A Maximum Likelihood Approach to Multiple Fundamental Frequency Estimation From the Amplitude Spectrum Peaks

Zhiyao Duan
Department of Automation
Tsinghua University
Beijing, China 100084
duanzhiyao00@mails.tsinghua.edu.cn

Changshui Zhang
Department of Automation
Tsinghua University
Beijing, China 100084
zcs@mail.tsinghua.edu.cn

Abstract

This paper presents a Maximum Likelihood approach to multiple fundamental frequency (F0) estimation in each frame of music signals in the frequency domain. The frequencies and amplitudes of the spectral peaks are viewed as observations, and the F0s are viewed as parameters to be estimated. The proposed method considers the potential errors in the peak detection algorithm and treats each peak as "true" and "false" separately. The likelihood models of the "true" and "false" peaks are learned from the monophonic training data, with the assumption that the statistics of the peaks in monophonic and polyphonic signals are similar. The proposed method also incorporates a rectified Bayesian Information Criteria (BIC) to estimate the number of the parameters, i.e. the polyphony. Evaluation is held on randomly mixed chords, which are generated from the previously unseen monophonic tones. Experimental results show the feasibility of this method.

1 Introduction

Multiple fundamental frequency (F0) estimation in polyphonic music signals, including estimating the number of concurrent sounds and their F0s, has been generally considered as one of the central problems in many music signal analysis applications, including automatic transcription, source separation and music information retrieval. Contrary to its importance, however, many conventional methods fell clearly behind human's ability in both accuracy and flexibility.

In recent years, some new methods with different techniques have been proposed. De Cheveigné [1], Tolonen and Karjalainen [2] proposed temporal cancelation methods based on modeling the human auditory system. Klapuri [3] also used auditory filterbanks as the front end, but estimated the F0s in an iterative spectral subtraction fashion. Poliner and Ellis [4] viewed the multiple F0 estimation as a multi-class classification problem, where each note in the piano is a class. Kashino and Murase [5] applied a Bayesian probability network to integrate musical context to address this problem. Davy, Godsill and Idier [6] proposed a generative signal model and employed reversible jump Markov Chain Monte Carlo (MCMC) to estimate the parameters including the F0s and polyphony. Goto [7], modeled the amplitude spectrum using mixture models and used a multiple-agent architecture to decide the birth and death of F0s. Kameoka, Nishimoto and Sagayama [8] also employed probabilistic spectral models and used some information criteria to decide the number of concurrent sounds.

However, all these methods above model the whole data (signal or spectrum), which is high dimensional and contains much useless information and noise. Some reductions of the data should simplify the problem and reduce irrelevant information. Goldstein [9] firstly proposed the method of probabilistic modeling the spectral peaks instead of the whole spectrum. He modeled the deviation of peak frequencies with a Gaussian but left the amplitude information unused. Thornburg, Leistikow and Berger [10] furthered this idea and incorporated the amplitude information of the peaks, aiming at melody extraction and musical onset detection. However, neither of them addressed the scenario of multiple pitches.

In this paper, we further the spectral peaks modeling idea into the multiple F0s scenario. We view multiple F0 estimation as a Maximum Likelihood estimation problem in the frequency domain in each frame, where the F0s are the parameters and the frequencies and amplitudes of the peaks are the observation. The proposed method incorporates the possibility of the peak detection errors into the likelihood function, and models the "true" and "false" peaks separately. The parameters of the models are learned from the monophonic training data, with the assumption that the statistics of the peaks in monophonic and polyphonic signals are similar. For estimating the number of parameters, i.e. the polyphony, a weighted Bayesian Information Criteria (BIC) is used. Experimental results on chords with polyphony ranging from 1 to 4, which are generated from previously unseen monophonic tones, show the feasibility of this method.

2 Modeling

Given a frame of polyphonic music, the multiple F0 estimation can be seen as a parameter estimation problem, where the spectrum is the observation and the F0s are the parameters. Here suppose the number of F0s is N, a Maximum Likelihood model can be formulated as:

$$\left(\hat{f_0^1}, \cdots, \hat{f_0^N}\right) = \arg\max_{f_0^1,\cdots,f_0^N \in \mathcal{F}} p\left(O|f_0^1, \cdots, f_0^N\right) \quad (1)$$

$$\overset{(assum.)}{=} \arg\max_{f_0^1,\cdots,f_0^N \in \mathcal{F}} p\left(f_1, A_1, \cdots, f_K, A_K|f_0^1, \cdots, f_0^N\right) \quad (2)$$

where $f_0^1, \cdots, f_0^N$ are the N logarithmic fundamental frequencies; $\mathcal{F}$ is the possible range of F0s; O represents the observation spectrum; $f_1, \cdots, f_K$ are the K logarithmic frequencies of the peaks in the amplitude spectrum, and $A_1, \cdots, A_K$ are their logarithmic amplitudes. In this paper, frequencies and amplitudes are all handled in the logarithmic scale (MIDI number and dB, respectively) [1] , for both ease of manipulation and accordance with human perception.

