Wavelet-based Front-End for Electromyographic Speech Recognition

Michael Wand, Szu-Chen Stan Jou, Tanja Schultz

International Center for Advanced Communication Technologies
Carnegie Mellon University, USA and Universität Karlsruhe, Germany
michael.wand@stud.uni-karlsruhe.de

Abstract

In this paper we present our investigations on the potential of wavelet-based preprocessing for surface electromyographic speech recognition. We implemented several variants of the Discrete Wavelet Transform and applied them to electromyographical data. First we examined different transforms with various filters and decomposition levels and found that the Redundant Discrete Wavelet Transform performs the best among all tested wavelet transforms. Furthermore, we compared the best wavelet transform to our EMG optimized spectral- and time-domain features. The results showed that the best wavelet transform slightly outperforms the optimized features with 30.9% word error rate compared to 32% for the optimized EMG spectral and time-domain features. Both numbers were achieved on a 108 word vocabulary test set using phone based acoustic models trained on continuously spoken speech captured by EMG.
Index Terms: Electromyography, Wavelets, Speech Recognition, Preprocessing

1. Introduction

Recognition of spoken language provides a natural way for humans to communicate with computers. While speech recognition based on acoustic speech has advanced substantially, this technology is limited to audible speech of reasonable SNR ratio. However, real-life situations often require the recognition of spoken speech even when the environment is extremely noisy or when the speech is not audible at all. This includes noisy and crowded environments as well as underwater or space operations. Non-audible speech is favored for the sake of privacy, for example when making a confidential phone call in public spaces. Last but not least, alternative input methods for speech recognition may be useful for patients with medical speech impairments.

One obvious way of achieving non-acoustic speech recognition is monitoring the physical process which creates an audible signal, namely the movements of the articulatory apparatus of the speaker, indicated by the activity of the facial muscles. To capture the activity, electrodes are attached to the speaker's face. This process is known as *surface electromyography*, for simplicity we abbreviate this in this paper to *electromyography (EMG)*. EMG recognition is possible on both audible and non-audible speech.

Recently, some successful attempts have been made to perform speech recognition on EMG data [1, 2]. Most research in EMG-based speech recognition is limited to a very small vocabulary and the recognition of isolated words. However, in [3] we presented first results on a phone-based EMG recognizer. The initial front-end of this recognizer was based on classical spectral features that showed limited success. After substantial engineering, we found a set of optimized EMG features, which decreased the word error rate to 32% on a 108 word vocabulary (see below).

The wavelet transform in all its variants has become a widely-used tool for signal processing and has been applied successfully to EMG recognition of isolated words [2]. In this paper we investigate the potential of wavelet transforms to EMG recognition of continuous speech. We implement various transforms and compare these wavelet-based features to our optimized spectral and temporal features.

2. Experimental Setup

2.1. Data Acquisition

EMG signals vary a lot across different recording sessions, even with the same speaker. In order to assure a controlled configuration of this research, in this paper we report results collected with the same data set that we used in [3]. In those experiments, data was collected from one male speaker in one session. The speaker read English sentences in normal audible speech, which were simultaneously recorded by an EMG recorder and a standard close-talking microphone. The signal borders were marked by the speaker pressing a start/stop button.

The corpus consisted of 38 phonetically balanced sentences for training and 12 sentences from news articles for testing. Each sentence was read 10 times, thus yielding a training set with a total duration of 45.9 minutes and a test set with a total duration of 10.6 minutes. Additionally, 10 special "silence" utterances were recorded.

The EMG signals were recorded with six pairs of Ag/Ag-CL electrodes attached to the speaker's skin capturing the signal of the articulatory muscles, namely the *levator angulis oris*, the *zygomaticus major*, the *platysma*, the *orbicularis oris*, the *anterior belly* of the *digastric* and the *tongue*. Eventually, the signal obtained from the *orbicularis oris* proved unstable and was dropped from the final experiments. The EMG signals were sampled at 600 Hz and filtered by a 300 Hz low-pass and a 1 Hz high-pass filter.

Details regarding the data acquisition setup can be found in [3].

2.2. The Audible Speech Recognizer

In order to forced-align the audible speech recordings, we used a Broadcast News (BN) speech recognizer trained with the Janus Recognition Toolkit (JRTk). The recognizer is HMM-based, and makes use of quintphones with 6000 distributions sharing 2000 codebooks. The baseline performance of this system is 10.2% WER on the official BN test set (Hub4e98 set 1), F0 condition.

