SMOOTH PITCH TRACKER BASED ON HARMONIC AND NOISE MODEL

Ing. Igor Szöke, Doctoral Degree Programme (2)
Dept. of Computer graphics and multimedia, FIT, BUT
E-mail: szoke@fit.vutbr.cz

Supervised by: Dr. Jan Černocký

ABSTRACT

The paper deals with an approach to a smooth and precise fundamental frequency estimator with voiced/unvoiced decision. Our method is based on a parametric Harmonic and Noise Model used especially in speech modification and synthesis. Errors between speech signal and resynthesized ones with different estimations of fundamental frequencies are computed. Dynamic programing is used to retrieve the best path in an “error map”. It uses a constant for setting smoothness of the path. Reliable voiced/unvoiced segmentation can be done using thresholding of the error.

1 INTRODUCTION

Fundamental frequency estimation (*pitch tracking*) and voiced/unvoiced (*V/UV*) decision are one of the essential preprocessing steps in applications like speech synthesis, coding and parameterization. Our recent work was aimed at a text-to-speech synthesis using Harmonic and Noise Model (*HNM*) [3]. The quality of synthesized speech (especially using HNM) depends on reliable pitch tracker and V/UV decision. HNM uses pitch synchronous windowing. Unvoiced parts are windowed too, but they have no pitch, so windows have constant length (as was presented in [2]). In a case V/UV decision is estimated incorrectly and/or a track of estimated fundamental frequency is discontinuous, disturbing artefacts can be present in the synthesized speech.

Common systems for fundamental frequency estimation using correlation function are fast, but not precise. This lead us to propose a reliable pitch tracker with V/UV decision.

2 THE ALGORITHM

Speech signal is windowed (e.g. 10 *ms* length rectangular windows). If we want to find voiced parts of speech, we have to find harmonic component. Harmonic component is composed of harmonics F_0^n of fundamental frequency. So we can synthesize speech signal $S(t)$ of n harmonics of fundamental frequency F_0. But we don’t know the F_0. It usually lies in interval $\langle 50\ Hz, 350\ Hz\rangle$ for adults. So we can try all frequencies and we will look for minimal difference of original and synthesized speech signal $E(t) = O(t) - S(t)$ (Figure 1).

2.1 ERROR COMPUTATION

We try to minimize error between original speech signal $O(t)$ and synthesized harmonic speech signal $S_{F_0}(t)$ (for given fundamental frequency F_0). We have to know parameters of harmonic components (amplitudes and phases) to do this. The determination of parameters is done by FFT. For a given fundamental frequency F_0 we pick for example the first 10 harmonics parameters and we synthesize speech signal $S_{F_0}(t)$ from them:

$$S_{F_0}(t) = \sum_{i=1}^{n} A_i \cos(iF_0 + \varphi_i), \tag{1}$$

where amplitudes A_i and phases φ_i are given by FFT. Next step is to minimize the error between $O(t)$ and $S_{F_0}(t)$ to avoid imprecisions caused by different energies of those two signals. We can write

$$G_{F_0} = \frac{\sum_t O(t) S_{F_0}(t)}{\sum_t S_{F_0}^2(t)}, \tag{2}$$

where G_{F_0} is a gain and we obtain the error for given F_0:

$$E_{F_0} = \sum_t O(t) - S_{F_0}(t) G_{F_0}. \tag{3}$$

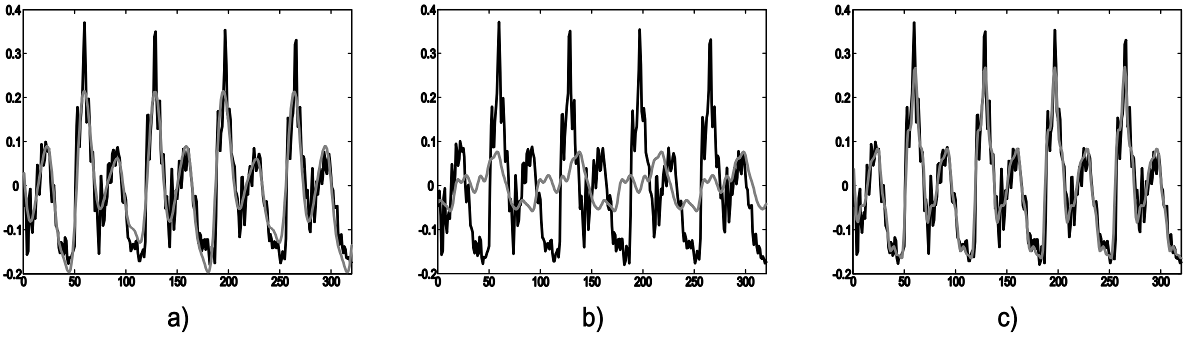

Figure 1: Original speech signal $O(t)$ (black) and synthesized speech signal $S_{F_0}(t)$ (gray). The signal was synthesized from 10 harmonics. a) $S_{F_0}(t)$ for fundamental frequency $F_0 = 118\ Hz$ (incorrect (halved) F_0). b) $S_{F_0}(t)$ for fundamental frequency $F_0 = 200\ Hz$ (incorrect F_0). c) $S_{F_0}(t)$ for fundamental frequency $F_0 = 235\ Hz$ (correct F_0).

We can compute errors E_{F_0} for fundamental frequencies in interval $F_0 \in \langle 50\ Hz, 350\ Hz\rangle$. Errors for one frame and some number of harmonics $n \in \langle 1, 5\rangle$ are shown in Figure 2. As the number of harmonic components in synthesized speech increases, local minima in harmonics of pitch appear. Examples of error curves for voiced (contains harmonic component) and unvoiced (only noise component) are given in Figure 2. Variances of noise error line are negligible.

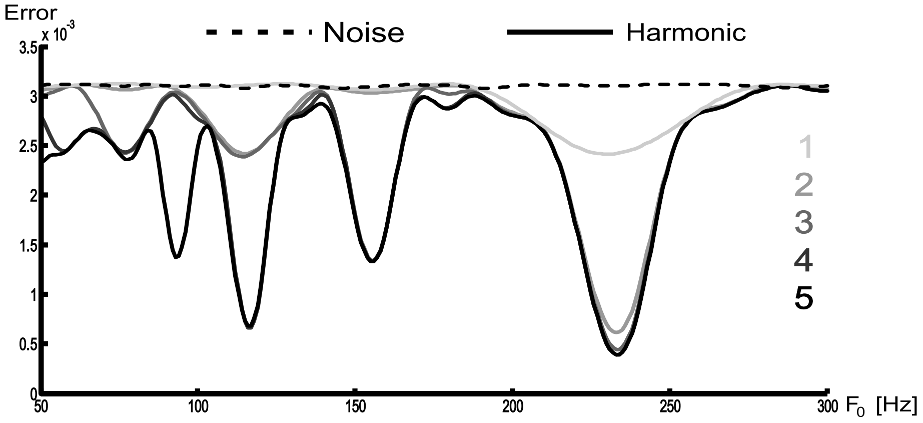

Figure 2: Error lines for harmonic signal synthesized using 1, 2, 3, 4 and 5 harmonics (solid lines) and error line for noise signal synthesized using 3 harmonics (dashed line).

As we can see in Figure 2, there are some local minima in harmonics of pitch period (halves of fundamental frequencies). Sometimes (for some number of harmonics) it happens that this local minimum overrides the "correct" minimum. To avoid this, we sum error lines for all numbers of harmonics from 1 to 5:

$$E^s_{F_0}(t) = \sum_{n=1}^{5} E^n_{F_0}(t), \tag{4}$$

where $E^n_{F_0}(t)$ denotes error line for synthesized speech signal containing n harmonics. A "map with valleys and ridges" originates from all $E^s(t)$ for all frames (see Figure 3).

