Inferring prosody from facial cues for EMG-based synthesis of silent speech

Christian Johner, Matthias Janke, Michael Wand, Tanja Schultz

Cognitive Systems Laboratory
Karlsruhe Institute of Technology, Germany
christian.johner@student.kit.edu,
{matthias.janke, michael.wand, tanja.schultz}@kit.edu

ABSTRACT

In this paper we introduce a system which is able to detect prosodic elements in a spoken utterance based on signals from the facial muscles. The proposed system can augment our surface electromyography (EMG) based *Silent Speech Interface* in order to make synthesized speech more natural. Having shown in (Nakamura, Janke, Wand, & Schultz, 2011) that it is possible to produce understandable synthesized speech from EMG signals, our current interest is to improve the quality and expressivity of the synthesis.

We show that a standard phonetically balanced German speech corpus with only a few additional utterances is sufficient to train a system that can discriminate yes/no questions from normal speech and also distinguish between normal and emphasized words in an utterance.

For the detection of prosodic information in facial muscle movement we extend our EMG based speech synthesis system with two additional EMG channels, recording the movements of the facial muscles *musculus corrugator* and *musculus frontalis*. Our classification method uses a frame-based SVM classification, followed by a majority vote to classify a whole word.

Our system achieves F-scores of up to 0.68 for the recognition of emphasized words and 1.0 for the classification between questions and normal utterances although the results show large variations depending on the feature combination used for training.

Keywords: EMG, synthesis, prosody, speech recognition

1 INTRODUCTION

Speech is the most convenient and natural way for humans to communicate. Beyond face-to-face talk, mobile phone technology and speech-based electronic devices have made speech a wide-range, ubiquitous means of communication. Unfortunately, voice-driven communication systems suffer from several challenges which arise from the fact that the speech needs to be clearly audible and cannot be masked: first, understanding the speech becomes very difficult in the presence of noise, for both humans and computers. Second, confidential communication in public places is difficult if not impossible, and even if privacy is not an issue, the audible speech frequently disturbs bystanders. Third, speech-disabled people may be unable to talk to other persons or to use voice-controlled systems.

These challenges may be alleviated 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 (Denby, Schultz, Honda, Hueber, & Gilbert, 2010). Over the past few years, we have developed a Silent Speech Interface (Schultz & Wand, 2010) based on surface electromyography (EMG). This technique captures articulatory activity related to speech production from the speaker's face using surface electrodes. Rather than recording acoustics, EMG captures muscle activity and therefore does not require any kind of acoustic signal: The speaker can speak *silently*, which means that the words are mouthed without any sound.

In this paper we report on using this technique to directly synthesize speech based on the electromyographic signals. This approach is particularly suited to human-human communication. However, speech carries more information than the pure meaning of words. Prosodic information codes whether an utterance should be understood as an urgent request, a question or a mundane statement. Current text-to-speech systems are capable of modulating the produced voice to match different sentence types based on prosodic annotations (Silverman, et al., 1992). The goal of our work is to empower our EMG-based speech synthesis system with the same capabilities – i.e. speech synthesized from EMG signals should exhibit varying intonation, emphasis, and prosody, reflecting the speaker's intentions.

In (Toth, Wand, & Schultz, 2009) we showed that understandable speech may be synthesized from electromyographic signals, using a voice conversion technique. However, these initial systems suffered from an unnatural quality of the synthesized voice. This paper draws on (Nakamura, Janke, Wand, & Schultz, 2011), where we showed for the first time that F0 contours may by recognized from EMG signals, and that this information may be incorporated into an EMG-to-speech system. Beyond the F0 contour generation, in this paper we focus on two aspects of prosody which we perceive to be of utmost importance for conveying meaning by speech: Firstly, we recognize emphasized words from EMG signals of spoken utterances, and second, we distinguish questions from normal statements.

In this work we only report results on recordings of audibly spoken speech. For our experiments we extend the setup used by (Toth, Wand, & Schultz, 2009) with two additional electrodes placed on the *musculus corrugator* and the *musculus*

frontalis, using a total of seven EMG channels.

