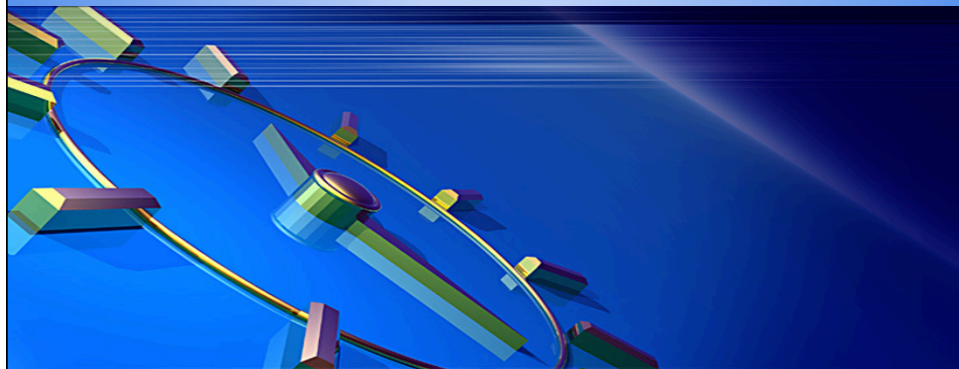


# Symbolic and Structural Representation of Melodic Expression



Authored by Christopher Raphael • Presented by Bo Li

## Contents

- Introduction & Problem formulation
- The Theremin
- Representing Musical Interpretation
- Estimating the Interpretation
- Results



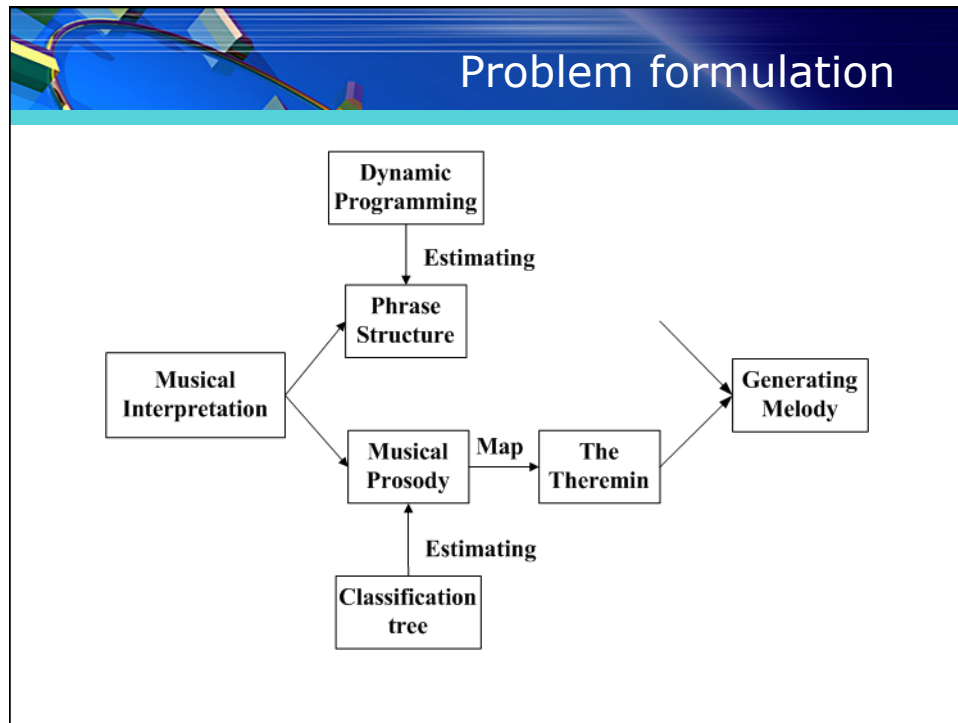
## Introduction

- Most past work on expressive synthesis has concentrated on piano music which is attractive for one simple reason: a piano performance can be described by giving the onset time, damping time, and initial loudness of each note.
- The synthesis of melody finds its richest form with “continuously controlled” instruments, which can produce wide variety of tone color, dynamics and other musical elements by simultaneously modulating many different parameters.
- In recent research work, the performance parameters are computed directly from the observable score attributes with no real attempt to describe any interpretive goals such as repose, passing tone, local climax, etc.



## Problem formulation

- Previous work does not represent the interpretation, but rather treats the consequences of this interpretation, such as dynamic and timing changes.
- The research work presented in this paper, explicitly tried to represent the interpretation itself in two ways.
  - a) Using a tree-like structure decomposition to make various levels of repetition or parallelism in the melody.
  - b) Introducing a hidden sequence of variables which represent the prosodic interpretation itself.



