INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Deep Neural Network Frontend for Continuous EMG-based Speech Recognition

Michael Wand and Jürgen Schmidhuber

Istituto Dalle Molle di Studi sull'Intelligenza Artificiale (The Swiss AI Lab IDSIA),
USI & SUPSI, Manno-Lugano, Switzerland

{michael, juergen}@idsia.ch

Abstract

We report on a *Deep Neural Network* frontend for a continuous speech recognizer based on Surface Electromyography (EMG). Speech data is obtained by facial electrodes capturing the electric activity generated by the articulatory muscles, thus allowing speech processing without making use of the acoustic signal. The electromyographic signal is preprocessed and fed into the neural network, which is trained on framewise targets; the output layer activations are further processed by a Hidden Markov sequence classifier. We show that such a neural network frontend can be trained on EMG data and yields substantial improvements over previous systems, despite the fact that the available amount of data is very small, just amounting to a few tens of sentences: on the *EMG-UKA* corpus, we obtain average evaluation set Word Error Rate improvements of more than 32% relative on context-independent phone models and 13% relative on versatile *Bundled Phonetic feature* (BDPF) models, compared to a conventional system using Gaussian Mixture Models. In particular, on simple context-independent phone models, the new system yields results which are almost as good as with BDPF models, which were specifically designed to cope with small amounts of training data.

Index Terms: Silent Speech interface, Deep Neural Networks, Electromyography, EMG-based Speech Recognition

1. Introduction and Related Work

Spoken language is of tremendous importance in our daily lives, for example for transmission of vital information, for communicating desires and intentions, and for social interaction. However, several issues arise due to the fact that speech must be pronounced audibly, including interference with the environment (bystanders are disturbed, private communication is impossible) and exclusion of speech-disabled people (for example laryngectomees, whose voice box has been removed) [1].

In this study we use a continuous speech recognizer based on Surface Electromyography (EMG) to overcome these problems: electric activity of the user's facial muscles is captured by surface electrodes, allowing speech to be processed without using the acoustic signal and thus enabling a *Silent Speech interface* [1]. Our baseline is an EMG-based speech recognizer developed by the first author at Karlsruhe Institute of Technology [2], with a Gaussian Mixture Model (GMM) frontend and a Hidden Markov Model (HMM) backend using either context-independent phone models [3] or *Bundled Phonetic Feature models* [4] (see section 4.2). The latter are a powerful generalization of conventional context dependent phone models for situations where very little training data is available. The contribution of this study is a hybrid classifier using a Deep Neural Network (DNN) frontend instead of the GMM: we demonstrate that such a system can be trained and significantly improves the recognition accuracy of the EMG-based speech recognizer, despite the fact that we use only a few minutes of training data per system. In addition, we can do away with the Bundled Phonetic Feature models, i.e. simplify the modelling paradigm, with only a very small loss of accuracy. This paper substantially extends our previous work [5], which dealt with neural network classification and analysis of EMG data at the frame level.

The paper is organized as follows: section 2 presents related work, sections 3–4 lay out the foundations of this study, and sections 5–6 present experiments and results on the new DNN-HMM system. Section 7 concludes the study.

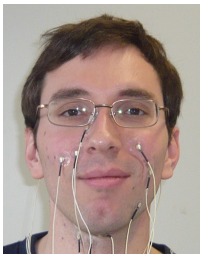
2. Related Work

EMG-based speech recognition started in the 80's with the studies of Sugie and Tsunoda [6, 7], extended by Jorgensen to non-audible speech [8]. The first system to recognize *continuous* speech from EMG signals was presented in 2006 [3]; our *Bundled Phonetic Feature* models [4] substantially improve this setup, yielding Word Error Rate (WER) reductions of more than 33% relative. With rising interest in nonacoustic communication, various groups have been investigating topics such as optimized signal processing [9] and acquisition [10], the discrepancy between audibly spoken and silently mouthed speech [11, 12], language-specific challenges [13], direct synthesis of speech from EMG signals [14, 15, 16], and session adaptation [17]. Other forms of non-acoustic speech processing include the relatively new method of permanent magnetic articulography [18] and visual systems using facial images and/or ultrasound recording of the vocal tract [19, 20].

