AN IMPROVED HYBRID TIME-FREQUENCY ALGORITHM FOR TIME-SCALE MODIFICATION OF SPEECH/AUDIO SIGNALS

Cristian NEGRESCU, Amelia CIOBANU, Dragoș BURILEANU, Dumitru STANOMIR

Time-Scale Modification
Basic approaches

• Time domain
  – Eg. SOLA (Synchronized Overlap-Add) family
  – Simple, efficient, good results for correlated signals

• Hybrid approach – Eg. The algorithm in the present paper

• Frequency domain
  – Eg. Phase vocoder, Sinusoidal Model
  – Good results for signals with complex spectral composition
Classical SOLA approach (for $\alpha<1$)

Classical SOLA approach for $\alpha<1$

- “Infamous” overlap-add:
  - Good results only for single pitch correlated signals
  - Frequency and phase discontinuities are preserved
  - Unnatural attenuation / amplification of some spectral components
  - The signal becomes dull
  - Generates beating effects
  - The sound becomes “rough”

**Solution?**
- Preserving segmentation, synchronization and concatenation
- Replacing OLA with a spectral smoothing procedure

The proposed hybrid algorithm

Frequency domain smoothing

Segmentation

Segment Alignment & Concatenation

Frame Selection

SM Analysis via IF

SM Synthesis

IF Attractors

Matching & Track Generation

Replacement

$\omega, A, \phi$

Time-scaled signal
The Signal’s Model

- **Traditional approach**
  - Sinusoidal Model (SM) based on STFT analysis.
    \[
    \hat{s}(n) = \sum_{p=1}^{P} s_p(n) = \sum_{p=1}^{P} A_p(n) \cos \theta_p(n), \quad n = 0, 1, \ldots, N_{sys}-1
    \]
  - Synthesis requires \( P \) controlled harmonic oscillators
  - **Control functions:**
    - Instantaneous amplitudes \( A_p(n) \)
    - Instantaneous phases \( \theta_p(n) = \omega_p(n) + \theta_p(0) \)
  - Smooth transitions requires interpolation (Quatieri/McCauley):
    - Amplitudes: linear interpolation based on \( A_p(0), A_p(N_{sys}) \)
    - Phases: constrained cubic polynomial interpolation based on \( \phi_p(0), \phi_p(N_{sys}) \)

- Traditional SM Analysis:
  - STFT decomposition
  - Peak picking
  - Track generation
Proposed approach for spectral smoothing

- **Traditional approach**
  - Sinusoidal Model (SM) based on STFT analysis.

- **Proposed algorithm**
  - Sinusoidal Model (SM) based on Instantaneous Frequency (IF) attractors.

- **Pros**
  - Improved accuracy in partial’s instantaneous frequency estimation
    - non-uniform frequencies grid
    - avoiding numerical computing of the derivative
  - Reducing the number of partials avoiding spurious local maxima introduced by analysis window

---

SM Analysis Using IF

- **SM model for the analog speech/audio signal**
  \[
  s(t) = \sum_{p=1}^{P} s_p(t) = A_p(t) \cos(\theta_p(t)) \left[ \sigma(t-t_{sp}) - \sigma(t-t_{ep}) \right]
  \]

- **Instantaneous frequency is defined as**
  \[
  \omega_p(t) = \frac{d}{dt} \theta_p(t) \Rightarrow \theta_p(t) = \int_{t_{sp}}^{t} \omega_p(\tau) \, d\tau + \phi_p
  \]

- **Task:** How can be estimated \( A_p(t), \omega_p(t), \phi_p \) from the shape of the corresponding \( s_p(t) \) given on a sample by sample basis?

- **A possible solution:**
  - build the analytical signal \( \tilde{s}_p(t) = s_p(t) + j \cdot H\{s_p(t)\} \)
  - extract the parameters
  \[
  A_p(t) = |\tilde{s}_p(t)|, \quad \omega_p(t) = \frac{d}{dt} \arg\{\tilde{s}_p(t)\}, \quad \phi_p(t) = \arg\{\tilde{s}_p(t)\} \bigg|_{t=0}
  \]

- **Subsequent problem:** How to extract a single partial \( s_p(t) \) from the whole mixture \( s(t) \)?
SM Analysis Using IF

- SM model for the analog speech/audio signal
  \[ s(t) = \sum_{p=1}^{P} s_p(t), \quad s_p(t) = A_p(t) \cos \left( \int \omega_p(\tau) d\tau + \phi_p \right), \quad t \in [t_{sp}, t_{ep}] \]

