TOWARDS SPEAKER-ADAPTIVE SPEECH RECOGNITION BASED ON SURFACE ELECTROMYOGRAPHY

Michael Wand and Tanja Schultz
Cognitive Systems Lab, University of Karlsruhe, Am Fasanengarten 5, Karlsruhe, Germany

Keywords: Speech recognition, Electromyography, Silent speech.

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 paper we present speaker-independent and speaker-adaptive training methods which for the first time allows us to use a large corpus of data from many speakers to reliably train acoustic models. On this corpus we compare the performance of speaker-dependent and speaker-independent acoustic models, carry out model adaptation experiments, and investigate the impact of the amount of training data on the overall system performance. In particular, since our data corpus is relatively large compared to previous studies, we are able for the first time to train an EMG recognizer with context-dependent acoustic models. We show that like in acoustic speech recognition, context-dependent modeling significantly increases the recognition performance.

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, the use of bone-conducting and throat microphones for more reliable recognition in noisy environments or the recognition of whispered speech (Jou et al., 2005) for confidential conversations in the public or for quiet communication that does not disturb bystanders. Other approaches include using optical or ultrasound images of the articulatory apparatus, i.e. (Hueber et al., 2007), or subvocal speech recognition (Jorgensen and Binsted, 2005).

In this paper, 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 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, (Chan et al., 2002) showed that myoelectric signals can be used to discriminate a small number of words. Other related works reported success in several different aspects of EMG speech recognition (Jorgensen and Binsted, 2005; Maier-Hein et al., 2005). In 2006, (Jou et al., 2006b) 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

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

Wand M. and Schultz T. (2009).
TOWARDS SPEAKER-ADAPTIVE SPEECH RECOGNITION BASED ON SURFACE ELECTROMYOGRAPHY.
In *Proceedings of the International Conference on Bio-inspired Systems and Signal Processing*, pages 155-162
DOI: 10.5220/0001549601550162

preprocessing (Wand et al., 2007), and advances in acoustic modeling using articulatory features in combination with phone models (Jou et al., 2006a). However, the described experiments were based on relatively small amounts of data, and consequently were limited to speaker-dependent modeling schemes. In (Maier-Hein et al., 2005), 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 paper reports for the first time EMG-based recognition results on continuously spoken speech comparing speaker-dependent, speaker-adaptive, and speaker-independent acoustic models. We investigate different signal preprocessing methods and the potential of model adaptation. For this purpose 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 32% WER on a testing vocabulary of 108 words (Jou et al., 2006b). The training data of this baseline recognizer consisted of 380 phonetically-balanced sentences from a single speaker, which is about 10 times larger than the training set we use for the speaker-dependent systems reported in this paper (see below for details on the training data).

The paper is organized as follows: In section 2, we describe the used data corpus and the method of data acquisition. 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 accuracy of the different methods and section 5 concludes the paper.

2 DATA ACQUISITION

For data acquisition, 13 speakers were recorded. Each speaker recorded two sessions with an in-between break of about 60-90 minutes, during which the electrodes were *not* removed. The recordings were collected as part of a psychobiological study investigating the effects of psychological stress on laryngeal function and voice in vocally normal participants (Dietrich, 2008; Dietrich and Abbott, 2007). The sentence recordings were obtained at the beginning and at the very end of the stress reactivity protocol. Participants were recruited at the University of Pittsburgh, Carnegie Mellon University, and Chatham University for a speech recognition study, but were also confronted with an impromptu public speaking task.

One session consisted of the recording of 100 sentences, half of which were read audibly, as in normal speech, while the other half were mouthed *silently*, without producing any sound. In order to obtain comparable results to previous work, we report recognition results from the *audibly spoken* sentences only.

Each block of audible and mouthed utterances had two kinds of sentences, 40 individual sentences that were distinct across speakers and 10 “base” sentences which were identical for each speaker. We used the individual block for training and the “base” sentences as test set.

The corpus of audible utterances had the following properties:

Speakers	13 females speakers aged 18 - 35 years with no known voice disorders
Sessions	2 sessions per speaker
Average Length (total) (training set) (test set)	 231 seconds per session 179 seconds 52 seconds
Domain	Broadcast News
Decoding vocabulary	101 words

The total duration of all audible recordings was approximately 100 minutes (77.5 minutes training set, 22.5 minutes test set).

During any session, “base” and individual sentences were recorded in a randomized order.

