Converting Speaking Voice into Singing Voice

1st place of the **Synthesis of Singing Challenge 2007**:

"Vocal Conversion from Speaking to Singing Voice using STRAIGHT"

by Takeshi Saitou et al.

STRAIGHT

Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum

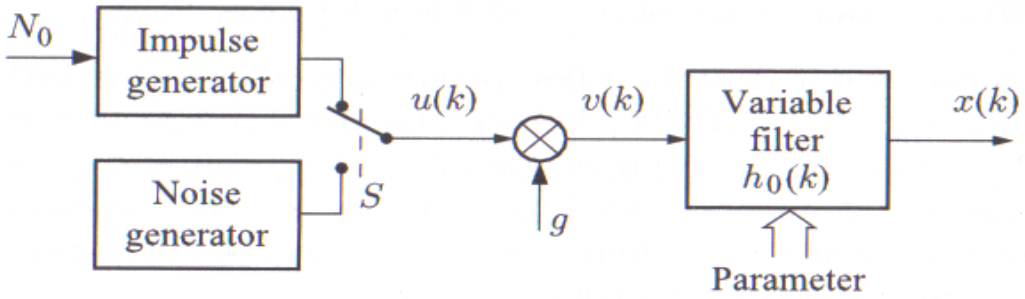
... is a set of simple procedures to estimate speech parameters, i.e. F0 and spectral information, proposed by Kawahara et al. in 1998

Idea:

- analysis of speech parameters
- manipulation of speech parameters
- re-synthesis of speech

Basic idea: Channel Vocoder

Vocoder



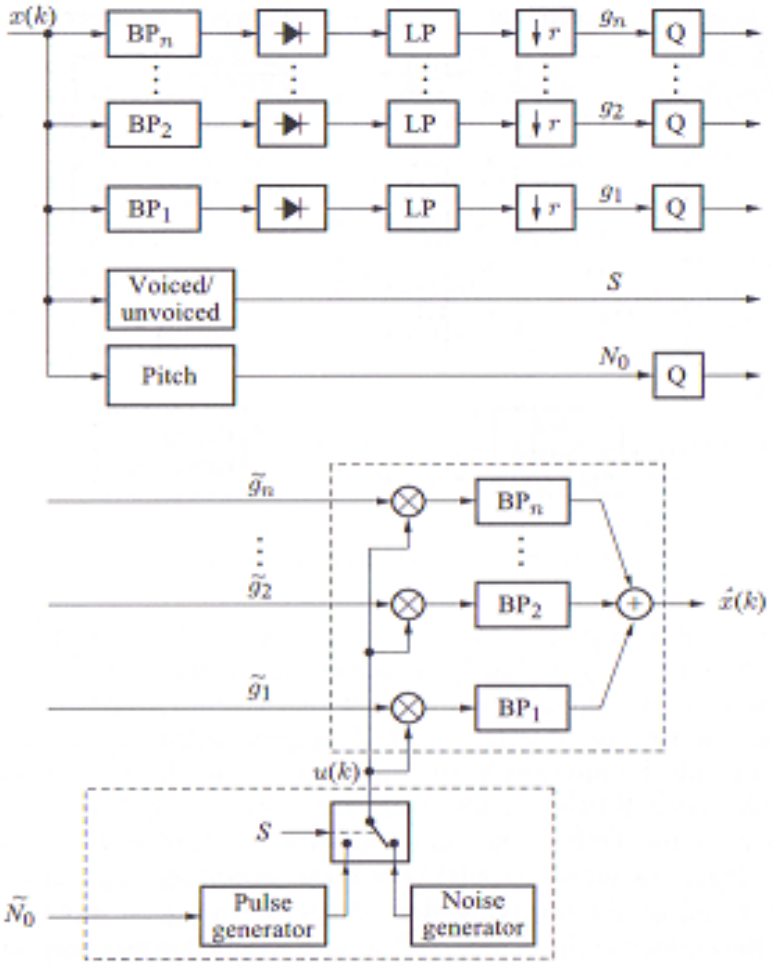
PRO

- simple
- easy to understand
- intelligible speech quality
- flexible in parameter manipulations

