INFORMATICA, 1998, Vol. 9, No. 3, 343–364

Preprocessing of Folk Song Acoustic Records for Transcription into Music Scores

Gailius RAŠKINIS
Vytautas Magnus University, Faculty of Computer Science
Daukanto 28, 3000 Kaunas, Lithuania
e-mail: idgara@vdu.lt

Received: May 1998

Abstract. This paper describes a preliminary algorithm performing the mapping of sound to music score. Our procedure is constructed over signal-extracted energy and fundamental frequency traces alone. The algorithm is tested on real songs of average complexity. Although results seem to be promising, their detailed examination reveals some shortages of our approach as well as the set of application specific problems. It appears that musical analysis can not be entirely dissociated from phonetic processing. Further work should be oriented towards integration of knowledge of music as well.
Key words: vocal signal analysis, pitch extraction, sound to music score mapping.

1. Introduction

Lithuanian ethnology-related repositories store more than 170 000 Lithuanian folk song records accumulated over eighty past years. Actually, this treasure of Lithuanian culture is subjected to conservation, analysis and editing processes. The algorithm presented in this paper intends to be used as a tool of folk song analysis. It is designed to automate the transcription of acoustic song records (Fig. 1) into corresponding music scores (Fig. 2), the task currently requiring efforts of multiple professional musicians.

Lithuanian folk song melodies belong to the family of tone-based music. Nevertheless, folk singing is never rigorously harmonic in the sense of theory of music. It is also characterized by the abundance of ornamentation. These particularities of natural singing

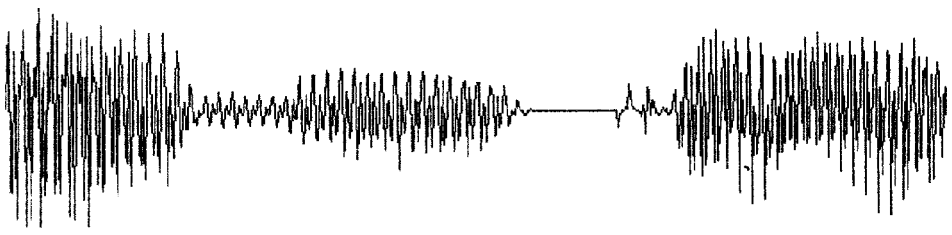

Fig. 1. Example of song signal.

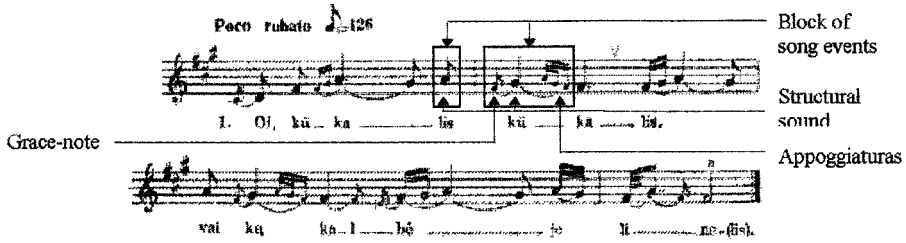

Fig. 2. Example of musical notation used for folk song transcription.

can not be expressed by the customary musical notation. For this, song transcription experts employ an extended notation system enriched by special signs (Fig. 2).

This paper describes our preliminary approach to the problem of sound to music score mapping. We call it preprocessing as current folk song analysis is based on acoustic data alone. In fact, music score is far from being a sequence of duration-tone pairs of random nature. Therefore, musical analysis might be effectively guided by domain knowledge. This issue is viewed as an object of our future research.

We also restrain this work by assuming several simplifications. First, we limit our analysis to monodic style of singing, i.e., to folk songs performed by one person at a time. Second, we consider only some of all possible ornamentation patterns. Finally, our algorithm is not required to show real-time performance.

The rest of this paper is organized as follows. Section 2 briefly presents some related research. Part 3 explains our preprocessing algorithm. In Part 4 we examine the results of application of this algorithm. The paper is finished by our conclusions and discussion about future research directions.

2. Related Research

The problem of sound to music score mapping received limited attention in comparison to other issues tackled by the community of acoustic signal processing research. We shall mention two recent works in this field.

Quiros and Enriquez (1994) describe real-time pitch-to-MIDI[1] converter. It focuses on the extraction of fundamental frequency[2] from singing voices and acoustic musical instruments. Fundamental frequency is transformed into musical data and converted into MIDI messages by means of a neural network which encodes fuzzy logic rules. This algorithm showed accurate performance on a few short phrases of singing voice and instrumental music taken from a CD record.

IRCAM[3] developed an application which follows the progress of musician's performance referring to the musical part provided beforehand. This system also encompasses

[1] Musical Instrument Digital Interface.

[2] *Fundamental frequency* or *pitch.* is a physical property of the waveform. It should not be equated to the perceptive tonal height which is a perceptive variable.

[3] Institut de Recherche et Coordination Acoustique / Musique.

a module performing signal to MIDI score conversion (Doval, 1994). Here, the waveform is transformed into frequency and amplitude traces that are later used for the detection of note boundaries. Algorithm was tested on a piece for clarinet and exhibited perfect results if no attention to processing delay was paid. However, authors have acknowledged that applicability of their method is restricted to instrumental records.

Our preprocessing approach differs from both procedures described above as it is essentially application oriented. Our algorithm has to be adapted for the analysis of low quality records of authentic entire songs. It should be able to deal with musical ornamentation and harmonic deviation phenomena as well. We raise no objective to outperform any existing algorithm of similar sort. Our goal is to transcribe an acoustic song record into several concurrent music scores for the subsequent knowledge-based processing.

3. Algorithm Description

3.1. *Basic Assumptions*

We assume that a song is perceived as a temporal succession of connected song events, where song events comprise structural sounds and musical adornments characterized by a certain duration. The most important and the most familiar song event is a structural sound. Structural sounds alone determine the melody. Our algorithm also takes into account two types of adornments: grace-notes and appoggiaturas. Together with structural sounds they constitute almost complete[4] subset of song events common to Lithuanian folk music. In reality, song events are organized in blocks. Each block is defined by a single structural sound and by the presence or absence of dependent ornamentation (Fig. 2). Unfortunately, our algorithm is unable to discriminate different types of song events yet so it assumes to face linear event sequence.

Music score represents a symbolic notation for the perceptive sequence of events. For example, a structural sound is represented by the traditional sign of a note which indicates its duration and tone. Duration and tone are the main attributes of every event considered in this work. Following the majority of related research we assume that they can be identified just using amplitude and fundamental frequency traces extracted from the physical record.

3.2. *System Architecture*

Our algorithm can be best described as a chain of transformation steps (Fig. 3). During each step some specific information processing is performed. Every transformation procedure takes a lower level song description for its input and provides the input of the next

[4]The extended folk song notation system also includes special signs for vibrato, portamento and glissando, signs for a slight rising or falling of the tone, droning sign, breathing duration markers, etc. These adornments are considered as additional attributes of some song event but not as events themselves. They are ignored in this work.

