Speech Prosody 2016
31 May - 3 Jun 2106, Boston, USA

Acoustic Cues to Perceived Prominence Levels – Evidence from German Spontaneous Speech

Stefan Baumann[1], *Oliver Niebuhr*[2] *& Bastian Schroeter*[3]

[1] I*f*L Phonetik, University of Cologne, Germany
[2] Dept. of Design and Communication, IRCA, University of Southern Denmark, Denmark
[3] Institut für Psychologie, CAU Kiel, Germany

stefan.baumann@uni-koeln.de, olni@sdu.dk, stu87088@mail.uni-kiel.de

Abstract

A corpus study on German spontaneous speech proved the robustness of a meaningful perceptual distinction between two levels of prominence, namely fully-fledged versus secondary accents, and its effective use in prosodic annotation. We found that the two levels are characterized by prosodic profiles that mainly differ in gradient phonetic features (F0 range, duration, intensity) and are only to some extent influenced by the choice of phonological accent type. Furthermore, the distinction between the two levels turned out to be largely independent of medial or final accent positions in the phrase.

Index Terms: prominence, perception, annotation, German, intonation, spontaneous speech, pitch accent, phrase accent

1. Background

Only few prosodic annotation models include a tier for perceived prominences. Recent exceptions are the *RaP* model (originally proposed for American English [1]) or the DIMA model for German intonation [2]. However, the first annotation model that systematically integrated different prominence levels already in the early 1990s was the *Kiel Intonation Model* (KIM [3,4,5]). The prosodic annotation of speech data in KIM includes four prominence levels: no accent (level 0), secondary (reduced) accent (level 1), default (fully-fledged) pitch accent (level 2), emphatic pitch accent (level 3). Importantly, the distinction between these levels is considered phonological, i.e. meaningful.

The sentence *Er ist ins Kino gegangen* ('He went to the cinema') in Figure 1 gives an example of the functional relevance of the different accent levels. The first utterance is realized with a sequence of accent levels 2 and 1 on noun and verb. It is informationally neutral in that it puts the focus on the entire utterance. In contrast, leaving the verb completely unaccented (level 0) shifts the focus to the first prominent element, and in this way creates a (narrow-focus) contrastive interpretation of the utterance in the sense of "He went to the cinema and not to the theater". Similarly, a fully-fledged accent (level 2) on the verb puts it in the limelight so that the (narrow-focus) contrastive interpretation of the utterance changes to "He walked to the cinema instead of taking his car". A further increase from level 2 to level 3 on noun or verb maintains the respective narrow-focus interpretation and additionally intensifies the respective piece of information. For example, changing the level 2 accent on *Kino* into an emphatic level 3 accent could signal "Oh, I envy him so much! I also wanted to go there today!"

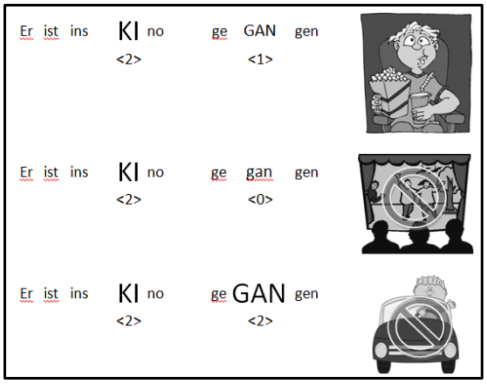

Figure 1: *Differences in the interpretation of the sentence* Er ist ins Kino gegangen *('He went to the cinema') caused by accent levels 0, 1 and 2 on the verb* gegangen *('went'). The example is inspired by a similar one in Kohler [3].*

Many studies investigated the phonetic cues that have an impact on prominence perception. F0 variation was found to be a particularly important cue [6,7] but also duration [8] and intensity [9], or 'total amplitude' (a combination of duration and intensity [10]) are involved in perceptual prominence, see Terken & Hermes [11] for an overview. All prosodic parameters show a positive correlation with prominence, i.e. the higher the parameter level, the more the corresponding element (syllable or word) stands out in perception.

