

SPEECH SYNTHESIS

The present invention is concerned with the automated generation of speech (for example from a coded text input). More particularly it concerns analysis-synthesis methods where the "synthetic" speech is generated from stored speech waveforms derived originally from a human speaker (as opposed to "synthesis by rule" systems). In order to produce natural-sounding speech it is necessary to produce, in the synthetic speech, the same kind of context-dependent (prosodic) variation of intonation that occurs in human speech. This invention presupposes the generation of prosodic information defining variations of pitch that are to be made, and addresses the problem of processing speech signals to achieve such pitch variation.

One method for pitch adjustment is described in "Diphone Synthesis Using an Overlap-add Technique for Speech Waveforms Concatenation", F.J. Charpentier and M G Stella, Proc. Int. Conf. ASSP, IEEE, Tokyo, 1986, pp. 2015-2018. Sections of speech waveform each representing a diphone are stored, along with pitchmarks which (for voiced speech) coincide in time with the greatest peak of each pitch period of the waveform and thus correspond roughly to the instant of glottal closure of the speaker; or are arbitrary for unvoiced speech.

A waveform portion to be used is divided into overlapping segments using a Hamming window having a length equal to three times the pitch period. A global spectral envelope is obtained for the waveform, and a short term spectral envelope obtained using a Discrete Fourier transform; a "source component" is obtained which is the short term spectrum divided by the spectral envelope. The source component then has its pitch modified by a linear interpolation process and it is then recombined with the envelope information. After preprocessing in this way the segments are concatenated by an overlap-add process to give a desired fundamental pitch.

Another proposal dispenses with the frequency-domain preprocessing and uses a Hanning window of twice the pitch period duration ("A Diphone Synthesis System based on Time-domain Prosodic Modification of Speech", C. Hamon, E. Moulines and F. Charpentier, Int. Conf. ASSP, Glasgow, 1989, pp. 238-241).

As an alternative to applying the time-domain overlap-add process to a complete speech signal it may be applied to an excitation component, for example by using LPC analysis to produce a residual signal (or a parametric representation of it) and applying the overlap-add process to the residual prior to passing it through an LPC synthesis filter (see "Pitch-synchronous Waveform Processing Techniques for Text-to Speech Synthesis using Diphones", F. Charpentier and E. Moulines, European Conference on Speech Communications and Technology, Paris, 1989, vol. II, pp. 13-19).

The basic principle of the overlap-add process is shown in Figure 1 where a speech signal S is shown with pitch marks P centred on the excitation peaks; it is separated into overlapping segments by multiplication by windowing waveforms W (only two of which are shown). The synthesised waveform is generated by adding the segments together with time shifting to raise or lower the pitch with a segment being respectively occasionally omitted or repeated.

According to the present invention there is provided a speech synthesis apparatus including means controllable to vary the pitch of speech signals synthesised thereby, having:

- (i) means for separating the speech signals into a spectral component and an excitation component;
- (ii) means for multiplying the excitation component by a series of overlapping window functions synchronous, in the case of voiced speech, with pitch timing mark information

corresponding at least approximately to instants of vocal  
excitation, to separate it into windowed speech segments;  
(iii) means to apply a controllable time-shift to the  
segments and add them together; and  
5 (iv) means for recombining the spectral and excitation  
components wherein the multiplying means employs at least  
two windows per pitch period, each having a duration of  
less than one pitch period. Preferably the windows consist  
of first windows, one per pitch period, employing the  
10 timing mark portions and a plurality of intermediate  
windows, and the intermediate windows each have a width  
less than that of the first windows.

In another aspect the invention provides a speech  
synthesis apparatus including means controllable to vary  
15 the pitch of speech signals synthesised thereby, having:  
(i) means for separating the speech signals into a  
spectral component and an excitation component;  
(ii) means for temporal compression/expansion of the  
excitation component, by interpolating new signal samples  
20 from input signal samples; and  
(iii) means for recombining the spectral and excitation  
components wherein the multiplying means employs at least  
two windows per pitch period, each having a duration of  
less than one pitch period. Preferably the  
25 compression/expansion means is operable in response to  
timing mark information corresponding at least  
approximately to instants of vocal excitation to vary the  
degree of compression/expansion synchronously therewith  
such that the excitation signal is compressed/expanded less  
30 in the vicinity of the timing marks than it is in the  
centre of the pitch period between two consecutive such  
marks.

Some embodiments of the invention will now be  
described, by way of example, with reference to the  
35 accompanying drawings.