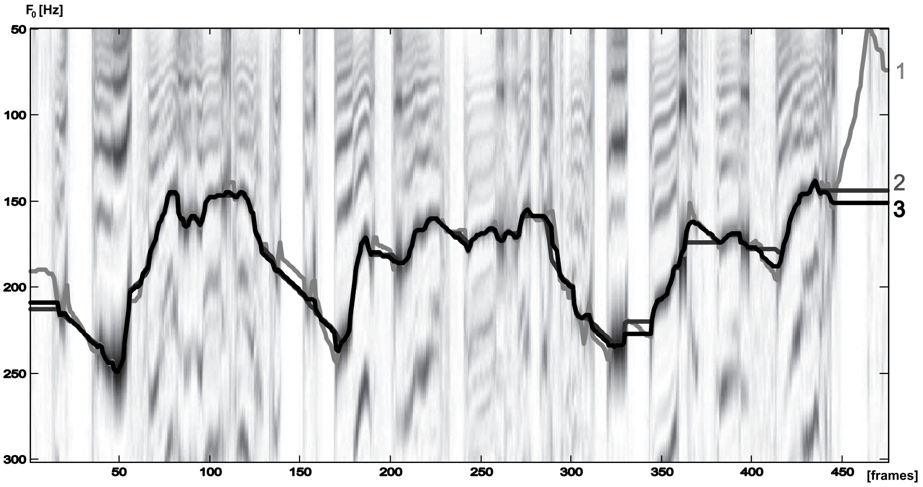

Figure 3: 2D map of error. X-axis is time (frames), y-axis is fundamental frequency F_0. Black color means low error between original speech signal $O(t)$ and synthesized signal $S_{F_0}(t)$ (for given F_0), white color means high error. 1) Path with the lowest level priority. 2) Optimal path. 3) Path with the shortest priority.

2.2 BEST PATH SEARCHING

Fundamental frequency F_0 can be found by looking for minimal error $\min_{F_0}\{E^s_{F_0}(t)\}$ in each voiced frame independently. Sometimes minimal error can be unfortunately half of F_0. We also need a pitch (sampling frequency) for unvoiced parts of speech. Dynamic programming can solve these two problems [1].

It has two passes, a *forward pass* and a *backward pass*. In the **forward pass** we sum error for all possible paths. To each $F_0(t+1)$, we can go from $F_0'(t) \in \langle F_0(t+1) - \Delta F_0, F_0(t+1) + \Delta F_0 \rangle$ where ΔF_0 is a maximal possible change of fundamental frequency (e.g. 20 Hz). We select the winner $W_{F_0}(t+1)$ for each $F_0(t+1)$ and remember from which $F_0'(t)$ it comes. The criterion for winning is the minimal path cost defined as:

$$W_{F_0}(t+1) = W_{F_0'}(t) + E^s_{F_0}(t+1) + C\sqrt{(F_0'(t) - F_0(t+1))^2 + 1} \quad \frac{E^s_{F_0'}(t) + E^s_{F_0}(t+1)}{2}, \quad (5)$$

where:

$W_{F_0}(t+1)$ is path cost to actual frame $t+1$ and fundamental frequency F_0

$W_{F_0'}(t)$ is path cost to the winner (previous frame t) and fundamental frequency F_0'

$E^s_{F_0}(t+1)$ is actual error (for $(t+1, F_0)$)

C is weighting coefficient

$\sqrt{(F_0'(t) - F_0(t+1))^2 + 1}$ is path length from (t, F_0') to $(t+1, F_0)$

$\frac{E^s_{F_0'}(t)+E^s_{F_0}(t+1)}{2}$ is average altitude (error)

The error function E^s was normalized to interval $\langle 0, 1 \rangle$ for stable results with coefficient C. The coefficient C adjusts weight between the lowest path and the shortest path (Figure 3). The core of equation 5 is the average altitude. When we are in voiced part of speech (we go in valley – the error (altitude) is low), the average altitude switches to lowest path priority. When we go on a ridge (we are in unvoiced part of speech – the error (altitude) is high), the average altitude switches to shortest path priority. This is exactly what we need: precise pitch detection and smooth straight path between voiced parts (Figure 3). The **backward pass** starts at the end (last frame). We find the minimum winner at the end. Now we go back through the map from the winner till the start. This is the reason we have to remember from which point (t, F_0') we went to point $(t+1, F_0)$.

