Fixed Point Representations for Very High-Quality Speech and Sound Modification System

Hideki Kawahara
Wakayama University, Japan
Summary

- Functional (computational after Marr) approach is important and productive.
- Fixed points provide feature values as well as their reliability indices.
  - Using within channel information
- Fixed point concept may provide clue to integrate Fourier based concept and wavelet-Mellin transform based concept.

Reference system

“Local” center of gravity
STRAIGHT: demo

spectral envelope conversion ratio
1.4
1
0.7
1/3

F0 conversion ratio
1

natural
unnatural
3

original
STRAIGHT demo: morphing

neutral

angry

extrapolation

interpolation

extrapolation

Word: /hai/ ("Yes" in Japanese)
Background

n "Auditory Brain" Project by CREST
  – Short term goal: speech processing systems based on functional models of auditory functions.
  – STRAIGHT: a very high-quality speech manipulation system
  – Fixed point based algorithms (alternative way of dimensional reduction of auditory representations)
  – Long term goal: "computational" theory of audition

n Frustrations in ways how auditory models are used in ASR and how speech processing systems are evaluated.
“Auditory Brain” Project

To develop a very high quality speech and/or sound manipulation system based on perceptually relevant parameters and it does not preserve phase/waveform information.

- ?? Are distance based quality measures relevant??
- ?? Why does periodic sound sounds smoother and richer (in Auditory Fovea)??
- ?? Is it relevant to test highly nonlinear speech perception using elementary sounds ??
Why high quality?

- Ecological approach for investigating highly nonlinear system, Human

  Not necessarily be predictable from elementary test signals

  Necessary to use ecologically valid stimuli

  Naturalness
Key issue: compatibility

Background figure is removed.
Please visit Hans Moravec’s page for the original figure.
Faster than exponential growth in computing power
(Chapter 3: Power and Presence, Page 60)
http://www.frc.ri.cmu.edu/~hpm/book98/

Hans Moravec: Robot, 2000, Oxford
“Auditory Brain” Project

- Computational theories of speech/auditory perception
  - ecological constraints on evolution
  - It cannot be ad hoc.
  » When there is an elegant and reasonable algorithm and it does not violate ecological (biological and environmental) constraints, there is no reason to deny that the algorithm shares the common underlying principles with our auditory system.
“Auditory Brain” Project

- Computational theories of speech/auditory perception
  - Periodicity: time-frequency sampling grid
  - Periodicity: stable reference point for wavelet-Mellin transform
  - Log-linear frequency axis
    » Wavelet-Mellin transform: shape and size
  - Why two ears? ICA
  - Long term correlation (structure)
    » ASR, music
STRAIGHT a core technology

- Conceptually simple architecture
  - Channel VOCODER
  - Source filter model
- Graded parameters (vs binary decision)
  - Sensitivity analysis
  - Morphing
- Reliability / Transparency
  - No post-processing
  - Weakly constrained model
STRAIGHT: architecture

F0-adaptive time-frequency smoothing to eliminate periodicity interferences

Instantaneous-frequency-based F0 and source information extractor

Group delay manipulation to add artificial naturalness

STRAIGHT is a very high-quality VOCODER.
STRAIGHT a core technology

- Conceptually simple architecture
  - Channel VOCODER
  - Source filter model
- Graded parameters (vs binary decision)
  - Sensitivity analysis
  - Morphing
- Reliability / Transparency
  - No post-processing
  - Weakly constrained model
STRAIGHT: structure

F0-adaptive time-frequency smoothing to eliminate periodicity interferences

Spectral envelope estimation

Instantaneous-frequency-based F0 and source information extractor

Group delay manipulation to add artificial naturalness

STRAIGHT is a very high-quality VOCODER.
Weakly constrained spectral envelope estimation

- Time window
- Waveform

Interferences in the time domain

Interferences in the frequency domain
Weakly constrained spectral envelope estimation

Reduction of edge discontinuity

Rectangular window

Hamming window

Hanning window

Reduction of periodicity interference

Composite window

Smoothing by spline basis

Pitch synchronized Gaussian (PSG) window

Complementary PSG window

STRAIGHT (complementary PSG)
Time-frequency smoothing
(current implementation)

\[
\begin{align*}
w_p(t) &= e^{-\pi \left( \frac{t}{\eta t_0} \right)^2} \odot h(t/t0) \\
h(t) &= \begin{cases} 
1 - |t| & |t| < 1 \\
0 & \text{otherwise,}
\end{cases} \\
w_c(t) &= w_p(t) \sin \left( \pi \frac{t}{t_0} \right) \\
P_r(\omega, t) &= \sqrt{P_o^2(\omega, t) + \xi P_c^2(\omega, t)}
\end{align*}
\]
Compensation of over-smoothing

\[ v(\omega) = \int_{-\infty}^{\infty} W(\omega - \lambda) h(\lambda) d\lambda \]

\[ u_l = \sum_{k=-N}^{k=N} c_k u_{l-k} \]

where

\[ u_l = \begin{cases} 1 & (l = 0) \\ 0 & \text{(otherwise)} \end{cases} \]
Compensation of over-smoothing

