Temporal stability of long-term measures of fundamental frequency

Pablo Arantes[1], *Anders Eriksson*[2]

[1]Languages and Linguistics Department, São Carlos Federal University, Brazil
[2]Department of Linguistics, Stockholm University, Sweden

pabloarantes@gmail.com, anders.eriksson@ling.su.se

Abstract

We investigated long-term mean, median and base value of F0 to estimate how long it takes their variability to stabilize. Change point analysis was used to locate stabilization points. In one experiment, stabilization points were calculated in recordings of the same text spoken in 26 languages. Average stabilization points are 5 seconds for base value and 10 seconds for mean and median. Variance after the stabilization point was reduced around 40 times for mean and median and more than 100 times for the base value. In another experiment, four speakers read two different texts each. Stabilization points for the same speaker across the texts do not exactly coincide as would be ideally expected. Average change point dislocation is 2.5 seconds for the base value, 3.4 for the median and 9.5 for the mean. After stabilization, individual differences in the three measures obtained from the two texts are 2% on average. Present results show that stabilization points in long-term measures of F0 occur earlier than suggested in the previous literature.

Index Terms: fundamental frequency, long term measurements, forensic phonetics

1. Introduction

The study of statistical measures of location or preferred value of the voice fundamental frequency (F_0) of an individual has at least two applications. The first one is the development of F_0 contour normalization procedures [1][2][3], that are used to factor out the most common value of a particular speaker so that variation due to linguistic components of the contour becomes more evident. In the other important usage, knowledge of preferred F_0 value of individual speakers is often relevant to speaker comparison in forensic case work (see [4] and references therein), speech technology applications [5] and the development of security systems that require user authentication by voice, for instance (see [6] for a recent survey of the field).

When it comes to the estimation of preferred F_0 value, important issues are (i) what long-term statistical measures are better suited to do it and (ii) how long should a speech sample be in order to the estimated value be representative of an individual's speech? Regarding the choice of estimator, the literature mostly cites the mean and standard deviation (for an overview of the subject see [4]). Given that long-term F_0 distributions are usually skewed towards higher values, it seems advisable to compare the mean to alternative measures that make no assumptions regarding normal distribution of sample values. To fill this gap, we propose to compare the mean to two other measures, the median and the base value. The base value can be conceived as a neutral and speaker-specific value of F_0 below which maintaining phonation becomes difficult and to which speakers return to after excursions of linguistic or expressive value (see [7] for a detailed explanation and section 2.4 for the definition implemented in this study). Both the base value and the median are quantile-based measures that are robust to the presence of skewness or extreme values in the F_0 sample, although the median does not share all of base value's properties.

Regarding the appropriate minimum length of speech sample required for long-term mean to stabilize, the literature shows no definite consensus. Eriksson [4] makes reference to five different estimates, ranging from 14 seconds to two minutes. There is also no agreed upon objective way of estimating when a long-term measure has reached a stable point, most researchers resorting to visual inspection of trajectories of cumulative mean. In this study we explore a statistical technique called change point analysis as a way to objectively compare the performance of the three measures studied here.

2. Materials and Methods

2.1. Language effect

To study the possible effect of language on the variability of measure of location of F_0, a set of recordings of speakers of 25 languages reading the "The North Wind and the Sun" passage were analyzed. The recordings are publicly available on the website of the International Phonetic Association (IPA). One recording of a Brazilian Portuguese speaker reading the same text was included in the sample. Sixteen speakers are male. The sample includes languages of eight linguistic families: Afro-Asiatic, Sino-Tibetan, Indo-European, Uralic, Niger-Congo, Altaic, Tai-Kadai and Turkic.

2.2. Text effect

To test the possible effect of the text being read on the time it takes the long term measures of F_0 to stabilize, recordings of four speakers reading two different texts were analyzed.

The texts are the "North Wind and the Sun" passage translated to Brazilian Portuguese and a passage of "A Menina do Narizinho Arrebitado" by Brazilian writer Monteiro Lobato. The first is coded text 1 and the second text 2. Text 2 is phonetically balanced in the sense of having all Brazilian Portuguese (BP) phonemes, while in translating and adapting the "North Wind and the Sun" passage to BP the goal was to be faithful to the semantic content and not to make sure all the phonemes in the language were being used.

