EXPLOITING FREQUENCY, PERIODICITY AND HARMONICITY USING ADVANCED TIME-FREQUENCY CONCENTRATION TECHNIQUES FOR MULTIPITCH ESTIMATION OF CHOIR AND SYMPHONY

Li Su, Tsung-Ying Chuang and Yi-Hsuan Yang
Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan
lisu@citi.sinica.edu.tw, jasonmail04@gmail.com, yang@citi.sinica.edu.tw

ABSTRACT

To advance research on automatic music transcription (AMT), it is important to have labeled datasets with sufficient diversity and complexity that support the creation and evaluation of robust algorithms to deal with issues seen in real-world polyphonic music signals. In this paper, we propose new datasets and investigate signal processing algorithms for multipitch estimation (MPE) in choral and symphony music, which have been seldom considered in AMT research. We observe that MPE in these two types of music is challenging because of not only the high polyphony number, but also the possible imprecision in pitch for notes sung or played by multiple singers or musicians in unison. To improve the robustness of pitch estimation, experiments show that it is beneficial to measure pitch saliency by jointly considering frequency, periodicity and harmonicity information. Moreover, we can improve the localization and stability of pitch by the multi-taper methods and nonlinear time-frequency reassignment techniques such as the Concentration of Time and Frequency (ConceFT) transform. We show that the proposed unsupervised methods to MPE compare favorably with, if not superior to, state-of-the-art supervised methods in various types of music signals from both existing and the newly created datasets.

1. INTRODUCTION

The ability to identify concurrent pitches in polyphonic music is considered admirable by most people. Throughout history, such an ability has been symbolic of a music genius, with the most popular legendary story possibly being Mozart's transcription of Allegri's *Miserere* at the age of fourteen. An interesting question is then whether computers can also possess the ability and perform automatic music transcription (AMT). A great deal of research has been done in the music information retrieval (MIR) community to develop AMT algorithms, but to date it is still an unsolved problem [4].

AMT is challenging for multiple reasons. One such challenge has to do with the creation of labeled multipitch data with diversity, for the labeling process requires considerable expertise and is usually time-consuming [24]. Existing multipitch datasets are often small in size and limited in diversity, and in combination they still cannot represent the rich variety found in music performances. For example, to our knowledge, there is no labeled multipitch data for choir, one of the most common type of music through the ages and cultures and also known as the theme featuring the legendary story of Mozart. As the evaluation of AMT algorithms requires labeled data, the transcription of choir music remains largely unexplored.

The rich variety of music also poses challenges in designing features robust to variations in timbre, genre, and type of performance. For example, it is difficult to design a feature that performs equally well in characterizing the pitch information in both piano and choir music, for they are fairly different — the latter involves a group of people singing in unison but each having her or his own vocal characteristics and control of pitch. This specific issue of possible imprecision in pitch has rarely been dealt with in the literature, possibly due to the scarcity of related labeled data. The shift-invariant Probabilistic Latent Component Analysis (PLCA) algorithm [3] can support non-ideal tuning and frequency so might be able to partially address this issue, but such an evaluation has not been reported before. Moreover, while PLCA is a supervised algorithm that demands the availability of data, we are interested in unsupervised algorithms.

This work attempts to address the aforementioned issues for multipitch estimation (MPE), a sub-task of AMT. Specifically, we propose new datasets and discuss the characteristics and distinct technical issues of MPE for choir and symphony music. Besides, by extending a previous work [23] we propose an unsupervised approach that interprets a pitch event in three dimensions — frequency, periodicity and harmonicity. Moreover, we introduce recent advance in time-frequency (TF) analysis, including the Synchrosqueezing Transform (SST) and the Concentration of Time and Frequency (ConceFT) method, to improve the stablization and localization of pitch information in our feature representation. Result shows that that the proposed unsupervised methods compare favorably with state-of-

Attribution: Li Su, Tsung-Ying Chuang and Yi-Hsuan Yang. "Exploiting frequency, periodicity and harmonicity using advanced time-frequency concentration techniques for multipitch estimation of choir and symphony", 17th International Society for Music Information Retrieval Conference, 2016.

the-art supervised methods in various types of music. Finally, a simple decision fusion framework also shows the effectiveness of combining multiple MPE methods.

2. PROBLEM DEFINITION

To facilitate our discussion on feature representation, we focus on the feature design for *frame-level* transcription of polyphonic music, namely the MPE sub-task. Other transcription sub-tasks such as note tracking and timbre streaming [7] are not discussed in this paper.

We refer to a multipitch signal as a superposition of multiple "perceptually mono-pitch" signals. This particular type of mono-pitch signal can be produced either by a single performer, with rather well-defined pitch and loudness, or by a group of performers playing instruments or singing in unison. The latter case, often referred to as "chorus" or "ensemble" sounds, has quite different signal-level characteristics from the former. The major difference lies in the small, independent variations in the fundamental frequency (F0), a.k.a. the *voice flutter* phenomenon [25]. For example, early research in choral music showed that the "dispersion of F0" (measured as the bandwidths of partial tones) among three reputable choirs varied typically in the range of 20–30 cents [17]. It is also found that the *pitch scatter* (i.e., the standard deviation of F0 across singers, averaged over the duration of each tone [17, 25]) among choir basses is 10–15 cents [26]. Previous work on the synthesis of chorus/ensemble effect also adjusted pitch scatter parameters in similar ranges [12, 18].

