FUNDAMENTAL FREQUENCY NORMALIZATION AND STATISTICAL POWER: AN ASSESSMENT OF 15 NORMALIZING TECHNIQUES

Jérémy Genette[1], Jo Verhoeven[1,2], Steven Gillis[1]

[1] Antwerp University, [2] City, University London
jeremy.genette@uantwerpen.be; jo.verhoeven@city.ac.uk; steven.gillis@uantwerpen.be

ABSTRACT

The effect of 15 normalization procedures on the power of statistical tests is assessed. The difference in the F0 in Hz between the high and low vowels (N=1471), i.e. the intrinsic vowel pitch, produced by 47 native speakers of Dutch was assessed by means of a t-test. The same test was applied to the F0 values normalized by means of 15 conventional procedures. The power of the tests was registered to assess the effect of the normalization procedures. The results show that statistical power is influenced by applying normalization, but the difference in power of the procedures is levelled out as a function of the number of observations. The effects of the normalization procedures are interpreted in terms of how the variability of the data is accounted for.

Keywords: normalization, methodology, statistics, fundamental frequency, intrinsic vowel pitch.

1. INTRODUCTION

In addition to linguistically relevant information, speech also conveys information about the speaker, such as their sociolinguistic and physiological characteristics [1]. The latter can be studied for their own sake, such as the effect of physiological differences on acoustic differences [2]. But in other areas of research, those characteristics are considered as noise. For instance, in studying F0, individual anatomical differences – whether or not related to gender differences – are treated as noise which researchers typically try to eliminate by applying normalization.

Similarly to the normalization of vowel formants, the normalization of F0 consists of (non-) linearly rescaling raw frequency values in order to level out anatomical differences between speakers. Research on normalization techniques have studied how well they preserve (sociolinguistic) variation while eliminating idiosyncratic physiological noise [3], the degree of overlap between normalized vowel spaces [4], and their ability to reduce the within-vowel category variability while enhancing the between-vowel category variability [5].

However, the *a posteriori* rescaling of the original frequency measurements in Hz can lead to a rescaling of the variability in the data. As such, normalization potentially affects hypothesis testing and the power of statistical tests, i.e. the probability of correctly rejecting the null hypothesis. Research in genetics [6] and neurology [7] has revealed the impact of normalization procedures on statistical power. However, to the best of our knowledge, F0 normalization has not yet been investigated in this respect.

The main objective of this paper is to study the effect of F0 normalization on statistical power. It addresses the question whether normalization affects the statistical power of statistical tests. For this purpose, 15 conventional normalization procedures were applied to the same dataset, which consisted of F0 measurements of 1471 vowels produced by 47 speakers. This corpus was selected because it contains all monophthong vowels of Dutch produced by many speakers in carefully controlled consonantal contexts. The difference in F0 between the high and low vowels, i.e. the so-called "intrinsic vowel pitch", was then tested on both the unnormalized and normalized data. The power of each statistical test was computed and compared.

2. EXPERIMENT

2.1. Participants

Audio recordings were made of 90 Belgian Dutch speaking children imitating Dutch (non-)words. The mean chronological age of the children was 6 years (min. 5, max. 7). They all attended their first year of primary school in their region of birth.

2.2. Speech materials

Each child imitated 36 monosyllabic CVC stimuli. The vowel nucleus of each stimulus contained one of the 12 monophthongs of Belgian Standard Dutch, i.e. /i, ʏ, ɪ, ɛ, ɑ, ɔ, u, yː, eː, øː, aː, oː/ [8]. Each vowel occurred in three consonantal contexts: (i) /p_t/, (ii) /l_t/ and (iii) /t_r/, thus yielding phonotactically well-formed (non-)words.

The stimuli were read by a trained phonetician and the recordings of these were presented aurally to the participants who were asked to imitate her speech samples. Children's speech was recorded in a quiet room on a TASCAM DAT recorder by means of a head-mounted MicroMic II. The audio files were converted to WAV files by means of a TASCAM US 428 Digital Control Surface. The recording sessions with the children yielded a total of 7,985 speech samples.

