Prosodic Manipulation

Advanced Signal Processing, SE

3.12.03

David Ludwig

homer@sbox.tugraz.at

Contents

- Introduction
- SOLA, PSOLA
- LP-PSOLA
- RELP
- Sinusoidal/harmonic-plus-noise modeling
- MBROLA
- Application

Introduction

Definition

- **prosody** *(noun)*

  1. the study of poetic meter and the art of versification

  2. the patterns of stress and intonation in a language
  Synonyms: inflection

  3. a system of versification
  Synonyms: poetic rhythm, rhythmic pattern

→ **pitch, duration, amplitude (gestures)**

→ **Function:**

→ **Stress, non – lexical information, discourse, emotion**

Pitch ´n Time

(Robot: His name is R1D1...)

- Playful_Time:
  - Random number between 10 and 400 milliseconds and use that for the phone duration.
- Serious_Time:
  - same duration value to each phone.
- Playful_Pitch:
  - random melody for the sentence
- Serious_Pitch:
  - same pitch values, monotone

SOLA, PSOLA

(pitch-)synchronous overlap-and-add

SOLA

- Time-segment processing
  - Segmentation of x[n] into overlapping frames
  - Shifting according to scaling-factor α
  - Repositioning, Overlap/Add
  - Cross-Correlation in the overlap interval
  - Maximum of CC
  - Fade in / fade out
  - Flexible time lag