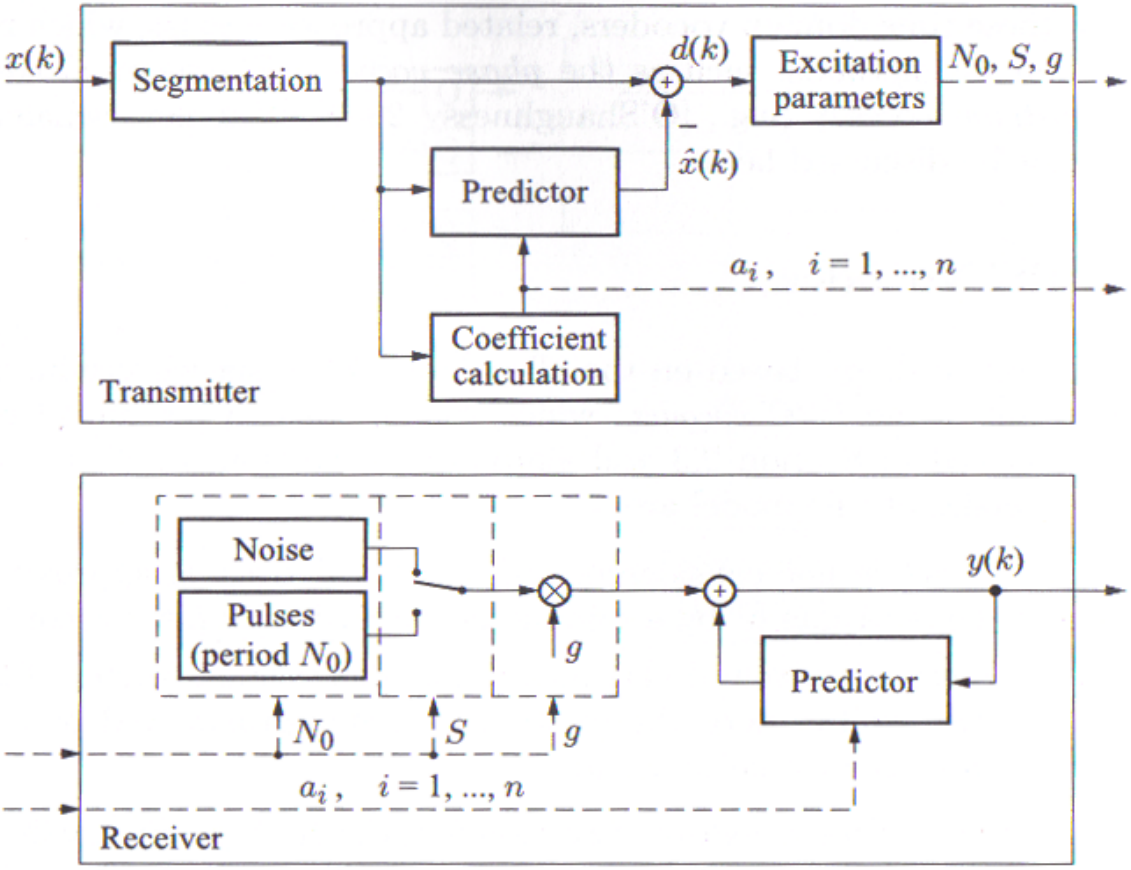
CON

- lousy quality
  the attribute “vocoder quality” is normally not a compliment
  (still the vocoder is often used as style instrument. For example: musical group “Air”)

Channel Vocoder



LPC Vocoder


x(k)
Segmentation
d(k)
Excitation parameters
N0, S, g
Predictor
x̂(k)
ai , i = 1, ..., n
Coefficient calculation
Transmitter
Noise
Pulses (period N0)
g
y(k)
Predictor
N0
S
g
ai , i = 1, ..., n
Receiver

Vocoder

Main Problems:

