Chapter 3

Chunking an unfamiliar language: Results from a perception study of German listeners

Nele Ots[a] & Piia Taremaa[b]

[a]Goethe-University Frankfurt [b]University of Tartu

This study investigates the impact of prosodic boundary phenomena and syntactic clause boundaries on native and non-native speech chunking. German and Estonian listeners were asked to listen to spontaneous utterances spoken in Estonian and to mark in corresponding written transcripts when they perceived any sort of a break between the words. Estonian listeners were the strongest guided by the clause boundaries whereas German listeners were sensitive to all of the prosodic boundary phenomena but resistant to the presence of clause boundaries. In particular, both German and Estonian listeners utilized longer pauses and rising F0 contour as cues for chunk boundaries. German listeners additionally employed phrase-final lengthening and intensity drop. These results suggest strong bottom-up effects in non-native speech processing, and both bottom-up effects and top-down effects in native processing of speech. Thus, the well-known prosodic boundary phenomena trigger bottom-up processing in on-going spontaneous speech comprehension.

1 Introduction

Speech comprehension starts with and depends on the extraction of discrete sequential units from continuous speech flow. In order to discern and maintain these units in working memory, listeners interpret smaller units detected in the context of larger ones, that is, in (speech) chunks (Dahan & Ferreira 2019, Christiansen & Chater 2016). Speech chunking operates across multiple levels of

Nele Ots & Piia Taremaa. 2023. Chunking an unfamiliar language: Results from a perception study of German listeners. In Fabian Schubö, Sabine Zerbian, Sandra Hanne & Isabell Wartenburger (eds.), *Prosodic boundary phenomena*, 87–117. Berlin: Language Science Press. DOI: 10.5281/zenodo.7777530

linguistic representation and is related to top-down as well as bottom-up processing (see, e.g., Dahan & Ferreira 2019). Top-down processing concerns world and linguistic knowledge (including lexical, semantic and syntactic knowledge), whereas bottom-up processing relates to sensory input from acoustic signals. Native listeners, when extracting discrete units from speech, are known to exploit their top-down knowledge about lexical-semantic and syntactic information (Mattys et al. 2005). Thus, top-down processing is decisive for chunking a continuous stream of speech into units. This speeds up spoken language comprehension by helping listeners to rapidly recognize and process segments of language that are syntactically, lexically and semantically coherent and plausible, given the context.

In certain aspects of linguistic structures, the extent of bottom-up processing in speech chunking is unclear. In particular, signal-driven prosodic cues (e.g., lengthening of the segments) have proven to be highly functional for the recognition of words (e.g., White et al. 2020). Whether listeners are also able to recognize chunks at a higher level, i.e., intonational phrases, is only vaguely understood. Recently, Ordin et al. (2017) pointed out that listeners may apply phrase-level prosody alongside word-level prosody for the generation of so-called prosodic frames (Keating & Shattuck-Hufnagel 2002, Schild et al. 2014, Silbert et al. 2014). Understanding the role of phrase-level prosody in chunking processes is necessary because prosodic frames are proposed to take part in encoding as well as decoding processes in language production and perception (Keating & Shattuck-Hufnagel 2002, Schild et al. 2014, Silbert et al. 2014).

A handful of phonetic perception studies have indicated that signal-driven prosodic information (e.g., durational or tonal discontinuities in terms of pausing, lengthening and pitch reset) ranks rather low in native speech chunking (Cole et al. 2010, Christodoulides et al. 2018, Duez 1985). In contrast, several psycholinguistic studies have demonstrated the significant role of phrasal prosody in recognizing and remembering novel words (Langus et al. 2012, Ordin et al. 2017). This, in turn, drives our investigation of speech chunking in non-native listeners in comparison with native listeners (for a similar approach, see Himmelmann et al. 2018, Riesberg et al. 2020). The current study asks to what extent speech chunks are accessible from signal-driven prosodic information.

1.1 Signal-driven prosodic boundary cues

A type of well-known prosodic unit is the so-called tone group (Halliday 1967), also known as the intonation unit (Chafe 1987) or, more commonly, an intonational phrase (Pierrehumbert 1980, Ladd 2008). In intonational phonology, intonational phrases (IPs) constitute abstract phonological units that are composed of discrete abstract categories of pitch accents and boundary tones (Ladd 2008, Pierrehumbert & Hirschberg 1990). The identification of phonological categories of pitch accents and boundary tones in spoken sentences usually follows from phonological analysis. As such, the well-known concept of IP constitutes top-down information about the phonological structure of a language. For the purposes of this study, it is interesting that IPs frequently correspond with concrete acoustic regularities directly observable in the speech signal.

IPs are frequently characterized as units of tonal coherence (Bois et al. 1992, Breen et al. 2012, Buhmann et al. 2002, Himmelmann et al. 2018). An underlying acoustical phenomenon of the significant percept of tonal coherence is the continuous decline of fundamental frequency (F0, acoustic approximation of sentence intonation) from the beginning to the end of an IP. This decline was traditionally measured considering the F0 maxima in phonological pitch accents (see, e.g., Ladd 1988, Liberman & Pierrehumbert 1984, Pierrehumbert 1979). More recent research has found a more automatic way to fit a straight line to the F0 contour as a function of time (Yuan & Liberman 2014). F0 declination is a global component of the F0 contour, and it should interact only mildly with local F0 movements determining pitch accents and boundary tones (Fujisaki 1983, Fujisaki & Hirose 1982). The regularity of F0 declination underlies the readily audible cue of pitch reset, which means that the continuous decline of the F0 contour is disrupted by setting the level of F0 much higher than predicted by an on-going F0 decline, e.g., by stepping up the pitch. In intonation research, the pitch reset has often been utilized as a valuable cue signaling the right edge of an IP (Cooper & Sorensen 1981, Couper-Kuhlen 2001, Himmelmann et al. 2018, Ladd 1988, Schuetze-Coburn et al. 1991, Thorsen 1985).

In the stream of speech, IPs are most easily defined by their boundaries. The IP boundaries in spoken language are associated with a battery of phonetic boundary phenomena encompassing systematic changes in duration, intensity and F0. In particular, a durational discontinuity that involves slowing down speech, or, more specifically, lengthening speech segments, constitutes a type of signal-driven prosodic cue that is frequently referred to as pre-boundary or phrase-final lengthening (Berkovits 1994, Fon et al. 2011, Nakai et al. 2009, Petrone et al. 2017, Wightman et al. 1992; for this cue in German, see also Schubö & Zerbian 2023 [this volume], Huttenlauch et al. 2023 [this volume], and Wellmann et al. 2023 [this volume]). In terms of prosodic boundary cues, intensity has attracted interest to a lesser degree. However, some studies have indicated that an intensity curve within words may also function as a boundary cue. The increasing intensity difference between the initial and final syllable in a word constitutes the phenomenon of intensity drop (Trouvain et al. 1998, Wagner & McAuliffe 2019). Finally, IP boundaries are well known to be indexed by intonational movements (boundary tones at the abstract level of intonational phonology), which are indexed by falling or rising F0 contours at the ends of IPs (e.g., consider falling intonation in statements and rising intonation in questions; on the role of rising F0 contours, see Petrone et al. 2017, Huttenlauch et al. 2023 [this volume], and Wellmann et al. 2023 [this volume]). Thus, signal-driven prosodic cues, such as phrase-final lengthening, intensity drop, pitch reset and F0 movements, define IPs in spoken utterances. These cues have certain acoustic correlates in the speech signal and these can be investigated as input for bottom-up processing.

1.2 Perception of phrasal prosody

In the auditory processing of language, the discontinuities of duration and F0 have been shown to be highly functional. Namely, phrase-final lengthening and continuous F0 declination can help listeners to discover long-distance dependencies between words, or tentatively, clausal relationships (de la Cruz-Pavía et al. 2019, Langus et al. 2012, Ordin et al. 2017). For example, in an experiment with Italian listeners, Langus et al. (2012) created a novel language by defining words and long-distance semantic dependencies between them through systematically manipulating the probability distributions of sounds and syllables. Importantly, the stipulative sentences of the novel language were additionally accompanied by pre-boundary lengthening and continuous F0 declination. Langus et al. (2012) were able to demonstrate that long-distance dependencies between the words were only discovered in the presence of prosodic cues. Moreover, they found that while F0 declination is useful for detecting dependencies at the level of a stipulative sentence or a clause, phrase-final lengthening induces the listener to perceive a stipulative syntactic phrase. Thus, they were able to separate the functions of the two types of prosodic cues at two different linguistic levels – a stipulative phrase vs. a stipulative clause. Altogether, the results from Langus et al. (2012) demonstrate that the presence of phrase-final lengthening and F0 declination clearly enforces perception of a sort of language chunk. For additional functionality in infant language acquisition see Wellmann et al. 2023 [this volume].