We assume that every sound contributing to the mono-pitch signal of interest is composed of a series of sinusoidal components which are with *nearly* integer multiples of the F0, i.e., every sound has *low inharmonicity*. In this way, the bandwidth of each partial is mainly determined by the amount of the frequency variations of every sound. Besides, we ignore issues of *missing fundamental* or *stacked harmonics* found in real-world polyphonic signals, for they both have been discussed in our previous work [23]. In summary, we assume the signal under analysis has discernable F0s and small but nonzero degree of inharmonicity and pitch scatter.

3. RELATED WORK

3.1 Pitch saliency features

For a signal $x(t)$ with multiple periodicities, a pitch candidate is determined by 1) a *frequency representation* $V(t, f)$ that reveals the saliency of every fundamental frequency and its harmonic frequencies (i.e., its integer multiples) in a signal, 2) a *periodicity representation* $U(t, q)$ that reveals the saliency of every fundamental period and its integer multiples in a signal, [1] and 3) the *constraints on harmonicity* described as follows: at a specific time t_0, a pitch candidate $f_0 = 1/q_0$ is the true pitch when there exists $M_v, M_u \in \mathbb{N}$ such that [23]:

1. A sequence of prominent peaks found at $V(t_0, f_0)$, $V(t_0, 2f_0), \ldots, V(t_0, M_v f_0)$.
2. A sequence of prominent peaks found at $U(t_0, q_0)$, $U(t_0, 2q_0), \ldots, U(t_0, M_u q_0)$.

An MPE algorithm following this approach has been found useful in transcribing a wide variety of music, including complicated music signals like symphony [23]. The frequency representation being used is the short-time Fourier transform (STFT). For a window function $h(t)$, the STFT of $x(t)$ is formulated by

$$V_x^{(h)}(t, f) = \int x(\tau)h(\tau - t)e^{-j2\pi f(\tau - t)}\mathrm{d}\tau\,. \quad (1)$$

For periodicity representations, an important one for MPE is the the *generalized cepstrum* [16, 28]:

$$U_x^{(h,g_\xi)}(t, q) = \int g_\xi(V_x^{(h)}(t, f))e^{-j2\pi qu}\mathrm{d}u\,, \quad (2)$$

where q is referred to as *lag* or *quefrency*, and $g_\xi(\cdot)$ is a nonlinear scaling function defined by either $g_1(y) = |y|^\gamma$ or $g_2(y) = (|y|^\gamma - 1)/\gamma$, $0 < \gamma \leq 2$. We remark that when $\gamma = 2$, $U_x^{(h,g_1)}$ becomes the ACF according to the Wiener-Khinchin theorem, and when $\gamma \to 0$, $U_x^{(h,g_2)}$ approximates to the *real cepstrum*. $U_x^{(h,g_1)}$ and $U_x^{(h,g_2)}$ are different merely in the scale and the zero-quefrency term, so for simplicity we use $U_x^{(h,g_1)}$ in this paper. We will omit ξ in the notation and simply denote $U_x^{(h,g_\xi)}(t, q)$ as $U_x^{(h)}(t, q)$.

3.2 Multi-taper time-frequency analysis

In conventional STFT, the spectrum is estimated by only one window function. In contrast, multi-taper TF analysis estimates the spectrum by averaging of the spectral estimation of multiple windows (i.e., tapers) [27]. The main purpose of multi-tapering is to stabilize the spectrum estimation by reducing the variance due to noises or boundary of the segments. The tapers are basically orthogonal to each other, and their estimates are approximately uncorrelated. Therefore, the average of them can reduce the variance. To be more specific, given $\mathbf{z} = [\nu_1, \ldots, \nu_J]$, a set of J-taper window with good concentration in the TF plane, where $\nu_i \in \mathbb{R}^T$ is a window of length T for $i = 1, 2, \ldots, J$, and $\mathbf{z}$ forms an orthonormal basis in $\mathbb{R}^T$, the multi-taper STFT is given by $\frac{1}{J}\sum_{j=1}^{J} V_x^{(\nu_j)}(t, f)$, the average of every $V_x^{(\nu_j)}(t, f)$.

Although rarely seen in the literature of MIR, the multi-tapering method has been found useful in several different applications in speech processing such as speaker identification and emotion recognition [1, 14], because of its stable output of feature representation.

3.3 Synchrosqueezing Transform (SST)

SST is a special case of *time-frequency reassignment* [2], a class of nonlinear TF analysis techniques. In a nutshell, it aims at moving the spectral-leakage terms caused by Heisenberg-Gabor uncertainty to the center of mass

[1] The fundamental period of a periodic signal $x(t)$ is defined as the smallest q such that $x(t + q) = x(t)$. Since q is measured in time, we refer to q as in the *lag domain*, to distinguish it from t in the *time domain*.

of true component, and therefore sharpens the harmonic peaks and achieves high localization [5]. SST uses the *frequency reassignment vectors* estimated by the *instantaneous frequency deviation* (IFD). In music processing, such a method can better discriminate closely-located components, and applications have be found in music processing tasks such as chord recognition, synthesis, and melody extraction [11, 13, 20, 22].

