A REAL-TIME FIR-BASED FILTERBANK, AS THE ACOUSTIC FRONT END OF A SPEECH RECOGNIZER

Paul van Alphen & Louis C.W. Pols
Institute of Phonetic Sciences, University of Amsterdam, Herengracht 338, 1016 CG Amsterdam, The Netherlands

ABSTRACT
This paper describes the design of a digital filterbank, that runs in real-time on a single TMS 320 processor. The designed filterbank uses Finite Impulse Response (FIR) filters, which enables exact time alignment for all frequencies. Furthermore, the output channels of the filterbank have a high frequency selectivity. This high selectivity has certain advantages, but also potential disadvantages, for instance with respect to low harmonics positioned inside or outside a specific filter. To ensure robustness of the filterbank to changes in f0, an algorithm is presented that eliminates most of these effects. Results for vowel identification are presented.

1 INTRODUCTION
In our Automatic Speech Recognition project, we want to adapt certain concepts of auditory modelling. As a front end for our recognition system, we have developed a digital filterbank, that runs in real time. The advantages of a real-time acoustic front end are obvious: large amounts of data can be processed, so one can investigate a wider variety of speech.

In designing a real-time digital filterbank on a signal processor, frequency shifting and decimation (down-sampling) can be powerful tools [1]. Using these techniques, it is possible to do the calculations of a 15-channel filterbank, including energy estimation, in a single TMS 32010 processor. An outline of the implemented filterbank is given in Section 2.A.

As a first recognition test for this filterbank, vowel identification experiments have been performed. The spectra of steady-state vowels are compared with reference spectra. The influence of spectral smoothness and the importance of a good choice for the distance metric will be demonstrated. The two metrics used (level metric / slope metric) are discussed in Section 2.B.

Tests and results of this recognition system are presented in Section 3. These tests include synthetic as well as natural vowels selected from spoken sentences.

2 RECOGNITION SYSTEM

2.A FILTERBANK
In a previous report [1], a detailed description has been given of the design of the filterbank, using the frequency shifting and decimation techniques. The conceptual advantages can be summarized as follows:
- adjacent channels are processed together wherever possible.
- filtering that is specific for one output channel is done at the lowest possible sample frequency, which entails little (calculational) effort. Because of this effective calculation we can use FIR filters throughout the system, which guarantees linear phase.

The total processing scheme for the filterbank is drawn in Fig. 1, where the processing blocks are explained at the bottom of the figure. For more details we refer to the above mentioned paper.

![Diagram of the processing scheme for the filterbank with 15 output channels. On top, the used sample frequency fs per section is indicated.](image-url)
The implemented (nonuniform) filterbank has a constant bandwidth for the low-frequency region, and a 1/3 octave spacing for the higher frequencies (roughly Bark scale). For all 15 channels the filter characteristics are plotted in Fig. 2(a). In Fig. 2(b) it can be seen that the composite spectrum is essentially flat over the frequency range of the filterbank. The overall energy can easily be calculated by taking the sum over all channels (provided that good time correction for each channel has been applied).

A side effect of the decimation technique is that the slopes of the filters are very steep. This has as a consequence that the output channels of the filterbank have an extremely high frequency selectivity. The spectrum (from filterbank outputs) can thus have peaks and dips that are very sharp. Because we are primarily interested in the spectral envelope, we prefer to further process this raw spectrum. This processing can be split up into two parts, which will be referred to as "filling" and "smoothing". By "filling" we mean the removal (by interpolation) of sharp dips in the low-frequency channels, caused by the harmonic structure of voiced speech. Because the position of the lower few harmonics in voiced speech always causes dips in specific channels, we use this as a criterion for applying filling or not. Spectra of unvoiced speech thus remain unprocessed. "Smoothing" means that we take the average of every two adjacent channels (this is relevant for distance metrics based on slopes). For our identification testing we use 3 different versions of the bandfilter spectrum:

- The direct spectrum. We include this to serve as a baseline against which the other data can be seen.
- Filled spectrum. The 0 effect should diminish.
- Smoothed spectrum (also filled). As a result slopes change more gradually.
- In the second experiment we also use cepstral coefficients of the direct spectrum.

