THE BI-LEVEL INPUT PROCESSING MODEL OF FIRST AND SECOND LANGUAGE PERCEPTION

by

IZABELLE GRENON

BA in English Studies, Université Laval, 2003

MA in Linguistics, Université Laval, 2005

A Dissertation Submitted in Partial Fulfillment

of the Requirements for the Degree of

DOCTOR OF PHILOSOPHY

in the Department of Linguistics

University of Victoria

Supervisory Committee

THE BI-LEVEL INPUT PROCESSING MODEL OF FIRST AND SECOND LANGUAGE PERCEPTION

by

IZABELLE GRENON
BA in English Studies, Université Laval, 2003
MA in Linguistics, Université Laval, 2005

Supervisory Committee

Ewa Czaykowska-Higgins (Department of Linguistics)
Supervisor

Sonya Bird (Department of Linguistics)
Co-Supervisor

John Esling (Department of Linguistics)
Departmental Member

Jim Tanaka (Department of Psychology)
Outside Member

Abstract

Supervisory Committee

Ewa Czaykowska-Higgins (Department of Linguistics)
Supervisor

Sonya Bird (Department of Linguistics)
Co-Supervisor

John Esling (Department of Linguistics)
Departmental Member

Jim Tanaka (Department of Psychology)
Outside Member

The focus of the current work is the articulation of a model of speech sound perception, which is informed by neurological processing, and which accounts for psycholinguistic behavior related to the perception of linguistic units such as features, allophones and phonemes. The Bi-Level Input Processing (BLIP) model, as the name suggests, proposes two levels of speech processing: the *neural mapping level* and the *phonological level*. The model posits that perception of speech sounds corresponds to the processing of a limited number of acoustic components by neural maps tuned to these components, where each neural map corresponds to a contrastive speech category along the relevant acoustic dimension in the listener's native language. These maps are in turn associated with abstract features at the phonological level, and the combination of multiple maps can represent a segment (or phoneme), mora or syllable. To evaluate the processing of multiple acoustic cues for categorization of speech contrasts by listeners, it may be relevant to distinguish between different types of processing. Three types of processing are identified and described in this work: additive, connective and competitive.

The way speech categories are processed by the neurology in one's L1 may impact the perception and acquisition of non-native speech contrasts later in life. Accordingly, five predictions about the perception of non-native contrasts by mature listeners are derived from the proposals of the BLIP model. These predictions are exemplified and supported by means of four perceptual behavioral experiments. Experiments I and II evaluate the use of spectral information (changes in F1 and F2) and vowel duration for identification of an English vowel contrast ('beat' vs. 'bit') by native North American English, Japanese and Canadian French speakers. Experiments III and IV evaluate the use of vowel duration and periodicity for identification of an English voicing contrast ('bit' vs. 'bid') by the same speakers. Results of these experiments demonstrate that the BLIP model correctly predicts sources of difficulty for L2 learners in perceiving non-native sounds, and that, in many cases, L2 learners are able to capitalize on their sensitivity to acoustic cues used in L1 to perceive novel (L2) contrasts, even if those contrasts are neutralized at the phonological level in L1. Hence, the BLIP model has implications not only for the study of L1 development and cross-linguistic comparisons, but also for a better understanding of L2 perception. Implications of this novel approach to L2 research for language education are briefly discussed.

Table of Contents

List of Tables

List of Figures and Illustrations

List of Symbols, Abbreviations and Nomenclature

Symbol	Definition
AM	Amplitude-modulated component
ANOVA	Analysis of variance
β	Standardized regression coefficient
BLIP	Bi-Level Input Processing model
NB	Noise burst
CF	Constant frequency component
F1, F2, F3, F4, F5	First Formant, Second Formant, Third Formant, Fourth Formant, Fifth Formant
FM	Frequency-modulated component
L1	First language
L2	Second language
PAM	Perceptual Assimilation Model
PRIMIR	Processing Rich Information from Multi-dimensional Interactive Representations
RT	Response Time
SDRH	Similarity Differential Rate Hypothesis
SLM	Speech Learning Model
VOT	Voice Onset Time

Acknowledgments

Over the past few years, as a graduate student and researcher, I have discovered the unavoidability of Murphy's Law, including but not limited to:

If nothing can go wrong, something will. (Surely, I pushed the record button… no?)

Nothing is as easy as it looks. (Analyzing the results? Piece of cake! Pfff!)

Everything takes longer than you think. (I've been a university student for *10 years*?)

If everything seems to be going well, you have obviously overlooked something. (Not now, please, not now…)

Nevertheless, I have survived academic life so far, and have even enjoyed a great deal of it. But it would be naïve to assume that I made it to this point all by myself. Many people were there to support me, academically, financially and morally, and I do not believe it would have been possible to accomplish this work without their help. Accordingly, I would like to take the time (and space) to thank each and every one of them, though too briefly, for their non-negligible contribution to this work and to my growth, as a speech scientist, and as a person.

First of all, I would like to thank my co-supervisors, Dr. Ewa Czaykowska-Higgins and Dr. Sonya Bird, for their great patience, diplomatic criticism, and, at times, greatly needed moral support. I would also like to extend my thanks to Dr. John Esling for being on my committee, and also for providing me with the opportunity to attend different academic events in Europe and to work on a different research project while doing my Ph.D. Thanks to Dr. Jim Tanaka for generously agreeing, at the very last minute, to serve on my committee, and to Dr. Yue Wang for agreeing to serve as my

external committee member. Thanks also to Dr. Jessica Maye, for being on my committee part of the way. Thanks to all the committee members and other fellows who attended my oral examination for the interesting and challenging questions during the discussion period. Given the multidisciplinary nature of my work, it was both exciting and energizing to exchange ideas with people who bring with them different areas of expertise. I am also very grateful to Dr. Chris Sheppard and Dr. Yoshinori Sagisaka from Waseda University who facilitated the recruitment of participants and provided me with the facilities to conduct my experiments in Japan (I also thoroughly enjoyed the Christmas party, *kampai!*) Special thanks to Dr. Darlene LaCharité and Dr. Johanna-Pascale Roy at Laval University for kindly offering me access to the necessary facilities to conduct my experiments in Québec City. I would also like to extend my thanks to the secretaries at the University of Victoria, Maureen and Gretchen, for their problem solving abilities and issue resolution skills.

I want to extend my utmost gratitude to everyone in the Department of Linguistics at the University of Victoria for a wonderful academic experience. More specifically, I would like to say a special thanks to my closest friends and colleagues, Allison, Janet, and Lyra, I don't know what I would have done without you, and to Carly, Dale, Laura, Matt, Nick, Pauliina, Qian, Rebeca, Scott, Sunghwa, Thomas, Ya, and many others for their moral support, animated academic discussions, help with all sorts of things and great parties! Merci maman et Nathalie pour vos encouragements et pensées positives. And special thanks to Willy, Mommy, Bethy and Fanny for years of comforting and purring. Finally, thanks to the FQRSC, SSHRC and the University of Victoria (various departments) for their generous financial contributions.

Dedication

Dédié à la mémoire de mon père, Fernand Grenon, qui a, sans l'ombre d'un doute, contribué à l'accomplissement de cet ouvrage par sa confiance rassurante et inébranlable dans mes aptitudes académiques.

Epigraph

For the most wild, yet most homely narrative which I am about to pen, I neither expect nor solicit belief.

- Edgar Allan Poe, "The Black Cat"

Chapter One: Introduction

Language is generally regarded as one of the most distinctive features of the human species. However, the exact mechanisms used by humans for language processing remain mostly elusive. Since a better understanding of speech processing may have important implications for second language education, language pathology, speech technology and for deepening our knowledge of the functioning of the human brain, the current work attempts to bridge the gap between psycholinguistic behavior related to the perception of linguistic components (i.e. features, allophones and phonemes) and neural processing by proposing a model of speech perception informed by previously documented experimental research in neural processing.

Extensive research in the fields of phonetics, linguistics and psycholinguistics has provided valuable information about the acoustic characteristics of speech sounds and about how these characteristics are perceived by humans and other species. Recent research in the field of neurophysiology and neuroethology has provided valuable insight into the functioning of isolated neurons in response to various types of simple and complex sounds. Building on findings from both research streams, a few neural-based models have emerged that have begun to bridge the gap between neural processing and speech perception. Sussman (1986) proposed a neural-based model for vowel normalization, while Sussman and colleagues (Sussman 1999; Sussman 2002; Sussman, Hoemeke & Ahmed 1993; Sussman, McCaffrey & Matthews 1991) argued for a neural model of stop place of articulation based on locus equations, a concept shown to be consistent with descriptions of cortical organization documented in animal studies. Bauer,

Der and Herrmann (1996) and Guenther and Gjaja (1996) proposed neural-based accounts of the perceptual magnet effect—a phenomenon documented by Kuhl and colleagues (e.g. Kuhl & Iverson 1995), while Guenther and colleagues (Guenther & Bohland 2002; Guenther, Husain, Cohen & Shinn-Cunningham 1999; Guenther, Nieto-Castanon, Ghosh & Tourville 2004; Guenther, Nieto-Castanon, Tourville & Ghosh 2001) extended the neural-based account proposed by Bauer, Der and Herrmann (1996) to study the effect of type of training on the development of auditory cortical maps in the brain. Using computer simulations, this later approach was argued to be consistent with native Japanese speakers' inability to perceive the English /r/-/l/ contrast (Guenther & Bohland 2002), and therefore, to have crucial implications for the study of second language (L2) perception and acquisition.

Despite these recent contributions, there is still a considerable gap between our understanding of neural processing and perception of speech sound contrasts. This work is intended to contribute to addressing this gap by articulating a linguistic model that is neural-based in the sense that the assumptions of the model are founded upon neural processing as documented in animal studies and upon neurolinguistic experiments with humans. The main research questions this work attempts to answer are:

1. What is the possible correspondence between neural processing and linguistic concepts, such as features, allophones and phonemes?
2. How are multiple cues processed in relation to one another?
3. How does speech sound processing differ cross-linguistically?
4. How does speech sound processing in L1 impact on the perception and acquisition of non-native sounds later in life?

Although the neurolinguistic aspect of the current work is primarily theoretical, it yields important implications for future research on speech perception, and formulates specific and testable predictions, some of which were tested in four behavioral experiments reported in Chapter 4.

The proposals presented in this work build on the above-mentioned neural-based models as well as on additional findings in the fields of neurolinguistics and neuroethology. These proposals are articulated into a conceptualized model of speech perception, referred to as the Bi-Level Input Processing (BLIP) model. The BLIP model defines two distinct levels of speech processing[1]—the neural mapping level and the phonological level—which are meant to account for the fact that results of behavioral experiments may vary significantly depending on the type of task used (e.g. auditory discrimination versus picture identification) and testing conditions (e.g. inter-stimulus interval). In particular, it is demonstrated that the levels posited by the BLIP model have important implications for the study and better understanding of L2 perception and acquisition. To serve as a convenient springboard for L2 studies, the BLIP model makes specific predictions about the perception and acquisition of non-native speech contrasts. These predictions are empirically tested and supported by the results of four behavioral experiments evaluating the perception of acoustic correlates of English speech contrasts by native North American English, Canadian (Québécois) French and Japanese speakers.

---

[1] The fact that neurons are generally organized into a hierarchy with neurons at different stages performing different functions is commonly accepted in the field of neuroscience and thus, this idea is not new (see for instance the neural-based speech processing model proposed by Greenberg 2006 and the model proposed by Sussman et al. 1991). However, it appears that these levels have never been clearly defined in relation to the processing of speech sounds by humans to account for seemingly contradictory perceptual results, particularly in L2 studies.

Outline

Chapter 2 of this work describes and discusses the general assumptions of the BLIP model concerning neural processing and speech perception, based on previous behavioral and neurological experiments, and resulting models. Specifically, section 2.1 describes perceptual/behavioral research suggesting that infants may initially extract statistical distribution information from the speech input for building the speech categories relevant to the language to which they are being exposed (2.1.1). Additional experiments indicate that adults are also sensitive to statistical distribution in the input, and may be able to use this information to form new categories (2.1.2). However, other studies reveal that exposure is not always sufficient to trigger the formation of new speech categories (2.1.3) and that perception does not always exactly mirror the statistical distribution found in the input (2.1.4). Contradictory results in L1 and L2 experiments also suggest that there is likely more than one level of speech processing, but how many levels and what these levels correspond to remain unresolved issues (2.2). Section 2.3 describes the basic properties and functions of neurons that are most likely to play a role in the categorical processing of speech contrasts (2.3.1). General theories about how neurons may be organized in the human auditory cortex are presented, especially in relation to the phenomenon referred to as the perceptual magnet effect, since the way neurons are organized may greatly impact on the perception and acquisition of native as well as non-native contrasts (2.3.2). Arguments suggesting that there is sufficient invariance in the input to enable the creation of invariant parameters (or neural maps) by the neurology is discussed (2.3.3) since this issue is argued to impact on psychological percepts such as

the notion of features or phonemes, and is crucial to the foundation of the model of speech sound perception presented in chapter 3. The following section (2.3.4) provides a review of different types of neurons identified in non-human animals that are believed to be active in the human brain as well, and to play a crucial role in human speech processing. A short summary of what is currently known or assumed about the neural processing of speech categories is presented in section 2.4.

Chapter 3 presents and discusses the proposed model of speech processing which aims at capturing the link between neural processing and abstract linguistic concepts. This model is referred to as the Bi-Level Input Processing model (BLIP). Section 3.1 summarizes the assumptions and general principles of the model. Section 3.2 describes the first level of processing posited, referred to as the neural mapping level. The mechanisms of the neural mapping level are described and exemplified with the processing of fricatives (3.2.1), vowels (3.2.2), stops (3.2.3) and suprasegmental elements such as lexical stress and tones (3.2.4). Section 3.3 describes the second level of processing posited, referred to as the abstract phonological level. The interaction between the two levels of processing—from neural maps to phonological features—is exemplified (3.3.1). Hypotheses about how listeners cope with speaker variability (3.3.2) and with incomplete or misleading information (3.3.3) are also presented and discussed. Section 3.4 explains how the different levels posited by previous models to account for varying results obtained depending on task type or task condition can be reconciled within the BLIP model. Finally, section 3.5 summarizes the major claims and mechanisms posited by the BLIP model.

Chapter 4 discusses the implications of the BLIP model for the study of L2 perception and acquisition, as compared with previous models such as the Perceptual Assimilation Model (PAM) and the Speech Learning Model (SLM). The chapter begins by describing the notion of cross-linguistic perceptual similarity used by previous L2 models to evaluate L2 perception or acquisition, along with the shortcomings of this approach (4.1). Section 4.2 presents the predictions of the BLIP model for the perception and acquisition of non-native contrasts by adult language learners. The BLIP model is intended to provide a different way of looking at the difficulties encountered by language learners in L2 perception, by assessing how the processing of acoustic cues *and* the way those cues are associated with abstract percepts in L1 may interfere or help with the perception of L2 contrasts. Sections 4.3, 4.4, 4.5 and 4.6 report four behavioral perceptual experiments evaluating the perception of English sound contrasts by native North American English, Canadian (Québécois) French and Japanese speakers that support the five predictions derived from the BLIP model. Section 4.7 summarizes the predictions of the BLIP model and supporting experiments. A general discussion concludes chapter 4 in section 4.8 by summarizing the additional contributions provided by the BLIP model as compared to the neural, L1 and L2 models introduced throughout this work.

Chapter 5 provides a summary of the proposals put forward by the BLIP model (5.1), discusses the implications of the BLIP approach for second language research and education (5.2), and outlines future directions that need to be explored for further development of the BLIP model (5.3).

Chapter Two: The neural grounding of speech processing

Language is generally regarded as one of the most distinctive features of the human species. It is not yet clear, however, which mechanisms, if any, are unique or essential for language processing and development. Pertinent to the current work, speech sounds are generally characterized by a combination of spectral and timing components, such as noise bursts, spectral peaks and so on, to which both humans and various animals have been shown to be sensitive. Of particular interest, some non-human species are able to categorize several human speech sounds in a way comparable to human performance, possibly because the acoustic components used in human communication are also found in the communication system of non-human animals. Accordingly, the processing of speech sounds by humans and other animals most likely underlies similar mechanisms (e.g. types of neurons, neural function and organization) only adapted to the needs of each species. It remains to be understood, however, how those mechanisms work, and to determine in which way these are human-specific, or potentially language-specific. Using a multidisciplinary approach, I attempt to address these issues in this work by bridging the gap between what we know about human speech processing based on behavioral studies and non-invasive neurolinguistic experiments and what we have learned about neurons and neural processing from animal studies.

The first part of this chapter (2.1) presents evidence indicating that infants and adults are sensitive to the statistical distribution of acoustic components in the input used to contrast speech sounds, and that this distribution presumably shapes the way speech sounds are perceived. That does not imply, however, that the human brain is simply a

passive receiver. Factors other than input distribution play a crucial role in the development of novel speech categories, such as the listener's level of attention and type of training. In addition, the use of different testing conditions has also been found to yield divergent perceptual performance, suggesting that more than one independent level of processing may need to be accounted for in a model of speech perception.

The second part of this chapter (2.3) reviews the literature suggesting that the statistical distribution in the input might shape the neurology into neural maps or invariant parameters corresponding to coarse speech categories. The neural processing of speech sounds, however, differs from many other tasks by its specific goal to categorize rather than simply discriminate similar stimuli, and therefore, is thought to involve neural mechanisms that partly depart from those attested in discrimination tasks. This section also puts together a summary of the spectral and timing components that appear most relevant for speech perception, along with the type of neurons or neural responses tuned to these components as identified in a number of species. This information is particularly pertinent in establishing the neural grounding for the model presented in the next chapter.

To summarize the various assumptions discussed in this chapter, section 2.4 presents a short scenario illustrating the mechanisms involved in first language acquisition from a neurology point of view, followed by central questions that must be addressed and investigated. This task is tackled with the proposal put forward with the BLIP model, described, exemplified and justified in the following chapters.

2.1 Language acquisition

A wide range of studies have demonstrated that infants are born with perceptual primitives that allow them to roughly discriminate most, if not all, contrastive sounds used in human languages, a phenomenon referred to as categorical perception (Aslin, Pisoni, Hennessy & Perey 1981; Best & McRoberts 2003; Eimas 1975; Eimas & Miller 1992; Eimas, Siqueland, Jusczyk & Vigorito 1971; Kuhl 1983; Kuhl & Miller 1975b; Liberman, Harris, Hoffman & Griffith 1957; Liberman, Harris, Kinney & Lane 1961; Tsao, Liu & Kuhl 2006; Werker & Lalonde 1988; Werker & Tees 1984; etc.) These primitives, however, are not restricted to human infants. The ability to discriminate frequency-related components (e.g. pure tone contrasts) and temporal features (e.g. Voice-Onset-Time) in a way comparable to humans has been observed in non-human animals (*chinchilla* = Kuhl 1981; Kuhl & Miller 1975a, 1978; *monkey* = Kuhl & Padden 1982, 1983; Sinnott & Brown 1997; Sinnott, Brown & Borneman 1997; *quail* = Kluender, Diehl & Killeen 1987). In addition, the perceptual mechanisms used for categorical perception by human infants are not restricted to the discrimination of speech sounds. Categorical perception has been observed with non-speech sounds (e.g. Jusczyk et al. 1977), as well as in other modalities, including spatial representations (Quinn 2004), colors (Franklin & Davies 2004; Franklin, Pilling & Davies 2005), shapes (Catherwood, Crassini & Freiberg 1989), and facial discrimination (Webster, Kaping, Mizokami & Duhamel 2004).

The perceptual categorical boundaries of speech sounds are not fixed, but rather, altered or refined as newborn infants are exposed to a specific language during the first months of life (see Kuhl 2007 for an overview). Experiments with adult L2 learners

suggest that categorical perception continues to be alterable throughout the life span (e.g. Maye & Gerken 2000, 2001; Wang & Munro 2004). How these changes are achieved is not yet fully understood, but empirical evidence highlights the possible role of various factors and levels of processing, which will be discussed in turn in the following subsections.

2.1.1 Infants' sensitivity to statistical distribution

To learn their first language, infants must be able to extract relevant information from the continuous speech stream. Although infant-directed speech (a.k.a. motherese) may sometimes consist of short, simple phrases spoken at a relatively slow speech rate compared to normal adult speech, infants must still deal with multiple strings of sounds that usually lack well-defined pauses or other acoustic cues denoting segment or word boundaries. Various studies conducted over the past two decades point to infants' computational abilities, which may facilitate language acquisition (Anderson, Morgan, & White 2003; Aslin, Saffran, & Newport 1998, 1999; Maye 2000; Maye & Weiss 2003; Maye, Weiss & Aslin 2008; Maye, Werker & Gerken 2002; White, Peperkamp, Kirk & Morgan 2008). For instance, infants have been shown to be able to segment the continuous speech stream into pseudo-lexical items by computing statistical information related to the transitional probability of syllables (Aslin, Saffran & Newport 1998, 1999; Saffran, Aslin & Newport 1996; Saffran, Newport & Aslin 1996). Aslin, Saffran & Newport (1998) presented 15 eight-month-old infants with random sequences of four synthesized trisyllabic nonsense words (e.g. *pabiku, tibudo, golatu, daropi*), presented in a continuous loop without any pauses or other acoustic cues to word boundaries. The

string of randomized words can be exemplified as: *pabikugolatudaropitibudodaropi* [...]. The assumption for this experiment is that syllables that form a word will appear more consistently together in the input than syllables across word boundaries. After only three minutes of familiarization using this procedure infants exhibited significant looking time preferences for combinations of syllables that appeared consecutively in the unsegmented string (e.g. proto-words) than to the actual nonsense words. As with other experiments with infants, longer looking time generally indicates that infants perceive the token as a novel item. These results appear to provide evidence for infants' ability to compute statistical distribution in speech segmentation tasks. An experiment by Gerken, Wilson & Lewis (2005) further showed that infants can use distributional cues to form syntactic categories. A series of experiments conducted by Maye and colleagues (2002, 2003, 2008), reported below, revealed that infants' sensitivity to statistical distribution extends to acoustic categories potentially relevant for language-specific speech contrasts as well.

Newborn infants' natural ability to perceive speech sounds categorically is altered after only a few months of contact with the language to which they are exposed. Infants become attuned to the sounds of their native language by six months for vowels and ten months for consonants; at these points, they also lose the ability to distinguish non-native contrasts (Kuhl 1993a, 1993b; Kuhl, Stevens, Hayashi, Deguchi, Kiritani & Iverson 2006; Kuhl, Williams, Lacerda, Stevens & Lindblom 1992; Tsushima et al. 1994; Werker & Tees 1984; etc.) Maye hypothesized that exposure to a unimodal (i.e. non-contrastive) distribution of a given acoustic cue would inhibit listeners' perception of a contrast, whereas exposure to a bimodal (i.e. contrastive) distribution of the same cue would enhance perception of the same contrast. Maye, Werker and Gerken (2002)

experimentally tested this hypothesis with 24 six-month-old and 24 eight-month-old infants from English-speaking homes. The infants were presented with tokens along a [da] - [ta] continuum that varied in terms of prevoicing duration and the first and second formant transitions into the vowel (since none of the sounds were aspirated, this contrast differs from the one used in English). Half of the infants were presented with a bimodal distribution of the tokens as represented by the dotted line in Figure 2–1, whereas the other half were presented with a unimodal distribution, illustrated by the plain line in the same figure.

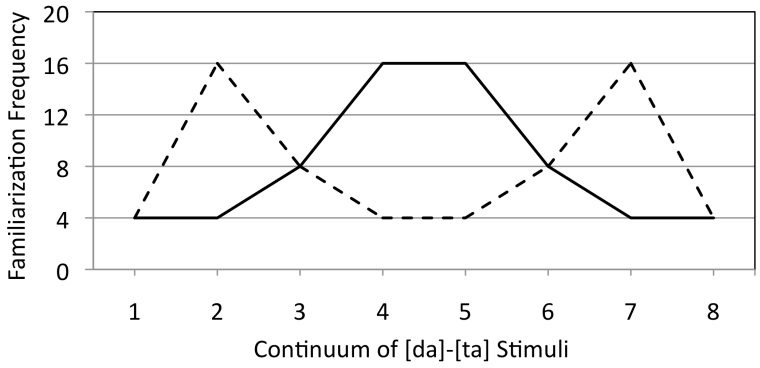

Figure 2–1 Bimodal vs. Unimodal distribution of [da]-[ta] stimuli during familiarization. The continuum of speech sounds is shown on the abscissa, with Token 1 corresponding to the endpoint [da] stimulus, and Token 8 the endpoint [ta] stimulus. The ordinate axis plots the number of times each stimulus occurred during the familiarization phase. The presentation frequency for infants in the Bimodal group is shown by the dotted line, and for the Unimodal group by the solid line. (Figure reproduced from Maye, Werker & Gerken 2002: B104).

For instance, token 1, which corresponds to the endpoint [da] stimulus, was presented four times in each of the distributions, whereas token 4, which corresponds to

an intermediate value between [da]-[ta], was presented only four times in the bimodal distribution but 16 times in the unimodal distribution context. Each infant was therefore presented with an equal number of stimuli, but the distributional frequency of those stimuli diverged according to the type of distribution the infant was exposed to.

After familiarization with one distribution of stimuli, which lasted two minutes, infants were tested on their ability to discriminate stimuli 1 and 8 in a series of alternating and non-alternating trials. A significant effect of the distribution condition (i.e. unimodal vs. bimodal) was observed irrespective of age group, indicating that infants at both six and eight months are sensitive to the statistical distribution of acoustic cues for sound discrimination after only two minutes of exposure to this distribution. A similar experiment conducted by Maye & Weiss (2003) and Maye, Weiss & Aslin (2008) with eight-month-olds provided further evidence for infants' sensitivity to distributional information, by showing that infants could not only apply this information to the discrimination of a previously difficult contrast (e.g. [da]~[ta]), but also that they could extend this ability to an untrained contrast that exhibited the same acoustic feature (e.g. [ga]~[ka]).

2.1.2 Adults' sensitivity to statistical distribution

Sensitivity to statistical distribution is not restricted to early infancy, but appears to persist into adulthood. In acoustic experiments, after nine minutes of exposure to a bimodal distribution, English-speaking adults were able to discriminate allophonic contrasts ([d] as in *day* from [t] as in *stay*) (Maye & Gerken 2000, 2001) that are not generally perceived categorically by English speakers (Pegg & Werker 1997).

Importantly, adults' ability to perceive the novel acoustic contrast after training in the bimodal distribution condition was achieved without the use of minimal pairs (i.e. the syllables used for the previous experiments were not associated with any semantic contrast). Training experiments using manipulated (Iverson, Hazan, & Bannister 2005) and non-manipulated (Bradlow, Akahane-Yamada, Pisoni, & Tohkura 1999; Bradlow, Pisoni, Akahane-Yamada, & Tohkura 1997; Logan, Lively, & Pisoni 1991) minimal pairs contrasting English [r] and [l] produced by various native English speakers demonstrated that adult native Japanese speakers could improve their perception of this non-native contrast, even though the ability to discriminate those sounds has been shown to dramatically decline around ten months in Japanese infants (Kuhl et al. 2006). Similar results were obtained for the discrimination of English vowel spectral contrasts (changes in F1 and F2), as perceived by native Mandarin speakers after extensive computer-based training with the English vowels (Wang & Munro 2004). Hence, categorical perception appears to remain alterable throughout the life span as long as adults are exposed to the appropriate contrastive distributional pattern.

In sum, infants and adults can learn to categorize speech sounds after a relatively short exposure to a contrastive statistical distribution of the acoustic components, even when these components are not presented in minimal pairs nor participants explicitly told that the sounds presented are contrastive. The behavioral studies summarized in this section demonstrate that speech categories can be formed prior to lexical acquisition, and therefore, are likely embedded in neural organization without necessitating prior lexical encoding.

2.1.3 When exposure is not enough

Although exposure to contrastive statistical distribution may trigger changes in the perception of sound categories in a controlled laboratory setting, as illustrated above, simple exposure to the natural environment in which the categories are contrasted is not necessarily correlated with better discrimination, at least in the case of adult L2 learners (Grenon 2006). Some studies emphasize the role of attention for successful statistical learning (e.g. Toro, Sinnett, & Soto-Faraco 2005), while other studies ensured participants' attention was directed to listening to the statistical distribution in their experimental design by asking adults to check an empty box on a sheet of paper for each word they heard (Maye & Gerken 2000, 2001) or by presenting a short video clip to children while delivering the auditory training stimuli[2] (Maye & Weiss 2003). Hayes-Harb's experiment with adults (2007) showed that the use of minimal pairs in the training task leads to better perceptual accuracy of a novel contrast than statistical information alone. That is, L2 learners appear to perform better on the learning task if supplemented with meaningful semantic information emphasizing the need for categorical distinction of the L2 contrasts.