Let $V_x^{(h)} = |V_x^{(h)}|e^{\Phi_x^h}$. The IFD, $\Omega_x^{(h,\theta_v)}$, is defined as the time derivative of the instantaneous phase term Φ_x^h:

$$\Omega_x^{(h,\theta_v)}(t,\eta) := \frac{\partial \Phi_x^{h_n}}{\partial t} = -\Im \left. \frac{V_x^{(\mathcal{D}h)}(t,\eta)}{V_x^{(h)}(t,\eta)} \right|_{\mathfrak{N}_v} , \quad (3)$$

where $\Im$ means the imaginary part and $\mathfrak{N}_v := \{f : |V_x^{(h_n)}(t,f)| > \theta_v\}$ gives a threshold so as to avoid computation instability when $|V_x^{(h_n)}(t,f)|$ is very small. This formulation (3) can be derived by definition. The SST is therefore represented as

$$S_x^{(h,\theta_v)}(t,f) = \int_{\mathfrak{N}_v} V_x^{(h)}(t,\eta)\delta\left(|f - \Omega_x^{(h)}(t,\eta)|\right) \mathrm{d}\eta \, . \quad (4)$$

As the result of our analysis is not sensitive to the value of the parameter θ_v, we set θ_v to 10^{-6} of the root mean square energy of the signal under analysis throughout the paper. For convenience, we also omit θ_v in the notation and simply denote $\Omega_x^{(h,\theta_v)}$ and $S_x^{(h,\theta_v)}$ as $\Omega_x^{(h)}$ and $S_x^{(h)}$.

3.4 ConceFT

The main drawback of the TF reassignment is the spurious terms contributed by inaccurate IFD estimation resulting from correlations between noise and the window function. To achieve both localization and stability at the same time, a solution is the *multi-taper SST*, the average of multiple SST computed by a finite set of orthogonal windows [30]. Recently, Daubechies, Wang and Wu improved this idea and proposed the ConceFT method [6]. ConceFT emphasizes the use of *over-complete* windows rather than merely orthogonal windows, by assuming that a spurious term in a specific TF location just appears sparsely for the TF representations using different windows. Theoretical analysis proves that the ConceFT leads to sharper estimates of the instantaneous frequencies for signals that are corrupted by noise. Experiments also showed that ConceFT is useful in estimating the instantaneous frequencies with a fluctuated trajectory [6], a case similar to pitch scatter.

The over-complete window functions for ConceFT is generated from $\mathbf{z}$. This set of window functions $\mathbf{h} = [h_1, \ldots, h_N]$ with N windows is constructed as $h_1(t) = \nu_1(t)$ and $h_n(t) = \sum_{j=1}^{J} r_{nj}\nu_j(t)$ for $j = 2, \ldots, J$, $n = 2, \ldots, N$, and $\mathbf{r}_n = [r_{n1}, \ldots, r_{nJ}]$ is a random vector with unit norm. In ConceFT we need $J > 1$ and $N \geq J$. In contrast, a single window TF analysis requires $J = 1$ and $N = 1$, where $\mathbf{h} = h_1$. ConceFT is represented by

$$C_x^{(\mathbf{h})}(t,f) = \frac{1}{N}\sum_{n=1}^{N} S_x^{(h_n)}(t,f) \, . \quad (5)$$

We refer the reader with interest to [6] for a summary of the current progress in this direction.

For simplicity, we use $J = 2$ in this paper. Specifically, we use the Hamming window $0.54 + 0.46\cos(2\pi t/T)$ for ν_1, and the sine window $\sin(2\pi t/T)$ for ν_2. Obviously, ν_1 is orthogonal to ν_2, and the spectrum of ν_1 is concentrated to zero frequency whereas ν_2 has a zero at $f = 0$.

4. PROPOSED METHOD

4.1 Combining frequency and periodicity

An intuitive way to combine the frequency and periodicity representations is to multiply $V_x^{(h_n)}$ and $U_x^{(h_n)}$, after mapping the latter from time-quefrency into the TF domain:

$$W_x^{(h_n)}(t,f) = |V_x^{(h_n)}(t,f)|U_x^{(h_n)}\left(t, \frac{1}{f}\right) . \quad (6)$$

This approach has been mentioned in previous work on single pitch detection [19], where the F0 is determined simply by $f_0(t) = \arg\max_f W_x^{(h_n)}(t,f)$. Please note that here we only consider the co-occurrence of salience in the frequency and periodicity representations, and so far the constraints of harmonicity have not been included. A threshold on either $V_x^{(h_n)}$ or $U_x^{(h_n)}$ for removing unwanted terms is also critical to system performance. We will consider these issues below.

4.2 Constraints on harmonicity

To identify the location of the harmonic components in the STFT, one may use *pseudo-whitening*, a preprocessing step of estimating the spectrum envelope [15]. This method, however, is unreliable for a spectrum whose envelope is not smoothly varying or not supported by a large number of harmonics. This happens to be the case in choral music, since the singers tend to sing with more power in the F0 region rather than in the singer's format region [21].

