INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Quantifying fundamental frequency modulation as a function of language, speaking style and speaker

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

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

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

Abstract

In this study, we outline a methodology to quantify the degree of similarity between pairs of f_0 distributions based on the Anderson-Darling measure that underlies its namesake goodness-of-fit test. The procedure emphasizes differences due to more fine-grained f_0 modulations rather than differences in measures of central tendency, such as the mean and median. In order to assess the procedure's usefulness for speaker comparison, we applied it to a multilingual corpus in which participants contributed speech delivered in three speaking styles. The similarity measure was calculated separately as function of speaking style and speaker. Between-speaker variability (different speakers, same style) in distribution similarity varied significantly between styles — spontaneous interview shows greater variability than read sentences and word list in five languages (English, French, Italian, Portuguese and Swedish); in Estonian and German, read sentences yield more variability. Within-speaker variability (same speaker, different styles) levels are lower than between-speaker in the style that exhibit the greatest variability. The results point to the potential use of the proposed methodology as a way to identify possible idiosyncratic traits in f_0 distributions. Also, they further demonstrate the effect of speaking styles on intonation patterns.

Index Terms: fundamental frequency, speaking style, cross-language comparison

1. Introduction

Research on speaking styles stems from a general scientific interest in discovering the effects of non-structural linguistic factors on the process of speech production. There is a growing body of research on the subject, a representative sample of which is presented in review papers (see [1, 2] and references therein). Knowledge about this subject, nonetheless, can be brought to bear on more applied endeavors, such as in the field of forensic phonetics research and practice. In this context, questioned and reference samples can be produced in different styles. A questioned sample can be an informal (often bugged) phone conversation and the reference sample can be an interview conducted at a police station on a more formal and tense setting or a recorded public speech taken from a YouTube video[1]. This mismatch can be a problem for speaker comparison protocols, since the same speaker can vary some of his/her speech patterns quite substantially across different styles (as they would, also, when speaking in different emotional states), raising the likelihood of false negatives.

In the field of prosody, interest in speaking styles led to research, among other things, on the effects of speaking style on the production of a number of acoustic parameters usually taken as correlates of word stress and how they relate to stress perception, such as the work reported in [3, 4, 5, 6]. Here we take the same speech material used in these studies and look into the effects of speaking style on voice quality, a dimension still unexplored in the context of the corpus analyzed in the cited papers.

In a previous work, we analyzed this same corpus in search for effects of language, speaking style and speaker sex on global measures of fundamental frequency [2]. There we focused on measures of central tendency and variability of f_0 and a number of statistically significant effects were uncovered. When analyzing the data, we found a number of cases of speakers that made extensive use of non modal phonation, specially creaky voice (in the most extreme cases, up to 30% of all voiced analysis frames in the audio could be associated with creakiness). In such cases, an interaction between creaky voice and f_0 emerged in the form of strongly non-unimodal f_0 distributions, given that one of the correlates of creakiness is a sudden drop in f_0 (see [7] and others). This result is in line with reports on the literature about the fact that it is usual for f_0 distributions to be non-normal in the statistical sense, although the most common cause pointed out for this is skewness, not lack of unimodality. These findings suggest that it is worth exploring the effect of factors such as speaking style on f_0 beyond measures of central tendency and dispersion. Here we follow up on these findings and present a procedure designed to compare the overall shape of f_0 distributions and the effect of speaking style on these differences with the intent of assessing the effect of speaking style on the fine-grained differences in f_0 histograms and their indexical properties, results that may have implications for forensic phonetics.

2. Material and methods

2.1. Speakers and speech material

The speech material is 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 minimize variation due to regional variation and age. 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

[1] Using publicly available speech recordings as reference samples in speaker comparison tasks is becoming a common scenario in the casework done at the Brazilian Federal Police Crime Lab according to personal communication with an expert working there.

http://dx.doi.org/10.21437/Interspeech.2019-2857

with respect to educational background and were native speakers of the languages. Recordings were all made at universities in the countries where 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 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 refer to [3, 4, 5, 6][2].

