Between physics and perception
signal models for high level audio processing

Axel Röbel
Analysis / synthesis team, IRCAM

DAFx 2010 – iem – Graz
Overview

- Introduction
- High level control of signal transformation algorithms
- Signal transformation – a short review
  - Sinusoidal model
  - Source-filter model
- Current State
- Research directions
Introduction
Transformation of sound signals

- Signal transformation
  - Change perceived signal characteristic: Volume, duration, pitch, timbre, etc.

- General objectives:
  - Simple control.
  - High quality, no artefacts.
  - Robust operation.

- ALL perceptual qualities leave us with an ambiguous description of the desired transformation.
  - “Duration” merely describes “length in seconds”.
  - Exact specification of how to achieve the desired duration is left to the algorithm.
Introduction
Transformation of sound signals

- Simple Control?

- Desired signal modifications need to be:
  - Easy to achieve and to control.

- Many users do not understand signal processing concepts.

- Algorithms should be controlled intuitively.

- Important especially for timbre transformation.
Introduction
Intuition and high-level control

- Intuitive control:
  - control parameters should relate directly with our experience in the physical world.
  - Categories related to physical and signal domain: pitch and duration.
  - Categories related to physical domain: description of the physical source, playing style, age and gender of speaker, instrument type, etc.

- High level control:
  - Use categories related to the physical domain to control signal transformations
  - Can be obtained most easily if algorithms have a direct link with the physical sound objects that we know in our everyday life.
High-level control

- What do we want to control:
  - instrument type, remix instruments, playing style and ornamentation, voice type, speaker characteristics, etc.

- Required
  - Mapping between the physical properties and the timbre.
  - Very complex and often nonlinear.
High-level control

- Physical models
  - Best candidate, but still requires extensive research to achieve high quality models.
  - Difficult to learn automatically from data.
  - Models are often very specific, general approach not yet available

- Signal models
  - Design a model that covers perceptually relevant properties of physical sound sources.
High-level control
Signal models

- 2 signal models are especially successful in establishing a link to the physical world.

- Source-Filter Model:
  - Independent representation of excitation source and resonator (body) structure.

- Sinusoidal Model:
  - Representation of the individual vibration modes of the excitation source.
Sinusoidal and source-filter model
A short history

- Mechanical speech models
  - Wolfgang von Kempelen, Speaking machine [1773]: using periodic excitation and a resonator filter
Sinusoidal and source-filter model
A short history

- Joseph Faber's "Euphonia», shown in London [1846]: periodic and noise input source
Sinusoidal and source-filter model

A short history

- Channel vocoder Homer Dudley [1939].
- First complete analysis/synthesize of speech,
- First electrical device.
  - Excitation switched between periodic pulse train (pitch controlled from analysis) and noise.
  - Energy distribution measured and controlled in ten 300Hz channels.
  - Manual control of channel energy.
Sinusoidal and Source-Filter model
A short history

28 Operators trained for 1 year
Voder greeting
Voder singing
Sinusoidal and Source-Filter Model
A short history

- Some important steps
  - Dudley [1939]: monolithic periodic excitation signal,
  - Flanagan/Golden [1966]: amplitude/frequency representation of a DFT spectrum phase vocoder, time stretching, in-harmonic signals, (Ex)
  - Moorer [1978]: phase vocoder+ LPC for transposition with timbre preservation, (Ex)
  - McAulay/Quatiery [1986]: Sinusoidal representation of source, independent analysis of vibrating modes, noise modelled as collection of sinusoids,
  - Smith/Serra [1990]: Distinct analysis/treatment of sinusoidal and noise components,
  - Quatiery/McAulay [1992]: Shape invariant speech model,
  - Laroche/Dolson [1999]: Phase vocoder with intra sinusoidal phase synchronization (Ex).
Sinusoidal and Source-Filter Model
Current state (IRCAM)

- Phase vocoder often used as efficient implementation of the sinusoidal model.
- Preservation of transients sufficient for time stretching, slightly worse for transposition. (Röbel DAFx 2003)
- Sinusoidal and noise components can be separated and modified independently (Zivanovic/Röbel/Rodet DAFx 2004 and 2007).
- Independent source and filter transformation allows high level control of age and gender for speech (Röbel/Rodet DAFx 2005, Röbel DAFx 2010).
Outlook

Extension of high-level control:

- Voice Conversion (convert between given speaker identities)
- Source-Filter model using non white source signals
- Source separation and polyphonic signal modification
- Expressive signal manipulation (Voice and instrument)
Outlook
Source-Filter Model

- Source and filter component separation is aiming to separate excitation oscillator and resonator filter.
- Physically reasonable excitation signal is not white!
- Coherent excitation signal estimates will:
  - improve perceptual relevance of controls, and
  - add new controls.
- Recent research:
  - Music: Klapuri [2007], Hahn [2010] (DAFx), Sample Orchestrator 2…
Source-Filter Model
Estimate LF glottal excitation signals (G. Degottex)
Source-Filter Model
Estimate glottal excitation signals (G. Degottex)

