Using amplitude envelope modulation spectra to capture differences between rhetorical and information-seeking questions

Friederike Hohl, Bettina Braun

Department of Linguistics, University of Konstanz, Germany
{friederike.hohl, bettina.braun}@uni-konstanz.de

Abstract

Rhetorical questions (RQs) and information-seeking questions (ISQs) differ in their pragmatic function (the first making a point, the second requesting information). They may have identical surface forms. For human-computer interaction, but also for interaction between humans, it is important to decode the intended function. Laboratory experiments have established that RQs are longer, more often realized with breathy voice quality than ISQs and differ in intonational realization. However, annotation is labor-intensive. Here we test whether the prosodic differences between RQs and ISQs are evident in the amplitude envelope modulation spectra. These capture the slow-changing energy distribution over utterances and do not demand manual annotation. Since amplitude envelope modulation spectra are sensitive to rhythmic differences between languages, they may be well-suited to capture the duration differences. We compare RQs and ISQs in three closely-related languages (English, German, Icelandic) to investigate whether RQs have different amplitude envelope modulation spectra than ISQs and whether these differences are language-specific. The results show differences between RQs and ISQs but, depending on language, in different frequency bands. We show that the differences cannot be explained by durational differences between RQs and ISQs alone, but that the amplitude envelopes capture the signal more holistically.

Keywords: question, prosody, amplitude envelopes, cross-linguistic, general additive mixed models

1. Introduction

The analysis of amplitude envelopes has become a widely used method in the speech sciences, language acquisition and neurolinguistics (Frota et al., 2022; Gross et al., 2013; Leong & Goswami, 2015; Poeppel, 2014; Poeppel & Assaneo, 2020). Amplitude envelopes track the amplitude distribution over an utterance and hence represent the part of the signal that is relevant to convey rhythm (Arvaniti, 2009). Furthermore, the method is easy to apply without demanding manual annotation (cf. Gibbon, 2021 for discussion of advantages of modulation-theoretic methods). Despite the increasingly wide-spread usage across disciplines, there is little research on *which aspects* of the speech signal influence the amplitude envelopes in *what way*. Cross-linguistic research has shown that a stress-timed language (German) led to lower power around 2Hz and to higher power between 7 and 10Hz than more syllable-timed Brazilian Portuguese (Frota et al., 2022), cf. Tilsen & Arvaniti (2013). Others have explored the use of amplitude envelopes for differences in speech style (Gibbon, 2021). In this paper, we test whether amplitude envelope modulation spectra can distinguish also between rhetorical vs. information-seeking questions.

Rhetorical questions do not seek information from the addressee but serve to make a make a point and commit the interlocutor to the presupposition expressed by the RQ. For instance, the questions in (1), uttered as rhetorical questions, attempt to commit the interlocutor to the statement that nobody likes phonetics (Biezma & Rawlins, 2017)

(1)	Who likes phonetics?	wh-question
	Does anyone like phonetics?	polar question

Since the questions in (1) can also be uttered to seek information (e.g. to find a suitable student assistant), it is sometimes only the prosodic realization that can help disambiguate between the two meanings. This disambiguation is not only important for human communication, but also for human-computer interaction and sentiment analysis.

Previous production data have shown consistent differences between RQs and ISQs across intonation languages such as English (Dehé & Braun, 2019), German (Braun et al., 2019, 2020), and Icelandic (Dehé et al., 2018), cf. Dehé et al., (2022) for an overview: In all of these intonation languages, RQs have longer constituent durations. Less consistent is the greater use of non-modal voice quality (breathy, glottalized) in RQs compared to ISQs. Differences in the intonational realization are language-specific. English speakers more often produced the nuclear (last) accent on the subject pronoun 'anyone' in RQs (but not in ISQs), followed by a high plateau, while the nuclear accent was typically produced on the final noun in German and Icelandic. In Icelandic, the boundary tone was always falling in both RQ and ISQs. Icelandic speakers more often produced an early-rise in RQs (i.e. the rise started early in the final noun), but the difference was not strong. In German, speakers more frequently produced a prominent rising accent (L*+H) in RQs (compared to a low accent, L*, in ISQs). Using classification and regression trees, German questions could be classified as RQ or ISQ with an accuracy of 87.5% with these parameters (Braun et al., 2018).