Phonetic studies of perceptual speech chunking further indicate that there is an unavoidable syntactic component in the perceptual chunking of language. For example, Duez (1985) presented listeners with natural, distorted and synthesized speech and asked them to explicitly mark silent pauses. Remarkably, the results show that listeners detected significantly fewer pauses in distorted and synthesized speech than in normal speech. Thus, the study indicates that listeners, even when explicitly detecting signal-based information, may rely more strongly on syntactic-semantic information than acoustic information or may ignore the latter altogether (for a replication of these results, see Simon & Christodoulides 2016). In a more recent study, Cole et al. (2010) asked native listeners of American English to listen to broadcasted conversations and to mark in written transcripts where they heard some sort of a break or juncture. The results show that clause boundaries had the strongest impact on boundary perception; phrase-final lengthening or a duration cue ranked lower, and F0 did not play any role. A study by Christodoulides et al. (2018) employed slightly different methodology by asking French listeners to press a button when they heard some sort of a break in speech. The timeline of button presses was synchronized with the stream of speech, and without having any written input, the outcome was nevertheless that syntactic clause boundaries most strongly contributed to boundary perception. These results further demonstrate how influential the access to clausal information is in the metalinguistic tasks. As the clause boundaries constitute the linguistic knowledge, they can be taken as input for the top-down processing. As such, the existing studies demonstrate pervasive top-down processing in native speech comprehension.

Strikingly, Riesberg et al. (2020) found that lexical and syntactic variables participate even in non-native perception of speech chunks. Their study employed the same methodology as in Cole et al. (2010) and asked native speakers of German to listen to short stories spoken in Papuan Malay and mark in written transcripts where they heard some sort of a break. Speakers of Papuan Malay were presented with short stories in German with the same task. Both language groups also judged the stories spoken in their native language. For listeners from both language groups, clause boundaries were the second strongest factor that contributed to the perception of chunks in the unfamiliar languages, while pauses were the strongest cue.

This result becomes less surprising when considering language production. Specifically, several studies have found that syntactically defined segments, such as clauses and phrases, are often accompanied by acoustic discontinuities (see, e.g., Cutler et al. 1997, Petrone et al. 2017). For example, Féry & Ishihara (2009) demonstrated in a reading experiment that speakers tended to reset pitch and start a new declination trend for F0 at the beginning of embedded subclauses. This indicates that IPs, or the signal-driven prosodic cues of duration, and F0 in particular, tend to strongly associate with the syntactic representation of language. In other words, syntactic elements such as clauses are produced as prosodically coherent speech chunks, and the results from Riesberg et al. (2020) suggest they will also be perceived as such regardless of the listener's language background.

1.3 The current study

The aim of this study is to determine the impact of signal-driven prosodic cues (i.e., bottom-up processing) separately from syntactic-semantic information (i.e., top-down processing) in speech chunking. For this, we investigate how non-native speakers perceive an unfamiliar natural language. When processing an unfamiliar language, semantic-syntactic cues are not available to the listener. Arguably, this forces non-native listeners to rely on signal-based acoustic cues whilst chunking speech (Himmelmann et al. 2018, Riesberg et al. 2020). By investigating chunking of speech flow in an unfamiliar language, we are able to examine the role of signal-based prosodic information in bottom-up processing of language.

To assess the influence of prosodic information on speech chunking, we conducted a chunking experiment based on Rapid Prosody Transcription (RPT; Cole et al. 2010, 2011, Mahrt 2016) in which Estonian and German listeners had to chunk excerpts of spontaneous utterances spoken in Estonian. We investigated the impact of signal-driven prosodic cues (i.e., phrase-final lengthening, intensity drop, rate of F0 declination, and pause duration) against the clausal structure of spontaneously spoken utterances and the listeners' language background. Signal-driven prosodic information serves as input for bottom-up processing, whereas the clausal structure provides input for top-down processing. Crucially, the clausal structure of Estonian utterances is not available for German listeners who are unfamiliar with the Estonian language. Therefore, we hypothesized that the impact of bottom-up information in speech chunking is modulated by the listener's language background. Based on the notion of top-down processing, we expected the German listeners to be less affected by the clausal structure and to be more sensitive to the signal-driven prosodic cues, whereas the Estonian listeners were expected to use both clausal and acoustic cues. The alternative prediction relies on the results in Riesberg et al. (2020). Namely, the German listeners could perform similarly to Estonian listeners in terms of clausal cues. This outcome would indicate a strong relationship between prosodic information and clausal structure in Estonian speech production because, arguably, the German listeners would rely on the prosodic cues that are tightly associated with clausal structure.

2 Materials and method

For our experiment, we applied the methodology of Rapid Prosody Transcription (RPT), in which listeners are typically asked to listen to excerpts of speech and mark the words that they perceive as prominent or that stand before some sort of a break (Cole et al. 2010, 2011, Mahrt 2016).

2.1 Participants

Altogether, 47 Estonian listeners (average age 30.0 years) took part in an earlier experiment (Ots & Taremaa 2022). They originated from various regions of Estonia. Given their age, they most likely speak Standard Estonian, and the dialectal variation in Estonia is probably not that pronounced in young speakers.

For this study, 90 native speakers of German were recruited through a crowdsourcing marketplace designed for conducting research (Prolific). They were paid about £2.50 to complete the task, which took about 20 minutes. The average age of the participants was 28.8 years (with 0.03 percent of participants not reporting). 48.9 percent of participants were female, and 46.7 percent were male (with 0.04 percent of participants not reporting). All participants reported German to be their first language. 86.7 percent of participants reported having knowledge of some other language, most frequently English. None reported having knowledge of Estonian.

2.2 Stimuli

We extracted 396 excerpts of spontaneous speech (4727 words altogether) from 10 native Estonian speakers (5 male and 5 female speakers with an average age of 25.3 years) from the phonetic corpus of spoken Estonian (Lippus et al. 2016). Auditive analysis did not reveal any distinctive dialectal characteristics in these speakers. They appeared to use Standard Estonian as it is taught in schools. The excerpts constituted a stretch of fluent speech between silent pauses of 400 ms or longer. The excerpts contained 18 to 24 syllables, yielding an average duration of 3300 ms. For the experiment with Estonian listeners, the 396 excerpts were randomly distributed between 4 different lists, each containing 99 excerpts in total. The lists for German listeners were kept shorter, as their task was to listen to non-native language. Thus, the 396 excerpts were randomly distributed between 9 lists, with each list containing 44 excerpts in total.

2.3 Procedure

The Estonian excerpts were presented to native speakers of German, unfamiliar with the Estonian language. The study was conducted over the internet using LMEDS software (Mahrt 2016). Based on RPT methodology (see, e.g., Cole et al. 2010, 2011, Riesberg et al. 2020), the participants were asked to listen to speech excerpts and identify the chunks of words ("kõnejupp" in Estonian, "Wortgruppierung" in German) in the written transcripts appearing on the screen. Technically, they needed to click on the words that they perceived as occurring at some sort of a break. In essence, the task was to make a binary choice to either place a boundary or not at each consecutive pair of words in an excerpt. No additional instructions on what exactly this break might be were provided. The Estonian listeners were allowed to listen to the excerpts two times, the German listeners were able to listen to the excerpts as many times as they needed.

As this task requires listening to speech excerpts and simultaneously reading written transcripts, it is recognizably difficult for a non-native listener to perform. However, it has already been successfully administered with languages that are typologically far apart in a study by Himmelmann et al. (2018), in which German listeners were asked to chunk speech excerpts from Indonesian languages, and speakers of Indonesian languages were asked to chunk speech excerpts in German. Riesberg et al. (2020) followed a similar procedure with German and Papuan Malay speakers. Both studies yielded interpretable and plausible results. The researchers' justification for this procedure was based on the shared orthographic conventions of the languages.

Estonian orthography is phonemic, and therefore, it should be easily accessible to a German listener/reader. Except for some contrasts in phoneme length, each symbol is encoded by exactly one sound, and most of the graphemes correspond to symbols in German. The survey conducted after the completion of the task indicated that the participants were happy to take part in the study: the average satisfaction on a scale of 0 to 100 was 78.7 (SD = 20.7). 13.3 percent of participants claimed to have difficulties with mapping speech sounds to written words, and 11.1 percent of participants even reported having fun listening to a language that they did not know.

We did not manage to present the lists to equal numbers of participants, as the LMEDS software does not have the option to define different lists of experimental stimuli. Unfortunately, our own solution for extending the LMEDS with this feature did not work properly. Thus, the number of listeners per excerpt varies across the lists, ranging from a total of 6 to a total of 12 listeners per list.

The participants' responses were encoded at the final boundary of every word, using 0 when no boundary was placed and 1 when a boundary was placed. Altogether, the Estonian results consisted of 55,541 data points, and the German results consisted of 47,257 data points (number of words multiplied by the number of listeners). We did not instruct the participants to listen for breaks in the very last words of excerpts, and therefore, the final words of each excerpt were excluded from the evaluation of effects, leaving us with 50,889 data points for the Estonian data and 43,291 data points for the German data.

2.4 Test variables

Four test variables capturing the variation in duration (syllable duration, pause duration), intensity (intensity difference) and F0 (F0 proportion) were automatically extracted from all words in the excerpts. The absolute duration of the last syllable of every word (syllable duration in milliseconds) was taken to index preboundary lengthening. An utterance was defined to be a stretch of fluent speech between silent pauses of 400 ms or longer. Thus, the selected utterances did not contain pauses that were longer than 400 ms. However, they did contain silent and filled pauses shorter than 400 ms (352 instances (0.07%) in a corpus of 4372 words). The duration of these silent and filled pauses was collected as the second durational variable after syllable duration (pause duration in milliseconds).