This paper is organized as follows: In Section 2 we describe our experimental setup, the data corpus, and EMG feature extraction and SVM training setup. Section 3 details our experiments for question and emphasis classification. In Section 4, we draw conclusions and outline future work.

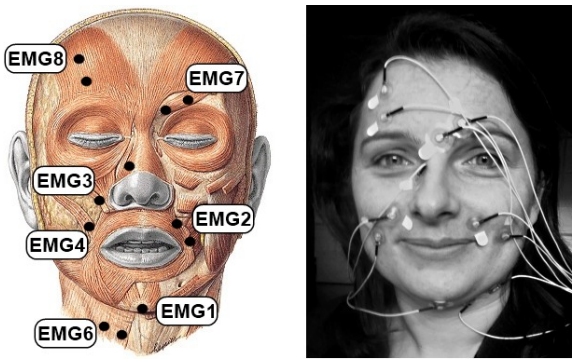

Figure 1: Illustration of facial muscles and electrode positions

2 EXPERIMENTAL SETUP

2.1 Data Acquisition and Corpus

Our data corpus is based on the Berliner Sätze (Sotschek, 1984), which is a phonetically balanced German speech corpus containing 102 utterances. We extended this corpus with 13 additional utterances to be realized as normal statements, 35 utterances with one emphasized word and 20 questions, both of which were recorded twice, and 10 exclamations, which were not used for this study, so that each subject, in each recording session, recorded 235 sentences. The sentences beside the Berliner Sätze corpus are not phonetically balanced.

The EMG signals were recorded with 14 Ag/Ag-Cl surface electrodes attached to the skin, as depicted in figure 1. The 14 electrodes were positioned in order to pick up the signals of major articulatory muscles: the *levator angulis oris* (EMG2,3), the *zygomaticus major* (EMG2,3), the *platysma* (EMG4), the *anterior belly of the gastric* (EMG1), the *tongue* (EMG1,6) and relevant facial muscles: *corrugator* (EMG7) and *frontalis* (EMG8). Channels 2, 6, 7, and 8 used bipolar derivation, whereas channels 1, 3, and 4 were unipolarly derived, with the reference attached to either the nose (EMG1) or to both ears (EMG3,4). Note that the electrode positioning follows (Maier-Hein, Metze, Schultz, & Waibel, 2005) for the articulatory (lower facial) muscles and (Pruzinec, 2010) for the upper facial muscles. EMG channel 5 remains unused.

EMG data was recorded with a multi-channel EMG recording system (Varioport, Becker Meditec, Germany). EMG responses were differentially

amplified, filtered by a 300 Hz low-pass and a 1Hz high-pass filter and sampled at 600 Hz.

We recorded seven sessions with 4 different speakers (three male, one female) all of them native Germans. In order to avoid inconsistencies due to slightly different electrode positioning and speaker properties, we trained *session-dependent systems.* Since all male speaker recorded two sessions, we nonetheless have a means of asserting that classifier settings remain consistent across recording sessions.

The recording protocol was as follows:

In a quiet room, the speaker read German sentences in normal audible speech, which were recorded with a parallel setup of an EMG recorder and a USB soundcard with a standard close-talking microphone attached to it. An analog marker signal was used for synchronizing the EMG and the speech signals.

We gave each speaker a short introduction about the purpose of the recording and encouraged them to be as natural during the recording as possible. It should be noted that none of the subjects was a professional actor or speaker.

2.2 EMG Feature Extraction

The feature extraction is based on time-domain features (Jou, Schultz, Walliczek, Kraft, & Waibel, 2006). Here, for any given feature **f**, $\mathbf{M_f}$ 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(**f**,n) is the stacking of adjacent frames of feature **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[k] is defined as:

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

The high-frequency signal is p[n] = x[n] - w[n], and we define the rectified high-frequency signal r[n] = |p[n]|. The final base feature is:

$$TDn = S(f_2, n), \; where \quad f_2 = \lfloor M_w, P_w, P_r, z_p, M_r \rfloor$$

We evaluated the final feature TDn with a stacking width of n = 1, 3, 7, 10 and 15. Frame size and frame shift were set to 27ms respectively 10ms. The feature extraction was applied for each EMG channel separately. In contrast to (Schultz & Wand, 2010) we did not apply any dimension reduction.

