INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Three Dimensions of Sentence Prosody and their (Non-)Interactions

Michael Wagner, Michael McAuliffe

McGill University, Canada

chael@mcgill.ca, michael.mcauliffe@mail.mcgill.ca

Abstract

Prosody simultaneously encodes different kinds of information, including the type of speech act of an utterance (e.g., falling declarative vs. rising interrogative intonational tunes), the location of semantic focus (via prosodic prominence), and syntactic constituent structure (via prosodic phrasing). The syntactic/semantic functional dimensions (speech act, focus, constituency) are orthogonal to each other, but to which extent their prosodic correlates (tune, prominence, phrasing) are remains controversial. This paper takes a 'bottom up' approach to test for interactions, and reports evidence that contrary to many current theories of sentence intonation, the cues to the three dimensions are often orthogonal where interactions are predicted.

Index Terms: intonation, focus, pitch scaling, phrasing, tunes, question intonation

1. Introduction

One idea of how different syntactic/semantic factors affect prosody is that they each independently influence the overall prosodic pattern of an utterance but do not interact with each other, as is assumed in so-called 'overlay' models [1, 2, 3, 4]. For example, with such an account, the choice of lexical pitch accent on a word and the choice of intonational tune on a sentence can each independently affect the intonational contour of the utterance [1], and [4] argues focus to be an additional function that independently affects the F0 contour.

A very different conception of how these factors combine to shape the overall prosody of an utterance is assumed in most phonological models of sentence prosody, e.g. [5, 6, 7], and embodied in the ToBI transcription system [8, 9]. In these models, the effects of different factors are mediated by a shared auto-segmental phonological representation (the 'AM-Model', cf. [10]). Such phonological models of sentence prosody embody specific assumptions about **interactions** between different factors influencing prosody. For example, many studies report that post-focal material remains unaccented in English. According to ToBI and other AM-models, phrasing distinctions should not be possible within such unaccented stretches, since "[...] there must be at least one pitch accent somewhere in every (prosodic) phrase [...]" [11]. Hence in the unaccented, post-focal domain phrasing distinctions should be neutralized. An earlier test of this prediction [12], however, found durational cues for phrasing in the post-focal part of sentences. We were interested to see whether F0 cues to phrasing are indeed absent in the post-focal domain, and more generally whether prominence and phrasing really interact in the way many current theories predict.

Prosodic phrasing, which in turn reflects reflects syntactic bracketing, affects the scaling of pitch accents. This is often attributed to an adjustment of a reference line depending on prosodic phrasing, relative to which tonal targets are scaled [13]. There have been varying accounts of F0 scaling in terms of such a reference or register line [14, 15, 16, 17, 18, 19]. To our knowledge the question of whether pitch scaling is used to convey prosodic phrasing in the post-focal domain has not been explored in the prior literature, although there are studies that suggest that the often-made claim that there are no tonal targets in the postfocal domain might not be accurate [4, 20]. The way phrasing encodes bracketing might also interact with the choice of intonational tune: Questions with a rising intonation are said to typically involve L^* instead of H^* accents. The question whether and how pitch scaling is used in questions has also not been previously explored as far as we know.

Another type of interaction that is expected by many accounts of phrasal phonology is that the F0 reflexes of focus should vary depending on the choice of pitch accent. If the prosodic effects of focus are reflect hyper-articulation compared to the unfocused rendition of a word due to increased prominence [21], then we might expect an increase in F0 under focus for declaratives (which involve H^* accents) and a decrease in F0 in questions (which involve an L^* accent) [22]. Similar predictions may follow from accounts that view the effect of focus on F0 as being an adjustment of F0 range, such that the range is widened for focused constituents and narrowed for post-focal material [4].

The present study reports first results from a study that attempts to look for the acoustic effects of different functional dimensions and their interactions in a 'bottom up' way. By crossing the dimensions (speech act, focus, constituency) in a factorial design we are able to test their phonetic import on the signal is independent of each other or whether they interact. Of particular interest for the present study is whether the contribution of all three functional dimensions can be successfully retrieved from specific acoustic parameters.

2. Methods

Participants first read the target sentences silently, and were asked to then 'say the sentences as naturally as possible, as if you were saying them to a friend in an everyday conversation.' They were not aware that our main interest was the intonation of the sentences. Each utterance consisted of two parts, a set-up sentences, and a second target sentence. We crossed three factors: type of speech act, syntactic constituent structure, and focus.

