IMPROVING THE REPRESENTATION OF TIME STRUCTURE IN FRONT-ENDS FOR AUTOMATIC SPEECH RECOGNITION

Wendy J. Holmes
20/20 Speech Ltd., Malvern Hills Science Park, Geraldine Road, Malvern, Worcs., WR14 3BS, UK.
E-mail: w.holmes@2020speech.com

ABSTRACT
This paper describes investigations into the use of ‘excitation-synchronous’ spectral analysis to provide acoustic features for automatic speech recognition. Within each 10 ms frame the region of maximum power is located and used as the centre for the window in a subsequent Fourier transform. The method has been found to be effective in locating stop bursts and vocal-tract responses to glottal closures. This excitation-synchronous analysis has been compared with the more conventional fixed-interval analysis for window lengths ranging from 5 to 25 ms. In connected-digit recognition experiments using mel-cepstrum features, the excitation-synchronous analysis with a window length of 10 ms gave a 10% improvement in recognition performance when compared with the best of the fixed-window conditions.

1. INTRODUCTION
It is usual to derive acoustic features for automatic speech recognition by computing some form of spectral representation over a specified time interval at a given time spacing. Typically, a Fourier transform is used and analysis windows are about 25 ms in length (with a bell-shaped window function) and are spaced at 10 ms intervals. During voiced speech, the effect of the glottal source is to cause a large variation in the short-time energy of the speech signal depending on the location within a pitch period. A window of 25 ms will usually include several pitch periods and so should smooth out this position-dependent variation. However, even with windows as long as 25 ms, it has been reported that the position of the windows can affect the analysis and subsequent recognition performance [1].

In addition, the use of long overlapping windows introduces temporal smoothing which, although probably desirable for fricative regions, inevitably blurs rapid events such as stop bursts and fast formant transitions. Better time resolution can be achieved by using shorter analysis windows. There will obviously be a reduction in the frequency resolution due to the time-frequency trade-off that is inherent in the short-time Fourier transform. However, for recognition it is usual to convert the Fourier output to a mel scale [2], for which the frequency resolution is around 100 Hz at low frequencies and decreases considerably at high frequencies. This scale is motivated by knowledge of the frequency-resolving power of the human auditory system. The frequency resolution provided by the 25 ms time-window of the original analysis is therefore probably more than is needed for the recognition feature set.

Various analysis methods have been suggested as alternatives to the short-time Fourier transform. These alternatives are aimed at more effective capture of both the temporal and the frequency characteristics of speech [3]. However, even when performing a conventional short-time Fourier analysis, it should be possible to capture transient events in the speech signal without harmful loss of frequency resolution by using shorter analysis windows, provided that these windows are positioned on the most informative regions of the speech signal. For releases of plosive consonants, the windows should be positioned on the stop burst and, in the case of vocalic regions, the positions for the windows should correspond to regions of maximum energy immediately following glottal closure. Thus some form of excitation-synchronous analysis is appropriate.

Although a variety of methods have been proposed for detecting instants of glottal closure (e.g. [4]), it is often difficult to identify the exact time of glottal closure reliably and excitation-synchronous analysis has not generally been used when deriving features for automatic speech recognition. However, for recognition all that is really required is to locate the relevant region approximately in order to obtain the most useful estimate of the spectrum for each frame. It should be possible to obtain this estimate simply by searching for the region of maximum local power in each frame, and using this region for the centre of the window for that frame. Even if occasional analysis windows are not located very accurately, overall the results should be much more representative of the evolving speech properties than results obtained using any length of window at a fixed position in each frame. This paper describes a method for locating regions to use for analysis and presents the results of investigations comparing this analysis with a fixed-interval analysis for different window lengths.