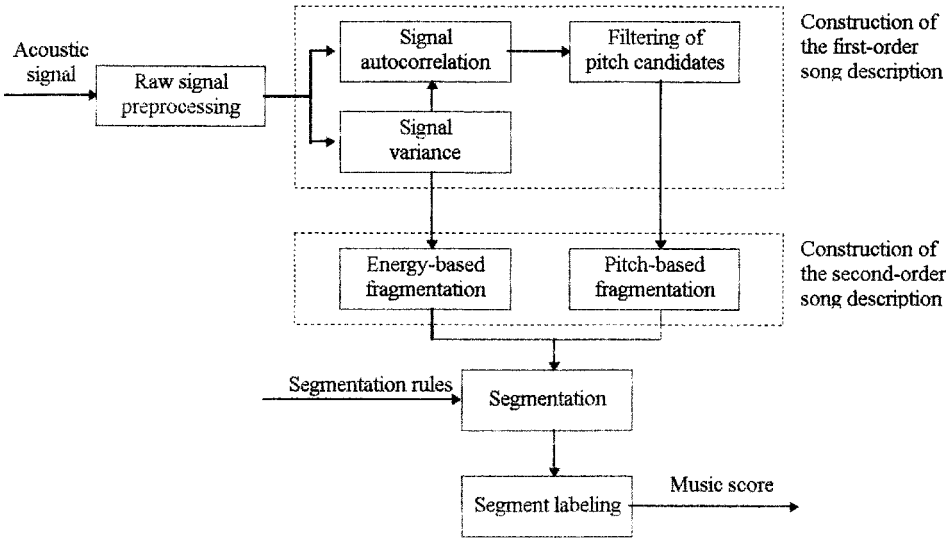

Fig. 3. General processing schema.

transformation procedure with higher level or more abstract description. An acoustic signal is considered to be a song description of the lowest level. The requested music scores represent the most abstract one. In reality, these two levels of abstraction are bound by four transformations:

- construction of the first-order song description,
- construction of the second-order song description,
- segmentation,
- segment labeling.

More detailed presentation of transformation steps follows in a few next sections.

3.3. *Construction of the First-Order Song Description*

The first processing step transforms the broadband non-stationary acoustic signal (Fig. 1) into the first-order song description. This description consists of two series of locally obtained estimates of signal energy and period (pitch).

Let $\{s_i^{raw}\}_{i=0,N-1}$ be the sequence of discrete signal samples. First of all, $\{s_i^{raw}\}_{i=0,N-1}$ is transposed giving $\{s_i\}_{i=0,N-1}$, the signal with zero-valued average.

$$s_i = s_i^{raw} - \bar{s}^{raw}, \quad \text{where} \quad \bar{s}^{raw} = \frac{1}{N}\sum_{k=0}^{N-1} s_k^{raw}, \quad i = 0, \ldots, N-1. \tag{1}$$

The signal $\{s_i\}_{i=0,N-1}$ is divided into n partially overlapping frames displaced by l samples ($n = N/l$). Frame length is fixed to $2 \cdot T_{\max}$, where $T_{\max}$ is the period corresponding to the lower bound of a possible range of vocal frequencies [5].

[5] Our experiments were carried out with frame size set to 30–40 ms and frame displacement to 10 ms.

3.3.1. *Energy-Based First-Order Description*
Each frame is assigned one energy estimate. Let $\{e_i\}_{i=0,n-1}$ denote the first-order series of energy estimates. The energy of the ith frame is defined as a local variance of signal samples:

$$e_i = \sqrt{\frac{1}{l}\sum_{k=0}^{l-1} s_{l-i+k}^2}, \quad i = 0, \ldots, n-1. \tag{2}$$

Experiences showed that frames with low e_i values as well as frames characterized by the presence of abrupt energy changes are later often assigned false pitch estimates. We simply omit such frames from further processing assigning them empty pitch labels. Formally, the ith frame is transferred to the pitch extraction routines only if two conditions are satisfied:

$$e_i > \varepsilon_1 \bar{e}, \quad \text{where} \quad \bar{e} = \frac{1}{n}\sum_{k=0}^{n-1} \quad \text{and} \quad 0 < \varepsilon_1 < 1, \tag{3}$$

$$\frac{1}{\varepsilon_2} < \frac{e_i}{e_{i+1}} < \varepsilon_2. \tag{4}$$

Here, parameters ε_1 and ε_2 define the notion of "low energy" as well as the range of permitted energy changes.

3.3.2. *Pitch-Based First-Order Description: Construction of a Candidate Set*
Each frame is assigned one pitch estimate. Let $\{p_i\}_{i=0,n-1}$ denote the first-order series of pitch estimates. In fact, $\{p_i\}_{i=0,n-1}$ is obtained in two steps. First, the sequence $\{\Psi_i\}_{i=0,n-1}$ of sets of candidates for pitch estimates is constructed. Afterwards, each set Ψ_i is filtered by leaving only one estimate per frame.

The sequence $\{\Psi_i\}_{i=0,n-1}$ is constructed by means of a temporal pitch tracking method based on signal autocorrelation[6]. It is as follows.

Let $A_i(t)$ be an autocorrelation function of the ith frame:

$$A_i(t) = \frac{\sum_{k=0}^{\max(l,t)} S_{l\cdot i+k} \cdot S_{l\cdot i+k+t}}{\sqrt{\sum_{k=0}^{\max(l,t)} S_{l\cdot i+k}^2} \cdot \sqrt{\sum_{k=0}^{\max(l,t)} S_{l\cdot i+k+t}^2}}, \tag{5}$$

$$0 < t < 2\cdot T_{\max}, \quad i = 0, \ldots, n-1.$$

For a quasi-periodic signal, $A_i(t)$ shows a pattern of evenly spaced peaks (Fig. 8a). The set $\Psi_i = \{t_k\}_i$ is defined as the set of arguments for which the autocorrelation func-

[6]The problem of pitch tracking was tackled by many researchers and a lot of methods were proposed on this topic (Hess, 1983). The whole set of methods can be divided into two broad categories, the distinction depending on whether temporal or spectral frame representation is used for signal analysis (Doval, 1994).

tion $A_i(t)$ attains its local maxima. In fact, we take into account only certain maxima. Particularly, we require the local peak $A_i(t_k)$ to be sufficiently distinct, i.e., greater than a certain specified value α_1, and close enough to the global maximum $A_i^{\max}$, the required closeness being defined by the numerical parameter α_2. The candidate peak is also required to have repeated maxima at the multiples of the period in question. Formally, every member t_k of the set Ψ_j must satisfy the following conditions:

$$A_i(t_k) > \max(\alpha_1, A_i^{\max} - \alpha_2), \tag{6}$$

where $A_i^{\max} = \max_i A_i(t)$ and $0 < \alpha_1,\ \alpha_2 < 1$,

$$\exists t_j > t_k \quad \text{such that } t_j \text{ satisfies (6) and } \left|1 - \frac{x}{[x]}\right| \text{ is a small value,} \tag{7}$$

where $x = \dfrac{t_j}{t_k}$.

It is assumed that the set Ψ_i includes the right pitch estimate. If Ψ_i is empty, the corresponding frame is assigned empty pitch label.