Artificial Neural Networks (ANNs) have been successfully used in speech recognition since the 1990's [21], more recently incorporating unsupervised pre-training [22, 23]. Current research targets two main directions: First, the feature preprocessing frontend is gradually replaced by networks trained on raw waveforms [24], and second, the HMM backend is replaced by a *recurrent neural network* architecture [25] usually implemented as LSTM (*Long Short-Term Memory* [26]), leading to end-to-end neural network systems. A first study on applying ANNs to EMG-based speech recognition was published by the first author in 2014 [5]; with experiments at the frame level, it was shown that ANNs not only improve the recognition accuracy, but also allow for "extracting and visualizing distinctive EMG features".

This research was supported by the FP7 Marie Curie Initial Training Network PROTOTOUCH (grant #317100).

http://dx.doi.org/10.21437/Interspeech.2016-340



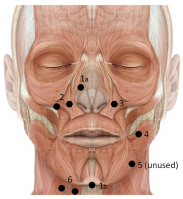


Figure 1: Electrode positioning (from [27]) with chart of the underlying muscles (muscle chart adapted from [28])

3. Data Corpus

We use the *EMG-UKA* Corpus [27], which is the most comprehensive publicly available corpus for EMG-based speech recognition. It consists of surface electromyographic and acoustic recordings of read speech in English language, from the Broadcast News domain. Data was recorded as normal (audible) speech as well as whispered and silently mouthed speech; in this study, as in [5], only EMG data from audible speech was used since only for this kind of data, high-quality phone-level time alignments are available[1]. Furthermore, only the "small" sessions from the EMG-UKA corpus distribution were used.

Figure 1 shows the recording setup, consisting of 6 EMG channels capturing data from major facial muscles according to [29], namely levator anguli oris, zygomaticus major, platysma, depressor anguli oris, anterior belly of the digastric, and the tongue. Channel 5 yields unstable signals and is not used. Recordings were performed with 600Hz sampling rate. The acoustic signal was recorded synchronously with a standard headset microphone; acoustic data was used to generate phone-level time alignments which are part of the official corpus distribution. Otherwise, the audio signal is not used in this study. Note that whenever audibly spoken speech is available as training data, the creation of data alignments is a straightforward task for any off-the-shelf conventional speech recognizer.

Each session comprises 50 sentences, a *BASE* subset of 10 sentences which is the same for each session and used for testing, and a *SPEC* subset of 40 sentences, which varies between sessions and is used for training. All experiments are *session-dependent*, i.e. separate systems are trained on each session of each speaker. We divide the entire (small-session, audible speech) data of the EMG-UKA corpus into a development and an evaluation set according to the official corpus distribution: The sessions available in the free *trial* corpus are considered development data and used for parameter tuning, the remaining sessions from the *full* corpus distribution are used for evaluating the final systems with the best settings. Thus, the development data consists of 12 sessions, and the evaluation data consists of 49 sessions. The corpus statistics are summarized in table 1.

Set	Number of Speakers	Number of Sessions	Avg session length	Total length
Dev	4	12	3:19	39:47
Eval	7	49	3:06	2:31:47

Table 1: Statistics of the data corpus

[1] Refer to [12] for a report on analyses of and remedies for the discrepancy between speaking modes.

4. Baseline System

4.1. Feature Extraction

We compute standard EMG *time-domain features* [3] according to the recipe in [27]: For any time-domain signal $\mathbf{x}$, define the frame-based time-domain mean $\bar{\mathbf{x}}$, the frame-based power $\mathbf{P_x}$, and the frame-based zero-crossing rate $\mathbf{z_x}$, with a frame size of 27ms and a frame shift of 10ms. Also, for a framewise feature $\mathbf{f}$, $S(\mathbf{f}, n)$ is the stacking of $2n+1$ adjacent frames, with time indices from $-n$ to n.

