THE ROLE OF RHYTHMIC CHUNKING IN SPEECH: SYNTHESIS OF FINDINGS AND EVIDENCE FROM STATISTICAL LEARNING

Annie C Gilbert[a,b], *Victor J. Boucher*[a] *& Boutheina Jemel*[b,c]

[a]Laboratoire de Sciences Phonétique, Université de Montréal, Canada;
[b]Laboratoire de Recherche en Neurosciences et Électrophysiologie Cognitive, Hôpital Rivière-des-Prairies, Canada;
[c]Centre de Recherche Fernand-Séguin, Département de Psychiatrie, Université de Montréal, Canada
annie.gilbert@umontreal.ca; www.phonetique.info

ABSTRACT

Our presentation summarizes evidence showing that listeners chunk speech in terms of rhythm groups. We discuss previous work involving both behavioral and EEG observations, which suggest an on-line segmentation of speech in rhythmic groups. A brief experiment is presented that further supports the view that statistical learning effects operate by reference to rhythmic chunks.

Keywords: prosody; speech chunking; speech segmentation; rhythm groups; statistical learning

1. INTRODUCTION

A central issue in language acquisition research is to find out how language learners manage to extract and learn forms like lexemes from the speech stream. These operations imply that heard speech must first be segmented into chunks that conform to constraints on memory and attention. But no universal segmentation cue has yet been found. Previous research has focused on different segmentation indices (for a review: Christophe, et al. (2003) [3]). One popular approach refers to probabilistic aspects of sound distributions and "statistical learning". On this view, numerous reports using small artificial languages have demonstrated that listeners are sensitive to transitional probabilities (TPs) between speech sounds and that these TPs serve the detection of the boundaries of "artificial words" (AWs). TPs between two sounds or syllables, *x* and *y*, are calculated by dividing the number of occurrences of *x* immediately preceding *y* by the total number of occurrences of *x*.

TP = no of occ.of sequence *xy*/total no. occ. of *x* (1)

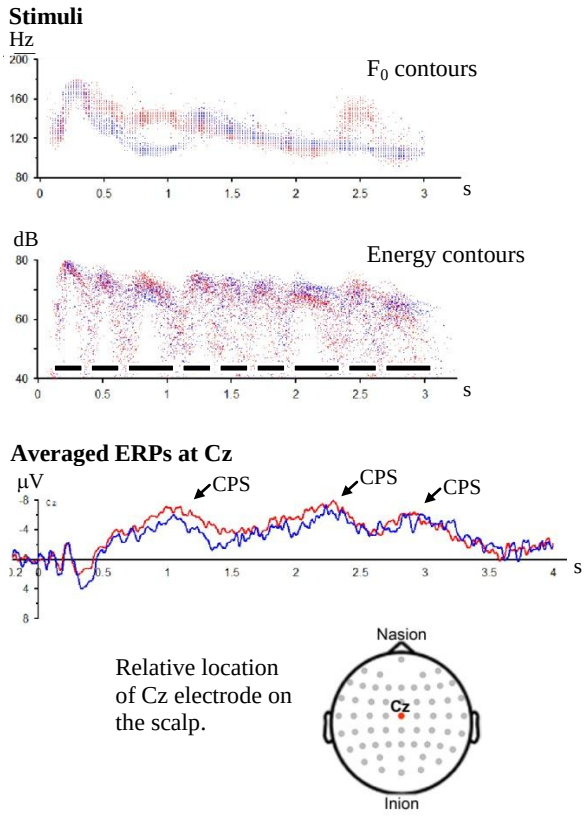
The higher the TP between two elements, the more likely these two belong to the same unit (such as a lexeme). Thus, a TP of 1 means that *x* is *always* followed by *y*. However, few studies have examined the joint effect of prosodic structures and TPs. Shukla, et al. (2007) [9] explored the effects of placing AWs with characteristic TPs within and across intonation phrases (IPs). Interestingly, their results showed that when AWs straddle the boundaries of IPs they are not recognized. Another study by Christophe, et al. (2003) [3] reported that prosodic groups marked by a final lengthening are used on-line by listeners in locating the boundaries of forms. In particular, it was shown that listeners (adults and infants) do not attempt lexical access on elements straddling a group marked by final lengthening. This suggests that the speech stream is initially chunked using lengthening marks. However, it is not clear whether these length marks relate to IPs or smaller frames. On the other hand, it should be recognized that *learning novel linguistic forms requires serial memory of speech sounds and that this imposes a frame of a given size*. Therefore, whatever this chunk is, it must follow certain natural constraints on serial memory. Our observations show that the chunking is associated to rhythmic groups (RGs) and not to intonation units like IPs. A RG is characterized as a series of (generally) 3 or 4 syllables that are bordered by a final long syllable [13]. These groups occur independently of word stress and appears cross-linguistically, for instance in the recall of digits like phone numbers [7].

