COMPENSATION OF RECORDING POSITION SHIFTS FOR A MYOELECTRIC SILENT SPEECH RECOGNIZER

Michael Wand, Christopher Schulte, Matthias Janke, and Tanja Schultz

Cognitive Systems Lab, Karlsruhe Institute of Technology, Germany

ABSTRACT

A myoelectric *Silent Speech Recognizer* is a system which recognizes speech by capturing the electrical activity of the human articulatory muscles, thus enabling the user to communicate silently. We recently devised a recording setup based on *electrode arrays* with multiple measuring points. In this study we show that this allows to compensate for shifts of the recording position, which happen when the array is removed and reattached between system training and application. We present a method which determines the amount of recording position shift; compensation is performed by linear interpolation. We evaluate our method by running recognition experiments across recording sessions and obtain a Word Error Rate improvement of 14.3% relative on the development set and 12.9% relative on the evaluation set, compared to using classical session adaptation.

Index Terms— Silent Speech Interfaces, EMG, EMG-based Speech Recognition, Adaptation, Signal Interpolation

1. INTRODUCTION

Humans use speech as their most natural method of communication: It is easily produced, allows to convey a large amount of information in a short time, and has additionally become a means of controlling technical devices. However, all these usages require speech to be clearly audible, which incurs disturbance for bystanders, compromised privacy, lack of robustness in noisy environments, and exclusion of speech-disabled people. Over the past few years, we have developed a *Silent Speech Recognizer* based on surface electromyography (EMG), where the electrical activity of the articulatory muscles is captured by EMG electrodes attached to the subject's face [1]. This allows to process speech even when no acoustic signal is produced, alleviating the issues of conventional speech communication mentioned above.

Recently, we introduced a recording setup using *electrode arrays* [2], which are grid structures with multiple EMG measuring points. Here the goal is to substantially enhance the EMG signal processing and feature extraction, and this paper is part of these efforts. We tackle *session discrepancies*: Ideally, a user can pre-train the system at his or her convenience, and then immediately apply it when desired. In [3] we showed that such *session-independent* systems are feasible and that they can be further improved by *session adaptation*. However, the systems presented in [3] gain robustness from being trained on many sessions, no explanation or correction for inter-session variations is given.

In this study we use the multi-channel EMG array data to tackle session discrepancies on a signal-oriented level: We consider pairs of sessions, namely the *source* session, used to train the recognizer, and the *target* session, on which recognition is performed. We compute the difference between the array positionings of source and target session and then *interpolate* the EMG signals of the target session so that the position shift is compensated for. This yields significant improvements on *cross-session* recognition tasks.

2. RELATED WORK

Silent Speech recognition is an emerging technology, and a lot of research on different aspects on the topic is currently taking place. Technologies under investigation include *Permanent Magnetic Articulography (PMA)* [4], where small magnets are glued to the subject's articulators, processing of ultrasound and/or optical images of the articulatory tract [5], as well as enhancement of very quiet speech signals, e.g. by using a stethoscopic microphone [6]. Investigations regarding the EMG-based approach include the application of electromyography in special circumstances, e.g. for firefighters who may be prevented from speaking because they wear a breathing apparatus [7], recognition of disordered speech [8], language-dependent challenges [9], and within the context of our array-based recording setup, signal source decomposition to retrace signal sources and remove artifacts [10].

3. DATA CORPUS

Our recording setup follows [2] (where it is named "Setup B"). We use the multi-channel EMG amplifier *EMG-USB2* produced and distributed by *OT Bioelettronica*, Italy. Two EMG arrays are used, a chin array with a row of 8 electrodes

This research project was partially funded by the German Research Foundation (DFG), Research Grant *MAPS - Myoelectric Array-based Processing of Speech*.

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

Fig. 1. EMG array positioning

with 5 mm inter-electrode distance (IED), and a cheek array with 4×8 electrodes with 10 mm IED, see figure 1. Sampling is performed at 2048Hz.

During the recordings, which took place over the course of several weeks, the supervisors were instructed to attach the array as accurately as possible. Yet, it proved all but impossible to hit a position to millimeter precision, or equivalently, to measure the repositioning between sessions at a millimeter scale. Therefore the corpus does *not* contain exact information about the repositioning between sessions: We focus entirely on our algorithmic solution to this problem.

During recording sessions, we frequently observed that the signals from one or several EMG channels exhibited strong superimposed artifacts, which harm our position shift estimation algorithm. Therefore after a recording session, we manually checked the signals and marked those channels which did not contain useful EMG signals (there also exist automatic heuristics for this task [10]).

