PROSODIC CONSTRAINTS ON STATISTICAL STRATEGIES IN SEGMENTING FLUENT SPEECH

By

MOHINISH SHUKLA

Supervised by

JACQUES MEHLER

Submitted in partial fulfillment of the requirements for

Doctor of Philosophy in Cognitive Neuroscience

Scuola Internazionale Superiore di Studi Anvanzati -

International School for Advanced Studies

Trieste, Italy

Jury

Richard Aslin, *University of Rochester.*

Anne Christophe, *LSCP, Paris.*

Marina Nespor, *University of Ferrara.*

Mathew Diamond, *SISSA, Trieste.*

Tim Shallice, *SISSA, Trieste, and ICN, London*

Main publication from the thesis

Shukla, Mohinish, Nespor, Marina, and Mehler, Jacques (2006) An interaction between prosody and statistics in the segmentation of fluent speech. *Cognitive Psychology*, in press.

doi:10.1016/j.cogpsych.2006.04.002.

Contents

List of Tables

List of Figures

Abbreviations and symbols

ν ..... Utterance

IP ..... Intonational Phrase

ϕ ..... Phonological Phrase

C ..... Clitic group

ω ..... Phonological Word

Σ ..... Foot

σ ..... Syllable

μ ..... Mora

C ..... Consonant

V ..... Vowel

IPA ..... International Phonetic Alphabet

PWC ..... Possible Word Constraint

TP ..... Transition Probability

N_σ ..... Number of unique syllables in an artificial stream

ADH ..... Acoustic Distance Hypothesis

NIRS ..... Near-Infrared Spectroscopy

OT ..... Optical Topography

LH ..... Left Hemisphere

RH ..... Right Hemisphere

Oxy-Hb ..... Oxygenated Hemoglobin

Deoxy-Hb ..... De-oxygenated Hemoglobin

Summary

Learning a spoken language is, in part, an input-driven process. However, the relevant units of speech like words or morphemes are not clearly marked in the the speech input. This thesis explores some possible strategies to segment fluent speech.

Two main strategies for segmenting fluent speech are considered. The first involves computing the distributional properties of the input stream. Previous research has established that adults and infants can use the transition probabilities (TPs) between syllables to segment speech. Specifically, researchers have found a preference for syllabic sequences which have relatively high average transition probabilities between the constituent syllables.

The second strategy relies on the prosodic organization of speech. In particular, larger phrasal constituents of speech are invariably aligned with the boundaries of words. Thus, any sensitivity to the edges of such phrases will serve to place additional constraints on possible words.

The main goal of this thesis is to understand how different strategies conspire together to provide a rich set of cues to segment speech. In particular, we explore how prosodic boundaries influence distributional strategies in segmenting fluent speech.

The primary methodology employed is behavioral studies with Italian-speaking adults. In the initial experimental chapters, a novel paradigm is described for studying distributional strategies in segmenting artificial, fluent speech streams. This paradigm uses artificial speech containing *syllabic noise*, defined as the presence of syllables that do not comprise the target nonce words, but occur at random at comparable frequencies. It is shown that the presence of syllabic noise does not affect segmentation. This suggests that statistical computations are robust.

We find that, although the presence of the noise syllables do not affect TP computations, the placement of nonce words with respect to each other does. In particular, 'words' with a clumped distribution are better segmented than 'words' with an even spacing. This suggests that even the process of statistical segmentation itself is constrained.

The syllabic noise paradigm is utilized to create speech streams as sequences of *frames*: syllabic sequences of fixed length. 'Words' can be placed at arbitrary positions with respect to these frames; the remaining positions are occupied by noise syllables. By adding pitch and length characteristics of Intonational Phrases (IPs, which are large phrasal constituents) from the native language, the frames can be turned into prosodic 'phrases'. Thus, nonce words can be placed at different positions with respect to such 'phrases'. It is found that 'words' that straddle such 'phrases' are not preferred over non-words, while 'phrase'-internal 'words' are. Removing the prosodic aspects from the frames abrogates this effect.

These initial experiments suggest that prosody carves speech streams into smaller constituents. Presumably, participants infer the edges of these 'phrases' as being edges of words, as in natural speech. It is well known that edge positions are salient. This suggests that 'words' at the edges of the 'phrases' should be better recognized than 'words' in the middles. The subsequent experiments show such an *edge effect* of prosody.

The previous results are ambiguous as to the whether prosody *blocks* the computation of TPs across phrasal boundaries, or acts at a later stage to *suppress* the outcome of TP computations. It is seen that prosody does not block TP computations: under certain conditions one can find evidence that participants compute TPs for both 'phrase'-medial and 'phrase'-straddling 'words'. These results suggest that prosody acts as a *filter* against statistically cohesive 'words' that straddle prosodic boundaries. Based on these results, the *prosodic filtering* model is proposed.

Next, we examine the generality of the prosodic filtering effect. It will be shown that a foreign prosody causes a similar perception of 'phrasal' edges; the edge effect and the filtering effect are both observed even with foreign IPs. Phonologists have proposed that IPs are universally marked by similar acoustic cues. Thus, the results with foreign prosody suggest that these universal cues play a role in the perception of phrases in fluent speech. Such cues include final lengthening and final pitch decline; further experiments show that, at least in the experimental paradigm used in this thesis, pitch decline plays the primary role in the perception of 'phrases'.

Finally, we consider the possible bases for the perception of prosodic edges in

otherwise fluent speech. It is suggested that this capacity is not purely linguistic, but arises from acoustic perception: we will see that time-reversed IPs, which maintains pitch breaks at 'phrasal' boundaries, can still induce the filtering effect.

In an annex, the question of how time-reversed (backward) speech is perceived in neonates is addressed. In a brain imaging (OT) study with neonates, we find evidence that forward speech is processed differently from backward speech, replicating previous results.

In conclusion, the task of finding word boundaries in fluent speech is highly constrained. These constraints can be understood as the natural limitations that ensue when multiple cognitive systems interact in solving particular tasks.

Acknowledgments[1]

First, special thanks to Jacques Mehler for scientific, ideological, monetary, equipmental and victual support and lots of patience. Special thanks also to Marina Nespor for (complementary) scientific and ideological support. Thanks to all the close collaborators on the projects that eventually wound up in the thesis: Luca Bonatti, Ansgar Endress, Judit Gervain and Ágnes Kovács. Superspecial thanks to Marcela Peña, specially for getting me started on imaging, both EEG and OT.

The writing of the thesis benefited greatly by comments, mainly from Marina, Jacques, Judit (positive comments) and Ansgar (negative comments). Despite their best efforts, I'm sure there are errors of omission and commission, which of course belong to me. In addition, lots of people directly or indirectly helped in reaching these final pages – I'm very grateful to la Triestina Debora and all friends and family members scattered around the world for emotional support during the very hard task of writing a thesis.

I'm grateful to various people at SISSA for academic, administrative and technical support. In particular, to John Nicholls and to Mathew Diamond and his lab for getting me into SISSA and into cognitive neuroscience, and to Tim Shallice and Henry and Lila Gleitman for interesting conversations and to Sven Mattys for sharing ideas. Thanks also to various academics at the University of Trieste, in particular Nicola Bruno and Pino Longobardi. A big thanks to Luigi Rizzi at the University of Siena for providing a very clear understanding of linguistics and language acquisition.

Finally a big thanks to Milind Watve for pushing me towards a career in science, and to Vidyanand Nanjundiah for starting a chain of events that led to this dissertation.

[1] The extended version can be found at http://mohinish.s.googlepages.com/acknowledgements

Part I

Introduction

'Where shall I begin, please your Majesty?' [the White Rabbit] asked.
'Begin at the beginning,' the King said gravely, 'and go on till you come to the end: then stop.'

Alice in Wonderland,
Lewis Carroll

Chapter 1

Segregating the blooming from the buzzing

How does an infant acquire its native language? In the view advocated by Noam Chomsky, language acquisition is the transformation of an innately specified "initial state" of the language faculty into a final, mature state. This process is intimately guided by a *Language Acquisition Device* (LAD), that draws extensively on experience to bring about this transformation (e.g., Chomsky, 1995, 2000). The 'experience' is the spoken corpus that the the infant is exposed to.

Considering spoken language as a mapping between sound and meaning, the task of the infant is to build a representation of the speech stream (phonology), assign it syntactic structure (syntax), and to arrive at the meanings of the utterances (semantics). While phonology, syntax and semantics are all intensely researched and controversial topics, they all hold in common that speech is built out of finite elements, the words (morphemes) of the language (Pinker, 1994; Chomsky, 1995; Baker, 2001; Bresnan, 2001; Prince & Smolensky, 2004; Chomsky, 2005).

However, it has been long appreciated that the words themselves are not overtly marked in fluent speech, as is clearly noticeable in listening to speech in an unknown language. Words are not consistently preceded or followed by pauses or other distinct acoustic signals (e.g., Harris, 1955; Cole, Jakimik, & Cooper, 1980; Saffran, Aslin, & Newport, 1996; Brent & Cartwright, 1996). This thesis examines this one specific aspect of language acquisition: *speech segmentation*, wherein fluent speech is transformed into a series of words.

1.1 What does the infant perceive?

Fluent speech is not readily available to the infant. Sound consists solely of the variations of air pressure in time. Thus, an entire rock concert, complete with vocals, lead guitar, bass, drums, keyboards and an appreciative audience can be captured in the oscillations of a single groove of a gramophone record. Similarly, speech is embedded in pressure variations at the eardrum that also contain contributions from a variety of incidental, irrelevant environmental noises. What does the infant make of its acoustic input?

More than a century ago, William James (1890) assumed that the perceptual world of the baby was "one great blooming, buzzing confusion". In this view, the mind of the neonate is a blank slate upon which the senses draw a chaotic, meaningless, holistic pattern. Only with experience does the baby learn to analyze and segregate the world into different (perceptual) objects.

One of the fundamental findings from the past century was that neonates bring with them a rich mental toolkit to analyze the input from the various senses (e.g., Mehler & Dupoux, 1994; Baillargeon, 1995; Gopnik, Meltzoff, & Kuhl, 1999). Very young infants have been shown to have remarkable cognitive capacities, including some basic numerical ability (Starkey, 1992; Wynn, 1996), physical concepts (Baillargeon, 1995) and an appreciation of biological motion (Bertenthal, 1993).

Similarly, several studies have documented early capacities of infants for language. For example, Colombo and Bundy (1983) showed that infants respond preferentially to speech streams as compared to other noises. To better understand this preference for speech, Mehler et al. (1988) contrasted spoken utterances with the same utterances played backwards. Although such *backward* utterances are matched with the *forward* utterances on a variety of acoustical parameters like volume, duration and frequency content, infants nevertheless preferred the forward utterances. More recent studies using imaging methods have shown that the brains of neonates (Peña et al., 2003) and 3-month-olds (Dehaene-Lambertz, Dehaene, & Hertz-Pannier, 2002) react differentially to forward and backward speech. In an annex to this thesis, I present additional imaging evidence supporting these fundamental observations (page 141).

Other behavioral studies have demonstrated that neonates (Moon, Cooper, &

Fifer, 1993) and two-month-olds (Christophe & Morton, 1998) prefer their native language to a foreign language, and can discriminate languages that belong to different rhythmic classes[1] (Nazzi, Bertoncini, & Mehler, 1998; Ramus, Hauser, Miller, Morris, & Mehler, 2000). Further, Bertoncini and Mehler (1981a) and Bertoncini, Bijeljac-Babic, Blumstein, and Mehler (1987) showed that very young infants already show some sensitivity to a fundamental building block of speech, the syllable.

These studies suggest that the acoustic percept of the neonate is far from a chaotic mélange of sounds. Instead, the neonate is capable of appropriating speech from ambient acoustic stimuli and organizing it in a manner conducive to acquiring language.

1.2 Segregating the input

The experimental psychologist Alvin Liberman and his colleagues proposed that speech is a code to which the human mind hold the key (e.g., Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967). Similarly, the Hungarian/British polymath, Arthur Koestler suggested that speech is just variations in air-pressure, unless there is a human nervous system to decode it. How is the speech code deciphered?

In his book "The Ghost in the Machine" (1967), Koestler offers a speculative account, wherein perception is seen as interlocking hierarchies of *filters* or *scanners*. The centripetal progress of a stimulus, from the sensorium to the cortex, is proposed to be through separate series of filters, wherein the primary data is de-coded, analyzed and summarized in progressively more abstract forms within each.

For example, from the acoustic flux of a rock concert, a particular hierarchy might selectively recover the lyrics of the song. At each successive level of the hierarchy, irrelevant information like the noise of the crowd, the accompanying

[1] The rhythmic classes, originally proposed by linguists (Pike, 1945; Abercrombie, 1967; Ladefoged, 1975; Dauer, 1987), sort the languages of the world on the basis of the basic perceived rhythmic unit into the stress-timed languages and syllable timed languages. See Nespor (1990); Ramus, Nespor, and Mehler (1999); Grabe and Low (2002); Galves, Garcia, Duarte, and Galves (2002) for recent re-appraisals.

instruments, and the incidental, auditory characteristics of the words are progressively filtered out to leave just the lyrics as the end-point of the hierarchy of filters. A second hierarchy might recover the chord succession of the lead guitar. Thus, different aspects of the input from a single modality are extracted and stored by separate hierarchies with different criteria of relevance.

Similarly, we can hypothesize that different aspects of speech are extracted by separate hierarchies. For example, imagine that a mature language user overhears speech in an unfamiliar language. The listener would not be able to tell us what was said, but might be able to report that the speaker was male or female, and perhaps even if (s)he was angry or happy. Thus, the words, the emotions, and the gender of the speaker are transmitted simultaneously.

1.3 Dividing, conquering, and reuniting

Let us make the assumption that the neonate, like the mature language user, does not perceive speech as an undivided, monolithic whole, but as a compendium of different sources of information. Such an assumption is not without foundation. For example, as noted before (Section 1.1), very young infants are sensitive both to the syllable, a basic unit of speech; and to linguistic rhythm, a global property of spoken language. Thus, infants can not just segregate speech from the acoustic input, but can also process the different aspects of speech in parallel.

To summarize, the neonate has innate mechanisms to segregate speech from the sumness of its acoustic input. It then segregates and separately analyzes the different aspects of the speech input. Eventually, the outputs of the various sources of information must all be put back together.

This thesis aims at building a specific model along these general lines. The specific task under consideration is the segmenting out of words from fluent speech. We will examine two sources of information: the statistical properties over the syllables, and the melodic (prosodic) organization of phrases.

Both the prosodic organization of language and the distributional properties over the syllables have been extensively studied. Their role in segmenting fluent speech in infants and adults has been explored by numerous researchers. The main concern in this thesis is how these sources of information interact.

The empirical method employed in this thesis is the behavioral responses of adults exposed to novel artificial 'languages'. Such experiments thus simulate the condition of the neonate confronted with its ambient language. In addition, the artificial languages allow the precise manipulation of different cues to word boundaries.

In the following chapter, we will see what is known about the prosodic organization of speech, and how it can help in segmenting fluent speech. Subsequently, we will examine studies that explore the contribution of statistical information.

'Now, first I put my head on the top of the gate–then I stand on my head–then the feet are high enough, you see–then I'm over, you see.'

Through the looking glass,
Lewis Carroll

Chapter 2

The prosodic organization of language

Speech is not merely a chaining together of sounds like beads on a string. Arthur Koestler, in "The Ghost in the Machine" (1967), suggests that:

> Melody, timbre, counterpoint, are patterns in time—as phonemes, words and phrases are patterns in time. None of them make sense—musical, linguistic, semantic sense—if considered as a linear chain of elementary units. The message of the air-pressure pulses can only be de-coded by identifying the wheels within wheels, the simpler patterns integrated into more complex patterns like arabesques in an oriental carpet.

In this chapter, we will examine the hierarchical nature of speech. In doing so, it will become clear that the organization of speech implies that fluent speech does contain cues that can aid in finding word boundaries.

2.1 Wheels within wheels

Language is a mapping between meaning and sound. Thus, Chomsky and colleagues consider language to be, minimally, a computational system that generates internal representations that are mapped onto the sensory-motor interface on the one hand, and onto the conceptual-intentional system on the other (e.g., Chomsky,

1995; Hauser, Chomsky, & Fitch, 2002). In this view, many of the properties of spoken language derive not from the syntactic component of language, but from the interface conditions between the core generative computations and the output system (be it speech or sign language). This implies that the rules that govern syntactic computations may be divorced from those that govern speech.

Indeed, such a conclusion was reached by *phonologists* in the late 1970s, studying the organization of spoken language. The rules of syntax were found to be insufficient to account for the organization of spoken utterances (e.g., Liberman & Prince, 1977; Goldsmith, 1976).

As an example, consider Figure 2.1. In this figure a single sentence has been broken down into its prosodic constituents.[1]

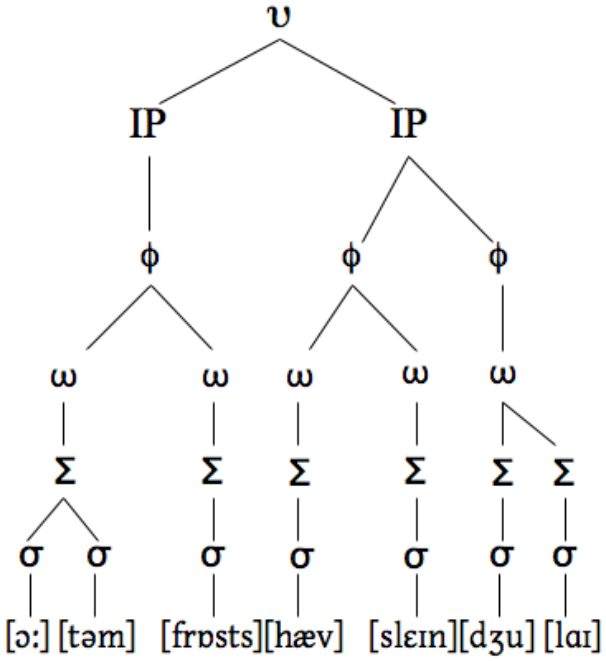


Figure 2.1: The structure of the utterance (υ) formed from the line "*Autumn frosts have slain July.*". IP: Intonational Phrase; ϕ: Phonological Phrase; ω: Phonological Word; Σ: Foot; σ: Syllable. The final line shows the syllabification of the utterance in IPA. The complete prosodic hierarchy is shown in Figure 2.2

There are two noteworthy things to consider in Figure 2.1. The first is the

[1] In this thesis, I will use the terminology and assumptions in Nespor and Vogel, 1986.

hierarchical nature of the constituents (the constituents themselves are explained below). The second is the fact that nowhere is there a mention of syntactic notions like 'verb', 'noun' or 'Verb Phrase'. That is, the various constituents of the utterance depicted in the figure are not syntactic domains, but are *prosodic domains*.

2.2 The prosodic hierarchy

According to prosodic theory, the mental representation of speech is divided into hierarchically arranged chunks, the prosodic domains. These domains have the following two important properties:

1. The chunks are domains for the application of phonological (and phonetic) rules[2].

2. Different chunks draw on different aspects of phonology and morpho-syntax but, crucially, are not necessarily isomorphic with chunks generated by any other grammatical process.

The first property can be viewed as the process of discovery and delineation of prosodic constituents: these are the chunks within which a particular phonological rule applies. The second is the main motivation for such chunks as independent prosodic components: they are not necessarily co-extensive with constituents of other components of grammar, like syntax or morphology.

However, morphology deals with the structure of words. Therefore, if we are to understand how the organization of speech contributes to placing word boundaries, we must understand the relation between prosody and morphology.

What is a 'word'? In written English, we recognize a word as text surrounded by white spaces or punctuation (a scheme used in this thesis), for example 'dog'. The word 'dog' represents the link between a certain sound pattern and a certain meaning in the mind of the English listener.

[2]Although not all phonological rules make reference to prosodic constituents. For example, in English, the choice of the indefinite articles 'a' or 'an' depend upon the initial vowel of the following noun.

However, linguists recognize the *morpheme* as the smallest unit of meaning. For example, in the English sentence in Figure 2.1, the word 'autumn' is a single morpheme, while the word 'frosts' is made up of two morphemes, the *stem*, 'frost', and the *suffix* 's' that marks plurality. In *agglutinating* languages, several suffixes can be added to a stem. So, from 'çocúk' (child), one can derive the words 'çocuklár' (children), 'çocuklarimíz' (our children) and 'çocuklarimizín' (of our children). Thus, in morphology, a word consists of a stem plus its affixes.

In prosodic theory, the constituent that most closely corresponds to a morphological word is the *Phonological word (ω)*. As described by Nespor and Vogel (1986, pg. 109), ω is the lowest constituent of the hierarchy that reflects an intimate relation between phonology and morphology.

The precise contribution of morphological information to the prosodic constituent ω varies from one language to another. Nevertheless, it is clear that cross-linguistically, ωs correspond at most to a lexical stem plus its affixes. This implies that both the right and the left edges of ωs are also the edges of (one or more) morphemes. Indeed, in edge-based theories of the syntax-phonology interface, the edges of lexical roots are aligned with the edges of phonological constituents, in particular ω (e.g., Cohn, 1989; Selkirk, 1986; McCarthy & Prince, 1993; Selkirk, 1996).

Figure 2.2 presents a hierarchy of prosodic constituents (adapted from Nespor and Vogel, 1986 and Selkirk, 1996; see also Figure 2.1).

Since we are interested in how words are segmented from fluent speech, let us take ω as our starting point. We can then examine the prosodic hierarchy in two ways:

- What are the constituents that make up the ωs?
- How are ωs put together into larger constituents?

2.2.1 Prosodic constituents smaller than ω

Notice from Figure 2.2 that the smallest prosodic unit we consider is the syllable. However, the syllable itself is made up of consonants and the vowels, together

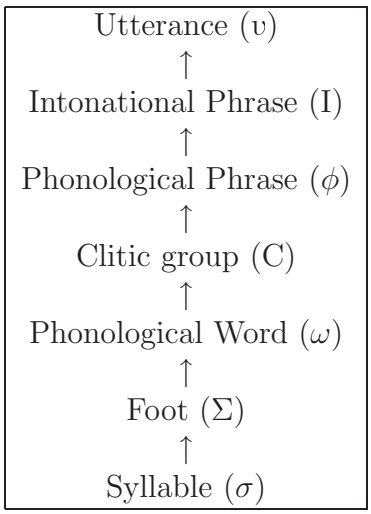


Figure 2.2: A hierarchy of prosodic constituents

called the *phones*. The phone is thus the minimal unit of speech[3]. Words in different languages are made up of different sets of phones. For example, in English, changing the phone [l] in the word 'lip' to the phone [r] changes the word. In contrast, in Japanese the [l]→[r] change has no effect on the status of a word (but see Cutler, Weber, & Otake, 2006). The set of phones which, when changed to another, change the word to another word (or to a non-word, for example from [lip]→[mip]) are called the *phonemes*.

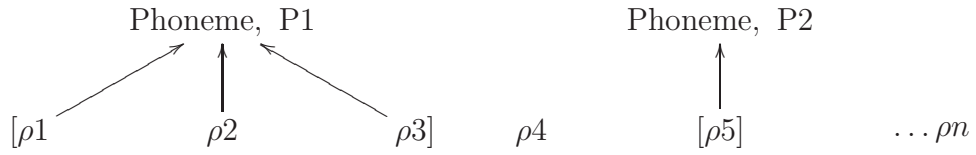
The mapping from phones to phonemes is not 1-to-1; each phoneme can be realized as one or more phones. For example, the phoneme /t/ in English[4] emerges as the aspirated phone [t^h] when it is foot-initial, and as the unaspirated [t] otherwise. All the phones of a language that map onto a single phoneme in that language are called the *allophones*. A variety of morphological and phonological rules determine the choice of one allophone or another in specific contexts. Thus, while changing a phoneme changes the meaning of a word, changing an allophone makes the word sound 'foreign', or from a different dialect, but does not change the meaning.

Schematically (below), if $\rho 1 \ldots \rho n$ represent all the possible phones, then for a

[3]Phones themselves are distinguished from one another by acoustic *distinctive features*. Also, phones are organized into another sub-syllabic constituent, the *mora* (μ). For ease of exposition, we will exclude the discussion of distinctive features and moras in this thesis

[4]Conventionally, phones are marked in square brackets and phonemes in slashes.

particular language, one phoneme like P1 might correspond to several phones (ρ1, ρ2 and ρ3), which together constitute the allophones of the phoneme P1. Another phoneme like P2 might correspond to only a single phone (ρ5).

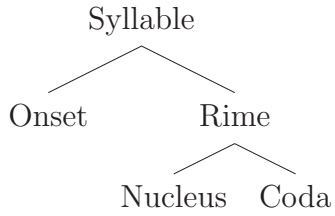


Thus, we will first look at how consonants and vowels (segments) are organized into syllables. Next, we will consider the prosodic constituent larger than the syllable, the foot. We will see that the construction of feet reflects some general principles of how smaller prosodic constituents are organized into larger ones.

The Syllable (σ)

In phonology, the syllable can be seen to be the domain of certain phonological processes (or constraints). For example, in English, a *t* preceding an *r* is alveopalatized (the point of articulation of the *t* moves from alveolar to alveopalatal), only when the *t* is syllable-initial. Thus, alveopalatization can occur in the *t* in 're.***t***rieve', but not in 'nigh***t***.rate' (see Nespor & Vogel, 1986; Blevins, 1995, for a thorough discussion).

The internal structure of the syllable The syllable is an organization of consonant and vowel segments as follows (e.g., Selkirk, 1982; Blevins, 1995):



The nucleus is the most *sonorous* of the segments that make up the syllable. Sonority refers to the relative loudness of one sound compared to another (Giegerich, 1992). Thus, vowels are more sonorous than consonants, and within the consonants, the nasals are more sonorant than the stops. Consequently, in the

majority of the languages of the world, the nucleus is vocalic. Typically, sonority decreases away from the nucleus (e.g., Blevins, 1995). This implies that for each syllable, there is at most one sonority peak. Indeed, the number of peaks in sonority in an utterance roughly correspond to the number of syllables in that utterance.

Cross-linguistically, the following generalizations can be drawn regarding the structure of the syllable:

1. The onset can be optional, but never completely disallowed.

2. The nucleus is obligatory and is the most sonorant segment.

3. The coda can be disallowed or optional.

4. Languages can allow for complex syllables by permitting that each of the constituents of the syllable *branch*, that is, have more than one segment.

Putting (1), (2) and (3) together, we see that a single vowel is the smallest possible syllable. Nevertheless, as seen from (1), there are no languages in which the onset is absent. Consequently, the CV syllable is found in all languages. It has been suggested that the CV syllable represents the most unmarked structure of speech sounds, probably being a precursor of modern speech (see MacNeilage, 1998; MacNeilage & Davis, 2001).

The syllable in psycholinguistics Psycholinguists have proposed the syllable as the fundamental building block of speech, both in production (e.g., Levelt, 1989) and in perception (e.g., Mehler, 1981). Importantly, from the point of view of this thesis, the syllable has been shown to be processed by very young infants (see, for example, Bertoncini & Mehler, 1981a; Bertoncini, Floccia, Nazzi, & Mehler, 1995; Ooijen, Bertoncini, Sansavini, & Mehler, 1997).

For example, using the high-amplitude sucking procedure, Bertoncini and Mehler (1981a) showed that infants could discriminate two syllables that differed

only in the serial order of their constituents, e.g. *PAT* and *TAP*. Moreover, infants failed to discriminate similar sequences *PST* and *TSP*, which, as we saw above, are not well-formed syllables. However, when such sequences were converted into legal syllables by inserting them in the context of vowels, e.g., *UPSTU* and *UTSPU*, discrimination ability was restored.

Further, (Bertoncini, Bijeljac-Babic, Jusczyk, Kennedy, & Mehler, 1988) found that neonates could discriminate a change in syllables when the vowel, but not the consonant was changed. By two months of age, infants could discriminate to a change in both a vowel or a consonant. Finally, Bijeljac-Babic, Bertoncini, and Mehler (1993) showed that infants could discriminate lists of CVCV bisyllables (like *maki*, *nepo*) from lists of CVCVCV trisyllables (like *makine*, *posuta*), regardless of whether the items differed or were matched in duration. These results were extended by Ooijen et al. (1997), who showed that neonates could discriminate bisyllables lists from monosyllable lists, even when one of the syllables of the bisyllables was phonologically weak.

A second source of evidence for the role of the syllable comes from experiments investigating the *Possible Word Constraint* (PWC), according to which, parses that leave behind isolated consonants are disfavored, since these can never be possible words (Norris, McQueen, Cutler, & Butterfield, 1997; McQueen, Otake, & Cutler, 2001; Cutler, Demuth, & McQueen, 2002; Yip, 2004). For example, in a word-spotting experiment, Norris et al. (1997) showed that English listeners found it much easier to spot the word 'apple' in 'vuffapple' than in 'fapple'; the latter leaves a single consonant stranded.

We saw earlier that phonological words are above the syllable in the prosodic hierarchy. We will see later (v, pg. 19) that the nature of prosodic hierarchies implies that higher constituents must contain at least one unit from all the lower constituents. Therefore, we can generalize the PWC as: *parses that do not leave behind at least one syllable are disfavoured.*

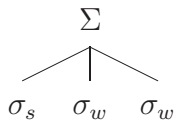
In a telling experiment, Cutler et al. (2002) examined word-spotting in the Bantu language Sesotho, in which not just single consonants, but also single *syllables* cannot be words. Sesotho speakers found it more difficult to spot words when a consonant was left stranded than when either a bisyllable or a monosyllable

were left stranded. That is, even though monosyllables are impossible words in Sesotho, they pattern like the bisyllables, rather than like the consonants. These results confirm the view that the PWC is universal; stranded consonants are unacceptable, but stranded syllables are acceptable, even if there are no monosyllabic words in the language.

More recently, Johnson, Jusczyk, Cutler, and Norris (2003) used the head-turn preference procedure with 12-month-old infants and found that when familiarized with, for example, 'win', they listened longer to sentences containing 'win' in a possible context (e.g., 'window') than in an impossible context (e.g., 'wind'). These results suggest that infants, like adults, observe the PWC in parsing fluent speech (see also, Johnson, Jusczyk, Cutler, & Norris, 2000).

The Foot (Σ)

Syllables are grouped together into ***feet***, Σ. Each Σ can be considered as a sequence of one relatively strong, and any number of relatively weak syllables (Nespor & Vogel, 1986). Thus, the Σ node dominates a flat structure:



The precise location of the strong syllable depends on language-specific factors. The foot determines the placement of secondary stress, which typically falls on the relatively strong syllable.

The foot as a phonological domain can be seen by considering aspiration in English: the voiceless stops, *p*, *t* and *k* are aspirated foot-initially, and unaspirated elsewhere. Thus, the *t* in 'satire' ([sa]$_\Sigma$[tire]$_\Sigma$) is aspirated, while the *t* in 'satyr' ([satyr]$_\Sigma$) is not (Nespor & Vogel, 1986).

The structure of the foot presented above reflects some general principles of prosodic constituents:

i. *Construction of prosodic constituents:* Join into an n-ary branching X^p all X^{p-1} included in a string delimited by the definition of the domain of X^p.

That is, a unit of a non-terminal level of the prosodic hierarchy is made up of the linear arrangement of units of the immediately lower level that fall within its domain. For example, each Σ is made up of one or more σs that are within its domain. Further, there are no extra (abstract) levels between a Σ and the σs (as shown above).

Also, the fact that there is only a single relatively strong syllable in each foot is itself due to a general principle:

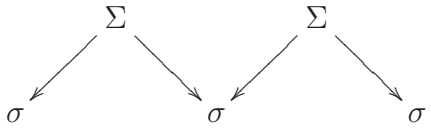
ii. The relative prominence relation defined for sister nodes is such that one node is assigned the value strong (s) and all the other nodes are assigned the value weak (w).

The geometry of the overall organization of prosodic units is accounted for by the following two principles:

iii. A unit of a given level of the hierarchy is exhaustively contained in the superordinate constituent of which it is part.

iv. Each non-terminal unit of the prosodic hierarchy is composed of one or more units of the immediately lower category.

Principle (iii) rules out structures like the following, wherein one σ belongs to two different Σs:



A consequence of this principle is that if, for example, an utterance is parsed into ωs, each containing all the Σs in their domain, there are no σs left over.

Principle (iv) can be seen from the fact that each Σ is composed not of μs or phonemes, but of (at least one) σ. This has also been termed the Strict Layer Hypothesis (Selkirk, 1984; see also Beckman and Pierrehumbert, 1986). A consequence of this principle is that the prosodic constituents cannot display recursion. Thus, structures like $[_{\omega} \ldots [_{\Sigma} \ldots [_{\omega} \ldots] \ ] \ ]$ are disallowed (in this particular case because the Σ contains a ω, which is a higher category).

The next larger unit in the prosodic hierarchy after the foot is the phonological word, ω. From principle (iii) above, it is clear that each ω groups together into an n-ary branching structure all the feet that are in its domain. As noted earlier, the domain of each ω is intimately linked to morphology, so that the edges of ωs are also the edges of one or more morphemes.

The principles (iii) and (iv) discussed above, imply the following principle (Neelman & Koot, 2006):

> v. *Proper containment*: A boundary at a particular level of the prosodic hierarchy implies all weaker boundaries.

A consequence of (v) is that ***a 'higher' constituent of the prosodic hierarchy must be co-extensive with at least one unit from all the lower constituents.*** Since ωs (roughly) correspond to words, this implies that larger constituents must contain at least one ω each. Thus, it follows that ***the edges of larger prosodic constituents are also the edges of words.*** This gives the primary motivation for considering the role of larger prosodic constituents in segmenting fluent speech.

In the remaining part of this chapter, we will look at how ωs are put together into the larger prosodic constituents.

2.2.2 'Larger' prosodic constituents

Spoken language is not a series of isolated words. The clearest constituent of spoken language is the *utterance*, a stretch of speech bounded by silent pauses. In prosodic theory, (phonological) words are not merely chained together into utterances. Instead, they are organized into clitic groups (C). Cs are further organized into phonological phrases (ϕ) that are in turn organized into intonational phrases (IPs). Utterances consist of one or more of such IPs.

The Clitic Group (C)

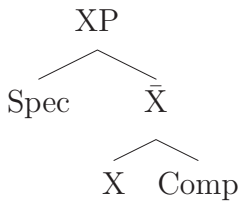
Clitic groups consist of at least one ω that contains an independent word, plus any adjacent ωs containing a clitic (Nespor & Vogel, 1986). *Clitics* are words that

syntactically function as free morphemes, but phonetically appear as bound morphemes. For example, in English, *enclitics* (coming after an independent word) include the abbreviated forms of 'be' (as in *I**'m***, *you**'re***, *she**'s***) or of auxiliaries (as in *they**'ll*** or *they**'ve***); while *proclitics* include the articles (as in ***a*** *boy*).

Let us now look in greater detail at two larger prosodic constituents. The first of these, the phonological phrase, makes reference to syntactic constituency. However, the second, the intonational phrase, is affected not only by syntactic information, but also by semantic and pragmatic information as well as performance factors like speech rate and style, and is possibly universal. These two constituents are given special consideration because, as we shall see, they play important roles in different aspects of language acquisition.

The Phonological Phrases (ϕ)

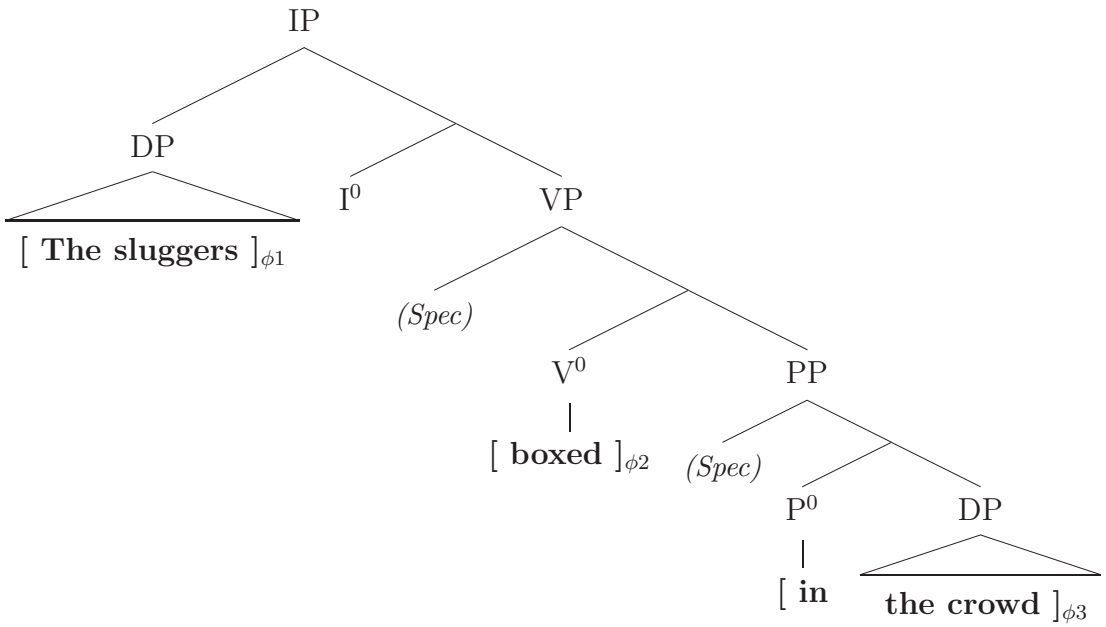
Phonological phrases appear to be tightly constrained by syntax. Nespor and Vogel (1986) propose that the domain of ϕ "...consists of a C[litic group] which contains a lexical head (X) and all C's on its nonrecursive side up to the C that contains another head outside of the maximal projection of X." In X-bar theory (Chomsky, 1986), all lexical and functional categories (X) project essentially the same structure. For right-recursive languages such as English, this can be represented as:



The *Spec*ifier and *Comp*lement can be considered as the external and internal arguments respectively of the lexical or functional head, X; while $\bar{\text{X}}$ is the (abstract) node that dominates the lexical head X and its complement. As can be seen from the syntactic tree above, English is a right-branching language. That is, the internal arguments of the phrasal head (X) occur to the right of X.

Thus, for English, a phonological phrase would consist of a C containing a lexical head, plus all the elements to its left, upto the next C containing a lexical

head. For example, the sentence "The sluggers boxed in the crowd" (from Nespor & Vogel, 1986), consists of three ϕs, going leftwards from the verb and the two nouns, as shown[5]:



Since the ϕ is based on syntax, it shows variation according to the syntax of the language. We saw that in English, lexical heads occur ϕ-finally. In contrast, in a left-branching language like Japanese, lexical heads occur at the left edges of ϕs (Nespor & Vogel, 1986).

Recall that one of the principles of the prosodic hierarchy states that only one of the sister nodes inside a prosodic unit can be strong (Principle ii, page 18). Nespor and Vogel (1986) show that prominence at the phonological phrase level depends on the syntax of the language: right-branching (*Head-initial*) languages have ϕ-final prominence, while left-branching (*Head-final*) languages have ϕ-initial prominence.

That is, prominence at a prosodic level (the ϕ) reflects and thus signals a major difference in syntactic variation in the languages of the world. This has clear implications for language acquisition: if infants are sensitive to ϕs in speech, they could use this (prosodic) information to discover a syntactic feature of their

[5]The syntactic tree is drawn with only the relevant branches; the *(Spec)* are empty. The IP in this tree refers to the *Inflection Phrase*, see, e.g., Pollock, 1989. Technically, the verb *boxed* is at the position marked I^0.

native language (see also Nespor, Guasti, & Christophe, 1996).

Indeed, Christophe, Guasti, Nespor, and Ooyen (2003) showed that 6- to 12-week-old infants could discriminate carefully matched sentences from Turkish (Head-final) and French (Head-initial). Thus, ϕs might play a role in the acquisition of syntax. In the next chapter we will examine evidence that ϕs can also be used in segmenting fluent speech.

In the next chapter, we will look at evidence that ϕs are used in speech segmentation.

The Intonational Phrase (IP)

The Intonational Phrase (IP) is the prosodic constituent that groups together ϕs. IPs are intonationally defined and are the domain of a perceptually coherent intonational contour (Pierrehumbert, 1980; Nespor & Vogel, 1986; Shattuck-Hufnagel & Turk, 1996). IPs are delimited by pauses, phrase-final lengthening and pitch movement.

All languages have intonation. Intonation has been described as a language universal (e.g., Hockett, 1963), both because pitch variations convey some linguistic or paralinguistic sense in all languages, and also because intonational systems appeared to be be shared by very different languages. For example, in many languages, a raised pitch can be used in contrast with lower pitch to indicate that an utterance is intended as a question, rather than a a statement (Hirst & Di Cristo, 1998).

Each IP is characterized by one *nuclear accent* attached to a stressed syllable with a full vowel. The nuclear accent is a pitch pattern that lends prominence to the syllable that bears it. Such a pitch pattern can be either a pitch movement, a jump in pitch or the point of a change in direction of the pitch contour. In addition, each IP ends with a *boundary tone*, typically marked by a decrease in pitch (Beckman & Pierrehumbert, 1986; Hayes & Lahiri, 1991).

While overall the pitch tends to decline over the course of an utterance, it is *reset*, especially at the borders of the IP (Maeda, 1974; Pijper & Sanderman, 1994; Swerts, 1997; Yu-fang & Bei, 2002)

IPs are thus perceptually salient, and they account for natural break points in speech. That is, being domains of perceptually coherent intonational contours,

pauses can be inserted at IP boundaries without disturbing the pitch contour. In 1, square brackets mark IPs (from Nespor & Vogel, 1986):

(1) [Lions,]$_{IP}$ [as you know,]$_{IP}$ [are dangerous.]$_{IP}$

IPs appear to be obligatorily required for certain syntactic constituents such as parentheticals (as in the example above), unrestrictive relative clauses, preposed adverbials, tag question, expletives and vocatives (e.g., Selkirk, 1978; Nespor & Vogel, 1986). However, as with all prosodic constituents, syntax is not sufficient to account for IPs.

For example, the length of an utterance determines the number of IPs: for the same syntactic structure, the longer the utterance, the more the number of IPs. However, the quicker the speech rate, the fewer are the IPs. Consequently, speech styles which result in slower speech, lead to more IPs in an utterance. Thus, the assigning of the different ϕs in a sentence to IPs might be based on physiological mechanisms, like breath capacity (Lieberman, 1967; Nespor & Vogel, 1986; Vaissière, 1995).

In addition, not all ϕs can be IPs. For example, the sentence "Three mathematicians in ten derive a lemma" cannot be divided into IPs as follows:

* [Three mathematicians]$_{IP}$ [in ten derive a lemma]$_{IP}$

That is, IP boundaries tend to occur at the end of an NP, but not after nouns within NPs. Indeed, Selkirk (1984) proposed that each IP is a *Sense Unit*: two constituents C_i and C_j form a sense unit if C_i depends on C_j (either a modifier or a complement).

