Cognition 171 (2018) 130–150

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Original Articles

Pre-linguistic segmentation of speech into syllable-like units

Okko Räsänen[a,*], Gabriel Doyle[b], Michael C. Frank[b]

[a] Department of Signal Processing and Acoustics, Aalto University, P.O. Box 12000, Aalto, Finland
[b] Department of Psychology, Stanford University, Stanford, CA 94305, United States

ARTICLE INFO

Keywords:
Early language acquisition
Speech perception
Syllables
Speech segmentation
Oscillatory entrainment
Sonority

ABSTRACT

Syllables are often considered to be central to infant and adult speech perception. Many theories and behavioral studies on early language acquisition are also based on syllable-level representations of spoken language. There is little clarity, however, on what sort of pre-linguistic "syllable" would actually be accessible to an infant with no phonological or lexical knowledge. Anchored by the notion that syllables are organized around particularly sonorous (audible) speech sounds, the present study investigates the feasibility of speech segmentation into syllable-like chunks without any a priori linguistic knowledge. We first operationalize sonority as a measurable property of the acoustic input, and then use sonority variation across time, or speech rhythm, as the basis for segmentation. The entire process from acoustic input to chunks of syllable-like acoustic segments is implemented as a computational model inspired by the oscillatory entrainment of the brain to speech rhythm. We analyze the output of the segmentation process in three different languages, showing that the sonority fluctuation in speech is highly informative of syllable and word boundaries in all three cases without any language-specific tuning of the model. These findings support the widely held assumption that syllable-like structure is accessible to infants even when they are only beginning to learn the properties of their native language.

1. Introduction

Theories of early language acquisition often assume that infants perceive speech in terms of syllabic units, even before they can extract the words of their native language. For instance, many artificial language learning experiments have been conducted using stimuli whose statistics are manipulated at the syllabic level and where the success in word learning is measured in terms of the learner's ability to capture statistical regularities connecting adjacent (Saffran, Aslin, & Newport, 1996) or non-adjacent (Newport & Aslin, 2004) syllables. Similarly, studies on artificial grammar learning have often used syllables as the representational level upon which the grammar operates (e.g., Gomez & Gerken, 1999; Marcus, Vijayan, Bandi Rao, & Vishton, 1999). The authors of these early behavioral studies were careful not to specify any specific type of representation underlying the statistical or rule-like computations capturing the syllable-level manipulations, simply referring to "statistical cues" and "speech sounds" (e.g., Saffran, Johnson, Aslin, & Newport, 1999; Saffran et al., 1996; Thiessen & Saffran, 2003), or "rules" and "variables" (Marcus et al., 1999). Nevertheless, later research has adopted the concept of syllable as a representational unit for pre-linguistic speech more explicitly (e.g., Frank, Goldwater, Griffiths, & Tenenbaum, 2010; Gambell & Yang, 2006; Meylan, Kurumada, Börschinger, Johnson, & Frank, 2012; Perruchet & Tillman, 2010; Perruchet & Vinter, 1998; Swingley, 2005; see also Swingley, 2005, for a related discussion). For example, both SARAH (Mehler, Dupoux, & Segui, 1990) and WRAPSA (Jusczyk, 1993; see also Jusczyk & Luce, 2002) models of early speech perception assume that infants are capable of segmenting speech into syllable-like segments before further phonological and lexical analysis. Many Bayesian models of word segmentation also assume that syllable boundaries and identities are known as precursors to word recognition (Doyle & Levy, 2013; Phillips & Pearl, 2012).

The general approach of assuming syllables is consistent with empirical findings suggesting that infant speech perception is better characterized in terms of syllabic frames than phonemic segments (Bertoncini, Bijeljac-Babic, Jusczyk, Kennedy, & Mehler, 1988; Jusczyk, Bertoncini, Bijeljac-Babic, Kennedy, & Mehler, 1990; Jusczyk & Derrah, 1987; Jusczyk, Kennedy, & Jusczyk, 1995) and of holistic rather than analytic representations (Dupoux, 1993; see Hallé and Christia (2012), for an overview).[1] But despite this belief in the importance of syllables in language acquisition, adult-like syllabification depends on knowledge of phonological structure and of the specific language being used, neither of which is available to a child in the early stages of language acquisition. Existing developmental research has not been clear on what

* Corresponding author at: Department of Signal Processing and Acoustics, Aalto University, P.O. Box 12200, FI-00076 Aalto, Finland.
E-mail address: okko.rasanen@aalto.fi (O. Räsänen).

