Available online at www.sciencedirect.com
SCIENCE DIRECT®
Pattern Recognition 38 (2005) 1135–1138
PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY
www.elsevier.com/locate/patcog

Rapid and brief communication

Automatic detection of vibrato in monophonic music

Hee-Suk Pang*, Doe-Hyun Yoon

DVC Gr., DM Research Lab., LG Electronics, Yatap-Dong 360-5, Bundang-Gu, Sungnam-Si, Kyunggi-Do, 463-828, Republic of Korea

Received 1 November 2004; received in revised form 22 November 2004

Abstract

Vibrato is one of the most common techniques in musical performances for enriching the sound. We propose a novel method that automatically detects vibrato in monophonic music. It is based on modeling the probability of vibrato existence using three vibrato parameters, i.e., the vibrato rate, extent, and intonation. Experiments using various musical instrument tones and the solo performance show the effectiveness of the method. The proposed method can be applied to music recognition such as the wav-to-midi conversion.

Keywords: Vibrato; Monophonic music; Probability of vibrato existence

1. Introduction

For music recognition, musical tones are first transformed into a time–frequency representation and then musical events are detected. For example, audio signals are converted to midi data based on the detection of starts, transitions, and ends of notes by musical instruments or singers [1]. Though these events are sufficient for simple musical note transcription, more complex analysis methods are required for further understandings of a variety of musical performances.

Vibrato is one of the most common techniques to enrich the sound in musical performances. Researches have been concentrated on analysis methods of the vibrato tones assuming that the vibrato is successfully detected [2,3]. The detection is usually accomplished by human inspectors and automatic detection of vibrato has not been studied much in the literature. In monophonic music, most tones can be categorized into three parts, i.e., normal part, vibrato part, and note transition part. In this paper, we propose a method that automatically detects the vibrato part in monophonic music.

2. The method

For analysis of a monophonic signal, we first perform a short-time Fourier transform (STFT) and find the fundamental frequency track. Then we detect the voiced region based on the energy and the fundamental frequency range in the track. We further apply frequency enhancement techniques such as the phase difference approximation [4] to estimate the fundamental frequency accurately. The note transition is detected by observing the abrupt change of the fundamental frequency. The threshold for the change is basically 12% or 200 cents, which corresponds to the interval of two semitones.

The fundamental frequency track of a vibrato tone is often not a form of a sinusoid in real cases but, at least, the sinusoidal component is the strongest in the track. The sinusoidal component can be represented as

$$f(n) = a(n) + b(n)\sin\left(2\pi\frac{f_v(n)}{f_{frame}}n + \theta\right), \quad (1)$$

* Corresponding author. DVC Gr., DM Research Laboratory, LG Electronics, 16 Woomyeon-Dong, Seocho-Gu, Seoul 137-724, Republic of Korea. Tel.: +82 31 789 4212; fax: +82 31 789 4207.
E-mail addresses: hspang@lge.com (H.-S. Pang), dhyoon@lge.com (D.-H. Yoon).

0031-3203/$30.00
doi:10.1016/j.patcog.2005.01.001

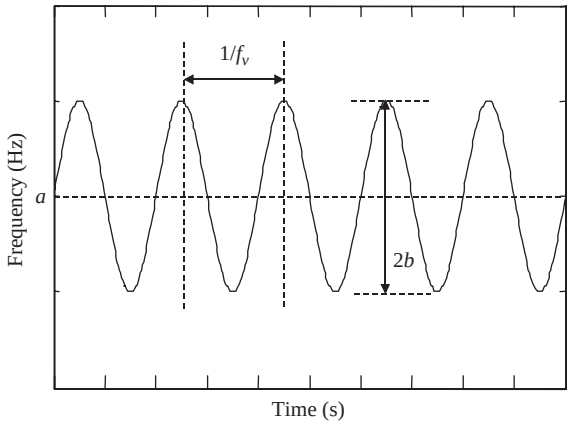

Fig. 1. Example of the fundamental frequency track of a vibrato tone.

where $a(n)$, $b(n)$, and $f_v(n)$ are considered as the intonation, extent, and rate of the vibrato and f_{frame} is calculated by dividing the sampling frequency f_s by the hop size in the STFT. An example of the fundamental frequency track is shown in Fig. 1 for a vibrato tone, where the three parameters are constant.