However, manual annotation of prosody is cost-intensive. In this paper, we therefore test whether RQs and ISQs also differ in terms of amplitude envelopes. Amplitude envelopes capture the wideband energy distribution and therefore capture suprasegmental differences such as differences in duration or voice quality (resulting in lower energy in high frequency areas). Modulation frequencies can be extracted from the speech signal in a number of ways (Poeppel & Assaneo, 2020). Most procedures first filter the sound into a number of frequency bands (spaced either logarithmically or such that they are equidistant on the cochlea), typically in the range between 100 and 8,000Hz

(or 10,000Hz). These signals are then filtered to remove the high-frequency components, leaving frequencies in the range of 0 to approximately 10Hz. These narrowband envelopes are then summed and the modulation frequencies are derived by Fourier analysis. The result is a spectrum, i.e. power values across frequency. We then compare the patterns holistically, rather than extracting single parameters (Tilsen & Arvaniti, 2013).

2. Data

2.1. Methods

The data were collected in separate production experiments. For all three languages (English, German, Icelandic), participants saw a context description, which was constructed to trigger a rhetorical or information-seeking intention (illocution). They then produced a visually presented question so that it fit the respective context.

2.1.1. Participants

For English, 21 participants (mean age 22.5 years, 14 female, 7 male), for German 12 participants (mean age 21 years, SD = 2.3 years, 10 female, 2 male,) and for Icelandic, 32 participants (aged 20–65, 20 female and 12 male) took part in the data collection for a small fee. All participants gave informed consent.

2.1.2. Materials

We constructed 11 *wh*-interrogatives that fitted both a rhetorical and an information-seeking reading (e.g., *Who likes celery?*). To this end, we used predications that – out of context – may be true for some people and false for others (e.g., 'liking celery'). From these *wh*-interrogatives, we derived polar questions by replacing the *wh*-word by the indefinite pronominal subject *anyone* and adapted the syntactic structure to verb-first (V1).

For each polar question, we constructed two contexts, one triggering an information-seeking interpretation of the interrogative and one triggering a rhetorical one. An example of polar question contexts is given in Table 1. To control for information structure and specifically to avoid effects of information structure on nuclear accent position and type, each context introduced the predication expressed in the sentence radical (e.g., *liking celery* in Table 1), rendering the referents of the constituents in the verb phrase discourse-given (see Braun et al., 2019 for more details).

ISQ	**RQ**
You cooked a dish with celery. You would like to know whether your guests like this vegetable and will eat it or not. You say to your guests:	In the canteen they have casserole with celery on the menu. However, you know that nobody likes this disgusting vegetable. You say to your friends:
'Does anyone like celery?'	

Table 1. Example contexts for information-seeking (ISQ, left) and rhetorical questions (RQ, right)).

The rhetorical contexts contained a sentence stating that it is generally known (or that the speaker knows) that nobody agrees with a certain proposition (e.g., *you know that nobody likes celery*). The information-seeking contexts differed from the rhetorical contexts in that they stated that the speaker was looking for some piece of information.

Additionally, 24 fillers with different syntactic structures were added to reduce awareness of the experimental manipulation. The materials were first designed for German, and then translated into English and Icelandic, with minor adaptions to account for cultural and phonological differences.

2.1.3. Procedure

Recording. Each participant produced both the rhetorical and the information-seeking version of each target interrogative in randomized order. Each experiment started with four familiarization trials, followed by a short break in which participants were allowed to ask questions if anything was unclear. The experiment was controlled using the experimental software *Presentation* (Neurobehavioral-Systems, 2000). Each trial started with the visual display of the context, which the participant had to read silently, followed – upon button press – by the target interrogative on the next screen. The target sentence had to be produced aloud and was recorded onto disk (44,100Hz, 16Bit).