Children’s imitations were perceptually assessed by 6 expert listeners who identified all speech samples in which the vowels were correct renditions of the target vowels. From these, a subset of samples was selected according to the following criteria. The speech samples containing the corner vowels /i, u, a/ were selected from children who had produced each corner vowel at least twice and who had produced each of the 12 vowels at least once. The final data selection consisted of a total of 1,471 vowels produced by 47 children. For each of those 47 speakers, one randomly selected vowel per vowel category (N=564) was retained because of the requirements of some of the normalization techniques.

2.3. Acoustic analysis

The F0, F1, F2 and F3 of all the vowels were measured using a Python script through the Parselmouth API [9] of PRAAT [10]. F0 was determined using the standard autocorrelation algorithm. The maximum number of candidates was set to 15, the silence threshold to 0.03, the voicing threshold to 0.45, the octave cost to 0.01, the octave-jump cost to 0.35 and the voiced/unvoiced cost to 0.14. To minimise the number of potential octave jumps, the pitch floor and ceiling were set at 175 Hz and 425 Hz respectively after visual inspection of the pitch data with PRAAT’s standard parameters. PRAAT’s “Kill octave jumps” function was also applied. In addition, the F1, F2 and F3 of the vowels were measured. The number of formants was set to 5 and the formant maximum was set to 5500 Hz. Formant measurements were needed since one normalization technique (as described in [11]) requires them as variables.

2.4. Normalization procedures

The F0 normalization procedures used in the present study are listed in Table 1. The procedures were taken from published research on acoustic vowel normalization [3, 4, 5, 12] and in research dealing specifically with intrinsic vowel pitch [13, 14, 15].

Vowel normalization procedures require information about different vowels (i.e., vowel-extrinsic normalization) or only information about the vowel to be normalized (i.e., vowel-intrinsic normalization) [3]. Four types of F0 normalization techniques can be distinguished:

- rescaling methods transform the physical Hz scale to a perceptually relevant scale. Although they do not perform proper normalization [16], some scholars (e.g. [3, 15]) consider them as potential auditory normalization techniques, hence they were included in the present study;
- range normalization procedures normalize each vowel with respect to all the vowels of individual speakers [12];
- centroid normalization procedures normalize values relatively to a central tendency of the total distribution per participant [12];
- log-mean normalization techniques subtract a reference value from the log-transformed Hz value [12].

UNNORMALIZED FREQUENCY	
HZ	base condition
SCALE CONVERSION – VOWEL-INTRINSIC	
BARK	Bark-conversion of the Hz scale [17]
MEL	Mel-conversion of the Hz scale [18]
ERB	ERB-conversion of the Hz scale [19]
LN	Natural logarithmic conversion of the Hz scale [20]
ST-1	Semitone-conversion (ref. 1Hz) of the Hz scale [15]
ST-50	Semitone-conversion (ref. 50Hz) of the Hz scale [21]
ST-100	Semitone-conversion (ref.100Hz) of the Hz scale [22]
RANGE NORMALIZATION – VOWEL-EXTRINSIC	
GERSTMAN	Range normalization [23]
H & H	Rescaling between 0 to 1 [14]
LCE	Linear compression/expansion method (cf. [5])
CENTROID NORMALIZATION – VOWEL-EXTRINSIC	
LOBANOV	Z-score transformation [5]
TO MEDIAN	Normalization to median [13]
TO IQR	Normalization to IQR [13]
LOG-MEAN NORMALIZATION – VOWEL-EXTRINSIC	
NEAREY1	Single log-mean procedure [11]
NEAREY2	Shared log-mean procedure [11]

Table 1: Selected normalization procedures divided by types according to [3] and [12].

