Spectral Modeling Synthesis

February 6, 2019

1 Spectral Modeling Synthesis

Prof. Xaviers Thesis

- Motivation - To obtain **musically useful** intermediate representation for sound transformations by modelling the spectral characteristics of sound
- Underlying Assumption

$$x = x_{sine} + x_{stochastic}$$

Where, $x_{sine} = \Sigma_i A_i[n] sin(\omega_i[n] + \phi_i[n])$ is a sinusoid captured by time varying amplitude, frequency and phase and $x_{stochastic}$ is the stochastic(non-deterministic) component

What constitutes a **good** transformation? - Flexibility(ease of transformation) - Computationally Efficient - Should faithfully reproduce the original sound with as good quality as it can

1.0.1 Background on some Synthesis Techniques

- Historical Background -
  1. Tape Recorders
  2. Analog tapes(Music Concrete)
  3. Digital
- Techniques borrowed from Speech Analysis -
  1. Vocoder

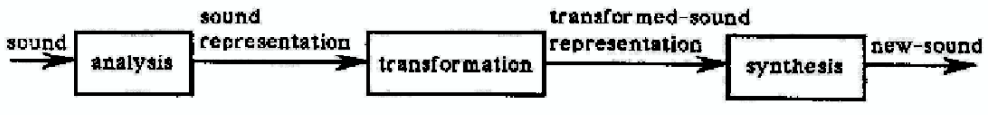


Figure 1: Diagram of a general analysis/synthesis system.

title

    - Modeling of speech by an excitation waveform(sound source) which is filtered(vocal tract)
    - Were able to obtain interesting sound effects(pitch modification, timbre morphing)
  2. Linear Predictive Coding
    - Linear time varying filtering
  3. Phase Vocoder
    - Representing signals by the short time phase and amplitude spectrum
    - Major motivation to move towards the Short Time Fourier Transform(STFT)

- Synthesis Methods -
  1. LPC based synthesis
    - wide variety of transformations because of the decomposition
    - works well when analyzed sounds have **clear formant structure**
  2. Analysis based synthesis -
    1. Heterodyne filtering
      - breaks input waveform into pseudoperiodic segments and then estimates the pitch of each pseudoperiodic segment
      - Similar to STFT, analyzes signal at multiple, evenly-spaced time points
      - The Application of Heterodyne Filter Analysis and Linear Predictive Coding using cSound's ADSYN, LPREAD, and LPRESON Opcodes
    2. Phase Vocoder
      - Manipulate Temporal and Spectral Features independently(decouple them)
    3. Formant wave-function synthesis
      - Directly modelling the time domain amplitude
      - Time Domain FoF
      - Final Project: Formant-Wave-Function Synthesis
    4. VOSIM
      - model with sinc pulses of variable amplitudes, delays
      - paper
    5. Wavelet transform
      - Wavelets as analysis functions

1.0.2 Short Time Fourier Transform

- Why perform analysis in the spectral domain?
  - Our ear is like a harmonic analyzer, thus spectral analysis mimics the bahaviour of the ear
  - Cochlea is likened to a set of narrow band pass filters, thus it performs some kind of FT
- How our ear is different?
  - Our ear obtains a **log scale spectrum** as opposed to the linear spectra obtained by conventional FT
  - Time and Frequency domain masking
  - Amplitude perception relative to frequency

