Codebook Clustering for Unit Selection based EMG-to-Speech Conversion

Lorenz Diener, Matthias Janke, Tanja Schultz

Cognitive Systems Lab, Karlsruhe Institute of Technology (KIT)
Karlsruhe, Germany
matthias.janke@kit.edu

Abstract

This paper reports on our recent advances in using Unit Selection to directly synthesize speech from facial surface electromyographic (EMG) signals generated by movement of the articulatory muscles during speech production.

We achieve a robust Unit Selection mapping by using a more sophisticated unit codebook. This codebook is generated from a set of base units using a two stage unit clustering process. The units are first clustered based on the audio-, and afterwards on the EMG feature vectors they cover, and a new codebook is generated using these cluster assignments. We evaluate different cluster counts for both stages and revisit our evaluation of unit sizes in light of this clustering approach.

Our final system achieves a significantly better Mel-Cepstral distortion score than the Unit Selection based EMG-to-Speech conversion system from our previous work while, due to the reduced codebook size, taking less time to perform the conversion.

Index Terms: electromyography, silent speech interface, unit selection

1. Introduction

Silent Speech Interfaces [1] are systems that process speech, but do not rely on an audible acoustic signal to do so. Compared to audio-based speech interfaces, they have various advantages:

1. robustness in the presence of noise,
2. less or no disturbance of bystanders,
3. better protection of privacy and confidential information,
4. usable by speech-disabled persons (e.g. Laryngectomees).

Over the last few years, different modalities for Silent Speech Interfaces have been proposed (e.g. [2], [3], [4]). Our method of processing speech signals relies on surface electromyography (EMG) [2], where the activation potentials of the facial articulatory muscles emitted during speech production are recorded with surface electrodes. This approach works even when an acoustic speech signal is not actually present - as it only relies on the activity of the articulatory muscles, merely mouthing the words is sufficient.

Our approach is based on the *direct conversion* of EMG signals to speech [5], which has several advantages compared to recognition-based Silent Speech Interfaces (i.e. systems that recognize and then synthesize speech, as done in e.g. [2], [6]). In contrast to these, our system does not suffer from vocabulary restrictions and is able to retain paralinguistic information such as speaker mood, allowing for a more natural communication.

In [7], we initially proposed a direct mapping based on *Unit Selection* [8]. In that paper, the approach is to first build a codebook of small units containing segments of synchronously recorded EMG and audio features. To synthesize speech, the units whose EMG segments best fit the input EMG frames are selected from this database and an audio signal is reconstructed from these units audio segments. Figure 1 illustrates this mapping process.

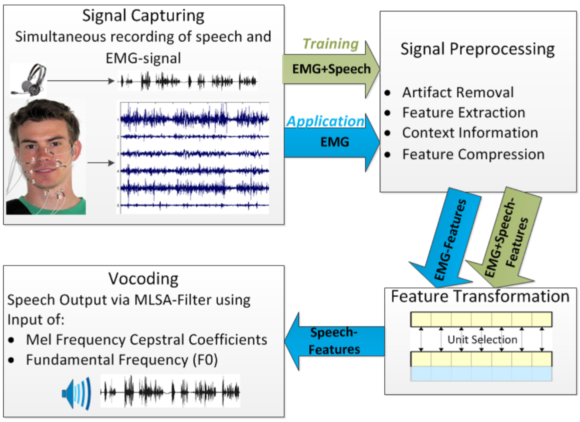

Figure 1: *Process of mapping from electromyographic input to speech output.*

In this paper, we introduce a method to create a codebook of units that make the Unit Selection process more robust. We perform k-means clustering of the units based on the segments of audio and EMG features contained therein and then build mean units based on the computed cluster assignments. This reduces the codebook size from a large number of base units to a low number of units that are more prototypical than the units they were created from. Due to that prototypicality, we expect this approach to allow for a more robust conversion that is not as sensitive to outliers, a common problem in Unit Selection based speech synthesis systems.

2. Data corpus information

To compare the clustering approach to our previous work, we selected the same recording sessions we already used in [7], which contain more than 500 utterances of EMG signals recorded during audible speech.

