THE ROLE OF BOUNDARY TONE IN CHUNKING SPONTANEOUS SPEECH

Lucas Weidinger[1], Xinyuan Wan[1], Liisa Koivusalo[1], Aleksandra Dobrego[1]

[1]University of Helsinki, Finland
lucas.weidinger@helsinki.fi xinyuan.wan@helsinki.fi liisa.koivusalo@helsinki.fi aleksandra.dobrego@helsinki.fi

ABSTRACT

This study aimed to explore how listeners segment spontaneous speech of different nature and what role boundary tone plays in the process. Participants (n = 41) were asked to listen to extracts of scientific and conversational speech while seeing corresponding transcripts on a screen and to segment the speech into chunks by placing markers between orthographic words. No explanation on what chunks are or should be was given to participants. The speech samples were annotated for silent pauses and boundary tones. Agreement on chunks between participants was high, with scientific samples showing stronger agreement than conversational ones (Fleiss' kappa, 0.56 vs 0.396, $p<.05$). Boundary tone appears to be a significant cue in both types of speech. The level of tone did not matter. The study suggests that spontaneous speech processing is affected by boundary tone independently of the nature of speech, and listeners use boundary tones as cues for chunking.

Key words: chunking, spontaneous speech

1. INTRODUCTION

The speech input we encounter every day is mostly spontaneous. It is filled with disfluencies, false starts, hesitations, and repetitions, but they nevertheless do not affect people's ability to effectively understand it. One of the prerequisites for understanding is chunking - the fast and largely automatic process by which the listener determines where one meaningful unit ends and the next begins [1]. What these chunks are, whether listeners agree on them, and what cues mark a chunk, are debatable issues.

Here, speech is studied from the perspective of Linear Unit Grammar (LUG) [2], which assumes that language is linear, unfolds in real time and is segmented into chunks by a listener or a reader. These chunks then can be treated as cohesive building blocks that carry meaning and give the speech structure. Even without a thorough explanation of what chunks are, people tend to intuitively chunk speech in the same way [3][4][5]. This was shown on English [3], Finnish, Russian and Swedish materials [6]. The participants listened to the speech extracts while simultaneously seeing the text on screen and were asked to mark boundaries between chunks (ChunkitApp, [7]). The researchers assessed comprehension, agreement, and segmentation strategies of different groups of speakers. It was found that participants agreed on chunks in their places and size and that their segmentation strategies did not differ. Interestingly, speakers of all languages studied (English/Finnish/Russian/Swedish) and all language levels (native/non-native, [5]) used prosodic cues rather than syntactic when choosing where to place a boundary in speech.

This paper sets out to explore this crucial role of prosody in chunking by looking at boundary tone (a rise or fall in pitch). The initial idea was to see whether boundary tone is important in chunking different kinds of speech. extracts of conversational and scientific speech were chosen, and participants' agreement and segmentation choices were investigated using ChunkitApp. ChunkitApp is a tablet application that displays transcripts and plays corresponding extracts simultaneously. The agreement served as a method validation: if, as in previous studies, the agreement was high, then this method did capture intuitive segmentation that could be analysed. If the agreement was low, some other method would need to be found to better capture participants' choices in segmentation.

It was expected that the boundary tone would be a significant indicator for scientific speech, but not for conversational speech. This hypothesis stems from the expectation that the prosody of the more formal, scientific speech is more premeditated than the prosody of the purely conversational speech. When the structure and prosody of the uttered sentence are planned slightly ahead of time, chunk boundary markers, like changes in pitch for example, would be placed more deliberately by the speaker. It was expected to be the case for what is here described as *scientific speech*: recordings of lectures, in which a lecturer knows what they are going to say ahead of time.

However, it was found that there was no difference between two types of speech, spontaneous and scientific, and the boundary tone was a reliable cue in chunking both. This suggests that the role of pitch fluctuations in processing spontaneous speech is independent of the setting and the purpose of the talk: Whether it is a predominantly monologic lecture or a conversation, dialogistic in nature.

2. METHODS

2.1. Experimental setup

The experiment was conducted using ChunkitApp [7]. Participants listen to the extract and marks chunk boundaries by pressing on a tilde symbol (~) between orthographic words in the transcript on screen. When pressed, the tilde symbol turns into a vertical line (|). This is then considered a boundary. After each extract, participants must answer a comprehension question. Participants were free to place boundaries wherever they felt they belonged, and all boundary placements were considered valid for analysis, given high enough agreement between participants on the location of the boundary. They received no instructions as to how many and where to place boundaries, with the exception that they should place at least one boundary in each extract.