2.B DISTANCE METRICS

The next step in the postprocessing of the filterbank spectra is to apply a level normalization to compensate for variations in speech level. For each spectrum the total energy is normalized, yielding the level normalized spectral components \( L_X(i) \). This is done by subtracting the total energy (in dB) from every individual output channel \( i = 1, N \), and then adding an arbitrary level of 72 dB. Using these level normalized spectra, the following distance metrics are evaluated:

A. Level Metric (LM), i.e. Euclidean distance of the spectra:

\[
d_{LM} = \sum_{i=1}^{N} (L_1(i) - L_2(i))^2
\]

B. Spectral Slope Metric (SM). With this is meant a reduced form of Klatt's Weighted Spectral Slope Metric (WSM) [2]:

\[
d_{SM} = \sum_{i=1}^{N} (S_1(i) - S_2(i))^2
\]

Where the spectral slopes \( S_X(i) \) (first difference of \( L_X(i) \)) are defined as:

\[
S_X(i) = L_X(i) - L_X(i+1)
\]

3 EXPERIMENTS

3.A SYNTHETIC VOWELS

3.A.1 STIMULI

As a first experiment synthetic vowels are tested. Three groups of five vowels are generated, positioned in the F1-F2 space around the extreme vowels /i/, /a/ and /u/ (see Fig. 3). Formants F3, F4 and F5 have fixed values of 2500 Hz, 3500 Hz, and 4500 Hz respectively. Each of the 15 different vowels is generated with 13 different fundamental frequencies \( f_0 \), ranging from 80 to 200 Hz, in steps of 10 Hz. Using these variables, the total number of stimuli becomes 15 * 13 = 195.

Fig. 2 Frequency response of the filterbank. Filter characteristics of the 15 individual output channels (a), and composite spectrum (b).

Fig. 3 The vowel stimuli of the first experiment plotted according to their first and second formant frequencies (in Hz).
Using a vowel synthesizer [3], the stimuli are generated digitally with the following properties:
- 16000 Hz sampling rate.
- 12-bit amplitude quantization.
- five-formant vowels using a cascade formant synthesizer.
- Q-factor is 10. (Q-factor is the ratio between mid-frequency and -3 dB bandwidth)
- polynomial glottal pulse (defined by Rosenberg [4]):

\[
3 \cdot (t/p)^2 - 2 \cdot (t/p)^3 \\
1 - ((t-p)/t_n)^2 \\
\text{for } 0 \leq t \leq t_p \\
\text{for } t_p < t \leq t_p + t_n \\
\text{where } t_p, t_n, T = 0.40, 0.16 \\
\text{+6 dB per octave spectral tilt simulating sound radiation from the lips.}
\]

Of all these vowels a steady-state approximation is synthesized. These signals are then analyzed by the filterbank. Because the middle of the vowel segment is the best choice as steady-state approximation, we define the corresponding filterbank output to be the vowel spectrum. Additional filling (f0 correction) and smoothing can be performed. As references for recognition the 15 spectra with a fundamental frequency of 100 Hz are used.

3.A.2 RESULTS

Testing consists of identifying the remaining 180 vowel spectra. Ideally, all test spectra at different f0 should be closest in distance to the reference spectrum of the same vowel (with f0 = 100 Hz). We hope of course that the f0 correction procedure (filling) will take care of “irrelevant” differences in the spectra. The error measure is defined to be the percentage of incorrect top candidates (= closest distance). In Table 1 the error rates for five different distance measures are listed. Most of the problems arise for the vowel /u/ (u0, u1, u2, u3, u4), where the absolute formant differences are smallest.

TABLE 1. Identification error rates for 15 synthetic vowels at 12 different f0 values. A level (LM) and slope (SM) metric are compared together with several manipulations on the bandfilter spectra (see text).

<table>
<thead>
<tr>
<th>metric manipulation</th>
<th>error percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>total</td>
</tr>
<tr>
<td>LM</td>
<td>none</td>
</tr>
<tr>
<td>LM</td>
<td>filled</td>
</tr>
<tr>
<td>LM</td>
<td>+smoothed</td>
</tr>
<tr>
<td>SM</td>
<td>filled</td>
</tr>
<tr>
<td>SM</td>
<td>+smoothed</td>
</tr>
</tbody>
</table>

One of the reasons why LM is less accurate is shown in Fig. 4, where two level normalized spectra are displayed that have to be compared. After the level normalization the spectra are "parallel". In case of SM parallel lines yield a zero distance, while LM then makes the "same" error for every point. By establishing this advantage of the slope metric, we have to keep in mind the restriction of this test. The limitation of this synthetic vowel identification is that only for the low-frequency channels there is an apparent effect of f0, whereas in natural speech the variation can be everywhere. In the next section we will therefore describe another test of the distance metrics but this time with more variation introduced by natural speech.

