A Maximum Likelihood Approach to Multiple F0 Estimation From the Amplitude Spectrum Peaks

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

Music, Mind and Cognition workshop of NIPS07
Whistler, Canada, Dec. 7, 2007

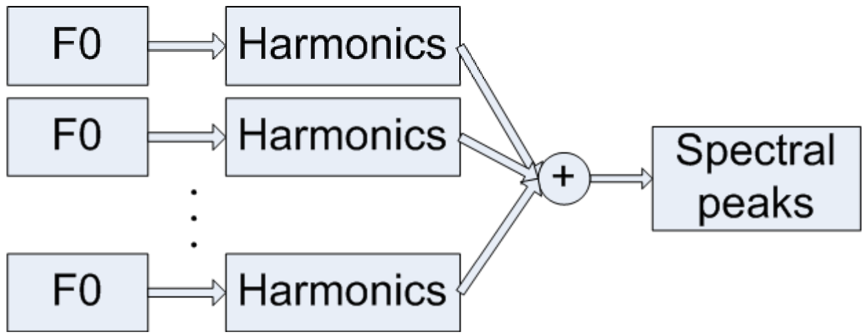
Neural Information Processing Systems Conference

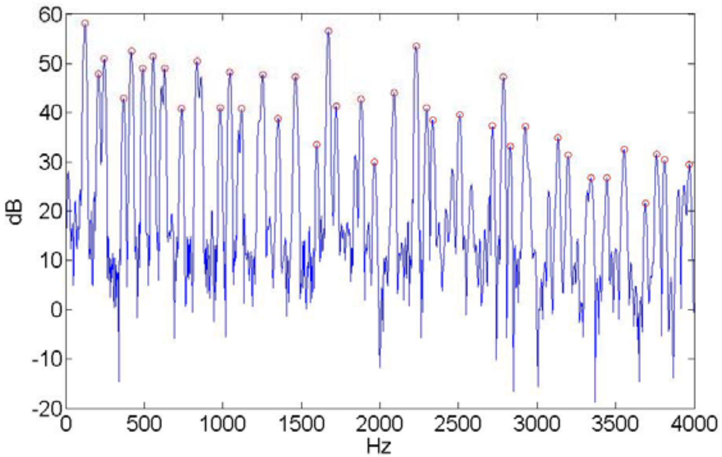
Multiple F0 Estimation

- A sound with mixed tones, tone 1 (F3), tone 2 (C4)
  - Estimate the polyphony (number of tones)
  - Estimate the frequencies of these tones
- How do musicians do this?
  - Analyze the frequency components by ears
  - Infer the frequencies by the brain
- Can computers also do this?
  - Analyze the frequency components by STFT
  - Infer the frequencies by a Maximum Likelihood method

Problem Formulation

- Parameters to be estimated
  - Number of F0s: N
  - F0s: $f_0^1, \cdots, f_0^N$
- Observation
  - frequencies and amplitudes of the peaks in the amplitude spectrum





Likelihood Function

$$\begin{aligned}
\mathcal{L}(\theta) &= p\left(f_1, A_1, \cdots, f_K, A_K \mid f_0^1, \cdots, f_0^N\right) \\
&= \sum_{I_1,\cdots,I_K} p\left(f_1, A_1, I_1, \cdots, f_K, A_K, I_K \mid f_0^1, \cdots, f_0^N\right) \\
&\overset{(assum.)}{=} \sum_{I_1,\cdots,I_K} \prod_{i=1}^{K} p\left(f_i, A_i, I_i \mid f_0^1, \cdots, f_0^N\right) \\
&= \prod_{i=1}^{K} \boxed{\sum_{I_i} p\left(f_i, A_i, I_i \mid f_0^1, \cdots, f_0^N\right)}
\end{aligned}$$

- A peak
  - "True": $I_i = 1$: generated by a harmonic
  - "False": $I_i = 0$: caused by detection errors

Likelihood Function (a peak)

$$\sum_{I_i} p\left(f_i, A_i, I_i | f_0^1, \cdots, f_0^N\right)$$

$$= \{\underbrace{p\left(f_i, A_i | I_i = 1; f_0^1, \cdots, f_0^N\right)}_{\text{“true” peak part}} p\left(I_i = 1\right) + \underbrace{\left(f_i, A_i | I_i = 0\right)}_{\text{“false” peak part}} p\left(I_i = 0\right)\}$$

- Learn the parameters from the training data
  - Training data: the monophonic note samples
  - Easy to know whether a peak is “true” or “false”
  - $p\left(I_i = 1\right)$ = 0.964