The stimuli were extracts of spontaneous speech and their transcripts. They were taken from the AE minicorpus [8] in case of the conversational speech, while the scientific speech samples were taken from open-source lecture recordings. A full list of the sources for the "scientific speech" samples can be found in the Appendix. 20 stimuli of each type, 40 in total, were included in the experiment. Additionally, one test stimulus and two training stimuli were used to familiarise the participants with the task. The durations of the stimuli ranged from 10 to 15 seconds, and they consist of between one and three sentences. All extracts were in English.

The transcripts were created manually to get them to reflect the speech samples as accurately as possible. An automatic speech-to-text system was found to be unsuitable for this task, as machine generated transcripts contained too many errors. They omitted some of the information specifically needed for this experiment, like repetitions and non-starts. It was ultimately faster and easier to transcribe by hand than to fix the machine generated transcripts.

ChunkitApp was also used to collect informed consent and background information of participants. Participants were asked about their education, language background, proficiency in English, and whether they had diagnosed reading or hearing disorders. Questions regarding native languages, proficiency in English, and reading and hearing disorders were obligatory to be able to control the participant group. Personal information, such as participants' names or voices, was not collected, and the participants could not be identified from their answers in any phase. Participants could take part in the study using their own device and a pair of headphones. They were allowed to perform the experiment at their own discretion and withdraw at any time.

2.2. Participants

Participants were recruited through the University of Helsinki Faculty of Arts mailing list as well as social media. The total number of recruited participants was 89, however, only results from 41 participants were used in the final analysis (see 3.1 Outliers). All of them were naïve to the purpose of experiment. Prior to the main task, participants were asked to self-report their English skills in terms of the six CEFR levels of English proficiency for adults (A1-C2). The description of these were given to the participants as a guideline to help them self-report their own English skills. All participants included in the analysis reported at least an upper intermediate level of English skill (CEFR B2). Five participants were English native speakers, 36 were L2- English speakers. The distribution of self-reported English skills is shown below. Participants were asked to self-report their English skills on a six-point Likert-scale. Figure 1 shows that 34 participants indicated their overall English skills as either advanced (5) or proficient (6).

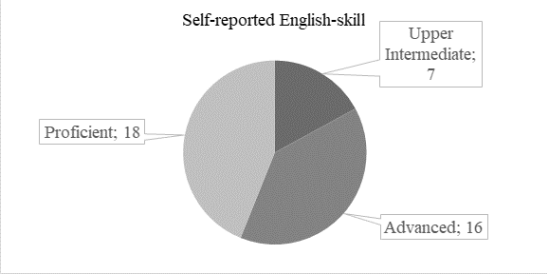

Figure 1: Participants' self-reported English skills, given in CEFR levels

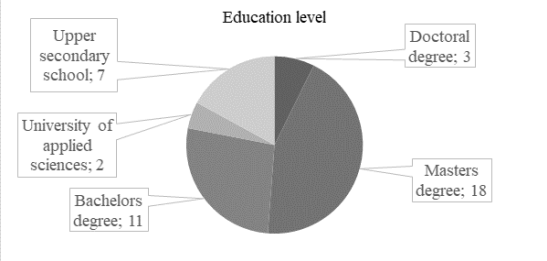

Figure 2: Participants' highest completed level of education

Participants were also asked to report the highest level of education they completed. Most of the participants (34 people) had some description of a tertiary degree. The exact distribution of completed degrees by participants can be seen in Figure 2.

2.3. Annotations

All speech samples were annotated using three tiers in Praat [9]: (1) orthographic, (2) break index, and (3) tonal. The orthographic tier was an interval tier, the others point tiers. On the orthographic tier, all utterances were marked using the transcripts. These utterances were generally separated by pauses; this made annotation of the extracts easier. Silent pauses 50-250 ms (SP) and above 250 ms (LSP) in duration were marked on the break index tier. On the tonal tier, either a high (H%) or a low (L%) tone was marked, based on the F0 pitch just before a silent pause. The full scheme for the annotations can be seen in Table 1.

The annotations were done by two annotators independently before being reviewed by all four authors.

Tier	Annotations		Count
Break index	silent pause	SP	65
	long silent pause	LSP	114
Tonal	high	H%	65
	low	L%	227

Table 1: Annotation scheme

3. RESULTS

3.1. Outliers

Participants were excluded from the analysis based on their background and the minimum number of boundaries they had put in total.