Altogether we use data from 21 recorded sessions of 4 speakers, the 5 sessions from speaker 4 are set aside for statistical evaluation. Each session consists of 160 training sentences, 20 adaptation sentences, and 20 test sentences, which are in English language and read in normal, audible speech; additional recordings of silently mouthed speech are available but are not used in this study since it is difficult to obtain exact phone-level alignments for them [11], complicating application of our position shift estimation algorithm. However note that we ran several studies regarding the difference between audibly spoken and silently mouthed speech, and how to deal with it [11–14]. The training sentences may differ between sessions, the adaptation sentences and the test sentences are the same for each session. The audio signal is always parallely recorded with a close-talking microphone, EMG and audio signal are aligned with a 50ms delay according to [15].

As in [2], our initial recordings used *bipolar* derivation, where the potential difference between two adjacent channels in a row is measured. In contrast, *unipolar* derivation means that the difference between each measuring point and a neutral reference (in our experiments, the back part of the user's neck) is captured. Bipolar recording reduces in particular common-mode artifacts, but one also loses access to the "raw" signal: It is possible to compute a bipolar signal from unipolar recordings, namely by application of a simple spatial filter, but not the other way round.

Subset	Av. session length (sec)	# of sessions / speakers	Total data length (mm:ss)
Development corpus			
Training	494	16/3	131:48
Adaptation	70	8/3	9:21
Test	63	8/3	8:26
Total	150 minutes		
Evaluation corpus			
Training	581	5/1	48:25
Adaptation	80	3/1	4:00
Test	71	3/1	3:33
Total	56 minutes		

Table 1. Data corpus. Training sessions were recorded either unipolarly or bipolarly, sessions which are used for adaptation and testing were always recorded in unipolar configuration.

Our features are based on bipolar EMG signals, see section 4.3. During the course of our experiments, it became clear that bipolar signals are unsuited for interpolation, so we changed our recording setup to unipolar derivation. Therefore we have two sets of sessions: Sessions where we used bipolar recording can only be used as source sessions, i.e. for recognizer training. Target sessions, on which position shift compensation, adaptation, and testing of the recognizer is performed, must have been recorded unipolarly, of course, these session are also used as source sessions. This yields a total of 36 pairs of training session and test session in the development corpus and 12 pairs in the evaluation corpus. Table 1 summarizes our data.

4. POSITION SHIFT ESTIMATION AND COMPENSATION

In this section we present the main algorithm of this paper. We first describe how we compensate for a given position shift between the source session and the target session, then we describe how we determine the amount of the position shift.

4.1. Position shift compensation by linear interpolation

We intend to compute a rotation and shift for the 8×4-channel cheek array, so that the EMG channels of the target session match the channels of the source session as closely as possible. We emphasize that it is not the goal to actually remove and reattach the array, we rather correct the misplacement on a signal processing level by computing a "virtual shift". Preliminary experiments indicated that interpolating the signals of the 8-channel chin array does not improve the recognition results; this is most likely due to the shape of this array: All measuring points are in a single line, which makes it difficult to robustly interpolate the signal for position shifts perpendicular to this line. We therefore concentrate on the 4×8-channel cheek array and do not change the signals of

the 8-channel chin array in any way. Only the signals of the target session are processed, the source session remains untouched.

We expect the following array positioning distances between source and target session: The array may be shifted maximally 10mm horizontally and vertically, with a step size of 1mm. In addition, rotations around the center of the shifted array of up to 5° are allowed, in steps of 1°. Altogether, this yields $21 \cdot 21 \cdot 11 = 4851$ different possibilites. Compensating for a shift requires to compute hypothetical EMG signals at positions in-between measuring points, which is done by *linear interpolation*: Assume that we need to compute the signal $x_p[n]$ at position $p = (p_x, p_y)$, located within a square formed by four adjacent measuring points $m_1, \ldots, m_4$. If $x_1[n], \ldots, x_4[n]$ are the measured signals at $m_1, \ldots, m_4$, we estimate $x_p[n]$ by

$$x_p[n] = \sum_{i=1}^{4} \frac{1}{||m_i - p||} \cdot x_i[n],$$

where $||m_i - p||$ is the distance between points p and m_i. We note here that this assumes that the amplitude of an EMG signal is antiproportional to the distance between the source and the measuring point, which is physiologically inexact, but may nonetheless serve as a first approximation. Also note that near the borders of the array, we must rely on a smaller set of data points for interpolation.

4.2. Estimating the Position Shift

In order to estimate the amount of shift and rotation between the source and the target session, we use the 20 *adaptation sentences* of both sessions and proceed as follows: First, we forced-align the parallely recorded acoustic signal in order to obtain phone-level alignments of the EMG signals. This is one of our standard methods, originally proposed in [15]. Then we compute the interpolated raw EMG signals for each possible combination of shift and rotation for the 20 adaptation sentences of the target session.

