WAVEFORM INTERPOLATION TECHNIQUE FOR TEXT-TO-SPEECH SYNTHESIS

Mikel Larreategui and Rolando A. Carrasco
School of Engineering, Staffordshire University
Beaconside, PO 333, ST18 ODF, Stafford, UK.
TEL: +44 1785 353366; FAX: +44 1785 353552
e-mail: mikel@staffs.ac.uk

ABSTRACT

The waveform interpolation (WI) technique has recently been proposed by Kleijn [5][6] for speech coding applications. However, there are no known published works in the open literature concerning the application of the WI method for high-quality text-to-speech (TTS) synthesis. The original contribution of this paper is to study and evaluate the performance of the WI technique in the context of TTS systems.

1 INTRODUCTION

In the last few years, the harmonic/stochastic (H/S) model has proven to be a successful approach for high-quality TTS [3]. The H/S model provides great flexibility for segment concatenation, voiced/ unvoiced (V/U) processing and prosodic modification. In order to estimate suboptimally the model parameters, several heuristic approaches have been proposed throughout the literature, such as the multiband excitation model (MBE) [4], the hybrid harmonic model [10], and sinusoidal transform coding. A common problem of these heuristic algorithms is their sensitivity to non-stationarities, as outlined by Dutoit in [2].

Alternatively, we have adopted the WI method, proposed by Kleign [5][6], to estimate the parameters of the H/S model. The main difference of the WI technique with respect to the previous heuristic approaches, lies in the length of the analysis segments. A WI analysis segment is one pitch period long, and therefore, it may be expected to give a more accurate representation of voiced speech and less affected by pitch-frequency variation and other non-stationarities. A second major difference is the way V/U decomposition is obtained. In the next section, the main characteristics of the WI technique for TTS synthesis are described and its performance evaluated and compared to other leading TTS systems.

2 THE WI TECHNIQUE

2.1 WI Analysis

A prototype waveform (PW) is a segment of speech of length equal to one pitch period (Figure 1). The PWs are extracted on a regular basis. At each update point, the pitch period and the pitch mark [8] is determined and the corresponding PW extracted from the sampled speech signal. For unvoiced segments, the value of the pitch period is arbitrary.

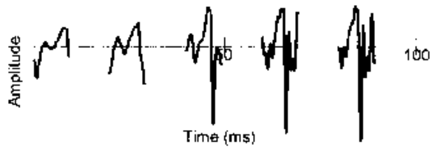

Figure 1. Prototype waveforms for WI.

The Fourier series (FS) coefficients of each PW are obtained using the DFT. However, prior to the estimation of the FS coefficients, it is convenient to apply the following transformation on each PW [5]:

1. *Spectral flattening*: the residual signal for each PW is obtained by means of LPC inverse filtering. The term prototype waveform will now refer to the prototype residual segment.

2. *Time alignment*: in the time alignment procedure (see Figure 2), the newly extracted prototype residual is circularly shifted so as to make its FS coefficients as close as possible to the FS coefficients of the previous update point. In practice, the PWs are aligned in such a way that the corresponding pitch mark is located at the origin of the current PW.

3. *Pitch normalisation*: the pitch-period of the PWs is normalised to a specific value K in order to have the same number of FS coefficients in all the PWs. The normalisation process is performed by a simple linear interpolation of the PW.

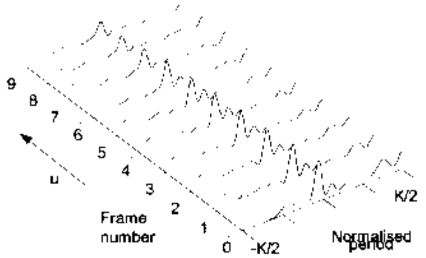

Figure 2: Alignment of prototype residual segments.

In this paper, the normalised pitch-period K is set to 128 samples (for f_s=10KHz). Note that this value allows the use of the more efficient FFT algorithm.

