Automatic English Stop Consonants Classification using Wavelet Analysis and Hidden Markov Models

Marco Kühne, Roberto Togneri

Centre for Intelligent Information Processing Systems (CIIPS)
School of Electrical, Electronic and Computer Engineering
The University of Western Australia
marco@ee.uwa.edu.au, roberto@ee.uwa.edu.au

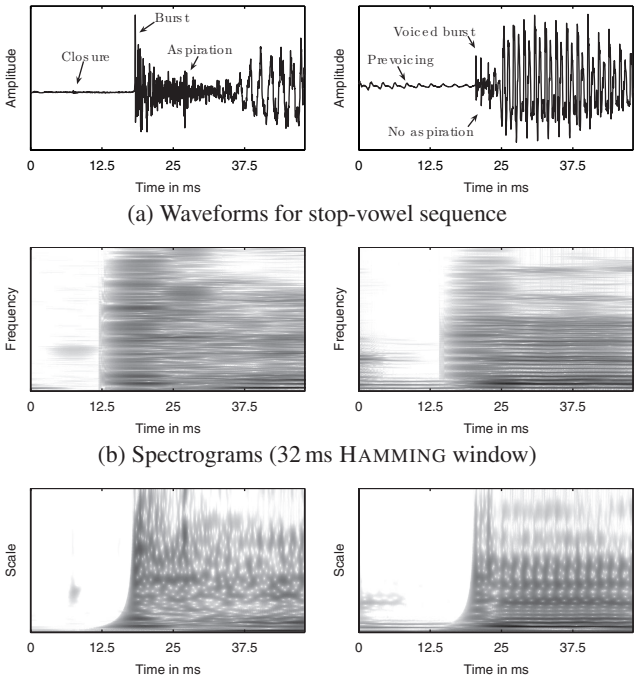
Abstract

This paper compares wavelet and STFT analysis for a speaker-independent stop classification task using the TIMIT database. In the designed experiment the HMM classifier had to assign each test token to one of the following stop classes [d,g,b,t,k,p,dx]. On 6332 stops the wavelet features obtained an overall accuracy of 86 % which corresponds to a 14 % relative error reduction compared to the STFT baseline system. Furthermore an analysis of the HMM misclassifications revealed that voiced stops were highly confused with their voiceless unaspirated counterparts.

Index Terms: speech recognition, wavelet analysis, Hidden Markov Models.

1. Introduction

The stop consonants can be characterized as highly non-stationary sounds with short durations usually articulated in the vocal tract in three diverse phases. These phases, which occur sequential in time, are denoted by silence, plosion and aspiration. A stop is denoted as voiced or voiceless depending on the state of the glottis. If the vocal cords are vibrating during the closure or even through the burst the stop is called voiced (e.g. [g],[d],[b]) otherwise it is called voiceless or unvoiced (e.g. [k],[t],[p]). In order to extract useful features for stop sounds an automatic speech recognition (ASR) system has to extract voicing information and formant transitions with sufficient high frequency resolution while on the other side very short events like the burst and closure points must be captured with high time resolution. The dominant signal analysis technique in current state-of-the-art speech recognizers is still based on the Short-Time-FOURIER-Transform (STFT) which uses a window function to analyze a signal. It is well known that the properties of the resulting local time-frequency analysis heavily depend on the chosen type and length of the window function. High frequency resolution can be achieved by long windows while time resolution can be increased by choosing shorter windows. Usually a HAMMING-window with duration between 20 ms - 40 ms is chosen as a compromise. It is often argued that the fix window size has serious consequences for the feature extraction of stops [1, 2]. Ideally, different speech sounds should be analyzed with different window lengths depending on the characteristics of the underlying signal. Exactly this kind of analysis is offered by the so-called Wavelet Transform (WT). In contrast to the STFT, WTs use long windows to measure low frequencies in the signal and short windows to capture high frequency components. This property of the wavelet transform makes it an ideal candidate for the analysis of stop consonants as long windows are useful for detecting voicing while the use of short windows allows to identify the place of articulation. Furthermore they have an infinite set of possible basis functions which are not limited to sine and cosine functions. Figure 1 shows an example of the different kind of analysis offered by WT and STFT.



