

Tracking pitch contours using minimum jerk trajectories

D. Neiberg, G. Ananthkrishnan, J. Gustafson

Department of Speech, Music and Hearing, Royal Institute of Technology (KTH), Sweden

neiberg@speech.kth.se, agopal@kth.se, jocke@speech.kth.se

Abstract

This paper proposes a fundamental frequency tracker, with the specific purpose of comparing the automatic estimates with pitch contours that are sketched by trained phoneticians. The method uses a frequency domain approach to estimate pitch tracks that form minimum jerk trajectories. This method tries to mimic motor movements of the hand made while sketching. When the fundamental frequency tracked by the proposed method on the oral and laryngograph signals were compared using the MOCHA-TIMIT database, the correlation was 0.98 and the root mean squared error was 4.0 Hz, which was slightly better than a state-of-the-art pitch tracking algorithm included in the ESPS. We also demonstrate how the proposed algorithm could be applied when comparing with sketches made by phoneticians for the variations in accent II among the Swedish dialects.

Index Terms: pitch tracking, Constant-Q, Swedish accent II

1. Introduction

The field of estimating the pitch of an audio signal has always been challenging. A number of algorithms has been proposed to solve the problem (for surveys see [1, 2]). Most of the algorithms estimate the fundamental frequency (F0) of the speech signal, which is highly correlated to the perceived pitch in most cases. With the advent and advance of F0 tracking algorithms, phoneticians have started adopting software programs which can present an estimate of the pitch in the form of fundamental frequencies. The automatic estimates of pitch have been useful in several instances, such as in understanding intonation [3], rhythm [4], stress [5], overall prosody [6], studying tone based languages [7] and human-human conversation [8].

There are some limitations to using such F0 detection based algorithms to model pitch conveniently. One of the problems is the minute differences in pitch that are detected by these algorithms may not be perceived by listeners and may not even be intended [9]. Secondly, an F0 value is not computable in the unvoiced regions of speech, such as in the silence before the bursts in plosives. In order to model linguistic phenomena, such as lexical stress as well as para-linguistic phenomena such as emotion, the derivative of the frequency is often used along with the pitch contours. This results in problems while modeling pitch, both mathematically [10] as well as conceptually [11]. Some researchers perform a post-processing step in order to interpolate and smoothen the estimated F0 values, before modeling it as piece-wise linear functions [3], conceptually similar to how trained phoneticians were drawing smooth trajectories by hand-drawn sketches [5]. We intend to model pitch with smooth trajectories, which provide a reasonable visual representation and a good conceptual basis for understanding both lexical as well as supra-segmental phenomena in speech. Since hand-drawn sketches of pitch contours, as shown in Fig. 1, are

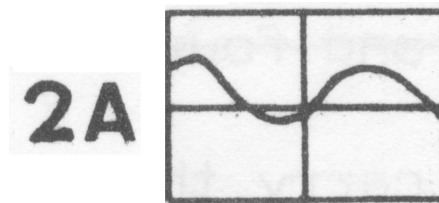


Figure 1: Example of a hand-drawn pitch contour to model the Accent II in a certain Swedish dialect (Reproduced from [5] with permission).

the motivation for our model, we apply a model for hand motor-movement, namely, the minimum jerk (the derivative of acceleration) trajectory model [12]. This phenomenon has also been observed in other muscle movements where it is proposed that resultant motion tends to follow a minimum jerk trajectory. The pitch is altered by changing the tension of the muscles in the vocal folds. Although there is no strong evidence that movements of these muscles also follow the minimum jerk trajectory model, we propose that the pitch is intended to be produced in such a way that it minimizes jerk in the log-frequency space. We believe that this sort of modeling not only provides a strong basis for conceptual modeling pitch, but also improves the reliability of the pitch tracker. In any case, since we want to apply our pitch tracking algorithm to model the way phoneticians sketch pitch contours, the sketches definitely would tend to follow minimum jerk trajectories, which we try to include in our model.