1.1. Neurological and behavioural evidence of the rhythmic chunking of speech

It is well known in memory research that rhythm chunks facilitate recall (e.g. [2]). Furthermore, Boucher (2006) [1] demonstrated that speakers tend to produce utterances with RGs that match those that facilitate recall (i.e., groups rarely exceed 4 syllables). Other experiments attempted to determine whether listeners attend to rhythm or

intonation groups in segmenting speech. In one study, listeners were asked to recall rhythm and intonation of repeated syllables. It was found that the participants privileged the reproduction of rhythm over intonation [5]. A subsequent experiment made use of event-related potentials (ERPs) and showed that the perception of the lengthened syllable marking the end of a RG elicits a *Closure Positive Shift* (CPS), [6] a neuro-physiological response previously associated with the perception of IPs [8, 10-12]. See Figure 1.

Figure 1 : Adapted from Gilbert, et al. (2010) [6] Stimuli and averaged ERPs at Cz. Note that the CPSs conform to the number of RGs marked by lengthenings in the stimuli.



1.2. The present study

The above findings suggest that RGs play a central role in the chunking of heard speech and may well constitute an essential framework for statistical learning. To demonstrate this, we used a paradigm adapted from Shukla, et al. (2007) [9] in which we manipulated RGs independently of IPs. As noted, previous studies demonstrate that listeners use TPs in learning AWs. In their paper, Shukla and colleagues [9] showed that listeners are able to pick-up on the TPs of AWs words presented *within* an intonation phrase (IP), but not when the AW straddles an IP boundary. However, if the initial or first-pass chunking of speech is based on RGs, as we hypothesize, then we should get similar results regarding the effects of TPs within RGs. Specifically, we would predict diminished recognition of AWs that straddle a RG boundary compared to AWs appearing within a RG.

2. METHODS

2.1. Participants

20 native speakers of French were recruited at the Université de Montréal (19 to 41 years of age; average 25.5 years; 7 men). All were right-handers and presented normal hearing in terms of a standard audiometric evaluation. Also, all participants presented a normal memory span according to the digit span test of the WAIS [14] (overall, average normalized score: 10.16, std dev.: 2.4).

2.2. Stimuli

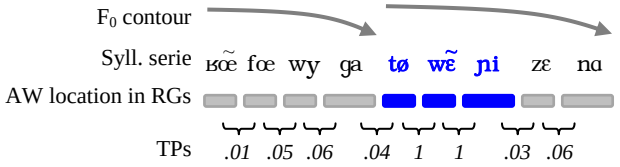
2.2.1. Stimuli design

The design reflected a test in two phases: participants first listen to series (a learning phase) and then are asked, in a separate forced-choice task, to identify recognized AWs. We used a limited set of French non-sense CV syllables in elaborating carrier sequences of 9 syllables and two AWs of 3 syllables. In all, 60 target series and 560 filler series were created each with three RGs. The central manipulation consisted of placing one of the AWs within a RG and the other across a RG boundary, as illustrated in Figure 2. Both rhythm and intonation of the sequences was controlled using a particular procedure (see 2.2.2. *Stimuli recording*). Every target series was presented twice during the experiment so that both AWs were heard a total of 60 times each.

