On the Normalization of Syllable Prominence Ratings

Christopher Sappok [1], *Denis Arnold* [2]

[1] German Studies, University of Duisburg-Essen, Germany
[2] Speech and Communication, University of Bonn, Germany
christopher.sappok@uni-due.de, dar@sk.uni-bonn.de

Abstract

The instructions under which raters quantify syllable prominence perception need to be simple in order to maintain immediate reactions. This leads to noise in the rating data that can be dealt with by normalization, e.g. setting central tendency = 0 and dispersion = 1 (as in Z-score normalization). Questions arise such as: Which parameter is adequate here to capture central tendency? Which reference distribution should the normalization be based on? In this paper 16 different normalization methods are evaluated. In a perception experiment using German read speech (prose and poetry), syllable prominence ratings were collected. From the rating data 16 complete "mirror" data-sets were computed according to the 16 methods. Each mirror data-set was correlated with the same set of measures from the underlying acoustic data, focusing on raw syllable duration which is seen as a rather straightforward acoustic aspect of syllable prominence. Correlation coefficients could be raised considerably by selected methods.

Index Terms: syllable prominence, syllable duration, perception experiment, normalization, read speech, German

1. Introduction

Perception experiments as introduced in [1] are an important foundation of research concerning syllable prominence. One problem in this connection is inter- and even intra-rater variability [2, 3]. Our starting point was an experimental setting based on [4]. Listeners had to rate syllable prominence in 8 speech signals which were 30 syllables long, being confronted with 30 vertical slide controls to adjust on a 0-30 scale (see below, Figure 2). Observations during the experimental sessions led to the following a posteriori hypotheses:

1. Listeners project an imaginary horizontal base-line onto the arrangement of slide controls and display *intersubjective* differences concerning the exact positions of these base-lines.
2. Base-line positions are also prone to *intrasubjective* shifting in the course of going through signal 1 to 8. (We even found hints that base-lines decline within the rating of one and the same signal.)
3. Resulting noise can be reduced by setting central tendency = 0 (by subtracting central tendency, e.g. mean, from each individual measure) with reference to the distribution of one listener's rating of one single signal (n = 30 phonetic syllables).
4. As to the specific parameter, the median is more suitable than the mean, because each base-line thus manifests itself in the form of straight zeroes.
5. Listeners differ in terms of rating-"generosity". Example: With a common base-line of 15, one "greedy" listener would rate a very prominent syllable 18, whereas another, "generous", listener would rate the same syllable, say, 28. But here, resulting noise cannot be met by setting dispersion = 1 with reference to the intra-rating distribution as above (see hypothesis 3). This would imply that the signals themselves do not differ in terms of broadness of prominence variation. Thus, the appropriate reference distribution for dispersion normalization (by dividing each individual measure by dispersion, e.g. mean deviation) is the total amount of ratings given by one specific listener (m = 8*30 = 240 phonetic syllables).

To test our hypotheses, we systematically varied a number of factors considered relevant (see below, Figure 3). For each combination of factor levels, a complete set of normalized rating data was computed (a mirror data-set). For each mirror data-set, Pearson product-moment correlation coefficients were computed with respect to the same set of underlying syllable duration measures. (Pitch and intensity measures can be considered similarly relevant, but measuring duration is more independent of specific concepts.)

2. The "Gold"-Corpus

The corpus was recorded in an experimental setting designed to elicit maximally different prominence distributions in repeated readings of the same wording. The wording was from a stanza of rhymeless, metrical poetry, presented once in the original poetry-layout and context and once stripped of verse-conform line-breaks and embedded in a prose context. Further conditions added up to a hierarchy of 4 factors (Figure 1).