The majority of overall computational efforts are spent on constructing the sequence of autocorrelation functions $A_i(t)$. Even if computational load is considered to be of secondary importance we gain some speed-up by restricting the domain of argument t. It may often be restricted to the neighborhood of the precedent pitch estimate.

3.3.3. *Pitch-Based Description: Filtering of a Candidate Set*

The selection of the final pitch estimate p_i among all candidates in Ψ_i is based on contextual information, i.e., on neighboring candidate sets. We assume that the pitch curve is continuous in time. The algorithm simply removes pitch candidates which violate continuity constraints.

Let $d_f(p_a, p_b)$ denote the distance (in half-tones) between two period values p_a and p_b:

$$d_f\left(p_a, p_b\right) = \left|\log_{\sqrt[12]{2}} \frac{p_a}{p_b}\right|. \tag{8}$$

First, the set of so called "islands of confidence" is identified. "Islands of confidence" correspond to frame subseries initially having single coherent pitch estimate. Formally, the subsequence $\{\Psi_i\}_{i=beg,end}$ is called an "island of confidence" if it satisfies the following conditions:

$$\begin{aligned} &\Psi_i = \{t_k\}_i = p_i \quad \text{for all } i \in [beg, end], \\ &L > \lambda_1 \quad \text{and} \quad \frac{1}{L}\sum_{i=beg}^{end} A_i(p_i) > \alpha_2, \quad \text{where } L = end - beg + 1, \\ &d_f(p_i, p_{i+1}) < \rho_1 \quad \text{for all } i \in [beg, end). \end{aligned} \tag{9}$$

Here, parameters λ_1, α_2 and ρ_1 define minimum requirements for duration, average correlation and smoothness of an "island of confidence", respectively.

Secondly, every subsequence of candidate sets situated between two "islands of confidence" is processed. Let $\{\Psi_i\}_{i=beg,end}$ be such subsequence. The algorithm propagates continuity constraints from the left and from the right and builds two smooth series $\{p_i^L\}_{i=beg,end}$ and $\{p_i^R\}_{i=beg,end}$:

$$\begin{aligned} p_i^L &= \left\{t_k | t_k \in \Psi_i \text{ and } d_f(t_k, p_{i-1}^L) < \pi_3\right\} \\ p_i^R &= \left\{t_k | t_k \in \Psi_i \text{ and } d_f(t_k, p_{i+1}^R) < \pi_3\right\} \end{aligned}, \quad i = beg, \ldots, end. \tag{10}$$

The final series of pitch estimates $\{p_i\}_{i=beg,end}$ is taken according to:

$$p_i = \begin{cases} p_i^L, & \text{if } p_i^L = p_i^R, \\ \text{empty}, & \text{otherwise}, \end{cases} \qquad i = beg, \ldots, end. \tag{11}$$

3.4. *Construction of the Second-Order Song Description*

The second-order song description represents a more general picture of volume and pitch transitions. It consists of two series $\{E_i\}$ and $\{P_i\}$ of records, where each record describes a sound fragment lasting for multiple frames. Sequences $\{E_i\}$ and $\{P_i\}$ are constructed by independently fragmenting sequences $\{e_i\}$ and $\{p_i\}$ and by assigning to each fragment a set of specific features. The fragmentation is constrained by allowing the subparts only of some predefined types.

3.4.1. *Energy-Based Second-Order Description*

Preliminary analysis of the sequence $\{e_i\}$ suggested us to distinguish three types of subsequences called the fragments of rising, falling and low volume.[7] Our fragmentation procedure used the following definitions for the recognition of volume-based fragment types.

The subsequence $\{e_i\}_{i=beg,end}$ is called the fragment of low volume if:

- $e_i < \varepsilon_1 \overline{e}$ for all $i \in [beg, end]$ (see (3)). (12)

Otherwise it is said to be of rising (falling) volume if:

- e_{beg} is a minimum (maximum) value and e_{end} is a maximum (minimum) value of $\{e_i\}$;
- there is an index $k < beg$ such that

$$\frac{e_{beg}}{e_k} < \frac{1}{\mu_1} \ (> \mu_1) \text{ and } e_{beg} - e_i < 0 \ (> 0) \tag{13}$$

for all $i \in (k, beg)$.

[7] The fragments of rising and falling volume carry the vibration of vocal chords. The fragments of low volume correspond to pauses, breathing, unvoiced explosive and fricative consonants (k, t, s, sh, tch).

- there is an index $l > end$ such that $\frac{e_l}{e_{end}} < \frac{1}{\mu_1}\ (> \mu_1)$ and $e_i - e_{end} < 0\ (> 0)$ for all $i \in (end, l)$.

Here, parameter μ_1 has plays the role of threshold of sensitivity. Energy changes less than μ_1 are disregarded.

Every energy-based fragment $\{e_i\}_{i=beg,end}$ is described by the set of *features* $E_j = (E^{type}, E^{beg}, E^{end}, E^{min}, E^{max}, E^{dif}, E^{rat})_j$, where:

$$E^{type} \in \{rise, fall, low\},$$

$$E^{beg} = beg, \quad E^{end} = end, \quad E^{min} = \min(e_{beg}, e_{end}),$$

$$E^{max} = \max(e_{beg}, e_{end}), \quad E^{dif} = E^{max} - E^{min}, \quad E^{rat} = \frac{E^{max}}{E^{min}}.$$

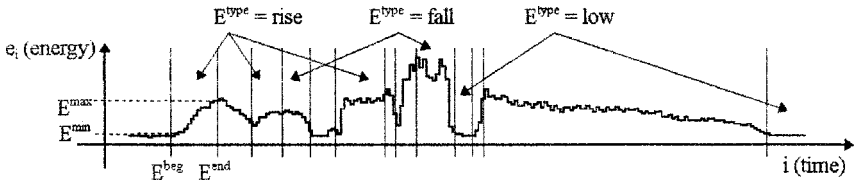

Fig. 4. Illustration of the second-order energy-based features.

3.4.2. *Pitch-Based Description*

Preliminary analysis of the sequence $\{p_i\}$ showed the necessity of distinguishing at least two types of subsequences called the fragments of steady and transitional pitch. The procedure of pitch-based fragmentation used the following definitions for the recognition of fragment types:

Let $d_f^{ext}(\{p_i\}; [a, b])$ denote the frequency extent of the subsequence $\{p_i\}_{i=a,b}$:

$$d_f^{ext}(\{p_i\}; [a, b]) = d_f\left(\max_{i\in[a,b]} p_i, \min_{i\in[a,b]} p_i\right), \quad \text{where } d_f \text{ is given by (8).} \tag{14}$$

The subsequence $\{p_i\}_{i=beg,end}$ is recognized as a steady pitch fragment if:

- $end - beg > \mu_2$ (15)
- $d_f^{ext}(\{p_i\};\ [a, b]) < \mu_3$ and $d_f^{ext}(\{p_i\}; [a-1, b]) > \mu_3$ and $d_f^{ext}(\{p_i\}; [a, b+1]) > \mu_3$.