89 participants were recruited. 31 participants were excluded due to their education background in linguistics, applied linguistics and some other programs with a focus on language studies. Recruiting linguists for the task was avoided since they often have an elaborate knowledge of language science, and they could stand out from the target sample. Additionally, 4 participants were excluded because they were diagnosed with dyslexia and hearing disorders. 13 participants had put less than 40 boundaries per 40 extract, and were excluded as well. Since the task was to put at least one boundary in each extract, in their case the task was considered incomplete. Therefore, in total, 41 users were included in the final analysis.

3.2. Analysis of agreement

All following analysis was done using programming language R version 4.0.5 and RStudio [10][11]. The resulting boundary data from ChunkitApp was a data frame with all possible boundaries as rows and participants as columns. A boundary was marked with 1 and no boundary with 0. Agreement was used as a measure to ensure the observed behaviour was not just random but was motivated for all participants with certain factors. Agreement was calculated similarly to [3]. Fleiss' kappa [12] was chosen as a measure to assesses the reliability of agreement between three or more raters. The measure ranges from 0 to 1, with 0 indicating no agreement and 1 indicating absolute agreement. Several studies have shown that Fleiss' kappa values are comparable across populations and experimental conditions [3][12], and results from previous studies show that it is conceptually possible to use Fleiss' kappa to investigate participants' agreement rate [3][4][5]. It was found that Fleiss' kappa for the boundary agreement on conversational speech is 0.396 ($p < .05$), and Fleiss kappa for the agreement on scientific speech is 0.56 ($p < .05$). The result suggests that participants agree on boundaries in scientific speech more, since probably conversational speech has more variation in syntactic and prosodic structure.

As there was some agreement for both types of speech, it was valid to proceed with investigating the boundary tone.

3.3. Analysis of boundary tone

This analysis was done to understand how the boundary tone influences the segmentation choices in different types of speech. The hypothesis was that boundary tone will be the predominant cue for segmentation in scientific speech, whereas for conversational speech the boundary tone will be not significant.

The dependent variable was boundary frequency, i.e., how many people marked a boundary in a particular place. The independent variable was the boundary tone (high / low / none).

Boundary tone was fitted to a binomial logistic regression with mixed effects (fixed + random intercept [extract]). The full model included one fixed effect: boundary tone as a factor with three levels (high, low and none), with no tone as reference level; and extract as random intercept. The model was fitted using the lme4 package [13], separately for each kind of speech. The summaries of the model coefficients can be seen in Table 2 (conversational) and Table 3 (scientific).

	Estimate	Std. Error	z value	Pr (>|z|)
(Intercept)	-6.443	1.001	-6.438	1.21e-10 ***
tonalH	6.063	1.064	5.701	1.19e-08 ***
tonalL	5.125	1.022	5.017	5.26e-07 ***

Table 2: Logistic regression coefficients for conversational speech.

	Estimate	Std. Error	z value	Pr (>|z|)
(Intercept)	-4.9819	0.5017	-9.930	< 2e-16 ***
tonalH	4.9213	0.6108	8.057	7.79e-16 ***
tonalL	4.8641	0.5468	8.896	< 2e-16 ***

Table 3: Logistic regression coefficients for scientific speech.

The regressions revealed a significant influence of boundary tone on placing a boundary, whether it is high, low, or none, for both scientific and conversational speech. TukeyHSD comparison of means was used to assess the differences in influence between these levels. The results are in Table 4 (conversational) and Table 5 (scientific).

	Estimate	Std. Error	z value	Pr (>|z|)
H - NONE == 0	6.0632	1.0635	5.701	< 1e-04 ***
L - NONE == 0	5.1253	1.0217	5.017	< 1e-04 ***
L - H == 0	-0.9378	0.4145	-2.262	0.0549

Table 4: TukeyHSD for conversational speech.

The results of TukeyHSD show that in both types of speech, boundary frequency is significantly affected by the presence [high and low] and absence [none] of the boundary tone. However, there are no differences between the low and high tones, i.e., it does not matter how prominent the boundary tone is to place a boundary.

	Estimate	Std. Error	z value	Pr (>|z|)
H - NONE == 0	4.92127	0.61078	8.057	< 1e-06 ***
L - NONE == 0	4.86411	0.54676	5.896	< 1e-06 ***
L - H == 0	-0.05716	0.41054	-0.139	0.989

Table 5: TukeyHSD for scientific speech.

