35th Annual International Conference of the IEEE EMBS
Osaka, Japan, 3 - 7 July, 2013

Artifact Removal Algorithm for an EMG-based Silent Speech Interface

Michael Wand, Adam Himmelsbach, Till Heistermann, Matthias Janke, and Tanja Schultz

***Abstract*— An electromygraphic (EMG) *Silent Speech Interface* is a system which recognizes speech by capturing the electric potentials of the human articulatory muscles, thus enabling the user to communicate silently. This study deals with improving the EMG signal quality by removing artifacts: The EMG signals are captured by *electrode arrays* with multiple measuring points. On the resulting high-dimensional signal, Independent Component Analysis is performed, and artifact components are automatically detected and removed. This method reduces the Word Error Rate of the silent speech recognizer by 9.9% relative on a development corpus, and by 13.9% relative on an evaluation corpus.**

I. INTRODUCTION

Speech is the most natural way of interaction between humans and has also become a means of wide-range communication and machine interaction due to the advent of telephone technology and speech-based electronic devices. However, classical voice-based communication requires speech to be clearly audible, which incurs lack of robustness in noisy environments, disturbance for bystanders, compromised privacy, and exclusion of speech-disabled people.

These challenges are tackled by Silent Speech Interfaces, which are systems enabling speech communication to take place without the necessity of emitting an audible acoustic signal, or when an acoustic signal is unavailable [1]. 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. This allows speech to be recognized even when it is produced silently, i. e. mouthed without any vocal effort [2].

Our current system uses *electrode arrays* for the recording of facial electromyographic signals. In [3] we present first results using this new system and introduce *Independent Component Analysis (ICA)* as a means to improve the recognition accuracy. This paper extends those results by an algorithm which automatically detects and removes artifact components in the ICA decomposition of the input EMG signal. Artifacts may arise from technical sources (e.g. power line noise) or biological sources (e.g. movement, ECG) and are known to be a major source of recognition errors. We additionally investigate the spatial distribution of the ICA components, thus giving evidence that ICA actually extracts localized sources of EMG activity.

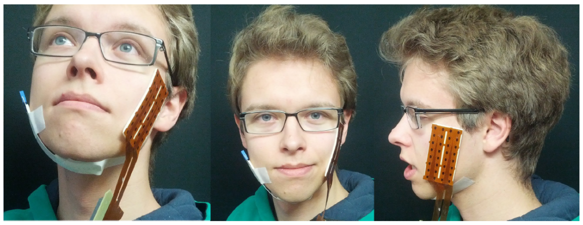

Fig. 1. EMG array positioning

II. RELATED WORK

Using EMG for speech recognition dates back to the 1980s, however competitive performance was first reported by [4], who achieved an average word accuracy of 93% on a 10-word vocabulary of English digits. [5] reported good performance even for silently spoken words.

The first EMG-based speech recognizers were usually whole-word recognizers with a relatively small vocabulary, ranging around 10 words (e.g. [4], [5], [6]). Current work still occasionally focuses on whole-word classification tasks, e.g. for specific problems like nasality [7]. In 2006, initial results were found on the usage of smaller modeling units, in particular context-independent phonemes [8] and phonetic features [9], which represent properties of phonemes like place or manner of articulation. Phonetic feature recognition gave rise to our *Phonetic Feature Bundling* algorithm, which reduces the Word Error Rate of the EMG-based speech recognizer by more than 33% relative [2]. During the past few years, we have been using the Bundled Phonetic Feature recognizer as a basis to tackle issues such as session independency [10] or discrepancies between audibly spoken versus silently mouthed speech [11].

All recognizers described so far use rather elementary preprocessing methods, e.g. time-domain features [7], [8] or frequency- or wavelet-based features [5], [6]. While time-domain features appear to be superior to frequency-based features [8], they are still incapable of separating signals from different EMG activity sources, or EMG signals and artifact noise. Our array-based system strives to enable versatile methods such as ICA or beamforming prior to normal feature extraction in order to achieve an improved signal preprocessing; in particular, this paper deals with ICA-based artifact removal.

