Resolving Octave Ambiguities: A Cross-dataset Investigation

Li Su, Li-Fan Yu and Yi-Hsuan Yang
CITI, Academia Sinica, Taipei, Taiwan
lisu@citi.sinica.edu.tw

Hsin-Yu Lai
Dep. EE, National Taiwan University, Taipei, Taiwan
b99901087@ntu.edu.tw

ABSTRACT

Octave dual-tone is one of the most difficult patterns to identify in multipitch estimation (MPE), as the spectrum of the upper note is almost masked by the lower one. This paper investigates the potential for a supervised binary classification framework to address this issue, and whether such a framework is adequate for diverse real-world signals. To this end, a new dataset comprising of 3,493 real single notes and octaves recorded by two pianists and guitarists are constructed to facilitate an in-depth analysis of this problem. The dataset is available to the research community. Performance of synthetic and real-world octave dual-tones using various spectral-, cepstral- and phase-based features are studied systematically. Our experiments show that the instantaneous frequency deviation (IFD) represents the most reliable feature representation in discriminating octave dual-tones from single notes. Based on this new dataset and the RWC dataset, we present a series of experiments to offer insights into the performance difference between synthetic and real octaves, piano and guitar notes, as well as studio recordings and home recordings. As the proposed method holds the promise of resolving octave dual-tone, we envision that it can be an important module of a multipitch estimation system.

1. INTRODUCTION

Consonant intervals are a very common part of music in our everyday lives. With frequencies of simple integer ratio, consonances have more coincident partials (overlapping harmonics) than dissonances. From a signal processing perspective, this makes it more difficult to specify all the pitch occurrences from the spectrum [1]. Accordingly, resolving overlapping harmonics poses a great challenge in multipitch estimation (MPE) tasks for multi-source music signals [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12]. Among the basic music intervals, the most challenging one might be the *octave dual-tone* (perfect 8th intervals), as the fundamental frequency (f0) of the upper note is exactly twice of the lower note. For instance, the spectrum of an octave dual-tone `A4+A5` would be much similar to that of a single note `A4`, as Fig. 1 shows. While the lower note `A4` can usually be identified by MPE algorithms, identifying the presence of the co-occurring `A5` is challenging. We refer to this confusion between a single note and its corresponding octave dual-tone (usually with its upper octave) as the *octave detection error* (instead of the confusion between a single note and the other single note an octave higher or lower to it).

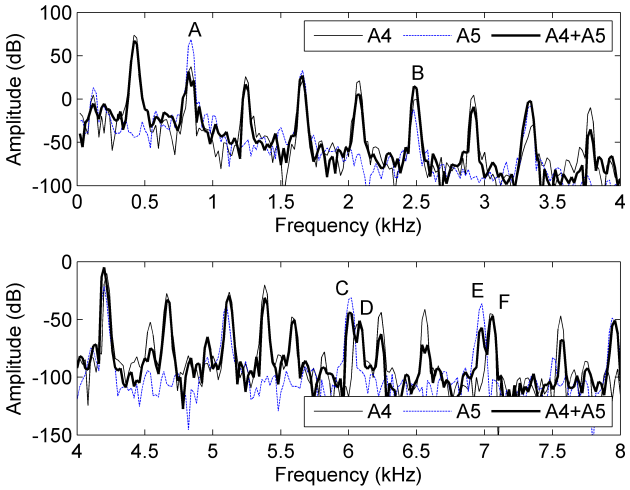

Figure 1. Spectra of a lower single note (`A4`, f0=440 Hz), an upper single note (`A5`) and an octave dual-tone (`A4+A5`). All three samples are played by the same musician and with the same piano. Upper: 0 to 4 kHz; lower: 4 to 8 kHz.

To circumvent this issue, a number of solutions have been proposed, encompassing those based on smoothness approximation [1, 3, 2], linear or non-linear interpolation from non-overlapping harmonics [5, 7, 6], and probabilistic models for spectral envelope (SE) modeling [3, 4], amongst others. Recently, matrix decomposition-based MPE algorithms have also been proposed, such as non-negative matrix factorization (NMF) or sparse coding (SC) [7, 3, 8, 9]. However, the performance of many prior arts is limited by the linear superposition assumption of the magnitude spectra and the requirement of obtaining non-overlapping harmonics as a reference. Moreover, oftentimes the evaluation of existing methods is performed on synthesized data alone, whose acoustic properties can be largely different from real-world signals.