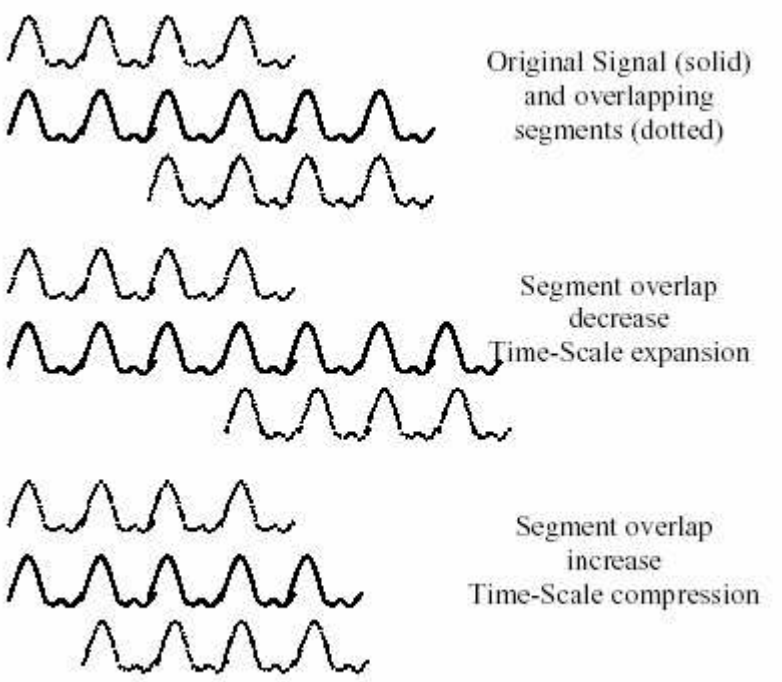
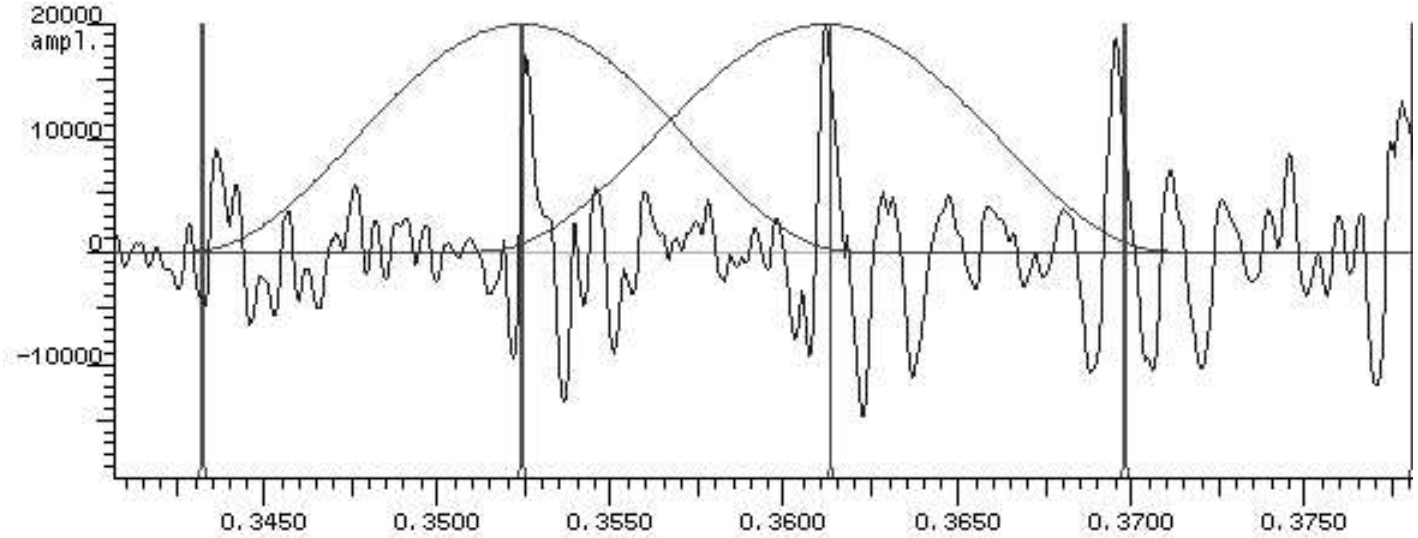
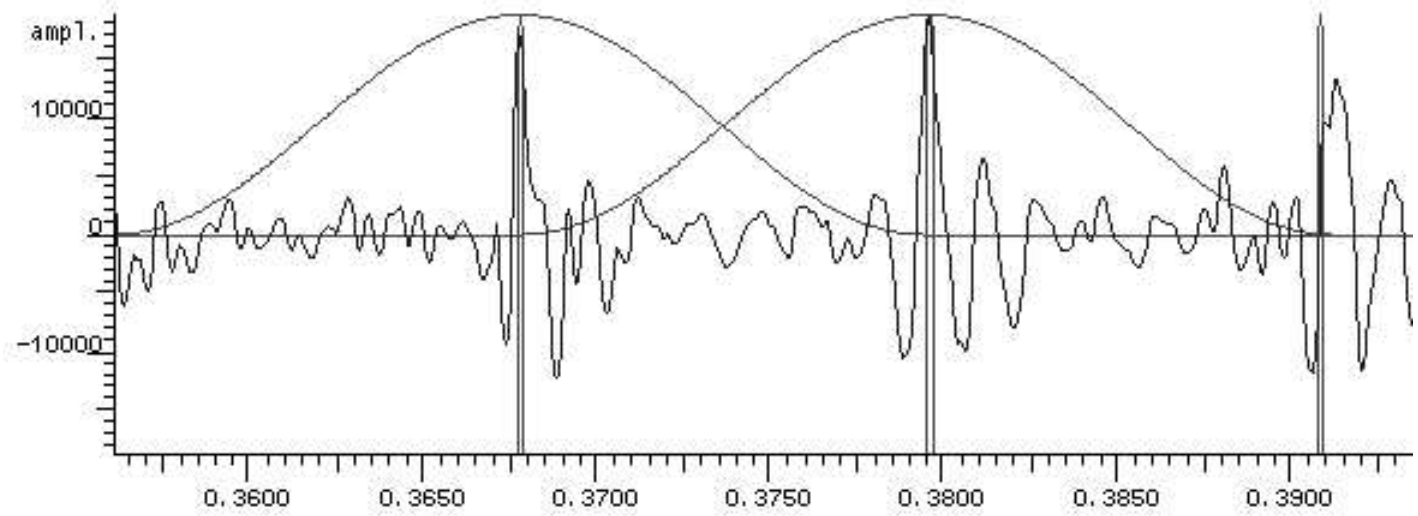


Figure 3 The Synchronized Overlap-Add Method (from [8])