In total the corpus contains four recording sessions, with data from two male speakers. While the speakers are non-native, their English pronunciation skills range from good to very good.

For the recording of the EMG signals, we used two different types of setups: a *single electrode* setup and a novel *electrode array* setup. For the single electrode setup, we used a computer-controlled 6-channel EMG data acquisition system (Varioport, Becker-Meditec, Germany). We captured signals from 1) the levator anguli oris, 2) the zygomaticus major, 3) the platysma,

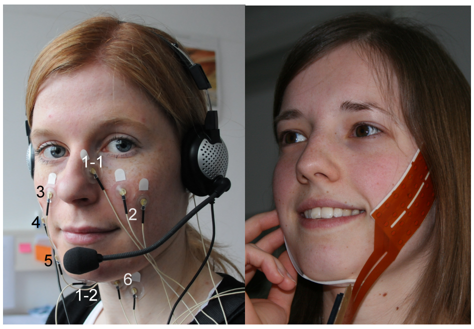

Figure 2: ***left****: Single electrode positioning, black numbers indicate unipolar derivation with reference electrodes behind the mastoid bone (except channel 1), white numbers indicate bipolar derivation.* ***right****: Electrode array positioning, one large array is positioned on the cheek, one small array under the chin. See text for details.*

4) the anterior belly of the digastric and 5) the tongue, see Figure 2 (left) for the electrode positioning. All EMG signals were sampled at 600 Hz and filtered with an analog 1 Hz high-pass filter. The electrode positioning which yielded optimal results was adopted from [9].

The electrode array acquisition device (EMG-USB2, OT Bioelettronica, Italy) recorded the EMG signals using a large electrode grid of four rows of eight electrodes each with 10 mm inter-electrode distance (IED) and a second smaller array with one row of eight electrodes with 5 mm IED. As illustrated in Figure 2 (right) the large array was placed on the subject's cheek - similar to the positioning of a cell phone - while the smaller one was positioned under the chin to ensure the recording of the tongue. The array signals were sampled at 2048 Hz, using a bipolar derivation, where the activation differences between two adjacent channels in a row are calculated. We therefore obtain a total of 35 signal channels out of the 4×8 cheek electrodes and the 8 chin electrodes [10].

In addition to the EMG signal, we simultaneously recorded the acoustic speech signal with a standard close-talking microphone at a sampling rate of 16 kHz. The audio signal is synchronized to the EMG signal using an additional analog marker channel.

The recorded text corpus is based on [11] and consists of phonetically balanced English sentences which originated from the broadcast news domain.

Each session was split into a *train* and *eval* set. The latter contains at least 10 different test sentences (plus repetitions), which are kept fixed across all sessions. For recording the data, the speaker read all prompted utterances in normal, audible speech in randomized order. This was supervised by a recording assistant to assure proper pronunciation and to guarantee a stable signal quality.

Table 1 lists the durations of the six recorded sessions and the number of utterances per session.

Since the EMG signal shows high inter-individual differences, we only build *session dependent* systems at this point.

3. Unit selection approach

3.1. Basic approach

Our Unit Selection approach attempts to convert EMG signals to audible speech by building a speech signal from short units of audio selected from a codebook according to the minimum cosine distance calculated on a set of EMG features covered by that same unit. To build such a codebook, we use a corpus of simultaneously recorded EMG and audio data. We extract sequences of feature frames from these raw signals (For details about the features used in this work, refer to section 4). We then extract overlapping windows of a certain length, the unit size w_u, from these sequences. This is done with a *unit shift* of $s_u = 1$ frame (i.e. with an overlap of $w_u - s_u$ frames) to get a large codebook. One pair of parallel EMG and audio segments make up a single unit $u = [u_{audio}, u_{emg}]$ in this codebook.