For our manipulation of constituent structure we used coordinated names [14, 23, 16, 17, 19]. The intended bracketing was indicated by the placement of commas, (1) illustrates an example with left-branching ([AB]C) and (2) illustrates right-branching (A[BC]).

(1) Declarative, Focus on Conjunct B, Left-Branching:
You said that Megan and Dillon, or Morgan would help. But in fact we were told that Megan$_A$ and Lauren$_B$, or Morgan$_D$ would help.

(2) Declarative, Focus on Conjunct B, Right-Branching

http://dx.doi.org/10.21437/Interspeech.2017-1500

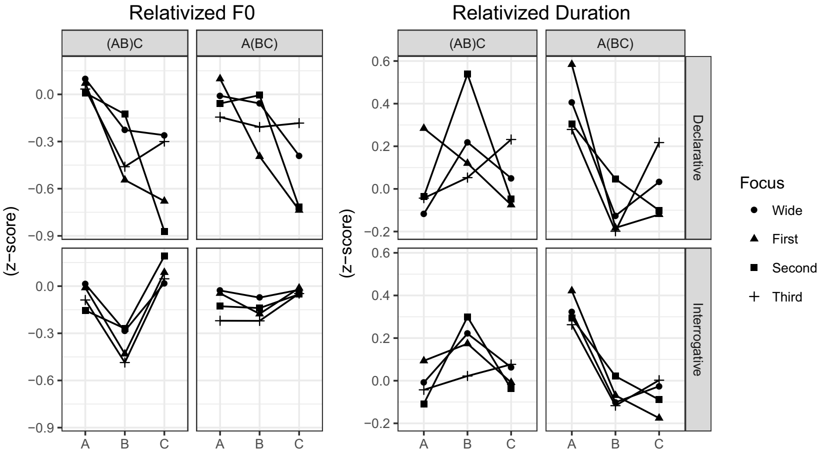

Figure 1: *Relativized F0 and duration measures of the three target words (A, B, C) by tune (declarative vs. interrogative) and constituency and focus. Note that y-axes are not shared between the F0 and duration plots.*

> You said that Megan, and Dillon or Morgan would help. But in fact we were told that Megan$_A$, and Lauren$_B$ or Morgan$_C$ would help.

We varied the type of speech act simply by using a period at the end of the target utterance (1–2), indicating that it was intended as an assertion, or a question mark (3), indicating that it was intended as an interrogative. We expected participants would produce a rising question intonation in the latter case:

(3) Interrogative, Wide Focus, Left-Branching
You said that Dillon would help. But now it turns out that Megan$_A$ and Lauren$_B$, or Morgan$_C$ would help?

For the manipulation of focus, we varied which parts of the first part of the utterance contrasted with the material in the second part. For example, in (1) and (2) the second part contrasts with respect to the choice of the second name ('focus on B'), while (3) the entire coordinate structure is the focus and contrasts with a single name in prior the set-up ('wide focus').

In total, the experiment involved two different types of speech acts (question vs. declarative), four focus conditions (focus on conjunct A, B, or C, or wide), and a manipulation of 2 phrasings (A[BC] vs. [AB]C), for a total of 16 conditions. The experiments involved four different sets of 16 sentences varying in lexical materials (which involved different sets of names, and used slightly different wordings in both setup and target), so every participant was recorded on 64 sentences. The manipulation of speech act was done between two experimental sessions of 32 trials each. About half were first run on the questions, and half were first run on the declaratives. In each session trial order was pseudo-randomized maximizing the distance between repeated conditions and repeated trials from the same item set. A total of 26 native speakers of North American English participated.

The recorded sentences were manually checked for speech errors, disfluencies, and hesitations, as well as recording errors. Only fluent utterances were kept. This step resulted in the exclusion of 24% of the data, which were more or less evenly distributed across the various conditions. The high exclusion rate is likely due to the length and complexity of the stimuli.

The materials were force aligned using the Montreal Forced Aligner [24] using models trained on LibriSpeech [25]. The aligned dataset was loaded into PolyglotDB [26], and F0 was calculated for each file using Praat [27][1].