Crucially, the type of training may also impact categorical perception. Guenther, Husain, Cohen & Shinn-Cunningham's (1999) perceptual experiment compared *discrimination* training of a series of narrow-band filtered samples of white noise with different center frequencies that were not perceived categorically prior to the experiment,

---

[2] The videoclip presented during the training session only present visual information, while the stimuli are delivered as the only auditory input. Hence, the videoclip and audio stimuli tap into two different modalities at the same time.

with *categorical* training of the same set of stimuli. Discrimination training requires listeners to distinguish tokens within a given category. Consequently, based on Bauer, Der & Herrmann's (1996) model, discussed in section 2.3, the researchers predicted that such training would improve listeners' ability to discriminate small differences within that category. Conversely, categorical training requires listeners to ignore differences between tokens within a given category. Accordingly, this type of training was predicted to lessen listeners' ability to discriminate tokens within that category. Participants assigned to the discrimination condition were indeed found to be better at discriminating the stimuli after training, while participants in the categorical condition became worse at discriminating the same set of stimuli, even though the statistical distribution of stimuli used during training was the same in both conditions.

Although infants and adults are sensitive to statistical distribution of relevant information for categorical perception of acoustic and phonemic elements, sensitivity to distributional information alone fails to explain the whole story when it comes to speech learning and processing, such as the fact that differences in discrimination of a novel contrast may depend on the type of training to which learners have been exposed – discrimination or categorical training; with or without minimal pairs; etc.

2.1.4 When perception does not mirror statistical distribution

To the extent that perception mirrors the distribution of acoustic attributes in the input, one would expect that more frequent attributes should be perceived more easily and with higher accuracy than low frequency attributes. A study by Tucker and Warner (2007)

evaluated the perception of reduced and unreduced American English flap,[3] as in the word 'puddle', by thirty native American English speakers. In a previous study, the reduced form used in the experiment was found to occur more frequently in the daily use of American speakers than the unreduced form (Warner & Tucker 2007). Yet, participants in Tucker and Warner's study encountered greater difficulties in identifying the reduced flap (the most frequent form), as reflected in longer response times and less accuracy than for the unreduced form (the less frequent form). Hence, the frequency of occurrence of a given acoustic value in the input is not necessarily positively correlated with better perception, unlike the findings in experimental settings as discussed in the previous sections.

A study by Goldstein, Nam, Kulthreshtha, Root & Best (2008) compared the distribution of tongue tip articulations of coronal stops in English and Hindi, which presumably impact their acoustic realization. A female Hindi speaker was recorded reading a story while her tongue movements were tracked and measured. The distribution of tongue movements in the Hindi speaker was compared with data from English speakers drawn from the Wisconsin X-ray database. The English data revealed no bimodal distribution in the production of the English coronal stop, as illustrated in Figure 2–2. The Hindi data exhibited a sharp distribution for production at the tongue tip (advanced), but the distribution of the retracted form, which corresponds to the retroflex stop in Hindi, was more uniform across the retracted region, as shown in Figure 2–2.

---

[3] An unreduced flap is defined by Tucker and Warner (2007) as having a burst, a clear stop closure, and a large drop in intensity, whereas a reduced flap is defined conversely as having no clear burst or closure boundaries, and only a small dip in intensity. In both cases, the formants continue throughout the flap.

Assuming that the distribution in tongue tip articulation impacts accordingly on the statistical distribution of the acoustic characteristics of the stop produced, this study suggests that although the Hindi input does not replicate the clear bimodal distribution used in laboratory experiments, Hindi speakers succeed in creating two categories presumably based on an acoustic distribution similar to the one shown in Figure 2–2.

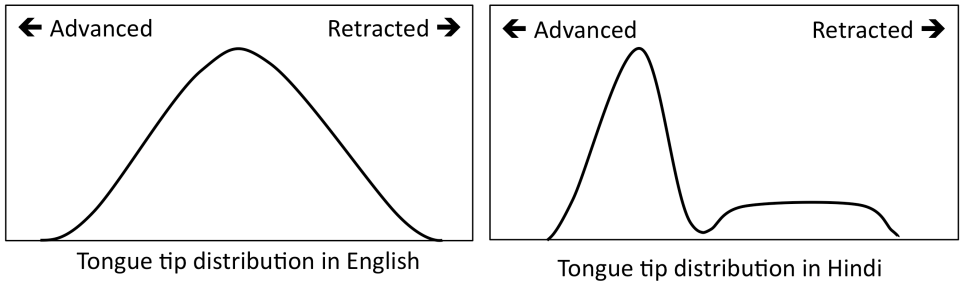

Figure 2–2 Adapted graphical representation of the histogram distribution of tongue tip horizontal positions in Hindi and English reported in Goldstein et al. 2008 (see original paper for accurate values).

The point is that perception performance does not necessarily reflect the input distribution; some studies have shown the role of attention and type of training on perceptual learning as discussed previously. Furthermore, the input distribution may be impoverished, and yet, humans are capable of forming distinct speech categories based on this input. Accordingly, the brain appears to be actively engaged in the learning process, rather than a mere passive receiver. This may have important implications especially for second language acquisition and education, as discussed in more details in chapter 4. The testing conditions and type of task used in experimental settings have also been shown to affect perceptual results. The next section presents some psycholinguistic

models that have endeavored to capture these facts by positing different levels of speech processing.

2.2 How many and what kind of levels of speech processing are there?

Thus far, it has been shown that although the statistical distribution in the input appears to be crucial for the development of speech categories during both L1 and L2 acquisition, other factors, such as the type of training and the use of minimal pairs may also play an important role in the development of these categories. In addition, task type (e.g. ABX discrimination task vs. picture identification task) and task conditions (e.g. changes in inter-stimuli interval) used in the experimental settings have been shown to trigger different responses, presumably because these factors tap into different levels of speech processing. Accordingly, various levels[4] of speech processing have been posited by previous linguistic models to account for different experimental results obtained by varying either the task conditions or type of task. A brief review of some of these proposals along with their respective justification is presented below, and serves to justify the fact that speech processing is best captured by positing two levels of speech processing (in addition to lexical encoding). The Bi-Level Input Processing model

[4] The term *factor* (Werker & Logan 1985), or *plane* (Werker & Curtin 2005) is sometimes preferred to *level,* presumably because the latter suggest a hierarchical organization among the different levels (either bottom-up or top-down). The term *level* in this subsection is used, for lack of a common term, as a generic term to denote that something is happening at a given stage without implying that the processing that takes place at a given level must occur before or after another level. Processing of different levels may occur concurrently. However, the term *level* in the BLIP model does imply a bottom-up (i.e. hierarchical) processing following a biological hierarchy.

proposed in the following chapter is meant to reconcile the different proposals described in this section by providing a neural-grounded account of these levels.

Table 2–1 exemplifies divergent, though not mutually exclusive, speculations about the kind and number of levels involved in speech processing. Although this list is non-exhaustive, it suffices to introduce concepts related to the need to posit different levels of processing in the first place, and to tackle the debate of how many levels a model of speech processing should include. Admittedly, Table 2–1 fails to do justice to the listed proposals; even though two levels may appear on the same row, they are usually dissimilar in non-negligible respects. A more detailed description of the levels proposed by each contributor is provided subsequently. Notwithstanding this limitation, general observations can be drawn by grouping these proposals into a comparative table. First, all the proposals include at least two levels of processing, though it is not entirely clear in Exemplar-based models, as represented here by Pierrehumbert's work in 2001, if the acoustic level and lexical level are really separate in those models (discussed below). Second, nearly all the proposals posit a level of representation for lexical items (although Werker & Logan did not specifically propose a lexical level in their 1985 paper, their data do not preclude the inclusion of one). Third, none of the proposals agree on the term assigned to the first level posited, labeled as *auditory*, *surface*, *acoustic*/*phonetic* or *general perceptual*. Incongruence in the labels associated with the first level of processing also reflects different views about the kind of processing achieved at this level, mostly related to the behavioral/perceptual data it was posited to account for in the respective studies. Fourth, the level traditionally referred to as *phonemic* has been the center of some controversy; many researchers have questioned the need for its existence

in a model of speech processing. Simple exemplar-based models, for instance, traditionally do not include a phonemic (or phonological) level. Recently, however, this view has been challenged, as discussed below.

Table 2–1 Speculations about the levels/factors/planes involved in speech processing

	Reference			
Processing of:	Werker & Logan (1985)	Curtin, Goad & Pater (1998)	Pierrehumbert (2001)	Werker & Curtin (2005)
Fine acoustic details	Auditory		Acoustic/ phonetic	General Perceptual
Categorical acoustic information	Phonetic	Surface		
Abstract segmental information	Phonemic			Phonemic
Abstract lexical/morphemic information		Lexical	Lexical	Word Form

Except for the levels posited to process lexical or morphemic information, most other levels posited by the different models summarized in Table 2–1 aim at capturing humans' percepts of sound contrasts, whether as allophones or phonemes. Linguistic descriptions of languages traditionally include a compilation of a language's phonemic inventory along with the possible allophonic variants that occur in the language. Both phonemes and allophones are concepts that refer to a sound *category* since the acoustic realization of speech sounds is not clearly delineated; each phoneme and allophone may encompass an infinite number of variants resulting from linguistic, individual, or sociolinguistic factors. Yet, listeners are able to ignore those variations and classify sounds into discrete categories.

Werker & Logan (1985) conducted a series of experiments comparing the perception of consonant contrasts by native English adult speakers. The stimuli for their experiment were a set of within- and between-category variants of the Hindi voiceless dental and retroflex stops. The dental and retroflex stops are used in Hindi to distinguish minimal pairs, but these sounds are not used contrastively in English. Stimuli were presented in three types of pairs: (1) physically identical instances; (2) Hindi within-category variants; and (3) (Hindi) between-category variants. The experiment also tested three inter-stimulus intervals (ISI): 250ms, 500ms and 1500ms. Participants in each of the ISI conditions had to judge whether the syllables containing those sounds were the same or different by completing an AX discrimination task. English participants exhibited a significant effect of ISI and type of pairing. The stimulus pairs corresponding to Hindi between-category variants were perceived as more similar as the duration of ISI increased, whereas stimulus pairs corresponding to identical stimuli and Hindi within-category variants were generally perceived as more dissimilar as the ISI increased. That is, each ISI in their experiment triggered a change in perception of at least one of the stimulus-pair. Accordingly, based on results showing that ISI conditions affected performance differentially, the researchers proposed three levels of processing: *auditory*, *phonetic* and *phonemic*. In Werker & Logan's (1985) hypothesis, the phonemic level corresponds to listeners' ability to distinguish acoustic characteristics of speech sounds that are contrastive in their own language; the phonetic level corresponds to listeners' ability to distinguish acoustic distinctions that are not phonemic in their language but that are phonemic in other languages; and the auditory level is the ability to discriminate differences that are not contrastive in any language.

More than a decade later, Curtin, Goad, & Pater's study (1998) argued for two main levels of processing, the so-called *surface* level and the *lexical* level, to account for divergent results obtained in their study of English and French speakers. In this study, English and French listeners responded differently depending on the type of task used for distinguishing the three-way voiced-plain-aspirated contrast in Thai. In a picture identification task where participants were presented aurally with one word and had to choose which of two pictures the word referred to, English and French speakers both performed better on the voiced-plain contrast – the contrast that is phonemic in their native language – than on the plain-aspirated contrast. In the second condition, an ABX task, participants only heard three words (no picture was presented). The first two words were different, and the participants had to decide if the third word was closer to the first or second word; each of the three words was uttered by a different speaker. In this task, English speakers performed equally well on the plain-aspirated contrast and the voiced-plain contrast, presumably because they were able to use their sensitivity to variations that occur at the allophonic level in their L1 to perceive the Thai plain-aspirated contrast. French speakers, on the other hand, still performed better on the voiced-plain contrast than on the plain-aspirated contrast, presumably because French lacks any plain-aspirated contrast at the phonemic or allophonic level. The authors reasoned that English and French listeners were probably using their lexical level of representation to complete the first picture-identification task. In the ABX condition, the L2 listeners could rely on their sensitivity to surface allophonic variations used in their L1 to perform the task, since in this condition, listening to words without any pictures would not necessarily entail lexical access. The surface level posited by the researchers in this paper corresponds to neither

the auditory nor the phonetic level posited by Werker & Logan (1985). Rather, the surface level corresponds to *allophonic* realization used in the speaker's *native* language, whereas Werker & Logan's phonetic level corresponds to *phonemic* categories used in *other* languages.[5] The lexical level posited by Curtin et al. was presumed to encode phonemic information. Hence, under this view, having a phonemic level in addition to a lexical level is unnecessary, as is the case in simple exemplar-based models discussed below (note that the model developed by Werker and Curtin in 2005 does include both a lexical level and a phonemic one, as discussed shortly).

Many researchers, particularly in the field of psycholinguistics, argue (or argued) that phonology is an artifact of lexical representations (i.e. phonology and phonological rules are not represented separately from the lexicon). The most influential exemplar-based models, such as the one described in Pierrehumbert (2001), do not assign a specific role to phonology[6]. Exemplar-based models that do not posit a distinct phonological level of processing are referred to here as "simple" exemplar models, following the terminology used by Pierrehumbert (2006). In this framework, lexical items are stored directly in the cognitive system with all their acoustic details. Exemplars are grouped according to their acoustic similarity in the cognitive perceptual space; similar exemplars are mapped together and appropriately labeled. Within this approach, phonological information can be inferred by analogy from the grouping and distribution of exemplars

---

[5] The speculation by Werker and Logan that listeners are sensitive, under some testing conditions, to phonemic contrasts in other languages has generally been taken as supporting the idea that human infants are born with universal perceptual categories that may or may not be activated with exposure to a given language (e.g. Brown 1997).

[6] However, see Pierrehumbert (2006) for a discussion of the need to posit a phonological level, and Pierrehumbert (2002) for proposals adding a phonological level to exemplar-based models.

across the cognitive perceptual space. Hence, in this model, the exemplars are the mechanisms by which speech input is processed.

Exemplar theory has crucially contributed to the modeling of frequency and gradiency effects, and has provided a plausible account for sociolinguistic factors and individual-specific variance (see Pierrehumbert 2006 for a review). Frequency refers to repeated occurrence of the same exemplar, whereas gradiency refers to the acoustic variability in the realization of exemplars within the same category. Exemplars are assumed to be stored with their acoustic details, thus accounting for listeners' ability to draw upon this information to discriminate indexical information associated with specific voices, genders, dialects, etc. For instance, when a listener perceives an exemplar uttered by a given voice, this perception will then activate previous exemplars uttered by the same speaker by assigning more *weight* to the exemplars previously perceived as belonging to the same individual. Exemplar models are also successful at accounting for word frequency effects that might be related, at least in the case of production, to processes such as lenition and deletion (Pierrehumbert 2001). Bybee (2000) noticed, for instance, that schwa reduction before /r/ or /n/ occurs more systematically in high frequency words such as *every* and *evening* than in low frequency words such as *mammary* and *artillery*.

The fact that direct exposure is not readily correlated with better perception or that perception does not, in some cases, directly mirror the input distribution may appear to counter exemplar-based models. However, exemplar theory takes into consideration the role of other cognitive factors for the organization of exemplar clusters, such as the role of attention and memory. Nonetheless, there are still some discrepancies in speech

processing that simple exemplar models are unable to capture. One of these discrepancies is discussed in Pierrehumbert (2006) and can be summarized as follows: Lexical neighborhood density, defined as the number of words that are minimally different from a given real word, is generally correlated with phonotactics probability, where words that have many close neighbors generally also exhibit high-probability phonotactics (as a general example, frequent words often exhibit the very common CV syllable pattern, as opposed to the low probability CCCVC syllable pattern). However, these two variables – lexical neighborhood density and phonotactics probability – were found to correlate with speech processing in opposite directions. Studies conducted by Vitevich and Luce (1998) and Vitevich, Luce, Pisoni, & Auer (1999), in which these conditions were independently varied, revealed that words with high-probability phonotactics are recognized faster, whereas words with many competitive neighbors are recognized more slowly. Although this may appear intuitively logical, from the point of view of simple exemplar-based models, which relies exclusively on frequency effects,[7] this outcome cannot be accounted for since these models predict that both words with high-probability phonotactics and words with high neighborhood density should be perceived relatively fast since these two factors are correlated in terms of frequency of occurrence. To reconcile the two phenomena, Pierrehumbert (2006) argues that hybrid models, which include a level devoted to processing the phonology separately, are needed. Werker & Curtin (2005) presented such a model with PRIMIR.

---

[7] Gradiency effects are irrelevant to this particular scenario.

The PRIMIR model of speech perception was designed to account specifically for seemingly contradictory results obtained in various studies with infants at different developmental stages (i.e. L1 acquisition). Mainly, the model highlights how 14-month-old infants' ability to discriminate minimally different words in a picture identification task depends on the exact procedure involved in the task and the familiarity of the infants with the words used. Hence, the three levels (planes) posited by the authors – *General Perceptual*, *Phonemic*, *Word Form* – propose a disconnection between lexical access and phonemic knowledge. The general perceptual level in PRIMIR is assumed to encode more detailed acoustic information than any of the previously described levels posited by the other models. Specifically, the general perceptual level in PRIMIR deals with the perception of detailed phonetic and indexical information, and also captures phonetic contrasts, to which they refer to as *phonetic features* (or *general perceptual features*). The phonemic level in PRIMIR deals with the encoding and perception of entire phonemic representation. No role is explicitly given to phonological features (e.g. voice) in this model since, as argued by Werker and Curtin (2005), phonetic features (e.g. VOT) coupled with the distribution of (lexical) exemplars provide sufficient information necessary to extract phonemic contrasts. The word form level in PRIMIR encodes information pertaining to entire words that may be associated with a meaning. Based on the specifics of the language task and the degree to which the child is paying attention, one or more of the levels will be activated to perform the task according to PRIMIR.

Although PRIMIR presents many advantages derived from its exemplar-based approach, the exact functioning of its phonemic level still appears to depend mostly on the categorization and labeling of stored exemplars. As a result, PRIMIR is unable to

account for the issue of how different types of training – discrimination versus categorization training – can yield differences in perception if the two training paradigms use the same input distribution (i.e., the same sets of exemplars), as discussed in section 2.1.3, since in the PRIMIR model the acquisition of speech categories only takes into consideration the words distributional frequency.

The psycholinguistic models such as the ones described above have made significant contributions to the study of language processing. However, these models still lack grounding in the neural mechanisms that underlie speech processing. In this dissertation, I propose a model of speech processing informed by neural processing as described in the following sections that can also account for the psycholinguistic behaviors described in the previous sections.

2.3 Neural processing of acoustic cues

The processing of acoustic stimuli involves the transmission, transformation and coding of information pertaining to the acoustic signal by various structures and sets of nerve cells in the inner ear (specifically in the cochlea), auditory nerve, and cortical regions of the brain. At each of these steps, the acoustic signal is processed differently, and may be transformed, in the sense that informational details pertaining to an acoustic stimulus might be either enhanced (e.g. frequency information related to vowel identification) or reduced and ultimately lost (e.g. background noise). The mechanisms involved at the different steps are also diverse. The cochlea, for instance, captures frequency and timing-related information through displacement of hair cells distributed along the basilar membrane, whereas in the cortex, information pertaining to spectral information is

captured by populations of neurons tuned to specific acoustic components, while timing information may be partly conveyed by discharge rates of those neurons (see Eggermont 2001 for a review). While the functioning of the cochlea and auditory nerve is fairly well documented and understood, the exact functioning of the auditory cortex and related areas in the human brain is still mostly unknown (Nelken 2008).

Important pieces of information relevant to understanding how speech is processed come from perceptual or behavioral studies, as described in the previous section. Other important pieces in the puzzle of speech processing have been claimed to come from neural experiments with non-human subjects. While experiments conducted on various (non-human) species provide valuable information pertaining to individual neuron responses to specific acoustic cues common to both human and animal vocalizations (e.g. burst noises, frequency components), the fact that the communication systems used by animals and humans are not readily comparable imposes caution in generalizing such results to the human brain.

A further consideration is that in many studies, the animals were under anesthesia during data collection (Angelo & Moller 1990; Clarey et al. 2004; Krebs, Lesica & Grothe 2008; Krishna & Semple 2000; Langner & Schreiner 1988; Palombi, Backoff & Caspary 2001; Razak & Fuzessery 2006; Rees & Moller 1983; and others), although a growing number of studies have been conducted with animals which were awake (Bendor & Wang 2005; Langner 1983; Langner, Albert & Briede 2002; Phan & Recanzone 2007; Qin, Chimoto, Sakai & Sato 2004) or on human patients using intracranial electrodes (Bitterman, Mukamel, Malach, Fried & Nelken 2008). As discussed by Langner, Albert and Briede (2002), in some cases, such as the periodicity coding of high frequencies,

neural responses differ depending on whether the animal was under anesthesia or awake at the time of testing. More fundamentally, it is questionable to what extent animal perception of human speech sounds reflects the processing that occurs in the human brain, given that the auditory cortex is tuned specifically to distinguish species-specific stimuli (Suga 2006). Hence, experiments on people using natural speech stimuli should provide the most accurate and reliable information about how speech correlates are processed by humans.

Experimental designs providing the most definite knowledge about neuronal responses in the auditory cortex, which involve complex surgical procedures, including removal of part of the animal's skull (e.g. Clarey et al. 2004; Langner, Albert & Briede 2002), cannot be conducted on human subjects for obvious ethical reasons. Experiments using a range of recently available non-invasive technologies, mainly positron emission tomography (PET), functional magnetic resonance imaging (fMRI), electroencephalography (EEG), which measures event-related potentials (ERP), and magnetoencephalography (MEG), which measures event-related magnetic fields (ERF), can potentially provide a better understanding of the underlying neural mechanisms responsible for categorical perception of speech sounds. However, these techniques are not without their limitations. PET and fMRI, for instance, measure brain activity by recording properties of blood changes – the former measures changes in blood flow, while the latter measures changes in oxygen content – which are known to be closely correlated with neural activity (Buckner & Logan 2001). EEG records the electrical activity produced by the firing of neurons, whereas MEG records the magnetic field

resulting from this electric activity (Gallen, Hirschkoff & Buchanan 1995). None of these techniques, however, provides direct information about the activity of isolated neurons.

Keeping these limitations in mind, however, behavioral and neural experiments, taken together, serve as a reasonable springboard to speculate about the underlying mechanisms of speech perception. In this section I present a general overview of the types of neurons and neural organization hypothesized by various researchers to play a role in human speech processing, and extrapolate on possible further implications of these findings. It is not my intention to present a complete and detailed overview of the matter here. Rather, I will focus on the research findings that are most relevant in accounting for the behavioral data reported in previous subsections, and to describing the cornerstones necessary for articulating the model presented in the next chapter.[8]

2.3.1 Neural properties and functions

The exact role of the primary auditory cortex is still debated, but it is generally agreed that higher levels of speech processing, particularly of sound categories or contrastive features, must somehow be encoded in cortical regions (Nelken 2008). Thus, the properties and functions of neurons and their organization in the auditory cortex and related areas are of particular interest to the current work, since they might uncover the key to categorical perception of speech sounds.

---

[8] I attempted, to the extent possible, to present the concepts in this section using terms accessible to the non-specialist. Hence, some apparent discrepancies may surface between formal neurological descriptions (and terms used in that field) and the description provided here.

A neuron is defined as a cell in the nervous system that can transmit information to another cell. Neurons communicate with one another via synaptic connections. The organization of neurons in the brain may be compared to the intricate threads of a gigantic canvas, each neuron having on average 7,000 synapses, for a total of about 100 to 500 trillion synapses in the adult human brain. This total is a significant decrease from the 10^{15} (1 quadrillion) synapses estimated to exist in the brain of a three-year-old child (Drachman 2005). Although the number of neurons present at birth presumably remains equal throughout adulthood (implying they cannot be replaced if they are damaged), synaptic connections can be altered. One might be tempted to speculate that this decrease in synaptic connections could partly account for the refined tuning to a limited set of speech categories that occurs early in life and the progressive "loss" of perceptual ability in distinguishing non-native speech sound categories (e.g. Sussman 1986). Even if this is the case, loss of synaptic connections does not entail that learning new categories or skills is impossible for the adult learner, since it is known that new synaptic connections can be created throughout the life span (Rose 2008).

Based on extensive work carried out on the auditory system of bats (O'Neill & Suga 1979, 1982; Suga 1969, 1973, 1978; Suga & O'Neill 1979; Suga, O'Neill, Kujirai & Manabe 1983; Suga, O'Neill & Manabe 1978, 1979) and other animals (Feng, Simmons & Kick 1978; Margoliash 1983; Mudry, Constantin-Paton, & Capranica 1977), Suga (1982, 1988, 2006) has made a series of observations and speculations about the types and functions of neurons in the central auditory cortex and its periphery. First, it has been extensively documented that neurons are fine-tuned to a given stimulus, such as a constant frequency component or combination of frequencies. Of particular relevance, a

recent study with human subjects found that neurons in the auditory cortex have ultra-fine frequency tuning — far narrower than that described in most studies with mammals (with the exception of bats) when exposed to speech input (Bitterman et al. 2008). The importance of auditory cortex mechanisms for processing fine-grained frequency information is intuitively evident when one thinks, for instance, about the processing of vowels, which differ from one another primarily by variations in their frequency components.

Second, based on neuroethological[9] data, Suga draws attention to the species-specific nature of neurons (and their tuning) in the auditory cortex, a specialization that presumably stems from the specific auditory behavioral needs of different species. For instance, significant variations in response properties of range-tuned neurons were found across different species of bats (O'Neill 1995). Hence, the human auditory cortex is expected to have its own finely tuned sets of neurons, and it is worth considering whether these neurons – or the patterns of synaptic connections that bind them – also differ cross-linguistically (this issue is discussed in more details in the next chapters).

Third, Suga proposes a distinction between the neural processing of *information-bearing elements* (IBEs) and *information-bearing parameters* (IBPs). She hypothesizes "that the central auditory system has specialized neurons tuned to each of the three types of IBEs [discussed below] for the processing of species-specific complex sounds (Suga 2006: 159 from Suga 1973)." The three types of IBEs Suga refers to are basic acoustic

---

[9] Neuroethology is a multidisciplinary field integrating neurobiology (study of nervous system) with ethology (study of animal natural behavior) and is devoted to elucidating how the processing of stimuli in the central nervous system impacts instinctive or innate behavior (e.g. speech perception).

correlates observed in both animal and human vocalizations: constant frequency (CF) components such as vowel formants, frequency-modulated (FM) components such as formant transitions,[10] and noise burst (NB) components like those produced with the release of a stop consonant. Thus, IBE neurons are tuned to relevant species-specific basic elements such as specific frequencies and will fire (i.e. be activated) when these frequencies are encountered in the input.

IBP neurons are higher order neurons that process information received from lower order neurons, for instance from IBE neurons. These neurons compute possible combinations of elements, such as CF-CF — observed, for instance, in the brains of mustached bats (Suga 1984), birds (Margoliash 1983), and primates (Olsen 1994). For example, IBP neurons may fire when a constant frequency component occurring at time 1 (CF_1) is followed by another constant frequency component at time 2 (CF_2). In this case, the IBP neurons will somehow compute the time difference between the occurrences of the two components.[11] The CF_1-CF_2 neurons are thought to be at the core of the bisonar system used by bats to echolocate their prey (Suga 2006). Humans use a comparable mechanism for sound localization by computing the minute interaural time and level differences based on the time of arrival of the sound at the two ears (see Eggermont 2001 for a review).

---

[10] Formant transitions correspond to what is generally described in animal studies as "sweeps" and consist of only one element rather than a combination of two (such as onset and offset of the transition). Neurons responsive specifically to the direction and steepness of these sweeps have been documented in various species (see Table 2–2 towards the end of this chapter for references). Hence, formant transitions are considered IBE rather than IBP.

[11] This description is a simplified view of the processes that are taking place, but my main concern here is to explain the basic (and most well documented) principles underlying the processing of complex sounds.

Neurons that process two or more elements at a time or consecutively are called *combination-sensitive neurons*. Many types of combination-sensitive neurons have been observed in the auditory system of different species (*amphibians* = Fuzessery & Feng 1982, 1983; *avians* = Margoliash & Fortune 1992; *mustached bats* = Edamatsu, Kawasaki & Suga 1989; Edamatsu & Suga 1993; Horikawa & Suga 1986; Mittman & Wenstrup 1995; Olsen & Suga 1991a, 1991b; Suga & Horikawa 1986; Tsuzuki & Suga 1988; Yan & Suga 1996; *horseshoe bats* = Schuller, O'Neill & Radthe-Schuller 1991; *little brown bats* = Maekawa, Wong & Paschal 1992; Tanaka, Wong & Taniguchi 1992; Wong, Maekawa & Tanaka 1992; *big brown bats* = Dear, Simmons & Fritz 1993; Dear & Suga 1995; and *primates* = Bartlett & Wang 2005; Olsen 1994; Srivatsun & Wang 2009). For instance, while some neurons in the auditory cortex of the squirrel monkey show weak responses to single-component calls (i.e. calls which include only one element, roughly comparable to either a rising tone, falling tone or level tone in human speech), the same neurons respond strongly when the call comprises all three components presented consecutively (Olsen 1994). Three main points emerge from these observations: 1) combination-sensitive neurons are activated by the presence of more than one component; 2) they have been documented in a large variety of species; and 3) they appear to have emerged as a result of the need for species-specific auditory behavior (Suga 2006). Therefore, it is reasonable to expect that various types of combination-sensitive neurons are likely embedded in the human auditory cortex as well.