III. DATA ACQUISITION AND CORPUS

For EMG recording we use the multi-channel EMG amplifier *EMG-USB2* produced and distributed by *OT Bioelettronica*, Italy (`http://www.otbioelettronica.it/`),

The authors are with Cognitive Systems Lab, Karlsruhe Institute of Technology, Karlsruhe, Germany. Corresponding author: `michael.wand at kit.edu`

This research project has been funded by the DFG Research Grant "MAPS - Myoelectric Array-based Processing of Speech".

978-1-4577-0216-7/13/$26.00 ©2013 IEEE

together with a set of electrode arrays also acquired from *OT Bioelettronica*.

The EMG array configuration for our experiments is shown in figure 1. We use two arrays: A chin array with a row of 8 electrodes with 5 mm inter-electrode distance (IED), and a cheek array with 4×8 electrodes with 10 mm IED. In order to minimize common-mode artifacts, we chose a *bipolar* measurement configuration, where the potential difference between two adjacent channels in a row is measured. This means that out of 4×8 cheek electrodes and 8 chin electrodes, we obtain $(4 + 1) \cdot 7 = 35$ signal channels. EMG signals are sampled at 2048 Hz. The audio signal is parallely recorded with a close-talking microphone.

The recording protocol follows [2]. We use 15 sessions by 6 speakers, where each session consists of 50 English sentences: a set of 10 “BASE” sentences, which is kept fixed across sessions and used for testing, and a set of 40 training sentences which varies across sessions. The sentences are read in normal, audible speech. Note that our array corpus also contains recordings of silently mouthed speech, which were not used in this study. The setup which is used in this paper corresponds to the “B-1” setup from [3], however we have extended the corpus: The 7 sessions which we used in [3] form a development set which we used to optimize the parameters of our artifact removal algorithm. From that corpus we extracted 8 further sessions, which were set aside to be used as an evaluation set. The speakers overlap between the development corpus and the evaluation corpus. In this study we only perform session-dependent experiments, i.e. training and testing is performed separately for each session. The following table summarizes our corpus.

Corpus	# of Speakers /	Average data length in sec.		
	Sessions	Training	Test	Total
Development	6 / 7	149	42	191
Evaluation	5 / 8	135	40	175

IV. BASELINE SYSTEM AND RESULTS

A. Feature Extraction, Training, and Decoding

For each EMG channel, we perform framing and compute five *time-domain features*, see [8] for details. ICA is applied on the 35 *raw* EMG signal channels *before* feature extraction, see section IV-B. The final feature is created by performing a stacking of adjacent feature vectors with context width 5 according to the optimal result from [3], thus yielding up to 35 · 5 · (5+1+5) = 1925 features if the full EMG channel set is used. After this step, we compute a Principal Component Analysis (PCA) on the resulting extended feature vectors, reducing their dimensionality to 700. This step is followed by Linear Discriminant Analysis (LDA) to obtain a final feature vector with 32 coefficients. In [3] we show that the PCA step is necessary in order to obtain robust results; otherwise, the LDA computation would be inaccurate.

The recognizer is based on three-state left-to-right fully continuous Hidden-Markov-Models. All experiments use bundled phonetic features (BDPFs) for training and decoding [2]. For decoding, we use the trained acoustic model together with a trigram Broadcast News language model. The test set perplexity is 24.24. The decoding vocabulary is restricted to the words appearing in the test set, which results in a test vocabulary of 108 words incl. variants. Note that we do *not* use lattice rescoring. For more details see [3].

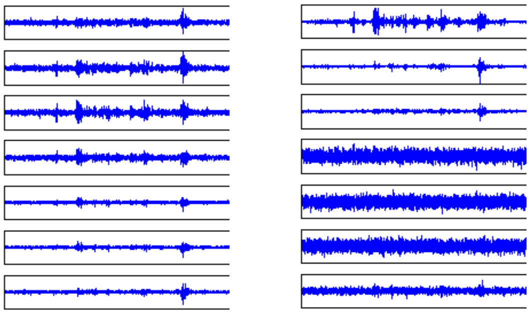

Fig. 2. EMG Signals of the chin array before ICA processing (left) and after ICA processing (right). The ICA decomposition shows visibly distinct EMG signal components and artifact noise.

