INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Effect of Language, Speaking Style and Speaker on Long-term F0 Estimation

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

[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, suskagutz@gmail.com

Abstract

In this study, we compared three long-term fundamental frequency estimates – mean, median and base value – with respect to how fast they approach a stable value, as a function of language, speaking style and speaker. The base value concept was developed in search for an f_0 value which should be invariant under prosodic variation. It has since also been tested in forensic phonetics as a possible speaker-specific f_0 value. Data used in this study – recorded speech by male and female speakers in seven languages and three speaking styles, spontaneous, phrase reading and word list reading – had been recorded for a previous project. Average stabilisation times for the mean, median and base value are 9.76, 9.67 and 8.01 s. Base values stabilise significantly faster. Languages differ in both average and variability of the stabilisation times. Values range from 7.14 to 11.41 (mean), 7.5 to 11.33 (median) and 6.74 to 9.34 (base value). Spontaneous speech yields the most variable stabilisation times for the three estimators in Italian and Swedish, for the median in French and Portuguese and base value in German. Speakers within each language do not differ significantly in terms of stabilisation time variability for the three estimators.

Index Terms: acoustic phonetics, speech acoustics, forensic phonetics

1. Introduction

The fundamental frequency base value was derived in the spirit of the Modulation Theory of Speech (MTS) [1] by Traunmüller and Eriksson [2, 3]. It arose from a need for producing stimuli for an experiment aimed at determining an appropriate scale for the perception of fundamental frequency variation [4]. Excursion size should be varied in equal steps from almost monotonous to very lively leaving everything else unchanged. This kind of control cannot be accomplished by a human speaker so it was necessary to find a way of producing naturally sounding stimuli by speech synthesis.

It is well known that simply expanding the variation around the mean quickly produces highly unnatural stimuli. It is also known that f_0-distribution is normally skewed towards higher frequencies. So the question arose, if there is anything in the fundamental frequency produced by a given speaker that remains reasonably unchanged when prosody is varied - a kind of eigenvalue for f_0 variation. Based on a number of studies (for references see [3]) it was possible to derive an eigenvalue or a "base value" as it was called. The procedure was first described in [2]. It also appears in an unpublished manuscript [3] available on-line. In these papers the base value was defined as 1.43 Standard Deviations below the mean. It corresponds to the 7th percentile which is the definition used here. Further information may be found in the paper describing the experiment for which the base value was used [4].

If the base value is, at least in theory, independent of excursion size it should be better suited as a speaker specific measure for forensic purposes than the mean or median. Given how the base value was derived one would predict that it would be less sensitive to fundamental frequency variation caused by variation in emotional attitude, for example. This prediction has been tested in [5], where it was shown to vary minimally as a function of simulated attitude in contrast to the median and mean. In the same study, the base value was also shown to be more robust with respect to variation in transmission channel quality. It was also shown to be more robust to moderate vocal effort variation.

In a recent study [6], the base value has been tested in automatic speaker comparison. The database used in the experiment was the Forensic Brazilian Portuguese Corpus (CFPB), built by the Federal Police of Brazil, containing studio quality recordings by 206 male and 50 female speakers. Equal Error Rates (EER) were calculated using Multivariate Kernel-Density for a number of long term fundamental frequency parameters such as base value, mean, median, standard deviation etc. and combinations thereof. When used separately the base value produced the lowest EERs (16.1%) followed by mean, mode and median (All 22%). The best combination (13.1%) was base value + median.

In another recent study [7] that looked into the effect of language, speaking style and speaker sex on a number of statistical measures of f_0, the base value and the median deviation from the base value were found to be fairly robust against the influence of these factors.

Voice samples in casework, are often short and there are therefore many studies where the purpose has been to estimate how long samples need to be for a reliable speaker-specific f_0 estimate. Suggested minima vary considerably ranging between 14 seconds [8] and 2 minutes [9]. In a study of a Chinese dialect, Rose [10] found that values stabilised after about 20 seconds. Braun [11] based on a literature review, suggests that 15–20 seconds should be sufficient. For details see [12, pp. 49–50].