- buzziness introduced by plosive excitations
  (there are methods to reduce these, not mentioned here)

- estimation errors of the spectral information due to interferences
  introduced by periodicity in the signal (voiced sounds)

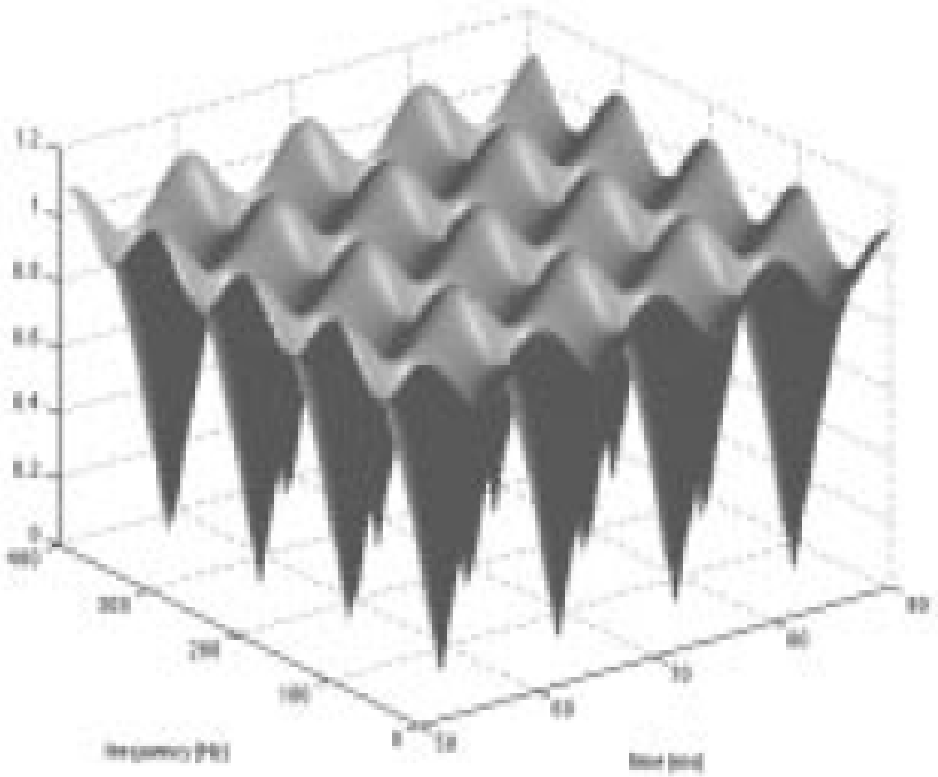
Periodic interferences

The Channel Vocoder actually uses the Power Spectrum to model the vocal tract

Spectrogram: graphical representation of the short term Fourier Transform

The spectrogram of a periodic (voiced) speech signal shows periodic interferences in the time domain as well as in the frequency domain, due to spectral smearing effects

