Four Approaches to Extracting Gradient of Intonation Downtrends in Czech

Jan Volín, and Tomáš Bořil
Institute of Phonetics
Faculty of Arts, Charles University
Prague, Czech Republic
jan.volin@ff.cuni.cz

***Abstract*—Linguistic descriptions of speech often exclude the concept of naturalness since they usually focus on ideal forms. As a consequence, they sometimes miss essential features of speech structure which not only contribute to its natural sound, but also support the coding of the meaning. That is the case of intonation downtrends, which received the deserved level of attention only relatively recently. However, there are still uncertainties concerning the methodology of their quantification. The present study evaluates four different approaches across two prosodic domains and two speech genres. We processed over 2400 prosodic phrases and over 850 declination units produced by 24 speakers. The chief objective was to map behavior of downtrends in Czech spoken texts, and suggest a suitable method of their quantification facilitating future cross-linguistic comparisons. Moreover, reference data are provided on intonation downtrends for Czech, a West Slavic language of Central Europe.**

***Keywords*—declinaton; downtrend; fundamental frequency; prosodic phrase; speech unit**

I. Introduction

Intonation downtrends are broadly defined as a gradual decrease in the height of comparable melodic events in the course of an intonation whole. A certain melodic event manifests higher absolute f_0 values at the beginning of a unit than a phonologically equivalent event in the middle or at the end of the unit. If the downtrend is sought in a relevant prosodic unit, its surface manifestation can be generalized into the expression $\Delta f_0/\Delta t < 0$. John Ohala is cited in [1] saying that the phenomenon can be observed in most and perhaps all languages of the world. Similarly, [2] lists downtrends among the three principal intonation universals.

Past decades have established that intonation downtrends are not just a plain physiological consequence of breathing (e.g., [3], [4] and [5]). In fact, they fulfil important roles in structuring spoken texts and in guiding the listener through the semantic contents of utterances (e.g., [6], [7], [8], and [9]). Although numerous questions of their functions in communication have been addressed, there are languages that still lack comprehensive description of this aspect of prosodic structure. The Czech language, the mother tongue of the two authors of this study, is one of them.

Even the languages in which downtrends have been studied do not possess descriptions of various speech styles, genres or communicative situations. Although [10] warned about 40 years ago that various speech styles might produce unequal values of downtrend gradients, most of the research has been done on isolated sentences or the so-called 'laboratory speech'. Our current study uses speech with communicative intent.

A crucial issue in speech descriptions is the methodology of parametrization. Not only should it allow for a wider use in the given research field, but it should also reflect the substantial properties of the phenomenon without too much noise. We consider the method proposed by [11] promising in this sense. By exploring four variants of this method (four approaches to its use), we would like to contribute to the debate concerning the speech units to be measured, and the approach to the data extraction. Ultimately, this debate should identify the methodology that is both linguistically meaningful and technologically expedient.

The research questions to be answered in the current study:

- How high is the occurrence of downtrends across the two genres of speech and the two prosodic domains?
- Does the frequency of occurrence differ substantially according to the approach to data extraction?
- What are the mean values of f_0 trends across the two genres and two prosodic domains?
- If only one of the four approaches were to be used, which would represent the others best?
- Does the size of the examined speech unit matter?
- Are there any speaker idiosyncrasies in the material?

II. Method

A. Speech samples

Two types of material were used: poetry reciting (POR) and news-reading (NWS). These represent two quite different genres, although the speech style could be conceived as essentially the same: clearly articulated monologuing based on a written text. An important aspect of the material is that it represents two genres with undisputable communicative intent: the speakers wish to be understood and appreciated.

This research was supported by Grant Agency of the Czech Republic, project no. *GAČR 20-15650S.*

The recordings representing the news reading (NWS) were authentic news-bulletins from a national broadcaster (channels Czech Radio 1 and Czech Radio 2). The current Czech Radio news readers are expected to represent civilized normative speech without any colloquialisms, salient idiosyncrasies or fashionable mannerisms: they are considered guarantors of the model Czech speech production. Our NWS sample consisted of 12 such experienced professionals (6 female + 6 male). The news bulletins are typically 3 to 4 minutes long (with voices of correspondents between paragraphs excluded). A typical extent of a news bulletin in our sample was 40 sentences, which is about 500 words.