The evaluation is carried out by comparing the performance of the algorithm when applied on laryngograph data with the case when it was applied on vocal acoustic data. We hold the estimate on the laryngograph data as ground truth. Using this evaluation method, we first show that applying the minimum jerk model to the pitch tracking problem provides a reasonable pitch estimation algorithm, by comparing our algorithm with a standard pitch tracking software using the Entropics Signal Processing Software (ESPS) [13] algorithm. We then show that the proposed method models the perception of pitch suitably by comparing the trajectories estimated by the algorithm with hand-sketched pitch contours made at a time when the use of F0 estimating software was not common among phoneticians. We hope that this method of pitch tracking is a more suitable tool for phoneticians to model pitch contours.

2. Algorithm

The algorithm basically applies the minimum jerk trajectory model to a variant of the subharmonic summation [14] (SHS) algorithm. There are thus two stages to the algorithm 1) Initial F0 estimates based on the SHS algorithm 2) iterative re-estimation

by applying the equations of the minimum jerk trajectory on multiple hypotheses of the successive estimates.

2.1. Initial F0 estimation

The initial estimate is a variation of the simple SHS algorithm. The main difference is in the use of a Constant-Q filter bank [15] with a corresponding Q factor of $1/(2^{1/12} - 1)$ or 16.8 which corresponds to the 12 semitones per octave in a musical scale. This was considered as the proper psychoacoustic pitch scale to study intonation by [16]. We first compute $X(k, t) : 1 \leq k \leq K, 1 \leq t \leq T$, the output of the Constant-Q filter bank for the time frame t and filter k . Since the scale is logarithmic, the harmonics are no longer integral multiples of the fundamental. The number of Constant-Q filters between the fundamental frequency, f , and its i^{th} harmonic is given by

$$N_f(i) = \text{round}(12 \log_2 i) \quad (1)$$

An approximation for tone in noise separation is used here which classifies all frequencies with amplitudes below 10 dB from the highest amplitude frequency component as noise. So any local maximum above this threshold occurring in the output of the filter-bank is considered as tones. Thus, we obtain \hat{k} the lowest frequency that may considered a tonal component in the entire utterance.

Several frequency bins are created, for which the amplitudes of the harmonically related frequencies are summed. This gives the harmonic strength, S_h , at every time frame in utterance.

$$\forall k : \hat{k} \leq k \leq K, \quad S_h(k, t) = \sum_{i=1}^{m_h} |X(k + N_k(i), t)|^2 \quad (2)$$

where m_h are the number of harmonics to be considered.

The initial estimate for the instantaneous F0 (F_{in}) is the maximum among the summed harmonic frequency bins.

$$F_{in}(t) = \arg \max_{\hat{k} \leq k \leq K} S_h(k, t) \quad (3)$$

2.2. Minimum Jerk Interpolation on Multiple Hypotheses

Hogan [17] noted that smoothness can be quantified as a function of jerk, \ddot{x} , which is the time derivative of acceleration and the third time derivative of position variable, x

$$\ddot{x}(t) = \frac{d^3 x(t)}{dt^3} \quad (4)$$

In our context, at the n^{th} iteration, the position variable is assumed to be the n^{th} estimate of the fundamental $F_n(t)$. The value that F_n can assume at time t , is to be selected from among several hypotheses, namely the various frequency bins, $[1, 2 \dots K]^T$

The smooth trajectory exhibiting minimum jerk for a time window $[t - w_s, t + w_s]^T$, (from here on, expressed as a vector \bar{t}) can be estimated by finding the minimum mean square error (MSE) solution to the following equation.

$$(\Xi - \Gamma * P)^T * \text{diag}(\Phi) * (\Xi - \Gamma * P) = 0 \quad (5)$$

where $P^{5 \times 1}$ are the parameters of the minimum jerk trajectory

(matrix dimensions are in superscript), $\Xi^{3*(2*w_s+1)*\bar{K} \times 1}$ is

$$\Xi = \begin{bmatrix} \bar{1} \\ \bar{0} \\ \bar{0} \\ \bar{2} \\ \bar{0} \\ \bar{0} \\ \vdots \\ \bar{K} \\ \bar{0} \\ \bar{0} \end{bmatrix} \quad (6)$$

where $\bar{c} \in \mathbb{I}$ denotes a constant integer column vector, the same length as \bar{t} .