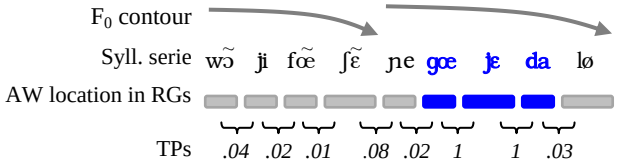
We applied several additional control measures in elaborating the sequences. In particular, the sequences were varied such that no consecutive syllables shared a common point of articulation or repeated vowels (to prevent confounding effects on recognition recall). Further, no sequence contained recognizable multi-syllabic words. Finally and most important, TPs between syllables in a carrier sequence were kept as low as possible (0.045 on average) whereas the TPs between syllables in an AW was 1 (see Fig. 2).

Figure 2: Schematic representation of stimuli showing: IPs (arrows); example of syllable series with AW in blue characters; boundaries of RGs (lengthened boxes); AW location (blue boxes); and TPs.

RG-internal condition



RG-straddling condition



In short, the design aimed at determining if participants used RGs to chunk heard speech (therefore creating a frame in which TPs are computed). To determine whether RGs interfered with the effects of TPs, we designed a recall task involving four "dummy words" created by changing only the final syllable of the original AWs. These dummy words were not presented in the learning phase but only in the recall phase (see *2.3.2*). Using such dummy words with only a single syllable difference served to evaluate listeners' *complete serial recall* of all 3 syllables of AWs. This precaution was crucial for the RG-straddling condition in that it served to determine whether the final syllable [da] of the AW was memorized as part of the AW when it straddled a RG.

2.2.2. Stimuli recording

Syllable series and AWs were recorded using a pacing technique where an individual produces contexts while listening to series of pure tones providing a metronome-like signal. This technique serves to guide a speaker in producing specific rhythms and intonations. Using headphones, a native speaker of French listened to a continuous playback of the metronome while he repeated each series. The pacers served to obtain productions of RGs with constant durations (4-syll RG = 1,150 ms, 3-syll. RG = 900 ms, 2-syll. RG = 650 ms.) marked by a lengthening of the last syllable (1.6 times longer than non-final syllables [4]).

Recordings were performed in a sound-treated booth using an external sound card (M-Audio *Fast-Track Pro*, 44,1kHz, 16 bits, mono). Every syllable series was saved in an individual sound file and amplitudes were normalized.

The AWs used in the recall phase were recorded separately (not simply spliced from the presentation phase) using the above pacing technique, which mimicked normal prosodic patterns. This was especially important since the goal of the present study is to find out if participants extract and learn the segmental information, not the prosodic contour in which they occurred. Therefore, the words presented in the recall phase are different from the version in the presentation phase.

2.3. Procedure

2.3.1. Presentation (learning phase)

In a first phase, participants heard 18 random blocks of sequences (6-7 target and 31 fillers.) Each sequence was followed by a target syllable and participants were asked to determine if the target was present in the preceding series or not via key-press (this task is irrelevant and served to maintain the subject's attention). Listeners were unaware of the hidden AWs present in some series. Playback amplitude was kept bellow 74 dBA at the headphones.

2.3.2. Recall phase

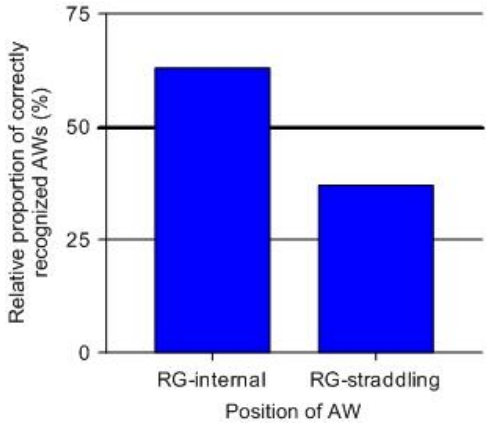
A forced-choice recognition task was performed following the above learning tasks. In the recognition test, participants were presented pairs of AWs (one target word, one dummy) and were asked to say which of the forms was "more familiar" using a key press. Each target AW was presented once with each of his dummy word. In 50% of the presented pairs the AW was presented first and the order of presentation was randomized across participants.

