Stop Consonant Recognition by Temporal Fine Structure of Burst

Seppo Fagerlund, Unto K Laine

Department of Signal Processing and Acoustics, Aalto University
Seppo.Fagerlund@aalto.fi

Abstract
The automatic classification of the unvoiced stop consonants is widely considered as a difficult task for traditional frequency domain and even time-frequency methods. Main reason for this is their short duration and diverse temporal structure. In this paper we present a novel method for stop consonant recognition. The method is based on statistical properties of short temporal fine structure of burst part. Classification is also evaluated with simple frequency domain method.

Index Terms: stop consonant recognition, temporal fine structure, permutation transformation, feature extraction.

1. Introduction
A great majority of the present signal analysis methods operate in the frequency domain. Attempts to perform analysis directly in the time domain are often doomed to fail. One critical aspect is the possible sensitivity of the method to the phase variations caused for instance by time varying channel parameters or signal reflections. Variation in the waveform due to signal phase variation is often considered irrelevant and the analysis method should not react on that. Another characteristic in time domain methods, such as autocorrelation, is that larger amplitude values have greater impact to the result that smaller ones.

In order to be of high quality, the frequency domain analysis has some premises, too. At least the signal under investigation should be sufficiently stationary within the analysis window. However, the stationarity assumption is not true for many time series, especially when speech and audio signals are considered.

In this paper we introduce a novel time-domain method for the analysis of non-stationary signals whose time structure may be rich and noisy. The method is based on the assumption that especially in non-stationary signals important structural information can be found in very short time windows. In our approach, the information related to the temporal fine structure of the signals is not lost during the transformation into the frequency or other domain (assuming phase properties are not considered), where the analysis window is generally too long for detailed temporal analysis. In order to illuminate the power of the method a detailed analysis of TIMIT stop consonant bursts is given.

In speech recognition stop consonants are generally considered as difficult to classify with current automatic speech recognition algorithms [1]. The main reason for this is that the stop consonants are short in duration and they have a rich temporal structure. Most of the studies of consonant recognition are based on spectral information of the burst and formant information of the following vowel [2,3,4]. Furthermore it is generally assumed that phonetic context has large impact on the stop consonant properties and that it provides even more information for stop consonant classification than the consonant burst itself. On the other hand Nathan and Silverman [5] have argued that there is not a clear consensus what are the best features for unvoiced stop consonant recognition. Also Halle et al. [6] have noted already in 1957 that burst part has important role in stop consonant classification. In this work we test stop consonant recognition solely based on the burst of the stop consonants.

In the recent literature, Ali et al. [1,7] have used acoustics-phonetic characteristics as features for classification of stop consonants. These features include burst frequency and formant information of the following vowel. They achieved average recognition rate of 86% for stop consonant classification using TIMIT database. Pican et. al. [3] achieved 78% accuracy for stop consonant recognition by using MFCC features and ANN for classification. Hidden Markov Model was used to bring contextual information to the model. DeMori and Flammia [8] used also formant information and Neural Networks for consonant classification and they achieved recognition rate of 73% for stop consonants.

Methods that are based on temporal fine structure have been studied relatively little, but during recent years interest towards these methods have increased within different kinds of time series analysis. Methods have been used for, e.g., detecting abnormalities in aircraft engines [9], analysis of EEG signals [10], and as complexity measures for time series [11]. A common aspect to these applications is that the signals under investigation are somewhat nonstationary or the interesting signal events are short in duration.

2. Speech Material
In this work we have used TIMIT-corpus for evaluation of our method. The predefined training and testing sets of TIMIT were used for these purposes correspondingly.

Stop consonant recognition was tested for plosives in different vowel contexts. We determined that only those CV pairs, for which TIMIT contains at least five pairs in the training data set and at least one corresponding pair in the testing data set, were qualified for this study. With this requirement we found 18 different cases presented in Table 1.

3. Method
Our method is based on rank ordering (permutation) of signal values in a very short time window. Permutation of a set is an ordered arrangement of its elements. If a set consists of $n$ different elements it can be ordered in $n!$ different ways and thus it has $n!$ different orderings, or permutations (Keller et. al. [7] use term ordinal pattern for a single permutation). In permutation transformation of time series, the absolute values of the samples are not relevant. Instead, the ranking of the amplitude values in a short time window is analyzed and the sample values are replaced by their rank numbers. In this procedure the continuous amplitude values in the window are quantized simultaneously and for each moment of time (window location) a new permutation index is created. Permutation transformation is thus a scale free method to perform local waveform quantization of a signal.
Permutation of real valued signal $x(t)$ is denoted by $\pi^r_n$, where $r$ is time delay and $n$ is size of the permutation window. The permutation window size denotes the number of samples in permutation window that are analyzed concurrently and the time delay is the window step size between adjacent signal samples in original time series (see Figure 1). Single permutation of a time series $x(t)$ is defined as

$$\pi^r_n(t) = \begin{pmatrix} 1 & 2 & 3 & \cdots & n \\ r_1 & r_2 & r_3 & \cdots & r_n \end{pmatrix}$$