Session	Accumulated data length, in (mm:ss)		# of train/eval utterances	
	Train	Eval	Train	Eval
Spk1-Single	27:10	01:19	500	20
Spk2-Single	26:54	00:49	496	13
Spk1-Array	31:01	00:47	500	10
Spk2-Array	25:44	01:10	500	20
Total	110:49	04:05	1996	63

Table 1: *Data corpus information for the recorded utterances, including speaker/session breakdown.*

To perform EMG-to-speech conversion using such a codebook, the input EMG signal is similarly preprocessed: As above, a sequence of feature frames is extracted, and these are windowed into units of size w_u. This time, the unit shift is chosen according to experiment results from [7]. Given this input unit sequence i_t, units are selected from the unit codebook for each unit according to the minimal mean cosine distance between the input unit and codebook unit EMG frames to create an output unit sequence o_t. The mean cosine distance cd between two units a and b is defined as follows:

$$\mathrm{cd}(a,b) = \frac{1}{w_u} \sum_{f^a \in a_{emg}, f^b \in b_{emg}} \frac{f^a \cdot f^b}{\|f^a\| \, \|f^b\|}$$

The process of selecting units is simply an exhaustive search of the codebook for the unit with the minimal distance:

$$o_t = \underset{u_n \in codebook}{\arg\min} \; \mathrm{cd}(u_n, i_t)$$

To create an output audio frame sequence from this output unit sequence, the audio features $o_{t_{audio}}$ of overlapping frames of units from o_t are averaged to create a final output audio frame, as illustrated in Figure 3.

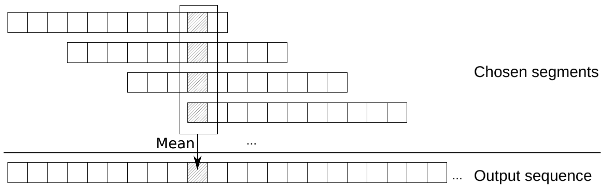

Figure 3: *Creating the output sequence from the chosen audio segments. Note that each box stands for an entire audio feature vector $g \in o_{t_{audio}}$.*

3.2. Unit clustering

The proposed Unit Selection approach taken in this work aims to improve the contents of the unit codebook to reduce audio artifacts resulting from the selection of wrong units. This is done by employing clustering to create units that are more representative of a single correspondence between EMG and audio signal than the units in the basic Unit Selection approach. This has the benefit of reducing sensitivity to outlier units, a single one of which can already greatly reduce intelligibility. Additionally, it eliminates redundancies in the codebook, which reduces computation time requirements for the conversion process.

A set of *base units* u_n is created in the same way as in the basic Unit Selection approach shown in section 3.1. These base units are then clustered in two stages, using the k-means algorithm. First, units are clustered according to the combined audio feature vector $[g \mid g \in u_{n_{audio}}]$ of all audio frames covered by each unit. Second, the units assigned to each audio cluster are clustered (separately for each cluster) according to the combined EMG feature vector $[f \mid f \in u_{n_{emg}}]$ covered by each of these units. Figure 4 illustrates this process.

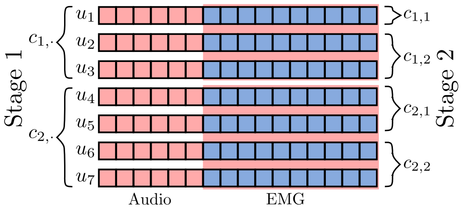

Figure 4: *The clustering process employed to create a more refined unit codebook, here exemplary for 7 units, clustering into two clusters in each stage. In stage 1, units are clustered according to audio features (red), in stage 2, they are clustered within the audio clusters according to EMG features (blue).*

Given these cluster assignments, a set of *cluster units* is created by taking calculating the mean audio and EMG feature frames over all units assigned to a cluster. These units are then used as the new codebook in the Unit Selection conversion process described above.

4. Experiment setup

4.1. Acoustic features

To represent audio data, we use a set of 25 Mel-Cepstral Coefficients (**MCEPs**) [12]. These are extracted for frames of 32ms length with a frame shift of 10ms. For the final speech synthesis, estimations of the fundamental frequency (F_0) are extracted from the reference audio for each frame.

For listening tests, we synthesize wave audio output from the MCEPs converted from the evaluation EMG signals and the extracted F_0s using Mel-Log Spectrum Approximation (MLSA) [13].

