INTERSPEECH 2009 BRIGHTON

Intonation segments and segmental intonation

Oliver Niebuhr

Laboratoire Parole & Langage, Université de Provence, UMR6057 CNRS, Aix-en-Provence, France
oliver.niebuhr AT lpl-aix.fr

Abstract

An acoustic analysis of a German dialogue corpus showed that the sound qualities and durations of fricatives, vocoids, and diphthongs at the ends of question and statement utterances varied systematically with the utterance-final intonation segments, which were high-rising in the questions and terminal-falling in the statements. The ways in which the variations relate to phenomena like sibilant/spectral pitch and intrinsic F0 suggest that they are meant to support the pitch course. Thus, they may be called segmental intonations.

1. Introduction

Perceived pitch is the essence of intonation. Different kinds of pitch movements as well as their timings relative to accented syllables or phrase boundaries convey communicative (e.g., attitudinal) meanings that refer to the associated words, the dialogue partner, or the dialogue itself. The phonological structuring and modelling of intonation focuses on the main acoustic counterpart of pitch, i.e. the fundamental frequency (F0), cf. the autosegmental-metrical (AM) theory [1] or the Kiel Intonation Model, KIM [2]. From this point of view, the segmental string appears at first sight as a mere troublemaker for the signalling of intonational meanings. Voiceless segments interrupt F0, and biomechanical interactions between tongue, jaw, and larynx, as well as changes in the supraglottal air pressure due to segment articulations create so-called microprosodic perturbations in the F0 course [3,4]. At a closer look, however, there are indications that the segmental string also contributes substantially to the signalling of the pitch patterns and communicative meanings associated with intonation. For example, speakers make use of the alternation between the low and high acoustic energy levels caused by continuous closing and opening of the vocal tract during the production of consonant-vowel sequences in order to differentiate meaningful intonational categories by suppressing or highlighting the perceptual salience of F0 sections, cf. [5,6]. This strategy then creates the timing and alignment variations that we observe for different intonation categories in different segmental contexts across languages, cf. [7,8].

Furthermore, the segmental string has the potential to contribute to the perceived pitch pattern. For example, the open-closed dimension of the vowel space, which is negatively correlated with the first-formant (F1) frequency, is directly related to F0. Everything else being equal, open vowels have a lower F0 than closed ones. This phenomenon, which is only partly compensated for in perception, is known as intrinsic F0. Moreover, back and front vowels (in combination with rounded and spread lips) lower and raise higher formants like F2. This change in the spectral energy distribution has an indirect effect on the perceived pitch. Everything else being equal, back and rounded vowels with a lower F2 yield lower pitch impressions than front and non-rounded ones with a higher F2, cf. [9,10]. Such a spectral pitch can even be evoked within a certain range by voiceless fricatives despite the lack of F0, solely based on the energy distribution within the noise spectrum, cf. [11] and "sibilant pitch", [12]. As in the case of the vowels, more back and / or rounded fricatives lower the pitch impression. But do speakers make use of such segment-based pitch potentials for conveying intonation and its meanings? For example, do the realizations and acoustic manifestations of speech sounds vary depending on the intonational context in order to create "lighter" or "darker" sounds?

Except for the study of [13] on American English, which seems to be more related to accentuation (i.e. the degree of salience) than to intonation, there are only few indications for pitch-related variations of speech sounds from tone languages yet, cf. [14, 15]. First more conclusive evidence that these indications are sound, and that they also hold in a comparable way for intonation languages was recently provided by [16]. The study found that the spectral energy patterns of sibilant-like /t/ aspirations at the end of short German utterances varied so that the resulting sibilant pitch patterns supported and continued the preceding nuclear and phrase-final intonation contours. Moreover, it was shown that the different utterance-final /t/ aspirations were linked with attitudinal meaning profiles that match with the ones of the nuclear intonation contours.

