Pattern Learning with Deep Neural Networks in EMG-based Speech Recognition

Michael Wand and Tanja Schultz

***Abstract*— We report on classification of phones and phonetic features from facial electromyographic (EMG) data, within the context of our EMG-based *Silent Speech interface*. In this paper we show that a *Deep Neural Network* can be used to perform this classification task, yielding a significant improvement over conventional Gaussian Mixture models. Our central contribution is the *visualization* of patterns which are learned by the neural network. With increasing network depth, these patterns represent more and more intricate electromyographic activity.**

I. INTRODUCTION

During the past decade, novel speech processing devices called *Silent Speech Interfaces* (SSI) [1] have been gaining more and more popularity. SSIs enable speech communication between humans and speech-based man-machine interaction even when an acoustic speech signal is not available. Application areas include confidential and undisturbing communication in public places, as well as the creation of assistive devices for speech-impaired persons.

This study is based on our Silent Speech interface using *surface electromyography* (EMG) [2], where the electrical potentials which emerge in the articulatory muscles during speaking are captured by surface electrodes. The system has been under development since 2005 and by now allows Hidden Markov Model-based recognition of continuous speech [3] with vocabularies of up to 2100 words [4], for both audibly spoken and silently mouthed speech [5].

Despite these successes, there exists only limited knowledge about what exactly the recognition system learns from its training data. One way to tackle this question from the classifier perspective is to forego the rather complex Hidden Markov Model framework which is used for continuous speech recognition and analyze recognition performance at the level of *single frames*. We follow our prior study [6] in assuming that each frame is uniquely assigned to a phone or *phonetic feature* (a phonetic property of a phone, see section III-A), and we evaluate the performance and characteristics of a classifier which learns this assignment.

The contribution of this study is the application of *deep neural networks* (DNN) for this frame-based classification task. We show that for phone classification, we achieve significant improvements over a baseline system using Gaussian Mixture Models (GMM), and we show in particular that we can *visually interpret* the activation patterns of the hidden nodes of the DNN in the input feature domain, leading to a better understanding about what constitutes discernible activity in the complex facial EMG activity generated during speaking.

Fig. 1. EMG array positioning

Our study is rooted in ongoing research on Silent Speech interfaces, see [1] for an overview. The EMG approach is rather well-developed and is being used to investigate a variety of specific challenges mostly related to applied Silent Speech recognition, for example recognition of disordered speech [7], language-dependent challenges like nasality [8] or tonality [9], as well as direct conversion from EMG signals to acoustics [10], with the purpose of generating natural intonation contours [11]. This study, with its more theoretical orientation, well complements these research topics.

II. DATA CORPUS

Our corpus consists of 25 sessions from 20 speakers, each comprising 200 read English-language utterances spoken in normal, audible speech. 11 sessions from 9 speakers were used as a development set for parameter tuning, the 14 sessions from the remaining 11 speakers were set aside for evaluation. As in [6], silent speech was not used since in this case obtaining phone-level alignments is problematic [5].

Data was taken from the current version of our *EMG-Array* corpus [2]. The multi-channel EMG amplifier *EMG-USB2* (OT Bioelettronica, Italy) was used together with two EMG arrays, see figure 1: A chin array comprising a single row of 8 electrodes with 5 mm inter-electrode distance (IED), and a cheek array with 4×8 electrodes with 10 mm IED. This data was recorded in bipolar fashion, where the difference of two adjacent channels is taken to reduce common mode artifacts, thus we finally got 35 ($5 \cdot 7$) EMG channels. Sampling was performed at 2048Hz.

Michael Wand is with The Swiss AI Lab IDSIA – Istituto Dalle Molle di Studi sull'Intelligenza Artificiale, University of Lugano & SUPSI, Manno-Lugano, Switzerland. Tanja Schultz is with Cognitive Systems Lab, Karlsruhe Institute of Technology, Karlsruhe, Germany. Work for this paper was performed while Michael Wand was at Cognitive Systems Lab. Contact: {michael.wand|tanja.schultz}@kit.edu.

This research project was funded by the German Research Foundation (DFG), Research Grant *MAPS - Myoelectric Array-based Processing of Speech*. Work was performed on the computational resource bwUniCluster funded by the Ministry of Science, Research and Arts and the Universities of the State of Baden-Württemberg, Germany, within the framework program *bwHPC*.

