Residual-Based Speech Modification Algorithms for Text-to-Speech Synthesis

M. Edgington and A. Lowry

BT Laboratories, Martlesham Heath, IPSWICH, IP5 7RE, U.K.
Email: mde@dwarf.bt.co.uk

ABSTRACT

This paper presents a set of novel algorithms for the signal modification component of concatenative text-to-speech systems. The algorithms described here are based around the LPC analysis/synthesis framework, and achieve prosodic modification by time-domain processing of the LPC residual. The modified residual is then recombined with the all-pole spectral estimate to synthesise the new speech signal.

The methods differ in the processing applied to the residual signal. The first method uses a modified version of TD-PSOLA, relying on assumptions of decorrelation and spectral flatness to avoid spectral distortion. The second method uses multiple windowing within each pitch period, enabling a given pitch modification to be realised by shifting several windowed segments by small amounts rather than a large shift of a single window. Again the aim is to reduce phase distortion introduced by the time-shifting process. The third method is based on a smoothly varying resampling of the residual, rather than windowed overlap-add.

TD-PSOLA and the residual-based methods were subject to informal listening tests both with pitch and time-scaled natural speech, and also integrated into the signal processing stage of the BT Laureate text-to-speech system.

1. INTRODUCTION

The basis of concatenative synthesis is to join short segments of speech, usually taken from a pre-recorded database, and then impose synthetic prosody (primarily pitch and duration) by appropriate signal processing. Both of these steps can introduce distortion to the synthetic speech:

- at the boundaries between speech segments by inappropriate selection or insufficient merging of segments, and
- by the prosodic modification process, due to an insufficiently robust speech modification model.

This paper restricts itself to the prosodic modification step.

The popular time-domain PSOLA (TD-PSOLA) method [1] is a simple and effective way of modifying the pitch and time-scale of speech, but is known to suffer from spectral and phase distortions. These are partly due to the time-domain nature of the processing, in that the spectral envelope cannot be adequately controlled. More complex methods, such as sinusoidal-model based approaches [2], are gaining in popularity, but tend to be computationally intensive, especially at the synthesis stage. Our aim here is to describe a set of algorithms that provide more flexibility, and less distortion, than time-domain techniques like TD-PSOLA, but are sufficiently simple to allow real-time implementation on non-specialised hardware.

The rest of this paper is split into four sections. Section 2 outlines the system framework and Section 3 describes the three residual modification algorithms in some detail. Section 4 presents some experimental results from a comparison of the methods with other speech modification methods.

2. SYSTEM FRAMEWORK

As an alternative to strictly time-domain techniques, the ubiquitous source-filter model of speech can be invoked. Prosody modification then becomes a task of separating the excitation and vocal tract components from the speech, modifying the excitation, and then recombining with the vocal tract component. In principle, this allows explicit control of the spectrum of the synthetic speech.

In an ideal (and non-existent) world, the analysis would separate out a physiologically motivated excitation signal, which could be modified independently of the tract response. In practice however, there is strong coupling between the excitation signal and tract response. As a practical compromise, the system attempts to separate the speech signal into a spectral shaping component, and an excitation (or residual) signal, which maintains much of the original signal's temporal detail. If the spectral estimate is accurate, then the excitation signal has the property of being spectrally flat, which gives a useful constraint on the excitation modification process.

Although there are a many well known techniques for spectral estimation, the algorithms described here are all based on the well known LPC analysis/synthesis framework [3]. Since LPC analysis attempts to produce an all-pole spectral model, the resynthesis filter is simply an all-pole filter, which is very easy to implement. Concatenative text-to-speech systems have the property that much of the original speech analysis can be performed as an off-line process, while the modification and synthesis steps are performed on-line, and the LPC model lends itself well to these requirements. As a further benefit, LPC analysis is a stage in many speech coding and compression algorithms, which can easily be incorporated into the speech analysis, reducing its storage requirements, although problems of tandeming with the speech modification algorithm must be considered.

2.1. Analysis

A set of LPC parameters is required for each pitch period in the original speech signal. This is achieved by pitch-synchronous analysis to derive an LPC model for each pitch period, followed by inverse filtering to generate the residual signal. (For unvoiced speech, equally spaced pseudo-pitchmarks are generated, and the prediction order is reduced.)

Analysis Framing. The analysis frames are centred at the midpoint between pitchmarks, as shown in Figure 1, rather than at the pitchmark itself. This reduces the effect of the glottal excitation, without having to resort to a closed-phase analysis and the associated problems of short data windows. Fixed length frames of twice the average pitch period were found to give more consistent results than pitch period dependent frame lengths.

