FINDING INTENSITIES AND TEMPORAL CHARACTERISTICS IN PIANO MUSIC

Wai Man Szeto, Kin Hong Wong, Chi Hang Wong

Department of Computer Science and Engineering
The Chinese University of Hong Kong
Shatin, N.T., Hong Kong
{wmszeto, khwong, chwong1}@cse.cuhk.edu.hk

ABSTRACT

Timing and dynamics are two important factors in music performance. Research on dynamics-related issues is comparatively rare because data on dynamics is difficult to obtain from music performance. Nevertheless, research of this kind is vital to the understanding of music performance and here we are investigating ways to identify the intensities of individual notes in a mixture of simultaneous notes. The approach to this problem is divided into two stages. The first stage consists of obtaining the magnitude of the fundamental frequency of an individual note to determine its intensity out of a mixture of simultaneous notes, on condition that the corresponding pitches of which are given. Two simultaneous notes one or two octaves apart are also included in this study. The second stage consists of generating, artificially, a mixture of notes from a recorded single-note database, subsequently referred to as "estimated mixture". The time lag between individual notes in the estimated mixture is adjusted, so that the residual between which and the input comes to a minimum. The proposed method is verified with real data and the result is satisfactory.

1. INTRODUCTION

Timing and dynamics are two important factors in music performance. The "tone-colour" of piano playing, for example, is jointly determined by these factors [1][1]. It is pointed out in [2], however, that research on dynamics-related issues, unlikely timing, is relatively rare because data on dynamics is difficult to obtain from real piano performances, as it is most unusual for a piece of piano music to be entirely monophonic, i.e. constituting only of single notes being played at a time. The norm, instead, is to have at least 2 notes being played a time, forming mixtures of simultaneous notes. The dynamic level is, in fact, a combination of the intensities of individual notes in a mixture. Existing research in this respect tends to obtain data from digital pianos or computer-monitored pianos, for there are optical sensors to detect key pressing speed or final hammer velocity (in the case of computer-monitored pianos) to determine intensities of individual notes, while studies on acoustic signals are confined to dynamics of the whole mixture. An overview of research on timing and dynamics in piano performance can be found in [2, 3].

This problem is closely related to the area of blind source separation. Certain source separation algorithms have been specially designed for music signals [4, 5, 6, 7] and their results are encouraging. Nevertheless, the algorithms in [4, 5] are applied to separate sources from different musical instruments, while the algorithms in [6, 7] focus on music transcription, which is closely-related to our study. A music transcription system aims at turning acoustic signals into a score-like representation including pitches, onsets and durations of notes being played. A detailed review of different systems can be found in [8]. Although the intensities of notes are usually either ignored or roughly estimated in these transcription systems, these systems provide valuable information of pitches, onsets and durations of notes, making the problem of finding intensities of individual notes trackable.

The objective of our study is to determine the intensities of individual notes in a mixture provided that pitches and the approximate onsets of notes in the mixture are given. The range of notes used in our experiments is from C2 (f_0 = 65.4 Hz) to C6 (f_0 = 1046.5 Hz), encompassing 4 octaves[2]. The intensity of a note is represented by its MIDI velocity which is an integer ranging from 0 to 127. A greater MIDI velocity indicates greater intensity. The magnitude of the fundamental frequency f_0 of an individual note is obtained to determine its MIDI velocity out of a mixture of simultaneous notes. For maximum reliability of experimental results, these mixtures and notes of the single-note database are recorded under identical acoustic condition and technical setup. Two simultaneous notes one or two octaves apart are also included in this study. The time lag between individual notes in the estimated mixture is adjusted, so that the residual between which and the input comes to a minimum. The proposed method is verified with real data and the result is satisfactory.

The rest of the paper is organized as follows: related work is reviewed in Section 2, the proposed method is presented in Section 3, experimental results are given in Section 4, before a conclusion drawn in Section 5.

2. RELATED WORK

An early attempt to investigate the intensities of mixtures is found in [9]. In this empirical investigation, it is shown that the maximum amplitudes are linearly proportional to piano hammer velocities. There is a linear relationship between the sum of the peak amplitude of the individual notes and the peak amplitude of the two-note mixtures. In addition, it is found that the sum of the peak amplitudes of the individual notes is slightly less than the peak amplitude of the two-note mixtures.

A later empirical investigation come in [10], which laid the foundation of our study. This work investigates whether the rel-

[1] In [1], the term "agogics" is used instead of "timing" but their meanings are basically the same.