Considering $f(n)$ as a discrete-time function, the problem is now estimation of two amplitudes and one frequency. We choose the maximum likelihood (ML) estimation [3] for estimating the parameters a, b, and f_v, which are the instantaneous estimates of $a(n)$, $b(n)$, and $f_v(n)$. The ML estimation is effective for automatic detection of vibrato since it applies directly to the fundamental frequency track and does not require additional processes such as the maxima detection in the Prame's method [2]. We briefly explain the ML estimation procedure for further progress. The cost function for the ML estimates of the frequencies is represented as

$$L(f_v) = \mathbf{x}_{mr}^{\mathrm{T}}\mathbf{E}(\mathbf{E}^{\mathrm{H}}\mathbf{E})^{-1}\mathbf{E}^{\mathrm{H}}\mathbf{x}_{mr}, \tag{2}$$

where $^{\mathrm{H}}$ and $^{\mathrm{T}}$ denote the complex conjugate transpose and the transpose of a matrix and $\mathbf{x}_{mr}$ is a mean removed version of $\mathbf{x} = [f(m)\ \ldots\ f(m+M-1)]^{\mathrm{T}}$, $\mathbf{E} = [\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3]$, $\mathbf{e}_n = [1, \exp(2\pi \mathrm{j} fn), \ldots, \exp(2\pi \mathrm{j} fn(M-1))]^{\mathrm{T}}$, the normalized frequencies $f_1 = 0$, $f_2 = f_v/f_{frame}$, and $f_3 = -f_2$, and M is the data length. The vibrato rate is estimated as the one that maximizes Eq. (2). Then the intonation and vibrato extent are calculated as $|a_{11}|$ and $|a_{21}|+|a_{31}|$, respectively, where $|a_{k1}|$ is the element of the kth row of $\mathbf{A}$ calculated as

$$\mathbf{A} = (\mathbf{E}_v^{\mathrm{H}}\mathbf{E}_v)^{-1}\mathbf{E}_v^{\mathrm{H}}\mathbf{x}, \tag{3}$$

where $\mathbf{E}_v$ is calculated by the ML estimate of the vibrato rate. Finally, the averaging is performed to reduce the estimation error. For more detailed procedures, refer to [3].

As shown in Fig. 1, it is obvious that the most distinct features for a vibrato tone are two parameters, the vibrato rate and extent. Since human perception is proportional to the log scale and the vibrato extent is proportional to the intonation, we use the vibrato rate and normalized vibrato extent, which is calculated by dividing the extent by the intonation, as features. We define the probability of vibrato existence as a multiplication of two probabilities as

$$f(x_r, x_e) = f_{rate}(x_r) \cdot f_{extent}(x_e), \tag{4}$$

where the former is the probability related to the vibrato rate and the latter to the normalized vibrato extent. We primarily compare $f(x_r, x_e)$ with a threshold value to determine whether the vibrato is present or not. The merit of this representation is that the boundary curve between vibrato tones and normal tones is represented as a smooth curve.

For the vibrato rate, there is an optimal range that human subjects prefer [2,3]. If the rate is too low, human subjects recognize not the vibrato but the periodicity of the pitch change. If it is too high, the sound becomes a rather unpleasant confusion of more than one tone. So we model the probability as a modified form of a gaussian probability function as

$$f_{rate}(x_r) = \exp\left(-\frac{(x_r - f_v)^2}{2\sigma^2}\right), \tag{5}$$

where x_r is the vibrato rate.

The probability of vibrato existence generally increases as the vibrato extent increases. In practice, the extent cannot be too large and its range should be restricted. For example, periodic note changes have characteristics similar to the vibrato but should not be detected as vibrato. Considering these two aspects, we model the probability as

$$f_{extent}(x_e) = \begin{cases} \dfrac{1}{1+\exp(-c(x_e - x_{thd}))} & \text{for } x_e < e_{thd}, \\ 0 & \text{otherwise}, \end{cases} \tag{6}$$

where x_e is the normalized vibrato extent.

The coefficients in Eqs. (5) and (6) should be determined considering human perception. We first set the threshold for $f(x_r, x_e)$ as 0.5, considering a value of $1/\sqrt{2}$ each for $f_{rate}(x_r)$ and $f_{extent}(x_e)$. The preferred range of the vibrato rate differs depending on the music style but experiments show that the rate is generally 5–7 Hz for western music [2,3]. We set f_v as 6 and σ^2 as $1/\log_e 2$ so that $f_{rate}(x_r) = 1/\sqrt{2}$, for $x_r = (6 \pm 1)$ Hz in Eq. (5). For Eq. (6), we first set c as 1000 in order to make $f_{extent}(x_e)$ abrupt at $x_e = x_{thd}$. In contrast to the vibrato rate, the perceptive threshold for vibrato extent is somewhat ambiguous. Instead, experiments on subjective perception of frequency modulated signals show that the threshold is constant at low-frequency regions but is proportional to the frequency at mid- and high-frequency regions, where the proportionality coefficient is about 0.0035 [5]. Since the ML estimation extracts only the sinusoidal component in the fundamental frequency track, we use a smaller value for the threshold, which is 0.003. This leads to $x_{thd} = 0.0021186$ in Eq. (6),