The reduction of the observation from the spectrum to the frequencies and amplitudes of the peaks has several reasons: First, in terms of human perception, phase spectrum has little effect while peaks in the amplitude spectrum have the very importance. Reserving only the peaks just cause little distortion of a harmonic sound [14]. Second, these peaks contain the most useful information for pitch estimation, because peaks appear and only appear at the integral times of the fundamental frequencies due to the physical property of harmonic instruments. Third, this reduction reduces the dimension of the observation significantly, and will make the parameter estimation much easier.

2.1 The Likelihood function

Ideally, peaks, representing the harmonics, are entirely caused by the F0s. However, some "false" peaks caused by other reasons, such as noise, inherent limitations of the peak detection method and overlapping partials of different F0s, may also be detected. So far there is not any peak detection method that can detect all the "true" peaks and withdraw all the "false" peaks. Therefore, the possibility of being one of the two cases for each detected peak should be considered. In our likelihood function $\mathcal{L}(\theta)$, $I_i = 1$ and $I_i = 0$ represent that peak i is "true" and "false", respectively.

$$\mathcal{L}(\theta) = p\left(f_1, A_1, \cdots, f_K, A_K|f_0^1, \cdots, f_0^N\right) \quad (3)$$

$$= \sum_{I_1,\cdots,I_K} p\left(f_1, A_1, I_1, \cdots, f_K, A_K, I_K|f_0^1, \cdots, f_0^N\right) \quad (4)$$

$$\overset{(assum.)}{=} \sum_{I_1,\cdots,I_K} \prod_{i=1}^{K} p\left(f_i, A_i, I_i|f_0^1, \cdots, f_0^N\right) \quad (5)$$

[1] MIDI number = 69+12× $\log_2$(Hz/440); dB = 20× $\log_{10}$(Linear amplitude)

$$= \prod_{i=1}^{K} \sum_{I_i} p\left(f_i, A_i, I_i | f_0^1, \cdots, f_0^N\right) \tag{6}$$

$$= \prod_{i=1}^{K} \sum_{I_i} p\left(f_i, A_i | I_i; f_0^1, \cdots, f_0^N\right) p\left(I_i | f_0^1, \cdots, f_0^N\right) \tag{7}$$

$$\overset{(assum.)}{=} \prod_{i=1}^{K} \sum_{I_i} p\left(f_i, A_i | I_i; f_0^1, \cdots, f_0^N\right) p\left(I_i\right) \tag{8}$$

$$= \prod_{i=1}^{K} \left\{p\left(f_i, A_i | I_i = 1; f_0^1, \cdots, f_0^N\right) p\left(I_i = 1\right) + \left(f_i, A_i | I_i = 0\right) p\left(I_i = 0\right)\right\} \tag{9}$$

From Eq.(4) to Eq.(5), the conditional independence of the peaks given F0s is assumed. This is reasonable for "true" peaks, because they are caused only by the F0s. For "false" peaks, because they can be influenced by other peaks, this assumption is not true. But it is commonly used in the spectral probabilistic modeling literature [6] for the ease of manipulation.

$p\left(I_i | f_0^1, \cdots, f_0^N\right)$ in Eq.(7) represents the probability of a detected peak is "true" or "false", i.e. the peak detection accuracy, given F0s. Although this accuracy may be influenced by the number and values of F0s, this influence is neglectable compared with the inherent property of the peak detection algorithm. Therefore, from Eq.(7) to Eq.(8), the independence between I_i and F0s is assumed.

As the prior information, $p\left(I_i\right)$ is learned from the monophonic training data. In each frame of the training data, YIN [11], a robust single F0 detection algorithm, is used to detect the exact F0 around the labeled note name. Then the deviation of a detected peak from the nearest harmonic position of the F0 is calculated. If the deviation is less than half a semitone range, the peak is treated "true", otherwise "false". The portion of the "true" peaks is 0.964 and is used as $p\left(I_i = 1\right)$.