There might not be feasible non-overlapping peaks for the masked, higher notes (e.g., `A5`) in real-world signals. As Fig. 1 exemplifies, although the second harmonic of `A4` and the f0 of `A5` coincide at position 'A', the amplitude of `A4+A5` at 'A' is smaller than `A4` and `A5`, possibly due to destructive interference or nonlinear effects. At po-

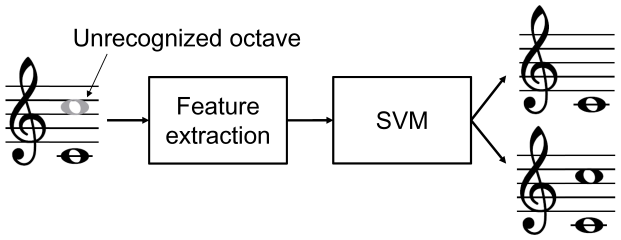

Figure 2. The scheme for octave classification — discriminating a single note from the corresponding octave dual-tone.

sition 'B', the peak amplitude of A4+A5 is larger than the sum of A4 and A5, which might also be a nonlinear phenomenon. Due to inharmonicity [13, 14] and longitudinal string modes [15, 14], the harmonic series in higher frequency start to deviate from the integer multiples and get irregular. Peak pairs 'C, D' and 'E, F' of A4+A5 are concerned by such effects. The additivity assumption of the magnitude spectra turns out to be too strong, in view of the interference with phase or nonlinear effects accompanying an octave dual-tone.

Since most MPE algorithms have stronger ability in recalling the lower note of an octave dual-tone, the octave ambiguity problem may be posted in an alternative way: *how to distinguish between a single note and its corresponding octave dual-tone*? As a preliminary attempt, we formulate this as a binary classification problem and experiment with a variety of spectral-, cepstral- and phase- derived features in the context of such an *octave classification* problem. In addition, a new real-world octave dataset of piano and guitar sounds with various dynamics and playing styles is collected for a systematic evaluation. Both real and synthesized signals are considered on a single-note level. Although we understand that it is important to evaluate the effect of the proposed octave classifier directly in the context of MPE, we opt for leaving this as a future work and concentrate on the issue of octave classification in this paper.

We carefully design six different experiments to gain insights from this study. In what follows, we describe the proposed system and evaluation framework in Sections 2 and 3, the experiment results in Section 4, and we draw conclusions in Section 5.

2. APPROACH

Octave classification can be understood either in the context of timbre classification or as a special case of classification based MPE [16]. The basic idea is to learn a audio-based binary classifier that discriminates a single note and its octave dual-note counterpart. As depicted in Figure 2, the system mainly contains a feature extraction step and a classification step. We use linear support-vector machine (SVM) [17] for classifier training here.

In particular, we assume that in a practical MPE system octave classification can be used as a refining and calibration procedure after a set of note candidates has been selected [5, 6]. For example, after an A4 has been detected, octave classification further verifies the presence of the upper note A5 in the signal. Therefore, in the proposed scheme we assume that the f0 of the lower note is known *a priori*, and train a binary classifier for each pair of single note (with a specific f0) and the corresponding octave dual-tone.

To identify audio features relevant to the proposed task, we evaluate the following features. The most fundamental one is the spectrogram (SG) $\|\mathrm{M}_x^h\|^2$, which is the squared magnitude of the short-time Fourier transform (STFT):

$$\mathrm{S}_x^h\left(t,\omega\right) = \mathrm{M}_x^h\left(t,\omega\right) e^{j\Phi_x^h(t,\omega)}\,, \tag{1}$$

where x and h denote the time domain signal and the window function, respectively. We use the dB-scaled discrete spectrogram $\mathrm{X}\left(n,k\right) \in \mathbb{R}^{N\times K}$, where n is the time index, k is the frequency index, N is the number of short-time frames, and K is the number of frequency bin. In addition, we consider two variants of spectral-based features — the frame-level *autocorrelation* function of the magnitude spectrum (SG-ACF):