The assumption of a secondary prominence level next to fully-fledged pitch accents is actually quite wide-spread. Usually, however, secondary prominences are structurally subordinated to fully-fledged pitch accents, either distributionally or phenomenologically. For example, approximate equivalents of level 1 prominences have been described as *phrase accents* (only in postnuclear position [12]), *post-focal prominences* indicated by a reduced pitch register (also postnuclear [13]), *rhythmically determined accents* (only prenuclear, e.g. [14]), or, more generally, *post-lexical stresses* (both in pre- and postnuclear position [15]). A secondary status may also be attributed to those fully-fledged accents that are inherently less prominent due to their accent *type*: Low accents (L*, following a GToBI notation, see [16]) or early peak accents with a falling onglide (H+L*/H+!H*) were perceived to be less prominent than accents with a rising onglide (e.g., L+H*) in German [17,18].

 doi: 10.21437/SpeechProsody.2016-146

For a long time, KIM was the only intonation model in which level 1 and level 2 accents were equivalent oppositions in the same phonological prominence paradigm. That is, the Kiel model assumes that instances of both levels can occur at *any* position in the phrase and link up with *any* type of pitch accent. It was also for these reasons that annotators of the *Kiel Corpus of Spontaneous Speech* [19,20] – which was at the same time driving force and main field of application of KIM – received no specific instructions as to how levels 1 and 2 should be used. The only training that annotators received were example sentences like those in Figure 1 in combination with the explicit instruction that the accent level decision was to be made solely by ear.

This knowledge-by-acquaintance approach of KIM (cf. [21]) and the hundreds of accent level decisions made on this basis represent a unique testbed for advancing our understanding of the perception and prosodic manifestation of pitch accents in general and accent levels in particular. Therefore, we address the following two questions, focusing on the non-emphatic and most frequent level 1 and level 2 accents in the German *Kiel Corpus*:

1. Is a distinction between prominence levels (annotated on a purely perceptual basis) systematically correlated with particular acoustic and phonological profiles?

2. How do the differences between prominence levels 1 and 2 look like? Do they involve all major correlates of perceived prosodic prominence? Are there differences due to position in the intonation phrase (medial vs. final)?

2. Method

2.1. Data

The *Kiel Corpus of Spontaneous Speech* was recorded in an appointment-making scenario. The corpus consists of 118 dialogues by 26 speaker pairs (52 Standard German speakers, 29 males, 23 females), adding up to about four hours of speech. The recordings have been segmentally and prosodically annotated. Crucially, all accents are annotated for prominence levels 1 to 3. The manual perception-based annotation took almost a decade and was done by ten different phonetically trained research assistants in the course of time. Critical cases were discussed and decided by a majority vote of the respective annotators.

For our investigation, we selected two frequent and clearly defined prosodic patterns, as is shown in Figure 2: (a) The target accent was in *phrase-medial* (prenuclear) position and concatenated through F0 valleys with regular pitch accents on both sides, (b) the target accent was in *phrase-final* position, separated by a boundary tone from the end of the phrase and by an F0 valley from a preceding fully-fledged pitch accent. Accent clash conditions were excluded. Note that in approaches other than KIM the target accent in condition (b) would be considered a nuclear *or* postnuclear accent, depending on whether or not it is realized as a fully-fledged pitch accent.

A script-based search for target accents in the corpus using Scilab [22] yielded a sample of 738 items in phrase-medial position, i.e. for prosodic pattern (a) (level 1: 263, level 2: 475), and an additional sample of 453 items in phrase-final position, i.e. for prosodic pattern (b) (level 1: 82, level 2: 371).