Importantly, although neurons might be tuned to a specific acoustic component (such as a downward frequency modulation) they may react to variations of this component and in that sense neurons are said to be broadly tuned to the mapped

parameter (Seung & Sompolinsky 1993). However, the firing rate of the neurons will vary for any given stimulus, and therefore, groups of neurons tuned to the same parameter will provide, together, high-resolution information about the stimulus (Eggermont 2001).

2.3.2 Role of neural processing in speech perception

As discussed in the previous section, human infants and adults are able to compute statistical acoustic information, which in turn affects their perception of stimuli contained within that distribution. Changes in perception appear to be reflected in neural changes (e.g. Pienkowski & Eggermont 2009). I consider below some hypotheses about how neural changes occur.

Neurons that process similar information, such as closely related frequencies, are generally spatially concatenated. In this sense, they are said to form a *neural map*. Tonotopic (neural) maps refer specifically to populations of neurons in a limited neighborhood that process closely related frequencies. Among mustached bats, for instance, areas in the auditory cortex are divided into frequency-frequency coordinates, where one delineated area consists of neurons sensitive to CF_1/CF_2, while another area, spatially separated from the latter population of neurons, is tuned to CF_1/CF_3 (Suga 1984 revised in 2006).

The relative number of nerve cells (neurons) involved in the processing of a stimulus is called *cell density*. Kohonen's self-organizing feature maps (1982, 2001) in computational modeling of neural network present a *magnification factor* that reflects the fact that cell density activation is increased (a.k.a. magnified) at frequently stimulated

cortical regions of the input space, in line with documented cortical processing in visual and somatic modalities. For instance, kittens raised in an environment filled with vertical lines have a significantly larger area of the visual cortex devoted to the perception of vertical contours than kittens reared in a normal visual environment (Rauschecker & Singer 1981). If more cells are activated during the perception of a given type of stimulus, the perceiver is able to discriminate that stimulus in more detail and with better accuracy. The magnification factor is exemplified in Figure 2–3. As more stimuli in the input are distributed towards the center of the relevant dimension (top), more cells are assigned to the processing of those acoustic attributes (bottom).

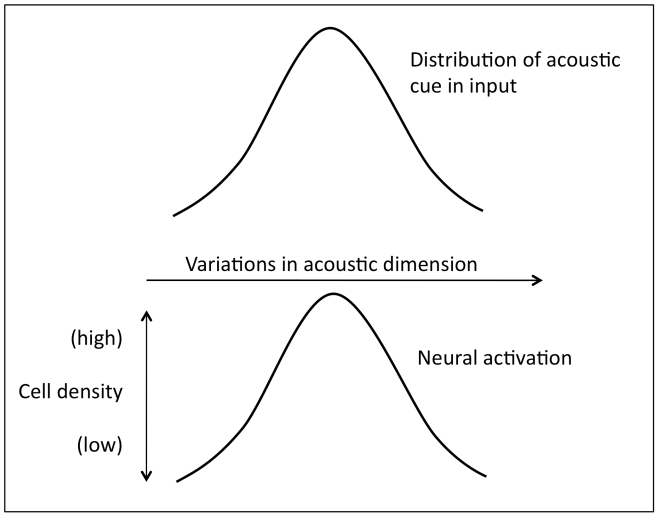

Figure 2–3 The magnification factor hypothesis: Cell density activation (bottom) increases proportionally as a result of the input distribution reflected here by a normal distribution along a given acoustic dimension (top).

In the case of speech sound processing, however, human infants appear to *lose* sensitivity for the perception of sounds within a given category after extensive exposure to the sounds of a particular language (Aslin, Pisoni, Hennessy & Perey 1981; Eimas, Siqueland, Jusczyk & Vigorito 1971; Kuhl et al. 2006; Werker & Tees 1984; see also Jusczyk 1987 or Werker 1995 for reviews). Hence, Bauer, Der and Herrmann (1996) posit an *inverted magnification factor* in which the learning of phonological categories in one's first language (L1) affects the density of cell activation in auditory cortical maps by *decreasing* the density of cell activation around categorical centers (a.k.a. prototypes). This hypothesis is illustrated in Figure 2–4. In other words stimuli corresponding to an acoustic value around the categorical center should activate fewer cells than stimuli at the edges of the same category.

Guenther and colleagues adopt the inverted magnification factor hypothesis in their neural network model (1999, 2001, 2002, 2004) and expand the model to make predictions about the effect of different types of training on neural activation. While discrimination training may lead to an increase in density of cell activation (magnification factor) and improve the perceiver's ability to discriminate minute differences between stimuli within the same category, categorization training may conversely trigger a decrease in density of cell activation (inverted magnification factor) and consequently lead to a decline in the perceiver's ability to discriminate stimuli within the same category. This hypothesis was tested and confirmed with a perceptual (behavioral) experiment, as discussed previously in Guenther et al. (1999) (see section 2.1.3). An ERP experiment by Tremblay, Shahin, Picton & Ross (2009) also found that auditory training may indeed trigger a change in neural activation in human subjects

exposed to a voice-onset-time contrast – in this case between 'mba' and 'ba'. Experiments with animals show that passive exposure to sounds within a given frequency range can result in reorganization of frequency tuning in both immature and mature animals; the animal becomes less sensitive to frequency variations corresponding to the frequencies presented during exposure (*newborn rats* = Chang & Merzenich 2003; *adult cats* = Noreña, Gourévitch, Aizawa, & Eggermont 2006; Pienkowski & Eggermont 2009).

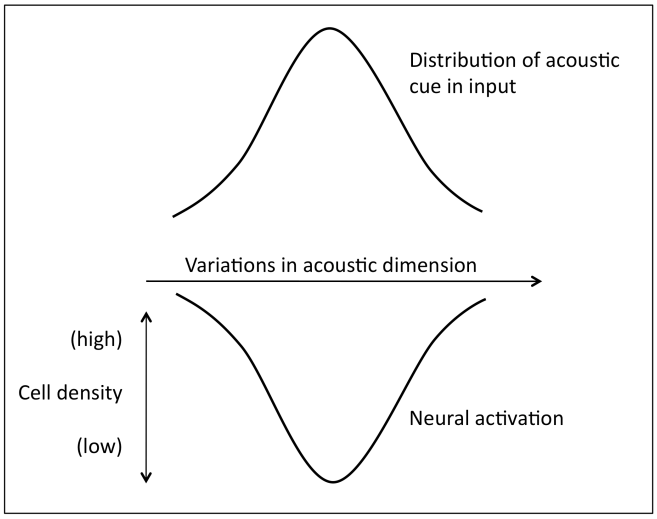

Figure 2–4 The inverted magnification factor hypothesis: Cell density activation (bottom) decreases at the categorical center of the acoustic dimension after intensive exposure to the input distribution (top). (Note: The inverted curve represents a decrease in cell density, not a negative value.)

At this point, it is worth asking what determines whether the cell density will be increased or decreased as a result of exposure to the same stimuli distribution. From a neuroethological point of view, whether an increase in the number of stimuli in a given category will trigger a magnification or an inverted magnification of cell density is likely

related to the behavioral needs associated with those stimuli. Let's compare neural processing with a hypothetical recycling facility. This facility may be divided into two sections, one dealing with plastic materials and the other dealing with glass (= the stimuli). Since there is great variety in the types of plastic, each plastic container has a different number (from 1 to 7) written on it (at least in North America). Now, because those plastic types are chemically different, they must also undergo different recycling processes (= behavioral need). As the hundreds of thousands of plastic containers arrive at the recycling facility, one can imagine that many employees (= neurons) will be required to sort the containers by number. Moreover, as more plastic containers arrive, more employees will be needed (= increase in cell density). On the other side of the recycling plant, in the glass department, there is no need to categorize the different types of glass, since no matter the size, color or shape of the glass material, they can all be recycled together using the same process. Accordingly, a few employees may suffice to process the glass containers. Since there is no need to separate the glass containers, as more glass containers are sent to the recycling facility, the manager may decide to make the process more cost-efficient by firing everybody (granted the union's approval, of course), with the exception of one or two spared employees (= decrease in cell density) who will be in charge of operating the newly acquired automated system that processes the glass from its time of arrival at the facility to its entry into the recycling machine. Thus, even if the number of plastic and glass containers arriving at the recycling facility is the same, more employees will be needed in the plastic department than in the glass department, because the recycling needs associated with each product are different. Hence, the plastic department is an example of discrimination behavior that requires an

increase in employees (= increase in cell density), whereas the glass section is an example of categorical behavior that requires fast and efficient treatment of the product by as few employees as possible (= decrease in cell density).

The specific phenomenon for which the inverted magnification factor was designed to account is the putative *perceptual magnet effect*, as first described by Kuhl and colleagues (Iverson & Kuhl 1995; Kuhl 1991; Kuhl & Iverson 1995; Kuhl et al. 1992). In some instances, it has been observed that stimuli are not equally discriminated across the perceptual space, even within the same category. For instance, pairs of tokens featuring slightly different versions of the same vowel, e.g. various acoustic instances of /i/, were perceived as similar by native American English listeners when they were acoustically close to the vowel corresponding to the putative best exemplar (a.k.a. prototype or categorical center) of that category, but were perceived as different when they were farther away from that prototype (Kuhl 1991). In other words, listeners were shown to be insensitive to differences between tokens surrounding the prototype of a given vowel, though they could perceive differences of similar magnitude if the tokens were farther away from that prototype. This phenomenon is known as the perceptual magnet effect, and has been observed with the perception of some native vowel categories (Kuhl 1991; Kuhl & Iverson 1995; Sussman & Lauckner-Morano 1995), the liquids /r/ and /l/ (Iverson & Kuhl 1995), and a Mandarin alveolo-palatal affricate-fricative distinction (Tsao 2001). A study comparing American and Swedish infants showed a perceptual magnet effect for vowels present in the infants' ambient language after (but not before) six months of age, presumably as a result of linguistic experience (Kuhl et al. 1992). Although the perceptual magnet effect was first argued to be a human

speech-specific phenomenon based on the fact that no similar distortion of the perceptual space was found in a study with monkeys (Kuhl 1991), this view has since been challenged (e.g. Lacerda 1995). A study by Guenther et al. (1999), in particular, has shown that it is possible to shrink the perceptual space of non-speech stimuli, while Barrett (1999) has provided similar results with music categories.

To account for the observed unevenness in the discriminability of tokens within the same category, Kuhl and colleagues proposed the Native Language Magnet (NLM) theory (Iverson & Kuhl 1996; Kuhl 1991; Kuhl & Iverson 1995; Kuhl et al. 1992). This model was meant to account for discoveries that adult listeners were generally able to identify the best instance of a given phonetic category, referred to as *prototype*, and that this prototype "perceptually pulls other members of the category toward itself [so that] the perceptual distance between outlying sounds and the prototype is reduced" (Kuhl & Iverson 1995: 123), hence the magnet effect metaphor. However, the NLM account, and the existence of a magnet effect itself, have been fervently debated (e.g. Lotto 2000; Lotto, Kluender & Holt 1998; Sussman and Gekas 1997; Thyer, Hickson & Dodd 2000), as the results from Kuhl and colleagues' experiments have proved difficult to replicate. Experiments by Sussman and Gekas (1997) failed to find a perceptual magnet effect for the lax English vowel [ɪ] comparable to the one found for the vowel [i] in previous experiments by Sussman & Lauckner-Morano (1995). BharrathSingh (2001) also found that although prototypes of the English vowels [i], [u] and [a] seem to perceptually attract stimuli in the direct neighborhood, the distribution of the "magnet effect" was not as symmetrical as suggested by the NLM account, corroborating findings by Sussman and Gekas (1997), who found a variety of individual differences in the pattern of the magnet

effect. Lotto, Kluender & Holt (1998) and Lotto (2000) argue that the perceptual magnet effect is the result of a methodological confound. These authors suggest that differences in the methodology used to assess identification judgments (evaluating the best prototype) and to assess discrimination judgments (evaluate the perceptual distance between exemplars) produce the phenomenon of the supposed perceptual magnet effect. They claim that this effect is simply a further demonstration of the phenomenon of categorical perception, which yields better discrimination for between-category exemplars than for within-category exemplars.

Although the perceptual magnet effect may not stand as a phenomenon specific to human language, and may not be as systematic and constant across individual results as originally thought, some evidence suggests that the effect is more than a mere methodological confound. From a neurological viewpoint, as proposed by the inverted magnification factor hypothesis, it is possible that the cell density devoted to the perception of sounds near a categorical boundary is *higher* than the cell density devoted to the perception of sounds at categorical centers, creating, in some instances, the illusion of a perceptual magnet effect. A recent fMRI study conducted by Guenther and colleagues (2002, 2004) demonstrate that listening to tokens of the same category might indeed trigger different sizes of cortical activation, depending on the tokens' proximity to the categorical center. In their experiment, native English speakers' perception of a putative prototypical English /i/ exhibited less cortical activation than perception of a non-prototypical /i/. Figure 2–5 illustrates the difference in the size of the area activated, depending on the proximity of the stimuli to the perceptual categorical center. The relative size of the activated cortical area presumably reflects the number of nerve cells

involved in processing, as illustrated at the very bottom of the figure. However, one important question remains: Why would more cells be activated at categorical boundaries than at categorical centers? Suga (2006) has suggested that more cells may be activated to resolve the possible competition between two close categories, implying that the cells activated would presumably belong to centers devoted to two separate categories. This hypothesis has received support from an fMRI study conducted by Myers (2007), in which greater activation (in bilateral inferior frontal areas) was revealed for [da] and [ta] tokens with VOT values near the categorical boundary, than for [da] and [ta] with VOT values corresponding to categorical centers.

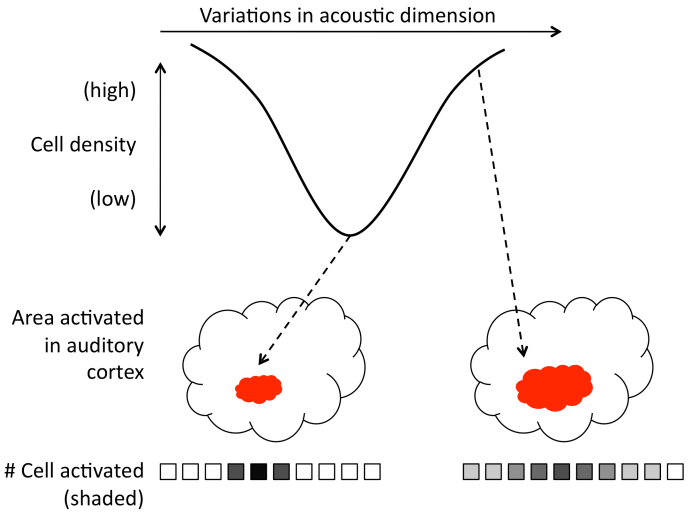

Figure 2–5 An acoustic component corresponding to a categorical center generates less neural activity (left) than an acoustic component near a categorical boundary (right). Consequently, the stimulus at the categorical center yields a smaller area of cortical activation (i.e. activates fewer neurons) in the auditory cortex than the stimulus near the categorical boundary.

To return to the recycling facility analogy, let's suppose the facility also recycles metal (i.e. cans) in addition to glass and plastic. Cans are like glass, in the sense that no special treatment is necessary for different types of metal container (= categorical processing). Containers that are made of glass will be processed by employees in the glass department, while containers made of metal will be processed by employees in the metal department. These two categories are clearly defined, and as long as the container that arrives is clearly made of glass or metal (= corresponds to the categorical center), the relevant department will process it quickly and effectively — that is, without requiring many employees (= low level of cell density activation). However, if a container arrives that is partly made of glass and partly made of metal (= at a category boundary), it cannot automatically be sent to one or the other department. To resolve this ambiguity, we can imagine that employees from both departments might be called to decide which department would be best suited to handle the confusing container. One can imagine that in resolving such a dilemma, more employees (= high level of cell density activation) would be involved and that the processing time would also be slower. Hence, it can also be expected that more neurons will be required for the processing of stimuli that do not correspond to the best category prototype, whether the stimulus is close to a categorical boundary or whether it simply deviates from the categorical center. This view appears reasonable in reconciling the different findings discussed previously, that stimuli across categorical space are not processed equally, either in terms of density of cell activation or processing time.

2.3.3 Resolving the invariance problem

In the previous section, we saw that neural organization is affected by the statistical distribution of stimuli in the input *and* by the behavioral need (discrimination or categorization) associated with these stimuli: neural maps used for discriminating fine details between stimuli appear to undergo some kind of magnification factor that *increases* the density of cell activation at frequently activated regions of the input space, whereas neural maps used for categorization of stimuli, such as in speech perception, appear to undergo an *inverted* magnification factor that *decreases* the density of cell activation at the most frequently activated regions of the input space (i.e. at categorical centers). While infants and adults have been shown to be sensitive to the statistical distribution of stimuli for categorical perception of relevant acoustic speech contrasts (as discussed in sections 2.1.1 and 2.1.2), it is questionable if the statistical distribution in the input provides sufficient invariance for the formation of distinct neural maps. The fact is that the acoustic envelope of a given sound may vary considerably as a function of the sounds that surround it (contextual effects), its prosodic context, and the speaker's age, gender, or dialect (indexical variations). Despite this variability, listeners rarely encounter major difficulties in understanding their native language in normal noise conditions. The lack of correspondence between speech units perceived as separate entities (e.g. phonemes) and their acoustic realization is known as the *invariance problem* (a.k.a. lack of invariance).

Various proposals have attempted to explain how humans are capable of perceiving sounds categorically without the presence of clearly identifiable and reliable categorical boundaries in the input. One notable account is the one put forward by

exemplar-based models. Although exemplar-based models are not neural-based (they are cognitive rather than neural-based models), they rely on the notion of episodic traces, which are memories of specific events (in this case, acoustic events) presumably encoded by the neurology. Thus, the proposal of these models to account for the invariance problem seems worth considering.

According to these models, the acoustic details of the speech input are directly stored as traces (episodic traces)[12] in the listener's cognitive system (i.e. in episodic memory). These traces are organized according to their similarity with previously encountered traces; then, new input is compared with the organized clusters or clouds of traces for interpretation of new or previously encountered voices, phonemes, lexical items, and so forth (e.g. Ashby & Maddox 1993; Goldinger 1996, 1998; Johnson 1997; Lacerda 1995; Medin & Schaffer 1978; Nosofsky 1984, 1986; Nosofski & Zaki 2002; Pierrehumbert 2001). In these models, putative phonological categories (the source of the invariance) consist of the average values of the traces within a restricted neighborhood of exemplars. A weighting scheme ensures that exemplars collected from a given voice will be attributed more importance during processing of exemplars uttered by that same voice to facilitate word recognition and processing. Similar schemes allow the identification of indexical information such as gender, age or dialectal differences. In these models, the "variance" is therefore embedded in the averaging schemes, which can be context-

[12] Note that episodic traces and neural maps are assumed to be different concepts (i.e. different phenomena), and therefore, the neural-based model proposed in this work departs significantly from exemplar-based proposals. The proposal by Hebb (1949) and others (e.g. Allport 1985) that memories simply correspond to patterns of neural connections has been seriously challenged by recent work showing that encoding of memories involves a wider range of spatially and temporally dynamic processes than discrete neural connections (e.g. Rose 2008; Silva et al. 2009), as discussed later in this section.

specific and voice-specific. Exemplar models assume that memories decay, possibly accounting for changes in perception/production over time, and for the fact that the full range of encountered exemplars does not necessarily end up stored in memory. Thus, the theory takes attention and capacity for memorization into consideration (cf. Pierrehumbert 2006 for a discussion of the virtues and shortcomings of exemplar models). In sum, in exemplar-based models, the invariance in speech perception is captured by clusters of memories organized based on their similarity to one another, where similar episodic traces are encoded in the same neighborhood.

Unfortunately, despite the elegance of this sensible proposal, recent findings in molecular and cellular biology do not concur with the averaging schemes of exemplar-based models, which (the averaging schemes) presumably assume that similar exemplars are encoded in the same neural circuits. Instead, allocation of new memories in neural circuits has been found to be highly dependent on the time at which the memory is encoded (rather than on their acoustic or semantic similarity), where stimuli encountered within minutes or hours of each other are more likely to be encoded in the same population of neurons.[13] Silva, Zhou, Rogerson, Shobe and Balaji (2009) describe two models of memory allocation based on recent molecular and cellular approaches: one model based on activation of neuronal populations and another based on activation within dendritic trees (which are part of the cell synaptic connection, and therefore, the activated

[13] Understandably, exemplars uttered by a given speaker during a conversation may be encoded in the same group of neurons, and may contribute to form some sort of clusters based on the similarity between these exemplars. However, that also means that if two speakers with different voice register speak within minutes of each other, the exemplars produced by the two speakers will be clustered together, irrespective of whether the exemplars share any similarity to one another.

dendritic trees end up activating different neurons). For example, the first model posits that recent activation of a cell (which triggers an increase in the activity and level of a responsive element-binding protein called CREB) makes this cell most likely to be involved in the encoding of the next memory. Accordingly, this implies that activation of memory A stored within hours (neuronal population model) or within minutes (dendritic tree model) of memory B will also activate memory B (and vice-versa), irrespective of whether or not the two memories share any acoustic or semantic similarities. That is, memories acquired within hours or minutes of each other will result in strong co-recall, whereas memories acquired at different time scales will result in weak co-recall. Hence, biological research on memory allocation to date seriously challenges the averaging schemes proposed by exemplar models, which are heavily based on similar memories being encoded in the same neighborhood (by forming "clusters" of similar memories).

Crucially, exemplar-based models are not grounded in neural processing, but are said, instead, to be cognitive models (i.e. based on the behavioral processing of information, in this case, based on the role of memory encoding). In addition, previous research suggests that the processing of acoustic details by the neurology is a distinct phenomenon from the encoding of specific acoustic episodes (Goldinger, Kleider & Shelley 1999, see also footnote 8). In fact, it appears that episodic memory is located in, or at least involves, different areas of the brain than the neural processing of acoustic cues: the encoding of episodic memory is believed to be centered overwhelmingly in the prefrontal cortex of the left hemisphere, while the right hemisphere appears to play an important function in the retrieval of episodic traces (see Tulving 2002 for a review of PET and fMRI studies supporting these claims). Hence, taken together these findings

strongly suggest that the invariance in speech (i.e. speech categories) is *not* (solely) captured by the encoding of the putative episodic traces in memory. Provided that invariant paradigms are already present in the speech input, it is possible that neural processes in the auditory cortex (or closely related areas) may, on their own, shape the perception of constant features of speech (e.g. voicing distinction, place of articulation, VOT). If so, speech sound categories (a.k.a. phonetic or phonemic categories) may somehow be encoded in neural organization (e.g. in neural maps) rather than in the organization of episodic traces in memory.

From a neuroethological perspective, it may also be advantageous to have some kind of normalization mechanisms at the neural level – that is prior to or instead of relying on the averaging of memorized episodes – to accelerate the processing of speech across speakers, while still enabling the coding of speaker-specific information for individual identification. Below, I review proposals and some experimental evidence suggesting that the invariance problem as well as some acoustic normalization may indeed be resolved by the neurology.

First, it must be understood that the "lack of invariance" problem stems from the speculation that the acoustic realization of speech sounds is presumably too variable across contexts, speakers and speech rates to provide any reliable cue for their categorization in discrete categories. To cite only three of many possible examples, a shift in formant patterns has been observed for English round vowels in alveolar contexts (Hillenbrand, Clark & Nearey 2001; Stevens & House 1963); the VOT value of stop consonants is affected not only by speech rate, but also by place of articulation, stress, and position in the sentence (Lisker & Abramson 1967); and the resonant frequencies of

English vowels vary considerably across speakers as a function of vocal tract size (Hillenbrand, Getty, Clark, & Wheeler 1995).

A perception study by Cooper, Delattre, Liberman, Borst, & Gerstman (1952) using synthesized speech stimuli in which F2 transition steepness, F2 direction, and vowels were systematically varied has shown that native English listeners may perceive the same falling F2 transition as [d] or [g], depending on the following vowel. However, the same study also found that listeners could identify the place of articulation of a stop consonant (b-d-g or p-t-k), based on the change between the onset of the F2 transition and the F2 of the vowel. The center frequency of the consonant burst release provided additional information to extract place of articulation. In addition, a study by Ladefoged and Broadbent (1957) showed that vowels can be identified by taking into consideration the relationship between the formants of a vowel and that of other vowels pronounced by the same speaker, pointing to the fact that our perception, especially of vowels, is relative rather than absolute.

Sussman, McCaffrey & Matthews (1991) further demonstrated that in natural (as opposed to synthetic) production, there is a surprisingly robust linear relationship between the F2 at onset and the F2 value taken at the mid-vowel nucleus across context, gender, and dialect; and, moreover, that this linear relationship differs according to place of articulation in terms of slope and intercept, as illustrated in Figure 2–6. This study was based on the recording and spectral analyses of /bVt/, /dVt/ and /gVt/ tokens produced with ten vowels by ten male and ten female speakers of different English dialects (Texas, California, New York, and Midwest). The linear relationship between two coordinates (resulting in a straight regression line) is referred to as *locus equation*. The concept of

locus equation was first conceived by Lindblom (1963). Locus equations for English stop transitions similar to those reported by Sussman et al. (1991) were reported by Nearey & Shammass (1987) using ten native Canadian English speakers, further demonstrating the robustness of locus equations across dialects of the same language. Interestingly, these locus equations are robust to speaker and dialectal variability, but at the same time, can still reflect cross-linguistic differences. The slopes and intercepts of the locus equations for place of articulation observed with Swedish (Krull 1989; Lindblom 1963), Thai, Cairene Arabic, and Urdu speakers (Sussman, Hoemeke & Ahmed 1993) differed from those obtained with English speakers. The relevance of the F2 onset and F2 vowel locus equations in perception was assessed and confirmed in a perceptual experiment by Fruchter & Sussman (1997).

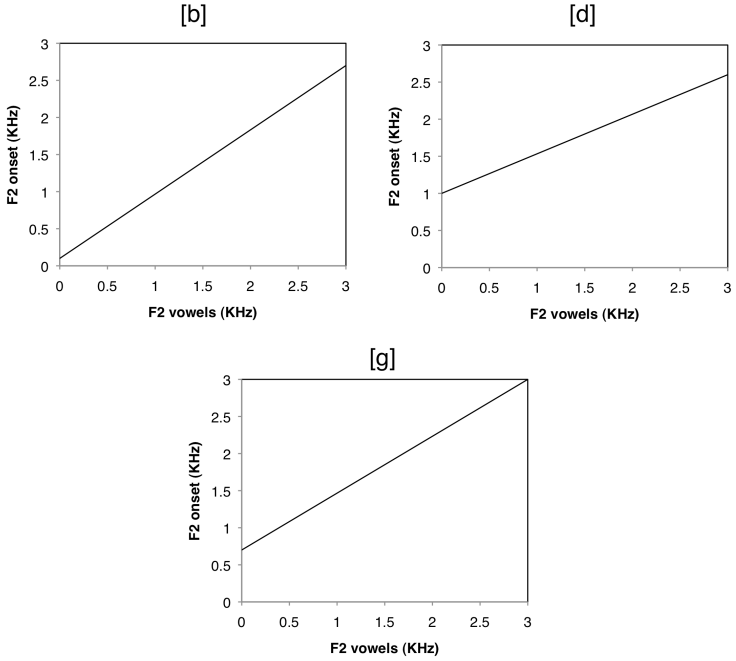

Figure 2–6 Locus equations for /b/, /d/, and /g/ combining male and female speakers (adapted from Sussman et al. 1991: 1314).

It is important to note that while some overlap still exists in the locus equations, when these equations are combined with additional information, mainly that provided by the centre frequency of the preceding burst, the possible resulting ambiguity disappears (Cooper et al. 1952; Sussman et al. 1991). Thus, simple combinations of acoustic information, such as those presented above, are potentially sufficient for identifying constant parameters of many sound contrasts in English and other languages, such as [b-d-g], [p-t-k], [m-l] (Cooper et al. 1952), and possibly many other contrasts.

