Comparative evaluation of F_0 estimation algorithms

Alain de Cheveigné

Ircam - CNRS
Paris, France
cheveign@ircam.fr

Hideki Kawahara

CREST, Wakayama University
Wakayama, Japan
kawahara@sys.wakayama-u.ac.jp

Abstract

This paper reports the comparative evaluation of several speech F_0 evaluation algorithms over a wide database of laryngograph-labeled speech. Included are several classic algorithms that are available in software on the net, as well as two new algorithms that offer greatly reduced error rates. Particular attention is given to the methodology of evaluation.

1. Introduction

The fundamental frequency (F_0) of a periodic signal is the inverse of its period. The period is defined as the smallest positive member of the set of time shifts that leave the signal invariant. This definition applies strictly only to a *perfectly* periodic signal, which is uninteresting because it cannot be switched on or off or modulated in any way without losing its perfect periodicity. Interesting signals such as speech or music are aperiodic in several ways, and the art of fundamental frequency estimation is to deal with them in a consistent and useful way.

Some applications give F_0 a definition closer to their purposes. For voiced speech, F_0 is usually defined as the rate of vibration of the vocal folds. Periodic vibration at the glottis may produce speech that is less perfectly periodic, due to changes in shape of the vocal tract that filters the glottal source waveform, making it hard to estimate fundamental periodicity from the speech waveform. Glottal vibration itself may also show aperiodicities, in the form of relatively smooth changes in amplitude, rate or glottal waveform shape (for example the duty cycle of open and closed phases), or intervals where the vibration seems to reflect several superimposed periodicities (diplophony), or where glottal pulses occur without obvious regularity of time interval or amplitude (glottalizations, vocal creak or fry) [11]. Arguably, in those cases there is no F_0 to estimate, and it is natural for an estimation algorithm to fail. However practical applications may require more graceful behavior than just random failure. All these factors conspire to make the task of obtaining a useful estimate of F_0 rather difficult.

If it can be reliably estimated, F_0 is useful for a wide range of applications. In speech, F_0 variations contribute to prosody, and in tonal languages they also help distinguish segmental categories. Attempts to use F_0 in speech recognition systems have met with mixed success, but this may in part be a consequence of the limited reliability of estimation algorithms. Several musical applications need F_0 estimation, such as automatic score transcription or real-time interactive systems, but again imperfect reliability is an obstacle. F_0 is a useful ingredient for a variety of signal processing methods, for example spectral envelope estimation [16]. Finally, a fairly recent application of F_0 is as metadata for multimedia content indexing, for example in the newly developped MPEG-7 standard [15, 7].

F_0 estimation has attracted, and continues to attract, much effort and ingenuity. The most comprehensive review remains that of [13] that cites literally hundred of methods. More recent reviews are [14] or [12]. A few examples of recent approaches are instantaneous frequency methods [1, 16, 2] , statistical learning and neural networks [4, 20, 9], auditory models [10, 6]. However a major weakness in this field is that it is often hard to judge the performance of the algorithms proposed, or make comparisons with other methods. In this respect, evaluation results are as useful as new methods.

2. Methods

2.1. Basic methodology

Evaluation is demanding. First, there must be some means to judge whether estimates are correct. Second, there must be enough data to be sure that results will generalize. Third, for others to benefit from the results, they must be able to compare them to methods that they are using or developping. Databases vary in difficulty, and therefore it is not enough to simply report figures. One option (followed for example by [3]) is to make the database freely available, another is to evaluate, together with the method being reported, other easily available methods.

Early F_0 extractors were evaluated informally, for example by checking visually for obvious breaks in the "pitch track". Coding applications evaluated subjective quality of resynthesized speech, but this method is expensive and the scores not very informative. Manual labeling of speech waveforms has been used for intonation studies, but it is time-consuming and error-prone. Recent studies use speech recorded together with the signal of a laryngograph, which measures the electrical resistance between electrodes on either side of the throat, function of the surface of contact between vocal folds. It is usually easier to estimate reliable F_0 values from this signal than from the acoustic waveform. Caveats are that a clean signal may be hard to obtain or maintain for certain morphologies or behaviors (large amplitude body movements), and may it be absent or weak in phonation modes for which there is little variation of vocal-fold contact surface.