2.2 Modeling the true peaks

A "true" peak may be generated by only one F0, or several F0s when they all have a harmonic at the peak position. In the later case, the linear amplitude of the peak is the summation of those of all the overlapping harmonics if they have the same phase, however, the logarithmic amplitude approximates the maximum of those of the harmonics. For example, a 50 dB harmonic overlapping a 30 dB one will cause at most a 50.8 dB peak. When the two harmonics have the same amplitude and phase, the logarithmic amplitude of the peak will have the biggest increase of 6 dB, but this case is rather rare and the amplitudes of most peaks are higher than 20 dB. Therefore, it is reasonable to assume that each peak is mainly generated by one F0, and is approximately independent from the other F0s. Then the likelihood of a "true" peak given F0s can be deduced as:

$$p\left(f_i, A_i | I_i = 1; f_0^1, \cdots, f_0^N\right) \overset{(assum.)}{=} p\left(f_i, A_i | f_0^{l(i)}\right) \tag{10}$$

$$= p\left(A_i | f_i, f_0^{l(i)}\right) p\left(f_i | f_0^{l(i)}\right) \tag{11}$$

where $l\left(i\right) \in \left\{1, \cdots, N\right\}$, is a function of i and means that peak i is mainly generated by $f_0^{l(i)}$. Note that it also implicates the condition $I_i = 1$. In practice, $f_0^{l(i)}$ is selected from the supposed N F0s to maximize Eq.(10). In Eq.(11) the likelihood of "true" peak i is split into two parts: the conditional distribution of amplitude and frequency, respectively.

In modeling $p\left(A_i | f_i, f_0^{l(i)}\right)$, we change the conditions to other equivalent conditions:

$$p\left(A_i | f_i, f_0^{l(i)}\right) = p\left(A_i | f_i, h_i(f_0^{l(i)})\right) \tag{12}$$

$$= \frac{p\left(A_i, f_i, h_i(f_0^{l(i)})\right)}{p\left(f_i, h_i(f_0^{l(i)})\right)} \tag{13}$$

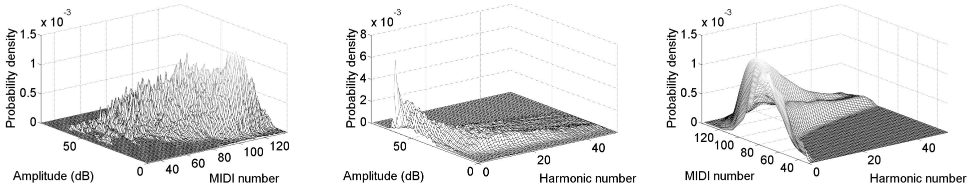
where $h_i(f_0^{l(i)}) = 2^{(f_i - f_0^{l(i)})/12}$, is the harmonic number of peak i, given its fundamental $f_0^{l(i)}$.

Table 1: Correlation coefficients between several variables of the "true" peaks of the monophonic training data. A, f, f_0 and h are the logarithmic amplitude, logarithmic frequency, logarithmic fundamental frequency and harmonic number of a peak, respectively. d is the logarithmic frequency deviation of a peak from the nearest harmonic position of its fundamental frequency.

	A	f	f_0	h	d
A	1.00	-0.72	-0.04	-0.61	0.00
f	-0.72	1.00	0.42	0.55	-0.00
f_0	-0.04	0.42	1.00	-0.40	0.01
h	-0.61	0.55	-0.40	1.00	-0.01
d	0.00	-0.00	0.01	-0.01	1.00

The reason of the replacement of the conditions is that A_i is much more correlated with $h_i(f_0^{l(i)})$ than with $f_0^{l(i)}$, because higher harmonics tend to have lower amplitudes while any fundamental frequency can have harmonics with very diverse amplitudes. This conclusion is proved by the statistics of the training data. In Table 1, it can be seen that the correlation between A and h is much stronger than that between A and f_0. Therefore, the joint probability density function $p\left(A_i, f_i, h_i(f_0^{l(i)})\right)$ is more convergent than $p\left(A_i, f_i, f_0^{l(i)}\right)$ and hence easier to learn.

