

Proceedings of Meetings on Acoustics

Volume 11, 2010

<http://acousticalsociety.org/>

160th Meeting
Acoustical Society of America
Cancun, Mexico
15 - 19 November 2010
Session 1pSC: Speech Communication

1pSC27. The importance of optimal parameter setting for pitch extraction

Evanini Keelan*, Catherine Lai and Klaus Zechner

***Corresponding author's address: Research and Development, Educational Testing Service, Rosedale Road MS R-11, Princeton, NJ 08541, kevanini@ets.org**

In this study we present a performance comparison for five pitch extraction algorithms: Auto Correlation, Cross Correlation, and Sub-Harmonic Summation (as implemented in PRAAT [Boersma and Weenick (2010)]), the Robust Algorithm for Pitch Tracking implemented in ESPS [Talkin (1995)], and SWIPE' [Camacho (2007)]. Recent research showed that SHS and SWIPE' outperformed the other algorithms on two speech databases with EGG reference values [Camacho (2007)]. That study, however, used a fixed search range of 40-800 Hz for all speakers, regardless of sex or speaker-specific pitch characteristics. In the current study, we adopt the parameter optimization strategy from De Looze and Rauzy (2009) to calculate specific pitch floor and ceiling values for each speaker. Our results show a substantial improvement in accuracy of the AC, CC, and RAPT algorithms when the optimized parameters are used (especially for the female speakers), and all five algorithms show similar performance. The gross error rate for all five algorithms ranges from 0.1% to 0.3% (N=18 098) on the FDA database [Bagshaw (1994)] and from 0.2% to 0.4% (N=11 527) on the Keele database [Plante et al. (1995)]. Our study thus highlights the importance of pre-processing the speech signal to determine optimal speaker-specific parameters for pitch extraction.

Published by the Acoustical Society of America through the American Institute of Physics

1 Introduction

Many studies have compared the performance of different F0 extraction algorithms, and dozens of algorithms have been published (see Chapter 2 in Camacho (2007) for a concise overview). These studies generally focus on the algorithms themselves, and often do not consider the effect of varying the parameters used by the algorithms in F0 extraction. Specifically, the *pitch floor* and *pitch ceiling* parameters (which specify the minimum and maximum F0 values that the algorithm will predict) are often given default values that may not produce the best results for each algorithm under consideration. For example, a recent study demonstrated that the SWIPE' algorithm outperformed 13 other F0 extraction algorithms, but the experiment used unrealistic values for the pitch floor (40 Hz) and pitch ceiling (800 Hz) parameters (Camacho 2007). Since the parameters used for that study were outside the range that would be expected in speech, pitch halving and pitch doubling errors would be more likely than if optimal values for these parameters were used. While using such a low pitch floor value and such a high pitch ceiling value may test the robustness of the algorithms to pitch halving and pitch doubling errors, a more practical approach would be to compare the algorithms with more realistic values for these parameters.

Therefore, this study provides a comparison of several standard F0 extraction algorithms using pitch floor and pitch ceiling values that are optimized for the individual speakers. We adopt a two-pass approach in which F0 values are first extracted using a broad range defined by default pitch floor and pitch ceiling values. Then, this first set of F0 values is analyzed to calculate optimal pitch floor and pitch ceiling values, and a second pass of F0 extraction is conducted using these parameters to extract the final results. Conceptually, this approach is similar to post-processing approaches for F0 extraction in which F0 halving and doubling errors are detected by comparing the pitch estimates to a distribution of pitch estimates from a larger sample of speech from the same speaker, as in Shriberg et al. (2000). In our case, however, we use at most 40 sec of speech data from each individual speaker to estimate the parameters (see Section 2 for details of the two speech corpora used in the experiments).

Using this approach, we show that all of the algorithms under comparison have similar accuracy. Thus, this two-pass approach to F0 extraction is recommended for all researchers using an F0 extraction algorithm that is less robust to a large default range for these parameters; most importantly, this

recommendation applies to the Auto-Correlation algorithm as implemented in the popular acoustics toolkit Praat.

