# A Cognition-Oriented Approach to Fundamental Frequency Estimation

Ulrike Glavitsch and Klaus Simon

EMPA, Swiss Federal Laboratories for Materials Science and Technology
Ueberlandstrasse 129, 8600 Dübendorf, Switzerland
ulrike.glavitsch@empa.ch, klaus.simon@empa.ch

**Abstract.** This paper presents an efficient, two-phase fundamental frequency detection algorithm in the time-domain. In accordance with the human cognitive process it first computes base fundamental frequency estimates, which are verified and corrected in a second step. The verification step proceeds from high-energy stable segments, where reliable estimates are expected, to lower-energy regions. Irregular cases are handled by computing a series of fundamental frequency variants that are evaluated for highest plausibility, in analogy with the hypothesis testing principle of human thinking. As a proof of concept, the algorithm was evaluated on a clean speech database where it shows significantly lower error rates than a comparable reference method.

## 1 Introduction

The fundamental frequency $F_0$ plays an important role in human speech perception and is used in all fields of speech research. For instance, the human brain is supposed to evaluate the positions of formants with respect to $F_0$ [1], and accurate estimates of $F_0$ are a prerequisite for prosody control in concatenative speech synthesis [2].

Fundamental frequency detection has been an active field of research for more than forty years. Early methods used the autocorrelation function [3], cepstral analysis [4] and inverse filtering techniques [5] for $F_0$ detection. In most of these approaches, threshold values are used to decide whether a frame is assumed to be voiced or unvoiced. More advanced algorithms incorporate a dynamic programming stage to calculate the $F_0$ contour based on frame-level $F_0$ estimates gained from either a conditioned linear prediction residual [6] or a normalized cross correlation function (NCCF) [7]. In the last decade, techniques like pitch-scaled harmonic filtering (PSHF) [8], Nonnegative Matrix Factorization (NMF) [9, 10] and time-domain probabilistic approaches [11] have been proposed. These achieve low error rates and high accuracies but at a high computational cost - either at run-time or during model training. These calculative approaches generally disregard the principles of human cognition and the question is whether $F_0$ estimation can be performed equally well or better by considering these.

In this paper, we propose an $F_0$ estimation algorithm based on the elementary appearance and inherent structure of the human speech signal. A period, i.e. the

inverse of $F_0$, is primarily defined as the distance between two maximum or two minimum peaks, and we use the same term to refer to the speech section between two such peaks. The speech signal can be divided into *stable* and *unstable* segments. Stable segments are those regions with a quasi-constant energy or quasi-flat envelope whereas unstable segments exhibit significant energy rises or decays. On stable segments, the $F_0$ periods are mostly regular, i.e. the sequence of maximum or minimum peaks is more or less equidistant, whereas the $F_0$ periods in unstable regions are often shortened, elongated, doubled, or may show little similarity with their neighboring periods. Speech signals are highly variable and such special cases occur relatively often. Thus, it makes sense to first compute $F_0$ estimates in stable segments and use this knowledge to find those of unstable segments in a second step. The $F_0$ estimation method for stable segments is straight-forward as regular $F_0$ periods are expected. The $F_0$ estimation approach for unstable segments, instead computes variants of possible $F_0$ continuation sequences and evaluates them for highest plausibility. The variants reflect the regular and all the irregular period cases and are calculated using a peak look-ahead strategy. We denote this $F_0$ estimation method for unstable segments as $F_0$ propagation since it computes and verifies $F_0$ estimates by considering previously computed values.

We regard the proposed algorithm as cognition-oriented inasmuch as it incorporates several principles of human cognition. First, human hearing is also two-stage process. The inner ear performs a spectral analysis of a speech section, i.e. different frequencies excite different locations along the basilar membrane and as a result different neurons with characteristic frequencies [12]. This spectral analysis delivers the fundamental frequency and the harmonics. The brain, however, then checks the information delivered by the neurons, interpolating and correcting it where necessary. Our proposed $F_0$ estimation algorithm performs in a similar way, in that the $F_0$ propagation step proceeds from regions with reliable $F_0$ estimates to ones where $F_0$ is not clearly known yet. We have observed that $F_0$ is very reliably estimated on high-energy stable segments, which typically represent vowels. Thus, we always compute $F_0$ for unstable segments by propagation from high-energy stable segments to lower-energy regions. Next, we have adopted the hypothesis testing principle of human thinking [13] for generating variants of possible $F_0$ sequences and testing them for the detection of $F_0$ in unstable segments. Lastly, human cognition uses context to decide a situation. For instance, in speech perception humans bear the left and right context of a word in mind if its meaning is ambiguous. In an analogous way, our algorithm looks two or more peaks ahead to find the next valid maximum or minimum peak for a given $F_0$ hypothesis. Special cases in unstable segments can often not be disambiguated by just looking a single peak ahead.