To address this issue, we propose to assess whether a component at (t, f) is a sinusoidal component by using the IFD, instead of the spectral envelope. The rational is: as small $|\Omega_x^{(h)}|$ implies that the corresponding component in STFT is close to the true component, we can assume that $|\Omega_x^{(h)}|$ is bounded by a positive value θ_s around the harmonic components in the STFT. We have accordingly the *constraint on harmonicity in frequency representation*:

$$\mathfrak{N}_s := \{f : \max\left[|\Omega_x^{(h_n)}(t,(1:M_v)f)|\right] < \theta_s\} , \quad (7)$$

where we use $(t, (1 : M_v)f)$ as the shorthand for the set of points $\{(t, f), (t, 2f), \ldots, (t, M_v f)\}$. That is, for an F0 at (t_0, f_0), we require that $|\Omega_x^{(h)}(t_0, f_0)|$ is smaller than θ_s at (t_0, f_0) and its integer multiples. Similarly, for a fundamental period event at (t_0, q_0), the amplitude of $U_x^{(h_n)}$ should be above a threshold θ_c not only at (t_0, q_0) but also the multiples of its period. This leads to the *constraint on sub-harmonicity in periodicity representation*:

$$\mathfrak{N}_c := \{f : \min\left[U_x^{(h_n)}(t,(1:M_u)q)\right] > \theta_c\} . \quad (8)$$

With (3), (7) and (8), we have a more succinct feature representation $Y_x^{(h_n)}(t,f)$ by removing most of the non-harmonic-related terms in $W_x^{(h_n)}(t,f)$:

$$Y_x^{(h_n)}(t,f) = W_x^{(h_n)}(t,f)\Big|_{\mathfrak{N}}\,, \tag{9}$$

where $\mathfrak{N} := \mathfrak{N}_v \cap \mathfrak{N}_s \cap \mathfrak{N}_c$. Moreover, to enhance localization of this multipitch feature, we consider synchrosqueezing operation on $Y_x^{(h_n)}$ (instead of on $V_x^{(h_n)}$):

$$S_x^{(h_n)}(t,f) := \int_{\mathfrak{N}} Y_x^{(h_n)}(t,\eta)\delta\left(|f - \Omega_x^{(h_n)}(t,\eta)|\right) \mathrm{d}\eta\,, \tag{10}$$

Finally, by modifying (4), the multi-taper or ConceFT-based multipitch feature is obtained from averaging either $Y_x^{(h_n)}$ or $S_x^{(h_n)}$ over $n = 1, 2, \dots, N$, respectively:

$$B_x^{(\mathbf{h})}(t,f) = \frac{1}{N}\sum_{n=1}^{N} Y_x^{(h_n)}(t,f)\,, \tag{11}$$

$$C_x^{(\mathbf{h})}(t,f) = \frac{1}{N}\sum_{n=1}^{N} S_x^{(h_n)}(t,f)\,. \tag{12}$$

In the experiments we will compare the performance of the four features formulated in (9)–(12).

4.3 Implementation issues

There are several ways to sample the value of $U_x^{(h_n)}(t, 1/f)$ from $U_x^{(h_n)}(t,q)$, the simplest way being assigning every components in q to the bin closest to $f = 1/q$. However, the problem is there are usually insufficient low-quefrency points in $U_x^{(h_n)}(t,q)$ to represent the high-frequency part in $U_x^{(h_n)}(t,1/f)$. For example, there are only 34 points in $U_x^{(h_n)}$ to represent frequencies ranging from 1 kHz to 4 kHz for a signal sampled at 44.1 kHz. A simple yet effective solution is to linearly interpolate $U_x^{(h_n)}(t,q)$ into a fine grid with 0.4 Hz spacing,[2] and then have $U_x^{(h_n)}(t,1/f) = \sum_{j\in\mathfrak{P}(f)} U_x^{(h_n)}(t,q_j)$, where $\mathfrak{P}(f) := \{j : 1/(f+0.5/T) < q_j < 1/(f-0.5/T)\}$. A short-pass lifter described in [23] is also applied to $U_x^{(h_n)}(t,q)$.

Another issue of this mapping scheme is that the low-frequency part could be overemphasized since the summation is over a wide quefrency range and thereby cannot reveal true salience of pitch. This is not a critical issue if the dynamic information of each note is not required in transcription, but such dynamic information is needed here for late fusion. To address this, in our implementation we use a *binarized* version of $U_x^{(h_n)}$ to treat it as a *mask* in filtering out unwanted harmonic peaks in $V_x^{(h_n)}$, by setting $U_x^{(h_n)}(t,q) = 1$ if $U_x^{(h_n)}(t,q) > \theta_c$ and 0 otherwise.

5. EXPERIMENT

5.1 Datasets

To provide available source for the research on transcribing music with pitch scatter, we propose two new datasets, directly named Choir and Symphony here, which contain 5 excerpts of choral music from 3- to 8-part, and 5 excerpts of symphony, respectively. The length of the excerpts ranges from 18 to 108 seconds, totaling 5 minutes and 40 seconds. Information of each note events, including onset, offset, pitch name and instrument, are annotated by a professional pianist using the annotation methodology proposed in [24]. The audio, annotation, and other detailed information will be made public through a website.[3]

To test the generalizability of the proposed method, we also experiment on another two commonly used MPE datasets — Bach10 [8] and TRIOS [9]. The former contains ten quartets of four different instruments, while the latter consists of five pieces of fully synthesized music of piano and two other pitched instruments. The sampling rate of all the audio files is 44.1 kHz.

