Hybrid Timbre Synthesis

James W. Beauchamp
Ailin Zhang

Objective

Modify the timbre of an input music signal so that it resembles a different instrument while preserving the input's pitch temporal variations, RMS amplitude and spectral centroid.

RMS measures the dynamic character

spectral centroid measures the brightness

Method

How to make the timbre resemble another instrument?

Replace input signal spectrums with the spectral envelope of the training signal.

timbre descriptor

Method

How to preserve pitches?

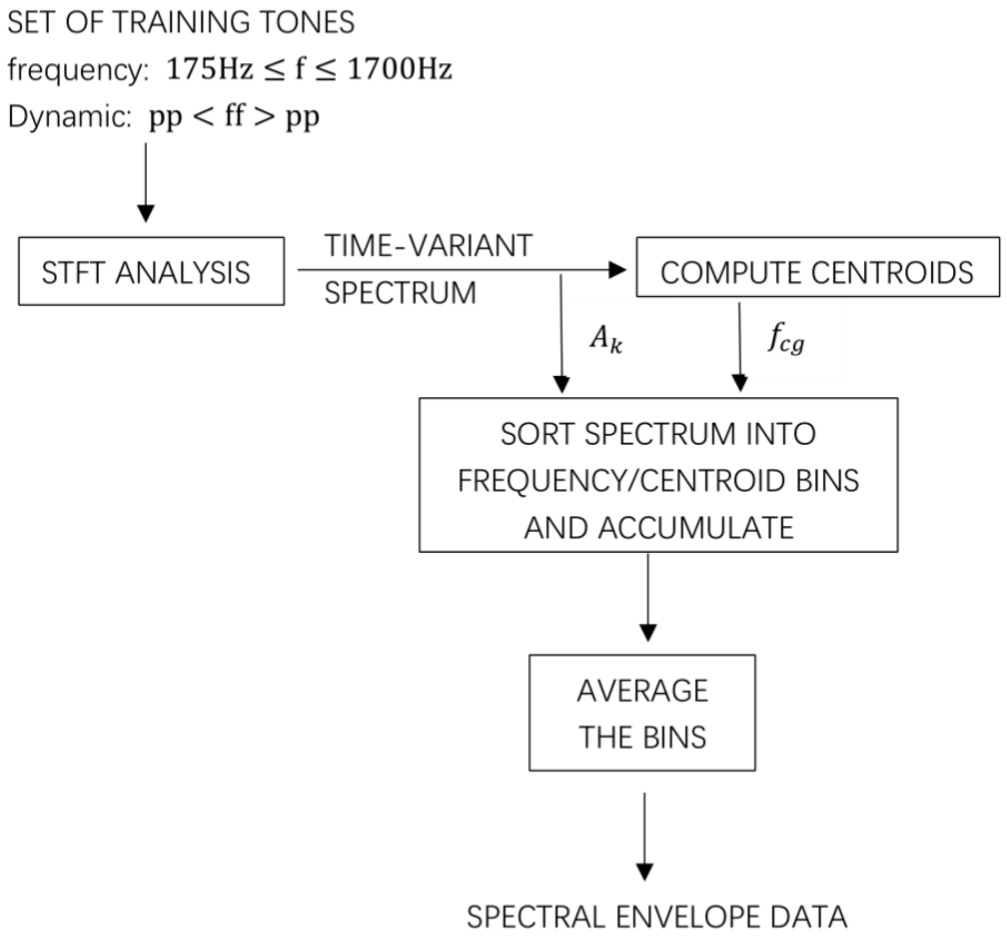
Perform pitch detection on the input to obtain its time-varying fundamental frequency. Sample spectral envelopes using the harmonic frequencies of the input.