2.2. Acoustic analysis

Before the f_0 extraction phase, audio files were preprocessed. Stretches that contained the speech of the experimenter, overlap between speaker and experimenter and non-speech events were silenced. This was done to minimize f_0 extraction errors.

f_0 contours were extracted using a Praat script[3] that implements a heuristic suggested by Hirst [9], that optimizes floor and ceiling values passed to Praat's *To Pitch (ac)* autocorrelation-based extraction function [10] by means of a two-pass procedure. 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:

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

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 optionally be set to 2.5 in case the speaker makes use of an extended range. 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 detection, usually in fricatives or transient noise in plosive releases. Cases such as incorrect unvoicing of frames, that can occur during glottalized phonation, had to be found by the analyst by comparing the f_0 contour with both the oscillogram and spectrogram.

2.3. Measuring f_0 distribution similarity

In order to quantify overall differences between the shape of two f_0 histograms, we designed a procedure that consists of applying the k-sample test based of the Anderson-Darling measure (A^2) of agreement between distributions. This test, described by [11], and implemented in the *kSamples* R package [12], can be used to test if two data samples follow the same underlying unspecified distribution. The Anderson-Darling (A-D) test is a modification of the well-known Kolmogorov-Smirnov test (K-S) that is used for the same purpose. They differ, however, in that A-D gives more weight to the tails of the distributions, while K-S is more sensitive to deviations in the center of the distribution. For our purposes, we take the A^2 value as an index of similarity between a given pair of histograms, so that the larger the value, the more different the two distributions are considered to be. We are not interested in the p-value provided by the test, only in the A^2 statistic.

Prior to test application, raw values in the f_0 contours were converted to the OMe scale, proposed by [13], using the formula

$$f_{ome} = \log_2\left(\frac{f_{hz}}{f_{med}}\right),$$

where f_{hz} is a value in Hz, f_{med} is the median value of the f_0 samples comprising the sample of interest and f_{ome} is the corresponding value in the OMe scale.

The scale conversion centers all histograms to the speaker-specific median value, such that differences in location among distributions are rendered uninformative. The rationale for performing the transformation is that the usefulness of differences in measures such as the mean or median f_0 for speaker comparison has already been quite explored before (see [14, 15]). Here we are more focused on finding out if factors such as speaker and speaking style have an effect on other aspects that contribute to the shape of f_0 distributions, such as dispersion, range, skewness and kurtosis.

After the scale conversion, each f_0 contour was transformed into a smoothed histogram called a kernel density estimate of the probability density of the data. This transformation was done by the `bkde` function included in the *KernSmooth* package for the R statistical environment [16]. The `dpik` function in the same package was used to select the optimal bandwidth for the kernel density estimate.

Figure 1 shows the two most similar and the two most dissimilar smoothed histograms according to the A^2 statistic metric, considering the entire corpus, but only within-language pairs. The statistic has a value of 1.1 for the pair in the left and 4878 for the right pair, a difference spanning three orders of magnitude. Histograms in the most similar pair are almost identical. Visual inspection of the least similar pair suggests differences in modal value location, modal density, spread, range, skewness and distribution "peakedness", the green one being more pointed and the blue one being more flat-topped.

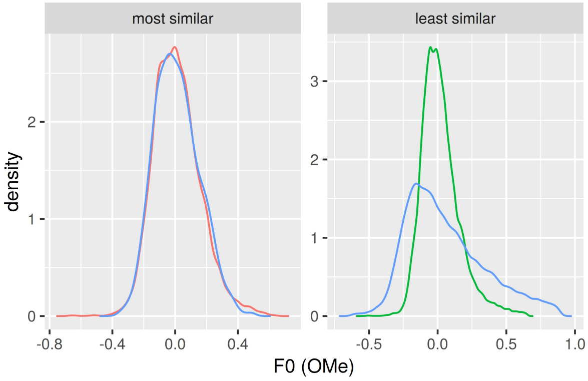

Figure 1: *Most similar and least similar pairs of histograms. On the left, the histograms represent one sentence reading and one word list reading contours, taken from the Swedish data. On the right, histograms are interview contours, from the Italian data.*

We generated plots of the five most similar and five most dissimilar histogram pairs of each language and a visual in-

[2] The texts in 2.1 and 2.2 are slightly modified versions of the corresponding sections in [8] where the same speech material was used, but for a completely different purpose.