Here, parameters μ_2 and μ_3 specify minimum required length of a steady fragment and the range of permitted pitch variations within it respectively.

The subsequence $\{p_i\}_{i=beg,end}$ is called the fragment of transitional pitch if it is not of steady pitch.

Every pitch-based fragment $\{p_i\}_{i=beg,end}$ is described by the set of features $P_j = (P^{type}, P^{beg}, P^{end}, P^{min}, P^{max}, P^{avg}, P^{var}, P^{rat})_j$, where:

$P^{type} \in \{stead, trans\}$,

$P^{beg} = beg,\ P^{end} = end,\ P^{min} = \min\limits_{i\in[beg,end]} p_i,\ P^{max} = \max\limits_{i\in[beg,end]} p_i,$

$P^{rat} = \dfrac{P^{max}}{P^{min}},\ P^{avg} = \dfrac{1}{L}\sum\limits_{i=beg}^{end} p_i,\ P^{var} = \sqrt{\dfrac{1}{L}\sum\limits_{i=beg}^{end}(p_i - P^{avg})^2}$ and

$L = end - beg + 1.$

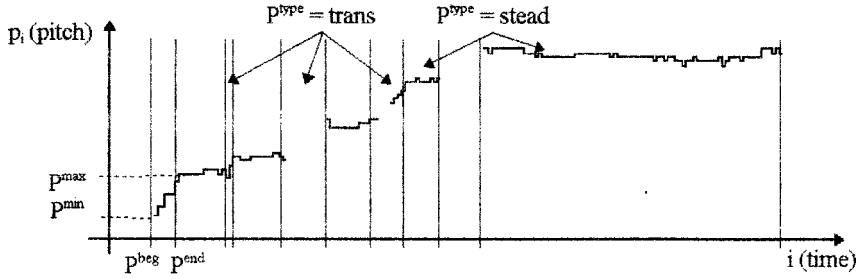

Fig. 5. Illustration of pitch-based second-order features.

3.5. *Segmentation*

The localization and separation of song events within an acoustic signal is realized by the segmentation procedure. The segmentation is performed on a basis of a second order description. It implements an ordered set of hand-defined rules. The procedure starts with an empty list of segmentation markers M and fills this list as it passes over the signal. Boundaries of song events are imprecise, so segmentation markers often correspond to time intervals rather than to exact time instants.

In general, segmentation markers are laid over all fragments of low energy. They are also laid over the fragments of transitional pitch if the numeric estimate of this transition is greater than a certain threshold ρ_{max}. Similarly, segmentation markers are inserted at the junction of fragments of falling and rising energy if only associated energy changes are greater than a threshold η_{max}. Algorithm may reduce this value to η_{min} if energy changes are accompanied by less distinct pitch transitions ρ_{min}. Finally, all segments are required to be of certain minimal length λ.

Schematized version of segmentation algorithm is given below:

$M = \{\oslash\}$

for all E_i

  if $E_i^{type} =$ low

    if $\exists k > i$ such that $E_k^{type} =$ low **and** $E_k^{beg} - E_i^{end} < \lambda$

      then $M = M \cup [E_i^{beg}, E_k^{end}]$.

      else $M = M \cup [E_i^{beg}, E_i^{end}]$.

else if $E_i^{type} =$ fall **and** $E_{i+1}^{type} =$ rise **and**

$\max(E_i^{rat}, E_{i+1}^{rat}) > \eta_{max}$ **or**

$\min(E_i^{rat}, E_{i+1}^{rat}) > \eta_{min}$ **and** $\exists k$ such that $P_k^{beg} < E_i^{end}$ **and** $P_k^{end} > E_{i+1}^{beg}$ **and** $P_k^{rat} > \rho_{min}$

then $M = M \cup [E_i^{end}, E_{i+1}^{beg}]$.

for all P_i

if $P_{i-1}^{type} = P_{i+1}^{type} =$ stead **and** $P_i^{type} =$ trans **and**

$\min(P_{i-1}^{end} - P_{i-1}^{beg}, P_{i+1}^{end} - P_{i+1}^{beg}) > \lambda$ **and** $d_f(P_{i-1}^{avg}, P_{i+1}^{avg}) > \rho_{max}$

then $M = M \cup [E_i^{beg}, E_i^{end}]$.

We define the sequence of segments $\{S_i\}$ as the sequence of song sections not covered by the set M. Every segment S_i is assigned two characteristics: duration S^{length} and average pitch S^{pitch}, i.e., $S_i = (S^{length}, S^{pitch})_i$.

3.6. *Segment Labeling*

Segment labeling is the last processing step which consists in assigning duration and tone labels to previously identified segments. This step is realized by temporal and tonal labeling procedures.

3.6.1. *Temporal Labeling*

The temporal labeling procedure has three tasks to be realized:

- resolve segment boundary imprecision, i.e., substitute all segmentation markers-intervals with segmentation markers-instants,
- define melody tempo T_0,
- assign segments duration labels.

Let $\{T_0 \cdot k_i\}$, $k_i \in K$, $K \subset N$ denote the set of possible durations of music notes.[8] Let also define $err_t(S^{length}; \{T_0 \cdot k_i\})$ as an approximation error of replacement of the actual segment duration S^{length} with the closest value from the set $\{T_0 \cdot k_i\}$:

$$err_t\left(S^{length}; \{T_0 \cdot k_i\}\right) = \min_{k_i \in K} \left| 1 - \frac{S^{length}}{T_0 k_i} \right|. \tag{16}$$

The substitution of segmentation markers-intervals with segmentation markers-instants is realized by an iterative minimization procedure. The typical case of marker substitution is illustrated by Fig. 6.

Let S_j andS_{j+1} be two segments separated by the segmentation marker-interval having duration m^{length}. Let $S^{*length}_{j}$ and $S^{*length}_{j+1}$ denote the respective durations of these segments after segmentation marker has been removed. The values of $S^{*length}_{j}$ and $S^{*length}_{j+1}$ are given by the following procedure:

[8] We used the set $K = \{1, 2, 3, 4, 6, 8, 12, 16\}$.

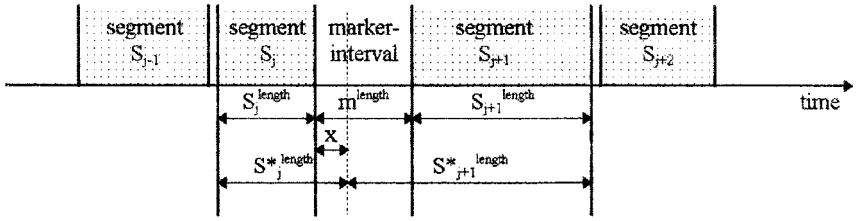

Fig. 6. Illustration of marker substitution.

$$x = \arg \min_{0 \leqslant y \leqslant m^{length}} \left(\left[err_t \left(S_j^{length} + y;\ \{T_0 k_i\} \right) \right]^2 + \left[err_t \left(S_{j+1}^{length} + m^{length} - y;\ \{T_0 K_i\} \right) \right]^2 \right); \tag{17}$$

$$S^{*length}_{\ j} = S_j^{length} + x; \quad S^{*length}_{\ j+1} = S_{j+1}^{length} + m^{length} - x.$$

The tempo T_0 is held to be locally constant, i.e., constant for a song section between two sufficiently long pauses. Local tempo is obtained by minimizing the squared approximation error over segments S_j belonging to this section.

$$T_0 = \arg \min_T \sum_j \left[err_t \left(S_j^{length}; \{TK_i\} \right) \right]^2. \tag{18}$$

Finally, one or two duration labels from the set *K* are assigned to each segment S_j. Two ordered label preferences are maintained if the approximation error of the best choice exceeds a certain desired value β_t, i.e., if:

$$err_t \left(S_j^{length}; \{T_0 k_i\} \right) > \beta_t. \tag{19}$$

3.6.2. *Tonal Labeling*

The tonal labeling procedure has two tasks to be realized:

- define tonal reference point F_0,
- assign segments tone height labels.