Extraction of amplitude envelope modulation spectra. All productions (N = 1000) were cut at utterance start and end. Average durations across conditions are shown in Table 2 and show lengthening of RQs as compared to ISQs.

Language	RQ	ISQ	Proportional lengthening of RQs
English	1.456	1.311	11.1%
German	1.551	1.330	16.7%
Icelandic	1.419	1.140	24.5%

Table 2. Average durations across languages (rows) and illocution types (columns) in seconds, including the proportional durational increase from ISQ to RQ.

Amplitude envelopes for all questions were extracted, following the descriptions in the literature (Chandrasekaran et al., 2009; Frota et al., 2022; Gross et al., 2013). First, we calculated the narrowband amplitude envelopes (He & Dellwo, 2016, 2017). The speech signal was first down-sampled to 22,050Hz and then filtered into nine frequency bands in the range from 100–10,000Hz, which are equidistant on the cochlear map (Gross et al., 2013). The cutoff frequencies were 100.5Hz, 250.7Hz, 458.6Hz, 748.8Hz, 1159.0Hz, 1449.0Hz, 2619.8Hz, 3954.2Hz, 6121.8Hz and 10000.8Hz. To remove high-frequency components, the signals were low-pass filtered (Hann filter between 0 and 10Hz with 1 Hz smoothing). The resulting narrowband envelopes were then added to compute the wideband amplitude envelope. These were spectrally analyzed in 100 0.1Hz steps. This approach is conceptually similar to approaches that do not compute narrowband envelopes (Gibbon, 2021; Tilsen & Johnson, 2008). The wideband envelope was spectrally analyzed in 100 0.1Hz steps (fast Fourier transform). All signal processing was done in Praat (Boersma & Weenink, 2018).

Statistical modeling. To model the effect of *language* and *illocution type* across frequency bands, we used generalized additive mixed models, GAMMs (Wieling et al., 2012; Wood, 2006, 2015; Wood & Saefken, 2016; Zahner et al.,

2019). They are well-suited to pinpoint in which frequency bands differences occur; taking into account non-linear relationships and auto-correlation. The response variable was log-normalized power. We modelled non-linear dependencies of *language* and *illocution type* first as separate smooth terms (e.g., s(fband_Hz, by = language, bs='tp', k = 20)). These smooth functions include a pre-specified number of base functions of different shapes, e.g., linear and parabolic functions of different complexity. The two factors *language* and *illouction type* were further added as parametric effects. Smooths for *speakers* (random intercept and over frequency bands) were also included (s(speaker, fband_Hz, by='re'). For model fitting, we employed the R package *mgcv* (Wood, 2015). The model was corrected for auto-correlation in the data using a correlation parameter, determined by the acf_resid() function. We use the function gam.check() to check whether the number of smooth functions (k) and the smoother (thin plate regression, 'tp') were adequate and adjusted if necessary.

2.2. Predictions

All the languages lengthened RQs compared to ISQs, most strongly in Icelandic (24.5%), see Table 2. This lengthening is expected to affect the amplitude envelopes in all three languages and is predicted to result in higher power in lower-frequency bands in RQs compared to ISQs. The differences are expected to be strongest in Icelandic and weakest in English, based on the extent of lengthening.

2.3. Results

For reasons of space, we do not show the spectra of the two illocution types separately, but directly present the *differences* in power spectra for RQs vs. ISQs (Figs. 1-3).