$p\left(A_i, f_i, h_i(f_0^{l(i)})\right)$ is estimated using Parzen window method, because it is very hard to give a parametric model. A $11 \times 11 \times 5$ Gaussian window with variance 4 is used to serve as a smoothing function and the parameters are selected without tuning. Figure 1 shows the three 2-D marginal density of the 3-D joint probability density function.



(a) $p\,(A, f)$ (b) $p\,(A, h)$ (c) $p\,(f, h)$

Figure 1: The three marginal density of the joint probability density function $p\left(A_i, f_i, h_i(f_0^{l(i)})\right)$, which is learned from the monophonic training data using Parzen window method.

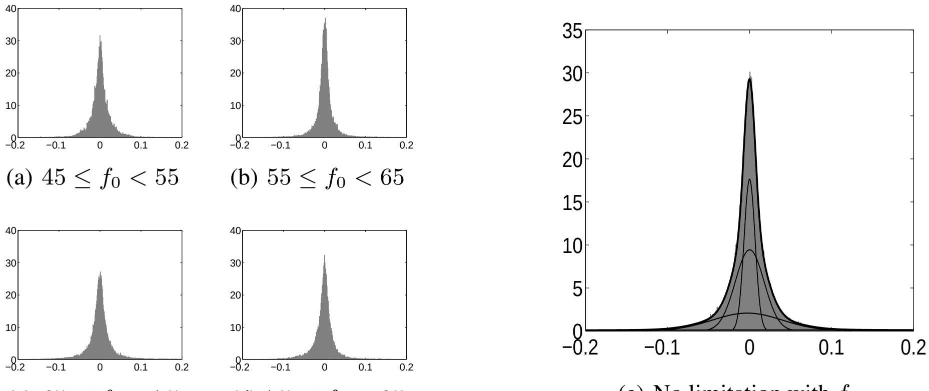
In modeling $p\left(f_i | f_0^{l(i)}\right)$, the second part of Eq.(11), we have:

$$\begin{aligned} p\left(f_i | f_0^{l(i)}\right) &\overset{(assum.)}{=} p\left(d_i | f_0^{l(i)}\right) && (14) \\ &\overset{(assum.)}{=} p\left(d_i\right) && (15) \\ d_i = f_i - f_0^{l(i)} &- \left[f_i - f_0^{l(i)}\right] && (16) \end{aligned}$$

where $[\cdot]$ denotes rounding to the nearest integer, and d_i is the frequency deviation in MIDI number of peak i from the nearest harmonic position of the given F0. Note that it always lies in $[-0.5, 0.5]$, i.e. a semitone range.

In Eq.(16), it is assumed that there always be a detected "true" peak in the semitone range around any harmonic position of its F0, so the integration of $p\left(f_i | f_0^{l(i)}\right)$ in each semitone range equals to 1, and the probability space of f_i is divided into each semitone range. With this assumption, $\left[f_i - f_0^{l(i)}\right]$ is constant with respect to f_i, therefore, f_i and d_i have the same distribution with only different average. This assumption is not true for some instruments, such as the clarinet, which has undetectable peaks around some even harmonic positions, however, it is acceptable by most instruments and can be mostly assured by loosening the detection conditions in the peak detection algorithm. In Eq.(15), the frequency deviation d_i is assumed to be independent from $f_0^{l(i)}$. This assumption is proved by the statistics in the training data, including both the small correlation coefficient between f_0 and d in Table 1, and the similar shape of the conditional densities $p\left(d_i|f_0^{l(i)}\right)$ in Figure 2(a)-(d).

Thus the normalized histogram of d_i is calculated using all the "true" peaks in the training data and plotted in Figure 2(e). It can be seen that this distribution is symmetric about zero, a little long tailed but not very spiky. A Gaussian Mixture Model (GMM) with four kernels is used to estimate this distribution. The probability density of the kernels and the whole is also plotted in Figure 2(e). The nearly concentric characteristics of the kernels shows the symmetry the distribution.



(a) $45 \leq f_0 < 55$

(b) $55 \leq f_0 < 65$

(c) $65 \leq f_0 < 75$

(d) $75 \leq f_0 < 85$

(e) No limitation with f_0

Figure 2: Illustration of modeling the frequency deviation of "true" peaks. The horizontal axis is the deviation in MIDI number, and the vertical axis is the probability density. (a)-(d) plot the conditional density, given the condition of a fundamental frequency range. (e) plots the density without this condition, and also plots the density of the Gaussian Mixture Model (four kernels in the thin curve and the whole in the bold curve), which is used to fit the deviation's distribution.