The recordings are by two male and two female speakers of two Brazilian states, São Paulo and Minas Gerais. Speakers from São Paulo and Minas Gerais are referred to by the sp1 and sp2 labels, respectively, followed by -f or -m to indicate if it is a female or male speaker.

10.21437/SpeechProsody.2014-219

2.3. Acoustical analysis

F_0 contours for every recording analyzed were extracted with the help of a Praat script that implements a heuristic suggested by Hirst [8] that tries to minimize extraction errors such as octave or fifth jumps by optimizing floor and ceiling values passed to Praat's auto-correlation F0 extraction algorithm[1]. Remaining errors were hand-corrected. Further processing of F_0 contours to obtain cumulative measures of location was done by a second Praat script written specifically for this purpose.

2.4. Measures of location

The following statistical measures of location were investigated:

- Arithmetic mean
- Median (50th-quantile of the sample F_0 values)
- Base value (7th-quantile of the sample F_0 values)[2]

All measures were taken cumulatively from the first voiced frame up to the last in non-overlapping steps of 200 ms. All F_0 values within each 200 ms interval are included in the computation of the measures. The number of F_0 samples contained in each 200 ms step depends on the floor parameter provided to Praat's F_0 extraction algorithm. In the IPA languages sample, the average minimum value for male speakers was 70 Hz and 120 Hz for female speakers, which gives us 20 values for male speakers and 32 for female speakers each 200 ms.

In the IPA languages sample, the median duration of recordings was 38 seconds with values ranging from 25 seconds (Galician) to 66 seconds (Thai). In the text effect experiment, recordings of text 1 have an average duration of 32 seconds and recordings of text 2 have an average duration of 41.3 seconds.

Mean and median values are usually close, but base value, by definition, is smaller than both. Since here we are more interested in how their variability changes over time and not specially in their absolute values, a normalization procedure was applied so that the three time series can be seen in the $[0, 1]$ interval. This was done by means of formula 1, where f_i is the i^{th} raw F_0 value in a given contour and $f_{\min}$ and $f_{\max}$ are respectively the minimum and maximum values in the contour:

$$(f_i - f_{\min})/(f_{\max} - f_{\min}) \tag{1}$$

The normalized values were used for visualization purposes only. The statistical analysis were carried out on the raw values (in Hz) of the cumulative measures.

2.5. Statistical analysis

Our main interest is to determine how long it takes for the time series defined by the cumulative measures of location of F_0 studied here to have its variability reduced to what could be considered a stable value. In most of the literature on the subject, what we are calling the stabilization point has been determined by visual inspection of the time series of cumulative measure of location. Although the visual inspection can be useful, it would be important to develop a less subjective and more automatic way of determining stabilization points.

A statistical technique called change point detection analysis was used to attain a greater level of objectivity in determining stabilization points. This technique estimates the point in time at which underlying statistical properties (mean, variance or both) of a time series change. A function of the R package changepoint [9] was used in the analysis to find a point that divides the time series in in two parts having different variances and tests the hypothesis that the two values are significantly different. We searched for single variance change points in cumulative mean, median and base value time series. Since distribution of cumulative mean, median and base value are highly skewed, an algorithm that does not assume that the values in the time series follow a normal distribution was used.

3. Results and discussion

3.1. Language effect

Figure 1 shows temporal evolution of cumulative normalized measures of location. Wide-range fluctuations in the three estimators are a general trend across languages, specially at the first seconds of the recordings. As estimators' values are computed over longer stretches of time, the range of fluctuations gets increasingly smaller, but with notable differences between the languages: in some cases, variability quickly drops (e.g, Arabic, German, Galician, Hungarian, Slovene and Swedish) whereas, in other cases, the reduction seems to be more gradual (e.g., Brazilian Portuguese, Catalan, French, Irish and Korean). For all languages, with the possible exception of Turkish, the cumulative value of the three estimators tend to reach a stable value at some point, usually within the first fourth of the recording duration.

Table 1 lists the temporal location of change points for the 26 languages investigated as determined by the change point analysis described in section 2.5 as well as ratio of variance before and after the change point. Figure 2 shows a box plot of change points broken down by typical measure type.

One of the main findings is that change points in this sample seem to happen in the low range of values suggested in most of the previous literature on the subject or even earlier than that (see section 1). The other main finding is that the base line tends to stabilize a little earlier (5 seconds) then mean and median (about 10 seconds). The base value change points are also less variable (median absolute deviation of 2.2) than mean and median (MAD of 6.2 and 7.6 respectively). Inspection of variance reduction factors in Table 1 suggests that in fact the points identified by the change point analysis can be considered stabilization points. Base value also has a superior performance in terms of variance reduction: base line has an average reduction factor of 120 and the mean and the median an average factor of 48.