$$\mathrm{ACF}\left(n,k\right) = \sum_{l=0}^{K-l} \mathrm{X}\left(n,l\right)\mathrm{X}\left(n,k+l\right)\,, \tag{2}$$

and the *second harmonic difference* (SG-D):

$$\mathrm{D}\left(n,k\right) = \mathrm{X}\left(n,k\right) - \mathrm{X}\left(n,2k\right)\,, \tag{3}$$

where $1 \le k \le \lfloor K/2 \rfloor$. The latter is a new feature we design for this task; it tries to capture the behaviors of the components which frequencies differ by two. We also consider the Mel-frequency cepstral coefficients (MFCC), a classic feature in instrument timbre classification [5]. The last feature of interest is the *instantaneous frequency deviation* (IFD), defined as the temporal derivative of the phase angle Φ_x^h of STFT:

$$\mathrm{IFD}_x^h\left(t,\omega\right) = \frac{\partial\Phi_x^h\left(t,\omega\right)}{\partial t} = \mathrm{Im}\left(\frac{\mathrm{S}_x^{\mathcal{D}h}\left(t,\omega\right)}{\mathrm{S}_x^h\left(t,\omega\right)}\right) \tag{4}$$

where $\mathcal{D}h\left(t\right) = h'\left(t\right)$. In discrete implementation, IFD in Eq. (4) indicates the deviation of frequency component from the discrete bin to the actual value. Therefore, IFD provides a good calibration under either low spectral resolution or high spectral leakage. Please refer to [18, 19] for more discussions and detailed derivation of IFD.

While ACF and IFD have been regularly applied in MIR problems such as onset detection and tempo estimation, the SG-D feature is relatively novel and can be viewed as a feature that is designed for octave classification. On the other hand, in our empirical evaluation we consider well-known features such as ACF and MFCC as the task of supervised octave classification has not been well studied before.

3. DATASETS

In view of the possible difference between real-world octaves and synthesized ones, a rich set of six different datasets featuring different characteristics is employed in this study. As Table 1 shows, we take two different piano subsets

Table 2. Experimental settings and the accuracy for each feature, with the best two highlighted for each experiment

	EXP #1	EXP #2	EXP #3	EXP #4	EXP #5	EXP #6
Training set	Piano 1 (Syn)	Piano 3 (Syn)	Piano 3 (Real)	Piano 3 (Syn)	Piano 1 (Syn)	Guitar 1 (Real)
Test set	Piano 2 (Syn)	Piano 4 (Syn)	Piano 4 (Real)	Piano 4 (Real)	Piano 4 (Real)	Guitar 2 (Real)
SG	**0.740**	0.595	0.612	0.566	0.450	0.722
SG-ACF	0.630	0.428	0.633	0.429	0.453	0.687
SG-D	**0.756**	**0.650**	**0.663**	0.629	**0.625**	0.719
MFCC	0.686	0.637	0.645	**0.636**	0.561	**0.768**
IFD	0.684	**0.686**	**0.658**	**0.651**	**0.645**	**0.832**
SONIC [21]	0.693	0.593	0.578	0.578	0.578	—

Table 1. Description and number of notes of the six datasets

Name	Record	#single notes	#synthetic octaves	#real octaves
Piano 1	studio (RWC [20])	792	684	-
Piano 2		792	684	-
Piano 3	home	525	453	684
Piano 4		788	680	683
Guitar 1	studio	264	-	177
Guitar 2	home	249	-	122

with different brands, referred to as 'Piano 1 and 2,' from the RWC instrument dataset [20]. RWC contains *studio-recorded* single notes of over 50 instruments with various brands, playing techniques, pitches and dynamics.

For other four remaining datasets we use in this study, four musicians with different professional levels are paid, and asked to play single notes and octave dual-tones in all possible pitch ranges of piano and guitar, with different musical dynamic levels (*forte*, *mezzo forte* and *piano*) and playing techniques (*normal*, *pedal* and *staccato* for piano and *normal* for guitar). Except for 'Guitar 1,' all these sound samples are home-recorded, meaning more noises, reverberation and other possible defects prevalent in real-world recordings. Because we are also interested in comparing real octaves and synthetic ones, we also synthesize octave dual-tones by additively mixing the single notes of the four piano datasets.