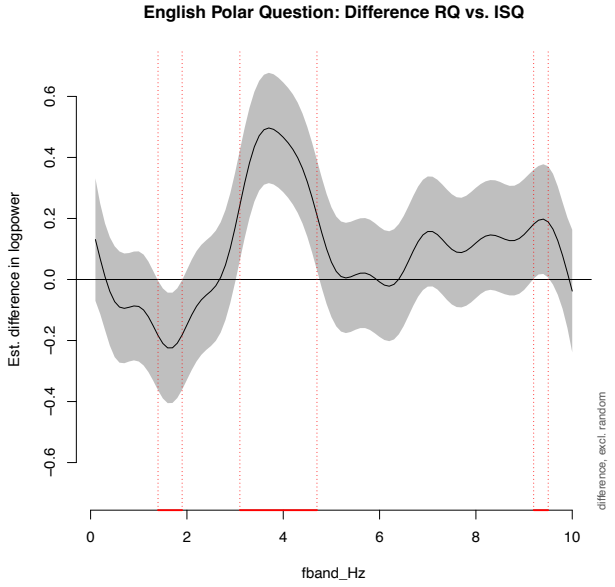

Fig. 1: Effect of illocution type (RQ minus ISQ) in English. Positive values indicate higher power for RQs than RQs. If the gray band of the confidence interval does not include 0, the difference is considered statistically significant at $\alpha = 0.05$.

The languages differ in how strongly illocution type affects the amplitude envelope modulation spectra. On the one hand, there were strong effects of illocution type on the English data (Fig. 1). English polar RQs had a lower power in the frequency range 1.4 – 1.9Hz and, prominently, higher power in the frequency range 3.1 – 4.7Hz.

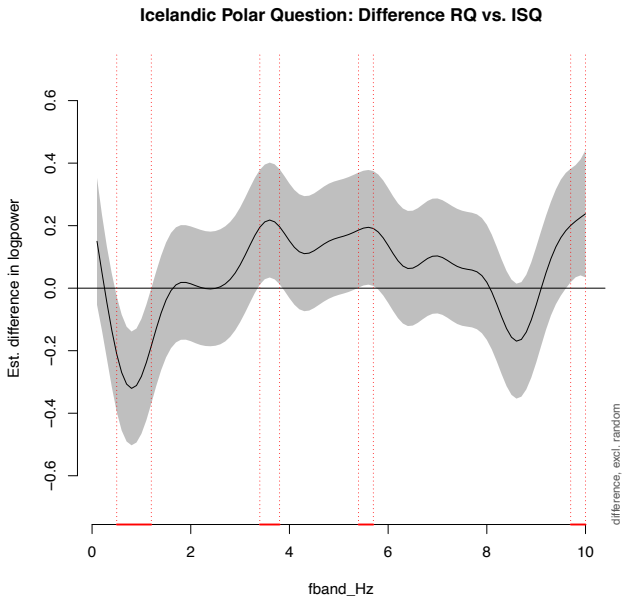

Fig. 2: Effect of illocution type (RQ minus ISQ) in Icelandic.

Icelandic (Fig. 2) shows differences as well, but in a smaller frequency range (0.5 – 1.2Hz and 3.4 – 3.8Hz) and with smaller differences in power. Furthermore, the differences occur in a slightly lower frequency band. German (Fig. 3) is again different: It exhibits a biphasic pattern very late, in the area between 7.3Hz and 9.1Hz, first lower power for RQs, then higher power for RQs. However, compared to English, the differences in power are small.

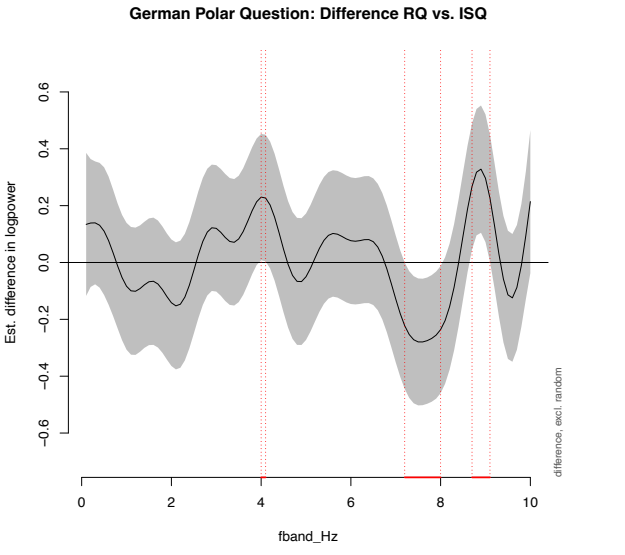

Fig. 3: Effect of illocution type (RQ minus ISQ) in German.