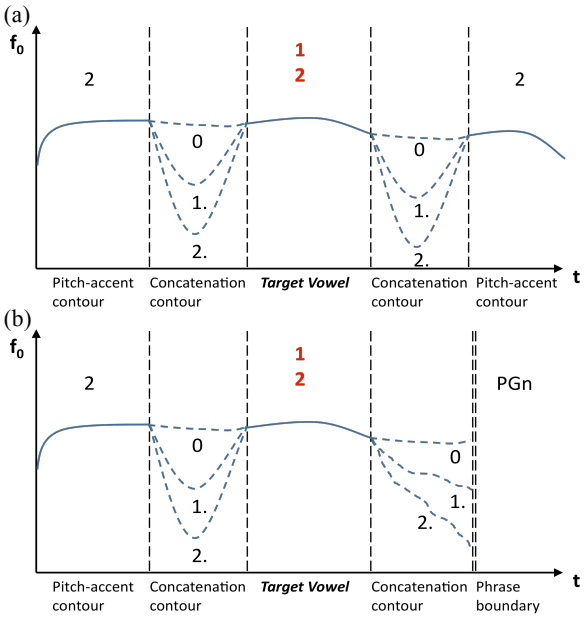

Figure 2: *(a) Target word (accented vowel) in phrase-medial position, (b) target word (accented vowel) in phrase-final position; <PGn> denotes a major phrase boundary, <1./2.> represent weak/strong dips in F0.*

Within the two context frames of Figures 2(a) and (b), we concentrated on the stressed vowel segments of the accented target words. Accented syllables would have been a more obvious measurement unit, but the KIM annotation does not provide syllable boundaries. This is due to the empirically supported theoretical assumption that phonological differences in pitch accent alignment are made in relation to vowel rather than syllable boundaries (cf. [23,24,25]).

We measured for the accented vowels how prominence levels 1 and 2 differed in terms of F0 (mean height and range), duration, and intensity (RMS). The F0 was calculated by Praat [26] using the algorithm of Boersma [27]. RMS was calculated using a 17 ms hamming window. Duration measurements used the given vowel boundaries that trained phoneticians set by combining visual cues from the waveform and the spectrogram, as is suggested by Skarnitzl & Machač [28]. In order to avoid measurement errors, we excluded all target vowels from our samples with a voiceless proportion of more than 20% and/or which showed an F0 range larger than 150 Hz.

Additionally, we counted the frequency of the individual pitch accent types provided by the phonological paradigm of KIM. The frequency counts were made with reference to the pitch accent labels of PROLAB [4].

2.2. Hypotheses and statistical analysis

Based on previous studies on perceptual prominence, some of which are cited in the introduction, we hypothesized that, compared to level 1 accents, level 2 accents are marked by higher F0 means and ranges, longer durations, higher intensity/energy levels, and more prominent accent types (such as high and rising accents). Furthermore, we expect more level 2 prominences on phonologically long and open vowels, which are intrinsically more sonorous than other vowels.

Our measurements were post-processed based on Gaussian Kernel density estimates with a bandwidth according to Silverman's "rule of thumb" [29]. Statistical testing included Student's t-tests, a Mann-Whitney U-test, and χ^2-tests for frequency counts. All calculations were done using R [30].

3. Results

3.1. Phonetic cues

The results of our analyses neither yielded significant differences between the mean F0 height of vowels at level 1 and level 2 in phrase-medial nor in phrase-final position (t_{575}=1.321, p=0.187 and t_{309}=0.145, p=0.885). However, the F0 ranges of level 1 accents differed significantly from those of level 2 in both positions, see Figure 3(b). The distributions in Figure 3(a) illustrate in addition that this difference primarily relies on a fatter right tail in phrase-final position and a longer right tail in phrase-medial position.

