pYIN: a Fundamental Frequency Estimator Using Probabilistic Threshold Distributions

Mauch, M; Dixon, S

PYIN: A FUNDAMENTAL FREQUENCY ESTIMATOR USING PROBABILISTIC THRESHOLD DISTRIBUTIONS

Matthias Mauch and Simon Dixon

Queen Mary University of London, Centre for Digital Music, Mile End Road, London

ABSTRACT

We propose the Probabilistic YIN (PYIN) algorithm, a modification of the well-known YIN algorithm for fundamental frequency (F0) estimation. Conventional YIN is a simple yet effective method for frame-wise monophonic F0 estimation and remains one of the most popular methods in this domain. In order to eliminate short-term errors, outputs of frequency estimators are usually post-processed resulting in a smoother pitch track. One shortcoming of YIN is that such post-processing cannot fall back on alternative interpretations of the signal because the method outputs precisely one estimate per frame. To address this problem we modify YIN to output multiple pitch candidates with associated probabilities (PYIN Stage 1). These probabilities arise naturally from a prior distribution on the YIN threshold parameter. We use these probabilities as observations in a hidden Markov model, which is Viterbi-decoded to produce an improved pitch track (PYIN Stage 2). We demonstrate that the combination of Stages 1 and 2 raises recall and precision substantially. The additional computational complexity of PYIN over YIN is low. We make the method freely available online[1] as an open source C++ library for Vamp hosts.

Index Terms— Pitch estimation, pitch tracking, YIN

1. INTRODUCTION

The estimation of the fundamental frequency (F0) from monophonic human voice signals is a prerequisite to comprehensive analysis of intonation in speech [1] and singing [2]. Since frame-wise pitch estimates are not completely reliable, post-processing is often used to clean the raw pitch track (see, e.g. [3]). Despite the high success rate of existing algorithms, this procedure is generally flawed because (potentially correct) frequency candidates that were discarded in the frame-wise stage cannot be recovered in the smoothing stage, and hence multiple frame-wise pitch candidates should be used before smoothing [4].

Several solutions to the problem of F0 estimation have been proposed in the area of speech processing [4, 5, 6, 7]. Among these, the YIN fundamental frequency estimator [7] has gained popularity beyond the speech processing community, especially in the analysis of singing [8, 9]. Babacan *et al.* [10] provide an overview of the performance of F0 trackers on singing, in which YIN is shown to be state of the art, and particularly good at fine pitch recognition.

The original YIN paper outlines a smoothing procedure that does not use the frame-wise estimate, but tracks low values in the underlying periodicity function. In this paper, we take YIN and modify its frame-wise variant in a probabilistic way to output multiple pitch candidates with associated probabilities, hence also reducing the loss of useful information before smoothing. We then employ a hidden Markov model which uses the modified frame-wise output to calculate a smoothed pitch track which retains YIN's well-known pitch accuracy and at the same time can be shown to provide excellent recall and precision.

2. METHOD

This section describes PYIN, our proposed method, which is divided into two stages: (1) frame-wise extraction of multiple pitch candidates with associated probabilities, and (2) HMM-based tracking of the pitch candidates into a monophonic pitch track. These stages will be addressed in turn.

2.1. Stage 1: F0 Candidates

The first stage of PYIN follows the same steps as the original YIN algorithm, differing only in the thresholding stage, where it assumes a threshold distribution, in contrast to YIN, which relies on a single threshold (see Fig. 1).

The YIN algorithm is based on the intuition that, in a signal $x_i,\ i = 1, \ldots, 2W$, the difference

$$d_t(\tau) = \sum_{j=1}^{W} (x_j - x_{j+\tau})^2, \tag{1}$$

will be small if the signal is approximately periodic with fundamental period $\tau = 1/f_0$. It can be shown [7] that this difference can be conveniently obtained by first calculating the

Matthias Mauch was funded by the Royal Academy of Engineering.

[1] `http://code.soundsoftware.ac.uk/projects/pyin`

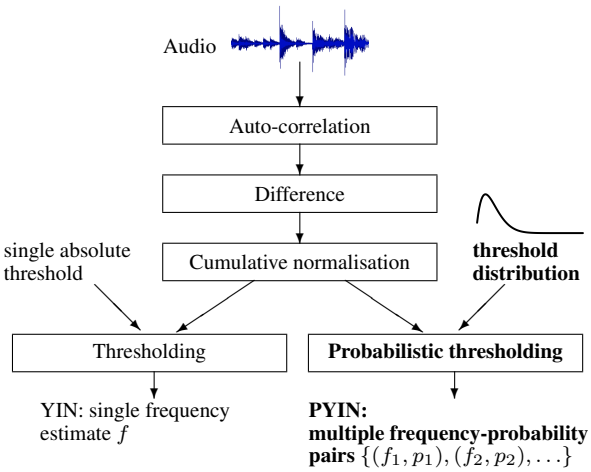


Fig. 1: Comparison of the first steps of the original YIN algorithm the proposed PYIN algorithm, with our contribution in bold print. Both have further steps for pitch refinement and pitch track smoothing, not pictured.