Provided that these locus equations are sufficiently transparent in the input to permit the formation of discrete categories, the next question is: Can the neurology deal with these parameters; and more importantly, how? Sussman and colleagues have proposed a neural account (Sussman 1989, 1999, 2002; Sussman et al. 1991). The authors speculate that the locus equations are captured by, or embedded in, similar columnar organizations of neurons or sets of neurons as those found in bats (Suga et al. 1983) and barn owls (Wagner et al. 1987). This proposal is illustrated in Figure 2–7. Combination-sensitive neurons organized in discrete columns are hypothesized to account for the processing of F2 values at transition onsets and midvowel nuclei. These columns correspond to the IBP filters posited by Suga, and to the percepts associated with distinct places of stop articulation (i.e. each column is associated with a different place).

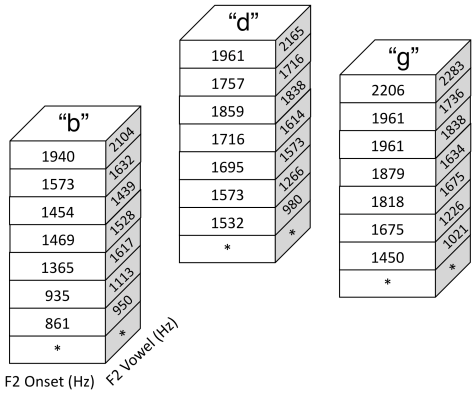

Figure 2–7 Hypothetical columnar organization of neurons encoding F2 values at onset and in the vowel (adapted from Sussman 2002: 9).

A brain-based model developed by Sussman and David Fruchter (as described in Sussman et al. 1991) explains how the different values can yield different outcomes by a stepwise processing of information pertaining to noise burst, F2 onset, and vowel formants. A simplified version of their model is presented in Figure 2–8.[14] In simple terms, the neurons that process the noise burst send this information to combination-sensitive neurons that process this information in conjunction with the $F2_{onset}$ value. Subsequently, sets of combination-sensitive neurons compile or evaluate the information received from the noise burst+$F2_{onset}$ neural map, the $F2_{onset}$ map, and $F2_{vowel}$ map and send the "winning" outcome for place of articulation to higher centers. At each stage, the set of neurons may project different possible outcomes. For instance, the $F2_{onset}$ value in

[14] While I have tried to provide a faithful representation of Sussman and Fruchter's model, I took the liberty of replacing some of their terms with those I consider to be equivalent, for consistency with the description and terminology introduced throughout this dissertation. In addition, I have reproduced only part of their figure here for the sake of simplicity. That said, I take full responsibility if these measures have led to a misrepresentation of their model.

the figure suggests the possibility of either a labial or alveolar place of articulation. The vowel formant information also yields the possibility of either a labial or alveolar plosive. However, the burst+$F2_{onset}$ neurons project only the possibility of an alveolar stop. Hence, in this case, 'd' is the "winner" because this option receives the most support (i.e. neural weight) in the transaction. taught

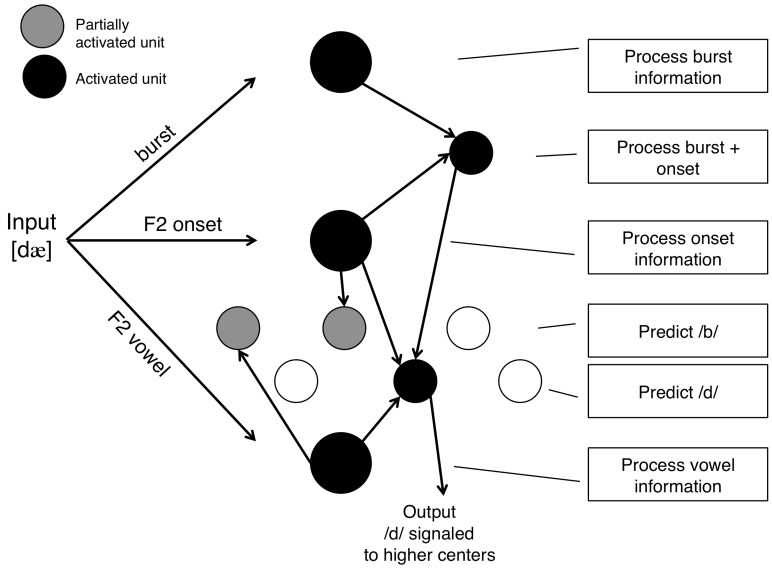

Figure 2–8 Schematic illustration of the brain-based model developed by Sussman and Fruchter (simplified and adapted version of the model presented in Sussman et al. 1991: 1324; please refer to footnote 14)

To sum up, despite its highly variable nature, the speech input appears to contain extractable invariant parameters (e.g. locus equations).[15] These parameters can be

[15] It remains to confirm that invariant parameters exist for all acoustic speech contrasts. This issue is discussed further in the following chapter with concrete examples and studies that have identified other acoustic correlates that may serve as invariant parameters.

captured in the neurology by sets of neurons tuned to series of combined component values (combination-sensitive neurons), which are likely organized in separate neural maps. These maps may be spatially organized into columns or other possible neural structures (there is no evidence for or against columnar organization in the human brain). The exact physical shape of the neural maps, however, is of no particular concern to the current work.

One final important issue that needs to be addressed is whether the neurology can perform normalization on the entering input to facilitate categorical processing. Unlike acoustic correlates of place of articulation, the formant patterns of vowels – particularly F1, F2, F3, which are critical for vowel identification – are highly variable and often overlap (e.g. Hillenbrand, Clark & Nearey 2001; Steinlen 2002). Vowels are particularly sensitive to contextual variation, and formant values are greatly affected by the pitch range of the speaker. Hillenbrand, Clark & Nearey's (2001) study of contextual effects on the production of English vowels by male and female speakers demonstrates that listeners' ability to discriminate vowels can be accounted for by modeling F0, duration, and the spectral coding of the formant pattern using two formant samples (arguably, in languages like French, coding of F3 may also be necessary). In a later perceptual study, Hillenbrand, Houde & Gayvert (2006) further demonstrated that information conveyed by spectral envelope peaks in the speech signal (including, but not limited to, perception of vowels) contribute sufficiently to speech intelligibility (that is, the presence of only the spectral peaks is sufficient for understanding speech), although the inclusion of spectral details (i.e. all harmonics) shows a measurable – though by no means large – advantage. Taken together, these results point to two conclusions: 1) listeners can extract speech

categories based on impoverished signals, provided that information related to spectral peaks (including, but not limited to, formants) is provided; and 2) listeners make use of all the information in the speech signal if it is provided. The second point might play an important role in normalization of the signal across speakers, an issue discussed in more detail below.

Sussman (1986) suggested that normalization of vowels can be achieved by the neurology and contribute to the extraction of abstract vowel categories. According to his model, the spectral peaks are first combined into three (or more) neural maps of combination-sensitive neurons processing F1/F2, F1/F3 and F2/F3 in conjunction. Higher-order sets of neurons perform normalization on the raw data. Various algorithms have been proposed to normalize vowels (see Johnson 2005 for a review), but the one used by Sussman for illustrative purpose is F_n / AVERAGE(F1+F2+F3) where $_n$ corresponds to each individual formant (which algorithm is most likely used at the neurological level is still unknown). Thus, each formant is normalized by taking into consideration the average value of the first three formants. The normalized formants are then grouped into some sort of neural map and associated with abstract representations corresponding to the percept of vowel categories. Sussman's model was tentative and simply meant to show that the neurology is perfectly capable of handling the normalization of formants thought necessary for categorical perception of vowels. A study conducted by Diesch and Luce (2000) evaluated the auditory-evoked neuromagnetic field elicited by single compared to two-formant vowel conditions. The response to vowels containing two formants was superadditive, suggesting that some blurring mechanism is applied on the spectral envelope. Diesch and Luce (2000)

conclude that this mechanism, the exact nature of which is not yet known, may extract some invariance from the input, in line with the normalization hypothesis proposed by Sussman (1986).

Hence, it appears that a neural mechanism performing some sort of normalization on the speech input is possible. This hypothesis concurs with perceptual data (the fact that despite the highly variable nature of vowels, listeners are able to instinctively perceive distinct categories) as well as with neural data obtained to date (Diesch & Luce 2000). It is also important to stress that it may not be necessary for the speech input to be invariant across all contexts. It has long been established that "any two cells or systems of cells that are repeatedly active at the same time will tend to become 'associated', so that activity in one facilitates activity in the other (Hebb 1949: 70)." This theory implies that exposure reinforces or magnifies the learning paradigm by creating associations between related neural maps, and that contrasts must only be distinctive in each context. Perceived similar contrasts across contexts can be mapped together (e.g. into a columnar organization) and associated with the same invariant parameter (a.k.a. feature, or phonetic or phonemic category).

2.3.4 Types of neurons relevant for perception of speech sounds

In the previous sections I demonstrated that experimental evidence to date suggests that there is sufficient invariance in the input for the neurology to create neural maps corresponding to human speech categories. However, it remains to uncover the type of acoustic components these maps specifically respond to, and how these components are specifically processed by the neurology when, for instance, multiple cues are available

for the same speech contrast. The model of speech perception presented in the following chapter is designed to address these questions by relying on the assumption that the types of neurons documented in animal studies are also active in the human brain. This assumption is based on the fact that humans and various non-human animals are sensitive to comparable acoustic components relevant for their respective communication system, and that neurons responding specifically to these components have been documented in invasive neural experiments with many of these non-human species. Hence, it is reasonable to assume that the types of neurons documented in animal experiments must also play an important role in human speech perception (e.g. Eggermont 2001; Suga 2006.) In this section, I explain the potential correlation between the types of neurons documented in various species and the perception of speech contrasts used in human communication.

Languages can be described by a limited set of sounds or sound contrasts, referred to as the phonemic inventory. Phonemes are percepts meaningful for lexical contrasts that are built up from a limited set of acoustic components. These components can be divided into two general classes: acoustic components related to spectral information (e.g. frequency and frequency-modulated components); and temporal and synchrony representation of sound contours (e.g. onset, offset, duration, voice-onset-time) (Eggermont 2001). Animal studies revealed that spectral information is processed by neurons tuned to specific frequencies, frequency modulations (slope), or combinations of frequencies. Temporal and synchronous components might be partly captured by the firing rate or discharges of the neurons as a response to spectral components, or by combination-sensitive neurons (e.g. Gehr, Komiya & Eggermont 2000).

The phonological contrasts and phonotactic constraints that distinguish each language may be derived from particular configurations of those few spectral and temporal components. Table 2–2 lists some of the components (acoustic cues) believed to contribute to the perception of linguistic categories, along with examples of speech contrast(s) each cue might serve to distinguish. This table also lists the type of neurons presumed (see Eggermont 2001 and Suga 2006 for a more detailed review) to selectively respond to that cue, and species in which this particular type of neuron has been observed. Understandably, many of those neurons are responsive to more than one component (combination-sensitive) or may contribute to different percepts at the same time by capturing, for instance, both frequency and timing information.

Cooper et al.'s (1952) perceptual (behavioral) study using synthetic CV stimuli showed that the center frequency of the release burst following the production of a stop consonant contributes to the identification of place of articulation for English listeners. High-frequency bursts are identified as corresponding to /t/, irrespective of the following vocalic context. Bursts with frequency values just above the second formant of the following vocalic segment are perceived as /k/, while other burst frequencies are perceived as /p/. Rauschecker, Tian & Hauser (1995) documented neural maps in the non-primary auditory cortex of macaques that respond to the center frequency versus the bandwidth of burst noises. Burst-like signals are present in the natural vocalizations of various other animals, such as the white-crown sparrow (Margoliash 1983). Thus, similar neural maps may be present in humans, possibly accounting for the categorical perception of place of articulation based on the center frequency (and possibly bandwidth) of the consonant noise burst.

Table 2–2 Hypothesized correspondence between acoustic cue, linguistic percept and type of neural response

Acoustic cues	E.g. Speech sound/contrast	Neuron type (or selectivity to)	Species attested (reference)
		Spectral components	
Noise bursts	Stop place contrast	center frequency and bandwidth of noise burst (NB)	monkey (R et al. 1995)
Spectral peaks (e.g. formants)	Vowels, liquids, fricatives, glides	Constant frequency components (CF)	cat (W & W 1942; Q et al. 2004), mustached bat (S et al. 1983) human (B et al. 2008)
Modulated spectral peaks (e.g. formant transitions)	Stop place and voicing contrast	Frequency-modulated components (FM)	mustached bat (S & O 1979), myna bird (S 1991), pallid bat (R & F 2006)
		Temporal or synchronous components	
Timing cues	Durational contrasts, Voice-Onset-Time (VOT)	Phase or Time-locked (discharges) and/or combination-sensitive neurons	mustached bat (S & O 1979), monkey (S et al. 1995)
Periodicity	Voicing contrasts	Amplitude-modulated components (AM)	cat (S & U 1986), mynah bird (H et al. 1987), chinchilla (L et al. 2002), gerbil (K et al. 2008), guinea pig (M 2008),
Pitch correlates (e.g. F0)	Intonation, lexical tone, stress, and accent	Amplitude-modulated components (AM)	marmoset monkey (B & W 2005)

Many sounds in the human communication system, such as vowels, fricatives, liquids, or glides, exhibit peaks in the sound envelope, including, but not limited to, formants (Ladefoged 2001). In acoustic terms, formants consist of a concentration of energy around a given frequency, and when this frequency remains relatively stable across a given time period, it is referred to as a constant (or characteristic) frequency component (CF). Formants (a.k.a. resonances) have long been used to describe vowels.

The relationship between the first and second formants of a vowel, as potentially processed by combination-sensitive neurons, usually suffice to discriminate most vowels, although the third formant may be required to distinguish some vowels and some liquids such as English /r/ and /l/. For instance, the cardinal vowels /i/ and /u/ contrast primarily in terms of the frequency of their second formants (F2 is higher for /i/). Both /i/ and /u/ contrast with /a/ in terms of the frequency of their first formant, which is lower for /i/ and /u/ than for /a/ (e.g. Ladefoged 2001).

In addition to playing a crucial role in distinguishing vowels, spectral peaks are also potentially important in distinguishing fricatives (e.g. Behrens & Blumstein 1988; Evers, Reetz & Lahiri 1998; Heinz & Stevens 1961; Hughes & Halle 1956; Jongman, Wayland & Wong 2000; Shadle 1990; Strevens 1960). Jongman et al. (2000) reported a noticeable decrease in the frequency of spectral peak as the place of articulation of the fricative moves further back in the oral cavity (labial /f, v/ = 7733 Hz; dental /θ, ð/ = 7470 Hz; alveolar /s, z/ = 6839 Hz; palato-alveolar /ʃ, ʒ/ = 3820 Hz). These values were averaged from 20 native American English speakers and in six vowel contexts. Importantly, unlike the burst noise of stop consonants, the spectral peaks of fricatives are generally not affected by the following vowel context, with the exception of alveolar fricatives, for which the spectral peak is significantly lower when followed by the back-rounded vowels /o/ and /u/. In any case, spectral peaks are crucial components in speech perception by humans and many other species. For instance, a perceptual experiment by Fitch & Fritz (2006) demonstrated that rhesus macaques could perceive changes in formant frequencies in conspecific (same species) vocalizations. Neurons sensitive to specific frequencies have been reported in studies with anesthetized (Woolsey & Walzl

1942) and awake cats (Qin, Chimoto, Sakai & Sato 2004), as well as in mustached bats (e.g. Suga et al. 1983). In addition, a very recent study confirmed the presence of neurons sensitive to fine frequency tuning in humans (Bitterman et al. 2008). The exact distribution and mapping of frequency in the human cortex is still unknown, but it is reasonable to speculate that it should somehow reflect the perceptual abilities observed in behavioral studies, a pattern that has been documented in various recent neuroethological data discussed throughout this chapter (see Eggermont 2001 and Suga 2006 for reviews).

Spectral peak components may exhibit a directional change and give rise to frequency-modulated components (FMs) such as formant transitions, which provide further information on place of articulation and voicing. F2 transitions have been shown to provide reliable cues for identification of place of articulation (locus equations) that are robust to speaker and context variability (e.g. Sussman et al. 1991, as discussed above). F1 transitions are also known to correlate with the voicing of stop contrasts in initial CV position in English and to provide a crucial cue for the voicing distinction in quiet (Lisker 1975; Summerfield & Haggard 1977) and noisy (Jiang, Chen & Alwan 2006) conditions. The change in F1 frequency of plosives in post-vocalic word-final position is also relevant to the voicing distinction, as tested with native Australian English speakers (Jones 2005). Changes in spectral peaks (a.k.a. frequency-modulated sweeps) play an important role in the communication system of many species. Neural mechanisms responding to the rate and direction of these sweeps have been documented in the mustached bat (Suga & O'Neill 1979), mynah bird (Scheich 1991), and pallid bat (Razak & Fuzessery 2006). Given the attested perceptual ability of humans to use frequency-modulated components for speech contrasts, it is plausible that neural maps

corresponding to these contrasts exist in the human brain, and this is the assumption that I make in this dissertation. For instance, Sussman et al. (1991) hypothesized that F2 transitions may be mapped by combination-sensitive neurons that may be organized into columns, corresponding to locus equations found to distinguish each place of articulation (discussed in section 2.3.3).

Temporal and synchronous components contribute, in concert with spectral components, to create additional perceptual contrasts. In Japanese, for instance, the duration of vocalic segments can be contrastively short or long, and the duration of stop closures can be associated with either singleton or geminate consonants (Akamatsu 1997; Vance 1987). In English and many other languages, the time difference between the onset of the stop burst release and the onset of the F1 transition (or onset of periodicity), known as VOT (voice-onset-time), can yield the perception of distinctive voicing contrasts (e.g. Ladefoged 2001; Lisker & Abramson 1967). Humans and other species, such as monkeys, are known to be particularly sensitive to VOT differences and to perceive the VOT continuum in terms of discrete categories (Sinnott & Adams 1987). Steinschneider, Schroeder, Arezzo and Vaughan Jr. (1995) found possible neuronal correlates of VOT in the auditory cortex of awake rhesus monkeys. They found that short VOT values, ranging from 0ms to 20ms (corresponding to perceived /da/ stimuli), elicited only one response peak (discharge) at stimulus onset from a particular set of neurons, while long VOT values of 40 and 60 ms (corresponding to perceived /ta/ stimuli) clearly elicited two response peaks, the first at stimulus onset, and the other at onset of periodicity. A neural discharge that occurs in synchrony with a particular event (e.g. onset of periodicity) is called a phase- or time-locked discharge (Johnson 1980; Wallace, Shackleton & Palmer

2002). Steinschneider et al. (1995) suggested that the differential phase-locked response patterns of the neurons may account for the categorical perception of VOT. In a MEG study, Simos et al. (1998) demonstrated event-related magnetic field differentiations that correspond to temporal patterns of neural organization for VOT in the human auditory cortex that are comparable to those found in monkeys (i.e. in agreement with the findings of Steinschneider et al. 1995). Thus, one can conclude that temporal or synchronous patterns may be encoded in humans by phase- or time-locked neurons that fire in response to the onset of spectral or other temporal elements (e.g. burst onset, onset of periodicity) or alternatively, by combination-sensitive neurons analogous to the neural maps for time interval and echo delay (distance) found, for instance, for mustached bats (Suga & O'Neill 1979).

The concepts of pitch, fundamental frequency, and periodicity overlap, since pitch is the perceptual correlate of fundamental frequency, and fundamental frequency is defined by periodicity. Accordingly, these terms will be discussed in parallel. The temporal regularity of a waveform constitutes its periodicity; in this sense, periodicity is a temporal cue (e.g. Langner 1992). Periodic and aperiodic waveforms are associated with voiced and voiceless speech sounds, respectively, and may serve to distinguish—among other potential cues— /s/ (aperiodic) from /z/ (periodic), for instance.

The frequency of a sound wave, such as that produced during speech production, corresponds to the number of occurrences of a repeating event (i.e. wave cycle) per time unit (e.g. 100 events per second = 100 Hz), which may or may not be periodic. The fundamental frequency (F0) and the derived harmonics (integer multiples of F0) in complex speech sounds are, however, periodic waveforms. The frequency of the

waveform increases proportionally as the repetition rate increases, and this results in the *perception* of a higher pitch. Some periodic signals, such as F0 in speech sounds, are also characterized by amplitude-modulated cycles, as exemplified in Figure 2–9. Periodicity (in speech) and F0 are referred to as amplitude-modulated components (from the perspectives of physics), which are not to be confused here with the definition of amplitude associated with the general percept of loudness in the field of phonetics and linguistics. Lower frequencies (top of Figure 2–9) exhibit a lower number of waves within the same time unit than higher frequencies (bottom), although the number of sound waves during one amplitude cycle may be the same. These waves vary in amplitude within their respective cycles, and lower frequencies have lower amplitude modulation (i.e. the change in amplitude between waves is less sharp) as a result of "wider" waves, as compared to high-frequency waves. Amplitude-modulated components (AMs) are generally defined by their modulation frequency viewed in temporal terms (e.g. Langner, Albert & Briede 2002, cf. Krebs, Lesica & Grothe 2008), and have been shown to play an important role in speech recognition (Shannon, Zeng, Famath, Wygonski & Ekelid 1995) and pitch perception (Rossing & Houtsma 1986). In turn, pitch perception plays a crucial role in the perception of speech contrasts, such as lexical tones in Mandarin, lexical accent in Japanese, and lexical stress in English. Pitch perception also contributes to the percept of information-bearing intonation contours at the phrase and sentence levels. Neural maps specially tuned to AM components have been documented in the cat (Schreiner & Urbas 1986), mynah bird (Hose, Langner & Scheich 1987), chinchilla (Langner, Albert & Briede 2002), gerbil (Krebs, Lesica, & Grothe

2008), and guinea pig (Middlebrooks 2008), and therefore, it is most probably the case that comparable neural maps are also present in humans.

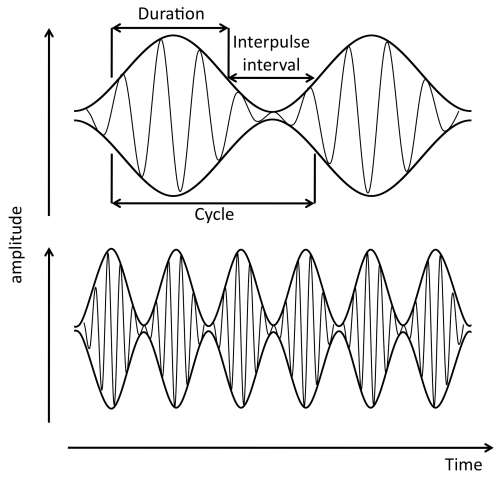

Figure 2–9 Examples of amplitude-modulated sine waves. The lower frequency wave (top) exhibits a fewer number of events per unit of time (over *x-axis*) and lower amplitude modulation than the higher frequency wave (bottom).

The perception of pitch is related to the perception of the fundamental frequency, as determined by the temporal regularity and repetition rate of the waveform. The fundamental frequency is typically higher in amplitude (i.e. louder) than its derived harmonics, and is, therefore, perceptually more salient than other spectral peaks (i.e. than the individual harmonics). However, the same pitch can be perceived even when the acoustic energy corresponding to F0 is removed, suggesting the existence of pitch-specific correlates that can still be captured by the neural system. Bendor & Wang (2005) found neurons in the auditory cortex of marmoset monkeys that respond to pure tones and missing fundamental harmonics of complex sounds with the same F0. Exactly what those neurons respond to is not known, but an experiment with gerbils conducted by Krebs,

Lesica and Grothe (2008) suggests that neurons sensitive to frequency modulation in the mammalian midbrain may also respond to other correlates associated with amplitude modulation components, such as the duration of the amplitude cycle. In any case, pitch is undoubtedly a predominant feature of the communication systems of humans and other species. While neurons that are able to capture some of the correlates associated with pitch have been found, their exact organization and function are still debated.

In sum, segmental and suprasegmental speech contrasts found in natural languages are derived from combinations of only a few spectral and temporal components. Although the exact functioning of the human brain remains elusive, neurons or neural maps sensitive to acoustic components similar to those used in human communication have been documented in other species. Given that the topography of neural organization usually reflects the acoustic behavioral needs of the species, it is probable that the same is true for humans. Hence, one would expect that the neural architecture in the human brain, especially in the midbrain and auditory cortex, should mirror the behavioral data observed in perceptual experiments, particularly in terms of the categorical perception of acoustic parameters.

Given what we know about speech perception (as summarized in sections 2.1 and 2.2,) and about neural processing (section 2.3), it should be possible to create a model that builds on those presented in section 2.2 but that is grounded in neural processing. In fact, some neural-based models have already been proposed: Sussman and colleagues (1991) proposed a neural-based model for identification of stop place of articulation (described in section 2.3.3), whereas Guenther and colleagues (1999; 2002) proposed a neural-based model of categorical learning of speech categories that account for the so-

called perceptual magnet effect and for the effect of different types of training on neural organization (as described in section 2.3.2). These models represent a significant step towards a multidisciplinary approach to speech sounds processing, but have not been developed fully enough (yet) to be firmly based on phonological facts, which would enable these models to be used for linguistic research. This is the gap that the BLIP model, presented in chapter 3, is meant to fill.

2.4 From neural processing to speech categories in a nutshell

Neural development during L1 acquisition may impact on the perception of novel L2 contrasts later in life, as discussed in chapter 4 of this work, hence the importance to assess and understand neural processing for L1 perception. Based on the review of literature presented in this chapter, it is possible to make a few assumptions about first language development. During the first few months (and years) of their life, normal-hearing infants are exposed to one or more languages. What infants are hearing, however, are not words or sounds, but series of spectral and temporal changes in the acoustic waveform (i.e. noise bursts, spectral peaks around a given frequency, silent gaps, periodic and aperiodic signals, etc.) Based on neuroethology data, infants are presumably born with neurons in their brain fully capable of processing at least basic acoustic components like noise bursts, constant frequency components and frequency modulated components as documented in other mammalian species (bats, cats, chinchillas, monkeys, and so forth), as well as neurons or combinations of neurons apt at processing timing information, an ability crucial, for instance, for sound localization. Whether those

neurons are pre-arranged in neural maps roughly corresponding to sound contrasts found in natural human languages is unknown, and not crucial to the current scenario.[16]

Acoustic cues in the input appear to generally have a statistical distribution roughly corresponding to their contrastive status in a particular context, to which infants (and adults) are sensitive. Neurons that fire to these combinations become associated and their associations strengthen through intensive exposure, giving rise to the creation of invariant neural parameters (or neural maps). The initial neural maps presumably correspond to rough (not necessarily adult-like) speech categories relevant for discriminating speech contrasts used in the infants' ambient language, and are later refined possibly as a result of lexical and motor development, and literacy (as discussed in the following chapter).

The creation of neural maps is also coupled with mechanisms working at reducing the processing time of the speech input and at enhancing the ability to categorize highly variable information more efficiently. This is potentially achieved by decreasing the number of neurons responding to stimuli within the perceived categories as a result of some kind of inverted magnification factor. In order to create invariant parameters, normalization may be performed on the speech input, especially on frequency components. The neurology is possibly capable of handling this task, though how exactly remains elusive.

---

[16] Although some universal arrangement of the neural network is likely, it is not necessarily the case that infants are born with neural maps corresponding to adult-like speech categories. It is imaginable that the discontinuities in perception observed in newborns (e.g. Kuhl 1993b) may be the result of perceptual discontinuities in the functioning of other parts along the auditory pathway, possibly in the cochlea or auditory nerve, prior to the signal being processed in the central auditory system.

Far from providing a fully satisfactory account of language development, this scenario brings up further questions that must be addressed. In particular, it is reasonable to wonder what is the correspondence between invariant parameters and linguistic concepts such as features, phonetic categories/allophones or phonemes? How are multiple cues, including but not limited to acoustic ones, processed in relation to one another, whether they contribute to providing similar or contradictory categorical information pertaining to the perception of a feature, phoneme, mora or syllable? If the neurology can normalize the input in order to deal with speaker variability, does that mean that speaker variability is totally ignored? How does neural mapping differ cross-linguistically? And how does neural mapping in an L1 impact on the perception and acquisition of an L2 later in life? In this thesis I attempt to provide tentative answers to these questions by articulating a model of speech perception informed by neurological processing that is linguistically sophisticated enough to deal with the kinds of questions that psycholinguistics have been addressing in their work. The principles of the model are presented in chapter 3, its implications for L2 perception specifically addressed and tested in chapter 4, by mean of four behavioral experiments with speakers of Canadian French, North American English and Japanese.