[2] To refer a specific pitch, C4 to B4 denotes the octave from middle C to B. C3 to B3 denotes the octave below middle C. C5 to B5 denotes the octave above the middle C and so on.

ative peak amplitude of recorded piano notes can be reliably inferred from the magnitudes of their fundamental frequency (first partial) and the second partial, which are measured in the spectrum near the note onset. It finds out that the peak RMS increases with the increase of MIDI velocity at a range of MIDI velocity. The magnitudes of their fundamental frequencies and the second partials generally increase linearly with the peak RMS. However, the peak RMS, and the magnitudes of their fundamental frequencies and the second partials vary substantially across pitches, even though the strings of different pitches are hit by the hammer at the same hammer velocity.

3. PROPOSED METHOD

3.1. Problem formulation

The time-domain signal of a single note with pitch p and intensity v is denoted as $x_{p,v}(t)$. The time-domain signal $y(t)$ is a mixture of n simultaneous notes and is modeled as the superposition of the time-domain signals of its individual notes

$$y(t) = x_{p_1,v_1}(t) + \cdots + x_{p_n,v_n}(t) \quad (1)$$

where p_i and v_i are the pitch and the intensity of the note i respectively. A note with a greater index i has a higher pitch, i.e., the pitch of note 1 is lowest while the pitch of the note n is highest. In our study, the pitch of each note in the mixture is known, i.e., each p_i is known, but each intensity v_i and each signal $x_{p_i,v_i}(t)$ are unknown.

In order to estimate each $x_{p_i,v_i}(t)$, there is a single-note database we recorded containing of all pitches from C2 to C6 played at a range of intensities. There are 7 intensity levels including the MIDI velocities 30, 40, 50, 60, 70, 80 and 90. The recording setup will be discussed in Section 3.2.

A note i with pitch p and intensity v in the database is denoted as $\hat{x}_{p_i,v_i}(t)$. An estimate of y, $\hat{y}$, is the superposition of the notes in the database:

$$\hat{y}(t, \hat{v}_1, \ldots, \hat{v}_n, \tau_1, \ldots, \tau_n) = \hat{x}_{p_1,\hat{v}_1}(t-\tau_1) + \cdots + \hat{x}_{p_n,\hat{v}_n}(t-\tau_n) \quad (2)$$

where $\hat{v}_i$ is the estimated intensity, and τ_i is the estimated time lag of the note i. The estimate ($\hat{y}$) has the same pitches of y. The reason of having the time lag variables is that the onsets of x and $\hat{x}$ may not be the same, because the strings are hit by the hammers at slightly different times. For example, it is shown in [2] that the melody note is consistently found to precede the other notes in the accompaniment by around 30 ms, even the pianist intended to play all notes simultaneously. This is because the melody note is usually played louder so the corresponding hammer will arrive at the strings earlier.

The error of the estimate is measured by the sum of the squared errors e:

$$e = \sum_{t=0}^{N} (y(t) - \hat{y}(t, \hat{v}_1, \ldots, \hat{v}_n, \tau_1, \ldots, \tau_n))^2 \quad (3)$$

where N is the time length of y and e is the power of the residue signal $y - \hat{y}$.

The problem is to find the optimal intensities $\hat{v}^* = \{\hat{v}_1^*, \ldots, \hat{v}_n^*\}$ and the optimal lags $\tau^* = \{\tau_1^*, \ldots, \tau_n^*\}$ to give an optimal $\hat{y}^*$ such that the sum of squared errors e is minimum. In other words, finding the optimal intensity of a note in a mixture is to classify it into one of the 7 intensity levels. The time lags τ are discrete sampling periods. A sampling period is equal to 1/(sampling frequency) = 1/(44100 Hz) = 0.0227 ms. We will find the optimal intensities and lags in two stages as shown in Figure 1. In these two stages, a single-note database is used. The proposed method is applied to two-note mixtures but it is extendable to tackle multi-note mixtures.

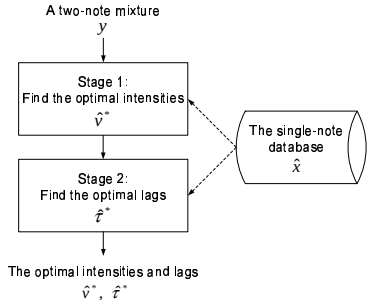

Figure 1: The flow of the proposed method.