PSOLA

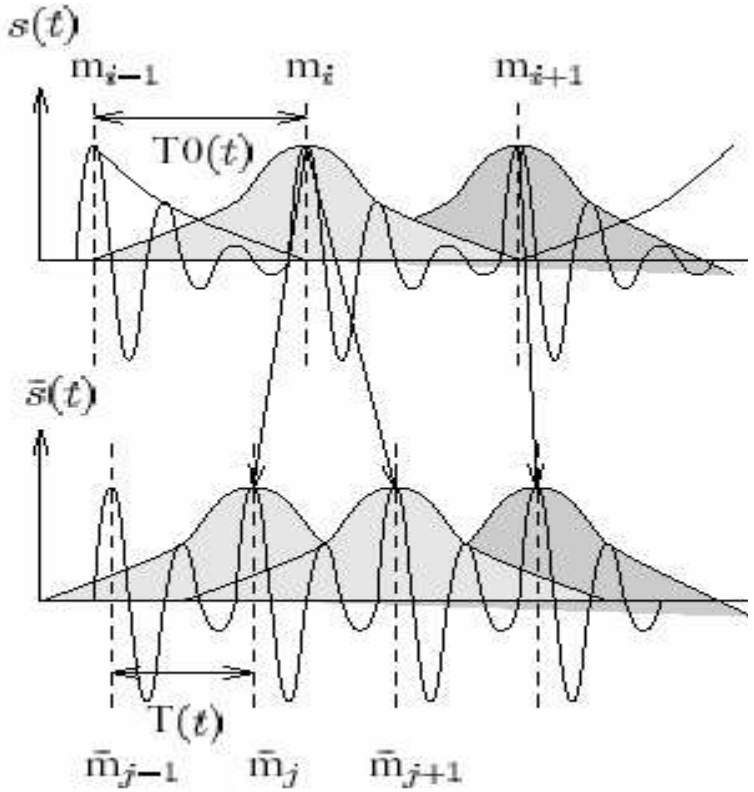
- Variation especially for voice processing
- Splitting signal ino overlapping windows
- Synchronized with fundamental frequency
  - Avoids pitch discontinuities
- Neccesitates preliminary **pitch marking**
- Analysis:
  - Pitch Period P(t) at pitch mark t_i
  - Segment extraction by windowing with pitch mark as its center


20000
ampl.
10000
0
-10000
0.3450
0.3500
0.3550
0.3600
0.3650
0.3700
0.3750


ampl.
10000
0
-10000
0.3600
0.3650
0.3700
0.3750
0.3800
0.3850
0.3900

Synthesis

- Time-Scaling:
  - ST-signals must be added (or suppressed) without altering the distance among adjacent pitch periods
- Pitch-Shifting:
  - synthesis time axis will have the same duration, but it will be necessary to scale the local pitch period
- ST-Signals might be discarded (compression/lower pitch)
- ST-Signals might be used twice (stretching/higher pitch)
- Artefacts:
  - Transient smearing, audible slices, Distortion due to phase errors



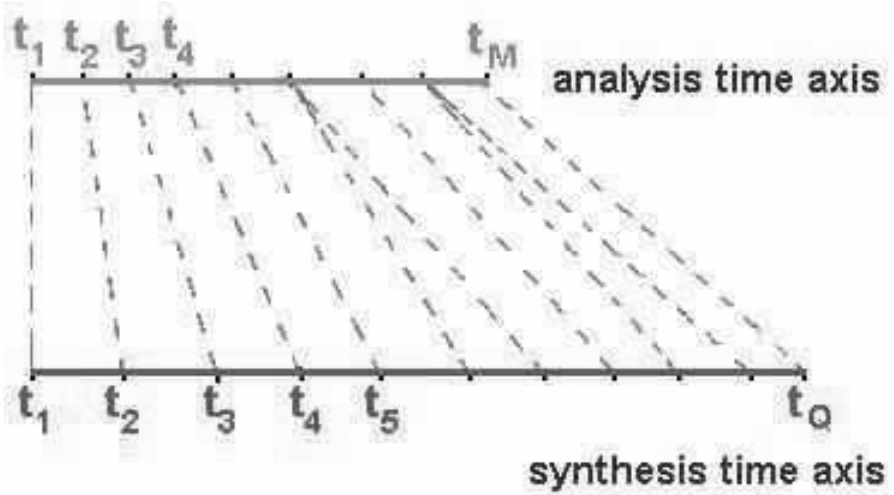


Figure 1: schematic representation of the mapping between synthesis and analysis time axis