5.2 Numerical illustration

For all the proposed features, we empirically set the window length T to 0.14 second, hop size H to 0.01 second and use $\theta_c = 2 \times 10^{-4}$. For $Y_x^{(h_1)}$ and $B_x^{(\mathbf{h})}$, we set $N = 1$ and $\theta_s = 3$ bins (21.43 Hz); and for $S_x^{(h_1)}$ and $C_x^{(\mathbf{h})}$, we set $N = 10$ and $\theta_s = 1$ bin (7.14 Hz), a narrower bandwidth to enhance localization. The nonlinear scaling factor of g_ξ in (2) is set to 0.15.

Figure 1 shows in row the ground truth, $V_x^{(h_1)}(t,f)$, $U_x^{(h_1)}(t,1/f)$, $W_x^{(h_1)}(t,f)$ and $C_x^{(\mathbf{h})}(t,f)$ of four 5-second excerpts sampled respectively from the four datasets introduced in Section 5.1. Notice that $U_x^{(h_1)}(t,1/f)$ shown here is after thresholding of (7) to avoid negative values. We apply a power scale $(\cdot)^{0.1}$ in drawing the figures.

The second and third columns show that $V_x^{(h_1)}(t,f)$ or $U_x^{(h_1)}(t,1/f)$ alone is not a good multipitch feature since the former suffers from unwanted harmonic terms and the latter from "sub-harmonic" ones. However, most of these terms are removed in $W_x^{(h_1)}$, as seen in the fourth column. Furthermore, in the rightmost column, $C_x^{(\mathbf{h})}$ achieves very sharp components with few noises; it also nicely localizes the pitches. Notably, we see that the STFT components in Choir and Symphony spread widely and are much more fluctuated than those in Bach10 and TRIOS (due to severe pitch scatter), but the sharpness of the components in $C_x^{(\mathbf{h})}$ of the four samples are almost the same.

It can also be seen that there are still some challenging cases in the symphony due to its high complexity. For example, the short pitch activation above 1kHz in the ground truth (*pizzicato* of the 1st violin) remains unrecalled even in $C_x^{(\mathbf{h})}$. This is a subject of future work.

5.3 Piano roll output and post-processing

To obtain the piano roll output, all the features are processed first by a moving median filter with length 0.21 second to enhance smoothness, and then by a peak peaking process to pertain local maxima and discard other non-peak terms. Then, the MPE result, represented in piano

[2] Pilot studies show that finer grid spacing results in smoother feature representation but provides no significant empirical gains in MPE.

[3] https://sites.google.com/site/lisupage/research/new-methodology-of-building-polyphonic-datasets-for-amt