For the third variable, intensity difference, the intensity as root mean square (RMS) amplitude of the very first and the very last syllable of a word was automatically extracted, and the intensity curve within a word was approximated by subtracting the RMS value of the last syllable from the RMS value of the first syllable (intensity difference). The intensity difference was calculated to index the intensity drop. The larger the intensity difference, the likelier it is that a word contains the intensity drop. A small or negative difference is an indication that a word does not contain an intensity drop.

F0 contours (Hz) were extracted from the excerpts in two passes with the help of the auto-correlation method available in Praat (Boersma & Weenink 2019). During the first pass, F0 tracks were extracted with Praat default settings for the lowest and highest F0, the "floor" and "ceiling" (75 Hz and 600 Hz, respectively). Then, the first and third quartiles of F0 (Q1 and Q3) were calculated for each speaker and recorded in a table. In the second pass, F0 contours were extracted with speaker-specific settings (0.75*Q1 for the floor and 1.5*Q3 for the ceiling). Finally, the resulting F0 contours were smoothed by 4 Hz and quadratically interpolated using the corresponding functions in Praat. Based on the F0 contours, F0 maxima (in Hz) were automatically identified in the vowels of the word-initial lexically stressed syllables. This identification procedure is well justified because, in Estonian, the high tone of the falling pitch accent is most frequently aligned with the first syllable (see, e.g., Asu & Nolan 1999). Therefore, relatively high F0 maxima from the word-initial syllables can be taken to index intonational pitch accents.

For the fourth variable – F0 proportion, F0 maxima were divided with the corresponding utterance's mean F0. As such, the F0 proportion was devised to approximate the height of a pitch accent relative to the utterance's mean F0. F0 proportion was calculated to normalize the speaker-specific and item-specific tonal variation in the utterances. Due to the well-known phenomenon of F0 declination, F0 maxima are higher at the beginnings of the corresponding domains (e.g., IP, clause, or a perceptual speech chunk) than at the ends of these domains (Cooper & Sorensen 1981, Liberman & Pierrehumbert 1984, Yuan & Liberman 2014). Therefore, F0 maxima decrease across the domain also relative to the utterance's mean. In other words, F0 proportion is smaller at the ends of corresponding domains than at the beginnings of these domains. Followingly, the F0 proportion should be smaller at the end of the perceived boundaries than at the non-boundaries if the non-expert perception of a break, or more generally, the perception of a speech chunk relies on the tonal coherence.

The material was also scored for the boundaries of clauses. This scoring was not devised in a particular syntactic framework but followed the functional approach provided in Erelt & Metslang (2017). A clause was defined as consisting of a finite verb together with elements that cluster around the verb and are not finite verbs themselves. Clauses were allowed to also consist of non-constituents, such as disclosures and interjections. In practice, conjunctions served as a frequent cue for the separation of utterances into smaller units of clauses (see rows 7 and 12 in Table 1). For clausal structure, the last word in a clause was encoded as being at the clause boundary.

2.5 Analysis

In our analysis, the continuous variables of syllable and pause duration, intensity difference, and F0 (F0 proportion) function as bottom-up information, whereas clause boundaries function as top-down information. In terms of the impact of continuous signal-based prosodic variables in perceptual chunking, we expected the likelihood of boundary perception to increase

1. together with increasing syllable duration,
2. together with increasing pause duration,
3. together with increasing intensity difference,
4. together with decreasing F0 proportion.

Table 1: Sample of the scoring of clause boundaries in conversational utterances.

Row	Transcription	Translation	Function	Clause boundary
1	*ja*	and	conjunction	no
2	*siis*	then	adverbial	no
3	*käisi-me*	went-we	verb	no
4	*seal*	there	adverbial	no
5	*iisraeli*	Israelian	adverbial	no
6	*muuseum-is*	museum-in	adverbial	yes
7	*kus*	where	conjunction	no
8	*see*	this	subject	no
9	*suur*	big	subject	no
10	*makett*	maquette	subject	no
11	*oli*	was	verb	yes
12	*mis*	which	conjunction	no
13	*oli*	was	verb	no
14	*päris*	pretty	predicative	no
15	*võimas*	awesome	predicative	yes

We predicted that the perception of both types of information would be modulated by the listener's linguistic background (familiar vs. unfamiliar) such that the effects of prosodic variables would be larger for German than for Estonian listeners and that the effect of clause boundaries would be larger for Estonian than for German listeners.

The effects of clause boundaries, syllable duration, intensity difference, F0 proportion and pause duration were estimated in relation to the language background in the general linear mixed effects regression analysis as provided in the `lme4` package (Bates et al. 2015) in R (R Core Team 2018). We defined five predictors of the binomially distributed response variable:

1. an interaction between clause boundaries and language,
2. an interaction between syllable duration and language,
3. an interaction between pause duration and language,
4. an interaction between intensity difference and language,
5. an interaction between F0 proportion and language.

Pause and syllable durations were logarithmically transformed with the base of 10. To maintain the interpretability and comparability of the slopes, all continuous variables were z-scored before entering the regression analysis. The generalized linear mixed effects model was defined to contain the number of listeners as an exposure variable because the four lists of excerpts in the Estonian experiment and the nine lists of excerpts in the German experiment were exposed to different numbers of listeners. The random effects structure included random slopes for listeners because we reasoned that listeners are highly likely to vary in their sensitivity to the clausal structure, syllable duration, pause duration, intensity difference and F0 proportion. We also included random slopes for excerpts because they originated from the conversations of 10 different speakers and displayed considerable and systematic variation in speech rhythm, intensity, and melody. The converging model fit was obtained by using the `optimx` optimizer (Nash 2014, Nash & Varadhan 2011).

3 Results

3.1 The impact of prosodic cues on non-native speech chunking

The aim of the analyses was to determine the impact of phonetic variation of duration, intensity and F0 as bottom-up information in non-native speech chunking. Before proceeding to the statistical evaluation, the explanatory variables were checked for correlations (see Table 2).

Table 2: Correlations between the explanatory variables as estimated by Pearson's r coefficient. The significance stars indicate how likely they are to be found in the whole population, given the sample means. ***: $p < 0.001$, **: $p < 0.01$, *: $p < 0.05$.

	Clause	Syl. dur.	Int. dif.	F0 prop.
Clause				
Syllable dur.	0.08***			
Intensity dif.	0.12***	−0.02		
F0 prop.	0.04**	0.04*	0.19***	
Pause dur.	−0.01	0.07	−0.06	0.03

The correlations between the selected variables in Table 2 are very close to zero. This indicates that they are appropriate as explanatory variables for the multiple regression analysis with mixed effects. The results of the analysis are presented in Table 3. The column "Est." contains the log odd estimates of the fixed effects clause, syllable duration, intensity difference, pause duration and F0 proportion in interaction with language. The third and the fourth column give the 95% confidence intervals of the estimates. The t-values and p-values can be found in the last two columns. The p-values are given together with the significance codes (asterisks).

Table 3: Log odd estimates and significance of the standardized variables in predicting boundary perception. ***: $p < 0.001$, **: $p < 0.01$, *: $p < 0.05$.

	Est.	2.5%	97.5%	t	p
(Intercept)	−12.22	−12.99	−11.46	−31.35	0.00***
Language [Ger]	1.84	0.97	2.71	4.15	0.00***
Clause [yes]	3.18	2.65	3.71	11.72	0.00***
Syllable dur.	0.04	−0.21	0.29	0.3	0.77
Intensity dif.	0.06	−0.31	0.44	0.34	0.73
Pause dur.	0.53	0.27	0.79	4.06	0.00***
F0 prop.	0.27	0	0.54	1.98	0.05*
Language [Ger]:Clause [yes]	−2.57	−3.11	−2.04	−9.41	0.00***
Language [Ger]:Syllable dur.	0.22	0.01	0.42	2.1	0.04*
Language [Ger]:Intensity dif.	0.32	0.08	0.56	2.57	0.01**
Language [Ger]:Pause dur.	0.06	−0.15	0.26	0.54	0.59
Language [Ger]:F0 prop.	0.16	−0.04	0.37	1.55	0.12
AIC					5770.47
R^2 (fixed effects)					0.15
R^2 (all effects)					0.78

The positive values of the log odd estimates indicate an increase in the probability of boundary perception, whereas the negative values suggest a decrease in the probability of boundary perception. Given that the variables were standardized before entering the regression analysis, the estimates enable us to see that the presence of a clause boundary is the factor that has the most profound effect on boundary perception. This is followed by the effect of the interaction between the language and clause and the main effect of the language. The lower-ranking effects stem from the signal-based prosodic variables. The main effect of the language is followed by the main effect of pause duration. The next strongest effect is the intensity difference in the interaction with language. This is followed by the main effect of F0 proportion. Finally, syllable duration also contributes to the boundary perception in the interaction with language. The main effects of syllable duration and intensity difference, and the interactions between language and pause duration and between language and F0 proportion did not turn out significant. The results of the linear-mixed effects regression analysis are illustrated in the effect plots in Figure 1. These plots highlight the predicted influences of clause boundaries, syllable duration, intensity difference, pause duration and F0 proportion on boundary perception.