[3] https://github.com/parantes/better-f0/

spection of them shows that the results are always very similar to the ones seen in figure 1. Carefully listening to the corresponding audio samples confirms that the most similar distributions present in general less f_0 modulation, as could be expected by the fact that the histograms such as the pair pictured in the left-hand part of figure 1 show almost no difference in the amount of spread around the central value. Listening to the pairs with the greatest dissimilarity degree confirmed that they differ markedly in terms of amount of f_0 modulation: in general, one of the samples has numerous cases of high amplitude f_0 excursions while the other has a much less varied contour. This situation is compatible with what is pictured in the right-hand side of figure 1, where the blue histogram is more asymmetrical and has a heavier positive tail (caused by cases of more extreme upward f_0 excursions) when contrasted with the green one, which is more symmetrical and less spread in comparison. The same conclusion can be drawn by looking at the contours in figure 2, where a representative stretch of the contours that generated the histograms on the right-hand side of 1 is shown. It can be easily seen that the blue contour has much wider excursions and spans a greater range than the green one.

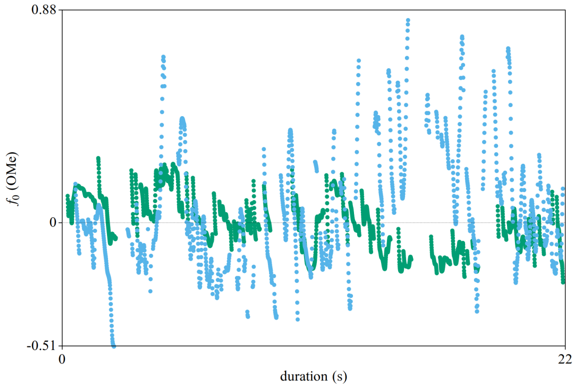

Figure 2: *Representative stretches of the f_0 contours of the least similar pairs of histograms. Colors are the same as in figure 1.*

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.

The dependent variable (DV) was always the value of the A^2 statistic obtained when comparing a pair of f_0 distributions. All pairwise combinations of style (three levels) and speaker (ten speakers) were compared, yielding 435 (= ${}_{30}C_2$) A^2 values per language. Comparisons were always within-language.

We looked at possible significant effects caused by STYLE and SPEAKER on the mean value of A^2. To test for significant effects on mean values, we used the Kruskal–Wallis test [17] when samples were heteroscedastic or Analysis of Variance if they were homoscedastic. The Fligner-Killeen test was used to test for homoscedasticity [18]. When the IV being tested had more than two levels, paired t-tests or Mann–Whitney U tests with Holm-corrected p-values [19] were used to check for differences among levels. An α level of 5% was adopted for all tests.

3. Results

3.1. Effect of style

In this analysis, the mean A^2 value for each style is the result of averaging the 45 values generated by the pairwise comparison of the ten individual f_0 histograms (one per speaker). Languages are analyzed separately. We take the results as an indication of the level of between-speaker variability in f_0 distribution similarity. A significant effect can be taken as evidence that f_0 distribution similarity is not uniform among styles. Table 1 shows the mean A^2 as a function of style and language.

Table 1: *Mean A^2 value and ±95% confidence interval of between-speaker comparisons as a function of style and language.*

Language	Interview	Sentences	Words
English	1047 (261)	118 (33)	103 (22)
Estonian	58.8 (19.3)	323 (90.4)	124 (34.4)
French	714 (222)	280 (71.1)	90.2 (19.7)
German	60.0 (19.6)	251 (69.0)	122 (25.0)
Italian	1334 (357)	465 (130)	284 (78.1)
Portuguese	784 (292)	405 (113)	144 (37.3)
Swedish	943 (269)	138 (37.4)	98.6 (25.7)

The samples of the seven languages are heteroscedastic (all Fligner-Killeen turned $p < 0.05$). The Kruskal–Wallis test pointed a significant effect of speaking style in each language (all $p < 0.05$). Multiple comparison tests indicate that languages can be grouped in two main groups. The largest one comprises English, French, Italian, Portuguese and Swedish. For this group, the spontaneous interview is the style with the highest A^2 values. There is no statistically-significant difference between sentences and words in English, Italian and Swedish. In Portuguese, there is no significant difference between interview and sentences, and both styles have mean values greater than the word reading style. In French, all differences are significant in the following order: interview > sentences > words. The second group has Estonian and German. For both, all differences are significant, but the order is: sentences > words > interview.