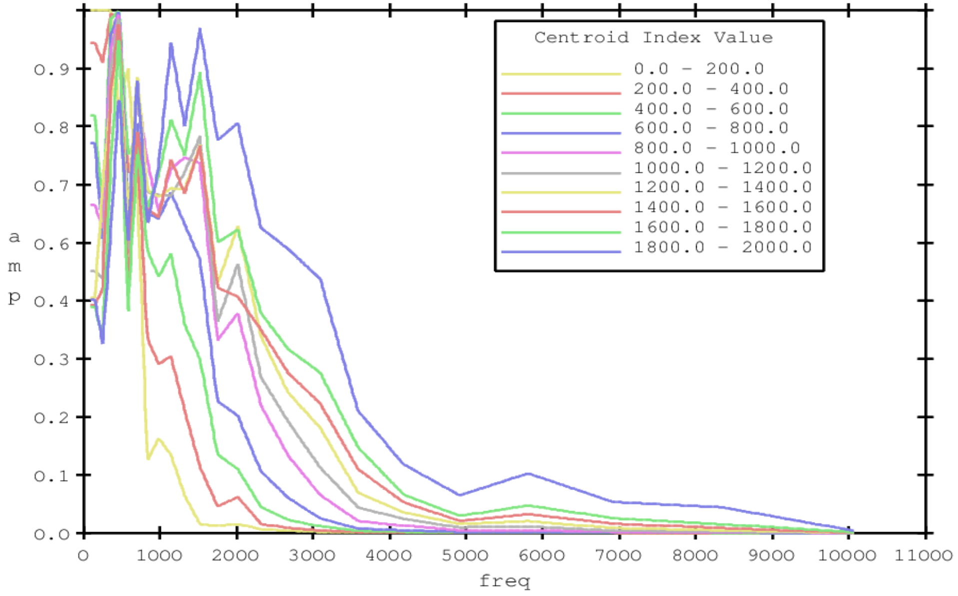
Method

How to create the output signal?

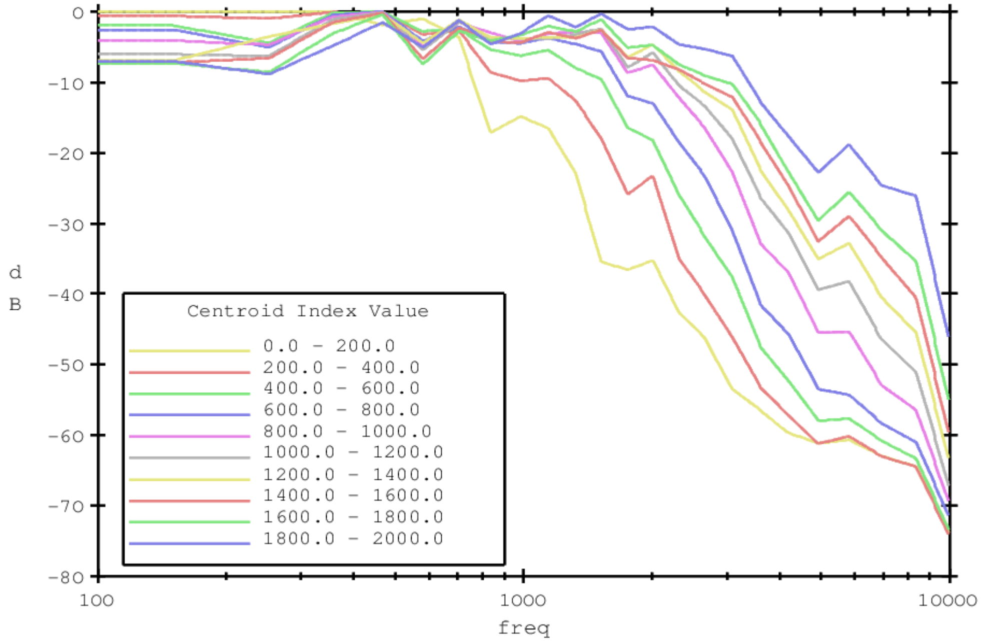
Do sine wave additive using harmonic frequencies.