Three analysis methods were investigated: autocorrelation, covarience and stabilised covarience. No perceptual difference in the resultant speech was detected, so autocorrelation was used as it guarantees a stable synthesis filter.

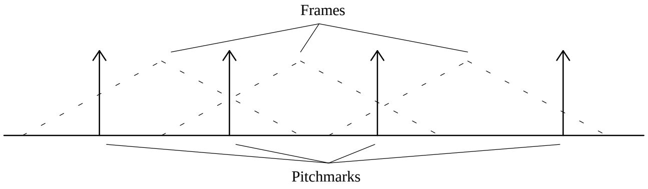

Figure 1: Pitch-synchronous framing centred on the midpoint of pitch periods. Triangular windows are shown for clarity; the actual window function depends on the analysis method used.

Parameter Interpolation for Inverse Filtering. In LPC techniques, discontinuities can occur in the synthetic speech due to abrupt changes in the parameters at frame boundaries. This often results in audible non-speech artefacts, such as clicks and pops, which are perceptually disturbing and greatly reduce the quality of the synthetic speech. To minimise these effects, the LPC parameters are interpolated at the speech sampling rate at both the analysis and synthesis phases. During analysis, inverse filtering is performed with a filter derived from a direct interpolation of the LPC parameters from consecutive frame centres using a raised cosine window. This has been found to produce no more distortion than interpolation of other parameters (e.g. LAR, LSP), but is not guaranteed to give a stable filter, although in this work no instability problems were encountered. In general, the estimated k^{th} filter coefficient a'_k for sample n is given by:

$$a'_k(n) = w\left(\frac{n-n_i}{n_{i+1}-n_i}\right)a^i_k + \left(1-w\left(\frac{n-n_i}{n_{i+1}-n_i}\right)\right)a^{i+1}_k \qquad (1)$$

where $a^i{}_k$ and n_i are the k^{th} filter coefficient and centre sample respectively, from frame i, the last frame before sample n, and $w(x)$ is the raised cosine window function, given by:

$$w(x) = 0.5 + 0.5\cos(x\pi) \qquad (2)$$

A Z-plane root loci plot of an example direct parameter interpolation between two frames is shown in Figure 2.

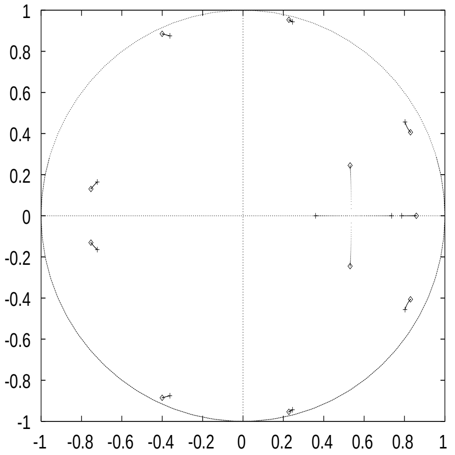

Figure 2: Z-plane root loci plot of direct filter parameter interpolation between two frame centres. The crosses and diamonds represent the roots at each frame centre.

2.2. Synthesis

The synthetic speech signal is regenerated by passing the modified residual signal through a synthesis filter. The coefficients of the synthesis filter are calculated in the same way as the inverse filter described above, using direct interpolation of the LPC parameters. As part of the speech modification process, a mapping is performed between original and modified pitchmarks, so the LPC parameters are selected from the appropriate pitch period in the original speech. Modifications to pitch and duration mean that the sequence of filters and pitch periods will in general be different from those used in the analysis, but interpolation still ensures a smooth variation in filter coefficients from sample-to-sample.

3. RESIDUAL MODIFICATION

The LPC residual (or error signal) has a number of advantages over the speech signal when it comes to modification: it is spectrally flat, and there is little correlation within each pitch period. The remainder of this section describes three algorithms which apply different processing to the residual signal.

3.1. LP-PSOLA

The most straightforward way to implement an LPC-based speech modification technique is to inverse filter the original speech signal, apply a time-domain process such as TD-PSOLA on the residual, and then resynthesise the signal. This approach is termed LP-PSOLA. Inverse filtering and resynthesis are linear operations, but as TD-PSOLA is a non-linear process the result is different to the time-domain equivalent. TD-PSOLA has a particular problem with pitch-lowering, since it cannot interpolate between time shifted sections of the original speech. The LP-PSOLA method offers an improvement, since the synthesis filter will produce an interpolated signal when the residual input is zero valued, giving a more realistic waveform.

