# A Simple Continuous Pitch Estimation Algorithm

Philip N. Garner, *Senior Member, IEEE*, Milos Cernak, *Member, IEEE*, Petr Motlicek, *Member, IEEE*

***Abstract*—Recent work in text to speech synthesis has pointed to the benefit of using a continuous pitch estimate; that is, one that records pitch even when voicing is not present. Such an approach typically requires interpolation. The purpose of this paper is to show that a continuous pitch estimation is available from a combination of otherwise well known techniques. Further, in the case of an autocorrelation based estimate, the continuous requirement negates the need for other heuristics to correct for common errors. An algorithm is suggested, illustrated, and demonstrated using a parametric vocoder.**

***Index Terms*—pitch estimation; Kalman smoother; speech coding; speech parameterisation**

## I. Introduction

PITCH estimation (pitch extraction or pitch tracking) refers to the process of discerning the fundamental frequency of the harmonic part of a signal. For instance, it is the entity described by the (height of the) notes in a musical score. Whilst pitch estimation has applications in radar and communications, this paper is concerned with audio in general and speech coding and synthesis in particular.

In the context of speech, pitch is associated with voicing; it is often referred to in the literature as $f_0$. $f_0$ normally carries information at a supra-segmental level (an exception being tonal languages). This means that in, e.g., automatic speech recognition, $f_0$ is of little use as the acoustic models are built at a segmental level. However, $f_0$ is important in text to speech synthesis (TTS) simply to make synthetic speech sound natural at that supra-segmental level. In statistical TTS, $f_0$ is both modelled by the hidden Markov model (HMM), and is used in the STRAIGHT vocoder of Kawahara et al. [1], which requires a pitch estimate in order to extract a spectral envelope.

## II. Background

### A. Pitch estimation

The estimate for STRAIGHT is usually provided by the TEMPO method [2]. Other notable work on pitch includes the YIN method of de Cheveigné and Kawahara [3]. This is based to an extent on the autocorrelation method described by Boersma [4]. More recently, model-based approaches such as those of Christensen and Jakobsson [5] and Nielsen et al. [6] promise higher accuracy. It is pertinent to note that $f_0$ extraction is by no means a solved problem. Yamagishi et al. [7] describe a three stage process:

> "$f_0$ is first extracted using a wide range over a whole database, then a range is determined for each speaker and $f_0$ is extracted again using three methods. Finally a median value of the three methods is chosen."

i.e., three well known $f_0$ extraction methods are known to produce different results; there is no oracle method.

A key issue in pitch estimation, at least for speech, is the handling of segments that are unvoiced; that is, where pitch cannot be observed. In HMM-based TTS, the multi-space distribution (MSD) of Tokuda et al. [8] is used; this involves building distinct models for voiced and unvoiced segments. Tokuda et al. cite the work of three other groups: Freij and Fallside [9], Jensen et al. [10] and Ross and Ostendorf [11]. In the first two of these, random values and zero respectively were assigned to $f_0$ when it could not be measured. This suited their stress and intonation recognition tasks, but is unsuitable for TTS because it would lead to synthesis of random or meaningless $f_0$. Ross and Ostendorf [11] use an appealing linear dynamical system model, but state that "values for $f_0$ in unvoiced regions are ignored" suggesting that the model in fact requires some MSD like structure to be used in practice.

A rather high level summary would be that the lack of voicing leads to difficulty or complexity.

### B. Continuous pitch

Recent work suggests that continuous pitch has advantages. Yu and Young [12] demonstrate that an HMM based TTS using a continuous $f_0$ produces more expressive $f_0$ contours than one based on the MSD. This follows at least in part from the ability to define dynamic features properly. Zhang et al. [13] introduce using the voicing strength in an otherwise continuous system to indicate an voiced/unvoiced decision. Latorre et al. [14] show that a voiced/unvoiced decision can in fact be left up to the aperiodicity features in a mixed excitation codec. In perceptual experiments, they show that this leads to fewer intrusive errors such as false unvoicing (hoarseness) and false voicing (buzziness).

Although $f_0$ is a characteristic of the excitation rather than the resonance, the estimation problem is analogous to that of the other formants in that they are also not necessarily present. It was shown by Garner and Holmes [15] that uncertainty about the presence of formants can be represented as a variance on their distributions. This can in turn be incorporated into HMM-based models. It is reasonable to suppose that the same method could apply to $f_0$ estimates.