Several researchers have attempted to give a comprehensive model of how IPs are constructed, and how they are utilized for online comprehension (e.g., Watson & Gibson, 2004; Frazier & Clifton, 1998). Such studies have shown that speakers tend to place IP boundaries before and after large syntactic constituents, and that listeners use IP boundary cues as signals to where syntactic phrases may attach (see also Frazier, Carlson, & Clifton, 2006).

Finally, several behavioral results have shown an effect of IP boundaries in processing fluent speech (for example, Watson and Gibson, 2004; see Cutler, Dahan, and van, 1997 for a review of prosodic effects in speech comprehension).

To summarize, the exact nature and occurrence of IPs depend on a multitude of factors. Nevertheless, what is of primary interest is that (a) IPs are clearly marked in fluent speech and (b) IPs are aligned with words; the edges of IPs are always also the edges of words.

Summary In this chapter, we looked at the prosodic organization of speech. We saw that

- A specific constituent of the prosodic hierarchy, the phonological word, corresponds roughly to our intuitive notion of a 'word'.
- Words can be thought of as sequences of syllables; each word is made up of at least one syllable.
- The nature of the prosodic hierarchy establishes that words are aligned with larger prosodic constituents like the phonological phrase and the intonational phrase.
- The intonational phrase is clearly marked in the signal.

Put together, it is clear that prosody may constitute an aid in segmenting fluent speech. We will examine speech segmentation in the next chapter.

Alice remained looking thoughtfully at the mushroom for a minute, trying to make out which were the two sides of it; and as it was perfectly round, she found this a very difficult question.

Alice's adventures in wonderland,
Lewis Carroll

Chapter 3

Implicit segmentation of fluent speech

In his *Cours de linguistique général*, the Swiss linguist Ferdinand de Saussure hypothesized:

> Even if we could record on film all the movements of the mouth and larynx in producing a chain of sounds it would still be impossible to discover the subdivisions in this sequence of articulatory movements; we would not know where one sound began and where another ended. *(Quoted in Jakobson, 1942)*

This intuition was subsequently confirmed by Menzerath and De Lacerda (1933), who made an X-ray film of the working of the vocal apparatus and showed that the act of speech is a continuous, uninterrupted articulatory gesture.

Thus, it has been long recognized that the continuous flux of fluent speech offers few obvious cues to word boundaries (e.g., Klatt & Stevens, 1973). In fact, in some cases, the origin of modern forms of words can be traced to the mis-segmentation of speech caused by ambiguous word boundaries. For example, the English words 'orange' and 'apron' are derived from the mis-segmentation of the Middle English 'narange' and 'napron' due to confusion with the indefinite article, for example in the phrase 'a narange' (Cole et al., 1980).

Nevertheless, for the most part, speech segmentation is an automatic and effortless process. A fundamental challenge in speech perception is to understand

how a continuous signal yields discrete percepts. In the previous chapter we saw how words are organized into a hierarchy of prosodic constituents to create spoken utterances. In this chapter, we will look at how a spoken utterance is broken down into a series of words.

Implicit Segmentation

The speech segmentation problem can be considered from two perspectives, the developmental and the mature. The developmental perspective tries to understand how an infant (or an adult hearing a language s/he does not understand) learns to parse fluent speech into discrete words of unknown size and constitution. The perspective from the mature, expert language user, tries to understand how speech is effortlessly broken down into, for the most part, a known series of words.

For an adult, recognizing an utterance as a sequence of (known) words is tantamount to segmenting it. Such a strategy, wherein an utterance is explicitly recognized as a series of words has been termed *Explicit Segmentation* (e.g., Cutler & Fodor, 1979; Norris, McQueen, & Cutler, 2000). Clearly, such a strategy is available only once an inventory of words (the *lexicon*) is acquired.

In contrast, *Implicit segmentation* refers to analyses of speech that incidentally lead to finding word boundaries. For example, recall from the previous chapter that the boundaries of phonological phrases are also word boundaries (see Section 2.2.2 on page 20). Later in this chapter, we will see that upon identifying a ϕ in fluent speech, neither infants nor adults attempt to look for words that span that ϕ boundary. Thus, a byproduct of identifying a ϕ in fluent speech is discovering the boundary of a word.

In this thesis we are interested in issues regarding language acquisition. Thus, we will examine different strategies for speech segmentation that have been proposed under the general rubric of implicit segmentation. These strategies have been grouped below into two broad categories: *Prosodic* cues and *Statistical* cues.

3.1 Prosodic cues

In the previous chapter we saw that the edges of words (morphemes) are aligned with the edges of larger prosodic constituents: the phonological phrase and the intonational phrase. How can this information be used to segment speech?

For an infant to be able to use prosody in segmenting fluent speech, three conditions must be satisfied (e.g., Jusczyk & Kemler, 1996):

1. There should be physical manifestations of the prosodic constituents or their edges.

2. Infants should be sensitive to such physical manifestations.

3. Infants should be able to use such physical manifestations in organizing the speech input and constraining lexical search.

What are the physical manifestations of prosodic constituents? It has been established that prosodic phrase boundaries are marked by a variety of acoustic cues that involve intonation, pausing, and duration. For example, several authors have found evidence for pre-boundary lengthening associated with major phrase boundaries (e.g., Klatt, 1976; Macdonald, 1976; Lehiste, Olive, & Streeter, 1976; Scott, 1982).

With the development of the theory of the prosodic hierarchy (see 2.2 on page 11), it was seen that pre-boundary lengthening of a segment was a function of its position within the prosodic hierarchy (e.g., Ladd & Campbell, 1991; Gussenhoven & Rietveld, 1992; Wightman, Shattuck-Hufnagel, Ostendorf, & Price, 1992; Cambier-Langeveld, 2000)[1].

Jun (1993) found that the *voice onset time* (VOT) of a Korean consonant depended on the position within the prosodic hierarchy, suggesting that the speech production system is sensitive to the prosodic hierarchy. Indeed, Fougeron and colleagues (Fougeron & Keating, 1997; Keating, Cho, Fougeron, & Hsu, 2003) showed that the *articulatory effort* for a segment was a function of its position in the prosodic hierarchy. Articulatory effort refers to the amount of lingual articulation; a greater lingual articulation results in greater linguopalatal contact. These

[1] These studies analyzed utterances in English or Dutch.

authors found that the higher up in the prosodic hierarchy was a constituent, the greater was the linguopalatal contact exhibited by its initial segment. So, for example, consonants at the onset of intonational phrases showed greater articulatory strengthening than those at the onset of phonological phrases (see also Fujimura, 1990).

Let us now consider data that suggests that both infants and adults can utilize acoustic cues that mark phonological phrases and intonational phrases.

3.1.1 Detecting and using ϕs

Can infants detect the acoustic cues to phonological phrase boundaries? Christophe, Dupoux, Bertoncini, and Mehler (1994) tested this with French newborns. These authors extracted bisyllables from speech that either did or did not contain a ϕ boundary. For example, the bisyllable *mati* was extracted either from inside a single word (like "mathé*mati*cien" or "cli*mati*sé") or from the junction of two words separated by a ϕ (like "panora*ma ty*pique" or "ciné*ma ti*tanesque"). Acoustically, the authors found pre-boundary lengthening as in previous studies. Behaviorally, 3-day-old infants were found to discriminate the two kinds of bisyllables.

However, in French, the last syllable of a word typically also carries stress, so that the word(ϕ)-internal and ϕ-straddling bisyllables differ due to the location of stress. In order to control for this potential confound, Christophe, Mehler, and Sebastián-Gallés (2001) replicated the previous results with French newborns, using Spanish stimuli. In this study, lexical stress was the same for both word(ϕ)-internal bisyllables (like *latí* from 'ge*latí*na' or 'escar*latí*na) and for ϕ-spanning bisyllables (like *latí* from 'Manué*la tí*mida' or 'goríla *tí*sico'). Again, French newborns discriminated the ϕ-internal from the ϕ-spanning bisyllables. Interestingly, in this study the authors did not find a significant lengthening of the pre-boundary vowel. However, the ϕ-initial consonant showed lengthening, the pre-boundary vowel showed a significantly higher pitch and the pre- and post-boundary segments both showed significantly higher amplitudes.

How does sensitivity to the acoustic cues that mark ϕ boundaries aid in processing speech? Christophe, Gout, Peperkamp, and Morgan (2003) proposed that

infants do not attempt lexical access on syllable sequences that span ϕ boundaries. Indeed, Gout, Christophe, and Morgan (2004) showed that 10- and 12.5-month-old infants are able to use ϕ boundaries to constrain on-line lexical access. In this study, the authors used a variant of the conditioned head-turn technique. Infants were first trained to turn their heads for isolated bisyllabic words (like 'paper'). Subsequently, infants were exposed to sentences in which the target bisyllable occurred either within a ϕ, or straddling it as in the examples below:

> ϕ-*internal:* [The scandalous **paper**]$_\phi$ [sways him]$_\phi$ [to tell the truth]$_\phi$.
> ϕ-*straddling:*[The outstanding **pay**]$_\phi$ [**per**suades him]$_\phi$ [to go to France]$_\phi$.

The authors found that, in the second phase, infants turned significantly more towards the sentences in which the target word did not straddle the ϕ.

More recently, Soderstrom, Seidl, Kemler Nelson, and Jusczyk (2003) have extended these results with single words to noun- and verb-phrases. These authors familiarized 6- and 9-month-old with sequences of words that were NPs and VPs. In a subsequent test phase, the infants reacted significantly to the presence of a familiarized syntactic phrases only when it corresponded to a phonological phrases, but not when it spanned a phonological phrase boundary.

The view that prosody constrains segmentation is strengthened by the finding that in adults, prosody constrains lexical access in a word recognition paradigm (Christophe, Peperkamp, Pallier, Block, & Mehler, 2004). In this study, French adults had to respond to the presence of a target word (for example, *chat* /ʃa/[2]) that could occur in a locally ambiguous context (e.g., *chat grincheux* /ʃagʀɛ̃ʃø/ where *chagrin* /ʃagʀɛ̃/ is a French word), or in a locally unambiguous context (like *chat drogué* /ʃadʀoge/; there is no French word starting with /ʃad/). The authors found that the word *chat* was responded to faster in the unambiguous than in the ambiguous context. However, this delay in detecting *chat* in an ambiguous context disappeared when a phonological phrase boundary occurred immediately after the target word (for example, *[le gros chat] [grimpait . . .]*, /ləgʀoʃa#gʀɛ̃pɛ/, wherein the possible word *cha#grin* is now interrupted by a phonological phrase boundary as indicated). In other words, phonological phrase boundaries appear to act as natural boundaries; lexical access is curtailed by such boundaries.

[2]Pronunciations are marked in IPA throughout.

3.1.2 Detecting and using IPs

Variations in pitch can indicate the boundaries of larger prosodic constituents, especially the intonational phrase. Recall (Section 2.2.2, pg. 22) that IPs correspond to single intonational contours.

Figure 3.1 shows a single (English) IP[3]. This IP contains two ϕs, marked with 'P' in the text transcription. In the figure, the end of the first ϕ is marked by a low boundary tone (L_P), while the end of the second ϕ, which is also the end of the IP, is marked by another low boundary tone (L_I). Notice the single intonational contour with a peak in F0 at the nuclear accent (H*) on the main stressed word, *Típperary.*

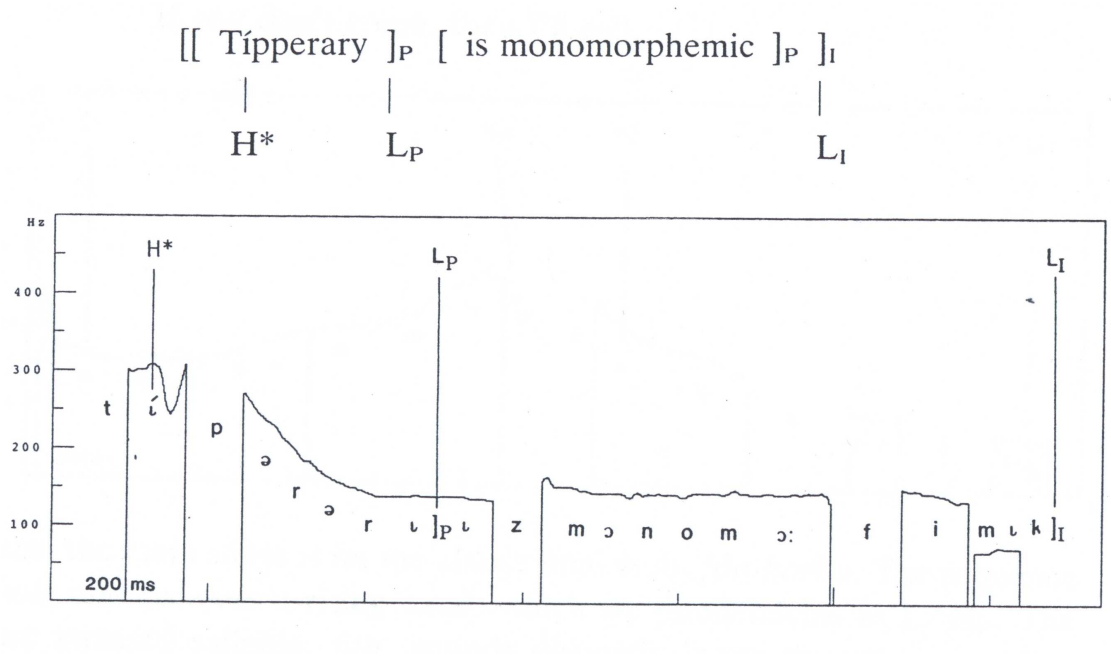


Figure 3.1: Single IP, showing the pitch contour. The x-axis is time, while the y-axis is pitch (Hz).

It has been proposed that both the IP as a constituent and some of its properties are universal. In general, several authors have proposed that the decline in pitch associated with large prosodic units has its underpinnings in physiology;

[3]The figure is from from Hayes and Lahiri, 1991, pg. 68. (Tipperary is the name of an Irish town.)

for example due to diminishing air supply in the lungs, causing a decrease in subglottal air pressure (e.g., Lieberman & Blumstein, 1988; Strik & Boves, 1995), or due to a collapsing ribcage that results in a downward pull of the larynx (Maeda, 1974, but see Ohala, Dunn, and Sprouse, 2004) Indeed, some of the characteristics of IPs are shared also with musical phrases (e.g., Jackendoff & Lerdahl, under review), suggesting that certain aspects of both music and language might derive from common, general cognitive mechanisms.

Are infants sensitive to IPs in fluent speech? Hirsh-Pasek et al. (1987) showed that 4.5 month-old (and 9 month-old) infants prefer utterances with artificially inserted pauses at clause boundaries, as opposed to utterances with pauses inserted in the middle of clauses; typically, clauses correspond to IPs (see also Jusczyk, Pisoni, & Mullennix, 1992; Kemler, Hirsh-Pasek, Jusczyk, & Wright-Cassidy, 1989; Morgan, 1994). Such preference was present even for low-pass filtered speech (Jusczyk, 1989; Jusczyk et al., 1992). Low pass filtering removes several fine acoustic details (making it difficult, for example, to identify the phonemes), but preserves broad intonational features, suggesting that infants are indeed sensitive to intonational features, which define IPs.

Morgan, Swingley, and Miritai (1993) inserted buzzing noises, rather than pauses at clause boundaries or in the middle of clauses. Even in this case, infants preferred clauses with buzzes at boundaries over clauses with buzzes inside them. Finally, (Morgan, 1994) showed that infants could *detect* non-linguistic noises best when these were placed at clause boundaries rather than inside clauses. These studies confirm the notion that infants perceive IPs as coherent wholes.

Adult ERP studies have shown that IP boundaries elicit a characteristic component, the Closure Positive Shift (CPS). The CPS is observed even when speech is low-pass filtered, suggesting that it is indeed a signature of prosody (e.g., Steinhauer, Alter, and Friederici; Steinhauer and Friederici; Friederici, Steinhauer, and Pfeifer, 1999; 2001; 2002; see Steinhauer, 2003 for an overview). Recently, the CPS signature has been observed even with very young infants (Pannekamp, Weber, & Friederici, 2006). Thus, electrophysiological studies provide converging evidence that young infants are sensitive to IP boundaries in speech.

Do infants use this information in organizing speech? Mandel and her colleagues showed that 2-month-olds were better able to remember phonetic proper-

ties of words that lay within a clause as opposed to words that spanned two contiguous clauses (Mandel, Jusczyk, & Nelson, 1994). In addition, Mandel, Kemler, and Jusczyk (1996) showed that 2-month-olds were able to detect a change in word order when a pair of words were part of a well-formed clause than when a prosodic boundary separated the two words. Subsequently, Nazzi, Kemler, Jusczyk, and Jusczyk (2000) showed that 6-month-olds could detect previously heard word sequences in fluent speech only if the sequence did not contain an IP boundary inside it. For example, infants were exposed to IP-internal or IP-straddling sequences excised from passages like in the following (sequences in bold: IP-internal, underlined sequences: IP straddling):

1. John doesn't know what rabbits eat. **Leafy vegetables taste so good.** They don't cost much either.

2. Many animals prefer some things. **Rabbits eat leafy vegetables.** Taste so good is rarely encountered.

In a subsequent test phase, infants showed a preference for those passages that contained the previously heard, well-formed sequence. In a subsequent experiment, these authors ruled out explanations based on acoustic similarity.

Taken together, these findings show that pre-linguistic infants are sensitive to acoustic cues that mark IP boundaries, and use this information in organizing fluent speech.

3.2 Statistical cues

It is an empirical fact that speech contains a wealth of statistical information (e.g., Charniak, 1993). For example, in the English word game *Scrabble*, the letter 'a' has a value 1, while the letter 'z' has a value 10, reflecting their respective frequencies in the language. In the Polish version of the game, instead, the letter 'z' has a value 1, since it is very common[4].

Several authors have proposed that the distributional properties of sub-lexical segments (like syllables or phonemes) can help in discovering word boundaries

[4]http://en.wikipedia.org/wiki/Scrabble_letter_distributions#Polish

(e.g., Harris, 1955; Brent and Cartwright, 1996; Gow, Melvold, and Manuel, 1996; Dahan and Brent, 1999; Batchelder, 2002).

In the following sections, we will look at four cues: allophone distributions, phonotactics, lexical stress and transition probabilities.

3.2.1 Allophone distributions

We saw in the previous chapter that each phoneme in a language might have one or more allophones. For example, in English, the voiceless stop consonants (/p/, /t/,/k/) have aspirated and unaspirated allophones (e.g., the [p^h] in 'pin' versus the [p] in 'spin'). The choice of the allophone depends on the context (Church, 1987). In English aspirated allophones occur foot-initially. Given the prosodic hierarchy, (Section 2.1 of the previous chapter), this implies that the unaspirated allophones will never occur at the beginning of an utterance. Thus, a learner might be biased to place a word-boundary preceding utterance-medial aspirated voiceless stops (e.g., Gow & Gordon, 1995).

Indeed, Hohne and Jusczyk (1994) showed that even two-month-olds are sensitive to allophonic information, being able to discriminate the allophonic versions of /t/ and /r/ in 'nitrates' versus 'night rates'. Further, Jusczyk, Houston, and Newsome (1999) showed that 10.5-month-olds can use this information in parsing fluent speech.

3.2.2 Phonotactics

Phonotactics refers to the restrictions in a language on the permissible combination of phonemes. Phonotactics can be interpreted as a set of constraints over possible phonemes at different positions within words, morphemes and syllables and their combinations. In English (but not, for example, in Dutch), the consonant sequence /kn/ cannot be a syllabic onset. Thus, there can be no words in English starting with /kn/. Therefore, a bias to place word boundaries between /k/ and /n/ will often (but not always, for example in the word 'hackney') lead to successful word segmentation (e.g., Church, 1987). Indeed, as the example shows, phonotactics need not be an all-or-none phenomenon. *Probabilistic phonotactics* refers to the fact that certain sound sequences, although not absent, are rare in

certain positions.

Several lines of research have revealed than infants are sensitive to phonotactic constraints, and they use these to discover word boundaries (e.g., Friederici & Wessels, 1993). By nine months of age, infants prefer to hear not only the appropriate phonotactics (Jusczyk, Cutler, & Redanz, 1993), but also the appropriate probabilistic phonotactics of their language (Jusczyk, Luce, & Charles-Luce, 1994). In addition, they can also use this information to segment fluent speech (Mattys & Jusczyk, 2001).

3.2.3 Stress patterns

The manifestation of stress is highly language dependent. In some languages, stress is fixed with respect to lexical items. For example, in Hungarian, stress always occurs on the initial syllable of a lexical item. In other languages stress is not fixed, but can be deduced from the phonological properties of the word. For example, in Latin polysyllabic words, stress occurs on the penultimate syllable when it is heavy, and on the antepenultimate syllable when the penultimate syllable is light. Finally, in some languages, stress is lexicalized. That is, the stress pattern cannot be deduced from a general rule, but must be memorized for each lexical item.

Nevertheless, even in languages with lexicalized stress, there can be strong statistical tendencies. For example, in English, stress is primarily on the first syllable in the words, while in Italian, stress tends to be on the penultimate syllable. For English, Cutler and Carter (1987) estimated that about 90% of common content words in conversational speech begin with a strong syllable. Based on this finding, Cutler and Norris (1988) proposed the Metrical Segmentation Strategy (MSS), wherein a word boundary is placed before each stressed syllable.

Indeed, (Jusczyk et al., 1999) found that 7.5-month-old infants were biased towards segmenting strong-weak bisyllables (like ‘*kingdom*) over weak-strong bisyllables (like *gui*‘*tar*) from fluent speech.

The three cues discussed above reflect language-specific distributional regularities over segments (allophone distributions), segment combinations (phonotactics)

and suprasegmental properties (stress).

However, it is possible that these are not purely statistical variations, but have their origin in the physiology of speech production and perception. For example, while the English phoneme /a/ is realized as an [a] in a word like *pat*, when it occurs between two nasal consonants (as in the word *man*), it is realized as the nasal allophone, [ã] due to coarticulation. Several studies have shown that the effect of coarticulation is blocked by phonological phrase boundaries (e.g., Hardcastle, 1985; Byrd, Kaun, Narayanan, & Saltzman, 2000). As a consequence, allophones that result from coarticulation (a physiological constraint) will have a different distribution at the edges and in the middles of phonological phrases (a distributional cue).

Similarly, certain phonotactic constraints might originate in the physiology of speech. For example, the phonotactic regularity that English utterances never begin with the sound sequence [lpk], might simply be due to the fact that such a sequence is very hard to produce. Again, notice that such a sequence can occur utterance medially, such as in the phrase 'he**lp k**ittens'.

Finally, it has been known that strong syllables are salient even for very young infants (e.g., Echols, 1993, 1996). Recall that Jusczyk et al. (1999) found that by 8 months of age English infants can segment trochaic bisyllables (the predominant English pattern). More recently, Nazzi, Iakimova, Bertoncini, Frédonie, and Alcantara (2006) examined French, wherein bisyllabic words are typically iambic. These authors found that, in a similar task, 8-month-old French infants did not show any evidence of segmenting iambic bisyllables from fluent speech. Even the segmentation of individual syllables from the bisyllablic items was delayed. Only by 12 months of age did the French infants show any evidence of recognizing single syllables, and this was limited to the stronger, second syllable.

Nazzi et al. (2006) propose that these cross-linguistic differences arise due to difference in linguistic rhythm in English and in French. An alternate possibility is that there might be an innate bias to place word boundaries before strong syllables. In English, such a strategy does yield words, while in French, such a strategy will mis-segment words. Thus in French, but not in English, stress-based cues will be mismatched with other word boundary cues, possibly explaining the delay in identifying (iambic) words in French as compared to English.

3.2.4 Transition Probabilities

Let us now examine a very general strategy for segmenting fluent speech. If we consider speech as a sequence of phonemes or syllables, it is clear that this sequence is far from random, and it has been proposed that distributional regularities in speech could help in segmenting it (e.g., Hayes & Clarke, 1970).

For example, Harris (1955) examined a corpus of utterances, transcribed as sequences of phonemes. Let P_n^u be the nth phoneme in the uth utterance. At each P_n^u, a count C_n^u was made of all the phonemes that occur in the position $(n+1)$, in all available utterances, following the string of phonemes from P_1^u to P_n^u. It was found that the positions within the utterances where C_n^u was high corresponded to the ends of morphemes. As a result of such a procedure, an utterance like /hiyzkwikər/ (*He's quicker*) is segmented as /hiy.z.kwik.ər/.

Intuitively, this procedure indicates the coherence of a string of phonemes. Within a coherent string, each individual phoneme strongly predicts the next. However, the last phoneme of a cohesive group can be followed by a variety of other phonemes, and thus at the last position, the C_n^u is high.

In order to gain insight into distributional strategies to segmenting words, Saffran, Newport, and Aslin (1996) considered utterances as sequences of syllables. Recall (Section 2.2.1 on page 14) that words contain at least one syllable. Thus, a statistical procedure for clustering syllables will lead to the discovery of words. Saffran et al. (1996) formalized the intuition that, within a multi-syllabic word, each syllable will be highly predictable of the next, compared to the predictability of any syllable following the last. The proposed the (forward) *transition probability* (TP) as an index of statistical coherence. The TP from any syllable x to another syllable y can be estimated by

$$TP(x \rightarrow y) = \frac{frequency(xy)}{frequency(x)} \tag{3.1}$$

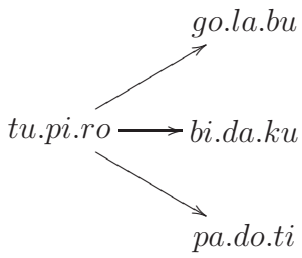
In general, the TP inside a word will be higher than the TP between one word and the next. Consider for example, the phrase "pretty baby". The first syllable, *pre* can be followed by a few other syllables (like in '**pri**ckle' or '**pri**mitive). However, the second syllable **ty** can be followed by scores of other syllables (for example those from all the words that can follow 'pretty'). Thus, whereas the TP

from the first to the second syllable of 'pretty' is high, there is a trough ('dip') in TP after the syllable 'ty'.

Saffran and colleagues (e.g., Saffran et al., 1996, 1996; Aslin, Saffran, & Newport, 1998) showed empirically that transition probabilities (TPs) between syllables can serve as cues to the segmentation of monotonous streams of continuous speech. In a series of experiments, they demonstrated that both infants and adults could use troughs ("dips") in TPs as cues to word boundaries to segment artificial speech streams that lack prosodic cues (see also Saffran, 2001; Thiessen & Saffran, 2003).

For example, Saffran et al. (1996) exposed infants to artificial speech streams constructed by randomly concatenating 45 tokens each of four trisyllabic nonce words, *tupiro*, *golabu*, *bidaku* and *padoti* (immediate repetitions were disallowed). There were no cues to the word boundaries, so a segment of the artificial speech stream can be orthographically represented as *. . . tupirobidakupadotibidakutupiro. . . .*

Diagrammatically, we can see that each word (e.g. *tupiro*) can be followed by one of the three other words (syllables are separated by dots):



Thus, the TP is 1.0, going from the first (or second) syllable of a nonce word to the next. However, from the last syllable of a nonce word, the TP to the next is 0.33. Saffran et al. (1996) showed that 8-month-old American infants familiarized to such streams were subsequently able to discriminate between a nonce word and a trisyllabic sequence that had never appeared during familiarization (a 'non-word', e.g., *dapiku* or *tilado*).

These authors further showed that 8-month-olds could also discriminate between such nonce words and 'part-words'; a part-word consisting of syllables from two adjacent nonce words. For example, from the sequence of words *tu.pi.ro#go.la.bu* (the # represents the word boundary), the part words can be *pi.ro#go* or *ro#go.la.*

That is, 8-month-old infants can discriminate between two trisyllabic sequences that had both occurred during familiarization, but which differed in their average TPs. However, note that there is a confound. A part-word like *pi.ro#go* can occur only when the 'word' *tu.pi.ro* is followed by *go.la.bu*. This implies that the frequency of a part-word in this experiment is a third the frequency of a 'word'.

In order to disentangle the effect of frequency and TP, Aslin et al. (1998) created an artificial speech stream similar to that in Saffran et al. (1996), but in which two trisyllabic nonce 'words' were twice as frequent as two other such 'words'. As a consequence, the low-frequency 'words' had the same absolute frequency as the high-frequency part-words, while the TPs were higher for such 'words' than for such part-words. Nevertheless, 8-month-olds reliably discriminated the two kinds of trisyllabic items. This experiment suggests that, over and above the frequency of occurrence, it is the high average TP that gives a certain coherence to the words.

Finally, Saffran (2001) showed that 8-month-olds indeed seem to represent high-TP trisyllables as words; after being familiarized with an artificial speech stream as before (e.g., Saffran et al., 1996), infants preferred such 'words' embedded in simple English contexts (e.g., "I like my *tibudo*").

Summary In this chapter, we looked at various implicit strategies for segmenting fluent speech. We concentrated on implicit strategies since these are most relevant at the earliest stages of word segmentation, when the lexicon of the infant (or of an adult learning a second language) lies empty. In the next chapter, we will look at how various cues might be put together to simplify the task of segmentation.

> . . . there was only one road through the wood, and the two finger-posts both pointed along it. 'I'll settle it,' Alice said to herself, 'when the road divides and they point different ways.'
>
> Through the looking glass,
> Lewis Carroll

Chapter 4

Towards an interactive model

In the previous chapters we looked at the prosodic organization of speech and various strategies for speech segmentation. We saw that there are several cues that can be utilized in segmenting speech. Some of these cues are derived from the input, while others might represent innate biases.

It is clear that the various cues to word boundaries do not occur in isolation. Whenever an utterance is heard, the listener presumably employs the entire suite of computations that can yield word boundaries. Indeed, several authors have shown that computational models of speech segmentation benefit greatly by the judicious, simultaneous use of all available cues (e.g., Brent & Cartwright, 1996; Batchelder, 2002). Similarly, developmental psycholinguists have explored how infants integrate multiple word boundary cues (e.g., Mattys, Jusczyk, Luce, & Morgan, 1999; Johnson & Jusczyk, 2001). Yet, there are no cognitive models to explain *how* two or more cues can interact in beginning to segment fluent speech.

The goal of this thesis is to build a model of how different cues to segmenting speech can interact. In particular, we will examine how distributional cues are affected by intonational phrase prosody. In this chapter, I describe the logic of the experiments and the organization of the empirical investigations.

4.1 Prosody and statistical cues

In the previous chapters, we saw that the perception of phrasal prosodic constituents can act as cues to word segmentation, since the edges of such phrasal

constituents are the edges of words. We also saw that sequences of syllables with a high transition probability (TP) between them are perceived as cohesive, word-like units. We would like to know *how* these two cues interact.

Previous studies have revealed that indeed prosodic and statistical cues interact in segmenting speech. For example, Morgan and Saffran (1995) demonstrated that English-learning 6-month-olds represented a pair of syllables as a coherent unit whenever there was a rhythmic regularity to the sequence, for example, they formed a strong-weak, *trochee*. Recall that in English, the vast majority of common words begin with a strong syllable (e.g., Cutler & Carter, 1987). However, 9-month-olds supported a grouping of a pair of bisyllables only when such a pair displayed both a rhythmic regularity and appeared in the same sequential order. That is, by 9-months of age, infants are able to put together statistical and rhythmic cues in forming multisyllabic percepts.

Johnson and Jusczyk (2001) provided further evidence for an interaction between various cues. In particular, these authors found that English 8-month-olds weigh stress and co-articulatory cues more heavily than statistical cues. More recently, Thiessen and Saffran (2003) pitted TPs against stress patterns in English-learning infants. In this study, artificial speech streams were created as an alternation of strong and weak syllables. However, the TPs were relatively higher going from a weak to a strong syllable than from a strong to a weak syllable. Thus, while the stress cue groups the syllables as a sequence of trochees, (in square brackets below), TPs group the syllables as *iambs* (weak-strong bisyllables, indicated by the overbraces):

$$\cdots [\sigma_s \overbrace{\sigma_w}][\sigma_s \overbrace{\sigma_w}][\sigma_s \overbrace{\sigma_w}][\sigma_s \cdots$$

The findings of these authors suggest that 7-month-old infants group the bisyllables according to TPs, so a coherent bisyllable is weak-strong, although in English strong syllables are typically word-initial. In contrast, for older, 9-month-old infants, the stress cues take precedence, and they consider strong-weak, low-TP bisyllables as coherent. Put together, the various findings suggest that by 9 months of age, infants are able to utilize and integrate multiple cues to word boundaries.

Turning to adults, Saffran et al. (1996) presented American adult participants with artificial speech consisting of a concatenated list of trisyllabic nonce words. In a subsequent test phase, participants showed a preference for the 'words' over non-words. In a second condition, these authors lengthened the vowel of either the first or the third syllable of the trisyllabic nonce words. They found that lengthening the final vowel influenced segmentation; participants in this condition outperformed participants in the initial-lengthening and the no lengthening conditions. Similarly, Bagou, Fougeron, and Frauenfelder (2002) evaluated the contribution of pitch and lengthening cues in adult Swiss French participants. They found that both pitch rise on the final syllable and final lengthening facilitated the segmentation high-TP trisyllables in artificial speech streams.

Toro, Mattys, and Sebastián-Gallés (submitted) made a comparative study of Spanish, English and French adults segmenting artificial speech made up of trisyllabic items. These authors introduced 'stress' in either the initial, the middle or the final syllable of the trisyllabic nonce words. They found that when such 'stress' was on the middle syllable, all three populations were at chance. However, when the 'stress' was on either the first or the last syllable, all three populations performed as well as in the absence of any prosodic cues. These authors suggested that prominent syllables might play a universal role in speech segmentation.

Thus, most such studies have shown an interaction between TP computations and the prosodic properties of constituents smaller than a phrase (lexical stress). We will now examine the outline of an experimental design to study the interaction between *phrasal* prosody and TPs

4.2 Outline of the empirical investigations

How can we study an interaction between phrasal prosody and TP between syllables? Let us begin by asking the question in the following manner: How is the segmentation of a nonce word affected by its location within a prosodic phrase?

Thus, we shall compare the segmentation of a nonce word in three different positions with respect to a prosodic phrase:

- Straddling a phrasal boundary.

- Aligned with the edge of a phrasal boundary.
- In the middle of a prosodic phrase.

Experiments that examine the role of TPs employ artificial speech that is made up of a few multisyllabic (e.g. trisyllabic) nonce words, concatenated at random (e.g., Saffran et al., 1996, Aslin et al., 1998; see also Section 3.2.4 on page 36 of the previous chapter). As a result, the location of 'words' inside the artificial speech stream is highly constrained. We would like to develop an experimental paradigm that allows the arbitrary placement of nonce words in artificial speech streams.

In the the first experimental chapter (Chapter 5 on page 47), we will see that Italian adults can recover high-TP nonce words from fluent artificial speech, even in the presence of *syllabic noise*. Syllabic noise refers to the presence of additional syllables besides the ones that make up the nonce words. For example, if $\boldsymbol{\sigma^a\sigma^b\sigma^c}$ is a trisyllabic nonce word, and $\boldsymbol{\sigma^d\sigma^e\sigma^f}$ is another, then a segment of artificial stream containing syllabic noise might look like:

$$\ldots \sigma_x\ \sigma_x\ \sigma_x\ \ \boldsymbol{\sigma^a\sigma^b\sigma^c}\ \ \sigma_x\ \sigma_x\ \ \boldsymbol{\sigma^d\sigma^e\sigma^f}\ \ \sigma_x\ \sigma_x\ \sigma_x\ \ldots$$

where σ_x stands for the *class* of noise syllables, consisting of syllables that do not contribute to any of the 'words'. The TPs between the noise syllables is kept low relative to the TP in the 'words'.

The finding that the presence of these random syllables does not interfere with the extraction of the high-TP 'words' implies that, by manipulating the relative positions of the 'words' and the random syllables, we can place the target 'words' in arbitrary locations. In Chapter 6 (pg. 59), this fact is put to use to create syllabic *frames*, consisting of a series of noise syllables. To each frame is associated the prosodic contour from one IP from the native language, such that each frame now represents an artificial IP. The trisyllabic 'words' can now be placed in various positions inside such frames; equivalent to placing them in different positions within an (artificial) IP. We will see that ***'words' in the middles of IPs are better extracted than 'words' straddling IPs.***

The results of Chapters 5 and 6 together constitute the basic experimental observation of an effect of phrasal prosody on the extraction of statistically defined

'words'. These results suggest that adult participants perceive the fluent artificial speech as being divided into 'phrases' due to prosody. Indeed, we will see in Chapter 7 (pg. 71) that ***'words' at the edges of such 'phrases' are better recognized than 'words' in the middles of the 'phrases'.***

In Chapter 8 (pg. 77), we will try to understand *how* prosody influences the extraction of statistically defined 'words'. This chapter thus represents the theoretical core of the thesis; evidence will be presented that suggests that ***prosody and statistical computations occur in parallel, and memory systems are involved in putting together the outputs of the two systems.***

Chapter 9 (pg. 89) examines alternate explanations for the results obtained in the previous chapters. In particular, we will see that the preference for 'phrase'-internal 'words' over 'phrase'-straddling 'words' is not merely an acoustic phenomenon; recourse to an abstract encoding of the artificial speech streams is required.

How robust are prosodic cues to phrase boundaries? Recall from Chapter 3 (Sections 3.1.1, pg. 28, and 3.1.2, pg. 30) that phrasal prosodic units are available even to young infants, suggesting that at least some of the properties of IPs might be universal. Thus, in Chapter 10 (pg. 103), we will look at the effect of a non-native prosody on adult participants. It will be seen that all the key findings using a native prosody are replicated using 'phrases' with a non-native prosody.

Finally, Chapter 11 (pg. 113) examines the contribution of various acoustic cues in determining 'phrasal' units in this artificial speech paradigm.

In the concluding chapter of the next part, we will look at the broader implications of these findings.

Part II

Empirical investigations

The Red Queen shook her head, 'You may call it "nonsense" if you like,' she said, 'but I'VE heard nonsense, compared with which that would be as sensible as a dictionary!'

Through the looking glass,
Lewis Carroll

Chapter 5

Segmenting using statistics under 'noisy' conditions

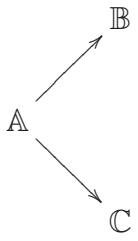
The goal of this chapter is to device an experimental paradigm that allows the arbitrary placement of trisyllabic nonce 'words' in an artificial speech stream. Let us first try and understand how exposure to a stream of syllables result in the computation of statistical regularities over them.

Recall that the TP from a syllable $\mathbb{A}$ to a syllable $\mathbb{B}$ is a measure of how frequently the sequence $\mathbb{AB}$ occurs compared to how frequently $\mathbb{A}$ alone occurs, that is,

$$TP(\mathbb{A} \rightarrow \mathbb{B}) = \frac{freq(\mathbb{AB})}{freq(\mathbb{A})}$$

Thus, if every time the syllable $\mathbb{A}$ occurs it is immediately followed by the syllable $\mathbb{B}$, the TP from $\mathbb{A}$ to $\mathbb{B}$ is 1.0: all occurrences of the syllable $\mathbb{A}$ are followed by the syllable $\mathbb{B}$.

Instead, imagine that the syllable $\mathbb{A}$ is followed either by the syllable $\mathbb{B}$ or by the syllable $\mathbb{C}$:



If both $\mathbb{B}$ and $\mathbb{C}$ are equally frequent, it follows that the TP from $\mathbb{A}$ to either $\mathbb{B}$ or to $\mathbb{C}$ will be 0.5. Thus, it would seem that in order to track TPs between syllables, one possible mechanism would be to track the frequency of all syllables and of all bisyllables.

However, it is clear that such a strategy would suffer a computational explosion as the number of syllables, N_σ, grows. For example, for $N_\sigma = 8$, the system would have to track the frequencies of 8 monosyllables and 64 bisyllables, while for $N_\sigma = 12$, the system would have to track the frequencies of 12 mono and 144 bisyllables.

Such a simplistic model of TPs would suggest that the efficacy of TP computations should decline with an increase in N_σ. However, one can think of several alternate algorithms that are functionally equivalent to computing TPs, but do not show a dependence on N_σ. But first, we must understand if, empirically, TP computations are indeed independent of N_σ.

Thus, we can ask: what would happen if the syllables that made up the (nonce) words comprised only a small fraction of all the syllables in the speech stream?

5.1 Pilot study: extracting 'words' from noise

In order to understand if TP computations are robust, it was decided to embed trisyllabic nonce words in a stream containing *noise syllables*. Noise syllables constitute a set of syllables distinct from those that make up the nonce 'words'. These noise syllables are randomly interspersed between the trisyllabic 'words', but themselves show no statistical structure. That is, the TP between any two noise syllables is low compared to the TP between the syllables that make up the 'words'. Importantly, the monosyllable frequencies are the same for the syllables that contribute to the noise and those that make up the 'words'.

If it can be shown that the presence of randomly interspersed syllables has little effect on the segmentation of embedded 'words', this would open up the possibility of being able to place 'words' in arbitrary locations with respect to themselves and to each other. This is tested in the following pilot study, the 'words' are embedded in a stream containing many random syllables.

In previous experiments (e.g., Saffran et al., 1996; Aslin et al., 1998), researchers have used between 6 and 18 syllables, all of which, in different combinations, form the nonce words. In this experiment, 52 syllables were used. Of these, only 12 syllables contributed to the four unique trisyllabic nonce words. The remaining 40 syllables were what we will refer to as the *'noise syllables'*[1]. Thus, in this experiment, N_σ is 52, which results in $2,704$ bisyllables. This is much higher than the 144 bisyllables that would need to be tracked for $N_\sigma = 12$ as in previous experiments. In this pilot study, *all* the syllables had the same absolute frequency.

Also, in previous adult experiments, researchers have typically used the two-alternative forced choice task (2AFC), comparing a 'word' against a part-word or a non-word in each trial. A *part-word* consists of a part of one word and a part of another. For example, if one 'word' is 'puliki' and another is 'beraga', a part-word would be 'kibera'. A *non-word*, instead, is a sequence that never occurs in the speech stream (for example, 'garali'). In this experiment instead, participants were asked to judge if individually presented trisyllabic tokens, 'words' or non-words, were heard during the familiarization phase (Appendix A).

The results from this experiment are displayed in Figure 5.1. From the figure, it is clear that participants rated the 'words' as being more familiar than they did the non-words.

One reason why segmentation is not affected by the presence of the interspersed noise syllables might be related to the distribution of TPs at the edges of 'words'. In previous experiments, where up to 6 'words' were concatenated at random, the TP from the last syllable of one 'word' to the first syllable of another was 0.2[2] (since immediate repetitions are not allowed, each 'word' can be followed by one of the other 5 'words'). In the present experiment, in contrast, the 40 interspersed noise syllables occur at random. Thus, the TP from the last syllable of a 'word' to any of the noise syllables is 0.025. At the left (leading) edge of 'words', any noise syllable can be followed by any of the other 39 noise syllable, but also by the first syllable of the four 'words'. Thus, the TP at the leading

[1]For details of this pilot experiment, please consult Appendix A.