The laryngograph signal is processed by an F_0-estimation algorithm tuned to the particularities of this signal: a lack of formant structure, sharp peaks in the derivative at glottal closure, and large-amplitude variations in "DC" offset. In addition to F_0, is also necessary to produce a mask, either automatically or manually, to indicate which portions correspond to regular glottis vibration. This corresponds roughly (not exactly) to a voiced-unvoiced decision. The F_0 estimate is then checked manually to eliminate errors. Two options are available: manual correction of estimate values, or manual adjustment of the mask. The latter option was chosen in this study. Two criteria

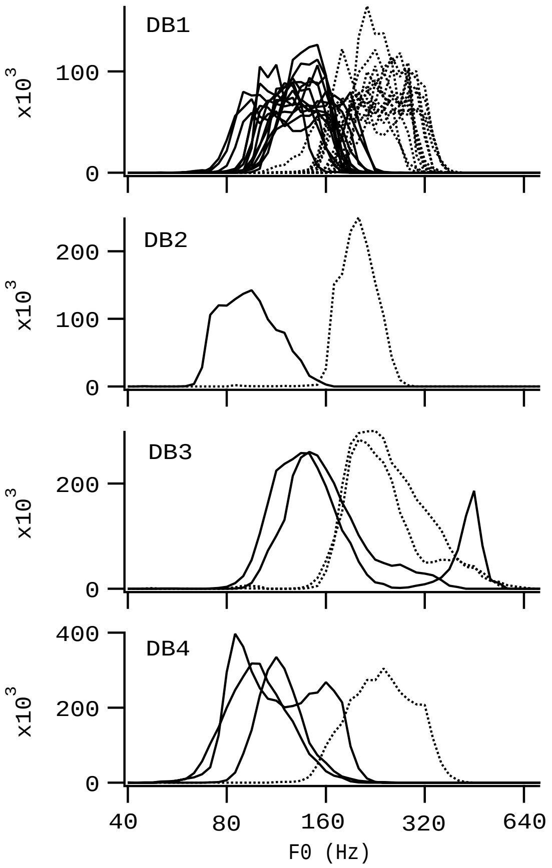

Figure 1: *Histograms of F_0 for each speaker and each database. Full lines are male speakers, dotted lines are female. Bin width is one semitone (one 12th of an octave). DB3 includes falsetto phonation, visible as an extension of the distributions to the right.*

were used, in this order: (1) any part for which the F_0 estimate is obviously incorrect is removed, (2) otherwise, any part for which there is any sign of glottal vibration is included. The first criterion ensures that all reference F_0 estimates are correct, the second includes as many "difficult" parts as possible. All these steps must be done with the utmost care.

It should be noted that this procedure eliminates portions of irregular glottal vibration (diplophony, creak, etc.). It is not indifferent how an algorithm treats them, but arguably there is little sense in setting a normative value where F_0 is inherently ambiguous.

Error rates are derived by counting the samples that differ from the reference by more than a criterion percentage (often 20%). This produces the "gross error rate". Some studies also report error rates for a "voiced-unvoiced" decision based on regularity, but as voicing cannot be reduced to mere regularity its detection is best treated as a separate issue.

Acoustic propagation from glottis to microphone, or implementation differences, may introduce a time shift between laryngograph- and microphone-based estimates. To avoid an unmerited penalty, the minimum error rate is taken over a range of time shifts. One must also be aware of a more insidious problem. Some estimation algorithms work (in effect) by comparing two windows of data that are shifted symmetrically in time with respect to the analysis point. Others work (in effect) by comparing a shifted window to a fixed window. An F_0-dependent corrective shift should be used in this case.

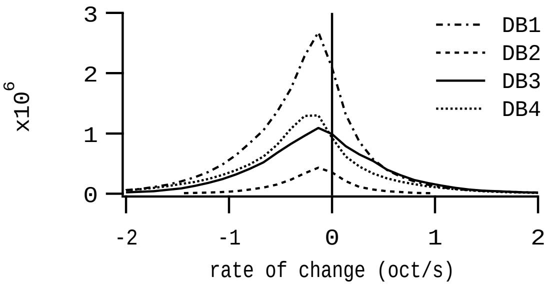

Figure 2: *Histograms of F_0 rate of change for each database, calculated over 25 ms intervals. Bin width is 0.13 oct/s.*

Methods being compared should be given similar parameters so that they are on a level ground. The search range must accomodate the values in the database, but wider ranges offer more options for error. Methods should be given the same range. Some methods produce estimates only when they judge speech to be "voiced", and are at a disadvantage with respect to methods that always offer an estimate. The voiced-unvoiced decision should be disabled. Other methods apply post-processing schemes to (hopefully) correct errors. This typically involves parameters and behavior that are hard to interpret and optimize. Post-processing is best evaluated separately, and should also be disabled when evaluating the basic algorithm. These recommendations can not always be followed, either because methods use radically different parameters, or because their implementation does not allow them to be controlled. Such differences must be kept in mind when comparing results.

