A multi-layer F0 model for singing voice synthesis using a B-spline representation with intuitive controls

Luc Ardaillon, Gilles Degottex, Axel Roebel

A multi-layer F0 model for singing voice synthesis using a B-spline representation with intuitive controls

Luc Ardaillon, Gilles Degottex, Axel Roebel
Analysis/synthesis team
IRCAM - UPMC - CNRS

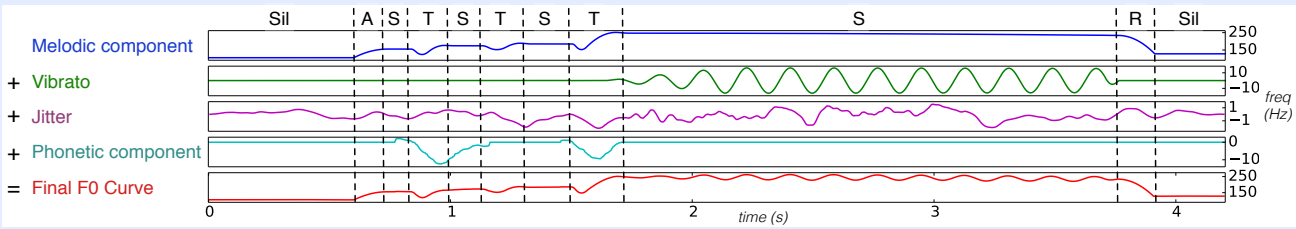
Objectives : Generate natural sounding and expressive F0 curves from a score for singing voice synthesis, with intuitive control.

Motivations : F0 carries not only melody, but also music style, personal expressivity and other characteristics of the voice production, and is thus critical for singing voice synthesis. For compositional purposes, it is necessary to provide an appropriate control over the expressive parameters.

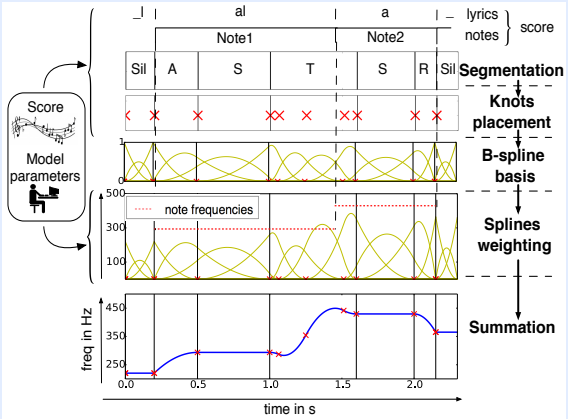
Outline : We propose a new multi-layer F0 curve model for singing voice, using B-splines to generate controllable expressive variations with a set of intuitive parameters. This model has been implemented in a concatenative singing voice synthesis system and evaluated through listening tests.

F0 Modeling :

The F0 curve is decomposed into 4 separate layers, and its temporal evolution is modeled as a succession of 5 basic sements :



Construction of the curve :



Model parameters and score information determine the curve's segmentation, knots positions, and weights of the B-splines

Melodic component :

- Represents attacks, transitions and releases
- Modeled using B-splines
- Each segment has intuitive control parameters :

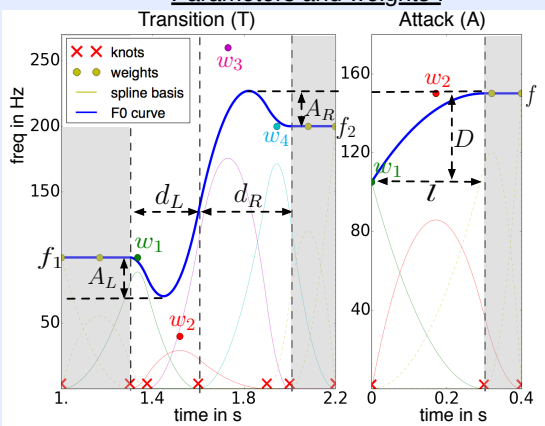
Transition parameters :

A_L : preparation amplitude
A_R : overshoot amplitude
d_R : left part duration
d_L : right part duration

Attack parameters :

D : depth
l : length

Parameters and weights :



Transition weigths :

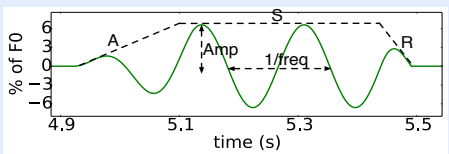
$w_1 = f_1 \; ; \; w_2 = f_1 - \triangle f \cdot A_L$
$w_4 = f_2 \; ; \; w_3 = f_2 + \triangle f \cdot A_R$

Attack weights :

$w_1 = f - D \cdot f$
$w_2 = f$

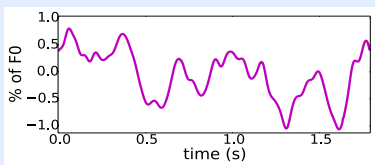
Vibrato :

- Sinusoid with fixed frequency scaled by an ASR curve
- Applied on sustain segments



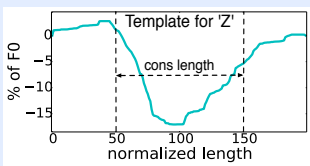
Jitter :

- Normalized templates extracted from sustained notes without vibrato
- Concatenated with 200ms crossfades



Phonetic component : (*microprosody*)

- Average templates for each voiced consonant. Normalized in time and frequency.
- Re-scaled at target phoneme length and F0 for synthesis

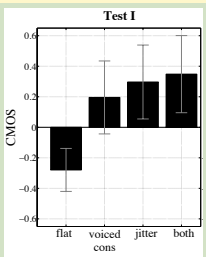


Listening tests :

Evaluate necessity and suitability of each layer, and overall perceived naturalness of the model.
➡ Compare pairs of similar synthesized examples with different F0 curves (CMOS test).

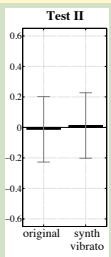
Material :
Simple melody with vs. without jitter & phonetic component (no vibrato)

Result :
➡ Jitter and phonetic components add naturalness to the synthesis



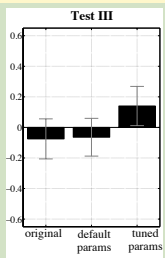
Material :
original F0 (from real singing) vs. synthetic vibrato

Result :
➡ No preference observed between real and synthetic vibratos



Material :
original F0 vs. model with default params vs. model with tuned params

Result :
➡ No preference observed between original and synthesized F0



Conclusion and future work :

The proposed model has proved to be capable of generating natural sounding F0 curves, while offering meaningful control parameters. But manual tuning of all parameters may still be fastidious. A next step in our work will be to provide automatic learning of the parameters in context, in order to model singing styles and alleviate the required amount of manual tuning.