2.3 SVM Training

The experiments described in this paper require a word segmentation of the data to function properly. For this study, we obtained these word alignments by forced-aligning the simultaneously recorded audio data with a German speech recognizer

trained with the Janus Recognition Toolkit (JRTK). We note that such time-alignments may be obtained just as well based on purely EMG data (Schultz & Wand, 2010).

For the training we split the set of recorded utterances into a training set and a test set. These sets were formed separately for the two recognition tasks.

Yes/no-questions have a significant different prosody compared to normal statements. Other types of questions do not necessarily use the same prosody. From

Figure 2: Illustration of training set generation. Each bar represents an utterance and each cube a frame. The red frames align with an emphasized word. The red frames were used as training data for the SVM for the emphasis-class (EMP class) and the blue and green ones for the no-emphasis-class (No EMP class)

the 40 questions in our corpus only 20 are yes/no-questions. We split this set into 14 questions for training and six questions for testing. From the remaining utterances we selected 16 sentences for the training. For testing we used the six yes/no questions and eight sentences from the normal utterances. Using only the last word of each utterance, this results in 30 words for training (14 questions/16 normal) and 14 words for testing (6 questions/8 normal).

For the training of the emphasis detector, we took 46 emphasized words and 36 normal words from the emphasized sentences, and 15 sentences containing of 81 words from the remainder corpus. For testing we used the remaining 24 sentences of the emphasized and 15 from the remaining corpus with a total of 226 words (24 emphasized/202 normal).

As described in 2.2, all our data is divided into frames with a frame shift of 10 ms. Every frame represents a data point for one of the two classes. We now train an SVM (Support Vector Machine) to perform a *frame-based* classification of the two classes in question, i.e. to discern questions from normal utterances or emphasized words from normally pronounced words. Figure 2 depicts, as an example, how the training data for the emphasis classifier is constructed, and Table 1 shows the number of training frames for each session. The results of the SVM classification are used as an input for a majority vote on a word base as described in Section 3.

Table 1: Number of frames for each of the seven sessions and classification tasks. Each frame has a duration of 10ms

	spk1-ses1	spk2-ses1	spk3-ses1	spk4-ses1	spk1-ses2	spk2-ses2	spk3-ses2
Emphasis	5500	5800	8366	5999	4748	5581	8396
Question	1972	2118	2375	2052	1902	1998	1858

3 EXPERIMENTS

We examined a large number of different parameter settings to find the best values for the classification of each task. For EMG-based speech recognition (Schultz and Wand, 2010) TD15 features have been shown to give good results. (Pruzinec and Schultz 2011) examined TD0 to TD5 features for EMG-based facial expression classification. We initially studied TD1, TD3, TD7, TD10 and TD15 features, but discarded TD1 and TD3 due to the poor results on the first classifications. We used all seven EMG channels for classification, but also investigated the impact of reducing the set of channels.

Our classifier is a combination of a SVM which was trained with frames for each class and a suffixed majority vote. We use the MATLAB 2011b SVM implementation with a radial basis kernel. As parameters we examine every combination of *sigma* in {1, 20, 50} and *boxconstraint* in {1, 10} and allow a violation of the Karush-Kuhn-Tucker condition of 5%. For convenience, parameter combinations of sigma and boxconstraint will be written in the remainder of this paper e.g. as s10-b1.

For the emphasis classifier every frame of every non-silent word in the test set the SVM returns a probability to which class the frame belongs. We sum the class probabilities and divide the result with the number of frames for a word. This results in a value between [0,1) which can be understood as the probability that a word is emphasized. We examine different thresholds above which a word is classified as emphasized. For each of these values or thresholds we count the number of *true positives* (TP) and *true negatives* (TN) and use the F-score (F1-score to be accurate) as an evaluation criteria.