The samples of poetry reciting (POR) were recorded in the sound treated studio of the Institute of Phonetics in Prague. The speakers (6 female + 6 male) were volunteering students of philology who expressed their inclination to poetry. They were given several poems, and were asked to get familiar with the contents and form of each of them, practice lines as long as they needed, and then read the poems out loud as if reciting for audiences. Three poems of similar structure (20 verses of 8 syllables) by each speaker were selected into the current set.

In both genres, our 24 speakers knew the texts ahead (with news readers actually composing or at least editing them). The sampling frequency and bit resolution of the recordings was considered irrelevant for the task since we were concerned with f_0 obtained with the autocorrelation algorithm.

B. Prosodic domains

The question of a downtrend domain has led to various candidates. It is also quite possible that downtrends operate at various levels simultaneously [2, p.35]. We opted for two units that are commonly accepted. First of all, it is the *prosodic phrase* (PP), also dubbed intonation phrase or tone unit in literature. PPs were established by expert auditory analysis guided by [12]. Second, the term *declination unit* (DU) suggested by [6] is used for two-verse stretches in POR speech and for individual sentences in NWS material.

C. Data extraction

The f_0 tracks were extracted with the autocorrelation method in Praat [13]. The window length was set individually for each speaker between 20 and 60 msec depending on their pitch range. The tracks were manually corrected for octave jumps, failures in creaky voice regions, etc. Resulting contours were smoothed with a 15 Hz bandwidth filter, interpolated through voiceless consonants, and converted into semitone (ST) values to approximate human perceptual framework.

The computational method as such was always the least-square linear fit of the f_0 tracks in the given speech unit (PP or DU – see above). Computations were implemented in [14]. Our four approaches differed in the data points from the contour that were chosen for the fit (Fig. 1). Specifically, they were:

- all-point approach (APT),
- all-syllable-nuclei approach (ASN),
- stressed-syllable-nuclei approach (SSN), and
- post-stress-syllable-nuclei approach (PSN).

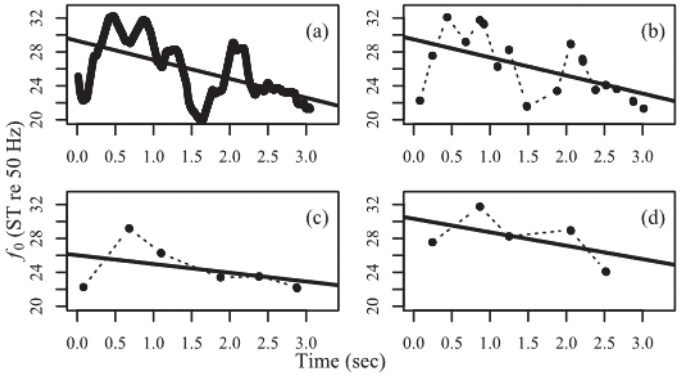

Fig. 1. Illustration of 4 approaches to downtrend quantification: panel (a) – APT, (b) – ASN, (c) – SSN, (d) – PSN. (DU by speaker PS01).

In APT (Fig. 1a) all points of a pre-processed f_0 track were used. The points were interspersed by 3 milliseconds. In ASN (Fig. 1b) only the values of syllable nuclei were used. These were represented by the values of three neighboring points in the middle of a vowel (or a syllabic liquid). The SSN approach (Fig. 1c) worked only with the stressed syllable nuclei. Finally, the PSN approach took only the vowels of syllable nuclei in post-stress position. This approach is motivated by the fact that typical stress-groups (SGs) in Czech exhibit a post-stress rise [15]. In other words, the most frequent pitch accent in Czech is L*+H (after [12]). It should be noted, that the number of post-stress vowels is sometimes lower than the number of stresses (Fig. 1d) due to unit-final monosyllabic words. Importantly, in all four approaches only PPs with at least 2 stress-groups (requirement of recurring events – see Section I, Introduction), and DUs with at least 2 PPs were considered (to avoid DU = PP).

III. Results

The arrangement of the presentation here follows the order of the research questions at the end of the Introduction above. Each of the questions listed will receive its own subsection apart from the first two that belong together.

A. Incidence of downtrends

Table I documents prevalence of speech units with negative gradient of the trendline, i.e., with the downtrend. Averaging across both prosodic domains (DU, PP) and both genres (POR, NWS), we see that about 80 % of the speech units are produced with a downtrend. The highest counts are in declination units of news reading (DU-NWS), whereas the lowest in prosodic phrases of the same genre (PP-NWS). In poetry reciting (POR) the percentages are roughly balanced between DUs and PPs. Individual approaches return quite similar results.