2.3 Modeling the false peaks

A "false" peak is that which is not generated by any F0s but detected by the peak detection algorithm. In monophonic training data, it is easy to classify these peaks, because they lie outside the semitone range of the nearest harmonic position of their F0s. These peaks are collected and used to estimate $p(f_i, A_i|I_i = 0)$ in Eq.(9). The shape of this probability density is plotted in Figure 3. It is a somewhat Gaussian-like distribution. Because the prior probability of "false" peaks is only 0.036, there is no need to model this density very precisely. Thus a 2-D Gaussian distribution is used here, whose means and covariance are calculated to be $(92.7, 20.3)$ and $\begin{pmatrix} 208.5 & -43.0 \\ -43.0 & 41.0 \end{pmatrix}$.

2.4 Model selection

So far the modeling of each part of Eq.(9) is done, but the number of F0s is still not estimated. Note that in Eq.(10), $f_0^{l(i)}$ is selected from the supposed F0s to maximize the "true" peak part likelihood of peak i, and in Eq.(9) the change of supposed F0s only causes the change of the "true" peak part likelihood. Therefore, if one more F0 is added to the existing supposed F0s, the "true" part likelihood will increase and cause the whole likelihood increase:

$$p(f_1, A_1, \cdots, f_K, A_K|f_0^1, \cdots, f_0^N) \leq p(f_1, A_1, \cdots, f_K, A_K|f_0^1, \cdots, f_0^N, f_0^{N+1}) \qquad (17)$$

This is the typical overfitting problem of Maximum Likelihood method and is usually alleviated by applying some model selection criteria. Here a rectified Bayesian Information Criteria (BIC) [12] is adopted. The original model penalization part is weighted by $2K^{0.45}$, which is found suitable in our

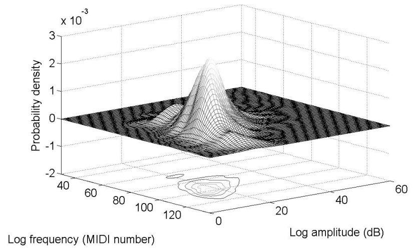


Figure 3: Illustration of the probability density of $p\left(f_i, A_i | I_i = 0\right)$, which is calculated from the "false" peaks of the training data. The contours of the density is plotted at the bottom of the figure.

Table 2: Algorithm flow

Training

1. Short time fourier transform (STFT) and peak detection;
2. Collect the "true" and "false" peaks;
3. Calculate $p\left(I_i\right)$, $p\left(A_i | f_i, f_0^{l(i)}\right)$, GMM parameters of $p\left(d_i\right)$, and Gaussian parameters of $p(f_i, A_i | I_i = 0)$.

Testing

1. Short time fourier transform (STFT) and peak detection;
2. Set $N = 1$, calculate and store the predominant fundamental, f_0^1, which maximizes Eq. (3);
3. $N = N + 1$;
4. Calculate and store f_0^N, which maximizes Eq. (3);
5. Repeat 3-4 until $N = 10$;
6. Select a value for N which maximizes Eq. (18). The estimated F0s are $f_0^1, \cdots, f_0^N$.

problem. Then the number of F0s and their frequencies are searched to maximize the BIC.

$$BIC = \ln p\left(f_1, A_1, \cdots, f_K, A_K | f_0^1, \cdots, f_0^N\right) - 2K^{0.45} \cdot \frac{1}{2} N \ln\left(2K\right) \tag{18}$$

3 Algorithm

In the proposed method, peak detection is the fundamental step. Here the peak detection algorithm in [13] is adopted. Basically, the smoothed amplitude envelope is calculated by convolving the amplitude spectrum with a moving Gaussian filter. Then the local maxima, which are higher than the envelope for a threshold (e.g. 8 dB) and whose amplitudes are also not lower than the global maximum for 50 dB, are detected as peaks. Finally, the peak amplitudes and frequencies are refined by quadratic interpolation [14]. The thresholds (e.g. 8 dB, 50 dB etc.) are the same for both training and testing data, to make the causes of "false" peaks as similar as possible.

Note that this Maximum Likelihood modeling tends to have the half F0 errors, because half F0s can have the same BIC value as the true F0s. In order to eliminate these errors and reduce the search space, the F0s are searched only around the 5 peaks who have lowest frequencies, and 5 peaks who have highest amplitudes. The search range is two semitones around these peaks and the step is 0.01 peak frequencies. Even though, the size of the search space is a combinational explosion problem when N is big. Here a greedy search strategy which starts from $N = 1$ is adopted. The whole algorithm flow including the training phase and testing phase is illustrated in Table 2.