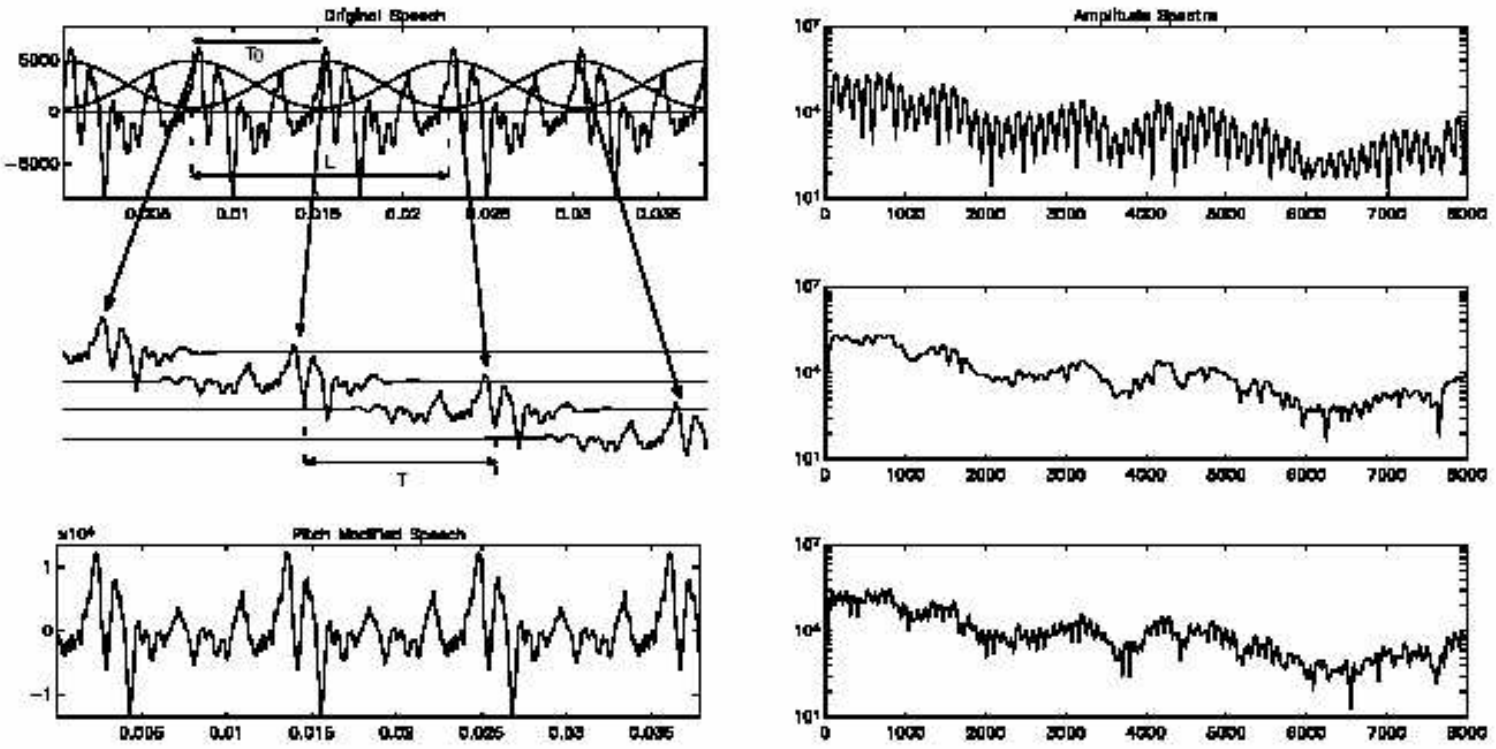


Fig 1. The TD-PSOLA Re-Harmonization process.

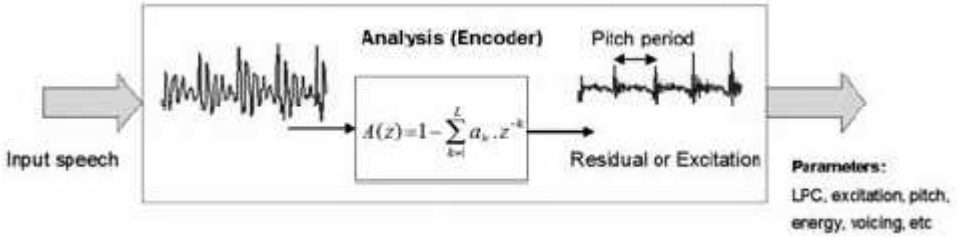
LP-PSOLA, RELP

LP-PSOLA

- LP-Residual or Error Function e(t) is used
  - spectrally flat
  - Separating excitation and vocal tract
  - Little correlation within each pitch period
- TD-PSOLA algorithm is applied to the residual part
- Advantages:
  - Control of spectral structure
  - No additional computation time

RELP

- Residual Excited LPC
- Vocoding technique for speech transmission (e.g. mobile phones)
- Residual Signal is compressed
  - Low-Pass Filtering
  - Downsampling
  - Re-Quantisation



