ICASSP 2007

Fundamental Frequency Estimation based on Pitch-Scaled Harmonic Filtering

Sergio Roa, Maren Bennewitz and Sven Behnke

Computer Science Institute, University of Freiburg, Germany
1 PSHF Based pitch estimation
   - PSHF
   - HTF
   - Pitch Tracking

2 Experiments
   - Dataset
   - Clean Speech
   - Noisy Speech

3 Conclusion
Objectives:

- Estimating the pitch (fundamental frequency) of speech for each time frame of processing.
Objectives:

- Estimating the pitch (fundamental frequency) of speech for each time frame of processing.
- Separating the harmonic and non-harmonic parts of the signal.
Introduction

Objectives:

- Estimating the pitch (fundamental frequency) of speech for each time frame of processing.
- Separating the harmonic and non-harmonic parts of the signal.
- Developing a robust algorithm for noisy speech (especially cocktail party noise).
Introduction

Objectives:

- Estimating the pitch (fundamental frequency) of speech for each time frame of processing.
- Separating the harmonic and non-harmonic parts of the signal.
- Developing a robust algorithm for noisy speech (especially cocktail party noise).
- Recognition of voiced and unvoiced frames, keeping track of the pitch of one speaker.
Algorithm based on Pitch-scaled harmonic filtering

How to estimate the pitch by using the STFT?

- The use of windows with a predetermined size centered at given frequency bins may lead to lost of information, because peak heights (multiples of the fundamental) may be found between these frequency bins.
Algorithm based on Pitch-scaled harmonic filtering

One could use pitch-synchronous analysis: use a rectangular window whose length is exactly one pitch period.
One could use pitch-synchronous analysis: use a rectangular window whose length is exactly one pitch period. But, we need to know the local pitch period!
One could use pitch-synchronous analysis: use a rectangular window whose length is exactly one pitch period. But, we need to know the local pitch period!

The pitch-scaled harmonic filter (PSHF) method allows us to use an analysis frame scaled to the pitch period, i.e., the analysis frame contains a small integer of pitch periods.
One could use pitch-synchronous analysis: use a rectangular window whose length is exactly one pitch period. But, we need to know the local pitch period!

The pitch-scaled harmonic filter (PSHF) method allows us to use an analysis frame scaled to the pitch period, i.e., the analysis frame contains a small integer of pitch periods.

The number $N$ of sample points will be inversely proportional to $F_0$. The harmonics of $F_0$ will be aligned with certain bins of the STFT.
A window function $w$ of length $N(m)$ is applied to the speech signal $x(n)$ centered at time $m$, to form $x_w(n) = w(n)x(n + m - \frac{N}{2})$. The spectrum $X_w(k)$ is computed by the DFT using $N = bP$ ($b$ pitch periods of length $P$ samples). Spectrum of a vowel /a/ spoken by a male with $F_0 = 110$ Hz, using a 4-pitch period window:

For $b$ pitch periods: $nb$th Fourier coefficient $\rightarrow nF_0$. Harmonic bins = \{ $b, 2b, 3b, \ldots, (N - 1)b$\}. 

Sergio Roa, Maren Bennewitz and Sven Behnke
How to estimate the pitch?

- Approach: Use of different window lengths at regular points of time in order to estimate $F_0$.
- Use of a cost function, in order to maximize energy concentrated at the harmonic bins and the estimated level noise at the surrounding beans.
- Example: Given a signal sampled with $F_s = 20 kHz$, take the STFT every 5 ms using 4-pitch period rectangular windows (200 to 1000 samples $\rightarrow$ 80 Hz to 400 Hz).
Harmonic template function

Let $X^{-d}_h$ and $X^{+d}_h$ be the energy of the surrounding bins for a given window at each harmonic $h \in \{1, 2, \ldots, H\}$ and at a bins-distance $d = 1, \ldots, b - 1$. Cost function of the surrounding bins for $b = 4$ at a time frame $m$:

$$S^4_h(N, m) = 15(|X^1_h| + |X^{-1}_h|) + 6(|X^2_h| + |X^{-2}_h|) + (|X^3_h| + |X^{-3}_h|),$$