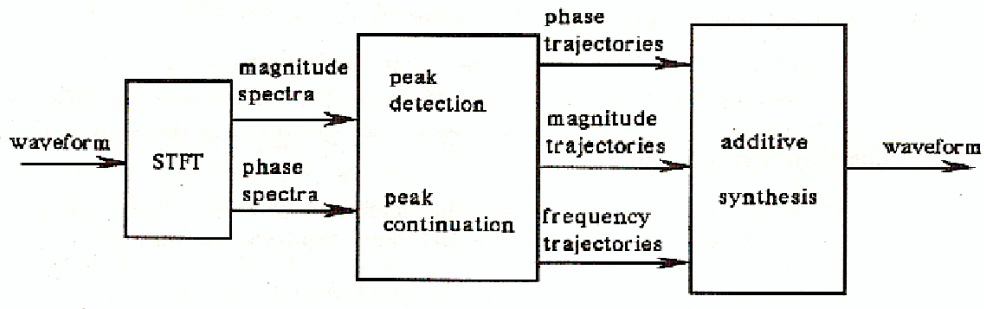


Figure 3.1: General block diagram of the sinusoidal system.

title

- Hearing and Perception

The STFT equation - $X_l(k) := \Sigma_{n=0}^{N-1} w(n)x(n+lH)e^{-j\omega_k n}$

2 important (controllable) parameters - 1. Analysis window w(n) - Determines time vs frequency resolution - Want narrow main lobe, low side lobe - For phase detection, constant phase spectrum obtained by using symmetric window 2. Hop size H - Depends on sound characteristics

Why move on? - Cannot manipulate sounds easily

Treat this as an intermediate step to obtain a more flexible representation

1.0.3 Sinusoidal Model

Model a signal as a sum of time varying sinusoids

$$s(t) = \Sigma_{r=1}^{R} A_r(t) cos(\theta_r(t)) \tag{1}$$

$$\theta_r(t) = \int_0^t \omega_\tau(\tau) d\tau + \theta_r(0) + \phi_r \tag{2}$$

Here, **R** is the number of sinusoidal components, $A_r(t)$ is the instantaneous amplitude and $\theta_r(t)$ is the instantaneous phase

The main steps in the parameter extraction are -

1. Spectral Peak Detection -
    - Peak detection
        - Local maxima in the magnitude spectrum at each time frame
        - Filtering the maxima with some threshold measure
        - For perceptual purposes, use knowledge of equal loudness contours
    - Peak interpolation
        - Return a better estimate for the frequency than the bin value
        - Fit a parabola to the frequency, and use the peak of parabola as estimate The output of this stem is the estimated magnitude, frequency and phase of the prominent peaks in the STFT for each time frame