For the question classifier our approach works analogously, but we only test the frames of the last word. We consider this a reasonable simplification because after a question the speaker normally stops talking, waiting for an answer or giving a listener a short pause to think about the said.

We did all the experiments using the first session of every speaker. The three additional sessions of the male speakers are used as an evaluation set.

3.1 Question Classification

Our test set contains 14 words, 6 from questions and 8 from normal utterances. The question words will be regarded as true positives and the normal words as true negatives for the purpose of F-Score calculation.

In our first series of experiments we investigate the performance of different SVM training parameters and TDn features using a stacking of all seven EMG channels for the final feature vector. As shown in Table 2, the classification achieves a very good overall average F-score of 0.87. First, these experiments indicate that SVM parameter combination s50-b1 and s50-b10 lead to slightly better results than the other parameters. Not shown in the table are the results for the

combination s1-b1 and s1-b10, which did not lead to a classification better than chance. This behavior is consistent throughout the experiments. Second, the results indicate that prosodic information can be extracted by using the TD7 and TD10 features which achieved 7 respectively 5 percentage points more than TD15.

Table 2: Question classification - Average F-Scores of the test sessions separated for different TDn features and SVM parameters using all seven EMG channels

	s10-b1	s10-b10	s50-b1	s50-b10	AVG
TD7	0.88	0.90	0.92	0.92	0.90
TD10	0.84	0.85	0.92	0.89	0.88
TD15	0.76	0.76	0.92	0.87	0.83
AVG	0.83	0.84	0.92	0.90	0.87

We now examined the impact of different thresholds. As explained in 3 each word will be assigned a value between zero and one. Depending on the threshold the word will finally be labeled as part of one of the two classes. As it can be seen in Table 3 best results are achieved with a threshold around 0.5. This is encouraging because it allows the conclusion that the SVM training results in good classifications for the frames.

Table 3: Question classification – Average F-scores over all test sessions with different thresholds, SVM parameters s50-b1 and TD10 features using all seven EMG channels

0.6	0.55	0.5	0.45	0.4	AVG
0.79	0.87	0.89	0.89	0.86	0.86

As demonstrated by (Wand & Janke, 2011) the EMG signals of the articulatory muscles differ strongly when changing the speaking mode from audible to silence. Achieving good classification results with the EMG channels from the non-articulatory muscles (EMG channels 7 and 8) would be desirable. With SVM parameters s50-b1 and TD10 features, the best average F-score when using only EMG channels 7 and 8 is achieved for a threshold of 0.4. The F-score for this parameter combination is 0.69, which is a difference of 0.17 compared to the best results with using EMG channels 1 to 8. Results improve when using SVM parameters s10-b10 to an average F-score of 0.77.

This indicates that using EMG channels 7 and 8 is not sufficient to get high recognition rates. Speaker 3 performs extremely poor using this EMG channel combination with a best F-score of 0.67. Speaker 4, on the other hand, achieves a perfect F-score of 1.0 for the parameter combination s10-b10, threshold 0.5 and 0.55 and TD7 feature.

Table 4: Question classification - F-scores for test and evaluation set of TD10 feature with threshold 0.45 and EMG channels 2, 6, and 8 (upper table) and all seven EMG channels (lower table)

	spk1-ses1	spk2-ses1	spk3-ses1	spk4-ses1	AVG	spk1-ses2	spk2-ses2	spk3-ses2	AVG
s50-b10	0.85	0.92	0.8	0.92	0.87	0.73	0.92	0.86	0.84
s50-b1	0.86	0.91	0.86	0.92	0.89	0.73	0.73	0.73	0.73

	spk1-ses1	spk2-ses1	spk3-ses1	spk4-ses1	AVG	spk1-ses2	spk2-ses2	spk3-ses2	AVG
s50-b10	0.77	0.92	0.91	0.83	0.86	0.77	0.86	0.8	0.81
s50-b1	0.77	0.83	0.83	0.83	0.82	0.73	0.83	1.0	0.85