B. ICA Application

The ICA algorithm computes a linear transformation on the input signal which maximizes the statistical independence between the estimated components. We use the Infomax ICA algorithm according to [12], as implemented in the Matlab EEGLAB toolbox [13], to compute a session-dependent ICA decomposition matrix from the 40 utterances of the training set of the respective session. In our baseline system we interpret ICA as a method of (blind) source separation [3], therefore we apply ICA and then perform feature extraction *on the estimated independent components*; this includes the PCA step and the LDA step. In this study we refine this method, see section V.

C. Baseline Results

We compute the Word Error Rate (WER) with and without ICA application for our baseline system. The following table shows the results: in all cases, ICA application reduces the WER, although the improvement is not statistically significant.

Corpus	WER without ICA	WER with ICA
Development	46.3%	45.3%
Evaluation	58.5%	54.7%

V. ICA-BASED ARTIFACT REMOVAL

A. Direct Approach and Backprojection

In the baseline system, we applied ICA without considering the nature of the estimated components. In particular, we fed both artifact components and EMG-like “target” components into our feature extraction and thus into the recognition system. From figure 2, we see that artifact components are typically vastly different from target components: the 7 original EMG channels of the chin array of one utterance (left) are decomposed into three “target” components which look like EMG signals, and four “noise” components. Therefore we expect that the removal of the noise channels improves the recognition results. We pursue two strategies:

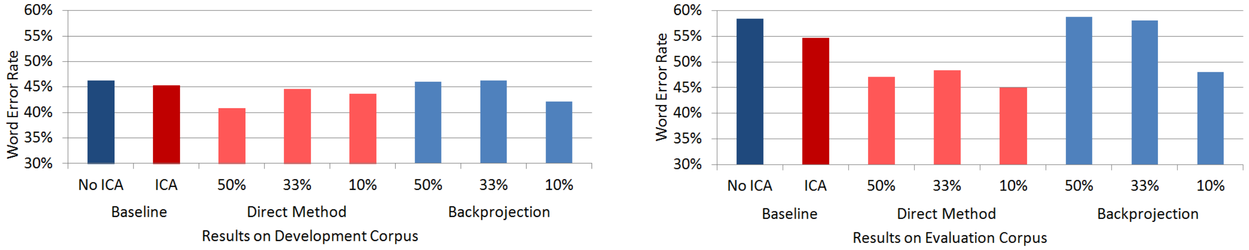

Fig. 3. Results for different ICA component removal strategies. See text for details.

- **Direct method:** We take the ICA components, identify and remove artifact components, and then compute the EMG features on the *remaining components*.
- **Backprojection:** We take the ICA components, identify and remove artifact components as before, and then back-project these components to the original signal. Mathematically, this can be described as applying the ICA decomposition, setting the artifact ICA components to zero, and then multiplying the altered set of ICA components with the *inverse* of the ICA matrix.

The ICA application in our baseline system [3] is a "direct method" approach without component removal.

B. Artifact identification

For each session, artifact components are identified by the following three measures, which are computed on the ICA decomposition:

- Autocorrelation measure: This method typically identifies very regular (periodic) artifacts, like power line noise. We compute the autocorrelation sequence of the input component and then take the value of the *first maximum* after the first zero of the sequence. If this value is greater than 0.5, this component is deemed an artifact.
- High-frequency noise detection: The surface EMG signal has frequency range of 0Hz - 500Hz [14]. Therefore a component with distinct high-frequency parts is considered an artifact. We compute the discrete-time Fourier transform of the input component and divide the frequency axis into two intervals: The "signal" interval from 0Hz to 500Hz, and the "noise" interval from 500Hz to 1024Hz (the Nyquist frequency). We then compute the areas of the amplitude of the Fourier transform over the two intervals and divide the "signal" area by the "noise" area. If the quotient is smaller than 1.3, this component is deemed an artifact.
- EMG signal range: The main energy of the EMG signal is found between 50Hz and 150Hz [14]. As before, we divide the frequency axis into two parts: A "signal" interval from 50Hz to 150Hz, and "noise" part from 0Hz to 50Hz and from 150Hz to 1024Hz. Then we divide the "signal" area by the "noise" area. If the quotient is below 0.25, we deem this component an artifact. For this measure, we found that the power spectral density yielded slightly more robust estimates than a standard Fourier transformation.