- Assumption: isolated partials

- SM Analysis using IF: - Band-pass filtering to extract \( s_p(t) \) from \( s(t) \)
  - Computing the analytical signal \( \tilde{s}_p(t) = s_p(t) + j \cdot H\{s_p(t)\} \)
  - Parameters extraction
    \[
    \omega_p(t) = \frac{d}{dt} \arg \{\tilde{s}_p(t)\} = \frac{d}{dt} \arctan \frac{\text{Im}\{\tilde{s}_p(t)\}}{\text{Re}\{\tilde{s}_p(t)\}}
    \]
    \[
    A_p(t) = |\tilde{s}_p(t)|, \quad \phi_p = \arg \{\tilde{s}_p(t)\}\bigg|_{t=0}
    \]

- Solution – Fixed BP filters with complete coverage
  - A bank of uniformly spaced bank of complex BP filters covers entire spectral domain
  - The bank is built by modulation of a causal and real LP prototype \( w(t) \)
  - Efficient implementation is performed via STFT analysis
SM Analysis Using IF

– Filtering stage –

• Goals
  – implementing a bank of equally-spaced BP filters to extract partials
  – computing the analytical $\tilde{s}_n(t)$ signals associated to the partials
• STFT is performed applying Fourier transform to the windowed signal
  $$S(\omega, \tau) = \int_{-\infty}^{\infty} s(t) w(\tau - t) e^{-j\omega t} dt, \quad \int_{-\infty}^{\infty} w^2(t) dt = 1$$
  
• For a fixed frequency, $\Omega_n$, and considering time as a variable STFT is represented in terms of linear filtering
  $$S(\Omega_n, t) = \left[ s(t) e^{-j\Omega_n t} \right] * w(t)$$
  
• We build an additional signal
  $$s_{f, \Omega_n}(t) = S(\Omega_n, t) e^{j\Omega_n t}$$
  
  • The output signal is an analytical signal!
  
• If a partial falls inside the bandwidth of the filter, the output is an estimator for the analytical signal associated to the respective partial!

SM Analysis Using IF

– IF computing –

• Goals
  – accurate estimation of IF
  – estimation of amplitudes and phases
  
  • Instantaneous frequency (IF) estimation
  $$\omega_n(t) = \frac{\dot{R}_{\Omega_n}(t) - R_{\Omega_n}(t) \dot{I}_{\Omega_n}(t)}{R_{\Omega_n}^2(t) + I_{\Omega_n}^2(t)}$$

  • Can we avoid time derivative of signals?
  $$\dot{s}_{f, \Omega_n}(t) = \frac{d}{dt} \left[ S(\Omega_n, t) e^{j\Omega_n t} \right] = e^{j\Omega_n t} \dot{S}(\Omega_n, t) + j\Omega_n S(\Omega_n, t) e^{j\Omega_n t}$$

  $$\dot{S}(\Omega_n, t) = \frac{d}{dt} \left[ s(t) e^{-j\Omega_n t} \right] * w(t) = \left[ s(t) e^{-j\Omega_n t} \right] * \dot{w}(t)$$

  $$\ddot{A}_n(t) = \lim_{\Delta \omega \to 0} \frac{1}{2\pi} \int_{\omega_n(t)}^{\omega_n(t) + \Delta \omega} S(\omega, t) e^{j\omega t} d\omega \quad A_n(t) = |\ddot{A}_n(t)|$$

  $$\phi_n(t) = \text{arg}\{\ddot{A}_n(t)\}$$
SM Analysis Using IF
– IF computing –

**Goals**
- accurate estimation of IF
- estimation of amplitudes and phases

**Remark:** We intend to apply the process for discrete time
\[
\omega_n(t) = \frac{I_{\Omega_n}(t)R_{\Omega_n}(t) - R_{\Omega_n}(t)I_{\Omega_n}(t)}{R_{\Omega_n}^2(t) + I_{\Omega_n}^2(t)}
\]
\[
\hat{A}_n(t) = \lim_{\Delta \omega \to 0} \frac{1}{2\Delta \omega} \int_s(t) e^{i\omega \Delta \omega} d\omega
\]

Amplitude and phase will be computed by interpolation based on adjacent STFT bins.