In order to compare our results with previous work, we additionally use the data set reported in (Jou et al., 2006b), which consists of a training set of 380 phonetically balanced sentences and a test set of 120 sentences with a duration of 45.9 and 10.6 minutes, respectively.

This results in a corpus of 14 speakers, where speaker 14 (with only one session) corresponds to the speaker from (Jou et al., 2006b) described above and is treated separately. In the results section, a result denoted with e.g. **3-2** means: Speaker 3 (out of 14), session 2 (out of 2).

The EMG signals were recorded with six pairs of Ag/Ag-CL electrodes attached to the speaker’s skin capturing the signal of the articulatory muscles, namely the *levator angulis oris*, the *zygomaticus major*, the *platysma*, the *orbicularis oris*, the *anterior belly* of the *digastric* and the *tongue*. Eventually, the signal obtained from the *orbicularis oris* proved unstable and was dropped from the final experiments. The EMG signals were sampled at 600 Hz and filtered with a 300 Hz low-pass and a 1 Hz high-pass

filter.

In addition, the audio signal was recorded simultaneously using a professional head-mounted microphone. Note that the microphone attachment did not interfere with the EMG recordings. Details regarding the EMG data acquisition setup can be found in (Jou et al., 2006b), most of the information therein also applies to this work. See picture 1 for an example of the electrode placement.

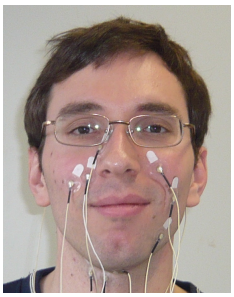

Figure 1: Electrode Positioning.

3 EMG-BASED SPEECH RECOGNIZER

The initial EMG recognizer was the same one as in (Wand et al., 2007), which in turn was set up according to (Jou et al., 2006b). It used an HMM-based acoustic modeling, which was based on fully continuous Gaussian Mixture Models. For the initial context-independent 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, it could be observed that most speaker dependent codebooks had about one to four Gaussians after the initial merge-and-split codebook generation.

The training concept works as follows: The time-aligned training data is used either for a full training run, or we use pre-trained models and adapt them to the current task. The latter is especially important since it allows setting up a recognizer with a very small amount of individual training data.

During the decoding, we use the trained acoustic model together with a trigram language model trained on Broadcast News data. The testing process consists of an initial testing run followed by a lattice rescoring in order to obtain optimal results.

In section 4.5, we present our investigations on using context-dependent modeling for the EMG recognizer.

3.1 Initialization

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 a Broadcast News (BN) speech recognizer trained with the Janus Recognition Toolkit (JRTk). The recognizer is HMMbased, 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 (Yu and Waibel, 2000).

Previously, (Jou et al., 2006a) demonstrated the anticipatory effect of EMG signals compared to audio signals and showed that taking this effect into account significantly improves performance. Accordingly, we modeled this effect by delaying the EMG signal for an amount of 0 ms to 90 ms (in steps of 10 ms), since in (Jou et al., 2006a) the optimal delay is found to be around 30 ms to 60 ms. The effect of this delaying is charted in section 4 and indicates that the best performance is to be found at 50 ms. Therefore, we report results for the remainder of experiments at 50 ms delay.

3.2 Feature Extraction

We compare two methods for feature extraction: *Time-domain Features* and *Wavelet Transform* as in (Wand et al., 2007).

For the time-domain features, we use the definitions following (Jou et al., 2006b): 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. $S(\mathbf{f},n)$ is the stacking of adjacent frames of feature $\mathbf{f}$ in the size of $2n+1$ ($-n$ to n) frames. 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 by (Walliczek et al., 2006).

In the above work, the best WER was obtained with the E4 feature defined as:

$$\mathbf{E4} = S(\mathbf{f2},5), \text{where } \mathbf{f2} = [\bar{\mathbf{w}}, \mathbf{P_w}, \mathbf{P_r}, \mathbf{z_r}, \bar{\mathbf{r}}].$$

For comparison, we use a Redundant Discrete Wavelet Transform (see e.g. (Shensa, 1992)) with a 14-tap q-shift filter according to (Kingsbury, 2000). We perform a wavelet decomposition to decompose level 5 and use both detail and approximation coefficients as features. The transformed signal is resampled to obtain a 10 ms frame shift as for the E4 feature.

In both training methods, we performed a stacking of the features from the five EMG channels to create a final "joined" feature consisting of the synchronized data from all channels.

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 labeled data. The diemnsionality of the final feature was set to 32 according to (Jou et al., 2006b). 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

We performed speaker adaptive training 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 ("cross-speaker training + adaptation").