For piano, the pitches of single-note data range from `A0` to `C8`, while octave data range from `A0+A1` to `C7+C8`. The pitch range of guitar is narrower, ranging from `E2` to `D6` for single notes, with some duplicated pitches played on different strings, and `E2+E3` to `D5+D6` for octave data. Not surprisingly, we have less guitar data than piano data.

In the course of this study, we have compiled a new dataset of 1,313 piano single notes, 1,367 piano octaves, 513 guitar single notes, and 299 guitar octaves contributed by four musicians. For reproduciblity and for calling more attention to this problem, the audio files of the sound recordings are publicly available online. [1]

[1] `http://mac.citi.sinica.edu.tw/PitchOctave`

4. EXPERIMENTS

The music signals are sampled at 44.1 kHz. As described in Section 2, five features are adopted: spectrogram (SG), SG-ACF, SG-D, MFCC and IFD. We use a Hanning window h of 2,048 samples and 20% hopping for STFT, normalize all the considered frame-level features by l_2-norm, and aggregate frame-level features to song-level ones by taking sum-pooling across time, before using the song-level features as input to SVM. We use the linear SVM implemented in the LIBLINEAR library [17]. Moreover, we optimize the SVM parameter C from a search range of $\{2^{-10}, 2^{-9}, \ldots, 2^{10}\}$ through a validation set and then fix the value to 2^{-2}. The features are processed by a sum-to-one normalization after aggregation. As pitch information is assumed to be known, we train a binary octave classifier for each pitch.

We conduct six experiments in a cross-dataset scheme by using different datasets for training and testing. The baseline accuracies of all experiments are all 0.5. The result is summarized in Table 2 with details discussed below.

4.1 Average behaviors

4.1.1 Comparison of features

By comparing the features across the experiments, we see that SG-D and IFD are the most reliable features for octave classification; the former obtains the best accuracy for EXP #1 and #3, while the latter represents the best one for the other four. SG-D consistently outperforms SG and SG-ACF for most cases, which is expected as it is designed specifically for this task. On the other hand, IFD represents a competitive alternative for both piano and guitar sounds, showing that phase-based feature can also be effective. Lastly, the performance of MFCC appears to be moderate.

4.1.2 Studio recording versus home recording

We can study the effect of recording environment by comparing the result for EXP #1 and #3. Table 2 shows severe accuracy degradation for SG and SG-D ($> 10\%$), but not severely so for IDF and MFCC ($< 5\%$). This shows the performance of spectral features is more sensitive to noises and other defects in a room environment. It also shows that only using studio recordings in an evaluation could lead to overly optimistic result. This justifies the use of the last four datasets.

4.1.3 Synthetic octaves versus real octaves

The effect of synthetic versus real octaves can be studied by comparing EXP #2 and #3, both of which use Piano 3 for training and Piano 4 for testing. We see that the performance difference between the two experiments seems to be mild ($< 3\%$), except for SG-ACF ($> 20\%$). EXP #4 and #5 further evaluate the result of using synthetic data to predict real-world data. Comparing these results with that of EXP #3 (i.e., using real-world data to predict real-world data), we observe that the accuracies of SG and SG-ACF drop severely (5–20%), whereas the result of SG-D and IFD does not alter much ($< 4\%$ and $< 2\%$, respectively). This result demonstrates the effect of dataset mismatch (in terms of synthetic versus real octaves) and shows that IFD stands for the best choice if we are only given synthetic data for training the octave classifier.

4.1.4 Piano versus guitar

By comparing EXP #3 and #6, octave classification is found easier for guitar than for piano, possibly due to less nonlinear effects as a result of narrower pitch and dynamic ranges for guitar. For example, the accuracy of IFD attains 0.658 in EXP #3 but 0.832 for EXP #6. Interestingly, we also find that SG-D does not work so well for guitar as for piano, possibly because the inharmonicity of guitar string is generally smaller than that of piano string, as reported before [14, 22].