Figure 1A further demonstrates how the significant main effect of clause boundaries is modulated by the significant interaction between clause boundaries and language background. In particular, we can see that the Estonian listeners are strongly affected by the presence of a clause boundary whereas the German listeners are insensitive to the presence of clause boundaries (compare blue points and whiskers to red points and whiskers). Figure 1B demonstrates that increasing duration of the last syllable contributes to the perception of a boundary for German (see the blue line and confidence intervals that are not overlapping from left to right) but not for Estonian listeners (see the red line and the red confidence intervals that are overlapping from left to right along the probability function). Similarly, Figure 1C indicates that the probability of hearing a boundary increases together with increasing intensity difference for German listeners (see the blue line and confidence intervals that are not overlapping from left to right) but not for Estonian listeners (see the red line and the red confidence intervals that are overlapping from left to right along the probability function). Figures 1D and 1E underscore the main effects of pause duration and F0 proportion. We can readily observe that regardless of the listener's language background, the probability of boundary perception increases as the pause duration and F0 proportion increase (see the rising probability functions and non-overlapping confidence intervals in blue and red from left to right along the probability functions).

3.2 Interrater agreement

To establish the interrater agreement, we calculated Fleiss' κ scores between the Estonian and German listeners according to the lists of excerpts (see Table 4).

The κ scores in Table 4 show fair agreement within Estonian listeners and within German listeners. While Estonian listeners of Lists 1 and 2 perform moderately, the scores for other lists remain below 40, yielding an average κ score of 0.38 for Estonians. The average κ score for German listeners is 0.28, also indicating fair agreement. It was not possible for us to calculate the κ scores between the Estonian and German listeners because the excerpts were distributed among the different lists (among four lists for Estonians and nine lists for Germans).

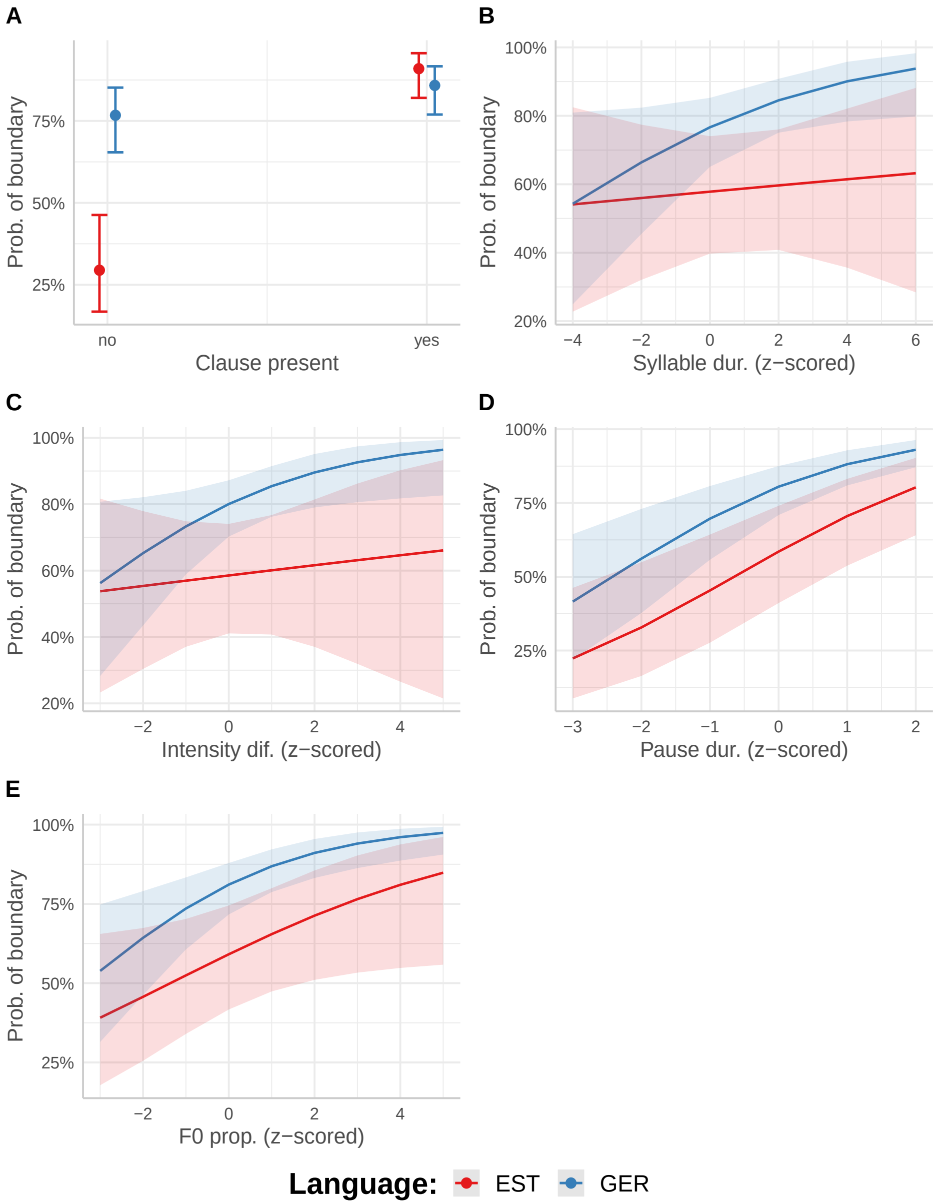

Figure 1: Predicted probabilities of boundary perception as a function of clause boundaries (A), syllable duration (B), intensity difference (C), pause duration (D) and F0 proportion (E) while holding other variables constant. The shadowed bands around the lines represent 95% confidence intervals of the estimates. The change in the probability function is significant when the confidence intervals do not overlap from left to right along the probability function.

Table 4: Fleiss' κ scores for boundaries in the familiar (Estonian) and unfamiliar (German) language conditions. N: Number of listeners. The κ values between 0–0.20 indicate *slight* agreement, 0.21–0.40 suggest *fair* agreement, 0.41–0.60 indicate *moderate* agreement, 0.61–0.80 indicate *substantial* agreement, and 0.81–1 suggest *almost perfect* agreement (see Landis & Koch 1977).

List	N	κ	95% CI	z	Agreement
Estonian					
1	13	0.47	(0.46, 0.47)	142.32	moderate
2	9	0.41	(0.40, 0.42)	85.67	moderate
3	14	0.39	(0.38, 0.39)	126.14	fair
4	11	0.27	(0.26, 0.28)	68.37	fair
Mean (SD):			0.38 (0.08)		(fair)
German					
1	10	0.29	(0.28, 0.30)	45.05	fair
2	10	0.34	(0.33, 0.36)	51.91	fair
3	9	0.33	(0.31, 0.34)	44.12	fair
4	11	0.25	(0.24, 0.26)	42.17	fair
5	13	0.25	(0.24, 0.26)	42.17	fair
6	12	0.28	(0.27, 0.29)	56.33	fair
7	7	0.23	(0.22, 0.24)	42.42	fair
8	6	0.32	(0.30, 0.34)	28.17	fair
9	12	0.27	(0.26, 0.28)	51.07	fair
Mean (SD):			0.28 (0.04)		(fair)

Therefore, we decided to investigate the perceptual chunks to see whether they show any similarities between the native and non-native speakers. The results of the regression analysis have strongly indicated that for the Estonian listeners, the boundaries of chunks correspond with clause boundaries. Additionally, they are guided by pause duration and F0 proportion. The German listeners, in contrast, are not affected by clause boundaries and rely more strongly on the acoustic characteristics of words (syllable duration, intensity difference, pause duration and F0 proportion). Therefore, we decided to investigate some lexical and prosodic characteristics of the chunks that were identified by the German and Estonian listeners. Firstly, we examined the length of the chunks in terms of duration (in milliseconds) and the number of words. There is an idea that chunking processes could be constrained by the capacity of working memory, which has been frequently measured in how many words a person is able to recall (Green 2017). The finding is that the working memory mostly spans from five to seven words (sometimes even nine words, Miller 1956). We speculated that the German listeners might be stronger constrained by the memory capacity than the Estonian listeners because language processing and memory of the Estonian listeners are supported by the semantic and syntactic information that is inaccessible to the German listeners. So, we expected the duration of non-clausal chunks that were perceived by German listeners to conform stronger with the memory constraint than the duration of the clausal chunks that were identified by Estonian listeners. In particular, we expected non-clausal chunks to be shorter and less variable than the clausal chunks.