The present study aims at continuing the line of research started by [16] on the basis of a detailed acoustic analysis that addresses further speech sounds and speakers of German. The selected three pairs of speech sounds, which constitute the research subject, have a particularly high potential to support and continue the intonation contour.

The first pair consists of the velar and postalveolar fricatives /x/ and /ʃ/. Both may be realized with a clear lip rounding in German, which is, however, not compulsory (i.e. distinctive) and can hence be adjusted to the intonational context. For example, in the case of /x/ the rounding is due to co-articulation, since the velar fricative is bound to preceding back, rounded vowels like [u], cf. [17]. Moreover, as long as the constriction in the oral cavity is formed by tongue blade or tongue body, respectively, /ʃ/ and particularly /x/ allow considerable variation in the exact place of articulation. The speech sounds of the second pair are [ə] and [ɐ]. Although they are vowel-like sounds, they have a phonologically different status than other vowels ([ɐ] is syllable-final allophone of /r/ in German, [17]). Therefore, they will be referred to as vocoids. The two vocoids are not in phonological contrast with other similar sounds, and since they are also known to have very context- and dialect-dependent qualities, they could be prone to vary according to the intonational context. Finally, /aʊ/ and /ɔʏ/ are included as the third pair of speech sounds, since diphthongs not only have the potential to vary with intonation in terms of sound qualities, but also in terms of movement patterns. That is, the diphthong transitions could be coordinated with the F0 transitions.

All six speech sounds are investigated in utterance-final position in the context of two different phrase-final "intonation segments", terminal-falling and high-rising contours. In these diametrically opposed contexts a systematic pitch-related variation in the production and the resulting acoustics of the speech sounds should show up very clearly. Specifically, the hypothesis is put forward that the acoustic representations of the speech sounds are adjusted to create

10.21437/Interspeech.2009-308

lower or higher pitches that match the phrase-final intonation segments. In the case of the fricatives, this hypothesis is tested by means of the centre-of-gravity (CoG) measure, i.e. the average of the spectral frequencies, which were multiplied (i.e. weighted) by their amplitudes. Although the CoG does not directly represent perceptual concepts like sibilant/spectral pitch, they are positively correlated with them. For the vocoids [ə] and [ɐ] the F1 and F2 frequencies are determined, aiming intrinsic F0 or spectral pitch, respectively. In the case of /aʊ/ and /ɔʏ/, only F2 measurements were taken, since the falling F1 was often difficult to distinguish from F0, particularly at the ends of the high-rising sentences. In terms of the hypothesis, CoG and F2 values should be higher, F1 values lower in the high-rising than in the terminal-falling contexts.

2. Method

The CoG measurements were taken on the basis of spectral slices in *Praat* (www.praat.org), resulting from a 30ms Hamming window (512 FFT points), which was shifted in steps of 7ms across the fricative. Frequencies between 1.5-15kHz were included. Then, the mean CoG and the CoG range were calculated for each fricative on the basis of its slice values. The whole procedure was done automatically by means of a praat script, which was written by Leonardo Lancia. F1 and F2 were measured manually in *Wave Surfer* (www.speech.kth.se). The values were determined by means of LPC spectra at three different points within the diphthongs and the vocoids: 20ms after the onset, in the centre, and 20ms before the offset. Finally, the durations of all sounds were measured.

The measurements were made in a corpus of natural-sounding German. It was labelled manually by the author with regard to the segmental boundaries of the six investigated speech sounds. Moreover, for each of the target words (see below) presence and type of pitch accent as well as the type of phrase-final intonation contour were labelled using the KIM-based PROLAB inventory, cf. [18].