(1)

satisfying

$$x_{r_{i+1}} \preceq x_{r_{i+2}} \preceq \cdots \preceq x_{r_{i+t}}$$

(2)

where $r$ is ranking of the sample and $r$ is time delay in samples between adjacent samples in permutation window. Note that if $e.g.$, $\tau = 2$, the permutation window consists of every second sample in the original time domain signal. Equal amplitude values in the original signal (in the permutation window) are assumed to be very rare but if they occur we define $r_i \succ r_{i+1}$. Figure 1 shows an example of one permutation window (black continuous line) of an arbitrary signal.

Table 1. Number of CV samples in training and testing sets.

<table>
<thead>
<tr>
<th>voc/diff</th>
<th>k</th>
<th>p</th>
<th>t</th>
<th>k</th>
<th>p</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>aa</td>
<td>170</td>
<td>104</td>
<td>87</td>
<td>38</td>
<td>36</td>
<td>16</td>
</tr>
<tr>
<td>ax</td>
<td>75</td>
<td>58</td>
<td>98</td>
<td>22</td>
<td>12</td>
<td>36</td>
</tr>
<tr>
<td>ac</td>
<td>149</td>
<td>48</td>
<td>112</td>
<td>48</td>
<td>9</td>
<td>32</td>
</tr>
<tr>
<td>ao</td>
<td>103</td>
<td>38</td>
<td>59</td>
<td>36</td>
<td>17</td>
<td>12</td>
</tr>
<tr>
<td>iy</td>
<td>77</td>
<td>73</td>
<td>128</td>
<td>17</td>
<td>25</td>
<td>40</td>
</tr>
<tr>
<td>ih</td>
<td>101</td>
<td>71</td>
<td>159</td>
<td>35</td>
<td>29</td>
<td>36</td>
</tr>
<tr>
<td>ix</td>
<td>225</td>
<td>70</td>
<td>407</td>
<td>55</td>
<td>34</td>
<td>146</td>
</tr>
<tr>
<td>eh</td>
<td>306</td>
<td>70</td>
<td>138</td>
<td>103</td>
<td>58</td>
<td>46</td>
</tr>
<tr>
<td>er</td>
<td>75</td>
<td>77</td>
<td>63</td>
<td>21</td>
<td>29</td>
<td>13</td>
</tr>
<tr>
<td>acri</td>
<td>44</td>
<td>145</td>
<td>124</td>
<td>12</td>
<td>66</td>
<td>39</td>
</tr>
<tr>
<td>ay</td>
<td>26</td>
<td>44</td>
<td>101</td>
<td>1</td>
<td>5</td>
<td>41</td>
</tr>
<tr>
<td>ah</td>
<td>122</td>
<td>21</td>
<td>62</td>
<td>54</td>
<td>9</td>
<td>13</td>
</tr>
<tr>
<td>aw</td>
<td>42</td>
<td>19</td>
<td>11</td>
<td>6</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>ey</td>
<td>85</td>
<td>92</td>
<td>89</td>
<td>18</td>
<td>18</td>
<td>58</td>
</tr>
<tr>
<td>ow</td>
<td>72</td>
<td>58</td>
<td>32</td>
<td>18</td>
<td>14</td>
<td>7</td>
</tr>
<tr>
<td>uh</td>
<td>33</td>
<td>20</td>
<td>28</td>
<td>10</td>
<td>10</td>
<td>8</td>
</tr>
<tr>
<td>uw</td>
<td>11</td>
<td>8</td>
<td>56</td>
<td>5</td>
<td>7</td>
<td>19</td>
</tr>
<tr>
<td>ux</td>
<td>35</td>
<td>5</td>
<td>240</td>
<td>6</td>
<td>2</td>
<td>56</td>
</tr>
<tr>
<td>Tot</td>
<td>1751</td>
<td>1021</td>
<td>1994</td>
<td>505</td>
<td>381</td>
<td>619</td>
</tr>
</tbody>
</table>

