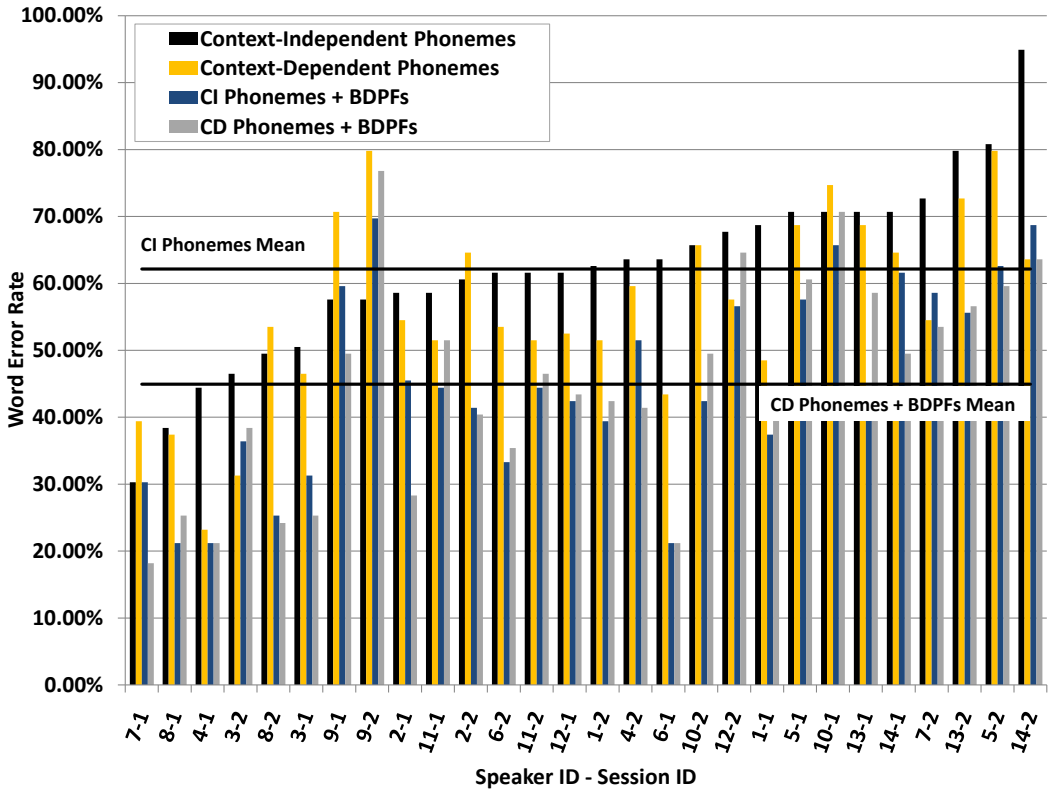

Figure 9: Combination of BDPFs and Context-Dependent and Context-Independent Phoneme Models in a Multi-Speaker Recognizer: Breakdown of Word Error Rates

The detailed results of these experiments are charted in Figure 9. The average results are given in Table 4.

It can be seen that both systems which use context-dependent BDPFs perform significantly better than the context-dependent phoneme system. This clearly shows that as in the speaker-dependent case, phonetic feature bundling significantly increases the modeling power of the system. We also see that the context-dependent phonemes + BDPF system performs only slighty better than the system with context-independent phonemes + BDPFs. The relative gain of 0.64% is not statistically significant. Moreover, it turns out that the optimal weighting of the phoneme stream, compared to the BDPF streams, is still quite low. This observation strongly suggests that indeed all information which the context-dependent phoneme stream yields is also present in the context-dependent BDPF streams.

Context	Model Unit	PFs	WER	rel. Gain
CI	Phoneme Recognizer	(= Baseline)	62.15	-
CD	Phoneme Recognizer		56.55	9.01*
CI	Phoneme Recognizer	with CD Bundled PF Streams	45.24	20.19*
CD	Phoneme Recognizer	with CD Bundled PF Streams	44.95	0.64

Table 4: Summary of Speaker Independent System Performances (averaged over Speakers)

5. Conclusions

In this article we have described the EMG-PIT corpus, a multiple speaker large vocabulary database collection of silent and audible EMG speech recordings. We implemented a new strategy of phonetic feature bundling for modeling coarticulation in EMG-based speech recognition and reported results on speaker-dependent and speaker-independent experimental setups. We could show that the appropriate modeling of the interdependence of phonetic features reduces the word error rate of our baseline system by over 33% relative in the speaker-dependent case, and by about 28% in the speaker-independent system. With this approach we achieved an average word error rate of 31.49% on a 101-word vocabulary task, bringing EMG-based speech recognition within useful range for silent speech applications.

Acknowledgments

The authors would like to thank Szu-Chen (Stan) Jou for his in-depth support with the initial recognition system, his help with the EMG-PIT collection and the data scripts. We also thank Maria Dietrich for recruiting all subjects and carrying out major parts of the database collection. Her study was supported in part through funding received from the SHRS Research Development Fund, School of Health and Rehabilitation Sciences, University of Pittsburgh.

References

Bahl, L. R., de Souza, P. V., Gopalakrishnan, P. S., Nahamoo, D., Picheny, M. A., 1991. Decision Trees for Phonological Rules in Continuous Speech. In: Proc. of the IEEE International Conference of Acoustics, Speech, and Signal Processing (ICASSP). Toronto, Ontario, Canada, pp. 185 – 188.