In the apparatus of Figure 2, portions of digital speech waveform S are stored in a store 100, each with corresponding pitchmark timing information P, as explained earlier. Waveform portions are read out under control of  
5 a text-to-speech driver 101 which produces the necessary store addresses; the operation of the driver 101 is conventional and it will not be described further except to note that it also produces pitch information PP. The excitation and vocal tract components of a waveform portion  
10 read out from the store 100 are separated by an LPC analysis unit 102 which periodically produces the coefficients of a synthesis filter having a frequency response resembling the frequency spectrum of the speech waveform portion. This drives an analysis filter 103 which  
15 is the inverse of the synthesis filter and produces at its output a residual signal R.

The LPC analysis and inverse filtering operation is synchronous with the pitchmarks P, as will be described below.

20 The next step in the process is that of modifying the pitch of the residual signal. This is (for voiced speech segments) performed by a multiple-window method in which the residual is separated into segments in a processing unit 104 by multiplying by a series of overlapping window  
25 functions, at least two per pitch period; five are shown in Figure 3, which shows one trapezoidal window centred on the pitch period and four intermediate triangular windows. The pitch period windows are somewhat wider than the intermediate ones to avoid duplication of the main  
30 excitation when lowering the pitch.

When raising the pitch, the windowed segments are added together, but with a reduced temporal spacing, as shown in the lower part of Figure 3; if the pitch is lowered, the temporal spacing is increased. In either  
35 case, the relative window widths are chosen to give overlap of the sloping flanks (i.e. 50% overlap on the intermediate

windows) during synthesis to ensure the correct signal amplitude. The temporal adjustment is controlled by the signals PP. Typical widths for the five windows are 2 ms. The use of multiple windows is thought to reduce phase distortion compared with the use of one window per pitch period. After the temporal processing, the residual is passed to an LPC filter 105 to re-form the desired speech signal.

The store 100 also contains a voiced/unvoiced indicator for each waveform portion, and unvoiced portions are processed by a pitch unit 104' identical to the unit 104, but bypassing the LPC analysis and synthesis. Switching between the two paths is controlled at 106. Alternatively, the unvoiced portions could follow the same route as the voiced ones; in either case, arbitrary positions are taken for the pitch marks.

As an alternative to overlap-add on the residual, another algorithm has been developed which aims to retain the shape of the residual, and further reduce phase distortion which may result from shifting and overlap-adding. The basic principle is to alter the pitch period by resampling the open phase, (that is to say, a portion of the waveform between pitchmarks, leaving the significant information in the vicinity of the pitchmark unchanged) retaining the high frequencies injected at closure and giving a more realistic overall shape to the excitation period. Typically 80% of the period may be resampled.

Resampling is achieved by mapping each sample instant at the original sampling rate to a new position on the time axis. The signal amplitude at each sampling instant for the resampled signal is then estimated by linear interpolation between the two nearest mapped samples. Linear interpolation is not ideal for resampling, but is simple to implement and should at least give an indication of how useful the technique could be. When downsampling to reduce the pitch period, the signal must be low-pass

filtered to avoid aliasing. Initially, a separate filter has been designed for each pitch period using the window design method. Eventually, these could be penetrated by table lookup to reduce computation.

5       As a further refinement, the resampling factor varies smoothly over the segment to be processed to avoid a sharp change in signal characteristics at the boundaries. Without this, the effective sampling rate of the signal would undergo step changes. A sinusoidal function is used,  
10       and the degree of smoothing is controllable. The variable resampling is implemented in the mapping process according to the following equation:

$$T(n) = n \left( \frac{N-1}{M-1} \right) - \alpha \left( \frac{N-1}{M-1} \right) \cos \left[ \frac{\pi(n-1)}{M-3} \right], n = 1, \dots, M-2$$

$$T(0) = 0$$

15        $T(M-1) = N-1$

where

$M$  = number of samples of original signal

$N$  = number of samples of new signal

$\alpha = [0, 1]$  controls the degree of smoothing

20        $T(n)$  = position of the  $n$ 'th sample of the resampled signal.

A major difference between this and single window overlap-add is that the change in pitch period is achieved without overlap-add of time-shifted segments, provided that  
25       the synthesis pitchmarks are mapped to consecutive analysis pitchmarks. If the pitchmarks are not consecutive, overlap-add is still required to give a smooth signal after resampling. This occurs when periods are duplicated or omitted to give the required duration.