2.5. Statistical analysis

The group difference between the high and low vowels was investigated by means of paired t-tests. The *pwr.t.test* function of the *pwr* package [24] in R [25] was used to compute the power of the t-test. Its effect size was expressed as Cohen’s *d* [26], i.e., the difference between the mean F0 of the high and low vowels divided by the pooled standard deviation for the two groups. The significance level was set to 0.05. The power was computed as a function of the number of vowels per sample. For this purpose, the sample size was increased stepwise from 1 to 100 vowels.

3. RESULTS

The power of the t-tests which aimed to detect a group difference between the F0 of high and low vowels is presented in Fig. 1 as a function of sample size and normalization technique. Table 2 provides the sample sizes required per sample (i.e., separately for both high and low vowels) to attain a statistical power equal to 0.8, which is often considered a good power level [28], and 0.99, which is close to the maximal power.

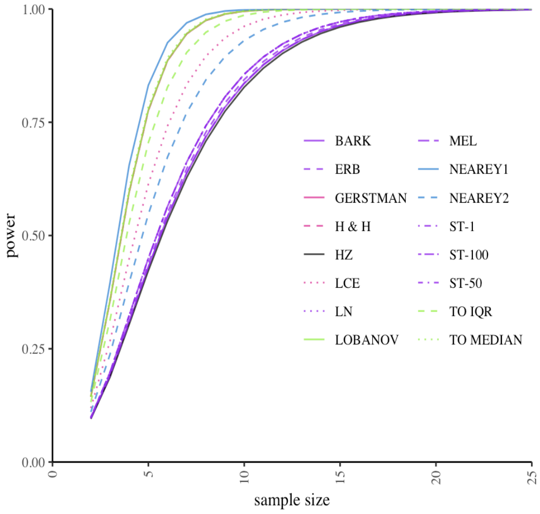

Figure 1: Results of the power analysis as a function of sample size per sample and normalization procedure.

Norm.	0.8	0.99	Norm.	0.8	0.99
HZ	9.5	19.3	GERSTMAN	5.2	9.1
BARK	9.3	19	H & H	5.2	9.1
ERB	9.1	18.5	LCE	6.6	12.5
LN	8.9	17.8	LOBANOV	5.2	9.1
MEL	9.3	18.9	TO MEDIAN	5.1	9
ST-100	8.9	18	TO IQR	5.7	10.5
ST-1	8.9	18	NEAREY1	4.8	8.2
ST-50	8.9	18	NEAREY2	7.4	14.3

Table 2: Sample size per sample needed to reach statistical power of 0.8 and 0.99 as a function of normalization procedure.

As can be seen in Fig. 1, the power curves of the different types of normalization clearly differ. The majority of the normalization techniques reach maximal power with a sample size close to 20. The unnormalized data in Hz (HZ) lead to one of the lowest power curves. Slightly higher power curves are achieved by scale conversion techniques (BARK, ERB, LN, MEL, ST-1, ST-100 and ST-50). It should be noted that the power curves of the three semitone conversions perfectly overlap. Range normalization procedures (GERSTMAN and H & H) result in higher power curves still. Relying on the same principle but differing in multiplication factor only, GERSTMAN and H & H exhibit the same power curves. The centroid normalization techniques (LOBANOV, TO MEDIAN and TO IQR) consistently lead to higher power curves: LOBANOV and TO MEDIAN yield the highest power curves of the three centroid normalization procedures. As far as log-mean normalizations (NEAREY1 and NEAREY2) are concerned, NEAREY1 represents one of the highest power curves. On the contrary, NEAREY2 exhibits a lower power curve.

Fig. 2 indicates that, relative to the unnormalized data, the difference in means/pooled *SD* ratio is very slightly modified by scale conversions. On the contrary, the log-mean, range and centroid normalization techniques clearly reduce the pooled *SD* relative to the difference in means.