2.3 VOICED/UNVOICED DECISION

After having the best path through the map, we must decide which parts of speech signal are voiced a which are not. The error between original speech signal $O(t)$ and synthesized speech signal $S(t)$ with fundamental frequency $F_0(t)$ is the best way. The error is related to portion of the voiced component in the speech signal. A simple threshold can be used. Frames with error lower than threshold are set voiced others are set unvoiced. Median filtering may be used for smoothing voiced/unvoiced decision function. Example can be seen in Figure 4.

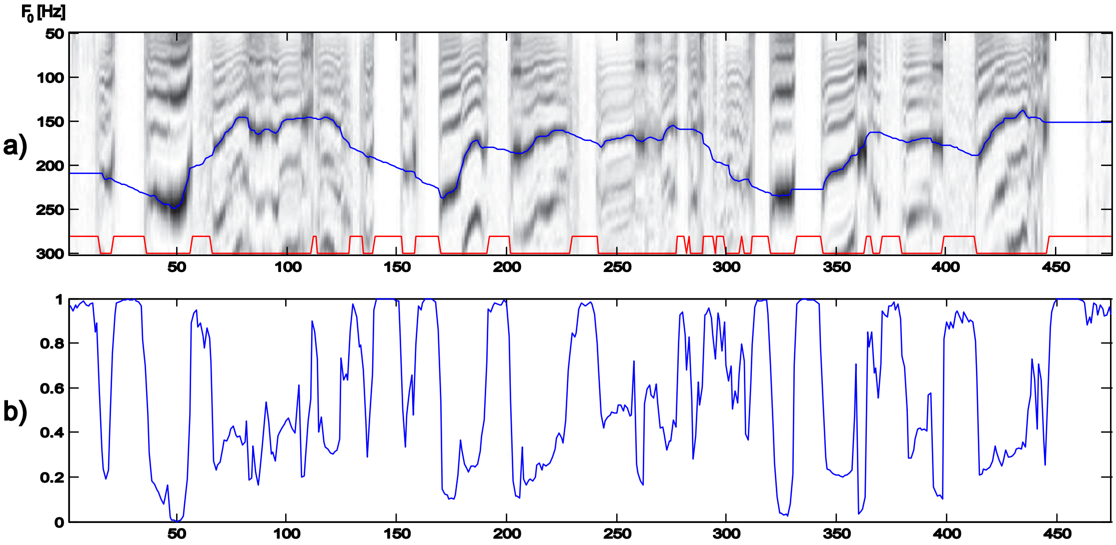

Figure 4: a) 2D map of error with the best path. b) Error run on the best path. After thresholding, we obtain V/UV decision function (bottom of panel a).

3 CONCLUSION

A precise algorithm to obtain fundamental frequency and voiced/unvoiced decision from sample of speech signal has been presented. The method has been successfully implemented and built into HNM synthesis [3] and some tools for pitch modifications based on HNM. The improvement against simple correlation based pitch detectors was significant. We are currently implementing the method for on-line pitch tracking and perform some optimizations for speed-up without sacrifying the quality.

ACKNOWLEDGEMENTS

This work was partially supported by Grant Agency of Czech Republic under project No. 102/05/0278. Jan Cernocky was supported by post-doctoral grant of Grant Agency of Czech Republic No. GA102/02/D108.

REFERENCES

[1] Huang, X. Acero, A. Hon, H. W.: Spoken Language Processing. Microsoft, Redmond, WA, USA, 1 edition, oct 2000.

[2] Stylianou, I.: Harmonic plus Noise Models for Speech, combined with Statistical Methods, for Speech and Speaker Modification. PhD thesis, École nationale supérieure des Télécommunications (ENST), Paris, jan 1996.

[3] Szöke, I.: Text-to-speech system with prosody. Master's thesis, Brno University of Technology, faculty of Information Technology, Czech Republic, jun 2003.