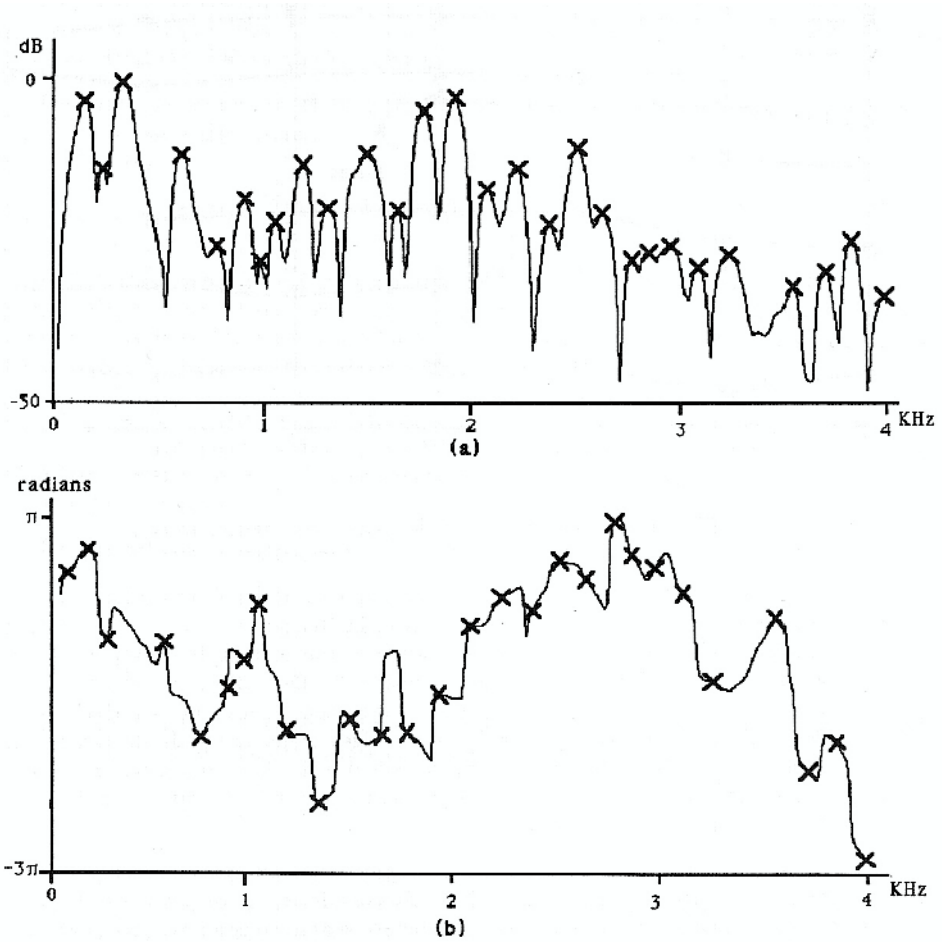


Figure 3.4: Peak detection on a spectrum of a piano attack sound: (a) magnitude spectrum, (b) phase spectrum.

title

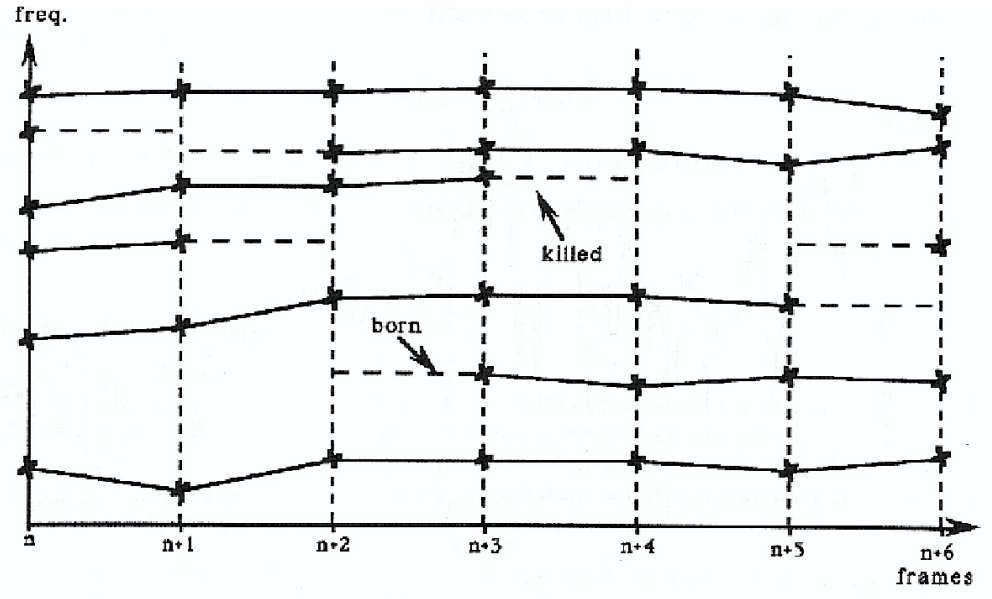


Figure 3.5: Illustration of the peak continuation process.

title

2. Spectral Peak Continuation -
    - Map the peaks at the $(n-1)^{th}$ time frame to the $(n)^{th}$ time frame
    - Find the peak in the $(n)^{th}$ time frame which is closest in frequency in the previous frame
    - Possible approaches - heuristic(rule based), probabilistic(hmm)

Once the parameters are obtained, the sound is synthesized by generating the sum of sinusoids for each time frame in the following way -