$\Gamma^{3*(2*w_s+1)*\bar{K} \times 6}$ is given by

$$\Gamma = \begin{bmatrix} \bar{1} & \bar{t} & \bar{t}^2 & \bar{t}^3 & \bar{t}^4 & \bar{t}^5 \\ \bar{0} & \bar{1} & 2\bar{t} & 3\bar{t}^2 & 4\bar{t}^3 & 5\bar{t}^4 \\ \bar{0} & \bar{0} & \bar{2} & 6\bar{t} & 12\bar{t}^2 & 20\bar{t}^3 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \bar{1} & \bar{t} & \bar{t}^2 & \bar{t}^3 & \bar{t}^4 & \bar{t}^5 \\ \bar{0} & \bar{1} & 2\bar{t} & 3\bar{t}^2 & 4\bar{t}^3 & 5\bar{t}^4 \\ \bar{0} & \bar{0} & \bar{2} & 6\bar{t} & 12\bar{t}^2 & 20\bar{t}^3 \end{bmatrix} \quad (7)$$

Vector Φ provides weight for each hypothesis from, \hat{K} , for each time instance in \bar{t} . How the weights are determined, is mentioned in the following sub-section.

Using the parameters, P , which minimize equation 5, the new smoothed trajectory $F_n(\bar{t})$ can be found by the following equation

$$F_n(\bar{t}) = [1 \bar{t} \bar{t}^2 \bar{t}^3 \bar{t}^4 \bar{t}^5] * P \quad (8)$$

We compute $F_n(\bar{t})$ by shifting the window by w_s frames. Since one would have two estimates for the overlapped regions, we apply a Bartlett window on every estimate, thus averaging the two estimates.

2.3. Iterative Interpolation

To increase robustness, an iterative method is used, by weighting subsequent F0 estimates with the previous estimates, starting with the initial estimate, $F_{in}(t)$. The 0th iteration is simply a mean smoothing, where every point in the window \bar{t} is replaced by its mean value, giving $F_0(t)$. For subsequent estimates, an exponential window is applied around the previous estimate for each time instants, t , as follows.

$$w_t(k) = 2^{-|k - \text{round}(F_n(t))|} \quad (9)$$

The weights corresponding to Φ are then obtained from the matrix of power amplitudes from the Constant-Q filter, $|X(k, t)|^2 w_t(k)$. The values are restructured (and repeated for velocity and acceleration parameters) in order to correspond to the matrix dimensions $\Phi^{3*(2*w_s+1)*\bar{K} \times 1}$. Using this weighting method, one can give more importance to the voiced components of the spectrum, both in frequency as well as the neighboring time frames while improving every subsequent estimate. This algorithm is thus less sensitive to the initial estimates and it offers a solution to interpolate F0 over unvoiced phonemes or silence. An example of the iterative estimates is found in Figure 2.

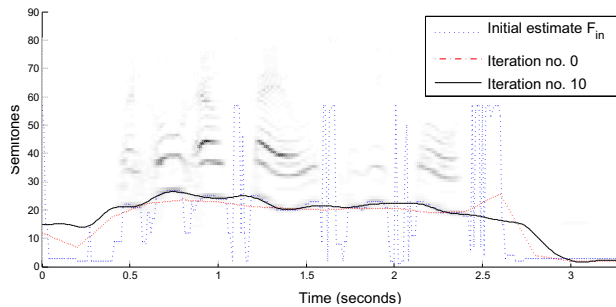


Figure 2: F0 tracking example of the sentence ‘When all else fails, use force’ uttered by subject FSEW, with $m_h = 4$ and $w_s = 200ms$ for different number of iterations, n . The F0 tracks are inlaid on the output of the Constant-Q filter-banks (X). Subsequent estimates provide both smoothing and interpolation over unvoiced regions as illustrated in the figure.