In a previous study by two of the present authors [13], the temporal stability of three different long-term measures of fundamental frequency, mean, median and base value, were compared. By temporal stability we mean how far into a given recording the values stabilise. Change point analysis [14] was used to locate stabilisation points. In one experiment, stabilisation points were based on recordings of a text read in 26 languages. Average stabilisation points were 5 seconds for the base value and 10 seconds for mean and median. In another experiment, four speakers read two different texts each. Stabilisation points for the same speaker across the texts do not exactly coincide as would be ideally expected. Average change point dislocations were 2.5 seconds for the base value, 3.4 for the median and 9.5 for the mean. The stabilisation points in the study occurred earlier than suggested in most previous studies. In [13] one speaker per language was tested. In the present study we have 10 speakers per language and they contribute more stabilisation estimates so that it is possible to investigate within- and

http://dx.doi.org/10.21437/Interspeech.2017-449

between-speaker variability. Also, speaking style is introduced as an independent variable in the experimental design.

2. Material and methods

2.1. Speakers and speech material

The speech material used here was a subset of a database of recordings used for a study of lexical stress in a number of languages. The data used in the present study were recordings in Brazilian Portuguese, British English, Estonian, French, German, Italian and Swedish by 5 male and 5 female speakers for each language. Great care had been taken in selecting the speakers to minimise variation due to regional variation and age. They all spoke a well-defined regional standard. Speaker age variation was the same for all languages within narrow margins. For the entire database ages ranged between 18 and 35, with most speakers in the 20-30 year range. The averages ranged between 23 and 26 for the different languages. The speakers were also closely matched with respect to educational background. All speakers were native speakers and the recordings were all made at universities in the countries were the studied languages are spoken. The data represent three different speaking styles – spontaneous speech, read phrases and read words. Spontaneous speech was elicited in informal interviews by a native speaker. Transcriptions of these recordings were used to produce manuscripts for the other two speaking styles. Phrases were selected where speech was fluent, had no speech errors and contained suitable target words. At a later stage, the speakers were then called back and asked to read the phrases and words they had produced in their spontaneous speech. This way we obtained identical linguistic content in all three speaking styles. For a more detailed description please consult [15, 16, 17, 18].

2.2. Acoustic analysis

Before the f_0 extraction phase, audio files were pre-processed. Stretches that contained the speech of the interviewer or experimenter, overlap between speaker and interviewer/experimenter and non-speech events were silenced. This was done to minimise f_0 extraction errors.

f_0 contours were extracted using a Praat script that implements a heuristics suggested by Hirst [19], that optimises floor and ceiling values passed to Praat's *To Pitch (ac)* autocorrelation-based extraction function [20]. f_0 extraction is a two-pass operation. The relevant parameters the algorithm manipulates are floor and ceiling f_0 values. In the first pass, the Pitch object is extracted using 50 and 700 Hz as floor and ceiling estimates. In the second pass another Pitch object is extracted using optimal values for floor and ceiling estimated from the voiced samples in the first Pitch object. The values are obtained using the following formulae:

$$\begin{aligned} f_{\text{floor}} &= 0.7q_1 \\ f_{\text{ceiling}} &= 1.5q_3, \end{aligned}$$

where q_1 and q_3 are respectively the first and third quartiles of the voiced samples in the first Pitch object. Hirst suggests that the constant for the ceiling value can be set to 2.5 in case the speaker makes use of an extended range. Another Praat script searched for abrupt value changes between consecutive frames, flagging them as potential errors[1]. Later they were checked individually and corrected by an analyst trained for the task. Most errors commonly detected by this procedure were octave halving or doubling and incorrect voicing, usually in fricatives or transient noise of plosives. Cases such as incorrect unvoicing of frames, that can occur during glottalised phonation, had to be found by the analyst by comparing the f_0 contour with both the oscillogram and spectrogram.

The next step was the definition of what parts of the audio file (and corresponding f_0 contour) were to be analysed. They were manually labelled in a dedicated tier of a TextGrid object. In the spontaneous style, all interpausal units were labelled. To do that, f_0 resetting and pauses longer than typical closure duration for articulating a plosive were used as cues. In both the sentence and word reading, all sentences and words were labelled. The start and end boundaries are placed close to their acoustic onset and offset.

After that, the following procedure was used to generate f_0 contours that were then analysed: a Praat script read the TextGrid created in the previous step and randomly sampled intervals among the non-empty ones until their combined duration reached at least 30 seconds. The f_0 contours corresponding to each separate interval were excised from the complete Pitch object and concatenated to make a new contour. For each speaker, 10 such sample contours were generated for each speaking style. In total, 30 samples were generated per speaker.