[2]These, and the following, are approximate values, since the speech streams are finite and randomly created.

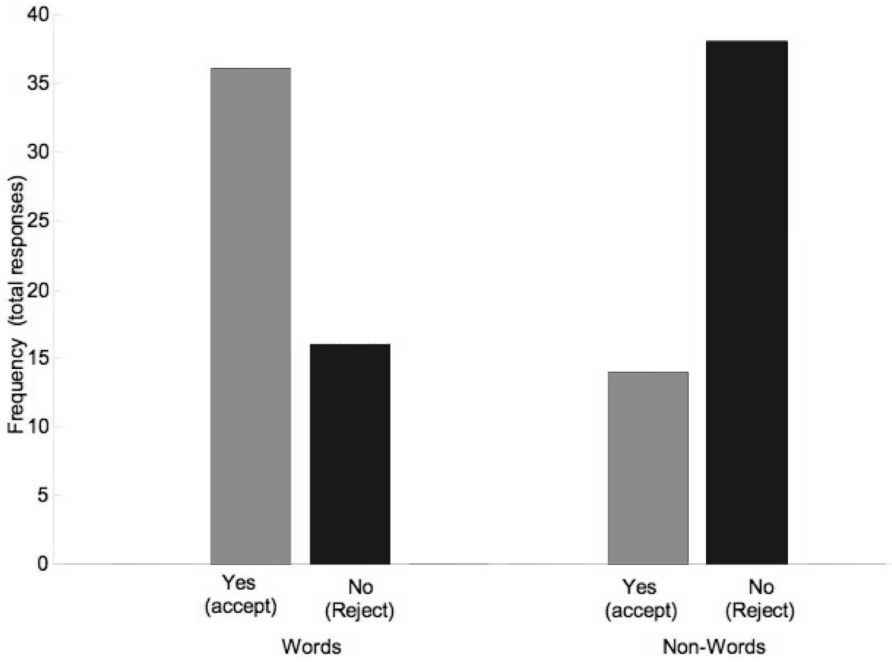


Figure 5.1: Frequency of 'Yes' (grey) and 'No' (black) responses, indicating accepting or rejecting having heard the 'words' (left) or the non-words (right) during familiarization. The graph shows the frequency of responses combined across all 13 participants. 'Words' are interspersed with noise syllables but are nevertheless segmented accurately. The overall segmentation score of 71.15% was significantly better than chance (two-tailed t-test, $p < 0.001$).

edge of 'words' is 0.023. In contrast, TPs between the syllables of the 'words' were always 1. Thus, possibly, the presence of a large amount of interspersed syllabic noise actually enhances the detection of words because of a large ratio (an approximately 40-fold difference in this case) between the word-internal TPs and the TPs at the edges of words.

5.2 The role of nearby repetitions

The results of the pilot study indicate that the TP computation mechanisms are robust. In particular, even when only a small fraction of bisyllable pairs comprise

the 'words' (5.3% in the pilot study), these are readily detected. Note that the frequencies of all the syllables was the same.

These results also argue against an algorithm that computes TPs by storing mono- and bisyllable frequencies. As we saw in Chapter 3 (Section 3.2.4), previous research has ruled out a frequency-based explanation for the finding that high-TP nonce words are preferred over part-words with relatively lower TPs. In particular, Aslin et al. (1998) showed that high-TP words are preferred over part-words (with lower TPs), even when the two are matched for absolute frequency. Recall that these authors used two kinds of trisyllabic nonce words. The high-frequency words occurred twice as frequently as the low-frequency words. The stream was so designed that the high-frequency *part-words* had the same absolute frequency as the low-frequency words. Thus, the high-frequency part-words and the low frequency words differed only in their TPs; the part-words had lower TPs than the words. Nevertheless, words were preferred over part-words, suggesting that frequency alone cannot account for the results; the participants must also compute the relative TPs.

However, there is another possible explanation for the results obtained by Aslin et al. (1998). The construction of artificial speech streams, for example those used by Saffran et al. (1996) and by Aslin et al. (1998) do not contain immediate repetitions of the trisyllabic nonce words. Indeed, unpublished results from our lab show that immediate repetitions 'pop-out' of the speech streams (Peña, M., pers. comm.). Thus, in similar studies, immediate repetitions have been fastidiously avoided (e.g., Peña, Bonatti, Nespor, & Mehler, 2002; Bonatti, Peña, Nespor, & Mehler, 2005).

It is nevertheless possible that repetitions at a close distance, even if they are not immediate, are processed preferentially. Imagine a situation during familiarization in the study by Aslin et al. (1998), in which a high-frequency word like 'tu.pi.ro' and a low-frequency word 'bi.da.ku' are in the configuration:

$$[\ldots \textit{tu.pi.ro.bi.da.ku.tu.pi.ro}\ldots].$$

If nearby (and not immediate) repetitions are also salient, the repeated word ('tu.pi.ro') will be extracted, leaving behind the low-frequency word:

$$[\ldots \boldsymbol{tu.pi.ro}.\textit{bi.da.ku}.\boldsymbol{tu.pi.ro}\ldots]$$

Thus, a high-frequency word would have the advantage of both high TPs and of being repeated at relatively 'close' distances. If participants are further sensitive to the fact that word boundaries cannot overlap, as they do not in natural speech (see also Principle (iii), pg. 18), then a consequence of parsing the high-frequency words will be that the high-frequency part-words are considered poor word candidates. The low-frequency words, in contrast, will never overlap with the high-frequency words.

But does the spacing between words really make a difference? That is, we know that immediate repetitions are highly beneficial for segmenting words from fluent speech. Is this benefit extended to word repetitions that are not immediate as well? We will test this in the first experiment.

5.3 Experiment 1: The effect of spacing on the computation of TPs

In this experiment we tested whether the spacing between nonce words influences their segmentation. Two groups of 'words' were created, the close-words and the far-words. Both groups of words had the same frequency in the speech stream. However, the close-words frequently recurred after 6 syllables, while the far-words only ever recurred after at least 24 syllables (see the Materials section below). Thus, if there is any processing advantage to nearby repetitions, close-words should be better recognized than far-words.

Noise syllables, as introduced in the pilot experiment, were inserted in order to manipulate the precise spacing between the nonce 'words'.

5.3.1 Materials and Methods

Participants

The participants were 25 Italian adults. In this and all subsequent experiments, the participants were undergraduate and graduate students and postgraduates, recruited from the local educational institutions. All participants were native Italian speakers, between 18 and 36 years of age, and na ive with respect to the

aims of the experiments. For all the experiments, participants were paid 3 Euros each, and they reported no auditory or language-related problems.

Materials

Two classes of 'words': close-words and far-words were defined. The far-words had an even distribution, and each far-word recurred after at least 24 other syllables (Figure 5.2). In contrast, the close-words had a clumped distribution: each close-word occurred in pairs, with 6-8 syllables separating the two tokens. Such pairs of close-words themselves recurred after 42 intervening syllables (on average), such that the overall frequency of occurrence was the same as that for the far-words.

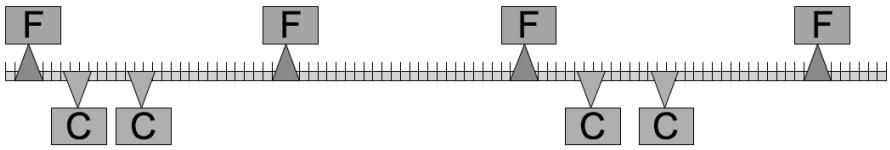


Figure 5.2: Sample timeline for the familiarization stream used in Experiment 1. The figure schematically represents 96 syllables (ticks) from the familiarization stream, with the relative placement of the trisyllabic far-words (F, upward-pointing arrowheads) and close-words(C, downward-pointing arrowheads). The close-words have a clumped distribution, but the overall frequency of the two kinds of 'words' is identical. The syllable slots not occupied by the 'words' contain random, noise syllables.

Notice that if we are to equate the frequencies of the noise syllables and the syllables that make up the 'words', the number of different noise syllables will constrain the maximum possible spacing. Thus, the choice of the close and far spacing were determined by the number of noise syllables.

Two groups of trisyllabic words and part-words were created as shown in Table 5.1[3]. In order to eliminate any phonetic cues that would distinguish the words from each other and from the non-words, all the words had a similar phonetic 'shape'; they all started with low sonority consonants in the first syllables, had

[3]Throughout the thesis, phonetic symbols are marked in the International Phonetic Alphabet (IPA), using the excellent TIPA LaTeX package by Rei Fukui.

Table 5.1: 'Words' and part-words used in the Experiment 1.

Group-1 'words'	Group-2 'words'	Non-words
/po-ru-ɮa/	/pu-ka-mu/	/t̪i-nɛ-ja/
/t̪ɛ-ni-d̪a/	/d̪o-nu-vi/	/t̪i-mɛ-fɛ/
/pa-mi-ve/	/tu-sɛ-vo/	/pi-mo-fo/
/ki-rɛ-d̪ɛ/	/d̪i-ri-fa/	/pɛ-no-jɛ/

the highest sonority consonants for the syllables in the middle and the last syllable had a middle-sonority consonant compared to the first two[4].

The noise syllables were /ku/, /no/, /pi/, /pu/, /fu/, /vu/, /mɛ/, /va/, /d̪u/, /fo/, /na/, /fi/, /mo/, /ko/, /ɮi/, /fɛ/, /pɛ/, /ɮo/, /ja/, /t̪a/, /kɛ/, /ro/, /jɛ/, /nɛ/, /su/, /ɮɛ/, /ma/, /ra/, /t̪o/, /t̪i/ and /sa/.

Two separate artificial speech streams were created. In Stream-1, the Group-1 'words' were the close-words, while the Group-2 'words' were the far-words. In Stream-2, Group-2 'words' were the close-words while the Group-1 'words' were the far-words. The streams were so designed that on average each of the random syllable occurred 50 times. All the 24 syllables that formed the words had a frequency of exactly 50 each. Care was taken to ensure that there were no bisyllables that sounded like English or Italian words. The same random list was used to construct the two streams. The only difference was the exact placement of the close- and far-words (because of the restrictions on certain bisyllables that sounded like words). The Group-1 and Group-2 words were placed in the random streams with the restriction that there were at least 6 syllables between the (pairs of) close-words, and at least 24 syllables between each repetition of a far-word. Note that in previous experiments (e.g., Saffran et al., 1996), the minimum allowed distance between two words was 3 syllables.

The syllable list thus created was converted to speech using the speech synthesis program MBROLA (Dutoit, 1997), using the Spanish male database (es1)[5], with all phonemes of the same duration of 125msec, with the F0 reaching maximum amplitude at 50% phoneme length. The resulting wav file was formatted at

[4]This 'shape' was chosen because the words so formed sounded pleasing to the ear

[5]Pilot studies indicated that Italian adults clearly perceived the phonemes in speech streams created using the es1 database, although the speech sounded 'foreign'. Thus this database has been used to construct all the speech streams in this thesis.

16kHz (16 bit stereo), and the edges of the wav were ramped using WaveWorks 1.23 (Innovative Solutions in Software, CA, USA).

For the test phase, words and non-words were similarly synthesised using MBROLA and WaveWorks (without ramping). The non-words were trisyllabic items that had never occurred during familiarization, and the syllables were chosen from the random stream (see Table 5.1).

Apparatus

The entire experiment was run, by PRESENTATION (Neurobehavioral Systems, Inc., CA, USA), which delivered all instructions and stimuli. The audio stimuli were delivered through headphones (Sony, MDR-CD280) attached to multimedia speakers (Harman/ Kardon Multimedia HK19.5) that were connected to the sound card (Sound Blaster Live! from Creative Technology Ltd.) on a computer running Windows 98^{TM}.

Procedure

The participants sat in a quiet room. There was no experimenter intervention; PRESENTATION ran the entire experiment and generated log files. Participants first heard the familiarization stream, which lasted 11 minutes. This was followed by the test phase in which subjects were presented, one at a time, pairs of words separated by 500ms of silence. All the close-words were compared all the part-words, with the close-words and the far-words occurring first or second with equal probability. The subjects had to respond in the following manner: if they thought they had heard the first word in the familiarisation phase, they were to press the z key and if they thought they had heard the second word during familiarisation, they were to press the / key. A response was coded as correct if the key-press selected a close-word or a far-word over a part-word.

5.3.2 Results

An ANOVA, with Stream (1 or 2) and Group (1 or 2) as fixed factors showed no effect of either (both $p > 0.25$). However, there was a significant interaction between the factors, $F(1, 46) = 13.8, p < 0.001$. Since there was no main effect of

either factor, the scores for close- and the far-words from both the streams were pooled. The results for the combined scores are presented in Figure 5.3. The

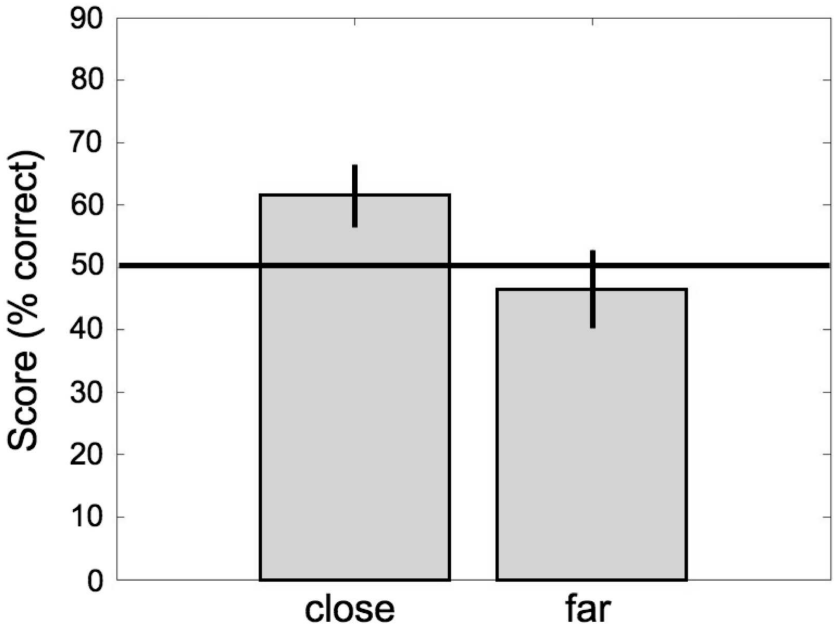


Figure 5.3: Results for Experiment 1. The means for close- and far-words are shown, collapsed across the two counter-balanced streams. Close-words are recognized, while far-words are not. Error bars represent 95% confidence limits of the mean.

combined score for close-words (61.5%, S.D. 11.8) was significantly different from chance, $t(24) = 4.9, p < 0.0001$, while the combined score for far-words (mean 46.5%, S.D. 15.2) was not, $t(24) = -1.15, p = 0.26$. The two groups differed significantly, $t(48) = 3.9, p < 0.001$.

5.3.3 Discussion

The results from this experiment suggest that TP computations are sensitive to the relative occurrence of the high-TP words. Any algorithm that merely computes TPs between syllables cannot account for these results. Another way of looking at these results is that, even if TPs *are* computed between syllables, this statistic interacts with other mechanisms that highlight possible word-like sequences in fluent speech streams, like nearby repetitions.

Nevertheless, the finding that the close-words are successfully extracted reaffirms the observation from the pilot experiment described earlier: the presence of syllabic noise (under appropriate conditions) does not hinder the extraction of statistically defined 'words' in artificial speech streams. These findings provide a rationale for the experiments to come. Recall that the central aim of this thesis is to understand how statistical information about word boundaries might interact with prosodic information about word boundaries. The results from this experiment (and the pilot study) suggest a possible way of examining the interaction between these two sources of information: prosodic phrases can be implemented as sequences of noise syllables. Then, nonce words can be placed at different locations within such 'phrases' in order to understand whether statistical 'words' in certain positions in such phrases are easier or harder to extract. The rest of the thesis will address this question in detail.

Alice laughed. 'There's no use trying,' she said: 'one *can't* believe impossible things.'

Through the looking glass,
Lewis Carroll

Chapter 6

Prosody vs. Statistics

In the previous chapter, a novel method was described for examining distributional strategies for segmenting fluent speech. In this method, high-TP 'words' are interspersed with noise syllables, which are matched in frequency with the syllables that constitute the 'words'. We saw that the presence of syllabic noise does not hinder the extraction of the 'words'.

Building on these observations, in this chapter we will establish a novel paradigm for studying an interaction between statistical computations and phrasal prosody in segmenting fluent speech. The rest of the chapters in the thesis rely on this methodology to explore how statistical computations and prosody interact.

6.1 Experiment 2: Segmenting 'words' in random frames

For this experiment, an artificial speech stream was conceived as a series of *frames*. Each frame was defined as a sequence of 10 CV (Consonant-Vowel) syllables (σ). A single frame can be represented as:

$$[\sigma 1 - \sigma 2 - \sigma 3 - \sigma 4 - \sigma 5 - \sigma 6 - \sigma 7 - \sigma 8 - \sigma 9 - \sigma 10]$$

The 10 syllabic 'slots' in each frame can be occupied by either the 'noise' syllables, or by the 'words'. The trisyllabic 'words' can be placed such that they lie within a frame, (for example at the position 4-5-6 or 5-6-7), or they can be placed such that

they straddle two frames (for example at the position 9-10-1′ or 10-1′-2′, where 1′ and 2′ represent syllabic slots from the successive frame), or they can be placed in edge positions (at the position 1-2-3). These three possibilities are depicted below; $\boldsymbol{\sigma^a\sigma^b\sigma^c}$ is a trisyllabic nonce word and the rest are noise syllables.

≻ Inside a 'phrase':

$[\sigma 1 - \sigma 2 - \sigma 3 - \underline{\boldsymbol{\sigma^a - \sigma^b - \sigma^c}} - \sigma 7 - \sigma 8 - \sigma 9 - \sigma 10]$

≻ Straddling two 'phrases':

$\ldots \sigma 5 - \sigma 6 - \sigma 7 - \sigma 8 - \underline{\boldsymbol{\sigma^a - \sigma^b}}] \; [\underline{\boldsymbol{\sigma^c}} - \sigma 2' - \sigma 3' - \sigma 4' \ldots$

≻ Aligned with the edge of a 'phrase':

$[\sigma 1 - \sigma 2 - \sigma 3 - \sigma 4 - \sigma 5 - \sigma 6 - \sigma 7 - \underline{\boldsymbol{\sigma^a - \sigma^b - \sigma^c}}]$

The reason for implementing the artificial speech stream as a series of such frames is that it allows to superimpose Intonational Phrase (IP) contours onto the frames. Upon adding prosody, each of the frames is turned into a 'phrase' (see Figure 6.4).

In the current experiment, we will place words in the middles of frames or straddling two frames. Further, we will ensure that there are no prosodic characteristics that mark the frames (or the 'words'). The results from the previous chapter suggest that, in such an absence of prosody, *all* the 'words' embedded in such a series of frames are correctly segmented. Once we obtain such a result, we can add prosody, turning the prosody-neutral frames into prosodic 'phrases' (Experiment 3).

In these experiments, we will maintain 'word'-internal TPs at 1.0, while TPs between any other pair of syllables will be kept at less than 0.3. Thus, statistically speaking, the 'words' represent coherent trisyllables in an otherwise random flux of syllables.

6.1.1 Material and Methods

Participants

Twenty adults (5 males and 15 females, mean age 24.8 years, range 19 - 40 years) participated in this experiment.

Materials

For this experiment, four trisyllabic nonce words were defined (see Table 6.1). The first two words were placed at frame-internal positions 4-5-6 or 5-6-7, and the other two were placed at frame-straddling positions 9-10-1′ or 10-1′-2′ (see Figure 6.1). This ensures that no two artificial 'words' can be adjacent; there is at least one noise syllable intervening between any two consecutive 'words'.

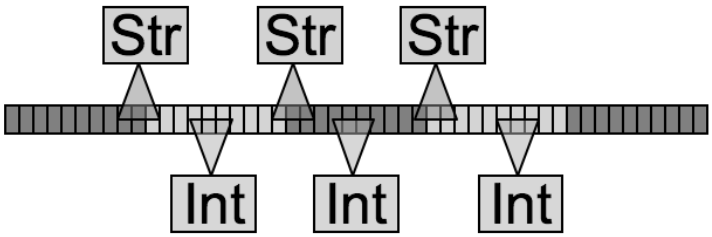


Figure 6.1: Schematic sample timeline for the familiarization stream in Experiment 2. Ticks represent individual syllables. Consecutive 'frames' are shaded dark/light. The placement of frame-straddling (Str) and frame-internal (Int) 'words' is indicated. The remaining syllabic slots are occupied by noise syllables.

There were 100 tokens of each 'word' in the familiarization stream. Each frame contained one contour-internal 'word', and between each pair of successive frames there occurred one contour-straddling 'word'. The remaining syllabic slots were occupied by one of eight different 'noise' syllables. These syllables were thus interspersed randomly between the 'words'. Care was taken to ensure that no bisyllabic sequence resembled an Italian or an English word. In addition, the average frequency of the noise syllables over the entire stream was 100. An algorithm, implemented in MATLAB (Mathworks, Inc.) generated the sequence

of frames. TPs between the syllables that formed the four 'words' were 1.0, while all other TPs were between 0.05 and 0.2, with a mean value of 0.1.

Recall from the previous chapter that 'words' which recur at short distances are preferentially accepted as compared to 'words' that recur at long distance (see Figure 5.3). In Experiment 1 (pg. 52), the 'close' distance was (at least) 6 syllables, while the 'far' distance was at least 24 syllables. Thus, in the present experiments, all the 'words' recurred at comparable distances, and the inter-'word' distance varied from 12 to 100 syllables, with a median value of 40 for all 'words' (except w2, for which the median inter-'word' distance was 38).

The resultant sequence of syllables was converted into a sequence of phonemes with a neutral prosody. Each phoneme was assigned a duration of 120 milliseconds and a constant pitch of 100 Hz. This sequence of phonemes was used to generate artificial speech using the diphone-based speech synthesizer, MBROLA (Dutoit, 1997) and the es1 (Spanish male) diphone database. The resultant was a 22.05kHz, 16-bit, mono wave file with a duration of 8 min, 2 sec. This file was converted into a stereo file, and the initial and final 5 sec were ramped up and down in amplitude, to remove onset and offset cues.

Trisyllabic sequences corresponding to the four 'words' and to four 'non-words' were separately created using MBROLA and the es1 diphone database. The non-words were trisyllabic sequences constructed by concatenating the last two syllables of one 'word' and the first syllable of another 'word' (see Table 6.1). Note that such trisyllables have been described as 'part-words' (e.g., Saffran et al., 1996), since they form part of one word and part of another. However, in previous such cases, the part-words had actually occurred in the artificial speech streams. Since in the present paradigm no two 'words' are ever immediately adjacent, all such part-words have a zero frequency, and are hence referred to as non-words. Nevertheless, note that each of these non-words do contain a (sub)sequence that was actually attested in the speech stream, and thus serve as more conservative foils than trisyllables wherein none of the sub(sequences) were ever encountered.

All the test items had phonemes of length 120 ms, and a constant pitch of 100Hz. All trisyllabic items were separately generated as 22.05kHz, 16-bit, mono wave files. These were converted to stereo files for use in the test phase.

The 'words' and non-words were pre-tested on 10 naïve participants. These

Table 6.1: 'Words' and non-words used in the experiments.

'Words'	Non-words
(w1) /pu-le-ʃa/	/le-ʃa-te/ (p1)
(w2) /ni-da-fo/	/da-fo-pu/ (p2)
(w3) /te-ki-me/	/ki-me-vo/ (p3)
(w4) /vo-ge-ʧu/	/ge-ʧu-ni/ (p4)

participants heard a fully randomized sequence of syllables for 2 minutes, followed by a test phase identical to that for this experiment (see below). All the syllables used to construct the artificial speech stream were included. The results of this pre-test are shown in Figure 6.2. As can be seen from the figure, there was no

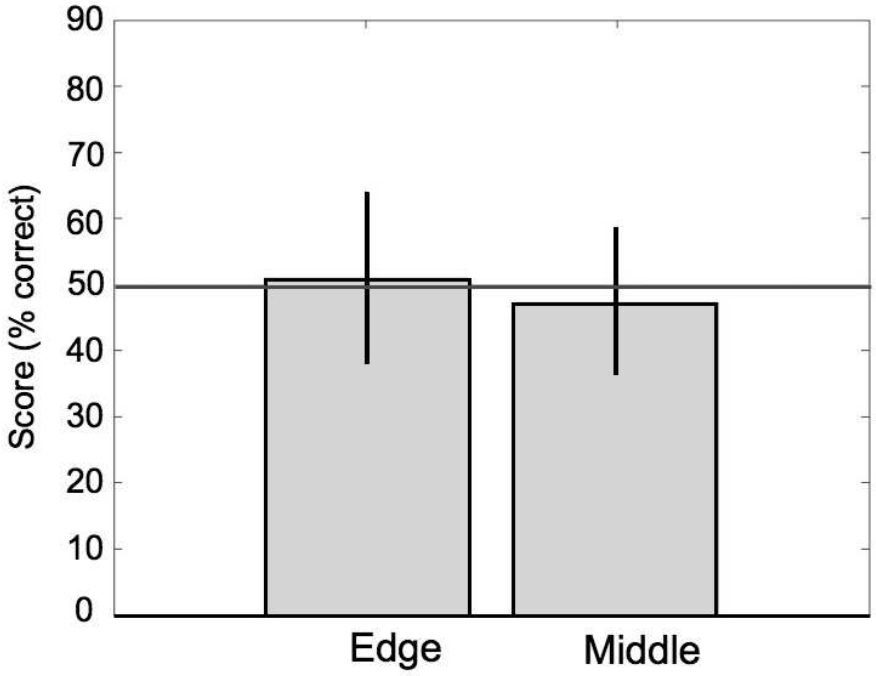


Figure 6.2: Mean scores (% correct) for 14 participants in the pre-test, separately for contour-internal 'words' and contour-straddling 'words'. Chance is 50% and indicates no preference for 'words' over non-words. The data indicates that the material presents no initial bias for the different 'words' and non-words. Error bars represent 95% confidence limits of the means.

preference for 'words' over non-words and no differences between frame-internal and frame-straddling 'words'.

Apparatus

The experiments were conducted in a sound attenuated room. The experimental design was prepared and delivered using E-Prime V1.1 (Psychological Software Tools, 2002) under the Windows 98^{TM} operating system. Sound was delivered through Sennheiser headphones attached to Harmon-Kardon speakers that themselves received input from SoundBlaster audio cards on the PCs. In the test phase, participants responded by pressing pre-marked keys on the E-Prime button box.

Procedure

Each participant was seated in front of a computer screen where instructions were displayed. In the first phase, participants were instructed to listen to a speech stream in an "invented" language and to try and pick up 'words' from this language.

At the end of the familiarization phase, participants were instructed to listen to 16 pairs of auditory test items. Each pair consisted of a 'word' (frame-internal or frame-straddling 'word') and a non-word, the 16 pairs represent the 16 combinations of all 'words' and non-words. After listening to each pair, participants had to press the left key on the button box if the first item of the pair was rated as more familiar and the right key if the second item was rated as more familiar. A response was coded as being correct if the key-press selected a frame-internal 'word' or a frame-straddling 'word' rather than a non-word. The order of 'words' and non-words was counterbalanced across trials, so that 'words' occurred equally often as the first or as the second item. All the (trisyllabic) 'words' and non-words were 720ms in length (120ms per phoneme) and had a constant pitch of 100 Hz. The two trisyllables in each trial were separated by a pause of 500ms.

6.1.2 Results

In this experiment 'words' were significantly preferred over non-words, (mean 65.6%, S.D. 13.37), $t(19) = 5.23, p < 0.001$ (all t-tests in this thesis are two-tailed). As can be seen from Figure 6.3, frame-internal 'words' were preferred over the non-words (mean 68.75%, S.D. 15.97), $t(19) = 5.25, p < 0.001$, as were the frame-straddling 'words' (mean 62.5%, S.D. 20.28), $t(19) = 2.76, p < 0.015$.

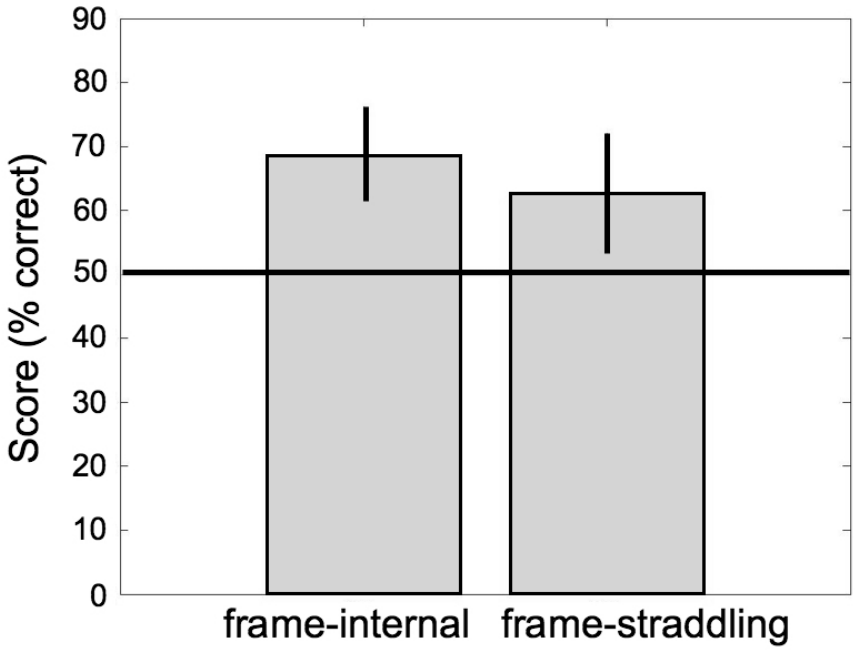


Figure 6.3: Mean scores (% correct) for 20 participants in Experiment 2, separately for frame-internal 'words' and frame-straddling 'words'. 'Words' are interspersed with noise syllables but are nevertheless segmented accurately. Chance is 50% and indicates no preference for 'words' over non-words. Error bars represent 95% confidence limits of the means.

In addition, the mean score for frame-internal 'words' was no different from the mean score for frame-straddling 'words', $F(1, 19) = 1.27, p = 0.27$.

6.1.3 Discussion

The results from this experiment demonstrate once again that the presence of syllabic noise does not hinder the extraction of statistically defined, trisyllabic 'words'. This experiment thus extends the results from the previous chapter by showing that under appropriate conditions, even when each 'word' recurs after a fairly long interval, these can be successfully segmented.

Experiment 2 paves the way to explore the interaction between prosodic and statistical cues. In the following experiments, we will examine the effect of adding prosody to the familiarization stream described in this experiment. In doing so,the

frames from Experiment 2 are converted into prosodic 'phrases' (contours).

6.2 Experiment 3: The effect of Italian prosody

The aim of this experiment is to introduce prosody to the familiarization stream described in Experiment 2, and examine the effect of such a manipulation on the segmentation of statistically well-formed 'words', that is, 'words' with high average TPs. The exact sequence of syllables as in the familiarization stream of Experiment 2 was used, so as not to alter the statistics over the syllables. In addition, for each frame, the pitch and duration characteristics were modified according to measurements from Italian IPs (described in the Methods section below). Thus, each frame was converted into a 'phrase' that is similar to an Italian IP. The test phase was identical to Experiment 2.

6.2.1 Material and Methods

Participants

Twenty adults participated in this experiment (9 males and 11 females, mean age 23.9 years, range 20 - 36 years).

Materials

A single Italian female speaker recorded nine short Italian declarative clauses, each one corresponding to a single IP[1]. These were embedded in carrier sentences (listed in Appendix B on page 161), and were between one and five words in length. The material was recorded with a Sony ECM microphone connected to a SoundBlaster sound card on a PC under Window 2000^{TM}. CoolEdit (Syntrillium Corp.) was used to record and digitally manipulate the speech waveforms. The speech segments corresponding to the IPs were digitally excised. For each IP, the pitch contour was extracted, smoothly interpolating across unvoiced segments using PRAAT (www.praat.org). A single pitch contour was converted into a vector of 400 pitch points. Thus, 20 pitch points per phoneme[2] could be used to shape

[1] I thank Silvia Pontin for these recordings

[2] MBROLA allows a maximum of 20 pitch points per phoneme.

each of the 20 phonemes (from 10 CV syllables) in a single frame. From the nine recorded IPs, nine different pitch contour vectors were thus obtained, of which eight were used for this experiment.

Next, the durations of the first and last syllables of each IP were measured. The durations were divided by the number of segments in the syllables, to get a normalized value. It was found that the average normalized duration of the phonemes of the last syllable (99.6 ms) was significantly different from the average normalized duration of the phonemes of the first syllable (79.9 ms), paired t-test, $t(8)=2.8$, $p=0.02$. Since in Experiment 2 phoneme durations of 120 ms were used, the phonemes of the initial syllable of each frame were shortened by 20 ms to a final value of 100 ms each. The phonemes of the final syllable in each frame were lengthened by 20 ms to 140 ms each. All the other phonemes in the frame were 120 ms in length. Thus, on average, all the phonemes in a frame had a length of 120 ms, as in Experiment 2.

The model of prosody elaborated thus consisted of eight pitch contours, randomly associated with frames of 10 syllabic slots that went from an initial syllable of 200 ms followed by 8 syllables of 240 ms and a final syllable of 280 ms. Figure 6.4 shows a schematic outline of the model of prosody that was implemented. The

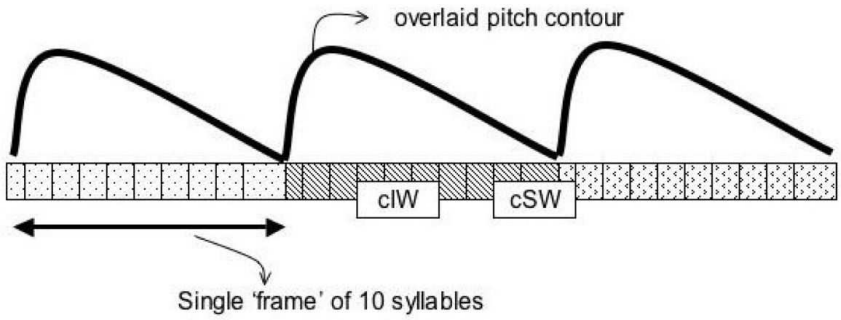


Figure 6.4: Schematic outline of the structure of the familiarization stream for Experiment 2. A series of three frames, each containing 10 syllable slots is shown. Duration and pitch characteristics of the phonemes are the suprasegmentals that define the overlaid prosodic contour. Possible positions of one contour-internal 'word' (cIW) and one contour-straddling 'word (cSW) are shown.

sequence of phonemes from Experiment 2 with the added prosodic characteristics was used to generate an artificial speech stream using MBROLA and the es1 diphone database as in Experiment 2. The 22.05kHz, 16-bit, mono wave file was converted to stereo and the initial and final 5 sec of the file were ramped up and down in amplitude.

In Experiment 2, 'words' and non-words in both the familiarization and the test phase all had phonemes of the same duration, and all had a constant pitch. Since in this experiment we have the same test phase, but have a prosodic familiarization, the 'words' heard during test are acoustically different from those heard during familiarization. In Chapter 9 we will look at several control experiments that establish that these differences do not contribute to the results we observe.

Apparatus and Procedure

These were identical to Experiment 2.

6.2.2 Results

The overall score, indicating correct segmentation of the speech stream was 56.56% (S.D. 13.37), compared to the score of 65.6% in Experiment 2, but it was still statistically significant, $t(19) = 2.195, p = 0.04$. However, from Figure 6.5, it can be seen that there appears to be a difference in the segmentation of contour-internal and contour-straddling 'words'.

An ANOVA with 'word' type (*Internal* or *Straddling*) as a (fixed) factor revealed a main effect of word type, $F(1, 19) = 12.93, p < 0.005$. A post-hoc Scheffe test revealed a significant difference between *Internal* and *Straddling* words, $p < 0.005$.

T-tests showed that the mean score of 68.13% (S.D. 22.39) for the contour-internal 'words' was significantly different from chance, $t(19) = 3.62, p < 0.002$, while the mean score of 45% (S.D. 16.42) for the contour-straddling 'words' was not, $t(19) = -1.36, p = 0.19$.

In order to compare the results from Experiment 3 with Experiment 2, a second ANOVA was run with word type (*Internal* or *Straddling*) as a within-subject factor and experimental condition ('flat' familiarization or prosodic fa-

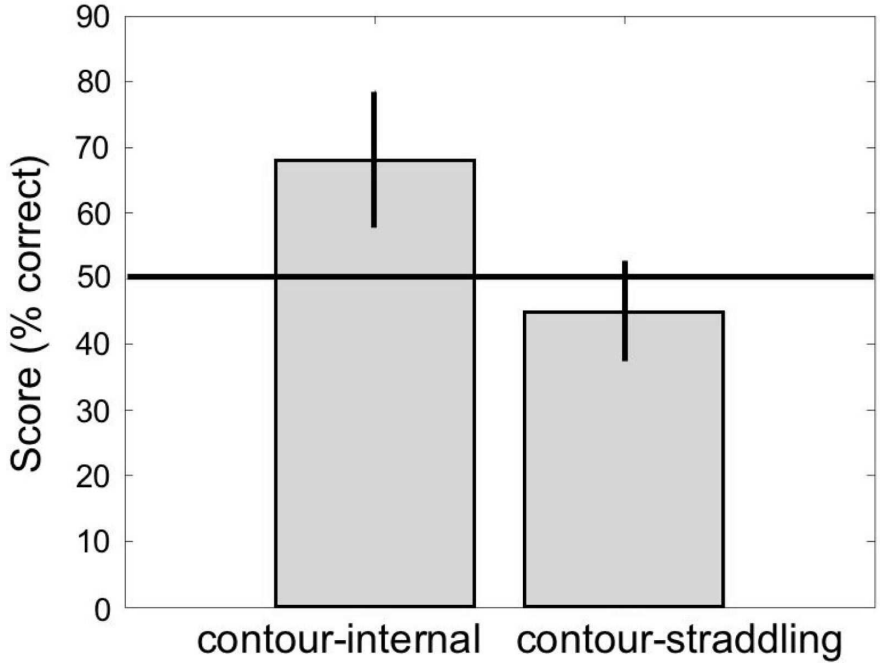


Figure 6.5: The mean scores (% correct) from Experiment 3. In the presence of IP prosody, (IP-)contour-internal 'words' appear to be correctly segmented. Error bars represent 95% confidence limits of the means.

miliarization) as a between-subject factor. There was a significant main effect of word type, $F(1, 38) = 11.95, p = 0.001$, as well as a significant word type X experimental condition interaction, $F(1, 38) = 3.95, p = 0.05$.

6.2.3 Discussion

Taken together, Experiments 2 and 3 demonstrate an effect of prosody on the segmentation of statistically defined 'words' in fluent speech. Experiment 2 established that in a monotonous speech stream, all statistically defined 'words' are correctly segmented. Experiment 3 demonstrated that when prosody is superimposed on the flat speech stream, only those 'words' that lie internal to prosodic 'phrases' appear to be segmented. That is, when the two cues are in conflict, prosodic cues appear to take precedence, such that prosodically "bad" syllabic sequences are rejected.

This result is in agreement with models that suggest that prosodic constituents

help segment speech (Nespor et al., 1996; Christophe, Nespor, Guasti, & van, 1997; Guasti, 2002). A corollary of such a view is that sequences that span prosodic constituents would be harder to detect. Our results thus suggest that prosody organizes the speech stream into'phrases', making contour-spanning 'words' harder to detect.

'All right,' said the Cat; and this time it vanished quite slowly, beginning with the end of the tail, and ending with the grin, which remained some time after the rest of it had gone.

Alice in Wonderland,
Lewis Carroll

Chapter 7

An *Edge Effect* in segmenting artificial, prosodic speech

The experiments reported in the previous chapter suggest that prosody serves to segment the speech stream. That is, instead of a continuous, unbroken sequence of syllables, listeners appear to perceive fluent speech as being divided into a series of phrases. In this chapter, we will examine evidence from a difference source that suggests that participants indeed perceive artificial, prosodic speech as a series of 'phrases'.

7.1 Edge phenomena

It is known from studies on human memory that learning an arbitrary sequence of verbal items is facilitated when the sequence can be perceptually chunked into subsequences (e.g., Hitch, Burgess, Towse, & Culpin, 1996; Burgess & Hitch, 1999). Moreover, edges of sequences are better recalled than their middles (e.g., Ebbinghaus, 1964; Miller, 1956), resulting in ∪-shaped recall curves (Baddeley, 1990; Brown, Preece, & Hulme, 2000). The edges are thought to be salient positions; the leading edge benefits from a primacy effect, while the trailing edge from a recency effect. Further, the edges are the only positions that are not *masked* on either side, that is, they are not flanked on both sides by other material.

In perceptually chunked verbal lists, such ∪-shaped recall is observed *even for*

each of the subsequences (Hitch et al., 1996, Henson, 1998, Burgess and Hitch, 1999; see Ng and Maybery, 2002 for a recent review).

We might perceive the artificial speech stream from Experiment 3 as a sequence of syllables containing subsequences: the prosodic 'phrases'. Thus, if prosody can divide a familiarization stream into phrases, we might expect that the edges of such phrases are more salient than their middles. In terms of the material from Experiment 3, this means that words at the edges of the contours ought to be better recalled than words in their middles. In other words, finding an advantage for the recall of trisyllabic 'words' at edges over trisyllabic 'words' in the middles would constitute further empirical evidence in favor of a model wherein prosody serves to segment the input.

7.2 Experiment 4: Empirical evidence for an Edge Effect

This experiment is aimed at establishing if there is an advantage for 'words' at the edges of prosodic contours over 'words' in the middle. The preparation of the speech stream was modified in several ways. While in Experiment 3 'words' occurred either contour-internally or straddling contours, in the current experiment all the 'words' occurred at contour-internal positions. However, two words were chosen to be placed at the edges of the prosodic contours, and two others in their middles.

We know from Experiment 3 that 'words' in the middle are correctly segmented. Also, the scores for the contour-internal 'words' were similar in the presence of prosody (68.13%) and in its absence (68.75%). Thus, the amount of familiarization was halved, providing 50 tokens of each 'word' instead of 100. This should make the task of segmentation more difficult, enhancing differences, if any, between 'words' at edges and 'words' in the middles.

The effect of the left and the right edges were tested separately. Thus, two separate groups of participants were exposed to streams with 'words' in the middles and at the edges; for one group the edge-'words' were at the left edge of 'phrases', while for the other group they were at the right edges. This was to ensure that

the streams were as similar as possible to that in Experiment 3 (pg. 66).

7.2.1 Material and Methods

Participants

Twenty-six Italian adults participated in this experiment. Fourteen (2 males and 12 females, mean age 23.9 years, range 18-30 years) were exposed to the stream with edge-'words' at the left edge. Twelve (6 males and 6 females, mean age 23.3 years, range 19-32 years) were exposed to the stream with edge-'words' at the right edge.