Spectral Envelope



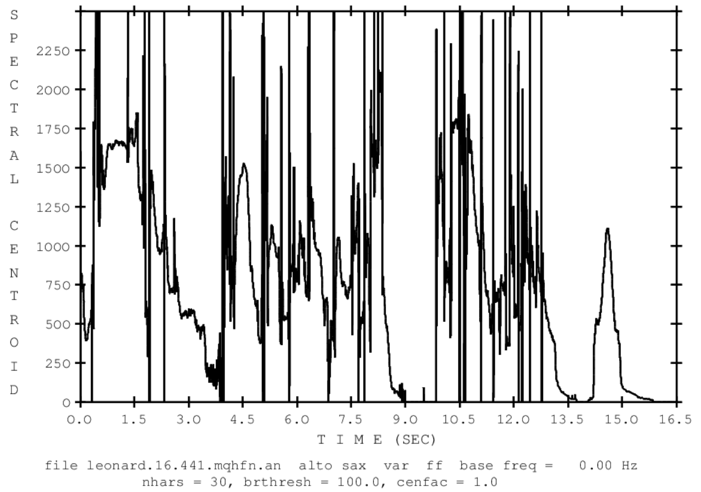


file trumpet.tf, degree = 8

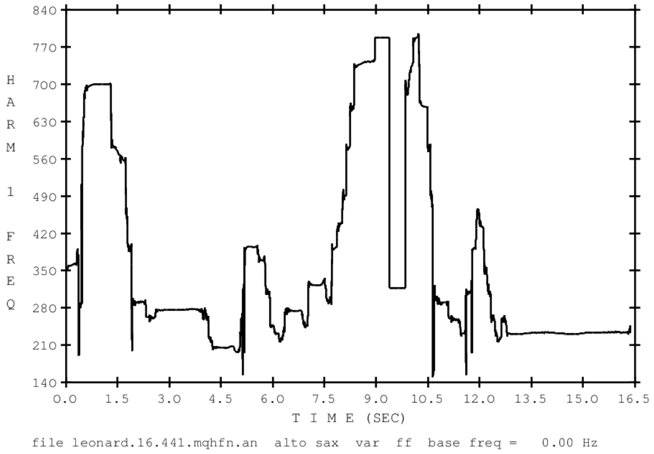


file trumpet.tf, degree = 8

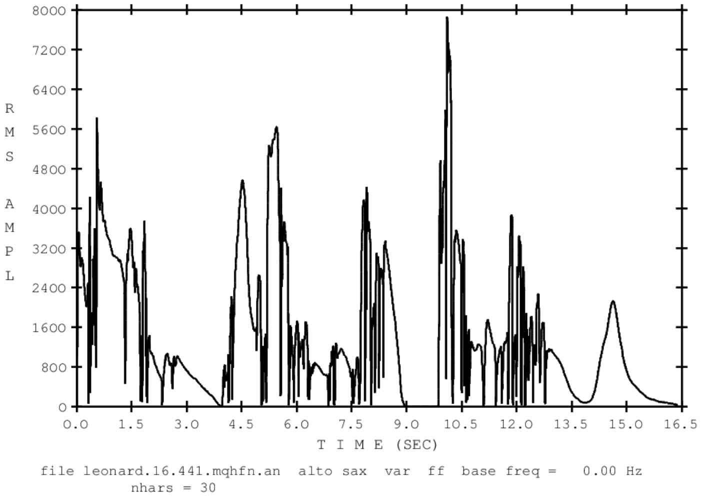
Key Point——
find the proper spectral envelope that matches input spectrum centroid, RMS and pitches



$f_{cg}(t)$



$f_1(t)$



RMS(t)