(a) Waveforms for stop-vowel sequence

(b) Spectrograms (32 ms HAMMING window)

(c) Scalograms (MORLET wavelet)

Figure 1: Comparison of FOURIER and wavelet analysis for the unvoiced stop [t] (left) and its voiced counterpart [d] (right).

Wavelets have previously been used for stop classification and the research community is in agreement that wavelet transforms offer clear theoretical advantages over STFT based feature extraction methods [2, 3]. In contrast, the experimental results did not reveal a clear superiority of wavelet features over STFT methods. Summarizing the reported results in the literature a non-conclusive situation about the use of wavelet transforms for automatic stop

classification was found [2, 3, 4, 5]. We believe this is mainly due to the use of different types of wavelet transforms, classifiers and databases. Even as most of the work was carried out on the TIMIT database, often only parts of the available data were used for the experiments.

This paper compares wavelet and STFT analysis for a speaker-independent stop classification task. Unlike in previous work stops were extracted from continuous speech of the TIMIT database using all available speakers from eight dialect regions. A standard Hidden MARKOV Model (HMM) paradigm was employed for classification. The used wavelet transform was kept quite similar to the STFT to isolate the effect of the different window size by simulating the mel-frequency scale in the filter spacing and by choosing an appropriate mother wavelet. Wavelet transforms have the potential to improve the stop recognition without sacrificing the performance of other phonemes. The described system is also applicable for other phoneme classes than stops as it basically follows the standard feature extraction used in speech recognition. It is therefore interesting to compare the achieved results with knowledge based systems specifically tuned for stop classification.

2. Wavelet analysis and feature extraction

This section introduces the wavelet feature extraction for English stop classification shown in the Figure 2.

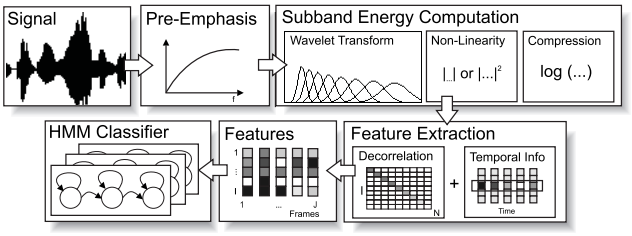


Figure 2: Flowchart of the involved signal processing steps for the stop classification experiments based on the mel-scaled wavelet filterbank.

2.1. Subband Energy Computation

Wavelet coefficients are obtained by correlating the signal with translated and dilated versions of a mother wavelet. This process can be considered as a filterbank analysis if we interpret the wavelet as the impulse response of a bandpass filter. In opposite to the STFT which is a constant bandwidth analysis the wavelet transform performs a so-called constant-Q analysis where the ratio of center frequency and bandwidth remains constant for all filters. This type of filterbank analysis is more appropriate than a constant bandwidth filterbank as it directly simulates the non-linear frequency perception in human sound processing.

In practice, the scaling and translation parameters as well as the speech signal itself have to be sampled. Hence, the realizable Continuous-Wavelet-Transform (CWT) of a signal x is defined by:

$$\mathfrak{W}(s,n) = \frac{1}{\sqrt{s}} \sum_{k=-\infty}^{\infty} x(k)\, \psi^* \left(\frac{k-n}{s} \right) \tag{1}$$

where s denotes the scaling parameter, n is the translation parameter and ψ^* is the complex conjugate wavelet function. The CWT was used because it produces frame synchronous coefficient vectors directly applicable for HMMs and allows an easy simulation of the mel-frequency scale by means of an appropriate sampling of the scale parameter. The scale parameter s can be related to the linear center pseudo-frequency f_{c_i} via:

$$s_i^{mel} = \frac{f_s}{f_{c_i}} F_c = \frac{f_s}{700 \cdot \left(10^{\mathfrak{f}_{c_i}/2595} - 1\right)} F_c \tag{2}$$

where f_s is the sampling frequency of the speech signal. F_c denotes the normalized center frequency of the mother wavelet and $\mathfrak{f}_{c_i}$ gives the equally spaced center frequencies

$$\mathfrak{f}_{c_i} = \mathfrak{f}_l + i \cdot \frac{\mathfrak{f}_h - \mathfrak{f}_l}{I+1} \tag{3}$$

along the perceptional mel-frequency scale:

$$\mathfrak{f}(f) = 2595 \cdot \log_{10}\left(1 + \frac{f}{700}\right) \tag{4}$$

where I is the total number of filters and $\mathfrak{f}_l, \mathfrak{f}_h$ are the lower and higher mel-frequency cut-offs of the entire filterbank. Here we employed a modulated Gaussian function, called the complex MORLET wavelet:

$$\psi(t) = \frac{1}{\sqrt{\pi F_b}} \cdot e^{-\frac{t^2}{F_b}} \cdot e^{2\pi i F_c t} \tag{5}$$

with bandwidth parameter F_b and center frequency parameter F_c [6]. Both parameters determine the characteristics of the initial wavelet filter which is then scaled to produce a set of filters to cover the frequency spectrum of interest. The F_c parameter defines the number of oscillations of the complex exponential while the bandwidth parameter controls the decay of the Gaussian modulation window. The constant-Q factor of the complex MORLET wavelet is given by:

$$Q(F_b, F_c) = \sqrt{\frac{\pi^2 F_b F_c^2}{2\ln(2)}} \tag{6}$$

The MORLET wavelet was chosen because it provides a good time-frequency resolution and it ensures that one can purely measure the effect of the varying window size as both STFT and WT transform use a (complex) sine-cosine basis.

2.2. Feature Extraction

To remove the linear correlation between the individual subbands a Principal Component Analysis (PCA) was applied to the logarithmic compressed wavelet subband energy vectors. The PCA is a statistical analysis tool which can be used to find a lower dimensional subspace whose orthogonal basis vectors correspond to the direction of the greatest variance in the original space. The PCA is the best linear transform in terms of decorrelation efficency and energy compaction.

It is well known that speech recognition systems purely based on static features fail to model the evolution of speech features over time. Especially for stop sounds it is important to model the feature trajectories as the discriminant information is mainly encoded in the transitions from the closure phase to burst and aspiration. The standard method for integrating dynamic information is to calculate the so-called delta coefficients [7] which approximate the time-derivatives of the feature trajectories. Usually the first and second order derivatives are appended to the static feature set.

3. Classification experiments

3.1. Database

Classification experiments were carried out on the TIMIT database using all available speakers and dialect regions. In particular, 17998 stops from 462 different speakers (326 male/136 female) were used for training while 6332 stops spoken by 168 different speakers (112 male/56 female) were extracted for evaluating the classifier's performance. Stop closures marked in the TIMIT data (e.g. [pcl]) were merged with the release label (e.g. [p]) to represent the stop as one segment (e.g. [p]). The individual composition of test and training set can be seen from Table 1. Besides the voiced and voiceless stops the flap [dx] was extracted from the TIMIT data as done in [8].

Table 1: Composition of stops in training and test set using the TIMIT database

Set	Voiced Stop			Voiceless Stop			Flap
	d	g	b	t	k	p	dx
Training	2432	1191	2181	3948	3794	2588	1864
Test	840	452	879	1367	1204	956	634

3.2. Model training and evaluation method

3.2.1. Training phase

The Hidden Markov Toolkit (HTK) toolkit was configured as a classifier [7]. The HMM topology followed standard left-to-right models without skips using nine emitting states. The observations were modeled by continuous Gaussian mixture probability density functions with diagonal covariance matrices. The HMM models were initialized by the HTK tools HCompV and HInit. Then the initialized models were trained using BAUM-WELCH re-estimation by means of the HTK tool HRest. Finally the model set was refined through mixture incrementing and several passes of BAUM-WELCH re-estimation. At this point new Gaussian mixture components were added one at a time followed by three rounds of BAUM-WELCH re-estimation. We used a maximum number of 8 mixtures per state.

3.2.2. Test phase

The classification accuracy was evaluated using the HTK tool HResults and the output of the HVite decoder. The percentage number of correctly classified stops was chosen to measure the performance.

To ensure that the observed performance differences between wavelet and baseline system are statistically significant MCNEMAR'S test was applied as it was suggested in [9]. The test requires that the errors made by an algorithm are independent, which is a valid assumption for isolated stop classification. Using the Normal distribution approximation the test was performed for different significance levels α. The lower the value of α the more the observed performance differences must diverge to be accepted as significant.