The corpus itself was recorded with reference to the method developed by [19]. That is, short dialogue texts were created on everyday topics (e.g., a lost book, shopping at the market). The dialogues comprised about 4-5 turns per speaker with 3-5 sentences each. Some of the sentences in the turns of speaker (B) ended in the mono- or disyllabic target words, which contained the investigated speech sounds in final position. Each of the target words occurred twice in different sentences. These sentences as well as the surrounding ones created semantic-pragmatic context frames that were designed to elicit a pitch accent on the target word, leading to an either terminal-falling or high-rising sentence-final intonation. For example, the difference between terminal-falling and high-rising was linked with (concluding) statement or (surprised) question, respectively. Thus, it was not necessary to give the subjects any explicit instructions as to the intended prosodic properties or the mere presence of the target words and sentences; and the dialogue texts were long enough to prevent the target words and sentences from attracting attention. Instead, the subjects were only instructed to produce the dialogues in a natural way with a colloquial speaking style. In order to facilitate this task, they were given the possibility to modify the wordings of the sentences slightly according to their personal tastes, for example, by adding, omitting, or replacing words or expressions. Moreover, speaker (A) was always the author himself. He served as an example and led the subject (speaker B) to take up his natural, colloquial speaking style.

Each investigated speech sound was represented by two target words in order to increase the reliability of the findings. The following words were used. /ʃ/: *"Tisch","Fisch"* (table, fish); /x/: *"Buch","Tuch"* (book, cloth); [ə]: *"Tage","Schramme"* (days, scratch); [ɐ]: *"lecker","Bäcker"* (tasty, baker), /aʊ/: *"Bau", "rau"* (building, rough); /ɔʏ/: *"neu", "scheu"* (new, shy). The excerpt below shows how the target words (underlined) were embedded in the dialogues.

A: "*Hier, schau mal, ich hab' mir einen Infrarot-Zollstock gekauft.*" (Look, I bought myself an infrared yardstick.)
B: *"Hm...lass mal sehen...ich glaub', ich hab' so was schon mal gesehen. Sind das die vom <u>Bau</u>?"* (Hm…let me see…I think, I have seen something like this before. Aren't these the ones they use at building sites?)
A: *"Ja, richtig. Und sogar von vielen Architekten wird diese Technik inzwischen benutzt."* (Yes, right. And even many architects use this technology already).
B: *"Tja, die Technik hält ja bekanntlich überall Einzug."* (Well, as we all know, technology arrives everywhere.)
A: "...und macht das Leben komplizierter." (...and makes life more complicated.)
B: *"Wem sagst Du das. Allein die Anleitungen...so dick wie ein <u>Buch</u>."* (You're telling me. The manuals already...as thick as a book.)
A: *"Ja, das gilt für dieses kleine Wunderding hier ganz genauso."* (Yes, this also holds fort the little marvel here.)
B: *"Ist der denn <u>neu</u>?"* (Is it new then?)
A: *"Nee, den hab ich von Ebay. Neu sind die Dinger noch viel zu teuer. So oft brauch' ich ihn ja auch gar nicht. Und er ist tadellos! Bis auf die kleine Schramme hier."* (No, I bought it on Ebay. These things are much too expensive new. I also don't need it very often. And, it is flawless! Except for this little scratch here.)
B: *"Ich seh' gar keine <u>Schramme</u>."* (I don't see a scratch.)

Each dialogue was produced four times in a row, and the last two productions were used for the acoustic analysis. To date, 5 male native speakers of German (25-35 years old) have been recorded at phonetic facilities in Kiel (IPdS) and Berlin (ZAS). The recordings were made digitally at 48 kHz. Each session took about 30 minutes. The created speech corpus comprises currently around 100 minutes.

3. Results

Each of the target sounds was produced in the expected context, i.e. at the end of the sentence-final target word, which had a pitch accent, followed by the corresponding terminal-falling or high-rising intonations in the statements or questions, respectively. The two intonation conditions yielded sentence-final average F0 values of 72Hz and 241Hz. So, none of the productions had to be excluded from the analyzed sample, which hence consisted of 20 tokens for each combination of speech sound and intonation context. Based on these 20 tokens, individual t tests for paired samples were calculated for each speech sound (and point of measurement), comparing the mean CoG, F1, and F2 values that were found in the two intonational contexts. Regarding the large number of t tests, the significance threshold was set to $p \leq 0.01$ (at $df=19$). The results of the acoustic analysis are summarized in Table 1 and Figures 1-3. Complementary sound examples that illustrate the findings are available online at www.ipds.uni-kiel.de/on/.