\[ c = (H'H)^{-1}H'u \]

\[ u = \begin{bmatrix} u_{-M}, & u_{-M+1}, & \ldots, & u_0, & \ldots, & u_{M-1}, & u_M \end{bmatrix}' \]

\[ c = \begin{bmatrix} c_{-N}, & c_{-N+1}, & \ldots, & c_0, & \ldots, & c_{N-1}, & c_N \end{bmatrix}' \]

\[ H = \begin{pmatrix} \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
-M & v_{-N-M} & \cdots & v_{l-M} & \cdots & v_{N-M} & \\
-k & v_{k-N} & \vdots & v_{k+l} & \vdots & v_{k+N} & \\
-M & v_{-N+M} & \cdots & v_{l+M} & \cdots & v_{N+M} & \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\end{pmatrix} \]
Weakly constrained spectral envelope estimation
Weakly constrained spectral envelope estimation

STRAIGHT (complementary PSG)
Fixed point based algorithms

- Fixed points in the frequency domain:
  - F0 extraction
- Fixed points in the time domain:
  - Excitation extraction
Fixed point of mapping

Examples

* Instantaneous frequency of a filter output around a sinusoidal component
* Energy centroid of a windowed signal around an event
Averaging and fixed point

Prominent component

Window locations

Fixed point

Average of windowed value

background
Averaging and fixed point

Prominent component

Parameters
(position, slope, [level])

Window locations

Fixed point

Average of windowed value
background
STRAIGHT: structure

F0-adaptive time-frequency smoothing to eliminate periodicity interferences

Instantaneous-frequency-based F0 and source information extractor

Group delay manipulation to add artificial naturalness

STRAIGHT is a very high-quality VOCODER.

F0 estimation
window selection for reliable representation of mapping

Refinement of Fo synchronous windows
Window with harmonic cancellation

\[ w_s(t, \lambda_c) = w(t, \lambda_c) \ast h(t, \lambda_c), \]

\[ w(t, \lambda_c) = e^{-\frac{\lambda_c^2 t^2}{4\pi \eta^2}} e^{i\lambda_c t}, \]

\[ h(t, \lambda_c) = \max \left\{ 0, 1 - \left| \frac{\lambda_c t}{2\pi \eta} \right| \right\}, \]
Fixed-point-based sinusoidal components extraction
Fixed-point-based sinusoidal frequency and C/N estimation

C/N information enables optimum F0 estimation based on multiple harmonic components
Approximate estimation of C/N

\[ \tilde{\sigma}^2(t, \lambda) = \int_{-T_w}^{T_w} |w(\tau, \lambda)| \tilde{\sigma}^2(t - \tau, \lambda) d\tau \]

\[ \tilde{\sigma}^2(t) = c_a \left( \frac{\partial \omega(t, \lambda)}{\partial \lambda} \right)^2 + c_b \left( \frac{\partial^2 \omega(t, \lambda)}{\partial t \partial \lambda} \right)^2 \]

\[ c_a = \frac{1}{\int_{-\infty}^{\infty} \left( \lambda_o \frac{dg(\lambda)}{d\lambda} \bigg|_{\lambda=\lambda_o} \right)^2 d\lambda_o} \]

\[ c_b = \frac{1}{\int_{-\infty}^{\infty} \left( \lambda_o^2 \frac{dg(\lambda)}{d\lambda} \bigg|_{\lambda=\lambda_o} \right)^2 d\lambda_o} \]
Reliable built-in mechanism for fundamental component selection

**Linear filter arrangement**

**Log-linear filter arrangement**

(mapping)

(filter output)
Fixed points on C/N map

Fundamental component
F0 evaluation based on EGG

W/O : 0.72%
with : 0.32%
Graded source Information

- Fixed point based Fo extraction (with C/N map)
- F0 trajectories (resolution: 1/F0)
- C/N for each fixed point

Graded aperiodicity information is also extracted
Fixed points in the time domain

- How to define auditory temporal events
  - Localized energy centroid

Alternative representation
Fixed points in the time domain

- Speech waveform
- Gaussian window
- Squared whitened signal
- Energy centroid
- Amount of energy concentration

Fixed points in the time domain refer to stable states or points of interest in the signal representation. The images illustrate the relationship between the waveform and its transformed versions, highlighting the energy concentration and its centroid.
Fixed point based event detection
Definition of event in the time domain

\[ \langle t(u) \rangle = \frac{\int t |x(t, u)|^2 dt}{\int |x(t, u)|^2 dt} \]

Mean time

\[ \sigma_t^2(u) = \frac{\int (t - \langle t \rangle)^2 |x(t, u)|^2 dt}{\int |x(t, u)|^2 dt} \]

duration

\[ \{ t_e \} = \{ u | \langle t(u) \rangle = u, \frac{d\langle t(u) \rangle}{du} < 1 \} \]

Event location

Windowed whitened signal
Windowed event location and the original event location

Gaussian window

Approximation of envelope

Windowed location

Window location

Original location

\[ w(t) = e^{-\frac{t^2}{2\sigma_w^2}}. \]

\[ |s(t)| = e^{-\frac{(t-t_e)^2}{2\sigma_s^2}}. \]

\[ \langle t(u) \rangle = \frac{\sigma_s^2 u + \sigma_w^2 t_e}{\sigma_s^2 + \sigma_w^2}. \]
Duration can be estimated from the geometrical parameter at the fixed point

\[ \sigma_s(t_e) = \sigma_w \sqrt{\frac{g(t_e)}{1 - g(t_e)}} \]

- duration
- Window parameter
- Slope at fixed point
Equivalence between the time domain definition and the frequency domain definition

Time domain definition

Frequency domain definition

Equivalence between the time domain definition and the frequency domain definition
Inverse problem: Where is the excitation?