3.2. Creation of the single-note database

The notes in the database were played by a computer-controlled piano which was a Yamaha Disklavier DU1A upright piano, Mark III series. During the recording session, both the top lid and the front face of the piano were open. The sound was recorded with a RØDE NT1000 condenser microphone placed approximately 20 cm above the keyboard and 18 cm in front of the C5 piano strings. This close-miking setup reduced the effect of room acoustics. The microphone was connected to an RME Fireface 800 Audio Interface, which acted as a microphone preamp and an A/D converter, and transferred the signals to a PC digitally through a firewire cable. The signals were stored in WAV format. The sampling frequency was 44.1 kHz and the number of bits per sample was 24. All notes from C2 to C6 were recorded. Each note was played at the MIDI velocities 30, 40, 50, 60, 70, 80 and 90 and lasted for 1 second. The notes played at MIDI velocities below 30 have very similar peak values to the MIDI velocity 30. The notes played at the MIDI velocities above 90 have very similar peak values to the MIDI velocity 90. Therefore, the database only contains the notes played in the range between 30 to 90. These 343 note samples constitute the single-note database. The reason of recording all notes is that the magnitudes of f_0 vary considerably across pitches as demonstrated in [10].

3.3. Stage 1: Finding the optimal intensities $\hat{v}^*$ of individual notes

It is shown in [10] that the magnitude of the fundamental frequency f_0 increases with the increase of the MIDI velocity. Similar result was obtained in our study. The f_0 magnitude of the C notes is shown in Figure 2. The f_0 magnitude generally increases monotonically with the MIDI velocity. Therefore, given the f_0 magnitude of a note recorded in the same experimental setup of recording the notes in the single-note database, its corresponding MIDI velocity can be uniquely determined. There are two cases of finding the MIDI velocity of each note in a mixture.

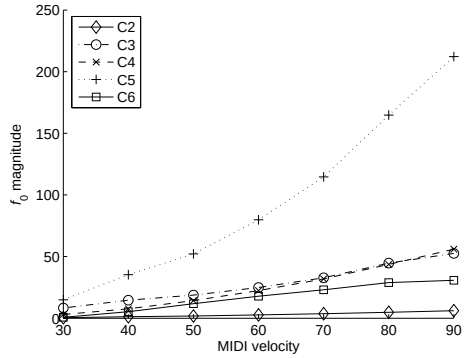

Figure 2: f_0 magnitude of C2, C3, C4, C5 and C6 against the MIDI velocity.

3.3.1. Case 1: Non-overlapping f_0

The sound of a piano note consists a series of frequency components. The lowest frequency component is the fundamental frequency (f_0) or first partial. Other components which are approximately the multiples of the f_0 are also called partials. For example, the f_0 of C4 is 262 Hz. Its second partial f_1 is approximately equal to $262 \times 2 = 524$ Hz, the third partial f_2 is $262 \times 3 = 786$ Hz, and so on. The first case of finding individual intensity is that the f_0 of a note does not overlap with the partials of other notes, such as G4 in a C4-G4 mixture. The f_0 of G4 is 392 Hz so it does not overlap with the frequency components of C4. Noted that in a mixture, the f_0 of the lowest note does not overlap with the partials of any other notes. Therefore, finding the optimal intensity of the lowest note always belongs to Case 1.

Before going into the details, the notation is presented first. The Fourier spectrum of y is Y. A pitch is equal-tempered if its f_0 follows the equal temperament tuning. The standard A4 is 440 Hz. The f_0 of the equal-tempered pitch p is denoted as $f_0(p)$ so $f_0(p)$ equals 440 Hz if p is A4. The $f_0(\hat{x}_{p,v})$ is the f_0 of $\hat{x}_{p,v}$. It is found by firstly finding the Fourier spectrum $\hat{X}_{p,v}$ of $\hat{x}_{p,v}$. Then in the magnitude spectrum $|\hat{X}_{p,v}|$, a peak is picked in the region of $\pm$ half semitone at the equal-tempered pitch p. The frequency of this peak is $f_0(\hat{x}_{p,v})$. The magnitude at $f_0(\hat{x}_{p,v})$ is $|\hat{X}_{p,v}(f_0(\hat{x}_{p,v}))|$ or simply $|\hat{X}_{p,v}(f_0)|$. If a mixture y contains the pitch p, the f_0 magnitude of p in Y is $|Y(f_0(p))|$. It is determined in the way that in the magnitude spectrum $|Y|$, a peak is picked in the region of $\pm$ half semitone at the equal-tempered pitch p. Then the peak is $|Y(f_0(p))|$. Here are the steps to determine the optimal intensity $\hat{v}_i^*$ of a note i with the pitch p_i:

1. Find the Fourier spectrum Y of y by FFT
2. Find the magnitude of f_0 of the note i in $|Y|$, denoted by $|Y(f_0(p_i))|$.
3. The optimal $\hat{v}_i^*$ is the $\hat{v}_i$ having the closest f_0 magnitude in the single-note database:

$$\hat{v}_i^* = \arg\min_{\hat{v}_i} \left| |Y(f_0(p_i))| - |\hat{X}_{p_i,\hat{v}_i}(f_0)| \right| \text{ for all } \hat{v}_i \quad (4)$$