4 Experiment

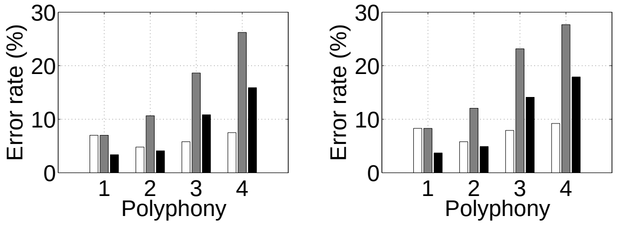
The proposed algorithm was evaluated on randomly mixed chords. The acoustic material consisted of note samples from the University of Iowa website [2]. There were in total 1500 note samples of dynamic "mf" and "ff" from 18 wind and arco string instruments with pitch ranging from C2 (65Hz, MIDI number 36) to B6 (1976Hz, MIDI number 95). Some of them had vibrato. 500 samples were randomly selected as the training data, and the others were randomly mixed with equal mean-square levels and no duplication in pitch into chords as the testing data. 1000 mixtures of one, two, three and four sounds were generated, totalling 4000 testing cases.

For both training and testing data, the sampling rate was 44.1 kHz and the frame length and hop size was 93ms and 46 ms, respectively. In training, all frames of each sample were used to detect and collect the peaks. In testing, one frame at the middle part of each chords is fed to the algorithm.

The results are illustrated in Figure 4 and 5. Figure 4(a) shows the results when the polyphony is given. White bars show the error rate of the *Predominant-F0 estimation*, in which an error is counted when the first estimated F0 mismatches all the true F0s. Here "match" means that two frequencies lie in the same semitone range of the Western musical scale. Gray bars show the error rate of the *Multiple-F0 estimation*, which is defined as the percentage of all F0s that are not correctly estimated in the input signals. Black bars present the error rate of multiple-F0 estimation without counting the octave errors. It can be seen that the multiple-F0 error rate increases fast with the polyphony, but that of the predominant-F0 remains almost the same. This indicates that the greedy search strategy in Table 2 is feasible. In addition, from the black bars, it can be seen that the octave errors take up almost the half of all the multiple-F0 errors. On one hand, this indicates the inherent limitations of our algorithm; on the other hand, these errors are not that annoying in some scenarios, such as chord recognition from the F0 estimates.

Figure 4(b) shows the results of substituting the GMM model of $p(d_i)$ in Figure 2(e) with a Gaussian model. The mean and variance of the Gaussian model is set to those of the histogram. It can be seen that this substitution makes the error rates of all the indexes increase. It indicates that the statistical information about the peaks in the monophonic training data is more helpful than a usually used non-informative Gaussian model in modeling the peaks in the polyphonic testing data.

Figure 5 shows the polyphony estimation results. In each panel the asterisk indicates the true polyphony, and bars show a histogram of the estimates. The results for polyphony 1 and 2 are acceptable, but are not satisfactory for polyphony 3 and 4. It indicates that the rectified BIC is still not a proper method for estimating the number of the concurrent sounds.



(a) $p(d_i)$ in a learned GMM (b) $p(d_i)$ in a non-informative Gaussian

Figure 4: Error rates of predominant-F0 (white), multiple-F0 (gray) and multiple-F0 without octave errors (black) estimations, given the polyphony.

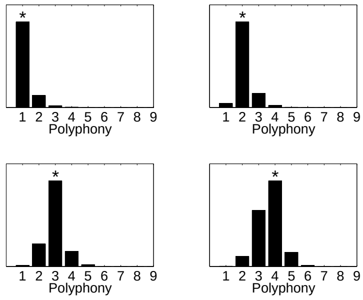


Figure 5: Histograms of polyphony estimates. The asterisks indicate the true polyphonies.

5 Conclusion and discussion

In this paper, a Maximum Likelihood approach in the frequency domain is proposed to address the multiple F0 estimation problem. Different from the other probabilistic spectral modeling methods,

[2] http://theremin.music.uiowa.edu/

the proposed method works on the frequencies and amplitudes of the spectral peaks. It considers the possibility of the errors in the peak detection algorithm and models the "true" and "false" peaks separately. Parameters of these models are learned from the peaks of the monophonic training data. The proposed method also incorporates a rectified BIC to estimate the number of the concurrent sounds. The combinational explosion problem in the search space is approached by a greedy search strategy. Experimental results on the randomly mixed chords show the feasibility of this method.