SPEAKER	**LAYM**								**PROF**							
	1		**2**		**3**		**4**		**5**		**6**		**7**		**8**	
TEXT	**LYR**				**PROS**				**LYR**				**PROS**			
T-ORDER	**1PROS**		**1LYR**		**1PROS**		**1LYR**		**1PROS**		**1LYR**		**1LYR**		**1PROS**	
R-ORDER	**P**	**S**	**P**	**S**	**P**	**S**	**P**	**S**	**P**	**S**	**P**	**S**	**P**	**S**	**P**	**S**
recordings	*01*	*03*	*05*	*07*	*09*	*11*	*13*	*15*	*17*	*19*	*21*	*23*	*25*	*27*	*29*	*31*
	02	*04*	*06*	*08*	*10*	*12*	*14*	*16*	*18*	*20*	*22*	*24*	*26*	*28*	*30*	*32*

Figure 1: *The factors (grey) and respective levels (white) underlying the "Gold"-corpus*

8 male Speakers participated. 4 were university professors of rhethoric, 4 were laymen (students or similar background) with no qualification or experience in professional speaking or reading (factor: SPEAKER, levels: LAYM vs. PROF).

Speakers were to read two texts, each two times in a row, in one individual recording session. One text looked like a poem, the other looked like prose, both were 123 syllables/75 words long (factor: TEXT, levels: LYR vs. PROS).

Randomly chosen, two of the laymen and two of the professors read TEXT LYR first, the others read TEXT PROS first (factor: T-ORDER, levels: 1LYR vs. 1PROS).

The first reading of each text was to be done on first sight ("prima vista"), the second a few seconds after the first reading was finished ("secunda vista"). Then the other TEXT was read in the same way (factor: R-ORDER, levels: P vs. S).

8*4 = 32 mono-signals were recorded, using the portable DAT-recorder SONY TCD-D100 and the SONY ECM-T140 microphone, at a sampling rate of 48 kHz (later sampled down to 16 kHz to avoid processing delays at activation in the perception experiment). Then we extracted a certain passage from each signal, the wording of which was identical in TEXT LYR and TEXT PROS. The "Gold"-corpus consists of these 32 extracts ("stanza 3"; 30 syllables/22 words).

TEXT LYR is part of the long epic poem "Bimini" [5] by German poet Heinrich Heine (1797-1857). It consists of 4 rhymeless stanzas à 4 verses à 4 trochaic feet (some verses lacking the last weak syllable). As a reading stimulus, TEXT LYR was reproduced in the original layout with line breaks after each verse and a blank line between stanzas. (1) is a reproduction of stanza 3 the way it appeared in TEXT LYR:

Gold war jetzt das erste Wort, (1)
Das der Spanier sprach beim Eintritt
In des Indianers Hütte -
Erst nachher frug er nach Wasser.

It may be translated into English fairly well preserving word order and meter:

Gold was now the prim'ry word (2)
That the spaniard spoke on ent'ring
In the native indian's shelter -
Only then ask'd he for water.

TEXT PROS is in part a reformulation of TEXT LYR. In stanzas 1, 2 and 4, word order was changed preserving syntactic structure in order to solely spoil the balanced metric organization of the original. Line breaks, now including stanza 3, were deleted, leaving line-organization to purely length-of-string based word processing. Verse-initial capitalization was modified according to regular German spelling. The result is a prose version of TEXT LYR with respect to word order and metrical organization, except for the embedded wording of stanza 3, and with respect to graphical organization entirely.

The main reason why TEXT LYR was selected as a basis for the corpus is the meter underlying the first 5 syllables of the fourth verse of stanza 3. According to the authors' native-speaker intuition, under condition TEXT LYR these syllables would preferably be read as indicated by (3), whereas under condition TEXT PROS they would preferably be read as indicated by (4) (prominent syllables represented by capitals):

ERST nachHER frug ER nach WASser (3)

erst NACHher FRUG er nach WASser (4)

Mainly because of this feature we believed that the "Gold"-corpus would contain sufficiently different prominence distributions from signal to signal. The following paragraphs describe the derivation of acoustic measures from the corpus. Afterwards, section 3 describes the derivation of perceptual measures from the corpus.