3.3.2. Case 2: Overlapping f_0

If the f_0 of a note overlaps with the partials of other notes, such as C5 in a C4-C5 mixture and G5 in a C4-G5 mixture, the f_0 cannot be used directly to determine the intensity. In the case of the C4-C5 mixture, The f_1 of C4 is approximately equal to the f_0 of C5 (Figure 3 (a) and (b)). The magnitude at the f_0 of C5 in the spectrum of the C4-C5 mixture is contributed by both the f_0 of C5 and the f_1 of C4. A naive method is to subtract the f_1 magnitude of C4 from the mixture at that frequency. However, unless they are exactly in phase, the magnitude at that frequency in the mixture spectrum is not equal to the addition of the f_0 magnitude of C5 in the C5 spectrum and the f_1 magnitude of C4 in the C4 spectrum. In Figure 3 (c), the spectrum of the C4-C5 mixture recorded under the same experimental setup is shown. The notes in the mixture and the single notes all were played at MIDI velocity 70. The magnitude at the f_0 of C5 (equivalently, f_1 of C4) in the spectrum is even lower than both the magnitude of the individual C4 and C5 at that frequency. In order to find the f_0 magnitude of C5, the C4 note must be removed from the mixture. To remove the lower

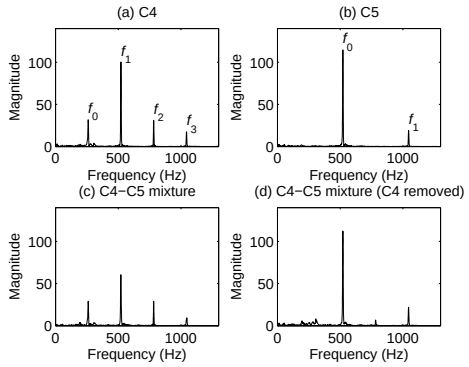

Figure 3: (a) The spectrum of C4. (b) The spectrum of C5. (c) The spectrum of the C4-C5 mixture. (d) The spectrum of the C4-C5 mixture from which C4 is removed. All notes were played at MIDI velocity 70. Hamming window was applied from the onsets and its length was 16384 (372 ms).

note in a two-note mixture, the intensity of the lower note is firstly estimated so the corresponding lower note signal in the database $\hat{x}_{p_1,\hat{v}_1}$ is subtracted from the mixture y to give $\tilde{y}$. However, before the subtraction can be done, the time lag $\tilde{\tau}_1$ between the mixture and the lower note signal must be found.

$$\tilde{y}(t,\tilde{\tau}_1) = y(t) - \hat{x}_{p_1,\hat{v}_1^*}(t-\tilde{\tau}_1) \quad (5)$$

If the lower note is successfully removed from the mixture, the magnitude of the f_0 of the lower note in the mixture spectrum has a very small value. Therefore, the lag $\tilde{\tau}_1$ is chosen if it gives the minimum f_0 magnitude of the lower note in the mixture spectrum for a lag range from $\tau_{\min}$ to $\tau_{\max}$.

$$\tilde{Y}(f,\tilde{\tau}_1) = \text{FFT}(\tilde{y}(t,\tilde{\tau}_1)) \quad (6)$$

$$\tilde{\tau}_1^* = \arg\min_{\tilde{\tau}_1} |\tilde{Y}(f_0(p_1),\tilde{\tau}_1)| \text{ for } \tau_{\min} \le \tilde{\tau}_1 \le \tau_{\max} \quad (7)$$

The $\tau_{\min}$ and $\tau_{\max}$ will be determined by the experiment in Section 4. After the lower note is removed, the intensity of the upper note can be estimated by using the method in Case 1. In Figure 3 (d), C4 is removed from the C4-C5 mixture by using the above method so only C5 is left. The f_0 of this C5 is close to the f_0 of the C5 in the single-note database.

3.4. Stage 2: Finding the optimal time lags τ^*

3.4.1. Method 1: Exhaustive search

The simplest way to find the optimal lags is to use exhaustive search which calculates the errors for all possible time lags. Let k be the number of samples in the range from $\tau_{\min}$ to $\tau_{\max}$, and n be the number of notes in the mixture. Then the search space contains k^n points. A better way is to use optimization techniques to reduce the search space. However, the error function is non-differentiable and discontinuous as the function depends on the series of signal values of y and $\hat{x}$, and the discrete lags τ. To solve this problem, we use the pattern search algorithm in [11] implemented by [12].