3. Evaluation of the algorithm

The F0 tracking algorithm is evaluated using the MOCHA-TIMIT corpus [18], which has simultaneous recordings of audio and laryngograph data for three speakers, FSEW (1 female), MSAK and MAPS (2 males). Each speaker read 460 phonetically rich sentences aloud. The first 20 sentences uttered by FSEW and MSAK are used as a development set for optimizing the three parameters in our algorithm. The number of harmonics to sum over, m_h , is varied between 2 and 10 in steps of 2, the window hop size, w_s is varied from 50 ms to 300 ms (i.e., 5 to 60 frames) in steps of 50 ms, and the number of iterations, n , is varied as 1, 5 or 10. We used a frame rate of 100 Hz. The performance is measured using two parameters, namely the correlation coefficient and the Root Mean Square Error (RMSE) between the F0 estimates on the audio and the laryngograph data. The results of the optimization procedure, using the RMSE criterion is shown in Table 1. The algorithm was least sensitive to the number of harmonics used for summation, m_h , which was selected to be 4 (not indicated in the table). The optimum window size is 410 ms (a hop size of 200 ms) and the optimum number of iterations are 10. These parameters are then used to evaluate the rest of the data.

Table 1: The optimization chart on the development data using the RMSE (Hz) as the evaluation metric. The number of summed harmonics, m_h , = 4

| No. Iter. (n) | Window hop size, w_s (ms) | | | | | |
|-------------------|-----------------------------|------|------|------------|------|------|
| | 50 | 100 | 150 | 200 | 250 | 300 |
| Initial Estimate | 12.3 | 12.3 | 12.3 | 12.3 | 12.3 | 12.3 |
| 1 | 13.8 | 15.0 | 14.2 | 13.0 | 9.2 | 8.9 |
| 5 | 6.1 | 4.2 | 4.9 | 5.4 | 6.2 | 7.0 |
| 10 | 4.1 | 5.3 | 4.9 | 3.6 | 4.4 | 5.0 |

It is clear from Table 1 that applying the Minimum Jerk modeling to the pitch detection, not only helps make conceptually suitable models, but in fact improves the estimates of the F0, quite substantially. The SHS algorithm was perceptually motivated, suggesting that the (inner ear) cochlea performs a harmonic summation of individual frequency components. We extended this to model include motor movements of the larynx, which helps in integrating information not only from several frequency sub-bands, but also from neighboring time frames in order to make better estimates of the F0 (or pitch).

The performance of the proposed algorithm was then com-

pared with the ESPS F0 tracker [13]. Since our algorithm includes a smoothing procedure, one always runs the risk of obtaining extremely flat contours on both the laryngograph as well as audio data. In order to verify that this was not the case, we also compare the performance of our algorithm to the ESPS F0 tracker on the laryngograph data. Since the ESPS algorithm provides F0 estimates only on voiced regions, we only compared the pitch contours detected as voiced by the ESPS algorithm. The results are shown in Table 2. One can see that while the proposed, *MinJerk* algorithm performs slightly better than the ESPS algorithm, the *MinJerk* algorithm also follows the F0 estimated on the laryngograph by the ESPS quite accurately.

Table 2: Evaluation of the proposed algorithm (in Hz), in comparison with the ESPS algorithm

| Audio | Lar. Ref. | Corr. | RMSE | Avg. Diff. | Std. Dev. |
|----------------|----------------|-------|------|------------|-----------|
| ESPS | ESPS | 0.93 | 5.6 | -0.4 | 19.5 |
| <i>MinJerk</i> | ESPS | 0.95 | 7.6 | 2.8 | 16.4 |
| <i>MinJerk</i> | <i>MinJerk</i> | 0.98 | 4.0 | 2.1 | 10.6 |

4. Application: Long Compound Word Accent II in Swedish dialects

The accent of Swedish words may take two contrastive forms, accent I and II. There are differences in the way the accent II is realized in the Swedish spoken at different parts of Sweden. Meyer [19] had made schematic pitch contours of pronunciations in Swedish dialects. He studied the pitch contours of several speakers from different regions of Sweden and Scandinavia and described them in the form of sketches. Gårding and Lindblad [5] used these sketches to make a topological analysis by dividing Sweden into four main dialect groups. These groups are based on the number of peaks and their timing relative to the stressed syllable, as shown in Figure 3a. Compound words in Swedish have two stresses, a primary stress on the first syllable and a secondary stress on the stressed syllable of the last member of the compound. Compound words are usually of accent II type, except for South Swedish where it might be a pitch contour of accent I.