Note that the thresholds were determined on the development set only. Our measures are first computed on each ICA component of *each utterance* of the training data set. In a second step, we combine the results: For a component to be considered an artifact, we require that *at least one* of the three methods considers this component an artifact on *a minimum percentage* of (training) utterances. This "threshold percentage" was varied between 10% and 50%, where a lower value causes more components to be removed. We observed that the threshold makes a difference when components vary across utterances, e.g. when the contact between electrode and skin deteriorates over time.

C. Results

The left part of figure 3 shows the Word Error Rates on the development corpus for different threshold percentages and for the two channel removal strategies. The direct method works better than backprojection. The best result is achieved with a 50% threshold percentage: Compared with the baseline ICA system without channel removal, the average WER improves from 45.3% to 40.8%, which is a relative improvement of 9.9%. For the backprojection method we obtain higher WERs, the best result with backprojection is 42.1% WER. Note that the intra-session variance between these results is rather high, so the results are only accurate within a confidence interval of around ±10%.

The difference between the direct method and backprojection is that in the former case, we compute features on a subset of the *ICA components*, whereas in the latter case we compute features on altered *original EMG signals*. Since in [3] we observed that computing features on ICA components yields better results than using the original EMG signals even without any kind of artifact removal, the superiority of the direct method had to be expected. We assume that besides isolating artifacts, ICA extracts EMG sources which are superimposed in the measured EMG signal; in section VI we present evidence supporting this claim.

Finally, we applied our methods to the evaluation corpus. The right part of figure 3 shows the results. The observations are similar to the development corpus: applying the direct method of artifact removal with a 50% threshold percentage reduces the WER from 54.7% to 47.1%, i.e. by 7.6% absolute or 13.9% relative. Using a threshold percentage of 10% is even a bit better (45.0% WER).

In order to statistically establish the validity of our procedure, we consider the WER reduction between the best baseline system (with ICA, no artifact removal) and the optimal system as determined on the *development* corpus (ICA, direct method, 50% threshold). Figure 4 shows the WER reduction for all session of the *evaluation* corpus: The average absolute WER reduction is 7.6%, with a 95% confidence interval ranging from 1.3% to 13.9%. So we can assert that our WER improvement is significantly greater than 0%. One outlier session is observed.

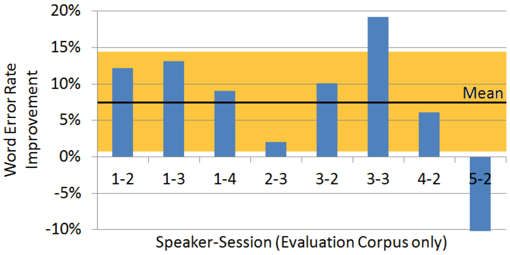

Fig. 4. Word Error Rate improvement on evaluation corpus, between baseline system with ICA and "direct method" artifact removal with 50% threshold percentage. The shaded area indicates the 95% confidence interval.

VI. EMG Spatial Patterns

So far we have considered properties of the artifact components which resulted from our ICA decomposition of the EMG signal. We have shown that removing these artifact components before feature extraction leads to improved results, in particular *when the features are computed on the remaining ICA components*. Why it is advantageous to compute EMG features on the ICA components, instead of the backprojected original EMG signals? It has to be assumed that the EMG-like components which result from the ICA decomposition capture properties of the articulatory apparatus better than the original EMG signals.

We propose that ICA extracts localized EMG activity sources, thus yielding a division of the input EMG signal into activities related to different muscles or motor units. In order to support this claim, one may consider the spatial distribution of these ICA components, as follows:

Assume that in the EMG signal, the horizontal axis represents time. Then left-multiplication with the ICA matrix yields the ICA decomposition of the signal, and left-multiplication of the ICA decomposition with the *inverse* of the ICA matrix yields, of course, the original EMG signals.