Chapter Three: The Bi-Level Input Processing Model

In the previous chapter, I reported and discussed studies conducted on humans and animals suggesting that speech contrasts and perception of these contrasts are based primarily on a limited number of *acoustic* (as opposed to *articulatory*) components, mainly spectral and timing components: noise bursts (NB), formants (constant frequency components-CF), formant transitions (frequency-modulated components-FM), periodicity and pitch correlates (amplitude-modulated components-AM). Crucially, neurons responding specifically to these components have been documented in various species and are believed to be active in the human brain as well (e.g. Eggermont 2001; Suga 2006). Although organization of these neurons in the human brain is still unclear, studies conducted on animals suggest that neural organization should generally reflect how the acoustic components are perceived by each species (Suga 2006).

Over the past decades, neural-based models of speech processing have begun to emerge in an attempt to better understand the exact functioning of the auditory cortex for speech processing by humans. Particularly relevant to the current work are the speculative neural-based model of vowel normalization proposed by Sussman (1986), and the neural-based account of the perceptual magnet effect proposed by Guenther and colleagues (Guenther & Bohland 2002; Guenther et al. 1999, 2004) described in chapter 2. Building on these models and on the assumptions about the neural grounding of speech processing as discussed in the previous chapter, the goal of chapters 3 and 4 is to consider the possible link between neural processing and specific linguistic components which are related to sound usage. That is, how are concepts such as features, allophones, and

phonemes, instantiated by the neurology?[17] In which way is the neural processing of acoustic components language-specific? And finally, can language-specific differences in the neural processing of acoustic components account for the impediments encountered by mature L2 learners with non-native speech contrasts? In an attempt to provide answers to these questions, I propose the Bi-Level Input Processing (henceforth BLIP) model, justified and exemplified in this chapter and the next.

The BLIP model endeavors to capture cross-linguistic differences in the processing of spectral and timing components, and to provide a framework for the study of language acquisition and development. In the current chapter, I am concerned specifically with how the neural mapping of speech categories emerges during L1 acquisition, and how the processing of acoustic components differs cross-linguistically. Implications of the model for L2 perception and acquisition are addressed and empirically tested in the next chapter.

In accordance with the emerging consensus that two levels of speech processing (prior to lexical encoding) are necessary and sufficient to account for the range of behavioral data observed in the linguistic and psycholinguistic literature, the BLIP model defines what these two levels may correspond to from a neural perspective: a neural mapping level and a phonological level. Given the prominent role that the neural mapping level plays in the development of speech categories, as argued in the BLIP model, and provided that this level is the main innovation of the model (as compared to

---

[17] To the best of my understanding, previous models do not specifically provide a neural-based account of all of these three linguistic concepts. For instance, the model proposed by Guenther and Bohland (2002) is said to account for *phoneme* category learning, while Sussman (1986) provides a short discussion of the phonological neurogenesis in which he refers to vowel *phones*.

previous psycholinguistic models), this work focuses mainly on the description of this level. The phonological level of processing is addressed briefly here since further work is needed to assess its exact functioning and neural grounding. I begin this chapter by laying out the general principles of the BLIP model (section 3.1). In section 3.2, the mechanisms of the neural mapping level (i.e. first level) posited by the BLIP model are exemplified as applied to the processing of fricatives (3.2.1), vowels (3.2.2.), stops (3.2.3), and suprasegmentals (3.2.4). The phonological level (second level) posited by the BLIP model is exemplified in section 3.3, where I explain how, under the current approach, neural maps are associated with abstract phonological units (3.3.1), how the model views the processing of speaker and dialectal variability (3.3.2), and how patterns of neural maps can deal with misleading or missing information (3.3.3). A summary of the main proposals of the BLIP model is provided in the last section (3.4). In the next chapter (chapter 4), I explain the implications of the model for the study of L2 speech acquisition, and exemplify experimental procedures used to test the predictions of the model that can serve to identify the source of language learners' difficulties with L2 speech contrasts. I will not, therefore, focus on these issues in the current chapter.

3.1 Assumptions and proposals of the BLIP model

The BLIP model endeavors to account for the categorical processing of speech sound contrasts in perception. It is not intended to account for lexical access or retrieval, nor to account for phenomena related to the production of those contrasts. Speech production involves other mechanisms that arguably go beyond those involved in perception;

therefore, a different model is required to account for these mechanisms.[18] The BLIP model differs from other linguistic or psycholinguistic models by its approach informed by neural processing, and differs from previous neural-based models by its attempt to link neural processing with specific linguistic concepts, mainly features and allophones (a.k.a. phonetic categories), and by proposing that acoustic cues are potentially processed by the neurology in three different ways for categorical perception of speech contrasts: additively, competitively and connectively. In other words, the assumptions of the BLIP model are grounded in neurological processing as described in chapter 2, but the BLIP model itself is primarily a psycholinguistic model and is designed for (L1 and L2) linguistic research.

The BLIP model assumes that speech contrasts are based on the perception of spectral and timing components such as noise bursts (NB), constant frequency components (CF), frequency-modulated components (FM), amplitude-modulated components (AM), or some combinations of these (e.g. CF-CF, AM-AM), and that the human brain has neurons sensitive to variations in these components. Neurons that are responsive only when two or more components are present in the input (e.g. CF-CF) are called combination-sensitive neurons and have been shown to be generally species specific. The BLIP model also assumes that the statistical distribution of acoustic cues in the input provides sufficient invariance to be captured by the neurology (provided that this information is successfully extracted by the perceiver) and contributes to shape

---

[18] The interested reader is referred to Guenther (1995, 2006) and Guenther, Ghosh, & Tourville (2006) for a neural-based approach to speech production (the DIVA model).

neural maps that roughly reflect this distribution. Finally, it assumes that speech processing is a categorical task that involves different neural processing and development (i.e. inverted magnification factor) than are required by a discrimination task.

As a model designed specifically to account for cross-linguistic differences in the processing of acoustic cues and to serve as a springboard for the study of language acquisition and processing (in L1 and L2), the BLIP model makes innovative proposals and hypotheses. First, it proposes that there are two levels of processing encoded separately by the neurology: the *neural mapping* level and the *phonological* level.

The neural mapping level consists of neural maps that are based on the statistical distribution of acoustic cues in the speech stream. The development of neural maps can be illustrated as in Figure 3–1, where each empty square represents a single neuron, and the lines going through the squares represent the strength of their interconnection. Prior to any reinforcement (that is, at birth), connections between neurons presumably conform to a general initial organization,[19] as shown on the left side of the figure.

[19] The exact initial organization of neurons and neural connections prior to language exposure is still unknown and not crucial to the current discussion. However, the changes in perceptual sensitivity documented in infants during the first year of life (e.g. Kuhl 2007) suggest that neural organization is *modified* through language exposure. This position is consistent with the (inverted) magnification factor hypothesis according to which the density of cell density activation is affected by the statistical distribution of stimuli in the input (e.g. Baur, Der & Herrmann 1996; Kohonen 2001).

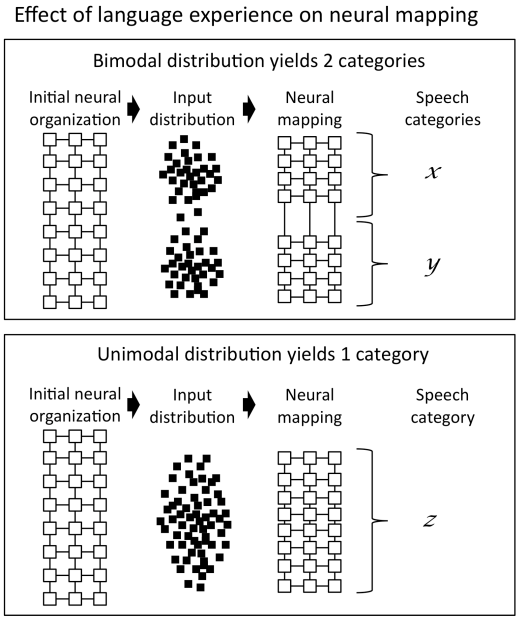

Figure 3–1 Neural mapping development during first language acquisition.

Neurons in the auditory cortex are generally tuned to respond (i.e. fire) to specific acoustic components.[20] For illustrative purposes, let's assume that the neurons in the figure are tuned to respond to constant frequency (CF) values. As explained in the previous chapter, neurons do not fire solely when they encounter their best frequency (i.e. frequency value to which they are specifically tuned), but react to closely related frequency values as well. As a hypothetical example, if a stimulus with a frequency of 1000Hz is detected, neurons in the cortex tuned to frequencies between, let's say, 950Hz and 1050Hz may fire to some degree, with neurons tuned to 1000Hz reacting most

[20] As mentioned in chapter 2, I am only concerned here with the processing of acoustic cues by neurons in the auditory cortex, since these neurons (or their organization) are believed to specifically encode speech categories (e.g. Nelken 2008; Suga 2006). Note that the processing of frequency components by hair cells in the cochlea of the inner ear or by nerve cells in the basilar membrane, though similar, significantly differs from the processing of the same frequency components by neurons in the cortex. Suga (2006: 166), in particular, observed that "the frequency axes in the auditory cortex are not exact copies of the frequency axis along the basilar membrane. Certain portions of it are shrunken or stretched."

strongly.[21] The important point here is that it is not only neurons tuned specifically to 1000Hz that will be activated, but neurons in their neighborhood as well. As such, neurons fire in groups. And when neurons repeatedly fire together, the strength of their interconnectivity increases (e.g. Hebb 1949). Accordingly, if a group of CF neurons is repeatedly activated by two sets of stimuli, one ranging in frequency from 900Hz to 1000Hz, the other ranging in frequency from 700Hz to 800Hz, neurons firing to the 900-1000Hz range will gradually become strongly interconnected, while the strength of their connectivity with the group of neurons firing to the 700-800Hz range will gradually decrease. Similarly, neurons firing to the 700Hz to 800Hz will also become strongly interconnected while the strength of their connection with the 900Hz to 1000Hz group will decrease. Hence, exposure to such a bimodal distribution will shape the neural organization into two separate neural maps, as illustrated in Figure 3-1 (top).

On the other hand, if the range of stimuli generally varies from 700 to 1000Hz, this unimodal distribution will shape the neural organization into a single neural map, as illustrated in Figure 3-1 (bottom). The grouping of neurons in Figure 3–1 does not represent a physical displacement of neurons through space, but rather an increase through exposure in the strength of the connections between them: neurons illustrated as closely spaced are more strongly interconnected than those depicted as more widely spaced. In general, neurons with the strongest connections fire together, hence the

[21] The range of frequency values provided here is for illustrative purposes and is not meant to represent any specific speech contrast or speech categories.

apparent disconnection of neurons into virtual separate neural maps in the bimodal distribution scenario.

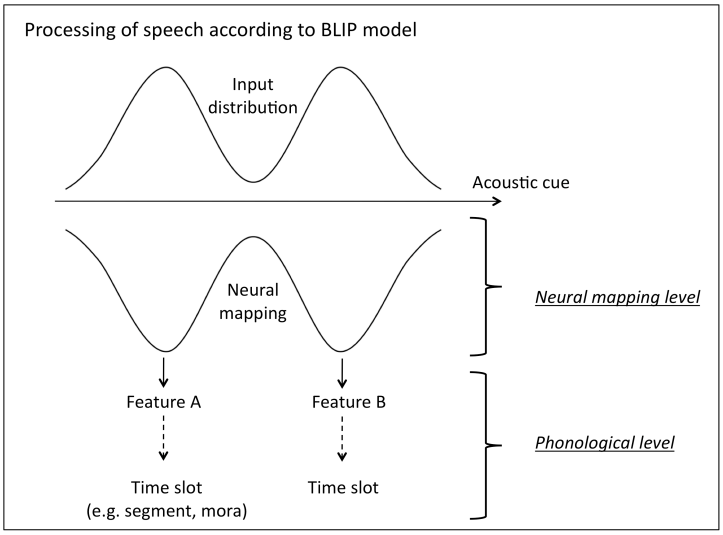

Figure 3–2 Processing of speech contrasts according to the BLIP model. The input distribution is processed by neural maps, which are in turn associated with contrastive behaviorally relevant features.

In the BLIP model, each neural map is posited to be associated with a phonological feature contrast, as show in Figure 3–2 below. The term feature is used here to refer to a contrast that is relevant for lexical distinction.[22] Features are speculated to be part of the phonology and to be encoded by higher order neurons (that is, by different neurons than those at the neural mapping level). While neurons at the neural mapping level are sensitive to physical characteristics of the sound wave (i.e. input) such as CF

[22] The definition of feature adopted here departs from both the phonetic view (a.k.a. "phonetic feature" as discussed in chapter 2 when presenting the PRIMIR model) and the traditional phonological view. How the definition of feature departs in this work from previous views is discussed in more detail later in this chapter.

and NB components, neurons at the phonology level encode information about the role of these maps for speech contrasts. In other words, neurons at the phonological level encode information about abstract and behaviorally relevant features that are meaningful to discriminate words or morphemes in a specific language.

In most cases, there is not a one-to-one correspondence between a neural map and a feature: more than one map can be associated with the same feature. However, one neural map *cannot be divided* into multiple features. This last point is crucial to the definition of feature adopted in this work. The term feature is used here, for lack of a better term, to refer to a distinctive, behaviorally relevant characteristic of a speech sound (i.e. relevant for lexical contrast) *that is captured by at least one neural map*. In the case of most consonantal sounds, the use of features in the current model is generally equivalent to the notion of features posited in phonological models such as Feature Geometry (FG), although the terms used in this work generally follow broad phonetic descriptions for clarity which may or may not correspond to a previously posited phonological feature (e.g. FG [voice] vs. BLIP |voice|; FG [labial] vs. BLIP |labio-dental|.) Note that at this point in the development of the BLIP model, the exact terms used to refer to a given feature is not crucial, since the idea of feature in this work is meant simply to refer to a contrast that is potentially relevant for meaning distinctions. In this sense, the features are phonological. However, in the case of vowels and suprasegmentals, the term feature as employed in this work differs considerably from traditional phonological views. For instance, following the proposal in Sussman's neuronal model of vowel normalization (1986), the BLIP model posits that the vowel

quality (e.g. |i| or |e|) *is* a feature, and this feature is not divisible (by the neurology) into features such as |high| and |back|, since neurons that process vowel contrasts respond to a combination of F1 and F2 values, rather than to only one value.[24] Despite this difference, the current view and the phonology view of features are not mutually exclusive; they only serve a different purpose. In the BLIP model, the term feature is used to refer to a *neurophysiological* contrast (i.e. which identifies the role of a neural map), whereas in phonological theory, the term feature refers to a *psychological* contrast that may play a role, for instance, in phonological processes. As discussed later in this chapter, one does not necessarily exclude the other, but this distinction is still crucial under the current approach. To emphasize this distinction, and to distinguish features from allophones, [25] features are presented between upright bars (e.g. |feature|) rather than in squared brackets, which are reserved to notate allophones (e.g. [allophone]). Under the current view, allophones are variations of a sound *in the input* (i.e. in the incoming signal) and are *not* represented at the phonological level (because they are captured at the neural mapping level instead), whereas features are abstract meaningful contrasts represented at the phonological level (i.e. encoded by neurons specialized to process a given feature). Higher levels of representation, such as abstract phonemic representations, are notated with the conventional slashes (e.g. /phoneme/). I will come back to the distinction of

---

[24] See section 3.3.1 for more details.

[25] Allophones are similar but different concepts from *phonetic categories* as used in psycholinguistic literature, and refer here simply to sounds that are acoustically different but that are not associated with different categories at the abstract, phonological level. That is, they may or may not be in complementary distribution.

feature as used in this model as opposed to the way this term is generally used in linguistic models later in this chapter with concrete examples and justification.

Another innovative proposal of the BLIP model is the hypothesis that acoustic cues relevant for speech contrasts may be processed by the neurology *additively*, *connectively*, or *competitively*. Two or more cues are processed *additively* when they are processed by different groups of neurons, for instance, one cue is processed by CF neurons while another cue is processed by AM neurons, and the cues are associated with *different* features (e.g. CF map → |labio-dental|, AM map → |voiced|). Two or more acoustic cues are processed *connectively* when they are processed *in relation to each other* by the same group of neurons and associated with only *one* feature (e.g. CF-CF map → |i|). Finally, two or more acoustic cues are processed *competitively* when they are processed by *different* groups of neurons but associated with the *same* feature (e.g. NB map → |alveolar| ← FM map). These different types of processing posited by the BLIP model are summarized in (1) below, and exemplified in the following sections.

(1) Speech-relevant cues can be processed:

A. Additively:	Two or more cues are processed *separately* by different groups of neurons and associated with *different* features;
B. Connectively:	Two or more acoustic cues are processed *in relation to each other* by the same group of neurons and associated with only *one* feature;
C. Competitively:	Two or more cues are processed *separately* by different groups of neurons and their relative relevance weighted in an attempt to associate them with only *one* feature value.

As illustrated throughout this chapter and the next, the two levels posited by the BLIP model (i.e. neural mapping and phonological) combined with the speculated three types of processing (i.e. additively, connectively and competitively) provide a framework suitable for the study of cross-linguistic differences in the categorical processing of speech contrasts, as well as for the study and better understanding of L2 speech perception. In particular, the two separate levels enable the BLIP model to account for the processing of allophones, and their role in L2 perception. The BLIP model follows Guenther and Bohland's (2002) original proposal suggesting that while two acoustically different sounds (e.g. /r/ and /l/) in a given language may be processed by two different maps along the relevant contrastive acoustic dimension (e.g. third formant), this cue may be processed by one overlapping map in a language in which these sounds are used as allophones (i.e. not contrastive). In the BLIP model, this theory is extended to account for the perception and neural mapping of vowel contrasts in different languages. In addition, the BLIP proposes that context-bound allophones are processed independently at the neural mapping level provided that their distribution in the input is contrastive enough to enable infants to forge distinct neural maps based on this distribution. Unlike for phonemic contrasts, the neural maps used to process context-bound allophones are associated with the same feature at the phonological level in the adult, mature brain.

In the next section, I begin by providing concrete examples of the concepts of the neural mapping level as proposed by the BLIP model. For the sake of simplicity and clarity, the neural maps are sometimes shown in this section to relate to specific sounds (e.g. /t/, /s/ or /v/), although as explained in section 3.3 the neural maps are posited to be associated first with a feature contrast rather than being associated with a phoneme.

3.2 Neural mapping level

The speech stream can be decomposed into a few spectral and temporal components: noise bursts with spectral peak frequency, constant frequency components (formants), frequency-modulated components (formant transitions), duration, periodicity, and pitch correlates, including but not limited to F0 (see chapter 2, section 2.3 for a review and relevant references). These components are particularly relevant because neurons that respond specifically to values of those components have been documented in various mammalian species and are believed to contribute to language processing in the human brain (Suga 2006). The organization of these neurons or neural maps for speech processing by humans is unknown. However, studies conducted on animals suggest that neural organization should generally reflect how speech-relevant acoustic cues are perceived (Suga 2006). Hence, the BLIP model was designed to conceptualize the possible relationship between neural organization and human speech perception.

This section details the mechanisms posited by the BLIP model to account for how spectral and temporal components may be captured by the neurology at the neural mapping level for the categorical perception of speech contrasts. In particular, this section exemplifies the neural mapping of acoustic components as applied to the perception of fricatives, vowels, stops, and suprasegmentals. I also use these examples to illustrate the different types of processing (additively, connectively and competitively) and the neural mapping of allophonic variations.

It is not the goal of the current work to illustrate the neural processing of *all* possible acoustic cues that may contribute to the identification of a contrast; nor do I aim

to provide an extensive review of cross-linguistic differences in the processing of those contrasts: English, French, Japanese, and Mandarin Chinese are the languages used for cross-linguistic comparison (although not for all contrasts) in this work. As a starting point for modeling speech processing, I assume here perception in ideal conditions, that is, when sounds or words are presented in isolation. How speech is perceived and processed in fluent speech is explored in section 3.3.3. Finally, although normalization may be captured by the neurology (see section 2.3.3), for the sake of simplicity I generally discuss the processing of spectral components without addressing this issue, keeping in mind that normalization may occur prior to processing at the phonological level.[26]

3.2.1 Mapping of fricatives

A neural approach to speech processing arguably implies that the primitives used by the perceiver (the brain) for sound contrasts are acoustic rather than articulatory, at least in the case of spoken language, since this is the information processed along the auditory pathway from the ears to neurons in the cortex. In most cases, speech sounds can be characterized and contrasted by more than one potential acoustic cue relating to their place and manner of articulation. Thus, the first step in understanding the neural processing of speech is to identify which of these acoustic characteristics serves as the basis for perceptual contrast, and second, to identify which aspects of the processing of

---

[26] The interested reader is referred to Sussman (1986) for an example of how the neurology may deal with normalization.

these cues are language-specific. I begin this section by using the processing of fricatives to illustrate the fact that one speech sound may have various distinctive acoustic characteristics, and that one must first identify the most reliable cue for speech processing to model speech perception. This cue must not only be reliable, it must also be captured by distinct neural maps (i.e. be captured categorically by the neurology). I use the case of fricatives to demonstrate how the BLIP model can capture cross-linguistic differences in the processing of speech contrasts.

Fricative place of articulation

As a concrete example, fricatives are generally contrasted according to their place of articulation, which corresponds to the location of the primary constriction in the oral cavity. Various acoustic correlates associated with turbulent air flow have been investigated as potentially relevant perceptual properties to distinguish fricatives: local spectral properties of the noise (e.g. Behrens & Blumstein 1988; Evers, Reetz & Lahiri 1998; Heinz & Stevens 1961; Hugues & Halle 1956; Jongman, Wayland & Wong 2000; Shadle 1990; Strevens 1960); spectral moments, which include the mean frequency, spectral tilt, and peakedness at different regions of the signal (e.g. Forrest, Weismer, Milenkovic & Dougall 1988; Jongman, Wayland & Wong 2000; Nittrouer, Studdert-Kennedy & McGowan 1989; Tomiak 1990); locus equations based on changes in the formant transition from vowel onset to midvowel nucleus (e.g. Fowler 1994; Jongman, Wayland & Wong 2000; Sussman 1994; Sussman & Shore 1996; Wilde 1993; Yeou 1997); overall noise amplitude (e.g. Behrens & Blumstein 1988; Jongman, Wayland & Wong 2000; Strevens 1960); relative amplitude, captured by the change in amplitude

from the fricative to the vowel (e.g. Jongman, Wayland & Wong 2000; Stevens 1985); and noise duration (e.g. Behrens & Blumstein 1988).

Noise duration has generally been found to be a relevant cue only for voicing contrasts (Behrens & Blumstein 1988), not place of articulation, while locus equations have yielded contradictory and inconsistent results in the case of fricatives. A study by Jongman, Wayland & Wong (2000) that compared all the above-mentioned acoustic properties, except for noise duration, found that spectral peak location, spectral moments (as defined in the previous paragraph),[27] and overall and relative amplitude can successfully distinguish all four fricative contrasts in English: labiodental /f, v/, interdental /θ, ð/, alveolar /s, z/, and palato-alveolar /ʃ, ʒ/. Their study is based on an analysis of the English fricatives produced by 10 male and 10 female native speakers of American English. The fricatives were produced in the initial position of CVC syllables produced with six different vowels and the same final consonant /p/ within a carrier phrase.

Although the overall amplitude and relative amplitude might be perceptually salient to distinguish the fricatives, particularly the sibilant from the non-sibilant series, amplitude differences that are associated with the perception of loudness (not pitch) (see footnote 40) may be captured by the firing rate of neurons tuned to other components such as constant frequency components (CF). Hence, it is possible that no separate neural

---

[27] Although the fricatives also exhibit significant differences in the peakedness of their noise distribution as captured by a measure of the kurtosis of spectral moments, this characteristic alone was unable to discriminate the alveolars /s,z/ from the labio-dentals /f,v/ in Jongman et al.'s analyses, as both had high kurtosis values indicating a clearly defined spectrum as opposed to a flat distribution for negative kurtosis values.

map exists to evaluate amplitude differences. In any case, for illustrative purpose, I assume here that fricative contrasts are distinguished instead by their spectral characteristics and analyzed by neurons sensitive to CF components, keeping in mind that amplitude information might be perceived as a change in the firing rate of the same neurons (i.e. within the same neural map).

The spectral peak location of the fricatives is defined in Jongman et al.'s (2000) analysis as the highest-amplitude peak of the fast Fourier transform (FFT) spectrum, examined through a 40-ms full Hamming window in the middle of the fricative noise. The peak frequency value decreases as the place of articulation of the fricative moves back in the oral cavity, yielding four possible distinct neural maps along the frequency axis according to the BLIP model, as illustrated in Figure 3–3. A distribution of English fricatives varying in terms of peak frequency along the *x*-axis and relative amplitude on the *y*-axis (i.e. difference between amplitude of the following vowel and amplitude of the fricative noise) is represented at the top of the figure, whereas the inverted curves (bottom) represent the neural mapping based on this distribution. The average spectral peak locations reported in Jongman et al.'s (2000) study and used as guidelines for this illustration are: 7733 Hz for labiodentals, 7470 Hz for dentals, 6839 Hz for alveolars, and 3820 Hz for palato-alveolars. Voiceless fricatives generally have a higher frequency peak than their voiced counterparts, whereas vowel context exerts a significant influence on peak location only for alveolars, for which the peak frequency is lower in the context of the vowels /o, u/.

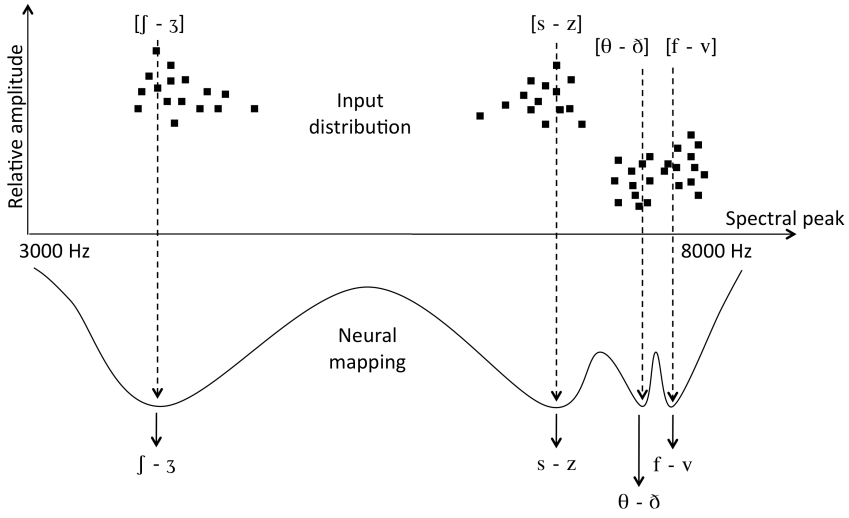

Figure 3–3 Hypothesized neural mapping of English fricatives based on spectral peak location according to the BLIP model (based on the values provided by Jongman et al. 2000).

The term *neural map* in this work is used to refer to a set of neurons that are tuned to distinguish acoustic contrasts, irrespective of whether the neurons are spatially close to one another. In this sense, the neural maps in the BLIP model are primarily *functional* maps (e.g. Eggermont 2001), since the BLIP model is designed to capture the correlation structure (i.e. strength of the interconnection) between neurons that serve to discriminate speech contrasts. The neural mapping level in the BLIP model involves sets of neurons that respond to related values of an acoustic component. Related values are determined by the statistical distribution of those values in the input, as illustrated in the figure above. In line with the inverted magnification factor hypothesis, since these neural maps are used for categorical processing (as opposed to discriminative processing), they are represented by inverted curves, reflecting a decrease in cell density activation around categorical centers. This representation is not intended to reflect any particular physical

shape of the neural maps in the brain, but rather to represent the relevant number of categorical contrasts perceived along a given acoustic dimension, as reflected in a decrease in cell activation at given intervals or values along this dimension.

A neural map is generally illustrated as bi-dimensional as in Figure 3–3 if the acoustic component captures a given contrast relatively independently from other acoustic cues. In other words, if the presence of one cue is sufficient to provide information about a given contrast, it is likely that this cue will be processed by neurons encoding only this information.

A prediction of the BLIP model is that proficient listeners do not attend to exact acoustic values to discriminate speech contrasts, but rather, attend more generally to the number of contrasts along a given acoustic dimension (understandably, within a limited region of this dimension). This proposal is justified with supporting evidence in 3.3.2 below and has implications for L2 perception as discussed in the next chapter. Relative to the current discussion, the important point is that the exact realization of a given contrast, especially in terms of categorical boundary, is not crucial for the perception of this contrast. For instance, even though French speakers' production (and incidentally, perception) of alveolar and palato-alveolar sounds is generally more fronted than English speakers', these sounds are neither predicted nor documented to be problematic for French learners of English.

However, the number of contrasts along an acoustic dimension may differ from one language to another, thereby accounting for cross-linguistic perceptual differences. For instance, it is well known that French speakers encounter difficulties when trying to discriminate the English interdentals from other related sounds. For instance, Canadian

French speakers were shown to encounter difficulties discriminating the English interdentals from English alveolar stops (e.g. LaCharité & Prévost 1999). This difficulty presumably stems from the fact that French lacks interdental contrasts. More specifically, though, the problem may be due to French speakers' use of only three neural maps (as opposed to English speakers' use of four) along the acoustic dimension used to contrast the fricative sounds, as illustrated in Figure 3–4.