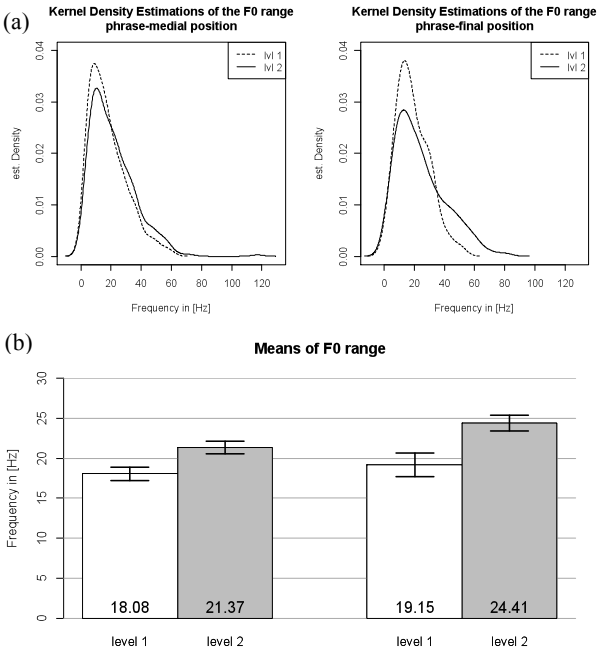

Figure 3: *(a) Kernel density estimations of F0 ranges in target vowels; (b) Means of F0 ranges in target vowels at level 1 and 2 in phrase-medial (left, t_{575}= 2.754, p=0.006) and phrase-final position (right, t_{309}=2.222, p=0.027).*

Furthermore, target vowels in the level 2 condition had longer durations both phrase-medially and phrase-finally, see Figure 4. According to t-tests, this difference was only significant in phrase-medial position, but, as the duration distributions had similar shapes and variances, we additionally conducted a Mann-Whitney U-test, which yielded a significant difference also in phrase-final position. We consider this a valid procedure accounting for the lack of a sufficient number of level 1 vowels in phrase-final position which presumably prevented the parametric t-test from reaching significance as well.

Regarding acoustic energy (RMS), the data set revealed significantly higher intensity values for level 2 accents relative to level 1 accents. Again, this difference applies to both phrase-medial and phrase-final position, see Figure 5.

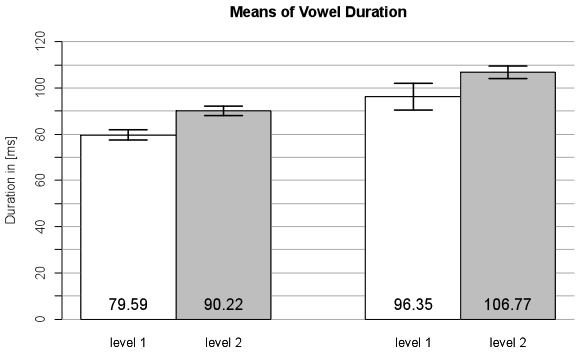

Figure 4: *Means of target vowel durations at level 1 and 2 in phrase-medial (left, t_{736}=3.365, p<0.001) and phrase-final position (right, t_{451}=1.610, p=0.108 and Md_1=77.19, Md_2=95.31, U=9435, p=0.022).*

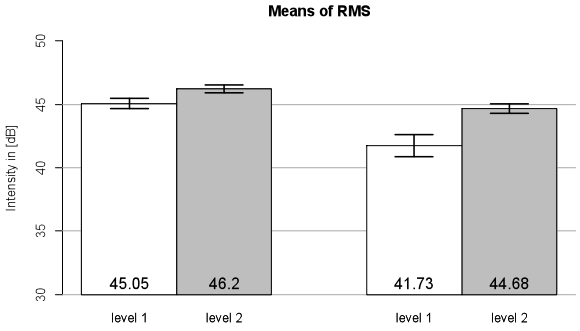

Figure 5: *Means of intensity (RMS) values of target vowels at level 1 and 2 in phrase-medial (left, t_{736}=2.228, p=0.026) and phrase-final position (right, t_{451}=1.516, p= 0.001).*

3.2. Phonological cues

The χ^2-tests showed that pitch accent types are differently distributed between the two prominence levels in phrase-medial (χ^2_5=47, p<0.001) but not in phrase-final position. Table 1 summarizes the phrase-medial distributions. Target vowels without separate tonal movements (i.e. post-lexical stresses) were much more frequently assigned to prominence level 1 than to prominence level 2 (13.6% vs. 2.7%). Among the other, melodic types of pitch accents, the medial peak (H*) is the most frequent type at both prominence levels (60%). Late peaks or rising accents (L+H*) occur more often with level 2 prominences (28.5% vs. 16.3%), whereas early peaks or falling accents (H+L*) are more frequently associated with level 1 prominences (8.7% vs. 5.5%). This is in line with the recent finding for German that rising accents are intrinsically more prominent [18].

