Automatic estimation of pitch range through distribution fitting

Johan Frid and Gilbert Ambrazaitis
SOL, Lund University

Abstract

In this paper we present a method for automatic selection of 'Floor' and 'Ceiling' values for pitch analysis. The method involves fitting a log-normal tied mixture model to distributions of pitch values and using the resulting parameter values. The method's ability to deal with pitch estimation errors in spontaneous speech is demonstrated.

Introduction

When performing pitch analysis, it is common to limit the range of possible values by providing the analysis method with a minimum ('Floor') and a maximum ('Ceiling') value. Setting these parameters so that they are close to the speaker's actual pitch range will greatly improve the analysis as many types of pitch estimation errors will be minimized.

It is, however, somewhat troublesome to select good values. For an unknown speaker or sound file, one really has no idea of what the actual range of the speaker is. If one knows the gender of the speaker, it may be possible to make a rough guess at which parameter values to use (e.g., female: 100 Hz– 500 Hz, male: 75 Hz– 300 Hz), but this is still rather error prone, and for fully automatic analysis, information about gender may not be available.

Our approach to this task will be the method of bootstrapping; we first perform an initial pitch analysis with a very wide range (allowing for both male and female voice ranges; children's voices were ignored for the context of this paper), then we analyse the results, which will result in a guess at the parameters values, which we then feed back into the pitch analysis in order to get the final pitch values.

Earlier work

Sönmez et al. (1997) develop a pitch analysis model that allows estimation of pitch statistics from pitch tracks which contain 'errors' such as doubling and/or halving. The model is a log-normal tied mixture (LTM) with three log-normal distributions with tied means and variances and they estimate the parameters of the model using the Expectation-Maximization algorithm.

Their system is part of a larger speaker recognition system and they unfortunately present very little experimental results of the LTM itself, only some plots of one female speaker are shown. Furthermore, they do neither give any 'Floor' and 'Ceiling' values for their pitch tracker nor present any information on how to guess the starting values for the EM algorithm. On the other hand it is an interesting way to treat the problem; pitch estimation errors are handled not by filtering away the errors, but rather by incorporating them in the model. The model itself is described in enough detail for us to use it in this study.

Edlund & Heldner (2007) analyse percentiles of pitch distributions and compare them with means and standard deviations. They also use a log (the 'semitone') scale. They conclude that percentiles indeed are useful for estimating a speaker's relative pitch.

Pitch estimation errors are handled by looking at the intensity and the modality of the pitch distribution. They state that a filtering procedure removes frames in the lower modes of the intensity and (in case of a bimodal F0 distribution) F0 distributions but further details like intensity thresholds and their method of detecting bi-modality are left out. Furthermore, they do not state which initial 'Floor' and 'Ceiling' levels were used. As these details are not provided we have not tried to replicate their method in this study.

Furthermore, whereas it is clear that filtering away intensity frames with low values should lead to 'better' pitch ranges and also might be justified from production and perception perspectives, we think that there might be

situations where it is advantageous to have a method that relies on one parameter – F0 – only.

De Looze & Hirst (2008) use a somewhat similar procedure. They also compare the percentiles (an alternative term for that is '100-quantiles' and the authors refer to them simply as 'quantiles') with manually estimated levels of maximum and minimum pitch.

They start off from an initial pitch estimate where Floor=60 Hz and Ceiling=750 Hz. This wide pitch range setting is necessitated by the need to initially accommodate both male and female pitch ranges (children may be another issue, but we ignore that in the context of this paper). They then present some empirical justification for setting Floor=0.75*q25 (25th percentile) and Ceiling=1.5*q75 (75th percentile) as minimum and maximum pitch values.

They do not seem to use the log or semitone scale and do not mention how to deal with bimodal F0 distributions and/or pitch halving/doubling. Later papers (De Looze & Hirst 2010) have suggested other variants of the formulae, for instance Floor=0.83*q15 and Ceiling=1.92*q65, but it is not clear whether this works better or not.