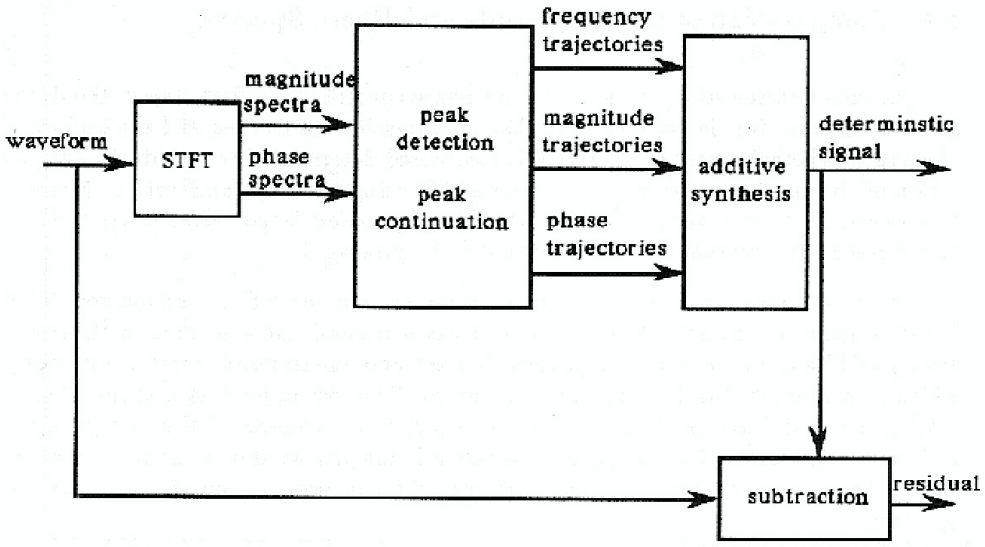
$$s^l(m) = \Sigma_{r=1}^{R} \hat{A}_r^l cos(m\hat{\omega}_r^l + \hat{\phi}_r^l) \tag{3}$$

Sound Effects - Can be easily achieved by playing around with the obtained parameters(scaling, interpolation, filtering etc.). For ease of manipulation, only the magnitude spectra is used as the ear is mainly sensitive to the spectral magnitude and not the phase

Why move on? - Difficult to model **noise** with sinsusoids(need a large number) - Because of the lack of noise modelling, the percieved quality is a bit artificial during transformations
This is motivation to model the noise in the signal as the next step

1.0.4 Deterministic + Residual Model

How is the **deterministic** component different from the previous? - As opposed to selecting any peak in the spectrum(like in the previous case), the deterministic models particularly model the partials in the sound - Thus, each sinusoid is assumed to model a **quasi-sinusoidal** component(piecewise linear amplitude and frequency variation) as opposed to any kind of sound



title

The **Residual** in this case is defined $ x_{original} - x_{deterministic} $. It usually models the energy that does not go into vibrations, or any component that is not inherently sinusoidal in nature.

The signal in this case is modelled as -

$$s(t) = \Sigma_{r=1}^{R} A_r(t) cos(\theta_r(t)) + e(t) \tag{4}$$

(5)

Here, e(t) is the residual

Most of the steps in extracting the parameters for the deterministic model are the same as the previous model. But, since the sinusoids are restricted to be partials only here, there is a modification in the **Spectral Peak Continuation** process. Using a heuristic(rule based) and some prior knowledge about the nature of the sound(harmonic, frequency range etc.), an algorithm is proposed which tracks only the clear and stable partials

The deterministic components are synthesized using the parameters obtained. The residual is obtained by subtracting this deterministic signal from the original signal.

An easier alternative is to subtract the frequency spectra of the two signals and ignoring the phase(perceptually unimportant)

Why move on ? - Residual is not flexible for performing transformations

This motivates to further study the residual, and approximate it with a model that can be easily played around with.