2. LOCATING REGIONS FOR ANALYSIS
The method used to locate regions for analysis was originally developed by J.N. Holmes for the formant analyser described in [5], and involves locating the most intense region within each 10 ms frame. The speech is sampled at 8 kHz, and is first pre-emphasized so that the average spectral power distribution is approximately uniform for vocalic regions. The low-frequency content is then further reduced by a spectral shaping filter designed just to weaken the lowest few hundred Hz of the signal. The intention is to ensure that the signal is really showing vocal-tract excitation rather than the low-frequency components of the glottal pulses. This filtered waveform is peak rectified before computing the sum of the signal samples within a 2.5 ms sliding window and searching for the maximum over the duration of each frame. The analysis window for a frame is then centred on the 2.5 ms region corresponding to the energy...
Figure 1: Processing waveforms to locate regions for analysis. Pre-emphasized speech waveforms are shown at the top, followed by the waveforms after filtering, after peak rectification, and after computing the sum over the samples in the previous 2.5 ms. For each frame (frame boundaries are marked with dashed vertical lines), the section that will then be used for the central region (5 ms) of the analysis window is shown by the horizontal bar at the top of the figure. Excitation points, corresponding to the onset of regions of maximum energy, are indicated by the short vertical lines at the bottom of the figure. (a) typical three-frame segment of a vowel, (b) frame containing the release of a plosive consonant, /t/.

maximum for that frame. The portion of the signal over which the search for the energy maximum takes place is chosen such that the centre of the analysis window will always be within the relevant frame.

Figure 1 illustrates the operation of the algorithm for three frames of a typical vowel and for a frame containing the onset of a plosive consonant. It can be seen that each analysis window will be centred on a high-energy portion of signal. The first sample of any 2.5 ms region corresponding to a local energy maximum can be regarded as an ‘excitation point’. In the case of the vowel (Figure 1(a)), this excitation point corresponds to the onset of a formant response to glottal closure, and for the plosive (Figure 1(b)), it is the onset of the stop burst.

The aim is to find the most prominent excitation point in each individual frame. However, it is obviously not appropriate for the detected excitation points in successive frames to be only a few samples apart, because the portion of signal being analysed would then be substantially the same for the two frames. This situation could arise if a region of high signal energy occurs near a frame boundary, but is avoided by imposing a minimum value (4 ms was used here) on the allowed spacing between successive excitation points. Due to this minimum spacing restriction, it is necessary to always look ahead some way into the next frame before making the decision for the current frame.

Overall the method has been found to be very effective in locating both the onsets of vocal-tract responses to glottal closures and the onsets of stop bursts. The method always finds one, and only one, excitation point for each frame. Thus, although for example the middle frame of the vowel segment shown in Figure 1(a) contains two excitation pulses, the one corresponding to the more intense region of signal is selected. Furthermore, an energy maximum will always be found, even for sounds such as fricatives for which the short-term signal energy varies much less (and not systematically) and in frames of silence (where in the extreme the only influencing factor will be the minimum spacing restriction).

3. SPECTRAL ANALYSIS

3.1. Method

The method described in Section 2 was used to specify the positions of regions for spectral analysis. A fast Fourier transform (FFT) was applied to different analysis windows, using window lengths ranging from very short (5 ms) to quite long (25 ms). For comparison FFT analyses were also performed using the same window lengths but always placed at the centre of each 10 ms frame.

The same analysis procedure was used for all the windowing conditions. Using speech sampled at 8 kHz, a Hanning window of the required length was applied at the specified position for each frame. The windowed signal was input to a 256-point FFT, after first padding with the appropriate number of zeros. Thus the output of all the analyses conformed to the same closely-spaced frequency sampling, although there was of course an increasing degree of smoothing as the window length decreased.

For the longer window conditions, where the window would usually include more than one glottal cycle in voiced speech, the harmonics of the fundamental were resolved in the FFT analysis. This influence of the excitation spectrum can mask the formant structure and is not generally considered desirable for speech recognition. The FFT output was therefore liftered to remove the components of the spectrum due to the excitation. So as to maintain consistency, the liftering was applied to all the conditions, although the effect of this process was negligible in the case of the short-window excitation-synchronous analysis.