As regards the voiceless fricatives /ʃ/ and /x/, the mean CoGs of were significantly higher and the CoG ranges were significantly smaller with a preceding high-rising than with a preceding terminal-falling intonation (cf. Tab.1). Both may clearly be seen in Figure 1. Perceptually, /ʃ/ and /x/ sound higher in combination with the high-rising intonation. In the case of [ɐ], the high-rising context led to significantly lower F1 and higher F2 values than the terminal-falling one. This holds for the measurements in the centre and towards the offset of [ɐ] (cf. Tab.1). In addition to the higher intrinsic F0 that may be assumed to follow from the lowered F1 (i.e. the less open quality), the greater F1-F2 difference, which is illustrated in Figure 2, shifts [ɐ] towards the perceptually lighter quality [ɛ̝]. Unlike [ɐ], no significant differences were found for the F1 values of /ə/. However, /ə/ showed a significant raise of the F2 offset value in connection with the high-rising intonation (Tab.1). This resulted in the impression of a diphthongized [əə̝] in many contexts.

Table 1. *Mean CoGs and averaged CoG ranges of the post-alveolar and velar fricatives (top) as well as mean F2 values measured after the onset (+20ms), in the centre, and before the offset (-20ms) of the two vocoids [ə] and [ɐ] (bottom) in the contexts of the terminal-falling (\) and the high-rising (/) sentence-final intonations (values are rounded to 50Hz). Significant dfferences between \ and / are marked by asterisks.*

n=20	/ʃ/	/x/		/ʃ/	/x/
mean CoG \	4.550*	2.400*	CoG range /	650*	800*
mean CoG /	5.300*	3.550*	CoG range \	1.150*	1.300*
	[ə]			[ɐ]		
n=20	onset	centre	offset	onset	centre	offset
mean F1 \	400	500	450	550	750*	850*
mean F1 /	450	450	500	450	500*	550*
mean F2 \	1.200	1.050	950*	1.600	1.250*	1.100*
mean F2 /	1.150	1.050	1.300*	1.650	1.550*	1.450*

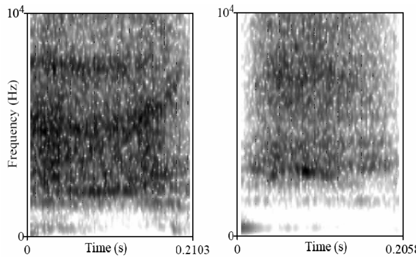

Figure 1. *Examples of two /ʃ/ spectrograms produced by the speaker RAI at the end of the target word "Fisch" with preceding terminal-falling (left) and high-rising (right) sentence-final intonation. In the latter condition the (darker) regions of high spectral energy are raised.*

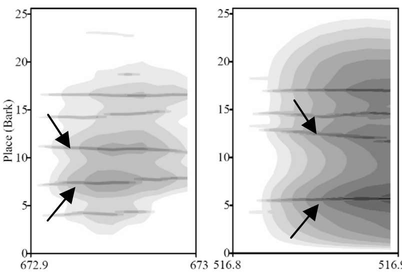

Figure 2. *Cochleargrams of two [ɐ] produced by the speaker MAT at the ends of "lecker" (tasty) in the contexts of terminal-falling (left) and high-rising (right) intonations. F1 and F2 are the dark lines around 5-7 and 10-13 Bark (cf. arrows). While F1 rises and F2 falls from the preceding /k/ in both cases, the formants are clearly pulled apart in the right one.*