To reduce the considerable differences in F0 between speakers and to control for any vowel-intrinsic effects on F0, we calculated a relativized measure of F0 in PolyglotDB. This relativized measure was a z-score of the F0 using per-speaker, per-segment means and standard deviations. The summary statistics were calculated based on all segments in the corpus, including those in words outside the target words.

Once relativized measures were calculated, measures for each of the target words (positions A, B, and C) were averaged per word. Relativized F0 was the mean z-score of F0 across the word, and relativized duration was the mean z-score of each segment's duration. These measures thus give a sense of, overall, for a given word, how much higher/lower the F0 or shorter/longer the duration is compared to the average production for that speaker. In addition to the relativized measures, each target word's position, orthography, speaker, focus condition (Wide, First, Second, or Third), tune (declarative or interrogative), and phrasing (A[BC] or [AB]C) were coded based on the sentence prompt.

In order to check whether our manipulations were successful, two RAs hand-annotated the data. They were asked to decide whether the intonation was falling or rising, what the bracketing of the coordinate structure was, and which of the 4 focus options was intended. For each dimension, they could also choose 'unclear' if they were not sure. Inter-annotator agreement was 'almost perfect' for the annotation of intonation (Cohen's kappa: 0.96), 'substantial' for constituency (Cohen's kappa: 0.73) and for prominence (Cohen's kappa: 0.63). Based on one annotator, the intonational tune (rising vs. falling) was as expected given our manipulation 96.3% of the time, the expected bracketing 61% of the time (with about one third of soundfiles marked as 'unclear'), and the correct prominence 38% of the time (with a large rate of confusion of wide vs. focus on C, and 21% of utterances marked as 'unclear'). The second annotator was slightly less accurate, but showed a similar pattern. We included all data, since we did not want to bias the

[1] Minimum and maximum F0 were 75 Hz and 300 Hz, respectively

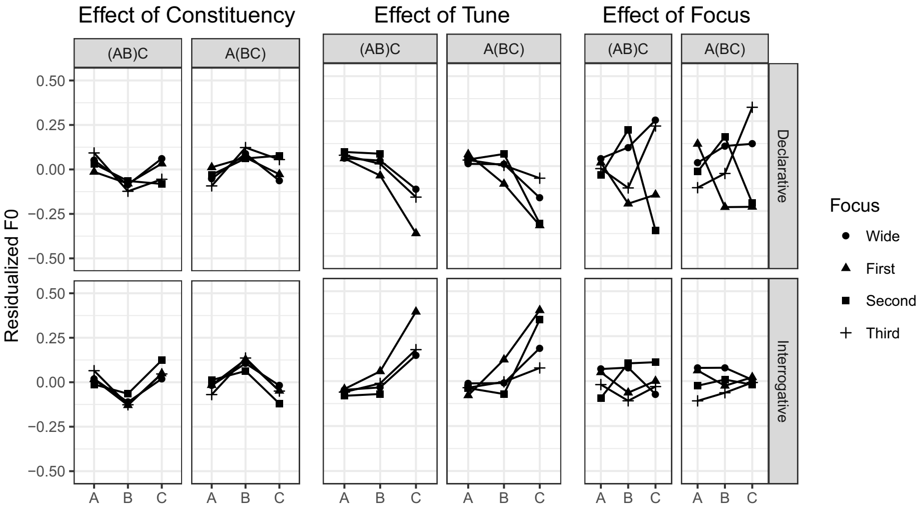

Figure 2: *Residual effects for the mean F0 over the word for each of the three functional dimensions.*

results based on prior expectations.

For the analysis reported here, we decided to collect a single measure of F0 and duration for each of the three words—the mean F0 of the word and the total duration. We took this coarse-grained approach in part because we wanted to stack the cards against the hypothesis that all three functional dimensions will be recoverable from the signal. Different dimensions are very likely to affect different parts of a word relatively more than others. For example, focus will exert its main influence on the stressed syllable, while durational effects of phrasing will be found mostly on the last syllable [28]. Looking at coarse-grained mean measures can serve as a proof-of-concept in that it is all the more surprising if we can disentangle different effects based on those, even if eventually more fine-grained analyses should also be applied to the data.

3. Results

We first look at the data based on the relativized measures. The relativization due segmental content and speaker allows for direct comparison across items and speakers, but will not help to tease apart the different dimensions we are interested in. Fig. 1 illustrates that there is substantial variation in how certain dimensions are conveyed depending on the others. For example, it is hard to see any systematic pattern of pitch scaling that correlates with phrasing that would be shared by the different focus conditions and/or shared across the two intonational contours.