True Peak Part

$$\left\{\underbrace{p\left(f_i, A_i | I_i = 1; f_0^1, \cdots, f_0^N\right)} p\left(I_i = 1\right) + \left(f_i, A_i | I_i = 0\right) p\left(I_i = 0\right)\right\}$$

$$\begin{aligned} p\left(f_i, A_i | I_i = 1; f_0^1, \cdots, f_0^N\right) &\overset{(assum.)}{=} p\left(f_i, A_i | f_0^{l(i)}\right) \\ &= \underbrace{p\left(A_i | f_i, f_0^{l(i)}\right)}_{\text{amplitude}} \underbrace{p\left(f_i | f_0^{l(i)}\right)}_{\text{frequency}} \end{aligned}$$

- Assume that each “true” peak is generated by only one F0
  - 50dB + 30dB = 50.8dB

True Peak Part (amplitude)

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

- Replace F0 with hi: harmonic number of the peak i

	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

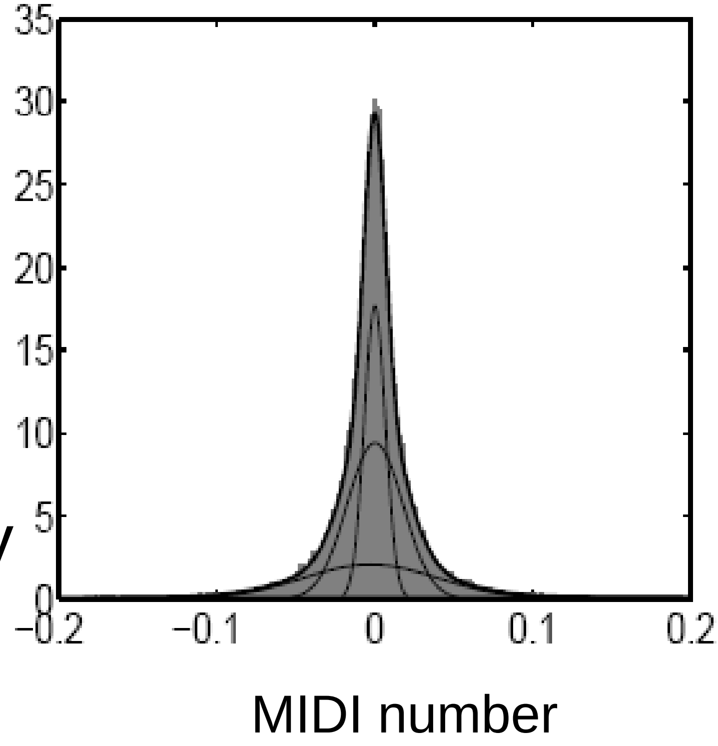
- Estimate $p\left(A_i, f_i, h_i(f_0^{l(i)})\right)$ from the training data
  - A Parzen window (11*11*5)

True Peak Part (frequency)

- Convert the peak frequency into the frequency deviation of the peak from the nearest harmonic position of F0

$$p\left(f_i | f_0^{l(i)}\right) \overset{(assum.)}{=} p\left(d_i | f_0^{l(i)}\right) \overset{(assum.)}{=} p\left(d_i\right)$$

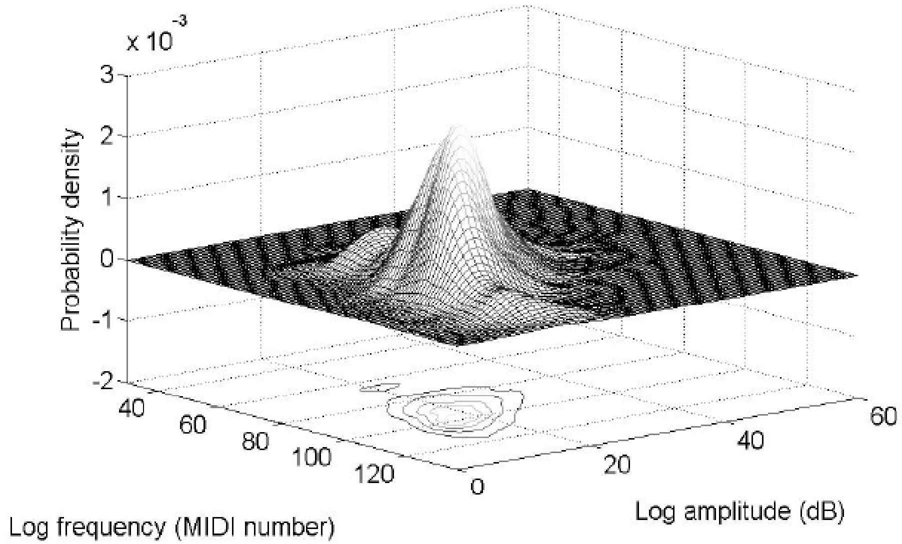
- Estimated from training data
- Symmetric, long tailed, not spiky
- A Gaussian Mixture Model (4 kernels)



False Peak Part

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

- Estimated from training data



- A Gaussian distribution
  - Mean $(92.7, 20.3)$
  - covariance $\begin{pmatrix} 208.5 & -43.0 \\ -43.0 & 41.0 \end{pmatrix}$