auto-correlation function (ACF)

$$r_t(\tau) = \sum_{j=t+1}^{t+W} x_j x_{j+\tau}, \tag{2}$$

from which (1) can be calculated as

$$d_t(\tau) = r_t(0) + r_{t+\tau}(0) - 2r_t(\tau). \tag{3}$$

These two calculations comprise the first two steps of the original YIN algorithm, as illustrated in Fig. 1. The third step of the original YIN algorithm is a normalisation of the difference (1) obtaining a 'cumulative mean normalised difference function' $d'(\tau)$ (we omit the subscript index t for simplicity) via a heuristic designed to compensate for low values at short periods (high frequencies) induced by formant resonances (for details see [7]).

The fourth step of the original YIN algorithm is to find the dip in the difference function d' that corresponds to the fundamental period. This is done by picking the smallest period τ for which d' has a local minimum and $d'(\tau) < s$ for a fixed threshold s (usually $s = 0.1$ or $s = 0.15$). In the case that $d'(\tau) > s$ for all τ (above threshold), the original YIN paper proposes $\arg\min_\tau d'(\tau)$ as the period estimate; alternatively this can be used to estimate the pitch as unvoiced [11]. We notate the period estimated by YIN as $Y(x_t, s)$.

Since both the choice of s and the strategy for handling the case that all values are above the threshold affect the result, we propose to abandon the use of a single absolute threshold and instead use a prior parameter distribution S given by $P(s_i)$, where $s_i,\ i = 1, \ldots, N$ are possible thresholds. We use thresholds ranging from 0.01 to unity in steps of 0.01, i.e. $N = 100$. The distributions used in our experiments are Beta distributions with means 0.1, 0.15, 0.2 ($\alpha = 1$ and $\beta = 18, 11\frac{1}{3}, 8$), as shown in Fig.2. Given such a distribution, and the prior probability p_a of using the absolute minimum strategy we can then define the probability that a period τ is the fundamental period τ_0 according to YIN as

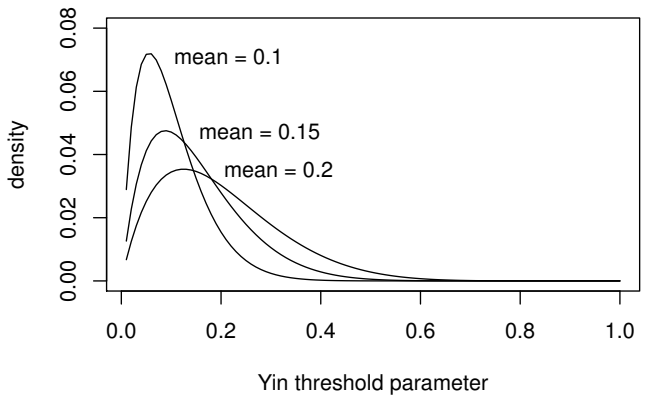


Fig. 2: The set of Beta distributions used as parameter priors for PYIN.

$$P(\tau = \tau_0 | S, x_t) = \sum_{i=1}^{N} a(s_i, \tau)\, P(s_i)\, [Y(x_t, s_i) = \tau], \tag{4}$$

where $[\,\cdot\,]$ is the Iverson bracket evaluating to unity for a true expression and to zero otherwise, and

$$a(s_i, \tau) = \begin{cases} 1, & \text{if } d'(\tau) < s_i \\ p_a, & \text{otherwise.} \end{cases} \tag{5}$$

We use $p_a = 0.01$. Note that if $p_a < 1$, then (4) does not necessarily sum to unity. The remaining probability mass can be interpreted as the probability that the frame is unvoiced.

Any τ for which $P(\tau = \tau_0 | S, x_t) > 0$ yields a fundamental frequency candidate $f = 1/\tau$. As in the original YIN algorithm frequency estimates are improved by parabolic interpolation on the difference function d'.

The set of fundamental frequency candidates along with their probabilities is the output of the first stage of the proposed PYIN algorithm. It has several appealing properties:

1. Any τ for which (4) is non-zero is a genuine YIN frequency estimate for some threshold s, at a minimum of the d' difference function.