978-1-4244-7929-0/14/$26.00 ©2014 IEEE

TABLE I
DATA CORPUS

	Development Corpus		Evaluation Corpus	
	Avg session length	# sessions	Avg session length	# sessions
Training	591 sec.		519 sec.	
Cross-Val.	85 sec.	11	76 sec.	14
Test	77 sec.		68 sec.	
Total amount of data (development): 138 minutes				
Total amount of data (evaluation): 155 minutes				

Acoustic data was simultaneously recorded with a standard close-talking microphone and synchronized to the EMG data with a hardware marker signal. According to [12], the EMG signal was delayed by 50ms to better match the audio signal. Phone-level alignments of the EMG signal, a prerequisite for performing our study, were then computed from the synchronized acoustic signal, as in [12]. We always consider these acoustic time alignments as ground truth.

Both development and evaluation sessions were subdivided as follows: 160 utterances were used as training set, 20 utterances were used for cross-validation (CV), and the remaining 20 utterances were used for testing the classifiers. Table I gives an overview of our data corpus.

III. EXPERIMENTAL SETUP

A. Classification of Phones and Phonetic Features

We evaluate seven different frame-based classifiers:

- First, we consider classification of *phones*. We partition our data into 45 phone classes; this is our standard for continuous speech recognition. Note that the amount of samples per phone varies greatly (for example, 6 versus 600 samples).
- Second, we perform classification of phonetic features (PF). The PFs were chosen along the main axes determining an (English) phone:
  - For consonants: *Position* (8 classes) and *Manner* (5 classes) of articulation, and *Voicing* (2 classes)
  - For vowels: *Frontness* (4 classes), *Openness* (4 classes), and (lip) *Rounding* (2 classes). For frontness and openness, diphthongs were placed in a separate class.

  For the classification of consonant PFs, all vowel frames are disregarded in both training and testing, and analogously, for vowel PFs consonant frames are ignored.

In all cases, silence which occurs at the beginning or end of utterances is ignored. The classifiers are always trained and tested on data from the same recording session. We measure the *balanced accuracy* or our recognizers: For each class, we separately compute the percentage of correctly recognized frames, then we average over all class-wise accuracies. Thus we compensate for the different number of frames per class, which implies that if there are n classes to be discriminated, the chance level of recognition is $1/n$.

B. GMM baseline system

The baseline for this study is the classifier presented in [6], which is based on Gaussian mixture models (GMM). However we adapted the EMG feature extraction to account for the purpose of this study, i.e. visualization and interpretation of learned EMG patterns[1]. We chose the *logarithmic power* as an easily interpretable feature: The EMG signal is subdivided into frames with a size of 27ms and a shift of 10ms, and for each frame, we compute the power and take the logarithm. We combine the log-power features for all 35 channels and finally stack adjacent frames with a context width of 5 (-5 ... 5), so that we get a feature size of $11 \cdot 35 = 385$. The features are then z-normalized.

Finally we compute a Linear Discriminant Analysis (LDA) transformation for dimensionality reduction. Since in our previous work, we observed that the optimal number of dimensions after LDA varies with the recording session and with the task at hand, we considered several parameter settings for the LDA transformation, namely, we experimented with 12, 22, or 32 retained LDA components, with an optional PCA step before LDA to reduce sparsity [2], retaining 900 dimensions after PCA. From these 6 settings, for each task we chose the *best-performing one* as our baseline, based on the result on the development set.

C. Training Deep Neural Networks

The key method which we apply in this study is training a Deep Neural Network (DNN), i.e. a network with relatively many hidden layers, on the input EMG features. We use the stacked 385-component log-power feature defined in section III-B as network input; since it is not needed here, we do *not* apply LDA. Then we use three logistic hidden layers, with varying sizes experimentally tuned on the development corpus, and a final softmax classification layer having as many nodes as there were classes to be distinguished.

For training we follow the method of Hinton et al. [13]: We first consider only the logistic hidden layers, on which we perform unsupervised greedy layer-wise pretraining with the *Restricted Boltzmann machine* (RBM) algorithm. We assume the input data to be Gaussian distributed and modify the RBM algorithm to get a *Gaussian-Bernoulli* RBM [14]. Our code is based on the original scripts by Hinton [15].