4.2. Electromyographic features

In our evaluation, we follow the approach used in [7] to allow for comparison of results and represent the EMG signal using a set of *time-domain features* [14]. For a given feature $\mathbf{f}$, $\bar{\mathbf{f}}$ is its frame-based time-domain mean. $\mathbf{P_f}$ is the corresponding frame-based power, and $\mathbf{z_f}$ is the frame-based zero-crossing rate.

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

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

The high-frequency signal is $p[n] = x[n] - w[n]$, and the rectified high-frequency signal is $r[n] = |p[n]|$.

Let $S(\mathbf{f}, n)$ denote the stacking of adjacent frames of the feature $\mathbf{f}$ in the size of $2n+1$ ($-n$ to n) frames, which is used in order to account for time-context information. With this, the EMG feature $\mathbf{TD15}$ is defined as follows:

$$\mathbf{TD15} = S(\mathbf{f2}, 15), \text{where } \mathbf{f2} = [\bar{\mathbf{w}}, \mathbf{P_w}, \mathbf{P_r}, \mathbf{z_p}, \bar{\mathbf{r}}].$$

We compute this feature for every channel in the EMG signal, with a frame size of 27ms and a frame shift of 10ms. We then fuse all EMG feature vectors for a frame into one large EMG feature vector. To reduce the dimensionality of this vector, we apply Linear Discriminant Analysis (LDA), trained on phoneme sub-state labels force-aligned to the parallel audio recording, before cutting the result to 32 dimensions, giving us our final EMG feature.

4.3. Experiment evaluation

To *objectively* evaluate our results, we employ the *Mel-Cepstral Distortion* (MCD) score [15], defined as a scaled Euclidean distance between MCEP vectors excluding the first coefficient. To reduce the effect of short misalignments on this measure, we perform a dynamic time warp alignment between reference MCEP sequence and evaluation MCEP sequence of each utterance before computing the mean MCD score over all frames.

The *subjective estimation* is evaluated using AB preference listening tests comparing the mapping output from our previous work [7] without codebook clustering to the synthesized speech from our proposed technique. Each participant listens to the original target audio file and compares the mapping outputs A and B to decide which one resembles the original speech. Each utterance is presented in randomized order. If no preference can be perceived, a third neutral option is available, so the listener is not forced to make a decision.

5. Experiment results

5.1. Cluster counts

Our clustering process has two free parameters: The number of clusters C_{aud} into which units are split according to MCEP features, and the number of clusters C_{emg} into which these clusters are sub-clustered using the post-LDA TD15 features. To determine good values for these parameters, we ran a series of experiments, varying both. This evaluation was performed with a unit size of w_u = 15 and a unit shift of s_u = 2, values which we have found to be good choices in previous experiments. The results of these cluster count experiments is shown in Figure 5. It can be seen that up to $C_{aud} = 2000$, using a larger number of audio units improves the MCD score. Improvements are large at first, with diminishing returns approaching the minimum. The effect of the EMG cluster count is comparatively minor - here, a lower number of clusters tends to be better. For the rest of our evaluation, we use cluster counts of $C_{aud} = 2000$ and $C_{emg} = 4$.

5.2. Unit size

In our previous work [7], we have found that shorter units and lower unit shift tend to improve Unit Selection performance. We therefore also performed the experiment described above with a reduced unit size ($w_u = 7, s_u = 1$). The results of this evaluation can be seen in Figure 5. The longer units perform slightly better than the short units for large C_{aud}, however, the difference is very small. For the rest of our evaluation, we decided to use long units.