Periodic interferences

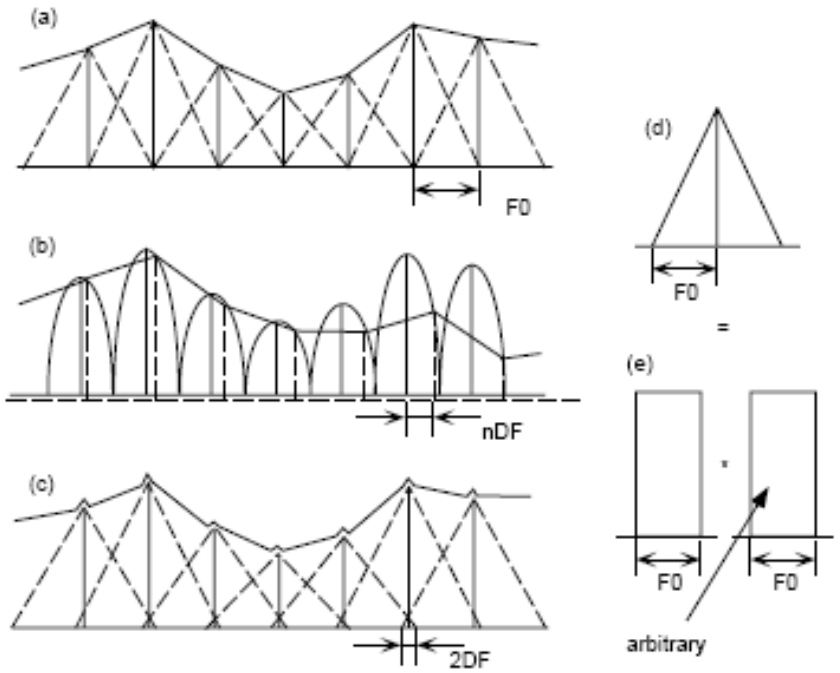


spectrogram of a regular pulse train with interferences caused by F0

Periodic interferences – first solution

- regard the spectrogram as 3D surface
- regard voiced excitation as sampling function on this surface, providing information every τ_0 in time domain and every f_0 in frequency domain
- the estimation of the spectrogram therefore yields in a surface reconstruction problem by using partial information (knot points)
- easiest method: connect knots with 1^{st} order polynomials

Surface reconstruction



1D case

Reducing phase interferences

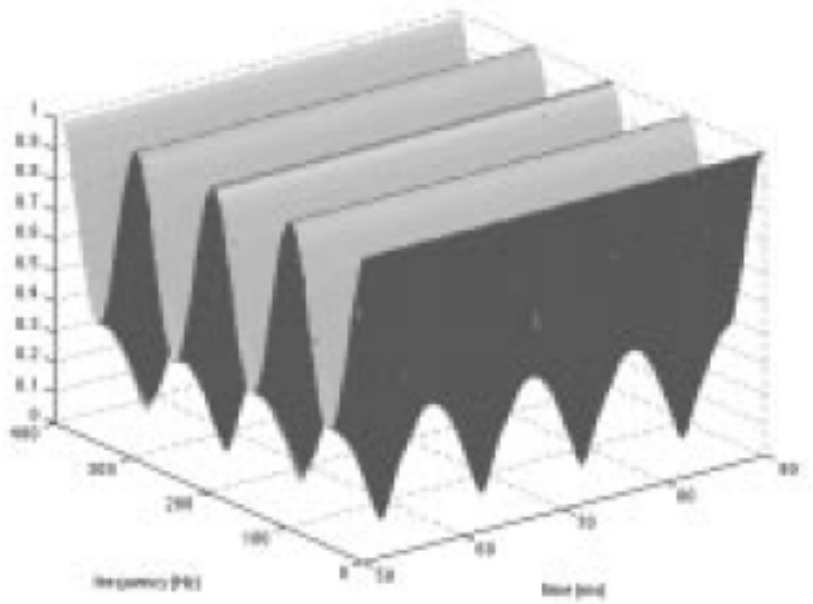
If we calculate the spectrogram pitch synchronous, i.e. using a window of length τ_0, we get rid of temporal interferences → need of exact f_0 estimation

used window:

$$w_p(t) = e^{-\pi\left(\frac{t}{t_0}\right)^2} \odot h(t/t0)$$

$$h(t) = \begin{cases} 1-|t| & |t|<1 \\ 0 & \text{otherwise} \end{cases}$$

Reducing phase interferences



using this window eliminates temporal interferences.

“holes” in the frequency domain remain due to phase extinction

Reducing phase interferences

Define a new window by modulating the old window:

$$w_c(t) = w_p(t) \sin\left(\pi \frac{t}{t_0}\right)$$

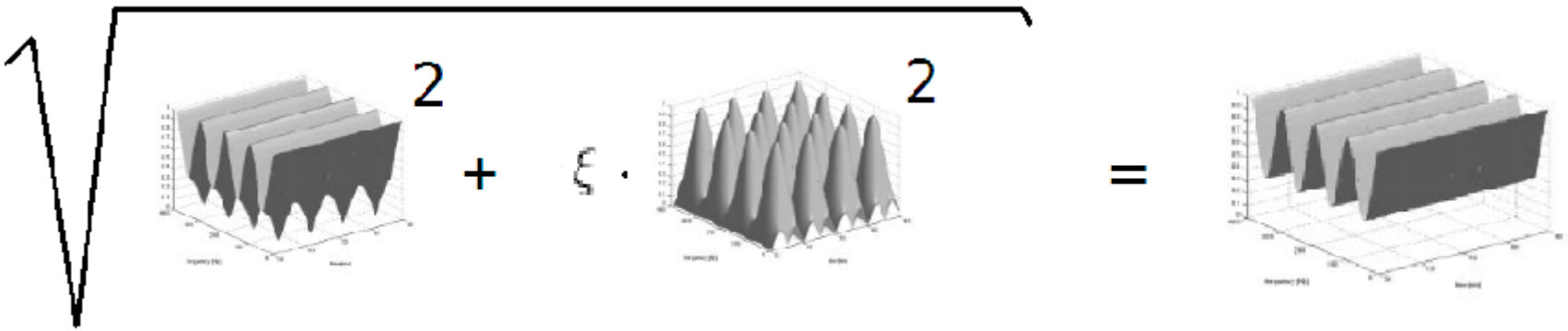
- harmonic components are shifted towards each other
- their phase is changed by π/2 in opposite direction
- → the resulting spectrogram has peaks, where the original spectrogram has holes

Reducing phase interferences

blend the original spectrogram with the compensating spectrogram to get the spectrogram with reduced phase interferences.

the blending factor ξ = 0.13655 was searched numerically

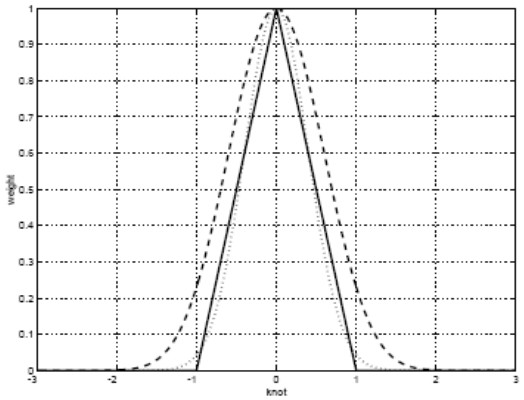
complementary windows:



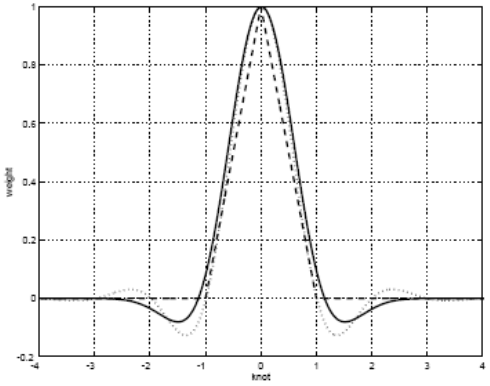
Over-smoothing

- The time window already smooths the spectrogram
- using the triangular smoothing kernel also smooths beside reducing phase interferences

- it is possible to design a kernel which compensates the over-smoothing effect



over-smoothed kernel



compensated kernel

Extracting F0

- normally done by measuring the fundamental period
- hard for speech
  - F0 changes with time
  - speech is unstable (pauses, voiced/unvoiced)
  - speech is not purely periodic