The 32 extracts were labeled on a segmental level by the first author and independently by a second labeler, using PRAAT [6] and following the "liberal phonemization"-standard of SAMPA-D-VMlex V1.0 [7], additionally documenting boundary phenomena such as pauses, pre-pause lengthening and lengthening without adjacent pause. Informal comparison showed only minute differences between the two labelers. Further steps were based on the first author's labels.

From the underlying TextGrid-files absolute syllable duration measures were derived, measuring [s] from the beginning of one onset-initial segment to the next, 32*30 = 960 measures altogether. Then all measures of pre-pause syllables and syllables with lengthening without adjacent pause were deleted manually, because in these cases duration cannot be taken to satisfactorily reflect prominence. 161 measures were affected. The 32 signal-specific vectors were concatenated to 8 speaker-specific vectors in an R-environment [8], each 4*30 = 120 positions long (deleted measures appearing as "NA"). Working with speaker-specific distributions helped to reduce noise which would have appeared in one "global" distribution resulting from, e.g., differences in speaking rate. The 8 speaker-specific vectors were the basis for the evaluation eventually carried out (see section 5).

3. The Perception Experiment

The experiment involved 64 listeners, each rating a selection of 8 out of the 32 signals of the "Gold"-corpus. Beforehand, the corpus was split up into 4 packages à 8 signals, making sure that each package contained one signal by each of the 8 speakers. All other factors (Figure 1) were neglected through randomization. Each listener was assigned one package at random, except for making sure that each package was treated 64:4 = 16 times. Within individual listening sessions, the order of the 8 signals in question was randomized every time by the experimental software.

64 students, mostly undergraduate but all with a certain amount of phonetic background, agreed to participate in a perception experiment. Several sessions with smaller groups took place in which each listener was seated at a computer work station equipped with headphones and with the experimental software already running: The screen displayed a greeting formula and a "Next"-button. On the next screen, the instructions appeared. Additionally, they were read out aloud to the listeners and questions could be asked. In similar contexts we had found that the German word "Silbenprominenz" (syllable prominence) is not familiar to most students. Therefore we referred to the more common concept of "Betonung" (highlighting pronunciation). The core instruction was: "Geben Sie zu jeder Silbe an, wie stark der Sprecher diese Silbe betont." ("In connection with each syllable, state how strongly the speaker pronounces this syllable.")

On the next screen, a practice arrangement appeared in which the listeners were to become familiar with the technical aspects of the experimental software. Afterwards the actual experiment started. It consisted of 8 rounds of the following procedure: On pressing the "Next"-button, the next signal sounded automatically one time. While the signal could be replayed without limit, it always sounded as a whole. Even though – given that each syllable was to be rated – the signals were quite long (ranging from roughly 5 to 11s), we refrained from letting listeners freely select parts of the signal, because too many types of action would be demanded and sequencing might affect prominence perception uncontrollably.

The screen displayed 30 vertical slide controls scaled 0-30, a "Play"-button, and a "Next"-button (Figure 2). There were no preset sliders in order to avoid bias resulting from any kind of default setting. Upon clicking anywhere on each slide control, a slider would appear at the spot of activation. These sliders could be moved and moved again in any order. The "Next"-button only worked if all sliders had been activated.

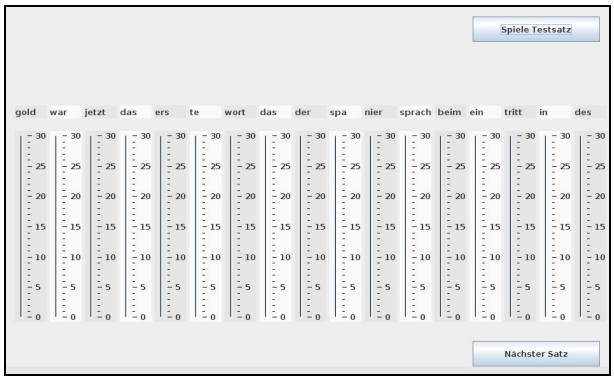

Figure 2: *First half of a screenshot of the arrangement used to rate one signal (sliders not yet activated, "Play"-button above, "Next"-button below)*