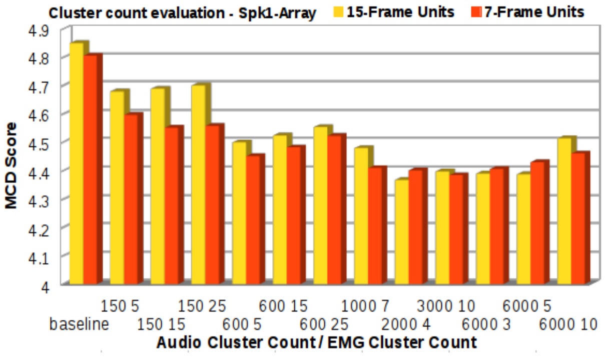

Figure 5: *MCD scores for Unit Selection performed with cluster units generated by clustering with different cluster counts. Lower is better.*

5.3. Output evaluation

Figure 6 shows spectrograms of the converted output with the proposed codebook clustering (bottom), as well as from the baseline Unit Selection approach (middle) [7] and the original target audio file (top) from the exemplary test utterance “The outages were apparently caused by system failure, not sabotage.”, taken from session *Spk2-Array*. The final output was synthesized using the MLSA filter, based on the converted MCEP output and the target F_0 information that was extracted from the synchronously recorded audio file. It can be seen that both conversion approaches show similar results, with a good general reconstruction but lacking in spectral details. Near seconds 1 and 3 (highlighted sections), phones are better reconstructed with the codebook clustering approach.

The baseline EMG-to-speech mapping system with the Unit Selection approach [7] achieves an average MCD of 5.17 on the four recording sessions. With our codebook clustering approach we obtain an average MCD of 4.71. This corresponds to a relative improvement of 8.92%.

Unit selection conversion requires, for every evaluation unit, the computation of the distance to every codebook unit (The effects of this multiplicative increase are obvious when comparing the conversion times of the two single-electrode sessions). Conversion time on the evaluation set (measured on a 4 x 2.66 GHz computer with 8GB RAM) improved, on average, from ca. 47.4 to 3 times realtime, an improvement of 93.7%, due to the reduction in codebook size from 119800 (Spk2-Array) —159987 (Spk1-Array) units down to 8000. Table 2 gives the MCDs and conversion time for each speaker/session.

Session	MCD Score		Time taken for conversion (mm:ss)	
	Baseline	Clustering	Baseline	Clustering
Spk1-Single	5.38	4.93	62:57	4:07
Spk2-Single	5.13	4.65	37:56	2:24
Spk1-Array	4.85	4.36	44:50	2:29
Spk2-Array	5.33	4.91	44:41	3:18

Table 2: *Mean evaluation set MCD scores and computation time for Unit Selection with base units versus* $C_{aud} = 2000$, $C_{emg} = 4$ *cluster units. Lower is better.*

Since objective MCD scores do not perfectly correspond to human acoustic perception, we perform an subjective AB preference listening test (adapted from [16]) between converted speech from the baseline Unit Selection based EMG-to-speech mapping [7] versus the proposed codebook clustering Unit Selection output. We also included a third neutral option, when no preference was perceived. Each participant listened to the target audio file and compares the two mapping outputs to decide which one is preferred. We randomly selected four utterances from the test set of each session, resulting in 16 utterances for the listening test, performed on 12 participants (total of 196 listened utterances). The codebook clustering approach was preferred in 100 (52.08%), the baseline system in 47 utterances (24.48%). Additionally, 45 times (23.44%) no clear preference could be made by the listener. This shows a strong preference to our proposed mapping technique.

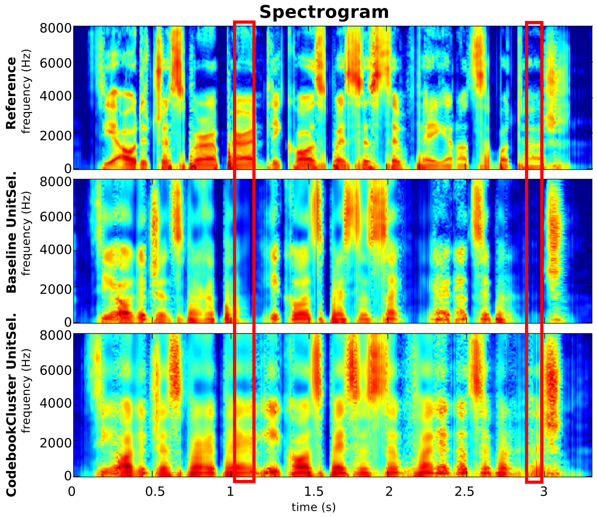

Figure 6: *Exemplary spectrograms of the reference target audio file, the baseline Unit Selection plus the proposed codebook clustering EMG-to-speech output (top to bottom) of the utterance: “The outages were apparently caused by system failure, not sabotage.”*