3.3. Experiments

In the designed experiment the classifier had to assign each test token to one of the following stop classes [d,g,b,t,k,p,dx]. The results were recorded for both the STFT baseline system consisting of the HTK-FBANK features followed by a PCA for decorrelation and the wavelet features as described in Section 2. Both filterbanks were configured using the following parameters:

- Number of channels: 24
- Frequency range: 125 Hz - 8 kHz
- Spacing of center frequencies: mel scale
- Number of static PCA coefficients: 15
- Pre-emphasis coefficient: 0.97
- Frame rate: 1 ms

The window size of the STFT was set to 32 ms. The window size of the CWT analysis varied between 2 ms and 40 ms. The MORLET wavelet had a constant-Q factor of about 3.3 by setting the center frequency parameter to $F_c = 0.5$ and the bandwidth parameter to $F_b = 6$. The chosen frame rate was motivated by the smallest wavelet filter length which ensured a high time resolution and generated at the same time enough frames for training the HMM models compared to the standard frame rate of 10 ms. It is clear that the STFT does not require such a high data rate but for comparison purposes both systems used the same frame rate. The static features were augmented with delta and acceleration coefficients obtained by HTK yielding 45-dimensional feature vectors.

3.4. Results

The results for the stop classification performance of the STFT baseline and the wavelet system are shown in Table 2.

3.5. Discussion

Looking at the overall stop classification rate the wavelet features with 86 % achieved a moderate but significant ($\alpha = 0.01$) higher accuracy than the STFT baseline system with 83 %. This corresponds to a 14 % relative error reduction. In one of the best reported results in the literature a knowledge based approach using acoustic-phonetic features was proposed for English stop classification [8]. The evaluation was carried out on 1200 stops of the TIMIT database (60 speakers) and achieved an overall accuracy of 86 %. In comparison to that our wavelet based system obtained an equivalent overall classification accuracy on a much larger dataset (168 speakers).

A detailed look on the errors made by the wavelet system revealed a strong asymmetry in the confusion of voiced and voiceless stop pairs (see confusion matrix in Table 2). About 54 % of the remaining errors occurred between voiced and their voiceless pendants. These findings are in agreement with previous reported results in [1]. This high confusion can partly be explained by the fact that the burst spectrum for a voiced stop and its voiceless counterpart is very similar [1]. Using a HMM framework another point to consider is the usually shorter duration of voiced stops compared to their voiceless pendants resulting in fewer training data and hence, less accurate acoustic models. Considering the composition of the database voiced stops are slightly underrepresented. Previous work [1, 8] has also shown that temporal measures like the voice onset time (VOT), defined as the passed time between the stop release and the onset of voicing, are better suited to distinguish voiced from unvoiced stops than spectral features. Voiceless

Table 2: Confusion matrix (%) for the classification of 6332 stops. The overall accuracy for the wavelet system (white columns) is 86 % while the STFT baseline system (gray columns) achieved an overall accuracy of 83 %. Since the flap [dx] is an allophone of [t] and [d] their confusions (marked as X) were not treated as errors [8].

	Classified as [t]		Classified as [d]		Classified as [k]		Classified as [g]		Classified as [p]		Classified as [b]		Classified as [dx]	
[t]	86.7	83.2	7.2	11.1	2.9	2.7	0.4	0.7	2.6	1.8	0.2	0.5	X	X
[d]	16.5	14.5	74.8	77.1	0.5	1.2	2.9	3.5	1.5	0.8	3.8	2.9	X	X
[k]	2.8	3.7	0.9	1.3	90.0	85.8	4.4	6.7	1.9	2.4	0.0	0.1	0.0	0.0
[g]	0.4	1.3	7.1	6.4	12.8	14.5	75.3	72.1	2.0	0.9	1.3	1.5	1.1	3.3
[p]	3.8	4.7	1.3	1.2	1.4	2.5	0.4	0.7	89.1	82.5	4.0	8.2	0.0	0.2
[b]	0.2	0.2	4.9	3.2	0.1	0.2	1.3	1.4	11.3	8.8	81.0	83.7	1.3	2.5
[dx]	X	X	X	X	0.0	0.2	0.6	1.1	0.2	0.2	0.3	1.9	98.9	96.6