2.3. The EMG Recognizer

The EMG speech recognizer used an HMM-based decoding algorithm. Since the training set is very small, we restricted the algorithm to a context-independent acoustic model. Furthermore, the decoding vocabulary was restricted to the words appearing in the test set, which contained 108 words in total. The training set contained 415 words, 35 of which were also part of the test set.

The EMG speech recognizer was bootstrapped with the help of the recordings of audible speech data. First of all, the forced-align labels of the audible speech data were generated with the BN speech recognizer mentioned above. This information was used as an initial labeling of the EMG data.

In our previous work [4], we showed that the EMG signal is ahead of the audio signal since the speech signal is a product of articulator movements and source excitation. We modeled this *anticipatory effect* by adding a delay of 0 ms to 90 ms to the EMG signal. Our experimental results are charted as a function of this delay.

3. Preprocessing Methods

In this section we first summarize the special optimized EMG features and then introduce the wavelet-based features. For details on the optimized EMG features and its performances the reader is refered to [3].

3.1. Spectral Features

Some of our initial experiments showed that the traditional spectral plus time-domain mean feature is very noisy. In particular, purely spectral preprocessing incurs a very high Word Error Rate (WER). Nonetheless, we compare our results with those obtained with a *Windowed (Short-Time) Fourier Transform (STFT)* during which the output frequencies were quantized into 9 separate subbands.

3.2. Special EMG Features

In order to extract features from EMG signals in a more robust manner, we designed special EMG features whose main properties are a better normalization and smoothing of the input signal [3].

The motivation behind this special feature design is that EMG signal is very different from speech acoustic signal. In a speech acoustic spectrogram, we can usually observe distinguishable phone characteristics, which is not the case in an EMG spectrogram. As a result, traditional spectral feature extraction does not work well for EMG signals. To solve this feature problem, we design this special EMG feature in order to reduce feature dimension while keeping the most useful information per frame. The reduced dimension also make it possible to stack feature with a wider context, which is beneficial for modeling long-range dynamics.

We use the following definitions: For any feature $\mathbf{f}$, $\bar{\mathbf{f}}$ 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. $S(\mathbf{f}, n)$ is the stacking of adjacent frames of feature $\mathbf{f}$ in the size of $2n+1$ ($-n$ to n) frames. In these computations, we used a frame size of 27 ms and a frame shift of 10 ms. These values are reported as giving optimal results by [5].

In our previous work, the best WER was obtained with the $\mathbf{E4}$ feature defined as:

$$\mathbf{E4} = S(\mathbf{f2}, 5), \text{where } \mathbf{f2} = [\bar{\mathbf{w}}, \mathbf{P_w}, \mathbf{P_r}, \mathbf{z_r}, \bar{\mathbf{r}}].$$

3.3. The Discrete Wavelet Transform

The *Discrete Wavelet Transform (DWT)* applies a set of linear time invariant (LTI) filters to the signal, thereby extracting certain properties of the original data. The main feature of the DWT is that it extracts *details and approximations* of the signal *on different scales*. This multi-scale analysis is achieved by either dilating the filters appropiately or downsampling the original signal, to which an adapted low-pass filter is applied before. The process of filtering the signal and changing the scale is repeated a fixed number of times. This number is called the *maximum decomposition level*.

There is a wide variety of filters which can be used for these transforms. Special consideration must be given to the choice of these filters since the results of the experiments vary considerably when the filter setup is changed. The applied filters are reported in the respective sections.

In each case, the basic high-pass filter which is used for extracting the details on a certain scale is a finite LTI filter of the form

$$H = (h_k)_{k \in \mathbb{Z}}. \tag{1}$$

This filter generates a corresponding low-pass filter G by the formula

$$G = (g_k)_{k \in \mathbb{Z}}, g_k = (-1)^k h_{1-k}. \tag{2}$$

This filter duality is a central property of all discrete wavelet-based algorithms.

3.4. The Redundant Discrete Wavelet Transform

The *Redundant Discrete Wavelet Transform (RDWT)* [6] is the most direct way of decomposing an input signal into different scales. The representation created is highly redundant, however it has got some desirable properties. In particular, this transformation is fully invariant towards small shifts of the input signal.