3.2. Text effect

Figure 3 shows cumulative values of long-term measures of typical F_0 for four speakers reading two different texts. Table 2 lists change points in the three measures as well as variance reduction factors.

For a given long-term estimator, the change point and the estimator value for a given speaker would ideally be the same, regardless of the text being read. Strictly speaking, that was not the case for the four speakers in our sample: average absolute difference between change point for text 1 and text 2 is 4.9 seconds, with a minimum of 0.4 and a maximum of 21.1 seconds. 75% of the differences are under 6 seconds. Considering

[1] Available at http://code.google.com/p/praat-tools/

[2] In [7], the authors say that the base value can "as a rule of thumb, be expected to be about 1.5 σ below his average F_0". To Define the base value in terms of a quantile is for all practical purposes equivalent to the definition based on standard deviation, assuming F_0 values are normally distributed.

Table 1: *Change point locations (in seconds) in the mean, median and base value time series for the 26 languages investigated. Ratio of variance before and after change point are shown in parentheses.*

language	mean	median	base value
Amharic	11 (44)	16.2 (137)	4.8 (69)
Arabic	4.2 (242)	4.6 (226)	4.6 (619)
Brazilian Portuguese	10.4 (33)	10.8 (33)	5.2 (41)
Bulgarian	16.2 (48)	15.6 (79)	3.8 (440)
Cantonese	6.2 (26)	7 (8)	6 (41)
Catalan	11 (40)	10.6 (33)	11.2 (31)
Croatian	0.8 (62)	12.4 (162)	0.6 (180)
Czech	6.8 (86)	4.8 (77)	8 (204)
Dutch	8.2 (50)	10.2 (36)	4.4 (778)
English	11.6 (48)	1.6 (5)	2.4 (63)
French	15 (7)	16 (8)	3.4 (22)
Galician	5.2 (28)	4.8 (67)	5.2 (100)
German	5.2 (141)	5.8 (135)	3.8 (26)
Hindi	10 (112)	16.4 (322)	10 (84)
Hungarian	2.4 (194)	3.8 (171)	4 (217)
Igbo	7.8 (19)	8.8 (2)	21 (222)
Irish	14.4 (7)	12.6 (17)	15 (10)
Japanese	6 (183)	11.6 (58)	0.4 (1772)
Korean	21.4 (18)	21.6 (13)	1 (3)
Persian	20.6 (27)	20.6 (21)	4.6 (3)
European Portuguese	11 (116)	6.4 (81)	5.8 (219)
Sindhi	15 (66)	15.4 (44)	6.4 (172)
Slovene	9 (141)	14.8 (21)	0.8 (232)
Swedish	6.2 (108)	2.8 (504)	15.4 (271)
Thai	22.2 (23)	24.4 (52)	27.6 (137)
Turkish	20 (11)	3.2 (3)	4.6 (16)

that the standard deviation of change point location on the IPA language sample is 6 seconds, the differences in change point location between text 1 and text 2 are less then what would be expected when comparing samples of different languages. A comparison of the differences between the raw values (in Hz) of the cumulative mean, median and base value of text 1 and text 2 at the time of change point shows that on average the difference is 2%, with a range going from 0 to 9%, the four speakers polled. 90% of the differences are under 4%, i.e., less than one semitone. These data indicate there is an effect due to text whose magnitude is slightly smaller than that due to language.

None of the texts yielded overall earlier change points or greater variance reduction factors. The only exception to that is the long-term median of text 2, whose change points are earlier than those of text 1 for all four speakers. It's not clear if the behavior of the median can be attributed to the fact that text 2 is phonetically balanced and why only the median should be affected by this particular feature of text 2.

4. Conclusions

We set out to compare three long-term measures of typical value for F_0, namely the mean, median and base value, in two respects: how long it takes for each measure to achieve a more or less stable level of variability and how much they are affected by language and text.

Our results indicate that long-term measures tend to stabilize at most 30 seconds after the begining of a recording of read speech, with median times around 10 seconds. Of the three measures, the base value seems to yield the earliest and less variable change points, confirming earlier findings suggesting its robustness against factors such as variation in speaker emotional state, vocal effort and channel quality.