Let $\{F_0 \cdot c^i\}$, $c = 2^{1/12}$, $i \in N$ denote the set of possible tone values. Let also define $err_f(S^{pitch}; \{F_0 \cdot c^i\})$ as an approximation error of replacement of the actual segment pitch S^{pitch} with the closest value from the set $\{F_0 \cdot c^i\}$:

$$err_f \left(S^{pitch}; \{F_0 c^i\} \right) = \min_{i \in N} d_f \left(S^{pitch}; F_0 c^i \right), \quad \text{where } d_f \text{ is given by (8).} \tag{20}$$

The tonal reference point F_0 is considered to be the same for the whole song. It is obtained by minimizing the squared approximation error over all song segments S_j:

$$F_0 = \arg\min_F \sum_j \left[err_f \left(S_j^{pitch}, \{Fc^i\} \right) \right]^2 . \tag{21}$$

Either one or two tone labels $i \in N$ might be assigned to every segment S_j. Two ordered label preferences are maintained if the approximation error of the best choice exceeds a certain minimum value β_f, i.e., if:

$$err_f \left(S_j^{pitch}; \{F_0 c^i\} \right) > \beta_f . \tag{22}$$

4. Algorithm Analysis

Our analysis had two primary goals. The first goal was a natural desire to assess the true performance of the procedure constructed by us. Secondly, we wanted to discover specific problems encountered by each processing step, to reveal existing limitations and to understand their causes. The results of such investigation are discussed in this part.

4.1. *Tested Song Collection*

The algorithm was tested on seven Lithuanian folk songs of average complexity.[9] Our records were selected to represent a variety of human performers as well as differences in tempo and recording quality. Table 1[10] briefly summarizes some of their main characteristics.

All songs were sampled at 11 kHz using 16 bit quantification. Records m1 and m2 were produced in an ordinary room using low-quality microphone. All signals but m1 and cd13 were accidentally subjected to more or less perceptible 50 Hz electric current interference.

4.2. *Performance Criteria*

Our application used the same set of parameter settings[11] for all seven songs. As a result it produced music scores which were given in form of diagrams of type illustrated by

[9] Song complexity was understood as relative complexity of its ornamentation patterns.

[10] Name – m1: *Ant kalno mūrai*
m2:*Du gaideliai.*
t15:*Lykie lietuli*, Lithuanian folk music I, Dzûkø dainos, Melodija, 1974.
t20:*Dveji traji meteliai*, the same.
t24:*Bėginėjo povelė po dvarą*, the same.
t31:*Broli, brolali sakalėli*, the same.
cd13:*Rūta žalioj, jau vakaras vakarėlis*, Lithuanian folk music, 33 Records, 1995.

[11] $\varepsilon = 0.2$, $\alpha_1 = 0.65$, α_2=0.30, α_3=0.0, λ_1=50 ms, λ_2=80 ms, λ_3=100 ms, η_1=2.0, η_2=2.0, η_{max}=3.0, η_{min}=2.0, ρ_1=1.0 ht, ρ_2=0.5 ht, ρ_{max}=2.0 ht, ρ_{min}=0.8 ht, β_t=0.05, β_f=0.25 ht.

Table 1
Tested song collection

Name	Recording source	Duration, s	Performer's age and sex	Frequency range, Hz
m1	microphone	24	24 male	85–180
m2	microphone	12	22 female	170–450
t15	tape	119	60 female	190–440
t20	tape	80	55 female	150–370
t24	tape	131	75 male	100–330
t31	tape	120	63 female	170–450
cd13	CD	123	56 female	260–410

Fig. 7. Such diagrams were convenient for separate analysis of specific transformation steps.

Particularly, we investigated the procedures of pitch trace extraction, segmentation and segment labeling. Accuracy estimation of lower level signal processing was somewhat arbitrary due to the lack of clear performance criteria. Segmentation and segment labeling accuracy was evaluated by comparing machine produced music scores with corresponding transcriptions provided by domain experts (Fig. 7). For all songs, except m1 and m2, we had transcriptions of original records. Songs m1 and m2 were contrasted with transcriptions of their different acoustic versions.

Detailed analysis of individual transformation steps follows in the coming sections.

4.3. *Pitch Trace Extraction*

Pitch trace extraction is the first non trivial task of low level signal processing. Although it is difficult to evaluate the processing accuracy of this step separately from later processing some partial observations can be formulated.

Impact of empty labels. Frames assigned empty pitch label seems to have no susceptible influence on results of subsequent processing. The relative number of empty labels increases with the decrease of sound quality. Statistical summary concerning such frames is shown in Table 2.

Precision of pitch assessment. The pitch is measured in discrete units since it corresponds to the integer number of waveform samples covered by one period. Still, the difference of one sample might be significant enough in terms of musical tones. Certain resolution could be gained by sampling signal at higher frequency. The better and more widespread solution to this problem is offered by the interpolation techniques being applied to the autocorrelation function $A_i(t)$.

Range of vocal frequencies. Any pitch tracking procedure is primarily concerned with frequency domain carrying the vibration of vocal chords. This domain lies somewhere between 50–1000 Hz (Lindsey and Norman, 1974). The raw signal embeds speech produced higher order spectral components as well. Unfortunately, these components modulate the autocorrelation function blurring its maxima (Fig. 8). The present algorithm

