Reading note

Moulines & Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones" (1990)

TSIA 206 – Speech and audio processing

Reference. E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones", *Speech Communication* 9(5/6), December 1990, pp. 453–467.

1 Context

At the time the paper was written, concatenative text-to-speech relied on assembling short pre-recorded units (typically diphones) into full sentences. The dominant method, LPC synthesis, modified pitch and duration through the parameters of an excitation model that was too simplified: a pulse train for voiced sounds and white noise for unvoiced ones produces speech that, in the authors' words, sounds "indistinct or fuzzy". The contribution of the paper is to separate decoding from prosodic modification. The diphone is first reconstructed as a natural waveform; the prosody is then changed directly on the waveform using a family of pitch-synchronous overlap-add (PSOLA) methods.

2 The PSOLA framework

The algorithm proceeds in three stages.

Analysis. The signal $x(n)$ is labelled with pitch marks t_m. On voiced regions one mark per glottal pulse; on unvoiced regions a constant rate of roughly one mark every 10 ms. Around each t_m a short-term signal is extracted by windowing,

$$x_m(n) = h_m(t_m - n)\, x(n),$$

with a Hann window h_m of length proportional to the local pitch period P, typically between $2P$ and $4P$.

Modification. A new stream of synthesis short-term signals $\tilde{x}_q$ is placed at new marks $\tilde{t}_q$. A mapping $\tilde{t}_q \mapsto t_m$ chooses, for each output mark, which input grain to copy. The pitch factor β and time factor γ determine the spacing of the synthesis marks and their total number. In TD-PSOLA the grain itself is left unchanged: only its placement varies.

Synthesis. The grains are summed back together. The paper presents a general least-squares formula due to Griffin and Lim and a simpler weighted-sum formula due to Allen. When the window length equals $2P$, the normalisation factor in the Allen formula becomes constant, and the synthesis reduces to

$$\tilde{x}(n) = \sum_q \tilde{x}_q(n).$$

This is the form we use in the TP.

3 Time-scale and pitch-scale modification

For a time factor γ, the synthesis marks are placed such that the warp $t \mapsto \gamma t$ takes them onto analysis marks. Slowing down ($\gamma > 1$) makes the algorithm duplicate some grains; speeding up causes deletions. The spacing between consecutive output marks stays equal to the local period, so the pitch is preserved.

For a pitch factor β, the marks are spaced by P/β rather than P, and the mapping between analysis and synthesis marks is one-to-one. The number of grains and the total duration are unchanged.

When both factors are applied, the construction splits into two consecutive warps (Fig. 2 of the paper): a joint (β, β) modification followed by a compensating γ/β time modification. In practice the two warps combine into a single mapping computed in one pass.

4 Distortions of TD-PSOLA

Section 2 of the paper is its analytical core. The authors model voiced speech as $d(m) + n(m)$ with d deterministic and periodic and n wide-sense stationary, then study what TD-PSOLA does to each component.

For the deterministic part, the Poisson summation formula yields

$$\tilde{d}(m) = \frac{1}{\beta P} \sum_k X_0(\Omega_k)\, e^{j\Omega_k m}, \qquad \Omega_k = \frac{2\pi k}{\beta P}.$$

The Fourier coefficients are samples of the original prototype spectrum at the new harmonic frequencies, so the spectral envelope is preserved. Formants do not move when pitch changes – which is precisely the property one wants from a pitch modifier.

For the stochastic part the picture is less clean. The autocorrelation of the synthesised noise is a periodic duplication of the original autocorrelation with period $(1-\beta)P$, and the power spectrum becomes a comb of shifted copies of $|H(\omega)|^2$. The synthesised noise is no longer white: a pseudo-harmonic structure appears that is audible as tonal noise, particularly on voiced fricatives.

The behaviour depends critically on the window length relative to P. The paper distinguishes two regimes.

Narrow band ($L \geq 4P$). The window resolves individual harmonics. Synthesis harmonics that do not coincide with original ones are attenuated, with the worst attenuations at multiples of

$$f_{\text{att}} = \frac{F_0\, \beta}{2|\beta - 1|}.$$

Even mild factors such as $\beta = 1.05$ create a deep notch near 1 kHz for a pitch of 100 Hz. The perceptual effect is described as reverberation.

Wide band ($L < 2P$). The window is too short to resolve harmonics. The bandwidth of formant peaks in the synthesised spectrum is dominated by the window rather than by the vocal tract, which broadens the formants. The paper points out that this is perceptually milder than narrow-band distortion: the just-noticeable difference for formant bandwidth is about 40 % on steady vowels. Our TP works in this regime ($L = 2P + 1$), which is also the regime under which the simplified synthesis formula above applies exactly.

A subtler issue concerns the alignment of pitch marks with the instant of glottal closure. With even-symmetric windows, the synthesised phase at each formant depends on the offset D between mark and excitation pulse. Experimentally the synthesised signal starts to sound hoarse when

D exceeds 30 % of the period. The TP avoids this question – the assignment explicitly states that marks are not aligned on glottal pulses – and accepts the resulting degradation.

5 FD-PSOLA and LP-PSOLA

To eliminate the narrow-band distortion, the authors propose performing the modification in the frequency domain. The DFT of each grain is resampled to the new pitch grid either by a linear frequency warp (compression–expansion) or by repeating and eliminating whole harmonics (elimination–repetition). The first technique is phase-coherent but mixes local spectral properties from one frequency region to another; the second preserves locality but breaks phase coherence. A hybrid using compression–expansion in the low band and elimination–repetition in the high band is suggested. FD-PSOLA is the most flexible variant but also the most expensive: in 1990 the authors quote 5 Mflop/s and 5 MB of memory.

LP-PSOLA applies TD-PSOLA to the LPC residual rather than to the speech signal. Because the residual is approximately flat, the spectral broadening of wide-band TD-PSOLA on the residual is much less harmful than on the signal, and the final LPC filter restores the envelope. Combined with low-rate residual coding (MPLPC, CELP), this gives an efficient real-time alternative.

6 Limitations and link to the TP

The paper's main weakness is the experimental evaluation: a single A–B preference test, 16 subjects, 10 French sentences, no per-phoneme breakdown. The PSOLA variants beat improved-excitation LPC, but the theoretical distortions of Section 2 are not measured directly. The method also assumes accurate pitch marks; with autocorrelation-based pitch detection the precision floor is on the order of one millisecond, which is close to the 30 % threshold for the glottal-closure offset, so the constraint is tight in practice. The model is locally stationary and does not handle plosive bursts well.

The TP implements TD-PSOLA in the wide-band regime, with $L = 2P + 1$ and the simplified synthesis formula. The function `AnalysisPitchMarks` extracts the marks of Section 1.1. The function `Synthesis` computes the OLA sum. `ChangeTimeScale` produces a constant warp $\gamma = 1/\alpha$ (the TP uses α for the duration ratio whereas the paper uses γ). `ChangePitchScale` produces the dual construction with $\beta_{\mathrm{TP}} = 1/\beta_{\mathrm{paper}}$ (the TP multiplies f_0 by β whereas the paper multiplies the period). The voiced/unvoiced distinction in `ChangePitchScale` is important: applying a non-unit scale to unvoiced grains would introduce exactly the pseudo-periodic structure with period $(1 - \beta)P$ that Section 2 identifies as the main TD-PSOLA artefact.