2.3. Change point analysis

A custom Praat script created a table listing cumulative mean, median and base value for the whole duration of each sample for each f_0 contour obtained as a result of the procedure described in section 2.2. All measures were taken cumulatively from the first voiced frame up to the last in non-overlapping steps of 200 ms. The time series defined by the cumulative rate estimates were then submitted to a statistical technique called change point analysis, implemented as the *changepoint* [21] package for the R statistical computing environment [22]. It detected the time point where a significant change in the underlying variance of the time series took place (for technical details see [14]).

2.4. Statistical analysis

The experimental design consisted of four independent variables (IV) with differing number of levels:

- LANGUAGE (7 levels): British English, Estonian, French, German, Italian, Brazilian Portuguese and Swedish.
- Speaking STYLE (3 levels): spontaneous interview, sentence reading and word list reading.
- SPEAKER (10 levels): 5 female and 5 male speakers per language.
- Statistical ESTIMATOR (3 levels): arithmetical mean, median, base value.

The dependent variable (DV) was always the stabilisation point for a given f_0 contour, measured in seconds from the start of the sample.

We looked at possible significant effects caused by the IV on both the variability and average location of stabilisation points. Homogeneity of variance tests were used to test the null hypothesis that there are no differences in variance among groups within IV. Fligner-Killeen test was chosen because it is

[1] Threshold was set to ± 0.4 octaves per analysis frame. In Praat's autocorrelation-based algorithm, frame duration is variable and determined by the quotient 0.75/floor.

robust against departures from normality [23]. Failure in rejecting the null hypothesis is interpreted as evidence that the IV tested have no effect on the observed variability of the DV. To test for differences in average stabilisation point location we used the Kruskal–Wallis test [24] if the sample is heteroscedastic or Analysis of Variance if the sample is homoscedastic. When the IV being tested had more than two levels, paired *t*-tests or Mann–Whitney *U* tests with Holm-corrected *p*-values [25] were used to check for differences among levels. An α level of 5% was adopted for all tests.

In order to compare within-speaker and between-speaker variability of stabilisation time we used the F-ratio, as suggested by [26]. It expresses numerically the ratio of the variance of speaker means to the mean of speaker variance. Values lower than 1 indicate that within-speaker variance is greater than between-speaker variance. Computation of the F-ratio followed formulas found in [27, p. 101] and [26].

3. Results

3.1. Effect of estimator

The effect of ESTIMATOR on stabilisation time was investigated by collapsing all other IV variables, resulting in three groups. Each group thus created has 2100 data points (= 7 languages $\times$ 3 styles $\times$ 10 speakers $\times$ 10 samples). Table 1 shows mean stabilisation time.

Table 1: *Mean stabilisation time in seconds (SD in parenthesis) of the estimators.*

Estimator	Value
Mean	9.76 (6.58)
Median	9.67 (6.58)
Base value	8.01 (6.32)

We compared the variances of the three estimators and the test yielded a significant result $[\chi^2(2) = 7.6, p < 0.05]$. As can be seen in the table, base value is slightly less variable compared to the other estimators. A Kruskal-Wallis test indicates that one the estimators is different from the others $[\chi^2(2) = 123.8, p < 0.001]$. Multiple comparison tests indicate that base value stabilises faster than both mean and median by almost two seconds ($p < 0.001$ in the two comparisons) and that mean and median do not differ significantly.

A comparison of stabilisation times as a function of sex shows no significant differences.

3.2. Effect of language

The effect of LANGUAGE and ESTIMATOR on stabilisation time was investigated by collapsing the other IV variables. Each group has 300 data points (= 3 styles $\times$ 10 speakers $\times$ 10 samples). Table 2 shows mean stabilisation time of the estimators for the seven languages.

Applying the homogeneity of variance test separately for each estimator we found significant results in all cases – mean $[\chi^2(6) = 57.4, p < 0.001]$, median $[\chi^2(6) = 41.4, p < 0.001]$ and base value $[\chi^2(6) = 49.8, p < 0.001]$ –, implying that stabilisation time variance across languages is not uniform.

Kruskal-Wallis tests applied separately for each estimator yield significant values for each estimator: mean $[\chi^2(6) = 83.4, p < 0.001]$, median $[\chi^2(6) = 63.1, p < 0.001]$ and base value $[\chi^2(6) = 39, p < 0.001]$. Average stabilisation times across languages are not uniform. No obvious groupings emerged from the multiple comparisons other than Estonian consistently having the shortest stabilisation averages considering the three estimators.