Given the factorial design we can look at the acoustic effects of a particular functional dimension by residualizing out the effects of all other dimensions. In order to create such residualized measures we fitted mixed-effects regression models which included the other factors as well as the position in the sentence (A, B, or C), and all interactions of these main effects. In addition, they included random effects for item set and participant (but no random slopes). For example, the residualized F0 measure we used to look for the effects of constituent structure were computed based on a model that had Focus, Tune, Position and their interactions as fixed effects. Fig. 2 shows the residualized F0 measures for each dimension, and Fig. 3 the duration measure.

When looking at the effect of constituency on F0 (the leftmost two columns in Fig. 2), the effect is remarkably consistent. Pitch scaling seems to work in a uniform way independent of intonational tune (top two panels are very similar to bottom two), and independent of focus (the four lines with each facet reflecting the focus manipulation look remarkably similar to each other).[2] Interestingly, the effect of phrasing on F0 is the same independent of tune, suggesting that both L* and H* are scaled higher or lower under the same circumstances, which speaks against an interpretation of hyperarticulation of phrase-final constituents, and is compatible with current accounts in terms of pitch register adjustment.[3]

The effect of intonational tune is also remarkably similar across focus conditions and across phrasings. The only apparent interaction is that the final fall/rise is more extreme when C is a separate constituent compared to when it forms a constituent with B and is more embedded.[4]

The effect of focus, however, is affected by the other dimensions. While constituency has some influence on the focus effects, it leaves most of the qualitative differences intact. For example, in the declarative intonation, the first constituent is realized with much higher F0 and the following ones with lower F0 compared to the control condition, irrespective of phrasing—even if the exact pattern is slightly different. Intonational tune has a much bigger influence on the focus effect and the interrogative contour diminishes it. It is clear though that even in the interrogative tune focus correlates with high F0, despite of the

[2] One exception seems to be that in the declarative contour, pitch scaling looks qualitatively different when the second word in the coordinate structure (constituent B) is focused. In this case, the F0 on the final word simply seems to copy the F0 level of the previous word, independent of phrasing.

[3] We note, however, that another possibility: The F0 effects of phrasing could be a passive reflex of speakers adjusting intensity for rhythmic reasons. Speakers raise F0 when aiming to talk louder for about a half semi-tone per db [29]. We do not have the space to discuss intensity and its relation with F0 here.

[4] Judging by the durational measures for phrasing in Fig. 1 and Fig. 3, this is not just a function of the length of the last constituent—it's length is rather unaffected by phrasing.

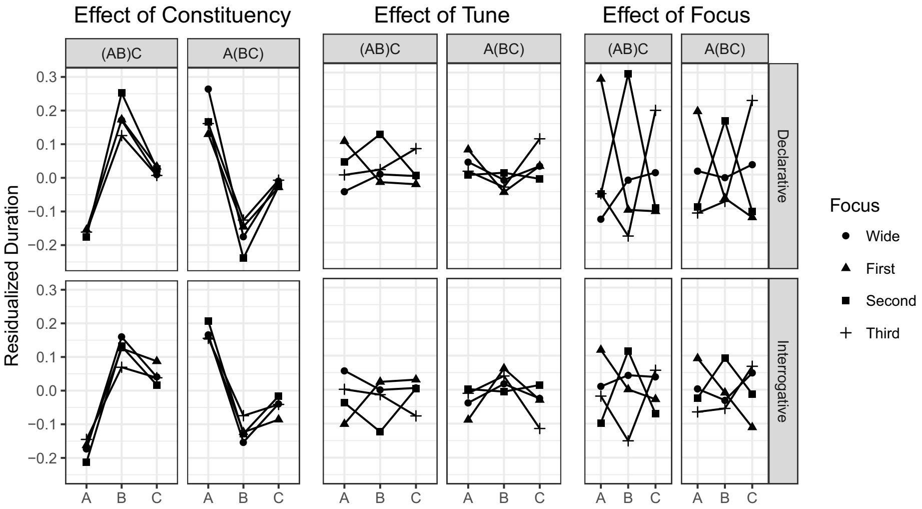

Figure 3: *Residual effects for the word duration measure for each of the three functional dimensions.*

fact that the accents involved are likely to be L* rather than H*. The hyperarticulation model of focus would predict the opposite pattern. However, since we did not qualitatively label the accents, we might be missing effects if speaker sometimes used H* and sometimes L* accents.