In the case of the two diphthongs /aʊ/ and /ɔʏ/, the intonational context had no strong effect on the F2 values at onset and offset, although – analogous to [ə] and [ɐ] – F2 at the end of /ɔʏ/ was higher in the high-rising than in the terminal-falling intonation context. Primarily, however, it may be seen from Figure 3 that the intonational context affected the diphthong dynamics and hence the central point of measurement. In connection with the high-rising intonation, the F2 transition from the start to the target quality set in earlier and was faster (observations suggest that the same holds for F1). Perceptually, this results in a noticeable discontinuity and a highlighting of the lighter target quality. Both is absent in the diphthongs produced in the terminal-falling intonation context.

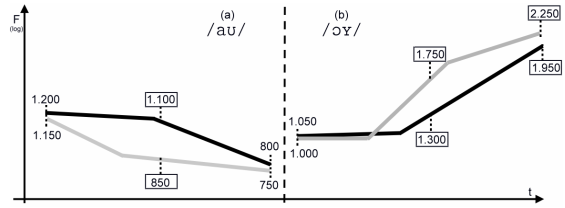

Figure 3. *Schematic representations of the characteristic F2 transitions observed around the mean values (rounded to 50Hz) after the onset (+20ms), in the centre, and before the offset (-20ms) of the diphthongs /aʊ/ (left) and /ɔʏ/ (right) in the contexts of terminal-falling (black) and high-rising (grey) sentence-final intonations. Significant differences between the means of the black and grey curves are put into boxes.*

4. Discussion

The hypothesis is supported by the results. Both voiced and voiceless sounds show differences in sound quality between the high-rising and terminal-falling intonation contexts. The quality differences raise or lower the perceived pitch of the sounds; either directly in terms of intrinsic F0, represented by an F1 change as in [ɐ] (lower F1 yields higher intrinic F0, [9,10]), or on an indirect spectral basis as represented by the F2 changes of [ə], [ɐ], /aʊ/, and /ɔʏ/, or the mean CoG changes of /ʃ/ and /x/ (higher F2 and CoG values cause higher pitch impressions, [11,12]). This phenomenon may represent *an adjustment of the "segmental intonation" to the coinciding intonation segments*. For /aʊ/ and /ɔʏ/, the adjustment also concerned the diphthong dynamics, i.e. the beginning and the abruptness of the F2 transition from the onset to the target quality. In this connection, it was striking that the beginning of the formant transition was closely coordinated with the beginning of F0 rise to the high intonation target. This suggests a coupling between articulatory and phonatory movement patterns, in line with findings of previous studies starting from [20,21]. An example for this coupling is provided by Figure 4.

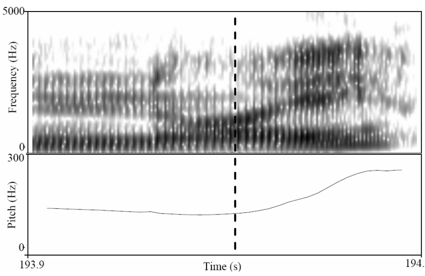

Figure 4. *Spetrogram (top) and F0 course (bottom) of a sentence-final "neu" (new, speaker TJA). The start of the high-rising F0 movement coincides with the start of the F2 transition from the onset to the target quality of the diphthong /ɔʏ/.*

From an articulatory angle, all observed changes in sound quality are likely related to changes in tongue position and/or lip rounding (fronting/raising of the tongue, spreading of the lips for higher pitch). For example, closer inspections of the /ʃ/ and /x/ spectra showed a rise of the energy maxima towards the end of the tokens that were preceded by the terminal-falling intonation. The changing spectra become similar to the (more stable) ones of the fricatives from the high-rising intonation condition (Fig.1). It may be assumed that the spectral change represents a spreading of the lips, which is already present from the very beginning in the fricatives of the high-rising sentences. Such a timing difference in the lip rounding during the fricative articulations in the terminal-falling and high-rising intonation contexts would also explain the significantly larger CoG ranges in the high-rising context. As already implied by this assumption, it is overall unlikely that the present findings are mere acoustic by-products of differences in larynx height, a known strategy to create F0 changes ([9,10], cf. also the not significantly different F2 values at the on- and offsets of /aʊ/, Fig.3, and the lack of F1 differences for [ə]).