3. RESULTS AND DISCUSSION

It should be recalled that in the above test, subjects heard 560 utterances and of this number 60 contained AWs. Also subjects were not made aware that they had to recall AWs. The results of the forced-choice task showed that the recognition rate of AWs across conditions (47.5%) was not significantly different from chance. However, 63% of correctly recognized AWs were forms presented within RGs, whereas only 37% of AWs that straddled the boundaries of RGs were recognized (see Fig. 3). We compared the frequency with which AWs were correctly identified across

subjects using a Wilcoxon test. The results showed that the recognition rate for AWs within RGs was significantly superior to recognition of forms placed across RGs (n = 20, Z = -2,24, p < 0,05). These results, though preliminary, suggest that RGs can influence statistical learning. The value of a further study should be evaluated by considering that RGs can constitute universal chunks by which series are learned. This is not only reflected in common observations such as the recall of digits like phone numbers. As noted, EEG observations show that listeners are detecting RGs in speech rather that IPs as such. Hence, rhythmic grouping or chunking is not only present in tasks of serial recall but also in listening to series of sounds in utterances.

Figure 3: Relative proportion of correctly recognized AWs (%) from RG-internal and RG-straddling conditions.



4. ACKNOWLEDGMENTS

This research was partly funded by SSHRC grant number 410-2008-1732 and by SSHRC and FQRSC scholarships awarded to the first author.

5. REFERENCES

[1] Boucher, V.J. 2006. On the function of stress rhythms in speech: Evidence of a link with grouping effects on serial memory. *Language and Speech* 49, 495-519.

[2] Broadbent, D.E., Broadbent, M.H.P. 1973. Grouping strategies in short-term memory for alpha-numeric lists. *Bulletin of the British Psychological Society* 26, 135.

[3] Christophe, A., et al. 2003. Discovering words in the continuous speech stream: The role of prosody. *Journal of Phonetics* 31, 585-598.

[4] Delattre, P. 1966. A comparison of syllable length conditioning among languages. *International Review of Applied Linguistics* 4, 183-198.

[5] Gilbert, A.C., Boucher, V.J. 2007. What do listeners attend to in hearing prosodic structures? Investigating the human speech-parser using short-term recall. In van Hamme, H., van Son, R. (eds.), *Proceedings of the Eighth Annual Conference of the International Speech Communication Association (InterSpeech2007)* Antwerp, Belgium, 430-433.

[6] Gilbert, A.C., Boucher, V.J., Jemel, B. 2010. Exploring the rhythmic segmentation of heard speech using evoked potentials. *Proceedings of the 5th Conference on Speech Prosody* 100334, 1-3.

[7] Nooteboom, S. 1997. The prosody of speech: Melody and rhythm. In Hardcastle, W., Laver, J. (eds.), *The Handbook of Phonetic Sciences*. Oxford: Blackwell, 640-673.

[8] Pannekamp, A., et al. 2005. Prosody-driven sentence processing: An event-related brain potential study. *Journal of Cognitive Neuroscience* 17, 407-421.

[9] Shukla, M., Nespor, M,. Mehler, J. 2007. An interaction between prosody and statistics in the segmentation of fluent speech. *Cognitive Psychology* 54, 1-32.

[10] Steinhauer, K. 2003. Electrophysiological correlates of prosody and punctuation. *Brain and Language* 86, 142-164.

[11] Steinhauer, K., Alter, K., Friederici, A.D. 1999. Brain potentials indicate immediate use of prosodic cues in natural speech processing. *Nature Neuroscience* 2, 191-196.

[12] Steinhauer, K., Friederici, A.D. 2001. Prosodic boundaries, comma rules, and brain responses: The closure positive shift in ERP's as a universal marker for prosodic phrasing in listeners and readers. *Journal of Psycholinguistic Research* 30, 267-295.

[13] Vaissière, J. 2006. Perception of intonation. In Pisoni, D. B., Remez, R. E., *The Handbook of Speech Perception*. Oxford: Blackwell, 236-261.

[14] Wechsler, D. 2008. *Wechsler Adult Intelligence Scale* (4th ed). San Antonio, TX: Pearson.