Sub-Word Unit based Non-Audible Speech Recognition using Surface Electromyography

Matthias Walliczek, Florian Kraft, Szu-Chen Jou, Tanja Schultz, Alex Waibel

Interactive Systems Labs
Universität Karlsruhe (TH), Germany and Carnegie Mellon University, USA
{walliczek|fkraft|ahw}@ira.uka.de, {scjou|tanja}@cs.cmu.edu

Abstract

In this paper we present a novel approach for a surface electromyographic speech recognition system based on sub-word units. Rather than using full word models as integrated in our previous work we propose here smaller sub-word units as prerequisites for large vocabulary speech recognition. This allows the recognition of words not seen in the training set based on seen sub-word units. Therefore we report on experiments with syllables and phonemes as sub-word units. We also developed a new feature extraction method that gains significant improvement for words and sub-word units.

Index Terms: silent speech, non-audible speech recognition, electromyography, sub-word unit comparison

1. Introduction

In the last decades, automatic speech recognition (ASR) has evolved into a state where it is able to work well in several scenarios. However, there are still two side effects which constrain its success in some special scenarios: First it performs significantly worse when there is a lot of background noise present and second a speaker needs to produce sound, which might disturb other persons in certain environments. In these cases electromyographic (EMG) speech recognition might help: Such a recognition system processes electric signals caused by the articulatory muscles in order to recognize non-audible (silent) speech which means that no acoustic signal is produced.

As state of the art EMG speech recognition systems still work with whole words, all words that should be recognized have to be trained, which requires a huge amount of training data or drastically restricts the recognizable vocabulary. A well known solution is to split the words in smaller units that are part of more than one word, since this allows to recognize all combinations of these units. In acoustic speech recognition, a phoneme turned out to be most suitable as sub-word unit for most cases. Since EMG speech recognition is based on capturing the muscle movements and the configuration of the vocal tract this may lead to longer range context dependency. This is why we also explored syllable units as an option for EMG speech recognition.

2. Related Work

Research in EMG based speech recognition has increased over the last 5 years, however it's practical use is still limited to small domains due to the fact that only full word models have been successfully applied. One attempt to work with a bigger corpus was done by Jorgensen et al. [1]. He investigated the recognition of non-audible speech by using surface EMG electrodes placed on the larynx and sub lingual areas below the jaw. In addition to working with control words and digits he trained and recognized eighteen vowels and twenty-three consonant phonemes. He reached a recognition rate of 50% using a Neural Network classifier. In [1] phonemes were recorded isolated and independently. The subjects were asked to think of a 'target' reference word while subvocalizing only the vowel in the given context with the correct pronunciation.

The usage of syllables as phonetic units has been investigated in small and large vocabulary speech recognition. Ganapathirajua [2] achieved a gain of 20% by using an acoustic syllable-based system instead of a comparable triphone system for the OGI Alphadigits task. However, for a vocabulary of 70 000 words 9 023 syllables were needed.

3. Sub-word Units for EMG-ASR

The usage of syllables has advantages and disadvantages. On one hand, it is closely connected to human speech perception and articulation. Especially in the context of EMG speech recognition it may better take into account previous planning of muscle movements in the cerebellum and also the overlapping of muscle activities. Previous work showed that the muscular activities run ahead up to 50 ms before the acoustic signals [3]. These factors make it more difficult to determine the exact time for the beginning and the end of each phoneme within a syllable, which in contrast can be regarded as an enclosed unit. On the other hand the training data amount increases. When using a phoneme based system 45 phonemes have to be trained, while for a syllable based one up to 9 000 syllables are required for English. When a large vocabulary should be covered, there are many syllables which appear only in a small number of words.

As we found significant improvement when training on same words in the same session, it is not feasible to train the full set of syllables occurring in each test word. Nevertheless, for a restricted vocabulary there is a choice between context dependent phoneme models and syllables.

4. Methods

4.1. Data acquisition

For our study we tried to minimize the number of syllables and selected a vocabulary of 32 English expressions consisting of only 21 syllables: *all, alright, also, alter, always, center, early, earning, enter, entertaining, entry, envy, euro, gateways, leaning, li, liter, n,*

navy, right, rotating, row, sensor, sorted, sorting, so, tree, united, v, watergate, water, ways. Each syllable is part of at least two words so that the vocabulary could be split in two sets each consisting of the same set of syllables.