2 Data

In order to evaluate the performance of different F0 extraction algorithms, it is necessary to compare the predicted F0 values with gold standard reference values. There are two publicly available speech corpora with electroglottograph (EGG) values that can be used for this purpose: Paul Bagshaw's database for evaluating Fundamental Frequency Determination Algorithms (FDA) and the Keele corpus.¹

The FDA corpus consists of 50 sentences each read by two speakers, one male and one female. There are a total of 37 declarative sentences and 13 interrogatives (4 yes-no questions and 9 wh-questions). The audio files in the corpus are sampled at 20,000 Hz with 16-bit quantization, and were downsampled to 11,025 Hz for this experiment. Both of the speakers were equipped with an electroglottograph (EGG) while reading the sentences. F0 estimates were taken from the EGG waveform approximately every 5 msec in regions of voiced speech, and these F0 values are used as the reference pitch values for this experiment. For more details about the FDA corpus, see Bagshaw et al. (1993).

The Keele corpus contains recordings of "The North Wind and the Sun" read by 5 female and 5 male speakers. The recordings are approximately 30 seconds in length, and have accompanying F0 measurements that were extracted from the EGG waveforms using an Auto-Correlation algorithm. The audio files in this corpus were also downsampled from 20,000 Hz to 11,025 Hz for this experiment. Since this corpus has several more speakers than the FDA corpus (10 versus 2), it provides a better test of the ability of the different pitch tracking algorithms to generalize across speakers. For more details about the Keele corpus, see Plante et al. (1995).

Table 1 summarizes the details of the two speech corpora used in this experiment.

¹Note that the procedures used to convert the raw EGG values F0 estimates are not perfect. Thus, the F0 reference values in both corpora contain a few gross errors.

Corpus	Speakers	Total Dur.	Mean Utt. Dur.	# Measurements
FDA	2	5 min 32 sec	3.32 sec	18,098
Keele	10	5 min 37 sec	33.7 sec	11,527

Table 1: Information about the two data sets used in the experiment

3 Methodology

Table 2 lists the five F0 extraction methods that were compared in the study. The first column presents an abbreviated name for each algorithm, the second column presents the algorithm's full name, and the third column provides a reference for each algorithm. The SWIPE' and SHS methods were selected because they were shown to be the two highest performing methods in Camacho (2007). The AC method was selected because it is the most widely used method in linguistics studies. Finally, the CC and RAPT methods were also selected because they are available in commonly used speech analysis toolkits. The sources for the algorithms are as follows: Praat (Boersma and Weenick 2010) was used for the SHS, AC, and CC algorithms; the *get_f0* command in the ESPS toolkit was used for the RAPT algorithm; and Kyle Gorman's C implementation of the Matlab code provided by Camacho (2007) was used for the SWIPE' algorithm.²

Method	Full Name	Reference
SWIPE'	Sawtooth Waveform Inspired Pitch Estimator	Camacho (2007)
SHS	Sub-Harmonic Summation	Hermes (1988)
AC	Auto-Correlation	Boersma (1993)
CC	Cross-Correlation	Atal (1968)
RAPT	Robust Algorithm for Pitch Tracking	Talkin (1995)

Table 2: F0 Extraction Methods

The parameter optimization approach conducts a first pitch extraction pass for each audio file using the values of 75 Hz for the pitch floor and 600 Hz for the pitch ceiling in order to determine a rough distribution of the pitch values from that speaker. These initial pitch floor and ceiling values are the

²The SWIPE' code was accessed from the following URL: <http://www.ling.upenn.edu/~kgorman/c/swipe>.

default values in Praat, and are intended to encompass the entire range of values that a given speaker would produce. Then, optimized pitch floor and ceiling values are calculated based on the following equations from De Looze and Rauzy (2009) (where q_{35} represents the 35th quantile for the extracted F0 values and q_{65} represents the 65th quantile):

$$\begin{aligned} pFloor &= q_{35} \times 0.72 - 10 \\ pCeiling &= q_{65} \times 1.9 + 10 \end{aligned}$$

These equations were hand-tuned by De Looze and Rauzy (2009), and were shown to outperform simpler heuristics for pitch floor and ceiling optimization. After the optimized $pFloor$ and $pCeiling$ values are obtained, a second pitch extraction pass is conducted using these optimized values.