After performing the above operations, which maximise the cross-correlation between successive PWs, the uth prototype waveform $e(t_a(u),n)$ at the update point $t_a(u)$ is then analysed to produce the Fourier series:

$$e(t_a(u),n) = \sum_{k=-K/2}^{K/2} c_k(t_a(u)) e^{jk\sigma(n)} \qquad (1)$$

where $c_k(t_a(u))$ is the time-varying FS coefficients and $\sigma(n)$ is the instantaneous phase for the first harmonic. Note that the prototype waveform $e(t_a(u),n)$ is a two-dimensional signal where $t_a(u)$ indicates the position of the PW in the speech signal and n is the axis along which the shape of the waveform is displayed.

2.2 Estimation of the Voiced and Unvoiced Components

When the PWs are time aligned, as illustrated in Figure 2, the changes that occur to their corresponding FS coefficients are indicative of the nature of the speech signal. Rapid changes occur for unvoiced speech, slow changes for unvoiced speech. As a result, a simple linear filter can be used to separate the effects of these changes [5]. By high-pass filtering the FS coefficients $c_k(t)$ in the t direction, the unvoiced component is obtained. Similarly, low-pass filtering leads to the periodic component of the speech signal.

The analysis parameters for V/U decomposition are as follows: firstly, the cutoff frequency for both lowpass and highpass filters is 25 Hz [5]; the PW extraction rate R, i.e., $t_a(u)=uR$, is set to 1000 Hz; and finally, the number of filter coefficients is 21 (the sampling frequency for the speech signal is assumed to be f_s=10 KHz). After filtering, the lowpass filtered FS coefficients are down sampled to 100 Hz since rapid fluctuations have been eliminated. Similarly, the highpass filtered coefficients are also down sampled without losing any perceptually relevant information.

2.3 WI Synthesis

The synthesis procedure consists of producing, separately, both the voiced and the unvoiced residual signals, and finally combining them with the vocal tract envelope to generate the complete synthetic signal. During voiced synthesis, the instantaneous excitation waveform $e(t,n)$ and the pitch period, evolve slowly over time. Let the synthesis update time instants defining the boundaries of the present interpolation interval be $t_s(u$-1) and $t_s(u)$. Then, the linear interpolation of the prototype waveforms, $e(t_s(u$-1),$n)$ and $e(t_s(u),n)$, renders a reconstructed excitation waveform

$$\begin{aligned}\hat{e}(n+t_s(u-1)) &= \hat{\alpha}(n)e(t_s(u-1),n) \\ &\quad +\alpha(n)e(t_s(u),n-N_s(u))\end{aligned} \qquad (2)$$

for $0 \le n < N_s(u)$, where $N_s(u)=t_s(u)-t_s(u-1)$ is the frame interval, $\alpha(n)$ is a monotonically increasing interpolation function which goes from $\alpha(t_s(u-1))=0$ to $\alpha(t_s(u))=1$, with $\hat{\alpha}(n) = 1-\alpha(n)$. The prototype waveforms, $e(t_s(u$-1),$n)$ and $e(t_s(u),n)$, are determined from the lowpass filtered FS coefficients and the instantaneous phase $\sigma(n)$. By definition, the instantaneous phase is given by the following differential equation:

$$d\sigma = 2\pi / \hat{p}(n)\, dn \qquad (3a)$$

where $\hat{p}(n) = \hat{\alpha}(n)p(t_s(u-1)) + \alpha(n)p(t_s(u))$ (3b)

which is the interpolated pitch period. The evolution of the instantaneous phase $\sigma(n)$ can be obtained by integrating the above differential equation:

$$\begin{aligned}\sigma(n) &= \frac{2\pi N_s(u)}{p(t_s(u)) - p(t_s(u-1))} \times \\ &\quad \ln\left[\frac{p(t_s(u)) - p(t_s(u-1))}{N_s(u)p(t_s(u-1))} n + 1\right]\end{aligned} \qquad (4)$$

If the pitch period remains constant over the interpolation interval, then the instantaneous phase is simply expressed as $\sigma(n) = 2\pi n / p(t_s(u))$.

As for the unvoiced component, a set of random bandpass signals, denoted as *narrow band basis functions* (NBBF) [10], are used to model the spectra contour of the highpass filtered FS coefficients.