Materials

Two new sequences of frames were created. In each frame, one of the two contour-internal 'words' from Experiment 3 was placed at positions 1-2-3 (left edge stream) or at positions 8-9-10 (right edge stream), and designated edge-'words'. The two contour-straddling 'words' from Experiment 3 were placed at positions 6-7-8 (left edge stream) or at positions 3-4-5 (right edge stream) inside each frame and were designated the middle-'words'. Schematically a single frame from the two streams can be depicted as follows (edge 'words': underbraces; middle 'words': overbraces):

- Left edge stream:
  $[\underbrace{\sigma 1_{ew} - \sigma 2_{ew} - \sigma 3_{ew}} -\sigma 4 - \sigma 5 - \overbrace{\sigma 6_{mw} - \sigma 7_{mw} - \sigma 8_{mw}} -\sigma 9 - \sigma 10]$

- Right edge stream:
  $[\sigma 1 - \sigma 2 - \overbrace{\sigma 3_{mw} - \sigma 4_{mw} - \sigma 5_{mw}} -\sigma 6 - \sigma 7 - \underbrace{\sigma 8_{ew} - \sigma 9_{ew} - \sigma 10_{ew}}]$

The edge-'words' and the middle-'words' occurred 50 times each during the entire stream. The remaining slots in all frames were filled with noise syllables with an average frequency of 50 across the entire stream. Each frame was randomly assigned one of eight prosodic contours from Experiment 3.

The two sequences of phonemes were fed to MBROLA, using the es1 (Spanish male) diphone database. The final output files were 22.05kHz, 16-bit, mono wave files of length 4 min. These file was converted into stereo files and the initial and

final 5 sec were ramped up and down to eliminate onset or offset cues to edge-'words' and middle-'words'. The test phase was identical to Experiments 2 and 3. Notice that in this experiment too, the non-words have zero frequency during familiarization.

Apparatus and Procedure

These were identical to Experiment 2.

7.2.2 Results

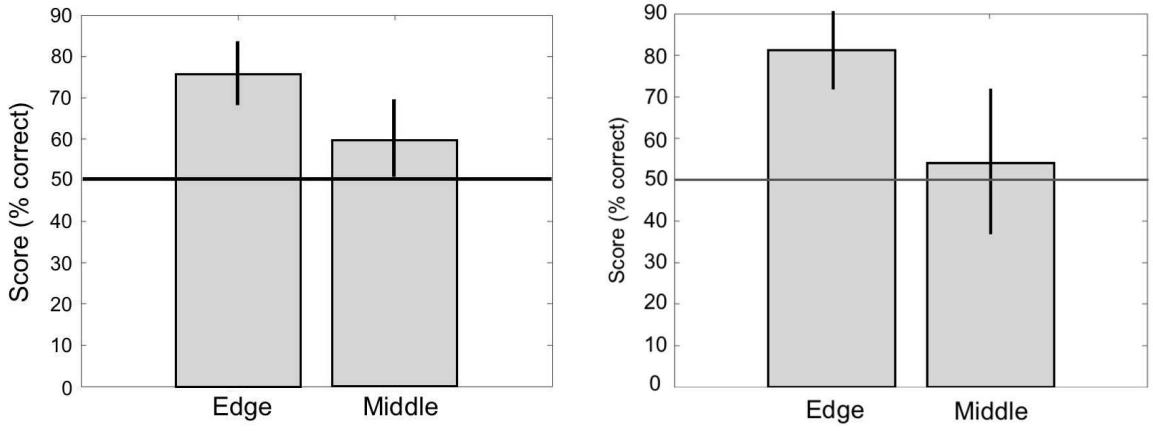
The overall scores for both streams were better than chance (left edge: mean 67.86%, S.D. 9.45, $t(13) = 7.07, p < 0.0001$; right edge: mean 66.25%, S.D. 19, $t(11) = 3.23, p < 0.01$). Table 7.1 and Figure 7.1, summarize the scores for edge-'words' and middle-'words' in the two streams.

Table 7.1: Scores for edge- and middle-'words' for the left and the right edge streams. ([a]Significantly different from chance, $p < 0.001$, [b]Significantly different from chance, $p = 0.05$).

'Word' type	Left edge stream	Right edge stream	Difference
Edge-'word'	75.9%[a]	81.3%[a]	n.s.
Middle-'word'	59.8%[b]	54.17%	n.s.
Edge vs Middle	$p < 0.03$	$p < 0.005$	

The edge-'words' at the left edge (mean 75.9%, S.D. 13.4) were recognized significantly above chance, $t(13) = 7.23, p < 0.0001$. Similarly, edge-'words' at the right edge (mean 81.3%, S.D. 15.54) were recognized significantly above chance, $t(11) = 6.97, p < 0.0001$. The middle-'words' in the two conditions were less well recognized, left edge: 59.8%, S.D. 17.11, $t(13) = 2.15, p = 0.05$, right edge: 54.17%, S.D. 27.87, $t(11) = 0.52, p = 0.62$.

Pooling the data in an ANOVA with factors Edge (left or right) and Position (edge-'word' or middle-'word') revealed a significant effect of Position, $F(1, 24) = 20.42, p < 0.001$. The Edge condition was not significant ($p > 0.9$), and neither was the interaction ($p > 0.2$). Post-hoc (Scheffe) tests revealed that the edge-'words' were recognized better than the middle-'words' in both groups (left edge, $p = 0.02$, right edge, $p < 0.001$).



(a) Left edge

(b) Right edge

Figure 7.1: Mean scores (% correct) for edge-'words' (EWs) and middle-'words' (MWs) from Experiment 4 in the left- and right-edge streams. In both streams, edge-'words' are efficiently segmented, while middle-'words' are segmented with much less efficiency. Error bars represent 95% confidence limits of the means

7.2.3 Discussion

The finding that edge-'words' are better recognized than middle-'words' provides further evidence that prosody serves to chunk the speech stream. These results are compatible with the aforementioned experiments in human memory, wherein chunking an arbitrary list of verbal items results in a 'multiply-bowed recall curve' (Ng & Maybery, 2002), that is, in ∪-shaped curves within each of the chunks. By analogy, 'words' placed at the edges of prosodic contours are better recalled than 'words' placed in their middles. This would be true *only* if prosody served to divide the fluent speech into a series of 'phrases'. Indeed, recent evidence from our lab suggests that the edges of a verbal list of items might be salient (Endress, Scholl, & Mehler, 2005). These authors found that the extraction and generalization of repetition-based structures was optimal when the repetitions occurred at edges as opposed to the middles of arbitrary seven-item syllabic sequences.

'If they would only purr for "yes" and mew for "no," or any rule of that sort,' she had said, 'so that one could keep up a conversation!'

Through the looking glass,
Lewis Carroll

Chapter 8

Possible models for an interaction between prosody and statistics

The previous two chapters have established that prosody serves to divide speech into phrases. Statistically coherent (high-TP) syllabic sequences that straddle such phrases are not preferred over trisyllables that never actually occurred in the speech stream (non-words). In contrast, phrase-internal high-TP syllables are significantly preferred over non-words.

Such a result is warranted if, as seen in the introductory chapters, words are aligned with larger prosodic constituents. That is, a word with a high TP between the syllables, but that straddles two prosodic phrases faces a conflict of cues: TPs indicate cohesion but prosody introduces a boundary. We saw in Chapter 6 that participants do not judge such items as possible words. However, cohesive syllabic sequences uninterrupted by a prosodic boundary, are readily preferred over non-words.

We now ask: at what level does prosody intercede? There are in theory at least two possibilities:

(a) higher level prosody might directly segment the syllabic representation, such that TPs are *computed within* prosodically defined syllabic chunks.

(b) prosody might act to *filter* the output of the TP system.

The two possibilities are shown schematically in Figure 8.1. Both possibilities

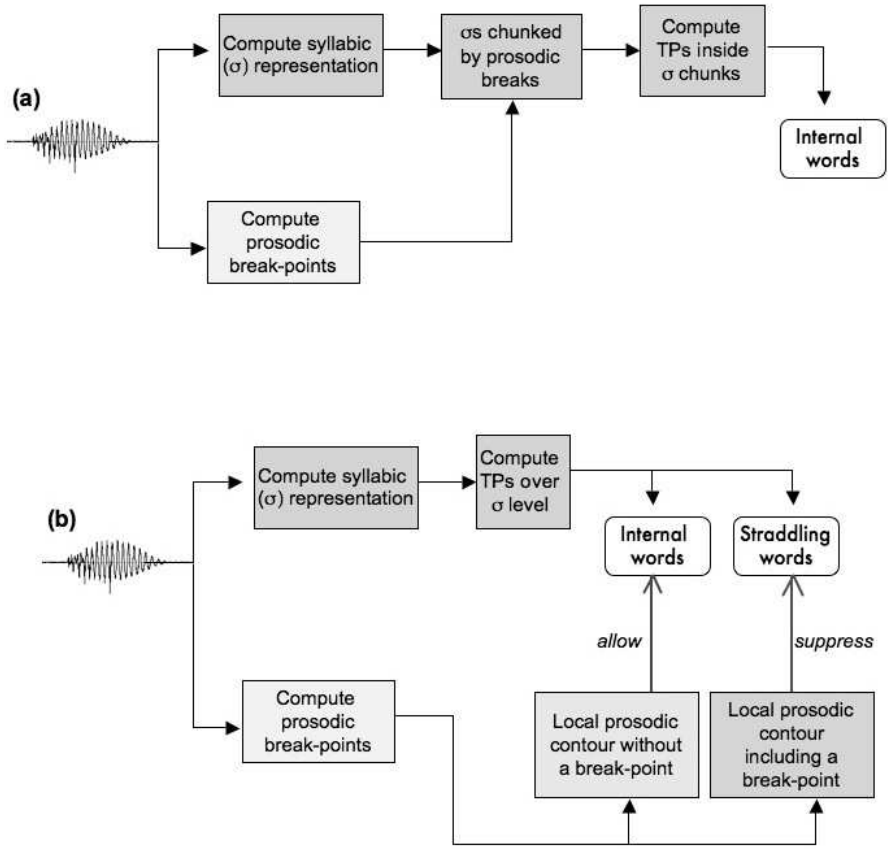


Figure 8.1: Two possibilities for an interaction between prosody and statistics. In (a), the speech stream is broken up by prosody into chunks (domains), and TPs are calculated only inside such domains. In (b), statistical analyses extract all possible 'words' from a syllabic representation. Prosody acts to suppress statistical 'words' that span prosodic boundaries.

outlined in Figure 8.1 make the same predictions for the experiments described so far. But, there is an underlying difference between the two with regards to Experiment 3. In Experiment 3, the familiarization stream contained both contour-internal and contour-straddling 'words', and we saw that only the contour-internal 'words' were correctly recognized in the test phase (see Figure 6.5 on page 69).

The two proposals make differing predictions for the failure of the contour-straddling 'words' to be recognized. According to proposal (a), the contour-straddling 'words' are not segmented *at all*. That is, since prosody carves the

input into chunks and TPs are restricted to such chunks, straddling 'words' do not enter TP computations. According to proposal (b) in contrast, both internal and straddling 'words' are correctly segmented, but straddling 'words' are suppressed due to prosody.

How can we distinguish between these possibilities? One way of addressing this question is to ask: are there conditions under which we find evidence that TPs are computed for *all* bisyllables and not only to phrase-internal ones?

In Chapter 5, we examined various hypotheses about how TPs between syllables are computed. The syllables themselves were assumed to be the basic units over which statistics are computed. Indeed, we saw in the introductory chapters that the syllable is regarded as a fundamental unit of speech (e.g., Bertoncini & Mehler, 1981a; Mehler, Segui, & Frauenfelder, 1981; Mehler, Dupoux, Nazzi, & Dehaene-Lambertz, 1996). However, as discussed in Chapter 2 (Section 2.2.1, pg. 14) the syllable itself is an abstract (prosodic) constituent, made up of the phonemic segments (see also Nespor & Vogel, 1986; Blevins, 1995).

Let us therefore assume two levels of representation. The first is the segmental level, wherein the segments are grouped together into the syllables. The second is the suprasegmental level, wherein properties associated to the segments like their pitch and their duration are represented. Such a distinction is in line with the autosegmental theories in phonology, wherein the segmental level is considered distinct from the suprasegmental level (e.g., Goldsmith, 1990).

Since we assume that TPs are computed over syllables, we predict that at the abstract level of the syllables, *all* high-TP sequences are equivalent. Thus, if, following the prosodic familiarization in Experiment 3 one could find a way to tap only the abstract, syllabic level of representation, then it might be possible to demonstrate recall of not only the internal 'words', but also the straddling 'words'. This hypothesis is tested in the next experiment.

8.1 Experiment 5: Distinguishing possible models for an interaction

How can prosodic effects be bypassed? In particular, how can we tap only into the abstract, syllabic level? One possibility is to follow prosodic familiarization by a *visual* test phase.

Reading is thought to involve, in part, a transformation from an orthographic into an abstract code, specially for psuedowords (e.g., see Lukatela & Turvey, 1994; Price, Wise, & Frackowiak, 1996; Frost, 1998; Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001). Behavioral studies have revealed that perceiving written words automatically activates representations of their spoken forms even when they are perceived subconsciously. Moreover, the written form appears to be converted to a phonological form prior to lexical access (e.g., Frost, 2003). Indeed, imaging studies have demonstrated that exposure to visually presented text automatically activates the left perisylvian area, presumably via a left inferior parietal, orthography-to-phonology transformation system (e.g., Price et al., 1996; Price, 1998).

More recently, Nakamura et al. (2006) found cross-modal repetition priming: a subliminal, visually presented word caused the priming of an auditory target, suggesting that the visually presented word activates an amodal representation. Indeed, Dehaene and Naccache (2006) speculate that the use of a highly regular script like the Japanese kana syllabary contributes to the ease of subliminal priming in the aforementioned study due to the automaticity of spelling-to-sound conversion afforded by the orthographic transparency of such scripts.

Speakers of Italian, which has a transparent orthography (e.g., Lepschy & Lepschy, 1981), would thus presumably read the pseudowords in a format that (a) is similar for all the 'words' and (b) possibly similar to the output of the TP computation system, before the intervention of prosody. At this stage, possibility (b) is offered as a hypothesis, in need of further empirical support. This manipulation would thus provide us the opportunity of evaluating if indeed straddling 'words' are extracted by the TP computation system before prosody intervenes.

There are different possible outcomes to this experiment. If TPs are computed over an abstract, syllabic representation, we would expect that both contour-

internal and contour-straddling 'words' are correctly segmented, as in Experiment 2 (familiarization without prosody). If, however, TP computations are limited to prosodically defined subsequences, then we expect that only the contour-internal 'words' are recognized, as in Experiment 3 (prosodic familiarization).

8.1.1 Material and Methods

Participants

Fourteen adults participated in this experiment (4 males and 10 females, mean age 24.6 years, range 20-32 years).

Materials

The familiarization phase used the same artificial speech file as that of Experiment 3. In the test phase, instead of the two trisyllabic sequences presented aurally in each trial, the same items were presented visually on the screen. The first word was displayed to the left and the second to the right of the screen centre. The same instructions as for the previous experiments were used.

Apparatus and Procedure

These were identical to Experiments 2 and 3.

8.1.2 Results

The overall score, indicating correct segmentation of the speech stream was 66.96% (S.D. 14.80), and was significantly different from chance, $t(13) = 4.29, p < 0.001$. Figure 8.2 shows the results of Experiment 5 separately for contour-internal and contour-straddling 'words'. Figure 8.2 illustrates that both the contour-internal as well as the contour-straddling 'words' were recognized at better than chance levels. An ANOVA with word type (contour-internal or contour-straddling) as a within-subjects factor indicated no differences between the two word types, $F(1, 13) = 0, p = 1$. Contour-internal 'words' had a mean score of 66.96% (S.D. 21.77), $t(13) = 2.92, p = 0.01$, while contour-straddling 'words' had a mean score of 66.96% (S.D. 20.57), $t(13) = 3.09, p < 0.01$.

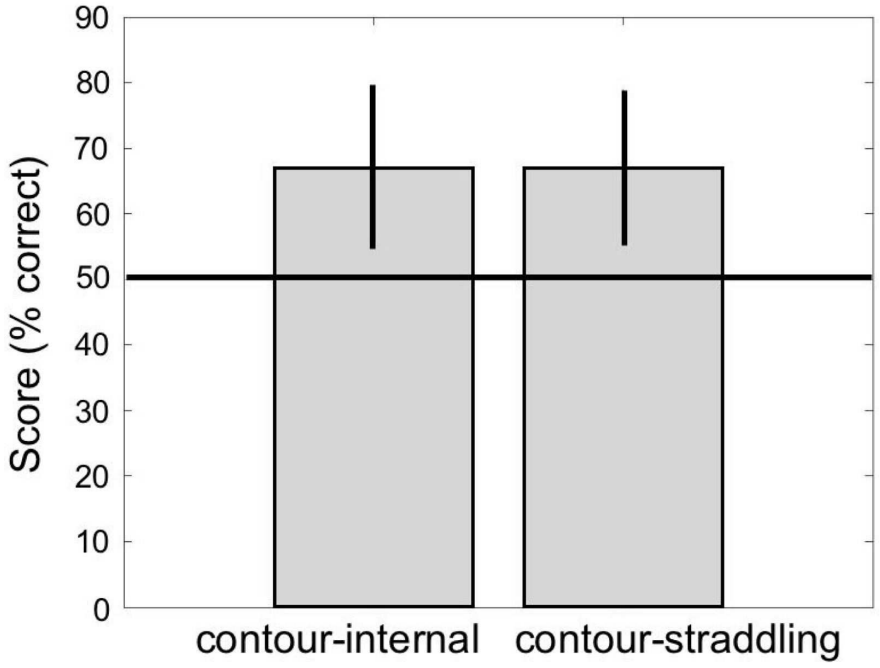


Figure 8.2: Mean scores (% correct) and standard errors for visually presented contour-internal and contour-straddling 'words'. Error bars represent 95% confidence limits of the means.

The pattern of results from this experiment was compared to that from Experiment 3 in an ANOVA with *Position* of the 'words' (contour-internal or contour-straddling) as a within-subjects factor and test *Modality* (auditory or visual) as a between-subjects factor. The results showed a main effect of *Position*, $F(1, 32) = 5.1, p < 0.03$, as well as a significant interaction between *Position* and *Modality*, $F(1, 32) = 5.1, p < 0.03$, suggesting that contour-internal 'words' are recognized better than contour-straddling 'words', and that this is due to contour-straddling 'words' being at chance in Experiment 3 and above chance in this experiment.

8.1.3 Discussion

The results from this experiment suggest that *all* statistically well-formed 'words' are extracted during prosodic familiarization. This result is difficult to reconcile with the possibility suggested in 8.1(a) (pg. 78). We had assumed that a visual

test phase might tap preferentially into the abstract phonological representation that participants build during familiarization. The finding that the straddling 'words' are recognized as well as the internal 'words' suggests that, in contrast to the model in 8.1(a), straddling 'words' *do* enter into TP computations, since they are segmented and recognized.

The results thus favor the model depicted schematically in 8.1(b): TPs are computed over a syllabic representation of the speech stream, and prosody filters the output of the TP computational system. This view is also coherent with the observation (Saffran, Newport, Aslin, Tunick, & Barrueco, 1997) that TP computations appear to be implicit and automatic. Thus, TP computations over syllabic representations of speech might be an encapsulated, automatic system that is itself unaffected by other properties of the speech stream.

However, it is possible that the results from this experiment are not due to a lack of the prosodic filtering effect, but due the nature of the test phase itself. That is, a visual test phase might give the observed results independent of the familiarization stream. In order to ensure that the visual test phase relies on information gathered during the familiarization phase, a control experiment was run, in which the two kinds of 'words' are expected to be recognized differently[1].

8.2 Experiment 6: Control: The visual test phase is not insensitive to prosodic familiarization

8.2.1 Material and Methods

Participants

Fourteen adults participated in this experiment (3 males and 11 females, mean age 27.8 years, range 21-36 years).

Materials

The familiarization phase used the artificial speech file from Experiment 4, wherein 'words' were aligned with the right edges of 'phrases'. In the test phase, instead of

[1] I thank Chuck Clifton for this suggestion.

the two trisyllabic sequences presented aurally in each trial, the same items were presented visually on the screen. The first word was displayed to the left and the second to the right of the screen centre. The same instructions as for the previous experiments were used.

Apparatus and Procedure

These were identical to Experiment 4.

8.2.2 Results

The overall score, indicating correct segmentation of the speech stream was 69.6% (S.D. 16.6), and was significantly different from chance, $t(13) = 4.43, p < 0.001$. Figure 8.3 shows the scores separately for the edge-words and the middle-words.

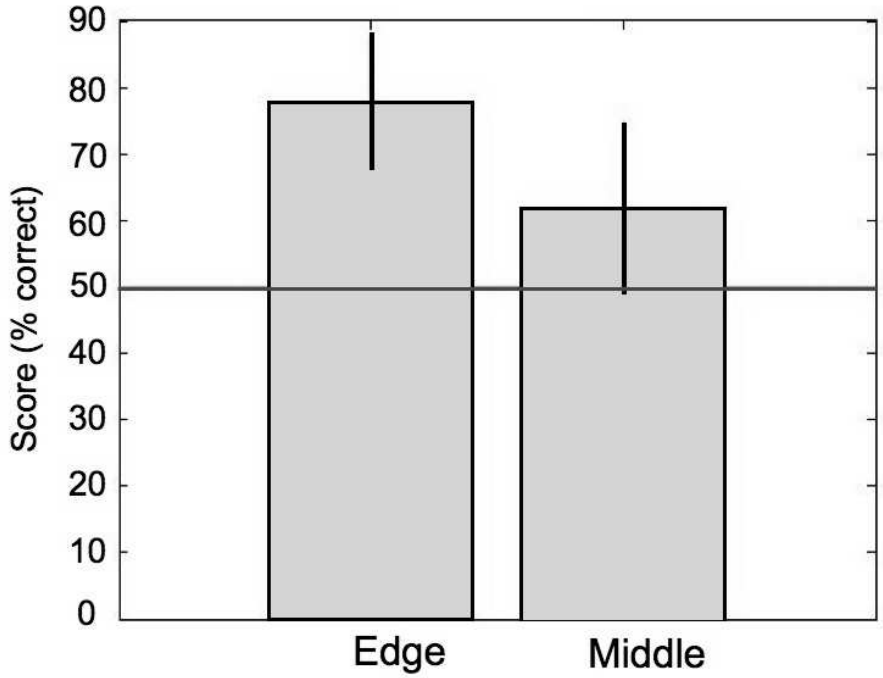


Figure 8.3: Mean scores (% correct) and standard errors for visually presented (right) 'edge-words' and 'middle-straddling 'words' following prosodic familiarization. Error bars represent 95% confidence limits of the means.

The edge-words were recognized significantly better than chance, 77.68% (S.D. 17.8), $t(13) = 5.82, p < 0.0001$, while middle-words were not, 61.6% (S.D. 22.7), $t(13) = 1.91, p = 0.078$. In addition, the edge-words were significantly better recognized than middle-words, $t(26) = 2.08, p = 0.047, d = 0.79$.

An ANOVA compared the results from this experiment with those obtained with the right-edge stream in Experiment 4. The factors were *Position* (edge-words or middle-words), and *Test Type* (auditory or visual). There was a main effect of *Position*, $F(1, 48) = 13.11, p < 0.001$; edge-words were better recognized than middle-words. The factor *Test Type* was not significant, and neither was the interaction between *Position* and *Test Type*.

8.2.3 Discussion

The results from this control experiment indicate that the visual test is indeed sensitive to some aspects of the prosodic familiarization. The previous experiment showed that the rejection of contour-straddling 'words' is not obtained with a visual test. In contrast, in this control experiment, the visual test phase replicates the edge effect, wherein 'words' at the edges of IPs are better recognized than 'words' in the middles. The comparison between this experiment and Experiment 4, where an auditory test phase was used, indicates that there is no difference in the pattern of results for the edge effect when the test phase is in different modalities.

These results are compatible with the hypothesis that the visual test phase preferentially taps into an abstract representation. Indeed, if syllables at the edges of IPs are in salient positions, we would expect that they are better processed, and hence better recalled, even in the visual modality.

We now need to explain *how* prosody can affect the output of the TP computations. The model proposed in Figure 8.1(b) suggests that the input is analyzed along two parallel pathways, one that computes TPs over the syllables, and the other that detects the edges of constituents. What mechanism can bring together the different elements of an encoded stimulus? One possibility is episodic memory.

The role of memory

Traditionally, episodic memory is a form of long-term memory, which stores individuated 'snap-shots' of previous experiences. One of the features of episodic memory is that it assumes a multimodal code, and shows encoding specificity (e.g., Tulving, 2002). More recently, such a multimodal store, the *episodic buffer* has also been proposed for short-term (working) memory (e.g., Baddeley, 2000, 2003).

The existence of multi-modal stores is necessary to explain the effects of context during encoding (see Bouton, Nelson, & Rosas, 1999, for a review). For example, Godden and Baddeley (1975) found that the extrinsic environment during the encoding of a list of words (on land or underwater in their experiment) had an effect on recall such that it was most effective when encoding and recall environments were the same.

We can propose a similar account for the effect of modality in the test phase, on the filtering effect of prosody. Recall from Experiments 3 (pg. 66) and 5 (pg. 80) that, while contour-straddling 'words' are not recognized when the test phase is in the auditory modality, they are recognized as well as the contour-internal 'words' with the visual test phase.

The acoustic modality of the test items provides an appropriate context for the recall of their acoustic characteristics during familiarization. In the next chapter, we will look at evidence that the precise acoustic shape of the 'words' during familiarization and during test does not contribute to the recall of the test items. Instead, we hypothesize that the presence or absence of an acoustic/prosodic break is recalled. 'Words' misaligned with such breaks are rejected as possible 'word' candidates.

The visual modality, in contrast, does not provide an appropriate context for the recall of acoustic characteristics. Instead, the phonological representation of the test items predominates. If, as we proposed, distributional analyses are carried out over such a phonological level, we expect that all high-TP syllable sequences are recalled, which is what we find in Experiment 5.

In sum, while distributional analyses might find several high-TP multisyllabic sequences, only those that are in prosodically appropriate contexts are considered

as possible lexical items.

An interactive lexicon

Notice that we assume that the lexical items are stored in an abstract, syllabic representation. Indeed, several researchers have proposed *phonological theories* of the lexicon, wherein lexical items are based on underlying abstract forms (e.g., Klatt, 1979; Lahiri & Marslen-Wilson, 1991; Pallier, Colome, & Sebastian-Galles, 2001; Eulitz & Lahiri, 2004). In such models, the lexical entry consists of a phonological (abstract) form, for example, as a sequence of syllables like we assume. Incoming speech is progressively stripped of incidental acoustic features like the timbre or intensity to arrive at an abstract form that corresponds to the stored phonological representations in the lexicon.

However, such theories do not explain the fact that we can retain acoustic details of word tokens, and that these details influence perception. For example, speaker recognition is possible when voices are played backwards, compressed or even converted to *sine-wave speech* (Van Lancker, Kreiman, & Emmorey, 1985b, 1985a; Remez, Fellowes, & Rubin, 1997). Several studies have found that the implicit memory for spoken words retains detailed acoustic information like vocal characteristics and intonation contours (e.g., Schacter and Church; Church and Schacter, 1992; 1994; see also Palmeri, Goldinger, and Pisoni; Goldinger, 1993; 1996).

Thus, since episodic traces of words persist in memory and affect subsequent processing, several authors have proposed that such traces might be all that constitute the mental lexicon. That is, in such *episodic theories*, lexical items are generalizations over stored episodes (e.g., Jacoby & Brooks, 1984; Goldinger, 1998; Pierrehumbert, 2003, amongst many others).

The prosodic filtering model proposed above in Figure 8.1(b) on page 78 suggests a possibility to reconcile the proposed phonological nature of the lexicon from the observation that episodic traces of words affect processing.

The prosodic filtering model proposes the separation of an abstract representation of the speech stream and the computation of acoustic/phonetic characteristics that mark phrasal boundaries. TPs are computed over an abstract level, and the output of such computations are (distributionally) coherent, high-TP syl-

lable sequences. Such sequences are (mis)aligned with prosodic edges in episodic memory.

Thus, an implication of the prosodic filtering model is that we can explain episodic effects found in word recognition, while maintaining that the lexicon is phonologically specified. That is, lexical items are proposed to be in an abstract (phonological) form, but are linked to incidental acoustic properties via episodic memory. Recall (above) the experiments in Godden and Baddeley (1975), wherein the context (underwater or on land) affected the recall of words. We propose that these results reflect the same underlying processes as those that show the recall of, for example, the voice characteristics (e.g., Palmeri et al., 1993). In both cases, episodic memory links the lexical items to their respective 'episodes'; be they the surrounding environment or the acoustic cues that distinguish one voice from another. In the experiments reported in this thesis, the same mechanism links 'words' to the presence or absence of prosodic edges.

The snoring got more distinct
every minute, and sounded more
like a tune: at last she could
even make out the words...

Through the looking glass,
Lewis Carroll

Chapter 9

Controlling for acoustic similarity

We have seen that, with an auditory test phase, a 'word' that straddles a 'phrasal' boundary is not preferred over a trisyllabic sequence that never occurred. In contrast, a 'word' internal to a 'phrase' is preferred over a non-word. As noted earlier (Section 6.2.1, pg. 66), in the experiments involving a prosodic familiarization, there is an acoustic difference between the 'words' during familiarization and during test. While the 'words' during familiarization have changes in pitch and duration, the 'words' in the test phrase are synthesized with a neutral prosody.

An alternate explanation of the results is that the acoustic difference between a middle 'word' during familiarization and during test is small, relative to such an acoustic difference for a contour-straddling 'word. That is, if the neutral-prosody test items are more similar to their intonated, middle counterparts during familiarization than to their straddling counterparts, this might drive participants to choose middle 'words' and not straddling 'words', over non-words. In this chapter we will examine evidence that suggests that acoustic similarity is not sufficient to explain the observed pattern of results.

Is there reason to believe that the contour-straddling 'words' are acoustically more different during familiarization and test? Recall that IPs evidence final lengthening (see Section 2.2.2 in the introductory Chapter 3), which we find in the IPs used in these experiments (Chapter 6, Section 6.2.1, pg. 66). Thus, in going from the last syllable of one contour to the first syllable of the next, there

is a large difference in duration. As a consequence, the syllables that constitute the contour-straddling 'words' differ in duration during familiarization, but not during test. In contrast, the syllables that constitute the contour-internal 'words' have the same durations during familiarization and during test.

The difference between familiarization and test is evident even in the pitch contours of the contour-internal and contour-straddling 'words'. In Figure 9.1, we can see the changes in pitch going from one 'phrase' to the next .

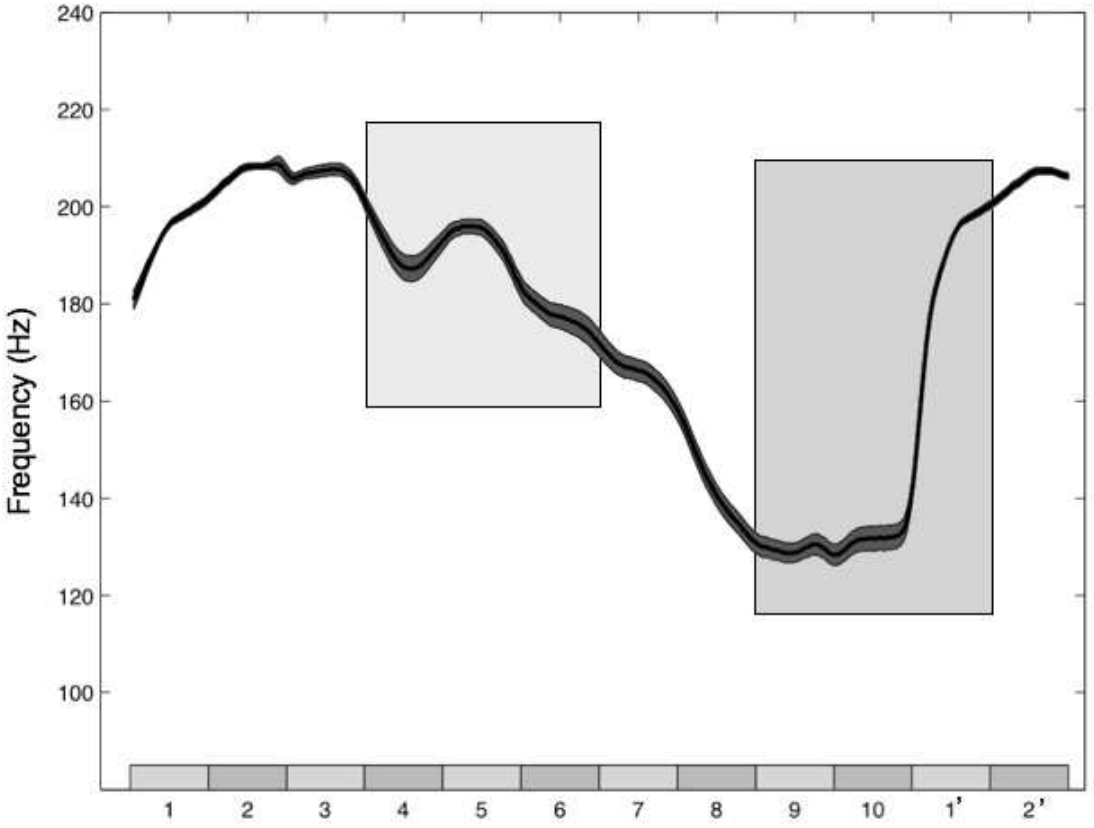


Figure 9.1: This figure shows the pitch contour of 12 syllables (alternating light and dark rectangles at the bottom), corresponding to the 10 σs from the first 'phrase', followed by the first two σs of the subsequent 'phrase'. The values are averages ($\pm$ 1 S.E.) from the first 20 such 12-σ sequences, taken from Experiment 3. The left rectangle delineates the pitch contour of an internal 'word' (at position 4-5-6), while the right rectangle delineates the pitch contour of a straddling 'word' (at position 9-10-1'). The y-axis represents frequency (Hz)

From Figure 9.1 it is clear that there is a larger pitch variation in a contour-straddling 'word' (at position 9-10-1, for instance) than in a contour-internal 'word' (e.g., at position 4-5-6). Since the test items have a neutral prosody, and thus no variation in pitch, they are more similar to the contour-internal 'words' than contour-straddling ones.

Thus, one might conclude that the contour-straddling 'words' during test are acoustically more distant from their counterparts during familiarization than are the contour-internal 'words'. Let us call this the *Acoustic-Distance Hypothesis* (ADH).

How much does ADH actually explain the pattern of results observed in the previous chapters? As we saw above, contour-straddling 'words' have large variations in pitch and duration. To simplify, since edges are accompanied with changes in pitch and duration, contour-straddling 'words' that contain two such edges are less well recognized than contour-internal 'words'.

However, this explanation does not predict the edge effect, as reported in Chapter 7. There, we had seen that 'words' aligned with both the left and the right edges of 'phrases' were better recognized than 'words' in the middles. But, the edge-'words' contain at least one edge. So, by the ADH, they should be *less* well recalled than the middle 'words', which is contrary to the observed results.

Nevertheless, we require direct evidence that the ADH does not account for any of the observed results. Thus, in the next experiment, we ask: what happens when the test items bear the same prosodic characteristics as they did during familiarization?

9.1 Experiment 7: Controlling for acoustic differences I - Using Familiarization prosody during test

In order to satisfy the aims of this experiment, it was necessary to change the familiarization stream in several ways. The main difference was that instead of being associated with several IP contours, each 'word' is now associated with only a single IP. Thus, each 'word' is precisely associated with only a single acoustic

shape. Notice that in the previous experiments, it would seem implausible that there is a *single* acoustic shape associated with each 'word', since the 'words' occur in different acoustic milieus. By associating each 'word' with only a single IP in this experiment, each 'word' has a well-defined acoustic shape during familiarization, which can be used in the test phase. Thus, the test words are intonated and bear the same duration and pitch characteristics as counterparts during familiarization.

This experiment thus directly tests the contribution of acoustic similarity during the recall of contour-internal and contour-straddling 'words'. There are some straightforward expectations about the outcome. If indeed acoustic similarity plays a role in recognition of the trisyllabic 'words' during test, then we would expect to find an improvement in the scores for the contour-straddling 'words'. If instead the crucial element is the presence of a prosodic edge during familiarization, then simply equalizing for acoustics should not play a major role during the test phase, and straddling 'words' would not gain any advantage. Thus, contour-internal 'words' are expected to be recognized better than chance, while straddling 'words' might be recognized at or better than chance.

9.1.1 Material and Methods

The overall logic of the experiment was identical to the previous ones. However, instead of eight IP contours, only two were used. 'Words' 1 and 2 were placed at fixed positions inside two IP contours (call them A and B), while the 'words' 3 and 4 were placed at fixed positions straddling contours $A - B$ and $B - A$. Thus two 'words' have unique, contour-internal prosodies associated with them, while two others have unique, contour-straddling prosodies.

Participants

Sixteen adults participated in this experiment (9 males and 7 females, mean age 26.1 years, range 21-35 years).

Table 9.1: Placement of 'words' in Experiment 9.1

'Word'	Frame(s)	Position (10-σ frames)
'Word' 1	Frame A	4-5-6
'Word' 2	Frame B	5-6-7
'Word' 3	Frames A-B	8-9-1'
'Word' 4	Frames B-A	9-1'-2'

Materials

An algorithm implemented in MATLAB first generated a (quasi) random sequence of two contours A and B, such that there were at least 100 $A-B$ and $B-A$ pairs. The four 'words' were placed as shown in Table 9.1. The remaining positions were filled with noise syllables as before. The entire sequence was generated by an algorithm implemented in MATLAB, which generated an MBROLA file. The MBROLA file was converted to a 12'06" sound file using the es1 (Spanish, male) database as in the previous experiments.

The trisyllabic test items were the same as those used in the previous experiments. However, instead of having a flat prosody, the prosody of each trisyllable ('words' and non-words) during familiarization was grafted onto the test trisyllables. Separate MBROLA files were created for the test items and sound files were generated from these using the es1 database. Notice that generating the test items in this manner is not the same as extracting them from the familiarization stream. This is because, since MBROLA is a diphone - based synthesizer, the exact first phoneme of a 'word' during familiarization would depend on the preceding phoneme. However, this is a minor difference, which is the same for all the 'words'. Indeed, precisely because of the diphone - based synthesis, it would have been very difficult to identify the precise onset of the phonemes, specially for trisyllables beginning with continuants.

Apparatus and Procedure

This was identical to the previous experiments.

9.1.2 Results

In Figure 9.2, the scores for the contour-internal and the contour-straddling 'words' is presented. As can be seen, *neither* kind of 'word' is recognized. The contour-

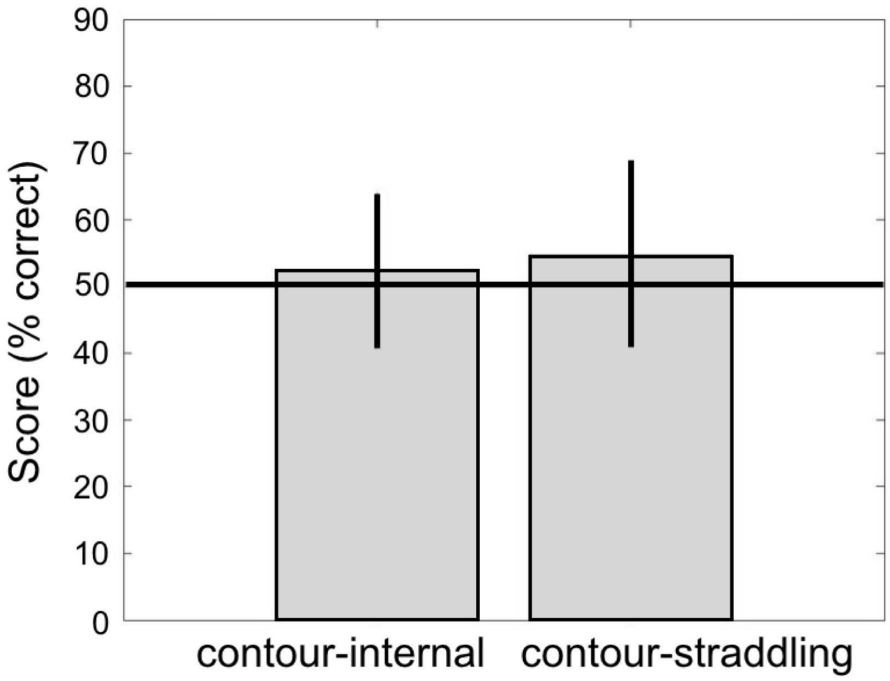


Figure 9.2: Results for Experiment 7. Neither contour-straddling nor contour-internal 'words' are preferred over non-words.

internal words were not preferred; the mean score was 52.3438%, which was not different from chance, $t(15) = 0.4263$, $p = 0.676$. Neither were the straddling words preferred over chance; the mean score was 54.6875%, $t(15) = 0.7165$, $p = 0.4847$. The two groups did not differ, $t(30) = -0.2742$, $p = 0.7858$.

9.1.3 Discussion

The results of this experiment show a surprising failure for both contour-internal and contour-straddling 'words' to be recognized. In order to interpret these results, it is first necessary to show that these are not due to the different material used in Experiment 7.

9.2 Experiment 8: Control for the material in Experiment 7

In Experiment 7, we were interested in understanding the role of acoustic similarity in obtaining the results described in Chapters 6 and 7. Thus, instead of eight IP contours, as in Experiment 3, only two IP contours were used. This was done to ensure that each 'word' occurred in a well-defined acoustic (prosodic) environment.

In the test phase, all the test items carried the acoustic / prosodic characteristics that they had during familiarization. We saw that neither the contour-internal nor the contour-straddling 'words' were chosen over non-words.

Is this failure to prefer 'words' over non-words due to the decreased variability of prosodic contours used in this experiment? In order to test for this possibility, a control experiment was run, wherein the test items all carried a neutral prosody. In doing so, we mimic Experiment 3 (Chapter 6), with the sole difference that we use two, instead of eight IPs.

9.2.1 Material and Methods

Participants

Twelve adults participated in this experiment (7 males and 5 females, mean age 23.1 years, range 18-34 years).

Materials

The familiarization phase used the artificial speech stream from the previous experiment (Experiment 7), while the test items were those from Experiment 3 (Chapter 6).

Apparatus and Procedure

This was identical to the previous experiments.