Table 2 provides the distribution of accent types in phrase-final position. In accordance with the non-significant outcome of the χ^2-test, there is no obvious difference between frequencies of accent types in level 1 and level 2 prominence conditions. If at all, we can see slightly more rising accents in combination with level 2, which fits in well with the distributional difference in phrase-medial position.

With regard to possible effects of the intrinsic prosodic characteristics of different phonological vowel qualities on accent-level annotations, we did not find any significant variation both phrase-medially and phrase-finally.

Table 1. *Distribution of phrase-medial pitch accent types (in KIM terms, with GToBI equivalents) in absolute numbers and percentages.*

Accent Type	Level 1 Accent	Level 2 Accent
Level contour or (*)	36 (13.6%)	13 (2.7%)
Early peak or H+!H*	23 (8.7%)	26 (5.5%)
Medial peak or H*	159 (60.2%)	293 (61.4%)
Late peak or L+H*	43 (16.3%)	136 (28.5%)
Early valley or L+H*	2 (0.8%)	2 (0.4%)
Late valley or L*(+H)	0	5 (1.1%)

Table 2. *Distribution of phrase-final pitch accent types (in KIM terms, with GToBI equivalents) in absolute numbers and percentages.*

Accent Type	Level 1 Accent	Level 2 Accent
Level contour or (*)	2 (3.6%)	11 (3.9%)
Early peak or H+!H*	13 (23.2%)	67 (23.8%)
Medial peak or H*	36 (64.3%)	157 (55.9%)
Late peak or L+H*	5 (8.9%)	46 (16.4%)

3.3. Summary

Table 3 summarizes the results of the inferential statistics we performed, i.e. t-tests plus a Mann-Whitney U-test for the gradient phonetic cues and χ^2-tests for the discrete phonological cues.

Table 3. *Significant differences in phonetic and phonological cues between prominence levels 1 and 2.*

Phonetic/ Phonological Cue	Phrase-medial	Phrase-final
F0 height	n.s.	n.s.
F0 range	2 > 1	2 > 1
Duration	2 > 1	2 > 1 (U-test)
Intensity (RMS)	2 > 1	2 > 1
Accent type	2: More rising accents, 1: More post-lexical stresses	n.s.
Vowel quality	n.s.	n.s.

We did neither find a significant gender difference in the distribution of accent levels (χ^2_1=2.6981, p=0.1005) nor for position in the phrase (χ^2_1=2.4283, p=0.1192).

4. Discussion and Conclusions

Our corpus analysis revealed that KIM's two prominence levels are associated with significantly different prosodic profiles. These profiles include all known phonetic prominence cues – i.e. F0 range, duration, and intensity. Moreover, differences are in the expected direction and in accord with established knowledge in that prosodic parameters were higher/larger for level 2 than for level 1 accents.

Only to some extent, accent level 1 and 2 decisions were also influenced by pitch accent type in that the inherently more prominent rising accents were often assigned level 2 (cf. [17,18]). In contrast, intrinsic segmental differences in F0, duration, and intensity, i.e. differences between phonetic and phonological vowel features, did not affect the annotators' perceptual accent level decisions. This is worth noting as there is evidence that intrinsic segmental prosodies can have an effect on prominence perception [31]. The lack of such an effect in our study could be interpreted as supporting the assumption of a perceptual compensation of intrinsic segmental prosodies, particularly under real psychophonetic rather than artificial psychoacoustic listening conditions [32,33]. Alternatively, it could mean that effects of intrinsic segmental prosodies are too small to exceed the phonological – i.e. meaning-oriented – difference limen between accent levels 1 and 2. From that perspective, the effects of pitch accent type on prominence level could also be due to meaning differences rather than intrinsic prominence differences. Investigating the two alternatives is a task of follow-up studies. For us, the crucial point is that accent level decisions were not just made on the basis of other linguistic factors. Rather, they were made systematically in relation to external reference categories and in an empirically reasonable way across listeners. In this sense, they represent an additional layer of prosodic information. As a confirmation of this conclusion, Kügler et al. [2] found that prominence levels 1 and 2 – which were adopted for the DIMA model – can be reliably annotated, even more so in spontaneous than in read speech, i.e. when prominences can be assumed to cover a larger range and are generally more variable.