Their method may appear somewhat simpler than the other models. On the other hand, it is described in enough detail for us to replicate it.

Pitch distributions

Our strategy for estimating suitable Floor and Ceiling parameters involves looking at the distribution of pitch values for a given speaker. The distribution, roughly, provides us with a profile of a speaker's most used pitch levels and enables us to make a reasonable estimate of the speaker's actual pitch range.

The distribution of any variable may be analysed by constructing a histogram of the data. This assesses the probability distribution of a given variable by depicting the frequencies of observations occurring in certain ranges of values. The histogram is sometimes criticized (Wilkinson 1992) for being strongly affected by the choice of the number of bins (or sub-ranges) and also of being ineffective at depicting the true shape of the distribution. An alternative which often works better is to use the kernel density instead. Figure 1 shows the histogram and the density estimates of the distribution of log(F0) of one male speaker. F0 was estimated using Praat's AC method (Boersma 1993) with Floor=60 Hz and Ceiling=750 Hz. Histogram and density estimates were calculated using R (R Development Core Team 2009); the 'truehist' function in the MASS package (Venables & Ripley 2002) and the 'density' function in the (standard) stats package. The higher values (in the right hand side of the figure, about > 6.0) correspond to pitch levels above 400 Hz and are most likely pitch estimation errors.

It may be worthwhile here to point out that it of course would be easy to get rid of these errors by setting an arbitrary 'Ceiling' at, say, 250 Hz – but – the idea in this paper is exactly NOT to use arbitrary 'Floor' and 'Ceiling' values, but rather to estimate them automatically based on an analysis of all obtained pitch values.

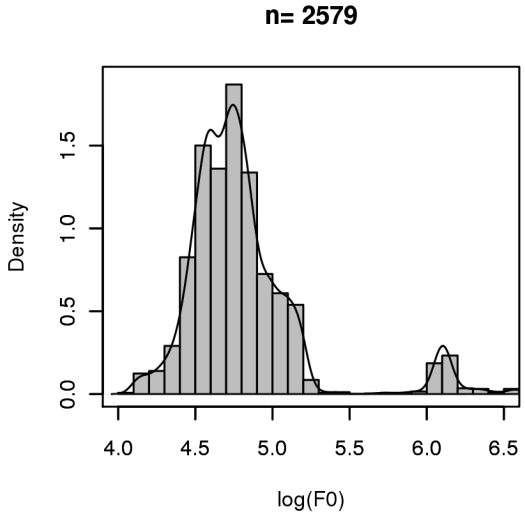

Figure 1. Histogram (vertical bars) and density (solid line) estimates of the distribution of log(F0) of 2579 voiced frames (one speaker).

Distribution fitting

The distribution estimate obtained with the kernel density described above is examined using the method of distribution fitting. Data (in this case density values) may be fitted by proposing a model that provides a description of the data. Some parameters of the model are allowed to vary (they are 'free') and an algorithm then tries to minimize the difference between the predictions of the model and the observed data by choosing suitable values. One common example of a model is the Gaussian (or normal) distribution, that has two free parameters, location and scale, whose starting values sometimes are estimated by the arithmetic mean and standard deviation. Since the distribution of

log(F0) often is more similar to a normal distribution than the distribution of F0 and since we want to deal with pitch halving/doubling we will adopt the log-normal tied mixture of Sönmez et al (1997).

Starting values

Minimization algorithms need reasonable starting values. Since we use a wide range, the result it is likely to contain pitch estimation errors. Therefore the arithmetic mean and the standard deviation will sometimes be quite bad starting values.

Robust estimators

We therefore employ some more robust measures of location and scale of the data that can ignore some of the outliers in the data. Some common robust techniques include an 'M-estimator' for location and the 'Bi-weight mid-variance' for scale (Herrington 2002). Figure 2 shows the difference between model predictions of a log-normal model with arithmetically versus robustly calculated values.