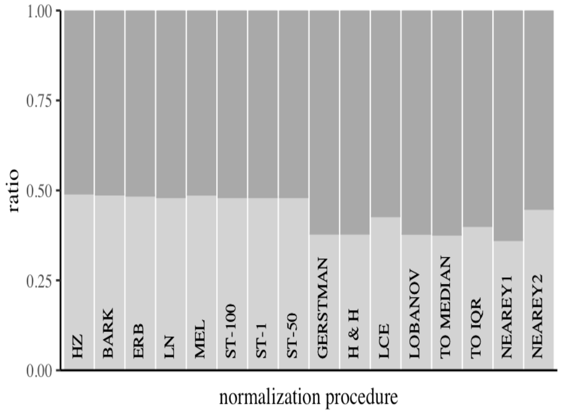

Figure 2: Ratio between the difference between the average F0 of high and low vowels (dark grey) and the pooled SD (light grey) as a function of normalization procedure.

4. DISCUSSION

This study aimed to investigate the effect of 15 normalization techniques on statistical power. 1,471 vowels from 47 speakers were analysed for F0. The inferential test that constituted the benchmark for this investigation was a paired t-test assessing the difference in F0 between high and low vowels.

The most important conclusion from this study is that data normalization influences the power of a statistical test. The results show that some normalization techniques achieve higher power with smaller data sets than others. The unnormalized and scale-converted normalized data tend to require larger sample sizes to achieve higher power than log-mean, centroid, and range normalized measurements.

The differences between the normalization techniques can be explained by how the data are rescaled. Specifically, if the variability of the data is rescaled to a certain extent and if the subsequent effect size is rescaled to the exact same extent, the power is kept unchanged. If both the difference in means and the pooled *SD* are rescaled proportionally, the ratio between them remains unchanged. Consequently, the power of a statistical test based on normalized data is not affected. However, it is clear that some normalization techniques can significantly affect the difference in means/pooled *SD* ratio and as a consequence, statistical power.

The differences between the scale conversion techniques as opposed to the log-mean, range or centroid normalization techniques can be explained in terms of whether the effect size is rescaled to the same extent as the variability of the data, see Fig. 2.

Scale conversions [15, 17, 18, 19, 20, 21, 22] non-linearly transform the data so that only perceptually relevant differences are reflected on the normalized scale. However, larger differences between high and low vowels due to a higher average F0 remain larger on the normalized scale. The variability of the data is therefore not affected much.

Range [5, 14, 23], centroid [5, 13] and log-mean [11] normalization methods use per-subject reference levels so that the variability between subjects is reduced. The centroid, range or log-mean reference thus help to reduce the per-speaker variability. As such, with respect to the reduced pooled *SD*, the difference in means is proportionally more important, hence the increased power. This is why a given power level can be achieved with fewer data than with unnormalized measurements, despite the data being the same.

This means that normalizing F0 values can reduce the possibility that the null hypothesis is erroneously rejected. To put it another way, the increased power indicates that, while keeping effect size constant, the *a posteriori* reduced variability in normalized data can lead to a more confident rejection of the null hypothesis than what untransformed data would suggest. This should therefore be a matter of careful consideration when analysing the results of inferential tests. However, as the number of observations per sample increases, the statistical power is less affected. It should also be noted that a lower significance level, a lower effect size or increased variability in the data set is expected to increase the effect of the normalization procedures.

5. CONCLUSIONS

In the case of a large enough sample size, the choice of an adequate normalization procedure is only a matter of choosing a normalization that best levels out physiological variation while preserving phonemic and sociolinguistic information. However, if smaller sample sizes are used, the effect of normalizing on power should be considered because log-mean, range and centroid normalizations reduce variability and consequently enhance the statistical power as compared to unnormalized or perceptually rescaled values. Even if similar results are expected, a natural progression of this work is to analyse the effect of formant normalization methods on statistical power.

6. ACKNOWLEDGMENTS

Thanks are due to the participants of this study. The research reported in this paper was supported by a research grant from the Research Foundation – Flanders (FWO) [grant G004321N].

7. REFERENCES

[1] Ladefoged, P., Broadbent, D. E. 1957. Information conveyed by vowels. *The Journal of the Acoustical Society of America* 29(1), 98–104.