[1] Some research has even gone further and suggested that syllables remain the core unit in adults' representations (Nasukawa, 2007; van der Hulst, 2005).

https://doi.org/10.1016/j.cognition.2017.11.003
Received 12 April 2016; Received in revised form 23 October 2017; Accepted 10 November 2017

form of "syllabic units" would be available to pre-linguistic infants, nor how they could be identified at an age where children only have their generic perceptual capabilities to bootstrap their language learning. Is there language-independent information potentially available to a child in the speech signal itself that would permit the extraction of syllables or syllable-like objects as an early step in language learning?

Since young infants have not yet mastered the sound system of their native language, it is unlikely that the syllable-like chunks they perceive – which are our current focus here – would correspond precisely to the phonological syllables of the language in question. Phonological syllables are defined at the level of formal linguistic representations and typically include a number of language-specific rules and constraints (e.g., Clements, 1990; Goldsmith, 2011; Redford & Randall, 2005; Tesar & Smolensky, 1998), calling for significant experience with the language in question. Further complicating matters, there is some debate about the definition of a syllable among phoneticians dealing with speech perception and production in the wild. This debate is a product of the observation that phonological definitions do not always have clear counterparts in the measurable acoustic or auditory structure of speech available to the listeners (Ladefoged, 2000; Malmberg, 1963; Ohala, 1990; Palmer, 1978), and boundaries of phonological and phonetic syllables may differ in certain situations (e.g., French mute *e*; Fudge, 1969). Hence, the "syllable" is not a unanimous concept in the context of speech perception, even in the case of linguistically proficient adults (Price, 1980). Given this complex situation, the perceptually-available units we focus on here are not traditional phonological syllables, and therefore we will refer to them as *acoustic chunks*.

We explore the idea that these prelinguistic acoustic chunks are derived from sonority, the relative audibility of speech sounds. Sonority connects phonological and phonetic characterizations of syllables and – critically for our purposes – also is available as a possible cue for the creation of acoustic chunks for prelinguistic infants. All definitions of syllables, regardless of level, agree that they are related to the rhythmic fluctuation of speech sonority. A syllable minimally consists of a local sonority maximum (the nucleus), which is typically a vowel, and optionally of less sonorous sounds in the onset and the coda. Sonority also (typically) decreases monotonically from the nucleus towards the edges of the syllable (Hooper, 1976). Researchers may disagree on whether sonority can be defined in terms of physical or perceptual properties of speech (Clements, 2009; Galves, Garcia, Duarte, & Galves, 2002; Parker, 2002), whether it is a purely phonological abstraction based on structural description of languages similarly to phonemes (see Clements (1990), for a discussion), and whether sonority is relevant to phonological theories of language (Harris, 2006).

Irrespective of this debate, sonority has a strong correlational relationship with measurable properties of speech such as intensity and voicing (Section 1.2), and it was originally proposed as the perceptual audibility of different speech sounds (e.g., de Saussure, 1916; Jespersen, 1920; Whitney, 1874). Several existing automatic algorithms for speech syllabification are based on representations of acoustic speech that are closely related to sonority, such as low-pass filtered amplitude or energy envelopes (Mermelstein, 1975; Obin, Lamare, & Roebel, 2013; Wang & Narayanan, 2007). Correlates of sonority are also used in a number of recent neurophysiological models of speech perception, where the neural oscillations of the auditory cortex are believed to phase-lock to the rhythmic properties of the speech envelope, providing timing for more detailed analysis of speech sounds within the resulting syllabic frames (Ghitza, 2011; Giraud & Poeppel, 2012; Hyafil, Fontolan, Kabdebon, Gutkin, & Giraud, 2015). Even without a unanimous and physically precise definition for sonority (if such definition can ever exist), we can still investigate how syllabic structure is represented by rhythmic properties of speech by tracking measurable correlates of sonority as a property of the acoustic signal.