In each recording session, twenty instances of each vocabulary word and twenty instances of silence were recorded non-audible. Two subjects, S1 and S2 (one female and one male), with no known speech disorders participated in the study. We recorded five sessions for each speaker and got a total data amount of 6523 seconds, which was split into training and test sets as explained below.

The order of the words was randomly permuted and presented to the subjects one at a time. A push-to-talk button which was controlled by the subject was used to mark the beginning and the end of each utterance. Subjects were asked to begin speaking approximately 1sec after pressing the button and to release the button about 1sec after finishing the utterance. They were also asked to keep their mouth open before the beginning of speech, because otherwise the muscle movement pattern would be much different whether a phoneme occurs at the beginning or the middle of a word.

To identify the beginning of speech within the recording and thereby minimizing the effect of the varying delay between pushing the button and the beginning of speech, we used a pseudo-word silence. For the pseudo-word silence the speakers had to keep all facial muscles relaxed for approximately 2sec.

EMG signals were collected for both subjects using seven pairs of Ag/Ag-Cl electrodes. A self-adhesive button electrode placed on the left wrist served as a common reference. The detailed setup is explained in [4].

All electrode pairs were connected to a physiological data recording system [5]. EMG responses were differentially amplified, filtered by a 300 Hz low-pass and a 1 Hz high-pass filter and sampled at 600 Hz. In order to avoid loss of relevant information in the signals we did not apply a 50 Hz notch filter which can be used for the removal of line interference.

4.2. Feature extraction

4.2.1. Baseline system

As baseline system we used the configuration resulting from the work of Maier-Hein et al. [4]. The signal data for each utterance were transformed into a sequence of feature vectors. For each of the 7 channels, 18-dimensional channel feature vectors were extracted from 54 ms observation windows with 4 ms overlap. In order to obtain channel feature vector o_{ij} for channel j and observation window i the windowed Short Time Fourier Transform (STFT) was computed. Delta coefficients served as the first 17 coefficients of o_{ij}. The 18th coefficient was the mean of the time domain values for the given observation window. The resulting feature vector o_i for the observation window i is the concatenation of all channel feature vectors o_{ij}.

4.2.2. Smaller observation window sizes

To allow a better recognition of phonemes, we changed the window size from 54 ms to 27 ms, since the shortest phonemes last only about 40 ms and would be mixed with previous or following phonemes otherwise. This reduction of the window size resulted in less STFT coefficients: With a window size of 27 ms we got 9 STFT coefficients. By adding the time domain mean this sums up to 10 coefficients per feature vector.

4.2.3. Time domain context feature

To optimize the recognition rate for sub-word units, we added a new feature extraction method: Since the EMG signal depends on the previous and succeeding position of the muscles, we added two time domain context (TDC) coefficients, which are given by the difference between the mean of the time domain values in the current observation window and the mean of the observation window 40 ms before and after the current one. For a window size of 54 ms this results in 20 coefficients per channel and in 12 coefficients for 27 ms windows respectively.

4.3. Sub-word division

To analyze the recognition with sub-word units, we tested three unit types with and without context dependency: word units, syllables and phonemes were explored. For a fair comparison we used three states for each phoneme and each model unit. One of the 21 syllables is comprised by one phoneme and modeled by three states. Ten syllables with two phonemes each are modeled by six states and the remainder are models by 9 states. The maximum number of phonemes for word models was ten. The different sub-word division methods are shown in figure 1.

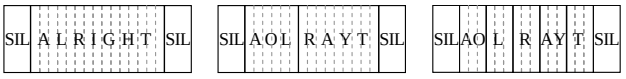

Figure 1: *The different sub-word division methods: word model (left hand side), syllabication (mid) and phoneme based division (right hand side); dashed lines correspond to HMM states.*

4.4. Model training

First order HMMs with Gaussian mixture models are used as classifiers as in most conventional ASR systems since they are able to cope with both, variance in the time-scale and variance in the shape of the observed data. Each training utterance was linearly segmented into a left-to-right Hidden Markov Model (HMM) starting with one silence state (SIL), followed by 3 states per phoneme and one succeeding silence state. A mixture of 14 Gaussians per state was initialized using kmeans. Afterwards the context independent (CI) models were trained using the Expectation Maximization (EM) algorithm. The number of iterations was chosen to be N = 4.