**IF Attractors**

**Goal**
- Selecting the partials avoiding spurious spectral peaks

**IF attractors are the frequencies that satisfy:**
\[
\mu(\Omega_n, t) = 0 \quad \mu(\Omega_n, t) = \omega_n(t) - \Omega_n
\]
\[
\frac{\partial \mu(\Omega_n, t)}{\partial \Omega_n} = \frac{\partial \omega_n(t)}{\partial \Omega_n} - 1 < 0
\]

**Algorithm for discrete frequencies:**
- For all available frequencies, compute \( \mu(\Omega_n) \)
- Find zero crossing for \( \mu \), using linear interpolation
- Using finite differences, keep frequencies which satisfy the slope condition
Matching and Track Generation

Generate $L_p(i)$

$\hat{s}_{11,i}$

Frame Selection

Segment Alignment & Concatenation

SM Analysis via IF

Matching & Track Generation

Frequency domain smoothing

IF Attractors

SM Synthesis

Repeat

Hard limitation

Generated $L_q(i)$

$\hat{s}_{11,i}$

SM Analysis Using IF

– Software implementation –

Universitatea Politehnica Bucureşti

Universitatea Politehnica Bucureşti
Experimental Results
– Test signals –

- Sounds: 19 high quality monophonic signals
  - Recording: 44.1KHz, lin. PCM, 16b/smp
  - Downsample to 32, 16, 8kHz >> 76 signals
  - Playback: PC >> SPDIF int. >> Sony STRDB780(QS)>>Yamaha NS10M

- Signals
  - Pure speech, clean background, single speaker, m/f – (4)
  - Pure speech, clean background two speakers, m+f – (1)
  - Speech with noisy background, multiple speakers – (1)
  - Speech with musical background, single speaker, m – (1)
  - Singing voice, clean background m/f – (2)
  - Singing voice, musical background (instrumental), m/f, – (4)
  - Singing voice, complex background (instrumental + choir), m – (1)
  - Pure music (instrumental) – (4)

- Quality evaluators: 5 trained listeners

---

Experimental Results
– Comparison –

- The proposed algorithm is an improved version of KBS-TSM algorithm

<table>
<thead>
<tr>
<th>Score (CMOS)</th>
<th>Comparative Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>Proposed method much better than KBS-TSM algorithm</td>
</tr>
<tr>
<td>2</td>
<td>Proposed method better than KBS-TSM algorithm</td>
</tr>
<tr>
<td>1</td>
<td>Proposed method slightly better than KBS-TSM algorithm</td>
</tr>
<tr>
<td>0</td>
<td>Proposed method equal to KBS-TSM algorithm</td>
</tr>
<tr>
<td>−1</td>
<td>Proposed method slightly worse than KBS-TSM algorithm</td>
</tr>
<tr>
<td>−2</td>
<td>Proposed method worse than KBS-TSM algorithm</td>
</tr>
<tr>
<td>−3</td>
<td>Proposed method much worse than KBS-TSM algorithm</td>
</tr>
</tbody>
</table>

- Improvements in the proposed algorithm:
  - T1: Improved version for correlation estimator, different frames (position, and correlation depth) for correlative matching an concatenation
  - T2: Different frame selection, decoupling (length, and position) of analysis and synthesis frames
  - T3: More refined spectral analysis based on IF instead of STFT
  - T4: Reducing the number of spurious peaks based on IF attractors

- Basic comparison idea:
  - For each scenario, the corresponding block from the proposed algorithm is replaced with the respective one from the KBS-TSM algorithm
Experimental Results
– Conclusions –

• The paper follows the trend of merging aspects of time domain with improved spectral analysis/synthesis methods

• We proposed a new algorithm which offers partial remedies of deficiencies met at the KBS-TSM algorithm:
  • Careful construction of the analysis frames used in correlative concatenation stage
  • More suitable correlation estimator
  • The entire timbre morphing procedure is revised
  • Using IF spectrogram we refined the spectral analysis and we perform a more accurate estimation of the limit conditions for the frequencies, amplitude and phases of the partials required by the SM model
  • The smearing artifacts were keep under control by reducing the length of the synthesis frame without sacrificing the spectral resolution due to the decoupling the analysis and the synthesis frames.
  • The efficiency of the solutions was confirmed by the results of the comparative listening tests.
  • The quality improvement offered by the proposed algorithm is significant for those TSM ratio where KBS-TSM algorithm is known not to operate properly.
  • In addition, we used IF attractors as an alternative for the common peak extraction routine. This allows us to decrease the arithmetic complexity by reducing the number of partials involved in SM synthesis.

• The proposed algorithm can be used to extend, with an acceptable implementation cost, the high quality operating range beyond the typical ±15% as it is for classical time domain algorithm