3.5 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 101 words. On the test sentences, the trigram-perplexity of the language model is 24.24.

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

4 EXPERIMENTAL RESULTS

4.1 Speaker-Dependent Training

Figure 2 shows the word error rates of the speaker-dependent recognition systems trained on the training data of *one* session and tested on test data from *the same* session. The average performance of the recognizer is 57.27% with the E4 preprocessing and 67.98% with RDWT (Wavelet) preprocessing. The E4 preprocessing seems to be consistently better than the RDWT across the speakers and sessions: The E4 preprocessing, which was introduced by (Jou et al., 2006b), remains the current state-of-the-art for EMG speech recognition. Wavelet preprocessing produces a higher WER in most cases. While the Wavelet performance generally follows the same trend as the E4 performance, we see a notable exception for speakers 10 and 11.

Overall we conclude from Figure 2 that the performance of EMG speech recognition varies considerably over speakers and even varies between different sessions of the same speaker. In general, the variance within speaker is smaller than across speakers, i.e. the two word error rates of the two sessions of the same speaker are similar. However, there are exceptions, such as for speaker 3, 4, 5, and 7. Furthermore, no pattern can be observed between order of session and performance. The variance within the data may be attributed to the challenges in longer duration EMG recording, namely that the EMG signal highly depends on the electrode contact, skin conductance, and environmental changes.

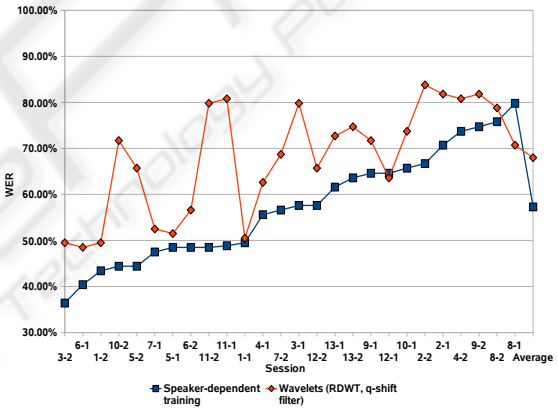

Figure 2: Speaker-dependent WER: E4 vs. RDWT.

4.2 Audio Delay

To investigate if the findings of (Jou et al., 2006a) on the anticipatory effect of the EMG signal carry over to multiple speakers, we trained the speaker-dependent recognizer with the E4 feature for all speakers and for EMG signal delays in the range of 0 ms to 90 ms.

The average performance of the recognizers for speakers 1 to 13 is charted in Figure 3. It can be seen that the optimal delay is achieved at about 50 ms, which confirms the results of (Jou et al., 2006b). For a single speaker, the curve may be less smooth than the average, but in almost all cases the optimal delay for each speaker was in the range of 30 ms to 60 ms. Experiments in (Jou et al., 2006a) indicated that the anticipatory behavior of the EMG signal may not be uniform, but depends on the muscle group involved in

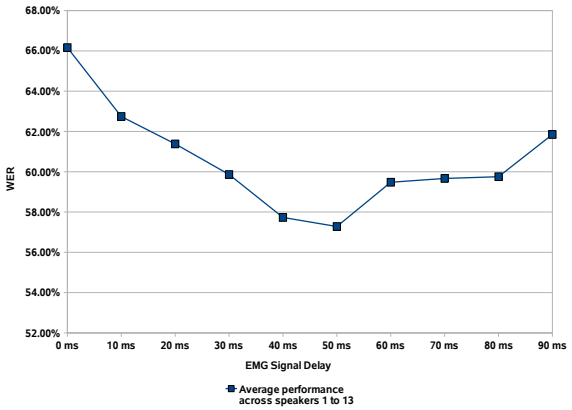

Figure 3: Comparison of Word Error Rates with Different EMG Delays.

producing the respective sounds; this may explain the variability of the optimal delay.

4.3 Cross-Speaker and Adaptive Experiments

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.
- Cross-Speaker Training + Adaptation: Like Cross-Speaker Training, but the resulting system is then adapted toward the test speaker using MLLR adaptation (Leggetter and Woodland, 1995) on the training data from one session. As above, testing is done on the test data from the same session.

Figure 4 shows the results of these experiments and indicates that the speaker-dependent and adaptive systems clearly outperform the speaker-independent 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 8 and 9 and to some extent 2 and 4, the performance of the adapted system does not degrade similarly and may outperform the speaker-dependent system. Further investigations are necessary to find appropriate adaptation schemes for the purpose of EMG speech recognition.