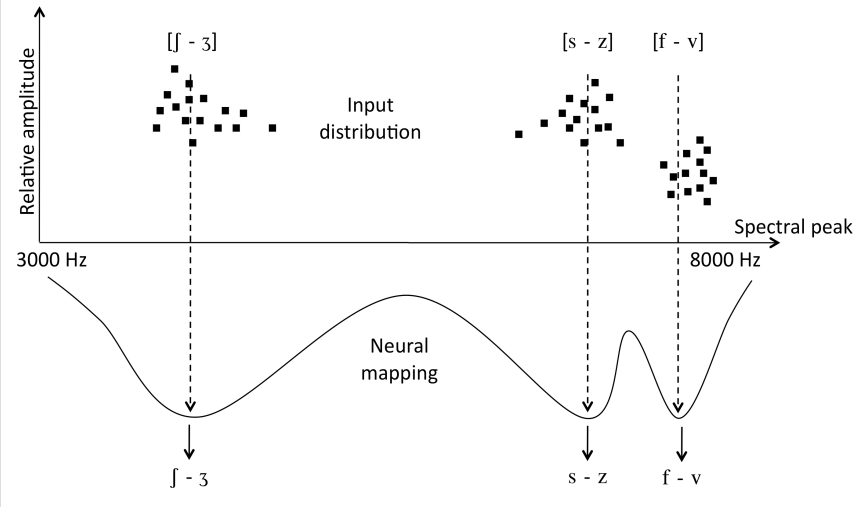

Figure 3–4 Hypothesized neural mapping of French fricatives based on spectral peak location according to the BLIP model.

Thus, even though the categorical boundary between palato-alveolar and alveolar sounds may not be set at the exact same location in English and French, French speakers should nonetheless be able to distinguish English palato-alveolar and alveolar sounds. By contrast, French speakers should initially fail to discriminate the interdental sounds from

neighboring sounds[28] in terms of peak frequency location because they make use of only two, rather than three, neural maps *around the same region* of this acoustic dimension, that is, between the area mapping alveolars and labiodentals. Hence, the number of maps along a given acoustic dimension may partly account for cross-linguistic differences in speech processing.

Fricative voicing contrast

In addition to variations in place of constriction, fricatives may contrast in terms of their voicing status. Based on studies with different types of animals, a periodicity map orthogonal to (i.e. distinct from) the frequency axis has been proposed to exist (see Langner, Albert & Briede 2002 for a review). Neurons in the periodicity maps are sensitive to amplitude-modulated (AM) components (related to fluctuations in the waveform), as described in the previous chapter (refer to section 2.3.4 for a review). However, it is difficult and possibly impossible to conceive of a neural map that would encode the absence of a periodic signal, since neurons generally fire in response to a stimulus. Thus, the voicing contrast for fricatives can be conveyed by the presence or absence of a periodic signal (among other possible cues). Assuming a continuum in the input from no periodicity to full periodicity (depending on the relative proportion of pitch

[28] Why specifically Canadian French speakers appear to confuse the English interdentals with /t, d/, whereas their European counterparts appear to confuse the English interdentals with /s, z/ needs further investigation, since I am unaware of any research that has tested Canadian French speakers' ability to discriminate English /s/ from the voiceless interdental when the spectral peak frequency was progressively manipulated from one category to the other. It is possible that these speakers may also have difficulties discriminating these two sounds in a way similar to European French speakers based on that cue, as suggested by the BLIP model.

periods over the total duration of the sound production), the neural mapping of the voicing contrast may be conceptualized as in Figure 3–5, where there is a neural map dedicated to capturing periodicity. In this sense, and to employ phonological terminology, voicing (periodicity) can be seen as the marked feature, in that it triggers more neural activation than voicelessness. In addition to capturing voicing contrasts, neurons in the periodicity map are believed to encode information pertaining to the perception of pitch (and F0), since they are especially tuned to variations of the waveform corresponding to low frequencies (e.g. Langner, Albert & Briede 2002). The mapping of pitch-related phenomena is discussed in section 3.2.4.

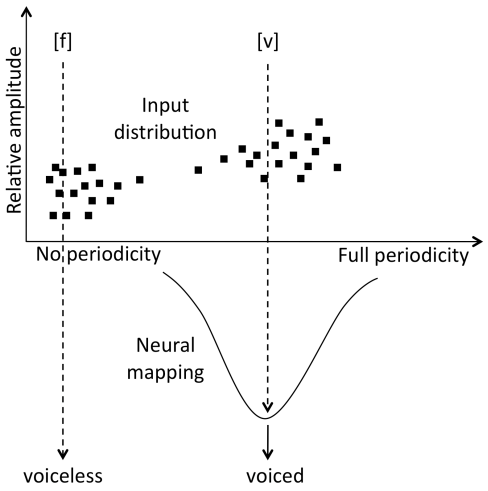

Figure 3–5 Hypothesized neural mapping of periodic contrasts (only periodic waveforms are encoded by neurons along the periodicity axis).

In short, discrimination of fricative contrasts may be partly captured by variations in peak frequency location (or alternatively, spectral moments)—to differentiate place of articulation—and by the presence or absence of periodicity in the signal—to differentiate voiced from voiceless fricatives. The number of maps along the acoustic dimension may

account for cross-linguistic perceptual differences, and at least partly explains the perceptual difficulties encountered by L2 learners, for instance, the challenges experienced by French speakers in discriminating English interdentals. Understandably, additional acoustic dimensions may be used to make a categorical distinction about these sounds. Thus, interdentals will not necessarily be assimilated to one of the neighboring sounds along this dimension. On the other hand, the fricative voicing contrast, based on the presence or absence of periodicity, is more straightforward, since neural overlapping between periodic and aperiodic signals is unlikely, if we assume that neurons are attuned specifically to process periodicity, rather than to distinguish between periodicity and aperiodicity. Consequently, as long as L2 learners are sensitive to the presence of periodicity for fricative contrasts, they should be able to perceive any other fricative contrast based on the presence or absence of periodicity, even if these fricatives are not used in their L1.

3.2.2 Mapping of vowels and their allophonic variations

The previous section demonstrated that fricatives may contrast in terms of voicing and place of articulation, and that these contrasts can be captured by neural maps processing periodicity (AM components) and spectral peak frequency (CF components) separately and independently. Consequently, these maps were illustrated as bi-dimensional. On the other hand, some sounds may require the concomitant processing of at least two acoustic values, one of which may be insufficient to identify the contrastive feature of the speech sound. If one acoustic component must be evaluated in relation to another, the cues are processed *connectively* by combination-sensitive neurons (see chapter 2 for a definition

of combination-sensitive neurons and appropriate references). This type of neural mapping can be illustrated by three-dimensional neural maps, as exemplified in Figure 3–6 (left) or alternatively, by an aerial view of the three-dimensional map (right). Vowels are likely processed by combination-sensitive neurons, as justified below. In this section, I describe how the processing of vowels can be captured by the neural mapping level of the BLIP model, and how allophonic variations of those vowels can give rise to *overlapping maps* in Japanese and *context-bound neural maps* in Canadian French. Finally, the processing of vowel durational contrasts will be explored.

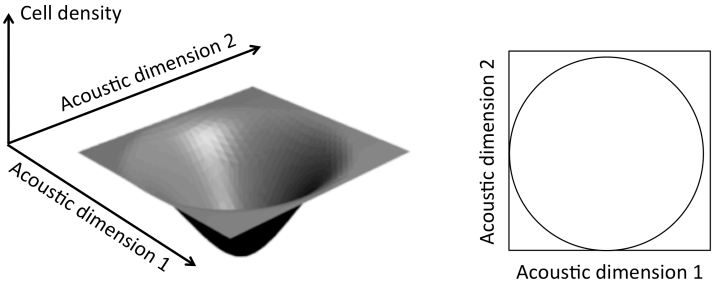

Figure 3–6 Example of a three-dimensional neural map involving the processing of two acoustic cues connectively by combination-sensitive neurons. Full three-dimensional view is shown on the left and an aerial view of the same map is at right.

Vowel quality

Formants correspond to resonances of the vocal tract, and serve as critical acoustic cues for many speech contrasts (Fant 1960; Ladefoged 2001; Lieberman & Blumstein 1988; Titze 1994), particularly vowels, which can generally be distinguished by their first and second formants (Delattre, Liberman & Cooper 1951; Ladefoged 2001).

While it is possible to perceive changes in formant values when only one formant is presented at a time, experimental evidence suggests that the processing of multiple formants differs from the combined processing of single formants. Recordings of isolated neurons in cats, for instance, reveal that the sum of the responses to the low and high frequency parts of species-specific calls (meows) is larger than the neural response to the natural call (containing both), suggesting that processing multiple formants differs from the separate processing of those formants (Gehr, Komiya & Eggermont 2000). Experiments with gerbils indicate that the distance between F1 and F2 is topographically represented in the primary auditory cortex of this mammal (Ohl & Scheich 1997), concurring with the hypothesis that the processing of vowel-like sounds involves a computational comparison of at least two formants by combination-sensitive neurons. The third formant may also provide relevant information for the identification of some human vowels (Ladefoged 2001; Vaissière 2006). However, it is unknown whether F3 is processed in comparison with F1 or F2, or in comparison to F1 and F2 simultaneously. For this reason, I include only the first two formants in the model presented below, keeping in mind that a similar process may be used to compute the value of F3 in relation to other formants.

It is worth considering why each vowel formant should not be processed separately, like the spectral peak frequency of fricatives. Understandably, the peak frequency of a formant is highly affected by the size of the vocal tract, which varies considerably from one speaker to another. Consequently, it may be difficult, especially from the perspective of a newborn, to build reliable categories based on the absolute values of formants. Alternatively, the ratio between F1 and F2 is more stable and reliable

when comparing, for instance, male and female speakers (e.g. Martin 2004). Since combination-sensitive neurons are able to compute such ratios, these neurons are likely more efficient at capturing vowel contrasts, and therefore, are speculated to play an important role in the processing of vowels (e.g. Sussman 1986). As a concrete example, the distribution of high front vowels in English should give rise to two separate auditory cortical maps in the brains of English infants. In the BLIP model, these neural maps are argued to form the initial speech categories developed by infants based on the statistical distribution of formant ratios in the input, as schematized in Figure 3-7. For convenience, the neural mapping of vowels follows the same schematic representation used by the International Phonetic Alphabet (IPA), where the arrows associated with F1 and F2 indicate the direction from low to high frequency.

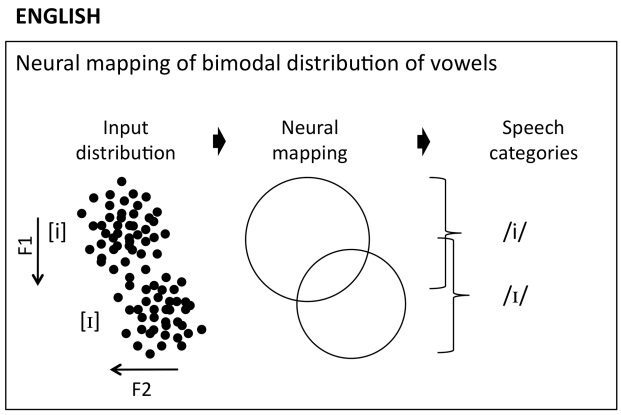

Figure 3–7 Hypothesized neural mapping development of the high front English vowels by L1 learners.

Conversely, in Japanese, since there is only one high front vowel category (Akamatsu 1997, Vance 1987), the Japanese neural map is free to range over the F1 and

F2 vowel space occupied by two vowels in the English neural map.[29] The relationship between the Japanese and English neural maps with respect to this particular contrast is referred to, in the BLIP model, as an *overlapping map* and is illustrated in Figure 3–8. The idea of an overlapping map was first proposed by Guenther and Bohland (2002) to account for native Japanese speakers' difficulties in perceiving the contrast between English /r/ and /l/. Japanese speakers are generally insensitive to variations in F3, which are used by native American English listeners to discriminate the English liquids (Iverson et al. 2003). Guenther and Bohland suggest that the neural map for the Japanese flap overlaps the area along the F3 dimension that features the /r-l/ contrast because Japanese does not have any /r/ and /l/ contrast (hence, the F3 value of the Japanese flap can vary freely along the dimension of F3 generally associated with English /r/ and /l/). Assuming that neural maps are subject to a decrease in cell density activation (which results in a decrease in perceptual sensitivity) around their categorical center, the overlapping map for the Japanese flap is expected to exhibit its strongest decrease in cell density activation around the critical F3 value for the English contrast. Similarly, the overlapping map posited here for Japanese accounts for Japanese speakers' impediments in discriminating the high front English vowels based on spectral information (e.g. Morrison 2002). An

---

[29] Speakers of a language with a relatively small vowel inventory (American English) are insensitive to consonantal context effects on vowels in a language with a comparatively large inventory (North German) for perceptual assimilation in native categories (Strange, Bohn, Nishi & Trent 2005). Conversely, speakers of a language with a relatively large inventory (Danish) are very sensitive to consonantal context effects on vowels in a language with a comparatively small inventory (Southern British English) for perceptual assimilation (Bohn & Steinlen 2003). These findings suggest that speakers of languages with small vowel inventories are more permissive about variations in vowel quality. From a perceptual/neurological point of view, this means that the topographic map is free to range over the F1/F2 space unoccupied by other vowels.

overlapping map means that a single neural map in a given language (in this example Japanese) covers the space occupied by two speech categories (i.e. by two neural maps) in another language (here English). Consequently, the notion of "overlapping map" is dependent on the languages being compared.

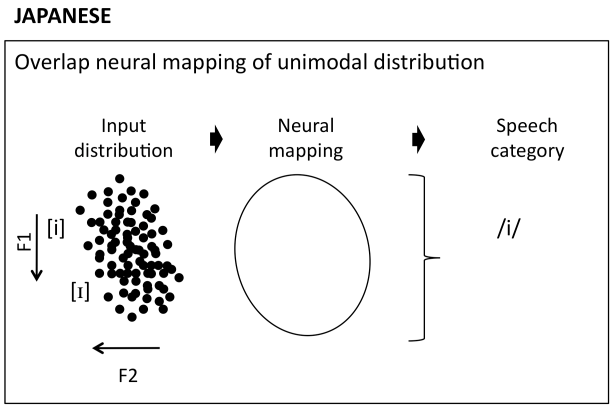

Figure 3–8 Hypothesized neural mapping development of the high front Japanese vowel by L1 learners.

Still a different scenario may occur with Canadian French L1 learners. Speakers of this particular dialect of French are known to consistently produce a high front tense (unrounded) vowel in open syllable position, as in *petit* [pətˢi] 'small, masc.', but to produce a high front lax (unrounded) vowel in closed syllable context, as in *petite* [pətˢɪt] 'small, fem.' (e.g. Martin 1996). Provided that this alternation is consistent and yields a contrastive, albeit context-bound, distribution in the input, infants exposed to this dialect should develop two neural maps based on this distribution, as illustrated in Figure 3–9. Note that this scenario is mostly identical to the neural development of the same vowels by native English speakers, except that here the vowels are context-bound in the input, and the neural maps are associated with only one speech category at the phonological

level (the phonological level is discussed in section 3.3). Provided that the statistical distribution of those vowels does not change during the period the infant is exposed to this dialect, there is no reason for these neural maps to be altered. It is reasonable to suppose that other factors may contribute to the association of the two neural maps with a single speech category in the course of the infant's phonological development. However, the two neural maps are postulated to remain active and to play an important role in speech perception; for instance, these maps may aid Canadian French speakers in the perception of dialectal variations, since this allophonic contrast does not exist in most European French dialects. The latter issue is discussed further when presenting the phonological level of processing. Within the BLIP model, neural maps based on context-bound allophones are referred to as *context-bound neural maps* and are argued to have important implications for L2 perception, as discussed in chapter 4.

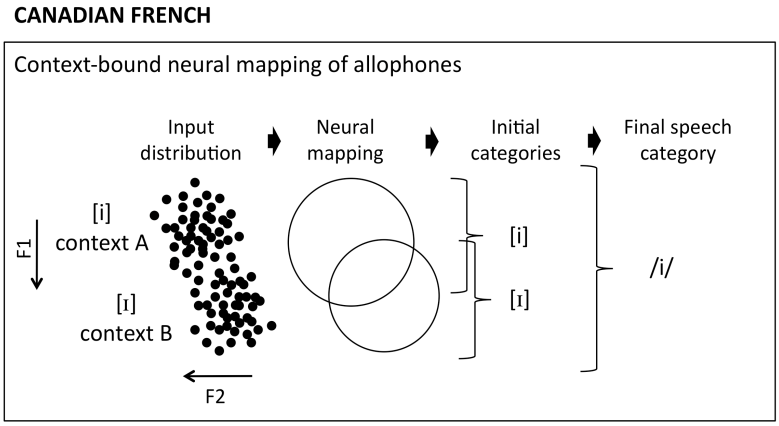

Figure 3–9 Hypothesized neural mapping development of the context-bound high front unrounded vowels in Canadian French by L1 learners.

Vowel duration

Vowel duration is another cue that may be processed by combination-sensitive neurons, especially among speakers of languages such as Croatian, Czech, Hausa, Hungarian, Japanese, Korean, and Thai, who use vowel length contrastively (Handbook of the IPA 1999). For instance, in Japanese the word *chizu* [ʨizɯ] 'map' contrasts with *chiizu* [ʨiːzɯ] 'cheese'. This is not to say, however, that speakers of other languages are totally insensitive to durational variations or that sensitivity to durational changes is restricted to the perception of speech stimuli.

There are three types of neural activation response to stimuli:[30] phasic (or time-locked), tonic, and phasic-tonic. Phasic and tonic responses are particularly relevant to the modeling of durational contrasts. As described in the previous chapter (section 2.3.4), phasic or time-locked discharges occur in synchrony with a particular event, such as the onset or offset of a stimulus. That is, the neural discharge is activated at the onset of the stimulus and then subsides or stops, even if the stimulus persists. Tonic responses, on the other hand, may not be activated quickly in response to the onset of a stimulus, but are generally activated (i.e. produce action potentials) over the entire duration of the stimulus (Randall, Burggren & French 1997). Accordingly, tonic responses intrinsically encode information about the duration of a stimulus, such as a vowel or a ring tone. As a result,

---

[30] The type of neural activation (tonic, phasic or phasic-tonic) is not to be confounded with type of processing as proposed in this work (e.g. additively or connectively) or with types of neurons (e.g. CF or NB neurons). These are three different concepts. For instance, a CF neuron can respond with a phasic or tonic discharge. That is, all three types of response can be used to perceive the same kind of information. For instance, phasic, tonic, and phasic-tonic activation have been observed in response to periodicity information in mammals (e.g. Langer, Albert & Briede 2002).

speakers of any language should be sensitive to durational variations, as captured by the tonic responses of neurons tuned to the acoustic component of speech or non-speech sounds.

Tonic responses to vowel duration may not, however, be sufficient for the *categorical* processing of vowel duration. In fact, the neural activity involved in the processing of vowel duration appears to differ depending on whether this cue is used contrastively in the language. For instance, native Finnish speakers, for whom vowel length is phonemic, exhibit a shift in neural activation (captured by modulation of the mismatch negativity (MMN)) towards the left hemisphere in reaction to changes in vowel duration. This shift is not observed with native German speakers, for whom vowel length is not phonemic (Kirmse et al. 2008). It is possible, thus, that the *categorical* processing of vowel duration is best captured by a combination of two phasic discharges, one at the onset of the vowel and the other at the vowel offset, rather than by tonic responses to a single cue. The onset and offset can be processed connectively by combination-sensitive neurons that compute the time difference between the two. Provided that the combination-sensitive neurons are different from the neurons responsible for capturing vowel quality, this account may at least partly explain the differences in neural activity observed in Finnish as opposed to German speakers in Kirmse et al.'s study (2008).

In terms of modeling, this contrast is schematized in the BLIP model as depicted in Figure 3–10. Note that even though the BLIP model speculates that duration is processed by combination-sensitive neurons, since the initial onset (Time 1) is always the same (it corresponds to 0 ms on a time scale), duration is in fact processed along only one dimension, that is, *time*. Accordingly, bi-dimensional neural maps are sufficient to model

the processing of vowel duration, as illustrated in this figure. For speakers of languages that do not use duration contrastively, these maps are potentially unnecessary (and possibly non-existent), since durational variations can still be captured by the duration of tonic responses to vowel spectral information or periodicity, though presumably not with the same efficiency and accuracy.

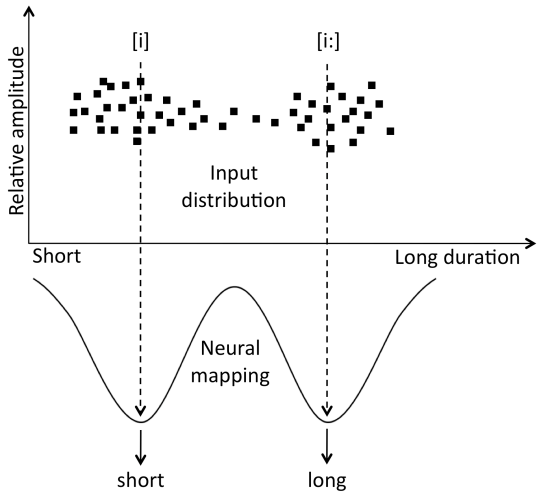

Figure 3–10 Neural mapping of vowel duration by speakers of languages known to use vowel duration contrastively.

To sum up, the neural mapping development of vowel categories by infants learning their L1 may be seen as based on the statistical distribution of at least the first two formants (F1 and F2) processed in relation to each other, that is, processed *connectively*. These formants may be processed by combination-sensitive (CF-CF) neurons, which can be modeled through the use of three-dimensional neural maps, plotting each formant on a separate axis. Provided that L1 neural development is based solely on the statistical distribution of formants in the input, irrespective of phonotactic information, the neural mapping level can account for important cross-linguistic and

cross-dialectal variations in the perception of acoustically close vowels. In some cases, a single *overlapping map* can encompass the area of the input space corresponding to two distinct vowels in another language. In other cases, where the same two vowels are not contrastive at the phonological level (i.e. they correspond to only one vowel in the phonology), but exhibit a strong context-bound bimodal distribution, the neural mapping development should give rise to two *context-bound neural maps*. Vowel duration contrasts may similarly be processed by combination-sensitive neurons that compute the time difference between the vowel onset and vowel offset. However, since duration varies along only one dimension (i.e. time), durational contrasts can be captured by bi-dimensional maps.

3.2.3 Mapping of multiple acoustic cues related to stop contrasts

The perception of stop contrasts, like that of fricatives and vowels, involves the processing of multiple acoustic cues. When different cues are processed separately (i.e. processed by distinct groups of neurons), each contributing to the identification of a different feature, the acoustic components are processed additively. The processing of fricatives provides a good example: spectral peak frequency components, as processed by CF neurons, serve to identify the fricative place of articulation, while periodicity information, as processed by AM neurons, contributes to identifying the voicing status of the same fricative. By contrast, some acoustic components may need to be evaluated in relation to one another. In these cases, the components are processed *connectively* by the same group of combination-sensitive neurons and contribute to the identification of a single feature. Vowel identification requires the concomitant processing of F1 and F2 by

the same group of combination-sensitive neurons.[31] It is nonetheless possible that multiple cues may be processed independently by different groups of neurons, while contributing to the identification of the same feature. In this scenario, the acoustic components may be processed *competitively*. Competitive processing does not necessarily imply that the different neural groups provide contradictory information, but simply that the value of each cue may be compared and even weighed (through the relative strength of the neural interconnections, or the level of discharge activated in each neuron) to identify the perceived or "winning" feature.

This section exemplifies the processing of competing (or complementary) cues as applied to the identification of the stop place of articulation and the stop voicing contrast in different positions. The processing of noise bursts and locus equations for the identification of stop place of articulation have already been the focus of a plausible model based on neural processing in the mature (adult) brain (Sussman et al. 1991). The BLIP model builds on the principles and assumptions discussed in Sussman et al.'s (1991) model, but approaches the issue from the developmental perspective of first language acquisition. Below, the processing of components that contribute to the voicing status of the stops, along with possible cross-linguistic differences in the perception and processing of voicing cues in word-initial and word-final position, are described.

---

[31] Listeners are also sensitive to variations in only one of the formants, which may result in a different vowel quality. Nevertheless, from a processing point of view, even if the F1 varies, the vowel quality is still processed by evaluating F1 and F2 in relation to each other. That is, processing of both formants connectively does not preclude sensitivity to changes in the value of a single formant, a capacity which may be important in phonological processes such as vowel harmony.

Stop place of articulation

As proposed by Sussman and colleagues (1991) two main cues contribute to the identification of stop place of articulation in initial position: information provided by the burst and F2 trajectory cues (locus equations). According to a study by Cooper et al. (1952), English listeners most consistently perceive a voiceless alveolar stop consonant (/t/) when the peak frequency of the burst is high, irrespective of the quality of the following vowel. However, when the peak frequency burst is relatively low (i.e. below ca. 3000Hz in their experiment), identification of the plosive as either bilabial (/p/) or dorsal (/k/) depends on the quality of the vowel that follows. In Sussman and colleagues' model, information provided by the burst is processed in conjunction with information pertaining to the F2 transition for identification of the stop consonant. Hence, while a single cue may be in some cases insufficient to provide a definitive or reliable categorical decision pertaining to the stop place of articulation, the combination of multiple cues concur to facilitate the categorical decision.[32] One may wonder, however, how infants are able to build the neural maps necessary to make proper categorical decisions in the first place, without any prior knowledge of phonological categories. That is, how can they learn to categorize the stop consonants if they do not know how many categories a language includes? I demonstrate below that the input actually provides sufficient non-contradicting information to enable infants to forge proper neural maps prior to lexical development.

---

[32] The processing of multiple cues in the way explained in this section is labeled as *competitive* rather than, for instance, *combinatory*, because even though the information provided by these cues may be complementary during L1 acquisition, they are still processed competitively, in the sense that they may not always complement each other (they may sometimes points to different categories) as discussed later.

During the first months of life, newborn infants face the daunting task of building relevant speech categories, relying solely on the statistical distribution of acoustic components.[33] Initially, the infant's brain, like that of any mammal, is equipped to process at least basic *information-bearing elements* (to employ Suga's terminology): noise bursts (NB), constant frequency (CF) components and frequency-modulated (FM) components. Based solely on these groups of neurons, it is possible to hypothesize how the neural mapping of stop place of articulation begins to emerge in the absence of prior phonological knowledge and contextual information.

Neurons sensitive to noise burst process only the peak frequency of the noise; thus, this information is unlikely to be processed in conjunction with vowel information, at least in the earliest stages of acquisition. Accordingly, infants may only have the statistical distribution of the noise burst to form tentative initial NB neural maps. As suggested by Cooper et al.'s study (1952), the spectral peak frequency of the burst does not provide a reliable cue for all stop places of articulation in English across vowel contexts. However, at least one category can be reliably identified irrespective of the quality of the following vowel: the alveolar stop.[34] The input distribution to which infants from English-speaking homes are exposed potentially resembles that illustrated in Figure 3–11, where the production of the alveolar stop yields a noise burst concentrated at a high

---

[33] Arguably, infants may also rely on motor development and articulatory awareness to develop these speech categories. To what extent and exactly how development in production impact on the neural development used for perception of speech contrasts is an interesting, though still unresolved question.

[34] The stops used in Cooper et al.'s study for evaluating perception based on noise burst center frequency were voiceless. Provided that this component is primarily affected by the place of constriction in the oral cavity, which should be the same for voiced and voiceless stops in English, I assume that roughly the same contrastive values apply to voiced English stops.

frequency, whereas the noise burst of other stops is generally spread over the lower frequencies. Based on this distribution, infants are able to extract at least two categories, one corresponding to English /d/ and the other corresponding to either /b/ or /g/.

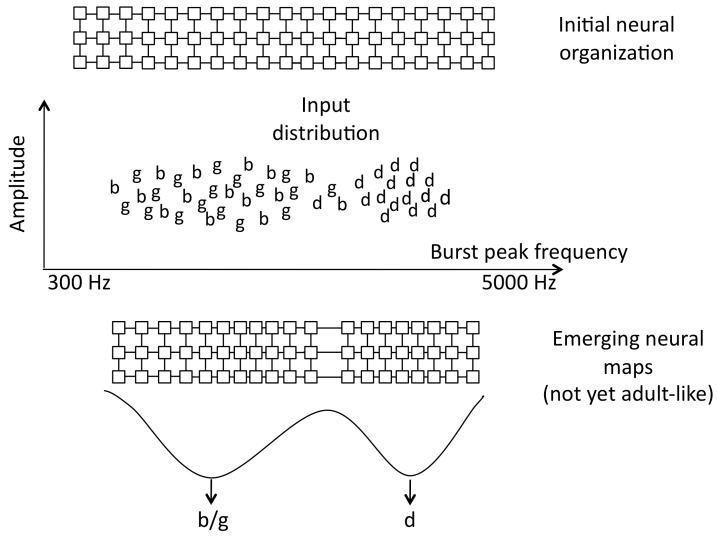

Figure 3–11 Emerging neural maps based on noise burst information in infants from English-speaking homes.