The remaining sections detail a Bayesian approach to pitch estimation that naturally yields estimates for unvoiced segments, along with variances for all estimates. An algorithm is described, and it is shown that the continuous pitch requirement has (positive) implications for the pitch extraction process. The resulting algorithm leads to an intuitive illustration and a persuasive demonstration using a parametric vocoder.


The authors are with Idiap Research Institute, Martigny, Switzerland.

## III. Bayesian approach

### A. The intuition barrier

The fundamentally unintuitive concept of assigning a value to an $f_0$ that does not exist can be resolved by a Bayesian approach. This approach requires a hypothesis that there is some underlying state variable of which $f_0$ is indicative. It could be something physical such as tautness of vocal cords, or something intangible such as the speaker's intent. This has an appealing analog in phonetics where a phoneme is the underlying intent and a phone is the acoustic realisation. It seems reasonable to use *pitch*, or $\rho$, to refer to the underlying state, and $f_0$ as the acoustic realisation. Mathematically,

$$p\left(\rho \mid f_0\right) \propto p\left(f_0 \mid \rho\right) p\left(\rho\right) \tag{1}$$

The Bayesian approach yields the pitch as being modelled using a probability density function, $p\left(\rho \mid f_0\right)$; an estimate of pitch is then available as the maximum or expectation of this density. Intuitively, where there is a clearly observable $f_0$, the density function of pitch should be narrow (with a small variance). Conversely, where $f_0$ cannot be measured the pitch density should have a wider variance. Where $f_0$ is not observable, information about the pitch is available from prior information, $p\left(\rho\right)$.

### B. Choice of prior

Depending upon the type of signal, different priors might be appropriate. For instance, a singing voice has certain constraints defined by music theory. In the case of speech, it is reasonable to assume that the pitch is a continuous contour. If $\rho_t$ is the pitch at time $t$, a first order relationship would define $p\left(\rho_t\right) \propto p\left(\rho_t \mid \rho_{t-1}\right)$. Modelling both this and the likelihood terms as normal distributions,

$$p\left(\rho_t \mid \rho_{t-1}\right) \sim N(\rho_{t-1}, \phi^2), \tag{2}$$

$$p\left(f_{0,t} \mid \rho_t\right) \sim N(\rho_t, \sigma^2), \tag{3}$$

where "$\sim$" is taken to mean "is distributed as" and $N(\mu, \sigma^2)$ is the normal distribution with mean $\mu$ and variance $\sigma^2$. Equations 2 and 3 constitute a linear dynamical system, the solution to which is the Kalman smoother. This is the same model used, albeit for TTS, by Ross and Ostendorf [11].

### C. Parameters

The dynamical system model introduces two standard deviation parameters. Of these, $\phi$ is a system-wide parameter; it must be set either heuristically or trained. The other, $\sigma$, is a function of the $f_0$ extraction, and is discussed below.

## IV. Probabilistic pitch estimation

### A. Observation variance

Any $f_0$ estimation technique will yield some estimate of $f_0$ whether voicing is present or not; the requirement here is to also produce some measure of how accurate the estimate is. As pointed out by Boersma [4], the autocorrelation based method not only yields an estimate of $f_0$, but also a harmonics-to-noise ratio (HNR). Boersma defines the HNR as

$$\mathrm{HNR} = 10 \log_{10} \frac{r'(\tau_{\max})}{1 - r'(\tau_{\max})}, \tag{4}$$

where $\tau_{\max} > 0$ and is the lag associated with the peak in the autocorrelation, $r(\tau)$ is the autocorrelation, and $r'(\tau) = r(\tau)/r(0)$. For purely harmonic signals the HNR is infinite; for noise it is minus infinity.

Notice that the reciprocal of the term inside the logarithm of equation 4 is zero for harmonic signals and infinite for noise. This is the same as the requirement for $\sigma$; it follows that a *heuristic* but *intuitively reasonable* definition would be:

$$\sigma \propto \frac{1 - r'(\tau_{\max})}{r'(\tau_{\max})}. \tag{5}$$

This leads to the distribution $p\left(f_{0,t} \mid \rho_t\right) \sim N(\rho_t, \sigma_t^2)$, where the variance, now dependent upon $t$, is small for harmonic signals and larger for noisier ones.

### B. Corollary

Although the dynamical system model arose simply to make use of prior information when $f_0$ cannot be observed, it has two other advantages in the context of autocorrelation based pitch estimation:

1) Because of the coarse granularity of the autocorrelation, better accuracy is sometimes sought via interpolation within frames. The dynamical system performs the same task implicitly over time.
2) If a small value is used for $\phi$, effectively over-smoothing the pitch contour, the resulting contour is robust to the wrong choice of peak in the autocorrelation.