1.0.5 Deterministic + Stochastic Model

Observations from the previous models - - Not necessary to preserve phase - Can model the residual as some kind of stochastic signal

Modelling the residual as a stochastic signal helps in easily transforming the signal

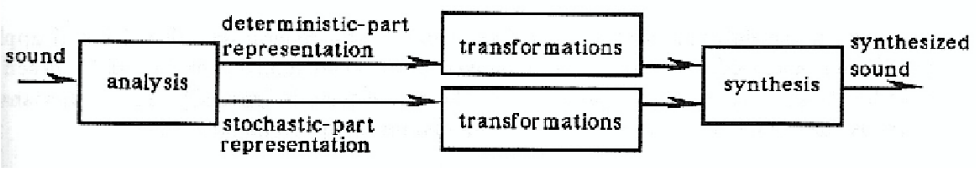


Figure 5.1: General diagram of the deterministic plus stochastic system.

title

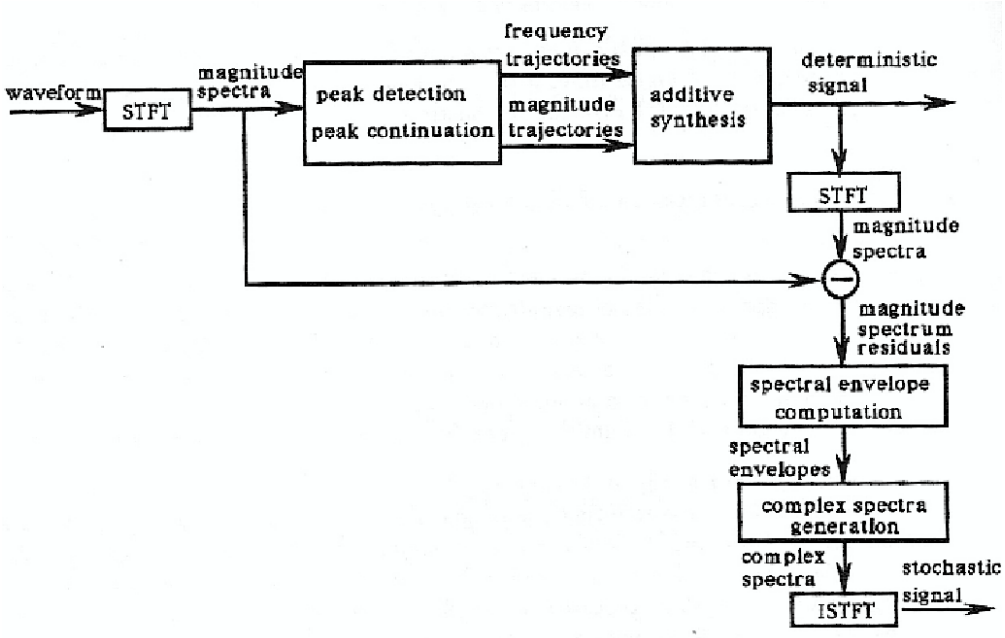


Figure 5.2: Block diagram of the deterministic plus stochastic system.

title

The representation obtained is similar to the previous case, just that the residual e(t) is modelled as a stochastic signal, thus allowing to write as the action of a Linear Time Variant system on white noise.

$$\hat{e}(t) = \int_0^t h(t, t-\tau)u(\tau)d\tau \tag{6}$$

Here, u(t) is white noise and h(t,t') is the filter.

The deterministic component is calculated in the same way as the previous. The parameters are set in such a way as to extract the partials as accurately as possible(to prevent them from appearing in the residual)

Since we assume the residual to be a stochastic signal, it is characterized by its amplitude and frequency.