Each mean-normalized EMG channel $x[n]$ is first low-pass filtered by twice applying a nine-tap unweighted moving average filter, yielding the low-frequency part $w[n]$ of the EMG signal. The high-frequency part is the residual $p[n] = x[n] - w[n]$, and its rectification is $r[n] = |p[n]|$. The framewise feature *before* context stacking is defined as $\mathbf{TD}0 := [\bar{\mathbf{w}}, \mathbf{P_w}, \mathbf{P_r}, \mathbf{z_p}, \bar{\mathbf{r}}]$, the final feature[2] is $\mathbf{TD}5 := S(\mathbf{TD}0, 5)$.

Finally, all channels are stacked, and a Linear Discriminant Analysis (LDA) dimensionality reduction is applied. The LDA matrix is computed on the training data of each session, with 127 target classes corresponding to three substates for each of 42 phones plus one silence model, using the frame-level alignments included in the EMG-UKA corpus. The LDA cutoff dimension varies, see section 6.

4.2. GMM-HMM Recognizer

The recognizer uses an established pattern [2, 4] with three-state left-to-right fully continuous HMMs with a GMM emission probability frontend, implemented with the software BioKIT [30]. Two model structures are employed, namely *context-independent phone* models and *Bundled Phonetic Feature* (BDPF) models; the latter were developed by the first author as an efficient method to generate a data-adapted model structure even when the amount of data precludes classical context-dependent models [4]. BDPF modeling consists of training a number of phonetic decision trees [31] whose roots correspond to (binary) phonetic features, like the place or manner of articulation or broader features like *voicedness*, and then averaging over frame-level scores of different phonetic decision trees to obtain a final emission probability for the HMM. In contrast to standard GMM training, the BDPF criterion is discriminative, using an entropy measure to perform a top-down decision tree generation. For this study, we use eight equally-weighted BDPF "streams", which correspond to eight decision trees whose roots are the most common phonetic features established in [2]. Note that each such stream contains its own set of BDPF models. Refer to [4] for a detailed description of the BDPF model structure and to [27, 2] for the details of the setup used in this study.

Training the phone-based recognizer requires computing Gaussian means and variances on 127 classes (substates of phones plus silence). This is done by merge-and-split training followed by 6 iterations of expectation maximization (EM). We use the acoustics-based time alignments from the corpus distribution, *not* recomputing alignment paths as in Viterbi-style training, since the acoustic paths have turned out to be of higher quality, and since we wish to use the exact same alignments for GMM and DNN training. In the case of BDPFs, the decision trees are computed using the recipe in [4], again followed by merge-and-split and EM training.

[2] In [27] we used the similarly defined $\mathbf{TD}10$ instead, however with a slightly different phone structure (45 instead of 42 phones). In preliminary experiments for this study, we obtained insignificantly better results with the $\mathbf{TD}5$ feature.

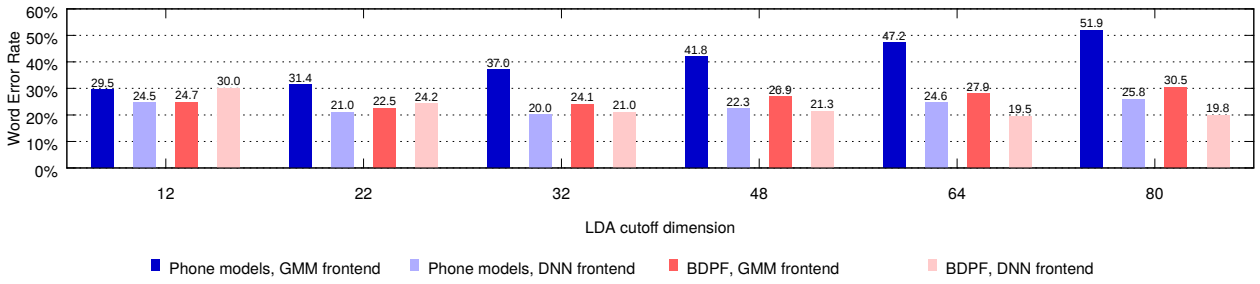


Figure 2: Word Error Rates on the development data, with different LDA cutoff dimensions