The fact that there is segmental variation which is able to contribute to the pitch course of utterances might be one of the reasons why intonation appears to be *"subjectively continuous"* ([22]:275). Moreover, such variations suggest to revisit the F0-based concepts concerning the truncation and compression of intonation contours, cf. [23]. However, contributing to the pitch courses of utterances might not be the only source of the segmental variation. For example, the aspiration noises investigated by [16] did not only vary in the spectral energy distribution, but also in duration and intensity, which might only be indirectly associated with perceived pitch, cf. [24]. Also in the present study /aʊ/, /ɔʏ/, and [ɐ] were found to be significantly longer in the high-rising than in the terminal falling intonation context. Overall, such findings offer the additional or alternative interpretation that the variations in sound quality are (also) "*articulatory metaphors"* of meanings which have so far been associated with intonation alone. For instance, the meaning difference between 'surprise'/'disbelief' on the one and 'concluding'/'assuring' on the other hand is related to insecurity or self-confidence, respectively. Besides the intonational pitch(-accent) difference that has been associated with this meaning difference in German (cf. [2]), long, high-effort vs. short (and rounded) low-effort articulations could be a further means to express this difference in meaning, cf. the 'effort code' in [25]. If this alternative or additional interpretation holds, which would require us to revisit the concept of 'intonational meaning' ([25]), then the domain of the variations in sound quality should not be related to local pitch levels, but to larger meaningful units like the accented word. And in fact, first measurements of the F2 levels in the vowels /ɪ/ and /u/ that immediately precede the fricatives /ʃ/ and /x/ in *"Tisch", "Fisch"* and *"Buch", "Tuch"* did yield significant trends (i.e. $0.05 \leq p \leq 0.1$) towards higher values in connection with the high-rising intonation (t tests for paired samples).

However, this effect could also be a simple by-product of pitch-related sound changes in the fricatives. Particularly, if these changes are mainly due to a difference in lip rounding as assumed above, it is likely that this difference is already anticipated in the preceding vowels, cf. [26]. It must also be kept in mind that the (bundles of) parameter variations that characterized the speech sounds in the present study and in [16] seem to exist cross-linguistically in comparable pitch contexts, also in tone languages, cf. [14,15]. Moreover, they are used to convey pitch contours in whispered speech, [12]. Thus, the constant source of the variation seems to be pitch rather than the expression of particular meanings by articulatory metaphors. Yet, follow-up studies must account for the possibility of segmental sound changes as articulatory metaphors of meanings that are traditionally viewed as 'intonational'. This further-leading research question is a good example for the importance to approach speech from a perspective that accounts for meaning and the listener, and that integrates segmental and prosodic events. As for the present study, the next steps will be to test the perceptual relevance of the acoustic findings based on meaning-related judgements ([16]), and to extent the scope to non-final speech sounds that are surrounded by F0.

5. References

[1] Pierrehumbert, J., "The phonology and phonetics of English intonation", PhD Diss, MIT, Mass., USA.

[2] Kohler, K. J.," Prosody in speech synthesis: the interplay between basic research and TTS application", J.Phon, 19: 121-138, 1991.

[3] Kohler, K. J., "Macro and micro F0 in the synthesis of intonation", In J. Kingston and M. E. Beckman [Eds], Papers in Laboratory Phonology I, 115-138, Cambridge Univ. Press, 1990.