4.4 Impact of Training Data Amount

In this section we investigate the impact of the amount of training data on the performance of the EMG recognizer. For this purpose we compare the two setups Speaker-Dependent Training and Cross-Speaker Training + Adaptation as described above. The difference lies in the fact that we run training and test on speaker 14, who has recorded a larger training and test set than the other speakers. In total we have 380 sentences for training and 120 sentences for testing of this speaker.

Figure 5 shows the results of this experiment. Obviously, if more than 10 sentences of training material are given, the development of speaker-dependent models if giving a gain. This may result from the fact that the recognizer uses context-independent phones and that the training set comprises of phonetically balanced sentences, thus allowing us to update few numbers of Gaussians on a very small set of training sentences.

It is notable that a training data set beyond 120 words does not significantly improve the recognition accuracy of the recognizer. This applies to both speaker-dependent and speaker-adaptive systems.

4.5 Context-Dependent Modeling

In this section we report on the effects of using *context-dependent modeling* for the acoustic models of the EMG recognizer. From the field of acoustic speech recognition, it is known that modeling a phoneme depending on its right and left neighboring phonemes drastically increases the recognition accuracy, provided that the training data corpus is large enough to offer sufficient training samples for the increased number of acoustic models. It is expected that the recorded data for a given pronounced phoneme in the EMG signal also depends strongly on the context in which the phoneme is spoken.

With speaker-independent and speaker-adative training in place, we now have a large enough training corpus to allow for context-dependent modeling. To the best of our knowledge, this is the first report ever on context-dependent modeling for an EMG recognizer.

We used a context-dependent recognizer setup based on (generalized) triphones sharing 600 codebooks. This means that we create a set of acoustic models, each of which takes into consideration

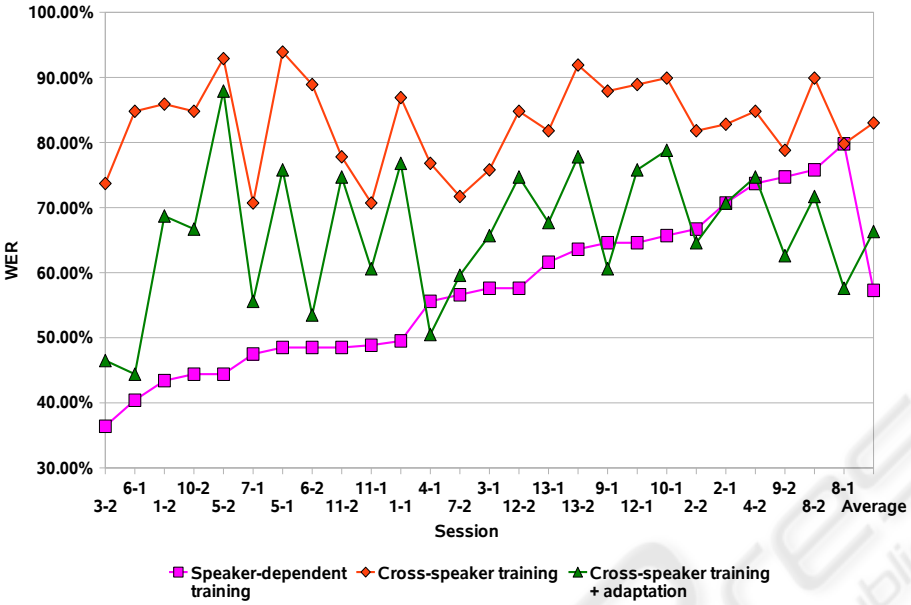

Figure 4: Comparison of Word Error Rates with Different Adaptation Methods.

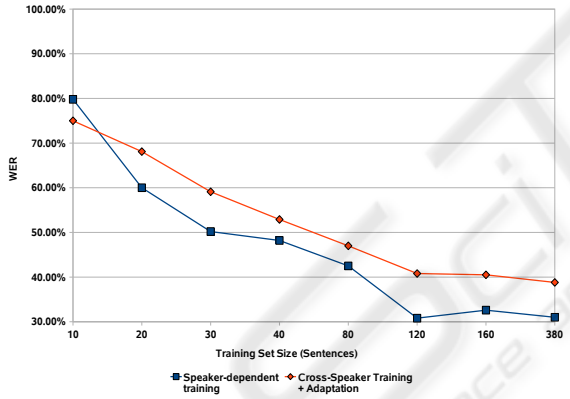

Figure 5: Comparison of Word Error Rates with Different Number of Training Sentences.

not only the current phoneme, but also the left and right neighboring phonemes (hence the name *triphones*). However it would be unfeasible to train separate acoustic models for *each* combination of three phonemes, therefore we use automatic context clustering to merge contexts which have similar effects on the current phoneme.