3.2. Analysis results

Visual comparisons were made between spectrographic representations of the output of different FFT analyses for a variety of speech sounds. With a long window of 20 or 25 ms, the analysis produced similar results whether performed excitation-synchronously or at fixed intervals. A typical fixed-
Figure 2: Wideband Spectrogram of an utterance “three six six”, shown time-aligned with spectrographic displays representing the output of a 256-point FFT applied to different analysis window conditions. Interval analysis is shown in the middle plot of Figure 2, and it can be seen that there is blurring of the stop release for both of the occurrences of /kJ/. In contrast, a shorter 10 ms excitation-synchronous window (top plot in Figure 2) captures these rapid onsets very effectively. When the same length of window was applied at fixed intervals (bottom plot in Figure 2), the accuracy with which stop bursts were captured depended on their position in relation to the positions of the windows. Of the two examples shown in Figure 2, the first one is by chance represented reasonably well, while the second one is captured more weakly.

When a short fixed window was used, during vowel regions there was frame-to-frame fluctuation depending on the position of the analysis window relative to the glottal cycle. However, when this length of window was used excitation synchronously, in vocalic regions the spectral changes from frame to frame were generally as smooth as when a longer window was used at fixed intervals. Although the frequency resolution is coarser when using a shorter window, the formant structure is seen very clearly with the excitation-synchronous analysis.

When the pitch was very low (less than 100 Hz) there were occasional analysis frames within a vowel region that contained no genuine excitation point, so causing the spectrum measured for that frame to be at an artificially-low level. However, the situation could be easily detected by comparing the spectral level with that in the two neighbouring frames. Such a test was incorporated into the analysis and, when a low-level frame was detected between two frames of similar spectral shape, the low-level frame was replaced by a repetition of one of its neighbours.

4. SPEECH RECOGNITION EXPERIMENTS

4.1. Feature set for recognition

The FFT outputs were first converted to a mel scale. Although the shorter-window FFTs have lower frequency resolution, the difference between the conditions is much smaller after conversion to the mel scale. For analysis windows of less than 10 ms duration there is some loss of accuracy at low frequencies, but otherwise the frequency resolution on the mel scale is similar for all the conditions.

A cosine transform was applied to the mel-scale spectrum, and the first eight cepstrum coefficients and an overall energy feature were used as the feature set for recognition experiments.

4.2. Experimental set-up

The recognition task involved speaker-independent connected-digit recognition using a simple hidden Markov model (HMM) framework. The test data were four lists of 50 digit triples spoken by each of 10 male speakers. The training data were from 225 different male speakers, each reading 19 four-digit strings taken from a vocabulary of 10 strings.
Experiments are now needed to see whether the findings also demonstrated on a simple digit-recognition task, and shorter 10 ms analysis window. These improvements have been synchronous analysis gave the lowest error rate when using a rather than using an analysis at fixed 10 ms intervals. Furthermore, whereas the fixed-interval analysis performed best when using long windows of 20 or 25 ms, the excitation-synchronous analysis gave the lowest error rate when using a shorter 10 ms analysis window. These improvements have been demonstrated on a simple digit-recognition task, and experiments are now needed to see whether the findings also apply to other tasks. A more difficult task may give more scope for improvement from better positioning of analysis windows.

Displays of spectral analyses have demonstrated the ability of a short time window excitation-synchronous analysis to capture dynamic information in the speech signal, including rapid formant transitions and sudden onsets of stop bursts. There may therefore be a greater advantage from using this analysis with models, such as trajectory-based segmental HMMs [6], which provide a better representation of speech dynamics. It may also be possible to gain more advantage for conventional HMMs by using appropriate time-derivative features.

When using excitation-synchronous analysis, the short time window has been found to be very successful for analysing vocalic sounds and stop bursts. However, a short window is also sensitive to the more random frame-to-frame fluctuations that are characteristic of fricative sounds. For these regions, the temporal smoothing provided by longer windows is probably beneficial. There may therefore be greater advantage to be gained by varying the size of the analysis window depending on properties of the speech. Initial experiments have indicated that such a combination of short and long analysis windows can give better performance than only short (or only long) windows.