2. The probabilistic estimate can be obtained in one original YIN loop with minimal computational overhead.

3. The conventional Yin estimate is among the candidates, i.e. PYIN covers at least as many true fundamental frequencies as the original YIN.

This last point is the main motivation for our method. In the original YIN algorithm, once an estimate is erroneous, the true value cannot be recovered. Our modification is based on

	orig.	noise	sound	live rec.	phone rec.	clip.
full cand.	**0.993**	**0.993**	**0.990**	**0.750**	**0.953**	**0.985**
YIN .10	0.988	0.980	0.886	**0.648**	**0.880**	**0.979**
YIN .15	**0.989**	**0.988**	0.934	0.644	0.843	0.958
YIN .20	0.980	0.986	**0.953**	0.632	0.803	0.931

Table 1: Recall by YIN parameter and degradation.

the observation that in many cases there exists an unknown threshold for which YIN will output the correct pitch, which is illustrated in Table 1: PYIN's set of candidates always has a higher pitch recall than any YIN parametrisation. We will show later that this substantially boosts the effectiveness of pitch tracking in Stage 2 of our proposed PYIN method.

2.2. Stage 2: HMM-based Pitch Tracking

The pitch tracking step consists of choosing at most one pitch candidate at every frame. We divide the pitch space into $M = 480$ bins ranging over four octaves from 55Hz (A1) to just under 880Hz (A5) in steps of 10 cents (0.1 semitones).

Such pitch bins can be modelled as states in a hidden Markov model (HMM). The model would then directly use the probabilities of pitch candidates obtained in the first PYIN step as observation probabilities: the probability of each observed pitch candidate is assigned to the bin closest to its estimated frequency; this results in a sparse observation vector $p^*_m, m = 1, \ldots, M$, where the only non-zero elements are those closest to pitch candidates. We use this idea but develop a more realistic HMM with one voiced ($v = 1$) and one unvoiced ($v = 0$) state per pitch (i.e. with $2M$ pitches), inspired by an existing note tracking method [9]. Assuming that the prior probability of being in either a voiced or an unvoiced state is $P(v = 1) = P(v = 0) = 0.5$, we define our model's observation probability as

$$p_{m,v} = \begin{cases} 0.5 \cdot p^*_m, & \text{for } v = 1 \\ 0.5 \cdot (1 - \sum_k P^*_k) & \text{for } v = 0. \end{cases} \quad (6)$$

Transition probabilities in this model have two main purposes: to favour natural (smooth) pitch tracks over discontinuous ones, and to favour few changes between unvoiced and voiced states. We encode this behaviour in two distributions pertaining to voicing transition and pitch transition:

$$p_v = P(v_t \mid v_{t-1}) = \begin{cases} 0.99, & \text{if no change} \\ 0.01 & \text{otherwise, and} \end{cases} \quad (7)$$

$$p_{ij} = P(\text{pitch}_t = j \mid \text{pitch}_{t-1} = i). \quad (8)$$

The latter is realised as a triangular weight distribution which encodes that a pitch jump can be at most 25 'bins', which corresponds to 2.5 semitones per frame. The highest likelihood peak is at 0 semitones. The window is always normalised to sum to 1. For more information refer to the source code. Assuming independence between voicedness and pitch, the actual transition probability between two states defined by pitch and voicedness is simply the product of the two individual probabilities. The initial probabilities are set to be uniformly distributed over the unvoiced states, and the HMM is decoded using an efficient version of the Viterbi algorithm that exploits the sparseness of the transition matrix.