2.4 Speech Modification

The speech modification module includes both time-scale and pitch-scale modification algorithms to produce the desired prosody pattern in the synthetic speech. Time-scale modification in the WI context can be easily performed by changing the synthesis frame interval $N_s(u)$. As for pitch scaling, a magnitude envelope and an unwrapped phase envelope are first obtained from the discrete spectrum specified by the lowpass filtered FS coefficients. Then, the new complex FS coefficients are estimated by sampling both envelopes at the new harmonic frequencies. In Figure 3 an example of speech modification is given.

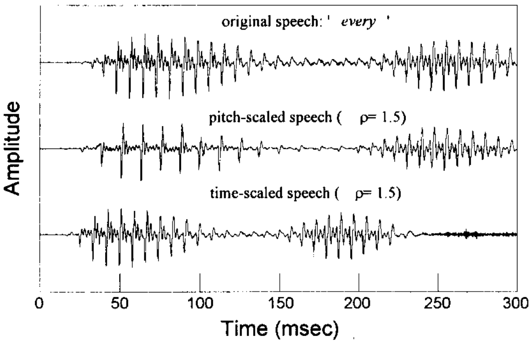

Figure 3: speech modification using the WI technique.

2.5 Phase Coherency

Due to the time alignment process (Figure 2) and the changes in both pitch and time scales, phase coherency between successive frames is lost. A lack of phase coherency leads to reverberant synthetic speech. In order to keep phase coherency, the notion of *pitch pulse onset time*, introduced by Quatieri and McAulay in [12], is used. The essence of this approach is to calculate a new set of synthesis pitch pulses taking into account the pitch and time-scale modification factors. At the synthesis stage, the PWs are aligned according to the new calculated synthesis pitch pulses. In practice, for a given synthesis frame, this alignment is performed by modifying the phase of the FS coefficients with a linear function, the slope of which is dependent on the offset of the corresponding synthesis pitch pulse. This can be expressed as

$$\arg(c_k'(t_s(u))) = \arg(c_k(t_s(u))) - n_o(u)\frac{2\pi k}{p(t_s(u))} \quad (5)$$

where $n_o(u)$ is the location of the synthesis pitch pulse relative to the origin of the current frame.

3 EVALUATION AND COMPARISON

In order to evaluate the performance of the WI-based TTS system, a comparative study, which includes a total number of four leading synthesis algorithms, was carried out. For all the investigated synthesisers, the same segments database and input data (diphones and prosody) were used.

3.1 Synthesis Candidates

1. *The linear prediction (LP) model*: the classical LP TTS system was implemented, with a predictor order of 12 [9]. Synthesis was performed with an all-pole filter, the coefficients of which were updated every 6.4 ms. Speech segments were concatenated by linearly interpolating the LSP parameters. The LP model constitutes the most inexpensive algorithm in terms of both computational load and database size, and was taken as ground quality.

2. *The H/S model*: decomposition of the speech segments into both voiced and unvoiced components was performed with the MBE model [4]. At the synthesis stage, the harmonic amplitudes were linearly interpolated over successive frames, while the harmonic phases used a cubic polynomial as an interpolation function [12]. As for the unvoiced synthesis, a set of narrow band basis functions, which model the contour of the stochastic power spectra, were used.

3. *The TD-PSOLA*: this non-parametric "model" has drawn considerable attention due to its exceptional efficiency and simplicity for speech synthesis and prosody modification [11]. However, it does not have any segment concatenation capability to minimise mismatches at the segment boundaries [1], resulting in a decrease in the quality performance. For database compression, the Regular Pulse Excitation approach is used [7].

3.2 Quality Assessment

Two kinds of quality assessments were performed, namely, the CVC test to assess intelligibility and the MOS test to evaluate the naturalness of the synthesisers, as well as the segment concatenation efficiency [3]. For the CVC test, fifty phonetically balanced CVC nonsense words were synthesised by each TTS system and ten listeners were asked to identify the phonemes. As for the MOS test, seven long sentences were generated by each system and ten listeners were asked to rate the naturalness according to the MOS punctuation scale. Results are shown in Figure 4 and discussed in the next section.