To find a good parameter combination for all speakers, we examined in a last experiment all EMG channel combinations with up to four different channels. Using as a constraint that one of the EMG channels should be either EMG7 or EMG8, the combination with EMG channels 2, 6, and 8 showed the most promising results. The best average F-Score for the four sessions was 0.89 with SVM parameters s50-b1 and 0.87 for s50-b10 both times with TD10 features and a threshold of 0.45, which were indicated by the previous experiment with seven channels to provide good results. We evaluate these parameters with our three unused sessions. As shown in Table 4 the average F-score for SVM parameter s50-b10 of the three sessions is nearly as good as the one with the four test sessions.

3.2 Word Emphasis Classification

The classification of emphasized words in an utterance showed to be a challenging task. We have 24 emphasized words as possible true positives and 202 words as true negatives. It should be noted that this discrepancy will normally lead to low F-Scores.

Table 5: Emphasis classification - Average F-scores for stacking all seven EMG channels

	s10-b1	s10-b10	s50-b1	s50-b10	AVG
TD07	0.45	0.45	0.46	0.46	0.45
TD10	0.38	0.41	0.45	0.47	0.43
TD15	0.37	0.35	0.47	0.5	0.42

We approached the task in the same way we did for classifying questions. First we investigated the results for EMG stacking of all seven EMG channels. TD1 and TD3 showed bad results from the start and were discarded. The behavior of our four test sessions was not consistent. Speaker 3 performed very poorly on TD10 with a maximum F-score of 0.3. Using only the other three sessions, best average F-scores were achieved with s50-b10 as can be seen in Table 5 and TD15 features.

Table 6: Emphasis classification - F-scores for EMG channels 7 and 8 with threshold 0.5, TD15 features

	spk1-ses1	spk2-ses1	spk3-ses1	spk4-ses1	AVG
s50-b1	0.4	0.38	0.25	0.33	0.34
s50-b10	0.42	0.32	0.23	0.38	0.34

EMG channels 7 and 8 showed lower performance than the full channel set, as for the question classification. The average F-score decreases to 0.34 for the best parameter combination, as shown in Table 6. Again Speaker 3 performs poorly.

We did not obtain a channel/parameter combination with good classification rates over all sessions. As can be seen in Table 7, the parameters for the best F-score results with TD15 features vary for each session. Compared to the best results with seven channels the sessions gained between 0.01 for session 1 of Speaker 2 and 0.15 for session 1 of Speaker 4. A positive conclusion, however, is that for all best results either EMG7 or EMG8 are involved, which indicates that we are on the right track.

Table 7: Emphasis classification - Best F-scores for each session with TD15 features and up to four different EMG channels

	spk1-ses1	spk2-ses1	spk3-ses1	spk4-ses1	spk1-ses2	spk2-ses2	spk3-ses2
Channels	2-6-8	6-8	1-3-8	1-5-7	1-2-3-7	1-6-7-8	1-2-7-8
Parameters	s50-b10	s10-b10	s10-b1	s50-b1	s50-b1	s10-b1	s50-b10
Threshold	0.5	0.5	0.6	0.55	0.55	0.55	0.5
F-Score	0.67	0.44	0.42	0.64	0.68	0.55	0.4

4 CONCLUSION AND FUTURE WORK

We showed that it is possible to detect prosodic information in EMG signals. Our approach achieves high classification rates for yes/no questions, and we showed that for this task, the optimal parameter combination remained stable across different speakers and sessions. On the evaluation set, the average F-score is 0.86.

The detection of emphasized words in a complete sentence showed to be a somewhat more challenging task. The best F-score was achieved for Speaker 1 with 0.68, but large variations over different speakers could be noticed.