When the last of the group had finished the experiment, the session was over and everybody received a little bag of sweets for their cooperation. Unfortunately, it turned out that not everybody had been so cooperative after all (e.g., setting all 30 sliders to the same value). Fortunately, the number of replays per signal had been recorded by the experimental software. Based on the listener-specific sum of replays over all 8 signals, we decided to discard two listeners per signal-package (replay-sums ranging from 2 to 14). Due to different replay sum distributions from package to package, this solution was not all balanced (in comparison, discarding the 8 listeners with the smallest replay-sum regardless of packages would have led to the same outcome except for 1 case).

4. Normalization Methods and Data Preparation

From the raw rating data, 16 complete mirror data-sets were computed (all computation carried out in R [8]). 16 is the result of a hierarchy of 3 factors taken into account (Figure 3).

CT	MEAN								MEDIAN							
CTRD	NONE			IRAT			ILIST		NONE			IRAT			ILIST	
DISPRD	NONE	IRAT	ILIST	NONE	IRAT	ILIST	NONE	IRAT	ILIST	NONE	IRAT	ILIST	NONE	IRAT	ILIST	NONE	IRAT	ILIST
methods		*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*		*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*

Figure 3: *The factors (grey) and their levels (white) underlying the 16 normalization methods (crossed-out fields represent unnormalized data)*

The first factor concerned the question which parameter the normalization of central tendency was based on, mean or median (factor: CT, levels: MEAN vs. MEDIAN).

The second factor concerned central tendency reference distribution, i.e. the distribution within which CT was set to 0. We distinguished between: no reference distribution (absence of CT-normalization), the intra-rating distribution of the 30 measures delivered in the rating of one signal, and the intra-listener distribution of the 240 measures delivered by one listener (factor: CTRD, levels: NONE vs. IRAT vs. ILIST).

The third factor concerned dispersion reference distribution, i.e. the distribution within which mean deviation was set to 1. (The dispersion parameter employed was mean absolute deviation from CT with MEAN as well as MEDIAN, since standard deviation would not have worked well with MEDIAN.) This factor's levels were again: no reference distribution (absence of dispersion-normalization), intra-rating distribution, and intra-listener distribution (factor: DISPRD, levels: NONE vs. IRAT vs. ILIST).

Note that methods 4, 8, 12, and 16 (see above: Figure 3) represent varieties of "classical" Z-score normalization, except for the use of mean deviation instead of standard deviation.

As with the duration data, each resulting mirror data-set was split up into 8 speaker-specific subsets and concatenated to 8 vectors. As a consequence of the "packaging-strategy" in the perception experiment (see above, section 3), each of the four signals per speaker had been rated by a different set of listeners, so ratings by all listeners were present in one speaker-specific vector (4*30*14 = 1680 measures).

Since each phonetic syllable had been rated by 14 listeners, each rating vector was 14 times longer than the respective duration vector. To match numbers for the correlation of prominence measures with acoustic measures, every duration vector was "pumped up" by writing each of its measures 14 times in a row. (It must be kept in mind, though, that prominence normalization in general is carried out in order to receive just one measure per phonetic syllable.)

For each mirror data-set, 8 speaker-specific correlation coefficients (r-values) with respect to syllable duration were computed. Due to the fact that the duration vectors had been affected by boundary phenomena to different degrees (see section 2), the number of sample-pairs effectively included in the individual correlation analyses ranged from 1316 to 1512.

Evaluation was carried out after transforming all coefficients in the following way: Firstly, r-values were Fisher-z-transformed to bring them on an interval scale. Secondly, from each resulting z-value, we subtracted the z-value of the analysis of the respective unnormalized speaker specific set (see below, Figure 4). The resulting parameter, z-diff, represents the gain brought forth by the normalization method in question in terms of z-units. Finally, the 8 speaker-specific z-diff values were averaged to arrive at one single mean-measure per method. To apply more sophisticated statistics to compare these means (ANOVA) seemed not indicated, because the individual z-diff values were associated with different sample-sizes as well as different p-values from the underlying correlation analysis.