For the initialization of the context dependent (CD) triphone syllable and phoneme models, the labels generated by the CI models were used. These models were also trained using four iterations of the EM algorithm. Phoneme models were clustered (CL) to reduce the number of models and thereby model the context dependency in a more robust way. The clustered context dependent models were also initialized using the CI labels and final four EM iterations were applied.

4.5. Recognition

To recognize an utterance the corresponding sequence of feature vectors o_k was computed. Next, the Viterbi alignment for each vocabulary word W_j was determined and the word corresponding to the best Viterbi score was output as the hypothesis. Feature extraction, HMM training, and classification were performed using the Janus Recognition Toolkit (JRTk) [6].

5. Experiments and Results

5.1. Test on seen words

5.1.1. Optimized feature extraction methods

First the new feature extraction methods were tested. Therefore, all recordings of each word were split into two equal sets, one for training and the other for testing. This means that each word of the word list was trained on half of the recordings and tested on the other half. After testing sets were swapped for a second iteration. All combinations of the new feature extraction methods were tested, a window size of 54 ms and 27 ms, with and without time domain context feature.

We tested the different feature sets on a word recognizer, a recognizer based on syllables as well as phonemes.

	54 ms	27 ms	54 ms TDC	27 ms TDC
Words	64.8	78.1	71.2	**82.9**
Syllables (CI)	51.0	64.6	58.0	**73.3**
Syllables (CD)	57.1	71.2	63.6	**79.3**
Phonemes (CI)	57.3	57.7	65.3	**67.3**
Phonemes (CD)	64.0	62.9	71.7	**72.6**
Phonemes (CL)	68.5	72.4	75.5	**79.8**

Table 1: *Average word accuracies (in %) for round-robin testing.*

The results in table 1 show that both changes (4.2.2 and 4.2.3) of the feature set cause a significant improvement in all cases. The largest improvement ranging from 6.5 to 9.6 percent points gain in word accuracy was caused by the new time domain context feature. This stresses the impact of the time domain for EMG signal preprocessing in general and with respect to context dependency in particular. We found in our experiments that a smaller window size causes a notable improvement in nearly all cases, especially for word units and syllable units (up to 15.7 % abs.).

Since the systems with a window size of 27 ms and usage of the TDC feature performed best, these features were used in all following experiments.

5.1.2. Sub-word unit comparison

In table 2 the results for the optimized system are shown per speaker to compare the different sub-word division methods.

Because the phonemes occur more often in the words of the training set (avg. 60.43 training instances per model) than the syllables (avg. 27.62 training instances per model), at the first part of this experiment (Phonemes A) the number of recordings that were used for training of the phoneme models was reduced so that the phoneme models got the same number of training examples as the syllable models.

The second part of the experiment (Phonemes B) focused more on an application based point of view considering the limited amount of training data caused by session dependency. In this experiment both model types were trained with the same number of recordings, so that phoneme models could perform better because of the higher number of training instances.

We discriminate two context dependent systems, CD and CL. CL is the clustered system where each context dependent unit has its own codebook and mixture weights, while the CD system shares the codebooks of the context independent models. We give

	cbs	dss	avg. (min / max)
Words	427	427	82.9 (63.0 / 95.0)
Syllables (CI)	154	154	73.3 (58.3 / 87.5)
Syllables (CD)	154	545	79.3 (62.2 / 92.2)
Phonemes (CI, A)	137	137	60.5 (41.9 / 78.7)
Phonemes (CD, A)	137	541	66.3 (49.7 / 83.9)
Phonemes (CL, A)	311	311	69.1 (48.3 / 87.3)
Phonemes (CI, B)	137	137	67.3 (49.9 / 79.7)
Phonemes (CD, B)	137	541	72.6 (54.3 / 85.9)
Phonemes (CL, B)	311	311	79.8 (58.4 / 93.3)

Table 2: *Averaged number of codebook set entries (cbs) and distribution set entries (dss) and word accuracies (in %) for round-robin testing.*

numbers for the phoneme CD system to compare with the numbers of the syllable CD system.