Procedure

1) Spectral envelope data
2) Input signal analysis

Input signal analysis

Perform STFT and pitch detection on the input sound signal. At each frame, compute:

fundamental frequency $f_1(t)$

Harmonic frequency amplitudes $A_k(\mathrm{t})$

Spectrum centroid

$$f_{cg}(t) \ = f_1(t)(\sum_{k=1}^{N(t)} kA_k(t) \,/ \sum_{k=1}^{N(t)} A_k(t) - 1)$$

RMS

$$\mathrm{RMS(t)} = \mathrm{sqrt}(\sum_{k=1}^{N(\mathrm{t})} A_k(\mathrm{t})^2)$$

Procedure

1) Spectral envelope data
2) Input signal analysis
3) Centroid match

Centroid match

Find spectral envelope number n that satisfies

$$\hat{f}_{cg_n}(t) < f_{cg}(t) < \hat{f}_{cg_{n+1}}(t)$$

where

$$\hat{f}_{cg_n}(t) = f_1(t)\left\{\frac{\sum_{k=1}^{N(t)} kSP_n[kf_1(t)]}{\sum_{k=1}^{N(t)} SP_n[kf_1(t)]} - 1\right\}$$

is the centroid value of the nth spectral envelope

Centroid match

Then do linear interpolation between the nth and (n+1)th spectral envelope to obtain the optimal spectral envelope SP, SP satisfies

$$\frac{SP(f_k) - SP_n(f_k)}{SP_{n+1}(f_k) - SP_n(f_k)} = \frac{f_{cg}(t) - \hat{f}_{cg_n}(t)}{\hat{f}_{cg_{n+1}}(t) - \hat{f}_{cg_n}(t)}$$

for all harmonic numbers, where

$$\frac{SP_n(f_k) - SP_n[m]}{SP_n[m+1] - SP_n[m]} = \frac{f_k - f[m]}{f[m+1] - f[m]}$$

where $f[m]$ and $f[m+1]$ are the middle band frequencies that straddle f_k, f_k is the kth harmonic frequency at each frame.

Centroid match

Problem:

it is possible that

$$0 < f_{cg}(t) < \hat{f}_{cg_1}(t)$$

or

$$f_{cg}(t) > \hat{f}_{cg_M}(t)$$

where M is the number of spectral envelopes

Centroid match

Solution:

1) Add 0-centroid spectral envelope SP_0 and interpolate between SP_0 and SP_1 when $0 < f_{cg}(t) < \hat{f}_{cg_1}(t)$

$$SP_0(f) = \begin{cases} 1 & f = f_1(t) \\ 0 & otherwise \end{cases}$$

Centroid match

Solution:

2) Use frequency threshold

at each frame we use $N(t)$ harmonics, where $N(t)$ is the maximum harmonic number with harmonic frequency below the Nyquist frequency of the training signals.

Procedure

1) Spectral envelope data
2) Input signal analysis
3) Centroid match
4) Rescale to match RMS

Rescale to match RMS

Compute the spectral envelope RMS

$$\mathrm{RMS(t)}_{SP} = \mathrm{sqrt}(\sum_{k=1}^{N(\mathrm{t})} SP_n[kf_1(t)]^2)$$

Compute input signal's spectrum RMS

$$\mathrm{RMS(t)} = \mathrm{sqrt}(\sum_{k=1}^{N(\mathrm{t})} A_k(\mathrm{t})^2)$$