The context clustering works by creating a *context decision tree* (see e.g. (Finke and Rogina, 1997)), which classifies triphones by asking linguistic questions about the triphones. The set of all possible questions is predefined, examples of these categorical questions are: *Is the left-context phone a back vowel?* or *Is the right-context phone a fricative?*. The context tree is created from top to bottom, i.e. the initial set of acoustic models consists of the context-independent models, and 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. The process ends when a predetermined 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 600 tree leaves, corresponding to 600 independent acoustic models, is generated, since this number was experimentally found to yield optimal results.

So the general training process is as follows:

- First, an ordinary context-independent EMG recognizer is trained on *all* available data (including the training data of the speaker to be tested). We call this setup "Speaker-Independent Training" as opposed to the cross-speaker experiments above. We use the performance of this preliminary recognizer as a baseline during these experiments.
- In a second step, the context decision tree is grown as described above.
- The final context-dependent EMG recognizer is trained using the 600 acoustic models defined in the previous step.

Figure 6 shows the recognition results of the context-dependent recognizer. The overall average

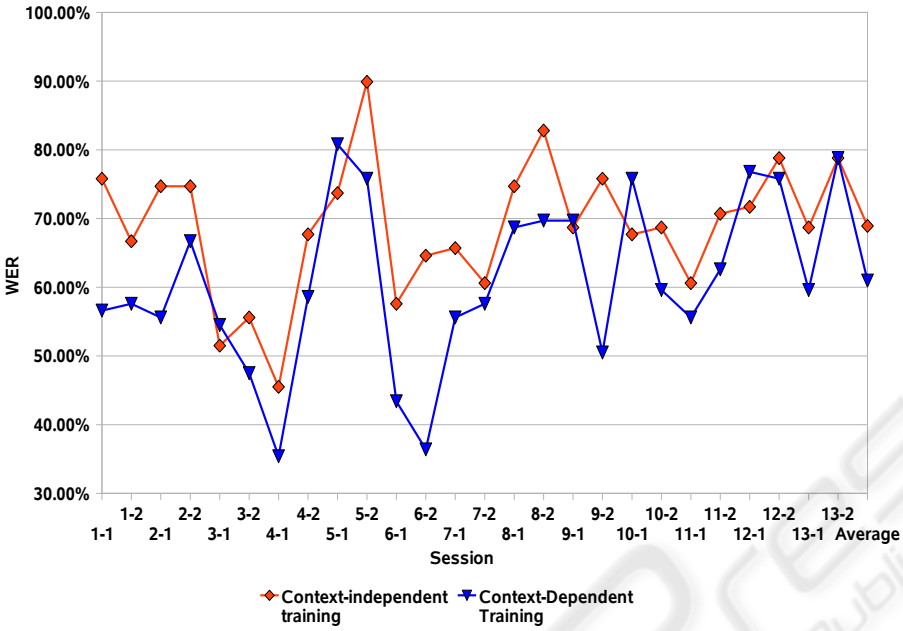

Figure 6: Comparison of Word Error Rates on a Context-Dependent and a Context-Independent System.

performance of the context-independent system has a WER of 68.92%, which by context-dependent modeling drops to 60.97%. This is a relative improvement of 11.5%. As can be seen, for the majority of speakers, context-dependent modeling significantly improves the recognition performance.

On this basis, it is worthwhile to take a closer look at the context decision tree. Since the context questions are asked in order of entropy decrease, the fact that a question occurs close to the root of a context decision tree means that this question distinguishes two representations (contexts) of a phoneme which create relatively *distinct* EMG signals.

A first manual inspection indeed shows a certain pattern in the context questions: For *vowels*, in many cases the first context questions (corresponding to high entropy decreases) ask whether the right or left neighboring phoneme is a (bi-)labial or labialized consonant. This suggests that the lip position not only is well picked up by the EMG electrodes, but also strongly influences the articulation of adjacent phonemes. For *consonants*, in many cases an early context split asks whether the current phoneme is preceded or followed by silence. Therefore we may assume that the articulation of such consonants differs depending on its position in the middle or at the end of a word.