5. Results

The starting point for the evaluation was the correlation of the unnormalized rating data with the acoustic data. The average absolute z-value over all speakers is .11 (with r = .11 as well). Figure 4 gives an overview over the underlying distribution (all p-values well below .01 except speaker 2, where $p < .05$).

SPEAKER	LAYM				PROF			
	1	2	3	4	5	6	7	8
z-abs	*.09*	*.05*	*.10*	*.09*	*.12*	*.14*	*.16*	*.13*

Figure 4: *Starting point: speaker-specific absolute z-values resulting from correlation of unnormalized rating data with raw acoustic data (syllable duration)*

One general outcome was that dispersion normalization without central tendency normalization (Figure 3: methods 1, 2, 9, 10) would even reduce the correlation to near zero or negative values. This can be attributed to the fact that with mean deviation set to 1, differences in the position of the base-lines (see section 1, hypothesis 1) which are greater than 1 must lead to inter-rating incommensurability. Therefore, Figure 5 presents only the results for the remaining methods.

CT	MEAN									MEDIAN								
CTRD	NONE			IRAT			ILIST			NONE			IRAT			ILIST		
DISPRD	NONE	IRAT	ILIST	NONE	IRAT	ILIST	NONE	IRAT	ILIST	NONE	IRAT	ILIST	NONE	IRAT	ILIST	NONE	IRAT	ILIST
methods				3	4	5	6	7	8				11	12	13	14	15	16
mean z-diff				.10	.12	.11	.05	.02	.07				.09	.11	.10	.05	.02	.06

Figure 5: *Average gain (mean z-diff) resulting from selected normalization methods (crossed out fields represent unnormalized data or ineffective methods)*

Figure 5 shows that CTRD = IRAT clearly leads to better results than CTRD = ILIST. With respect to hypotheses 1 to 3 (see section 1), this indicates that listeners' base-lines indeed vary in the course of rating 8 rather long signals. The noise caused can be neutralized to a certain extent by IRAT-based normalization of CT. Concerning the remaining factors we can thus confine to the methods associated with CTRD = IRAT (methods 3, 4, 5, 11, 12, 13).

Concerning CT, MEAN appears to be slightly superior to MEDIAN. With respect to hypothesis 4 (see section 1), it can be said that zeroes in MEDIAN/IRAT ratings do not directly reflect the listener's base-line. This may in part be due to declination phenomena. The next evaluation step relates to CT = MEAN, but in general it does not seem to matter much which of the two parameters is applied for CT normalization.

The question regarding DISPRD appears to be whether to carry out dispersion normalization at all. The fact that DISPRD = IRAT is not advisable (see section 1, hypothesis 5) leaves us with DISPRD = NONE vs. ILIST (method 3 vs. method 5). To get a more detailed picture, Figure 5 shows the underlying distributions of speaker-specific z-diff values.

SPEAKER	LAYM				PROF			
	1	2	3	4	5	6	7	8
z-diff (method 3)	.08	.05	.12	.12	.09	.10	.13	.11
z-diff (method 5)	.08	.06	.13	.14	.12	.11	.14	.11

Figure 5: *Speaker-specific gain (z-diff) without dispersion normalization (method 3) and with dispersion normalization (method 5)*

Comparing the distributions yielded by method 3 and method 5 shows that inter-listener differences in rating-"generosity" create much less noise than intra-listener differences in terms of base-line. The smallness of the differences indicates that dispersion normalization may even be neglected altogether. This assumption is supported further by the analysis shown in Figure 6 (see below).