9.2.2 Results

In Figure 9.2, the scores for the contour-internal and the contour-straddling 'words' is presented. Contour-straddling 'words' are not preferred over non-words, while contour-internal 'words' are. The contour-internal words were preferred over

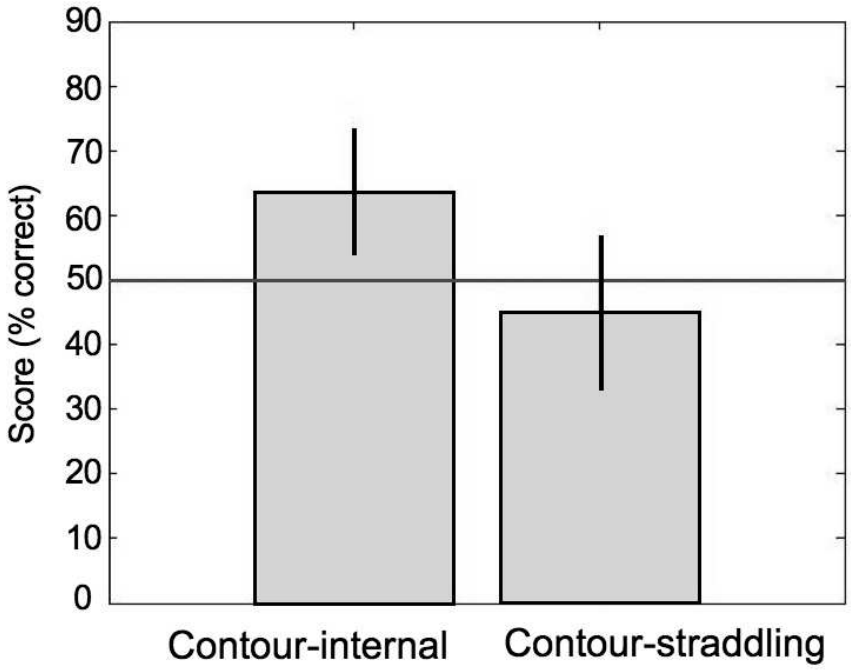


Figure 9.3: Results for Experiment 8. Only contour-internal 'words' are preferred over non-words, replicating the results from Experiment 3. Error bars are 95% confidence limits of the means.

the non-words; the mean score was 63.54%, which was different from chance, $t(11) = 3.03$, $p = 0.012$. The straddling words were not preferred over chance; the mean score was 44.79%, $t(11) = -0.96$, $p = 0.36$. The two groups were significantly different, $t(22) = 2.27$, $p = 0.014$.

9.2.3 Discussion

The results from this experiment replicate the filtering effect of prosody observed in Experiment 3: contour-internal 'words' are recognized, while contour-straddling 'words' are not.

Let us now consider Experiment 7 in light of these results. The results from Experiment 7 showed that neither contour-internal nor contour-straddling 'words' were recognized when the test items carried the acoustic patterns that they had during familiarization. The results from Experiment 8 reassure us that this failure is not due to inability to extract 'words' from the modified speech stream used in these experiments. Thus, we can safely reject the ADH as a possible explanation for the selective recovery of only the contour-internal 'words' in Experiments 8 and 3. But then, how do these results square with the prosodic filtering model that we considered in the previous chapter (see Figure 8.1, pg. 78)? In particular, why are the contour-internal 'words' not recognized when they carry the same prosody as they did during familiarization?

In the introductory chapters, we examined the prosodic organization of spoken language (Chapter 2). We saw that an utterance, defined as a stretch of speech bounded by silence, is composed of hierarchically nested prosodic constituents (see Figure 2.2, pg. 13). Further, we saw that as a consequence of *Proper Containment* (Principle v, pg. 19), any utterance contains at least one of all the lower prosodic constituents. Thus, a word produced in isolation, being a single utterance, would also be expected to have the characteristics of an IP, of a ϕ and so on, down to a syllable.

Subsequently, a single word spoken in isolation cannot simply carry the prosody of a *portion* of an IP; it's prosody must respect the factors that define how *any* utterance may be produced. This suggests if a 'word' that is part of an IP (a contour-internal 'word', for example) is presented in isolation, it must have the prosodic characteristic of a full IP, not just part of one.

Indeed, a few studies have shown that words excised from fluent speech are often unintelligible (e.g., Pollack & Pickett, 1964; Bard & Anderson, 1983, 1994). In addition, the intelligibility of words depends on contextual factors. For example, Bard and Anderson (1994) found that word intelligibility in fluent speech was inversely related to word predictability, and that this was worse for child-directed speech than for adult-directed speech.

Thus, we can explain the failure of contour-internal 'words' being recognized due to the unnatural prosody that such 'words' carry in the test phase. This

explanation makes the interesting prediction that the test items can have *any* prosody, as long as it is a 'natural' one. Thus, in order to support this explanation for the results from Experiment 7, we will test such a prediction in the next experiment.

9.3 Experiment 9: Controlling for acoustic differences II - List prosody during test

The results from Experiments 7 and 8 suggest that merely having the same acoustic / prosodic characteristics during familiarization and test are not sufficient to recall even the contour-internal 'words' correctly. It was suggested that the test items need to have a 'natural' prosody. Thus, in this experiment, we will examine the effect of using a natural prosody in the test phase. To do so, we will replicate Experiment 3 (pg. 66), with the test items bearing a natural prosody.

9.3.1 Material and Methods

Participants

Fourteen adults participated in this experiment (5 males and 9 females, mean age 23.4 years, range 20-32 years).

Materials

The familiarization material was the same as for Experiment 3. To recapitulate, two trisyllabic 'words' were placed inside of Italian IP-contours realized over *frames*, each 10-syllables in length. Two others were placed straddling consecutive frames, and thus straddling IP-contours (see page 66 and Figure 6.4, pg. 67, for further details).

To prepare the test phase, the pairs of triplets from each test trial were first recorded by a female, naive Italian speaker[1]. She was instructed to read pairs of triplets, in a natural manner. This resulted in what is termed *list prosody*, which is characterized by a pitch decline on the last (in this case the second) item

[1]I thank Silvia Pontin for these recordings

of the list. In Figure 9.4, the (time normalized) pitch contours for the triplets, both 'words' and non-words in the first and the second position for each trial is plotted. As can be seen, all 16 contours start out at approximately the same pitch

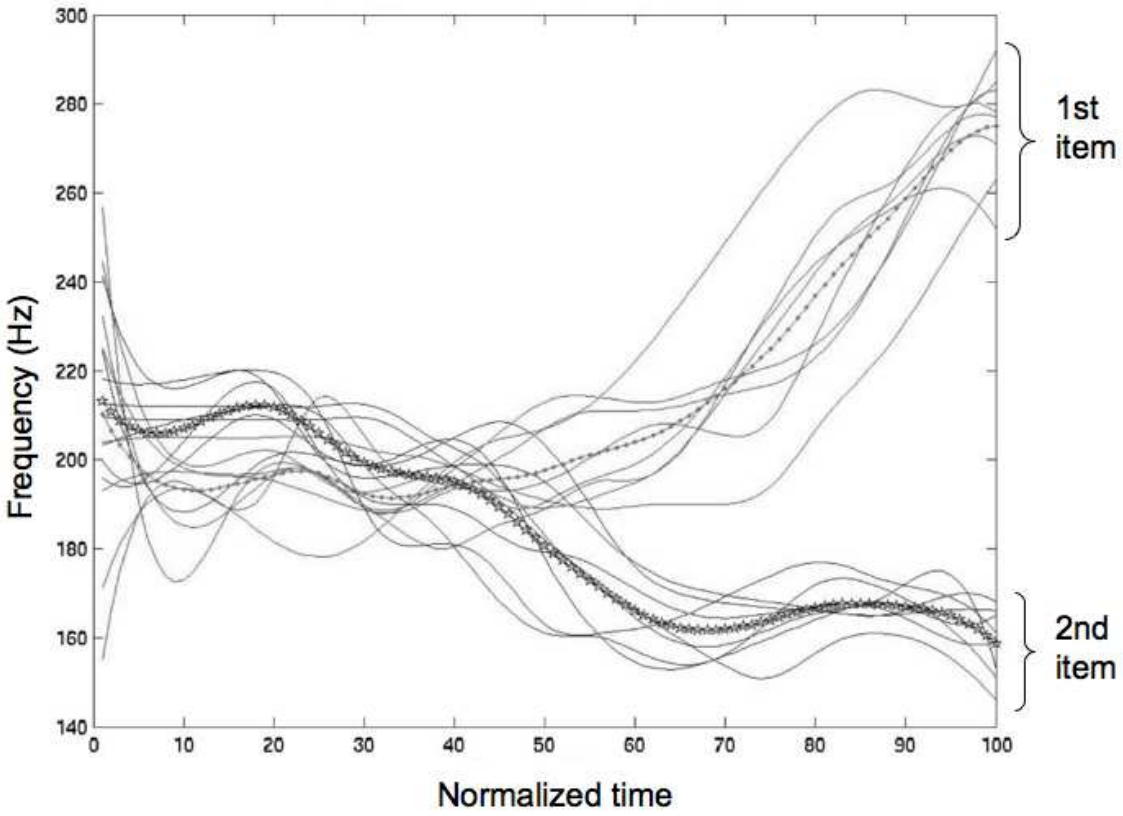


Figure 9.4: Pitch contours for the first and second triplet in each trial, spoken by a naive, female Italian speaker. The items in the first position end with a rising pitch, while the items in the second position end with a falling pitch.

levels, but then the eight that correspond to the first item in each list (trial) form a cluster with a relatively higher pitch, while the eight that correspond to the second item in each trial form a cluster at a relatively lower pitch level. Recall from Experiment 6.2 that each 'word' and non-words in the test phase occurs in both the first and the second positions.

In Figure 9.5, the duration of the individual phonemes for all the triplets is shown. From the figure, it is clear that there are no systematic differences in the

durations of the phonemes in the two positions. This contrasts with the pitch contours, which are strongly position dependent (Figure 9.4 on the preceding page).

The pitch and duration characteristics of all the test items were used to create prosodic test tokens in MBROLA, again using the es1, Spanish male database.

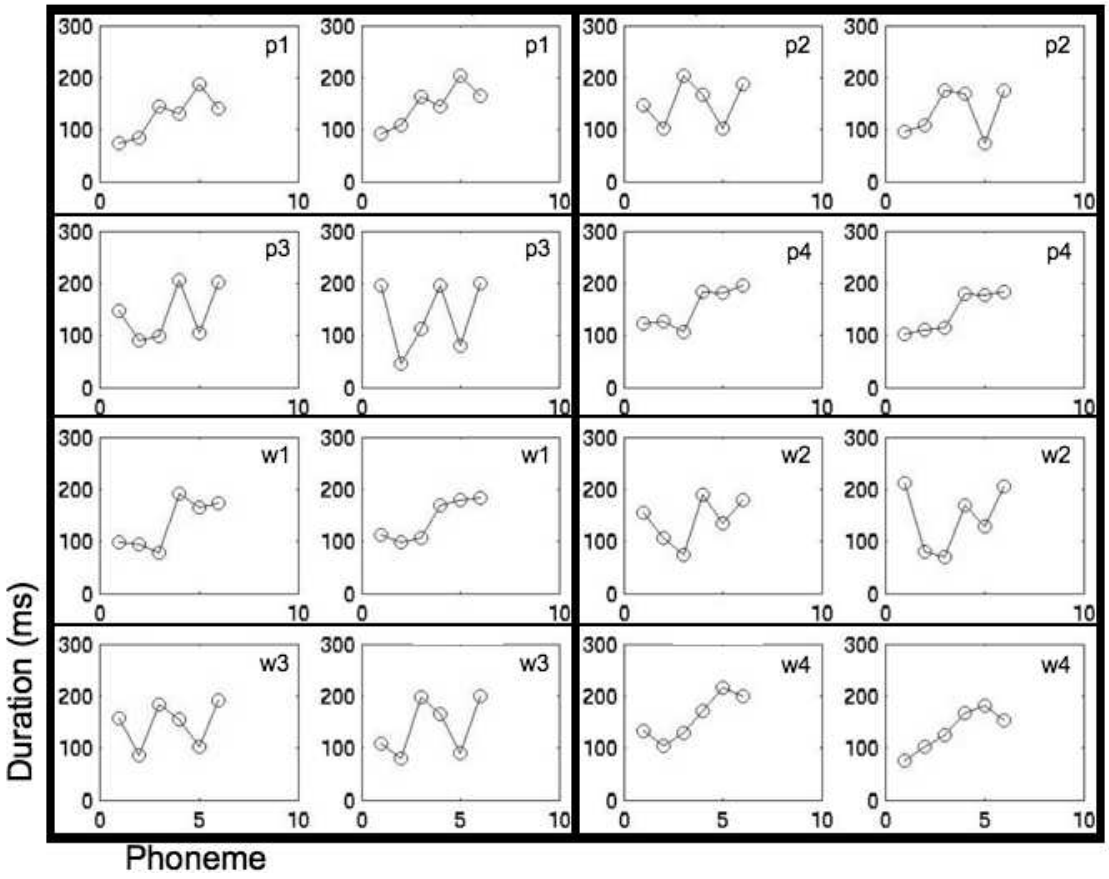


Figure 9.5: The figure shows phoneme durations of the six phonemes making up the eight trisyllabic test items. For each triplet, the figure to the left represents the token in the first position, while the figure to the right represents the token in the second position in each trial. p1: /da-fo-pu/; p2: /ge-ʧu-ni/; p3: /ki-me-vo/ ; p4: /le-ʎa-te/; w1: /ni-da-fo/; w2: /pu-le-ʎa/; w3: /te-ki-me/; w4: /vo-ge-ʧu/. (See also Table 6.1 on page 63)

Aparatus and Procedure

These were identical to Experiment 3 on page 66.

9.3.2 Results

In Figure 9.6, the scores for the contour-internal and the contour-straddling 'words' is presented. These results replicate those in Experiment 3, in that the contour-internal 'words' are recognized better than chance, while the contour-straddling 'words' are not. Overall segmentation (61.16%, S.D. 8.9) was significant, $t(13) = 4.69, p < 0.001$. Contour-internal 'words' were recognized better than chance (mean 68.75%, S.D. 11.76), $t(13) = 5.97, p < 0.0001$. Straddling 'words' were not recognized at better than chance levels (mean 53.57%, S.D. 10.32), $t(13) = 1.3, p = 0.22$. The internal and straddling 'words' differed significantly, $t(26) = 3.63, p < 0.002$.

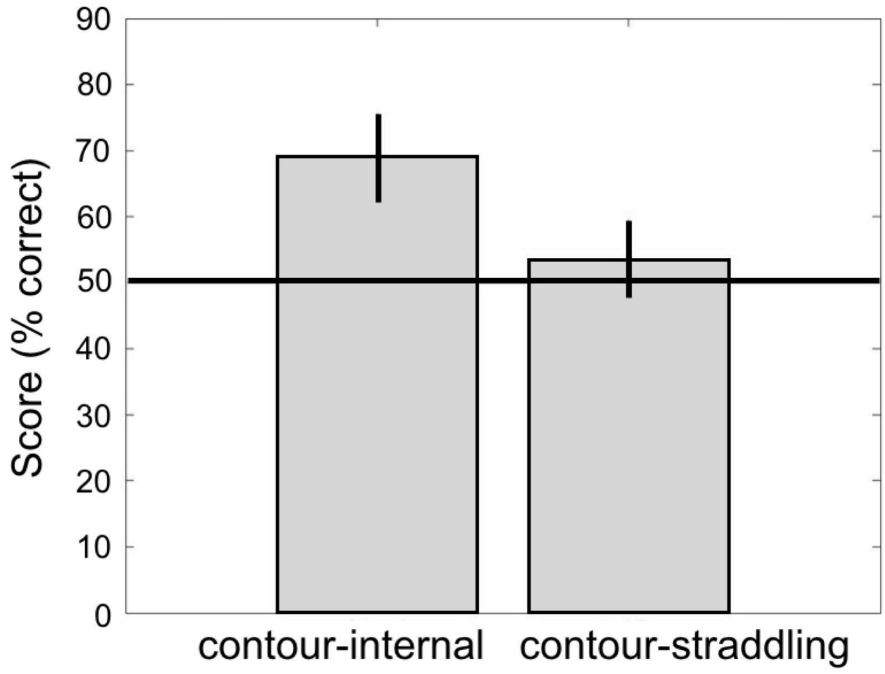


Figure 9.6: Results for Experiment 9. Contour-straddling are preferred over non-words, while contour-internal 'words' are not.

9.3.3 Discussion

In this chapter, we have looked at evidence that supports the view that acoustic similarities cannot account for the filtering effect of prosody.

First, we found (Experiment 7) that presenting the same prosody during familiarization and during test does not enhance recall of the 'words'; instead, performance is *degraded* for the contour-internal 'words'. We hypothesized that test items bearing the same prosody as they did during familiarization are unnatural, because a portion of an IP is not itself a well-formed IP in isolation. In support of this view, we saw that the filtering effect is obtained even when the test items bear an unrelated, but natural, list prosody (Experiment 9).

These experiments further support the model proposed in Chapter 8 (see Figure 8.1, pg. 78). In this model we proposed that statistics over the abstract, syllabic level, are computed in parallel with the prosodically determined 'phrases'. In support of this model, we showed that if we could tap into only the abstract level of representation, we would find that both contour-internal and contour-straddling 'words' are preferred over non-words.

Taken with the findings from this chapter, we can further speculate that not only are the TPs computed over the abstract syllabic representation, but also that the output of such computations are 'words' stored as abstract, phonological forms.

'Let's hear it,' said Humpty Dumpty. 'I can explain all the poems that were ever invented–and a good many that haven't been invented just yet.'

Through the looking glass,
Lewis Carroll

Chapter 10

The filtering effect of non-native prosody

In this thesis, we have asked how prosodic cues can interact with statistical cues (TPs) in segmenting out words from fluent speech. In the previous chapters we saw that TPs are computed over the syllabic representation, and high-TP syllabic sequences are considered as possible lexical candidates. However, prosody has a filtering effect, such that high-TP syllabic sequences that straddle a prosodic phrase are considered poor lexical candidates. We examined the suggestion that the putting together of statistical and prosodic information is accomplished by memory systems; when the prosodic trace is weakened, the filtering effect is no longer obtained.

We spent the entire previous chapter in showing that just the acoustic characteristics that accompany the various positions in relation to prosodic phrases (e.g., contour-internal or contour-straddling) do not explain the observed results. However, there is yet another possible confound. In all the experiments reported thus far, Italian participants were exposed to artificial streams bearing Italian IP prosody. It is possible that through years of experience in their native language, our Italian participants were sensitive to subtle acoustic cues that mark Italian IPs, and hence provide cues to the 'words' in our artificial speech streams.

Clearly, if we are to consider phrasal prosody as an early cue to segmentation, it should be available even with little or no experience, as is the situation for a neonate. Thus, in this chapter, we will examine the filtering effect of Japanese

prosody on Italian adults.

10.1 Experiment 10: An interaction between prosody and statistics using Japanese IP characteristics

In order to examine the effect of a non-native prosody, we will create artificial speech streams comparable to those used in the previous chapters, but bearing Japanese instead of Italian IP prosody.

In this first experiment, we will replicate, using Japanese prosody, the results from Experiment 3 (Chapter 6), where we first observed an interaction between prosody and statistics using Italian prosody.

10.1.1 Material and Methods

Participants

Fourteen adults participated in this experiment (5 males and 9 females, mean age 25.3 years, range 19-36).

Materials

In order to get Japanese IPs, a set of sentences were constructed[1]. Each set of sentences was constructed such that there was one clear IP corresponding to a single simple declarative clause, and it was flanked by IPs on either side. The list of sentence sets if given in Appendix B on page 161.

A single Japanese female speaker recorded the entire material[2]. The material was recorded with an Audio-Technica ATR20 microphone connected to a SoundBlaster sound card on a PC under Window 2000^{TM}. CoolEdit (Syntrillium Corp.) was used to record and digitally manipulate the speech waveforms. The

[1]I'm extremely grateful to Yuki Hirose at the department of Human Communication, The University of Electro-Communications, Tokyo, Japan, and Hifumi Tsubokura at the Tokyo Women's Medical University, Tokyo, Japan, for help in the preparation of the material.

[2]Many thanks to Yoko Imai for recording the set of Japanese sentences

speech segments corresponding to the IPs were digitally excised. As in the Italian case, for each IP, we measured the pitch contour, smoothly interpolating across unvoiced segments using PRAAT (www.praat.org). A single pitch contour was converted into a vector of 400 pitch points. Thus, 20 pitch points per phoneme could be used to shape each of the 20 phonemes (from 10 CV syllables) in a single frame. From the nine recorded IPs, we thus obtained nine different pitch contour vectors. Figure 10.1 shows a comparison of Italian and Japanese pitch contours. Although there are differences, both show a clear downward-going pitch. Note that Japanese has a larger difference between the initial and the final pitch levels.

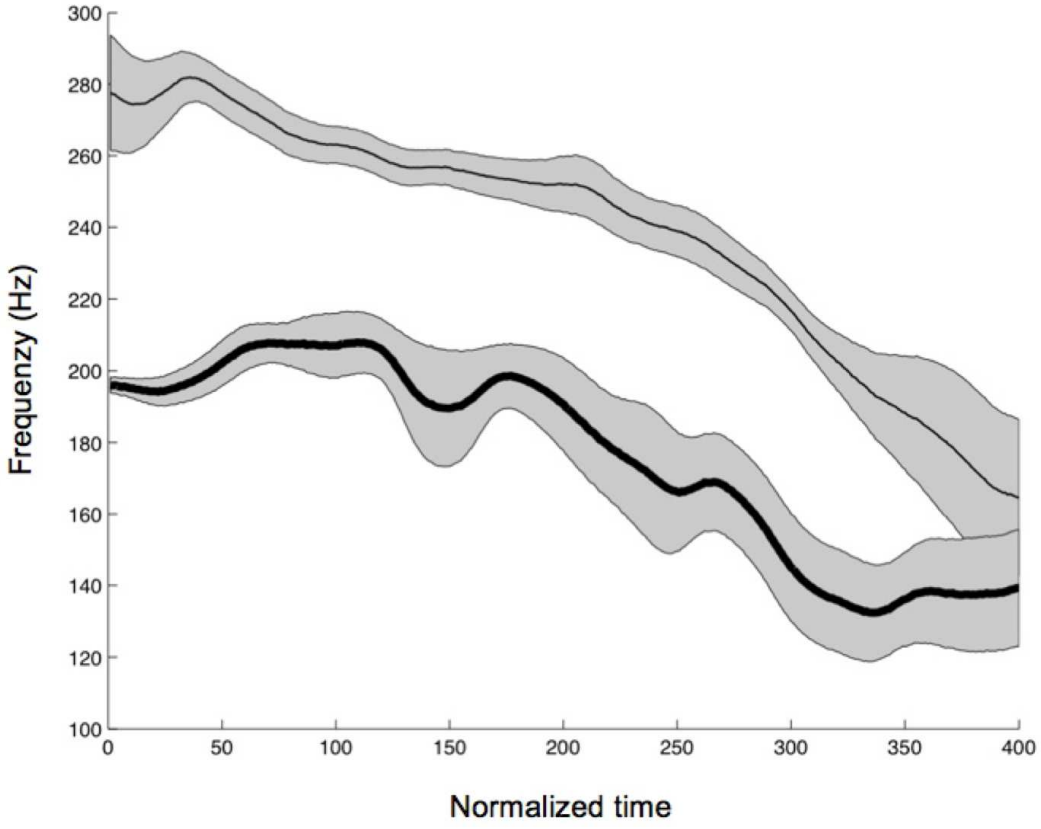


Figure 10.1: Comparison of Italian and Japanese pitch contours. The shaded regions represent ±1 S.E.s around the means: thin line: Japanese, thick line: Italian. The x-axis represents (normalized) time, the y-axis is frequency.

Next, the durations of the first and last syllables of each IP were measured.

Table 10.1: Comparison of initial and final phoneme durations for Italian and Japanese

	Italian	Japanese
Initial	79.9 msec	73.2 msec
Final	99.6 msec	99.8 msec

The durations were divided by the number of segments in the syllables, to get a normalized value. The average normalized length of the phonemes of the last syllable (99.8 ms) was significantly greater than the average normalized length of the phonemes of the first syllable (73.19ms), paired t-test, $t(8) = 3.72$, $p < 0.005$. These values correspond rather closely to those obtained for Italian, as described in Table 10.1 below: Thus, it was decided to keep the length of the final phonemes unchanged, while the length of the initial phoneme was decreased to the (rounded) difference, by 5msec. Thus, in these experiments, the length of the phonemes comprising the first syllable was 95ms (as opposed to 100ms for Italian), and the length of the phonemes comprising the last syllable was 140ms, as for Italian (see page 66).

In order to see if Italian adults subjectively show any evidence of perceiving the Japanese IPs, another eight naïve participants were exposed to a 2 minute sequence of 10-syllable frames bearing the Japanese IP characteristics as derived above. These frames contained a completely random sequence of syllables. The participants were instructed that they would hear speech in a foreign language, and were instructed to tap the space-bar every time they heard a 'sentence' in the artificial language.

In Figure 10.2 on the next page, we see the distribution of ΔTap, the time interval between one tap of the space-bar and the next, for all the eight participants. From the figure, it is clear that Italian participants do indeed perceive some kind of grouping of the random syllables that is congruent with the Japanese IPs. The time interval between one tap of the space bar and the next shows a mode at 2390 ms (the first peak in Figure 10.2), corresponding to one 'phrase', and another at 4780 ms (the second peak), corresponding to two 'phrases'.

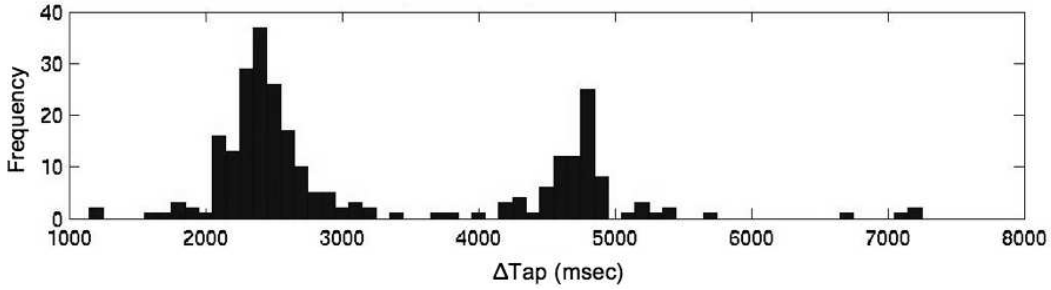


Figure 10.2: ΔTap reveals that Italian adults do perceive Japanese IPs. The histogram of ΔTap (the difference between consecutive presses of the space-bar, see text) shows peaks primarily at time periods corresponding to one (2390 ms) or two (4780 ms) 'phrases'.

To prepare the familiarization stream for the main experiment, the same MBROLA file as used for Experiment 3 was used, and the eight Italian contours were each replaced by one of the Japanese contours. This ensures that the statistical properties of the two sound streams (with Italian and with Japanese prosody) are largely matched. That is, they are identical for all the distributional properties at the level of the syllable, and the order of appearance of the IP contours (though not their identity, naturally) are identical. Thus, any difference in results could be attributed solely to the prosodic characteristics of the Japanese IPs.

The resulting MBROLA file was converted to a 22.05 kHz mono wave file using the es1, Spanish male database. This file was converted to a stereo file and the first and last 5 seconds were ramped in amplitude to remove onset and offset cues.

Apparatus and Procedure

This was the same as for Experiment 6.2.

10.1.2 Results

In Figure 10.3, the results from this experiment are displayed. It is clear that Japanese prosody appears to have the same effect on Italian adults as does Italian prosody: contour-internal 'words' are significantly preferred over the non-words, while the straddling 'words' are not. Overall segmentation was evidenced by a

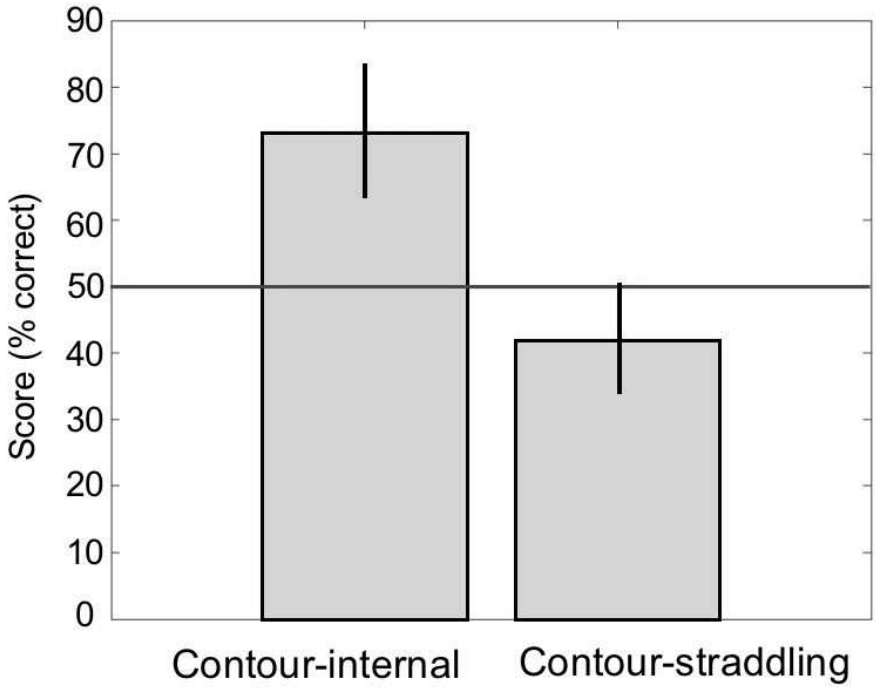


Figure 10.3: Results for Experiment 10. Contour-internal are preferred over non-words, while contour-straddling 'words' are not.

significant score of 57.59%, $t(13) = 2.36, p = 0.035$. The mean score for the contour-internal 'words' was 73.21%, $t(13) = 4.94, p < 0.0005$, while the mean score for the straddling 'words' was 41.96%, $t(13) = -2.09, p = 0.057$. The score for the straddling 'words' thus shows a marked tendency to be *below* chance. This is equivalent to saying that the straddling 'words' in this experiment had a tendency to be rejected. The two groups were themselves significantly different, $t(13) = 5.15, p < 0.0001$.

An ANOVA compared this experiment with its counterpart, Experiment 3 (page 66), wherein Italian prosody was used. *Language* (Italian or Japanese) was one fixed factor, while *Position* was the other. The ANOVA revealed a main effect of *Position*, $F(1, 64) = 36.48, p \leq 0.0001$, while there was no main effect of *Language*, $F(1, 64) = 0.05, p = 0.82$. Also, there was no significant interaction between the two , $F(1, 64) = 0.815, p = 0.37$.

10.1.3 Discussion

The results from this experiment suggest that Japanese IP characteristics evince the same processing in Italian adults as does Italian prosody. These results strongly suggest that participants rely on universal acoustic / prosodic characteristics of IPs, rather than subtle cues that characterize them, in segmenting speech in this experimental paradigm.

It is interesting to note that the straddling 'words' show a tendency to being *rejected*. Indeed, in the model we consider (see Figure 8.1(b), pg. 78), the statistics over the syllables recovers both contour-internal and contour-straddling 'words'. The tendency for the contour-straddling 'words' to be rejected in this experiment implies that, possibly due to an enhanced filtering effect, participants successfully recognize the contour-straddling 'words' are *not* being likely lexical candidates. We will see more evidence for this in the following chapter.

For now, let us buttress our findings that Japanese prosody is indeed perceived and utilized by Italian participants.

10.2 Experiment 11: Looking for an 'Edge effect' with Japanese prosody

In this experiment, we replicate, using Japanese prosody, the edge effect observed with Italian prosody in Chapter 7 (pg. 71). As in the Experiment 4 on page 72 of that chapter, two groups of participants were exposed to two different streams, one containing 'words' at the left edges and in the middles of IPs, and the other containing 'words' at the right edges and in the middles of IPs.

10.2.1 Material and Methods

Participants

Fourteen adults were exposed to the stream with edge-words at the left edge (4 males and 10 females, mean age 23.5 years, range 20-28 years). A separate group of twelve adults were exposed to the stream with edge-words at the right edge (1 male and 11 females, mean age 22.9 years, range 19-27 years).

Materials

The preparation of the familiarization stream for this experiment paralleled that of the previous one. The MBROLA files from Experiment 4 (in which the edges of Italian IPs were examined) were used as a starting point, and each Italian IP contour in those files was replaced by one Japanese contour. Again, as in the previous experiment, this ensures that the distributional properties with respect to the syllables is matched in the present experiment and in Experiment 4.

The entire sequences of phonemes were fed to MBROLA, using the Spanish male diphone (es1) database. The final output files were 22.05 kHz, 16-bit, mono wave files of duration 4 min each. These files were converted into stereo files and the initial and final 5 sec were ramped up and down to eliminate onset or offset cues to edge-words and middle-words. The test phase was identical for both groups of participants and was identical to Experiment 4.

Apparatus and Procedure

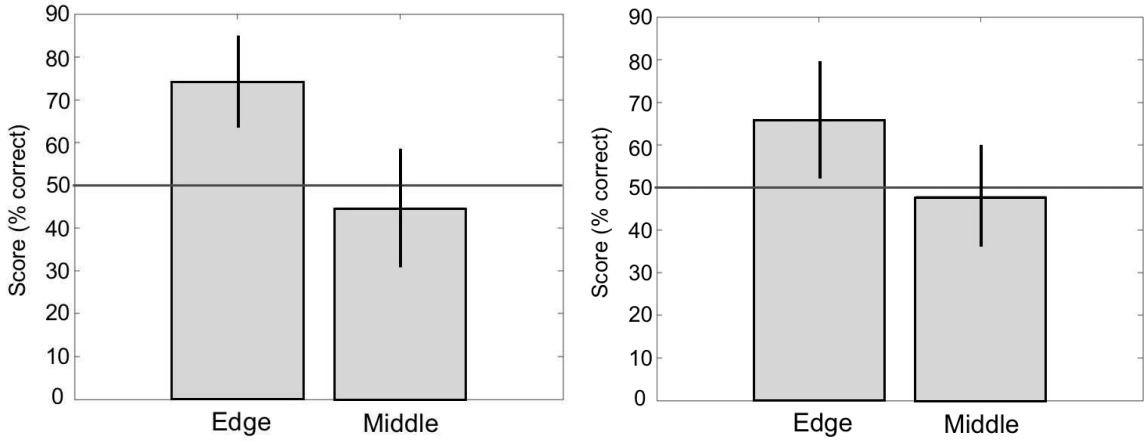
These were identical to Experiment Experiment 4.

10.2.2 Results

The overall score for the left-edge stream group (mean 59.38%, S.D. 11.43) was significantly different from chance, $t(13) = 3.07, p < 0.01$.However, for the right edge stream group, the overall score (mean 56.77%, S.D. 13.18) was not significantly different from chance, $t(11) = 1.78, p = 0.1$.In Figure 10.4, the scores for edge-words and middle-words for the left- and right-edge groups are shown separately.

The edge-words at the left edge (mean 74.11%, S.D. 18.65) were recognized significantly above chance, $t(13) = 4.84, p < 0.001$. Similarly, edge-words at the right edge (mean 65.63%, S.D. 22.06) were recognized significantly above chance, $t(11) = 2.45, p < 0.05$. The middle-words in the two conditions were less well recognized, left edge: 44.64%, S.D. 24.37, $t(13) = -0.82, p = 0.43$, and right edge: 47.92%, S.D. 19.09, $t(11) = -0.38, p = 0.71$.

Pooling the data in an ANOVA with factors Edge (left or right) and Position (edge-word or middle-word) revealed a significant effect of Position, $F(1, 24) =$



(a) Left edge

(b) Right edge

Figure 10.4: Mean scores (% correct) for edge-'words' and middle-'words' from Experiment 11. In (a), edge-'words' occurred at the left edge of IPs, while in (b), edge-'words' occurred at the right edge of IPs. Edge-'words' are efficiently segmented, while middle-'words' are segmented with much less efficiency. Error bars represent 95% confidence limits of the means.

$11.99, p = 0.002$.The Edge condition was not significant ($p > 0.6$), and neither was the interaction ($p > 0.3$). Post-hoc (Scheffe) tests revealed that the edge-words were recognized better than the middle-words for the left edge group, $p < 0.01$, but not for the right edge group, $p = 0.09$.

Since the ANOVA revealed no main effect of Edge or interaction between Edge and Position, we collapsed the data from the left- and right-edge groups. The combined data revealed that edge-'words' were recognized better than chance, mean 70.19% (S.D. 20.3)$t(25) = 5.1, p < 0.0001$, while the middle-'words' were not, mean 46.15% (S.D. 21.7), $t(25) = -0.9, p > 0.3$. In addition, the score for the combined data for the edge-'words' was significantly different from the score for the combined data for the middle-'words', $t(50) = 4.12, p < 0.001$.

A separate ANOVA compared the results from this Experiment (edges of Japanese contours) with Experiment 4 (edges of Italian contours). The factors were Language (Japanese or Italian), Edge (left or right) and Position (edge-word or middle-word). There was a main effect of Position, $F(1, 96) = 32.5, p \leq 0.0001$

and a main effect of Language, $F(1, 96) = 6, p = 0.016$, while the factor Edge was not significant. None of the two- or three-way interactions were significant. A post-hoc Scheffe test revealed that participants in the Italian condition performed better than those in the Japanese condition (overall, 9.7% greater accuracy in the Italian condition), $p = 0.016$.

10.2.3 Discussion

The results using Japanese IPs replicate the pattern of results obtained with Italian IPs. In both cases, internal 'words' are recognized while 'straddling' words are not (Experiments 3 and 10), and 'words' at the edges of IPs are better recognized than 'words' in their middles (Experiments 4 and 11).

Notice that Japanese is geographically, historically and structurally very different from Italian. Despite these dissimilarities however, the overt realization of IPs from both languages contain cues that signal 'phrases' in otherwise fluent speech. In the experimental paradigm described in this thesis, these are indexed both by an advantage of IP-internal 'words' over straddling 'words', as well as an advantage for edge 'words' over middle 'words'.

Comparing Experiments 4 and 11 (edge-words against middle-words with Italian or Japanese IPs) revealed a significantly better performance with Italian IPs. The better performance with Italian IPs was not observed while comparing Experiments 3 and 10 (internal 'words' against straddling 'words' with IPs from the two languages). Thus, although in some tasks familiarity with native prosody results in an advantage, nevertheless, the overt realization of IPs from both languages appear to contain cues that signal 'phrases' in otherwise fluent speech.

They were in such a cloud of dust, that at first Alice could not make out which was which: but she soon managed to distinguish the Unicorn by his horn.

Through the looking glass,
Lewis Carroll

Chapter 11

Acoustic contributions to prosodic phrases

The overt realization of prosodic phrases is accompanied by acoustic cues. In the introductory chapters, we examined some such cues, including final lengthening and a pitch decline-reset at the end of prosodic phrasal constituents. Indeed, we find such acoustic cues even in the Italian and Japanese IPs used in this thesis, as can be seen in Table 10.1 (pg. 106) and Figure 10.1 (pg. 105). The results from the previous chapter suggest that these cues might be sufficient to cause the filtering effect, since they are obtained for both Italian and Japanese IPs, with Italian adults.

In Chapter 9 we saw that acoustic properties of 'words' heard during familiarization *per se* did not play any major role in their subsequent recall in the test phase. Thus, we assume that the suprasegmental, acoustic / prosodic properties of the syllables constitute a parallel source of information; one that groups syllables into 'phrases' through the identification of the edges of such 'phrases'. In this chapter, we will tease apart the contributions from the durational and the pitch cues.

11.1 Experiment 12: 'Filtering' of contour-straddling words through final lengthening

First, let us consider final lengthening. From Table 10.1 (pg. 106), we see that both Italian and Japanese IPs show an increase in duration for the phonemes of the last syllable, compared to the phonemes of the first syllable, as seen in previous studies (e.g, Marotta, 1985 for Italian, Fisher and Tokura, 1996; Ueyama, 1999 for Japanese). Thus, in this experiment, we will examine the effect of final lengthening as the sole prosodic cue to phrasal boundaries.

11.1.1 Material and Methods

Participants

Twelve adults participated in this experiment (4 males and 8 females, mean age 23.8 years, range 19-29).

Materials

As can be seen from Table 10.1 (pg. 106), the durations for the initial and final phonemes are comparable for Italian and Japanese in our corpus. Thus, the Japanese values were chosen, since these show a slightly greater difference in duration between the initial and final syllables.

Thus, the speech stream was derived from that used in Experiment 10 on page 104, where we observed the filtering effect of Japanese prosody. In the MBROLA file used to create the speech stream for that experiment, all the pitch information was removed, and each phoneme had a constant pitch of 100 Hz (similar to that used in the prosodically 'flat' familiarization stream in Experiment 2, pg. 59).

The resulting file was converted into a 22.05 kHz sound file using the es1 Spanish male database as before.

Apparatus and Procedure

These were identical to the previous experiments.

11.1.2 Results

The results from the experiment are shown in Figure 11.1. As can be seen, the

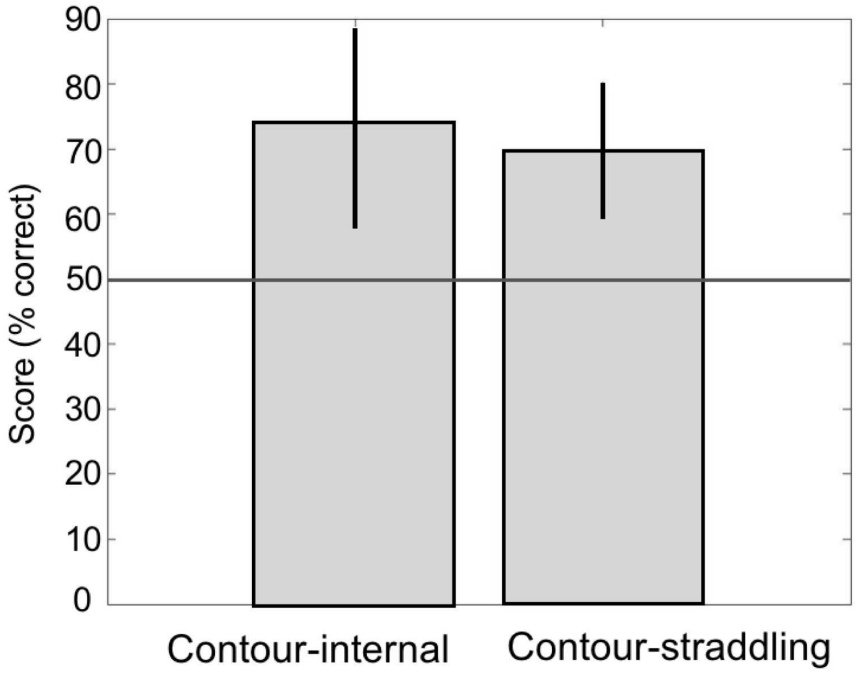


Figure 11.1: Results for Experiment 12. The results indicate that both internal and spanning 'words' are correctly segmented, as in Experiment 2.

results mirror those obtained when there were no prosodic cues during familiarization at all, in Experiment 2 on page 59. Overall segmentation was attested by a significant score of 71.88% (S.D. 17.78), $t(11) = 4.26, p < 0.005$. Internal 'words' had a score of 73.96% (S.D. 22.9), $t(11) = 3.62, p < 0.005$, and straddling 'words' had a score of 69.8% (S.D. 16.39), $t(11) = 4.18, p < 0.005$.