3. RESULTS

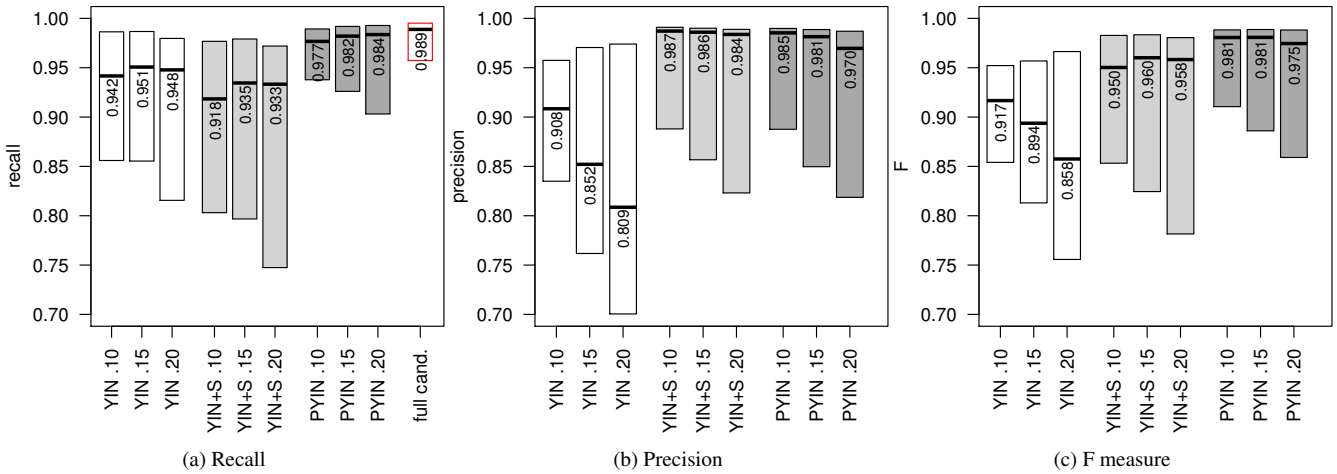
We synthesised singing from the F0 pitch tracks available with the RWC Music Database [12] and saved them as linear PCM wav files at a sample rate of 44.1kHz. The tracks cover the 100 full-length songs of the popular music subsection of the database. In order to simulate a more realistic situation without such clean data, we degraded the audio using five presets of the Audio Degradation Toolbox (ADT [13]) in addition to the original wav files. The complete data comprises in excess of 30 hours of audio. All results given are on this complete data set.

We ran three different versions of the proposed PYIN method with the Beta parameter distributions introduced in Section 2.1 (means 0.10, 0.15, 0.20, see Fig. 2). For comparison we also ran three versions with a parameter distribution S that has only one non-zero element s_i (here, too, these were 0.10, 0.15, 0.20). Since this effectively simplifies to original YIN plus smoothing, we refer to these as YIN+S. The baseline is the conventional YIN, also with three different versions of the original YIN method with threshold parameters $s = 0.10, 0.15, 0.20$. All methods are run in a Vamp plugin implementation with a step size of 256 (5.8ms) and a frame size of 2048 (46.4ms).

3.1. Quantitative Analysis on Synthetic Data

Recall. In the context of intonation analysis, it is desirable to have a correct frequency estimate for as much of the voiced data as possible, i.e. to obtain high recall. We calculate recall as the proportion of actually voiced frames (according to the ground truth) which the extractor recognises as voiced *and* tracks with the correct frequency. We follow [10] and accept a frame as correctly tracked if the estimate is within one semitone of the true frequency.

Figure 3a shows that the proposed method with any of the beta distributions tested (PYIN .10, .15, .20) clearly outperforms the original YIN estimate. The PYIN median recall values (0.977, 0.982, 0.984) approach much more closely the upper bound given by the number of correct pitches covered by the full candidate set (0.98) than any of the conventional YIN methods (medians all below 0.951). Note that this is true despite the fact that the YIN methods have an advantage due



(a) Recall (b) Precision (c) F measure

Fig. 3: Box plots of performance measures providing median and 1st and 3rd quartiles.

to not making a voicing decision, i.e. they always provide an estimate.

Note that the YIN+S estimates have a worse recall than the original YIN method. This is because there is only a single YIN estimate per frame to smooth over, and hence recall is bounded by the performance of YIN.

Precision and F Score. In addition to recall, any analysis of pitch and intonation requires good precision, that is: a high proportion of correct pitch estimates in frames marked by the extractor as voiced. As is shown in Figure 3b, all PYIN and YIN+S estimates perform better than the original YIN, even though—again—our evaluation favours YIN by using only those frames for evaluation that YIN recognises as voiced. Note that the YIN+S methods have the best precision, by virtue of recognising fewer frames as voiced. The optimum tradeoff between precision and recall as measured by the F score $F = 2\frac{\text{precision}\times\text{recall}}{\text{precision}+\text{recall}}$ is obtained by using the proposed PYIN algorithm with a Beta parameter distribution, as illustrated in Figure 3c.

Octave Errors and Voicing Detection. The PYIN methods also provide excellent robustness against octave errors (0.5%, 0.9% and 1.7%, respectively) and very good voicing detection recall (92.5%, 94.1% and 95.0%) and specificity (91.9%, 90.6% and 88.9%).

3.2. Real Human Singing: Qualitative Example

Figure 4 shows pitch tracks of the last four notes of the song 'Happy Birthday' extracted by YIN .15 and PYIN .15. The recording was sung by a female singer for a study on singing intonation [14]. While the first two notes (up to 59 seconds) are tracked almost perfectly by both pitch trackers, many octave errors occur in the YIN pitch track. This may have been caused by the unusual breathy timbre of the singer. PYIN is robust to these errors and extracts the correct pitch track.