In the testing phase, the trained GMM-HMM is used together with a trigram Broadcast News language model whose evaluation set perplexity is 24.24. As in previous experiments, we limit the decoding vocabulary to the 108 words appearing in the test set. Note that this constraint is mainly due to the small amount of training data; we published several experiments with larger training data amounts and larger vocabularies, as well as with session-independent setups [17, 2].

5. Deep Neural Network Setup

The contribution of this study is the replacement of the GMM emission models with a discriminatively trained frontend based on Deep Neural Networks (DNN), for which we used our in-house toolkit PyLSTM. DNNs are trained using a frame-based accuracy criterion; when DNN training is finished, the test data is ran through a single feed-forward pass of the neural network, and the activations of the final softmax layer are directly used as state-level probabilities for the backend Hidden Markov model. As before, all training is done session dependently.

The DNNs are set up as follows. After initial experiments, we chose networks with four hidden layers, each of which consists of 200 neurons with a tanh nonlinearity. The neurons of the final softmax layer correspond to the possible state models of the HMM backend. Therefore, there are two substantially different training setups: For the phone-based system, we have 127 phone substate models as in the GMM case; for the BDPF system, we have around 100 state models *per BDPF stream*, and we train DNNs for each stream separately. The decision tree structure is derived from the GMM-based system according to the method described in section 4.2, so that between the GMM-based system and the DNN-based system, only the last stage of training differs. Note that it is also possible to compute DNN-based phonetic decision trees without using GMMs [32].

All DNNs are initialized using a Gaussian distribution with a standard deviation of 0.1 (thus, no pretraining is used) and trained by stochastic gradient descent on the Multiclass Cross Entropy criterion using minibatches of size 30, with a learning rate of 0.005. Regularization is not applied, except of early stopping when the training data accuracy has not improved for 5 epochs (due to the small amount of session-dependent training data, we refrained from splitting off a separate validation set).

6. Results and Analysis

6.1. GMM frontend versus DNN frontend

We now have four different experiments, characterized by model structure (phone-based or BDPF) and frontend (GMM or neural network). Figure 2 shows the average Word Error Rates (WER) on the development sessions, for different LDA cutoff dimensions. Note that we did not investigate GMM systems with less than 12 dimensions since it was shown in [2] that this does not improve the recognition accuracy.

We make the surprising observation that the optimal LDA cutoff dimension vastly differs between the GMM and the DNN systems: For GMMs, retaining more than 12 dimensions (for the phone-based system) resp. 22 dimensions (for the BDPF system) causes deteriorating results, more so for the phone system than for the BDPF system. This may be due to the way GMMs are generatively estimated from *only* the training data corresponding to a particular model: since some phones occur sparsely in the training data, these models are in danger of being undertrained when the feature dimensionality rises. The problem is alleviated by BDPF models, since they are generated subject to a constraint on the available amount of data. Yet, even for BDPFs it is not helpful to go beyond 22 dimensions.

Using the DNN frontend, we can and must use much higher LDA cutoff dimensionalities. For the phone-based system, the optimal result is 20.0% WER with 32 features after LDA: compared to the best phone-based GMM system, this is an improvement of more than 32% relative. With BDPF models, the best result is obtained with 64 features; the improvement from 22.5% to 19.5% WER is less drastic, but still a noticeable 13% relative. The WER with the DNN frontend is almost always lower than the WER with the GMM frontend, with the notable exception of the BDPF system at low LDA cutoffs: this is not yet fully explained, but may be related to the fact that the model structure was generated using discriminability with GMMs at an LDA cutoff dimension of 22 as an optimization target.