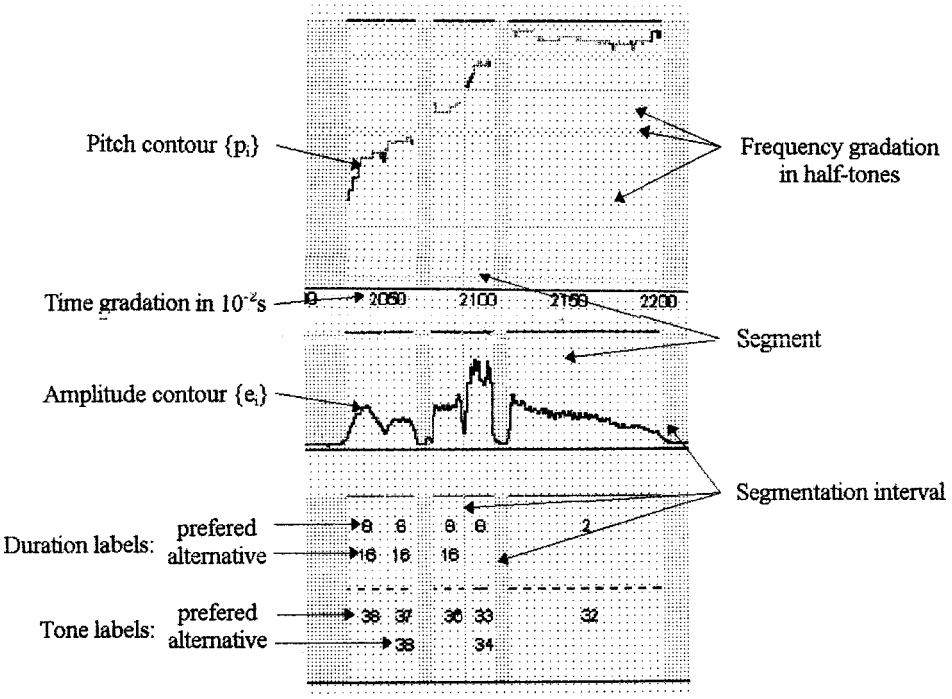

Fig. 7. Example of application produced diagram.

Table 2

Frames assigned empty pitch label. Letters L (low energy), A (abrupt energy alteration), P (lack of prominent pitch candidate), M (peaks at multiples absent) designate empty labels assigned due to conditions (3), (4), (6) and (7), respectively. Symbol R stands for labels removed by postprocessing

Name	L	A	P	M	R	Empty labels	Total Frames
m1	392 16,1%	62 2,6%	15 0,6%	52 2,1%	4 0,2%	525 21,6%	2430
m2	211 18,5%	14 1,2%	6 0,5%	27 2,4%	1 0,1%	259 22,7%	1139
k15	1821 15,3%	22 0,2%	192 1,6%	227 1,9%	89 0,7%	2351 19,7%	11927
k20	1506 16,8%	14 0,2%	174 1,9%	308 3,4%	336 3,8%	2338 26,1%	8955
k24	2491 19,1%	57 0,4%	235 1,8%	372 2,8%	139 1,1%	3294 25,2%	13056
k31	2565 21,4%	54 0,5%	657 5,5%	777 6,5%	238 2,0%	4291 35,8%	11972
cd13	1998 16,2%	68 0,6%	115 0,9%	196 1,6%	88 0,7%	2465 20,0%	12328

attempts to compensate high frequency related distortions by additional computational efforts. Nevertheless, low-pass signal filtering might represent a valuable solution.

Pitch candidate selection. Postprocessing algorithm of the type used in this work has a weakness of “locking ” on falsely identified “islands of confidence ”. This results in undesired pitch trace shift by octave for more or less extended part of a song. Fortunately,

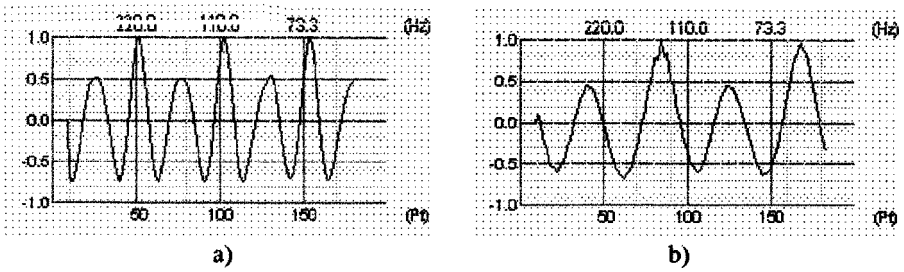

Fig. 8. Regular (a) and modulated (b) autocorrelation functions.

scientific literature proposes some useful reflections on this point (Secrest and Doddington, 1982). In practice, misidentification errors appear to be not numerous. The overall number of such cases is given in Table 3.

Table 3
Errors of pitch misidentification

Name	Errors	Total labeled (non-empty)
m1	0 0,0%	1905
m2	0 0,0%	880
k15	0 0,0%	9576
k20	108 1,6%	6617
k24	21 0,2%	9762
k31	10 0,1%	7681
cd13	0 0,0%	9863

Erroneous labels were hand corrected before submitting pitch contour to posterior processing.

4.4. *Segmentation*

We registered two possible types of segmentation errors: segmentation marker omission and needless marker generation.[12] The processing accuracy related to these two types of errors is given by the following pair of tables.

Table 4 shows the relative weight of detected and missed segmentation markers over the set of all required markers, i.e., markers placed by an expert. For better understanding of our results required segmentation markers are grouped into structural and ornamentation categories. Segmentation markers of structural type separate blocks of song events. Markers internal to such structural blocks and usually inserted between a structural sound

[12]Significant positional shift of a segmentation marker was counted twice: once as a missing and once as a needless marker.

Table 4
Accuracy of marker detection

Name	Segmentation marker type	Detected markers	Missed markers	Required markers
m1	ornamentation	5 83,3%	1 16,7%	6
	structural	56 100,0%	0 0,0%	56
	Total	61 98,4%	1 1,6%	62
m2	ornamentation	0	0	0
	structural	37 90,2%	4 9,8%	41
	Total	37 90,2%	4 9,8%	41
t15	ornamentation	38 74,5%	13 25,5%	51
	structural	126 98,4%	2 1,6%	128
	Total	164 91,6%	15 8,4%	179
t20	ornamentation	13 68,4%	6 31,6%	19
	structural	135 93,8%	9 6,3%	144
	Total	148 90,8%	15 9,2%	163
t24	ornamentation	26 51,0%	25 49,0%	51
	structural	180 98,4%	3 1,6%	183
	Total	206 88,0%	28 12,0%	234
t31	ornamentation	20 31,7%	43 68,3%	63
	structural	180 95,2%	9 4,8%	189
	Total	200 79,4%	52 20,6%	252
cd13	ornamentation	26 57,8%	19 42,2%	45
	structural	106 100,0%	0 0,0%	106
	Total	132 87,4%	19 12,6%	151

and its adornments are named ornamentation markers. This distinction seems to be reasonable. The failure to detect a structural segmentation marker is always an important error whereas missing ornamentation marker might be considered as a less serious mistake. This is due to the fact that song parts containing ornamentation patterns are exactly the places where transcribing musicians often take very subjective decisions. Notwithstanding, it appears that musical adornments are hard to find. They are shorter and less distinct than structural sounds.

Table 5 summarizes marker generation accuracy by indicating the part of needless segmentation markers in the whole set of generated markers.

Detailed analysis of segmentation errors conducted us to following remarks:

Definition of musical sensitivity. It appears that transcribing experts employ different notions of temporal, tonal and volume sensitivity. Moreover, sensitivity related judgments of the same person may change within a single song. The sensitivity of our algorithm is implicitly coded under the uniform set of numeric parameters which is not adjusted to a

Table 5
Accuracy of marker generation

Song name	Necessary markers	Needless markers	Generated markers
m1	61 100,0%	0 0,0%	61
m2	37 100,0%	0 0,0%	37
t15	164 87,7%	23 12,3%	187
t20	148 93,7%	10 6,3%	158
t24	206 86,2%	33 13,8%	239
t31	200 93,9%	13 6,1%	213
cd13	132 85,7%	22 14,3%	154

particular song. This lack of correspondence resulted in a great number of segmentation errors of both types.