The resulting algorithm is very efficient, thoroughly extensible, easy to understand and has been evaluated on a clean speech database as a proof of concept. Recognition rates are clearly better than those of a reference method that uses cross-correlation functions and dynamic programming. In addition, it delivers a

segmentation of the speech signal into stable and unstable segments that may be useful for an automatic speech recognition component.

The outline of the paper is as follows. Section 2 describes the preprocessing steps of peak detection and energy computation. Section 3 presents the $F_0$ computation on stable segments. Section 4 describes the $F_0$ propagation stage. Section 5 outlines the post-processing step of computing $F_0$ on entirely unstable voiced segments. The evaluation of the algorithm is presented in Section 6. Finally, we draw conclusions and give an outlook for future work in Section 7.

## 2 Preprocessing

The first preprocessing step is the extraction of signal peaks. A peak is defined as either a local minimum or a local maximum in the sequence of signal samples. For each peak $p$, we maintain a triple of values $\langle x, y, c \rangle$ where $x$ and $y$ are the peak coordinates and $c$ is the peak classification - either a minimum or a maximum peak. In a second step, we compute the mean energies for all signal frames. A frame is a small section of the signal where successive frames overlap by some extent. We selected a frame length of 20 ms and an overlap length of 10 ms. For periodic signal parts, the mean energy must be computed on an integer multiple of a period to be meaningful. However, as the period of a frame is not known at this point in time, we therefore compute the mean energy on a scale of window lengths each of which corresponds to a different period length. An optimization step then finds the best window length for each frame. This procedure is similar to pitch-scaled harmonic filtering (PSHF) [8] where an optimal window length is calculated for finding harmonic and non-harmonic spectra. The window lengths are selected such that periods of $F_0$ between 50 and 500 Hz roughly fit a small number of times into at least one of these lengths. The selected window lengths correspond to fundamental frequencies of $50, 55, 60, \ldots, 95$ Hz. Each window length is centered around the frames middle position. The optimal window length is the one where the mean energies of a small number of frames around the frames middle position show the least variation.

## 3 $F_0$ Estimation on Stable Segments

The estimation of the fundamental frequency is first performed on stable voiced segments, i.e. voiced speech sections with a quasi-constant energy. For this purpose, the speech signal is segmented into voiced and unvoiced parts and into stable and unstable sections within the voiced regions. The method to estimate $F_0$ on stable segments is relatively straight-forward as we mainly expect regular periods. The $F_0$ estimates of stable segments are grouped into sequences of roughly equal $F_0$ values in order to provide anchor points for the $F_0$ propagation described in Section 4.

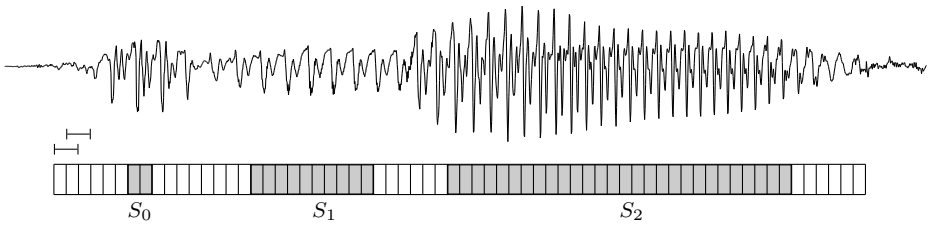


**Fig. 1.** *Overlapping frames of voiced segment of speech signal "the north" uttered by a male speaker which contains three stable segments $S_0$, $S_1$ and $S_2$ of lengths* 1, 9 *and* 27.