We remark that using a smaller number of layers does not cause substantial accuracy degradation, yet it is notable that even with the small amount of training data, the training procedure is robust enough to allow training a large number of layers, which in our opinion bodes well for future experiments with more complex setups (e.g. session- and speaker-independent training). We emphasize that the DNN performance with phone models is only slightly worse than with the much more sophisticated BDPF models. Finally, we remark that due to the stochastic nature of DNN training, some variability in the single DNN results has been observed and should be expected.

6.2. Why does the DNN frontend mitigate the disadvantage of phone models?

Having established the potential of Deep Neural Networks for EMG-based continuous speech recognition, it is a striking result that the DNN frontend partly mitigates the disadvantage of phone-based models compared to BDPF models, and we ask for

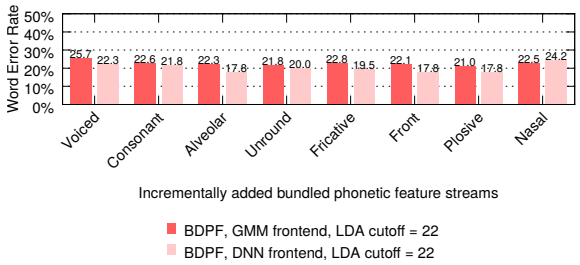


Figure 3: Word Error Rate evolution when Bundled Phonetic Feature streams are *incrementally* added, on the development data. Labels indicate the root phonetic feature, note that *each stream* comprises Bundled Phonetic features.

the reason. There are two main properties of BDPFs in which they differ from phone models (see section 4.2), namely 1) the data-driven discriminative optimization of the model structure, 2) the combination of scores from different streams.

Figure 3 shows the evolution of the WER when BDPF streams are incrementally added, in order of frequency of the underlying phonetic feature. For the GMM systems, we confirm the well-known result [4, 2] that just a few, but definitely more than one BDPF stream are required to obtain optimal results. When the DNN frontend is used, the tendency becomes less clear: While at least three BDPF streams are required to obtain the optimal WER, this result should not be assumed to be generalizable since there is no consistent behavior of the WERs when more streams are added. Yet conversely, this suggests that when the DNN frontend is used, multi-stream modeling could be discarded, which makes the modeling structure simpler and also should allow us to compare the GMM and DNN frontends with just one single BDPF stream.

We ran a further set of experiments using just *one* BDPF stream, corresponding to the phonetic feature of voicedness, the most frequent one in our corpus. The resulting WERs are displayed in figure 4, where we observe the very same pattern as in figure 2. Thus, it is proven that DNN training is indeed more effective on phone models than on BDPF models *even when only a single BDPF stream is used*: so this must be due to the fundamentally different (discriminative) structure of BDPF models.

Next, we show that the discrepancy between phone and BDPF models is *not* due to the different number of models in these systems. As described above, the phone systems have exactly 127 models, now we trained a set of systems on the BDPF stream for voicedness, tuning the BDPF tree generation

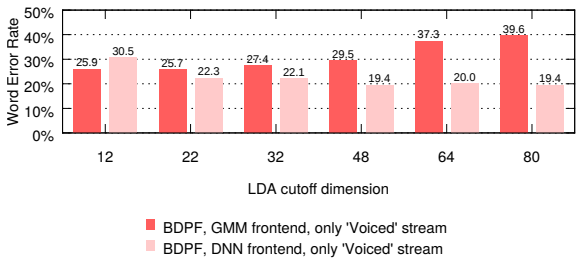


Figure 4: Word Error Rates with different LDA cutoff dimensions using only one bundled phonetic feature stream (the 'Voiced' stream), on the development data