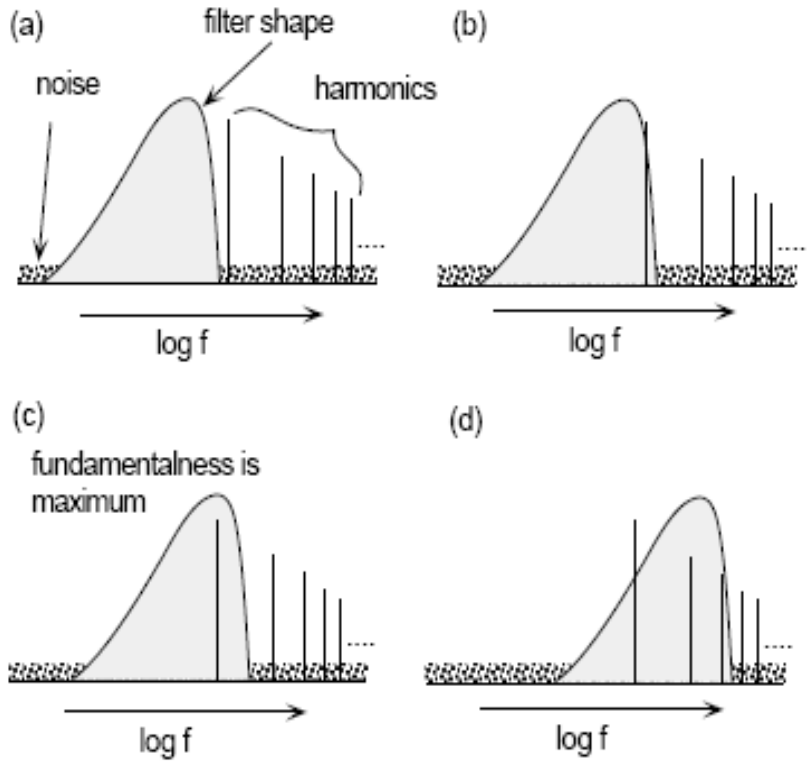
- proposed speech representation:

$$s(t) = \sum_{k \in N} \alpha_k(t) \sin\left(\int_{t_0}^{t} k(\omega(\tau) + \omega_k(\tau)) d\tau + \phi_k \right)$$

  i.e. a superposition of AM(α_k) and FM(ω_k) modulated sinusoids

- definition of the new term “fundamentalness”: fundamentalness is high, when AM and FM magnitudes are low

Fundamentalness



We scan the frequency domain with a special filter and define the F0 to be the frequency, were "fundamentalness" is highest

Definition of the filter:

$$g_{AC}(t) = g(t - 1/4) - g(t + 1/4)$$

$$g(t) = e^{-\pi\left(\frac{t}{\beta}\right)^2} e^{-j2\pi t}$$

Fundamentalness

$$D(t,\tau_c) = |\tau_0|^{-\frac{1}{2}} \int_{-\infty}^{\infty} s(t) g_{AG}\left(\frac{t-u}{\tau_c}\right) du$$

$$\begin{aligned} M_c = & -\log\left[\int_\Omega \left(\frac{d|D|}{du} - \mu_{AM}\right)^2 du\right] \\ & -\log\left[\int_\Omega \left(\frac{d^2 \arg(D)}{du^2} - \mu_{FM}\right)^2 du\right] \\ & +\log\left[\int_\Omega |D|^2 du\right] + \log\Omega(\tau_0) \\ & +2\log\tau_0 \end{aligned} \quad (17)$$

$$\mu_{AM} = \frac{1}{\Omega}\int_\Omega \left(\frac{d|D|}{du}\right) \quad (18)$$

$$\mu_{FM} = \frac{1}{\Omega}\int_\Omega \left(\frac{d^2 \arg(D)}{du^2}\right) \quad (19)$$

We decompose the speech signal into a set of channels, with the characteristic period τ_0