### 3.1 Segmentation of Speech Signal

A signal frame is voiced if its mean energy exceeds a certain threshold, the absolute height of the frames maximum or minimum peak is above a given level, and the number of zero crossings is greater or equal to some configurable number. Consecutive voiced frames represent a so-called voiced segment. A voiced frame is classified as stable if its mean energy, as computed in Section 2, does not deviate more than a given percentage from the mean energy of both the previous and the next frame. The current value is 50 %, i.e. we allow some energy deviation between neighboring frames but not too much. Consecutive sequences of stable frames form stable segments. In Fig. 1, a voiced segment of a speech signal with stable segments $S_0$, $S_1$ and $S_2$ is depicted.

### 3.2 $F_0$ Estimation Method

The $F_0$ estimation method for stable segments finds a quadruple of peaks $P = \langle p_L, p_0, p_1, p_R \rangle$ of either maximum or minimum peaks $p_i$, $i = L, 0, 1, R$, such that the center position of the frame is between $p_0$ and $p_1$. The $F_0$ estimate is the inverse of the mean of the period lengths found in $P$, i.e. the mean of the distances between peaks $p_L$ and $p_0$, $p_0$ and $p_1$ as well as $p_1$ and $p_R$. The tuple $P$ is selected among a series of possible candidate peak tuples according to some similarity score. Furthermore, it is checked whether the peak tuple is not a multiple of the supposedly true $F_0$ period, otherwise, a different peak tuple is selected. In the following, we describe the algorithm to find such a peak tuple $P$ for each stable frame.

We start by finding the peak in the frame that has the highest absolute value. We then look for candidate peaks that have a similar absolute height and whose distance from the highest peak is within the permissible range of period lengths. The search for candidate peaks is performed in the direction of the center position of the frame. Given a peak pair of the highest absolute peak and a candidate peak, the algorithm looks for the peaks to the left and right of the given pair to complete the quadruple. We select the peak with the highest absolute value above some threshold and within a tolerance range on the time axis to the left and the right of the candidate peak pair. The peak tuple $P$ may

reduce to a triple peak sequence if such a peak at one side cannot be found. Each such candidate peak quadruple or peak triple is scored and the tuple with the highest score is selected as the tentatively best candidate.

The proposed score measures the uniformness of the peaks in the peak tuple with respect to their distances and their absolute heights. The score $s$ for peak tuple $P = \langle p_L, p_0, p_1, p_R \rangle$ is the product of partial scores $s_x$ and $s_y$. The value $s_x$ measures the equality of the peak intervals whereas $s_y$ is a measure of the sameness of the absolute peak heights. The partial score $s_x$ is defined as $1 - a$, where $a$ is the root of the mean square difference between the peak distances at the tuple edges from the peak distance of the two middle peaks $p_0$ and $p_1$. Similarly, partial score $s_y$ is given as $1-b$, where $b$ is the root of the mean square difference of the absolute peak heights from the maximum absolute peak height. The equations below show how the score $s$ is computed for a peak quadruple as defined above. The formulas are easily adapted for tuples with only three peaks.

$$s = s_x s_y \tag{1}$$

The partial score $s_x$ is defined as follows:

$$s_x = 1 - a \tag{2}$$

$$a = \sqrt{(b_0^2 + b_1^2)/2} \tag{3}$$

$$b_0 = \frac{d_1 - d_0}{d_1},\ b_1 = \frac{d_1 - d_2}{d_1} \tag{4}$$

$$d_0 = x_0 - x_L,\ d_1 = x_1 - x_0,\ d_2 = x_R - x_1. \tag{5}$$

The value $x_i$, $i = L, 0, 1, R$ refers to the x-coordinate of peak $p_i$ as mentioned in Section 2.

The partial score $s_y$ is given by:

$$s_y = 1 - b \tag{6}$$

$$b = \sqrt{\frac{1}{4}(g_L^2 + g_0^2 + g_1^2 + g_R^2)} \tag{7}$$

$$g_i = (|y_i| - y_{max})/y_{max}, i = L, 0, 1, R \tag{8}$$

$$y_{max} = max(|y_i|\,|i = L, 0, 1, R). \tag{9}$$

Similarly, $y_i$ denotes the peak height of peak $p_i$ in tuple $P$, $i = L, 0, 1, R$. The score $s$ delivers exactly 1 if the peak heights and peak intervals are equal and less than 1 if they differ.

The peak tuple with the highest score may be a multiple of the true period or it may also be half of a period if the signal has a strong first harmonic. The case of a multiple period candidate is checked by testing the existence of equidistant partial peaks within the peak pair. If such partial peaks are found, we look for a candidate peak tuple with the partial peak distance and install it as the currently best candidate. We then check whether the best candidate tuple is only half of

a true period, by comparing the normalized cross correlation function (NCCF) [7] of the currently best candidate tuple with the NCCF of the tuple with the double period. If the NCCF of the former is significantly smaller than that of the latter, we install the peak tuple with the double period as the best candidate.