Permutation pair frequency matrix

In order to build statistical models for stop consonant bursts, Permutation Pair Frequency (PPF) matrix is constructed from the permutation code sequences. It describes the statistics of permutation transitions (permutation pairs) found in the sequence. The PPF matrix can also be created using different time lags, which describe the time delay between two permutations (corresponding window hop size) in the permutation code sequence and delay (in samples) between permutation windows in original signal. PPF matrices are then used as statistical models of the signals in the classification tests. With permutation window size five, the transition frequency matrices have 120x120 elements. PPF matrix can be defined as

$$A(\pi_i,\pi_j) = \frac{n(\pi_i,\pi_j)}{N}$$

(3)

where $n(\pi_i,\pi_j)$ is total number of permutation code pairs $\pi_i$ and $\pi_j$ in chosen time lag of the sequence and $N = \sum n(\pi_i,\pi_j)$ is a normalization coefficient that forces the sum of the elements of $A$ to be exactly 1.

In general PPF matrices become very sparse since most of the possible permutation pairs are not present in the sequence. This may cause numerical problems in classification phase. However, the sparseness of the PPF matrix can be reduced by using a spatial filter. The spatial filter is not applied directly to the PPF matrices but through the Kendall’s tau $\tau_k$ [12]

$$\tau_k(\pi_i,\pi_j) = 1 - \frac{2d_k(\pi_i,\pi_j)}{d_{k\min}}$$

(4)

where $d_k$ is Kendall’s distance or metric. It is defined as the minimum number of local, elementary permutations needed to reorganize a permutation $\pi_0$ to form an equal list with permutation $\pi_k$. The elementary permutation means an operation where the neighboring elements are interchanged, e.g., $\{1\ 2\ 3\ 4\} \rightarrow \{2\ 1\ 3\ 4\}$ has $d_k = 1$. The term $d_{k\min}$ is the minimum number of elementary permutations

Note that if the permutation window is moved only one sample after each transformation the signal is quantized from 16 bit/sample to $\log_2 120 = 6.9$ bit/sample.
needed to organize a permutation to its reverse permutation, e.g. \( \{ 2 \ 1 \ 3 \ 4 \} \rightarrow \{ 4 \ 3 \ 1 \ 2 \} \). Term \( d_{\text{K}} \) is given by
\[
d_{\text{K}} = \frac{n(n-1)}{2}
\]
where \( n \) is size of the permutation window. Thus Kendall's metric is not geometric in nature, but algorithmic.

Each PPF matrix element is composed of two different permutations and every permutation has four closest neighbors \( d_{\text{K}} = 1 \), thus each element pair (corresponding to one point in the frequency matrix) is affected by eight other elements that are at a \( d_{\text{K}} = 1 \) distance from the original elements of the pair.

New values for the matrix elements are obtained by having a weighted sum of nine neighboring values in the Kendall space (in our tests neighboring elements have half of the weight related to the original value). This can be regarded as low-pass filtering of the matrices in Kendall-space. In classification experiments we found that optimal results were obtained when the smoothing was repeated three times. Recall that the closest neighbor of a permutation corresponds to interchanging two adjacent samples in a permutation window. Thus smoothing improves robustness of the method by allowing local change of sample ordering (minor amplitude variation).

In the following simulations, the smoothing method is applied to the statistics of the PPF matrices. We integrate the number of events inside a certain Kendall distance to form a smoothed (averaged) statistical image of occurrences of certain permutation pairs in the permutation code sequence.

4. Recognition tests

In this work the classification of the [k], [p] and [t] consonants was performed entirely based on the temporal fine structure of their burst parts. Recognition was performed with and without the context information of the upcoming vowel. For comparison the classification test was performed with the same training and testing material also in the FFT domain.

Three different preprocessing filtering methods were studied: pre-emphasis, flat (no filtering), and low-pass with cut-off frequency at 7.2kHz. Also lower cut-off frequencies were tested but they had a negative impact on the classification accuracy. Pre-emphasis filter is a standard preprocessing procedure in different speech and audio analysis and recognition tasks. Here we used a pre-emphasis filter with \( \alpha = 0.9 \). In the low-pass case we used a 10th order low-pass FIR-filter.

4.1. Training

In the training phase, one model for each consonant in different vowel context was created as an average of all corresponding tokens in the training set. In context invariant test, one model for each consonant was created from all tokens listed in Table 1. Consonant models were created from the beginning of consonant burst part only. We used variable frame lengths for the model construction in order to study the temporal distribution of the distinctive information along the burst and its effect to the classification. The starting point of the window was always fixed at the beginning of the burst. If the model frame length was greater than the length of the actual burst, we used the burst part only for the model construction. Spatial filtering was applied to each PPF matrix model after collecting the data from all corresponding training signals.

Models for the FFT experiments were also created from the beginning of the burst parts of the consonants. FFT-frame size was equal to the frame size used for the permutation model creation.