Secondly, we analysed the lexical constituency of the chunks. Concerning words, we expected that the chunks identified by the Estonian listeners are more likely to begin with conjunctions and the so-called clausal connectors (e.g. *et*, 'that'; *aga*, 'but'; *kui*, 'if/when'; etc.) than the chunks identified by the Germans. This is because conjunctions signal the beginning of a new clause (in our analysis) and only the native speakers have access to this syntactic information. Thus, it is not likely that the German listeners would consistently identify conjunction-initial chunks. Finally, we explored the tonal coherence of the perceptual chunks. As discussed in the Introduction, the tonal coherence can be approximated by the decline of F0 across the respective domain (e.g., IP, clause or perceptual chunk). Thus, we visually estimated the degree of tonal coherence of the perceptual chunks by observing the averaged F0 contours of the native and non-native language chunks. We speculated that the non-native chunks (non-clausal chunks) exhibit tonal coherence to a larger degree than the native chunks (clausal chunks) because the German listeners were stronger guided by the bottom-up prosodic cues than were the Estonian listeners.

3.3 Any shared characteristics between the native and non-native chunks?

We examined the chunks identified by the Estonian and German listeners considering the chunks' length (in duration and number of words, Table 5), lexical characteristics (Table 6) and tonal coherence (Figure 2).

The averages of duration and length in words in Table 5 indicate that the perceptual chunks do not differ in duration or the number of words between the two language groups. In other words, listeners with Estonian and German backgrounds identify chunks of the same length and size. The difference is that the chunks identified by Estonian listeners are more likely to form a clause than the chunks identified by the German listeners.

Table 5: Lengths of chunks in German and Estonian listeners as estimated by duration (ms) and number of words in chunks.

	Av. duration (ms)		Length in words	
Language group	Mean	SD	Mean	SD
Estonian	1452	850	5.85	3.35
German	1417	749	5.86	3.24

For the lexical characteristics in Table 6, we identified words that appeared most frequently in the first, second, third and final positions in the chunks. The aim was to see if the lexical content of the chunks differs between the two language groups.

Table 6 reveals no differences in the lexical constituency of Estonian and German chunks. The word frequencies reflect the nature of spontaneously spoken Estonian, in which the connectors (*et* 'that', *ja* 'and') and the pronouns (*ma* 'I', *see* 'this') have the highest frequency (see Lippus 2019).

Furthermore, we investigated the tonal coherence of the perceptual chunks that were identified by German and Estonian listeners. For this, we extracted F0 contours of each excerpt identified by each listener and categorized them based on their position within the excerpt: (i) at the beginning of the excerpts, that is, first chunk, (ii) following the first chunk, that is second chunk, (iii) at the end of the chunk, that is final, (iv) and all others between the second and the last chunk within the excerpt. There were 3999 three-chunk excerpts (46.9 percent of all the chunked excerpts), 2615 four-chunk excerpts (30.6 percent of all the chunkings) and only 1099 two-chunk excerpts (12.9 percent of all the chunkings). F0 contours of the perceptual chunks were then time-normalized by extracting 32 F0 measures, equally distributed within a respective perceptually identified chunk. The 32 measurements of F0 were then averaged by their position (see Figure 2). The different panels in Figure 2 enable us to follow the decline of F0 in the excerpt-initial chunks, in the chunks of second position, the chunks of excerpt-medial position, and the chunks of the excerpt-final position.

We can observe a continuous decline in F0 over the entire excerpt but also over the chunks identified at the different positions in the excerpts. Tonally, the chunks identified by German and Estonian listeners are comparable, and no major differences occur.

Table 6: The 10 most frequent words in the first, second, third and final positions of chunks identified by Estonian and German listeners. FR: frequency ranking

FR	First pos.		Second pos.		Third pos.		Last pos.	
Estonian								
1	*et*	‘that’	*siis*	‘then’	*on*	‘is’	*et*	‘that’
2	*ja*	‘and’	*ei*	‘no’	*ei*	‘no’	*noh*	‘well, uhm’
3	*siis*	‘then’	*ma*	‘I’	*oli*	‘was’	*on*	‘is’
4	*ma*	‘I’	*see*	‘this’	*et*	‘that’	*siis*	‘then’
5	*aga*	‘but’	*on*	‘is’	*me*	‘we’	*see*	‘this’
6	*see*	‘this’	*oli*	‘was’	*see*	‘this’	*ka*	‘too’
7	*kui*	‘if, when’	*et*	‘that’	*nagu*	‘like’	*oli*	‘was’
8	*või*	‘or’	*ta*	‘(s)he’	*ma*	‘I’	*jah*	‘yes’
9	*ei*	‘no’	*me*	‘we’	*seal*	‘there’	*ja*	‘and’
10	*mingi*	‘some’	*seal*	‘there’	*kui*	‘if, when’	*seda*	‘this [PART]’
German								
1	*et*	‘that’	*siis*	‘then’	*on*	‘is’	*et*	‘that’
2	*ja*	‘and’	*ma*	‘I’	*ei*	‘no’	*see*	‘this’
3	*siis*	‘then’	*ei*	‘no’	*oli*	‘was’	*ja*	‘and’
4	*ei*	‘no’	*et*	‘that’	*et*	‘that’	*siis*	‘then’
5	*on*	‘is’	*see*	‘this’	*see*	‘this’	*nagu*	‘like’
6	*ma*	‘I’	*on*	‘is’	*ma*	‘I’	*mingi*	‘some’
7	*see*	‘this’	*oli*	‘was’	*me*	‘we’	*on*	‘is’
8	*oli*	‘was’	*seal*	‘there’	*kui*	‘if, when’	*seda*	‘this [PART]’
9	*noh*	‘well, uhm’	*ja*	‘and’	*nagu*	‘like’	*oli*	‘was’
10	*aga*	‘but’	*kui*	‘if, when’	*siis*	‘then’	*noh*	‘well, uhm’

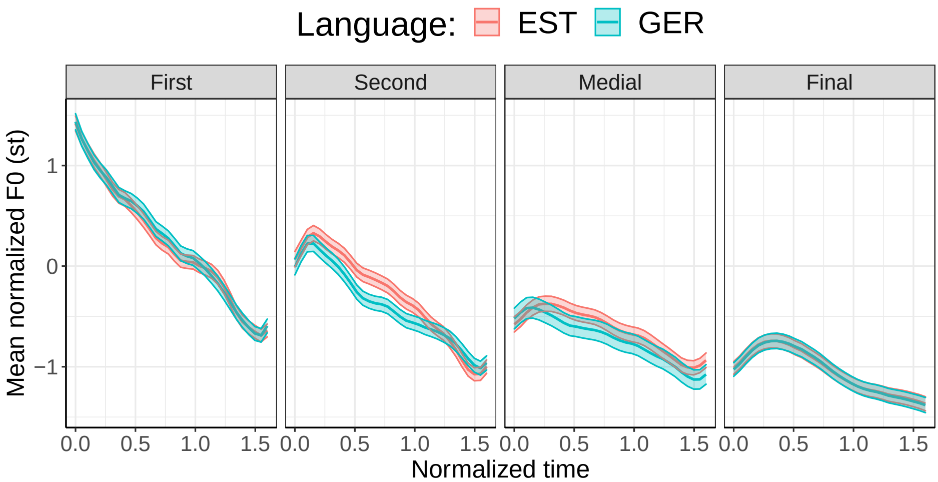

Figure 2: Time-normalized F0 contours converted into semitones (st) by position in excerpts and by language background of listeners. "First" refers to the excerpt-initial position and "Second" to the second position in an excerpt. "Medial" incorporates all other positions except the final position, and "Final" indicates the excerpt-final chunks.

4 Discussion

This study set out to investigate the impact of signal-driven prosodic cues on speech boundary perception in listeners of an unfamiliar language. We employed an RPT experiment in which German listeners unfamiliar with the Estonian language were asked to chunk spontaneous utterances spoken in Estonian. The results of the experiment were compared to the results of a previous experiment in which Estonian listeners were asked to perform a similar task listening to the same speech excerpts (Ots & Taremaa 2022). We examined the duration of word-final syllables, the duration of pauses, intensity curves and F0 (F0 maxima relative to average F0 of respective sentences) as set against the clausal structure at the chunk boundaries identified by German and Estonian listeners.

The results show that German listeners appear to use all the phonetic cues of syllable duration, intensity, pause duration and F0, and to ignore the clausal information. In contrast, Estonian listeners mostly utilize top-down information, as they largely relied on clause boundaries in the chunking task. After the clausal information, Estonian listeners also used the phonetic cues of pause duration and F0 but not syllable duration and intensity difference.

More specifically, the results demonstrate for German listeners that the probability of boundary perception increased together with increasing duration of the word-final syllable. The longer syllable duration corresponding with the chunk boundary resembles the well-known prosodic boundary cue – the pre-boundary or phrase-final lengthening (Berkovits 1994, Fon et al. 2011, Nakai et al. 2009, Petrone et al. 2017, Wightman et al. 1992; for German, see also Schubö & Zerbian 2023 [this volume], Huttenlauch et al. 2023 [this volume], and Wellmann et al. 2023 [this volume]). Thus, it seems that the German listeners are guided by phrase-final lengthening while chunking an unfamiliar language. Furthermore, the analysis indicates that boundary perception became likelier among the German listeners as the intensity difference between the first and last syllable in the word increased. This suggests that German listeners interpreted the intensity drop as an additional cue for a chunk boundary.