The final step in the $F_0$ estimation of a stable frame finds the peak tuple in the center of the frame that has the same period length as the best candidate tuple. This is achieved by looking for peaks to either the left or right side of the best candidate in the distance of the period length until a peak tuple is found where the frame's center position is between the two middle peaks.

### 3.3 Equal Sections

The last step of this stage is the detection of sequences of roughly equal $F_0$ estimates within a stable segment. These sequences are referred to as *equal sections*. The $F_0$ estimates of the frames in an equal section must not deviate by more than a given threshold from the mean $F_0$ of the equal section. The longest such equal section with a minimum length of 3 is stored as the equal section of the stable segment. The remaining equal sections of the stable segment are maintained in a list for eventualities.

## 4 $F_0$ Propagation

The $F_0$ propagation is the second major stage of the proposed $F_0$ detection algorithm. Its purpose is to calculate and check $F_0$ estimates in regions where no reliable $F_0$ estimates exist. This mainly affects unstable regions, but also portions of stable regions where e.g. the $F_0$ estimates do not belong to an equal section. The main idea is that the $F_0$ propagation starts at the stable segment with the highest energy from where it proceeds to the regions to both its left and right side. It always progresses from higher-energy to lower-energy regions. Once a local energy minimum is reached, it continues starting from the next stable segment in propagation direction that is a local energy maximum. For the verification and correction of calculated $F_0$ estimates we have developed a peak propagation procedure that computes the most plausible peak continuation sequence given the peak tuple of a previous frame. The most plausible peak sequence is found by considering several variants of peak sequences that reflect the regular and irregular period cases. In the following, we describe the control flow of the $F_0$ propagation and explain the particular peak propagation procedure.

### 4.1 Control Flow

The propagation of $F_0$ estimates is performed separately for each voiced segment. The first step in this procedure is the definition of the propagation order and the propagation end points. The propagation starts with the stable segment that contains the frame with the highest mean energy in its equal section. From this equal section the propagation flows first to the left and then to the right side.

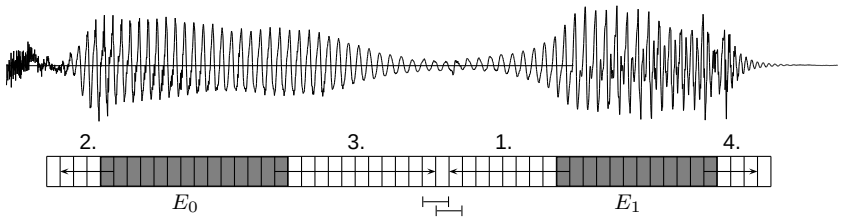


**Fig. 2.** *Propagation order, directions and end points of a voiced segment "disput**ing wh**ich" (bold part) uttered by a female speaker. Propagation start and end points are marked at the center of the corresponding frames. Propagation starts from equal section $E_1$ as it has frames with higher energies than $E_0$.*

For each stable segment containing an equal section we define the right and left propagation end points. They are the start and end frame of the voiced segment if there is only one stable segment in the voiced segment. The propagation end point is a local energy minimum frame, or its direct neighbor frame if there is a local energy minimum region between two stable segments. Fig. 2 shows the propagation directions, order and end points of a voiced segment that contains two stable segments with equal sections $E_0$ and $E_1$.

After identifying the propagation anchor points, directions and end points we compute candidate $F_0$ values for the unstable frames following the method presented in Section 3.2 but with a restricted allowable range for $F_0$. We allow an $F_0$ range of more than an octave lower and two thirds of an octave higher than the mean $F_0$ of the equal section where the propagation starts. In contrast with the $F_0$ estimation method of Section 3.2, the check for multiple periods and strong harmonics is omitted, since it would hardly work in unstable regions with potentially strongly varying peak heights.