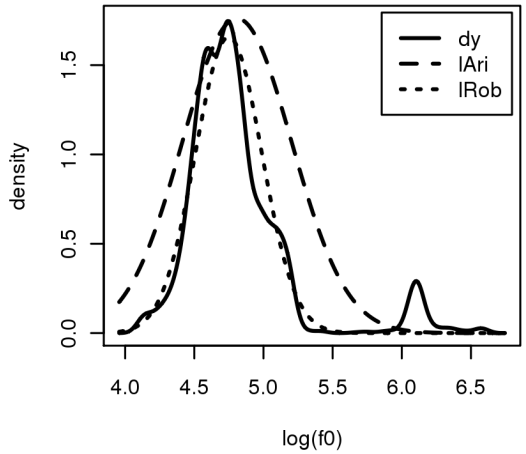

Figure 2. Density (dy) and predictions made by log-normal models with arithmetically (lAri) and robustly (lRob) calculated values.

Note how the robust ('lRob') model follows the log(F0) density values that we deem 'correct' much more closely than the arithmetic ('lAri') model. In this respect, we can say that the robust values are better since the model based on them provides a better description of the data than the model based on arithmetic values. In addition, note that the log-normal models appear normal in Figure 2, since the x-scale is logarithmic.

Mode and half-width-at-half-max

In some cases of bimodal pitch distributions, neither the arithmetic nor the robust method will give good starting values. We therefore also use the mode (the value that occurs the most frequently in a data set or a probability distribution) for location and the half-width-at-half-max (HWHM) for spread.

Since a bimodal distribution may arise either from halving doubling of pitch, any one of the peaks in such a distribution may be the 'correct' one. So we also need use the mode and the HWHM of the second-highest peak.

Outline of the method

Our proposed 'Distribution Fitting' model uses Praat for pitch analysis and R for models and distribution fitting. Here follows an outline:

- Perform an initial pitch analysis with a wide pitch range, accommodating for the pitch ranges of adult men and women. Following De Looze and Hirst (2008), we currently use Floor=60 Hz and Ceiling=750 Hz.
- Get a robust estimate of location and scale of log(F0).
- Estimate the mode and half-width-at-half-max of log(F0) (both raw data and first residual)
- Fit log-normal tied mixture models to the log(F0) distribution using non-linear least-squares with:
  1. robust location and scale
  2. mode and hwhm
  3. mode and hwhm of residuals of 2

  as starting values and 'plinear' and 'port' algorithms (provided by R)
- Reject 2 and/or 3 if their estimated 'location' is:
  - too far away from the robust estimate of location (currently location +/- 0.9*scale)
  - < 120 or > 375
- Return the model with the lowest residual sum-of-squares and calculate a new log-normal distribution with its location and scale parameters
- Calculate Floor and Ceiling from this new model: currently Floor=0.75 * 1st quantile, Ceiling=1.5 * 3rd quantile. These calculations are similar to De Looze and Hirst (2008).

Testing the method

We performed a comparison of 1) the method of De Looze & Hirst (2008), 2) a model based on the robust estimators and 3) the 'Distribution fitting' model described above. Method 2 differs from Method 1 in that instead of using actual quantiles, we estimate the 1st and 3rd quantiles of a log-normal distribution with the parameters values that the robust estimations result in. Below, we will refer to them as the 'lAri', 'lRob' and 'DF' models.

Estimation of 'Floor' and 'Ceiling'

We estimated Floor and Ceiling values with all three methods for 429 speakers (both male and female) from the public part of the Swedia 2000 (Aasa et al. 2000) database. Speech files contain roughly 30s of spontaneous speech. It is somewhat problematic to assess the performance of the methods as we do not have any notion of what the 'correct' values would be for these speakers (no 'Gold standard'). Figures 3-6 show some examples where the DF model makes a good choice but seem problematic for the non-fitting models. In each figure, the vertical lines show the 'Floor' and 'Ceiling' levels, the observed distribution is drawn with a solid line and the predicted distributions are shown using different line types.