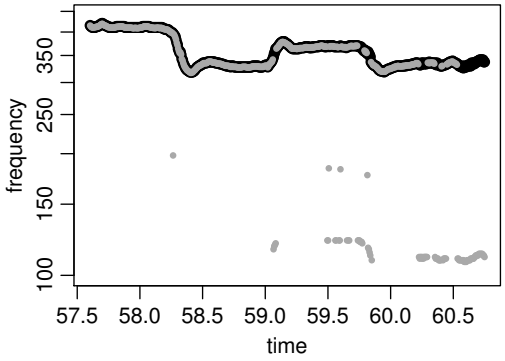


Fig. 4: Pitch tracks of PYIN .15 (black) and YIN .15 (grey) on an example of real human singing (last four notes of 'Happy Birthday' sung by a female singer).

4. CONCLUSION

We presented the PYIN algorithm, a modification of YIN which jointly considers multiple pitch candidates based on a probabilistic interpretation of YIN. A prior distribution on the YIN threshold parameter yields a set of pitch candidates with associated probabilities, computed using YIN. The procedure effectively turns the popular frame-wise YIN algorithm into a probabilistic machine which outputs pitch candidates with associated probabilities. Temporal smoothness is obtained by using the candidate probabilities as observations in an HMM, which is Viterbi-decoded to produce the final pitch track. We demonstrated that PYIN has superior precision and recall to YIN on a database of over 30 hours of synthesised singing. Since PYIN is parametrised by a distribution of YIN thresholds, the algorithm is more robust to choice of distribution than YIN is to its choice of threshold.

5. REFERENCES

[1] Fang Liu and Yi Xu, "Question Intonation as Affected by Word Stress and Focus in English," in *Proceedings of the 16th International Congress of Phonetic Sciences*, 2007, pp. 1189–1192.

[2] Johanna Devaney and Daniel P W Ellis, "An Empirical Approach to Studying Intonation Tendencies in Polyphonic Vocal Performances," *Journal of Interdisciplinary Music Studies*, vol. 2, no. 1, pp. 141–156, 2008.

[3] Baris Bozkurt, "An Automatic Pitch Analysis Method for Turkish Maqam Music," *Journal of New Music Research*, vol. 37, no. 1, pp. 1–13, Mar. 2008.

[4] David Talkin, "A Robust Algorithm for Pitch Tracking," in *Speech Coding and Synthesis*, pp. 495–518. 1995.

[5] Paul Boersma, "Praat, a system for doing phonetics by computer," *Glot International*, vol. 5, no. 9/10, pp. 341–345, 2001.

[6] Hideki Kawahara, Jo Estill, and Osamu Fujimura, "Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT," *Proceedings of MAVEBA*, pp. 59–64, 2001.

[7] Alain de Cheveigné and Hideki Kawahara, "YIN, a fundamental frequency estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, 2002.

[8] Johanna Devaney and Dan Ellis, "Improving MIDI-Audio Alignment with Audio Features," in *2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2009, pp. 18–21.

[9] Matti P Ryynänen, *Probabilistic Modelling of Note Events in the Transcription of Monophonic Melodies*, Ph.D. thesis, Tampere University of Technology, 2004.

[10] Onur Babacan, Thomas Drugman, Nicolas D'Alessandro, Nathalie Henrich, and Thierry Dutoit, "A Comparative Study of Pitch Extraction Algorithms on a Large Variety of Singing Sounds," in *Proceedings of the 38th International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2013)*, 2013, pp. 7815–7819.

[11] Joren Six and Olmo Cornelis, "Tarsos - A Platform to Explore Pitch Scales in Non-Western Music," in *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)*, 2011, pp. 169–174.

[12] Masataka Goto, Hiroki Hashiguchi, Takuichi Nishimura, and Ryuichi Oka, "RWC Music Database: Popular, Classical, and Jazz Music Databases," in *Proceedings of the 3rd International Conference on Music Information Retrieval (ISMIR 2002)*, 2002, pp. 287–288.

[13] Matthias Mauch and Sebastian Ewert, "The Audio Degradation Toolbox and its Application to Robustness Evaluation," in *Proceedings of the 14th International Society for Music Information Retrieval Conference (ISMIR 2013)*, 2013, pp. 83–88.

[14] Matthias Mauch, Klaus Frieler, and Simon Dixon, "Intonation in Unaccompanied Singing: Accuracy, Drift and a Model of Intonation Memory," *under review*, 2013.