Although pauses are usually infrequent in conversational utterances (Biron et al. 2021), they are known to be accessible and reliable cues for boundary perception (Himmelmann et al. 2018, Riesberg et al. 2020, Petrone et al. 2017). In our study, we observe that listeners of both familiar and unfamiliar language backgrounds benefited from the presence of longer, rather than shorter, pauses: the longer the pause, the likelier the perception of a chunk boundary. Similarly, both language groups benefited from variation of F0. However, the tonal cue was interpreted in the opposite direction from what was predicted. The higher the word-initial stressed syllable was relative to the sentence's mean F0, the likelier it was that the listener would perceive a boundary after this word.

At first sight, the result concerning intonation is somewhat puzzling. As an explanation, we consider that the F0 maxima in our materials index a sort of rising boundary tone and not pitch accents. Theoretically, the stressed syllable preceding a rise is pitched low, and the F0 maxima in the final unstressed syllables should index the rise right before the phrase boundary. The lexical makeup of the identified chunks in Table 6 demonstrates that the most frequent words at the ends of chunks were monosyllabic words. Monosyllabic words are considered to carry lexical stress, but they tend to become reduced and unstressed in unaccented positions of spoken utterances (Lehiste 1960: 54). As such, they may well serve as carriers of the phrase-final tonal rise. Thus, our listeners, irrespective of their language background, interpreted increasing F0 contour as a cue for a chunk boundary. As such, the result corroborates the findings in Petrone et al. (2017) and in Kentner & Féry (2013), who have found for German that the F0 in the first and last syllables of phrase-final words at the IP-medial positions is high, that is, the IP-medial phrase-final words have a strong tendency to carry a tonal rise. Our study, together with Petrone et al. (2017), establishes that tonal rises are interpreted as boundary cues also in the perception of spontaneous speech natively and non-natively.

In comparing the two language groups, we discovered that while phrase-final lengthening and intensity drop functioned as boundary cues for German listeners, they did not for Estonian listeners. It is possible that this difference might relate to the differing prosodic profiles of these languages. For example, steep F0 falls accompanied by a deep intensity drop are quite common for German declarative sentences (Peters 1999, Ulbrich 2002). Thus, German listeners might be attuned to hearing large intensity drops accompanied by tonal falls as boundary cues. Similarly, phrase-final lengthening is most frequently attested in German and English. However, the lengthened segments signal the three-way quantity contrast of phonological feet that distinguishes between morpho-lexical functions in Estonian (Eek 1990, Lehiste 1960, 1997). Although the phonological variation of duration does not directly confine the phenomenon of phrase-final lengthening in production, Estonian listeners might nevertheless concentrate on aspects of segmental lengthening differently from German listeners. Thus, the results on intensity drop and pre-boundary lengthening indicate that the crosslinguistic applicability of prosodic boundary cues depends on the prosodic characteristics of the crossed languages.

Clause boundaries, phrase-final lengthening, intensity drop and rising boundary tone performed well in explaining the distribution of boundary marks in the logistic mixed-effects analysis, but the concordance within the two groups of participants showed that the listeners demonstrated only fair agreement in identifying the presence of a boundary. The Fleiss' κ scores compared to the κ scores reported in previous studies were considerably lower (see, e.g., Himmelmann et al. 2018, Riesberg et al. 2020). This holds true especially for the Estonian listeners who attended to their native language. On the one hand, this result might arise from the nature of the materials the participants were asked to listen to. The utterances were extracted from a corpus of dialogues that were held among friends or acquaintances on a freely chosen topic. Although they were recorded in an unnatural recording situation (in a professional sound-attenuated recording studio), these utterances represent highly conversational speech. The low agreement numbers most likely reflect the high acoustic variability characteristic of conversational speech. Also, the selected utterances probably display several different combinations of acoustic boundary cues in which pauses, pause duration, pre-boundary lengthening, intensity drop, and increasing F0 contour are produced at varying strengths. Rising F0 movement is usually accompanied by a decrease in intensity difference. As such, the rising boundary cue might counteract the cue of intensity drop. On the other hand, the low concordances suggest that listeners vary greatly in their cue weighting. For example, Baumann & Winter (2018) found that German listeners in a similar chunking task were divided into two groups: those who attend to pitch-related cues (such as pitch accent type, mean and maximum F0) and those who instead rely on duration and lexical and syntactic information. Most likely, the participants of the experiment made sense of numerous combinations of boundary cues in many different ways, which also explains the low agreement scores.

In the final part of the analysis, we compared the lexical and acoustic characteristics of the speech chunks identified by the German and Estonian listeners. The native and non-native speech chunks displayed a number of shared characteristics. Specifically, the chunks were comparably long in duration and in the number of words. They also displayed very similar lexical variation, common for spontaneous speech in general. More importantly, the average F0 contours demonstrate that the speech chunks identified by both language groups conform to the concept of tonal coherence. Regardless of position in the excerpts, F0 was gradually declining across the native as well as non-native speech chunks. Thus, the chunks identified by the German and Estonian listeners differed from each other neither prosodically nor lexically.

The Estonian chunks, however, corresponded more frequently with the syntactic clauses. To stay within the boundaries of the current study, we must refrain from further examination of the chunks that the German listeners identified. However, we find it very interesting that the German participants clearly found types of speech chunks that are not clauses but show prosodic coherence and high comparability with the clauses detected by native listeners. For future research, we propose to investigate what types of chunks German listeners identify in terms of semantic and pragmatic coherence and whether these could be helpful for language learners when decoding a second language.

Overall, the study provides evidence that the two language groups – German and Estonian listeners – employed longer pauses and rising F0 contour in a speech chunking task. In other words, we have found crosslinguistic application of pausing and F0. As non-native listeners, Germans additionally utilized preboundary lengthening and intensity drop. Thus, while German listeners made use of all acoustic variables we investigated here, Estonian listeners applied only a few of them and relied mainly on the presence of clause boundaries.

We categorized phonetic variables (duration, intensity and F0) as bottom-up information and clause boundaries as top-down information. We predicted less influence from clausal information but more influence from signal-based prosodic information for German listeners than for Estonian listeners. As discussed above, the results support this prediction. As expected, in chunking Estonian speech, German listeners unfamiliar with the Estonian language make use of bottom-up information only, whereas Estonian listeners mostly utilize top-down information, as they largely relied on clause boundaries in the chunking task. This outcome runs counter to the results in Riesberg et al. (2020) and demonstrates that the production of prosody in Estonian spontaneous speech is not too tightly bound to the clausal structure. Nevertheless, the results reflect well on the bottom-up and top-down processing mechanisms.

Clearly, when a listener has no knowledge of a language, prosodic boundary cues are the primary source of information for making sense of speech in an unfamiliar language. Native listeners, however, mainly employ semantic and syntactic knowledge, that is, top-down information, but, as we have seen, benefit from prosodic information as well. We speculate that the role of prosodic information is even greater but in the type of RPT task, it is flooded with semantic and syntactic information which emerges from lexical sources. Therefore, the role of prosody in the earliest stages of spoken language processing might be better established by using more sensitive methods being able to tap into the ongoing decoding processes (see Wellmann et al. 2023 [this volume] for boundary perception in infants). Nevertheless, our study of non-native listeners in comparison to the previous study of native listeners has successfully demonstrated both bottom-up and top-down effects in the processing of spontaneous speech.

We probably see top-down processing somewhat overriding bottom-up processing in native speech processing. This is understandable because top-down processing, together with prediction, is an efficient way to reduce the cognitive load, as it enables one to avoid processing every single aspect of information available in the environment (Bar et al. 2006, Clark 2016, Engel et al. 2001). We believe that the phenomenon of top-down processing also explains the results of previous phonetic perception experiments in which boundary perception in native listeners has been shown to be mediated mainly by syntactic and lexical variables (e.g., Cole et al. 2010, Christodoulides et al. 2018, Baumann & Winter 2018). To demonstrate the impact and functions of bottom-up information – signal-driven prosodic boundary cues in particular – for native listeners, future studies should involve more rigorous research techniques that can assess on-going comprehension.

5 Conclusion

In this study, we investigated the impact of signal-driven prosodic cues on chunking excerpts of a natural language. For this, we utilized RPT methodology and asked non-native listeners (Germans) to identify speech chunks in excerpts spoken in an unfamiliar language (Estonian). We examined the acoustic variation at the boundaries of chunks identified by German listeners with reference to chunk boundaries detected in the same excerpts by Estonian listeners in an earlier experiment. The results show that German listeners, having no access to the semantic-syntactic structure of Estonian, largely rely on signal-driven prosodic information and utilize syllable duration, intensity curves, pause duration and rising F0 contour when dividing a continuous stream of speech into smaller chunks. Estonians, on the contrary, rely mainly on the presence of clause boundaries, but they additionally apply pause duration and rising F0 contour for the identification of speech chunks. The results demonstrate the importance of signal-driven prosodic boundary cues in bottom-up processing of spoken language and highlight the interaction between bottom-up processing (sensory input from speech acoustics) and top-down processing (linguistic knowledge about clause structure) in native speech comprehension.