30       An alternative implementation involves resampling of the whole signal rather than a selected part of each pitch period. This presents no problems for pitch raising provided that appropriate filtering is applied to prevent aliasing, since the harmonic structure still occupies the

whole frequency range. When lowering pitch, however, interpolation leaves a gap at the high end of the spectrum. In a practical system aimed at telephony applications, this effect could be minimised by storing and processing the speech at a higher bandwidth than 4kHz (6kHz for example).  
5 The "lost" high frequencies would then be mostly out of the telephony band, and hence not relevant.

Both variations of the resampling technique suffer from the high computational requirements associated with interpolation/decimation, particularly if the resampling  
10 factor is not a ratio of two integers. The technique will become more attractive with continuing development of DSP technology.

Returning to the LPC analysis, as mentioned above, this is synchronous with the pitch markings. More  
15 particularly, one set of LPC parameters is required for each pitchmark in the speech signal. As part of the speech modification process, a mapping is performed between original and modified pitchmarks. The appropriate LPC  
20 parameters can then be selected for each modified pitchmark to resynthesise speech from the residual.

In LPC techniques, discontinuities can occur in the synthesised speech due to abrupt changes in the parameters at frame boundaries. This can result in clicks, pops, and  
25 a general rough quality, all of which are perceptually disturbing. To minimize these effects, LPC parameters are interpolated at the speech sampling rate in both analysis and synthesis phases.

The LPC analysis may be performed using any of the conventional methods, when using covariance or stabilised  
30 covariance method, each set of LPC parameters would be obtained for a section of the speech portion (analysis frame) of length equal to the pitch period (centred on the midpoint of the pitch period rather than on the pitch  
35 mark), or alternatively longer, overlapping sections might

be used which has the advantage of permitting the use of an analysis frame of fixed length according to pitch.

Alternatively with an autocorrelation method, a windowed analysis frame is preferred, as shown in Figure 4.

5        Although the frames in Figure 4 are shown with a triangular window for clarity, the choice of window function actually depends on the analysis method used. For example, a Hanning window might be used. The frame centre is aligned with the centre of the pitch period, rather than  
10       the pitchmark. The purpose of this is to reduce the influence of glottal excitation on the LPC analysis without resorting to closed-phase analysis with short frames. As a result, each parameter set is referenced to the period centre rather than the pitchmark. The frame length is  
15       fixed, as this was found to give more consistent results than a pitch-dependent value.

With short frame lengths, the stabilised covariance method would be preferable in terms of accuracy. With the longer frames used here, no perceptual difference is  
20       observed between the three methods, so the autocorrelation method is preferred as it is computationally efficient and guaranteed to give a stable synthesis filter.

Having determined the LPC parameters, the next step is to inverse filter the speech on a pitch-synchronous basis.  
25       As mentioned above, the parameters are interpolated to minimise transients due to large changes in parameter values at frame boundaries. At the centre of each pitch period, the filter corresponds exactly to that obtained from the analysis. At each sampling instant between  
30       successive period centres, the filter is a weighted combination of the two filters obtained from the analysis. Preferably the interpolation is applied directly to the filter coefficients. This has been shown to produce less spectral distortion than other parameters (LAR's, LSP's  
35       etc), but is not guaranteed to give a stable interpolated

filter. No instability problems have been encountered in practice.

In general, at sample  $n$  the filter coefficients are given by

5 
$$a_n(i) = \alpha_n a_l(i) + (1 - \alpha_n) a_r(i), \quad i = 0, \dots, P$$

where  $p$  is the order of the LPC analysis,  $\alpha_n$  is the value of a weighting function at sample  $n$ .  $a_l$  and  $a_r$  represent the parameter sets referenced to the nearest left and right period centres. To ensure a smooth evolution of filter coefficients, the weighting function is a raised half-cosine between successive period centres, given by

10 
$$\alpha(i) = 0.5 + 0.5 \cos(\pi i / N), \quad i = 0, \dots, N-1$$

where  $N$  is the distance between period centres, and  $i=0$  corresponds to the centre of each period.

15 The filter coefficients for the re-synthesis filter are calculated in the same way as for inverse filtering. Modifications to pitch and durations mean that the sequence of filters and the period values will be different from those used in the analysis, but the interpolation still ensures a smooth variation in filter coefficients from sample-to-sample.

For the first pitchmark in a voiced segment, filtering starts at the pitchmark and no interpolation is applied until the period centre is reached. For the last pitchmark in a voiced segment, the period is assumed to be the maximum allowed value for the purposes of positioning the analysis frame, and filtering stops at the pitchmark. These filtering conditions apply to both analysis and re-synthesis. When re-synthesising from the first pitchmark, the filter memory is initialised from preceding signal samples.