Modified LP-PSOLA. In this approach a modified version of TD-PSOLA was used, which maintains the temporal structure of the original speech in the more important closed-phase of voiced speech, at the cost of a shorter overlap between consecutive windows. In a direct time-domain implementation, the adverse effects of the shorter overlap become apparent when synthesis pitchmarks become widely separated when lowering pitch, as there is a longer ‘null region’ where no signal is regenerated. Within the LP-PSOLA approach, this problem is alleviated, and the synthetic speech has a slightly clearer quality than LP-PSOLA with standard TD-PSOLA.

3.2. Multiple Window Processing

Each period of the LPC residual can be modified in duration by an overlap-add process involving several pseudo-pitchmarks within the period, in addition to the main one at glottal closure. This approach means that any given pitch modification is realised by shifting several windowed segments by small amounts, rather than a single large shift of a single window, reducing phase distortion introduced by the time-shifting process. The basic process is shown in Figure 3.

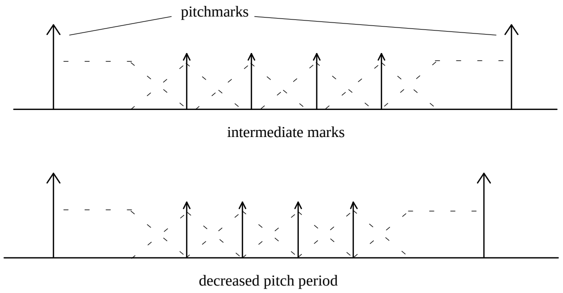

Figure 3: Pitch modification by multiple sub-period windowing; location of pseudo-pitchmarks within a pitch period.

It is important to avoid duplication of the main glottal excitation, so the intermediate marks are not spaced throughout the whole pitch period, but are concentrated towards the centre. The pitch modification is achieved by adjusting the spacing between the intermediate pseudo-pitchmarks only. Window lengths are chosen to give 50% overlap during synthesis, ensuring that the overlap-add signal has the correct amplitude.

3.3. Residual Resampling

The third technique described in this paper also aims to retain the shape of the residual signal, by reducing the phase distortion which can result from time shifting and overlap-adding. From analysis of residual signals, it is evident that they are not entirely decorrelated, but still contain structure related to the open/closed phases of the glottis. The resampling algorithm attempts to retain this structure by preserving the time-domain 'shape' of the residual within the period. The region of major excitation around glottal closure is not altered, and the resampling is applied to the latter part of each period, which ensures that the high frequencies injected by glottal closure are retained.

Resampling is achieved by mapping each sample instant at the original sampling rate to a position on the new time axis at the new sampling rate. The signal amplitude at each new sampling instant is then estimated by interpolation. Linear interpolation from the two nearest neighbours was used in this work, which although sub-optimal from a signal processing point of view, reduced the computational requirements of the implementation. When down-sampling to reduce the pitch period, the signal must be low pass filtered to avoid aliasing. The characteristics of the anti-aliasing filter are dependent on the ratio of the original and synthetic pitch periods, and can be pre-stored in a table to reduce on-line computation.

Smooth resampling. As a further refinement, the resampling factor varies smoothly during the segment 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, and the degree of smoothing is controllable by a single parameter. The variable resampling is implemented in the sample time-mapping function *T*(*n*), which gives the position of the nth sample at the new sampling rate:

$$\begin{aligned} &T(n) = n\left(\tfrac{N-1}{M-1}\right) - \alpha\left(\tfrac{N-M}{M-1}\right)\cos\left[\tfrac{\pi(n-1)}{M-1}\right] \\ &T(0) = 0 \\ &T(M-1) = N-1 \end{aligned} \qquad (3)$$

where M and N are the number of samples at the old and new sampling rates respectively, and α is the smoothing factor, ranging from 0 to 1. The main advantage of the resampling approach is that changes in pitch period are achieved without recourse to the overlap-add of time-shifted segments, and the associated phase distortion, provided that the synthesis pitchmarks are mapped to consecutive analysis pitchmarks. If the pitchmarks are not consecutive when periods are duplicated or omitted to produce a required duration, overlap-add is still required to give a smooth signal after resampling.

Continuous resampling. A slight variant of the approach described above involves resampling the whole signal, rather than a selected part of each pitch period. This presents no major problems for pitch raising, providing that appropriate filtering is in place to prevent aliasing, since the harmonic structure still occupies the whole frequency range. When lowering pitch, however, the interpolation is unable to fill in extra harmonics introduced at the high end of the spectrum. In a practical system aimed at band-limited applications, e.g. telephony, this effect can be minimised by simply storing and processing the speech at a higher bandwidth, so that higher harmonics are already present in the original speech and are scaled down into the required bandwidth by the resampling process.