After this pretraining, we add the discriminative softmax layer and perform standard backpropagation training on the resulting pre-initialized network. Parameters for this step include a minibatch size of 1000 frames, a maximum of 300 epochs, and a linearly decaying learning rate ($10^{-4} \ldots 10^{-5}$). After each epoch, we compute the balanced accuracy on the cross-validation data, the network weights which are used for classifying the test data are chosen to maximize the balanced accuracy on the cross-validation set.

IV. CLASSIFICATION RESULTS

Figure 2 charts the recognition accuracies for our different experiments, averaged over sessions of the development or

[1]Also note that the PF classes in [6] were defined slightly differently, so that those results are not directly comparable to the ones obtained here.

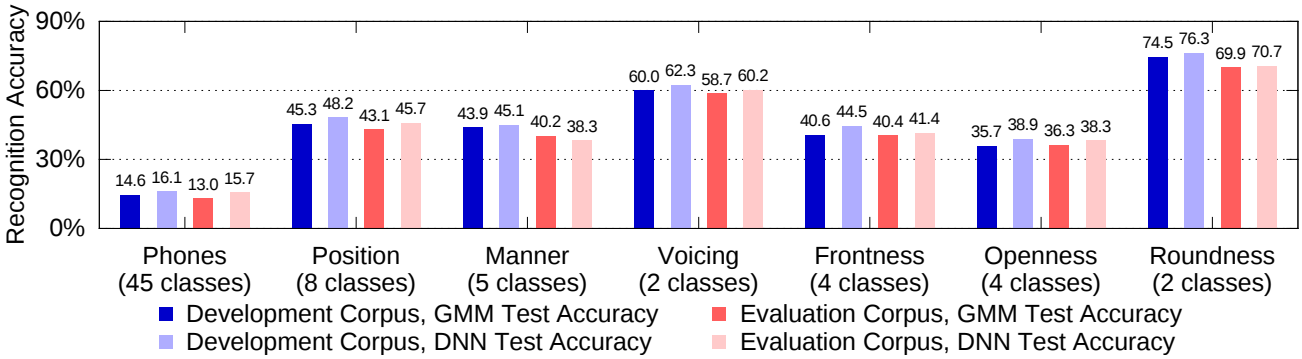

Fig. 2. Average accuracies of the different systems on the test utterances of the development and evaluation corpus, in percent.

evaluation corpus, respectively. For better readability we only display results on the *test* sentences. The results on the CV set are not fundamentally different.

The key observation is that the DNN training yields accuracy improvements in almost all cases, only for the classification of the manner of articulation, GMMs perform better on the evaluation corpus. The improvement is largest for phones (more than 20% relative on the evaluation corpus), it is also highly significant (one-sided t-test for paired samples, $p = 2.8 \cdot 10^{-6}$). The relative improvements for phonetic features are lower (however the baseline accuracy is higher), and the t-test frequently does *not* prove significance.

With the exception of lip rounding, the vowel features achieve relatively low accuracies (typically around 40%, at 25% chance level), which might show that these features, which only differ slightly in articulatory movements, are generally hard to recognize from EMG data. Yet, English vowel pronunciation allows a lot of variation (particularly with both native and non-native English speakers in our corpus), so that only limited accuracy should be expected *in principle*. The manner of articulation is harder to classify than the position of articulation, no matter which method is used; this confirms results from [6].

This result supports using DNNs as a method for phone/PF classification from EMG data (suggesting subsequent usage for continuous EMG-based speech recognition as well). For the purpose of this paper, we can expect our DNN pattern visualization to yield reasonable results.

V. VISUALIZATION OF HIDDEN NODES

Assume that we have finished training a multi-layer neural network. Then we can use this network for classification by "forward propagating" a test pattern x through the network, computing node activations on the way. Clearly, for any node N, its activation is a function of x: $N = N(x)$. The weights do no more appear as parameters, since they are now fixed.