Having a large discrepancy between true positives (#24) and true negatives (#202), a weak classifier improving the ratio could yield significantly higher recognition rates. A quick examination of our results with low thresholds showed that while preserving over 20 of the true positives, more than 50 percent of the true negatives could be discarded. The best result achieved was for session 2 of Speaker 2. For a specific parameter combination 24 true positives and 124 true negatives could be investigated. This is a reduction of the true negatives of over 60 percent. Defining a classifier for the remaining data could be something worth researching.

Both classifications showed that using only EMG channels 7 and 8 for classification is not enough to get good results. Our future work includes

reevaluating the results of this paper on EMG signals of *silent* speech. EMG channels 7 and 8 may become more important for silent speech EMG.

It should be stated that both classifiers still need some kind of labeling process providing the word boundaries even if used with EMG-based voice conversion. To get rid of the necessity of word boundaries is an issue for further research.

REFERENCES

Black, A., & Lenzo, K. (2000). Building voices in the Festival speech synthesis system. http://festvox.org/bsv/.

Chan, A., Englehart, K., Hudgins, B., & Lovely, D. (2001). Myoelectric Signals to Augment Speech Recognition. *Medical and Biological Engineering and Computing , 39*, 500-506.

Denby, B., Schultz, T., Honda, K., Hueber, T., & Gilbert, J. (2010). Silent Speech Interfaces. *Speech Communication , 52* (4), 270-287.

Dupont, S., & Luettin, J. (2000). Audio-Visual Speech Modeling for Continuous Speech Recognition. *IEEE Transactions on Multimedia , 2*, 141-151.

Eide, E., Aaron, A., Bakis, R., Hamza, W., Picheny, M., & Pitrelli, J. (2004). A corpus-based approach to <AHEM/> expressive speech synthesis. *5th ISCA ITRW on Speech Synthesis*, (pp. 79-84).

Ekman, P., & Frisen, W. (1978). The facial action coding system (FACS): A technique for the measurement of facial action. *Consultion Psychologists Press* .

Grice, M., & Baumann, S. Deutsche Intonation und GToBi.

Jou, S.-C., Schultz, T., Walliczek, M., Kraft, F., & Waibel, A. (2006). Towards Continuous Speech Recognition Using Surface Electromyography. *Proc. Interspeech*, (pp. 341-353).

Maier-Hein, L., Metze, F., Schultz, T., & Waibel, A. (2005). Session-independent non-audible speech recognition using surface electromyography. *Proc. ASRU*, (pp. 331-336).

Nakamura, K., Janke, M., Wand, M., & Schultz, T. (2011). Estimation of fundamental frequency from surface electromyographic data: EMG-to-F0. *Proc. ICASSP*, (pp. 573-576).

Pruzinec, M. (2010). Facial Expression Recognition using Surface Electromyography. *Diploma Thesis, Karlsruhe Institute of Technology* .

Schultz, T., & Wand, M. (2010). Modeling coarticulation in large-vocabulary EMG-based speech recognition. *Speech Communication , 52* (4), 341-353.

Silverman, K., Beckman, M., Pierrehumbert, J., Ostendorf, M., Wightman, C., Price, P., et al. (1992). ToBI: A Standard Scheme for Labeling Prosody. International Conference of Spoken Language. *Proc. of International Conference of Spoken Language*, (pp. 867-869).

Sotschek, J. (1984). Sätze für Sprachgütemessung und ihre phonologische Anpassung an die Deutsche Sprache. *Tagungsband DAGA: Fortschritte der Akustik*, (pp. 873-876).

Toth, A., Wand, M., & Schultz, T. (2009). Synthesizing Speech from Electroyography using Voice Transformation Techniques. *Proc. Interspeech.*

Walliczek, M., Kraft, F., Jou, S.-C., Schultz, T., & Waibel, A. (2006). Sub-Word Unit based Non-Audible Speech Recognition using Surface Electromyography. *Proc. Interspeech.*

Wand, M., & Schultz, T. (2011). Session-Independent EMG-based Speech Recognition. *Proc. Biosignals.*