Acknowledgments

We are extremely thankful to the Estonian volunteers and the German participants who took part in our experiments. We further appreciate the warm and supportive audience of the workshop "Prosodic boundary phenomena" at the 43rd annual meeting of the DGfS (Deutsche Gesellschaft für Sprachwissenschaft), who strongly motivated a study with non-native listeners.

Funding information

This work was supported by research funding awarded to the first author by Fritz Thyssen Stiftung in Germany (10.18.2.040SL, "Planning sentences and sentence intonation cross-linguistically") and by funding from the European Union through the European Regional Development Fund (Centre of Excellence in Estonian Studies) and from the research fund of Kadri, Nikolai, and Gerda Rõuk, both of which were awarded to the second author.

References

Asu, Eva Liina & Francis Nolan. 1999. The effect of intonation on pitch cues to the Estonian quantity contrast. In *Proceedings of the 14th ICPhS*, 1873–1876. https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS1999/papers/p14_1873.pdf.

Bar, Moshe, Karim S. Kassam, Avniel Singh Ingh Ghuman, Jasmine Boshyan, Ann Marie Schmidt, Anders M. Dale, Matti S. Hämäläinen, Ksenija Marinkovic, Daniel L. L. Schacter, Bruce Robert Rosen & Eric Halgren. 2006. Top-down facilitation of visual recognition. *Proceedings of the National Academy of Sciences USA* 103(2). 449–454. DOI: 10.1073/pnas.0507062103.

Bates, Douglas, Martin Mächler, Ben Bolker & Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67. 1–48. DOI: 10.18637/jss.v067.i01.

Baumann, Stefan & Bodo Winter. 2018. What makes a word prominent? Predicting untrained German listeners' perceptual judgments. *Journal of Phonetics* 70. 20–38. DOI: 10.1016/j.wocn.2018.05.004.

Berkovits, Rochele. 1994. Durational effects in final lengthening, gapping and contrastive stress. *Language and Speech* 37. 237–250. DOI: 10.1177/002383099403700302.

Biron, Tirza, Daniel Baum, Dominik Freche, Nadav Matalon, Netanel Ehrmann, Eyal Weinreb, David Biron & Elisha Moses. 2021. Automatic detection of prosodic boundaries in spontaneous speech. *PLOS ONE* 16(5). 1–21. DOI: 10.1371/journal.pone.0250969.

Boersma, Paul & David Weenink. 2019. *Praat: Doing phonetics by computer*. http://www.praat.org/.

Bois, John W. Du, Susanna Cumming, Stephan Schuetze-Coburn & Danae Paolino (eds.). 1992. *Discourse transcription*, vol. 4. Santa Barbara: Department of Linguistics, University of California.

Breen, Mara, Laura C. Dilley, John Kraemer & Edward Gibson. 2012. Inter-transcriber reliability for two systems of prosodic annotation: ToBI (Tones and Break Indices) and RaP (Rhythm and Pitch). *Corpus Linguistics and Linguistic Theory* 8(2). 277–312. DOI: 10.1515/cllt-2012-0011.

Buhmann, Jeska, Johanneke Caspers, Vincent J. van Heuven, Heleen Hoekstra, Jean-Pierre Martens & Marc Swerts. 2002. Annotation of prominent words, prosodic boundaries and segmental lengthening by non-expert transcribers in the spoken Dutch corpus. In *Proceedings of the third international conference on Language Resources and Evaluation (LREC'02)*. http://www.lrec-conf.org/proceedings/lrec2002/pdf/96.pdf.

Chafe, Wallace. 1987. Cognitive constraints on information flow. In Russell S. Tomlin (ed.), *Coherence and Grounding in Discourse: Outcome of a Symposium* (Typological Studies in Language 11), 21–51. Amsterdam: Benjamins. DOI: 10.1075/tsl.11.03cha.

Christiansen, Morten H. & Nick Chater. 2016. The Now-or-Never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences* 39. e62. DOI: 10.1017/S0140525X1500031X.

Christodoulides, George, Anne Catherine Simon & Ivana Didirkova. 2018. Perception of prosodic boundaries by naive and expert listeners in French. Modelling and automatic annotation. In *Proceedings of Speech Prosody*, 641–645. DOI: 10.21437/SpeechProsody.2018-130.

Clark, Andy. 2016. *Surfing uncertainty: Prediction, action, and the embodied mind*. Oxford: Oxford University Press. DOI: 10.1093/acprof:oso/9780190217013.001.0001.

Cole, Jennifer, Yoonsook Mo & Soondo Baek. 2010. The role of syntactic structure in guiding prosody perception with ordinary listeners and everyday speech. *Language and Cognitive Processes* 25(7–9). 1141–1177. DOI: 10.1080/01690960903525507.

Cole, Jennifer, Yoonsook Mo & Mark Hasegawa-Johnson. 2011. Signal-based and expectation-based factors in the perception of prosodic prominence. *Laboratory Phonology* 1(2). 425–452. DOI: 10.1515/labphon.2010.022.

Cooper, William E. & John M. Sorensen. 1981. *Fundamental frequency in sentence production*. New York: Springer.

Couper-Kuhlen, Elizabeth. 2001. Intonation and discourse: Current views from within. In Deborah Schiffrin, Deborah Tannen & Heidi E. Hamilton (eds.), *The handbook of discourse analysis*, 13–34. Malden, MA: Blackwell. DOI: 10.1002/9780470753460.ch2.

Cutler, Anne, Delphine Dahan & Wilma van Donselaar. 1997. Prosody in the comprehension of spoken language: A literature review. *Language and Speech* 40(2). 141–201. DOI: 10.1177/002383099704000203.

Dahan, Delphine & Fernanda Ferreira. 2019. Language comprehension: Insights from research on spoken language. In Peter Hagoort (ed.), *Human language: From genes and brains to behavior*, 21–33. Cambridge, MA: MIT Press.

de la Cruz-Pavía, Irene, Janet F. Werker, Eric Vatikiotis-Bateson & Judit Gervain. 2019. Finding phrases: The interplay of word frequency, phrasal prosody and co-speech visual information in chunking speech by monolingual and bilingual adults. *Language and Speech* 63(2). 264–291. DOI: 10.1177/0023830919842353.

Duez, Danielle. 1985. Perception of silent pauses in continuous speech. *Language and Speech* 28(4). 377–388. DOI: 10.1177/002383098502800403.

Eek, Arvo. 1990. Units of temporal organisation and word accents in Estonian. *Linguistica Uralica* 26(4). 251–263.

Engel, Andreas K., Pascal Fries & Wolf Singer. 2001. Dynamic predictions: Oscillations and synchrony in top-down processing. *Nature Reviews Neuroscience* 2(10). 704–716. DOI: 10.1038/35094565.

Erelt, Mati & Helle Metslang. 2017. *Eesti keele süntaks [eng. Estonian syntax]* (Eesti keele varamu). Tartu: Tartu Ülikooli Kirjastus.

Féry, Caroline & Shinichiro Ishihara. 2009. How focus and givenness shape prosody. In Malte Zimmermann & Caroline Féry (eds.), *Information structure: Theoretical, typological, and experimental perspectives*, 36–63. Oxford: Oxford University Press.

Fon, Janice, Keith Johnson & Sally Chen. 2011. Durational patterning at syntactic and discourse boundaries in Mandarin spontaneous speech. *Language and Speech* 54(Pt 1). 5–32. DOI: 10.1177/0023830910372492.

Fujisaki, Hiroya. 1983. Dynamic characteristics of voice fundamental frequency in speech and singing. In Peter F. MacNeilage (ed.), *The productions of speech*, 39–55. Heidelberg: Springer.

Fujisaki, Hiroya & Keikichi Hirose. 1982. Modelling the dynamic characteristics of voice fundamental frequency with applications to analysis and synthesis of intonation. In *Preprint of papers. Working group on intonation, 13th International Congress of Linguistics*, 57–70.

Green, Clarence. 2017. Usage-based linguistics and the magic number four. *Cognitive Linguistics* 28(2). 209–237. DOI: 10.1515/cog-2015-0112.

Halliday, Michael A. K. 1967. Notes on transitivity and theme in English: Part 2. *Journal of Linguistics* 3(2). 199–244. DOI: 10.1017/S0022226700016613.

Himmelmann, Nikolaus P., Meytal Sandler, Jan Strunk & Volker Unterladstetter. 2018. On the universality of intonational phrases: A cross-linguistic interrater study. *Phonology* 35(2). 207–245. DOI: 10.1017/S0952675718000039.

Huttenlauch, Clara, Marie Hansen, Carola de Beer, Sandra Hanne & Isabell Wartenburger. 2023. Age effects on linguistic prosody in coordinates produced to varying interlocutors: Comparison of younger and older speakers. In Fabian Schubö, Sabine Zerbian, Sandra Hanne & Isabell Wartenburger (eds.), *Prosodic boundary phenomena*, 157–192. Berlin: Language Science Press. DOI: 10.5281/zenodo.7777534.

Keating, Patricia A. & Stefanie Shattuck-Hufnagel. 2002. A prosodic view of word form encoding for speech production. *UCLA Working Papers in Phonetics* 101. 112–156. https://escholarship.org/uc/item/1qf5f44k.