To obtain the general shape of the residual spectrum, we approximate the envelope of the residual spectrum, which is obtained by subracting the deterministic spectra from the original spectra. This

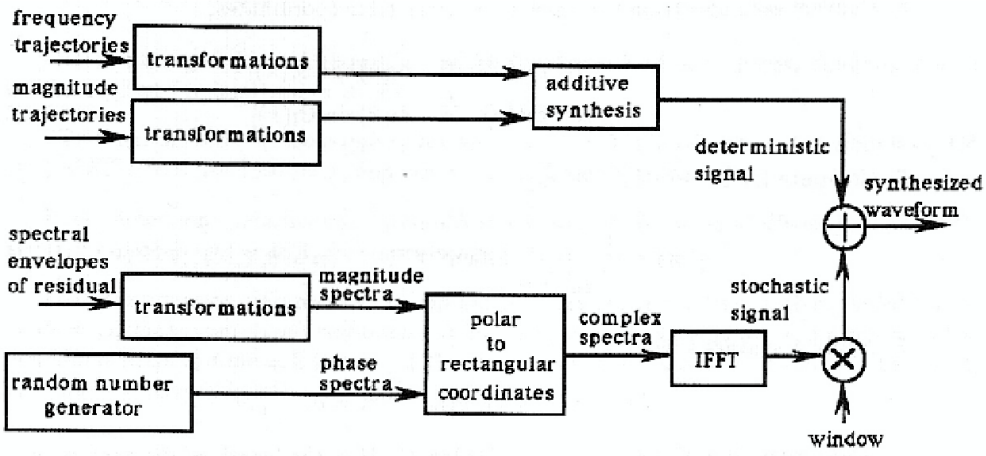


Figure 5.9: Block diagram of the synthesis part of the deterministic plus stochastic system.

fig

is because only the shape of the envelope contributes to the sound characteristics. The envelope is approximated by **curve fitting** or **LPC**.
Once the envelope is obtained, we generate the stochastic signal by using this as our amplitude and generate random numbers as phase

$$\hat{e}(t) = IFT(A(k)e^{j\Theta(k)}) \tag{7}$$

Here, A(k) is the envelope, and $\Theta(k)$ is the phase(random)

Transformations - Can be separately applied to the deterministic and stochastic components. - Deterministic - Similar transformations like before - Stochastic - Envelope shaping, filtering etc.

1.0.6 Examples of sound effects using the above model (Refer 4.pdf)

1. Filtering
2. Pitch Scaling, transposition and discretization
3. Vibrato, tremolo
4. Spectral shape shifting
5. Gender changing
6. Harmonizing
7. Hoarseness
8. Morphing

Musical Instrument Sound Morphing Guided by Perceptually Motivated Features

---

For sound examples, visit this page

What is **Morphing**?
- Blurring Distinction between **Source** and **Target** - Somewhat like creating **hybrid** musical instruments
- Would like to idelly perform **Perceptually Linear** transformations - The morphed sound should not simply sound like a mixture of sounds(the ear can distinguish in such cases). It should rather sound like a single **entity**

How is it done?

- Obtain some kind of reprsentation of the sound, and then have an interpolation function that gradually interpolates these representations from one sound to the other.

- Control the whole morphing process(algorithmically and perceptually) with a single coefficient α , the interpolation factor
- You would ideally want to vary the interpolate the parameters so that the morphed sound vary **perceptually linearly**

In this work, the authors have proposed to seek sound parameters that favor Perceptually Linear transformations

Work done previously
- Mostly interpolate parameters/features without caring much about the perceptual impact

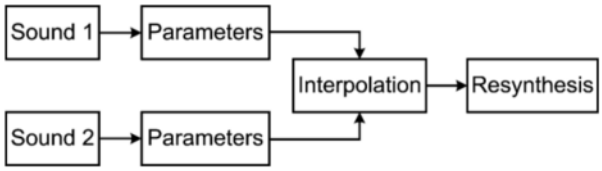


Fig. 2. Depiction of the classic morphing scheme using the interpolation principle, which assumes that perceptually intermediate representations possess intermediate parameter values.