Since all functions used in the neural network are differentiable, so is $N(x)$. For understanding the behavior of this node, it is therefore possible to maximize $N(x)$ over x:

$$\hat{x} = \arg\max_x N(x). \quad (1)$$

This method, proposed by Erhan et al. [16], yields the *maximum activation* pattern $\hat{x}$ for the given node N. Here it is necessary to place a norm constraint on x, we thus require $||x||_2 = 1$. The constrained optimization in equation 1 is performed with the Matlab `fmincon` function.

For display purposes, we chose a small DNN trained for phone classification, with hidden layers of sizes 30, 20, and 30. Figure 3 displays several typical resulting patterns for randomly chosen nodes from the three hidden layers. Each *row* contains one single maximum activation pattern, arranged to show the shapes of the two EMG arrays (1×7 and 4×7). The 11 plots per row stem from the 11 ($2 \cdot 5 + 1$) *context* frames which form the input feature vector (see section III-B), thus each row contains the *time evolution* of a maximum activation pattern. Also note that since we use a log-power feature, activation patterns directly correspond to EMG energy levels. We show sample patterns for session 2 of speaker 2 (development corpus), these patterns repeat over and over for a variety of speakers and sessions.

Subfigure 1), on the left-hand side, contains activation patterns for the first hidden layer, which directly connects to the input EMG feature. We see that some nodes (in particular, the ones in rows a to c) exhibit maximum activation patterns which extract *localized sources* of EMG activity (in the case of row a, the strongest activation is in the small chin array). These patterns strongly resemble those in [17, fig. 2], where Independent Component Analysis (without considering context information) was used to extract EMG activity sources. The activation patterns in rows d to f are less clear, yet such patterns are also frequently observed. Totally smooth patterns are not to be expected, since the raw EMG signal is known to contain artifacts, even including disconnected electrodes which yield channel features which the network must completely ignore.

Subfigures 2) and 3) display patterns from the deeper layers 2 and 3 of the network, respectively. Again, one can see that patterns emerge which correspond to evolving EMG activity, this is observed best in rows 2a, 2b, 3a, and 3b. We made the interesting observation that these patterns tend to become more intricate than in the first layer, although such a statement is somewhat hard to quantify.

The observed EMG patterns are quite similar for phone

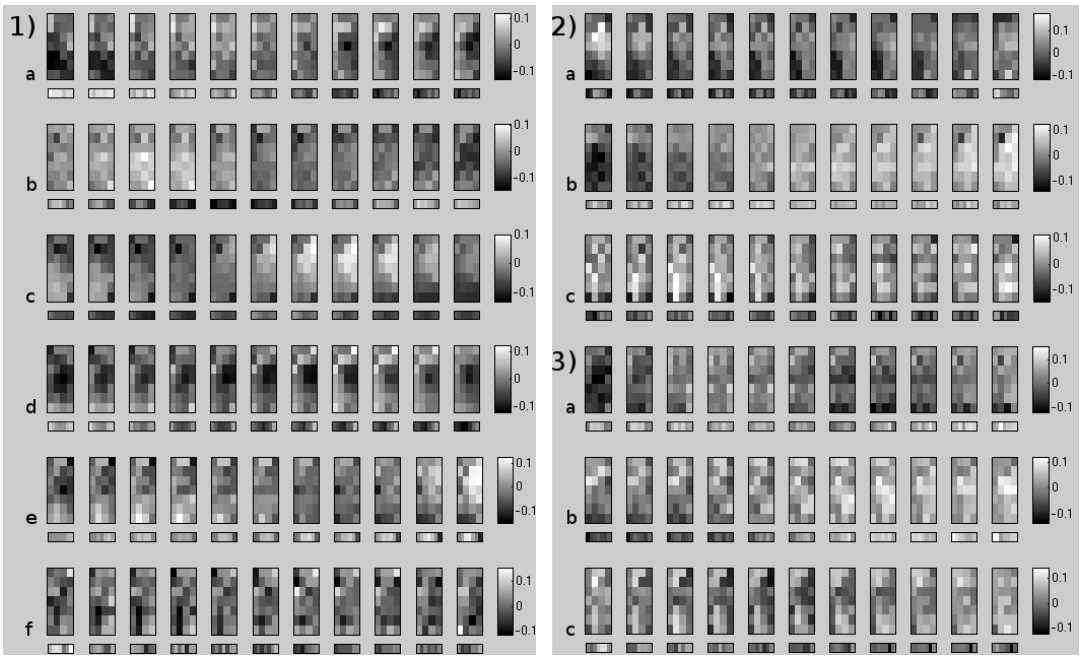

Fig. 3. Visualization of (array-shaped) input features learned by nodes in the hidden layers 1, 2, and 3. See text for details.