Event as the energy centroid

Minimum phase response

Excitation (impulse)

compensation
Equivalence in definitions

Frequency domain definition of the event location

\[ \langle t(u) \rangle = - \int \psi'(\omega, u)|S(\omega, u)|^2 d\omega \]

\[ \sigma_t^2(u) = \int \left( \frac{B'(\omega, u)}{B(\omega, u)} \right)^2 B^2(\omega, u) d\omega + \int (\psi'(\omega, u))^2 \langle t(u) \rangle^2 B^2(\omega, u) d\omega \]

\[ S(\omega, u) = \frac{1}{\sqrt{2\pi}} \int x(t, u)e^{-j\omega t} dt = |S(\omega, u)|e^{j\psi(\omega, u)} = B(\omega, u)e^{j\psi(\omega, u)} \]

Assuming causality
Group delay of a minimum phase response

\[
\tau_\phi(\omega, u) = -\frac{d}{d\omega} \left( \text{imag} \left[ \frac{1}{\sqrt{2\pi}} \int C(q, u)e^{j\omega q}dq \right] \right) \tag{14}
\]

\[
C(q, u) = \begin{cases} 
2c(q, u) & q > 0 \\
c(q, u) & q = 0 \\
0 & \text{その他の場合}
\end{cases}
\]

\[
c(q, u) = \frac{1}{\sqrt{2\pi}} \int \log B(\omega, u)e^{-j\omega q}d\omega \tag{15}
\]
Compensation based on minimum phase group delay

\[ \langle \tilde{t}(u) \rangle = - \int (\psi'(\omega, u) + \tau_\phi(\omega, u)) |S(\omega, u)|^2 d\omega \]

Compensated event location

\[ \tilde{\sigma}_P^2(u) = \int (\psi'(\omega, u) + \langle t(u) \rangle + \tau_\phi(\omega, u))^2 |S(\omega, u)|^2 d\omega \]

Compensated event duration

- \[ -\psi'(\omega, u) \] Observed group delay
- \[ \tau_\phi(\omega, u) \] Causal group delay
example

Observed group delay
Minimum phase group delay
Compensated group delay
Excitation estimation based on fixed point based event detection

- Excitation
- Energy centroid
- Vocal fold closure
- Compensated group delay

Event based concentration
Excitation based concentration

9-10 August, 2002 Computational Audition 48
Excitation extraction accuracy

Standard deviation
Multiple resolution display of events (fixed points)

Estimated excitation

Speech waveform
Multiple resolution display of events (fixed points)

Fixed points due to one excitation aligns on a straight line

demo
Phase map of wavelet transform
Instantaneous frequency based fixed points
Instantaneous frequency based fixed points
Group delay based fixed points
Group delay based fixed points
STRAIGHT: structure

Source attribute control

Instantaneous-frequency-based F0 and source information extractor

Group delay manipulation to add artificial naturalness

STRAIGHT is a very high-quality VOCODER.
Group delay manipulated mixed-mode excitation source

..provides continuous coverage from pulse train to random noise
Summary

- Functional (computational after Marr) approach is important and productive.
- Fixed points provide feature values as well as their reliability indices.
  - Using within channel information
- Fixed point concept may provide clue to integrate Fourier based concept and wavelet-Mellin transform based concept.
Colleagues

- Haruhiro Katayose, Toshio Irino, Takanobu Nishiura (Wakayama Univ.)
- Minoru Tsuzaki, Hideki Iwasawa (ATR)
- Parham Zolfaghari (NTT)
- Kiyohiro Shikano, Hiroshi Saruwatari (NAIST)
- Fumitada Itakura, Kazuya Takeda, Shoji kajita, Hideki Banno (CIAIR, Nagoya Univ.)
- Masato Akagi, Masashi Unoki (JAIST)
- Seiichi Nakagawa (Toyohashi Inst. Tech)
- Shigeki Sagayama, Nobuaki Minematsu (Univ. Tokyo)
- Diane Kewley-Port (Indiana Univ. USA)
- Osamu Fujimura (Ohio state Univ. USA)
- Alain de Cheveigné (IRCAM, France)
- Roy D. Patterson (CNBH, UK)
References


For computational “Audition”

- Parts preparation
- F0 trajectory and frequency axis modification
- Mixing and level adjustment
Nonlinear time warping based on phase of the F0 component (FM pulse train)

without time warping

with time warping
Nonlinear time warping based on phase of the F0 component (vowel sequence /aiueo/)