Now for each such shift and rotation, we compute root mean square (RMS) features from the interpolated EMG signal, after processing it with a bipolar spatial filter. We compute *one* RMS value for each phone: We found this approach to give better results than using a fixed frame length and shift, or using the $\mathbf{TD5}$ features which we use for training the recognizer (see section 4.3). All computations are done channelwise, where we omit interpolated EMG channels affected by channels marked as noisy (see section 3).

We similarly compute RMS features of the 20 adaptation sentences of the source session and note that their textual content matches the content of the adaptation sentences of the target session: Thus we have a pairwise matching of the adaptation sentences from both sessions. Two sentences with the same textual content yield the same number of RMS feature frames, since they have the same number of phones.

Now we compute the *correlation* between the RMS features of the source session and the interpolated target sessions and average over all channels, except those marked as containing artifacts, and over the 20 adaptation sentences. This average correlation is a scalar value measuring the similarity between the source session data and the interpolated target session data, for a particular shift and rotation. We finally assume that the optimal shift and rotation between source and target session is obtained by maximizing the average correlation over all possible shifts and rotations. A preliminary study (see [16]) indicates that this is frequently the case.

Now for adaptation and testing of the EMG-based speech recognizer, all target session data is transformed by interpolation with the optimal shift and rotation, as determined by the above algorithm.

4.3. The EMG-based Speech Recognizer

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

Features for our recognition system are always computed from *bipolar* EMG signals. When position shift compensation is used, *all* unipolar target session data is first transformed by interpolation as described above, then we compute bipolar signals by applying a spatial filter.

Our feature set is taken from [2, 15]: For any given feature $\mathbf{f}$, $\bar{\mathbf{f}}$ is its frame-based time-domain mean, $\mathbf{P_f}$ is its frame-based power, and $\mathbf{z_f}$ 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.

For an EMG signal with normalized mean $x[n]$, the nine-point double-averaged signal $w[n]$ is defined as

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

The high-frequency signal is $p[n] = x[n] - w[n]$, and the rectified high-frequency signal is $r[n] = |p[n]|$. The final feature $\mathbf{TD}5$ is defined as follows:

$$\mathbf{TD}5 = S(\mathrm{TD}0, 5), \text{where } \mathrm{TD}0 = [\bar{\mathbf{w}}, \mathbf{P_w}, \mathbf{P_r}, \mathbf{z_p}, \bar{\mathbf{r}}],$$

i.e. we use a total of 11 context frames. PCA+LDA is used for dimensionality reduction [2], after PCA, 700 dimensions are retained, which are further compressed by LDA to a 12-dimensional feature vector. We note that this gives us better results than the 32 dimensions which we used in [2].

The **recognizer setup** follows a standard pattern: We use three-state left-to-right fully continuous Hidden-Markov-Models, where the emission probabilities are modeled with bundled phonetic features (BDPFs) [1]. The recognizer is trained using the training data of the source session, yielding a "background" *myoelectric model* which comprises all information which we can obtain from the source session.

Adaptation is performed by *Maximum Likelihood Linear Regression* (MLLR) [17]. The 20 adaptation sentences of the target session are used for estimating the MLLR transformation[1], which modifies the myoelectric model to better match the adaptation data.

For **testing** on the interpolated test data of the target session, we use the (adapted) myoelectric model together with a trigram Broadcast News language model. The decoding vocabulary is restricted to the words appearing in the test set, which results in a test vocabulary of 108 words incl. variants. The test set perplexity is 24.24. For details see [2].

5. EXPERIMENTS AND RESULTS

We run the following five experiments:

- **Direct application:** We train a recognizer on the source session and use it to decode the target session, without any adaptation at all, i.e. neither MLLR nor position shift compensation is used.
- **Shift compensation:** We interpolate the target session data as described in sections 4.
- **MLLR:** We use MLLR to adapt the myoelectric model from the source session towards the target session.
- **MLLR + Shift compensation:** We first interpolate the target session data, then we apply MLLR to the source myoelectric model, using the *interpolated* adaptation sentences of the target session for MLLR estimation.
- **Session dependent:** For comparison, we train a session-dependent system on the training data of the target session. This is expected to yield the best results, however it requires that 160 training sentences from the target session are available, which is not assumed for our cross-session systems. In this case, MLLR and position shift compensation are neither required nor useful.

Figure 2 depicts the results of these experiments on the development corpus, broken down by speakers. Our measure is the *Word Error Rate* (WER) on the test data sets of the target sessions, which we intend to minimize. The WERs are averaged over all 36 possible pairs of source and target session.

First of all, we observe that direct application of a recognizer trained on the source session towards a different target session does not work at all: The average WER always exceeds 90%. We also observe that MLLR adaptation, even with the small amount of adaptation data, helps a great deal: The average WER across sessions is reduced to 50.4%.