Results indicate that each language has some significant variation in f_0 distributions across styles. For a group of five languages (English, French, Italian, Portuguese and Swedish), spontaneous interview is the style that yields the highest levels of inter-speaker variation in distribution dissimilarity. For Estonian and German, on the other hand, sentence reading is the style with the greatest mean level of distribution dissimilarity. The effect size observed when the interview style has the highest mean A^2value is greater than when the sentence style has the highest mean value.

We listened to the audio files of the sentence reading pairs with the five largest A^2 values in Estonian and German and confirmed that it was always the case that one had a more level f_0 contour and the other showed excursions of greater amplitude. In both languages, spontaneous interview is the style with the lowest degree of f_0 contour dissimilarity between speakers. Contours in spontaneous speech are more symmetrical and show less modulation (narrower f_0 excursions and range).

This result points to two consistent effects: the first one, that some speaking styles tend to yield more between-speaker variability in terms of f_0 contour modulation; the second one, the effect of language on how different speech styles are implemented. Speakers of one group of languages tend to implement livelier (more modulated) f_0 contours when speaking spontaneously and speakers of the other group when reading sentences. We speculate that this may be related to different cross-cultural practices regarding spontaneous and read speech.

3.2. Effect of speaker

In this analysis, the mean A^2 value for each speaker is the result of averaging three data points, corresponding to the inter-style comparisons (interview–sentences, interview–words and sentences–words). Besides being an inter-style comparison, we also take this analysis as an indication of the level of within-speaker variability in f_0 distribution similarity since the data is presented as a function of each of the ten speakers. A significant effect can be taken as evidence that variation in f_0 distribution similarity caused by the three styles is not similar among speakers. Table 2 shows the overall (all ten speakers collapsed) mean A^2 as a function of language.

Table 2: *Overall mean speaker A^2 value and ±95% confidence interval of within-speaker/between-style comparisons as a function of language.*

Language	Mean
English	175 (62.1)
Estonian	74.2 (24.0)
French	141 (60.2)
German	99.3 (26.6)
Italian	433 (187)
Portuguese	207 (78.6)
Swedish	129 (49.5)

The samples of the seven languages are all homoscedastic (Fligner-Killeen tests turned non-significant). The ANOVA tests run separately for each language turned all non-significant as well, indicating that speakers' means are not statistically different. This set of results seems to indicate that the effect of speaking styles on the variation of f_0 distribution shape is uniform among speakers and this result holds for the seven languages studied here.

As we said earlier in this section, we consider this analysis to be a way to estimate within-speaker variability, even though the fact that the three samples each speaker contributes vary in terms of speaking style can be seen as a confounding factor. We interpret the results reported on the previous paragraph as an indication that the possible confounding influence of speaking style is not great in this case.

A second analysis used one-sample t-tests to compare, for each language separately, the overall A^2 mean for the speakers with the mean A^2 values observed in the analysis report on the previous section. This is a way to determine if the A^2 measure varies more due to within-speaker or between-speaker factors. To illustrate, we can take from table 2 the mean value of 175 corresponding to the overall A^2 mean for English speakers and compare it with the value 1047 taken from table 1, the mean A^2 value reflecting the between-speaker variability observed in the interview style in English. The one-sample t-test comparing both values yields a significant result $[t(29) = -27.508, p < 0.001]$. That puts mean within-speaker variability at a lower level when compared to between-speaker variability in the interview style. We get statistically significant results for the other four languages (French, Italian, Portuguese and Swedish) for which the interview style has the highest value in table 1. For the other two languages, Estonian and German, we get statistically significant results when comparing the values in table 2 with the mean values in table 1 for the sentence reading style.