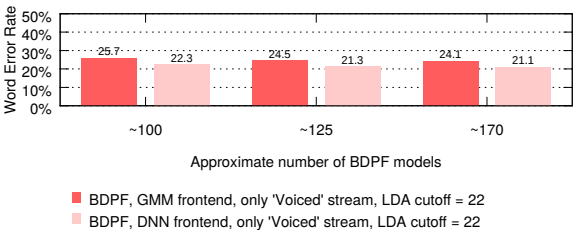


Figure 5: Word Error Rates with different number of BDPF tree leaves using the *voiced* BDPF stream, on the development data

criterion leading to systems with approximately 100 (the original), 125, and 170 models per stream (due to the tree generation criterion, it is impossible to *exactly* fix the model count). The resulting WERs are displayed in figure 5: the accuracy varies only slightly for different numbers of BDPF models, for both the GMM and the DNN system. This shows that the different behavior of phone models and BDPF models when replacing GMM with DNN is not due to the number of models.

6.3. Statistical test on the evaluation data

Finally, for statistical evaluation, we make two hypotheses, namely 1) the DNN frontend improves the phone-based system, 2) the DNN frontend improves the BDPF system. In all cases, we use the optimal LDA cutoff dimensionality as determined on the development data. The resulting WERs, averaged over the 49 sessions of the evaluation data set, are shown in table 2, together with the corresponding WERs on the development data for comparison. It can be seen that the tendency is similar on development and evaluation data; a significance test (one-tailed t-test with paired samples) shows that *both results are highly significant* ($p = 7.02 \times 10^{-8}$ for the phone-based system, $p = 1.17 \times 10^{-2}$ for the BDPF system).

	Phone Models		BDPF Models	
	GMM	DNN	GMM	DNN
Development	29.5%	20.0%	22.5%	19.5%
Evaluation	33.6%	26.5%∗	27.2%	23.8%∗

Table 2: Average Word Error Rates of the optimal systems on the development and evaluation data. Results marked with ∗ are significant improvements over the respective baseline.

7. Conclusion and Outlook

We showed that a hybrid DNN-HMM recognizer substantially improves EMG-based speech recognition compared to a GMM-based system, while allowing a much simpler modeling structure (context-independent phones instead of Bundled Phonetic features) with hardly any loss of accuracy. In particular, we proved that it is possible to train the neural network frontend even with very little training data: We believe this is crucial not only for this particular system, but also for other Silent Speech recognizers, since in all cases the available data corpora are much smaller than typical acoustic speech corpora.

This result paves the way towards further experiments and analyses, in particular, deep learning shall allow to generate optimal distributed representations of the EMG-to-speech mapping, for example when session- or even speaker-independent recognition is desired, or for speech synthesis [16] tasks.

8. 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, "Advancing Electromyographic Continuous Speech Recognition: Signal Preprocessing and Modeling," Dissertation, Karlsruhe Institute of Technology, 2014.

[3] 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.

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

[5] M. Wand and T. Schultz, "Pattern Learning with Deep Neural Networks in EMG-based Speech Recognition," in *Proc. EMBC*, 2014, pp. 4200 – 4203.

[6] N. Sugie and K. Tsunoda, "A Speech Prosthesis Employing a Speech Synthesizer – Vowel Discrimination from Perioral Muscle Activities and Vowel Production," *IEEE Transactions on Biomedical Engineering*, vol. 32, no. 7, pp. 485 – 490, 1985.

[7] M. S. Morse and E. M. O'Brien, "Research Summary of a Scheme to Ascertain the Availability of Speech Information in The Myoelectric Signals of Neck and Head Muscles using Surface Electrodes," *Computers in Biology and Medicine*, vol. 16, no. 6, pp. 399 – 410, 1986.

[8] C. Jorgensen, D. D. Lee, and S. Agabon, "Sub Auditory Speech Recognition Based on EMG/EPG Signals," in *Proc. IJCNN*, Portland, Oregon, 2003, pp. 3128 – 3133.

[9] G. S. Meltzner, G. Colby, Y. Deng, and J. T. Heaton, "Signal Acquisition and Processing Techniques for sEMG based Silent Speech Recognition," in *Proc. EMBC*, 2011, pp. 4848 – 4851.