Table 2: *Change point (seconds) in the mean, median and base value for the four speakers and the two texts. Ratio of variance before and after change point are shown in parentheses.*

speaker	text	mean	median	base value
sp1-f	1	9.8 (18)	9.6 (21)	9.6 (6)
sp1-f	2	15.2 (28)	5.2 (13)	5.6 (83)
sp1-m	1	10.2 (92)	10.2 (58)	6.2 (33)
sp1-m	2	4.8 (22)	4.8 (42)	5.2 (7)
sp2-f	1	7.2 (11)	7.2 (4)	9.4 (9)
sp2-f	2	28.2 (8)	5.6 (6)	9.8 (39)
sp2-m	1	10.4 (33)	10.8 (33)	5.2 (41)
sp2-m	2	24 (20)	8 (14)	13.8 (39)

The results also show that both language and the specific text being read seem to cause variability in stabilization points. The base value is less affected by the language effect than the mean and median measures. The text effect is slightly smaller than the language effect and the three measures seem to be equally affected by it.

In order to increase the accuracy of the results obtained in the present study, in follow-up studies we are going to increase the number of speakers in the languages investigated and the length of the recordings. Two promising avenues of investigation worth exploring are the effects of different speaking styles (reading vs. spontaneous speech, for instance) and non-contemporaneous recordings on the temporal stability of long-term measures.

5. References

[1] Jassem, W. "Normalisation of F0 curves", in Fant, G. and Tahtam, M. [Eds], Auditory Analysis and Perception of Speech, 523-530, Academic Press, 1975.

[2] Rose, P. "How effective are long term mean and standard deviation as normalisation parameters for tonal fundamental frequency?", Speech Communication 10, 229–247, 1991.

[3] Maidment, J. A. and Garca Lecumberri, M. L. "Pitch Analysis Methods for Cross-Speaker Comparison", Proceedings of ICSLP 1996, v. 4, 2247–2249, 1996.

[4] Eriksson, A., "Aural/Acoustic vs. Automatic Methods in Forensic Phonetic Case Work", in Neustein, A. and Patil, H. A. [Eds], Forensic Speaker Recogntion: Law Enforcement and Counter-terrorism, 41-69, Springer-Verlag, 2011.

[5] Ferrer, L, Shriberg, E. and Stolcke, A. A prosody-based approach to end-of-utterance detection that does not require speech recognition. Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2003.

[6] Müller, C. [Ed], Speaker Classification I: Fundamentals, Features and Methods, Springer-Verlag, 2007.

[7] Traumüller, H. and Eriksson, A. "The frequency range of the voice fundamental in the speech of male and female adults". Manuscript. Retrieved via http://www2.ling.su.se/staff/hartmut/f0_m&f.pdf.

[8] Hirst, D. "The Analysis by Synthesis of Speech Melody: from Data to Models", Journal of Speech Sciences 1(1):55-83, 2011.

[9] Rebecca Killick and Idris Eckley. "changepoint: An R package for changepoint alysis". R package version 1.1. http://CRAN.R-project.org/package=changepoint, 2013.