2.2. Databases

Four databases were used in this study. All were generously provided by their authors who are warmly thanked. Together they represent a total of 1.75 hours of speech, of which 47% were labeled as regularly voiced. They include speech from a total of 38 speakers (19 male, 19 female) of Japanese (30), English (4), and French (4). Each included a laryngograph waveform recorded together with the speech. Details of availability of these and other resources can be obtained from the URL <http://www.ircam.fr/cheveign/data/eusp2001/>.

1. DB1. Thirty Japanese sentences were each spoken by 14 male and 14 female speakers for a total of 0.66 hours of speech [2].

2. DB2. Fifty English sentences were each spoken by one male and one female speaker for a total of 0.12 hours of speech [3]. This databse has been used in several studies and is available for download. It includes a laryngograph-based F_0 estimate that was not used here (our error rates are thus not directly comparable to rates reported elsewhere for the same database).

3. DB3. Four speakers of French, 2 male and 2 female, each pronunced 45 to 55 sentences each for a total of 0.46 hours of speech. The database includes sentences pronunced according to several modes: normal (141), head (30) and fry (32) [22]. The fry mode was not used

for evaluation because it was often too irregular to derive an unambiguous reference F_0.

4. DB4. English speech was pronunced by two male speakers, and Japanese speech by one male and one female speaker. The database was created for the purpose of deriving prosody rules for speech synthesis [5]. An extract of 0.51 hours was used here.

The laryngograph data were processed with a version of the algorithm of [8], with compensation for variations of amplitude and DC offset. F_0 estimates were produced at the same sampling rate as the speech and laryngograph signals (16 kHz for databases 1,3 and 4, and 20 kHz for database 2).

2.3. Algorithms

We evaluated both algorithms implemented elsewhere and available as software on the net, and locally implemented methods. The former have the advantage that they were implemented and tuned independently, are easily available for replication or comparison, and are representative of tools in common use. The latter offer us better control over parameters.

1. pda. This program is part of the Edimburg Speech Tools Library (check the URL <http://www.ircam.fr/cheveign/data/eusp2001/> for pointers to this and other resources, as well as details of parameters used). It implements the "super resolution pitch determination algorithm" of [18] modified by [3]. Examination of the code suggests that the program uses continuity constraints to improve tracking.

2. pitch_tracker. This program is available for download (see URL). It is described as "loosely based" on the algorithm of [18]. Examination of the code suggests that the program uses continuity constraints to improve tracking.

3. fxac. This program is available as part of the Speech Filing System. Processing is described as: '(i) cubing waveform sample values, (ii) autocorrelation, and (iii) voicing and fundamental frequency decision', using 25 ms windows with a repetition time of 5 ms. Examination of the code suggests that the search range is restricted to 80-400 Hz. The program provides estimates only for speech that is judged "voiced".

4. fxcep. This program is also part of the the Speech Filing System. According to the documentation a '512 FFT is performed on 40 ms windows of the input speech to find the log spectrum, and then an FFT of that provides the ceptrum'. Then the rules of [19] are used to decide whether the speech is voiced, and if so to derive the F_0. Examination of the code suggests that the search range is restricted to 67-500 Hz.

5. sptk. This program is part of the Speech Signal Processing Toolkit (SPTK) v. 2.0. It implements a cepstrum-based method.

6. additive. This program is a locally available implementation of the probabilistic spectrum-based method of [9].

7. acf. This is a simple implementation of the autocorrelation method. The autocorrelation function was calculated as:

$$r_i(\tau) = \sum_{j=1}^{W} x_{j+i} x_{j+i+\tau} \qquad (1)$$

using a 25 ms square window. The function was multipled by a linear ramp that tapered its value to zero at 35 ms (tuned for best performance over DB1). The global minimum over the 40-800 Hz search range gave the period estimate.