TABLE I. Percentages of negative gradients in speech

	Approaches to extracting the gradient			
	APT	***ASN***	***SSN***	***PSN***
DU-POR (n = 356)	86.2	88.2	79.5	82.0
DU-NWS (n = 414)	90.6	91.1	87.0	81.4
PP-POR (n = 749; 537)	88.7	89.6	76.9	84.2
PP-NWS (n = 1327; 1136)	68.3	66.8	66.8	62.9

B. *Trendline gradients*

The quantification of the downward slopes is considered an important delivery of our study. Fig. 2 presents the outcomes. The mean values range roughly between –0.5 and –3 semitones per second (ST/sec). It seems that the situation is quite similar, even if not identical, for declination units in news reading and poetry (DU-NWS and DU-POR). Prosodic phrases in news reading (PP-NWS) are not far from the previous in the mean values, but the variation as expressed by standard deviation is much higher there (Fig. 2, bottom). This finding somehow resonates with the last line of Table I, which indicates the lowest relative incidence of downtrends in PP-NWS – below 70 % by all four approaches.

Prosodic phrases in poetry reciting (PP-POR) were produced with remarkably steeper downward slopes, yet the variability does not exceed that of PP-NWS.

The pattern of mutual ratios among values obtained by the four approaches is roughly parallel for DU-NWS and DU-POR for both the mean and standard deviation (see left part of both the top and the bottom section of Fig. 2). Contrary to that, the four approaches produced somehow equalized output under PP-NWS condition, and disparate values under PP-POR condition. A striking difference in the latter is visible especially between SSN and PSN approaches. We briefly return to the difference between stressed-syllable nuclei and post-stress-syllable nuclei with a comment in Section IV – Discussion.

C. *Correlations of outputs*

To measure the association among the values of downtrend gradients obtained by our four approaches, we computed Pearson correlation coefficients. They are presented in Table II. All the correlations were statistically significant at the level of $\alpha = 0.05$. The coefficients range between 0.343 and 0.986, and the values in bold indicate the highest coefficient in the given line of the table (naturally, with the self-correlated values where $r = 1$ excluded).

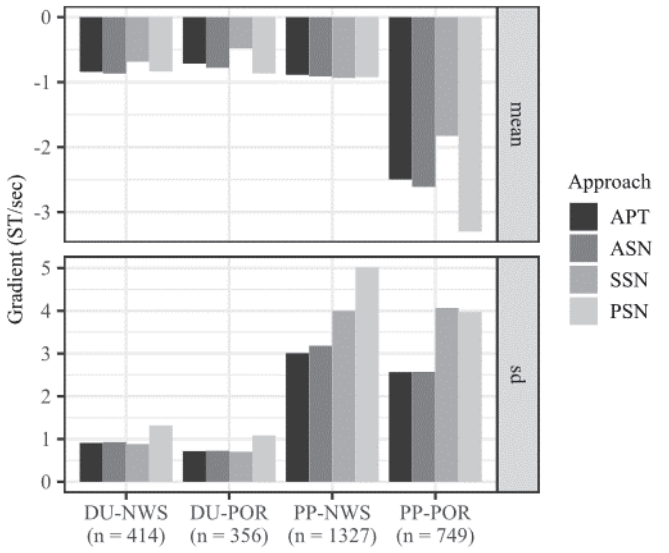

Fig. 2. Means and standard deviations of trendline gradients extracted by four approaches in four types of speech material.

TABLE II. CORRELATIONS AMONG GRADIENTS IN FOUR TYPES OF SPEECH MATERIAL

	Approaches to extracting the gradient			
DU-POR	***APT***	***ASN***	***SSN***	***PSN***
APT	1.000	**0.963**	0.628	0.730
ASN	**0.963**	1.000	0.720	0.757
SSN	0.628	**0.720**	1.000	0.531
PSN	0.730	**0.757**	0.531	1.000
DU-NWS				
APT	1.000	**0.986**	0.830	0.863
ASN	**0.986**	1.000	0.840	0.871
SSN	0.830	**0.840**	1.000	0.677
PSN	0.863	**0.871**	0.677	1.000
PP-POR				
APT	1.000	**0.920**	0.370	0.597
ASN	**0.920**	1.000	0.403	0.644
SSN	0.370	**0.403**	1.000	0.343
PSN	0.597	**0.644**	0.343	1.000
PP-NWS				
APT	1.000	**0.951**	0.440	0.669
ASN	**0.951**	1.000	0.438	0.654
SSN	**0.440**	0.438	1.000	0.403
PSN	**0.669**	0.654	0.403	1.000