An ANOVA, compared this experiment with Experiment 10, wherein both pitch and length characteristics of Japanese IPs were used. The fixed factors were *Familiarization Type* (pitch plus length or pitch alone) and *Position* (internal or straddling). The results indicated a main effect of *Familiarization Type*, $F(1, 24) = 5.899$, $p < 0.03$ as well as *Position*, $F(1, 24) = 20.691$, $p = 0.0001$. In addition, the interaction between the two was significant, $F(1, 24) = 12.1$,$p < 0.002$.

Post Hoc (Scheffe) tests indicated that the Straddling 'words' in this experiment and Experiment 10 were significantly differently perceived, $p < 0.0001$.

11.1.3 Discussion

These results suggest that phrase-final lengthening alone, at least within this experimental paradigm, cannot account for the observed pattern of results. The results obtained with only the presence of duration cues to prosodic phrases look no different from those obtained with a neutral prosody (Experiment 2 on page 59).

However, there is an alternate possible explanation. In Italian, lexical stress falls on the penultimate syllable (e.g., Nespor & Vogel, 1986). In particular, for open (CV) syllables, the main acoustic correlate of stress is an increase in vowel duration (e.g., Bertinetto, 1981; D'Imperio & Rosenthall, 1999; Santen & D'Imperio, 1999). Thus, due to years of experience with Italian, the participants in these experiments might have been predisposed to place a word boundary after the syllable following the lengthened syllable.

Recall that the contour-straddling 'words' were placed at positions 9-10-1' or 10-1'-2' (see the Methods sections for Experiment 2 and Experiment 3). This implies that the Italian participants perceived half of the contour-straddling 'words' having stress on the penultimate syllable, as is the most common pattern in Italian. Therefore, it is likely that the participants treated lengthening not as a cue to the end of a phrase, but as a cue to a lexical item, and hence they showed an absence of the prosodic filtering effect.

11.2 Experiment 13: Pitch alone can induce 'filtering' I - Italian

Let us now examine the effect of pitch for the prosodic filtering effect. In the previous experiment we found that final lengthening alone does not cause prosodic filtering. Therefore, we expect that both Italian and Japanese pitch contours alone should induce the prosodic filtering effect observed in earlier chapters. In this experiment, we first look at thhe effect of Italian pitch contours alone.

11.2.1 Material and Methods

Participants

Fourteen adults participated in this experiment (4 males and 10 females, mean age 25.2 years, range 21-31).

Materials

To prepare the artificial speech stream for this experiment, the MBROLA file from Experiment 3 was taken as the starting point. The durations of all the phonemes was set to a constant 120 ms, which is the average of all the phonemes. Thus, this experiment is identical to the Experiment 3 except that the 'phrases' do not show final lengthening.

Apparatus and Procedure

These were identical to the previous.

11.2.2 Results

The results are displayed in Figure 11.2. In contrast to the results from Experiment 3 (page 68), wherein overall segmentation was significant, we find that in this experiment the overall segmentation score of 53.13% (S.D.18.63) was not significantly different from chance, $t(13) = 0.63, p = 0.54$. The internal 'words' had a score of 67.86% (S.D. 24.86), which was significantly different from chance, $t(13) = 2.69, p < 0.02$. The straddling 'words' had a score of 38.39% (S.D. 24.25), which were at chance, $t(13) = -1.79, p = 0.097$. The internal and straddling 'words' were significantly different from each other, $t(26) = 3.174, p < 0.005$.

11.2.3 Discussion

We find that the prosodic filtering effect obtained in Experiment 3, wherein both Italian pitch cues and final lengthening were present, is found even with only the pitch cues. As we saw in the previous experiment, final lengthening by itself might be confounded with lexical stress cues. Thus, at least within the experimental

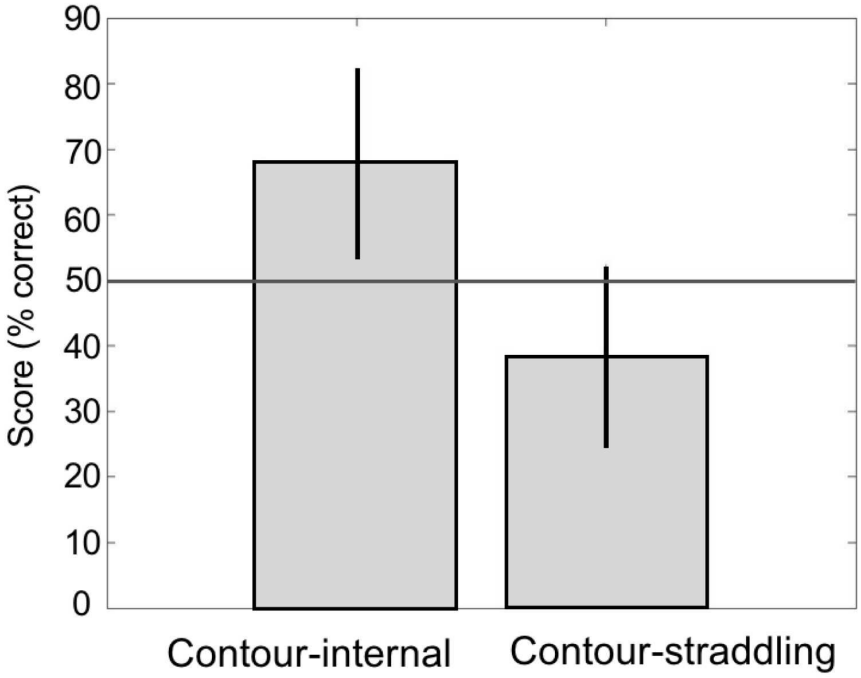


Figure 11.2: Results for Experiment 13. The prosodic filtering effect is observed with Italian pitch contours alone: 'words' at the edges of contours are preferred over non-words, while 'words' in the middles are not.

paradigm utilized in this thesis, pitch contours that accompany Italian IPs are sufficient to induce the prosodic filtering effect.

11.3 Experiment 14: Pitch alone can induce 'filtering' II - Japanese

In this experiment, we will look at the effect of Japanese pitch contours alone, in the absence of final lengthening. Given the results observed in the previous experiments in this chapter, we expect that Japanese pitch contours alone also cause the prosodic filtering effect.

11.3.1 Material and Methods

Participants

Twelve adults participated in this experiment (6 males and 6 females, mean age 24.3 years, range 20-32).

Materials

The material was constructed starting with the MBROLA file from Experiment 10, where we saw the prosodic filtering effect of Japanese prosody. In the MBROLA file, all phoneme durations were set to 120 ms as in the previous experiment. The pitch contours were left untouched. Thus the only difference between the familiarization stream of this experiment and Experiment 10 is that there are no variations in phoneme durations in this experiment.

Apparatus and Procedure

These were identical to the previous experiments.

11.3.2 Results

The results are presented in Figure 11.3 The overall segmentation score of 53.65% was not significant, $t(11) = 1.05, p = 0.32$. This was because, while the internal 'words' were recognized better than chance, the straddling 'words' were significantly rejected. The score for internal 'words' was 76.04%, $t(11) = 5.23, p < 0.001$, while the that for the straddling 'words' was 31.25%, $t(11) = -4.45, p < 0.001$. The two groups differed markedly from each other, $t(22) = 6.87, p < 0.00001$.

11.3.3 Discussion

In this experiment we find, as in previous experiments, that contour-internal 'words' are significantly preferred over non-words. However, in contrast with previous experiments, we find a significant *rejection* of the contour-straddling 'words'. That is, the choice of the participants indicates that they judge trisyllabic items

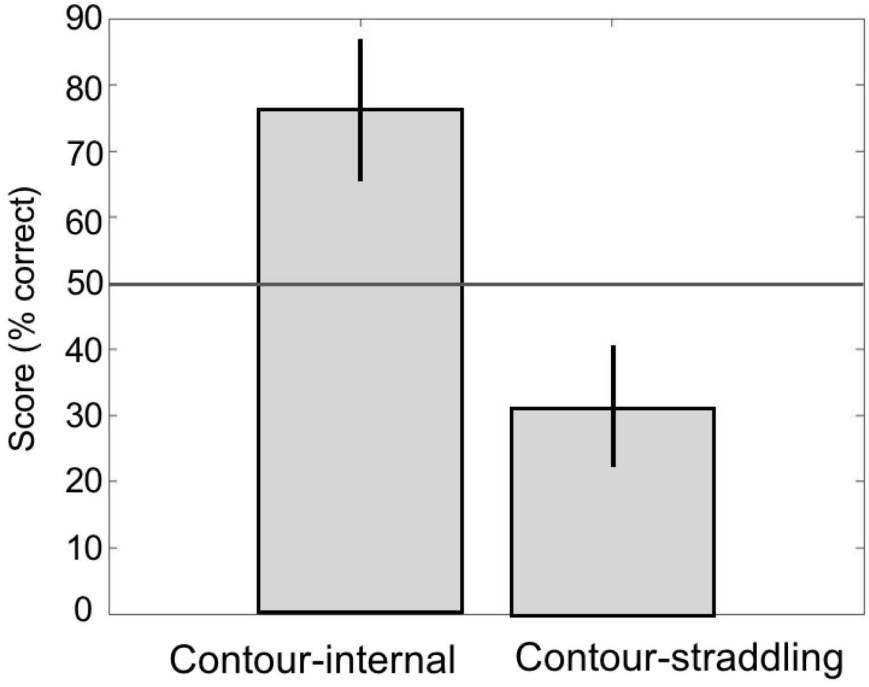


Figure 11.3: Results for Experiment 14. Internal 'words' are recognized better than chance, whereas straddling 'words' appear to be *rejected*.

they have never encountered to be possible 'words', compared to 'words' that actually occurred (but straddled a phrasal boundary).

These results provide very strong evidence for the model presented in Figure 8.1b. To recapitulate, the input is hypothesized to be analyzed in parallel by two systems. The first computes statistics over the (abstract) syllabic representations, while the other detects phrasal boundaries. In Experiment 5, we found the first piece of evidence that, at the syllabic level of representation, even contour-straddling 'words' are extracted since they have high TPs between the constituent syllables. In this experiment, when the memory trace for the syllabic level was preferentially strengthened by visual presentation of the test phase, both contour-internal and contour-straddling 'words' were preferred over non-words.

Thus, we can explain the results of this experiment as follows: finding that the contour-straddling 'words' are significantly rejected implies that (a) they are actually recognized and (b) they are judged as not being likely lexical items. That

is, on the one hand we proposed that *all* high-TP sequences are recovered. On the other hand, when such sequences are associated in memory with the prosody they bore during familiarization, those that contain a prosodic edge are filtered. Clearly, if the filtering effect is sufficiently large, we *expect* that contour-stradding 'words' are actually rejected.

So, why is the filtering effect most potent in this particular experiment? Of all the experiments that show the filtering effect, this is the only one with both of the following properties:

1. The use of Japanese IP characteristics.

2. No differences in phoneme durations across all the positions.

As noted before (see Figure 10.1, pg. 105), Japanese IPs show a larger difference in pitch levels between the beginning and the end of IPs in our corpus. Thus, 1 suggests that a greater pitch decline-reset might be associated with more potent filtering effects.

As for 2, recall from Experiment 12 above that final lengthening can be construed as lexical stress, causing contour-straddling 'words' to be judged as possible lexical items. Thus, final lengthening might actually work *against* the pitch patterns in these experiments. Therefore, since there are no durational differences in this experiment, the filtering effect of pitch is most clearly seen.

Notice that, with Japanese IPs, even when both pitch and duration differences were present, there was a statistical tendency for the contour-straddling 'words' to be rejected (Experiment 10, pg. 104). Taken with the findings from this experiment, we can hypothesize that the variations in pitch which accompany IPs are potent cues that define boundaries of 'phrases' in these experiments. Comparing the results using Italian IPs with those obtained with Japanese IPs, we can conclude that the strength of prosodic edges might be a function of the magnitude of the acoustic 'break'. Thus, the Japanese IPs in our corpus, which have larger pitch reset values at 'phrasal' boundaries show greater filtering effects than do the Italian IPs.

Recall that IPs are thought to be based on physiological mechanisms (Chapter 2). Might even the perception of an IP be a physiological response? That is, since we find evidence that an acoustic variable, the extent of the pitch reset at

the 'phrasal' boundary, determines the strength of the filtering effect, we might conclude that it is the acoustic break itself, rather than a prosodic boundary, that causes the filtering effect. We will test this hypothesis in the next experiment, the final experiment in this thesis.

11.4 Experiment 15: 'Filtering' by time-reversed IPs

We would like to test the hypothesis that the vital ingredient that generates the filtering effect of prosody is the perception of acoustic breaks, aligned with the edges of prosodic constituents.

In order to test the hypothesis, the previous experiment was re-run, but with all the pitch contours reversed. This maintains acoustic breaks at 'phrasal' boundaries, although within each 'phrase' the pitch rises instead of declining, and the reset is from high to low pitch, instead of from low to high. Also, there are no duration differences.

11.4.1 Material and Methods

Participants

Twelve adults participated in this experiment (5 males and 7 females, mean age 22.6 years, range 19-33).

Materials

For this experiment, the starting point was the MBROLA file from Experiment 14, wherein only pitch information from the Japanese IPs was used. Starting from this file, each IP was reversed, keeping the same order of the phonemes. The resulting file was converted into a sound file using the es1 Spanish male database as before.

Apparatus and Procedure

These were identical to the previous experiments.

11.4.2 Results

The results from this experiment are shown in Figure 11.4. Overall segmentation

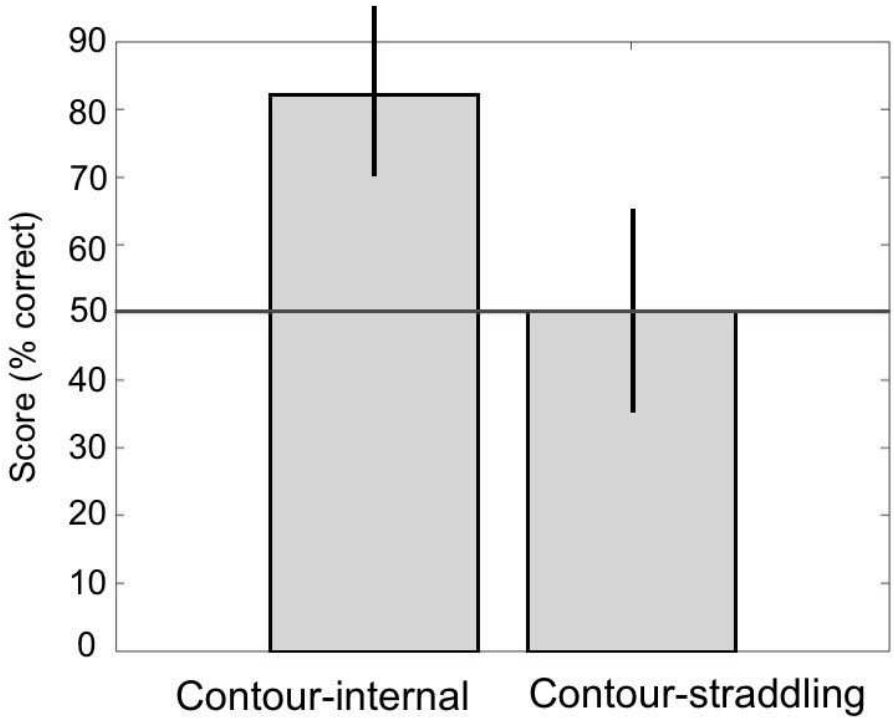


Figure 11.4: Results for Experiment 15. Even when the Japanese pitch contours are reversed, only internal 'words' are correctly segmented, and are different from the straddling 'words', as in Experiment 14. However, the straddling 'words' are not significantly rejected, as in the previous experiment.

was attested by a significant score of 66.15% (S.D. 18.74), $t(11) = 2.98, p < 0.02$. Internal 'words' had a score of 82.29% (S.D. 19.55), $t(11) = 5.72, p < 0.001$, and straddling 'words' had a score of 50% (S.D. 23.84), $t(11) = 0, p = 1$. The two 'word' types were significantly different from each other, $t(22) = 3.63, p < 0.002$.

An ANOVA compared the results from this experiment with reversed Japanese pitch contours with the previous experiment without reversed pitch contours. The main factors were Word Type (Internal or Straddling) and Pitch Contours (Normal or Reversed). There was a significant main effect of Word Type, $F(1, 22) = 76.25$, $p \leq 0.0001$, while the factor Pitch Contours was not significant, and nei-

ther was the interaction of the factors. A post-hoc Scheffe test confirmed that contour-internal 'words' were better recognized than contour-straddling 'words' $p \leq 0.0001$.

11.4.3 Discussion

We find a difference in the pattern of results when we compare this experiment (reversed Japanese IP pitch contours) with the previous Experiment 14 ('forward' Japanese IP pitch contours). In both experiments, contour-internal 'words' are preferred over non-words, and the scores for the contour-internal 'words' are significantly higher than those for the contour-straddling 'words'.

However, while in Experiment 14 the contour-straddling 'words' were significantly rejected, in the current experiment they are at chance. Although the ANOVA failed to detect a significant difference between the contour-straddling 'words' in the two conditions, it is interesting to note that only in the previous experiment do we observe a clear demonstration of the filtering effect of prosody: the significant rejection of contour-straddling 'words'. This finding suggests that acoustic cues that mark the edges of (forward) prosodic phrases might be more potent cues to edges than the same cues backwards either because (a) through experience they have come to be associated with phrases or (b) they reflect an asymmetry in general auditory processes.

Nevertheless, we can tentatively hypothesize that, rather than the prosodic breaks themselves, it is the fact that phrasal constituents are accompanied by significant acoustic events that mark the edges of such constituents that cause the perception of 'phrasal' boundaries and result in the filtering effect.

'It seems very pretty,' she said when she had finished it, 'but it's RATHER hard to understand!'

Through the looking glass,
Lewis Carroll

Chapter 12

General Discussion and Conclusions

How does an infant discover words from fluent speech? In the introductory chapters of this thesis, we saw that this is not a trivial task; nevertheless, there are several cues to word boundaries in speech, and an important research question is how these cues are extracted and utilized. We saw that distributional properties can aid in finding multisyllabic 'units' in speech (Chapter 3).

We also saw that speech is best described as a series of hierarchically nested *prosodic phrases*, rather than as a series of words. Moreover, morphology plays a crucial role in the construction of the phonological word, an intermediate prosodic constituent. As a result, the boundaries of larger (phrasal) prosodic constituents are also word boundaries. Therefore, the prosodic structure of spoken language itself provides word boundaries in fluent speech (see Chapter 2).

In this thesis, we have explored whether and how multiple cues can come together to ease the segmentation problem. Clearly, utilizing information from several sources ought to vastly simplify speech segmentation. Indeed, several studies have shown that various sources of information interact in signaling word boundaries even in infants (e.g., Mattys & Jusczyk, 2001). Yet, we lack explicit models of how such interactions might come about.

Thus, we set out to develop such a model. The model was developed by examining the response of adult participants exposed to carefully controlled artificial speech streams. The results of the empirical investigations with adult participants,

as described in the previous part, can be gathered together into four (inter-related) points:

Designing an experimental paradigm that reveals an interaction between TP computations and phrasal prosody. In Chapter 5, we examined some possible algorithms to compute TPs. Based on these, we predicted and observed (Chapter 6) that high-TP trisyllabic 'words' can be extracted even when they are embedded in *syllabic noise*: randomly interspersed, equifrequent syllables other than those that make up the 'words' . This observation allowed the creation of artificial speech streams wherein we could examine the extraction of trisyllabic nonce 'words' in various positions with respect to intonational 'phrases'. We saw that 'words' that straddled 'phrase' boundaries were not recognized, while those that lay inside such 'phrases' were (Chapter 6). Further, in Chapter 7, we found that 'words' at the edges of such 'phrases' were better recognized than 'words' in their middles. In sum, we found good evidence that prosody influences the extraction of words from fluent speech.

Providing evidence that a novel prosody can constrain statistical computations. The finding that prosody constrains statistical computations is of great importance for acquisition. There is evidence that pre-lexical infants are sensitive both to prosody and to distributional regularities. Therefore, a model of the interaction of these sources of information, as developed in this thesis, is of relevance even to infants learning their native language. In order to confirm that the effects of prosody observed with adults were not due to a sensitivity to learnt, language-specific cues, the effects of a prosody that the participants had never heard before was tested. We found that Italian adult participants used prosodic cues from Japanese IPs just as they used such cues from Italian IPs (Chapter 10). These results indicate that universal properties of IPs, such as the pitch decline, can be used to discover phrases in fluent speech, and thus might be of relevance even in pre-lexical infants.

Understanding the role of acoustic cues. In Chapter 11, we saw that the perception of 'phrases', as indexed by the failure to recognize contour-straddling

'words', might derive from general principles of auditory scene analysis, that are likely to be derived from the physiology of audition. That is, rather than only prosodic units, the pitch contours of IPs can also be perceived as purely acoustic 'units'. Nevertheless, while the global pitch patterns that define 'phrases' might be derived from general audition, we found in Chapter 9 that the filtering effect of prosody cannot be attributed to the acoustic differences between the 'word' tokens during familiarization and during test.

Developing a model of the cognitive basis for the interaction between prosody and statistical computations. In Chapter 8, we asked *how* the interaction between prosody and statistical computations comes about. We considered two possibilities. The first is that prosody defines the domain over which TPs are computed. According to this view, contour-straddling 'words' are never recovered. The second possibility is that TP computations are unaffected by prosody, so *all* 'words' are recovered; prosody plays a role at a later stage and suppresses contour-straddling 'words' (Figure 12.1 on the following page). We found evidence for the second possibility, suggesting that prosody acts as a *filter* that excludes contour-straddling sequences from being considered as possible lexical candidates. In the following sections, we will examine this model and its implications in greater detail.

12.1 The central, Prosodic Filtering model

In his book, *The Ghost in the Machine* (1967), Arthur Koestler writes in the concluding summary of the chapter "A Memory for Forgetting":

> "... we must assume the existence of multiple, interlocking hierarchies of perception which provide the multidimensionality or multicolouration of experience. In the process of storing memories each hierarchy strips down the input to bare essentials, according to its own criteria of significance.
>
> *Recalling* the experience requires dressing it up again. This is made possible, up to a point, by the co-operation of the hierarchies con-

cerned, each of which contributes those factors which it has deemed worthy preserving. The process is comparable to the superposition of colour-plates in printing – or of the wallpaper-maker's several stencils."

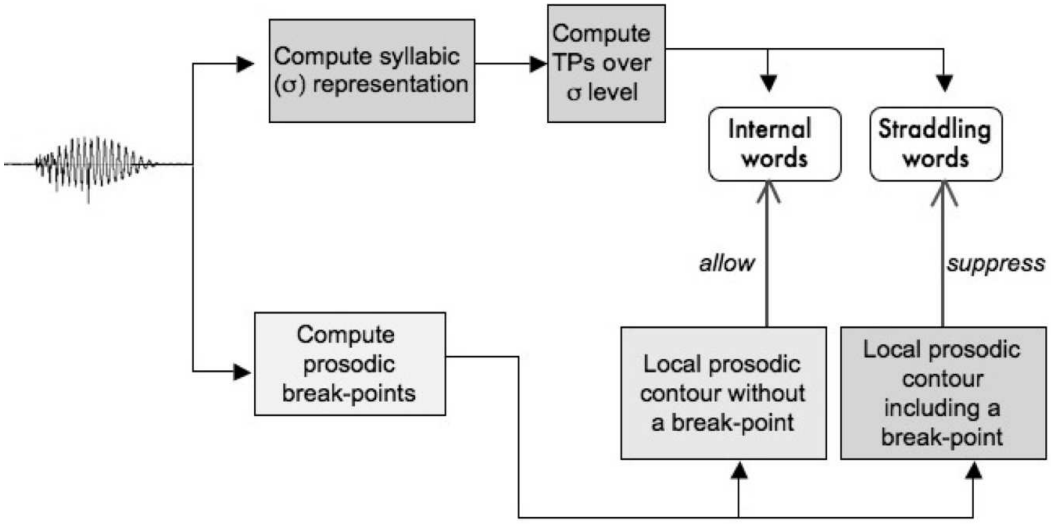


Figure 12.1: The central, prosodic filtering model proposed in this thesis (possibility *b* in Figure 8.1 on page 78). TPs are computed over the syllabic representation. Prosodic breaks are computed in parallel. The two are linked in episodic memory; 'words' misaligned with prosodic break-points are rejected.

Koestler's words reflect the central model in this thesis, that different computational mechanisms (*hierarchies*, in Koestler's sense) analyze the input according to their own criteria of significance. Thus, TPs are computed over the syllabic representation of the speech stream even as the prosody/acoustics suggest 'phrasal' groupings. The output of the different mechanisms are superimposed: statistically coherent syllabic sequences are aligned with prosodic groupings, such that only those coherent syllabic sequences that lie within prosodic domains are considered as possible word candidates.

Let us examine the model in two directions. In the first, the input is processed along separate processing streams. In the second, information from the separate streams is brought together.

12.1.1 Separate processing streams

The first important feature of the model in Figure 12.1 is the separation of TP computations and the perception of prosodic break-points.

Neuroscience is no stranger to the hypothesis that a single percept can be analyzed in parallel along distinct dimensions (e.g., Goodale & Milner, 1992). For example, in humans, the spatial information and the identity of objects (or sound patterns) has been proposed to be processed by two parallel streams both in vision (e.g., Haxby et al., 1991) and in audition (e.g., Alain, Arnott, Hevenor, Graham, & Grady, 2001).

Similarly, for speech, it has been proposed that, while temporal features that underlie lexical representations are processed preferentially by the left hemisphere, emotional prosody is preferentially processed by the right hemisphere (e.g., Pell, 1999; Hickok & Poeppel, 2000; Giraud et al., 2000; Zatorre & Belin, 2001; Blake, 2003). For example, recently, Boemio, Fromm, Braun, and Poeppel (2005) showed that sounds are analyzed at a faster timescale in the left hemisphere as compared to the right hemisphere, supporting the distinction between lexical analyses (that rely on fast auditory transitions) and prosody (that is encoded over a larger timescale, see also Benson and Zaidel, 1985).

Indeed, various studies have suggested that words are processed in the left hemisphere (see Démonet, Thierry, & Cardebat, 2005, for a recent review). Further, there is evidence that the units of words, the syllables, are perceived in the left hemisphere (e.g., Poldrack et al., 1999; Siok, Jin, Fletcher, & Tan, 2003). Coupled with the finding that prosody is represented in the right hemisphere, we thus find empirical support for the proposed separation between the perception of speech as a sequence of syllables and as a sequence of prosodic units.

Note that the precise involvement of the two hemispheres in various aspects of language perception and production are far from clear. Indeed, children with hemispherectomies can nevertheless master many, if not all aspects of language (e.g., Vargha-Khadem et al., 1997, Bates, Vicari, and Tauner, 1999, but see Curtiss and Bode, 2003). What is of importance is that, in the normal population, functionally different linguistic competences involve different brain areas, suggesting separate parallel processing. That these competences can be supported by

other brain tissue is an orthogonal question.

Let us now turn to the question of what goes on in the two processing streams.

TP computations

We hypothesized that, in one processing stream, TPs are computed independent of prosody (Chapter 8).

TP computations over speech streams have been demonstrated in human adults and infants (Saffran et al., 1996, 1996; Peña et al., 2002), monkeys (Hauser, Newport, & Aslin, 2001) and even in rats (Toro & Trobalón, 2005). Further, Saffran et al. (1997) showed that both adults and children can segment speech streams using TPs even in an incidental learning paradigm. While these results suggest that TP computations happen automatically, Toro, Sinnett, and Soto-Faraco (2005) showed that when attention is diverted away from the speech stream (in human participants), TP computations can be suppressed. Thus, we can conclude that, given sufficient attentional resources, the vertebrate brain automatically computes TPs.

What is the unit over which TPs are computed? The most general learning strategy would be to compute TPs over *all* units (features, phonemes, syllables). Indeed, Newport and Aslin (2004) demonstrated that under certain conditions, adult participants can extract dependencies over either the consonants or the vowels. Subsequently, Bonatti et al. (2005) showed that TPs are computed *preferentially* over consonants rather than over vowels. These investigations need to be extended to infants as well. Further, it still remains to be seen if TPs can be computed over different representations, for example over features (like voicing, place of articulation, etc.).

In this thesis, we have considered the syllable as the unit over which TPs are computed. Indeed, we know from psycholinguistic investigations with prelexical infants that they can represent syllables (e.g., Bertoncini & Mehler, 1981b; Bertoncini et al., 1987; Bijeljac-Babic et al., 1993). Such studies show that by 2 months of age, infants can detect a change in a syllable when either the vowel or the consonant changes. Thus, syllables might represent the predominant, initial unit over which TPs are computed, although further work is necessary to clarify this point. What is important for the prosodic filtering model is the fact that TPs

are computed over phonological units (see below).

Detecting prosodic groupings

In the other processing stream, we hypothesize a mechanism to detect prosodic break-points.

In Chapter 3, we examined previous evidence that infants can detect phonological phrases (e.g., Christophe et al., 1994; Gout et al., 2004; Soderstrom et al., 2003) and intonational phrases (e.g., Hirsh-Pasek et al., 1987; Jusczyk et al., 1992) in fluent speech. What drives the perception of larger prosodic constituents?

We saw in Experiment 15 on page 122 (Chapter 11) that backwards prosody showed the filtering effect: contour-straddling 'words' were not recognized, while contour-internal 'words' were recognized. Thus, the perception of 'phrases' in speech might rely on general auditory principles, which imposes acoustic break-points at abrupt pitch changes. Indeed, as noted by Bregman, "... [acoustic] units are formed whenever a region of sound has uniform properties and boundaries are formed whenever properties change quickly" (Bregman, 1983/1990, pg. 72).

Experiments with infants have revealed that infants can organize non-linguistic sounds in a manner reminiscent of how phrases are organized. Krumhansl and Jusczyk (1990) and Jusczyk and Krumhansl (1993) used a pause-detection paradigm with $4\frac{1}{2}$- and 6-month-olds and showed that even the younger infants perceived musical phrases as being defined by a pitch decline-reset at phrase boundaries and by a relatively longer final tone.

Thus, we see that a unitary speech stream can be separately analyzed for high-TP syllabic sequences and for prosodic break-points. How are these two sources of information put together?

12.1.2 Reconstructing the input

In Chapter 8 we appealed to the episodic memory system as a mechanism for putting together distributional and prosodic information (see Section 8.2.3, pg. 86). Recall from Experiment 3 (pg. 66) and Experiment 5 (pg. 80) that, while contour-straddling 'words' are not recognized when the test phase is in the auditory modal-

ity, they, as well as the contour-internal 'words', are recognized with the visual test phase.

According to the episodic memory hypothesis, the acoustic modality of the test items provides an appropriate context for the recall of their acoustic characteristics during familiarization. We saw in Chapter 9 that the precise acoustic shape does not appear to contribute to the recall of the test items. Instead, we hypothesize that the presence or absence of an acoustic/prosodic break is recalled. 'Words' misaligned with such breaks are rejected as possible lexical candidates.

The visual modality, in contrast, does not provide an appropriate context for the recall of acoustic characteristics. Instead, the phonological representation of the test items predominates. If, as we proposed, distributional analyses are carried out over such a phonological level, we expect that all high-TP syllable sequences are recalled, which is what we find in Experiment 5.

In sum we find that, while distributional analyses might find several high-TP multisyllabic sequences, only those that are in prosodically appropriate contexts are considered as possible lexical items.

12.2 Implications for acquisition

The prosodic filtering model developed in this thesis is aimed at understanding how multiple sources of information can plausibly contribute to speech segmentation in pre-lexical infants. Thus, we now examine a framework within which we can understand the implications of the model.

Several authors have considered the possibility that many or all of the computations that make up the capacity for language are present in infants just as in adults (e.g., Fodor, 1981). According to this *continuity hypothesis*, infants and adults share the same cognitive capacities (e.g., Gillette, Gleitman, Gleitman, & Lederer, 1999). Is the continuity hypothesis valid even for speech segmentation?

In the introductory chapters, we saw that speech segmentation can be lexically driven (explicit segmentation), or achieved using sub-lexical cues (implicit segmentation). The continuity hypothesis suggests that all cues are available at

all ages, while their relative importance changes depending on the amount of exposure to spoken language.

Such a view is implicit in Mattys, White, and Melhorn (2005), who suggest a hierarchical organization of segmentation cues, as shown in Figure 12.2

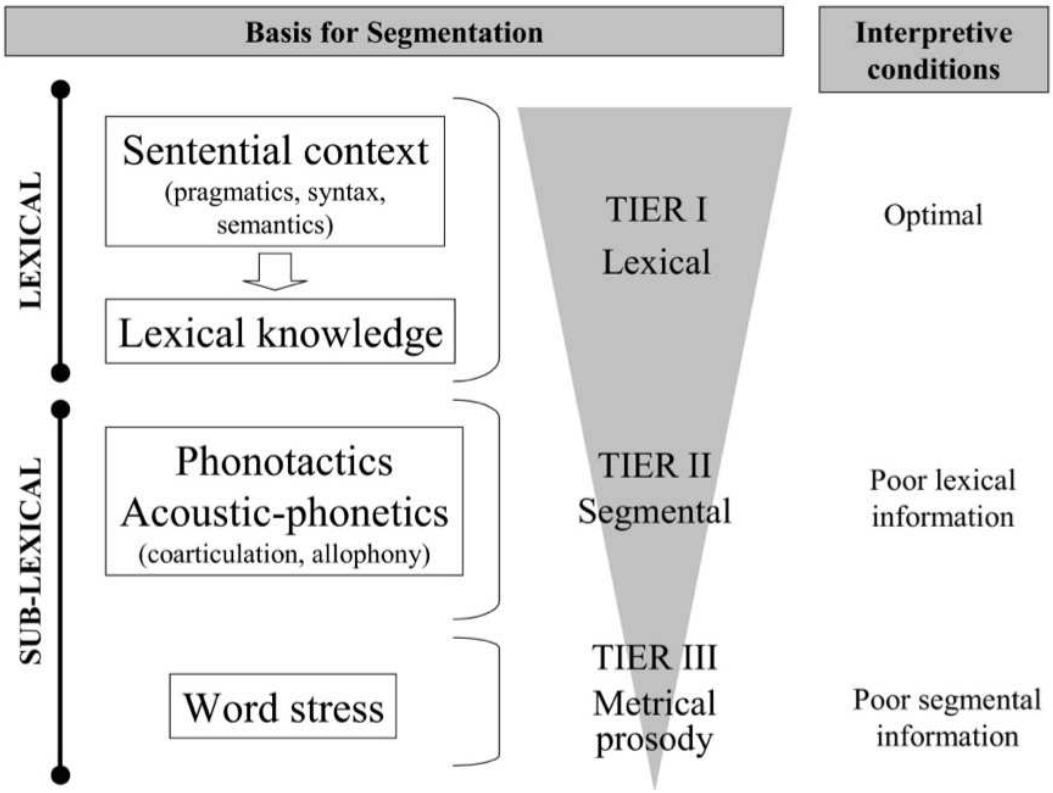


Figure 12.2: A hierarchical model of speech segmentation, taken from Mattys, White & Melhorn, 2005, with kind permission.

In the hierarchical model, the various cues to segmentation are assigned different weights depending on the listening conditions. For example, for an adult with a rich lexicon listening to clearly enunciated speech on a well-established theme, explicit (lexical) segmentation will dominate. In contrast, under poor listening conditions, sub-lexical cues like phonotactics or stress patterns will determine segmentation.

Clearly, Tier I (lexical) cues are unavailable to infants as they lack a substantial lexicon. However, recent evidence suggests that infant can use their meagre lexicons in segmenting speech: Bortfeld, Morgan, Golinkoff, and Rathbun (2005)

showed that 6-month-olds can use their own names and common, frequent words (like *Mommy*) to segment and recognize words following the familiar word. What about sub-lexical cues?

In the hierarchical model, sub-lexical cues like phonotactics and word stress constitute Tiers II and III. In the introductory chapters, we had grouped the cues that constitute Tiers II and III in the hierarchical model as 'statistical' cues. Such cues must be derived from the input. Indeed, several lines of evidence indicate that by $1\frac{1}{2}$-years of age, infants have acquired several of the sub-lexical segmentation cues such as lexical stress, allophones and phonotactics (e.g., Friederici & Wessels, 1993; Mattys et al., 1999; Jusczyk et al., 1999; Weber, Hahne, Friedrich, & Friederici, 2004).

Mattys et al. (2005) propose a developmental account wherein distributional strategies are utilized to acquire Tier II and Tier III cues. For example, Thiessen and Saffran (2003) find evidence that while younger ($7\frac{1}{2}$-month-olds) prefer statistically coherent nonce words, by 9 months of age, infants prefer nonce words that respect the strong-weak stress pattern of English (see Section 4.1, Chapter 4).

Let us therefore introduce a Tier IV to the hierarchical model in Figure 12.2. Distributional cues (like TPs), which are independent of the specific language, would count as Tier IV cues, and bootstrap the acquisition of Tier II and Tier III cues. In addition, in this thesis, we propose yet another Tier IV cue: the detection of phrases in fluent speech. Thus, Tier IV includes cues that are language-universal. Both distributional and prosodic cues can operate independent of the specific language. However, the two cues provide different kinds of information. While distributional information provides possible word candidates, prosodic information merely restricts this candidate list.

Thus, we might think of the interaction between prosody and statistics as the *constraint that prosody places on the output of statistical computations*. In the most general sense, we can thus view the interaction between various cognitive processes as the constraints that one process imposes on another. Let us look at the constraints on distributional strategies as a case in point. In doing so, we can consolidate the various empirical results obtained from previous studies and from this thesis.

12.2.1 Constraining distributional strategies in speech segmentation

We have seen that distributional strategies are general purpose mechanisms in segmenting speech. The key finding in this thesis is that prosody constrains distributional strategies. However, distributional strategies themselves might not be entirely general. As an example, we saw in Experiment 1 (pg. 52) that the spacing between 'words' that are otherwise (distributionally) identical influences their segmentation.

Thus, let us discriminate two kinds of constraints: those that are *internal* to the statistical computation system, and those that *external* to it.

Internal constraints on distributional strategies

Throughout the thesis, the terms 'statistical computations' and 'transition probabilities' have been used interchangeably. However, TPs capture only one kind of statistical regularity in the input. For example, several authors have suggested that 'chunk strengths' or the mutual information between mono-, bi- or tri-syllables might help extract words from fluent speech (e.g., Perruchet and Vinter, 1998, Swingley, 2005; see also Brent and Cartwright, 1996, Christiansen, Allen, and Seidenberg, 1998 for other possibilities). For example, in streams used in the experiments in this thesis, both TPs and the frequencies of tri-syllables would result in the 'words' being extracted. Thus, we might propose that the first constraint on distributional strategies is the *nature of the computation*.

Even if we assume that TPs over syllables constitute the appropriate statistical measure of the coherence of a multi-syllabic sequence, we saw in Chapter 5 that TPs might be more than computing mono- and bisyllable frequencies (see also Aslin et al., 1998). Thus, the second internal constraint is on the *implementation* of the distributional computation algorithms.

The third constraint is on the *units* over which statistics are computed. As discussed earlier, we assume that the unit is the syllable. However, other data suggests that, within syllables, TPs over the consonants contribute more to the identification of words than the TPs over vowels (Bonatti et al., 2005).

External constraints on distributional strategies

External constraints represent the interaction between the output of distributional strategies and other cognitive domains. Indeed, as mentioned earlier, the central model of the thesis, as presented in Figure 12.1 can be understood as a *constraint of prosody* on the output of TP computations.

The experiments wherein 'words' at the edges of IPs were contrasted with 'words' in their middles suggest a second constraint; one due to the *salience of edges* (Experiments 4 and 11). Thus, words at the edges of larger prosodic phrases gain a processing advantage due to their salient location.

We have argued that the IP, which is an instance of a clearly marked unit in speech, might be derived from general principles of auditory perception (Chapter 11). Thus, a third constraint comes from the *auditory perception of acoustic groups* (see also Creel, Newport, & Aslin, 2004).

Finally, *attention and memory* provide constraints from broad cognitive systems (Section 8.2.3, Chapter 8; see also Toro et al., 2005).

Taken together, the entire spectrum of internal and external constraints represents the complexity of distributional strategies in segmenting speech and their rich interactions with other cognitive systems.

Indeed, several researchers have proposed *bootstrapping* solutions to language acquisition[1], which involve the interaction of various cognitive systems. Bootstrapping thus provides an adequate framework to situate the findings in this thesis.

12.2.2 Bootstrapping

The general problem in acquisition is that a grammatical unit like a 'noun' or a 'verb phrase' or a 'subject' is not marked as such in the input (e.g., Pinker, 1995). A similar problem is faced in learning the meanings of words – the referents of words are not clearly marked in the input (e.g., Quine, 1960). Likewise, as we saw in the introductory chapters, words themselves are not clearly marked in fluent speech.

[1]Language acquisition is likened to trying to pull oneself up by the bootstraps.

Bootstrapping accounts of acquisition describe learning as a probabilistic, multiple-cue driven learning process. Importantly, learning in one domain uses multiple cues from different domains. For example, semantic regularities can drive the acquisition of grammatical categories (Semantic Bootstrapping, e.g. Pinker, 1984, 1995), syntactic frames can be used to constrain the meanings of words (Syntactic Bootstrapping, e.g. Gleitman, Cassidy, Nappa, Papafragou, & Trueswell, 2005), the number of noun phrases can be used to infer the transitivity of verbs (e.g., Lidz, Gleitman, & Gleitman, 2003; Lidz & Gleitman, 2004a), and prosody can be used to infer grammatical categories like open- and closed-class (lexical) items (Phonological Bootstrapping, e.g. Morgan, Shi, & Allopenna, 1996).

Second, learning is *probabilistic* because cues from one domain do not precisely map onto what is to be learnt in another domain. For example, although function words are typically phonologically 'weak', this is not true of all function words within or across languages (Morgan et al., 1996). Similarly, while the number of noun phrases indicate the argument number, they are not in a one-to-one correspondence (Lidz & Gleitman, 2004b).

From this thesis, it is clear that strategies for speech segmentation too can be seen as bootstrapping solutions. Finding a word boundary makes use of information from other domains. In this thesis, we examined the contribution from phrasal prosody.

12.3 Conclusions

Language has been described as the last evolutionary transition that life on earth has witnessed (Szathmary & Smith, 1995). Empirical and theoretical advances in the last century have established language as a complex cognitive process that relies on a conglomerate of disparate cognitive capacities. Such a complex system is acquired through mere exposure by the time a child turns three. This suggests that the process of acquisition is bootstrapped by cues from different domains acting in concert, such that the learning path is narrow and constrained.