The new kth harmonic amplitude is

$$A_k(t)' = SP_n[kf_1(t)] * \frac{\mathrm{RMS(t)}}{\mathrm{RMS(t)}_{SP}}$$

Procedure

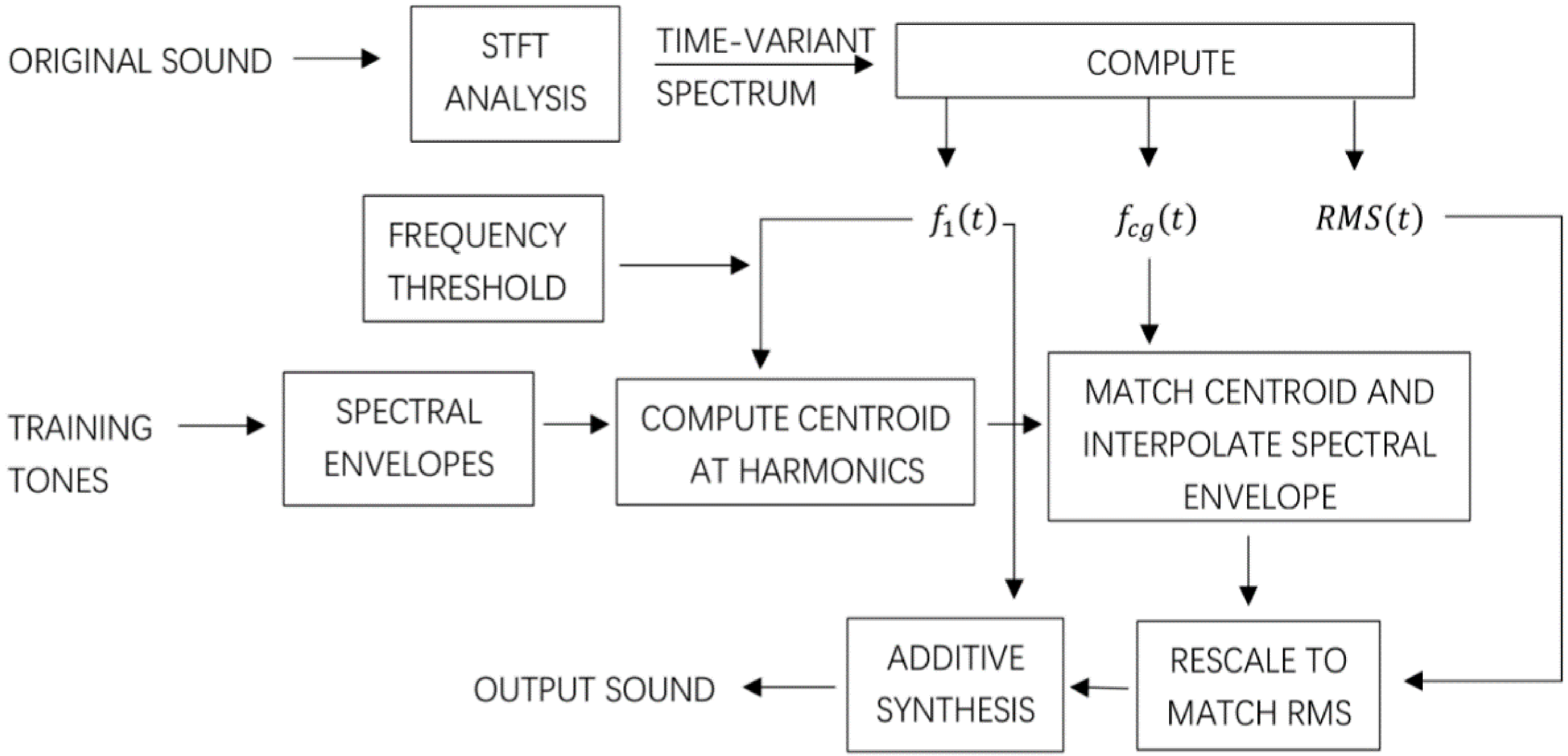
1) Spectral envelope data
2) Input signal analysis
3) Centroid match
4) Rescale to match RMS
5) Additive synthesis

Additive synthesis

The output sound signal is created by sine wave additive synthesis

$$\mathrm{s(t)} = \sum_{k=1}^{N(t)} A_k(t)' \sin(2\pi k f_1(t) t + \varphi(t))$$

Procedure



ORIGINAL SOUND
STFT ANALYSIS
TIME-VARIANT SPECTRUM
COMPUTE
$f_1(t)$
$f_{cg}(t)$
$RMS(t)$
FREQUENCY THRESHOLD
TRAINING TONES
SPECTRAL ENVELOPES
COMPUTE CENTROID AT HARMONICS
MATCH CENTROID AND INTERPOLATE SPECTRAL ENVELOPE
RESCALE TO MATCH RMS
ADDITIVE SYNTHESIS
OUTPUT SOUND