Meanwhile, FM neurons can process information related to formant transitions. F2 transitions into the vowel have been shown by Sussman and colleagues to provide a robust cue for identification of stop place of articulation, and can capture cross-linguistic variations in the realization of stop consonant contrasts (Fruchter & Sussman 1997; Sussman 1999; Sussman, Fruchter & Cable 1995; Sussman, McCaffrey & Matthews 1991). The distribution of these transitions across speakers and vowel contexts forms a strong linear relationship referred to as a locus equation. Neural maps based on these locus equations for English stops can be conceptualized in the BLIP model as being

captured by FM neurons that process the slope from F2 onset to F2 in the vowel, and which can be illustrated as in Figure 3–12.

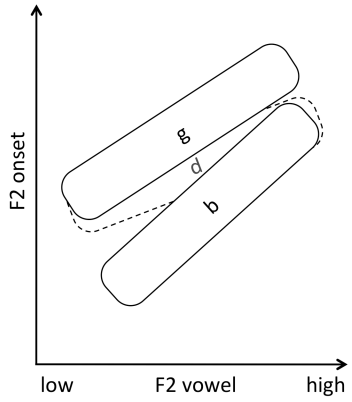

Figure 3–12 Emerging neural maps based on locus equations for English /b/, /d/ and /g/ reported by Fruchter & Sussman (1997, p. 3006) in infants from English-speaking homes. The maps illustrated with plain lines are presumed to emerge first in the BLIP model.

Noticeably, there are considerable areas of overlap between the locus equations for /b/ and /d/ and those for /d/ and /g/. Arguably, however, provided that infants are able to extract the category associated with /d/ from the center frequency of the initial burst, the categories that need to be dissociated with the locus equations are only those for /b/ and /g/, which do not overlap. Hence, while a single acoustic cue may be insufficient to distinguish all stop contrasts, it is possible to extract all stop contrasts by proceeding in a step-wise fashion, using information provided by the noise burst and locus equations. Crucially, neural maps corresponding to the stop categories can emerge despite speaker

variability, and without taking into consideration any contextual information.[35] Therefore, infants have the tools (appropriate neurons) and sufficient input (predictable acoustic patterns) to develop appropriate neural maps for the identification of all stop contrasts in English. With more experience and extensive exposure, infants would presumably be able to refine these maps, for instance by creating maps for /b/ and /g/ based on noise burst information, taking into consideration vowel information (provided that these categories are first extracted using locus equations). This end result may be achieved based on principles derived from cell assembly theory, as proposed by Hebb (1949), often summarized as "cells that fire together, wire together". This proposal was reformulated by Allport (1985) as:

> If the inputs to such a system cause the same pattern of activity to occur repeatedly, the set of active elements constituting that pattern will become increasingly strongly inter-associated. That is, each element will tend to turn on every other element in the inter-associated pattern and (with negative weights) to turn off the elements that do not form part of the pattern. To put it another way, the pattern as a whole will become *'auto-associated'*—it will come to cause itself as its own successor. (p. 44)

As a result, a set of NB (noise burst) neurons that fires repeatedly at the same time as a group of FM (frequency-modulated) neurons that fires to the /b/ category can become associated with this /b/ category, and may ultimately activate the /b/ category irrespective of which FM neurons are activated (e.g. in experimental settings, when these cues are manipulated to correspond to different categories). The point here is that

---

[35] Vowel quality itself does not need to be taken into consideration, since each formant transition can be processed separately. That is, it is not the vowel quality that is taken into account here, but only the slope and direction of the F2 transition.

multiple cues can contribute to the identification of the same feature, even if they are processed by different groups of neurons. In the case of stop place of articulation, the presence of more than one acoustic component appears to be essential for the creation of adult-like speech categories in the course of neural development. In the mature adult brain, each cue should continue to be processed independently (i.e. by different groups of neurons), with the difference that adult speech categories will become highly dependent on contextual information through learned, auto-associated patterns. Consequently, in the event that each cue points to a different feature value (e.g. in fast speech, foreign-accented speech, or an experimental setting), cues may end up *competing* with one another. That is, one of the cues will be given prominence (or alternatively, will simply be ignored or inferred) in the generation of a categorical decision. Importantly, the noise burst and transition information are not processed additively because they contribute to identify the *same* feature, and they are not processed connectively either because they are processed by *different* neurons (see definitions provided in (1) for each type of processing).

Stop voicing contrast

Another characteristic that distinguishes stop consonants that has been the focus of intensive research is the stop voicing contrast. In utterance-initial or pretonic position, this contrast is traditionally referred to as Voice-Onset-Time (VOT), referring to the time elapsed between the burst onset and the start of periodicity. The VOT value is positive, indicating a voiceless stop consonant, if periodicity starts after the burst release; conversely, the VOT value is negative, indicating a voiced consonant, if periodicity

precedes the burst onset. Depending on the position of the stop in the utterance, various other cues may contribute to the identification of the stop voicing contrast, particularly in English. These cues include F1 transition, intensity of the burst, duration of the preceding vowel, duration of the stop closure (in utterance-final position), and presence or absence of periodicity during the stop closure (e.g. Benkí 2001, see Kingston & Diehl 1994 for a review). Here, I am concerned only with the role of periodicity and vowel duration in the identification of the stop voicing contrast.

Previously, it was shown that voiced and voiceless fricatives can be distinguished, respectively, by the presence or absence of a periodic signal, and that only the periodic signal is mapped by the neurology. That is, neurons are insensitive to the absence of periodicity. Similarly, voiced stop consonants in onset position can be distinguished from voiceless ones by the presence of periodicity in the signal prior to the burst release (so-called negative VOT). This is the case in French, where voiced stops [b, d, g] exhibit periodicity during the stop closure, while voiceless stops [p, t, k] do not (Abdelli-Beruh, 2004; Caramazza, & Yeni-Komshian 1974; Caramazza, Yeni-Komshian, Zurif, & Carbone 1973; Laeufer 1996; Sundara 2005). By contrast, in North American English, the stops transcribed in the orthography as 'b, d, g' do not systematically exhibit periodicity prior to the burst release when produced in initial pretonic position, although the periodicity does sometimes occur, as documented in a study with Canadian English speakers (Sundara 2005). English stop consonants in this position contrast primarily in terms of short-lag VOT (unvoiced and unaspirated) versus long-lag VOT (aspirated) (Keating 1984; Lisker & Abramson, 1964, 1967; Sundara 2005; Zlatin, 1974). As a result, in terms of neural mapping of periodicity, French has a voicing contrast based on

the presence or absence of periodicity during the stop closure in initial position, while English does not.[36] However, given that English speakers are sensitive to the presence of periodicity for the voicing contrast in fricatives, they should be able to perceive stop voicing contrasts given the proper testing conditions, as demonstrated by Curtin, Goad and Pater (1998), described in section 2.2. Incidentally, speakers of both French and English should also be sensitive to the presence or absence of periodicity in utterance-final position for stop voicing contrasts, since the same neural map that captures periodicity would be activated irrespective of syllabic context.

On the other hand, English speakers are known to use vowel duration to discriminate coda consonants; a preceding long vowel is associated with a voiced consonant, and a short vowel with its voiceless counterpart (e.g. Flege 1993). Interestingly, vowel lengthening is coupled with shortening of the stop closure in the production of a voiced stop, as illustrated in Figure 3–13. In this figure, the words 'bit' (top) and 'bid' (bottom), as pronounced by the same female speaker of Canadian English, are viewed through a 900 ms window in Praat (Boersma & Weenink 2007). The figure shows that while the vowel is longer in the word 'bid,' the duration of the closure is shorter (and here the final consonant is also voiced throughout). Put another way, vowel lengthening may be seen as a means to produce a periodic signal during the time that would otherwise be occupied by the stop closure.

---

[36] English may still be argued to have a stop voicing contrast based on other cues, such as F1 transition, which was shown, unlike VOT, to be a critical cue for the voicing contrast in syllable-initial position in noise (Jiang, Chen & Alwan 2006).

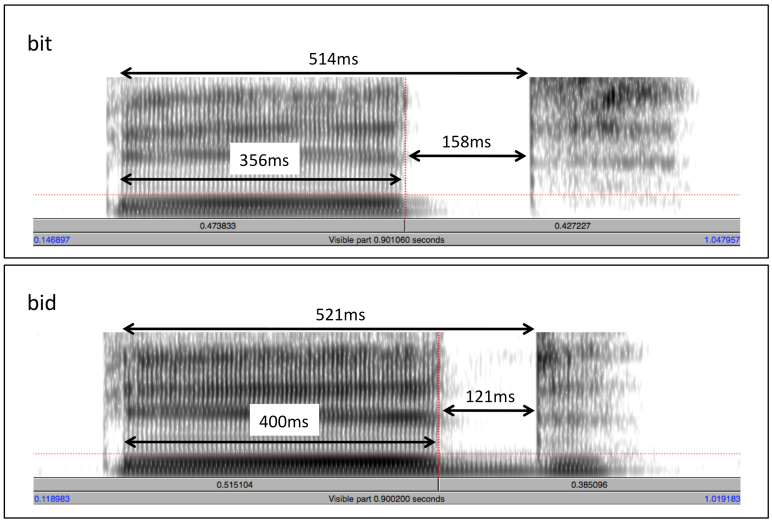

Figure 3–13 Spectrograms of the word 'bit' (top) and 'bid' (bottom) pronounced by a female Canadian English speaker, showing that when the vowel duration is lengthened, the stop closure duration is proportionally shortened.

However, a study with English children and adults suggests that the specific association between vowel duration and the voicing contrast is learned. The perceptual study conducted by Jones (2005) demonstrates that English-speaking adults rely more heavily on vowel duration than do children for the categorization of final stop voicing contrasts. Hence, while vowel lengthening may be a production strategy that contributes to the voicing contrast, the association between the vowel duration contrast and the coda voicing contrast must be acquired. The BLIP model proposes that English speakers map vowel duration to a short versus long vowel contrast in a manner comparable to that employed by speakers of languages that employ vowel length contrastively (discussed in 3.2.2). However, for English speakers, the neural maps are associated, at the phonological

level, with a consonant voicing contrast[37] (and also potentially with a stress-unstressed syllable contrast) instead of being associated with a vowel length contrast (this point is discussed further in 3.3.1, in the presentation of the phonological level of processing). In short, within the BLIP model, the presence of periodicity during stop closure and the length of the vowel preceding the stop may both contribute to determining the voicing status of a word-final stop. Accordingly, at least in English, these cues are processed competitively: periodicity is processed by neural maps sensitive to AM components, whereas vowel duration is processed by combination-sensitive neurons computing durational contrasts. Some kind of weighing schemes appear to be in place for resolving the categorical decision since people perceive sound almost always categorically even when the cues are ambiguous or contradictory (see for instance studies by Norris, McQueen, & Cutler 2003 and McQueen, Norris, & Cutler 2006). The exact factors (e.g. the listener's preference for a given cue) that play a role in these weighting schemes besides the basic neural mechanisms (e.g. level of discharge) are still unknown, and need to be further investigated.

3.2.4 Mapping of suprasegmentals

Acoustic cues can be processed by the neurology in three different ways: additively, connectively, or competitively, as introduced in the previous sections and discussed

---

[37] Note that even though the use of vowel duration to make a voicing contrast is context-bound (i.e. typically used in the "pre-stop" environment), the voicing contrast using vowel duration as the main cue is phonemic since it can be used to contrast words (whereas context-bound allophones do not usually distinguish words). For instance, although both /t/ and /d/ are realized as a voiced flap in the words *writer* and *rider*, the former typically exhibits a shorter preceding vowel than the latter, which is thought to impact on the perception of these otherwise homophonous words as different (see Kenstowicz 1994).

further in 3.3.1. The way acoustic cues are processed is more than a mere theoretical issue; it may have significant implications for the perception of non-native contrasts. In this section, I argue that English speakers' difficulties in perceiving and acquiring lexical tones stem from the different way in which F0 is used and processed in tone versus non-tone languages. While English speakers need to process only one F0 value at a time, speakers of tone languages, such as Mandarin Chinese, need to process *changes* in F0 (that is, at least two F0 values). As a result, English stress can be processed by neurons sensitive to F0 contrast along one dimension (competitively with vowel duration), whereas values of F0 for Mandarin Chinese tone identification must be processed connectively by combination-sensitive neurons able to encode changes in F0 during vowel production.[38]

Processing of lexical tone contrasts involves the processing of changes in pitch (F0) *within* a syllable. For example, to discriminate the Mandarin high-falling tone (tone 4) from the high tone (tone 1), one must be able to process changes in the slope and direction of F0 during the production of the syllable nucleus. This task can be performed by neurons that are sensitive to amplitude-modulated (AM) components, which encode pitch-relevant information, including F0, as documented for the marmoset monkey (Bendor & Wang 2005). More specifically, these neurons must be sensitive to general changes in the AM component (i.e. perceived change in the slope and direction of F0). For modeling purposes, the processing of Mandarin tones can be captured by four neural

---

[38] Alternatively, one could argue that it is the slope of the F0 rather than two F0 values that is processed in a way similar to the processing of FM components. However, this distinction is not crucial to the current discussion and would, in any case, yield similar results and conclusions.

maps across a two-dimensional plane that connectively process F0 onset and offset, as illustrated in Figure 3–14, where onset generally corresponds to the onset of the change in F0, referred here as *slope* for convenience. This implies that even if the F0 in a falling tone does not begin to decline before the middle of the nucleus, the sound should still be perceived as a falling tone, since neurons sensitive to the perception of the "falling" component will fire (more strongly) whenever this change occurs within the production of the vowel/syllable.

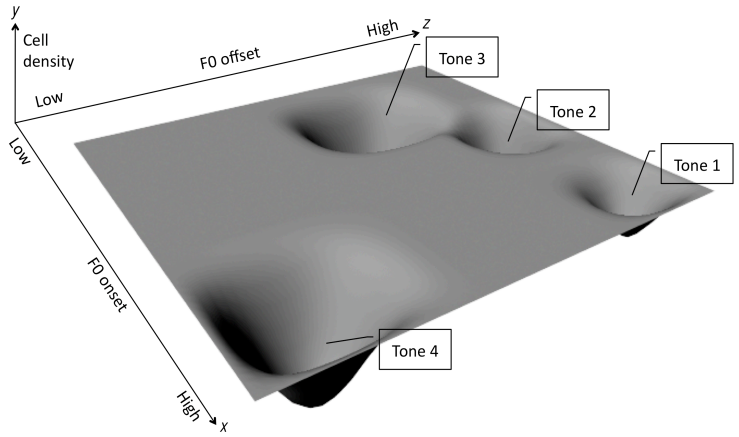

Figure 3–14 Schematic representation of the neural mapping of the four Mandarin tones. The x-axis represents F0 at onset whereas the z-axis represents F0 at offset. The y-axis represents cell density. Tone 1 is high-high (55), Tone 2 is mid-high (35), Tone 3 is low(falling)-high (214) and Tone 4 is high-falling (51). The relative size and potential overlapping of the neural maps are arbitrary in this figure.

Therefore, to identify a tone contour, the perceiver must be able to process F0 variations along two (or three, see footnote 39) dimensions: slope onset and offset. Importantly, these two dimensions must be processed *connectively* by combination-sensitive neurons (AM-AM neural maps). That is, the onset and offset of the tone contour

must be processed in relation to each other, the value of one or the other being in most cases insufficient for tone identification. This account yields the potential of 25 tone contrasts based on a 5-level bi-dimensional frequency system (i.e. without including the possibility of two consecutive slopes such as a rising-falling tone[39]), easily capturing the variety of tone contours found cross-linguistically. Other cues, such as voice quality, duration, and intensity, may be relevant for tone identification in various tone languages (e.g. Brunelle 2009; Kuo, Rosen & Faulkner 2008). However, since those cues also contribute to the perception of pitch contour, they are processed *competitively* with F0 information (although, as mentioned previously, intensity differences may be captured by neurons responsive to other components).

Given that lexical tones involve the processing of amplitude-modulated components (F0), and that native English speakers are generally sensitive to variations in F0 for stress identification, one might wonder why English speakers encounter difficulties in perceiving and learning tone contrasts as documented in various studies (e.g. Kiriloff 1969; Wang, Spence, Jongman & Sereno 1999; Xu, Gandour & Francis 2006). While the processing of tones requires the processing of acoustic cues connectively (AM-AM map), the processing of F0 for the identification of English lexical stress does not (AM map). This difference in neural processing is argued to be at the root of their difficulty.

---

[39] The same process can be applied to tone contours involving two changes in slope direction instead of one. Neurons that process a falling-rising tone presumably respond most strongly when all components are present (which may arguably be the situation for Tone 3 in the Mandarin Chinese example presented above), as documented in the squirrel monkey, where neurons were found to respond most strongly to the combination of three components, but to react very poorly to the presence of only one or two components (Olsen 1994, cited in Suga 2006).

The perception of English stress involves the processing of three main acoustic cues: amplitude (or intensity),[40] F0 contours, and vowel duration (Fry 1955, 1958; Lehiste 1970; Lieberman 1960; Mol & Uhlenbeck 1965; Morton & Jassem 1965; Wang 2008). For instance, stressed syllables are usually higher in pitch (F0) and exhibit higher amplitude and vowel duration than unstressed syllables. However, if the relative pitch of a syllable is high, but its amplitude is low and the vowel duration is not distinctively short or long, the perceiver must choose how to weigh these cues in deciding where stress falls, since each cue may independently signal stress (see Wang 2008 for a review). In this sense, the acoustic cues associated with English stress are processed *competitively*. Accordingly, each cue should have an independent cortical representation (except intensity, for reasons explained previously). Since stress identification does not necessitate sensitivity to changes in F0 *within* the syllable[41] but requires sensitivity to change in F0 *across* syllables, the neural mapping of F0 can be captured by neural maps that process F0 differences along only one dimension, as illustrated in Figure 3–15.

---

[40] From a psychophysical point of view, intensity and amplitude are not the same concepts. This difference, however, is not critical to the current discussion. Therefore, both terms are used here to refer to the perception of loudness.

[41] This does not prevent the production of F0 contours nor sensitivity to F0 modulations over an utterance. This only means that for stress identification, unlike tone identification, the processing of only one F0 value within a vowel is sufficient (or alternatively, the summation or average of F0 values during vowel production).

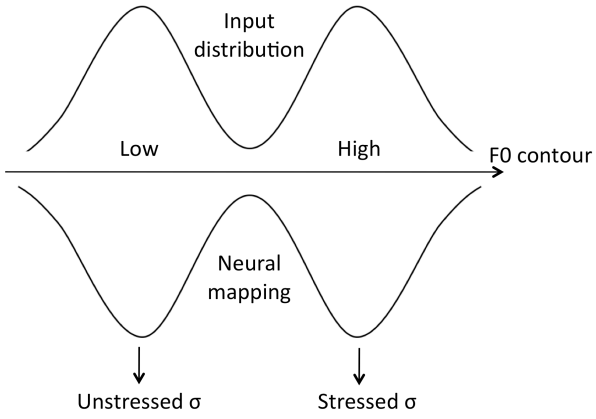

Figure 3–15 Neural mapping of F0 contours for stress identification in English.

Hence, English speakers' difficulties in perceiving Mandarin tones do not stem from an insensitivity to F0 changes, but from the way F0 is processed in the two languages: lexical stress may be processed by AM neural maps, whereas lexical tones must be processed connectively by combination-sensitive AM-AM neural maps. This implies that even though Tone 1 or 4 in Mandarin may arguably be equivalent to an English stressed syllable, the neurons processing the tones in Mandarin and stress in English are presumably different. However, in practice, English speakers should be able to perceive a high tone from a low tone (in tone languages featuring such a contrast) using the neurons used in English for stress identification, though this would not be efficient when trying to process more than two tone categories. The next section explains how the different types of processing impact phonological representations.

3.3 Phonological level of processing

In the previous chapter, it was argued that two separate levels of speech processing are necessary to account for the divergent results obtained depending on the task used to

assess perception of a given speech contrast. In the BLIP model, these two levels are posited as the *neural mapping level* and the *phonological level*. Given that within the current approach, the neural mapping level already captures speech categories based on the input distribution, and that technically, these categories may suffice to build the lexicon based on patterns of activated maps, it is worth considering why one needs the phonological level at all. What does the phonology correspond to, and how is it instantiated by the neurology?

It is reasonable to suppose that the phonology has emerged to palliate the complexity of the speech-processing task and to render it not only more efficient, but also to provide more latitude to both speakers and listeners. As discussed in the previous section, more than one cue can usually serve to identify the same speech contrast. For instance, a voicing contrast can be conveyed in English by the presence of periodicity, by the F1 transition, by the duration of the preceding vowel, etc. We can imagine that if there was no system in place that could relate each contrastive map to only one feature |voiced|, either each map would be perceived as a separate speech category (e.g. one category associated with the presence of a periodic signal, a different category associated with the F1 transition, etc.) or alternatively, all cues would have to be present and properly perceived to allow identification of the contrast. Also, without the phonology, lenition or coarticulation effects that result in an acoustically different sound would not be possible. For instance, native North American English listeners are able to relate the initial word in 'put this on' [pʊt̚.ðɪ.sɔn] to the same word containing the flap allophone in intervocalic context, as in 'put it on' [pʊ.ɾɪ.ɾɔn] virtually ignoring the acoustic differences. Therefore, perceivers must have found a way to associate different neural maps with

relevant speech categories (and this is what the phonology does). The phonological level posited by the BLIP model corresponds to the association of these neural maps with a common, meaningful contrast, referred to as a *feature*. A feature is defined here as a contrastive attribute commonly recognized by listeners and speakers of the same language as meaningful for the purposes of lexical identification.

Within this view, the initial *speech categories* that are later associated with distinctive features by the phonology *are* the neural maps, which are forged based on the unsegmented input (whether in words, syllables, or phonemes)—that is, prior to lexical learning. Once the neural maps are in place and able to perceive contrastive speech categories, these maps can be grouped in response to lexical development and articulatory awareness (through motor development as well as through the use of visual cues to identify common place and manner of articulation.) Lexical development promotes the organization of relevant groupings of neural maps based on phonotactic information while motor development provides relevant cues about similar places or manners of articulation, helping, for instance, to relate the stop in the sequence 'put this on' to the flap allophone in 'put it on,' since both sounds are produced in the same alveolar region (at least for English speakers who do not glottalize the stop in *put this on*).[42]

---

[42] To some extent, literacy may also serve to further refine the association between neural maps and features (a similar proposal is found in Werker & Curtin 2005). In the example of the stop-flap alternation, both sounds are transcribed with the same letter 't'. However, given that literacy is acquired after many years of language exposure, we can assume that the influence of orthography may be restricted to reinforcing such associations rather than having a crucial impact in creating the phonology. For instance, although the letters 'th' in English represent both the voiced and voiceless interdental fricatives, the speaker cannot interchange these sounds without potentially causing confusion for the native English listener. This tendency suggests that despite the same orthographical representation, the two sounds are perceived as distinctive at the phonological level.

Finally, the association between neural maps and phonological features may potentially be instantiated by the neurology through higher-order neurons, that is, neurons that compile and compare processing results obtained through the neural maps, in line with Nelken's (2008) conclusion:

> Recent progress in the processing of complex sounds suggests that the 'double personality' of the auditory system is reflected in a physiological hierarchy in which early stages encode sounds with exquisite sensitivity to their physical structure while later stages show selectivity to abstract, but behaviorally relevant, features. (p. 416)

It may not be possible to determine the exact functioning of the phonology in the auditory cortex until non-invasive technology is available to record the activity of single cells in the human brain. Nevertheless, building on the neural mapping level described in the previous section, it is theoretically possible to speculate about how neural maps are associated with features by relying on results of behavioral speech experiments. Since knowledge about the neural processing of abstract speech categories is still scarce, this section is succinct and is meant simply to explore how, within the current approach, neural maps might be associated with abstract feature representations.

The remainder of this section exemplifies the different types of association between neural maps and features. Below, I discuss the phonological processing of allophonic variations (3.3.1) and provide an explanation of how listeners may adjust their neural mapping to deal with speaker and dialectal variations (3.3.2). Finally, I address how language is processed in connected speech in natural contexts (3.3.3).

3.3.1 From neural maps to phonological features

The current framework posits that neural maps are associated with *features*. The role of features in speech processing is well established and supported by empirical evidence from perception (e.g. Benkí 1998; see Maye 2000 for a review) and production (e.g. Grenon, Benner & Esling 2007) studies. For instance, Bai, a Tibeto-Burman language spoken in China, does not contain any laryngeal contrasts at the segmental level, but uses laryngeal constriction contrastively in its tonal register system. However, unlike the production of English or Arabic infants towards the end of their first year, Bai infants still produce mostly laryngeal consonantal sounds at the same age period, rather than using laryngeal constriction at a syllabic/word level like Bai adults (Grenon, Benner & Esling 2007), even though they were able to use laryngeal constriction at the syllabic level at an earlier age (Benner, Grenon, & Esling 2007). Thus, Bai infants appear to notice that this *feature* is productive in the language, but have not yet learned to associate this feature to the proper higher level of representation (syllabic rather than phonemic). Maye, Weiss and Aslin (2008) also conclude that infants appear to index featural information, based on the fact that 8-month-old infants in their experiment could discriminate a novel untrained contrast after exposure with a contrast sharing the same acoustic information.

On the other hand, the role and abstract representation of phonemes have been disputed[43] (e.g. Greenberg 2006; Jusczyk 1993). The BLIP model remains agnostic on this latter issue, since within the current approach, features may be associated with

---

[43] For instance, Greenberg (2006) firmly argues, from the point of view of his neural-based model of speech processing, that "the phoneme is not a privileged unit; in fact, it does not even exist, except as a means of translating the output of other levels into a conventional linguistic form" (p. 412).

phonemes, mora, syllables, or entire lexical representations with equal ease. For illustrative purposes, however, features are shown to combine with phonemes, keeping in mind that units like phonemes may be the result of auto-associated patterns of neural maps activated at the same time, as proposed by cell assembly theory (e.g. Hebb 1949; Allport 1985; see quote by Allport in section 3.2.3 for a definition of auto-associated patterns). Thus, phonemes may not have a distinct representation (i.e. they may not be encoded specifically by a separate group of neurons tuned, for instance, to identify /t/, although phonemic representations may still be encoded by *patterns* of neural maps activated virtually simultaneously), as suggested by Greenberg (2006), among others.

Figure 3–2 illustrating the association between a neural map and a feature, as defined in this work, is reproduced below as Figure 3–16 for convenience. Conceptually, neural maps are the electrical circuits in the brain designed to capture the statistical distribution of an acoustic cue in the input, while features are labels indexing each map (i.e. they associate each map with a distinct meaningful contrast). Phonological development is the process of tagging each of these maps and cataloguing together those maps that serve a common purpose or feature contrast.

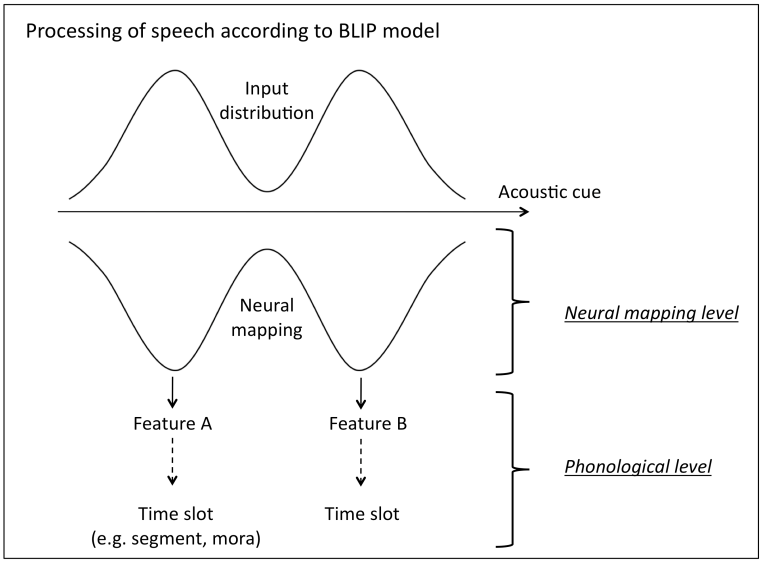

Figure 3–16 Processing of speech according to the BLIP model. The input distribution is processed by neural maps, which are in turn associated with contrastive features.