Fig. 4 Two level normalized spectra to be compared.

3.B SPOKEN DUTCH VOWELS

3.B.1 STIMULI

In this experiment all Dutch vowels are spoken in a carrier sentence by one male speaker. The Dutch vowel set consist of 12 monophthongs: [a] [a] [a] [a] [a] [a] [a] [a] [a] [a] [a] [a] where a pairwise long short opposition exists between: [a] [a] [a] [a] [a] [a] [a] [a] [a] [a] [a] [a] We use two different context sentences, in which the 12 different vowels are inserted at two places: "de V van pVt" ("the V of pVt") "de V van mVt" ("the V of mVt") From every sentence both realizations of the vowel V are processed. First vowel segments are selected manually. From the sequence of spectra corresponding to this segment, the spectrum with the smallest interpoint distance is defined to be the vowel spectrum. Every sentence is spoken twice for the reference set creating 8 reference spectra for each vowel (2 sentences; 2 realizations per sentence; spoken twice). The same sentences are spoken three more times for the test set, creating 144 test spectra (2 sentences; 2 per realizations sentence; spoken 3 times for all 12 vowels).
3.B.2 RESULTS

The testing procedure is the same as described in Section 3.A.2, which means that a test spectrum should be closest in distance to one of the 8 reference spectra of the same vowel in order to be correct. Table 2 lists the error rates for the different distance metrics. The lower half of this table includes test results while using 1 up to 16 cepstral coefficients and the LM distance. Cepstra are obtained by extending the 15 spectral components with zeros before applying a 32-point inverse Fourier Transform, which results in 16 cepstral coefficients (the zero-coefficient is dropped). Cepstral coefficients are used as an alternative way of spectral smoothing.

From this test it may be clear that the LM metric on the filled spectra has a high robustness to natural variance. However, the SM metric becomes comparable to the LM metric once a spectral smoothing has been applied. We also see that the usage of cepstral coefficients for spectral smoothing does not improve the recognition rate as much as the filling algorithm. A reason for this can be that the smoothing via cepstral coefficients effects the whole spectrum. Meaningfull dips in the high-frequency region are removed just as well as the unwanted dips in the low-frequency region.

TABLE 2. Identification error rates for 12 spoken Dutch vowels. A level (LM) and slope (SM) metric are compared together with several manipulations on the bandfilter spectra (see text).

<table>
<thead>
<tr>
<th>metric</th>
<th>spectral manipulation</th>
<th>error percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>total</td>
</tr>
<tr>
<td>LM</td>
<td>none</td>
<td>19.4</td>
</tr>
<tr>
<td>LM</td>
<td>filled</td>
<td>12.5</td>
</tr>
<tr>
<td></td>
<td>+ smoothed</td>
<td>14.6</td>
</tr>
<tr>
<td>SM</td>
<td>filled</td>
<td>20.8</td>
</tr>
<tr>
<td></td>
<td>+ smoothed</td>
<td>11.8</td>
</tr>
<tr>
<td>LM</td>
<td>cepstrum</td>
<td></td>
</tr>
<tr>
<td></td>
<td>number of coeff.:</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10</td>
</tr>
<tr>
<td></td>
<td></td>
<td>11</td>
</tr>
<tr>
<td></td>
<td></td>
<td>12</td>
</tr>
<tr>
<td></td>
<td></td>
<td>13</td>
</tr>
<tr>
<td></td>
<td></td>
<td>14</td>
</tr>
<tr>
<td></td>
<td></td>
<td>15</td>
</tr>
<tr>
<td></td>
<td></td>
<td>16</td>
</tr>
</tbody>
</table>

4 CONCLUSIONS

Two vowel identification experiments have been performed by using a digital filterbank based on frequency shifting and decimation. This filterbank design results in a high frequency selectivity per channel. Because of this high selectivity, the harmonic structure of voiced speech can be a problem for the low-frequency channels. Both experiments show that we get improved identification if we remove the harmonic-effect by filling the low-frequency dips before identifying the vowel spectra.

A simple Level Metric proves to have a high robustness to natural variance. The more sophisticated Slope Metric is comparable with the Level Metric once a general smoothing is performed. Whether this Slope Metric can actually improve recognition of natural speech remains a topic for further research.

5 REFERENCES