In this thesis, we have explored solutions to the problem of speech segmentation. In particular, we found that, while TPs between syllables extract 'words' from fluent speech streams, these 'words' are filtered by prosody. Thus, prosody

places a constraint on distributional strategies. We proposed that this interaction is mediated by episodic memory processes.

Further, we saw that prosody carves fluent speech into a series of phrases. The identification of such phrases relies on principles of auditory groupings, and a consequence of identifying such phrases is the enhanced processing of their edges, which are salient positions.

Thus, we find that a single aspect of acquiring language, segmenting words from fluent speech, involves the rich interplay of various cognitive processes.

Part III

Annexe

'I only wish *I* had such eyes,' the King remarked in a fretful tone. 'To be able to see Nobody! And at that distance, too! Why, it's as much as *I* can do to see real people, by this light!'

Through the looking glass,
Lewis Carroll

Chapter 13

Neonate perception of speech

In the introductory Chapter 1 we saw that the perceptual world of the neonate is not a disorganized chaos. Instead, neonates display richly structured initial biases (e.g., Mehler & Dupoux, 1994; Gopnik et al., 1999). In this annex, we will look at an example of such an initial bias, the representation of speech in neonates. We will examine, using a recently developed Near Infrared Spectroscopy (NIRS) system, whether the infant brain responds differentially to speech utterances versus non-speech. The non-speech stimuli used in this experiment are the same speech utterances, but played backwards, which are thus controlled for a variety of incidental acoustic properties.

13.1 Infant perception of speech

In the realm of language, it has been known that prosodic aspects of speech can be perceived by the fetus, such that they prefer their mothers' voice to the voice of a (female) stranger in the womb (Kisilevsky et al., 2003). This information is retained in neonates, such that they prefer their mothers' voice to that of a stranger ex-utero (DeCasper & Fifer, 1980). Indeed, neonates are capable of preferentially honing in onto speech sounds in their acoustic input, thus presumably privileging speech processing right from the start (e.g., Mehler et al., 1988).

Behavioral evidence suggests that speech sounds are special to the young infant. For instance, infants prefer speech to a variety of environmental sounds (Colombo & Bundy, 1983) and to speech attenuated below 3.5 kHz or contin-

uous, repetitive stimuli like heartbeats or even short speech phrases presented at the rate of heartbeats ("heartspeech"; Ecklund-Flores & Turkewitz, 1996). Strikingly, neonates even prefer speech utterances to the same utterances run backwards (Mehler et al., 1988). Not only do infants prefer *forward* speech to *backward* speech, but they can also discriminate languages belonging to different *rhythmic classes* only when they hear forward speech in these languages (Ramus et al., 2000; see also Chapter 1).

Thus, speech contains characteristics that induce infants to process it preferentially. This has led to the hypothesis that the neonate comes equipped with an apparatus that is dedicated to language from the outset (e.g., Mehler & Dupoux, 1994; Pinker, 1994). It is well known that in adults, language is principally processed in the left hemisphere (LH, e.g., Broca, 1861; Dehaene et al., 1997; Kim, Relkin, Kyoung-Min, & Hirsch, 1997). This functional lateralization of language is accompanied by an anatomical difference in LH and RH auditory cortical areas, at the level of the planum temporale, in both infants and adults (e.g., Wada, Clarke, & Hamm, 1975). Could it be, as some authors suggest, that language (or speech) is preferentially processed by the LH? In support of such a view, it has been observed that neonates display a right ear (LH) advantage for speech and a left ear advantage for music (e.g., Glanville, Best, and Levenson, 1977, Bertoncini et al., 1989, Segalowitz and Chapman, 1980; but see Best, Hoffman, and Glanville, 1982).

A slightly different viewpoint is that the speech rides piggyback on general acoustic processing. In support of such a view, Tincoff et al. (2005) found that, like human infants, cotton-top tamarin monkeys discriminate speech from languages belonging to different rhythmic classes only when exposed to forward and not to backward utterances. These results suggest that the preferential processing of forward speech in infants arises from pre-existing auditory capacities common to the primate lineage, and is not specific to language. General acoustic processing itself might shows hemispheric asymmetries. For example, anatomical asymmetries have been proposed to underly better processing of fast temporal events by the LH as compared to the RH (e.g., Zatorre, Belin, & Penhune, 2002). Indeed, understanding fluent speech requires the processing of phonetic information that occurs at a relatively fast time scale. However, how this is related to the neonate

preference for speech is not clear, since neonates primarily have access to prosodic features of speech, which occur over slower timescales (Moon & Fifer, 2000).

Most of the previous results were obtained by behavioral methods like high-amplitude sucking (e.g., Bertoncini et al., 1989) or foot-kicking responses (e.g., Segalowitz & Chapman, 1980). More recently, imaging methods have been developed to supplement such behavioral data and provide converging evidence for the preferential processing of speech and hemispheric lateralization in neonates. Such methods include high-density electrophysiology (e.g., Dehaene-Lambertz, 2000), functional magnetic resonance imaging (fMRI, e.g., Dehaene-Lambertz et al., 2002), NIRS (e.g., Peña et al., 2003) and magnetoencephalography (MEG, Kujala et al., 2004; Imada et al., 2006). Of these techniques, NIRS is the most practical and non-invasive technique, and has been used in an increasing number of experimental studies of the infant's brain and cognition (e.g., Baird et al., 2002; Peña et al., 2003; Taga, Asakawa, Maki, Konishi, & Koizumi, 2003; Wilcox, Bortfeld, Woods, Wruck, & Boas, 2005; Bortfeld, Wruck, & Boas, 2006).

13.1.1 NIRS

The NIRS technique relies on the differential absorption of near-infrared (NIR) light by human brain tissue. NIR light incident on the skull is reflected and absorbed to different extents by different brain tissue.

The change in intensity between the emitted and the recorded light can be related to neural activity. It has been observed that neural activity is accompanied by changes in hemodynamics, in particular an increase in the concentration of oxygenated hemoglobin, [Oxy-Hb][1], and a decrease in the concentration of deoxygenated hemoglobin, [Deoxy-Hb] (e.g., Obrig et al., 1996; Gratton, Goodman-Wood, & Fabiani, 2001). Thus, NIRS can be used to measure changes in cerebral blood oxygen saturation as an index of brain activation (e.g., Jobsis, 1977; Chance, Zhuang, Chu, Alter, & Lipton, 1993; Villringer & Chance, 1997; Strangman, Culver, Thompson, & Boas, 2002; Meek, 2002)

How do changes in light intensity indicate changes in concentration of [Oxy-Hb] and [Deoxy-Hb]? A beam of light passing through a medium is absorbed,

[1]Square brackets indicate concentrations.

reflected and refracted to different extents based on the properties of the medium. The extent to which light is absorbed is dependent both on the composition of the medium and the wavelength of the light. The *absorption coefficient*, ϵ is a measure of the relative absorbance of light for a particular medium at a particular wavelength. Figure 13.1 plots the ϵ for Oxy-Hb and Deoxy-Hb as a function of the wavelength of light. As can be seen from Figure 13.1, Oxy- and Deoxy-

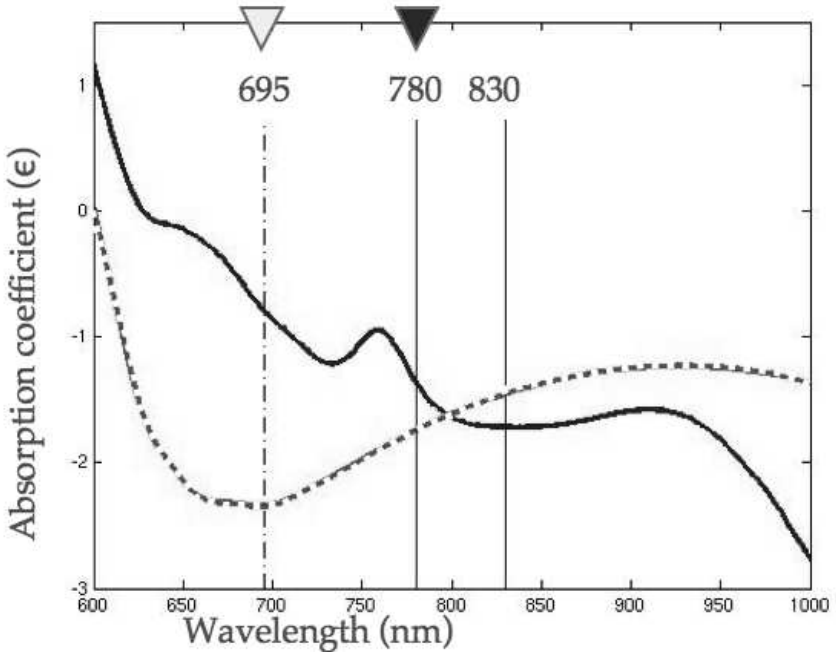


Figure 13.1: The absorption coefficients (ϵ) of oxygenated (broken line) and deoxygenated (solid line) species of hemoglobin as a a function of the wavelength of light. The vertical lines indicate the wavelengths used in this study and in Peña et al. (2003). In this study, the lower wavelength used was 695 nm (empty arrowhead), while in Peña et al. (2003), it was 780 nm (filled arrowhead). Both studies used an upper wavelength of 830 nm.

Hb absorb NIR light differentially at different wavelengths. Thus, measuring the change in intensity of NIR light at two different frequencies allows the simultaneous estimation of changes in both [Oxy-Hb] and [Deoxy-Hb].

This annex describes a NIRS study that attempts to replicate the results obtained by Peña et al. (2003), in which these authors examined the organization of the neonate brain for language.

13.2 Replicating Peña et al. (2003)

Peña et al. (2003) studied the organization of the neonate brain using the ETG-100 OT system. This device allows the simultaneous measurement of [Oxy-Hb] and [Deoxy-Hb] from 24 channels, organized in two sets of 12 each. In this study, 12 channels were placed on the scalp over peri-sylvian areas of the LH and the other 12 on symmetric locations of the RH.

Each 'channel' corresponds to the cortical path traversed by incident light from an emitter to a detector. The layout of the channels is shown in Figure 13.2. Each channel primarily captures the hemodynamic responses at the cortex 2-3

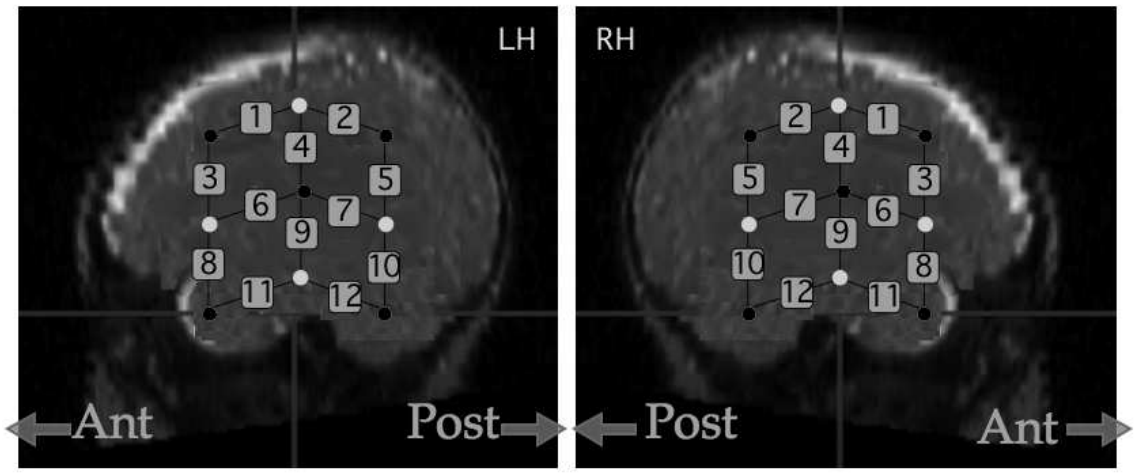


Figure 13.2: The placement of OT probes, overlaid on an infant MR scan, as used by Peña et al. (2003). Each number corresponds to a single 'channel', which marks the path between emitters (black dots) and detectors (grey dots). The vertical lines mark the vertex-tragus axis, while the horizontal lines mark the inion-nasion axis. Ant: Anterior, Post:Posterior.

cm below the scalp. The ETG-100 emits NIR light at 780 nm and 830 nm. The intensity of each wavelength is modulated at frequencies between 1 to 6.5 kHz, with a total power of 0.7 mW per channel. Solid-state lock-in amplifiers sample the reflected light at a frequency of 10Hz and separate the signals for the two wavelengths. These signals are stored along with 'marks' identifying time points corresponding to stimulation for later analysis.

Peña et al. (2003) tested neonates in three conditions: *forward* utterances in the maternal language, recorded by a female in a child-directed manner (FWD), the same utterances played *backwards* by digital reversal (BCK), and a silence condition (SIL). These authors found that in the FWD condition, the LH showed a greater increase in [Total-Hb][2]. The increase in [Total-Hb] in the FWD condition was greater in the LH as compared to the RH. Also, in the lower channels of the LH (corresponding to peri-Sylvian areas), [Total-Hb] was higher in the FWD condition as compared to BCK or to SIL. In particular, two channels close to the peri-sylvian areas (channels 9 and 11, see Figure 13.2), showed greater FWD-specific response in the LH as compared to the LH.

Thus, the results of Peña et al. (2003) show that the LH is attuned to speech in the brain of the neonate. Areas around the Sylvian fissure show greater activation to FWD as compared to BCK. These results provide converging evidence for the aforementioned behavioral findings that the neonate brain shows functional asymmetries in processing speech.

However, using fMRI, Dehaene-Lambertz et al. (2002) did not find clear lateralization of preferential speech processing in primary auditory areas. As discussed in Peña et al. (2003), various differences between the OT and fMRI studies must be highlighted, like the age of the infants (neonates vs 3-month-olds) or the use of the BOLD response, which is primarily driven by changes in [Deoxy-Hb], versus the changes in [Oxy-Hb] and [Deoxy-Hb] obtained in NIRS.

Given the theoretical and technical importance of the results in Peña et al. (2003), we decided to replicate the study, using a more recent version of the machine used previously.

13.2.1 Differences in the studies

Peña et al. (2003) used the ETG-100 (Hitachi Medical System). For the current study we used the more recent ETG-4000. The first difference between the machines is in the choice of the wavelengths. Both the studies used the same high wavelength of 830 nm. However, while the ETG-100 used a low wavelength of 780 nm, the ETG-4000 uses 695 nm. As shown in Figure 13.1, the separation of the

[2] [Total-Hb] is the sum of [Oxy-Hb] and [Deoxy-Hb].

ϵs for Oxy-Hb and Deoxy-Hb is larger at 695 nm than at 780 nm. Such a change in wavelength should improve the signal-to-noise ratio.

A second difference was the design of the probes. In the ETG-100, a silicon holder kept emitters and detectors in contact with the scalp of the neonate in a fixed geometry (as shown in Figure 13.2). All emitters and detectors could be attached independently to the silicon holder. In contrast, the probes of the ETG-4000 consist of the emitters and detectors embedded into thin, semi-rigid silicon strips arranged in a chevron shape. While this shape allows the easy placement of the probes behind the ear of the neonate, these probes are harder to place compared to those used in the previous study. While the structure of the probes differed between this experiment and the previous one, the relative placement of the emitters and detectors was comparable. Further, in both cases, the spacing between the emitters and detectors was 3 cm.

However, due to difficulties in placing the probes, and due to the fact that in the previous study a trained pediatrician (M. Peña) placed the probes on the neonates, the arrangement of the probes on the heads of the neonate differed slightly between the two experiments. With respect to the layout in Peña et al. (2003) (see Figure 13.2), in the current study the probes were slightly more dorsal, medial and inclined away from the vertex-tragus axis, towards the nasion, as shown in Figure 13.4.

Another source of difference was in the power output in the channels. Peña et al. (2003) used 0.7 mW per channel. In this study we used longer optical fibers (5 m) compared to the previous (3 m). Due to attenuation in the fiber optic cables, the net power output was 0.43 mW per channel.

13.3 Experiment

Here we used the same procedure as in Peña et al. (2003), with three conditions, FWD, BCK and SIL, whose order was balanced across participants, as in the previous study.

13.3.1 Material and Methods

Participants

Twenty-five full-term healthy Italian neonates ranging from 1 - 6 days were tested. All infants had an APGAR score ≥ 9 at 1 and 3 minutes after birth. All neonates were tested as they slept in their cribs in a quiet room at the Santa Maria della Misericordia hospital in Udine, Italy. The ethics committee of the University of Udine granted permission for the experiments. Parents received all relevant information, and signed a consent form.

Materials

The material was the same as that used in Peña et al. (2003). Two Italian mothers recorded utterances in an infant-directed style. Ten such utterances, with a mean duration of 15 sec ($\pm$1 sec) and (root-mean-squared) intensity equalized, were used in the *forward* (FWD) condition. The same utterances were digitally reversed to create the utterances for the *backward* (BCK) condition. In the SIL condition, the FWD sentence set was used, but the sound output was digitally set to zero.

Apparatus

We used the ETG-4000 OT system (Hitachi Medical) to measure changes in [Oxy-Hb] and [Deoxy-Hb] in sleeping neonates. The entire experiment was run by PRESENTATION (Neurobehavioral Systems, Inc., CA, USA) on a computer running Windows^{TM}2000. PRESENTATION delivered the stimuli via a sound card (SoundBlaster Live! from Creative Technology Ltd.) on the PC. The audio stimuli were delivered at comfortable levels through SoundSticks (Harman/Kardon), consisting of two 10-watt tweeters at the level of the neonate and one 20-watt subwoofer placed below the crib. PRESENTATION also sent timing signals that marked the onset and offset of the stimuli to the ETG-4000, over the serial port.

Procedure

The infants were tested individually in their cribs. The two silicon strips (probes) containing the emitters and detectors, corresponding to 12 'channel' each, were placed along the vertex-tragus line as shown in Figure 13.2. As noted earlier, the relation of the probes to the external landmarks differed between the two experiments. Figure 13.4 shows the placement of the probes in this experiment (compare with Figure 13.2). The distance between the emitters and detectors was 3 cm, as in the previous study.

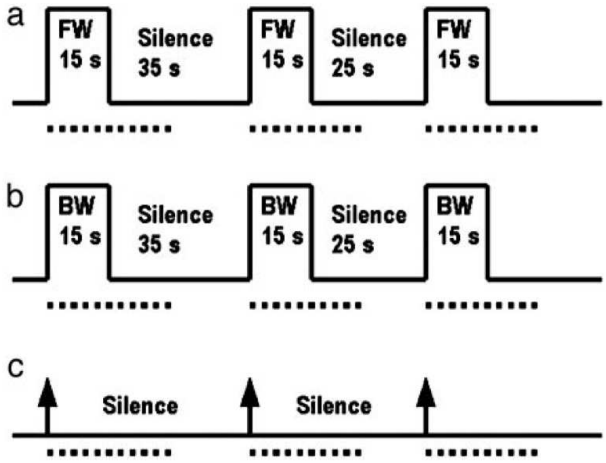


Figure 13.3: The testing protocol. Infants were exposed to three conditions, FWD (a), BCK (b) and SIL(c) in a random order. Each condition consisted of ten blocks (of which only three are shown in this figure). For (a) and (b), within each block, successive stimuli were separated by variable silent periods. The arrows in (c) indicate the onset of 15sec 'blocks' in the silent condition. Dotted lines indicate the period from each block that was considered during statistical analyses.

Infants were exposed to three conditions, FWD, BCK and SIL (see Figure 13.3). The order of the conditions was randomized across neonates. The FWD and BCK conditions consisted of 10 utterances of forward or backward utterances, separated by variable silence between 25 - 35 sec. The SIL condition was identical to the FWD, except the volume was set to zero. Thus, there was no stimulation, although the PC sent marks indicating silence onset and offset marks to the OT machine.

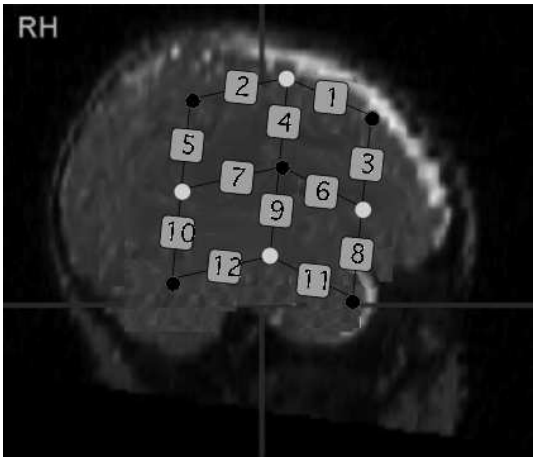


Figure 13.4: Probe placements in the current experiment. Notice the difference from the previous study (Figure 13.2).

Data analysis

The data analysis used the same procedure as was used in Peña et al. (2003). Briefly, the raw changes in absorbance at the two wavelengths registered by the ETG-4000 were converted into [Oxy-Hb] and [Deoxy-Hb] values. Since the ETG-4000 uses different wavelengths compared to the ETG-100 (Figure 13.1), we expected to find better signal-to-noise ratios. While the previous study looked at the variations in [Total-Hb] as the dependent measure, their values for [Total-Hb] corresponded closely to the values for [Oxy-Hb] due to the minor contribution from [Deoxy-Hb]. Indeed, [Oxy-Hb] alone replicates the results found with [Total-Hb] (M. Peña, pers. comm.). Thus, in this study, since we expected greater contribution from [Deoxy-Hb], we looked separately at the changes in [Oxy-Hb] and [Deoxy-Hb].

The raw time series recorded by the ETG-4000 for all the channels were band-pass filtered between 0.02 and 1 Hz to remove low-frequency components arising from heartbeat- and respiration-related cerebral blood flow changes, and high-frequency noise. Motion-related artifacts (signal variations > 0.1 mmol.mm) were detected and marked. Next, portions of the signal of duration 35 sec were extracted, corresponding to a period 5 sec prior to the onset of each stimulus and 30 seconds post-onset. For each such block that did not contain artifacts, a fit was computed between the first and the last 5 seconds. The principle dependent mea-

sure was the mean [Oxy-Hb] and [Deoxy-Hb] in the 25 sec period post stimulus onset. Thus, each infant contributed at most 1440 data points: 2 concentrations ([Oxy-Hb], [Deoxy-Hb]) $\times$ 3 conditions (FWD, BCK, SIL) $\times$ 24 channels $\times$ 10 blocks.

Statistical analyses were carried out using repeated measures ANOVAs, with DataDesk (DataDescription Inc.). The factors of interest were Condition (FWD, BCK, SIL), Hemisphere (LH, RH) and Position (Upper channels, Lower channels). The upper channels consisted of channels 1-6 in each hemisphere, while the lower channels were 7-12. The lower channels cover the temporal cortices close to the Sylvian fissure.

13.3.2 Results

Figure 13.5 shows the changes in [Oxy-Hb] for the 24 channels across the two hemispheres for each of the three conditions, FWD, BCK and SIL. Although we expected to find changes also in [Deoxy-Hb] as a result of the choice of wavelengths, the [Deoxy-Hb] data was very noisy and of too low an amplitude to reveal any effects.

We compared the values for [Oxy-Hb] across the three conditions and across the two hemispheres and the two positions. There was no main effect of any of the three factors, and all the interactions were non-significant (all $p > 0.2$). There were no significant results for [Deoxy-Hb].

In Peña et al. (2003), the authors found the most clear results for the lower channels 9 and 11. Thus, we decided to carry out a region-of-interest analysis for these channels using a repeated measures ANOVA, with Channels (9 and 11), Hemisphere (LH, RH) and Condition as fixed, within-subject factors. Looking at [Oxy-Hb] as before, we find a main effect of Condition, $F(2, 46) = 3.7, p = 0.032$, while the other factors were not significant. Post-hoc tests revealed that in the LH channels, FWD showed a greater activation than both BCK ($p = 0.0084$) and SIL ($p = 0.035$), while BCK and SIL were not different from each other. In the RH channels, a different pattern was observed: none of the conditions differed from each other (all $p > 0.35$).

From Figure 13.5, it can be seen that in addition to channels 9 and 11, there

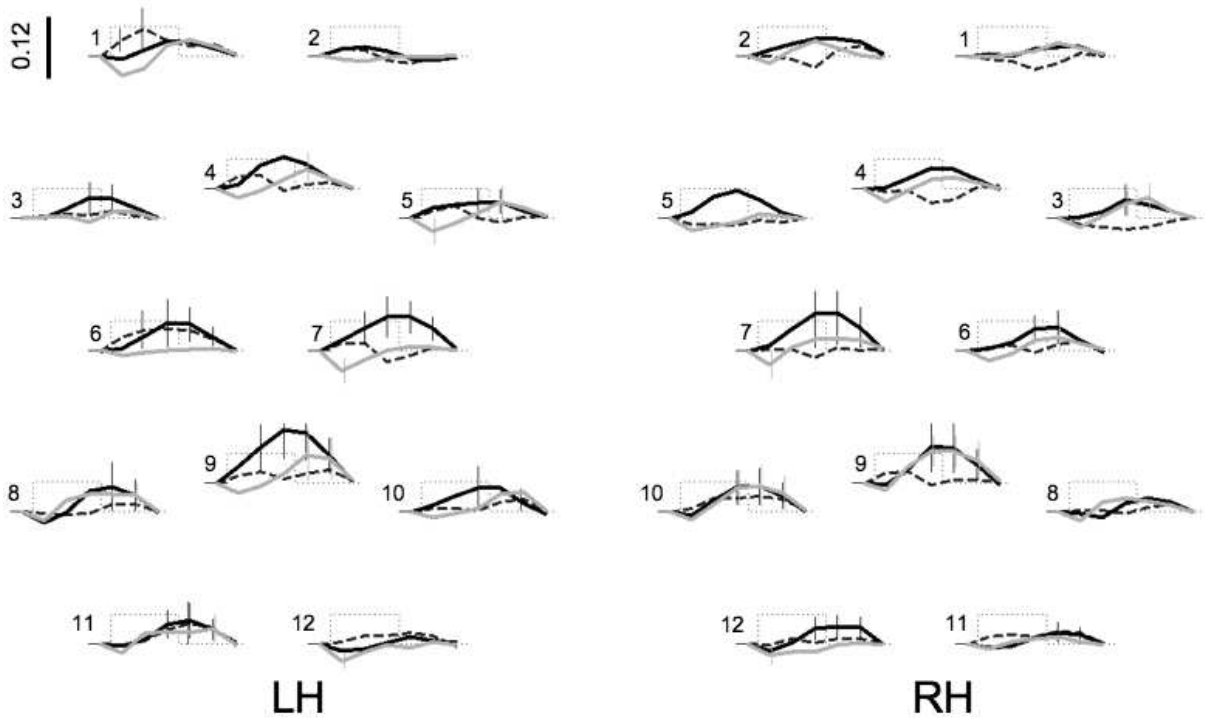


Figure 13.5: Values for [Oxy-Hb] for the three conditions FWD (black), BCK (grey) and SIL (broken). Subplots represent channels. Each curve is the grand average of the mean value from good blocks from 25 infants, for successive 5 sec windows. The first point is for the 5 sec window preceding stimulation. The dotted line marks the duration of the stimulation. Time points bearing vertical lines are significantly different from baseline; the vertical lines indicate 95% confidence limits of the mean in the 5 sec windows containing them. The vertical line on the top left of the figure indicates the scale (mmol.mm) for all the plots.

is activation also in channel 7. An ANOVA restricted to channel 7 reveals that, in the LH, FWD is larger than both BCK ($p < 0.001$) and SIL ($p = 0.023$).

As can be seen from Figure 13.5, for channel 9, the RH shows a significant activation during the plateau of the [Oxy-Hb] response, for both FWD and BCK (but not for SIL). However, in the LH, significant activation is seen only for FWD, and not for either BCK or SIL. Indeed, post-hoc tests restricted to channel 9 show that in this channel, FWD evokes a stronger response that SIL, while BCK is no different from SIL. Further, in this channel, for the FWD condition, LH shows

a greater activation than the RH ($p = 0.0052$), while in the BCK and the SIL condition, the two hemispheres do not show differential activation (both $p > 0.2$).

Similar analyses for [Deoxy-Hb] showed no significant results.

13.3.3 Discussion

The results from this study replicate partially those observed in Peña et al. (2003). Unlike in the previous study, we do not find any main effects of condition, hemisphere or channels. As mentioned earlier, the power of the NIR light used was smaller in this study as compared to Peña et al. (2003) (0.43 mW, compared to 0.7 mW per channel). Indeed, the size of the responses in this study were much smaller than those reported in Peña et al. (2003). Thus, one reason why we do not find main effects in the repeated measures ANOVA as in the previous study might be due to a smaller signal-to-noise ratio in this study. The smaller power used in this study might also be responsible for the lack of any effects with [Deoxy-Hb].

However, a region-of-interest analysis, restricted to channels 9 and 11, finds two important results. First, in the LH, FWD produced a greater activation than BCK, while in the RH there was no such difference. Further, in channel 9, [Oxy-Hb] was significantly greater in the LH compared to the RH only in the FWD condition.

The choice of channels 9 and 11 was motivated by the finding in Peña et al. (2003) that these showed the greatest degree of lateralization of the signal. Despite small differences in the positioning of the probes, channel 9 in both the studies shows a larger activation for forward utterances in the LH as compared to the RH. In addition, in this study, channel 7 too shows a greater activation for FWD as compared to the BCK and SIL, only in the LH.

Notwithstanding the differences between the studies, in this replication of Peña et al. (2003), we find evidence for a lateralized brain response to (forward) speech compared to backward, non-speech. Thus, we can conclude that natural speech contains certain cues which preferentially engage auditory areas in the LH of the neonate.

These findings further validate the use of NIRS in understanding the functional

organization of the neonate brain. Recent advances in imaging methods have demonstrated that the perisylvian language areas in adults are asymmetrically organized for the linguistic content of speech sounds (e.g., Dehaene-Lambertz et al., 2006). NIRS offers the possibility of examining the functional organization of such areas in the pre-linguistic infant.

Part IV

Appendices and references

Appendix A

Details of the pilot experiment from Chapter 5

For this pilot experiment, I chose the following four trisyllabic nonce words as targets: /rutuja/, /tamoki/, /suɮaka/ and /tɛrɛda/

The following syllables were used to construct the speech stream in the pilot experiment: /ku/, /no/, /pi/, /pu/, /fu/, /vu/, /mɛ/, /va/, /d̪u/, /fo/, /na/, /fi/, /mo/, /ko/, /ɮi/, /fɛ/, /pɛ/, /ɮo/, /ja/, /t̪a/, /kɛ/, /ro/, /jɛ/, /nɛ/, /su/, /ɮɛ/, /ma/, /ra/, /t̪o/, /t̪i/, /sa/, /po/, /ni/, /pa/, /mi/, /vɛ/, /d̪ɛ/, /mu/, /jo/, /d̪o/, /nu/, /vi/, /sɛ/, /vo/, /d̪i/, /ri/ and /fa/.

The familiarization stream was made as follows. Initially, 100 blocks of noise were created by randomly permuting the 40 syllables 100 times. Next, the four words were inserted once into each block at a random location and in random order. The permutations and placements were not completely at random; only those bisyllables were allowed that did not sound like an Italian or an English word, as all the people to be used as subjects spoke at least one of these languages. This was done by creating a lookup table for all the possible bisyllables; each time that a new syllable was chosen, the program checked using the table whether or not the current syllable could follow the previous one. Finally, the blocks were concatenated. The program was written in BASIC. On average, each of the random syllables could be followed by 20 other syllables. This gives a theoretical TP of 0.05 per bisyllable. However because of the stochastic nature of the algorithm that generated the random stream, the actual range of TPs was between 0.01 and

0.15 (which is much lower than the TP of 1.0 per bisyllable that makes a word) .

The concatenated list of syllables was fed into the speech synthesis program, MBROLA (Dutoit et al., 1996), using the Spanish (male, es1) database, with all phonemes of the same length (125msec), with the F1 reaching maximum (100%) amplitude at 50% phoneme length. The resulting wav file was formatted to 16kHz, 16 bit, stereo and the edges were ramped using WaveWorks 1.23. For the test phase, words and non-words were similarly synthesised using MBROLA and WaveWorks (without ramping). The non-words were trisyllabic units that had never actually occurred during familiarization.

Methods

Participants

Thirteen undergraduate and graduate university students and postgraduates between the ages of 20 and 30 participated in this experiment. All spoke at least one of Italian or English as a native language. They reported no auditory or language-related problems and were naïve with respect to the aims of the experiment.

Materials

Apparatus

The entire experiment was run by the software PRESENTATION (Neurobehavioral Systems, Inc., CA, USA), which delivered all instructions and stimuli. The audio stimuli were delivered through headphones (Sony, MDR-CD280) attached to multimedia speakers (Harman/kardon Multimedia HK19.5) that were connected to the sound card (SoundBlaster Live! from Creative Technology Ltd.) on the computer.

Procedure

Each participant was seated in front of a computer screen where instructions were displayed. In the first phase, participants were instructed to listen to a speech stream in an "invented" language and to try and pick up 'words' from this language.

At the end of the familiarization phase, participants were instructed to listen to trisyllabic auditory test items. The four words and four non-words were presented randomly for a total of 8 trialsAfter listening to each trisyllable, participants had to press the (premarked) 'z' key if they though that that they had heard the item during familiarization, and the '/' if not. A response was coded as being correct if the participant responded with a *yes* for a 'word' and *no* for a non-word.

Results and discussion

The total correct responses for all the participants, expressed as a percentage of the total possible responses, was 71.15%, and it was significantly different from chance, 2-tailed t-test, $p < 0.001$. Scores for the words alone (69.2%) and for the non-words alone (73.1%) were both different from chance (both $p < 0.01$). An item-wise analysis showed no difference between the different test items (see Figure 5.1 on page 50).

The results from this pilot experiment demonstrate that the presence of syllabic noise does not hinder the extraction of statistically defined, trisyllabic 'words'.

Appendix B

Sentences for making IPs

The following Italian sentences were spoken by a single female Italian speaker. In each set of sentences, a phrase corresponding to a single intonational phrase (IP) was embedded. The portions corresponding to the intonational phrases are underlined.

1. È gia tardi, devi andare a scuola. Il latte è caldo, bevilo. Tutte le mattine la solita storia! ("It's already late, you have to leave for school. The milk is hot. Drink it up. Every morning it's the same old story!")

2. Ti ho comprato lo sciroppo per la tosse. Bevilo tutto. Ti fa bene. ("I've bought you some syrup for your cough. Drink it all up. It will do you good.")

3. Ascolta. Bevilo lentamente. È molto caldo. ("Listen. Drink itslowly. It's very hot.")

4. Mi sembri un bambino di due anni. Non ridere quando bevi. Ti sbrodoli tutto. ("You look like a 2-year-old child. Don't laugh while drinking. You're making a mess all over you.")

5. Ti ho preparato un po' di Scotch. Mettici il ghiaccio e bevi. Ma non troppo, visto che devi guidare. ("I've fixed you some Scotch. Put in some ice and drink it. But don't overdo it, since you have to drive.")

6. Mi vergogno di te. Guarda come bevi. Sembri un bambino. ("I'm ashamed of you. Look how you drink. You look like a child.")

7. Questo libro mi è molto caro. Lo metto via. Altrimenti rischio di perderlo. ("This book is very dear to me. I'll put it away. I might lose it otherwise.")

8. Dicono che questo libro è molto bello. Lo penso anch'io. Anche se devo ancora finire di leggerlo. ("They say that this is a very nice book. I think so too. But I am yet to finish reading it.")

9. Giovanni vuole il mio libro. Lo terrò nascosto. Altrimenti, se lo prende, rischia di non restituirmelo più. ("John wants my book. I'm going to keep it hidden. If he takes it, he might never return it.")

The following Japanese sentences were spoken by a single female Japanese speaker. In each set of sentences, a phrase corresponding to a single intonational phrase (IP) was embedded. The portions corresponding to the intonational phrases are underlined.

1. Keito-wa kibun-ga warukatta. Kusuri-wo nonda. Kedo isha-ni itta. Kanojo-ha ima kibun-ga yokunatteiru. ("Keito was not feeling well. She took medicine. But she went to the doctor. She is feeling better now.")

2. Keito-ha atama-ga itakatta. Isha-ni itta. Kanojo-ha kusuri-wo moratta. ("Keito had headache. She went to the doctor. She got medicine").

3. Keito-ha netsu-ga atta. Kusuri-wo nonda kedo, isha-ni itta. Kanojo-ha haien datta. ("Keito had fever. Though she took medicine, she went to the doctor. She contracted pneumonia.").

4. Keito-ha isshoukenmei-ni benkyousita. Nyuusi-ni shippaisita. Kedo isha-ni naritakatta. Kanojo-ha saido chousensuru. ("Keito studied hard. She failed in an entrance examination. But she would like to be a doctor. She will challenge it again.")

5. Keito-ha igaku-ni kyoumi-ga atta. Isha-ni naritakatta. Kanojo-ha isshoukenmei-ni benkyousita. ("Keito was interested in medicine. She would like to be a doctor. She studied hard.")

6. Keito-ha benkyousinakatta. Nyuusi-ni shippaisitakedo, isha-ni narita katta. Kanojo-ha benkyousihajimeta. ("Keito didn't study. Though she failed in an entrance examination, she would like to be a doctor. She started to study. ")

7. Keito-ha Fukutsuu-ga sita. Naottato omotta. Kedo isha-ni denwasita. Kanojo-ha sugu isha-ni itta. ("Keito had abdominal pain. She thought that it got better. But she called a doctor. She went to the docter immediately.")

8. Keito-ha kega-wo sita. Isha-ni denwasita. Isha-ha byouin-he ikuyouni itta. ("Keito was injured. She called a doctor. He said that she should go to the hospital.")

9. Keito-ha keiren-wo okoshita.Naottato omotta kedo, isha-ni denwasita. Isha-ha annsei-ni suruyouni itta. ("Keito went into convulsions. She thought that they got better, but she called the doctor. He said that she should lie quietly.")

References

Abercrombie, D. (1967). *Elements of general phonetics*. Edinburgh: Edinburgh University Press.

Alain, C., Arnott, S. R., Hevenor, S., Graham, S., & Grady, C. L. (2001). "What" and "where" in the human auditory system. *Proc Natl Acad Sci U S A*, *98*(2), 12301-12306.

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of conditional probability statistics by human infants. *Psychological Science*, *9*, 321-324.

Baddeley, A. (1990). *Human memory*. Lawrence Erlbaum Associates Ltd. Hove, UK.

Baddeley, A. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences*, *4*, 417-423.

Baddeley, A. (2003). Working memory: Looking back and looking forward. *Nature Reviews Neuroscience*, *4*, 829-839.

Bagou, O., Fougeron, C., & Frauenfelder, U. (2002). Contribution of prosody to the segmentation and storage of "words" in the acquisition of a new mini-language. In B. Bel & I. Marlien (Eds.), *Proceedings of the speech prosody 2002 conference* (p. 59 - 62). Aix-en-Provence: Laboratoire Parole et Langage.

Baillargeon, R. (1995). Physical reasoning in infancy. In M. Gazzaniga (Ed.), *The cognitive neurosciences* (p. 181-204). Cambridge, MA: MIT Press.

Baird, A., Kagan, J., Gaudette, T., Walz, K., Hershlag, N., & Boas, D. (2002). Frontal lobe activation during object permanence: data from near-infrared spectroscopy. *Neuroimage*, *16*, 1120-1126.

Baker, M. (2001). *The atoms of language: The mind's hidden rules of grammar*. New York: Basic Books.

Bard, E., & Anderson, A. (1983). The unintelligibility of speech to children. *J Child Lang*, *10*, 265-292.

Bard, E., & Anderson, A. (1994). The unintelligibility of speech to children: effects of referent availability. *J Child Lang*, *21*, 623-648.

Batchelder, E. (2002). Bootstrapping the lexicon: A computational model of infant speech segmentation. *Cognition*, *83*(2), 167-206.

Bates, E., Vicari, S., & Tauner, D. (1999). Neural mediation of language development: Perspectives from lesion studies of infants and children. In H. Tager-Flusberg (Ed.), *Neurodevelopmental disorders* (p. 533-581). Cambridge, MA: MIT Press.

Beckman, M. E., & Pierrehumbert, J. (1986). Intonational structure in japanese and english. *Phonology Yearbook*, *3*, 255-309.

Benson, D. F., & Zaidel, E. (Eds.). (1985). *The dual brain: Hemispheric specialization in humans*. New York: Guilford Press.

Bertenthal, B. (1993). Infants' perception of biomechanical motions: Instrinsic image and knowledge-based constraints. In C. Granrud (Ed.), *Visual perception and cognition in infancy.* Hillsdale, NJ: Erlbaum.

Bertinetto, P. (1981). *Strutture prosodiche dellItaliano. accento, quantit, sillaba, giuntura, fondamenti metrici*. Firenze: Accademia della Crusca.

Bertoncini, J., Bijeljac-Babic, R., Blumstein, S. E., & Mehler, J. (1987). Discrimination in neonates of very short cvs. *J Acoust Soc Am*, *82*(1), 31-37.

Bertoncini, J., Bijeljac-Babic, R., Jusczyk, P. W., Kennedy, L. J., & Mehler, J. (1988). An investigation of young infants' perceptual representations of speech sounds. *J Exp Psychol Gen*, *117*(1), 21-33.

Bertoncini, J., Floccia, C., Nazzi, T., & Mehler, J. (1995). Morae and syllables: Rhythmical basis of speech representations in neonates. *Lang Speech*, *38*(Pt 4), 311-329.

Bertoncini, J., & Mehler, J. (1981a). Syllables as units in infant speech perception. *Infant Behavior and Development*, *4*, 247-260.

Bertoncini, J., & Mehler, J. (1981b). [perception in newborn infants]. *Rev Prat*, *31*(5), 387-8, 391-2, 395-6.

Bertoncini, J., Morais, J., Bijeljac-Babic, R., McAdams, S., Peretz, I., & Mehler, J. (1989). Dichotic perception and laterality in neonates. *Brain Lang*, *37*(4), 591-605.

Best, C., Hoffman, H., & Glanville, B. (1982). Development of infant ear asymmetries for speech and music. *Percept Psychophys*, *31*(1), 75-85.

Bijeljac-Babic, R., Bertoncini, J., & Mehler, J. (1993). How do four-day-old infants categorize multisyllabic utterances? *Developmental Psychology*, *29*, 711-721.

Blake, M. (2003). Affective language and humor appreciation after right hemisphere brain damage. *Semin Speech Lang*, *24*, 107-120.

Blevins, J. (1995). The syllable in phonological theory. In J. Goldsmith (Ed.), *The handbook of phonological theory* (p. 206-244). Oxford: Blackwell.

Boemio, A., Fromm, S., Braun, A., & Poeppel, D. (2005). Hierarchical and asymmetric temporal sensitivity in human auditory cortices. *Nat Neurosci*, *8*, 389-395.

Bonatti, L., Peña, M., Nespor, M., & Mehler, J. (2005). Linguistic constraints on statistical computations: The role of consonants and vowels in continuous speech processing. *Psychological Science*, *16*(6), 451-459.