Finally, we compared this result with the properties of a context decision tree of an *audible* speech recognizer, namely the BN speech recognizer described in section 3.1, and found that in the case of acoustic speech recognition, these findings do *not* hold (see also (Finke and Rogina, 1997)). Therefore we can conclude that our context decision tree actually captures EMG-specific articulation properties.

5 CONCLUSIONS

We have compared EMG speech recognition on a speaker-dependent, a speaker-independent and a speaker-adapted system on a newly developed corpus of training and test data. We compared the performance of the new corpus to the data used by (Jou et al., 2006b) and could support the findings therein: both in performance and in EMG delay properties, our new data set shows similar properties as the data set of (Jou et al., 2006b).

We reported first results on the performance of EMG speech recognition across multiple speakers and sessions. While we found that for the majority of cases a speaker- and session-dependent EMG system still performed best, we showed that the MLLR adaptation method is feasible for EMG speech recognition and generally yields good results, which makes the building of speaker-adaptive EMG recognition systems possible.

Finally, we successfully applied context-dependent phoneme modeling on EMG speech recognition and showed that it significantly increases

the recognition performance of an EMG recognizer trained on multi-speaker data.

ACKNOWLEDGEMENTS

We would like to thank Maria Dietrich for the collection of the data and her advisor Katherine Verdolini Abbott for her support. This study was supported in 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.

Many thanks go to Szu-Chen (Stan) Jou for his ever-patient help with the Janus Recognition Toolkit and the EMG decoding scripts!

REFERENCES

Chan, A., Englehart, K., Hudgins, B., and Lovely, D. (2002). Hidden Markov Model Classification of Myolectric Signals in Speech. *Engineering in Medicine and Biology Magazine, IEEE*, 21(9):143–146.

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

Dietrich, M. and Abbott, K. V. (2007). Psychobiological framework of Stress and Voice: A Psychobiological Framework for Studying Psychological Stress and its Relation to Voice Disorders. In: *K. Izdebski (Ed.): Emotions in the Human Voice (Vol.II, Clinical Evidence, pp. 159-178).* San Diego, Plural Publishing.

Finke, M. and Rogina, I. (1997). Wide Context Acoustic Modeling in Read vs. Spontaneous Speech. In *Proc. ICASSP*, volume 3, pages 1743–1746.

Hueber, T., Chollet, G., Denby, B., Dreyfus, G., and Stone, M. (2007). Continuous-Speech Phone Recognition from Ultrasound and Optical Images of the Tongue and Lips. In *Proc. Interspeech*, pages 658–661.

Jorgensen, C. and Binsted, K. (2005). Web Browser Control Using EMG Based Sub Vocal Speech Recognition. In *Proceedings of the 38th Hawaii International Conference on System Sciences*.

Jou, S.-C., Maier-Hein, L., Schultz, T., and Waibel, A. (2006a). 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*.

Jou, S.-C., Schultz, T., and Waibel, A. (2005). Whispery Speech Recognition Using Adapted Articulatory Features. In *Proc. ICASSP*.

Jou, S.-C., Schultz, T., Walliczek, M., Kraft, F., and Waibel, A. (2006b). Towards Continuous Speech Recognition using Surface Electromyography. In *Proc. Interspeech*, Pittsburgh, PA.

Kingsbury, N. G. (2000). A Dual-Tree Complex Wavelet Transform with Improved Orthogonality and Symmetry Properties. In *Proc. IEEE Conf. on Image Processing, Vancouver*.

Leggetter, C. J. and Woodland, P. C. (1995). Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer Speech and Language*, 9:171–185.

Maier-Hein, L., Metze, F., Schultz, T., and Waibel, A. (2005). Session Independent Non-Audible Speech Recognition Using Surface Electromyography. In *Proc. ASRU*.

Shensa, M. J. (1992). The Discrete Wavelet Transform: Wedding the À Trous and Mallat Algorithms. *IEEE Transactions on Signal Processing*, 40:2464–2482.

Walliczek, M., Kraft, F., Jou, S.-C., Schultz, T., and Waibel, A. (2006). Sub-Word Unit Based Non-Audible Speech Recognition Using Surface Electromyography. In *Proc. Interspeech*, Pittsburgh, PA.

Wand, M., Jou, S.-C. S., and Schultz, T. (2007). Wavelet-based Front-End for Electromyographic Speech Recognition. In *Proc. Interspeech*.

Yu, H. and Waibel, A. (2000). Streamlining the Front End of a Speech Recognizer. In *Proc. ICSLP*.