30 As a yet further alternative implementation of the pitch adjustment 104, a single-window overlap-add process may be used, with however a window width of less than two pitch period duration (preferably less than 1.7 e.g. in the range 1.25 - 1.6). With less than 100% overlap (i.e. 50%

each side) the window function necessarily has a flat top, moreover it is preferably asymmetrically located relative to the pitch marks (preferably embracing a complete period between two pitchmarks). A typical window function is  
5 shown in Figure 5, with a flat top having a length equal to the synthesis pitch period and flanks of raised half-cosine or linear shape.

With the window limited in duration as shown above, there is a potential problem when lowering pitch. When the  
10 synthesis pitchmarkers are sufficiently far apart, the windows will not overlap at all, and this situation will occur sooner with the shorter window than with standard pitch-synchronous overlap-add. The effect is to introduce a slight buzzy quality to the synthetic speech, but this  
15 only occurs when fairly extreme pitch lowering is requested by the TTS system. Pitch lowering is generally more difficult than pitch raising anyway, because of the need to generate missing data rather than cut out existing data. When raising pitch, the modified window produces better  
20 results due to the lower overlap period, and hence a shorter interval over which the signal is distorted.

This form of window is beneficial because a smaller temporal portion of the signal is constructed by the overlap-add process than with a longer window, and the  
25 asymmetric form places the overlap-add distortion towards the end of the pitch period where the speech energy is lower than immediately after the glottal excitation.

Use of the resampling and multi-window pitch control is envisaged (as shown in Figure 2) as operating on the  
30 residual signal (to avoid distortion of the formants), however, the short asymmetric window method may also be employed without separation of the spectrum and excitation, but directly on the speech signal.



CLAIMS

1. A speech synthesis apparatus including means controllable to vary the pitch of speech signals synthesised thereby, having:
  - 5 (i) means for separating the speech signals into a spectral component and an excitation component;
  - (ii) means for multiplying the excitation component by a series of overlapping window functions synchronous, in the case of voiced speech, with pitch timing mark information
  - 10 corresponding at least approximately to instants of vocal excitation, to separate it into windowed speech segments;
  - (iii) means to apply a controllable time-shift to the segments and add them together; and
  - (iv) means for recombining the spectral and excitation
  - 15 components;wherein the multiplying means employs at least two windows per pitch period, each having a duration of less than one pitch period.
2. A speech synthesis apparatus according to Claim 1 in
- 20 which the windows consist of first windows, one per pitch period, embracing the timing mark positions and a plurality of intermediate windows.
3. A speech synthesis apparatus according to Claim 2 in
- 25 which the intermediate windows each have a width less than that of the first windows.
4. A speech synthesis apparatus including means controllable to vary the pitch of speech signals synthesised thereby, having:
  - (i) means for separating the speech signals into a
  - 30 spectral component and an excitation component;





- (ii) means for temporal compression/expansion of the excitation component, by interpolating new signal samples from input signal samples; and
- (iii) means for recombining the spectral and excitation components.

5. A speech synthesis apparatus according to Claim 4, in which the compression/expansion means is operable in response to timing mark information corresponding at least approximately to instants of vocal excitation to vary the degree of compression/expansion synchronously therewith such that the excitation signal is compressed/expanded less in the vicinity of the timing marks than it is in the centre of the pitch period between two consecutive such marks.

6. A speech synthesis apparatus according to any one of claims 1 to 5 comprising:

- (a) a store containing items of data each defining a portion of speech signal waveform, and each including timing mark information corresponding at least approximately to the peak of the vocal excitation; and
- (b) driver means responsive to signals input thereto to provide addresses to read out items of data from the store and to provide pitch signals representing context-dependent pitch changes to be made to speech.

7. A speech synthesis apparatus according to any one of claims 1 to 6 in which the means for separating the spectral and excitation components comprises:

- (a) analysis means for receiving synthesised speech and generating parameters of a filter having a frequency response similar to the spectral content of the speech and of a filter having the inverse response; and
- (b) a filter connected to receive the parameters to filter the speech to produce a residual signal;





THIS SHEET HAS BEEN  
REPLACED

and the means for recombining them comprises;

(c) an inverse filter connected to receive the parameters and to filter the residual signal in accordance with the inverse response.

- 5 8. A speech synthesis apparatus -including means for  
controlling the pitch of a signal by multiplying the signal  
by a series of overlapping windows to separate it into  
segments and recombining the segments after subjecting them  
to a time shift, the windows being synchronous with timing  
10 marks representing instants of vocal excitation, wherein  
each window has a duration of less than twice the pitch  
period between timing marks and is asymmetric about the  
timing mark.

This Sheet Has Been Replaced