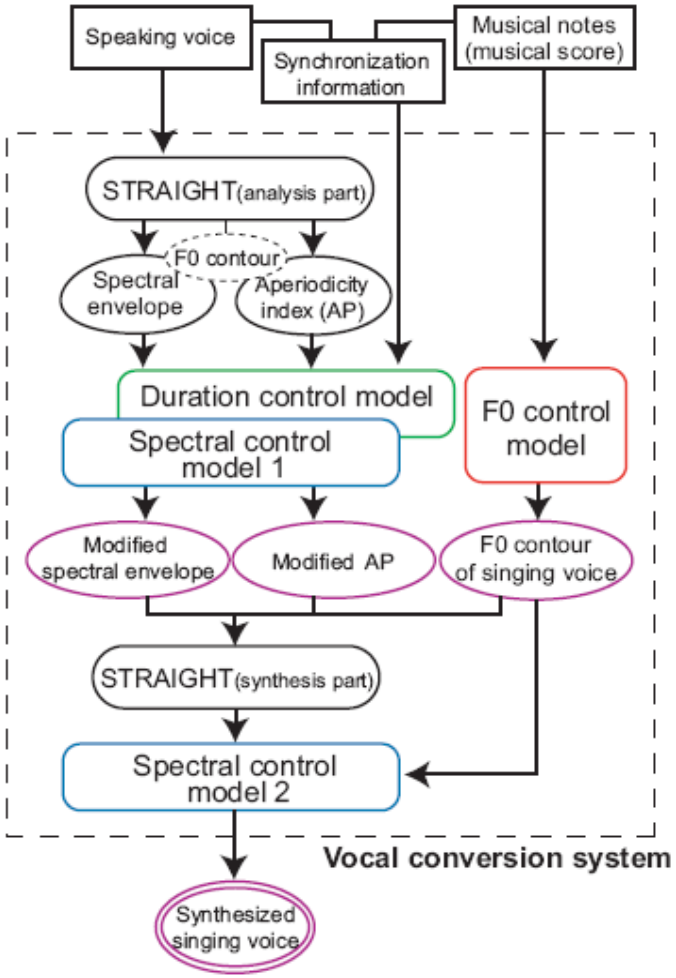
We calculate the fundamentalness index for each channel

The integration interval Ω is proportional to the size of g_{AG}

Fundamentalness – End of the STRAIGHT Part

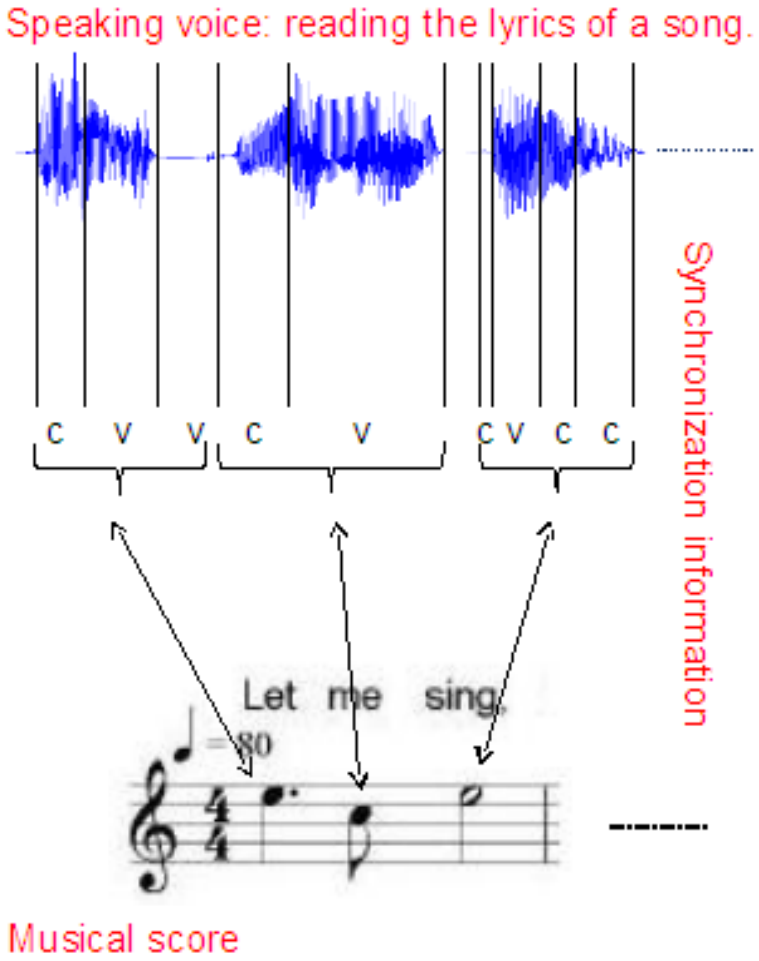
- The fundamentalness concept proved to be very robust and accurate
- The method can be applied to any “fundamental-like” signal, not only to speech
- proposed name: TEMPO (Time-domain Excitation extractor using Minimum Perturbation Operator)