8. nacf. Same as above, but the autocorrelation function was normalized according to:

$$r'_i(\tau) = r_i(\tau) / \sqrt{r_i(0) r_{i+\tau}(0)} \qquad (2)$$

9. TEMPO. This is an implementation of the instantaneous frequency method developped by the second author [17]. Briefly the method uses lower harmonic components to calculate F_0 combined with wavelet analysis. The search range was set to 40-800 Hz.

10. YIN. This is an implementation of a method developped by the first author [8]. Briefly, the method combines autocorrelation and AMDF methods[21] with a set of modifications that reduce errors. It does not require an upper limit on the F_0 search range. The lower limit was set to 40 Hz. Neither this method nor the previous one use post-processing.

3. Results and Discussion

The results are shown in Table 1.

Table 1: *Gross error rates for each method and each database. The last column shows the detail of too-high and too-low (often subharmonic) errors.*

	gross error (%)				
method	DB1	DB2	DB3	DB4	Avg. (high/low)
pda	6.7	8.3	9.9	12.0	**9.3** (1.4 / 7.8)
pitch_track	7.1	10.0	24.6	-	-
fxac	8.2	7.1	9.6	7.1	**8.1** (0.54 / 7.6)
fxcep	2.9	6.83	3.1	3.0	**3.2** (0.55 / 2.7)
sptk	3.3	7.6	18.8	12	**10** (9.3 / 0.71)
additive	1.5	1.7	2.2	1.5	**1.7** (0.3 / 1.4)
acf	0.28	0.94	4.0	5.1	**2.7** (2.6 / 0.13)
nacf	0.27	0.82	3.8	5.0	**2.6** (2.5 / 0.08)
tempo	0.61	1.5	4.9	1.1	**1.8** (1.5 / 0.28)
YIN	0.18	0.62	1.15	0.56	**0.55** (0.36 / 0.20)

Error rates vary widely according to methods and databases (missing data for pitch_track are due to a software incompatibility with the large size of files of DB4). The relatively large rates for the freely available methods might have several explanations. One is that several methods implement a voiced/unvoiced decision that we could not disable. "Unvoiced" samples are generally given a value of 0, and thus inflate the "two low" error count. The testing conditions may have been different from those that the methods were tuned for, and we may not have chosen optimal parameters. These factors should be kept in mind when judging the results.

The better rates for the more recent methods seem to generalize well across databases. This is good news, as it shows that progress can still be made in this field, and that there is hope for applications that require reliable estimates. A surprise came from the simple autocorrelation methods that provided extremely good performance over a relatively extensive (28 speakers) database. The fact that it did not generalize to others illustrates the need for large and diversified databases.

We used what we consider the best methodology for evaluation, but it is worth pointing out its limits. First, the need for a clean laryngograph signal may exclude speech modes that produce only weak variations of transglottal resistance, as well as certain morphologies and large amplitude body movement. Second, the need for regularity excludes portions that are clearly voiced, but either irregular (vocal fry, creaky voice, isolated glottal pulses) or with an ambiguous periodicity (diplophony). These are common for certain modes and speakers. They appear to carry intonation, and they must somehow be handled in practical applications, but obviously F_0 is not the best tool for that purpose. More research is required to characterize irregular voiced phonation.

4. Conclusion

We described a methodology for evaluation of F_0-estimation algorithms, and provided results for a set of methods including some freely available in software format on the network, as well as several newer methods. The evaluation was performed over an extensive laryngograph-labeled database aggregated from several sources comprising speech from a total of 38 speakers. The results showed large differences in ranking between methods for different databases, emphasizing the need for extensive test databases. The two most recent methods [17, 8] were nevertheless uniformly more effective than others. The improvement ranges up to an order of magnitude, suggesting that progress has been made in addressing the task of F_0 estimation.

5. Acknowledgments

This work would not have been possible without the laryngograph-labeled speech databases. Thanks are due to Y. Atake and coworkers for creating DB1 and making it available to the authors. Paul Bagshaw was the first to distribute such a database (DB2) freely on the Internet. Nathalie Henrich, Christophe D'Alessandro, Michèle Castellengo and Tuan Vu-Ngoc kindly provided DB3, and Nick Campbell DB4. Thanks are due to the many people who generously created, edited and distributed the software packages that were used for comparative evaluation. We acknowledge that our choice of parameters may not have done them justice, and offer our apologies for this eventuality.