The results in table 2 show that our context dependent phoneme based EMG speech recognition system works nearly as good as a word based system and slightly better than the syllable models when using all training data. But it should be considered that we only needed to train 21 syllables because of the special word list. That is why we only got 154 codebook set entries for the syllable system - if we had used a larger word list the number of syllables and their codebook set entries would have been much higher in comparison to the number of phonemes. For a fair comparison between phoneme and syllable units, where both units got the same number of training instances per model, syllables work better because they can model a larger context.

5.2. Test on unseen words

While in the previous tests seen words were recognized, we test in this section on words that have not been seen in the training (unseen words). Therefore, the vocabulary was split into two disjoint sets, one training and one test set. The words in the test set consist of the same syllables as the words in the training set, so that all phonemes and syllables could be trained.

For an acoustic speech recognition system training of phonemes allow the recognition of all combinations of these phonemes and so the recognition of all words consisting of these combinations. This test investigates whether EMG speech recognition performs well for context sizes used in ASR or whether the context is much more important and goes beyond triphones. To do so we tested both a phoneme based system and a syllable based system. While the syllable based system covers a larger context, the phoneme based system can obtain more training data per phoneme.

The optimized feature set (27 ms window size with time domain context feature) was used for this test.

When recognizing unseen words, it turned out that the syllable-based system does not improve when using context dependency in the cross test, while it does improve in the previous test. That is because, now that the test words do not occur in the training set and we hence only trained half the number of words, we used the same number of training recordings but these recordings covered less context variability. That is also the reason why the number of codebook set entries at the context dependent systems is lower than at the previous test (table 3).

Another difference to the previous test is that the phoneme

	alright	also	alter	always	center	early	enter	entertaining	entry	envy	euro	leaning	liter	navy	sorted	watergate	PERFORMANCE
alright	**60.5**	0	0	**38.0**	0	0	0	0	0	0	0	0	0	0.5	0.5	0.5	**60.5**
also	6.5	**75.5**	5.5	3.0	0	1.0	0	0	2.0	0.5	4.0	0	0.5	0.5	0.5	0.5	**75.5**
alter	4.5	5.0	**55.0**	2.0	5.5	6.0	9.0	2.0	1.0	0	0	0.5	3.0	0.5	2.0	4.0	**55.0**
always	0.5	0	0	**98.5**	0	0	0	0	0	0	0	0.5	0	0.5	0	0	**98.5**
center	3.0	1.5	1.0	5.0	**62.0**	1.0	1.5	2.0	1.0	0	1.0	9.0	3.5	0	8.0	0.5	**62.0**
early	6.5	3.5	**22.5**	13.5	0	**32.5**	0.5	4.5	4.5	0	1.0	3.5	2.5	0.5	1.5	3.0	**32.5**
enter	0.5	3.5	5.0	7.0	16.0	1.0	**41.5**	3.5	7.0	2.5	0.5	4.5	3.0	0.5	4.0	0	**41.5**
entertaining	2.0	0.5	0	1.0	0.5	0	1.5	**67.5**	1.0	0	0	**21.5**	0.5	0	0	4.0	**67.5**
entry	1.5	0.5	0	1.0	0.5	0.5	0	3.5	**86.5**	2.5	0.5	1.0	1.0	0.5	0	0.5	**86.5**
envy	2.5	0.5	0	8.0	0	0.5	0	0	9.5	**70.0**	1.0	0	0	7.0	0	1.0	**70.0**
euro	1.0	3.5	0	2.0	0	0	0	0.5	1.5	0	**89.0**	0	0	0	2.5	0	**89.0**
leaning	0.5	1.0	0	4.0	0	0.5	0	9.0	0.5	0	0	**81.0**	3.0	0.5	0	0	**81.0**
liter	4.5	5.5	15.5	2.0	6.0	3.0	0	0.5	3.5	1.5	4.0	18.5	**29.0**	0	6.5	0	**29.0**
navy	1.0	1.0	0.5	11.5	0	1.5	0	1.5	10.5	**35.5**	0.5	1.5	0.5	**32.0**	1.0	1.5	**32.0**
sorted	5.0	0	0.5	9.0	3.5	0.5	0	1.0	3.0	0	0	3.0	0.5	0.5	**72.0**	1.5	**72.0**
watergate	11.5	0	0	**39.0**	0	0	0	0.5	0	0	0.5	2.5	0	0	0.5	**45.5**	**45.5**
TOTAL	**7.0**	**6.3**	**6.6**	**15.3**	**5.9**	**3.0**	**3.4**	**6.0**	**8.2**	**7.0**	**6.4**	**9.2**	**2.9**	**2.7**	**6.2**	**3.9**	**62.4**