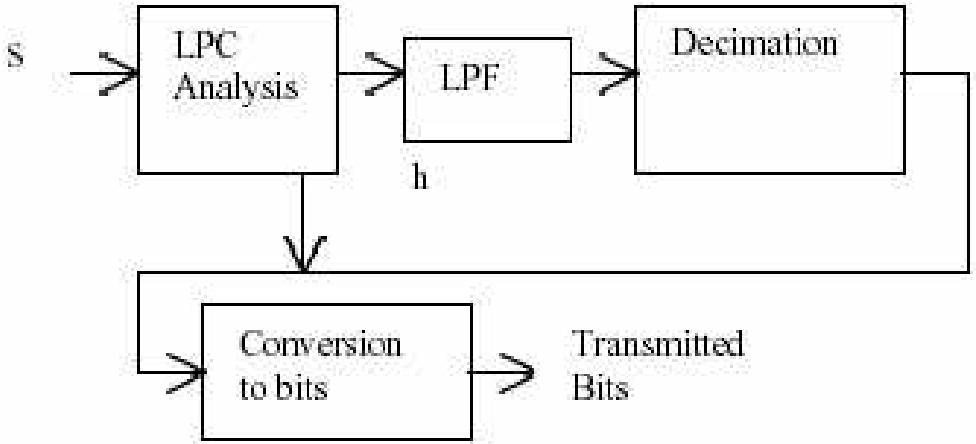


Figure 2: Transmitter part of RELP system

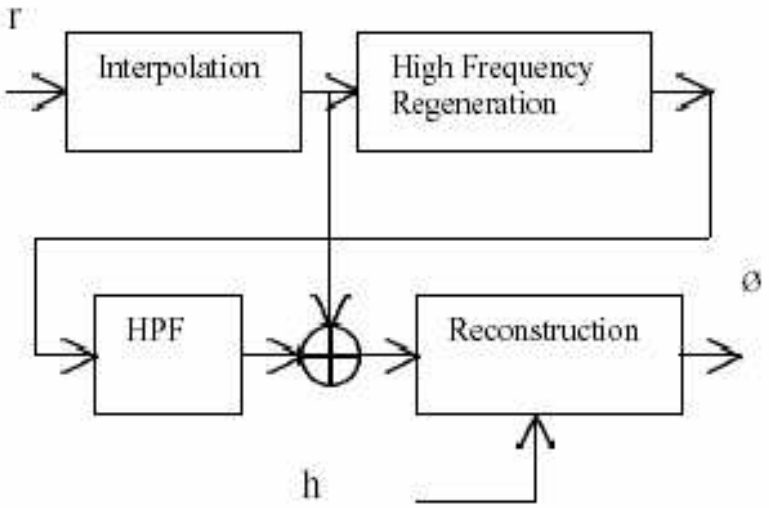


Figure 3: Receiver part of the RELP system

Source-Filter Models

- Source=oscillation of vocal chords
  - Voiced (Dirac-Impulses)
  - Unvoiced(Noise)
- Filter=TF of vocal tract
- LP → Approximation of spectral envelope
- Problem: Estimation of filter coefficients

Sinusoidal/Harmonic + Residual Model (HNM)