Using the speech material described by [20] we performed automatic F0 tracking which may enable us to compare the detected pitch contours with those sketched by [5]. The database consisted of four speakers (two male and two female, with a wide age span) from each of 18 regions in Sweden, read the sentence ‘‘Mobiltelefonen är nittioalets stora fluga, både bland företagare och privatpersoner’’, translated to English as ‘‘The mobile phone is the big hit of the nineties both among business people and private persons’’. In this study, the compound word ‘‘Mobiltelefonen’’, an utterance initial noun was manually segmented from the sentences. These recordings were carefully chosen by two linguistic researchers as a subset of the Swedish speech database SpeechDat [21], which contains recordings made over the telephone. Speakers from 16 of the regions were assigned to one of the groups that was described by [5]. The geographical assignment and the number of speakers per type were South: TYPE1A (N = 12); Gotland: TYPE1B (N = 4); East: TYPE2A (N = 24) and West: TYPE2B (N = 24). Figure 3 shows a comparison between the pitch tracks sketched for the four types and a model of the pitch contour, the prototypical F0, made automatically from the several utterances that were recorded from the respective regions. The prototypical F0 plot is obtained by first normalizing the F0 estimates along time

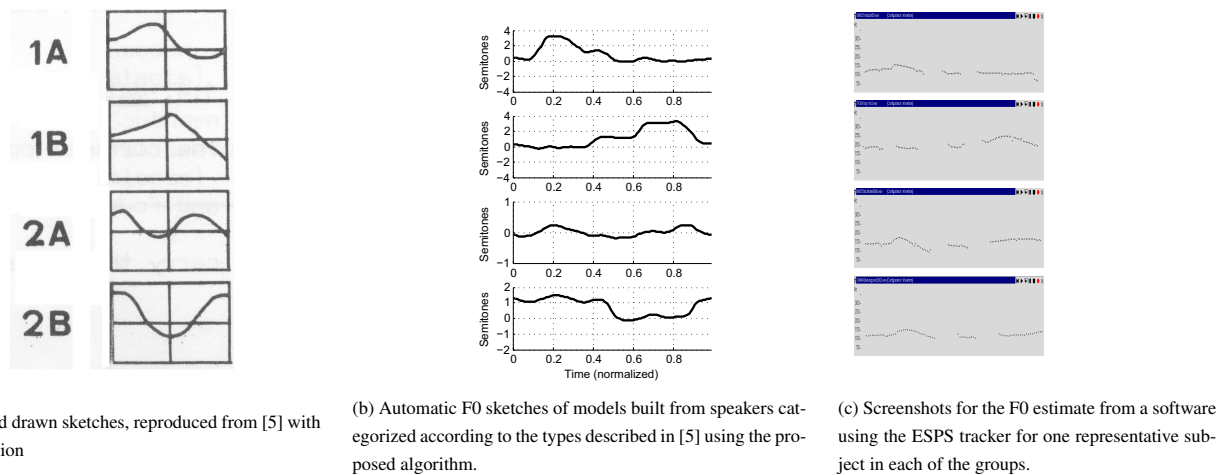


Figure 3: Comparison of the hand drawn sketches representing the various dialectal groups [5] with the automatic algorithms.

and frequency, then averaging over the several instances of the resulting spectrograms, as in [8]. Then automatic sketches are drawn using the proposed pitch tracking algorithm. One can see from Figure 3b, that the estimates from our proposed algorithm are close to the original sketched pitches. One needs a lot more imagination to find the correspondences between the output of the ESPS algorithm (Figure 3c) and the original sketches. The discontinuity in the pitch estimates, also makes modeling and subject independent generalization difficult.