3.4.2. Method 2: Pattern search

The pattern search algorithm is a class of direct search algorithms. At each iteration, the pattern search algorithm searches a set of points, called a mesh, around the current point. The algorithm constructs the mesh by adding the current point to a scalar multiple of a fixed set of vectors called a pattern. The scalar multiple is called the step-length parameter. If the algorithm finds a point in the mesh that improves the objective function at the current point, the new point becomes the current point at the next iteration of the algorithm, and the step-length parameter increases to expand the mesh. If the algorithm fails to find a point improving the objective function, the step-length parameter decreases to shrink the mesh and the algorithm does not change the current point at the next iteration. Further details can be found in [12, 11].

In this paper, the objective function is the residual power in Equation 3. The optimization variables are τ_1 and τ_2, the lags of the lower note and the upper note respectively. The patterns used are $(0,1)$, $(1,0)$, $(0,-1)$, and $(-1,0)$. This means that at each iteration, the algorithm searches in the direction of north, east, south and west. It tries to find a point in the mesh that best improves the objective function. If such point is found, the step-length parameter is multiplied by 2; otherwise, the step-length parameter is multiplied by $1/2$. The initial step-length parameter is set to 1. In the pattern search algorithm, a starting point $(\hat{\tau}_1, \hat{\tau}_2)$ is required. We calculate the starting point from the phase shifts. The phase of f_0 of the shifted single note from the database would be close to the phase of f_0 of that note in the mixture. Therefore, it is reasonable to guess the starting point by shifting $\hat{x}_{p_i,\hat{v}_i^*}$ by $\hat{\tau}_i$. Such that the phase of the f_0 of $\hat{x}_{p_i,\hat{v}_i^*}$ in the spectrum $\hat{X}$, $(\angle \hat{X}_{p_i,\hat{v}_i^*}(f_0))$, is equal to that of the f_0 of $\hat{x}_{p_i,\hat{v}_i^*}$ in the spectrum Y, $(\angle Y(f_0(\hat{x}_{p_i,\hat{v}_i^*})))$, so the starting point is $(\hat{\tau}_1, \hat{\tau}_2)$ where

$$\hat{\tau}_i = -\frac{\angle Y(f_0(\hat{x}_{p_i,\hat{v}_i^*})) - \angle \hat{X}_{p_i,\hat{v}_i^*}(f_0)}{2\pi f_0(\hat{x}_{p_i,\hat{v}_i^*})} \tag{8}$$

For the upper note in Case 2 (overlapping f_0), Y is replaced by $\tilde{Y}$.

Moreover, the error function is highly nonlinear as shown in Figure 4[3]. The error map is generated by the exhaustive search method. The reason for the nonlinearity is that shifting a period $(1/f_0)$ of $\hat{x}_{p_i,\hat{v}_i^*}$ gives a similar error value. If $\hat{x}$ has strong f_1 or even f_2, shifting a half period $(1/(2f_0))$ or one third of the period $(1/(3f_0))$ will also give a similar error value. If there is only one starting point, it is easily trapped into a local minimum. To avoid this problem, a set of starting points are generated. The new starting points are formed by shifting $\hat{\tau}_i$ by various $1/s_i$ periods in the range from $\tau_{\min}$ and $\tau_{\max}$ where s_i is an integer:

$$\mathcal{T}_i = \{\tau_i^1, \ldots, \tau_i^{m_i}\} \tag{9}$$

where $\hat{\tau}_i = \tau_i^k$ for some k, $\tau_i^{k+1} - \tau_i^k = 1/(s_i \cdot f_0(\hat{x}_{p_i,\hat{v}_i^*}))$, $\tau_{\min} \le \tau_i^k \le \tau_{\max}$ for all k, $\tau_i^0 < \tau_{\min}$, and $\tau_i^{m_i+1} > \tau_{\max}$.

A starting point $(\tau_1^{k_1}, \tau_2^{k_2})$ is in the Cartesian product of $\mathcal{T}_1 \times \mathcal{T}_2$. The integer variable s_i is equal to 3 for the pitch p_i of a note is equal to or below C3, and it is equal to 2 otherwise. A grid of starting points is shown in Figure 4. In the pattern search algorithm, bound constraints are added to avoid duplicated search. For the dimension τ_i, the bound of a starting point $(\tau_1^{k_1}, \tau_2^{k_2})$ is $\tau_i^{k_i} \pm 1/(2 \cdot s_i \cdot f_0(\hat{x}_{p_i,\hat{v}_i^*}))$. From each starting point, the algorithm searches for a local minimum. The estimated global minimum is the minimum of these local minima.