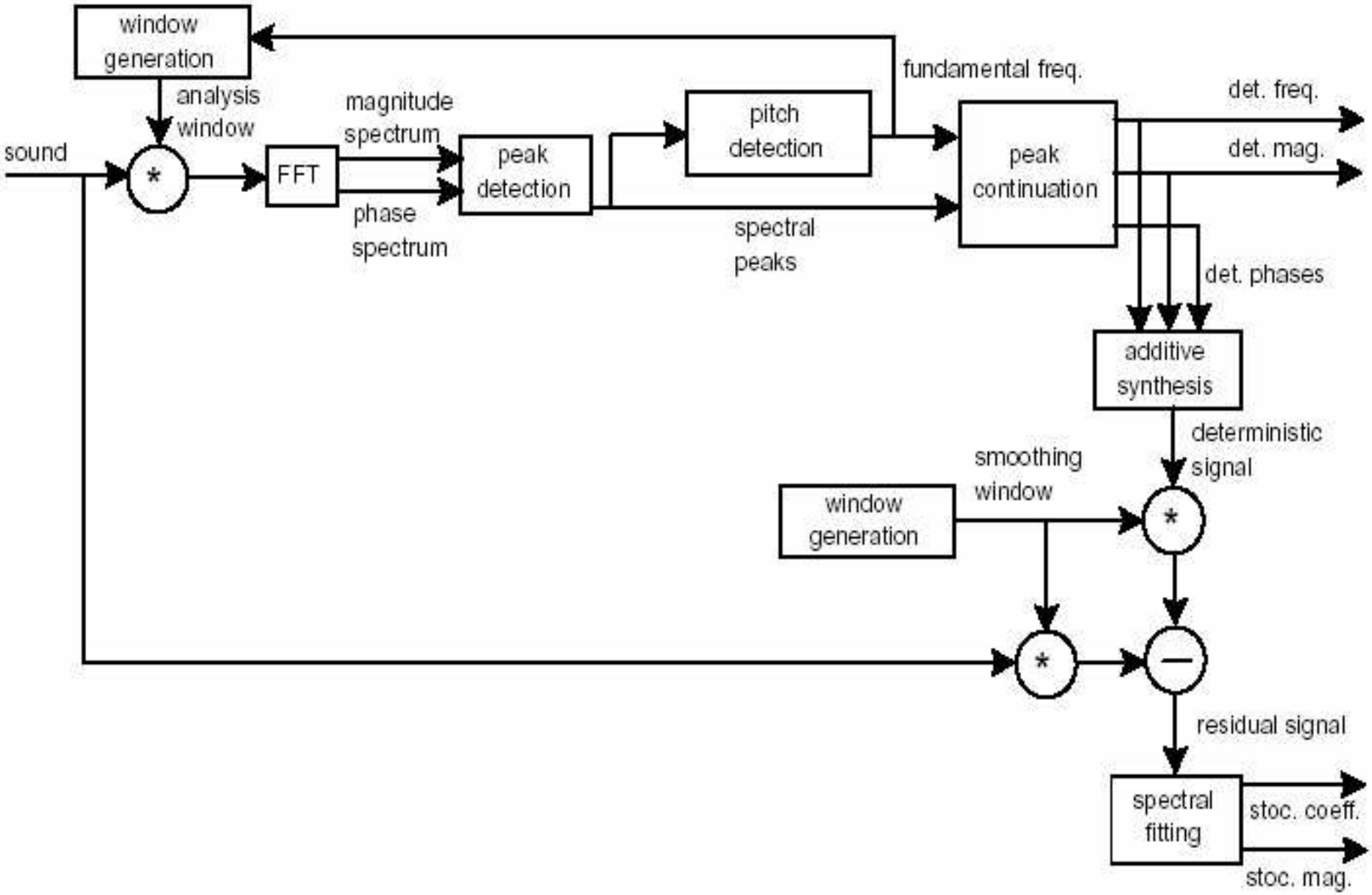
Analysis/Synthesis

- Signal is decomposed in harmonic+noise part:

$$s(t) = \sum_{k=-L(t)}^{L(t)} A_k(t) e^{jk\omega_0(t)t} + e(t)$$

$$e(t) = w(t)[h(\tau,t) * n(t)]$$

- Number of harmonics, fundamental frequency, time-variant amplitudes (harmonic model)
- Peak detection/continuation, pitch detection, Subtraction
- Residual~ time-pulsed,filtered noise
- Synthesis: Additive/Subtractive Synthesis


window generation
analysis window
sound
*
FFT
magnitude spectrum
phase spectrum
peak detection
pitch detection
fundamental freq.
peak continuation
det. freq.
det. mag.
det. phases
spectral peaks
additive synthesis
deterministic signal
smoothing window
window generation
*
*
—
residual signal
spectral fitting
stoc. coeff.
stoc. mag.

Features

- Voiced/unvoiced decision
- Crucial:
  - Pitch estimation
  - Peak continuation
    - McAulay + Quatieri Algorithm
  - Phase Relationships

$$E = \frac{\int_{0.7\omega_0}^{1.3\omega_0} (|S(\omega)| - |\tilde{S}(\omega)|)^2}{\int_{0.7\omega_0}^{1.3\omega_0} |S(\omega)|^2}$$

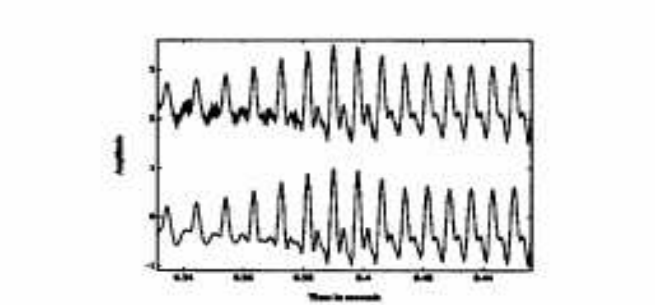


Figure 1: Original speech signal (top) and the corresponding deterministic signal (bottom)