## The Theremin

The theremin known as an early electronic instrument, is capable of producing a rich range of expression. The widely-used mathematical formulation is

$$s(t) = a(t) \sin(2\pi \int_0^t f(\tau) d\tau)$$

$a(t)$  and  $f(t)$  are amplitude and frequency modulated over time.

The above representation is modified to capture the tone color by defining it as a function of amplitude.

$$s(t) = \sum_{h=1}^H A_h(a(t), f(t)) \sin(2\pi h \int_0^t f(\tau) d\tau)$$

$\{A_h\}$  are hand-designed functions, monotonically increasing in the first argument.

## Representing Musical Interpretation

Only two important components are discussed in the paper to investigate the musical interpretation.

**Phrase Structure:**  
typically the hierarchical structure is captured by simple tree structures, often involving binary groupings at various levels of grouping.

**Musical Prosody:**  
the placing, avoidance, and foreshadowing of local stress and the associated low-level groupings that follow.

## Phrase Structure

Since the folk-like music treated in the paper is mostly composed of simple musical structure, with a high degree of repetition of musical elements, the tree structure is applied to make musical phrases correspond to levels of tree.

Example: God Save the Queen may be described (2-2-2)+((2-2)-(2-2)), each number represents a group of measures, "+" denotes concatenation and "-" denotes grouping.

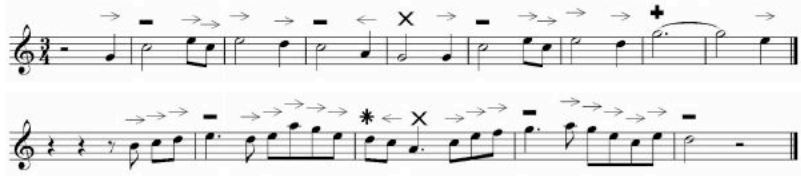
## Musical Prosody

The main focus of this paper is on musical prosody, which is represented by a series of symbols from a small alphabet.

$$A = \{l^{-1}, l^x, l^+, l^{\rightarrow}, l^{\leftarrow}, l^*\}$$

Stresses or points of "arrival"			Moving forward towards a future stress		Receding movement
$l^{-1}$	$l^x$	$l^+$	$l^{\rightarrow}$	$l^*$	$l^{\leftarrow}$
direct and assertive stress	"soft landing" stress	continues forward in anticipation of future unfolding	"garden-variety" passing tone	Reserved for the passing stress or highlight a recurring beat-level emphasis	Receding movement when a note is connected to the stress that precedes it

## Musical Prosody



The figure shows two staves of musical notation. The top staff is for 'Amazing Grace' and the bottom staff is for 'Danny Boy'. Above the notes, various symbols from the alphabet are placed to indicate musical prosody, such as arrows, crosses, and asterisks.

**Figure 1.** *Amazing Grace* (top) and *Danny Boy* (bot) showing the note-level labeling of the music using symbols from our alphabet.

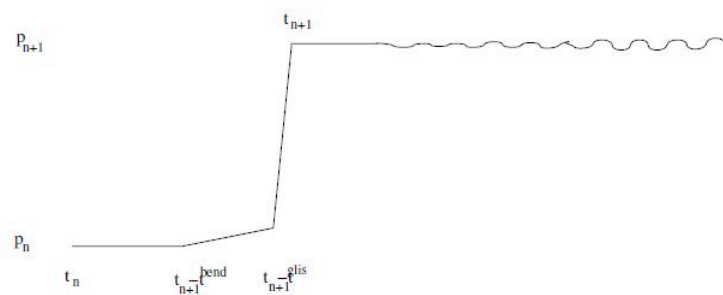
## Mapping from Labeling To Audio

A deterministic mapping is devised to translate the prosodic labeling of a melody, using symbols from A, to the amplitude and frequency function for sound synthesis.

The synthesis of  $f(t)$  and  $a(t)$  begins by modifying the literal interpretation of musical timing expressed in the score to include slowing down at the ends of phrases.

a) Modify  $f(t)$  to include vibrato to long and stressed notes, and bend each pitch in towards the following pitch with a final glissando to encourage a sense of legato.

## Mapping from Labeling To Audio



**Figure 2.** A graph of the frequency function,  $f(t)$ , between two notes. Pitches are bent in the direction of the next pitch and make small *glissandi* over the transitions.