Table 2: *Mean stabilisation time in seconds (SD in parenthesis) of the estimators as a function of language.*

Language	Mean	Median	Base value
English	9.08 (5.79)	8.80 (5.89)	6.80 (5.40)
Estonian	7.14 (4.90)	7.50 (5.38)	6.74 (5.00)
French	9.62 (5.54)	9.63 (5.67)	8.31 (5.42)
German	9.13 (5.44)	9.14 (5.69)	7.17 (5.51)
Italian	11.41 (7.83)	11.33 (7.57)	8.81 (6.98)
Portuguese	10.87 (6.46)	10.68 (6.51)	8.93 (6.66)
Swedish	11.03 (8.37)	10.57 (8.12)	9.34 (8.12)

The same effect of estimator described in section 3.1 – base line stabilising faster than either mean or median – holds true for the languages analysed separately (all $p < 0.05$), except for Estonian, for which there are no significant differences among estimators.

stabilisation times for male and female speakers show no significant differences as a function of sex if the languages are tested separately with the exception of mean and median for Italian (f < m) and median for Estonian (m < f).

3.3. Effect of speaking style

The effect of STYLE and ESTIMATOR on stabilisation time was investigated by collapsing the other IV variables. Each group has 700 data points (= 7 languages $\times$ 10 speakers $\times$ 10 samples). Table 3 shows mean stabilisation time of the estimators for the three speaking styles.

Table 3: *Mean stabilisation time in seconds (SD in parenthesis) of the estimators as a function of speaking style.*

Style	Mean	Median	Base value
Spontaneous	12.28 (7.88)	12.22 (7.63)	10.47 (7.31)
Sentences	8.63 (5.30)	8.68 (5.54)	6.79 (5.13)
Words	8.36 (5.51)	8.09 (5.57)	6.78 (5.56)

Homogeneity of variance tests made for each estimator indicate that styles are not homogeneous with respect to variance: mean $[\chi^2(2) = 41, p < 0.001]$, median $[\chi^2(2) = 41.4, p < 0.001]$ and base value $[\chi^2(2) = 49.7, p < 0.001]$. stabilisation time for spontaneous speech is more variable than for the other styles. Tests applied on a per language basis show that for Swedish and Italian the three estimators yield significantly different variances ($p < 0.001$). French and Portuguese show significantly different variances for the median (all $p < 0.05$) and German for the base value ($p < 0.05$). Thus, the global results shown in Table 3 are not the same in all languages.

As for average stabilisation time comparisons as shown in Table 3, tests run separately for each estimator indicate a significant effect of speaking style in all cases: mean $[\chi^2(2) = 121.41, p < 0.001]$ (spontaneous style has longer stabilisation intervals than the other two, all $p < 0.001$); median $[\chi^2(2) = 41, p < 0.001]$ (spontaneous > sentences > words, all $p < 0.05$); base value $[\chi^2(2) = 140, p < 0.001]$ (spontaneous style has longer stabilisation times than the other two,

$p < 0.001$). Tests run separately for each estimator and language show that in all cases there is at least one difference between styles for at least one of the estimators. The most common pattern found is the spontaneous style having slower stabilisation times than the others. This is true for the three estimators in English, Italian, Portuguese and Swedish. For French this is true for the median and base value and for Estonian and German this is true for the base value. For German and French the word reading style has the fastest stabilisation times for the mean and median.

Considering the speaking styles separately and comparing the variances of the three estimators (languages collapsed), the tests indicate that the sentence reading style is the only style where the variance of the estimators differ significantly: $[\chi^2(2) = 15, p < 0.001]$ (base value is the least variable estimator).

Tests comparing average values of the three estimators confirm that the advantage of the base value over the mean and median, reported in section 3.1, holds true for the speaking styles: spontaneous $[F(2, 2097) = 12.8, p < 0.001]$, sentences $[\chi^2(2) = 65.5, p < 0.001]$ and words $[F(2, 2097) = 16.2, p < 0.001]$. Pairwise comparisons for the three styles show the same pattern: base value has the lowest stabilisation times and there is no difference between mean and median (all $p < 0.001$).

Analysing languages separately, the common patterns observed were: in no language there is a difference between estimators in the spontaneous style, except Italian ($p < 0.05$). In all languages the base value has the shortest average stabilisation time either in the sentence reading or word reading or both styles, except French (no differences at all). In Estonian, French and Swedish no differences between estimators were found in the word reading style.

stabilisation times for male and female speakers show no significant differences as a function of sex in the preceding paragraphs analysis and do not contradict any of the results described there.