4.1.5 Comparison with an MPE algorithm — SONIC

Finally, we compare the result of the proposed classification-based approach against 'SONIC,' a state-of-the-art, publicly-available MPE algorithm specialized in piano music transcription [21]. Although it is virtually impossible to compare the performance absolutely fair, a prediction of SONIC is considered correct if it returns an octave combination (regardless of the number of notes returned) for an octave dual-note, and if it returns only one note (regardless of its pitch estimate) for an single note. As SONIC is designed for piano, we do not use it for EXP #6. As Table 2 shows, the performance of SONIC is generally inferior, especially when SG-D or IFD is used. The performance gap is more pronounced for real octaves (i.e., EXP #3–#5).

4.2 Pitchwise behaviors

Figures 3(a) and 3(b) show the average accuracy for each pitch (in MIDI number) of piano (EXP #3) and guitar (EXP #6), respectively. Here the pitch number denotes the actual pitch of a single note or the lower pitch of an octave dual-tone. The pitch information is provided in the upper and lower margins of the figures, with the range corresponding to the available notes for piano and guitar, respectively. We see non-smooth trends going from low to high pitches, but generally a bell-shape is observed for both instruments. This can be due to the insufficient frequency resolution for low-pitched signal and limited amount of harmonic information for high-pitched signal. The performance variation seems to be larger for guitar, which can be due to the nature of the instrument or the less amount of available data.

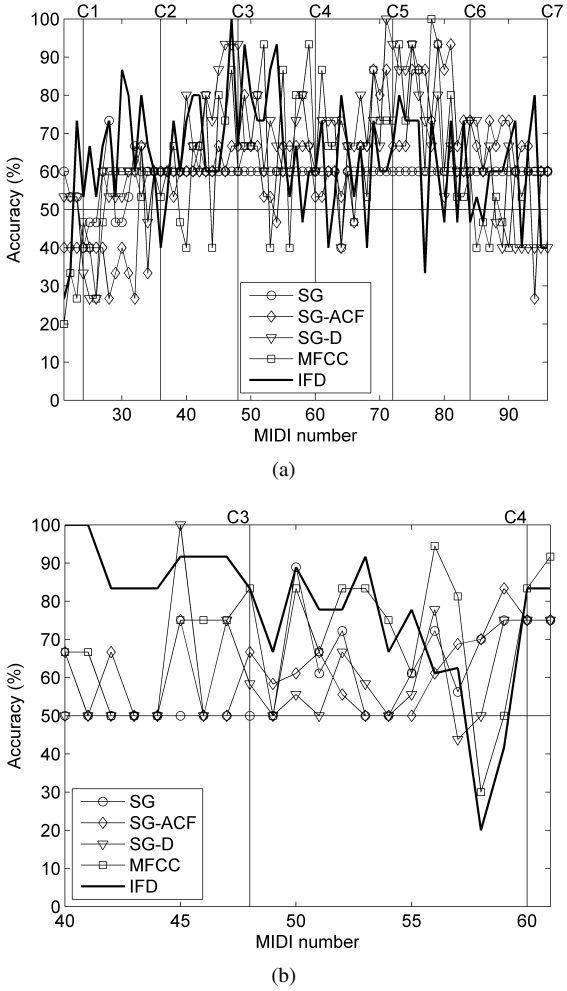

Figure 3. Pitchwise accuracies of (a) cross-dataset octave classification for piano (EXP #3) and (b) for guitar (EXP #6).

5. CONCLUSIONS

In this paper, we have presented a novel classification-based system for distinguishing between single notes and the corresponding octave dual-tones. A systematic performance study that investigates different audio features and test cases in a cross-dataset setting validates the effectiveness of the proposed approach for either synthetic or real-world signals. The best accuracy of octave classification for piano ranges from 64.5% to 75.6% across experiment settings, whereas the accuracy for guitar attains 83.2%. Relatively more reliable estimate is observed by using either the second harmonic difference of spectrogram or the instantaneous frequency deviation as features. For future work, we are interested in incorporating the idea to MPE, perhaps using more advanced features that are designed for this task. A study on other intervals (e.g., twelfths and double octaves) is also underway.

Acknowledgments

This work was supported by the Academia Sinica Career Development Award.

6. REFERENCES

[1] A. P. Klapuri, "Multiple fundamental frequency estimation based on harmonicity and spectral smoothness," *IEEE Speech Audio Process.*, vol. 11, no. 6, pp. 804–816, 2003.