Bias towards vowel-consonant separation. Our algorithm is predisposed to place segmentation markers at the junction of vowel and consonant sounds. On the one hand, the required segmentation marker separating two vowels or fragmenting the same vowel sound is sometimes missed because of the second order description showing little variability at the environment of this segmentation point. On the other hand, needless segmentation marker is generated between a vowel and a consonant sounds belonging to the same song event when this coincides with abrupt and undesired changes in the second-order description.

Irregular singing patterns. Both types of errors might occur when the algorithm passes over so called irregular singing patterns. An example of such pattern is an unexpected and substantial fall of singing volume followed by its rise. In fact, there exists no clear agreement permitting to distinguish melody adornments from transitional effects, accidental deviations or simply uninteresting particularities of individual song interpretation.

4.5. *Segment Labeling*

Actually, our algorithm is not capable of discriminating between different types of song events. All events are labeled according to the same procedure described in Section 3.4. However, transcribing professionals use different approach concerning temporal labeling of structural sounds and labeling of such ornamentation events as grace-notes and appoggiaturas. Ornamentation events are assigned labels describing their individual durations. Structural sounds are labeled as if their duration was extended by associated adornments, i.e., structural sound receives the label corresponding to the whole block of song events. Since temporal labeling accuracy is estimated by comparing man and machine produced scores, only songs m1 and m2, which have minimum of adornments, are used for this estimation.

The occurrence of severe temporal labeling mistakes is explained by widespread singers' tendency to truncate sounds that precede inhalation. Less important inaccura-

Table 6
Accuracy of temporal labeling

Name	Prefered label is correct	Alternative label is correct	Correct label absent	Total
m1	58 95,1%	2 3,3%	1 1,6%	61
m2	32 88,9%	2 5,6%	2 5,6%	36

cies are linked to slight shift of a segmentation marker at the neighborhood of a voiced consonant.

Melody tempo slightly oscillates along the song. The real extent of tempo variation is indicated by the table below.

Table 7
Oscillation of melody tempo

Name	Local parts	Length of quarter-note, ms
m1	5	46, 44, 45, 48, 48
m2	2	53, 55

Tonal labeling results seem to be affected by singing and recording quality. The following table summarizes tonal labeling accuracy of all seven songs. These estimations are based on segments of structural sounds correctly identified during the segmentation step.

Table 8
Accuracy of tonal labeling

Name	Prefered label is correct	Alternative label is correct	Correct label absent	Total
m1	49 81,7%	6 10,0%	5 8,3%	60
m2	21 58,3%	8 22,2%	7 19,4%	36
k15	111 78,7%	26 18,4%	4 2,8%	141
k20	95 68,8%	31 22,5%	12 8,7%	138
k24	167 90,8%	14 7,6%	3 1,6%	184
k31	139 74,3%	31 16,6%	17 9,1%	187
cd13	117 95,1%	6 4,9%	0 0,0%	123

Tonal labeling errors occurred due to the following phenomena:

Unsteady pitch curve. Fragments of pitch curve corresponding to song events are far from being perfectly horizontal. In contrary, they can take the most unexpected forms

sometimes extending in frequency for a few half-tones. This is speech-related phenomena which seem to be somewhat regular. Our algorithm attempts to extract the steady part of pitch curve. However, this is not always the best solution. We don't know any simple and universal recipe yet answering the question of what is the perceptive tone value in every such case.

Vowel dependent pitch shift. Two adjacent song events based on different vowel sounds and perceived as having the same tone[13] are sometimes depicted by clearly shifted pitch curves (Fig. 9). The detailed examination of this phenomenon goes beyond the scope of this paper. However, we should mention that a vowel dependent pitch shift shows some regularity.

Drift of tonal reference point. Tonal reference point appears not to be globally constant but drifts along the song. For certain songs (m2, k20, k31) the range of this drift exceeds one half-tone (Fig. 10).

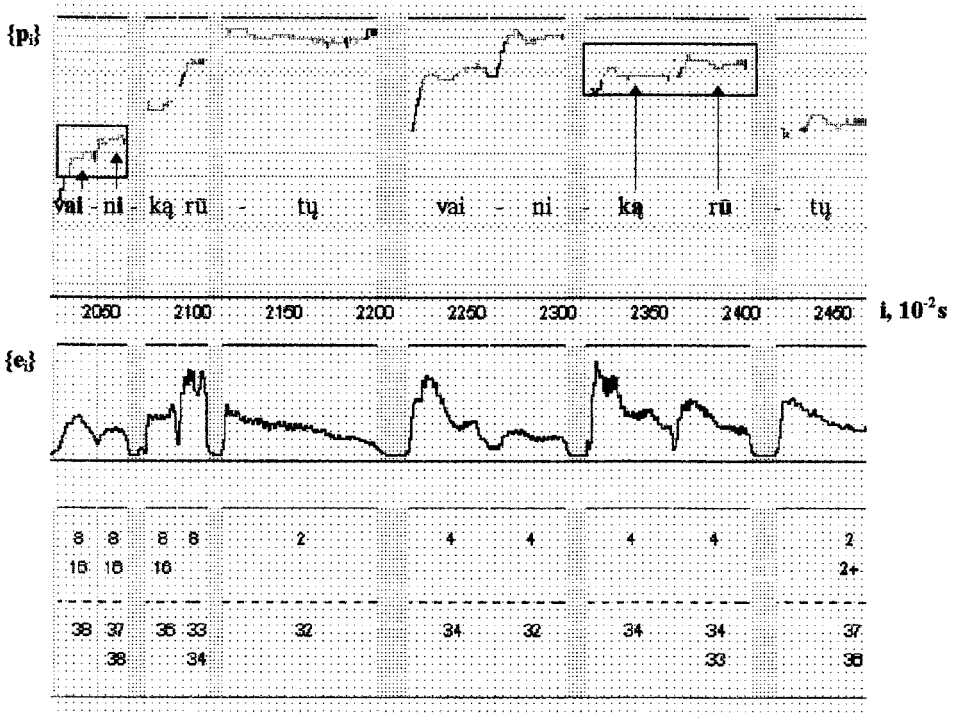

Fig. 9. Vowel-dependent pitch shift.

[13]Certainly, the perceptive difference exists but it requires special attention in order to be noticed.