3.4. Effect of speaker

In order to evaluate if SPEAKER has a sizeable effect on stabilisation time, we determined the F-ratio in two different ways: as a function of estimator (see Table 4) and as a function of speaking style (see Table 5). In both cases, values were generated for each language separately.

Table 4: *F-ratio of the estimators as a function of language.*

Language	Mean	Median	Base value
English	0.068	0.015	0.086
Estonian	0.077	0.074	0.100
French	0.051	0.052	0.043
German	0.067	0.033	0.023
Italian	0.051	0.059	0.080
Portuguese	0.062	0.047	0.057
Swedish	0.030	0.055	0.049

Collapsing all languages, average F-ratios for the mean, median and base value are 0.058, 0.048 and 0.063. As for speaking style, average values for spontaneous, sentence reading and word reading are 0.099, 0.1 and 0.099.

All values are well below 1, indicating that within-speaker variability is higher than between-speaker variability. Putting it another way, average stabilisation time is not a speaker-specific feature. In the forensic context, it means that estimates of stabilisation time shown here can in principle be applied to other samples of speakers as well.

Table 5: *F-ratio of the estimators as a function of speaking style.*

Language	Spontaneous	Sentences	Words
English	0.064	0.160	0.079
Estonian	0.104	0.076	0.169
French	0.083	0.132	0.066
German	0.063	0.075	0.143
Italian	0.137	0.096	0.075
Portuguese	0.144	0.091	0.098
Swedish	0.101	0.069	0.062

4. Discussions and conclusions

As it was described in the Introduction, the base value has been shown to be the most robust f_0 estimator under a number of different conditions: prosodic variation, transmission channel, vocal effort [5] and language [13]. The prediction that it should also be a better representation of the fundamental frequency level of a given speaker was borne out in an automatic speaker comparison test [6]. The present study adds further evidence for the robustness of the base value as an f_0 estimator by showing that it consistently converges earlier than the other estimators.

Stabilisation times as a function of language suggest that both their variance and average values are not uniform statistically speaking, although the differences that arise are too small to be considered language-specific.

Speaking style has a significant effect on stabilisation times. The spontaneous style delays stabilisation by about 4 seconds. This may be explained by the significantly greater variation in spontaneous speech. Standard deviation in semitones for the full data set are 2.6 st for the spontaneous versus 2.2 st for the other styles as measured in a study using the same corpus [7]. If the languages are analysed separately, the same pattern appears, SD is systematically greater in spontaneous speech. stabilisation time is shortest for the base value regardless of style.

Our analyses show that stabilisation time is not speaker-specific to a significant degree both for effects of estimator and speaking style.

In the analyses presented in section 3, it was shown that speaker sex very seldom made a significant difference and in the few cases it did, there was no systematic direction of the differences and the stabilisation time differences were quite small. We may therefore conclude that genuine sex differences, if they indeed exist, have negligible effects on stabilisation time. In particular, we may note that no analyses examining the base value showed any sex differences. This suggests, that when using the base value for forensic purposes the same settings can be used regardless of the sex of the speaker.

5. Acknowledgements

This work has been supported by a grant (IB2015-6488) from *The Swedish Foundation for International Cooperation in Research and Higher Education* (STINT). The third author was supported by a PIBIC grant by *Conselho Nacional de Desenvolvimento Científico e Tecnológico* (CNPq) between August 2014 and July 2015.

6. References

[1] H. Traunmüller, "Conventional, biological and environmental factors in speech communication: A modulation theory," *Phonetica*, vol. 51, pp. 170–183, 1994.

[2] H. Traunmüller and A. Eriksson, "F0-excursions in speech and their perceptual evaluation as evidenced in liveliness estimations," *PERILUS*, vol. XVII, pp. 1–34, 1993.

[3] ——, "The frequency range of the voice fundamental in the speech of male and female adults," (manuscript), 1994. [Online]. Available: http://www2.ling.su.se/staff/hartmut/f0_m&f.pdf

[4] ——, "The perceptual evaluation of f0-excursions in speech as evidenced in liveliness estimations," *Journal of the Acoustical Society of America*, vol. 97, no. 3, pp. 1905–1915, 1995.

[5] J. Lindh and A. Eriksson, "Robustness of long time measures of fundamental frequency," in *Proceedings of Interspeech 2007*, 2007, pp. 2025–2028.