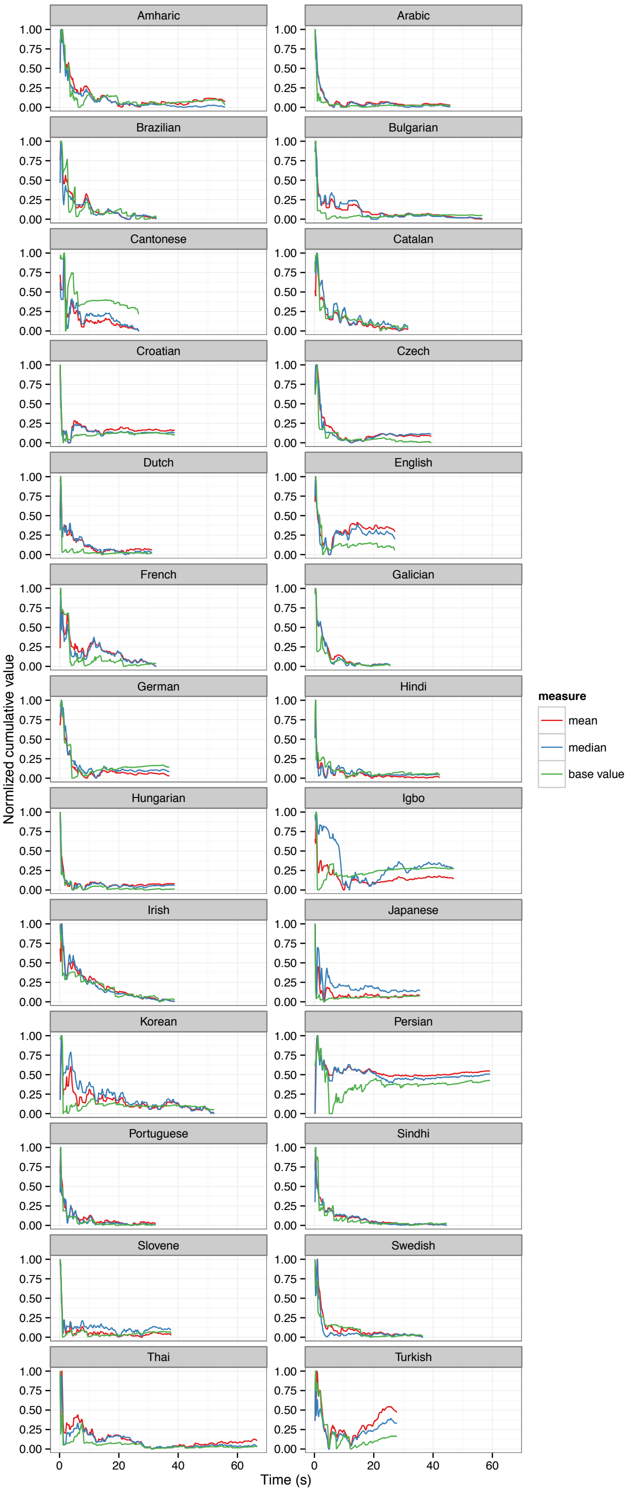

Figure 1: *Language effect on typical F_0 value. Vertical axis shows normalized cumulative mean, median and base value for 26 languages.*

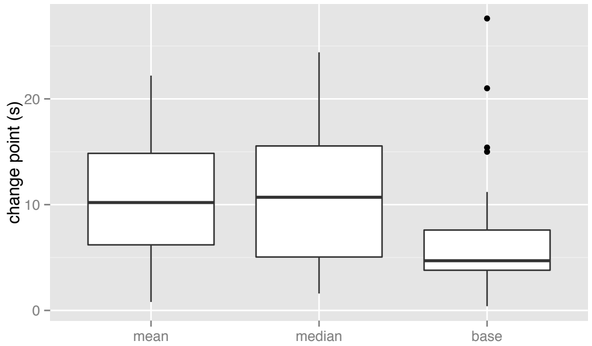

Figure 2: *Box plot of change points (seconds) of mean, median and base value time series for the 26 languages in the IPA sample.*

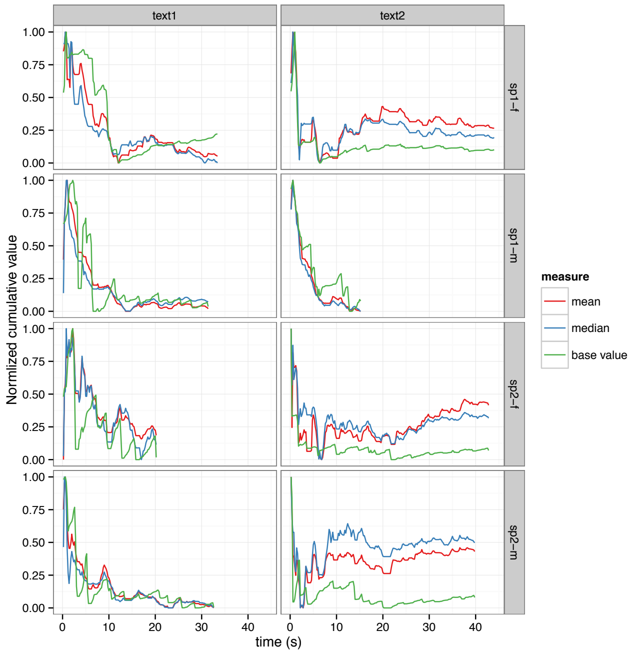

Figure 3: *Normalized cumulative mean, median and base value for the four speakers and the two texts.*