What are parameters,features? - Parameters - Coefficients obtained from sound analysis models(can resynthesize sound from them) - Features - Particular aspects of sound

Methods Used - 1. Parameter interpolation using Wigner Distributions(Time Frequency)
2. GMM models for parameter interpolation
3. **Model Sounds as dynamical systems with ANN**
4. Discrete Wavelet Transform(DWT) + Singular Value Decomposition(SVD)

The above don't consider perceptual factors and suggest suggest interpolation strategies with better perceptual corelations, like the ones below 1. Dynamic Frequncy Warping(DFW) to morph spectral envelopes 2. Multi Dimensional Scaling(MDS)

One important thing to consider in all the above cases is the need for the sound to be **temporally alligned**, or else some kind of smearing might occur, thus making the resultant sound artificial to hear

In this work, as opposed to interpolating parameters directly, the authors propose to first obtain relevant features from the parameters(which might have a more perceptual meaning than the parameters themselves), and then interpolate these features itself.

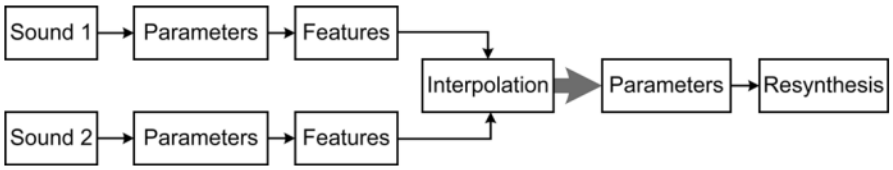


Fig. 3. Depiction of the morphing by feature interpolation principle adopted in this work, which advocates that perceptually intermediate representations present intermediate feature values rather than intermediate parameter values. Notice that the step represented by the grey arrow implies retrieving parameters from features.

However, obtaining parameters from features is difficult(It is not a one-one transformation!). Thus, instead of this approach, the authors propose to use parameters for whom the interpolated sounds features are close to the interpolated feature values(suitable evaluation scheme suggested, use parameters -> feauture values vary linearly when interpolating linearly)

The features the authors use in this work are obtained by finding **accoustic correlates of Timbre Spaces using MDS**(Essentially trying to mathematically describe the Timbre Space). The features are both temporal and spectral.

Temporal

1. log attack time 2. temporal centroid

Spectral 1. spectral centroid 2. spectral spread 3. spectral skewness 4. spectral kurtosis

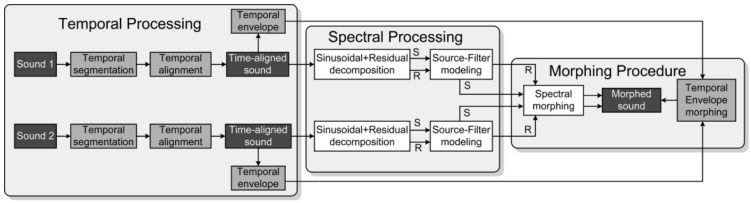


Fig. 4. Depiction of the general steps of the musical instrument sound morphing procedure. There are three distinct parts, temporal processing, spectral processing, and morphing procedure. Blocks with dark grey background represent waveforms, blocks with light grey background represent temporal feature extraction and processing, and blocks with white background represent spectral feature extraction and processing.

The authors proposed model -

Extraction of Parameter - 1. Temporal Segmentation - Segment into ADSR 2. Temporal Allignment - Boundaries should coincide 3. Temporal Envelope Extraction - True Amplitude Envelope(TAE) based on cepstral smoothing 4. Sinusoidal + Residual Model - To obtain the parameters 5. Source Filter Model -

Morphing - 1. Spectral Envelope Morphing - Shift in frequency peaks smoothly 2. Interpolation of partial frequencies 3. Temporal Envelope Morphing

Evaluation - Vienna Symphonic Library - Listening Test - Judge several Characteristics for each morph value - Complicated and very subjective - Proposed Objective error function(assuming linearity, essentially the MSE)