To investigate whether the differences across languages are significant, we fit a model with a smooth for the interaction *language* and *illocution type* and compared it with a model with smooths for the individual terms *language* and *illocution type*, using the package *itsadug* (van Rij et al., 2015). Model comparison showed that the model with the interaction-smooth provided a significantly better fit than the model without ($X2(14.00)=62.450$, $p<2e-16$). To corroborate the interaction between *language* and *illocution type* indicated in the model, we constructed additional models containing binary difference smooth terms that capture *the difference of the difference* over frequency band between two languages (English vs. German, English vs. Icelandic, German vs. Icelandic), closely following the procedure described in van Rij et al., (2019, pp. 11–13) and Wieling (2018, p. 109ff).

The results showed a number of frequency bands with *significant differences across language pairs*. These differences are as follows:

- English differed from German in the frequency band from 3 – 5Hz and 6.5 – 8.5Hz.
- English differed from Icelandic in the frequency band from 3 – 5Hz and 7.8 – 9.5Hz.
- German differed from Icelandic in the frequency band from 0.2 – 1.2Hz, 4.8 – 5.2Hz and 7 – 9.5Hz.

2.4. Discussion

All languages showed an effect of illocution type on the amplitude envelopes. Since information structure was controlled across illocution types (i.e. the same for ISQs and RQs), the differences cannot be related to that factor. There were significant differences between the three intonation languages on the frequency bands in which RQs differed from ISQs. The power differences were strongest in English, with a pronounced peak in energy around 4Hz. The amplitude envelope differences across languages do not mirror the durational lengthening (Table 1). Therefore, it is unlikely that the amplitude envelopes only track the durational differences between RQs and ISQs. Interestingly, the English and Icelandic power differences show a similar pattern in the lower frequency range, but the Icelandic differences are much smaller. The German data show a pronounced difference in the higher frequency band (from 7.5 – 10 Hz).

If duration is a bad predictor for these power differences between RQs and ISQs, we need to take a closer look at other prosodic cues that may explain the cross-linguistic differences. In terms of **voice quality**, German is the only language with differences in voice quality across illocution types (in German 36% of the first words in RQs were breathy, compared to 10% in ISQs, cf. Braun et al., 2019). This cue may explain the biphasic pattern in the high frequency bands, which is absent in English and Icelandic. These latter two languages do not show voice quality differences for polar questions (Dehé & Braun, 2019; Dehé & Wochner, 2022). **Intonationally**, Icelandic and English are similar in terms of accent placement: in both languages, the subject ('anyone' in (1)) has a higher probability of receiving an accent in RQs compared to ISQ (28.8% vs. 0% in English, 12.3% vs. 0.6% in Icelandic). This may explain the power differences below 2Hz and around 4Hz. On the contrary, the fact that both German and English end RQs with high plateau boundary tones (and ISQs with high rising boundary tones) does not seem to be reflected in the amplitude modulation. For automatic classification of questions as ISQ or RQ, parallel consideration of f0 may prove useful (Gibbon, 2021; Ludusan et al., 2011).

Taken together, amplitude envelope modulation spectra differ across illocution types and are most likely influenced by differences in voice quality and accent placement, and less by intonational contour.

3. General Discussion

We showed that amplitude envelope modulation spectra distinguish rhetorical and information-seeking questions in three closely related Germanic languages. We predicted that differences would be largest in Icelandic because this language showed the largest duration differences between RQs and ISQs. However, the amplitude envelope modulation spectra were not largest for Icelandic, but for English. Therefore, the amplitude envelope modulation spectra differences were not (or at least not only) caused by durational differences between RQs and ISQs. Relating amplitude envelope modulation spectra differences to prosodic differences across conditions suggests that voice quality differences and differences in accent placement may play a significant role. In particular, voice quality differences on the first word of the question (more often breathy in RQs in German) seem to have an effect on higher frequency bands, most likely because breathy voice reduces the spectral power of the words. Furthermore, English and Icelandic often placed an accent on the subject pronoun 'anyone', which affects the macro-prosodic rhythm of the utterance (Jun, 2012).