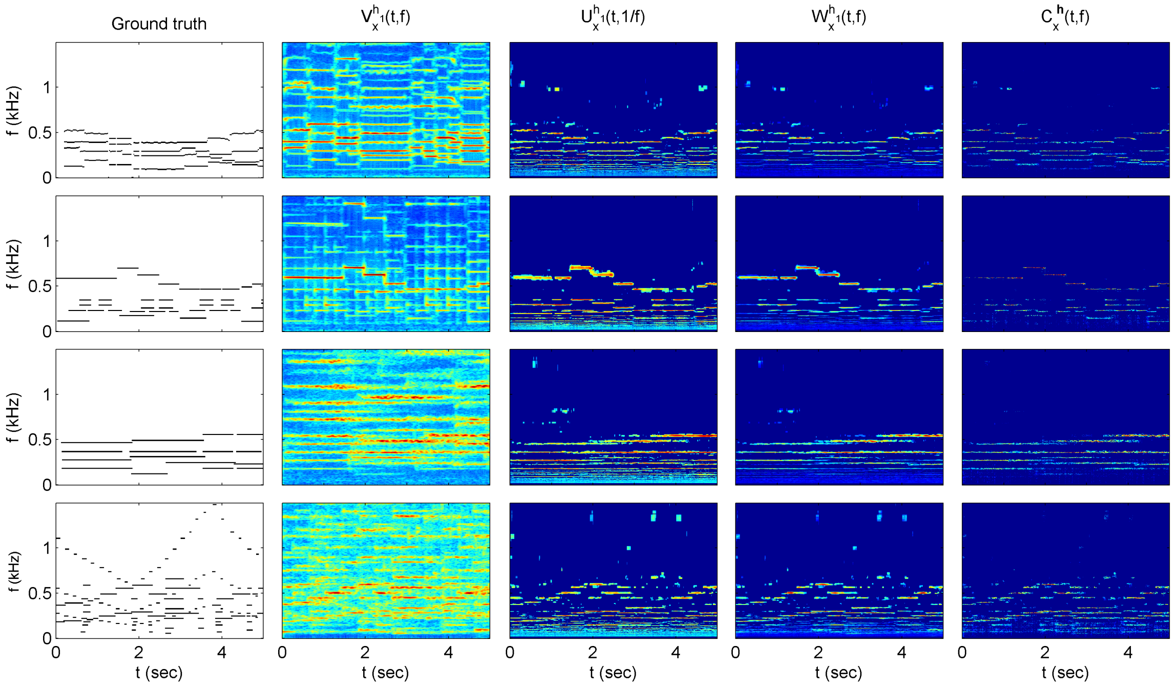

Figure 1. Illustration of the ground truth, $V_x^{(h_1)}(t,f)$, $U_x^{(h_1)}(t,1/f)$, $W_x^{(h_1)}(t,f)$ and $C_x^{(\mathbf{h})}(t,f)$ of four 5-second excerpts. $V_x^{(h_1)}$: STFT; $U_x^{(h_1)}$: generalized cepstrum; $W_x^{(h_1)}$: combination of STFT and generalized cepstrum; $C_x^{(\mathbf{h})}$: the proposed representation with ConceFT. First row: '01-AchGottundHerr.wav' (i.e. Bach's *Ach Gott und Herr, wie gros und schwer*, BWV 255) in quartet (violin, clarinet, saxophone and bassoon). Second row: Mozart's Trio in Eb major 'Kegelstatt', K.498, in piano, clarinet and viola. Third row: William Byrd, *Ave Verum Corpus*, in SATB choir. Fourth row: Tchaikovsky, Symphony No.6, Op.74 (*Pathetique*), Mov.2., in flute, oboe, clarinet, bassoon, horn, trumpet and strings.

roll $O(t,p)$, where $p = 13, 14, \ldots, 76$ is the piano roll number from `A1` (55 Hz) to `C7` (2,093 Hz), is obtained by $O(t,p) = \sum_{\mathfrak{F}(p)} X(t,f)$, where $\mathfrak{F}(p) = \{f : 440 \times 2^{(p-49-0.5)/12} \leq f < 440 \times 2^{(p-49+0.5)/12}\}$ (notice that A4 is the 49th key on the piano), and X denotes one of the feature representations described in (9)–(12).

The results of the proposed and baseline algorithms are refined by the same post-processing steps. The first step removes isolated pitches that are above `C5` and leave any other pitches in the affinity of 0.1 second by more than an octave. This is done because composers usually prefer smaller intervals within one octave in the high-pitch range. The second step is again a moving median filter with length also 0.21 second for smoothness.

5.4 Baselines and evaluation

Three baseline methods are considered. The first baseline is the *unsupervised* method proposed in our previous work [23], which also combines information of frequency, periodicity and harmony, but its harmonicity constraint was performed on the piano roll representation rather than directly on the TF representation. We did not use advanced TF analysis such as SST and ConceFT in the prior work [23]. Moreover, the method of computing the adaptive threshold of spectral representation is also different. We use the parameters suggested in [23], and the nonlinear scaling factor for generalized cepstrum is also set to 0.15.

The second one is an *unsupervised* method based on the Constrained Non-negative Matrix Factorization (C-NMF) algorithm proposed by Vincent *et al.* [29].[4] For the experiment on all datasets, we set $\beta = 0.5$ for computing the β-divergence and the value of $\vartheta = -32$ dB for thresholding the activation patterns in C-NMF.

The third baseline is a *supervised* method based on the shift-invariant PLCA proposed by Benetos *et al.* [3].[5] This approach uses labeled data to learn five templates for each pitch of each instrument and voice. The templates are learned from the single notes of the RWC instrument dataset [10], which contains various instruments as well as five vowels of human voice including soprano, alto, tenor and bass. We set the parameter for instrument activation $s_z = 1.3$, the parameter for source contribution $s_u = 1.1$ and the parameter for pitch shifting $s_h = 1.1$, all similar to [3]. To facilitate the comparison between the supervised and unsupervised approaches to MPE, we employ an *instrument-informed* setting that uses templates learned from different instruments for different music pieces, which might have given PLCA some advantages.

Moreover, to investigate cross-model behaviors of the algorithms, we experiment with a *late fusion* scheme that combines our method with PLCA. We first normalized

[4] `http://www.irisa.fr/metiss/members/evincent/multipitch_estimation.m`

[5] `https://code.soundsoftware.ac.uk/projects/amt_mssiplca_fast`

Table 1. Experiment result. $Y_x^{(h_1)}$, $B_x^{(\mathbf{h})}$, $S_x^{(h_1)}$ and $C_x^{(\mathbf{h})}$ are described in (9), (10), (11), (12), respectively. $Y_x^{(h_1)}$: single-window, without synchrosqueezing; $B_x^{(\mathbf{h})}$: single-window, with synchrosqueezing; $S_x^{(h_1)}$: overcomplete-window, without synchrosqueezing; $C_x^{(\mathbf{h})}$: overcomplete-window, with synchrosqueezing