Estimating the Polyphony

- The likelihood will increase with the number of F0s (overfitting)
- A weighted Bayesian Information Criteria (BIC)
  - K: number of peaks; N: polyphony

$$BIC = \underbrace{\ln p\left(f_1, A_1, \cdots, f_K, A_K \middle| f_0^1, \cdots, f_0^N\right)}_{\text{Log likelihood}} - \underbrace{2K^{0.45}}_{\text{weight}} \cdot \underbrace{\frac{1}{2} N \ln(2K)}_{\text{BIC penalty}}$$

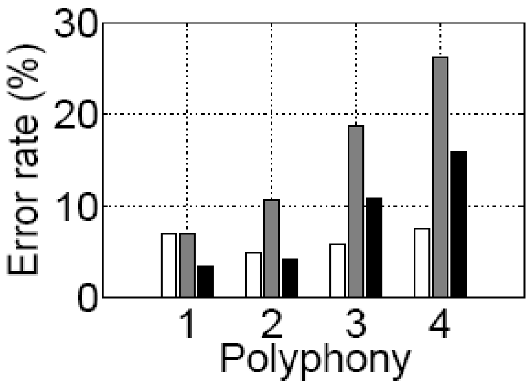
- Search the F0s and the polyphony to maximize BIC
  - A combinational explosion problem
  - Greedy search: Start from N=1; add F0 one by one

Experiments (1)

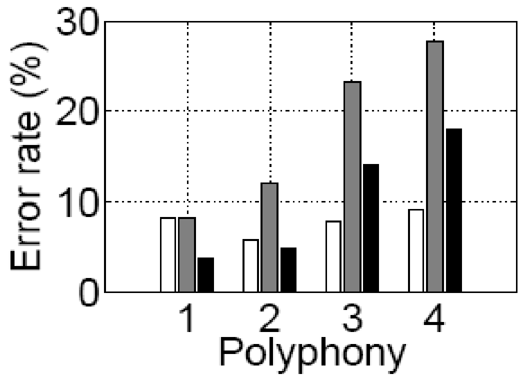
- Acoustic materials: 1500 note samples from Iowa music database
  - 18 wind and arco-string instruments
  - Pitch range: C2 (65Hz) – B6 (1976Hz)
  - Dynamic: mf, ff
- Training data: 500 notes
- Testing data: generated using the other 1000 notes
  - Mixed with equal mean square level and no duplication in pitch
  - 1000 mixtures each for polyphony 1, 2, 3 and 4.

Experiments (2)

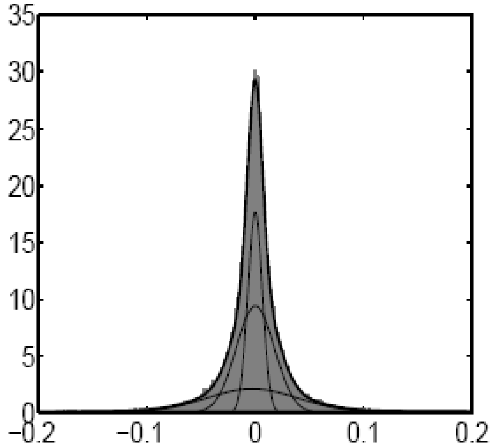
- Frequency estimation



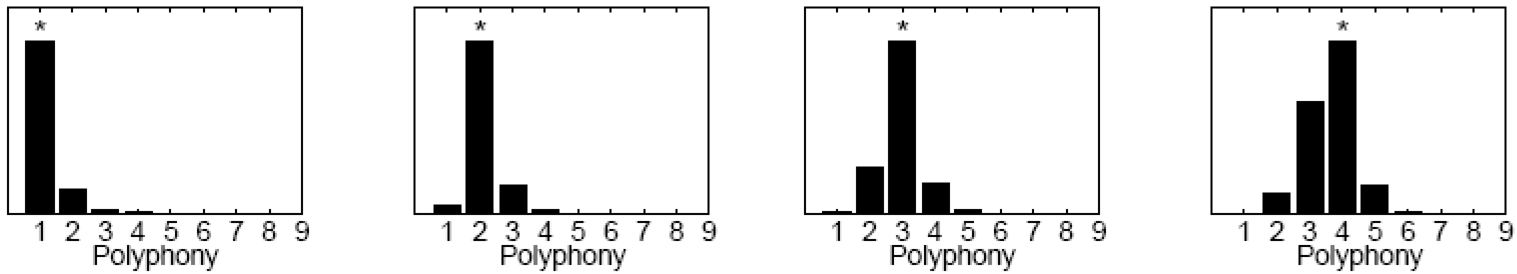
(a) $p(d_i)$ in a learned GMM



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



- Polyphony estimation



Thank you!

Welcome to my poster!