The method of evaluating the different F0 algorithms in this experiment closely follows the method described in Camacho (2007:89). Specifically, all values from time stamps where any of the algorithms produced an “undefined” value (i.e., the algorithm labeled that frame as unvoiced) were excluded. Thus only time stamps where all algorithms predicted the frame to be voiced are retained. Again following Camacho (2007:89), the main performance metric used in this study is the Gross Error Rate (GER). This represents the number of measurements where the pitch estimate differs from the reference (EGG) pitch by more than 20%. This metric thus reports the percentage of gross errors (almost always due to pitch-halving or pitch-doubling). To provide a somewhat different view of the performance, Root Mean Squared Error (RMSE) is also reported.

4 Results

Tables 3 and 4 present the results for the FDA and Keele corpora, respectively. In each table, the results using the default pitch floor (75 Hz) and pitch ceiling (600 Hz) are presented first in the columns labeled “Default”. Then, the results using the optimized pitch floor and ceiling values are presented in the columns labeled “Optimized.” In each case, the number of frames that were used in the evaluation is also presented (this number varies between the “Default” and “Optimized” conditions because only frames that were predicted as voiced by all algorithms were considered in the analysis, as described in the previous section).

Method	Default (N=20,167)		Optimized (N=18,098)	
	GER	RMSE	GER	RMSE
SWIPE'	0.2%	8.0	0.1%	7.7
SHS	0.2%	8.8	0.2%	9.9
AC	0.4%	10.3	0.1%	7.5
RAPT	0.6%	11.9	0.3%	11.5
CC	0.7%	11.8	0.1%	7.4

Table 3: Results from the FDA corpus

Method	Default (N=12,252)		Optimized (N=11,527)	
	GER	RMSE	GER	RMSE
SWIPE'	0.2%	5.2	0.2%	5.2
RAPT	0.3%	7.3	0.3%	6.6
SHS	0.4%	7.6	0.4%	6.8
AC	0.7%	8.4	0.3%	5.6
CC	1.1%	10.4	0.3%	5.7

Table 4: Results from the Keele corpus

In Tables 3 and 4 the algorithms are presented in order of decreasing accuracy (as measured by the GER metric) in the “Default” condition. This study thus replicates the result in Camacho (2007:90) that the SWIPE' algorithm performs the best in this condition for these two corpora. The crucial results for this study, however, are shown in the “Optimized” columns: after optimizing the values for the pitch floor and ceiling parameters, all 5 algorithms perform similarly with respect to the Gross Error Rate metric. The results with the optimized parameters for both corpora show a difference of only 0.2% in GER between the best-performing and worst-performing algorithms. When the RMSE metric is used for evaluation, the results also show that the parameter optimization procedure generally leads to all algorithms exhibiting similar accuracy (the results for the SHS and RAPT algorithms on the FDA corpus do not completely fit with this generalization).

The results in Tables 3 and 4 combined the results from the male and female speakers for each corpus. However, an analysis that takes the speaker's sex into account provides a more detailed view of the effect of the parameter

optimization. First, Figure 1 presents the results using the GER metric for the male speakers only. This figure shows that the values from the “Default” and the “Optimized” conditions are nearly identical in each corpus. Thus, there is no noticeable improvement when the optimized parameters are used. This is due to the fact that all of the algorithms perform quite well with the default parameters of 75 Hz (pitch floor) and 600 Hz (pitch ceiling) for the male speakers (the GER for the male speaker in the FDA corpus is 0.0% for all 5 algorithms).