Dataset	Proposed				Baseline			Proposed+PLCA (Late fusion)			
	$Y_x^{(h_1)}$	$B_x^{(\mathbf{h})}$	$S_x^{(h_1)}$	$C_x^{(\mathbf{h})}$	[23]	C-NMF	PLCA	$Y_x^{(h_1)}$	$B_x^{(\mathbf{h})}$	$S_x^{(h_1)}$	$C_x^{(\mathbf{h})}$
Bach10	**83.96**	83.29	79.18	82.13	81.97	79.78	70.57	82.39	82.14	**82.69**	82.04
TRIOS	**66.30**	**66.30**	60.23	66.26	64.09	59.40	64.93	**71.10**	70.79	70.35	70.57
Choir	57.44	59.71	51.29	**61.18**	44.98	45.62	61.07	64.36	64.88	64.06	**65.31**
Symphony	49.14	**50.44**	46.95	50.33	48.82	40.34	47.04	51.73	**52.46**	51.02	51.86

every frame of the piano roll output by its l_2 norm, then combine them through linear superposition, and finally discard the terms which are smaller than a threshold ϵ:

$$\bar{O}_{fusion} = (\alpha \bar{O}_{PLCA} + (1-\alpha)\bar{O}_{proposed} - \epsilon)_+ \,, \quad (13)$$

where $\bar{O}$ is the normalized piano roll output, $\alpha \in [0,1]$ controls the relative weights of the two methods, and $(x)_+ = \max(0,x)$ is a hard thresholding function.

We evaluate the accuracy of MPE using the micro-average frame-level F-score, which counts the number of true positives, false positives and false negatives over all the frames within a dataset and then calculates the harmonic mean of the precision and recall rates.

6. RESULT

Table 1 lists the F-scores on the four datasets using the proposed methods using features (9)–(12), three baseline methods and late fusion of proposed features with PLCA. The main findings are reported below.

First, the four proposed methods outperform the three baselines in general. Although the method [23] adopted the same approach of combining frequency and periodicity information as the proposed methods do, it was reported to be sensitive to γ, the nonlinear scaling factor in computing the generalized cepstrum. Besides, the method [23] cannot benefit from the constraint on harmonics, especially in the case of Choir, as the estimation of the spectral threshold and noise terms is rather inaccurate without IFD information. C-NMF also performs poorly for such challenging musical signals. In comparison to the two unsupervised baselines, PLCA performs fairly well in Choir and Symphony, perhaps because it uses supervised templates and allows template shifting in pitch [3].

Second, among the proposed methods, we find the multi-taper ones $B_x^{(\mathbf{h})}$ and $C_x^{(\mathbf{h})}$ do perform better than those use only one window, i.e., $Y_x^{(h_1)}$ and $S_x^{(h_1)}$, for datasets with pitch scatter (i.e., Choir and Symphony). However, for Bach10 and TRIOS, where most of the pitches are played with only one instrument, multi-tapering and SST do not give better performance, as there is no need to reduce the variance of a spectral peak of a single source.

Moreover, the method using single-window SST $S_x^{(h_1)}$ performs the worst among the proposed methods, as its nonlinearity usually gives rise to unwanted speckle terms, a major known drawback of SST [6]. This problem is nicely solved by ConceFT, as we can see that $C_x^{(\mathbf{h})}$ outperforms $S_x^{(h_1)}$ by around 10% in Choir and 3% in Symphony. This suggests the need to introduce multi-tapering to stabilize the estimation, when feature localization is an important requirement for the system.

Finally, a grid search over the four datasets shows that the optimal result of late fusion is achieved by setting $\alpha = 0.05$ (i.e. emphasizing the proposed method) and $\epsilon = 3 \times 10^{-5}$. Combining PLCA and $C_x^{(\mathbf{h})}$ achieves 65.31% for Choir, which amounts to more than 4% improvement over $C_x^{(\mathbf{h})}$. Combining PLCA and $Y_x^{(\mathbf{h})}$ further improves the F-score by 4.8%. However, less improvement is found in Bach10 and Symphony, possibly because that the former already has limited space for improvement and the latter is rather complicated such that some information cannot be well captured by both method. Although the weighting on PLCA is small, PLCA does capture some critical information missed by the proposed methods. This suggests the importance of fusing different MPE models, in particular unsupervised and supervised ones.

7. CONCLUSION

To improve the robustness of MPE algorithms in dealing with diverse music signals, we introduce and incorporate novel TF analysis tools including SST and ConceFT to enhance the stability and localization of multipitch features. The proposed unsupervised methods also measure pitch saliency by jointing considering frequency, periodicity and harmonicity. Result on two newly created datasets of choral and symphony music demonstrates the superiority of the proposed methods for MPE in music signals featuring pitch scatter. Slightly better result can be obtained by combining our methods and the supervised method PLCA.

8. ACKNOWLEDGEMENT

We thank Patricia Hsu for data labeling and Dr. Hau-Tieng Wu for fruitful discussions. Dr. Li Su was supported by the postdoctoral fellowship of Academia Sinica.

9. REFERENCES

[1] Y. Attabi, M. J. Alam, P. Dumouchel, P. Kenny, and D. O'Shaughnessy. Multiple windowed spectral features for emotion recognition. In *Proc. ICASSP*, pages 7527–7531, 2013.

[2] F. Auger and P. Flandrin. Improving the readability of time-frequency and time-scale representations by the reassignment method. *IEEE Trans. Signal Process.*, 43(5):1068 –1089, may 1995.

[3] E. Benetos and S. Dixon. A shift-invariant latent variable model for automatic music transcription. *Computer Music Journal*, 36(4):81–94, 2012.