Harmonic Template function HTF:

$$J^4(N, m) = \sum_{h=1}^{H} 44|X_w(bh + 1)| - S^4_h,$$
Harmonic template function for pitch estimation (right-hand side) using 4-pitch period windows of length 200 to 1000 for the utterance “The north” (male), and its corresponding spectrogram.
For certain frequencies factors or multiples of the fundamental, energy peaks are also found. These frequencies correspond to pitch intervals, like octaves (2 : 1) or fifths (3 : 2). Let us consider $U_h^b(N, m) = 44|X_w(bh + 1)| - S_h^b$.
Repercussion of scaling:

Mapping of 4-pitch period window length to $F_0$.

HTF contour as a function of $F_0$. 

$F_0$ Estimation based on PSHF
Improvements of the HTF

- Use of different values $b$ pitch periods.

<table>
<thead>
<tr>
<th>$b$</th>
<th>$F_0$ min (Hz)</th>
<th>$F_0$ max (Hz)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>50</td>
<td>100</td>
</tr>
<tr>
<td>3</td>
<td>75</td>
<td>150</td>
</tr>
<tr>
<td>4</td>
<td>100</td>
<td>200</td>
</tr>
<tr>
<td>5</td>
<td>125</td>
<td>250</td>
</tr>
<tr>
<td>6</td>
<td>150</td>
<td>300</td>
</tr>
<tr>
<td>7</td>
<td>175</td>
<td>350</td>
</tr>
<tr>
<td>8</td>
<td>200</td>
<td>400</td>
</tr>
<tr>
<td>9</td>
<td>225</td>
<td>450</td>
</tr>
<tr>
<td>10</td>
<td>250</td>
<td>500</td>
</tr>
</tbody>
</table>

$F_0$ ranges for window lengths from 400 to 800 samples.
Use of different values $b$ pitch periods.

<table>
<thead>
<tr>
<th>$b$</th>
<th>$F_0$ min (Hz)</th>
<th>$F_0$ max (Hz)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>50</td>
<td>100</td>
</tr>
<tr>
<td>3</td>
<td>75</td>
<td>150</td>
</tr>
<tr>
<td>4</td>
<td>100</td>
<td>200</td>
</tr>
<tr>
<td>5</td>
<td>125</td>
<td>250</td>
</tr>
<tr>
<td>6</td>
<td>150</td>
<td>300</td>
</tr>
<tr>
<td>7</td>
<td>175</td>
<td>350</td>
</tr>
<tr>
<td>8</td>
<td>200</td>
<td>400</td>
</tr>
<tr>
<td>9</td>
<td>225</td>
<td>450</td>
</tr>
<tr>
<td>10</td>
<td>250</td>
<td>500</td>
</tr>
</tbody>
</table>

$F_0$ ranges for window lengths from 400 to 800 samples.

Use of binomial coefficients to automatically generate HTFs. In general,

$$S_h^b(N, m) = \sum_{d=1}^{b-1} \frac{1}{K} \binom{2(b-1)}{d-1} \left( |X_h^d| + |X_h^{-d}| \right)$$

for $b = 2, \ldots, 10$,

$$J_h^b(N, m) = \sum_{h=1}^{H} |X_w(bh + 1)| - S_h^b$$

for $b = 2, \ldots, 10$,

where $K$ is a normalization constant.
Values returned by all HTFs are combined (interpolated and normalized) in order to obtain a final HTF.

- Use of a logarithmic scale for the range of $F_0$ for the interpolation.