### Table 1: Connected-digit recognition performance as a function of analysis window length, compared for ‘excitation-synchronous’ analysis versus ‘fixed’ analysis windows centred on each 10 ms frame.

<table>
<thead>
<tr>
<th>Window Length in ms</th>
<th>% Subs.</th>
<th>% Del.</th>
<th>% Ins.</th>
<th>% Error</th>
<th>% Subs.</th>
<th>% Del.</th>
<th>% Ins.</th>
<th>% Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>1.5</td>
<td>0.5</td>
<td>0.1</td>
<td>2.1</td>
<td>1.3</td>
<td>0.5</td>
<td>0.2</td>
<td>2.0</td>
</tr>
<tr>
<td>20</td>
<td>1.5</td>
<td>0.5</td>
<td>0.1</td>
<td>2.1</td>
<td>1.4</td>
<td>0.5</td>
<td>0.2</td>
<td>2.1</td>
</tr>
<tr>
<td>15</td>
<td>1.6</td>
<td>0.5</td>
<td>0.1</td>
<td>2.2</td>
<td>1.5</td>
<td>0.5</td>
<td>0.1</td>
<td>2.1</td>
</tr>
<tr>
<td>10</td>
<td>1.8</td>
<td>0.6</td>
<td>0.3</td>
<td>2.7</td>
<td>1.4</td>
<td>0.4</td>
<td>0.1</td>
<td>1.9</td>
</tr>
<tr>
<td>7.5</td>
<td>2.4</td>
<td>0.7</td>
<td>0.4</td>
<td>3.5</td>
<td>1.4</td>
<td>0.5</td>
<td>0.1</td>
<td>2.0</td>
</tr>
<tr>
<td>5</td>
<td>3.2</td>
<td>0.7</td>
<td>0.7</td>
<td>4.6</td>
<td>1.4</td>
<td>0.4</td>
<td>0.3</td>
<td>2.1</td>
</tr>
</tbody>
</table>

For each analysis condition, three-state context-independent monophone models and four single-state non-speech models were used, with single-Gaussian pdfs and diagonal covariance matrices. The model structure was a simple left-to-right one including self-loop transitions. Model means were initialized from a very small quantity of hand-annotated training data (twelve digits from each of two speakers), with model variances initialized to the same arbitrary value. All model parameters were trained with ten iterations of Baum-Welch re-estimation.

### 4.3. Recognition results

From the results shown in Table 1 it can be seen that recognition performance is very similar for window lengths of 25, 20 and 15 ms. In addition, for these window lengths the performance is not greatly affected by whether the analysis is performed at fixed intervals or excitation-synchronously. When the window length is reduced to less than 15 ms the performance degrades for the fixed-interval analysis. This drop in performance is not surprising, in view of the frame-to-frame variation that occurs depending on window position (see Figure 2) and the fact that some speech events may be missed altogether with such short fixed windows. In contrast, when the windows are positioned excitation synchronously good performance is maintained as the window length is made much shorter, with the best performance being for the 10 ms window. This error rate of 1.9% shows a worthwhile gain over all the fixed window conditions, for which the lowest error rate of 2.1% was obtained with a window length of 20 or 25 ms.

It is interesting that, when using the excitation-synchronous analysis, the overall level of recognition performance does not vary greatly across a wide range of window lengths. There were however differences in the pattern of errors, and further work is needed to investigate the relative merits of different window lengths for representing different properties of the speech signal.

### 5. CONCLUSIONS

A small but consistent improvement in recognition performance has been demonstrated by using analysis windows that are centred on the region of maximum energy in each 10 ms frame, rather than using an analysis at fixed 10 ms intervals. Furthermore, whereas the fixed-interval analysis performed best when using long windows of 20 or 25 ms, the excitation-synchronous analysis gave the lowest error rate when using a shorter 10 ms analysis window. These improvements have been demonstrated on a simple digit-recognition task, and experiments are now needed to see whether the findings also apply to other tasks. A more difficult task may give more scope for improvement from better positioning of analysis windows.

### 6. REFERENCES