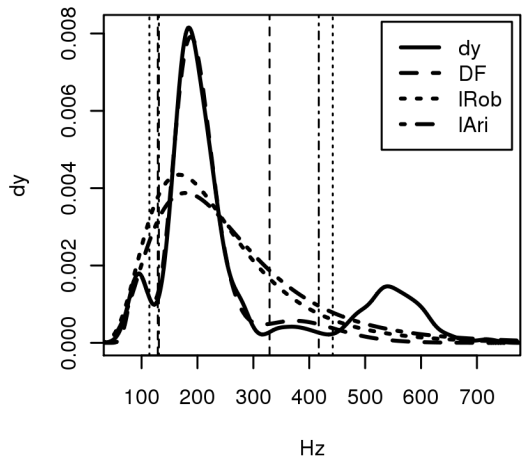

Figure 3. Female speaker; observed (dy) and predicted (DF, lRob and lAri) distributions; vertical lines are estimated 'Floor' and 'Ceiling' levels for each method.

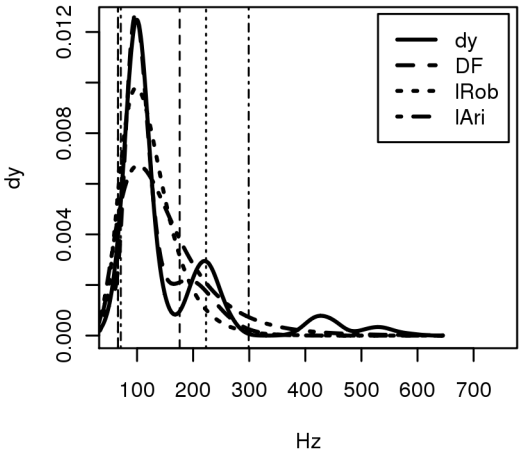

Figure 4. Male speaker; observed (dy) and predicted (DF, lRob and lAri) distributions; vertical lines are estimated 'Floor' and 'Ceiling' levels for each method. Distribution is influenced by pitch levels of a female interviewer.

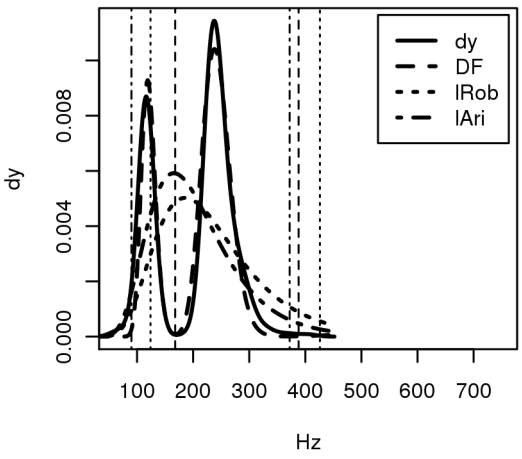

Figure 5. Female speaker; observed (dy) and predicted (DF, lRob and lAri) distributions; vertical lines are estimated 'Floor' and 'Ceiling' levels for each method.

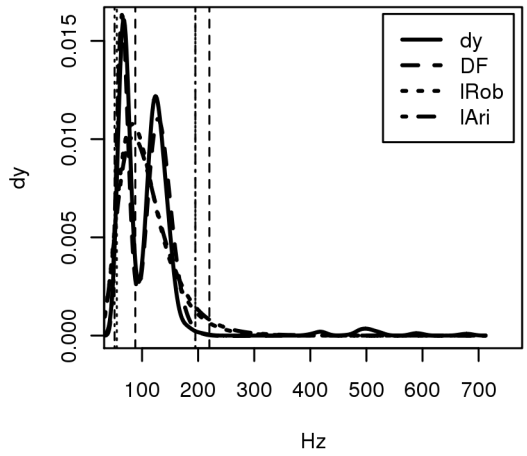

Figure 6. Male speaker; observed (dy) and predicted (DF, lRob and lAri) distributions; vertical lines are estimated 'Floor' and 'Ceiling' levels for each method.