Given this background, in the present paper we examine the nature of the *acoustic chunks* that are accessible through rhythmic sonority variation in speech. Our rationale is that sonority information is likely available to pre-linguistic infants, who are known to be sensitive to rhythmic properties of speech already at a young age (Jusczyk & Thompson, 1978; Mehler & Christophe, 1995; Mehler et al., 1988; Nazzi, Bertoncini, & Mehler, 1998), and hence sonority might be a cue that would allow infants to bootstrap acoustic chunks as inputs to other learning mechanisms. The method for our investigation is a computational model that instantiates this proposal of sonority-based segmentation into acoustic chunks. Before testing this proposal, we will motivate and clarify our goals by reviewing the evidence on the role of syllables in speech perception and on sonority in syllabic structure.

1.1. Role of syllables in speech perception

The idea that syllables[2] are central to pre-linguistic speech perception can be traced back to pioneering work by Mehler, Bertoncini, Jusczyk and their colleagues. In their seminal work, Bertoncini and Mehler (1981) showed that 2-month-old infants are better at discriminating acoustic differences between sequences that resemble syllables than sequences that do not, suggesting the role of syllable as a natural unit of speech perception. This result was later supported by similar findings with 4-day-old (Bijeljac-Babic, Bertoncini, & Mehler, 1993) and 3–4-month-old infants (Eimas, 1999). In addition, a series of studies revealed that 2-month-old infants are capable of extracting and retaining acoustic properties of syllables, and that the results are better understood in terms of syllabic, not phonemic, segmental units (Bertoncini et al., 1988; Jusczyk & Derrah, 1987; Jusczyk et al., 1990; Jusczyk et al., 1995). In parallel, a number of findings revealed perceptual primacy of syllables over phonemes in adult listeners (Mehler, Dommergues, Frauenfelder, & Segui, 1981; Segui, Frauenfelder, & Mehler, 1981; see also Mehler (1981), for a discussion). Studies with children (Liberman, Shankweiler, Fischer, & Carter, 1974) and illiterate adults (Morais, Bertelson, & Alegria, 1986; Morais, Content, Cary, Mehler, & Segui, 1989) also show that conscious access and mental manipulation of syllables is easier than access to phonological segments for participants who have not received formal language instruction. In general, speech perception research shows that phone perception is conditioned by the neighboring sound context, indicating that temporal units of perception must be greater than individual phones (see, e.g., Nusbaum and DeGroot (1991), for a discussion). It is also well-known that syllables often exhibit coarticulatory patterns where cues for different speech sounds overlap in time (e.g., plosive cues are located in the formant transitions of the following vowel in CV-syllables).

Psychoacoustic and neurophysiological data also support the idea of perceptual processing of speech at time-scales greater than individual phone segments. The low-level auditory system integrates signal information across durations approximately corresponding to syllable lengths (~250 ms) (see, e.g., Wagner, 2008, or Räsänen & Laine, 2013, for a review) and is most sensitive to amplitude modulations around 4–5 Hz (e.g., Dau, Kollmeier, & Kohlraus, 1997; Viemeister, 1979)—a typical syllable rate in continuous speech (Greenberg, Carvey, Hitchcock, & Chang, 2003). Oscillatory neural activity in the auditory cortices is known to phase-lock to the syllable-driven amplitude envelope of the incoming speech input (Gross et al., 2013; Luo & Poeppel, 2007). The strength of the coupling correlates with the general intelligibility of the stimuli (Gross et al., 2013), as well as one's subjective comprehension (Ahissar et al., 2001; Peelle, Gross, & Davis, 2013), with abnormalities in amplitude modulation tracking being associated with dyslexia or otherwise impaired phonological development (Leong & Goswami, 2014, 2015). These findings have led to neurophysiological models of speech perception where syllabic rhythm, cued by temporal modulations in the speech envelope and manifested as entrained theta-

[2] Here and below, we continue to use the term "syllable" to describe the units studied in prior research, even though we are – as noted above – agnostic about whether these are true syllables or merely acoustic chunks.