4. RESULTS

The residual resampling technique was not considered to be sufficiently developed to evaluate, so the performance of TD-PSOLA, modified LP-PSOLA and multiple window processing were investigated by informal listening tests in two contexts:

1. applying a fixed pitch and time scaling to natural speech utterances
2. as the output stage of a text-to-speech system

The first of these is a useful development strategy, as modifications can be applied in a controlled manner to real speech, and any perceived degradation is due solely to the modification algorithm. The second approach is a more stringent test of the algorithm robustness, as quite severe scaling functions may be demanded which can vary rapidly during an utterance.

4.1 Fixed Scaling of Natural Speech

Figures 4 and 5 show the spectra from a male /i/ vowel sampled at 12kHz, after pitch-scaling by a factor of 1.3 by the TD-PSOLA and modified LP-PSOLA techniques respectively. This vowel was chosen as it is characterised by high second and third formants which are close together. According to [1], these formants may be broadened and possibly merged by TD-PSOLA, and this effect can be seen in Figure 4. The modified LP-PSOLA spectrum in Figure 5 does not show formant broadening to the same extent. Observation of the waveforms confirmed that TD-PSOLA produces markedly different waveforms from the residual-based techniques, as predicted, especially when the pitch is lowered. Listening tests revealed a slight preference for the modified LP-PSOLA and multiple-window methods over TD-PSOLA, but a formal subjective assessment would be needed to give any statistical validity to claims of improvement.

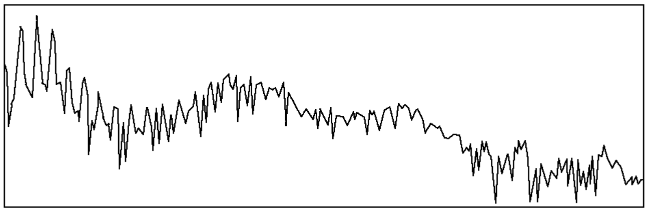

Figure 4: Spectrum of a male /i/ vowel sampled at 12kHz after pitch-scaling by a factor of 1.3 using TD-PSOLA. The spectrum shows characteristic formant merging between the second and third formant.

4.2 Text-to-speech Output

TD-PSOLA and the two residual-based techniques were integrated into the existing Laureate text-to-speech system developed at BT Laboratories [4]. A wide variety of material was synthesised, and again subjected to informal listening tests. Listeners did not report any gross differences between the methods, and rated the techniques very similarly in terms of "smoothness", which implies that LPC filter instability was not a problem. There was a slight preference for the two residual-based methods, with listeners reporting a "brighter" speech quality, possibly related to better preservation of the formant structure.

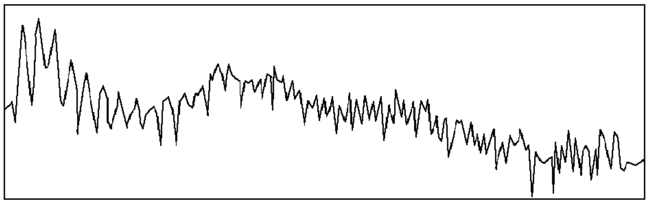

Figure 5: Spectra of the same /i/ vowel as Figure 4 after pitch-scaling by a factor of 1.3 using modified LP-PSOLA. The distinctness of the second and third formant is maintained.

5. CONCLUSIONS

A set of algorithms for the speech modification component of concatenative text-to-speech system has been presented. These algorithms are based on the LPC analysis/synthesis framework, and exploit properties of the LPC residual which could improve the speech modification performance. In speech coders, inaccurate coding of the LPC residual often dominate the quantisation noise, but this work has indicated that quite large modifications to the residual have not produced additional degradation. The problems of phase distortion prompted the development of a residual resampling technique, but at the current stage of development this technique does not give adequate performance to justify its computation requirements.

Informal listening tests indicate that there may be a preference for the residual-based methods over traditional TD-PSOLA. Residual-based methods also offer scope to model more aspects of human speech production, by explicit control of the spectral structure. Although the LPC methods require additional computation, the majority of this is in the analysis stage, which can be performed off-line and efficiently combined with compression of the speech database.

6. REFERENCES

1. Moulines, E., and Charpentier, F. "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones", *Speech Comm.*, Vol. 9, pp 453-467, Dec. 1990.
2. McAulay, R.J., and Quatieri, T.F., "Shape invariant time-scale and pitch modification of speech", *IEEE Trans. Signal Processing*, Vol. 40, pp 497-510, March 1992.
3. Makhoul, J., "Linear prediction: a tutorial review", *Proc. IEEE*, Vol. 63, pp 561-580, April 1975.
4. Page, J.H., and Breen, A.P., "The Laureate text-to-speech system - architecture and applications", *BT Technol. J.*, Vol. 14, pp 57-67, Jan. 1996.