Result

input : saxophone

training data: trumpet

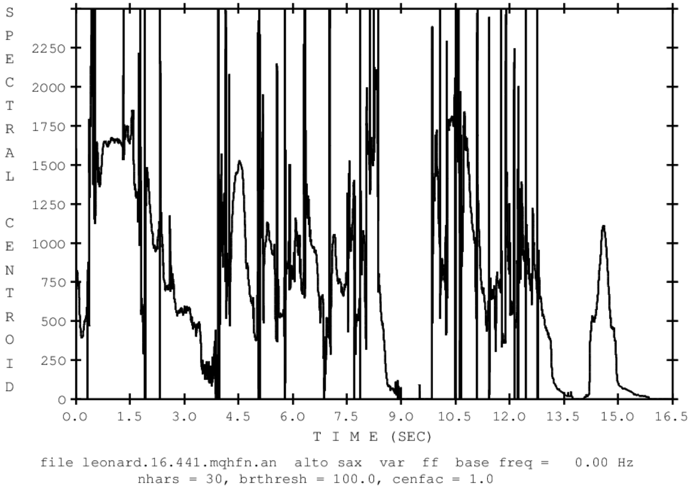
input : tenor

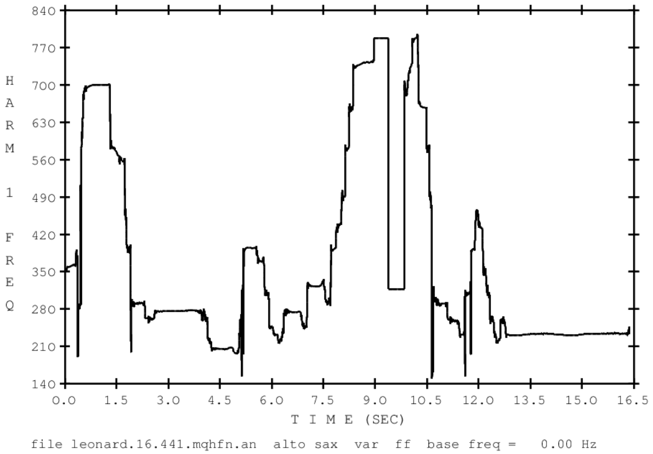
training data: trumpet

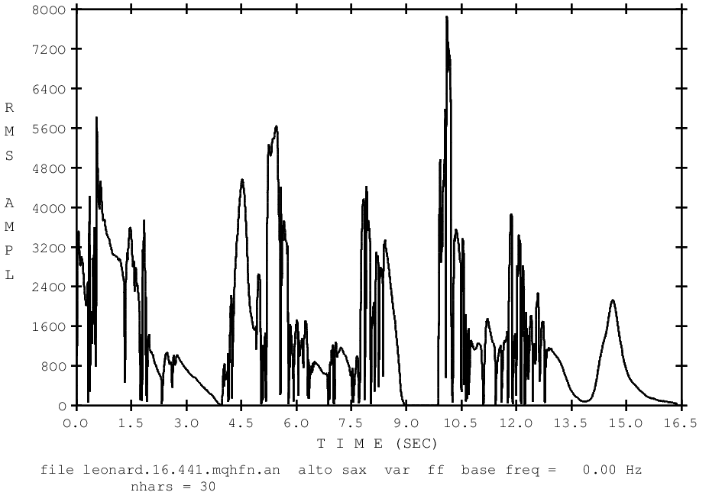
Result

$f_{cg}(t)$ | $f_1(t)$ | RMS(t)

INPUT







OUTPUT