The implication is that the pitch tracker no longer requires these components, and can thus be significantly simplified.

### C. Algorithm

The above intuition leads to the following algorithm for pitch estimation:

1) Frame the signal into possibly overlapping frames.
2) Window each frame.
3) For each frame calculate the autocorrelation and divide by that of the window as described by Boersma [4].
4) For each frame identify a peak, $\tau_{\max,t}$, in the normalised autocorrelation between limits defined by frequencies $f_{\mathrm{lo}}$ and $f_{\mathrm{hi}}$.
5) For each frame calculate the (heuristic) variance

$$\sigma_t^2 = \left( \frac{1 - r'(\tau_{\max,t})}{r'(\tau_{\max,t})} \times (f_{\mathrm{hi}} - f_{\mathrm{lo}}) \right)^2 \tag{6}$$

6) Using a value of $\phi^2 = 1000$ (i.e., pitch expected to remain within tens of Hz) and prior mean and variance $\mu_0 = \frac{f_{\mathrm{hi}} + f_{\mathrm{lo}}}{2}$ and $\sigma_0^2 = (f_{\mathrm{hi}} - f_{\mathrm{lo}})^2$, apply the Kalman smoother to the sequence of estimates and variances to give a sequence of pitch estimates.

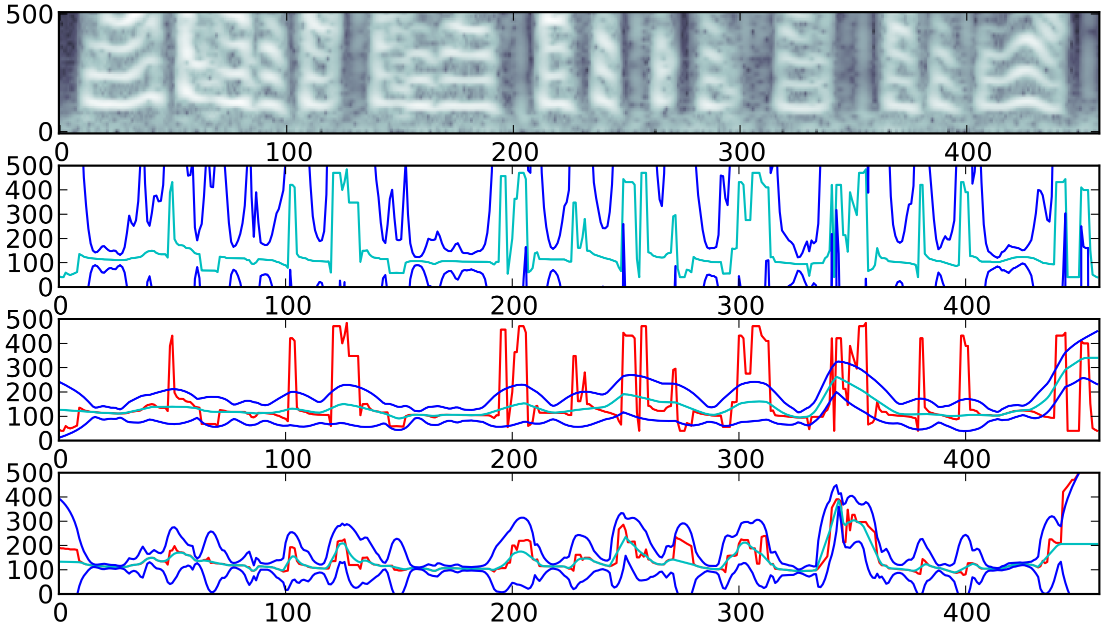


Fig. 1. Illustrative example of pitch estimation. The illustrations are, from top to bottom: The periodogram up to 500 Hz of the 16 kHz signal; the pitch estimate and HNR-implied standard deviation from the normalised autocorrelation; the first (over-)smoothed pitch estimate; and the second estimate with more relaxed smoothing. In each case, the dark blue lines are $\pm 1$ standard deviation. The reference pitch is clearly visible as the first harmonic in the periodogram.

7) Calculate new estimates as in step 4, except with time dependent frequency bounds:

$$f_{\text{lo},t} = 1.5\rho_t, \tag{7}$$
$$f_{\text{hi},t} = 0.75\rho_t, \tag{8}$$

i.e., within the pitch halving and doubling range.
8) Recalculate the variance as above, except using the time dependent frequency bounds.
9) Using a value of $\phi^2 = 10000$ (i.e., pitch allowed to vary hundreds of Hz), reapply the Kalman smoother to the sequence of estimates and variances to give a final sequence of pitch estimates.