It is noteworthy that correlations between APT and ASN always exceed 0.9, that is, regardless of the genre or the prosodic domain, these two approaches produce very similar outcomes. We consider this a useful bit of information since pre-processing of speech samples for both APT and ASN are labor extensive in a different manner, thus, researchers can decide which of the two is more convenient in their situation.

D. *Size of the speech unit*

The size of a speech unit can be expressed in smaller units that form it. Thus, we measured PPs in the number of stress-groups (SGs) they contained, and DUs in the number of PPs they contained. Fig. 3 shows how the mean gradient changes with the growing size of a unit. PPs are on the left, while DUs are on the right of the figure. The ASN approach is used.

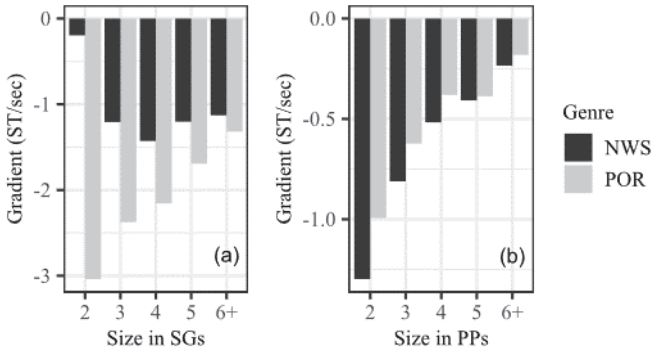

Fig. 3. Downtrend gradients in speech units of different sizes expressed in smaller constituent units. Panel (a) – PPs in stress-groups (SGs); Panel (b) – DUs in PPs .

Downtrends in DUs exhibit mutually similar effect in both POR and NWS. The DUs consisting of only two prosodic phrases have the steepest slopes. As the number of PPs in a DU increases, the downtrend becomes more moderate, down to only about 0.25 ST in units consisting of 6 or more PPs.

In the case of PPs measured in constituting SGs (Fig. 3a, i.e., left panel) the two genres did not produce parallel results. Poetry reflects the situation in Fig. 3b: the shortest units have the steepest downtrends, and the longer the unit, the more moderate the downtrend becomes. Interestingly, short PPs in news reading do the exact opposite: their mean gradient is close to 0. The phrases of 3 or more SGs do not exhibit any clear trend: they all decline by slightly over 1 ST per sec.

E. Individual speakers

Our current sample comprised speech production of 24 individuals (12 news readers + 12 poetry reciters). Their performance was far from uniform. In Fig. 4 they are ordered by their mean gradient. First of all, it is interesting to notice that producing the steepest downtrend in DUs does not mean the same in PPs. That can be observed in both NWS (i.e., NR speakers) and POR (i.e., PS speakers). Second, there is no monotonous trend in the size of standard error of the mean, which is captured by the whiskers in the graph, even though the NR03 produced the steepest downtrend and the largest standard error, while NR09 did the opposite.

IV. Discussion

Intonation downtrends were confirmed in our study as clearly present in the examined material. Also, the trends displayed reasonably nonrandom behavior. It can be suggested that the typical sound of news reading and poetry reciting in the Czech language entails gradual decrease in f_0 values throughout prosodic phrases and declination units. Admittedly, our current study leaves linguistic contents of the speech units for future.