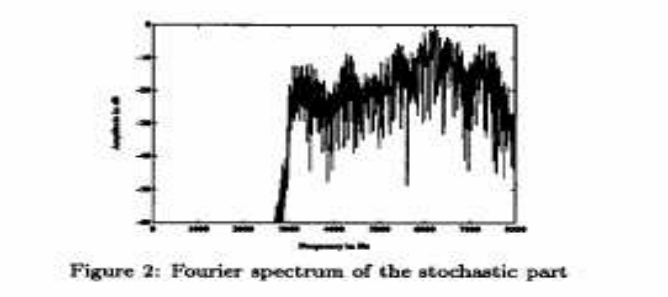


Figure 2: Fourier spectrum of the stochastic part

MBROLA

MBROLA

- Multiband Resynthesis OLA
- Faculté polytechnique de Mons (Belgium)
- Open source synthesizer
  - As many voices, dialects and languages as possible
  - Actually 27 languages !!
  - Diphone concatenation
- Time-domain approach (MBR-PSOLA)
- Smoothing of spectral discontinuities in the time domain enhances fludity

Examples

- German 
- U.K. English 
- U.S. English 
- Japanese 

Manipulation

- Manipulation in frequency domain
- Pitch-Shifting
  - Direct access to sinusoidal components → frequency shifting with/without formant preservation
- Time-Scaling
  - No change of Input/Output hopsize
  - Superior to phase vocoder
  - Computationally expensive

Application

Application

- Mean value <F0>
- Macro-Prosody D_{F0}
- Micro-Prosody M_{F0}
- Pitch Modification by:

$$\overline{F_0}(t) =< F_0 > +\alpha(t) D_{F_0}(t) + \beta(t) M_{F_0}(t)$$

$$\gamma(t) = \frac{\overline{F_0}(t)}{F_0(t)}$$

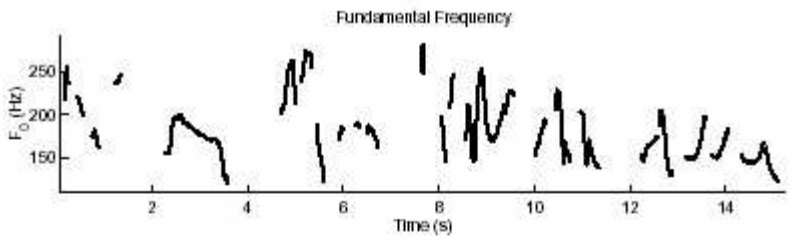


Figure 6: *Fundamental frequency of a spoken voice "lalula".*

$$F_0(t) =< F_0 > +D_{F_0}(t) + M_{F_0}(t)$$

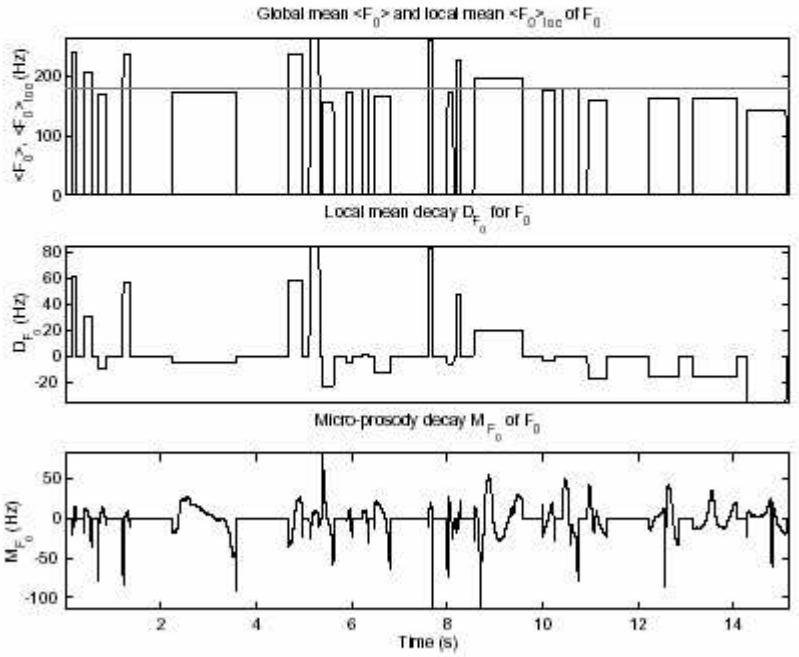


Figure 7: *Fundamental frequency decomposition using a simple voice/unvoiced mask.*