For an input signal $c^0 = (c^0_k)_{k \in \mathbb{Z}}$, we calculate the *detail coefficients* d^n and the *approximation coefficients* c^i, $i \geq 1$, by the following algorithm:

1. Start with the filters (h_k) and (g_k) as described above. Let L be the maximum decomposition level. Let $i := 1$.
2. Calculate c^i and d^i, $i \leq L$, by

$$\begin{aligned} c^i_l &= \sum_{k \in \mathbb{Z}} c^{i-1}_k h_{k-l} \\ d^i_l &= \sum_{k \in \mathbb{Z}} c^{i-1}_k g_{k-l} \end{aligned} \tag{3}$$

3. Upsample the filters (h_k) and (g_k), i. e.

$$\begin{aligned} h^{NEW}_k &= \begin{cases} h^{OLD}_{k/2} & \text{if } k \text{ is even} \\ 0 & \text{if } k \text{ is odd} \end{cases} \\ g^{NEW}_k &= \begin{cases} g^{OLD}_{k/2} & \text{if } k \text{ is even} \\ 0 & \text{if } k \text{ is odd} \end{cases} \end{aligned} \tag{4}$$

4. Jump to step 2 if $i < L$ and increment i by 1.

3.5. The Fast Wavelet Transform

The *Fast Wavelet Transform (FWT)* [7] is a variant of the RDWT which eliminates the redundancy of the output vectors. This is done by downsampling the output vectors by the factor two during each decomposition step. The algorithm avoids computing unused coefficients and is therefore very fast.

We define the operators $\tilde{H}$ and $\tilde{G}$, based on the original filters H and G, by

$$\begin{aligned} \tilde{H} : \ell_2 \longrightarrow \ell_2, (c_k) \mapsto \left(\sum_{m \in \mathbb{Z}} h_{m-2k} c_m \right) \\ \tilde{G} : \ell_2 \longrightarrow \ell_2, (c_k) \mapsto \left(\sum_{m \in \mathbb{Z}} g_{m-2k} c_m \right). \end{aligned} \tag{5}$$

Then for each $i > 0$, we can compute

$$c^i = \tilde{H}(c^{i-1}) \qquad \text{and} \qquad d^i = \tilde{G}(c^{i-1}) \tag{6}$$

until the desired maximum decomposition level is reached.

All information of the original signal is contained in the output vectors $\{(d^1), (d^2), \ldots, (d^L), (c^L)\}$, where L is the maximum decomposition level. However, for EMG recognition purposes, it is sometimes useful to consider *all* generated vectors.

3.6. The Double-Tree Complex Wavelet Transform

The major drawback of the FWT is that it is not robust with respect to small shifts of the input signal, i. e. such shifts may generate a very different output signal. The *Double-Tree Complex Wavelet Transform (DTCWT)* [8] was designed to overcome this problem without resorting to the highly-redundant RDWT. It achieves an approximate shift-invariance by computing two FWTs with phase-shifted filters in parallel. This creates two sets of detail/approximation coefficients $\left\{(c_A^i), (d_A^i), (c_B^i), (d_B^i) \,\middle|\, 1 \leq i \leq L\right\}$, which are considered as the real and imaginary parts of a wavelet transform with a *complex* wavelet. The final coefficients used in the classification process are then computed by:

$$\begin{aligned} (c^i)_k &= |(c_A^i)_k + j \cdot (c_B^i)_k| = \sqrt{(c_A^i)_k^2 + (c_B^i)_k^2} \\ (d^i)_k &= |(d_A^i)_k + j \cdot (d_B^i)_k| = \sqrt{(d_A^i)_k^2 + (d_B^i)_k^2} \end{aligned} \tag{7}$$

for each k, where $j = \sqrt{-1}$.

4. Experiments and Results

We used a two-step approach to evaluate the virtues of the DWT for EMG signal processing. In the first part, the transforms described above are compared to each other and to a classical STFT. In the second part, the strategies used in our previous work [3] to create special EMG features are applied to data which is first processed by a wavelet transform.

In all experiments, an LDA is applied to the final feature vectors, whose dimension is reduced to 32. The anticipatory effect of the EMG signals is modeled by delaying the speech signal for 0 ms - 90 ms.

4.1. Comparison of Plain DWT Features

We use the following features:

STFT Short-Time Fourier Transform with a window size of 27 ms and a frame shift of 10 ms. The resulting frequencies and quantized into 9 subbands.