6. References

[1] T. Abe, T. Kobayashi, and S. Imai, "Harmonics tracking and pitch extraction based on instantaneous frequency," Proc. IEEE-ICASSP, pp. 756-759, 1995.

[2] Y. Atake, T. Irino, H. Kawahara, J. Lu, S. Nakamura, and K. Shikano, "Robust fundamental frequency estimation using instantaneous frequencies of harmonic components," Proc. ICLSP, vol. II, pp.907-910, 2000.

[3] P. C. Bagshaw, S. M. Hiller, and M. A. Jack, "Enhanced pitch tracking and the processing of F0 contours for computer and intonation teaching,"Proc. European Conf. on Speech Comm. (Eurospeech), pp. 1003-1006 1993.

[4] E. Barnard, R. A. Cole, M. P. Vea, and F. A. Alleva, "Pitch detection with a neural-net classifier," IEEE Trans. Sig. Proc., vol. 39, pp. 298-307, 1991.

[5] N. Campbell, " Processing a Speech Corpus for CHATR Synthesis.", Proc. ICSP'97 (International Conference on Speech Processing).

[6] A. de Cheveigné, "Speech f0 extraction based on Licklider's pitch perception model," ICPhS, pp. 218-221, 1991.

[7] A. de Cheveigné and G. Peeters, "Core set of audio signal descriptors," ISO/IEC JTC1/SC29/WG11, MPEG00/m5885 2000.

[8] A. de Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," J. Acoust. Soc. Am., vol. submitted, 2001.

[9] B. Doval, "Estimation de la frquence fondamentale des signaux sonores," unpublished doctoral thesis, Université Pierre et Marie Curie, 1994.

[10] H. Duifhuis, L. F. Willems, and R. J. Sluyter, "Measurement of pitch in speech: an implementation of Goldstein's theory of pitch perception," J. Acoust. Soc. Am., pp. 1568-1580, 1982.

[11] P. Hedelin and D. Huber, "Pitch period determination of aperiodic speech signals," IEEE-ICASSP, pp. 361-364, 1990.

[12] D. J. Hermes, "Pitch analysis," in Visual representations of speech signals, M. Cooke, S. Beet, and M. Crawford, Eds. New York: John Wiley, 1993, pp. 3-25.

[13] W. Hess, Pitch determination of speech signals. Berlin: Springer-Verlag, 1983.

[14] W. J. Hess, "Pitch and voicing determination," in Advances in speech signal processing, Sadaoka Furui and M. M. Sohndi, Eds. New York: Marcel Dekker, 1992, pp. 3-48.

[15] ISO/IEC JTC 1/SC 29, "Information Technology Multimedia Content Description Interface Part 4: Audio,", ISO/IEC CD 15938-4, 2000.

[16] 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 reptitive structure in sounds," Speech Communication, vol. 27, pp.187-207, 1999.

[17] H. Kawahara, H. Katayose, A. de Cheveigné, R. D. Patterson: "Fixed Point Analysis of Frequency to Instantaneous Frequency Mapping for Accurate Estimation of F0 and Periodicity," Proc. EUROSPEECH'99, vol. 6, pp. 2781-2784, 1999.

[18] Y. Medan, E. Yair, and D. Chazan, "Super resolution pitch determination of speech signals," IEEE ASSP, vol. 39, pp. 40-48, 1991.

[19] A. M. Noll, "Cepstrum pitch determination," J. Acoust. Soc. Am., vol. 41, pp. 293-309, 1967.

[20] X. Rodet and B. Doval, "Maximum-likelihood harmonic matching for fundamental frequency estimation," J. Acoust. Soc. Am., vol. 92, pp. 2428-2429 (abstract), 1992.

[21] M. J. Ross, H. L. Shaffer, A. Cohen, R. Freudberg, and H. J. Manley, "Average magnitude difference function pitch extractor," IEEE Trans. ASSP, vol. 22, pp. 353-362, 1974.

[22] Vu Ngoc Tuan, C., and d'Alessandro, C. (2000). "Glottal closure detection using EGG and the wavelet transform.", Proc. 4th International Workshop on Advances in Quantitative Laryngoscopy, Voice and Speech Research, Jena, 147-154.