[2] K. Dressler, "Pitch estimation by the pair-wise evaluation of spectral peaks," in *Audio Engineering Society Conference: 42nd International Conference: Semantic Audio*. Audio Engineering Society, 2011.

[3] E. Vincent, N. Bertin, and R. Badeau, "Adaptive harmonic spectral decomposition for multiple pitch estimation," *IEEE Audio, Speech, Language Process*, vol. 18, no. 3, pp. 528–537, 2010.

[4] M. Bay, A. F. Ehmann, J. W. Beauchamp, P. Smaragdis, and J. S. Downie, "Second fiddle is important too: Pitch tracking individual voices in polyphonic music." in *Proc. ISMIR*, 2012, pp. 319–324.

[5] E. Benetos and S. Dixon, "Joint multi-pitch detection using harmonic envelope estimation for polyphonic music transcription," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 6, pp. 1111–1123, 2011.

[6] A. Pertusa and J. M. Iñesta, "Efficient methods for joint estimation of multiple fundamental frequencies in music signals," *EURASIP J. Advances in Signal Processing*, vol. 2012, no. 1, pp. 1–13, 2012.

[7] N. Bertin, R. Badeau, and E. Vincent, "Enforcing harmonicity and smoothness in bayesian non-negative matrix factorization applied to polyphonic music transcription," *IEEE Audio, Speech, Language Process*, vol. 18, no. 3, pp. 538–549, 2010.

[8] J. J. Carabias-Orti, T. Virtanen, P. Vera-Candeas, N. Ruiz-Reyes, and F. J. Cañadas-Quesada, "Musical instrument sound multi-excitation model for non-negative spectrogram factorization," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 6, pp. 1144–1158, 2011.

[9] N. Boulanger-Lewandowski, Y. Bengio, and P. Vincent, "Discriminative non-negative matrix factorization for multiple pitch estimation." in *Proc. ISMIR*, 2012, pp. 205–210.

[10] C. Yeh, "Multiple fundamental frequency estimation of polyphonic recordings," Ph.D. dissertation, Université Paris VI - Pierre et Marie Curie, France, 2008.

[11] V. Emiya, R. Badeau, and B. David, "Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle," *IEEE Audio, Speech, Language Process*, vol. 18, no. 6, pp. 1643–1654, 2010.

[12] E. Benetos, S. Dixon, D. Giannoulis, H. Kirchhoff, and A. Klapuri, "Automatic music transcription: challenges and future directions," *J. Intelligent Information Systems*, pp. 1–28, 2013.

[13] H. Järveläinen, V. Välimäki, and M. Karjalainen, "Audibility of inharmonicity in string instrument sounds, and implications to digital sound synthesis," in *Proc. Int. Computer Music Conf.*, 1999, pp. 359–362.

[14] N. H. Fletcher and T. D. Rossing, *The physics of musical instruments*, 1998.

[15] B. Bank and L. Sujbert, "Modeling the longitudinal vibration of piano strings," in *Proc. Stockholm Music Acoustics Conf*, 2003, pp. 143–146.

[16] G. E. Poliner and D. P. W. Ellis, "A discriminative model for polyphonic piano transcription," *EURASIP J. Advances in Signal Processing*, vol. 2007.

[17] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A library for large linear classification," *J. Machine Learning Research*, vol. 9, pp. 1871–1874, 2008.

[18] F. Auger and P. Flandrin, "Improving the readability of time-frequency and time-scale representations by the method of reassigment," *IEEE Trans. Signal Processing*, vol. 43, no. 5, pp. 1068–1089, 1995.

[19] S. W. Hainsworth and M. D. Macleod, "Time frequency reassignment: A review and analysis," Cambridge University Engineering Department and Qinetiq, Tech. Rep., 2003.

[20] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database: Music genre database and musical instrument sound database." in *Proc. ISMIR*, 2003, pp. 229–230, http://staff.aist.go.jp/m.goto/RWC-MDB/rwc-mdb-i.html.

[21] M. Marolt, "A connectionist approach to automatic transcription of polyphonic piano music," *IEEE Trans. Multimedia*, vol. 6, no. 3, pp. 439–449, 2004.

[22] H. Järveläinen and M. Karjalainen, "Importance of inharmonicity in the acoustic guitar," in *Proc. Int. Comp. Music Conf.*, 2005, pp. 363–366.