When looking at duration (Fig. 3), constituency again shows remarkably systematic effects across all conditions. The first constituent (A) is shortened while the second one (B) is lengthened for the left-branching structure ([AB]C), while the opposite is true for the right-branching structure (A[BC]). The effect of tune, however, seems less systematic. Indeed, tunes have not been linked to cause durational effects in the prior literature, so this is not surprising. The focal effects, on the hand, are vary parallel across the different intonational tunes and phrasings, even though the effect size seems lower for the intonational tunes. In other words, duration reliably seems to encode phrasing and focus, but not tune.

In order to further evaluate the data, we fitted regression models for each acoustic measure. We ran models for each position (A, B, C) on the relativized measures and added all fixed factors including position and their interactions, plus random effects. As an example, we report on a model for the mean F0 measure of the second conjunct (word B). Our model confirms that although non-residualized F0 does not look like a consistent cue to phrasing (Fig. 1), with very different patterns across tunes and foci, mixed effects regression model shows there is a significant effect of phrasing ($\beta = -0.19; SE = 0.06; t = -2.9$). The interactions with Focus predicted by AM-theory (e.g., since there should be no phrasing cues in the postfocal domain), on the other hand, did not come out significant ($\beta = -0.02; SE = 0.08; t = -0.2$). Additional models show a significant effect of phrasing on F0 scaling even (i) within the subset only involving utterances with focus on the first NP; and (ii), within the subset including only questions.

4. Discussion

Overlay models predict the prosodic reflexes of the three separate dimensions to be independent and recoverable from the signal, while most current phonological models predict interactions resulting in neutralizations between conditions. We used residualized measures of F0 and duration to explore whether the dimensions can be restored once variability due to other factors (e.g. Focus, Intonation, Position in the sentence, in the case of phrasing) is accounted for, and to look for interactions.

The finding that duration and F0 reliably encode phrasing across all focus conditions runs counter to the claim that focus causes postfocal phrasing distinctions to be lost, and suggests that focus and scaling for the most part are encoded independently of each other. The significant effect of cues where some AM-theories predict neutralization and the qualitative uniformity of the patterns where AM models predict interactions lend support to some aspects of overlay models. There is evidence for a sequential organization of intonation ([10], pace the overlay model in [2]) and for a mediation of phonetic reflexes by a phonological representation ([10], pace more recent sequential overlay models as in [3, 4]). But even such a sequential phonological approach would be compatible in principle with representations that keep the three dimensions investigated here more orthogonal than current AM-theories do. Our results motivate a reconsideration of some of our assumptions about how the three dimensions interact, but also suggest that a pure overlay model would miss potentially important interactions.

Future analyses of this data set should look at the entire signal rather than just a very sparse stylized compression into mean values for certain words of interest, and should also try to distill out the effect of position (for declaratives a declination across the utterance would be expected). It would also be important to explore whether the dimensions are recoverable without minimal pairs (unlike in our factorial design), which could be useful for syntactic and semantic parsing.

5. Acknowledgements

This project was funded by SSHRC grant 435-2014-1504 and funding through the Canada Research Chair program. Thanks to the members of Prosodylab, in particular Erin Olson and Thea Knowles for their help with data analysis.

6. References

[1] S. Öhmann, "Word and sentence intonation: A quantitative model," *KTH Department of Speech, Music and Hearing. Quarterly Status Report*, 1967.

[2] H. Fujisaki, "Dynamic characteristics of voice fundamental frequency in speech and singing. acoustical analysis and physiological interpretations," *KTH Department of Speech, Music and Hearing. Quarterly Status Report*, 1981.

[3] B. Möbius, *Ein quantitatives Modell der deutschen Intonation: Analyse und Synthese von Grundfrequenzverläufen*. Walter de Gruyter, 1993.

[4] Y. Xu, "Speech melody as articulatorily implemented communicative functions," *Speech Communication*, vol. 46, pp. 220–251, 2005.

[5] J. Pierrehumbert, "The phonology and phonetics of English intonation," Ph.D. dissertation, MIT, September 1980.