[6] R. R. da Silva, J. P. C. L. da Costa, R. K. Miranda, and G. Del Galdo, "Applying base value of fundamental frequency via the multivariate kernel-density in forensic speaker comparison," in *The 10th International Conference on Signal Processing and Communication Systems, ICSPCS'2016*, 2016.

[7] P. Arantes and M. É. N. Linhares, "Efeito da língua, estilo de elocução e sexo do falante sobre medidas globais da frequência fundamental," manuscript accepted for publication in *Letras de Hoje*.

[8] Y. Horii, "Some statistical characteristics of voice fundamental frequency," *Journal of Speech and Hearing Research*, vol. 18, no. 1, pp. 192–201, 1975.

[9] J. Baldwin and P. French, *Forensic Phonetics*. London: Pinter, 1990.

[10] P. Rose, "How effective are long term mean and standard deviation as normalisation parameters for tonal fundamental frequency?" *Speech Communication*, vol. 10, no. 3, pp. 229–247, 1991.

[11] A. Braun, *Fundamental frequency – how speaker specific is it?*, ser. Beiträge zur Phonetik und Linguistik. Trier: WVT Wissenschaftlicher Verlag, 1995, pp. 9–23.

[12] A. Eriksson, "Aural/acoustic vs. automatic methods in forensic phonetic case work," in *Forensic Speaker Recognition: Law Enforcement and Counter-terrorism*, A. Neustein and H. A. Patil, Eds. Springer, 2012, pp. 41–70.

[13] P. Arantes and A. Eriksson, "Temporal stability of long-term measures of fundamental frequency," in *Speech Prosody 2014*, N. Campbell, D. Gibbon, and D. Hirst, Eds., 2014, Conference Proceedings, pp. 1149–1152.

[14] R. Killick and I. A. Eckley, "changepoint: An R package for changepoint analysis," *Journal of Statistical Software*, vol. 58, no. 3, pp. 1–19, 2014. [Online]. Available: http://www.jstatsoft.org/v58/i03/

[15] P. A. Barbosa, A. Eriksson, and J. Åkesson, "On the robustness of some acoustic parameters for signaling word stress across styles in brazilian portuguese," in *Proceedings of Interspeech 2013*, 2013, pp. 282–286.

[16] P. Lippus, E. L. Asu, and M.-L. Kalvik, "An acoustic study of estonian word stress," in *Proceedings of Speech Prosody 2014*, 2014, pp. 232–235.

[17] A. Eriksson and M. Heldner, "The acoustics of word stress in english as a function of stress level and speaking style," in *Proceedings of Interspeech 2015*, 2015, pp. 41–45.

[18] A. Eriksson, P. M. Bertinetto, M. Heldner, R. Nodari, and G. Lenoci, "The acoustics of lexical stress in italian as a function of stress level and speaking style," *Proceedings of Interspeech 2016*, pp. 1059–1063, 2016.

[19] D. J. Hirst, "The analysis by synthesis of speech melody: from data to models," *Journal of Speech Sciences*, vol. 1, no. 1, pp. 55–83, 2011.

[20] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," *Proceedings of the Institute of Phonetic Sciences*, vol. 17, pp. 97–110, 1993.

[21] R. Killick, K. Haynes, and I. A. Eckley, *changepoint: An R package for changepoint analysis*, 2016, r package version 2.2.1. [Online]. Available: http://CRAN.R-project.org/package=changepoint

[22] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2016. [Online]. Available: https://www.R-project.org/

[23] W. J. Conover, M. E. Johnson, and M. M. Johnson, "A comparative study of tests for homogeneity of variances, with applications to the outer continental shelf bidding data," *Technometrics*, vol. 23, pp. 351–361, 1981.

[24] W. H. Kruskal and W. A. Wallis, "Use of ranks in one-criterion variance analysis," *Journal of the American Statistical Association*, vol. 47, no. 260, pp. 583–621, 1952.

[25] S. Holm, "A simple sequentially rejective multiple test procedure," *Scandinavian Journal of Statistics*, vol. 6, pp. 65–70, 1979.

[26] F. Nolan, "Intonation in speaker identification: an experiment on pitch alignment features," *Forensic Linguistics*, vol. 9, no. 1, pp. 3–21, 2002.

[27] ——, *The Phonetic Bases of Speaker Recognition*. Cambridge, UK: Cambridge University Press, 1993.