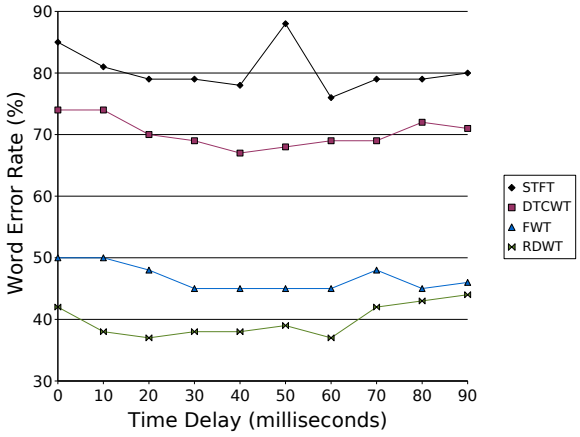
DTCWT DTCWT with a Daubechies-4 filter pair [9] for the first step and a 14-tap q-shift filter pair [10] for all subsequent steps.

FWT FWT with a 14-tap q-shift filter [10] for all steps.

RDWT RDWT with a 14-tap q-shift filter [10] for all steps.

In all wavelet experiments, the maximum decomposition level was 5. Both detail and approximation coefficients were used to generate the final feature vectors. Figure 1 shows that

Figure 1: Word Error Rate on Plain DWT Features



the RDWT performs best among the wavelet transforms investigated here. Clearly, due to the downsampling performed by both FWT and DTCWT, these tranforms fail to capture some essential information of the EMG signal. In particular, the localization of certain signal properties at lower frequencies may only be represented inaccurately.

A somewhat surprising result is the high WER when the DTCWT is used—in fact, even the FWT performs better than the DTCWT.

As expected, the STFT performs worse than all the DWT variants. In [3], we reported that time-domain smoothing and averaging of the EMG signal provides additional gain compared to purely spectral features. The fact that the approximation coefficients of the wavelet transforms contain such a time-domain smoothing of the input signal may explain the better performance of the DWT variants.

4.2. The Wavelet Transform and Special EMG Features

In our previous experiments in [3], a separation of low-frequency and high-frequency components of the signal is achieved by computing a nine-point double-averaged signal representing the low frequencies and then subtracting it from the original signal to get the high frequencies. These features are used as a base for calculating time-domain properties of the signal. The goal of this section is to investigate the effects of replacing this formula with the more elaborate multi-scale analysis performed by the wavelet transform.

For these experiments, we define three features:

X1 The signal is processed with a RDWT with a 14-tap q-shift filter [10] till level 5. Only the detail coefficients $\{(d^1), \ldots, (d^5)\}$ are used for further processing.

X2 The signal is processed with a RDWT with a 14-tap q-shift filter [10] till level 2. Detail and approximation coefficients $\{(d^1), (d^2), (c^1), (c^2)\}$ are used for further processing.

E4 The **E4** feature (see above).

For features **X1** and **X2**, the further steps are as follows: To each row r^i of detail and approximation coefficients (the latter

only in the case of $\mathbf{X2}$), we apply the transformation F defined by

$$F(r^i) = [P_{r_i}, \bar{r}_i, z_{r_i}].$$

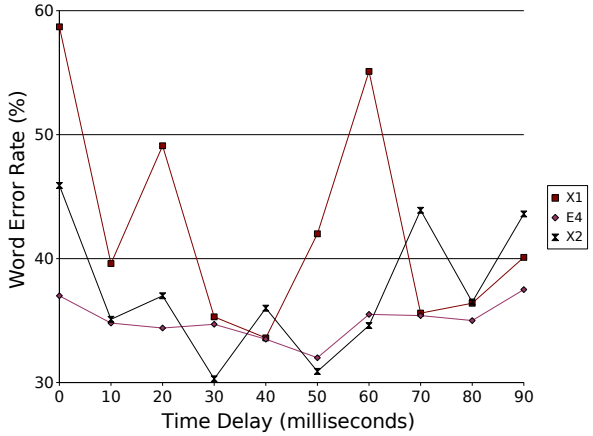
Thus we get the "preliminary" features $\tilde{\mathbf{X}}\mathbf{1}$ and $\tilde{\mathbf{X}}\mathbf{2}$ by

$$\tilde{\mathbf{X}}\mathbf{1} = [F(d^1), F(d^2), F(d^3), F(d^4), F(d^5)]$$
$$\tilde{\mathbf{X}}\mathbf{2} = [F(d^1), F(d^2), F(c^1), F(c^2)],$$

i. e. the processed rows are stacked upon each other. Finally, we obtain the features $\mathbf{X}\mathbf{1}$ or $\mathbf{X}\mathbf{2}$ by applying the Stacking filter:

$$\mathbf{X}\mathbf{1} = S(\tilde{\mathbf{X}}\mathbf{1}, 5)$$
$$\mathbf{X}\mathbf{2} = S(\tilde{\mathbf{X}}\mathbf{2}, 5).$$