3.3 Discussion

In Figure 4, the results of both the intelligibility and naturalness tests are depicted for each of the four synthesisers. As expected, the CVC results show that the worse performance corresponds to the LPC synthesiser. A closer examination of the CVC tests shows that vowels were correctly identified, whereas most of the consonant phonemes were completely misunderstood. MBE and TD-PSOLA have got similar behaviour, with a slight advantage for MBE. This can be attributed to the fact that MBE has got better concatenation capabilities than TD-PSOLA. Finally, WI is the most intelligible synthesis method. The superiority of WI over MBE is most likely due to the fact that WI is less sensitive to non-stationarities than MBE and, hence, V/U decomposition is more robust in the WI synthesiser.

Finally, the MOS results can be appreciated to follow basically the same trend as CVC ones, except that TD-PSOLA sounds more natural than MBE. This can be attributed to the fact that TD-PSOLA is much less sensitive to V/U analysis errors than MBE.

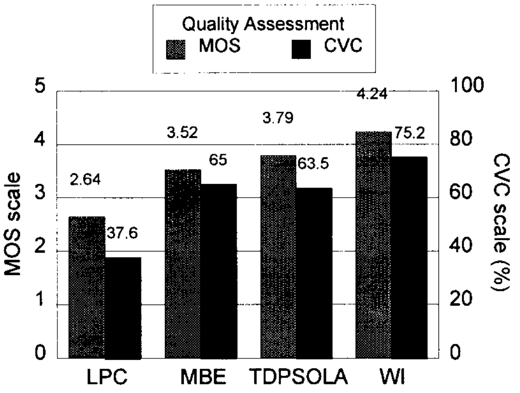

Figure 4: Quality assessment

4 CONCLUSIONS

In this paper, the application of the WI technique in the context of TTS synthesis has been addressed. The quality assessment tests demonstrated the superior performance of the WI over leading synthesisers, such as MBE and TD-PSOLA. However, the main drawback concerning WI is the computational load required for speech analysis and synthesis.

ACKNOWLEDGEMENTS

The authors wish to thank the Spanish blind organisation ONCE for their financial support.

REFERENCES

[1] T. Dutoit and H. Leich, 'Improving the TD-PSOLA text-to-speech synthesizer with a specially designed MBE re-synthesis of the segments database', *Proc. EUSIPCO 92, pp. 25-28, 1992.*

[2] T. Dutoit and H. Leich, 'An analysis of the performances of the MBE model when used in the context of a text-to-speech system', *Proc. EUROSPEECH 93, Berlin, pp. 531-534, 1993.*

[3] T. Dutoit, 'High quality text-to-speech synthesis: a comparison of four candidate algorithms', *ICASSP'94, vol. 1, pp. 565-568, 1994.*

[4] D. W. Griffin and J. S. Lim, 'Multiband excitation vocoder', *IEEE Trans. on ASSP, vol. 36, pp. 1223-1235, Aug. 1988.*

[5] W. B Kleijn and J. Haagen, 'A General Waveform-Interpolation Structure for Speech Coding", *EUSIPCO'94, Edinburgh, 1994.*

[6] W. B. Kleijn and J. Haagen, 'A Speech Coder Based on Decomposition of Characteristic Waveforms', *Proc. ICASSP'95, pp. 508-511, 1995.*

[7] P. Kroon, E. F. Deprettere and R. J. Sluyter, 'Regular-pulse excitation- A novel approach to effective efficient multipulse coding os speech', *IEEE Trans. on ASSP, vol. 34, no. 5, pp. 1054-63, 1986.*

[8] M. Larreategui, F. J. Ancin and R. A. Carrasco, 'An improved epoch detection algorithm based on sinusoidal modelling of speech', *Eurospeech '95, Madrid, vol. 1, 1995.*

[9] J. D. Markel and A. H. Gray, 'Linear prediction of speech', *Springer Verlag, New York, 1976* .

[10] J. S. Marques and A. J. Abrantes, 'Hybrid Harmonic Coding of Speech at Low Bit-Rates', *Speech Communication, 14, pp. 231-247, 1994.*

[11] E. Moulines and F. Charpentier, 'Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones', *Speech Communication, vol. 9, n. 5-6, 1989.*

[12] T. F. Quatieri and R. McAulay, "Shape Invariant Time-Scale and Pitch-Scale Modification of Speech", *IEEE Trans. on ASSP, vol. 40, no. 3, 1992.*