[4] Fowler, C.A. and Brown, J.M., "Intrinsic F0 differences in spoken and sung vowels and their perception by listeners", Perception & Psychophysics, 59: 729-738, 1997.

[5] Niebuhr, O., "The role of the accented-vowel onset in the perception of German early and medial peaks", Proc. 3rd Speech Prosody Conference, Dresden, Germany: 109-112, 2006.

[6] Niebuhr, O., "Categorical perception in intonation: a matter of signal dynamics?", Proc. Interspeech 2007, Antwerp, Belgium: 642-645, 2007.

[7] Pfitzinger, H., Mixdorff, H., and Peters, B., "Correspondences between KIM-based symbolic prosodic labels and parameters of the Fujisaki model", Proc. 10th Nordic Prosody Conference, Helsinki, Finland, 2009.

[8] Santen, J. van and Hirschberg, J., "Segmental effects on timing and height of pitch contours", Proc. ISLP, Yokohama, Japan: 719-722, 1994.

[9] Stoll, G., "Pitch of vowels: experimental and theoretical investigation of its dependence on vowel quality", Speech Communication, 3: 137-150, 1984.

[10] Niebuhr, O., "Intrinsic pitch in opening and closing diphthongs of German", Proc. 2nd Speech Prosody, Japan: 733-736, 2004.

[11] Higashikawa, M. and Minifie, F.D., "Acoustical-Perceptual Correlates of 'Whisper Pitch' in Synthetically Generated Vowels", J. Speech Lang. Hear. Res., 42: 583-591, 1999.

[12] Traunmüller, H., "Some aspects of the sound of speech sounds," In M.E.H. Schouten [Ed], The Psychophysics of Speech Perception, 293-305, Dordrecht, 1987.

[13] Hirschberg, J. and Ward, G., "The influence of pitch range, duration, amplitude and spectral features on the interpretation of the rise-fall-rise intonation contour in English", J.Phon, 20: 241-251, 1992.

[14] Kong, Y.Y., and Zeng, F.-G., "Temporal and spectral cues in Mandarin tone recognition", JASA, 120: 2830-2840, 2006.

[15] Whalen, D.H. and Xu, Y., "Information for Mandarin tones in the amplitude contour and in brief segments," Phonetica, 49: 25-47, 1992.

[16] Niebuhr, O., "The coding of intonational meaning beyond F0", JASA, 124: 1251-1263, 2008.

[17] Kohler, K.J., "Einführung in die Phonetik des Deutschen", Berlin: Erich Schmidt, 1995.

[18] Kohler, K.J., "Modelling prosody in spontaneous speech", In Y. Sagisaka, N. Campbell, and H. Higuchi [Eds], Computing prosody. Computational models for processing spontaneous speech, 187-210, Springer, 1997.

[19] Kohler, K. J. and Niebuhr, O., "The phonetics of emphasis", Proc. 16th ICPhS, Saarbrücken, Germany: 2145-2148, 2007.

[20] D'Imperio, M., Nguyen, N., and Munhall, K.G., "An articulatory hypothesis for the alignment of tonal targets in Italian", Proc. 15th ICPhS, Barcelona, Spain: 253-256, 2003.

[21] Xu, Y. and Liu, F., "Determining the temporal interval of segments with the help of F0 contours", JPhon, 35: 398-420, 2007.

[22] Jones, D., "An outline of English phonetics", Cam.: Heffer, 1918

[23] Grabe, E.., "Comparative intonational phonology: English and German", PhD thesis, Univ. Nijmegen, The Netherlands, 1998.

[24] Pike, K.L., "Tone Languages", Ann Arbor: Edwards, 1948.

[25] Gussenhoven, C., "Intonation and interpretation: phonetics and phonology", Proc. 1st Speech Prosody Conference, Aix-en-Provence, France: 47-57, 2002.

[26] Soli, S., "Second formants in fricatives: acoustic consequences of fricative-vowel coarticulation", JASA, 70: 976-984, 1981.