Table 4: *The combined results of the unseen words tests for clustered phoneme models, given in word accuracies (in %). On the left side the references are listed, on top the hypothesis. The bottom line shows the normalized occurrence rate (in %).*

	cbs	dss	avg. (min / max)
Syllables (CI)	154	154	54.1 (48.4 / 67.5)
Syllables (CD)	154	493	55.1 (48.1 / 66.3)
Phonemes (CI)	137	137	56.9 (44.7 / 70.3)
Phonemes (CD)	137	441	58.8 (43.8 / 68.8)
Phonemes (CL)	246	246	62.4 (51.3 / 70.6)

Table 3: *Averaged number of codebook set entries (cbs) and distribution set entries (dss) and word accuracies (in %) for testing of unseen words.*

based system performs better than the syllable based system even at the non clustered states. This may also be caused by the smaller number of words in the training set. Since several syllables occur only in one word, it is difficult to segment these syllables.

At the confusion matrix in table 4 it is noticeable that some words have bad recognition rates, i.e. *navy*. From the mapping between phonemes and muscle movements we derived that the muscle movement pattern for vocalizing the words *navy* and *envy* are quite similar (except the movement of the tongue, which is only barely detected using our setup). So the word *envy* is often falsely recognized as the word *navy*.

6. Conclusions and Future Work

Our tests show that signal processing for EMG speech recognition can be improved by paying more regard to time domain values. The tests also show that it is possible to build an EMG speech recognizer based on sub-word units. Therefore syllables are the better choice when using the same number of training instances per model, because they come closer to the EMG speech characteristic. However, syllables need more training data and when using the same number of training recordings, phoneme models perform slightly better. The second test shows that it is also possible to recognize unseen words with a recognition rate of 62.4%.

Since the time domain values are quite important for recognition, we expect further improvement by using better processing methods like in [7].

7. Acknowledgments

The authors wish to thank Peter Osztotics and Irene Hassinger for their valuable contributions to this study.

8. References

[1] C. Jorgensen and K. Binsted, *"Web Browser Control Using EMG Based Sub Vocal Speech Recognition"*, in Proc. of the 38th Hawaii International Conference on System Sciences, 2005.

[2] A. Ganapathiraju, J. Hamaker, J. Picone, M. Ordowski and G. R. Doddington, *"Syllable-Based Large Vocabulary Continuous Speech Recognition"*, in IEEE Transactions on speech and audio processing, Vol. 9, No. 4, May 2001.

[3] S.-C. Jou, L. Maier-Hein, T. Schultz and A. Waibel, *"Articulatory feature classification using surface electromyographie"*, in Proc. ICASSP 06, Toulouse; France, 2006, IEEE.

[4] L. Maier-Hein, F. Metze, T. Schultz and A. Waibel, *"Session independent non-audible speech recognition using surface electromyography"*, in Proc. ASRU, Costa Rica, Nov 2005.

[5] K. Becker, *"Varioport"*, http://www.becker-meditec.de.

[6] M. Finke, P. Geutner, H. Hild, T. Kemp, K. Ries and M. Westphal, *"The Karlsruhe Verbmobil Speech Recognition Engine"*, in Proc. ICASSP 97, München; Germany, 1997, IEEE.

[7] S.-C. Jou, T. Schultz, M. Walliczek, F. Kraft and A. Waibel, *"Towards Continuous Speech Recognition Using Surface Electromyography"*, in Proc. ICSLP 06, Pittsburgh; USA, 2006.