Vocal Conversion System



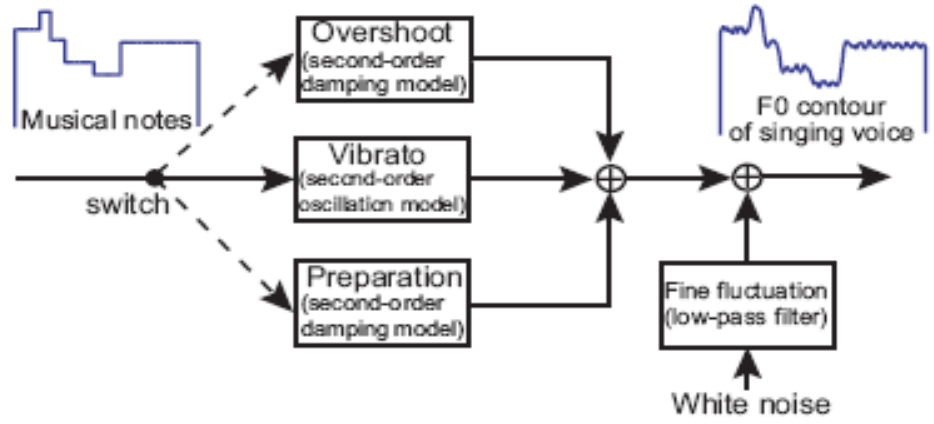
- speech parameters of spoken lyrics are analysed by STRAIGHT
- speech parameters are changed according to music score and empirical know-how
- resulting parameters are re-synthesized by STRAIGHT into singing voice

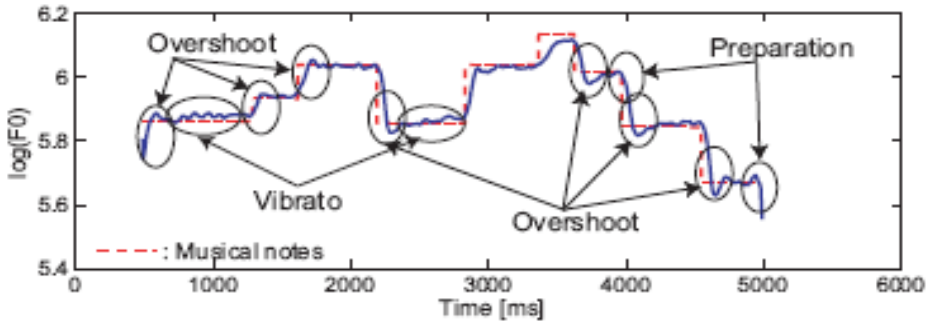
Synchronization



Synchronization between speech signal and musical score (done by hand)

Modelling F0





- Overshoot: exceeding of the target note
- Vibrato: frequency modulation(4-7 Hz) of the F0
- Preparation: deflection to the opposite direction before a note change
- Fluctuations: variations in the F0 contour (>10Hz)

Changing duration

- a consonant followed by a vowel is modelled as
  - consonant part
  - boundary part (last 10ms of the consonant, first 30 ms of vowel part = 40ms)
  - vowel part
- durations of parts are changed
  - consonant part by fixed rates
    - fricative: 1.28
    - plosive: 1.0
    - semivowel: 2.37
    - nasal: 1.43
    - /y/: 1.22
  - boundary part is kept unchanged
  - vowel part is changed that the whole combination fills the note length

Spectral changes

There are two features in singing voices implemented by the authors:

- a strongly present singing formant around 3kHz → emphasize a peak in the spectrogram
- AM of the formants synchronized with the vibrato of F0