Beyerlein, P., 2000. Diskriminative Modellkombination in Spracherkennungssystemen mit großem Wortschatz. Ph.D. thesis, RWTH Aachen.

Chan, A., Englehart, K., Hudgins, B., Lovely, D., 2001. Myoelectric Signals to Augment Speech Recognition. Medical and Biological Engineering and Computing 39, 500 – 506.

Denby, B., Schultz, T., Honda, K., Hueber, T., Gilbert, J., 2009. Silent Speech Interfaces. Speech Communication, to appear.

Dietrich, M., 2008. The Effects of Stress Reactivity on Extralaryngeal Muscle Tension in Vocally Normal Participants as a Function of Personality. Ph.D. thesis, University of Pittsburgh.

Frankel, J., Wester, M., King, S., 2004. Articulatory Feature Recognition Using Dynamic Bayesian Networks. In: Proc. of the International Conference on Spoken Language Processing (ICSLP). Jeju Island, Korea, pp. 1202 – 1205.

International Phonetic Association, 1999. Handbook of the International Phonetic Association. Cambridge University Press.

Jorgensen, C., Lee, D., Agabon, S., 2003. Sub Auditory Speech Recognition Based on EMG/EPG Signals. In: Proceedings of International Joint Conference on Neural Networks (IJCNN). Portland, Oregon, pp. 3128 – 3133.

Jou, S.-C., Schultz, T., Walliczek, M., Kraft, F., Waibel, A., Sep 2006. Towards Continuous Speech Recognition using Surface Electromyography. In: Proc. Interspeech. Pittsburgh, PA, pp. 573 – 576.

Jou, S.-C. S., Schultz, T., Waibel, A., 2007. Continuous Electromyographic Speech Recognition with a Multi-Stream Decoding Architecture. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). Honolulu, Hawaii, pp. 401 – 404.

Kirchhoff, K., 1999. Robust Speech Recognition Using Articulatory Information. Ph.D. thesis, University of Bielefeld.

Leveau, B., Andersson, G., 1992. Output forms: Data analysis and applications. In: Selected Topics in Surface Electromyography for Use in the Occupational Setting: Expert Perspective.

Maier-Hein, L., Metze, F., Schultz, T., Waibel, A., 2005. Session Independent Non-Audible Speech Recognition Using Surface Electromyography. In: IEEE Workshop on Automatic Speech Recognition and Understanding. San Juan, Puerto Rico, pp. 331 – 336.

Metze, F., 2005. Articulatory Features for Conversational Speech Recognition. Ph.D. thesis, University of Karlsruhe.

Metze, F., Waibel, A., Sep 2002. A Flexible Stream Architecture for ASR Using Articulatory Features. In: Proc. of the International Conference on Spoken Language Processing (ICSLP). Denver, Colorado, USA, pp. 2133 – 2136.

Morse, M. S., Day, S. H., Trull, B., Morse, H., 1989. Use of Myoelectric Signals to Recognize Speech. In: Proc. 11th Annual Conference of the IEEE Engineering in Medicine and Biology Society. pp. 1793 – 1794.

Morse, M. S., Gopalan, Y. N., Wright, M., 1991. Speech Recognition Using Myoelectric Signals with Neural Networks. In: Proc. 13th Annual Conference of the IEEE Engineering in Medicine and Biology Society. pp. 1877 – 1878.

Morse, M. S., O'Brien, E. M., 1986. Research Summary of a Scheme to Ascertain the Availability of Speech Information in The Myoelectric Signals of Neck and Head Muscles using Surface Electrodes. Comput. Biol. Med. 16 (6), 399 – 410.

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

Sugie, N., Tsunoda, K., 1985. A Speech Prosthesis Employing a Speech Synthesizer – Vowel Discrimination from Perioral Muscle Activities and Vowel Production. IEEE Trans. Biomed. Eng. 32 (7), 485 – 490.

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

Walliczek, M., Kraft, F., Jou, S.-C., Schultz, T., Waibel, A., Sep 2006. Sub-Word Unit Based Non-Audible Speech Recognition Using Surface Electromyography. In: Proc. Interspeech. Pittsburgh, PA, pp. 1487 – 1490.

Wand, M., Schultz, T., 2009a. Speaker-Adaptive Speech Recognition based on Surface Electromyography. BIOSTEC - BIOSIGNALS 2009 best papers. Communications in Computer and Information Science - to appear (CCIS) series by Springer, Heidelberg. To appear.

Wand, M., Schultz, T., 2009b. Towards Speaker-Adaptive Speech Recognition Based on Surface Electromyography. In: Proc. Biosignals. Porto, Portugal, pp. 155 – 162.

Wand, M., Toth, A., Jou, S.-C., Schultz, T., 2009. Impact of Different Speaking Modes on EMG-based Speech Recognition. In: Proc. Interspeech. Brighton, United Kingdom.

Yu, H., Schultz, T., 2003. Enhanced Tree Clustering with Single Pronunciation Dictionary for Conversational Speech Recognition. In: Proc. Eurospeech. Geneva, Switzerland, pp. 1869 – 1872.

Yu, H., Waibel, A., 2000. Streamlining the Front End of a Speech Recognizer. In: Proc. of the International Conference on Spoken Language Processing (ICSLP). Beijing, China, pp. 353 – 356.