The *columns* of the inverse ICA matrix are called *spatial patterns* [15]. The j-th spatial pattern characterizes the spatial distribution of the j-th ICA component, i.e. it indicates how much an ICA component is observable at the measuring points of the EMG array.

Figure 5 shows exemplary ICA patterns for the 28-channel signal measured from the *cheek array*, where we manually labeled the ICA components as artifacts or EMG-like. The visualization shows the 28-element column of the inverse ICA matrix, reshaped to the form of the EMG array. The right side shows three typical signal components, each exhibits a visible center where the respective ICA component is strongest, and declining strength away from this center. The left side shows three typical artifact components: either the spatial pattern of that component appears random (bottom left), or the pattern is overly strong for very few single components (top/middle left). We have observed such patterns for a large number of ICA components and take these results as a confirmation that ICA attains EMG source localization. Our future work will include making use of EMG spatial patterns to further improve the EMG feature extraction and the modeling of phonetic features.

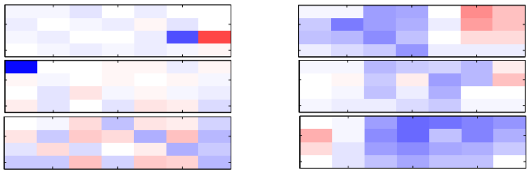

Fig. 5. Typical ICA patters for artifact components (left) and EMG signal components (right), on the 28-channel cheeck array. EMG signal components exhibit a visible peak and declining activity away from the center. Noise components show other patterns.

References

[1] B. Denby, T. Schultz, K. Honda, T. Hueber, and J. Gilbert, "Silent Speech Interfaces," *Speech Communication*, vol. 52, no. 4, pp. 270 – 287, 2010.

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

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

[4] A. Chan, K. Englehart, B. Hudgins, and D. Lovely, "Myoelectric Signals to Augment Speech Recognition," *Medical and Biological Engineering and Computing*, vol. 39, pp. 500 – 506, 2001.

[5] C. Jorgensen, D. Lee, and S. Agabon, "Sub Auditory Speech Recognition Based on EMG/EPG Signals," in *Proceedings of International Joint Conference on Neural Networks (IJCNN)*, Portland, Oregon, 2003, pp. 3128 – 3133.

[6] L. Maier-Hein, F. Metze, T. Schultz, and A. Waibel, "Session Independent Non-Audible Speech Recognition Using Surface Electromyography," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, San Juan, Puerto Rico, 2005.

[7] J. Freitas, A. Teixeira, and M. S. Dias, "Towards a Silent Speech Interface for Portuguese," in *Proc. Biosignals*, 2012.

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

[9] S.-C. S. Jou, T. Schultz, and A. Waibel, "Continuous Electromyographic Speech Recognition with a Multi-Stream Decoding Architecture," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2007.

[10] M. Wand and T. Schultz, "Session-independent EMG-based Speech Recognition," in *Proc. Biosignals*, 2011.

[11] M. Wand, M. Janke, and T. Schultz, "Decision-Tree based Analysis of Speaking Mode Discrepancies in EMG-based Speech Recognition," in *Proc. Biosignals*, 2012.

[12] A. J. Bell and T. I. Sejnowski, "An Information-Maximization Approach to Blind Separation and Blind Deconvolution," *Neural Computation*, vol. 7, pp. 1129 – 1159, 1995.

[13] A. Delorme and S. Makeig, "EEGLAB: An Open Source Toolbox for Analysis of Single-Trial EEG Dynamics including Independent Component Analysis," *Journal of Neuroscience Methods*, vol. 134, no. 1, pp. 9–21, 2004.

[14] H. Zhao and G. Xu, "The Research on Surface Electromyography Signal Effective Feature Extraction," in *Proc. of the 6th International Forum on Strategic Technology*, 2011.

[15] B. Blankertz, R. Tomioka, S. Lemm, M. Kawanabe, and K.-R. Müller, "Optimizing Spatial Filters for Robust EEG Single-Trial Analysis," *IEEE Signal Processing Magazine*, vol. 25, pp. 41 – 56, 2008.