From a more theoretically oriented perspective, our results are clearly in favour of KIM's concept of a phonological accent level paradigm that is independent and hence freely combinable with a phonological inventory of pitch accents. Level 1 accents can occur at every position in the prosodic phrase, and they differ from level 2 accents quantitatively rather than qualitatively. Thus, notions like phrase accents, rhythmic prominences, or post-lexical stresses that are distributionally and/or phenomenologically clearly subordinated to fully-fledged level 2 accents are inconsistent with our findings. As a matter of fact, "reduced accents" are not restricted to postnuclear prominences, and a nuclear pitch accent can be perceived as "secondary" as well.

Taken together, the findings clearly suggest that it is appropriate and useful to distinguish between fully-fledged and reduced pitch accents in intonational modelling. Yet, a number of questions remain open. In particular, it must be kept in mind that accent levels 1 and 2 are the result of the annotators' perception-based and meaning-oriented instructions. Different instructions could have created different findings. In fact, our results do not question the relevance of concepts like phrase accents, rhythmic prominences, or post-lexical stresses. They might just not be adequately covered by accent level 1 as introduced to the annotators in the present study, since they represent non-tonal and/or non-meaningful events that are outside a phonological paradigm of accent levels. Extending our line of research in this direction as well as taking accent level 3 into account will be important tasks for the future.

5. References

[1] L. Dilley, and M. Brown. *The RaP Labeling System*, v. 1.0, MIT, 2005.

[2] F. Kügler, B. Smolibocki, D. Arnold, S. Baumann, B. Braun, M. Grice, S. Jannedy, J. Michalsky, O. Niebuhr, J. Peters, S. Ritter, C. Röhr, A. Schweitzer, K. Schweitzer, and P. Wagner, "DIMA – Annotation guidelines for German intonation," *Proceedings 18th ICPhS,* Glasgow, Scotland, paper 317, pp. 1-5, 2015.

[3] K. Kohler, "A model of German intonation", *AIPUK 25*, pp. 295-360, 1991.

[4] K. Kohler, "Modelling prosody in spontaneous speech," in Y. Sagisaka, N. Campbell, and N. Higuchi (eds.), *Computing Prosody. Computational models for processing spontaneous speech.* New York: Springer, pp. 187-210, 1997.

[5] B. Peters, and K. Kohler, "Trainingsmaterialien zur prosodischen Etikettierung mit dem Kieler Intonationsmodell KIM," URL: http://www.ipds.uni-kiel.de/kjk/pub_exx/bpkk2004_1/TrainerA4.pdf, 2004.

[6] D.B. Fry, "Experiments in the Perception of Stress", *Language and Speech* 1, pp. 126-152, 1958.

[7] D. Bolinger, "A Theory of Pitch Accent in English", *Word* 14, pp. 109-149, 1958.

[8] J. Cole, Y. Mo, and M. Hasegawa-Johnson, "Signal-based and expectation-based factors in the perception of prosodic prominence", *Laboratory Phonology* 1, pp. 425-452, 2010.

[9] G. Kochanski, E. Grabe, J. Coleman, and B. Rosner, "Loudness Predicts Prominence; Fundamental Frequency Lends Little", J. *Acoustical Society of America* 11(2), pp. 1038-1054, 2005.