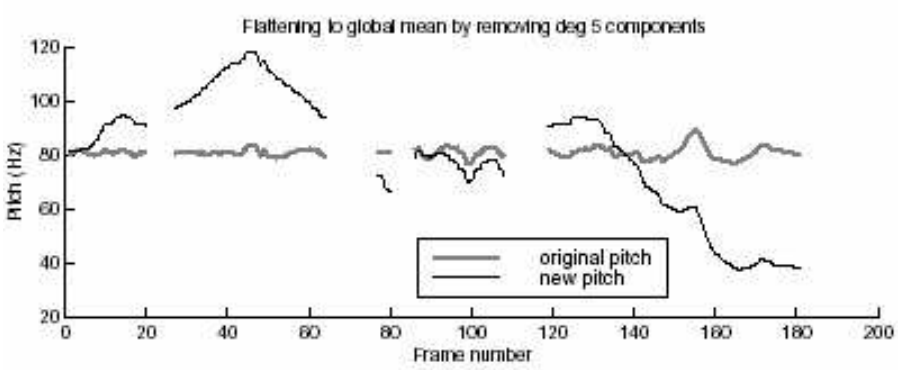


Figure 10: *Pitch-shifting that globally flattens the intonation, using* $\beta = 0$, $\alpha = 0$.

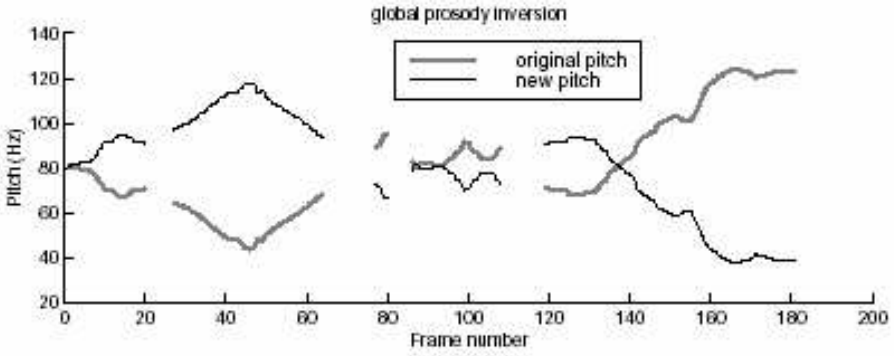


Figure 12: *Pitch-shifting that globally inverse the intonation, using* $\beta = -1$, $\alpha = -1$.

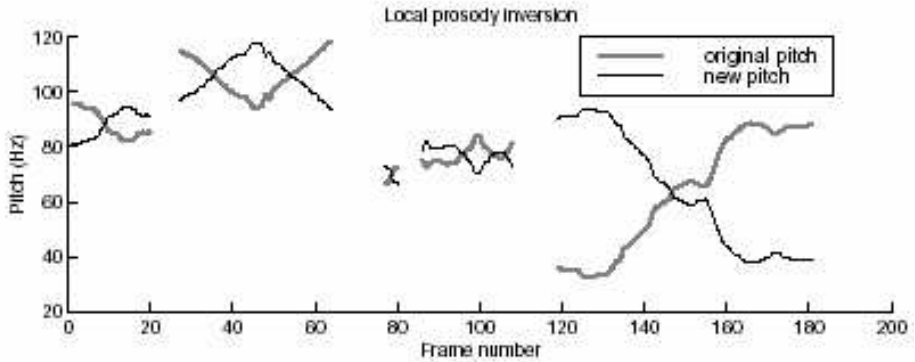


Figure 11: *Pitch-shifting that locally inverse the intonation, using* $\beta = -1$, $\alpha = 1$.

References

- PSOLA:
  - U.Zoelzer: DAFX Wiley, John & Sons, Incorporated

  - E. Moulines and F. Charpentier: Pitch-Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis using Diphones, Speech Comminucation, vol 9, pp 452-467, 1990.

- HNM:
  - J. Laroche, Y. Stylianou, and E. Moulines: HNS: Speech Modification Based on a Harmonic+Noise Model, Proc. of ICASSP 1993, vol.2, pp.550-553.

- MBROLA:
  - tcts.fpms.ac.be/synthesis/mbrola.html

THANK YOU