Bortfeld, H., Morgan, J. L., Golinkoff, R. M., & Rathbun, K. (2005). Mommy and me. familiar names help launch babies into speech-stream segmentation. *Psychological Science*, *16*(4), 298-304.

Bortfeld, H., Wruck, E., & Boas, D. (2006). Assessing infants' cortical response to speech using near-infrared spectroscopy. *NeuroImage*, in press.

Bouton, M. E., Nelson, J. B., & Rosas, J. M. (1999). Stimulus generalization, context change, and forgetting. *Psychological Bulletin*, *125*, 171-186.

Bregman, A. (1983/1990). *Auditory scene analysis*. Cambridge, MA: MIT Press.

Brent, M., & Cartwright, T. (1996). Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition*, *61*(1-2), 93-125.

Bresnan, J. (2001). *Lexical funnctional syntax*. Oxford: Blackwell.

Broca, P. P. (1861). [loss of speech, chronic softening and partial destruction of the anterior left lobe of the brain]. *Bulletin de la Société Anthropologique*, *2*, 235-238.

Brown, G., Preece, T., & Hulme, C. (2000). Oscillator-based memory for serial order. *Psychological Review*, *107*, 127-181.

Burgess, N., & Hitch, G. (1999). Memory for serial order: A network model of the phonological loop and its timing. *Psychological Review*, *106*, 551-581.

Byrd, D., Kaun, A., Narayanan, S., & Saltzman, E. (2000). Phrasal signatures in articulation. In M. Broe & J. Pierrehumbert (Eds.), *Papers in laboratory phonology v* (p. 70-67). Cambridge: Cambridge University Press.

Cambier-Langeveld, T. (2000). *Temporal marking of accents and boundaries*. Leiden: Holland Institute of Generative Linguistics.

Chance, B., Zhuang, Z., Chu, U., Alter, C., & Lipton, L. (1993). Cognition-activated low-frequency modulation of light absorption in human brain. *Proc Natl Acad Sci U S A*, *90*, 3770-3774.

Charniak, E. (1993). *Statistical language learning*. Cambridge, MA: The MIT Press.

Chomsky, N. (1986). *Knowledge of language: Its nature,origin and use*. New York: Praeger.

Chomsky, N. (1995). *The minimalist program (current studies in linguistics)*. Cambridge, MA: The MIT Press.

Chomsky, N. (2000). *New horizons in the study of language and mind*. Cambridge University Press Cambridge.

Chomsky, N. (2005). Three factors in language. *Linguistic Inquiry*, *36*(1), 1-22.

Christiansen, M., Allen, J., & Seidenberg, M. (1998). Learning to segment speech using multiple cues: A connectionist model. *Language and Cognitive Processes*, *13*, 221-268.

Christophe, A., Dupoux, E., Bertoncini, J., & Mehler, J. (1994). Do infants perceive word boundaries? An empirical study of the bootstrapping of lexical acquisition. *J Acoust Soc Am*, *95*(3), 1570-1580.

Christophe, A., Gout, A., Peperkamp, S., & Morgan, J. (2003). Discovering words in the continuous speech stream: the role of prosody. *Journal of Phonetics*, *31*, 585-598.

Christophe, A., Guasti, M. T., Nespor, M., & Ooyen, B. van. (2003). Prosodic structure and syntactic acquisition: The case of the head-complement parameter. *Developmental Science*, *6*, 213-222.

Christophe, A., Mehler, J., & Sebastián-Gallés, N. (2001). Perception of prosody boundary correlates by newborn infants. *Infancy*, *2*, 385-394.

Christophe, A., & Morton, J. (1998). Is dutch native english? linguistic analysis by 2-month-olds. *Developmental Science*, *1*(2), 215-219.

Christophe, A., Nespor, M., Guasti, M. T., & van, B., Ooyen. (1997). Reflections on phonological bootstrapping: Its role in lexical and syntactic acquisition. In G. Altmann (Ed.), *Cognitive models of speech processing: A special issue of language and cognitive processes.* Lawrence Erlbaum Mahwah, N.J.

Christophe, A., Peperkamp, S., Pallier, C., Block, N., & Mehler, J. (2004). Phono-

logical phrase boundaries constrain lexical access: I. adult data. *Journal of Memory and Language*, *51*, 523-547.

Church, B., & Schacter, D. (1994). Perceptual specificity of auditory priming: Memory for voice intonation and fundamental frequency. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *20*, 521-533.

Church, K. (1987). *Phonological parsing in speech recognition*. Dordrecht: Kluwer Academic.

Cohn, A. (1989). Stress in indonesian and bracketing paradoxes. *Natural Language and Linguistic Theory*, *7*, 167-216.

Cole, R., Jakimik, J., & Cooper, W. (1980). Segmenting speech into words. *J Acoust Soc Am*, *67*(4), 1323-1332.

Colombo, J., & Bundy, R. (1983). Infant response to auditory familiarity and novelty. *Infant Behavior and Development*, *6*, 305-311.

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychol Rev*, *108*(1), 204-256.

Creel, S., Newport, E., & Aslin, R. (2004). Distant melodies: statistical learning of nonadjacent dependencies in tone sequences me factors. *J Exp Psychol Learn Mem Cogn urnal Article*, *530*, 1119-1130.

Curtiss, S., & Bode, S. de. (2003). How normal is grammatical development in the right hemisphere following hemispherectomy? the RI stage and beyond. brain and language, 86, 193206. *Brain and language*, *86*, 193-206.

Cutler, A., & Carter, D. (1987). The predominance of strong initial syllables in the english vocabulary. *Computer Speech & Language*, *2*, 133-142.

Cutler, A., Dahan, D., & van, W., Donselaar. (1997). Prosody in the comprehension of spoken language: a literature review. *Lang Speech*, *40*(Pt 2), 141-201.

Cutler, A., Demuth, K., & McQueen, J. (2002). Universality versus language-specificity in listening to running speech. *Psychol Sci*, *13*(3), 258-262.

Cutler, A., & Fodor, J. (1979). Semantic focus and sentence comprehension. *Cognition*, *7*(1), 49-59.

Cutler, A., & Norris, D. (1988). The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human Perception and*

Performance, *14*, 113-121.

Cutler, A., Weber, A., & Otake, T. (2006). Asymmetric mapping from phonetic to lexical representations in second-language listening. *Journal of Phonetics*, *34*(2), 269-284.

Dahan, D., & Brent, M. (1999). On the discovery of novel wordlike units from utterances: An artificial-language study with implications for native-language acquisition. *J Exp Psychol Gen*, *128*(2), 165-185.

Dauer, R. M. (1987). *Phonetics and phonological components of language rhythm.* Paper presented at the 11th international congress of phonetic sciences, vol. 5 Talinn.

DeCasper, A., & Fifer, W. (1980). Of human bonding: newborns prefer their mothers' voices. *Nature*, *208*, 1174-1176.

Dehaene, S., Dupoux, E., Mehler, J., Cohen, L., Paulesu, E., Perani, D., van, P., de Moortele, Lehericy, S., & Le, D., Bihan. (1997). Anatomical variability in the cortical representation of first and second language. *Neuroreport*, *8*(17), 3809-3815.

Dehaene, S., & Naccache, L. (2006). Can one suppress subliminal words? *Neuron*, *52*, 397-399.

Dehaene-Lambertz, G. (2000). Cerebral specialization for speech and non-speech stimuli in infants. *Journal of Cognitive Neuroscience*, *12*, 449-460.

Dehaene-Lambertz, G., Dehaene, S., Anton, J.-L., Campagne, A., Ciuciu, P., Dehaene, G., Denghien, I., Jobert, A., LeBihan, D., Sigman, M., Pallier, C., & Poline, J. (2006). Functional segregation of cortical language areas by sentence repetition. *Hum Brain Mapp*, *27*, 360-371.

Dehaene-Lambertz, G., Dehaene, S., & Hertz-Pannier, L. (2002). Functional neuroimaging of speech perception in infants. *Science*, *298*(5600), 2013-2015.

Démonet, J.-F. c., Thierry, G., & Cardebat, D. (2005). Renewal of the neurophysiology of language: Functional neuroimaging. *Physiol Rev*, *85*, 49-95.

D'Imperio, M., & Rosenthall, S. (1999). Phonetics and phonology of main stress in italian. *Phonology*, *16*, 1-28.

Dutoit, T. (1997). *An introduction to text-to-speech synthesis.* Kluwer Academic Publishers.

Ebbinghaus, H. (1964). *Memory: A contribution to experimental psychology.* Dover Oxford, England.

Echols, C. (1993). A perceptually-based model of children's earliest productions. *Cognition*, *46*(3), 245-296.

Echols, C. H. (1996). A role for stress in early speech segmentation. *Signal to syntax: Bootstrapping from speech to grammar in early acquisition.*, 151-170.

Ecklund-Flores, L., & Turkewitz, G. (1996). Asymmetric headturning to speech and nonspeech in human newborns. *Dev Psychobiol*, *29*(3), 205-217.

Endress, A., Scholl, B., & Mehler, J. (2005). The role of salience in the extraction of algebraic rules. *J Exp Psychol Gen*.

Eulitz, C., & Lahiri, A. (2004). Neurobiological evidence for abstract phonological representations in the mental lexicon during speech recognition. *J Cogn Neurosci*, *16*, 577-583.

Fisher, C., & Tokura, H. (1996). Prosody in speech to infants: Direct and indirect acoustic cues to syntactic structure. In E. Morgan James L. & E. Demuth Katherine (Eds.), *Signal to syntax: Bootstrapping from speech to grammar in early acquisition.* (p. 343-363). Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc.

Fodor, J. (1981). The present status of the innateness controversy. In J. Fodor (Ed.), *Representations* (p. 257-316). Cambridge, MA: MIT Press.

Fougeron, C., & Keating, P. (1997). Articulatory strengthening at edges of prosodic domains. *Journal of the Acoustical Society of America*, *101*, 3728-3740.

Frazier, L., Carlson, K., & Clifton, C. (2006). Prosodic phrasing is central to language comprehension. *Trends Cogn Sci*, *10*(6), 244-249.

Frazier, L., & Clifton, C. (1998). Sentence reanalysis, and visibility. In J. Fodor & F. Ferreira (Eds.), *Reanalysis in sentence processing.* Dordrecht: Kluwer.

Friederici, A., Steinhauer, K., & Pfeifer, E. (2002). Brain signatures of artificial language processing: evidence challenging the critical period hypothesis. *Proc Natl Acad Sci U S A*, *99*(1), 529-534.

Friederici, A., & Wessels, J. (1993). Phonotactic knowledge of word boundaries and its use in infant speech perception. *Perception & psychophysics*, *54*(3), 287-295.

Frost, R. (1998). Toward a strong phonological theory of visual word recognition:

True issues and false trails. *Psychological Bulletin*, *123*, 71-99.

Frost, R. (2003). The robustness of phonological effects in fast priming. In S. Kinoshita & S. Lupker (Eds.), *Masked priming: The state of the art* (p. 173-191). New York: Psychology Press.

Fujimura, O. (1990). Methods and goals of speech production research. *Language and Speech*, *33*, 195-258.

Galves, A., Garcia, J., Duarte, D., & Galves, C. (2002). Sonority as a basis for rhythmic class discrimination. In B. Bel & I. Marlien (Eds.), *Proceedings of the speech prosody 2002 conference* (p. 323-326). Aix-en-Provence: Laboratoire Parole et Langage.

Giegerich, H. (1992). *English phonology: An introduction*. Cambridge, MA: Cambridge University Press.

Gillette, J., Gleitman, H., Gleitman, L., & Lederer, A. (1999). Human simulations of vocabulary learning. *Cognition*, *73*(2), 135-176.

Giraud, A.-L., Lorenzi, C., Ashburner, J., Wable, J., Johnsrude, I., Frackowiak, R., & Kleinschmidt, A. (2000). Representation of the temporal envelope of sounds in the human brain. *J Neurophysiol*, *84*(3), 1588-1598.

Glanville, B., Best, C., & Levenson, R. (1977). A cardiac measure of cerebral asymmetries in infant auditory perception. *Developmental Psychology*, *13*, 54-59.

Gleitman, L., Cassidy, K., Nappa, R., Papafragou, A., & Trueswell, J. (2005). Hard words. *Language Learning and Development*, *1*(1), 23-64.

Godden, D., & Baddeley, A. (1975). Context-dependent memory in two natural environments: On land and under water. *British Journal of Psychology*, *66*, 325-331.

Goldinger, S. (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. *J Exp Psychol Learn Mem Cogn*, *22*(5), 1166-1183.

Goldinger, S. (1998). Echoes of echoes? an episodic theory of lexical access. *Psychol Rev*, *105*(2), 251-279.

Goldsmith, J. (1976). *Autosegmental phonology*. Garland Press, 1979 Doctoral dissertation, MIT. New York.

Goldsmith, J. (1990). *Autosegmental and metrical phonology*. Blackwell Cam-

bridge.

Goodale, M., & Milner, A. (1992). Separate visual pathways for perception and action. *Trends Neurosci*, *15*, 20-25.

Gopnik, A., Meltzoff, A., & Kuhl, P. (1999). *The scientist in the crib: minds, brains, and how children learn*. New York: William Morrow and Co.

Gout, A., Christophe, A., & Morgan, J. L. (2004). Phonological phrase boundaries constrain lexical access: Ii. infant data. *Journal of Memory and Language*, *51*, 547-567.

Gow, D., & Gordon, P. (1995). Lexical and prelexical influences on word segmentation: Evidence from priming. *Journal of Experimental Psychology: Human Perception and Performance*, *21*, 344-359.

Gow, D., Melvold, J., & Manuel, S. (1996). How word onsets drive lexical access and segmentation: evidence from acoustics, phonology and processing. *Proc. ICSLP96*.

Grabe, E., & Low, E. L. (2002). Durational variabilty in speech and the rhythm class hypothesis. In *Papers in laboratory phonology 7*. The Hague: Mouton.

Gratton, G., Goodman-Wood, M., & Fabiani, M. (2001). Comparison of neuronal and hemodynamic measures of the brain response to visual stimulation: an optical imaging study. *Hum Brain Mapp*, *13*, 13-25.

Guasti, M. T. (2002). *Language acquisition: The growth of grammar*. The MIT Press Cambridge, MA, US.

Gussenhoven, C., & Rietveld, A. (1992). Intonation contours, prosodic structure and preboundary lengthening. *J. Phon*, *20*, 283-303.

Hardcastle, W. (1985). Some phonetic and syntactic constraints on lingual coarticulation in stop consonant sequences. *Speech Communication*, *4*, 247-263.

Harris, Z. (1955). From phoneme to morpheme. *Language*, *31*, 190-222.

Hauser, M., Chomsky, N., & Fitch, W. (2002). The faculty of language: what is it, who has it, and how did it evolve? *Science*, *298*(5598), 1569-1579.

Hauser, M., Newport, E., & Aslin, R. (2001). Segmentation of the speech stream in a non-human primate: statistical learning in cotton-top tamarins. *Cognition*, *78*(3), B53-64.

Haxby, J., Grady, C., Horwitz, B., Ungerleider, L., Mishkin, M., Carson, R., Herscovitch, P., Schapiro, M., & Rapoport, S. (1991). Dissociation of object

and spatial visual processing pathways in human extrastriate cortex. *Proc Natl Acad Sci U S A*, *88*, 1621-1625.

Hayes, B., & Lahiri, A. (1991). Bengali intonational phonology. *Natural Language and Linguistic Theory*, *9.1*, 47-96.

Hayes, J., & Clarke, H. (1970). Experiments on the segmentation of an artificial speech analogue. In J. Hayes (Ed.), *Cognition and the development of language.* Wiley New York.

Henson, R. (1998). Short-term memory for serial order: The Start End model. *Cognitive Psychology*, *36*, 73-137.

Hickok, G., & Poeppel, D. (2000). Towards a functional neuroanatomy of speech perception. *Trends Cogn Sci*, *4*, 131-138.

Hirsh-Pasek, K., Kemler, D. G., Nelson, Jusczyk, P. W., Cassidy, K. W., Druss, B., & Kennedy, L. (1987). Clauses are perceptual units for young infants. *Cognition*, *26*(3), 269-286.

Hirst, D., & Di Cristo, A. (1998). A survey of intonation systems. In D. Hirst & A. Di Cristo (Eds.), *Intonation systems: A survey of twenty languages.* Cambridge: CUP.

Hitch, G., Burgess, N., Towse, J., & Culpin, V. (1996). Temporal grouping effects in immediate recall: A working memory analysis. *Quarterly Journal of Experimental Psychology*, *49A*, 116-139.

Hockett, C. (1963). The problem of universals in language. In J. H. Greenberg (Ed.), *Universals of language.* Cambridge, MA: MIT Press.

Hohne, E., & Jusczyk, P. (1994). Two-month-old infants' sensitivity to allophonic differences. *Percept Psychophys*, *56*(6), 613-623.

Imada, T., Zhang, Y., Cheour, M., Taulu, S., Ahonen, A., & Kuhl, P. (2006). Infant speech perception activates brocas area: a developmental magnetoencephalography study. *Neuroreport*, *17*(10), 957-962.

Jackendoff, R., & Lerdahl, F. (under review). *The capacity for music: What is it, and whats special about it?* Retrieved July 28, 2006 from http://ase.tufts.edu/cogstud/incbios/RayJackendoff/index.htm.

Jacoby, L., & Brooks, L. (1984). Nonanalytic cognition: Memory, perception, and concept learning. In G. Bower (Ed.), *The psychology of learning and motivation, vol. 18* (p. 1-47). New York: Academic Press.

Jakobson, R. (1942). Lectures on sound and meaning.

James, W. (1890). *The principles of psychology.* New York: Holt.

Jobsis, F. (1977). Noninvasive, infrared monitoring of cerebral and myocardial oxygen sufficiency and circulatory parameters. *Science*, *198*(4323), 1264-1267.

Johnson, E., & Jusczyk, P. (2001). Word segmentation by 8-month-olds: When speech cues count more than statistics. *Journal of Memory and Language*, *44*, 1-20.

Johnson, E., Jusczyk, P., Cutler, A., & Norris, D. (2000). 12-month-olds show evidence of a possible-word constraint (a). *JASA*.

Johnson, E., Jusczyk, P., Cutler, A., & Norris, D. (2003). Lexical viability constraints on speech segmentation by infants. *Cognit Psychol*, *46*(1), 65-97.

Jun, S.-A. (1993). *The phonetics and phonology of Korean prosody.* Unpublished doctoral dissertation, Ohio State University.

Jusczyk, P., Cutler, A., & Redanz, N. (1993). Infants' preference for the predominant stress patterns of english words. *Child Dev*, *64*(3), 675-687.

Jusczyk, P., Houston, D., & Newsome, M. (1999). The beginnings of word segmentation in english-learning infants. *Cognit Psychol*, *39*(3-4), 159-207.

Jusczyk, P., & Krumhansl, C. (1993). Pitch and rhythmic patterns affecting infants' sensitivity to musical phrase structure. *J Exp Psychol Hum Percept Perform*, *9*(3), 627-640.

Jusczyk, P., Pisoni, D., & Mullennix, J. (1992). Some consequences of stimulus variability on speech processing by 2-month-old infants. *Cognition*, *43*(3), 253-291.

Jusczyk, P. W. (1989). *Perception of cues to clausal units in native and non-native languages.* Paper presented at the biennial meeting of the Society for Research in Child Development Kansas City, Missouri.

Jusczyk, P. W., & Kemler, D. G., Nelson. (1996). Syntactic units, prosody, and psychological reality during infancy. In J. L. Morgan & K. Demuth (Eds.), *Signal to syntax: Bootstrapping from speech to grammar in early acquisition.* (p. 389-408). Lawrence Erlbaum Associates, Inc Hillsdale, NJ, England.

Jusczyk, P. W., Luce, P., & Charles-Luce, J. (1994). Infants' sensitivity to phono-

tactic patterns in the native language. *Journal of Memory and Language*, *33*, 630-645.

Keating, P., Cho, T., Fougeron, C., & Hsu, C. (2003). Domain-initial articulatory strengthening in four languages. In J. Local, R. Ogden, & R. Temple (Eds.), *Phonetic interpretation* (p. 143-161). Cambridge, MA: Cambridge University Press.

Kemler, D., Nelson, Hirsh-Pasek, K., Jusczyk, P., & Wright-Cassidy, K. (1989). How the prosodic cues in motherese might assist language learning. *J Child Lang*, *16*(1), 55-68.

Kim, K., Relkin, N., Kyoung-Min, L., & Hirsch, J. (1997). Distinct cortical areas associated with native and second languages. *Nature*, *July 10388*, 171-174.

Kisilevsky, B., Hains, S., Lee, K., Xie, X., Huang, H., Ye, H., Zhang, K., & Wang, Z. (2003). Effects of experience on fetal voice recognition. *Psychol Sci*, *14*(3), 220 - 224.

Klatt, D. (1976). Linguistic uses of segmental duration in english: acoustic and perceptual evidence. *J Acoust Soc Am*, *59*(5), 1208-1221.

Klatt, D. (1979). Speech perception: A model of acoustic-phonetic analysis and lexical access. *Journal of Phonetics*, *7*, 279-312.

Klatt, D., & Stevens, K. (1973). On the automatic recognition of continuous speech: Implications from a spectrogram-reading experiment. *IEEE Transactions on audio and electroacoustics*, *21*(3), 210-217.

Koestler, A. (1967). *The ghost in the machine*. London: Hutchinson.

Krumhansl, C., & Jusczyk, P. (1990). Infants' perception of phrase structure in music. *Psychological Science*, *1*, 70-73.

Kujala, A., Huotilainen, M., Hotakainen, M., Lennes, M., Parkkonen, L., Fellman, V., & Näätänen, R. (2004). Speech-sound discrimination in neonates as measured with meg. *Neuroreport*, *15*(13), 2089-2092.

Ladd, R., & Campbell, N. (1991). Theories of prosodic structure: Evidence from syllable duration. *Proceedings of the XIIth International Congress of Phonetic Sciences, Aix-en-Provence*, *II*, 290-293.

Ladefoged, P. (1975). *A course on phonetics*. New York: Harcourt Brace & Jovanovich.

Lahiri, A., & Marslen-Wilson, W. D. (1991). The mental representation of lexical

form: A phonological approach to the recognition lexicon. *Cognition*, *38*, 245-294.

Lehiste, I., Olive, J., & Streeter, L. (1976). The role of duration in disambiguating syntactically ambiguous sentences. *JASA*, *60*, 1199-1202.

Lepschy, A., & Lepschy, G. (1981). *La lingua italiana*. Bompiani Milan.

Levelt, W. (1989). *Speaking: From intention to articulation*. Cambridge: MIT Press.

Liberman, A., Cooper, F., Shankweiler, D., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychol Rev*, *74*, 431-461.

Liberman, M., & Prince, A. (1977). On stress and linguistic rhythm. *Linguistic Inquiry*, *8*(2), 249-336.

Lidz, J., Gleitman, H., & Gleitman, L. (2003). Understanding how input matters: verb learning and the footprint of universal grammar. *Cognition*, *87*(3), 151-178.

Lidz, J., & Gleitman, L. (2004a). Argument structure and the child's contribution to language learning. *Trends Cogn Sci*, *8*(4), 157-161.

Lidz, J., & Gleitman, L. (2004b). Yes, we still need Universal Grammar (Discussion). *Cognition*, *94*, 85-93.

Lieberman, P. (1967). *Intonation, perception and language.* Cambridge, MA: MIT Press.

Lieberman, P., & Blumstein, S. (1988). *Speech physiology, speech perception, and acoustic phonetics*. Cambridge, UK: Cambridge University Press.

Lukatela, G., & Turvey, M. (1994). Visual lexical access is initially phonological: Evidence from phonological priming by homophones and pseudohomophones. *Journal of Experimental Psychology: General*, *123*, 331-353.

Macdonald, N. (1976). Duration as a syntactic boundary cue in ambiguous sentences. *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '76*, *1*, 569-572.

MacNeilage, P. (1998). The frame/content theory of evolution of speech production. *Behav Brain Sci*, *21*(4), 499-511; discussion 511-46.

MacNeilage, P., & Davis, B. (2001). Motor mechanisms in speech ontogeny: phylogenetic, neurobiological and linguistic implications. *Current Opinions in Neurobiology*, *11*, 696-700.

Maeda, S. (1974). A characterization of fundamental frequency contours of speech. *Quarterly progress report, MIT research laboratory of electronics*, *114*, 193-211.

Mandel, D., Jusczyk, P., & Nelson, D. (1994). Does sentential prosody help infants organize and remember speech information? *Cognition*, *53*(2), 155-180.

Mandel, D., Kemler, D., Nelson, & Jusczyk, P. (1996). Infants remember the order of words in spoken sentences. *Cognitive Development*, *11*, 181-196.

Marotta, G. (1985). *Modelli e misure ritmiche*. Bologna: Zanichelli.

Mattys, S., & Jusczyk, P. (2001). Do infants segment words or recurring contiguous patterns? *J Exp Psychol Hum Percept Perform*, *27*(3), 644-655.

Mattys, S., Jusczyk, P., Luce, P., & Morgan, J. (1999). Phonotactic and prosodic effects on word segmentation in infants. *Cognit Psychol*, *38*(4), 465-494.

Mattys, S., White, L., & Melhorn, J. (2005). Integration of multiple speech segmentation cues: a hierarchical framework. *Journal of Experimental Psychology: General*, *134*(4), 477-500.

McCarthy, J., & Prince, A. (1993). Generalized alignment. In G. Booij & J. v. Marle (Eds.), *Yearbook of morphology* (p. 79-153). Boston: Kluwer.

McQueen, J., Otake, T., & Cutler, A. (2001). Rhythmic cues and possible-word constraints in japanese speech segmentation. *Journal of Memory and Language*, *45*(1), 103-132.

Meek, J. (2002). Basic principles of optical imaging and application to the study of infant development. *Dev Sci*, *5*, 371-380.

Mehler, J. (1981). The role of syllables in speech processing: Infant and adult data. *Philosophical Transactions of the Royal Society*, *295*, 333-352.

Mehler, J., & Dupoux, E. (1994). *What infants know*. Cambridge: Basil Blackwell.

Mehler, J., Dupoux, E., Nazzi, T., & Dehaene-Lambertz, G. (1996). Coping with linguistic diversity: The infant's viewpoint. In E. Morgan James L. & E. Demuth Katherine (Eds.), *Signal to syntax: Bootstrapping from speech to grammar in early acquisition.* (p. 101-116). Lawrence Erlbaum Associates, Inc Hillsdale, NJ, England.

Mehler, J., Jusczyk, P., Lambertz, G., Halsted, N., Bertoncini, J., & Amiel-Tison, C. (1988). A precursor of language acquisition in young infants. *Cognition*, *29*(2), 143-178.

Mehler, J., Segui, J., & Frauenfelder, U. (1981). The role of the syllable in language acquisition and perception. In T. Myers, J. Laver, & J. Anderson (Eds.), *The cognitive representation of speech.* Amsterdam: North Holland.

Menzerath, P., & De Lacerda, A. (1933). *Koartikulation, steuerung und lautabgrenzung.* Bonn: Bonn.

Miller, G. (1956). The magical number seven plus or minus two: Some limits on our capacity for processing information. *Psychol Rev*, *63*(2), 81-97.

Moon, C., Cooper, R. P., & Fifer, W. (1993). Two-day-olds prefer their native language. *Infant Behavior and Development*, *16*, 495-500.

Moon, C., & Fifer, W. (2000). Evidence of transnatal auditory learning. *J Perinatol*, *20*(8 Pt 2), S37-44.

Morgan, J., & Saffran, J. (1995). Emerging integration of sequential and suprasegmental information in preverbal speech segmentation. *Child Dev*, *66*(4), 911-936.

Morgan, J., Swingley, D., & Miritai, K. (1993). Infants listen longer to speech with extraneous noises inserted at clause boundaries. *Paper presented at the Biennial Meeting of the Society for Research in Child Development, New Orleans, LA.*

Morgan, J. L. (1994). Converging measures of speech segmentation in preverbal infants. *Infant Behavior and Development*, *17*, 387-400.

Morgan, J. L., Shi, R., & Allopenna, P. (1996). Perceptual bases of rudimentary grammatical categories: Toward a broader conceptualization of bootstrapping. In E. Morgan James L. & E. Demuth Katherine (Eds.), *Signal to syntax: Bootstrapping from speech to grammar in early acquisition.* (p. 263-283). Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc.

Nakamura, K., Hara, N., Kouider, S., Takayama, Y., Hanajima, R., Sakai, K., & Ugawa, Y. (2006). Task-guided selection of the dual neural pathways for reading. *Neuron*, *52*, 557-564.

Nazzi, T., Bertoncini, J., & Mehler, J. (1998). Language discrimination by newborns: toward an understanding of the role of rhythm. *J Exp Psychol Hum Percept Perform*, *24*(3), 756-766.

Nazzi, T., Iakimova, G., Bertoncini, J., Frédonie, S., & Alcantara, C. (2006). Early segmentation of fluent speech by infants acquiring French: emerging

evidence for crosslinguistic differences. *(in press)*.

Nazzi, T., Kemler, D., Nelson, Jusczyk, P., & Jusczyk, A. (2000). Six-month-olds' detection of clauses embedded in continuous speech: Effects of prosodic well-formdness. *Infancy*, *1*(1), 123-147.

Neelman, A., & Koot, J. van de. (2006). On syntactic and phonological representations. *Lingua*, *116*, 1524-1552.

Nespor, M. (1990). On the rhythm parameter in phonology. In I. Roca (Ed.), *Logical issues in language acquisition.* Dordrecht: Foris.

Nespor, M., Guasti, M. T., & Christophe, A. (1996). Selecting word order: the rhythmic activation principle. In U. Kleinhenz (Ed.), *Interfaces in phonology* (p. 1-26). Berlin: Akademie Verlag.

Nespor, M., & Vogel, I. (1986). *Prosodic phonology.* Dordrecht: Foris.

Newport, E., & Aslin, R. (2004). Learning at a distance i. statistical learning of non-adjacent dependencies. *Cognit Psychol*, *48*(2), 127-162.

Ng, H., & Maybery, M. (2002). Grouping in short-term verbal memory: Is position coded temporally? *Quarterly Journal of Experimental Psychology*, *55A*, 391-424.

Norris, D., McQueen, J., & Cutler, A. (2000). Merging information in speech recognition: feedback is never necessary. *Behav Brain Sci*, *23*(3), 299-325; discussion 325-70.

Norris, D., McQueen, J., Cutler, A., & Butterfield, S. (1997). The possible-word constraint in the segmentation of continuous speech. *Cognit Psychol*, *34*(3), 191-243.

Obrig, H., Hirth, C., Junge-Hulsing, J., Doge, C., Wolf, T., Dirnagl, U., & Villringer., A. (1996). Cerebral oxygenation changes in response to motor stimulation. *J. Appl. Physiol.*, *381*, 1174-1183.

Ohala, J., Dunn, A., & Sprouse, R. (2004). Prosody and phonology. In B. Bel & I. Marlien (Eds.), *Speech prosody 2004, nara, japan* (p. 161-163). ISCA Archive.

Ooijen, B. van, Bertoncini, J., Sansavini, A., & Mehler, J. (1997). Do weak syllables count for newborns? *J Acoust Soc Am*, *102*(6), 3735-3741.

Pallier, C., Colome, A., & Sebastian-Galles, N. (2001). The influence of native-language phonology on lexical access: exemplar-based versus abstract lexical

entries. *Psychol Sci*, *12*(6), 445-449.

Palmeri, T., Goldinger, S., & Pisoni, D. (1993). Episodic encoding of voice attributes and recognition memory for spoken words. *J Exp Psychol Learn Mem Cogn*, *19*(2), 309-328.

Pannekamp, A., Weber, C., & Friederici, A. (2006). Prosodic processing at the sentence level in infants. *Neuroreport*, *17*(6), 675-678.

Pell, M. (1999). Fundamental frequency encoding of linguistic and emotional prosody by right hemisphere-damaged speakers. *Brain and Language*, *69*(2), 161-192.

Peña, M., Bonatti, L., Nespor, M., & Mehler, J. (2002). Signal-driven computations in speech processing. *Science*, *298*(5593), 604-607.

Peña, M., Maki, A., Kovacic, D., Dehaene-Lambertz, G., Koizumi, H., Bouquet, F., & Mehler, J. (2003). Sounds and silence: an optical topography study of language recognition at birth. *Proc Natl Acad Sci U S A*, *100*(20), 11702-11705.

Perruchet, P., & Vinter, A. (1998). Parser: A model for word segmentation. *Journal of Memory and Language*, *39*, 246-263.

Pierrehumbert, J. (1980). *The phonology and phonetics of english intonation.* MIT Linguistics Ph.D. thesis.

Pierrehumbert, J. (2003). Phonetic diversity, statistical learning, and acquisition of phonology. *Lang Speech*, *46*(Pt 2-3), 115-154.

Pijper, J. de, & Sanderman, A. (1994). On the perceptual strength of prosodic boundaries and its relation to suprasegmental cues. *JASA*, *96*, 2037-2047.

Pike, K. L. (1945). *The intonation of american english.* Ann Arbor, MI: University of Michigan Press.

Pinker, S. (1984). *Language learnability and language development.* Cambridge, Ma.: Harvard University Press.

Pinker, S. (1994). *The language instinct: How the mind creates language.* New York: William Morrow.

Pinker, S. (1995). Language acquisition. In L. Gleitman & M. Liberman (Eds.), *An invitation to cognitive science (vol. 1): Language.* Cambridge, MA: MIT Press.

Poldrack, R., Wagner, A., Prull, M., Desmond, J., Glover, G., & Gabrieli, J.

(1999). Functional specialization for semantic and phonological processing in the left inferior frontal cortex. *Neuroimage*, *10*, 15-35.

Pollack, I., & Pickett, J. (1964). The unintelligibility of excerpts from conversation. *Lang Speech*, *6*, 165-171.

Pollock, J. (1989). Verb movement, universal grammar and the structure of ip. *Linguistic Inquiry*, *20*, 365-424.

Price, C. (1998). The functional anatomy of word comprehension and production. *Trends Cogn Sci*, *2*, 281-288.

Price, C., Wise, R., & Frackowiak, R. (1996). Demonstrating the implicit processing of visually presented words and pseudowords. *Cereb Cortex*, *6*, 62-70.

Prince, A., & Smolensky, P. (2004). *Optimality theory: Constraint interaction in generative grammar*. Oxford: Blackwell.

Quine, W. (1960). *Word and object*. Cambridge, MA: MIT Press.

Ramus, F., Hauser, M. D., Miller, C., Morris, D., & Mehler, J. (2000). Language discrimination by human newborns and by cotton-top tamarin monkeys. *Science*, *288*(5464), 349-351.

Ramus, F., Nespor, M., & Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition*, *73*(3), 265-292.

Remez, R. E., Fellowes, J., & Rubin, P. (1997). Talker identification based on phonetic information. *Journal of Experimental Psychology: Human Perception and Performance*, *23*, 651-666.

Saffran, J., Aslin, R., & Newport, E. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926-1928.

Saffran, J., Newport, E., Aslin, R. N., Tunick, R., & Barrueco, S. (1997). Incidental language learning: Listening (and learning) out of the corner of your ear. *Psychological Science*, *8*, 101-195.

Saffran, J. R. (2001). Words in a sea of sounds: The output of infant statistical learning. *Cognition*, *81*(2), 149-169.

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word segmentation: The role of distributional cues. *Journal of Memory and Language*, *35*, 606-621.

Santen, J. van, & D'Imperio, M. (1999). Positional effects on stressed vowel duration in standard Italian. In *Proceedings of the XIVth International Congress of Phonetic Sciences* (p. 241-244). Aug 1-7, San Francisco, USA,

v.1.

Schacter, D., & Church, B. (1992). Auditory priming: Implicit and explicit memory for words and voices. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *18*, 915-930.

Scott, D. R. (1982). Duration as a cue to the perception of a phrase boundary. *JASA*, *71*(4), 996-1007.

Segalowitz, S., & Chapman, J. (1980). Cerebral asymmetry for speech in neonates: a behavioral measure. *Brain Lang*, *9*(2), 281-288.

Selkirk, E. (1978). On prosodic structure and its relation to syntactic structure. In T. Fretheim (Ed.), *Nordic prosody II.* Trondheim: TAPIR.

Selkirk, E. (1982). Syllables. In H. van der Hulst & N. Smiths (Eds.), *The structure of phonological representations, vol 2* (p. 337-383). Dordrecht: Foris.

Selkirk, E. (1984). *Phonology and syntax: The relation between sound and structure.* Cambridge, MA: The MIT Press.

Selkirk, E. (1986). On derived domains in sentence phonology. *Phonology Yearbook*, *3*, 371-405.

Selkirk, E. (1996). The prosodic structure of function words. In J. L. Morgan & K. Demuth (Eds.), *Signal to syntax: Bootstrapping from speech to grammar in early acquisition.* (p. 187-213). Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc.

Shattuck-Hufnagel, S., & Turk, A. E. (1996). A prosody tutorial for investigators of auditory sentence processing. *J Psycholinguist Res*, *25*(2), 193-247.

Siok, W., Jin, Z., Fletcher, P., & Tan, L. (2003). Distinct brain regions associated with syllable and phoneme. *Hum Brain Mapp*, *18*, 201-207.

Soderstrom, M., Seidl, A., Kemler Nelson, G., Deborah, & Jusczyk, W., Peter. (2003). The prosodic bootstrapping of phrases: Evidence from prelinguistic infants. *Journal of Memory and Language*, *492*, 249-267.

Starkey, P. (1992). The early development of numerical reasoning. *Cognition*, *43*, 93-126.

Steinhauer, K. (2003). Electrophysiological correlates of prosody and punctuation. *Brain Lang*, *86*(1), 142-164.

Steinhauer, K., Alter, K., & Friederici, A. (1999). Brain potentials indicate immediate use of prosodic cues in natural speech processing. *Nat Neurosci*,

2(2), 191-196.

Steinhauer, K., & Friederici, A. (2001). Prosodic boundaries, comma rules, and brain responses: the closure positive shift in erps as a universal marker for prosodic phrasing in listeners and readers. *J Psycholinguist Res*, *30*(3), 267-295.

Strangman, G., Culver, J., Thompson, J., & Boas, D. (2002). A quantitative comparison of simultaneous BOLD fMRI and NIRS recordings during functional brain activation. *Neuroimage*, *17*, 719-731.

Strik, H., & Boves, L. (1995). Downtrend in f_0 and p_{sb}. *Journal of Phonetics*, *23*, 203-220.

Swerts, M. (1997). Prosodic features at discourse boundaries of different strength. *JASA*, *101*(1), 514-521.

Swingley, D. (2005). Statistical clustering and the contents of the infant vocabulary. *Cognitive Psychology*, *50*, 86-132.

Szathmary, E., & Smith, J. (1995). The major evolutionary transitions. *Nature*, *374*(6519), 227-232.

Taga, G., Asakawa, K., Maki, A., Konishi, Y., & Koizumi, H. (2003). Brain imaging in awake infants by near-infrared optical topography. *PNAS PNAS*, *19100*, 10722-10727.

Thiessen, E., & Saffran, J. (2003). When cues collide: Use of stress and statistical cues to word boundaries by 7- to 9-month-old infants. *Dev Psychol*, *39*(4), 706-716.

Tincoff, R., Hauser, M., F, T., G.Spaepen, Ramus, F., & Mehler, J. (2005). The role of speech rhythm in language discrimination: further tests with a non-human primate. *Developmental Science*, *18*, 26-35.

Toro, J., Mattys, S., & Sebastián-Gallés, N. (submitted). Role of perceptual salience and positional information during the segmentation of speech.

Toro, J., Sinnett, S., & Soto-Faraco, S. (2005). Speech segmentation by statistical learning depends on attention. *Cognition*, *97*, B25-B34.

Toro, J., & Trobalón, J. (2005). Statistical computations over a speech stream in a rodent. *Percept Psychophys*, *67*, 867-875.

Tulving, E. (2002). Episodic memory: From mind to brain. *Annu Rev Psychol*, *53*, 1-25.

Ueyama, M. (1999). An experimental study of vowel duration in phrase-final contexts in Japanese. *UCLA Working papers in Phonetics*, *97*, 174-182.

Vaissière, J. (1995). Phonetic explanations for cross-linguistic prosodic similarities. *Phonetica*, *52*, 123-130.

Van Lancker, D., Kreiman, J., & Emmorey, K. (1985a). Familiar voice recognition: Patterns and parameters. Part II: Recognition of rate-altered voices. *Journal of Phonetics*, *13*, 39-52.

Van Lancker, D., Kreiman, J., & Emmorey, K. (1985b). Familiar voice recognition: Patterns and parameters. Part I: Recognition of backward voices. *Journal of Phonetics*, *13*, 19-38.

Vargha-Khadem, F., Carr, L., Isaacs, E., Brett, E., Adams, C., & Mishkin, M. (1997). Onset of speech after left hemispherectomy in a nine-year-old boy. *Brain*, *120*(Pt 1), 159-182.

Villringer, A., & Chance, B. (1997). Non-invasive optical spectroscopy and imaging of human brain function. *Trends Neurosci*, *20*(10), 435-442.

Wada, J., Clarke, R., & Hamm, A. (1975). Cerebral hemispheric asymmetry in human. Cortical speech zones in 100 adults and 100 infant brains. *Arch Neurol*, *32*, 239-246.

Watson, D., & Gibson, E. (2004). The relationship between intonational phrasing and syntactic structure in language production. *Language and Cognitive Processes*, *19*(6), 713-755.

Weber, C., Hahne, A., Friedrich, M., & Friederici, A. (2004). Discrimination of word stress in early infant perception: Electrophysiological evidence. *Cognitive Brain Research*, *18*, 149-161.

Wightman, C., Shattuck-Hufnagel, S., Ostendorf, M., & Price, P. (1992). Segmental durations in the vicinity of prosodic phrase boundaries. *J Acoust Soc Am*, *91*(3), 1707-1717.

Wilcox, T., Bortfeld, H., Woods, R., Wruck, E., & Boas, D. (2005). Using near-infrared spectroscopy to assess neural activation during object processing in infants. *J Biomed Opt*, *10*(1), 011010.

Wynn, K. (1996). Infants' individuation and enumeration of actions. *Psychological Science*, *7*, 164-169.

Yip, M. (2004). Possible-word constraints in Cantonese speech segmentation. *J*

Psycholinguist Res, *33*(2), 165-173.

Yu-fang, Y., & Bei, W. (2002). Acoustic correlates of hierarchical prosodic boundary in Mandarin. *Proc. Prosody 2002, Aix-en-Provence (France), April 2002*.

Zatorre, R., & Belin, P. (2001). Spectral and temporal processing in human auditory cortex. *Cereb Cortex*, *11*, 946-953.

Zatorre, R., Belin, P., & Penhune, V. (2002). Structure and function of auditory cortex: Music and speech. *Trends Cogn Sci*, *6*(1), 37-46.