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

[11] 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.

[12] M. Wand, M. Janke, and T. Schultz, "Tackling Speaking Mode Varieties in EMG-based Speech Recognition," *IEEE Transaction on Biomedical Engineering*, vol. 61, no. 10, pp. 2515 – 2526, 2014.

[13] J. Freitas, A. Teixeira, S. Silva, C. Oliveira, and M. S. Dias, "Velum Movement Detection based on Surface Electromyography for Speech Interface," in *Proc. Biosignals*, 2014, pp. 13 – 20.

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

[15] K.-S. Lee, "Prediction of Acoustic Feature Parameters using Myoelectric Signals," *IEEE Transactions on Biomedical Engineering*, vol. 57, pp. 1587 – 1595, 2010.

[16] L. Diener, M. Janke, and T. Schultz, "Codebook Clustering for Unit Selection Based EMG-to-Speech Conversion," in *Proc. Interspeech*, 2015.

[17] M. Wand and T. Schultz, "Towards Real-life Application of EMG-based Speech Recognition by using Unsupervised Adaptation," in *Proc. Interspeech*, 2014.

[18] J. A. Gonzalez, L. A. Cheah, J. M. Gilbert, J. Baib, S. R. Ell, P. D. Green, and R. K. Moorea, "A Silent Speech System Based on Permanent Magnet Articulography and Direct Synthesis," *Computer Speech and Language*, vol. (in press), 2016.

[19] T. Hueber and G. Bailly, "Statistical Conversion of Silent Articulation into Audible Speech using Full-covariance HMM," *Computer Speech and Language*, vol. 36, pp. 274 – 293, 2016.

[20] M. Wand, J. Koutník, and J. Schmidhuber, "Lipreading with Long Short-Term Memory," in *Proc. ICASSP*, 2016.

[21] H. Bourlard and N. Morgan, *Connectionist Speech Recognition. A Hybrid Approach*. Kluwer Academic Publishers, 1994.

[22] J. Schmidhuber, "Learning Complex, Extended Sequences Using the Principle of History Compression," *Neural Computation*, vol. 4, no. 2, pp. 234 – 242, 1992.

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

[24] N. Jaitly and G. Hinton, "Learning a Better Representation of Speech Soundwaves Using Restricted Boltzmann Machines," in *Proc. ICASSP*, 2011.

[25] S. Fernández, A. Graves, and J. Schmidhuber, "An Application of Recurrent Neural Networks to Discriminative Keyword Spotting," in *Proc. ICANN*, 2007, pp. 220 – 229.

[26] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, pp. 1735 – 1780, 1997.

[27] M. Wand, M. Janke, and T. Schultz, "The EMG-UKA Corpus for Electromyographic Speech Processing," in *Proc. Interspeech*, 2014, pp. 1593 – 1597.

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

[29] L. Maier-Hein, F. Metze, T. Schultz, and A. Waibel, "Session Independent Non-Audible Speech Recognition Using Surface Electromyography," in *Proc. ASRU*, 2005, pp. 331 – 336.

[30] D. Telaar, M. Wand, D. Gehrig, F. Putze, C. Amma, D. Heger, N. T. Vu, M. Erhardt, T. Schlippe, M. Janke, C. Herff, and T. Schultz, "BioKIT - Real-time Decoder for Biosignal Processing," in *Proc. Interspeech*, 2014.

[31] L. R. Bahl, P. V. de Souza, P. S. Gopalakrishnan, D. Nahamoo, and M. A. Picheny, "Decision Trees for Phonological Rules in Continuous Speech," in *Proc. ICASSP*, 1991, pp. 185 – 188.

[32] L. Zhu, K. Kilgour, S. Stüker, and A. Waibel, "Gaussian Free Cluster Tree Construction Using Deep Neural Network," in *Proc. Interspeech*, 2015.