## Mapping from Labeling To Audio

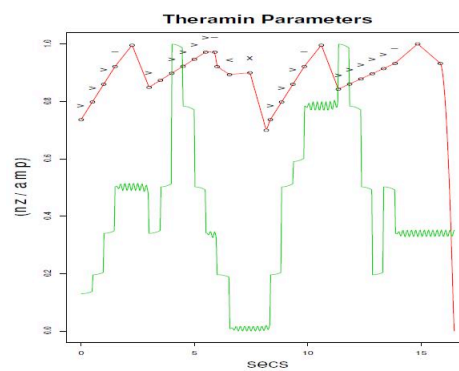
b) The heart of the transformation is in the construction of the amplitude function  $a(t)$ .

This function is created through a series of soft constraints that are placed on the amplitude defined at various "knot" locations over time, which are taken from the prosodically-annotated score and the structure representation.

i) Adding quadratic penalty to make phrase beginnings low in amplitude, high amplitude in stress notes etc.

ii) Computing the values at the knot locations by minimizing the quadratic penalty function, and interpolate the resulting amplitudes at the knot locations.

## Mapping from Labeling To Audio



**Figure 3.** The functions  $f(t)$  (green) and  $a(t)$  (red) for the first phrase of *Danny Boy*. These functions have different units so their ranges have been scaled to 0-1 to facilitate comparison.

## Estimating the Interpretation

The essential goal of this work is to algorithmically generate expressive renderings of melody. Having formally represented notion of musical interpretation, an expressive rendering by estimating phrase structure and musical prosodic labeling can be generated.

i) The estimation of melody structure is obtained by maximizing an objective function defined on the decomposition using dynamic programming.

ii) Viewing the label sequence (A) as a Markov chain, and estimating and modeling the conditional distributions by classification tree methodology. Then a optimizing labeling can be computed by dynamic programming.

## Estimating Phrase Structure

Dynamic programming approach is applied to choose the best scoring of all labelings as structure estimate.

a) Firstly, each note subsequence was labeled as a single group with no subdivisions, the score of each possible label for the subsequence is computed.

c) Secondly, grouping the measures which share the values, and piecing together  $k$  identically labeled segments and labeling the result as  $(k-k-\dots-k)$  by finding the optimal score.

c) For the final production phase, a concatenation operator is used to complete collection of melody notes which can not be captured by a regular tree structure.

## Estimating the Prosodic labeling

The fundamental modeling assumption views the label sequence as a Markov chain, given the data,  $y$ :

$$p(x | y) = p(x_1 | y_1) \prod_{n=2}^N p(x_n | x_{n-1}, y_n, y_{n-1}) = p(x_1 | y_1) \prod_{n=2}^N p(x_n | x_{n-1}, z_n)$$

Where  $z_n = (y_n, y_{n-1})$ . The feature vector  $y_n$  measures attributes of the musical score at the  $n$ th note.  $x_n \in A$

a) The score data were split into  $|A|$  groups,  $D_l = \{(x_{it}, z_{it})\}$  to train the conditional distributions  $p(x_n | x_{n-1}, z_n)$  by classification tree methodology.

b) Given a piece of music with feature vector  $z_1, \dots, z_N$ , the optimizing labeling is computed by dynamic programming

$$\hat{x}_1, \dots, \hat{x}_N = \arg \max_{x_1, \dots, x_N} p(x_1 | y_1) \prod_{n=2}^N p(x_n | x_{n-1}, z_n)$$


## Results

When computing the most likely labeling for each melody, a total of 678/2674 errors (25.3%) is found in the following figure.

The error rate of 15.3% indicates the forward-moving labels and stress labels are subject to interpretation, while using these categories.

	$l^*$	$l^{\rightarrow}$	$l^{\leftarrow}$	$l^-$	$l^\times$	$l^+$	total
$l^*$	135	112	0	18	2	0	267
$l^{\rightarrow}$	62	1683	8	17	0	0	1770
$l^{\leftarrow}$	3	210	45	6	2	0	266
$l^-$	49	48	4	103	15	0	219
$l^\times$	5	32	2	65	30	0	134
$l^+$	0	3	0	12	3	0	18
total	254	2088	59	221	52	0	2674

**Figure 4.** Confusion matrix of errors over the various classes. The rows represent the true labels while the columns represent the predicted labels. The block structure indicated in the table shows the confusion on the coarser categories of stress, forward movement, and receding movement



## Results

a) A subset of the most well-known melodies of the dataset and created audio files from the random, hand, and estimated annotations, were used to compare the perceived musicality of the performances.

b) The response from 23 subjects shows no preference for the hand annotations over the estimated annotations, while both of them were clearly preferred to the random annotations.

The example of Danny-boy

Estimated annotation: 🗣️

Hand annotation: 🗣️

Random annotation: 🗣️



# Thank You !