The core part of this stage is to check whether the $F_0$ estimate of a frame is in accordance with the $F_0$ of its previous frame and if not, to perform the peak propagation step (see Section 4.2) to find the most plausible peak continuation sequence from which the frame's actual $F_0$ estimate is derived. The $F_0$ estimate of a frame may deviate from the $F_0$ of its predecessor by a given percentage. As soon as the propagation end point is reached, we check whether the mean $F_0$ of the equal section of the next stable segment is similar to the mean $F_0$ of the most recently calculated values. Propagation continues normally from the next stable segment if this condition holds, otherwise the list of equal sections for eventualities (see Section 3.3) is searched for a better fitting equal section and the algorithm uses this as the new propagation starting point.

### 4.2 Peak Propagation

The peak propagation step computes a set of peak sequence variants that may follow the peak tuple of the previous frame and evaluates them for plausibility. Each peak sequence is computed by a look-ahead strategy for the next peak. In general, we look two peaks ahead before deciding on the next one.

The following peak sequence variants are considered:

- V1 (regular case): The peaks continue at about the same distance as the peaks in the previous frame.
- V2 (elongated periods): The periods are elongated and the peak distances become larger.
- V3 (octave jump down): The peaks follow at double distance as in the previous frame.
- V4 (octave jump up): The peaks follow at half the distance as in the previous frame.

These peak sequence variants are computed depending on the octave jump state of the previous frame. The octave jump state is maintained for each frame and its default value is 'none'. There are two additional values 'down' and 'up' for the state of $F_0$ that is an octave higher than normally, and the state of an $F_0$ estimate that has fallen by an octave. Variant V2 is used to detect extended periods that may not be captured by V1. However, V2 may easily deliver too large periods, e.g. in the case of spurious peaks, requiring additional checks. V3 is necessary to test the case of a sudden octave jump down but is only calculated in the case of an octave jump state of 'none'. V4 is considered only in an octave jump down state to check whether such a phase ends. Currently, for simplicity we forbid sequences of repeated octave jumps down and also sudden octave jumps up.

For each of these variants V1 to V4, we define interval ranges where subsequent peaks are expected. These ranges are defined relative to the last peak distance $D$, i.e. $D$ is the distance between the last two peaks in propagation direction of the previously computed peak sequence. The peak sequence starts with the peak tuple of the previous frame and adds peaks to the left or to the right as propagation proceeds. Each new peak in the peak sequence is searched for in the given interval, while at the same time checking whether a peak exists in the interval that follows. Each such peak pair is scored by computing their mean absolute height. The first peak in the pair which achieves the highest such score is installed as the definite next peak in the peak sequence. The peak propagation stops as soon as the center frequency of the addressed frame has been passed by two peaks or if no further peak is found. It is deliberate that the score for the peak propagation considers only the absolute peak heights. A measure that accounted also for the peak distances would deliver false peak sequences, owing to the irregular peak distances that we expect in unstable regions. Fig. 3 shows the look-ahead strategy for successor peaks in a left direction, starting at peak $p_0$ that is part of peak tuple $\langle p_0, p_1, p_R \rangle$. Peaks $p_{k(1)}$ and $p_{k(2)}$ are inspected in interval $I_0$ from $p_0$, peaks $p_{k(2,1)}$ and $p_{k(2,2)}$ in interval $I_{k(2)}$ from peak $p_{k(2)}$ gained in the first round. The peak pair $p_{k(2)}$ and $p_{k(2,2)}$ achieves the highest score, i.e. highest mean absolute value, thus $p_{k(2)}$ is installed as the next valid peak.

The final step in the peak propagation stage is the evaluation of peak sequence variants. In general, the variant with the highest score, i.e. with the

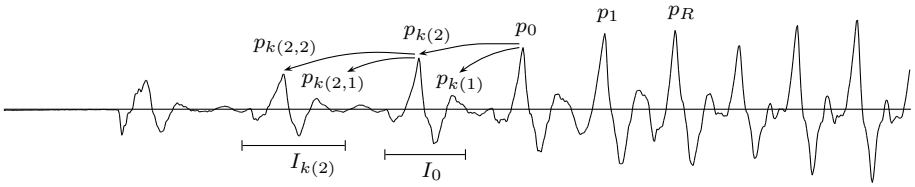


**Fig. 3.** *Peak propagation for V2 (extended period case) with tuple* $P = \langle p_0, p_1, p_R \rangle$ *in propagation direction to the left. Peaks* $p_{k(1)}$ *and* $p_{k(2)}$ *are inspected from* $p_0$ *in interval* $I_0$*, peaks* $p_{k(2,1)}$ *and* $p_{k(2,2)}$ *are found in interval* $I_{k(2)}$ *when starting from* $p_{k(2)}$