[1]It is remarkable that 20 sentences are already enough to obtain good results here, indeed we reported in [3] that at least 30 sentences are required: Side experiments show that this is due to the LDA dimensionality reduction. If substantially more than 12 LDA components are retained, more adaptation sentences are needed.

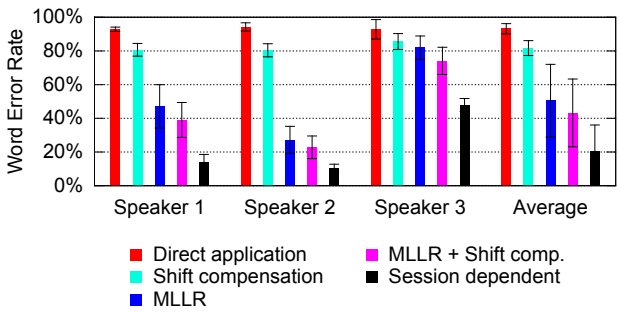

Fig. 2. Breakdown of results by speaker for the development corpus. Bars indicate standard deviation.

Our shift compensation algorithm substantially improves this result, yielding a final average WER of the cross-session system of 43.2%: An improvement over sole application of MLLR of 14.3% relative. This result is consistent across speakers. Using position shift compensation *without* an adaptation step also improves the WER of the cross-session system, namely from the original 93.2 % to 81.7%. This is still way above the WER obtained by MLLR alone, from which we conclude that the discrepancy between sessions is not due to the position shift alone. Session-dependent systems perform best, with an average WER of only 20.7%.

We evaluate our system on the held-back evaluation data from speaker 4. The results are as follows:

Experiment	Average WER
Direct application	92.69 %
Shift comp. without MLLR	82.19 %
MLLR	52.66 %
MLLR + Shift comp.	45.88 %
Session dependent	15.63 %

We see the same trend as on the development corpus (figure 2). The average *absolute* improvement between the MLLR and MLLR + Shift compensation setups is 6.8%, with a 95% confidence interval width of 1.6%; proving statistical validity of the improvement.

6. CONCLUSION AND FUTURE WORK

In this study we developed an algorithm to compensate for variation of the EMG recording position for our EMG-based speech recognizer. We showed that our algorithm significantly improves the recognition accuracy over established session adaptation, which underlines the advantages of using arrays for EMG signal acquisition. We expect that these results can be further improved by training a recognizer on multiple sessions, as we already proved for our single-electrode setup [3], paving the way towards robust and easy-to-use EMG-based speech recognition in the field.

7. REFERENCES

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

[2] Michael Wand, Christopher Schulte, Matthias Janke, and Tanja Schultz, "Array-based Electromyographic Silent Speech Interface," in *Proc. Biosignals*, 2013, pp. 89 – 96.

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

[4] Robin Hofe, Stephen R. Ell, Michael J. Fagan, James M. Gilbert, Phil D. Green, Roger K. Moore, and Sergey I. Rybchenko, "Small-Vocabulary Speech Recognition using a Silent Speech Interface based on Magnetic Sensing," *Speech Communication*, vol. 55, pp. 22 – 32, 2013.

[5] Thomas Hueber, Gérard Bailly, and Bruce Denby, "Continuous Articulatory-to-Acoustic Mapping using Phone-based Trajectory HMM for a Silent Speech Interface," in *Proc. Interspeech*, 2012.

[6] Viet-Anh Tran, Gérard Bailly, Hélène Loevenbruck, and Tomoki Toda, "Improvement to a NAM-captured Whisper-to-Speech System," *Speech Communication*, vol. 52, pp. 314 – 326, 2010.

[7] Charles Jorgensen and Sorin Dusan, "Speech Interfaces based upon Surface Electromyography," *Speech Communication*, vol. 52, pp. 354 – 366, 2010.

[8] Yunbin Deng, Rupal Patel, James T. Heaton, Glen Colby, L. Donald Gilmore, Joao Cabrera, Serge H. Roy, Carlo J. De Luca, and Geoffrey S. Meltzner, "Disordered Speech Recognition Using Acoustic and sEMG Signals," in *Proc. Interspeech*, 2009, pp. 644 – 647.

[9] Joao Freitas, Antonio Teixeira, and Miguel Sales Dias, "Towards a Silent Speech Interface for Portuguese," in *Proc. Biosignals*, 2012, pp. 91 – 100.

[10] Michael Wand, Adam Himmelsbach, Till Heistermann, Matthias Janke, and Tanja Schultz, "Artifact Removal Algorithm for an EMG-based Silent Speech Interface," in *Proc. EMBC*, 2013, pp. 5750 – 5753.

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

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

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

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

[15] 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, pp. 573 – 576.

[16] Christopher Schulte, "Kompensation unterschiedlicher Elektrodenpositionierungen in der EMG-basierten Sprachverarbeitung," M.S. thesis, Karlsruhe Institute of Technology, 2013.

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