6. Conclusion

The evaluation of normalization methods of syllable prominence ratings shows an effective and at the same time rather simple way of reducing noise in syllable prominence ratings: Set central tendency = 0 with reference to the distribution of measures delivered by one listener in connection with one signal. Whether to apply mean or median (method 5 vs. method 11, see Figure 5) appears to be hardly relevant concerning the factors taken into account here. Dispersion normalization is apparently unnecessary altogether.

What remains to be clarified is the extent of actual profit from normalization, when only one prominence rating per phonetic syllable is wanted and the normalized single ratings of one phonetic syllable are averaged. On this basis, Figure 6 shows regular Pearson's product-moment coefficients r, because this is one of the most conventional of parameters in correlation analysis.

SPEAKER	LAYM				PROF			
	1	2	3	4	5	6	7	8
r unnorm, after averaging	.38	.22	.42	.31	.35	.40	.51	.40
r meth.-3-norm, after averaging	.38	.21	.43	.43	.36	.46	.56	.52
r meth.-5-norm, after averaging	.36	.22	.42	.44	.39	.46	.55	.51

Figure 6: *Speaker-specific correlation coefficients r with respect to duration after averaging rating measures to receive one measure per phonetic syllable – upper box: unnormalized and method-3-normalized values, lower box: method-5-normalized values*

It is not surprising that the values all are much higher than the z-values (and also underlying r-values) discussed before, because this type of averaging is in itself a form of further normalization. The speaker-specific gains concerning method 3 ($r_{meth.-3-norm} - r_{unnorm}$) range from -.01 to .12, showing a tendency to be higher for the professional speakers. Respective gains concerning method 5 confirm the conclusion that the additional normalization of dispersion is unnecessary. (We also tried sd-based dispersion normalization, other things being equal to method 5, but the picture did not change.)

Further research points in two directions. One is about the "Gold"-corpus and what can be found out about, e.g., the hypothesis illustrated by text-samples (3) and (4) in section 2, using method-3-normalized prominence ratings. The other is about validating the outcome of the evaluation presented. Here, much work remains to be done: Other acoustic correlates and other types of speech signals as well as other languages should be included. In a first step, we are currently preparing the "Gold"- and other corpora of German read speech in order to analyze pitch and intensity measures, too.

7. References

[1] Fant, G. and Kruckenberg, A., Preliminaries to the study of Swedish prose reading and reading style. STR-QPSR, 2/1989, pp. 1–80, KTH, Stockholm, 1989.

[2] Eriksson, A., Thunberg, G. and Traunmüller, H., Syllable prominence: A matter of vocal effort, phonetic distinctness and top-down processing. In Proceedings of Eurospeech 2001, pp. 399–402, Aalborg, Denmark, 2001.

[3] Liljencrants, J., Judges of prominence. In Fonetik 99: Proceedings from the Twelfth Swedish Phonetics Conference (Andersson, R., Abelin, Å., Allwood, J. and Lindblad, P., eds), pp. 101–107, Department of Linguistics, Göteborg University, Göteborg, 1999.

[4] Arnold, D., Wagner, P. and Möbius, B., Evaluating different rating scales for obtaining judgments of syllable prominence from naïve listeners. In Proceedings of the 17th ICPhS 2011, Hong Kong, 2011.

[5] Heine, H., Historisch-kritische Gesamtausgabe der Werke. Bd. 3. Romanzero, Gedichte. 1853 und 1854, Lyrischer Nachlaß (Windfuhr, M., ed), Hamburg, Germany, 1992.

[6] Boersma, P. and Weenink, D., Praat: doing phonetics by computer (Version 5.1.31) [Computer program], 2011. URL: http://www.praat.org/

[7] Gibbon, D., SAMPA-D-VMlex Dokumentation V1.0, URL: http://coral.lili.uni-bielefeld.de/Documents/sampa-d-vmlex.html, Bielefeld, 1995.

[8] R Development Core Team, R: A language and environment for statistical computing (Version 2.12.2). R Foundation for Statistical Computing, Vienna, Austria. 2011. URL: http://www.R-project.org.