Liljencrants-Fant glottic source model + radiation

Time signal

Spectrum

Relaxed voice

Tense voice
Source-Filter Model
Estimate glottal excitation signals (G. Degottex)

Examples:
Transformation of glottic source parameters

Original voice:
Transform into relaxed:
Transform into tense:
Voice Conversion (P. Lanchantin)

- High-level control of speaker identity
- Transform a given source speaker into a given target speaker
- Context aware transformation has to be learned from data.
- Current approaches limit the transformation to the vocal tract (filter part).
- To be addressed:
  - Prosodic information
  - Voiced/unvoiced balance
  - Glottic source parameters
Voice Conversion (P. Lanchantin)

Training Data

SOURCE

TARGET

Temporal alignment

Data Selection

Linear regression

Statistical Model (GMM)

Synthesis

Output voice

9/9/2010

DAFx 2010 - Graz

Voice conversion involves the following steps:

1. **Temporal alignment**:
   - SOURCE and TARGET are aligned.

2. **Data Selection**:
   - F0, envelope, voicing, energy, phonetic, residual parameters are selected.

3. **Linear regression**:
   - Parameters are transformed.

4. **Synthesis**:
   - The transformed parameters are used to generate the output voice, denoted as s'(t).
Voice Conversion (P. Lanchantin)

Input Voice:

Target Voice:

Transformed Voice:

9/9/2010

DAFx 2010 - Graz
Transformation of Expressivity
Speech (C. Veaux)

- Text-to-Speech synthesis generally produces neutral speech.
- Expressive and emotional aspects are missing.
- Intended high-level control: expressive state (anger, sadness, happiness, fear)
Transformation of Expressivity
Speech (C. Veaux)

Expressive transforms must address the five components of prosody
[Pfitzinger, H.R., *Speech Prosody* 2006]

<table>
<thead>
<tr>
<th>Components</th>
<th>Features</th>
<th>Transforms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intonation</td>
<td>Pitch</td>
<td>Dynamic transposition</td>
</tr>
<tr>
<td>Intensity</td>
<td>Loudness</td>
<td>Dynamic scaling</td>
</tr>
<tr>
<td>Speech Rate</td>
<td>Syllabic duration</td>
<td>Time stretch</td>
</tr>
<tr>
<td>Vocal Quality</td>
<td>Open quotient, Roughness and Fry, Breathiness</td>
<td>Rd modification, Glottal pulse jitter and shimmer</td>
</tr>
<tr>
<td>Articulation</td>
<td>Relative position of formants</td>
<td>Envelope Warping</td>
</tr>
</tbody>
</table>

After a stylization process (Legendre polynomial fitting), the reduced representations of these features are clustered for each expressivity (modelization step)
Transformation of Expressivity
Speech (C. Veaux)

Examples:

- Original, neutral expression
- Angry
- Happy
- Sad
Transformation of ornamentation
Vibrato/Tremolo/Note Transitions

- Objective: independent manipulation of ornamentation and expressive playing style and pitch and duration.
- Recent versions of music samplers and other music software starts to integrate basic notion of expressivity transformation
Transformation of Ornamentation
Vibrato/Tremolo/Note Transitions

Example: Vibrato removal

Original

Pitch correction

Induced tremolo correction

9/9/2010

DAFx 2010 - Graz
Source Separation/Acoustic Scene Analysis

- Source separation and acoustic scene analysis are active research topics.
- Algorithms may use sparsity constraints or signal models.
- Applications: Polyphonic signal remixing and editing, signal restoration, automatic transcription, etc.
- First commercially available algorithm (Melodyne DNA) uses sinusoidal signal model.
Music Scene Analysis

- Integration of multiple signal analysis’:
  - Beat markers (G. Peeters)
  - Polyphonic fundamental frequency (C. Yeh)
  - Adaptive instrument models (R. Houzet)
  - Onsets (A. Röbel)

- Followed by an adaptive filtering and stream forming phase
Music Scene Analysis
Audio to note (C. Yeh)

Example:
Multi Pitch Analysis
Polyphonic Audio Transformation
(C. Yeh)

- High-level controls: Pitch and duration of individual notes of the polyphonic Music.

- Examples:
  - Spanish guitar (original)
  - Spanish guitar (some notes transposed)
  - Jazz trumpet (original)
  - Jazz trumpet (some notes transposed)
SUMMARY

- The desire and need to use everyday concepts to intuitively control sound transformation is one of the driving forces of the evolution of sound transformation algorithms.
- The underlying signal model concepts (sinusoidal and source-filter model) have hardly moved within the last 50 years.
- The high-level control concepts are becoming increasingly more complex.
- Many interesting questions are waiting to be solved.

Thanks for listening.