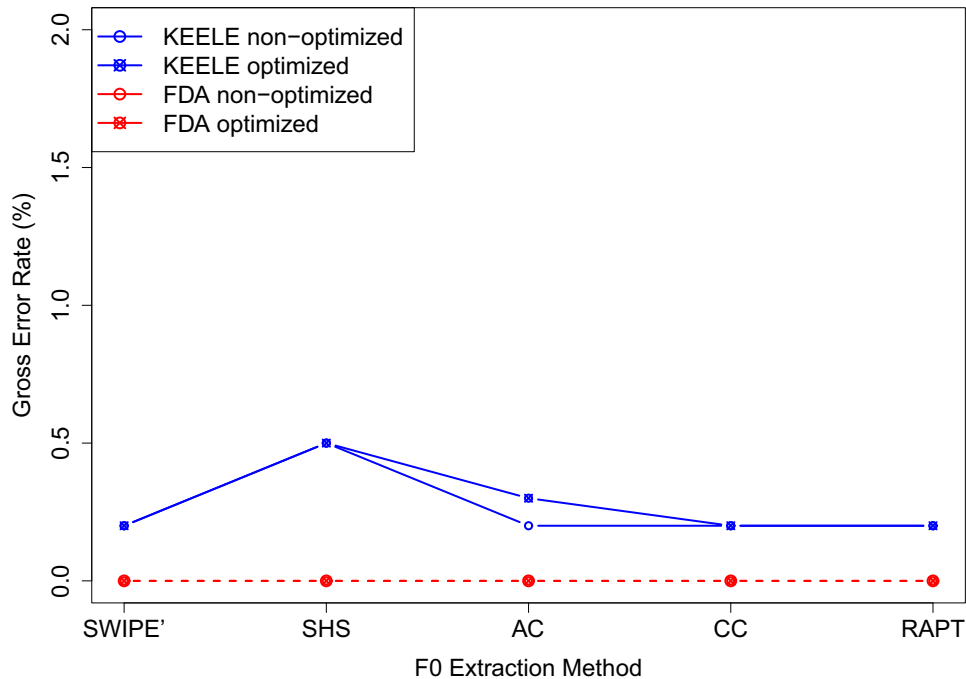


Figure 1: Results for male speakers

Figure 2 shows that the picture is different for the female speakers. For these speakers, the AC, CC, and RAPT algorithms show a large performance improvement after parameter optimization. After obtaining better pitch floor and pitch ceiling parameters, the performance of these three algorithms on the female speakers is nearly identical to the performance of SWIPE' and SHS.