Figure 2: Word Error Rate on Special EMG/Wavelet Features



We see that the feature $\mathbf{X}\mathbf{2}$ slightly improves the WER to a minimum of 30.9% compared to the $\mathbf{E}\mathbf{4}$ feature, whose optimal performance was 32%. While this is a relatively small improvement, the wavelet transform gives us an important means to *customize* the splitting of the EMG input signal into details on different scales. Using a specially adapted filter may be a lever to further increase the word accuracy, however choosing such a filter is beyond the scope of this paper.

$\mathbf{X}\mathbf{1}$ shows unstable results with respect to the time delay, however for certain delays its performance is comparable with $\mathbf{X}\mathbf{2}$. Note that for $\mathbf{X}\mathbf{1}$, we only used the detail coefficients of the RDWT, since the transformations contained in F perform a smoothing and averaging of the signal. Further experiments showed that adding the approximation coefficients to the feature $\mathbf{X}\mathbf{1}$ *significantly increases* the WER.

The Stacking filter S adds context information to the final feature. As we reported in [3], this context information is crucial to obtaining optimal recognition results. Experiments with other methods of adding context information showed comparable results. However, it also turned out that beyond a certain limit, adding more context information decreases the recognition accuracy. Therefore, the means of adding contextual information must be carefully chosen and optimized.

5. Conclusions

We investigated the potential of wavelet transforms for surface electromyographic speech recognition. Several variants of the Discrete Wavelet Transform were implemented and applied to electromyographical data. We found that the Redundant Discrete Wavelet Transform performs best among all tested wavelet transforms.

Furthermore, we used data processed by the Redundant Discrete Wavelet Transform as a base for the calculation of specially optimized time-domain features and achieved an improvement of the WER, which dropped from 32% to 30.9%.

Thus in this setup, the flexibility of the wavelet transform gives us a means of customizing the EMG preprocessing. For the future, we expect to further improve the feature extraction by using specially optimized filters.

6. References

[1] Lena Maier-Hein, Florian Metze, Tanja Schultz, and Alex Waibel. Session Independent Non-Audible Speech Recognition Using Surface Electromyography. In *Proc. ASRU*, 2005.

[2] Chuck Jorgensen and Kim Binsted. Web Browser Control Using EMG Based Sub Vocal Speech Recognition. In *Proceedings of the 38th Hawaii International Conference on System Sciences*, 2005.

[3] S.-C. Jou, T. Schultz, M. Walliczek, F. Kraft, and A. Waibel. Towards Continuous Speech Recognition using Surface Electromyography. In *Proc. Interspeech*, Pittsburgh, PA, September 2006.

[4] S.-C. Jou, L. Maier-Hein, T. Schultz, and A. Waibel. Articulatory Feature Classification using Surface Electromyography. In *Proc. ICASSP*, Toulouse, France, May 2006.

[5] M. Walliczek, F. Kraft, S.-C. Jou, T. Schultz, and A. Waibel. Sub-Word Unit based Non-audible Speech Recognition using Surface Electromyography. In *Proc. Interspeech*, Pittsburgh, PA, September 2006.

[6] M. J. Shensa. The Discrete Wavelet Transform: Wedding the À Trous and Mallat Algorithms. In *IEEE Transactions on Signal Processing*, vol. 40, pp. 2464 – 2482, October 1992.

[7] S. G. Mallat. A Theory for Multiresolution Signal Decomposition: The Wavelet Representation. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 7, pp. pp. 674–693, July 1989.

[8] I. W. Selesnick, R. G. Baraniuk, and N. G. Kingsbury. The Dual-Tree Complex Wavelet Transform. In *IEEE Signal Processing Magazine*, vol. 22, no. 6, pp. 123–151, November 2005.

[9] I. Daubechies. Orthonormal Bases of Compactly Supported Wavelets. In *Comm. Pure Appl. Math. 41*, 1988.

[10] Nick G. Kingsbury. A Dual-Tree Complex Wavelet Transform with Improved Orthogonality and Symmetry Properties. In *Proc. IEEE Conf. on Image Processing, Vancouver*, 2000.