highest mean absolute peak height, is the best peak continuation sequence. However, some checks still need to be performed to verify it. Here we describe the evaluation process for the case where the previous frame has no octave jump (regular case): a similar procedure is applied if the previous frame is in an octave jump down state. In the regular case, we first check whether V1 and V2 deliver the same peak sequence. If so, we keep V1 and discard V2. Otherwise, an additional peak propagation step for the next frame is performed to see whether V2 diverges and delivers periods which are too large. In this case, V2 is discarded and V1 is kept. In all other cases, we keep the variant with the larger score, i.e. the higher absolute mean peak height, in V1. Then, if V3 has a score greater than or equal to V1, we evaluate V1 against V3. V3 is installed and the frame's octave jump state is set to 'down' only if V3 has no middle peaks of sufficient heights, i.e. if the absolute height of the middle peak is smaller than a given percentage of the minimum of the absolute heights of the enclosing peaks. Otherwise, V1 is established.

## 5 Unstable Voiced Segments

Voiced segments without stable regions, or voiced segments that have no sufficiently large subsequences of equal $F_0$, are treated in a separate post-processing step. Basically, the same propagation procedure is applied, but the propagation starting point or anchor is found using looser conditions and additional information.

First, we compute the mean $F_0$ of the last second of speech. The mean $F_0$ is calculated by considering only those frames with a reliable $F_0$ estimate, i.e. frames of voiced segments where the verification step, i.e. the $F_0$ propagation, has been performed. We then compute candidate $F_0$ values for all unstable frames of the voiced regions in the range of the mean $F_0$. The permissible $F_0$ range is the same as described in Section 4.1. The anchor point for propagation is found by inspecting the equal section list of the stable segments in the voiced regions, or a small section around the highest-energy frame if no stable segment in the voiced region exists. The propagation starts from such a section if the mean of the $F_0$ estimates does not deviate too largely from the last second's mean $F_0$. If

no such section can be found, we leave the $F_0$ estimates unchanged. In this case, no propagation takes place.

## 6 Experiments and Results

Our $F_0$ estimation algorithm was evaluated on the Keele pitch reference database for clean speech [14] as a proof of concept. We measured the voiced error rate (VE), the unvoiced error rate (UE) and the gross pitch error rate (GPE). A voiced error is present if a voiced frame is recognized as unvoiced, an unvoiced error exists if an unvoiced frame is identified as voiced and a gross pitch error is counted if the estimated $F_0$ differs by more than 20 % from the reference pitch. The precision is given by the root mean square error (RMSE) in Hz for all frames classified as correct, i.e. as neither voiced, unvoiced, nor gross pitch errors. Results for the proposed algorithm - denoted as HCog (human cognition based) - are given in Table 1. We also cite results of other state-of-the-art $F_0$ estimation methods: RAPT [7] (one of the best time-domain algorithms based on cross correlation functions and dynamic programming), and the two frequency-domain algorithms PSHF Based [8] and Nonnegative Matrix Factorization (NMF) [9].

**Table 1.** *Results obtained on the Keele pitch reference database*

| | VE (%) | UE (%) | GPE (%) | RMSE (Hz) |
|---|---|---|---|---|
| HCog | 2.53 | 4.46 | 1.49 | 5.09 |
| RAPT | 3.2 | 6.8 | 2.2 | 4.4 |
| PSHF | 4.51 | 5.06 | 0.61 | 2.46 |
| NMF | 7.7 | 4.6 | 0.9 | 4.3 |

The results show that the proposed algorithm performs excellently in terms of VE and UE. None other of the cited algorithms shows such low VE and UE. The GPE, at 1.49 %, is clearly lower as for RAPT but not as low as for the frequency-domain algorithms. However, a GPE of 1.49 % in the presence of a VE of only 2.53 % is very low. A higher VE may also hide several gross pitch errors.

The RMSE, at 5.09 %, is higher than with the other algorithms. We see two reasons for this. First, maximum and minimum peaks often have an inclination - either to the left or to the right - and often, there is a set of close peaks around the maximum or minimum peak so that $F_0$ is not as accurately calculated as with other methods. Second, it may occur that the leftmost or rightmost peak of a peak tuple is not the true period end point due to thresholds selected and the fact that propagation is started from suboptimal peaks. However, accuracy can certainly be improved by adjustment procedures and smoothing.