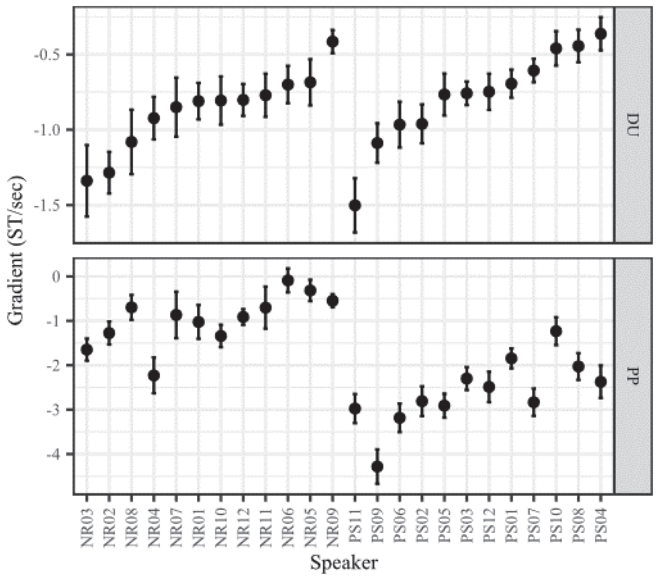

Fig. 4. Mean downtrend values produced by individual speakers. The whiskers capture std. error of the mean. NR = news reading, PS = poetry.

Apart from that, the future research should also address the difference between SSN and PSN. Due to the common post-stress rises in Czech, the SSN is probably paralleling the intonation *baseline*, while PSN the converging *topline*.

Multi-layered prosodic structure as suggested by counter-parallelisms in Fig. 3 and 4 signals that units of various prosodic levels are worth examining or evaluating jointly.

Finally, perceptual testing is indispensable if we want to find out how downtrends contribute to various communicative functions of speech including the esthetic one.

V. Conclusion

Two investigated communicative genres manifested high proportion of downtrends in both prosodic phrases and declination units. At the same time, however, downtrends in news reading behaved differently from those in poetry reciting. That was especially visible in the domain of prosodic phrases.

Of the four approaches to downtrend quantification, two are equally satisfactorily sensitive (APT, ASN), while the other two (SSN, PSN) will require further linguistically informed analyses.

References

[1] P. Wong C. M., "The effect of downdrift in the production and perception of Cantonese level tones", in Proceedings of ICPhS, San Francisco: UCLA, 1999, pp. 2395-2398.

[2] D. Hirst and A. Di Cristo, Intonation Systems: A Survey of Twenty Languages. Cambridge: Cambridge University Press, 1998.

[3] W. E. Cooper and J. M. Sorensen, Fundamental Frequency in Sentence Production. New York: Springer-Verlag, 1981.

[4] J. 't Hart, R. Collier, and A. Cohen, A Perceptual Study of Intonation: An Experimental-Phonetic Approach to Speech Melody. Cambridge University Press, 1990.

[5] J. Yuan and M. Liberman, "F0 declination in English and Mandarin Broadcast News Speech", Speech Commun., vol. 65, pp. 67–74, 2014.

[6] S. Schuetze-Coburn, M. Shapley, and E. G. Weber, "Units of intonation in discourse: a comparison of acoustic and auditory analyses", Lang. Speech, vol. 34, no. 3, pp. 207–234, 1991.

[7] D. R. Ladd, Intonational Phonology, 2nd ed. Cambridge: Cambridge University Press, 2008.

[8] T. H. Morrill, L. Dilley, and J. McAuley, "Prosodic patterning in distal speech context: Effects of list intonation and f0 downtrend on perception of proximal prosodic structure", J. Phon., vol. 46, pp. 68–85, 2014.

[9] K. Furukawa and Y. Hirose, "Boundary-driven downstep in Japanese", in Proceedings of 19th ICPhS, Canberra: ASSTA, 2019, pp. 1009–1013.

[10] N. Umeda, "F0 declination is situation dependent", J. Phon., vol. 10, no. 3, pp. 279–290, 1982.

[11] P. Lieberman, W. Katz, A. Jongman, R. Zimmerman, and M. Miller, "Measures of the sentence intonation of read and spontaneous speech in American English", JASA, vol. 77, no. 2, pp. 649–657, 1985.

[12] M. E. Beckman and G. Ayers Elam, "Guidelines for ToBI Labelling", version 3. Ohio State University: The Ohio State University Research Foundation, 1997.

[13] P. Boersma and D. Weenink, "Praat: doing phonetics by computer," http://www.praat. org/, 2019.

[14] R Core Team, R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, 2021. URL https://www.R-project.org/.

[15] M. Jůzová and J. Volín, "F0 Post-Stress Rise Trends Consideration in Unit Selection TTS", in Text, Speech, and Dialogue 2018, Springer Verlag, 2018, pp. 360–368.