![Graph showing the logarithmic scale for $F_0$ and interpolation index.](image-url)
For window lengths corresponding to half of the fundamental, peaks of energy are immediately followed by valleys. Using the adjacent HTF values $U_{b+1}$ and $U_{b-1}$ we apply the operation:

$$U_{b}^h(N, m) = |X_w(bh + 1)| - S_{h}^b = \min \left\{ \sum U_{b+1} + U_{b-1}, U_{b} \right\}$$

for $h = 2, \ldots, H - 1$. 

Harmonic coefficients for $F_0 = 176$ Hz
Subtraction of half of the energy contained at $F_0$ from the energy at its multiples.

The energy was erased up to a given $F_0$ at multiples $2F_0$ and $3F_0$. 
Pitch tracking

- Use of dynamic programming methods to track the pitch.
- Decision of voice onset/offset by using a pitch confidence ratio.

Pitch tracking for a male voice.
Keele Pitch Database

- speech signals
  - 5 female speakers
  - 5 male speakers
  - each speaking about 30-40s
  - 20 kHz sampling rate

- pitch reference
  - simultaneous recorded signal of laryngograph
  - hand made corrections
  - 100 Hz sampling rate (pitch estimation every 10 ms)
Error Measures

- **voiced error (VE)**
  - reference voiced
  - estimation unvoiced
- **unvoiced error (UE)**
  - reference unvoiced
  - estimation voiced
- **gross pitch error (GPE)**
  - reference and estimation differ by more than 20%
- **root mean square error (RMSE)**
  \[ \sqrt{\sum_t (\text{ref}(t) - \text{est}(t))^2} \]
Clean Speech. Some experimental results

Path comparison between ground truth (Keele database) and proposed algorithm.
Some experimental results

Multiple speakers

- Energy of the male voice is higher than of the female one, with SNR of 8 dB.

Pitch tracking for the female voice.

Pitch tracking with multiple speakers.
## Evaluation results against Keele database

<table>
<thead>
<tr>
<th></th>
<th>VE (%)</th>
<th>UE (%)</th>
<th>GPE (%)</th>
<th>RMSE (Hz)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSHF Based</td>
<td>4.51</td>
<td>5.06</td>
<td>0.61</td>
<td>2.46</td>
</tr>
<tr>
<td>NMF (*)</td>
<td>7.7</td>
<td>4.6</td>
<td>0.9</td>
<td>4.3</td>
</tr>
<tr>
<td>RAPT (*)</td>
<td>3.2</td>
<td>6.8</td>
<td>2.2</td>
<td>4.4</td>
</tr>
</tbody>
</table>

(*) Sha, Saul: *Real-Time Pitch Determination of One or More Voices by Non-negative Matrix Factorization*. NIPS, 2004
Noisy Speech. Some experimental results

“Cocktail party” noise recorded in the university restaurant, applying a Signal-to-noise ratio (SNR) of -1 dB:
Noisy speech

VE

![Graph showing the relationship between SNR (dB) and VE (%), with two lines representing White and Cocktail noise conditions.]

Sergio Roa, Maren Bennewitz and Sven Behnke

$F_0$ Estimation based on PSHF
Noisy speech

Estimation based on PSHF
Noisy speech

GPE

SNR (dB) vs. GPE (%)

- White
- Cocktail

SNR (dB):
0 2 4 6 8 10

GPE (%):
0 1 2 3 4 5 6 7 8 9 10

Sergio Roa, Maren Bennewitz and Sven Behnke

$F_0$ Estimation based on PSHF
Noisy speech

RMSE

---

SNR (dB)

RMSE (%)

White

Cocktail

---

Sergio Roa, Maren Bennewitz and Sven Behnke

$F_0$ Estimation based on PSHF
Conclusions

- PSHF-BA yields good results for clean speech.
- Results for noisy speech are also acceptable.
- Pitch tracking must be improved in presence of multiple speakers.
The End...

Thanks for your attention!

Sergio Roa, Maren Bennewitz and Sven Behnke

$F_0$ Estimation based on PSHF