In steps 6 and 9, although the value of $\phi^2$ has the indicated meaning, it also functions more generally as a weighting factor between the likelihood and prior distributions. The Kalman smoother is detailed in the appendix.

### D. Illustrative example

Fig. 1 shows the effect of the above algorithm on a real recording (utterance `EM1_ENG_0001_0` from the EMIME bilingual database [16]). Notice that the effect of the Kalman smoother is rather intuitive: In segments of clearly defined $f_0$, the estimated distribution has small variance; in less clear segments the variance is larger. The large variance is especially evident during the opening and closing silence.

Some pitch halving and doubling errors can be seen to be corrected. Reciprocally, there is a false high HNR around frame 340 during a region of otherwise low HNR and large variance.

## V. Vocoder

Whilst quantitative validation and incorporation into a TTS system are matters for future research, the technique described has been validated qualitatively by incorporation into a simple parametric vocoder (the signal is parameterised rather than coded). Such a vocoder is a prerequisite for HMM-based TTS.

|  | Male | Female |
|---|---|---|
| Original | `EM1_ENG_0001_0.16` | `EF2_ENG_0001_0.16` |
| Vocoded | `EM1_ENG_0001_0.vo` | `EF2_ENG_0001_0.vo` |

TABLE I
Recordings demonstrating vocoder performance (files have `.WAV` extension).

In the encoding part of the vocoder, 16 kHz speech is split into overlapping frames of 256 samples every 128 samples. Each frame is represented using 24$^{\text{th}}$ order auto-regression (AR) coefficients. The pitch estimator described above is also used to represent frames of 1024 samples, but at the same period as the AR. For each frame, a pitch estimate and HNR are recorded. In the decoder, frames of 256 samples are constructed using an impulse stream of the given frequency, and white noise. The impulses and noise are added in the ratio suggested by the HNR, and used to excite the AR filter. Frames are concatenated using overlap-add.

The vocoder relies on two effects to mask the harmonic component when none is required:

1) In segments of silence, the gain of the AR filter is small.
2) During unvoiced speech, the HNR is low.

The performance is evidence from the recordings included with this submission, again from the EMIME bilingual database, summarised in table I. Although the vocoder suffers from the “buzziness” associated with the simplistic excitation, and certainly contains artefacts, the speech is clear.

## VI. Conclusions

In this letter, it has been shown that the linear dynamical system and associated Kalman smoother allow a pitch extraction algorithm to generate continuous pitch estimates given discontinuous $f_0$ observations. In using the Kalman smoother, some heuristic aspects of the pitch extracter are rendered moot, enabling simplification.

The resulting pitch estimation has been validated both intuitively by illustration, and qualitatively using a vocoder. In doing so, the concept of a parametric vocoder with no voiced/unvoiced decision has been demonstrated.

No claims have been made about the quantitative accuracy of the algorithms; this is a matter for future research.

The algorithm is undoubtedly better suited to the Bayesian pitch estimation of Nielsen et al. [6]. That algorithm produces a distribution over pitch, yielding $p\left(\rho \mid f_0\right)$ directly, along with a distribution over HNR (it is their value $g$, which is in turn a simple function of signal to noise ratio). Although likely to be quantitatively more accurate, Nielsen's estimation is considerably more computationally intensive, justifying in part the more heuristic approach presented here.

The approach as described, or using another pitch estimation method, is suitable for the HMM modelling of Yu and Young [12]. It is also appropriate for incorporation into HMM training in the same way as formants via the algorithm presented by Garner and Holmes [15].

## Acknowledgement

This research was partly supported under the RECOD project by armasuisse, the Procurement and Technology Center of the Swiss Federal Department of Defence, Civil Protection and Sport.

## References

[1] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 3-4, pp. 187–207, April 1999.

[2] H. Kawahara, H. Katayose, A. de Cheveigné, and R. D. Patterson, "Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of F0 and periodicity," in *Proceedings of EUROSPEECH*, Budapest, Hungary, September 1999.

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

[4] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," in *Proceedings of the Institute of Phonetic Sciences, Amsterdam*. University of Amsterdam, 1993, no. 17, pp. 97–110.

[5] M. G. Christensen and A. Jakobssen, *Multi-Pitch Estimation*, ser. Synthesis Lectures on Speech and Audio Processing. Morgan & Claypool, 2009.