In future work, we plan to include typologically different languages, e.g., such as tone languages (Zahner-Ritter et al., 2022 for Chinese) or accentual phrase languages to get a better overview on the factors that influence amplitude envelope modulation spectra. Furthermore, we plan to use the parameters from the general additive mixed models for automatic classification of utterances as RQs vs. ISQs.

4. Conclusion

This paper adds to our understanding of the factors that influence amplitude envelope modulation spectra by testing three intonation languages. Previous research has shown differences between rhythmically different languages (stressed-timed German vs. more syllable-timed Brazilian Portuguese, cf. Frota et al., 2022). We show that even within one and the same language, amplitude envelope modulation spectra can differ quite extensively (in particular in English polar RQs vs. ISQs). The results showed that lengthening is a poor predictor of differences in amplitude envelope modulation spectra across languages. Differences in voice quality and accent placement also seem to play a role. Clearly, more analyses of carefully controlled materials from typologically different languages are necessary to understand better, which information is encoded in which way in amplitude envelope modulation spectra.

5. Acknowledgements

We thank Volker Dellwo for providing the praat-scripts for extracting the narrowband amplitude envelopes. Furthermore, we thank Antje Strauß and Sonia Frota for discussion on amplitude envelope modulation spectra. Data collection was funded by the German research foundation (BR 3428/4-1, 4-2).

6. References

Arvaniti, A. (2009). Rhythm, timing and the timing of rhythm. *Phonetica*, *66*, 46–63.

Biezma, M., & Rawlins, K. (2017). Rhetorical questions: Severing asking from questioning. In D. Burgdorf, J. Collard, S. Maspong, & B. Stefánsdóttir (Eds.), *Proceedings of SALT 27* (pp. 302–322).

Boersma, P., & Weenink, D. (2018). *Praat: Doing phonetics by computer*.

Bögel, T., & Braun, B. (2022). Rhetorical questions in Persian. *Phonetik Und Phonologie Im Deutschsprachigen Raum*, Phonetik und Phonologie im deutschsprachigen Raum. https://doi.org/10.11576/PUNDP2022-1042

Braun, B., Daniela Wochner, Zahner, K., & Nicole Dehé. (2018). *Classification of interrogatives as information-seeking or rhetorical questions*. 17th Speech Science and Technology Conference. Sydney, Australia.

Braun, B., Dehé, N., Neitsch, J., Wochner, D., & Zahner, K. (2019). The prosody of rhetorical and information-seeking questions in German. *Language and Speech*, *62*(4), 779–807. https://doi.org/10.1177/0023830918816351

Braun, B., Einfeldt, M., Esposito, G., & Dehé, N. (2020). *The prosodic realization of rhetorical and information-questions in German spontaneous speech*. Speech Prosody.Tokyo, Japan.

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., & Ghazanfar, A. A. (2009). The Natural Statistics of Audiovisual Speech. *PLOS Computational Biology*, *5*(7), e1000436. https://doi.org/10.1371/journal.pcbi.1000436

Dehé, N., & Braun, B. (2019). The prosody of rhetorical questions in English. *English Language and Linguistics*, *24*(4), 607–635. https://doi.org/10.1017/s1360674319000157

Dehé, N., Braun, B., Einfeldt, M., Wochner, D., & Zahner-Ritter, K. (2022). The prosody of rhetorical questions: A cross-linguistic view. *Linguistische Berichte*, *269*, 3–42. https://doi.org/10.46771/978-3-96769-175-7

Dehé, N., Braun, B., & Wochner, D. (2018). *The prosody of rhetorical vs. Information-seeking questions in Icelandic*. 403–407.