Figure 3 shows a case where background noise increases the 'Ceiling' estimate for the non-fitted models. In Figure 4, the main speaker is male, but there is also a female interviewer. This, of course, adds frames in the female pitch range. The DF model handles this well. Figures 5 and 6 are examples where pitch halving lowers the 'Floor' estimates.

Visual inspection of plots like Figures 3-6 for all speakers indicate that the DF model always makes a reasonable 'Floor' and 'Ceiling' estimate.

Log-normality

We will show another aspect of the DF model: it truncates the data so that it becomes closer to a normal or log-normal distribution, which may be advantageous for further statistical analysis, comparisons, standardisations and grouping of data from different speakers.

In this sub-study we used the Swedia material described above, but also material from TIMIT (Garofolo et al. 1990). This database consists of much more controlled speech material, has 'cleaner' recordings and consequently contains much less pitch estimation errors. We estimate Floor and Ceiling levels by the DF and the lAri models for each speaker in each database. Then we truncate the data and compare the distribution of the remaining values with the log-normal distribution. This is done by comparing the quantiles of the data with the quantiles of the log-normal distribution (this is the same procedure as doing a 'QQ-plot') and computing the correlation coefficients between these quantiles. The method is described in Johnson (2008). Then we compare the correlation coefficient in scatterplots, see Figures 7 and 8.

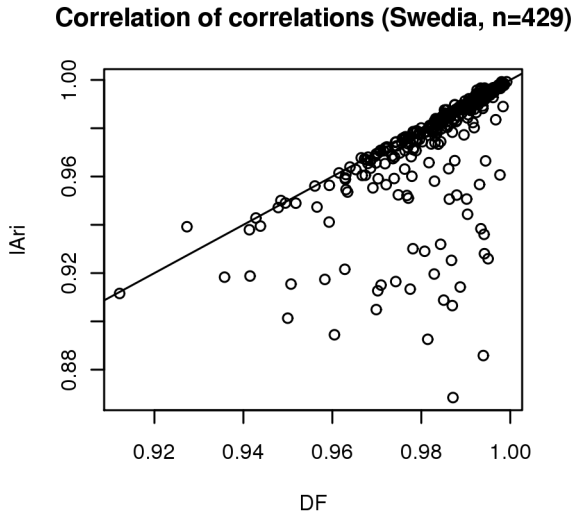

Figure 7. Scatterplot of correlation coefficients for the DF and lAri models with a log-normal distribution. Speakers from the Swedia database.

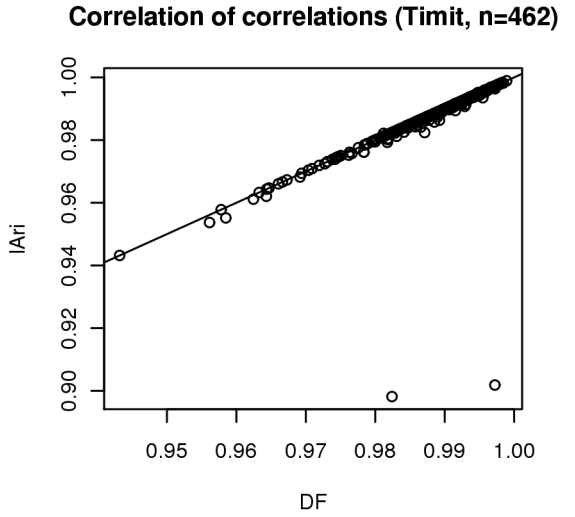

Figure 8. Scatterplot of correlation coefficients for the DF and lAri models with a log-normal distribution. Speakers from the Timit database.

In Figure 7 we see that in the majority of cases the two models show similar performace, more or less, but there is a also a noticeable amount of

cases (around 8% of the data) where the correlation is higher with the DF model.