This result is encouraging in terms of the usefulness of f_0 distribution shape in speaker comparison tasks because, if confirmed in further studies, it indicates that this feature has the hallmark of a good parameter in forensic speaker comparison: within-speaker variability is less than between-speaker variability.

4. Discussion and conclusion

In this study we report a method to compare pairs of f_0 distributions and quantify their difference in a way that emphasizes differences in the overall shape of the distributions' histograms. Using this metric, we have shown that speaking style has a significant effect on the shaping of f_0 distributions. Interview or sentence reading are the styles in which speakers differ the most in terms of distribution shape depending on the langugage. We have also shown that, according to our metric, f_0 contour by the same speaker vary less when speaking in different styles than the countours of different speakers that are speaking in the same style (specially the spontaneous style). This result is encouraging in terms of the usefulness of f_0 distribution shape in speaker comparison tasks because, if confirmed in further studies, it indicates that this feature has the hallmark of a good parameter in forensic speaker comparison: within-speaker variability is less than between-speaker variability.

In a speaker comparison scenario, forensic experts should concentrate on analyzing spontaneous styles, such as interviews, because it is more likely to find differences among speakers' f_0 distributions in those styles. We are measuring distribution similarity on the basis of a single number (A^2 statistic). Future work should tease out which statistical descriptors (range, asymmetry, kurtosis etc) correlate better with our metric. Also, it would be important in the future to measure within-speaker variability by means of different samples of the same speaker in the same style. Here we only had one sample in each style per speaker.

5. Acknowledgements

This work has been supported by a joint grant from The Swedish Foundation for International Cooperation in Research and Higher Education (STINT) and Brazil's Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) grant 88881.155645/2017-01.

6. References

[1] J. Llisterri, “Speaking styles in speech research,” in *ELSNET/ESCA/SALT Workshop on Integrating Speech and Natural Language*, Dublin, Ireland, 1992.

[2] P. Arantes and M. E. N. Linhares, “Efeito da língua, estilo de elocução e sexo do falante sobre medidas globais da frequência fundamental,” *Letras de Hoje*, vol. 52, no. 1, pp. 26–39, 2017.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] J. Laver, *The phonetic description of voice quality*. New York: Cambridge University Press, 1980.

[8] P. Arantes, A. Eriksson, and S. Gutzeit, “Effect of language, speaking style and speaker on long-term F0 estimation,” in *Interspeech 2017*. Stockholm: ISCA, 2017, pp. 3897–3901.

[9] 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.

[10] 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.

[11] F. W. Scholz and M. A. Stephens, “K-Sample Anderson-Darling Tests,” *Journal of the American Statistical Association*, vol. 82, no. 399, pp. 918–924, sep 1987.

[12] F. Scholz and A. Zhu, *kSamples: K-Sample Rank Tests and their Combinations*, 2017, r package version 1.2-7. [Online]. Available: https://CRAN.R-project.org/package=kSamples

[13] C. De Looze and D. J. Hirst, “The ome (octave-median) scale: A natural scale for speech melody,” in *Proceedings of the 7th International Conference on Speech Prosody*, N. Campbell, D. Gibbon, and D. Hirst, Eds., Dublin, 2014, pp. 910–914.

[14] Y. Kinoshita, S. Ishihara, and P. Rose, “Exploring the discriminatory potential of F0 distribution parameters in traditional forensic speaker recognition,” *The International Journal of Speech, Language and the Law*, vol. 16, no. 1, pp. 91–111, 2009.

[15] Y. Kinoshita and I. Shunichi, “F0 can tell us more: speaker verification using the long term distribution,” in *Proceedings of the Australasian International Conference on Speech Science and Technology 2010*, Melbourne, Australia, 2010, pp. 50–53.

[16] M. Wand, *KernSmooth: Functions for Kernel Smoothing Supporting Wand & Jones (1995)*, 2015, r package version 2.23-15. [Online]. Available: https://CRAN.R-project.org/package=KernSmooth

[17] 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.

[18] 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.

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