[6] J. K. Nielsen, M. G. Christensen, and S. H. Jensen, "An approximate Bayesian fundamental frequency estimator," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Kyoto, Japan, March 2012, pp. 4617–4620.

[7] J. Yamagishi, T. Nose, H. Zen, Z.-H. Ling, T. Toda, K. Tokuda, S. King, and S. Renals, "A robust speaker-adaptive HMM-based text-to-speech synthesis," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 17, no. 6, August 2009.

[8] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "Multi-space probability distribution HMM," *IEICE Transactions on Information and Systems*, vol. E85-D, no. 3, pp. 455–464, March 2002, (Invited paper).

[9] G. J. Freij and F. Fallside, "Lexical stress recognition using hidden Markov models," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, New York, New York, USA, April 1988, pp. 135–138.

[10] U. Jensen, R. K. Moore, P. Dalsgaard, and B. Lindberg, "Modelling intonation contours at the phrase level using continuous density hidden Markov models," *Computer Speech and Language*, vol. 8, no. 3, pp. 247–260, 1994.

[11] K. N. Ross and M. Ostendorf, "A dynamical system model for generating fundamental frequency for speech synthesis," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 3, pp. 295–309, May 1999.

[12] K. Yu and S. Young, "Continuous F0 modelling for HMM based statistical parametric speech synthesis," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 5, pp. 1071–1079, 2011.

[13] Q. Zhang, F. Soong, Y. Qian, Z. Yan, J. Pan, and Y. Yan, "Improved modeling for F0 generation and V/U decision in HMM-based TTS," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Dallas, Texas, USA, March 2010, pp. 4606–4609.

[14] J. Latorre, M. J. F. Gales, S. Buchholz, K. Knill, M. Tamura, Y. Ohtani, and M. Akamine, "Continuous F0 in the source-excitation generation for HMM-based TTS: Do we need voiced/unvoiced classification?" in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Prague, Czech Republic, May 2011, pp. 4724–4727.

[15] P. N. Garner and W. J. Holmes, "On the robust incorporation of formant features into hidden Markov models for automatic speech recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1, 1998, pp. 1–4.

[16] M. Wester, "The EMIME bilingual database," The University of Edinburgh, Report EDI-INF-RR-1388, September 2010. [Online]. Available: http://www.inf.ed.ac.uk/publications/report/1388.html

[17] L. L. Scharf, *Statistical Signal Processing. Detection, Estimation and Time Series Analysis*. Addison Wesley, 1991.

## Appendix

Although the Kalman smoother is well known (see, e.g., the book by Scharf [17]), the recursions are stated below for reference.

The forward *filter* maintains a mean, $M_t^+$, and variance, $V_t^+$; it is initialised as

$$M_1^+ = \frac{f_{0,1}\sigma_0^2 + \mu_0\sigma_1^2}{\sigma_0^2 + \sigma_1^2}; \quad V_1^+ = \frac{\sigma_0^2\sigma_1^2}{\sigma_0^2 + \sigma_1^2}. \tag{9}$$

With no offset, the first *predictor* has the same mean with variance

$$P_2 = \phi^2 + V_1^+. \tag{10}$$

That predictor then replaces the prior for a the second frame:

$$M_2^+ = \frac{f_{0,2}P_2 + M_1^+\sigma_2^2}{P_2 + \sigma_2^2}, \quad V_2^+ = \frac{P_2\sigma_2^2}{P_2 + \sigma_2^2}, \tag{11}$$

and an iteration is evident.

The backward *smoother* then updates these to a mean, $M_t^-$, and variance, $V_t^-$. The first term, this time at time $T$, is

$$M_T^- = M_T^+; \quad V_T^- = V_T^+. \tag{12}$$

The next backward term is then

$$M_{T-1}^- = \frac{V_{T-1}^+ M_T^-}{\phi^2 + V_{T-1}^+} + \frac{M_{T-1}^+\phi^2}{\phi^2 + V_{T-1}^+}, \tag{13}$$

$$V_{T-1}^- = \frac{V_{T-1}^+}{\phi^2 + V_{T-1}^+}\left(\phi^2 + \frac{V_{T-1}^+ V_T^-}{\phi^2 + V_{T-1}^+}\right), \tag{14}$$

and again a recursion is evident. At any time $t$, the posterior pitch distribution is

$$p\left(\rho_t \mid f_{0,t}\right) \sim N(M_t^-, V_t^-). \tag{15}$$