Dehé, N., & Wochner, D. (2022). Voice quality and speaking rate in Icelandic rhetorical questions. *Nordic Journal of Linguistics*, 1–10. https://doi.org/10.1017/S0332586522000014

Frota, S., Vigário, M., Cruz, M., Hohl, F., & Braun, B. (2022). *Amplitude envelope modulations across languages reflect prosody*. https://doi.org/10.21437/SpeechProsody.2022-140

Gibbon, D. (2021). The rhythms of rhythm. *Journal of the International Phonetic Association*, 1–33. https://doi.org/10.1017/S0025100321000086

Gross, J., Hoogenboom, N., Thut, G., Schys, P., Panzeri, S., Belin, P., & Garrod, S. (2013). Speech Rhythms and Multiplexed Oscillatory Sensory Coding in the Human Brain. *PLOS Biology*. https://doi.org/10.1371/journal.pbio.1001752

He, L., & Dellwo, V. (2016). *A Praat-based algorithm to extract the amplitude envelope and temporal fine structure using the Hilbert transform*. 530–534.

He, L., & Dellwo, V. (2017). *Amplitude envelope kinematics of speech signal: Parameter extraction and applications. Elektronische Sprachsignalverarbeitung 2017. Dresden: TUDpress, 1-8*.

Jun, S.-A. (2012). *Prosodic Typology Revisited: Adding Macro-Rhythm*. Speech Prosody, Shanghai, China.

Leong, V., & Goswami, U. (2015). Acoustic-emergent phonology in the amplitude envelope of child-directed speech. *PLoS One*, *10*(12), e0144411.

Ludusan, B., Origlia, A., & Cutugno, F. (2011). *On the Use of the Rhythmogram for Automatic Syllabic Prominence Detection*. Interspeech 2011. Florence, Italy.

Poeppel, D. (2014). The neuroanatomic and neurophysiological infrastructure for speech and language. *Curr Opin Neurobiol*, *28*, 142–149.

Poeppel, D., & Assaneo, M. F. (2020). Speech rhythms and their neural foundations. *Nature Review Neuroscience*, *21*, 322–334.

Tilsen, S., & Arvaniti, A. (2013). Speech rhythm analysis with decomposition of the amplitude envelope: Characterizing rhythmic patterns within and across languages. *The Journal of the Acoustical Society of America*, *134*(1), 628–639.

van Rij, J., Hendriks, P., van Rijn, H., Baayen, R. H., & Wood, S. N. (2019). Analyzing the Time Course of Pupillometric Data. *Trends in Hearing*, *23*, 233121651983248.

van Rij, J., Wieling, M., Baayen, R. H., & van Rijn, D. (2015). *itsadug: Interpreting Time Series and Autocorrelated Data Using GAMMs*. https://cran.r-project.org/package=itsadug

Wieling, M. (2018). Analyzing dynamic phonetic data using generalized additive mixed modeling: A tutorial focusing on articulatory differences between L1 and L2 speakers of English. *Journal of Phonetics*, *70*, 86–116.

Wieling, M., Margaretha, E., & Nerbonne, J. (2012). Inducing a measure of phonetic similarity from pronunciation variation. *Journal of Phonetics*, *40*(2), 307–314. https://doi.org/10.1016/j.wocn.2011.12.004

Wood, S. N. (2006). *Generalized additive models: An introduction with R*. Chapman & Hall/CRC Press.

Wood, S. N. (2015). *mgcv: Mixed GAM computation vehicle with GCV/AIC/REML smoothness estimation*.

Wood, S. N., & Saefken, B. (2016). Smoothing parameter and model selection for general smooth models. *Journal of the American Statistical Association*, *111*, 1548–1575.

Zahner, K., Kutscheid, S., & Braun, B. (2019). Alignment of f0 peak in different pitch accent types affects perception of metrical stress. *Journal of Phonetics*, *74*, 75–95.

Zahner-Ritter, K., Chen, Y., Dehé, N., & Braun, B. (2022). The prosodic marking of rhetorical questions in Standard Chinese. *Journal of Phonetics*, *95*, 101190.