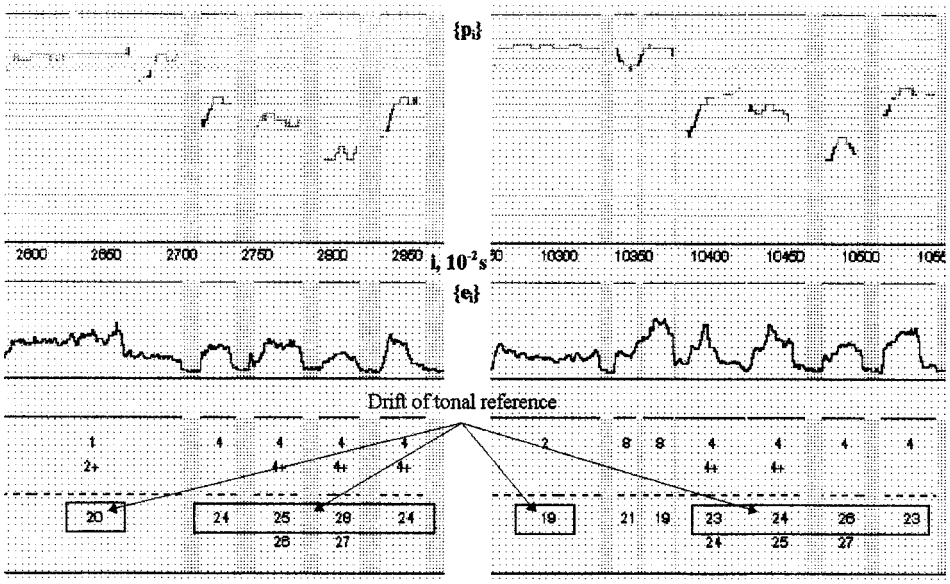

Fig. 10. Drift of tonal reference. Corresponding fragments of the third (left) and of the tenth (right) stanzas of the song k31.

5. Discussion

We consider the extraction of existing first-order features to be satisfactory. Nevertheless, our pitch tracking procedure can be improved in several ways as it was briefly discussed in Section 4.3. Furthermore, we think it makes sense to investigate alternative processing approaches: spectral, cepstral analysis or some other more sophisticated techniques. Our pitch tracking method appeared to be well suited to monodic songs. Possibility whether this method is more general remains questionable as it was tested neither on polyphonic melodies nor on instrumental parts. Similar methods are normally inappropriate for straight analysis of signals that mix multiple acoustic sources. The possibility of handling such records in future is not entirely excluded, however. The research on acoustic source separation (Bell and Sejnowski, 1995; Nakatani *et al.*, 1995) is of great interest from this point of view.

Energy and pitch based features appeared to be very useful but not completely sufficient for the purposes of this work. In fact, our algorithm without any phonetic song description available can hardly take into account such phenomena as presence of irregular singing patterns or vowel dependent pitch shift. We imagine necessary phonetic description as a succession of recognized phonetically meaningful signal segments. It is desirable but not obligatory to aim at obtaining precise pronunciation layout. The minimum of details required can be achieved if the subsystem of phonetic feature detection is able to separate vowels from consonants and to distinguish vowels among themselves. In the ideal case musical analysis can be additionally guided by higher-order linguistic (index of syllabic entries) knowledge.

We hope that many difficulties that our algorithm has encountered can be left to machine learning techniques. In fact, we do not pretend to be using complete and precise segmentation rule set. Human experts are incapable of establishing such rules either. This is true not only because of sensitivity problem but also as the musicians are used to explain their decisions in perceptual language. Therefore, automatic induction of segmentation rules based on real-world examples seems to be one of the best solutions in this case. The same inductive approach can be used for two other problems: recognition of tone label in case of a variable segment pitch and recognition of song event. Of course, application of machine learning techniques may lead to success only if empirical examples are described in an appropriate language which has a sufficient power of expression.

6. Conclusions

The algorithm described in this paper was designed to perform preliminary processing of acoustic musical records. It was tested on a number of real Lithuanian folk songs. Processing of relatively simple songs resulted in a music score that corresponded to our expectations. It was precise enough in order to render further application of theory of music possible. Nevertheless, difficulties related to the phenomenon of musical ornamentation were underestimated. They require further research to be pursued.

We believe that the present work was significant not as much for the accurate performance of our algorithm but for the revelation of many specific problems of folk singing. Furthermore, detailed comparison of music scores produced by machine and humans let us to state numerous suggestions of how to deal with subsisting difficulties. Certain minor problems require only minute algorithmic perfection. Others are essential and entail the choice of completely different approach. In particular, we determined that the spoken component of a song may not be simply ignored but can help in resolving uncertainties related to noisy data. We also concluded that mathematical methods we based this work on are limited. They need to be complemented with inductive knowledge acquisition procedures.

Acknowledgments

I would like to thank Prof. Laimutis Telksnys, Head of Recognition Processes Department at the Institute of Mathematics and Informatics, Prof. Jean-Gabriel Ganascia, Head of Knowledge Acquisition and Machine Learning (ACASA) Research Group at the University Paris VI in France and Prof. Romualdas Apanavičius, Director of the Institute of Ethnomusic, for their advice and useful comments about this work.

References

Bell, A.J., and T.J. Sejnowski (1995). Blind separation and blind deconvolution: an information-theoretic approach, *ICASSP*, 3415–3418.

Doval, B. (1994). *Estimation de la Fréquence Fondamentale des Signaux Sonores*, Thèse de doctorat de l'Université Paris VI, LAFORIA.

Hess, W. (1983). *Pitch Determination of Speech Signals*, Springer-Verlag.

Pachet, F., and G. Assayag (1995). *Bulletin de l'AFIA* (Association Française pour l'Intelligence Artificielle) **23**.

Quiros, F.J.C., and P. F.-C. Enriquez (1994). *Real-Time Loose-Harmonic Matching Fundamental Frequency Estimation for Musical Signals*, ICASSP, pp. 221–224.

Secrest, B.G., and G.R. Doddington (1982). *Postprocessing techniques for voice pitch trackers*, ICASSP, pp. 172–175.

Nakatani, T., T. Kawabata and H.G. Okuno (1995). *A Computational Model of Sound Stream Segregation with Multi-Agent Paradigm*, ICASSP, pp. 2671–2674.

Lindsey, P., and D. Norman (1974). *Pererabotka informacii u cheloveka*, Moscow (in Russian).

G. Raškinis was born in 1972. He obtained M.Sc degree in artificial intelligence and pattern recognition in 1995 by the University of Pierre et Marie Curie in Paris. He is now completing his Ph.D research at the Vytautas Magnus University in Kaunas on musical pattern recognition.

Pradinis algoritmas liaudies dainų įrašų užrašymui natomis

Gailius RAŠKINIS

Straipsnyje aprašomas pirminis algoritmas, kurio tikslas užrašyti muzikinį įrašą natomis. Aprašyta procedūra pagrįsta tik dviejų signalo charakteristikų – pagrindinio dažnio ir energijos – analize. Algoritmas išbandytas su tikromis vidutinio sudėtingumo lietuvių liaudies dainomis. Nors gauti rezultatai teikia nemažų vilčių, detali jų analizė atskleidžia kai kuriuos pasirinktos metodikos trūkumus, o taip pat ir nemažai su taikymo sritimi susijusių problemų. Pavyzdžiui, tampa aišku, kad muzikinės ir fonetinės analizės neįmanoma visiškai atskirti. Tyrimą numatoma tęsti muzikos teorijos žinių integravimo į šį algoritmą kryptimi.