Compared with other spectral probabilistic modeling methods, using only the peaks reduces the dimension of the observation and may prevent the effects of some noise. In addition, it makes that possible of modeling subtly the peaks, which serve important roles in human perception of sounds.

Considering the possibility of errors in the peak detection algorithm in the likelihood function is one of the core idea of this method, because it prevents the algorithm being trapped into finding F0s to increase the likelihood of the "false" peaks. Discriminating and learning the statistics of the "true" and "false" peaks, however, is hard to accomplished in the testing polyphonic data directly, therefore we use the statistics of the peaks in the monophonic training data. Though the multiple F0s may make the peak statistics different from the training data, the experimental results prove this model's feasibility. In the future, we would like to use the training data to start and then "bootstrap" the modeling of peaks in the testing data itself, that is, we want to iteratively learn the statistics and discriminate the "true" and "false" peaks in the testing data.

The current formulations limits the use of the method in harmonic sounds, but it is not hard to extend to quasi-harmonic sounds such as piano tones. The only change will occur in the calculation of the harmonic number of each peak. Another inherent limitation lies in the tendency of estimating the half F0s. We want to address this issue in the future by rectifying the likelihood function, such as increasing the spectral amplitudes at the harmonic positions of the F0s into the observation.

References

[1] A. de Cheveigné. "Separation of concurrent harmonic sounds: fundamental frequency estimation and a time-domain cancellation model of auditory processing," *J. Acoust. Soc. Am.*, vol. 93, no. 6, pp. 3271-3290, 1993.

[2] T. Tolonen and M. Karjalainen. "A computationally efficient multipitch analysis model," *IEEE Trans. Speech and Audio Processing*, vol. 8, no. 6, pp. 708-716, 2000.

[3] A. Klapuri, "Multiple fundamental frequency estimation by summing harmonic amplitudes," In *Proc. International Conference on Music Information Retrieval*, 2006.

[4] G. Poliner and D. Ellis, "A discriminative model for polyphonic piano transcription," *Eurasip Journal on Advances in Signal Processing*, vol. 2007, Article ID 48317, 9 pages, 2007.

[5] K. Kashino and H. Murase, "A sound source identification system for ensemble music based on template adaptation and music stream extraction," *Speech Communication*, vol. 27, pp. 337-349, 1999.

[6] M. Davy, S. J. Godsill and J. Idier, "Bayesian analysis of western tonal music," *J. Acoust. Soc. Am.*, Vol. 119, No. 4, pp. 2498-2517, Apr. 2006.

[7] M. Goto, "A real-time music scene description system: predominant-F0 estimation for detecting melody and bass lines in real-world audio signals," *Speech Communication*, vol. 43, no. 4, pp. 311-329, 2004.

[8] H. Kameoka, T. Nishimoto and S. Sagayama, "Separation of harmonic structures based on tied gaussian mixture model and information criterion for concurrent sounds," In *Proc. IEEE International Conference on Acoustic, Speech and Signal Processing*, 2004, pp. 297-300.

[9] J. Goldstein, "An optimum processor theory for the central formation of the pitch of complex tones," *J. Acoust. Soc. Am.*, vo. 54, pp. 1496-1516, 1973.

[10] H. Thornburg, R. J. Leistikow, J. Berger, "Melody extraction and musical onset detection via probabilistic models of framewise STFT peak data," *IEEE Trans. Audio Speech Language Process.*, vo. 15, no. 4, pp. 1257 - 1272, 2007.

[11] A. Cheveigné, and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *J. Acoust. Soc. Am.*, vol. 111, pp. 1917-1930, 2002.

[12] G. Schwarz, "Estimating the dimension of a model," *Annals of Statistics*, vol. 6, pp. 461-464, 1978.

[13] Z. Duan, Y. Zhang, C. Zhang and Z. Shi, "Unsupervised single-channel music source separation by average harmonic structure modeling", *IEEE Trans. Audio Speech Language Process.*, submitted.

[14] J. O. Smith, X. Serra "PARSHL: an analysis/synthesis program for non-harmonic sounds based on a sinusoidal representation," In *Proc. Internetional Computer Music Conference*, 1987.