[4] E. Benetos, S. Dixon, D. Giannoulis, H. Kirchhoff, and A. Klapuri. Automatic music transcription: challenges and future directions. *J. Intelligent Information Systems*, 41(3):407–434, 2013.

[5] I. Daubechies, J. Lu, and H.-T. Wu. Synchrosqueezed wavelet transforms: An empirical mode decomposition-like tool. *Appl. Numer. Harmon. Anal.*, 30:243–261, 2011.

[6] I. Daubechies, Y. Wang, and H.-T. Wu. Conceft: Concentration of frequency and time via a multitapered synchrosqueezed transform. *Philosophical Transactions A*, 374(2065), 2016.

[7] Z. Duan, J. Han, and B. Pardo. Multi-pitch streaming of harmonic sound mixtures. *IEEE/ACM Trans. Audio, Speech, Language Process.*, 22(1):138–150, 2014.

[8] Z. Duan, B. Pardo, and C. Zhang. Multiple fundamental frequency estimation by modeling spectral peaks and non-peak regions. *IEEE/ACM Trans. Audio, Speech, Language Process.*, 18(8):2121–2133, 2010.

[9] J. Fritsch. High quality musical audio source separation. Master's thesis, Queen Mary Centre for Digital Music, 2012.

[10] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka. RWC music database: Music genre database and musical instrument sound database. In *Proc. ISMIR*, pages 229–230, 2003.

[11] S. W. Hainsworth. *Techniques for the Automated Analysis of Musical Audio*. PhD thesis, University of Cambridge, UK, 2003.

[12] D. Kahlin and S. Ternström. The chorus effect revisited-experiments in frequency-domain analysis and simulation of ensemble sounds. In *Proc. EUROMICRO Conference*, volume 2, pages 75–80, 1999.

[13] M. Khadkevich and M. Omologo. Time-frequency reassigned features for automatic chord recognition. In *Proc. ICASSP*, pages 181–184, 2011.

[14] T. Kinnunen, R. Saeidi, F. Sedlák, K. A. Lee, J. Sandberg, M. Hansson-Sandsten, and H. Li. Low-variance multitaper mfcc features: a case study in robust speaker verification. *IEEE Trans. Audio, Speech, Language Proc.*, 20(7):1990–2001, 2012.

[15] Anssi P. Klapuri. Multiple fundamental frequency estimation based on harmonicity and spectral smoothness. *IEEE Speech Audio Process.*, 11(6):804–816, 2003.

[16] T. Kobayashi and S. Imai. Spectral analysis using generalized cepstrum. *IEEE Trans. Acoust., Speech, Signal Process*, 32(5):1087–1089, 1984.

[17] W. Lottekmoser and Fr. J. Meyer. Frequenzmessungen an gesungenen akkorden. *Acta Acustica united with Acustica*, 10(3):181–184, 1960.

[18] J. Pätynen, S. Tervo, and T. Lokki. Simulation of the violin section sound based on the analysis of orchestra performance. In *Proc. WASPAA*, pages 173–176, 2011.

[19] G. Peeters. Music pitch representation by periodicity measures based on combined temporal and spectral representations. In *Proc. ICASSP*, 2006.

[20] G. Peeters and X. Rodet. Sinola: A new analysis/synthesis method using spectrum peak shape distortion, phase and reassigned spectrum. In *Proc. ICMC*, 1999.

[21] T. D. Rossing, J. Sundberg, and S. Ternström. Acoustic comparison of voice use in solo and choir singing. *J. Acoust. Soc. Am.*, 79(6):1975–1981, 1986.

[22] J. Salamon and E. Gòmez. Melody extraction from polyphonic music signals using pitch contour characteristics. *IEEE Trans. Audio, Speech, and Language Processing*, 20(6):1759–1770, 2012.

[23] L. Su and Y.-H. Yang. Combining spectral and temporal representations for multipitch estimation of polyphonic music. *IEEE/ACM Trans. Audio, Speech, Language Process.*, 23(10):1600–1612, 2015.

[24] L. Su and Y.-H. Yang. Escaping from the abyss of manual annotation: New methodology of building polyphonic datasets for automatic music transcription. In *Int. Symposium on Computer Music Multidisciplinary Research*, 2015.

[25] S. Ternström. Perceptual evaluations of voice scatter in unison choir sounds. *Journal of Voice*, 7(2):129–135, 1993.

[26] S. Ternström and J. Sundberg. Intonation precision of choir singers. *J. Acoust. Soc. Am.*, 84(1):59–69, 1988.

[27] D. J. Thomson. Spectrum estimation and harmonic analysis. *Proc. IEEE*, 70:1055–1096, 1982.

[28] T. Tolonen and M. Karjalainen. A computationally efficient multipitch analysis model. *IEEE Speech Audio Process.*, 8(6):708–716, 2000.

[29] E. Vincent, N. Bertin, and R. Badeau. Adaptive harmonic spectral decomposition for multiple pitch estimation. *IEEE Trans. Audio, Speech, Language Process.*, 18(3):528–537, 2010.

[30] J. Xiao and P. Flandrin. Multitaper time-frequency reassignment for nonstationary spectrum estimation and chirp enhancement. *IEEE Trans. Signal Process.*, 55:2851–2860, 2007.