Figure 8 shows the result for TIMIT. Here we see much less difference, as might have been expected since pitch halving and doubling errors are scarcer. Yet, we see two cases of creaky speakers (in the lower-right hand corner of the plot), which are made to correlate well with the log-normal distribution with the DF model, but not with the lAri model.

Summary and conclusion

In this paper we have presented a procedure that determines reasonable upper and lower limits of F0 automatically from an initial F0 estimate with a very generous range. It is insensitive to the individual speaker's actual pitch range and accommodates both for men and women (we have not tested on children).

We show that compared to quantile-based models the DF model performs at least as well for clean, controlled speech and – based on a qualitative analysis – better for spontaneous speech in less controlled set-ups.

The method can be used for re-analysis of pitch with reasonable floor and ceiling values. It may also be used as a speaker's 'pitch range' to which one can 'normalize' or 'standardize' or determine a 'base-value' in the sense of Traunmüller & Eriksson (1995) (see also Traunmüller 1994 and Lindh & Eriksson 2007).

Further studies may include how to model the truncated values. A possible investigation is to see whether they are best modelled by a log-normal model or if there might be other models and/or transformations that describe the data better.

Acknowledgements

This work is supported by a grant from the Swedish Research Council (VR) project 2009-1566.

References

Aasa A, Bruce G, Engstrand O, Eriksson A, Segerup M, Strangert E, Thelander I & Wretling P (2000). Collecting dialect data and making use of them: an interim report from Swedia 2000. *Proceedings of Fonetik 2000*, University of Skövde, pp. 17-20.

Boersma P (1993) Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. In *IFA Proceedings* 17, pp. 97-110.

De Looze C & Hirst D (2008). Detecting changes in key and range for the automatic modelling and coding of intonation. *Proceedings of International Conference on Speech Prosody* (4 : 2008 avril 6-9 : Campinas, Brazil). 2008, pp. 135-138.

De Looze C & Hirst D (2010). Integrating changes of register into automatic intonation analysis. *Proceedings of Speech Prosody* (2010 : Chicago, USA) [Forthcoming].

Edlund J & Heldner M (2007). Underpinning /nailon/: automatic estimation of pitch range and speaker relative pitch. In C. Müller (Ed.), *Speaker Classification* II (Vol. LNAI 4441, pp. 229-242). Berlin, Germany: Springer-Verlag.

Garofolo J, Lamel L, Fisher W, Fiscus J, Pallett D & Dahlgren N (1990). DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CD-ROM. National Institute of Standards and Technology. NTIS Order No. PB91-505065.

Herrington R (2002). *Using Robust Mean and Robust Variance Estimates to Calculate Robust Effect Size* [WWW document]. URL http://www.unt.edu/benchmarks/archives/2002/july02/rss.htm

Johnson K (2008) *Quantitative Methods in Linguistics*. Oxford: Blackwell.

Lindh J & Eriksson A (2007). Robustness of Long Time Measures of Fundamental Frequency. *Proc. Interspeech* 2007, pp. 2025-2028.

R Development Core Team (2009). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Sönmez M K, Heck L, Weintraub M & Shriberg E (1997). A Lognormal Tied Mixture Model of Pitch for Prosody-based Speaker Recognition. *Proc. EUROSPEECH 97*, Rhodes, Greece, September 1997, Volume 3, pp. 1391 – 1394.

Traunmüller H (1994). Conventional, biological, and environmental factors in speech communication: A modulation theory. *Phonetica* 51, pp. 170-183

Traunmüller H & Eriksson A (1995). *The frequency range of the voice fundamental in the speech of male and female adults*. Unpublished Manuscript (can be retrieved from http://www.ling.su.se/staff/hartmut/aktupub.htm).

Venables W N & Ripley B D (2002). *Modern Applied Statistics with S*. Fourth Edition. Springer, New York. ISBN 0-387-95457-0.

Wilkinson L (1992). Graphical displays. *Statistical Methods in Medical Research*, 1, pp. 3–25.