and phonetic feature networks. Finally, we remark that the 30-20-30 network which we used for the above investigation is *not* the best one we can come up with: Larger network sizes, e.g. in the range 160-80-160, typically yield better results in both the phone and PF classification tasks. Figure 2 always displays DNN accuracies for the optimal network size, as determined on the development set. When we compare the maximum activation patterns for smaller and larger networks, we typically observe that irregular patterns, like the one in figure 3, row 1f), emerge with increasing frequency. It must be concluded that such patterns do contain important information for classification, yet the fact that small systems create more regular patterns indicates that these regular patterns are the most important ones.

VI. CONCLUSION

In this study we showed that Deep Neural Networks may be used for extracting and visualizing distinctive EMG features which play a role in the classification process. This result is a stepping stone towards improved understanding of the classification which occurs in our EMG-based silent speech recognizer, since it offers a way to extract EMG activities which hint at the difference between certain phones, or present and absent phonetic features. Future work will include an extended analysis of the intra-layer activity propagation in the DNN, with the goal of extracting a set of features which directly represent certain articulatory activities, and, in turn, can be linked to phonetic features.

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] M. Wand, C. Schulte, M. Janke, and T. Schultz, "Array-based Electromyographic Silent Speech Interface," in *Proc. Biosignals*, 2013, pp. 89 – 96.

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

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

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

[6] M. Wand and T. Schultz, "Analysis of Phone Confusion in EMG-based Speech Recognition," in *Proc. ICASSP*, 2011, pp. 757 – 760.

[7] Y. Deng, R. Patel, J. T. Heaton, G. Colby, L. D. Gilmore, J. Cabrera, S. H. Roy, C. J. D. Luca, and G. S. Meltzner, "Disordered Speech Recognition Using Acoustic and sEMG Signals," in *Proc. Interspeech*, 2009, pp. 644 – 647.

[8] J. Freitas, A. Teixeira, and M. S. Dias, "Towards a Silent Speech Interface for Portuguese," in *Proc. Biosignals*, 2012, pp. 91 – 100.

[9] N. Srisuwan, P. Phukpattaranont, and C. Limsakul, "Feature Selection for Thai Tone Classification based on Surface EMG," *Procedia Engineering*, vol. 32, pp. 253 – 259, 2012.

[10] A. Toth, M. Wand, and T. Schultz, "Synthesizing Speech from Electromyography using Voice Transformation Techniques," in *Proc. Interspeech*, 2009, pp. 652 – 655.

[11] C. Johner, M. Janke, M. Wand, and T. Schultz, "Inferring Prosody from Facial Cues for EMG-based Synthesis of Silent Speech," in *Proc. AHFE*, 2012, pp. 5317 – 5326.

[12] S.-C. Jou, T. Schultz, M. Walliczek, F. Kraft, and A. Waibel, "Towards Continuous Speech Recognition using Surface Electromyography," in *Proc. Interspeech*, 2006, pp. 573 – 576.

[13] G. E. Hinton and R. R. Salakhutdinov, "Reducing the Dimensionality of Data with Neural Networks," *Science*, vol. 313, pp. 504 – 507, 2006.

[14] K. Cho, T. Raiko, and A. Ilin, "Gaussian-Bernoulli Deep Boltzmann Machine," in *Proc. IJCNN*, 2013.

[15] Available online: http://www.cs.toronto.edu/~hinton/MatlabForSciencePaper.html.

[16] D. Erhan, Y. Bengio, A. Courville, and P. Vincent, "Visualizing Higher-layer Features of a Deep Network," University of Montreal, Tech. Rep., 2009.

[17] T. Heistermann, M. Janke, M. Wand, and T. Schultz, "Spatial Artifact Detection for Multi-Channel EMG-Based Speech Recognition," in *Proc. Biosignals*, 2014, pp. 189 – 196.