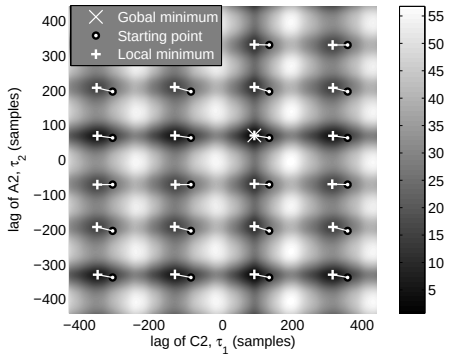

Figure 4: The error map of τ_1 and τ_2 of a C2-A2 mixture. A white line connects a starting point and its corresponding local minimum. $\hat{x}_{p_1,\hat{v}_1^*}$ is C2 at MIDI velocity 70 and $\hat{x}_{p_2,\hat{v}_2^*}$ is A2 at MIDI velocity 70. The periods of these C2 and A2 are 676 and 401 samples respectively. The variables s_1 and s_2 are equal to 3 so along each axis, adjacent starting points are separated by one third of the period. The period is the period of C2 for the x-axis and is the period of A2 for the y-axis. The residue-to-signal ratio at the global minimum (133, 63) is 0.0159.

4. EXPERIMENTS AND EVALUATIONS

Since the search range of Stages 1 and 2 depends on the variability of the computer-controlled piano, we need to determine the search range in order to obtain a better result. The procedure will be presented in Section 4.1. After the search range is found, we will test our proposed methods on real data. In Section 4.2, we will evaluate the method of finding optimal intensities in Stage 1. The exhaustive search and the pattern search methods in Stage 2 will be compared in Section 4.3. Finally, we will show the overall performance of our proposed methods.

4.1. Finding the search range and residue-to-signal ratio

In this experiment, the objective is to determine the search range, i.e., the minimum lag $\tau_{\min}$ and the maximum lag $\tau_{\max}$, and investigate the residue power of the same single notes. As the computer-controlled piano is a mechanical device, each playing of the same

[3] A full-colour error map is available at http://www.cse.cuhk.edu.hk/∼wmszeto/dafx05/demo.htm.

note at the same MIDI velocity is slightly different from time to time. In order to study this variability, all notes from C2 to C6 were played at the MIDI velocity 70 for 4 times except C2, C3, C4, C5 and C6 which were played 6 times. Similar to two-note mixtures, the residue power and the time lags are defined as below. $x^i(t)$ is the i-th time of a note being played. The aim is to find the optimal time lag $\tau_*^{i,j}$ such that the sum of squared errors (residue power) $e^{i,j}$ is minimum and the residue power is

$$e^{i,j} = \sum_{t=0}^{N} (x^i(t) - x^j(t - \tau^{i,j}))^2 \tag{10}$$

where $N = 22049$ (500 ms).

The average absolute lag is the average of all $|\tau^{i,j}|$ for $i \neq j$. The residue-to-signal ratio (RSR) of $x^i(t)$ and $x^j(t)$ is defined as

$$\text{RSR}^{i,j} = \frac{1}{2}\Big(\frac{e^{i,j}}{\sum_{t=0}^{N}(x^i(t))^2} + \frac{e^{i,j}}{\sum_{t=0}^{N}(x^j(t))^2}\Big) \tag{11}$$

All discrete time lags were tested in the range from -1323 samples (-30 ms) to 1323 samples (30 ms) to find the optimal lags . The mean and the standard deviation of the average RSRs were 0.0315 and 0.0405 respectively, while the mean and the standard deviation of the average absolute lag were 87.6 samples (2.00 ms) and 30.1 samples (0.684 ms) respectively. The minimum lag $\tau_{\min}$ and the maximum lag $\tau_{\max}$ are decided to be -441 samples (-10 ms) and 441 samples (10 ms) respectively. The residue-to-signal of single notes can be used to compare the residue-to-signal of two-note mixtures.

4.2. Stage 1: Finding the optimal intensities $\hat{v}^*$

A wide range of two-note mixtures was chosen as below. The lower note was the note C2. The upper note was selected in the way that the mixtures included all possible intervals in an octave and also a double octave. As a result, the mixtures were C2-C♯2, C2-D2, C2-D♯2, C2-E2, C2-F2, C2-F♯2, C2-G2, C2-G♯2, C2-A2, C2-A♯2, C2-B2, C2-C3 and C2-C4. This selection rule was repeated for the lower note C3, C4 and C5 except C5-C7 which was discarded because C7 was out of the investigating range C2 to C6. Therefore, there were 51 two-note mixtures consisting of 102 notes under test. These mixtures were recorded in the same experimental setup in Section 3.2 as the notes in the single-note database. Each note in all mixtures were played at MIDI velocity 70 and lasted for 1 second. The two notes in a mixture had the same onset and offset.