4.2. Classification

The PPF matrix construction for the classification starts with data located in a 19ms wide window placed just before the beginning of the burst. At each iteration, the window size is increased with 5 ms. Frequency matrix is updated and compared to the models. We used three different distance measures between the test sample and the models: relative Euclidean distance, maximum absolute difference, and Kulback-Leiber divergence measure. In relative Euclidean distance the models (PPF-matrices) are treated as feature representations and Euclidean distance is calculated as
\[
d_{\text{rel}}(p,q) = \sqrt{\sum_{i,j} (p_{ij} - q_{ij})^2}
\]

5. Results

Results of the stop-consonant recognition with and without matrix smoothing are presented in Table 2. Columns indicate recognition accuracy with different window sizes in model creation phase. Euclidean distance measure was used in results presented in Table 2. Results show that we need only a little data from beginning of the burst for efficient classification. Preprocessing procedure has only a little influence to the recognition accuracy while PPF matrix smoothing has increasing effect to the recognition accuracy.

<table>
<thead>
<tr>
<th>Window Size/ms</th>
<th>Smoothing ON</th>
<th>Smoothing OFF</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>HP</td>
<td>AP</td>
</tr>
<tr>
<td>3</td>
<td>65</td>
<td>72</td>
</tr>
<tr>
<td>6</td>
<td>74</td>
<td>78</td>
</tr>
<tr>
<td>13</td>
<td>74</td>
<td>78</td>
</tr>
<tr>
<td>22</td>
<td>77</td>
<td>77</td>
</tr>
<tr>
<td>60</td>
<td>67</td>
<td>72</td>
</tr>
</tbody>
</table>

Consonants in the training dataset have different amount of samples in different contexts. In our experiments, all
available data were used for model creation and effect of the amount of data was compensated with model normalization as explained previously in Eq. (3). In order to test the impact of different amounts of training data, we also applied a test for one CV pair (\(k\),\([p],[t]+[eh]\)) with equal number of training samples when creating the models. Data in this test was selected randomly from all possible samples and the test was repeated ten times using different samples in models each time. Average recognition result over 10 iteration using 22ms window from beginning of the burst for model creation was 84% correct with standard deviation 0.47 when using all available data accuracy was 83%. This result justifies our decision to utilize all available data.

Recognition results as function of length of the training data used for model creation are presented in Figure 2. The results are averages over recognition results in different vowel contexts. Results shows clearly that PPF-matrix smoothing improves recognition accuracy. Also context information has improving impact to the recognition accuracy.

Figure 2: Recognition results as function of frame length in model training phase.

6. Conclusions

In this work, the permutation transformation in combination with proper statistical methods is shown to be an efficient tool in classification of short-term signals with a rich temporal structure. The recognition results show that the permutation transformation method efficiently codes relevant structural information especially at low and middle frequencies. Recognition accuracy does not vary much if the signals are low-pass filtered while pre-emphasis reduces the classification accuracy.

As for the parameters, the best recognition results were obtained with time delay value of two. This result suggests that the current permutation time window could be increased in order to collect more required and relevant information needed for consonant classification. However, larger window size increases the number of possible permutations and the method becomes computationally infeasible. On the other hand, decreasing window size would lead to dense transition matrices with less discriminative elements.

Recognition results obtained by the time-domain permutation method can be compared with results produced by using spectral models (see e.g. [4,5]). However, unlike this work, information from the following vowel is typically utilized in consonant-vowel recognition. Niyogi and Sondhi [1] tested several different algorithms for detecting and recognizing stop consonants from continuous speech. Segmentation and consonant classification was divided into two separate problems and classification rates for consonant recognition ranged from 70% to 90%. Ali et al. [3,6] used several different acoustics-phonetic features for phoneme representation and rule based classifiers for recognition. They gained an 86% overall recognition rate for stop consonant recognition.

In this study, models for consonants were created from the consonant release (burst) part only and information from the upcoming transition to the vowel locus was not used. Recognition results were still found to be relatively high although it is often thought that the following vowel has a strong influence on the overall stop consonant perception (but see also [13]). Having separate models for vowels would probably still increase the recognition accuracy. Also, it is possible to extend the method to biphones (a set of two adjacent phones) by modeling burst-vowel combinations. In any case permutations have shown to be a promising method for analysis and classification of non-stationary short-term speech events directly in the time domain. The results also speak for the importance of the temporal fine structure and temporal order of the acoustic waveform.

7. Acknowledgement

This study has been partially financed by EU project Acquisition of Communication and Recognition Skills (ACORNS) and partially by the Academy of Finland.

8. References