[6] E. O. Selkirk, *Phonology and Syntax. The relation between sound and structure*. Cambridge, MA: MIT Press, 1984.

[7] ——, "Sentence prosody: Intonation, stress, and phrasing," in *Handbook of Phonological Theory*, J. A. Goldsmith, Ed. London: Blackwell, 1995, ch. 16, pp. 550–569.

[8] K. Silverman, M. Beckman, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg, "ToBI: A standard for labeling English prosody," in *Proceedings of the 1992 international conference conference of spoken language processing*, vol. 2, 1992, pp. 867–870.

[9] M. E. Beckman, J. Hirschberg, and S. Shattuck-Hufnagel, "The orginal ToBI system and the evolution of the tobi framework," in *Prosodic models and transcription: Towards prosodic typology*, S.-A. Jun, Ed. Oxford: Oxford University Press, 2005, pp. 9—54.

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

[11] M. Beckman, "The Parsing of Prosody," *Language and Cognitive Processes*, vol. 19, no. 11, pp. 17–67, 1996.

[12] E. Norcliffe and T. Jaeger, "Accent-free prosodic phrases? Accents and phrasing in the post-nuclear domain," in *Proceedings of Interspeech 2005*, 2005.

[13] D. R. Ladd, "Declination and 'reset' and the hierarchical organization of utterances," *JASA*, vol. 84, pp. 530–544, 1988.

[14] R. van den Berg, C. Gussenhoven, and T. Rietveld, "Downstep in Dutch: implications for a model," in *Papers in Laboratory Phonology, vol. II: Gesture, segment, prosody*, G. Docherty and R. Ladd, Eds. Cambridge: Cambridge University Press, 1992, pp. 335–58.

[15] C. Féry and H. Truckenbrodt, "Sisterhood and Tonal Scaling," *Studia Linguistica*, vol. 59, no. 2-3, pp. 223–243, 2005.

[16] C. Féry and G. Kentner, "The prosody of embedded coordinations in German and Hindi," in *Proceedings of Speech Prosody*, vol. 5, 2010.

[17] G. Kentner and C. Féry, "A new approach to prosodic grouping," *The Linguistic Review*, vol. 30, no. 2, pp. 277–311, 2013.

[18] H. Truckenbrodt and C. Féry, "Hierarchical organisation and tonal scaling," *Phonology*, vol. 32, no. 01, pp. 19–47, 2015.

[19] C. Petrone, H. Truckenbrodt, C. Wellmann, J. Holzgrefe-Lang, I. Wartenburger, and B. Höhle, "Prosodic boundary cues in german: Evidence from the production and perception of bracketed lists," *Journal of Phonetics*, vol. 61, pp. 71–92, 2017.

[20] C. Féry and F. Kügler, "Postfocal downstep in German," *Language and Speech (Online First)*, 2016.

[21] M. Breen, E. Fedorenko, M. Wagner, and E. Gibson, "Acoustic correlates of information structure," *Language and Cognitive Processes*, vol. 25, no. 7, pp. 1044–1098, 2010.

[22] J. Howell, "Focus placement on adjacent words in yes/no questions," in *Proceedings of the 18th International Congress of Phonetic Sciences (ICPHS) in Glasgow*, 2015.

[23] M. Wagner, "Prosody and recursion," Ph.D. dissertation, MIT, 2005.

[24] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal Forced Aligner [computer program]," https://montrealcorpustools.github.io/Montreal-Forced-Aligner/, 2017.

[25] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *Proc. ICASSP 2015*, 2015, pp. 5206–5210.

[26] M. McAuliffe, E. Stengel-Eskin, M. Socolof, and M. Sonderegger, "PolyglotDB [software package]," https://github.com/MontrealCorpusTools/PolyglotDB, 2017.

[27] P. Boersma and D. Weenink, "PRAAT, a system for doing phonetics by computer. report 132." 1996, institute of Phonetic Sciences of the University of Amsterdam.

[28] A. Turk and L. White, "Structural influences on accentual lengthening in English," *Journal of Phonetics*, vol. 27, no. 2, pp. 171–206, 1999.

[29] P. Gramming, J. Sundberg, S. Ternström, R. Leanderson, and W. H. Perkins, "Relationship between changes in voice pitch and loudness," *Journal of Voice*, vol. 2, no. 2, pp. 118–126, 1988.