stops normally exhibit a much larger release duration than voiced stops (see Figure 1). However, this only holds for stops occuring in syllable-initial position. If a voiceless stop occurs in a word-final (e.g. pit) or cluster position (e.g. stop) it is usually articulated unaspirated which leads to shorter and hence ambiguous VOTs. The experiments undertaken in this study used stops extracted from a large variety of positions complicating the voicing distinction based on VOT. To underline this argument we measured the release duration as an estimate of the VOT of all stops in the training set. We further measured the release durations for all stops being misclassified by the HMM with respect to their voicing property (e.g. [d] as [t] or [p] as [b]). The result shown in Figure 3 demonstrates that there exists a fairly large overlap of the release duration distributions of voiced and voiceless stops in contrast to the almost perfect separation reported in [1]. This is mainly due to unaspirated voiceless stops which exhibit a VOT very similar to voiced stops [10]. Furthermore it is interesting to see that most of the HMM misclassifications occur for release durations between 10 -40 ms which is exactly the region of the highest overlap. This indicates that an additional VOT rescoring as done in [1] would not result in much benefit as the HMM already models the VOT quite well.

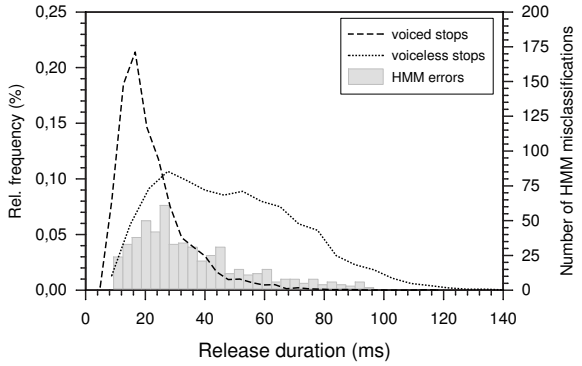


Figure 3: Release duration distributions for unvoiced and voiceless stops measured in the TIMIT training set. The bars show the according distribution of the HMM misclassifications between voiced and voiceless stop pairs based on the TIMIT test set.

4. Conclusions

The following conclusions about the use of wavelet analysis and HMMs for English stop classification can be drawn. Firstly, the multi-resolution property of WTs allows for a better modeling of stop consonants, in particular voiceless stops. A statistically significant 14 % relative error reduction was achieved on a large speaker-independent stop classification task compared to the STFT analysis. The obtained overall performance of the data-driven HMM approach was comparable to one of the best knowledge based systems tuned for stop classification. Secondly, the study has shown that the varying window size of the wavelet analysis can not prevent the confusions between voiced and voiceless unaspirated stops. Because in that case the VOT is not an reliable voicing indicator other (acoustic) features related to the state of the glottis may be needed. Also applying discriminative feature transformations instead of the PCA as well as integrating context information are likely to result in further improvements.

5. References

[1] P. Niyogi and P. Ramesh, "The voicing feature for stop consonants: recognition experiments with continuously spoken alphabets," *Speech Communication*, vol. 41, pp. 349–367, 2003.

[2] E. Lukasik, "Wavelet packets based features selection for voiceless plosives classification," in *ICASSP*, Istanbul, Turkey, 2000.

[3] Christophe Gerard et al., "A wavelet representation evaluation for stop-consonants classification," in *VIII European Signal Processing Conference*, Trieste, Italy, 1996.

[4] H. L. Cycon et al., "Stop consonant classification using wavelet packet transforms and a neural network," *Intelligent Engineering Systems Through Artificial Neural Networks*, vol. 5, pp. 733–738, 1995.

[5] Basilis Gidas and Alejandro Murua, "Classification and clustering of stop consonants via nonparametric transformations and wavelets," in *ICASSP*, Detroit, USA, 1995.

[6] Anthony Teolis, *Computational Signal Processing with Wavelets*, Birkhäuser, 1998.

[7] S. Young et al., *The HTK Book*, Cambridge University Engineering Department, 2005.

[8] A. M. A. Ali et al., "Acoustic-phonetic features for the automatic classification of stop consonants," *IEEE Transactions on Speech, and Audio Processing*, vol. 9, no. 8, 2001.

[9] L. GiIlick and S. Cox, "Some statistical issues in the comparison of speech recognition algorithms," in *ICASSP*, Glasgow, Scottland, 1989, pp. 532–535.

[10] R.N. Ohde, "Fundamental frequency as an acoustic correlate of stop consonant voicing," *Journal of the Acoustical Society of America*, vol. 75, no. 1, pp. 224–230, 1984.