The method proposed in Section 3.4 was tested to find the optimal intensities. The time length N of a mixture y is 32768 (743 ms) for C2-C♯2, C2-D2 and C3-C♯3, and N is 16384 (372 ms) for other mixtures. A longer time length is required for those three mixtures because the f_0 of the pair of the notes in the mixtures are very close. The Hamming window was applied before FFT.

The result is shown in Figure 5. The successful rate is 93.1%. 95 of 102 notes were correctly classified to MIDI velocity 70. The misclassified notes are C4 in C2-C4, C3 in C3-C♯3, C3 in C3-D3, C4 in C3-C4, C6 in C4-C6, D♯5 in C5-D♯5, and C6 in C5-C6. This may be due to the variability of the notes played in the computer-controlled piano as shown in Section 4.1.

4.3. Stage 2: Finding the optimal time lags τ^*

Given the correct MIDI velocity (70), the exhaustive search method in Section 3.4.1 was used to find the global minimum with the

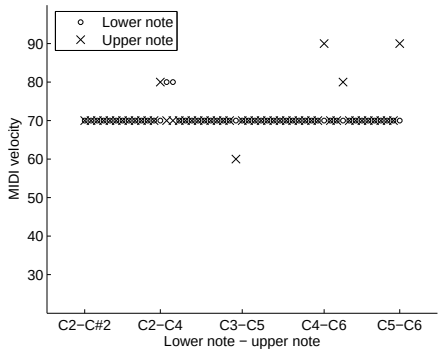

Figure 5: The MIDI velocity found. All notes were played at MIDI velocity 70. 95 of 102 notes were correctly classified to MIDI velocity 70. The successful rate is 93.1%. The pairs of the lower note and the upper note in the x-axis from left to right are C2-C♯2, C2-D2, C2-D♯2, ..., C2-B2, C2-C3, C2-C4, C3-C♯3, ..., C5-B5, C5-C6.

range of τ in $[\tau_{\min}, \tau_{\max}]$. Then given the same MIDI velocity 70, the method of the pattern search in Section 3.4.2 was used to estimate the global minimum. This experiment aims to verify the correctness of the pattern search method. If one of the local minima coincides with the global minimum, the global minimum is correctly estimated as shown in Figure 4.

The result is that the pattern search method correctly found the global minimum of all 51 two-note mixtures. This means that the pattern search method and the exhaustive search method gave the same result. We also investigated the reduction in the number of search points of the pattern search method comparing to the exhaustive search method. The reduction was calculated as

$$\text{Reduction in search points} = 1 - \frac{N_p}{N_e} \tag{12}$$

where N_p is the number of search points in the pattern search method and N_e is the number of search points in the exhaustive search method.

The result is shown in Figure 6. The average reduction is 99.1%. The number of search points in the pattern search method is much less than that in the exhaustive search method, which is $(441 \times 2 + 1)^2 = 779689$. There is a decreasing trend of the reduction for the higher pitches. This is because a higher pitch signal has a shorter period so the number of starting points increases.

4.4. Overall performance

We tested the overall performance by using both the optimal intensities and the optimal lags found in the previous sections. In the first testing condition, the optimal intensities found in Section 4.2, including the misclassifications, were used to find the optimal lags by the pattern search method proposed in Section 3.4.2. Therefore, the first testing condition shows the overall performance of our proposed method. In the second testing condition, The optimal intensities (MIDI velocity 70) were assumed to be correctly found and they were used to find the optimal lags by the exhaustive search method proposed in Section 3.4.1. This testing condition provides a benchmark for the overall performance. We compare these two testing conditions by the residue-to-signal ratio (RSR)

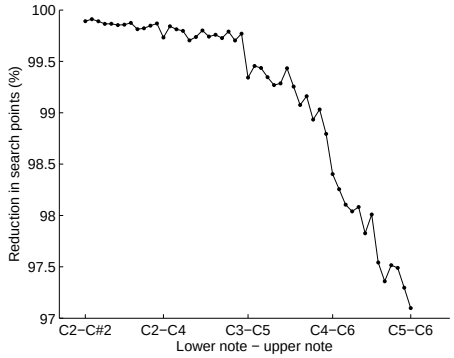

Figure 6: The reduction in search points of the pattern search method comparing to the exhaustive search method. The pairs of the lower note and the upper note in the x-axis from left to right are C2-C♯2, C2-D2, C2-D♯2, . . . , C2-B2, C2-C3, C2-C4, C3-C♯3, . . . , C5-B5, C5-C6.

of two-note mixtures: $RSR = e/\sum_{t=0}^{N}(y(t))^2$ where e is the residue power defined in Equation 3.