Kentner, Gerrit & Caroline Féry. 2013. A new approach to prosodic grouping. *The Linguistic Review* 30. 277–311. DOI: 10.1515/tlr-2013-0009.

Ladd, D. Robert. 1988. Declination “reset” and the hierarchical organization of utterances. *The Journal of the Acoustical Society of America* 84(2). 530–544. DOI: 10.1121/1.396830.

Ladd, D. Robert. 2008. *Intonational phonology*. Cambridge: Cambridge University Press. DOI: 10.1017/cbo9780511808814.

Landis, J. Richard & Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33(1). 159–174. DOI: 10.2307/2529310.

Langus, Alan, Erika Marchetto, Ricardo Augusto Hoffmann Bion & Marina Nespor. 2012. Can prosody be used to discover hierarchical structure in continuous speech? *Journal of Memory and Language* 66(1). 285–306. DOI: 10.1016/j.jml.2011.09.004.

Lehiste, Ilse. 1960. Segmental and syllabic quantity in Estonian. *American Studies in Uralic Linguistics*. Uralic and Altaic series (1). 21–82.

Lehiste, Ilse. 1997. Search for phonetic correlates in Estonian prosody. In Ilse Lehiste & Jaan Ross (eds.), *Estonian prosody: Papers from a symposium*, 11–35. Tallinn: Institute of Estonian Language.

Liberman, Mark Y. & Janet B. Pierrehumbert. 1984. Intonational invariance under changes in pitch range and length. In Mark Aronoff, Richard T. Oehrle, Frances Kelley & Bonnie W. Stephens (eds.), *Language and sound structure: Studies in phonology presented to Morris Halle by his teacher and students*, 157–233. Cambridge, MA: MIT Press.

Lippus, Pärtel. 2019. *Foneetikakorpuse sagedussõnastik*. DOI: 10.15155/re-62.

Lippus, Pärtel, Tuuli Tuisk, Nele Salveste & Pire Teras. 2016. *Phonetic corpus of Estonian spontaneous speech v.1.0.0*. DOI: 10.15155/1-00-0000-0000-0000-00074L.

Mahrt, Tim. 2016. *LMEDS: Language markup and experimental design software*. Computer Program.

Mattys, Sven L., Laurence White & James F. Melhorn. 2005. Integration of multiple speech segmentation cues: A hierarchical framework. *Journal of Experimental Psychology. General* 134(4). 477–500. DOI: 10.1037/0096-3445.134.4.477.

Miller, George A. 1956. The magical number seven plus or minus two: Some limits on our capacity for processing information. *Psychological Review* 63(2). 81–97. DOI: 10.1037/h0043158.

Nakai, Satsuki, Sari Kunnari, Alice E. Turk, Kari Suomi & Riikka Ylitalo. 2009. Utterance-final lengthening and quantity in Northern Finnish. *Journal of Phonetics* 37. 29–45. DOI: 10.1016/j.wocn.2008.08.002.

Nash, John C. 2014. On best practice optimization methods in R. *Journal of Statistical Software* 60(2). 1–14. DOI: 10.18637/JSS.V060.I02.

Nash, John C. & Ravi Varadhan. 2011. Unifying optimization algorithms to aid software system users: optimx for R. *Journal of Statistical Software* 43(9). 1–14. DOI: 10.18637/JSS.V043.I09.

Ordin, Mikhail, Leona Polyanskaya, Itziar Laka & Marina Nespor. 2017. Cross-linguistic differences in the use of durational cues for the segmentation of a novel language. *Memory and Cognition* 45(5). 863–876. DOI: 10.3758/s13421-017-0700-9.

Ots, Nele & Piia Taremaa. 2022. A perceptual study of language chunking in Estonian. *Open Linguistics* 8(1). 1–26. DOI: 10.1515/opli-2020-0182.

Peters, Benno. 1999. *Prototypische Intonationsmuster in deutscher Lese- und Spontansprache*. Magister Thesis, Universität Kiel. 1–171. https://www.ipds.uni-kiel.de/kjk/pub_exx/bp1999_1/aipuk34_bp-ma.pdf.

Petrone, Caterina, Hubert Truckenbrodt, Caroline Wellmann, Julia Holzgrefe-Lang, Isabell Wartenburger & Barbara Höhle. 2017. Prosodic boundary cues in German: Evidence from the production and perception of bracketed lists. *Journal of Phonetics* 61. 71–92. DOI: 10.1016/j.wocn.2017.01.002.

Pierrehumbert, Janet B. 1979. The perception of fundamental frequency declination. *Journal of Acoustical Society of America* 66(2). 363–369. DOI: 10.1121/1.383670.

Pierrehumbert, Janet B. 1980. *The phonology and phonetics of English intonation*. Massachusetts Institute of Technology. (Doctoral dissertation). http://dspace.mit.edu/handle/1721.1/16065.

Pierrehumbert, Janet B. & Julia Bell Hirschberg. 1990. The meaning of intonational contours in the interpretation of discourse. In Philip R. Cohen, Jerry Morgan & Martha E. Pollack (eds.), *Intentions in communication*, 271–312. Cambridge, MA: MIT Press. DOI: 10.7916/D8KD24FP.

R Core Team. 2018. *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. https://www.R-project.org/.

Riesberg, Sonja, Janina Kalbertodt, Stefan Baumann & Nikolaus P. Himmelmann. 2020. Using Rapid Prosody Transcription to probe little-known prosodic systems: The case of Papuan Malay. *Laboratory Phonology* 11(1)(8). 1–35. DOI: 10.5334/labphon.192.

Schild, Ulrike, Angelika B. C. Becker & Claudia K. Friedrich. 2014. Phoneme-free prosodic representations are involved in pre-lexical and lexical neurobiological mechanisms underlying spoken word processing. *Brain and Language* 136(100). 31–43. DOI: 10.1016/j.bandl.2014.07.006.

Schubö, Fabian & Sabine Zerbian. 2023. The patterns of pre-boundary lengthening in German. In Fabian Schubö, Sabine Zerbian, Sandra Hanne & Isabell

Wartenburger (eds.), *Prosodic boundary phenomena*, 1–34. Berlin: Language Science Press. DOI: 10.5281/zenodo.7777526.

Schuetze-Coburn, Stephan, Marian Shapley & Elizabeth G. Weber. 1991. Units of intonation in discourse: A comparison of acoustic and auditory analyses. *Language and Speech* 34(3). 207–234. DOI: 10.1177/002383099103400301.

Silbert, Lauren J., Christopher J. Honey, Erez Simony, David Poeppel & Uri Hasson. 2014. Coupled neural systems underlie the production and comprehension of naturalistic narrative speech. *Proceedings of the National Academy of Sciences USA* 111(43). E4687–E4696. DOI: 10.1073/pnas.1323812111.

Simon, Anne Catherine & George Christodoulides. 2016. Perception of prosodic boundaries by naïve listeners in French. In *Proceedings of Speech Prosody*, 1158–1162. DOI: 10.21437/SpeechProsody.2016-238.

Thorsen, Nina Grønnum. 1985. Intonation and text in Standard Danish. *The Journal of the Acoustical Society of America* 77(3). 1205–1216. DOI: 10.1121/1.392187.

Trouvain, Jürgen, William J. Barry, Claus Nielsen & Ove Kjeld Andersen. 1998. Implications of energy declination for speech synthesis. In *Proceedings of the 3rd ESCA/COCOSDA workshop on speech synthesis*, 47–52.

Ulbrich, Christiane. 2002. A comparative study of intonation in three standard varieties of German. In *Proceedings of Speech Prosody*, 671–674. https://www.isca-speech.org/archive/pdfs/speechprosody_2002/ulbrich02_speechprosody.pdf.

Wagner, Michael & Michael McAuliffe. 2019. The effect of focus prominence on phrasing. *Journal of Phonetics* 77. 100930. DOI: 10.1016/j.wocn.2019.100930.

Wellmann, Caroline, Julia Holzgrefe-Lang, Hubert Truckenbrodt, Isabell Wartenburger & Barbara Höhle. 2023. Developmental changes in prosodic boundary cue perception in German-learning infants. In Fabian Schubö, Sabine Zerbian, Sandra Hanne & Isabell Wartenburger (eds.), *Prosodic boundary phenomena*, 119–156. Berlin: Language Science Press. DOI: 10.5281/zenodo.7777532.

White, Laurence, Silvia Benavides-Varela & Katalin Mády. 2020. Are initial-consonant lengthening and final-vowel lengthening both universal word segmentation cues? *Journal of Phonetics* 81. 10098. DOI: 10.1016/j.wocn.2020.100982.

Wightman, Colin W., Stefanie Shattuck-Hufnagel, Mari Ostendorf & Patti J. Price. 1992. Segmental durations in the vicinity of prosodic phrase boundaries. *The Journal of the Acoustical Society of America* 91(3). 1707–1717. DOI: 10.1121/1.402450.

Yuan, Jiahong & Mark Y. Liberman. 2014. F0 declination in English and Mandarin broadcast news speech. *Speech Communication* 65. 67–74. DOI: 10.1016/j.specom.2014.06.001.