## 7 Conclusions

We have presented an $F_0$ estimation algorithm as an approximate model of the human cognitive process. The algorithm achieves very low error rates, outperforming the state-of-the-art correlation-based reference method in this respect. These results are achieved with little resources in terms of memory and computing power. Obviously, the strength and potential of the algorithm lie in the concepts which simulate human recognition of $F_0$.

The algorithm is thoroughly extensible, as new special cases are easily implemented. In this sense, the algorithm can also be applied to other tasks, e.g. spontaneous speech, by analyzing the new cases and modeling them. In this way it will become more and more generic. This procedure closely reflects human learning, which is said to function by adopting examples and building patterns independently of the frequency or probability of their occurrence [15]. For this reason, we have refrained from using weights or probabilities to favor one or another case but look ahead and evaluate until the case is decided.

Our algorithm delivers a classification of the speech signal into stable and unstable segments additionally to the $F_0$ contour. Automatic speech recognition systems may profit from this classification since the recognition of phonemes should preferably be started from stable segments as well. The spectral information required for phoneme recognition is certainly more reliably computed on those segments.

Future work will focus on extending the algorithm to other tasks and improving the accuracy of the $F_0$ estimates.

## 8 Acknowledgements

The authors wish to thank Prof. Jozsef Szakos from Hong Kong Polytechnic University for valuable comments on this paper. We thank Matthias Scheller Lichtenauer and Iris Sprow from Swiss Federal Laboratories for Materials Science and Technology, as well as Prof. Guy Aston from University of Bologna for their careful proof-reading.

## References

1. Traunmüller, H.: Paralinguistic Variation and Invariance in the Characteristic Frequencies of Vowels. Phonetica **45**(1) (1988) 1–29
2. Ewender, T., Pfister, B.: Accurate Pitch Marking for Prosodic Modification of Speech Segments. In: Proc. of Interspeech. (2010) 178–181
3. Rabiner, L.R., Cheng, M.J., Rosenberg, A.E., McGonegal, C.A.: A Comparative Performance Study of Several Pitch Detection Algorithms. IEEE Transactions on Acoustics, Speech, and Signal Processing **24**(5) (October 1976) 399–418
4. Noll, A.M.: Short-Time Spectrum and 'Cepstrum' Techniques for Vocal-Pitch Detection. J. Acoust. Soc. Am. **36**(2) (February 1964) 296–302
5. Markel, J.D.: The SIFT Algorithm for Fundamental Frequency Estimation. IEEE Transactions on Audio and Electroacoustics **20**(5) (December 1972) 367–377

6. Secrest, B.G., Doddington, G.R.: An integrated pitch tracking algorithm for speech systems. In: Proc. of ICASSP. (1983) 1352–1355
7. Talkin, D.: A Robust Algorithm for Pitch Tracking (RAPT). In Kleijn, W.B., Paliwal, K.K., eds.: Speech Coding and Synthesis, Elsevier Science B. V. (1995)
8. Roa, S., Bennewitz, M., Behnke, S.: Fundamental frequency estimation based on pitch-scaled harmonic filtering. In: Proc. of ICASSP. (2007) 397–400
9. Sha, F., Saul, L.K.: Real-Time Pitch Determination of One or More Voices by Nonnegative Matrix Factorization. In: Advances in Neural Information Processing systems 17, MIT Press (2005) 1233–1240
10. Peharz, R., Wohlmayr, M., Pernkopf, F.: Gain-robust multi-pitch tracking using sparse nonnegative matrix factorization. In: Proc. of ICASSP. (2011) 5416–5419
11. Achan, K., Roweis, S., Hertzmann, A., Frey, B.: A Segment-Based Probabilistic Generative Model Of Speech. In: Proc. of ICASSP. (2005) 221–224
12. Moore, B.C.J.: An Introduction to the Psychology of Hearing. Emerald, Bingley, United Kingdom (2008)
13. Kahneman, D.: Thinking, Fast and Slow. Farrar, Straus and Giroux, New York (2011)
14. Plante, F., Meyer, G.F., Ainsworth, W.A.: A pitch extraction reference database. In: Proc. Eurospeech '95. (1995) 837–840
15. Kuhn, T.S.: Second Thoughts on Paradigms. In: The Essential Tension. Selected Studies in Scientific Tradition and Change, The University of Chicago Press, Chicago and London (1977) 293–319