The residue-to-signal ratios of the first and second testing conditions are depicted in Figure 7. The first testing condition performed worse only in the case of the misclassifications of optimal intensities. For the first testing condition, the mean and the standard deviation of RSR were 0.0351 and 0.0705 respectively. For the second testing condition, the mean and the standard deviation of RSR were 0.0143 and 0.0091 respectively. The results are comparable to the case of the single notes. Demonstrations are available at http://www.cse.cuhk.edu.hk/∼wmszeto/dafx05/demo.htm

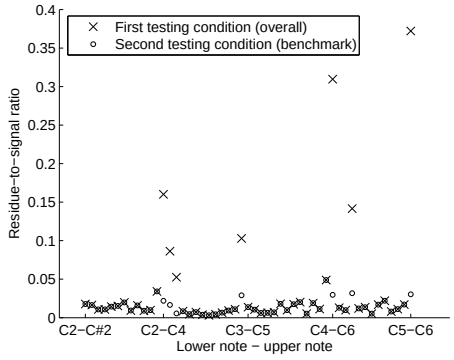

Figure 7: The residue-to-signal ratio. The pairs of the lower note and the upper note in the x-axis from left to right are C2-C♯2, C2-D2, C2-D♯2, . . . , C2-B2, C2-C3, C2-C4, C3-C♯3, . . . , C5-B5, C5-C6.

5. CONCLUSIONS

In this study, we formulate the problem of finding individual intensity and relative onset of notes in a mixture. A two-stage method is proposed to tackle this problem. The experimental result shows that 95 of 102 notes are correctly classified to MIDI velocity 70. The successful rate is 93.1%. All optimal time lags are found by the pattern search algorithm. Comparing to the exhaustive search method, there is a great reduction of the number of search points in the pattern search method. In the future, mixtures containing more than 2 notes will be examined. Moreover, mixtures containing the notes with the MIDI velocity not in the single-note database will also be investigated. Another possible extension is that, in order to improve the successful rate, if the ratio of the residual power to the signal power is greater than a certain threshold, the neighboring MIDI velocities can be searched iteratively.

6. ACKNOWLEDGEMENTS

We are thankful to Eos Cheng, Yiu Tong Chan, Hau Man Lo, Kai Tong Sun, and Central Music (H. K.) Company for their valuable comments.

7. REFERENCES

[1] J. Gát, *The Technique of Piano Playing*, Collet's (Publishers) Limited, 5th edition, 1980.

[2] W. Goebl, *The Role of Timing and Intensity in the Production and Perception of Melody in Expressive Piano Performance*, Ph.D. thesis, Karl-Franzens-Universitat Graz, 2003.

[3] A. Askenfelt, Ed., *Five Lectures on the Accoutics of the Piano*, Royal Swedish Academy of Music, 1990, Available at http://www.speech.kth.se/music/5_lectures/.

[4] T. Virtanen, "Separation of sound sources by convolutive sparse coding," in *Workshop on Statistcal and Perceptual Audio Processing (SAPA)*, Jeju, Korea, 2004.

[5] M. Davy and S. Godsill, "Bayesian harmonic models for musical signal analysis," in *Bayesian Statistics VII*. Oxford University Press, 2003.

[6] S. A. Abdallah and M. D. Plumbley, "An independent component analysis approach to automatic music transcription," in *Proceedings of the 114th Convention of the Audio Engineering Society*, Amsterdam, March 2003.

[7] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *IEEE Workshop on Application of Signal Processing to Audio and Acoustics*, New Paltz, NY, 2003, pp. 177–180.

[8] A. Klapuri, *Signal Processing Methods for the Automatic Transcription of Music*, Ph.D. thesis, Tampere University of Technology, 2004.

[9] C. Palmer and J. C. Brown, "Investigations in the amplitude of sounded piano tones," *Journal of the Acoustical Society of America*, vol. 90, no. 1, pp. 60–66, 1991.

[10] B. H. Repp, "Some empirical observations on sound level properties of recorded piano tones," *Journal of the Acoustical Society of America*, vol. 93, no. 2, pp. 1136–1144, 1993.

[11] R. M. Lewis and V. Torczon, "Pattern search algorithms for bound constrained minimization," *SIAM Journal on Optimization*, vol. 9, no. 4, pp. 1082–1099, 1999.

[12] MathWorks, "Genetic algorithm and direct search toolbox version 1.0," 2004.