[10] M.E. Beckman, *Stress and Non-Stress Accent*, Dordrecht: Foris, 1986.

[11] J. Terken, and D. Hermes, "The perception of prosodic prominence", in M. Horne (ed.), *Prosody: Theory and experiment,* Dordrecht: Kluwer, pp. 89-127, 2000.

[12] M. Grice, D.R. Ladd, and A. Arvaniti, "On the place of phrase accents in intonational phonology". *Phonology*, vol. 17, no. 2, pp. 143-185, 2000.

[13] F. Kügler, and C. Féry, "Postfocal downstep in German," *Language and Speech,* submitted.

[14] S. Calhoun, "How does informativeness affect prosodic prominence?", *Language and Cognitive Processes*, vol. 25, pp. 1099-1140, 2010.

[15] M. Grice, and S. Baumann, "Intonation in der Lautsprache: Tonale Analyse," in B. Primus, and U. Domahs (eds.), *Handbuch Laut, Gebärde, Buchstabe.* De Gruyter, Reihe Sprachwissen, to appear.

[16] M. Grice, S. Baumann, and R. Benzmüller, "German Intonation in Autosegmental-Metrical Phonology", in S.-A. Jun (ed.) *Prosodic Typology. The Phonology of Intonation and Phrasing*. Oxford: Oxford University Press, pp. 55-83, 2005.

[17] K. Kohler, and R. Gartenberg, "The perception of accents: F0 peak heigt versus F0 peak position", *AIPUK 25*, pp. 219-294, 1991.

[18] S. Baumann, and C. Röhr, "The perceptual prominence of pitch accent types in German", in *18th ICPhS, Glasgow, Scotland,* Proceedings, 2015, paper 298, pp. 1-5.

[19] K. Kohler, M. Pätzold, and A. Simpson, "From scenario to segment - The controlled elicitation, transcription, segmentation and labelling of spontaneous speech", *AIPUK* 29. 1995.

[20] B. Peters, "The Database 'The Kiel Corpus of Spontaneous Speech'," *AIPUK* 35a, pp. 1-6, 2005.

[21] B. Russell, "On Denoting", *Mind: New Series 14/56*, pp. 479-493, 1905.

[22] Scilab Enterprises, *Scilab: Free and Open Source software for numerical computation* (OS, Version 5.XX) [Software]. Available from: http://www.scilab.org, 2012.

[23] K. Kohler, "Timing and communicative functions of pitch contours", *Phonetica 62*, pp. 88-105, 2005.

[24] O. Niebuhr, "Categorical perception in intonation: a matter of signal dynamics?" *Proceedings Interspeech 2007,* Antwerp, Belgium, pp. 109-112, 2007a.

[25] O. Niebuhr, "The signalling of German rising-falling intonation categories - The interplay of synchronization, shape, and height", *Phonetica 64*, pp. 174-193, 2007b.

[26] P. Boersma, and D. Weenink (2015). Praat: doing phonetics by computer [Computer program]. Version 6.0.05, retrieved 8 November 2015 from http://www.praat.org/.

[27] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound", *IFA Proceedings* 17, pp. 92-110, 1993.

[28] R. Skarnitzl, and P. Machač, *Principles of Phonetic Segmentation*, Prague: Epocha, 2009.

[29] B.W. Silverman, *Density Estimation*, London: Chapman and Hall, 1986.

[30] R Core Team, "R: A language and environment for statistical computing," R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/, 2015.

[31] E. Rosenvold, "The role of intrinsic F0 and duration in the perception of stress," *ARIPUC*, vol. 15, pp. 147-166, 1981.

[32] C. Fowler, and J.M. Brown, "Intrinsic f0 differences in spoken and sung vowels and their perception by listeners", *Perception & Psychophysics 59*, pp. 729-738, 1997.

[33] O. Niebuhr, "Intrinsic pitch in opening and closing diphthongs of German", *Proceedings 2nd International Conference of Speech Prosody, Nara, Japan*, pp. 733-736, 2004.