5. Conclusions

In this paper, we have proposed and evaluated a pitch tracker, by integrating the estimation with the modeling of pitch as a minimum jerk trajectory. After incorporating the model, the proposed algorithm has a slightly better performance than the ESPS F0 tracker. As an application, we compared the pitch tracings for accent II in Swedish, sketched by a phonetician. The method is capable of producing pitch sketches which are quite similar to the conceptual models that were sketched by hand. We hope to apply this model for training second language learners and hearing impaired subjects to gain a better conceptual understanding of prosodic phenomena. This algorithm could also be used as a tool for phoneticians to explore and automatically model prosodic phenomena in speech. Another application is to see if perceptually motivated pitch contours are more suitable for natural sounding speech synthesis, rather than acoustically motivated pitch contours.

6. Acknowledgements

Funding was provided by the Swedish Research Council (VR) projects 2009-4291 and 2009-4599.

7. References

- [1] De Cheveigné, A., "Pitch perception models – a historical review," CNRS-Ircam, Paris, France, 2004.
- [2] Hess, W. et al., *Pitch determination of speech signals: algorithms and devices*, vol. 84, Springer-Verlag, 1983.
- [3] 't Hart, J., "A phonetic approach to intonation: from pitch contours to intonation patterns," *Intonation, Accent and Rhythm*, 1984.
- [4] Santen, J. and Hirschberg, J., "Segmental effects on timing and height of pitch contours," in *Third International Conference on Spoken Language Processing*, 1994.
- [5] Gårding, E. and Lindblad, P., "Constancy and variation in Swedish word accent patterns," in *Working Papers 7*, Lund, Lund University, 36–110, 1973.
- [6] Mertens, P., "The Prosogram: Semi-automatic transcription of prosody based on a tonal perception model," in *Proceedings of Speech Prosody*. Citeseer, 23–26, 2004.
- [7] Xu, Y., "Effects of tone and focus on the formation and alignment of F0 contours," *Journal of Phonetics*, 27(1):55–105, 1999.
- [8] Neiberg, D. and Gustafson, J., "The prosody of Swedish conversational grunts," in *Proc. of Interspeech*, Sept. 2010.
- [9] 't Hart, J., "Differential sensitivity to pitch distance, particularly in speech," *The Journal of the Acoustical Society of America*, 69(3):811, 1981.
- [10] Laskowski, K., Wolfel, M., Heldner, M., and Edlund, J., "Computing the fundamental frequency variation spectrum in conversational spoken dialogue systems," *Journal of the Acoustical Society of America*, 123(5):3427–3427, 2008.
- [11] Spaai, G. and Hermes, D., "A visual display for the teaching of intonation," *Calico Journal*, 10:19–19, 1993.
- [12] Viviani, P. and Flash, T., "Minimum-jerk, two-thirds power law, and isochrony: converging approaches to movement planning," *J Exp Psychol*, 21:32–53, 1995.
- [13] Talkin, D., *Speech Coding and Synthesis*, chapter A Robust Algorithm for Pitch Tracking (RAPT), Elsevier, 1995.
- [14] Hermes, D. J., "Measurement of pitch by subharmonic summation," *J. Acous. Soc. Am.*, 83(1):257–264, 1988.
- [15] Brown, J., "Calculation of a constant q spectral transform," *Journal of the Acoustical Society of America*, 89(1):425–434, January 1991.
- [16] Nolan, F., "Intonational equivalence: an experimental evaluation of pitch scales," in *Proc. International Congress of Phonetic Sciences*, 771:774–778, 2003.
- [17] Hogan, N., "Adaptive control of mechanical impedance by coactivation of antagonist muscles," *Automatic Control, IEEE Transactions on*, 29(8):681–690, Aug 1984.
- [18] Wrench, A., "The MOCHA-TIMIT articulatory database," Queen Margaret University College, Tech. Rep, 1999.
- [19] Meyer, E. A., "Die intonation im schwedischen, i: Die sveamundarten," in *Studies Scand. Philol.* Stockholm University, (10), 1937.
- [20] Beskow, J., Bruce, G., Enflo, L., Granström, B., and Schötz, S., "Recognizing and Modelling Regional Varieties of Swedish," in *Interspeech 2008*, 2008.
- [21] Elenius, K., "Two Swedish speechdat databases - some experiences and results," in *Eurospeech 99*, 2243–2246, 1999.