[2] van der Harst, S. 2011. *The vowel space paradox: A sociophonetic study on Dutch*. LOT.

[3] Adank, P., Smits, R., Hout, R. 2004. A comparison of vowel normalization procedures for language variation research. *The Journal of the Acoustical Society of America* 116(5), 3099–3107.

[4] Flynn, N. 2011. Comparing vowel formant normalisation procedures. *York Papers in Linguistics Series* 2(11), 1–28.

[5] Lobanov, B. M. 1971. Classification of Russian vowels spoken by different speakers. *The Journal of the Acoustical Society of America* 49(2), 606–608.

[6] Qiu, X., Wu, H., Hu, R. 2013. The impact of quantile and rank normalization procedures on the testing power of gene differential expression analysis. *BMC Bioinformatics* 14, 124.

[7] Tustison, N. J., Avants, B. B., Cook, P. A., Kim, J., Whyte, J., Gee, J. C., Stone, J. R. 2012. Logical circularity in voxel–based analysis: Normalization strategy may induce statistical bias. *Human Brain Mapping* 35(3), 745–759.

[8] Verhoeven, J. 2005. Belgian standard Dutch. *Journal of the International Phonetic Association* 35(2), 243–247.

[9] Jadoul, Y., Thompson, B., Boer, B. 2018. Introducing parselmouth: A python interface to praat. *Journal of Phonetics* 71, 1–15.

[10] Boersma, P. 2001. Praat, a system for doing phonetics by computer. *Glot. International* 5(9), 341–345.

[11] Nearey, T. M. 1978. *Phonetic feature systems for vowels*. Indiana University Linguistics Club.

[12] Heeringa, W., Velde, H. 2018. The implementation of methods in Visible Vowels. *Vignette of the R package 'visvow'*.

[13] Chen, W.–R., Whalen, D. H., Tiede, M. K. 2021. A dual mechanism for intrinsic f0. *Journal of Phonetics* 87, 101063.

[14] Hoole, P., Honda, K. 2011. Automaticity vs. feature–enhancement in the control of segmental F0. In: Clements, G. N., Ridouane R. (eds.), *Where do phonological features come from*. John Benjamins Publishing Company, 131–171.

[15] Whalen, D. H., Levitt, A. G. 1995. The universality of intrinsic F0 of vowels. *Journal of Phonetics* 23(3), 349–366.

[16] Clopper, C. G. 2009. Computational methods for normalizing acoustic vowel data for talker differences. *Language and Linguistics Compass* 3(6), 1430–1442.

[17] Traunmüller, H. 1981. Perceptual dimension of openness in vowels. *The Journal of the Acoustical Society of America* 69(5), 1465–1475.

[18] Stevens, S. S., Volkmann, J. 1940. The relation of pitch to frequency: A revised scale. *The American Journal of Psychology* 53(3), 329–353.

[19] Glasberg, B. R., Moore, B. C. 1990. Derivation of auditory filter shapes from notched–noise data. *Hearing Research* 47, 103–138.

[20] Miller, J. D. 1989. Auditory-perceptual interpretation of the vowel. *The Journal of the Acoustical Society of America* 85(5), 2114–2134.

[21] Rietveld, A.C.M., Van Heuven, V. J. 2009. *Algemene fonetiek*. Coutinho.

[22] Fant, G., Kruckenberg, A., Gustafson, K., Liljencrants, J. 2002. A new approach to intonation analysis and synthesis of Swedish. *Speech prosody 2002*, 283–286.

[23] Gerstman, L. 1968. Classification of self–normalized vowels. *IEEE Transactions on Audio and Electroacoustics* 16(1), 78–80.

[24] Champely, S. 2020. *pwr: Basic functions for power analysis*. https://CRAN.R–project.org/package=pwr

[25] R Development Core Team 2012. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

[26] Cohen, J. 1988. *Statistical power analysis for the behavioral sciences*. Lawrence Erlbaum Associates.

[27] Winter, B. 2019. *Statistics for linguists: An introduction using R*. Routledge.