4. DISCUSSION

Current study aimed to investigate the role of boundary type in chunking conversational and scientific speech. Extracts of spoken speech were played back to participants who were asked to mark segments in corresponding transcripts shown on a screen using ChunkitApp. Boundary tone (rise and fall in pitch) was found to be an important cue for pacing a boundary marker in both types of speech. This suggests that spontaneous speech processing is affected by the boundary tone independently of the nature of speech. Besides, it did not matter how strong the boundary tone is - if it is present at all, a listener would most probably mark it as a boundary, i.e., the beginning or the end of a chunk.

This study, of course, has its limitations. First, the annotation convention is largely based on the ToBI annotation system which marks the boundary tone if the word carrier is at the end of an intonational phrase or is followed by a detectable pause (i.e., BI No.4 in ToBI). Therefore, it is not certain that it is the boundary tones that participants used as the predominant cues for chunking since boundary tone and pause are always shown up together. A different annotation convention could be proposed to explore the roles of pauses versus boundary tones in chunking.

Secondly, theories in intonational phonology and discourse analysis often consider L-tone (i.e., L%, low static or falling pitch contours) as the default boundary tone. In turn, H-tone (i.e., H%, high static or rising pitch contours) is usually marked for indicating a question or the continuation of speech with pauses in between during turn taking in conversation. Due to this function of H%, it was initially hypothesised that participants were more likely to mark a boundary when it is associated with a H%, rather than L%. However, the difference between them was not found to be significant. An interesting continuation of this study would be to have the same task conducted the by musicians or linguists who regularly have much more experience with pitch fluctuations.

Finally, apart from acoustic cues, other linguistic factors such as words with indications may also be used as important cues in chunking. A potential future study could be to explore whether and how lexical indicators are used as boundary markers.

5. ACKNOWLEDGEMENTS

This study was conducted as a part of an Experimental Laboratory course at the University of Helsinki, Finland. We thank the Department of Digital Humanities and the course leader Juraj Šimko for the environment that inspired this study.

6. REFERENCES

[1] Christiansen, M. H. & Chater, N. (2016). The now-or-never bottleneck: A fundamental constraint on language. *The Behavioral and Brain Sciences, 39*, E62.
[2] Sinclair, J. M., & Mauranen, A. (2006). Linear unit grammar: Integrating speech and writing (Vol. 25). John Benjamins Publishing.
[3] Vetchinnikova, S., Konina, A., Williams, N., Mikušová, N., & Mauranen, A. (2022). Perceptual chunking of spontaneous speech: Validating a new method with non-native listeners. Research Methods in Applied Linguistics, 1(2), 100012.
[4] Anurova, I., Vetchinnikova, S., Dobrego, A., Williams, N., Mikusova, N., Suni, A., ... & Palva, S. (2022). Event-related responses reflect chunk boundaries in natural speech. *NeuroImage, 255*, 119203.
[5] Dobrego, A., Konina, A., & Mauranen, A. (2022). Continuous speech segmentation by L1 and L2 speakers of English: the role of syntactic and prosodic cues. *Language Awareness*, 1-21.
[6] Konina, A., Dobrego, A., Mauranen, A. (forthc.) Natural speech segmentation strategies: cross-linguistic evidence.
[7] Vetchinnikova, S., Mauranen, A., & Mikušová, N. (2017). ChunkitApp: Investigating the relevant units of online speech processing. In INTERSPEECH 2017: 18th Annual Conference of the International Speech Communication Association. pp. 811– 812. Interspeech 2017, Stockholm, Sweden.
[8] Cavalcante, F. A. & Ramos, A. C. (2016). The American English spontaneous speech minicorpus: Architecture and comparability. *CHIMERA. Romance Corpora and Linguistic Studies 3.2*, 99–124
[9] Boersma, P. & Weenink, D. (2021). Praat: doing phonetics by computer [Computer program]. Version 6.1.55, retrieved 25 October 2021 from http://www.praat.org/
[10] R Core Team. (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
[11] RStudio Team. (2022). RStudio: Integrated Development Environment for R. RStudio, PBC, Boston, MA. URL http://www.rstudio.com/.
[12] Fleiss, D. J., DiGiovanni, C. W., Sangeorzan, B. J., Hansen Jr, S. T., Kuo, R., Tejwani, N., & Price, R. (2003). Diagnostic tests for gastrocnemius tightness. *JBJS, 85*(4), 760.
[13] Bates, D., Mächler, M., Bolker, B. M., Walker, S. C. (2014). Fitting linear mixed-effect models using lme4, *Journal of Statistical Software*

7. APPENDIX

Link to the document containing links to all sources for scientific, practice and training stimuli:

https://drive.google.com/file/d/19dzSonA9aujUj1FMyKUhHg_KowzZDhns/view?usp=share_link