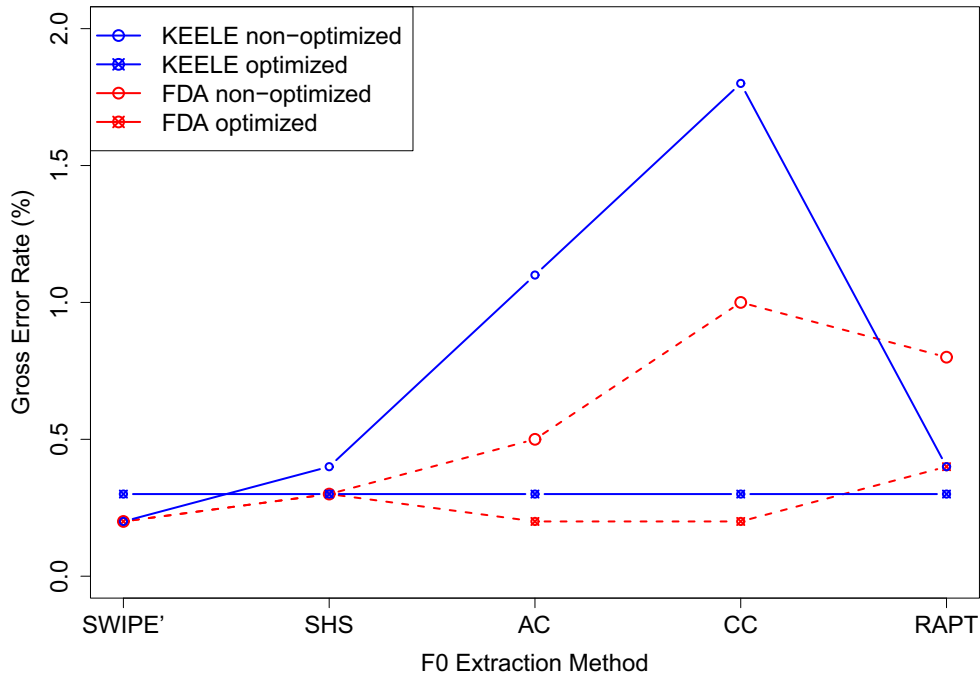


Figure 2: Results for female speakers

5 Discussion

The results presented in the previous section indicate that providing optimal values for the pitch floor and pitch ceiling parameters can substantially improve the overall performance of pitch extraction algorithms. While this result is not surprising, the finding that all 5 algorithms produced similar results after parameter optimization is important. Specifically, the AC, CC, and RAPT algorithms all improved to the level of the SWIPE' algorithm after parameter optimization. This result has practical implications, given that the Praat AC method of pitch extraction is by far the most commonly used method. This study thus indicates that practitioners can continue to use this method and obtain extremely accurate results if they adopt the two-pass approach for parameter optimization and pitch extraction described in this paper.

Another important result from this study was the performance difference

that was observed when the results for the male and female speakers were compared. This analysis showed that the AC, CC, and RAPT algorithms performed substantially worse on the female speakers, and that the main effect of the parameter optimization was to improve performance for these speakers. This suggests that the AC, CC, and RAPT algorithms are more susceptible to pitch halving errors than SWIPE' and SHS.

As discussed in Section 3, the experiment presented in this paper adopted the parameter optimization approach from De Looze and Rauzy (2009). This approach uses four pre-specified parameters for obtaining the optimal pitch floor and ceiling values: the quantiles from the F0 values extracted using the default parameters (q_{35} and q_{65}) and the multipliers used for scaling the quantiles (0.72 and 1.9). In future research we plan to systematically compare different combinations of values for these four parameters to see whether a different set of values could produce even more accurate F0 measurements.

One limitation of this study is that the results were only evaluated using the Gross Error Rate metric after the exclusion of frames that were not hypothesized as voiced by all algorithms under comparison. This approach to evaluation could potentially favor algorithms that hypothesize fewer overall F0 measurements. Therefore, future experiments will also use the voicedness prediction accuracy of each algorithm as part of the evaluation.

References

- Atal, Bishnu S. 1968. Automatic Speaker Recognition Based on Pitch Contours. Doctoral dissertation, Polytechnic Institute of Brooklyn.
- Bagshaw, Paul C., Steven Hiller, and Mervyn A. Jack. 1993. Enhanced pitch tracking and the processing of f0 contours for computer aided intonation teaching. In *Proc. Eurospeech*.
- Boersma, Paul. 1993. 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 17*, 97–119.
- Boersma, Paul, and David Weenick. 2010. Praat: Doing phonetics by computer, version 5.0.38. <http://www.praat.org>.
- Camacho, Arturo. 2007. SWIPE: A Sawtooth Waveform Inspired Pitch Estimator for Speech and Music. Doctoral dissertation, University of Florida.

- De Looze, Céline, and Stéphane Rauzy. 2009. Automatic detection and prediction of topic changes through automatic detection of register variations and pause duration. In *Proc. Interspeech*.
- Hermes, Dik J. 1988. Measurement of pitch by subharmonic summation. *J. Acoust. Soc. Am.* 83:257–264.
- Plante, Fabrice, Georg F. Meyer, and William A. Ainsworth. 1995. A pitch extraction reference database. In *Proc. Eurospeech*.
- Shriberg, Elizabeth, Andreas Stolcke, Dilek Hakkani-Tür, and Gökhan Tür. 2000. Prosody-based automatic segmentation of speech into sentences and topics. *Speech Communication* 32:127–154.
- Talkin, David. 1995. A Robust Algorithm for Pitch Tracking (RAPT). In *Speech Coding and Synthesis*, ed. W.B. Kleijn and K.K. Paliwal, 495–518. Elsevier.