6. Conclusions and future work

We successfully introduced a codebook clustering method to substantially improve our Unit Selection based EMG-to-speech conversion, where surface electromyographic (EMG) signals of the articulatory muscles are transformed to audible speech. An objective evaluation shows a relative improvement of 8.92% compared to our previous work, yielding an average MCD of 4.71, while the subjective listening test evaluation also gives clear preference to the proposed technique. The proposed reduction of codebook size additionally gives a substantial speedup in conversion time.

In the future we plan to evaluate different kinds of EMG input features, since the proposed TD15 feature is highly optimized for EMG-based speech recognition, rather than for synthesis. We also consider to integrate label information in the codebook database.

7. Acknowledgements

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

8. References

[1] B. Denby, T. Schultz, K. Honda, T. Hueber, J. M. Gilbert, and J. S. Brumberg, “Silent speech interfaces,” *Speech Communication*, vol. 52, no. 4, pp. 270–287, 2010.

[2] A. Chan, K. Englehart, B. Hudgins, and D. Lovely, “Hidden markov model classification of myoelectric signals in speech,” *Engineering in Medicine and Biology Magazine*, vol. 21, no. 5, pp. 143–146, 2002.

[3] T. Toda and K. Shikano, “Nam-to-speech conversion with gaussian mixture models,” in *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*, 2005, pp. 1957–1960.

[4] M. Fagan, S. Ell, J. Gilbert, E. Sarrazin, and P. Chapman, “Development of a (silent) speech recognition system for patients following laryngectomy,” *Medical Engineering and Physics*, vol. 30, no. 4, pp. 419–425, 2008.

[5] A. R. Toth, M. Wand, and T. Schultz, “Synthesizing speech from electromyography using voice transformation techniques,” in *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*, 2009, pp. 652–655.

[6] S.-C. Jou, L. Maier-Hein, T. Schultz, and A. Waibel, “Articulatory feature classification using surface electromyography,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, 2006, pp. 605–608.

[7] M. Zahner, M. Janke, M. Wand, and T. Schultz, “Conversion from facial myoelectric signals to speech: A unit selection approach,” in *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*, 2014.

[8] A. J. Hunt and B. A. W, “Unit selection in a concatenative speech synthesis system using a large speech database,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1996, pp. 373–376.

[9] L. Maier-Hein, F. Metze, T. Schultz, and A. Waibel, “Session independent non-audible speech recognition using surface electromyography,” in *2005 IEEE Workshop on Automatic Speech Recognition and Understanding*, 2005, pp. 331–336.

[10] M. Wand, C. Schulte, M. Janke, and T. Schultz, “Array-based electromyographic silent speech interface,” in *International Conference on Bio-inspired Systems and Signal Processing (BIOSIGNALS)*, 2013, pp. 89–96.

[11] T. Schultz and M. Wand, “Modeling coarticulation in emg-based continuous speech recognition,” *Speech Communication*, vol. 52, no. 4, pp. 341–353, 2010.

[12] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, “An adaptive algorithm for mel-cepstral analysis of speech,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, 1992, pp. 137–140.

[13] S. Imai, “Cepstral analysis synthesis on the mel frequency scale,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 8, 1983, pp. 93–96.

[14] S.-C. S. Jou, T. Schultz, M. Walliczek, F. Kraft, and A. Waibel, “Towards continuous speech recognition using surface electromyography,” in *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*, 2006, pp. 573–576.

[15] R. F. Kubichek, “Mel-cepstral distance measure for objective speech quality assessment,” in *IEEE Pacific Rim Conference on Communications, Computers and Signal Processing*, 1993, pp. 125–128.

[16] S. Kraft and U. Zoelzer, “Beaqlejs: Html5 and javascript based framework for the subjective evaluation of audio quality,” in *Linux Audio Conference*, 2014.