The current approach posits that speakers should at least be sensitive to *feature* contrasts (as defined in the current work) irrespective of their first language, without precluding the possibility of a different organization at higher levels—whether phonemic, moraic, syllabic, etc. Whether a universal organization exists between features and lexical representations remains to be investigated, and is of no concern for the BLIP model at this time.

Importantly, the notion of feature differs within the current approach from traditional views posited, for instance, by Feature Geometry, since the approach here is neither articulatory- nor auditory-based, but rather, neural-based. That is, this approach takes into consideration restrictions imposed by the neurology. For instance, since the spectral components of vowels (e.g. F1 and F2) are processed in conjunction by the same

group of neurons, these neurons must be associated with a single feature (i.e. vowel quality or spectral identity), rather than being decomposable into features such as |high|, |back|, and |low|, as proposed in articulator-based models (e.g. Sagey 1986), or |labial|, |coronal|, and |dorsal| as suggested in constriction-based models (e.g. Clements and Hume 1995). The end result is that the *quality* of the vowel—as opposed to characteristics of constriction or articulation—*is* the abstract phonological feature encoded by neurons at the phonological level, yielding features such as |i| or |ɪ| (these features are not to be confused with allophones, which only occur in the speech input). The logic is that the neural map of a vowel (and of any other speech contrasts that involve the processing of two or more acoustic cues connectively) is built along two (or more) acoustic dimensions. In order to associate the map of a vowel with a feature, the entire map (i.e. vowel quality) must be associated to this feature. This is not to say that characteristics of vowels such as their relative articulatory height and backness as captured by variations in F1 and F2 are totally ignored or do not have any impact. The issue between the notion of feature adopted here and the traditional phonological view of feature may be compared to the concept of "cold". From a physic perspective, cold is simply the absence of heat and therefore, cold does not in itself have a physical reality (i.e. it does not exist). Nevertheless, people can certainly *feel* cold, and the effect of cold temperatures has tangible effects and measurable impacts on both people and objects. Thus, even though "cold" itself does not have a physical reality, the *concept* of coldness can be felt and can be measured. I argue that the same analogy can be applied to the concept of some features previously posited such as those mentioned above and their existence in the neurology. Provided that the hypotheses of the BLIP model laid out in this model hold, a

given neural map cannot be divided to be associated with different features, and as a result, the only features that have neural correlates in the case of vowels correspond to vowel qualities such as |i|, |e|, |u|, etc. However, the *concepts* of vowel height and backness, for instance, can still be perceived *within* the tonotopic organization of neurons at the neural mapping level, and certainly have tangible and measurable impacts on phonological processes, and so forth.

The same principles would also apply to the mapping of tones. Since the neural mapping of tones like in Mandarin involves at least two dimensions across the AM-AM space, it is the entire tone (i.e. the entire map) that is associated with a feature at the phonological level. For instance, that means that the high and low parts of a high falling tone are not associated with two different features (such as |high| and |low|) but with a feature representing the entire tone contour. Nevertheless, this does not prevent speakers of Mandarin to perceive the separate components that make up the tone contour, since this information is contained within the neural map itself (it is simply not encoded separately at the phonological level). The notion of feature in this work departs significantly from previous phonological views, and further investigation and justification of the view adopted here is needed. However, I leave this for future development of the BLIP model since the main goal of the current work is the description of the neural mapping level and phonological level insofar as they can account for the processing of linguistic units such as allophones and abstract distinctive feature contrasts.

In the previous sections, three different types of processing, summarized previously in (1) but repeated as (2) below for convenience, were discussed and their implications at the neural mapping level demonstrated with concrete examples. In this

section I demonstrate how the neural maps are associated with phonological features depending on these three types of processing, using the same fricative, vowel, stop and suprasegmental examples used in the previous section.

(2) Speech-relevant cues can be processed:

A. Additively:	Two or more cues are processed *separately* by different groups of neurons and associated with *different* features;
B. Connectively:	Two or more acoustic cues are processed *in relation to each other* by the same group of neurons and associated with only *one* feature;
C. Competitively:	Two or more cues are processed *separately* by different groups of neurons and their relative relevance weighted in an attempt to associate them with only *one* feature value.

Figure 3–17 illustrates the different types of processing defined in (2): additively (left), connectively (center), and competitively (right). It is important to point out that these types of processing are meant to capture how two acoustic cues present in the acoustic realization of a given linguistic unit (i.e. segment, mora, or syllable) are processed in relation to each other by the neurology. While some cues contribute to identify the same abstract characteristic (i.e. feature) of a linguistic unit, some cues contribute to identify different features of that unit. The end result is that a given cue could theoretically be processed connectively with another cue, competitively with a third one, and additively with a fourth one. In other words, the types of processing posited here are not mutually exclusive. Figure 3–17 demonstrates that the additive processing of two acoustic cues (left) is performed by two independent neural maps and associated with two different features. By contrast, connective processing of two acoustic cues (center) is

performed by a single group of neurons (i.e. one neural map), and associated with only one feature. Finally, competitive processing of two acoustic cues (right) is achieved by two independent neural maps, like the additive processing of acoustic cues, except that in this case, the neural maps contribute to the identification of the same feature contrast rather than to two unrelated ones. When multiple cues point to different values of the same feature (e.g. one suggests a bilabial place of articulation, while the other suggests a velar articulation), these cues may be weighted to determine which feature value is the correct one. In this sense, the acoustic cues are said to *compete* with each other, as they may not always point to the same feature value. The processing of context-bound allophones essentially follows the processing of acoustic cues competitively, since two different neural maps are required to process the allophones, but the neural maps are associated with the same feature. This scenario is exemplified below with the case of vowels in Canadian French.

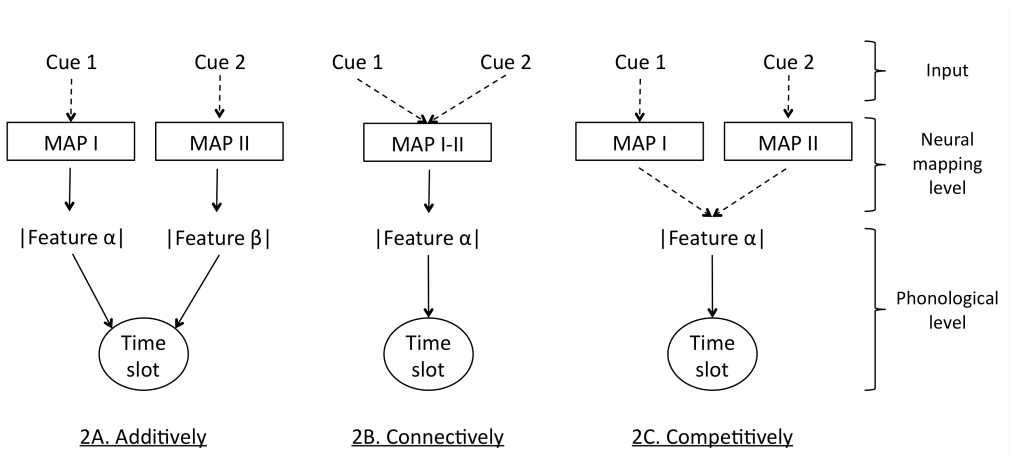

Figure 3–17 Association between neural maps and phonological features depending on type of processing: additively (left), connectively (center), and competitively (right).

Fricatives

As discussed previously, fricatives can be distinguished by at least two acoustic contrasts, one based on spectral peak frequency resulting from the configuration constriction in the oral cavity, and the other based on the presence of periodicity in the signal, as shown in Figure 3–18. These cues are processed by different types of neurons: spectral peak frequency by CF neurons, and periodicity by AM neurons. These maps are then labeled by the phonology as corresponding to specific features. For instance, the frequency map (CF neurons) that processes the fricative [v] is associated with the feature |labio-dental|, while the periodicity map (AM neurons) is associated with the feature |voiced|. The two features are linked to a common timing slot corresponding to the perception of the phoneme /v/. Since each cue is processed by different neural maps associated with distinct features, the acoustic cues are processed additively, corresponding to scenario 2A above.

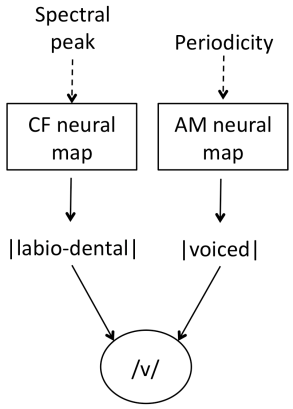

Figure 3–18 Additive processing of acoustic cues in identification of the voiced labio-dental fricative /v/ in English.

Vowels

Vowels can also be contrasted by various cues. However, F1 and F2 are most likely processed connectively and fit under scenario 2B of Figure 3–17. Accordingly, neither F1 nor F2 are directly associated with a feature; only the combined result of F1 and F2 as processed by combination-sensitive neurons (CF-CF maps) is associated with a feature, which corresponds to a vowel quality (or spectral identity). This scenario is shown in Figure 3–19 for the processing of the vowels [i] and [ɪ] in English (left).

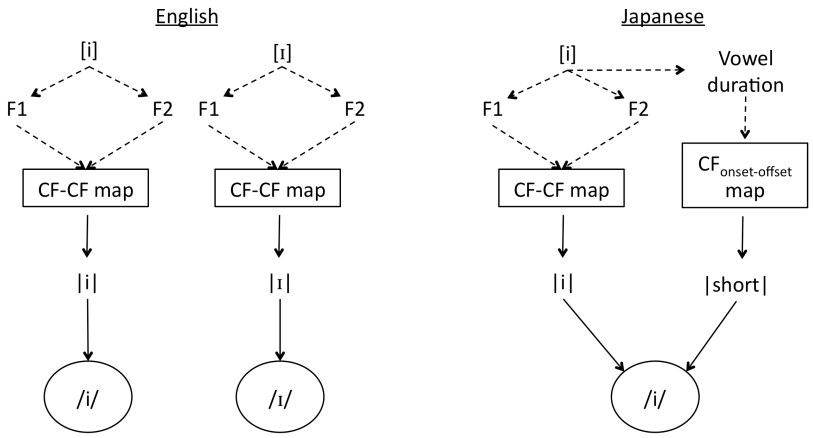

Figure 3–19 Processing of high front vowels in English and Japanese.

In Japanese, vowel duration—in addition to spectral contrast— is computed for vowel categorization, since short and long vowels are contrastive in the language. This cue is processed by a different neural map than the one used to identify vowel quality, and this map is associated with a different feature, in this case, with the feature |short|,[44]

[44] The names of features related to timing components (mainly AM and duration components) are tentative in this work, and are simply meant to reflect behaviorally relevant contrasts.

as shown in Figure 3–19 (right). Vowel quality is processed in addition to other cues relevant for vowel categorization such as vowel duration and voice quality (in languages where these cues are contrastive), each of which is associated with a different feature contrast.

Now we turn to the processing of allophonic variants. While allophones do not have a distinct phonological representation, they can be processed by distinct neural maps, provided that their distribution in the input is sufficiently contrastive, that is, provided that the allophones are strongly context-bound. The high front unrounded vowels in Canadian French illustrate this possibility. As discussed previously, the tense vowel [i] occurs in open syllables in this French dialect, while the lax vowel [ɪ] occurs in closed syllables. Within the BLIP model, the contrastive distribution of those vowels is proposed to give rise to the development of two neural maps, as illustrated in Figure 3–20 (right). However, at the phonological level, the two maps are associated with the same vowel quality feature, that is, |i|. Compared to other French dialects, Canadian French differs in the processing of the vowels at the neural mapping level, as shown in the same figure: while Parisian French only uses one neural map to process high front unrounded vowels (left), Canadian French uses two (right). Hence, the processing of context-bound allophones is comparable to the competitive processing of acoustic cues (scenario 2C above), except that in the case of allophones, the cues never compete with each other, since only one of the maps can be activated at a time in response to the input. Finally, only the neural map for the lax vowel is context-bound, capturing the fact that while the tense vowel can occur in both contexts (as it occurs in other French dialects, and also in the loanword *cheap* produced in Québécois French with a tense vowel to contrast the

word with the loanword *chip* pronounced with the original lax vowel), the lax vowel cannot, since the feature to which the maps are associated is that for the vowel |i|.

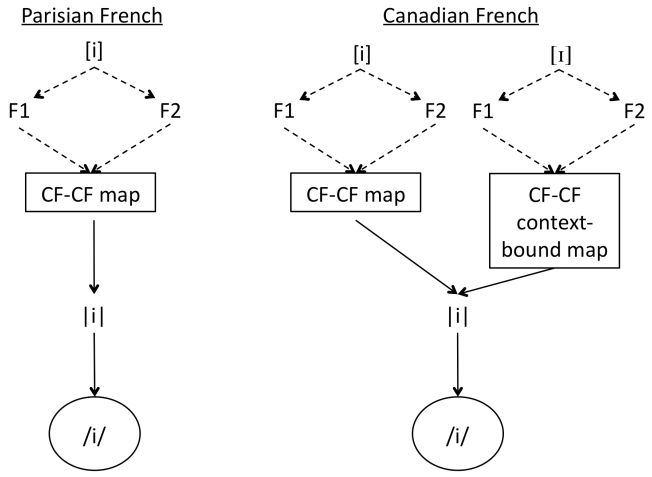

Figure 3–20 Processing of high front vowels and their allophonic variants by speakers of different French dialects: Parisian French versus Canadian French.

Stops

Four acoustic cues for the identification of stop consonants in English were discussed in the previous section: noise burst, F2 transitions, vowel duration, and periodicity. The center frequency of the noise burst as processed by NB neurons competes with F2 transitions processed by FM neurons for the identification of stop place of articulation (i.e. competitive processing), illustrated in Figure 3–21 as corresponding to the feature |alveolar|. In addition to these cues, vowel duration is processed by combination-sensitive neurons that compute differences between vowel onset and offset, possibly competing with the detection of periodicity as processed by AM neurons in the identification of the voicing status of the consonant, particularly in coda position (again,

these cues are processed competitively). Importantly, noise burst and F2 transitions are processed here additively in relation to vowel duration and periodicity, since the former and latter sets of acoustic cues contribute to the identification of two different features. Then, both features may be associated with a common timing slot, illustrated as corresponding to the phoneme /d/ in this figure. Interestingly, while vowel duration is associated with a vowel length feature in Japanese, as demonstrated in Figure 3–19, the same acoustic cue can be associated with a different feature in the case presented here: a coda voicing contrast. The same cue and neural map may also contribute to the identification of more than one feature. This is the case of vowel duration in English, which may contribute to the identification of the coda voicing contrast and the lexical stress contrast, as explained below.

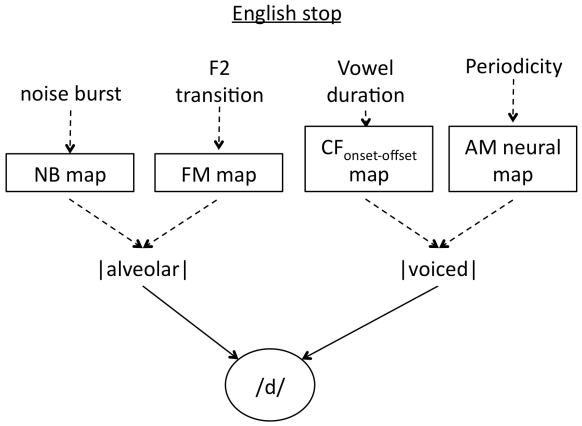

Figure 3–21 Processing of four different acoustic cues for identification of a stop consonant in English (see footnote 37 above).

Suprasegmentals

As discussed in the previous section, the processing of F0 differs significantly in English versus Mandarin Chinese, possibly accounting for English speakers' difficulties in acquiring lexical tone contours. In English, F0 is processed along only one dimension (i.e. relative perceived height) for lexical stress identification, whereas in Mandarin Chinese, F0 is processed along two dimensions (i.e. F0 onset and F0 offset) for lexical tone identification. This contrast is illustrated in Figure 3–22.

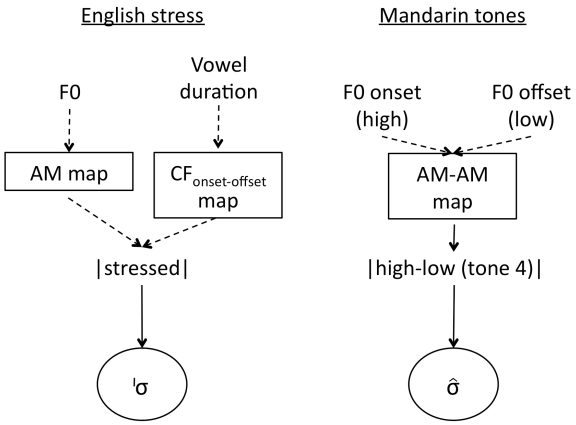

Figure 3–22 Processing of lexical stress in English versus processing of lexical tones in Mandarin Chinese.

In addition to these differences in processing, other cues may be used for stress and tone identification. In the case of English in particular, vowel duration is also known to contribute to stress identification. Hence, this cue is processed competitively with F0, since they both serve to identify the same feature contrast.

In summary, although speech sounds can be contrasted by a limited number of spectral and timing components, languages differ in the use and processing of those components in various ways. As demonstrated previously with the processing of vowels

and fricatives, languages and dialects can differ in the number of maps or features they employ to process the same acoustic component. Languages can also differ in the ways they associate neural maps and features; in some cases, the same acoustic component is associated with different features in different languages. For example, vowel duration is associated with a vowel contrast in Japanese, but with a coda voicing contrast or lexical stress contrast in English. Lastly, languages can differ in how an acoustic cue is processed by the neurology, and consequently, in the features associated with the same acoustic component, as seen in the processing of F0 in English versus Mandarin Chinese. Consequently, the fact that an acoustic cue is used in a language does not guarantee that non-native contrasts based on the same cue will be perceived by L2 listeners. The above-noted considerations may all play a paramount role in speech perception and acquisition, particularly L2 speech perception. This issue is discussed in detail in chapter 4.

3.3.2 Processing speaker and dialect variability

Given that neural maps correspond to neurons tuned to a range of acoustic values rather than to a single fixed acoustic value, these maps may partly suffice to encompass most speaker and dialectal variations in the realization of sound contrasts. One might wonder, however, whether these maps are, consequently, rigid entities that process speech contrasts in an inflexible manner. Importantly, we must also ask whether the brain is simply a passive receiver that blindly builds neural maps based on the input distribution. Based on behavioral data, and assuming that the neural mapping hypothesis described in this work holds, it appears most likely that the neural maps are in fact malleable, flexible entities, as discussed below. I also propose that the maps are built to *optimize* the

perception of speech contrasts, rather than strictly and rigidly *reflecting* the input frequency distribution.[45] In this sense, the brain is not viewed as a purely mechanical and passive receiver: the perceiver exerts some control (whether overt or covert) over neural development, which may lead to important disparities in the use or weighting of different cues for perception of a given speech contrast by speakers of the same language.

Research by Norris, McQueen, & Cutler (2003) and McQueen, Norris, & Cutler (2006) has shown that listeners can retune their interpretation of a perceptual boundary of an acoustic contrast virtually automatically, pointing to the malleability of the so-called categorical boundaries. Participants in these experiments were presented with an ambiguous sound intermediate between [f] and [s] in either an [f]-biased context or an [s]-biased context. Participants exposed to the [f]-biased context adjusted their acoustic boundary by labeling more tokens as containing [f] in a subsequent task in which they were asked to categorize sounds ranging along a continuum from [s] to [f]. Conversely, participants exposed to an [s]-biased context labeled more tokens as containing the sound [s] in the same subsequent task.

This ability can be seen, within the current neural approach, as an adjustment of the virtual boundary between neural maps that allows listeners to rapidly cope with new voices and dialects. The BLIP model, therefore, posits that the relative boundary between categorical centers is highly flexible and adaptable, so long as one does not jump over an entire category. The terms *virtual* and *relative* boundary are used here, because from the

---

[45] It is important to emphasize that this view is antithetical to an exemplar-based approach, which assumes, conversely, that the perception of speech contrasts is directly correlated with input frequency.

standpoint of biological endowment, there is no boundary per se: connections between neurons exist within and across related neural maps. One might recall that according to the inverted magnification factor hypothesis, a categorical center corresponds to a decrease in cell density activation around some acoustic values, whereas the in-between region exhibits an increase in cell density activation associated with an enhanced ability to perceive acoustic details. Hence, the so-called perceived boundary is mostly arbitrary, free to fluctuate based on the perceived contrasts as produced by a given individual. Information about the exact acoustic values—or alternatively, about the relative categorical boundaries—used by different individuals is potentially what is indexed in episodic memory, serving to identify previously encountered voices and dialects to assist in the perception of similar voices and dialects. On the other hand, information about patterns of neural maps and their associated phonological features is potentially encoded in lexical memory (i.e. if this memory component does not also encode information related to episodic traces, then lexical memory may only encode pattern of abstracted contrasts as captured at the phonological level) and permits the identification of previously encountered words and morphemes.

Although the input undoubtedly influences the shaping of the neural maps, as described in previous sections, it is doubtful and possibly even undesirable that these maps should be assumed to represent a direct copy of the input. First, as discussed in chapter 2, the input distribution does not always represent an ideal and straightforward contrast (e.g. Goldstein et al. 2008). If the neural mapping were designed to simply mirror the input distribution, it would make it potentially difficult to perceive natural speech contrasts, especially vowel contrasts (since the F1 and F2 values may overlap

considerably across speakers or even within the speech of the same individual). Second, the perception of an acoustic value is not always correlated with its frequency in the input: it is possible for the most frequent acoustic realization of a speech contrast to be perceived with less accuracy than its less frequent counterpart (e.g. Tucker & Warner 2007; Warner & Tucker 2007). Consequently, the categorical center of the neural maps may not correspond to the most *frequent* acoustic cue in the input, but to the most *contrastive* values of this cue, and this is the position adopted by the BLIP model. As illustrated in Figure 3–23, even if the most frequent tokens in a hypothetical bimodal distribution along an acoustic dimension are very close to one another (top), the neural mapping of this cue may exhibit categorical centers distinctly apart from one another (bottom). Accordingly, the BLIP model posits that categorical centers do not correspond to the most frequent acoustic input, but to the most perceptually salient contrastive realization of the acoustic component. That is, neural maps are intentionally built with categorical centers as far apart as possible within the possible range of realizations of this contrast. Incidentally, this organization potentially allows the relative categorical boundary to be displaced freely between the categorical anchors (or centers), providing more latitude to process speaker and dialectal variations.

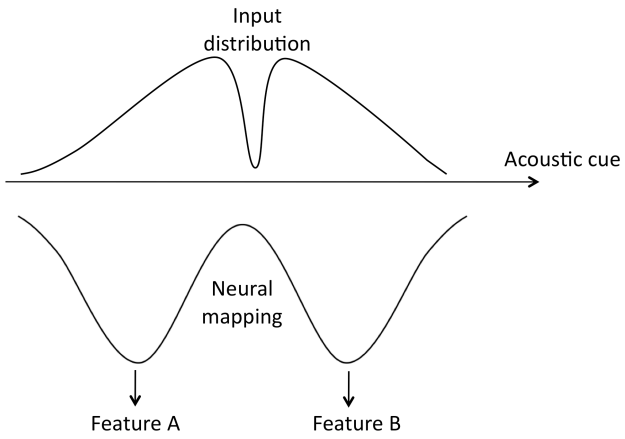

Figure 3–23 Hypothetical scenario demonstrating that the neural mapping of an acoustic cue is not based on the distribution frequency of this cue in the input, but on the most contrastive realization of this cue.

The fact that the perceiver is able to exert some control over neural development is supported by evidence (presented in chapter 2) showing that although adults can learn to perceive a novel contrast based solely on the input distribution (Maye & Gerken 2000, 2001), awareness of the contrast facilitates its acquisition (Hayes-Harb 2007). Research on placebo effects further confirms that expectations not only influence behaviour, but also impact biological processes, including synaptic activity (Scott, Stohler, Egnatuk, Wang, Koeppe, & Zubieta 2007). That is, synaptic connections (and therefore neural maps) may be altered without exposure to effect-induced external stimuli.[46] Based on such findings, we may predict that adult language learners' exposure to an L2 may never be enough to trigger neural reorganization if an acoustic cue is not *expected* to be

[46] This means that neurons can be activated and the strength of their connections altered without their being directly activated by acoustic stimuli.

contrastive, or if the learner is unable to identify the relevant contrastive cue. Therefore, the BLIP model does not conceive of the brain as a computing machine,[47] but rather as a living organism with its own volition that generally aims to model its surrounding (acoustic) environment in a way convenient for proper functioning and interaction with other individuals sharing the same language(s).

3.3.3 Processing misleading or incomplete information

One of the most impressive accomplishments of the human brain in regard to speech perception is its ability to "fill the gaps"—particularly in fast connected speech—when the acoustic information in the input is impoverished (e.g. reduced segments), inaccurate (e.g. a vowel that takes on the quality of another vowel), or simply missing (e.g. deleted vowels between voiceless segments). As an anecdotal example, in a talk at the University of Victoria, Dr. Natasha Warner (from the University of Arizona) presented an audience of linguists with a recorded sentence produced by a native female speaker of American English, which everyone perceived as *I bought a book yesterday*. However, when the word *bought* was extracted from the sentence and played in isolation, it was clear that the word produced corresponded acoustically to the word *but* rather than *bought*. Interestingly, even after knowing that fact, when the sentence was re-played, the word in

---

[47] Computer-based neural network models (e.g. Allport 1985), including exemplar-based models (e.g. Nosofsky & Zaki 2002), seem to view the brain mostly as a machine that computes mathematical equations and extrapolations with limited volition of its own. In the equation described by Nosofsky & Zaki (2002), for instance, individual variability in the processing of the acoustic input—deterministic probability—is quantified as a variable referred to as γ, which may vary from 0 to 1. Although this view is attractive and practical for many reasons (e.g. it may have useful applications for speech recognition technology), I argue that this view as applied to the functioning of the human brain is too limiting and fails to represent the conscious influence the perceiver may have over his/her own neural development, especially in L2 acquisition.

the sentence was still perceived as *bought*. It is possible that the perceiver's expectations of what the next word should be (whether the language has a strict word order or not) is at least partly responsible for such discrepancies in the perception of fluent speech versus isolated words.

It is important to point out that despite the descriptions provided about neural mapping in the last sections, the BLIP model does *not* assume that the acoustic input need always be processed in exhaustive detail by experienced listeners, since the processing of a few acoustic cues may be sufficient to relate the input to a set of maps that are usually activated at the same time. This view is consistent with general learning network models, such as the one posited by Allport (1985):

> The activation of only some elements of the learned pattern will tend to evoke each of the remaining elements of that pattern, since all of its missing elements receive positive connections from each of the elements already present, while currently active elements that are *not* part of the learned pattern are inhibited. (p. 44)

Assuming that the ability to ignore incomplete and misleading information is correlated with the presence of the kinds of strongly established patterns that result from intensive experience of a language, this ability may have crucial implications for non-native listeners, who may fail to properly perceive and recognize patterns that do not entirely fit their emerging patterns. In other words, while missing or misleading information may not generally affect L1 speech processing, because the processing of a few acoustic components may suffice to recover an entire pattern, the same gaps may impinge on L2 speech perception and processing, which may attempt to process the input in more detail.

3.4 Reconciling the speculated levels of speech processing

In the previous chapter, it was shown that the perception of speech contrasts might differ considerably depending on task type and testing conditions. Various levels of speech processing have been posited to account for divergent experimental results by previous linguistic and psycholinguistic models, as described in chapter 2, section 2.2. Researchers have given these levels different names (e.g. auditory, acoustic, phonetic, surface, general perceptual, phonemic, etc.) and it is not always clear how, if at all, they relate to one another. In this section, I argue that all the previously posited levels can be reconciled within the BLIP model, and I explain how different tasks and testing conditions tap into these levels, an issue that is crucial to the design of the L2 experiments reported in the next chapter.

Werker and Logan (1985) posited three distinct levels of speech processing to account for the different results obtained in their discrimination task with three different interstimulus intervals (ISIs): auditory, phonetic, and phonemic. In a typical discrimination task, two stimuli, separated by a given ISI, are presented to listeners, who are asked to decide if the two sounds they heard are the same or different. In Werker and Logan's experiment, the three ISIs used were 250ms, 500ms and 1500ms. Werker and Logan posited that the acoustic level would be activated with a very short ISI (250ms), allowing listeners to discriminate acoustic differences *within* a speech category. Consequently, a discrimination task combined with an ISI of 250ms is typically used to evaluate the perceptual magnet effect (e.g. Iverson & Kuhl 1995). From a neural perspective, the perceptual magnet effect, as discussed in chapter 2, corresponds to the

perception of differences between stimuli *within* a given neural map (e.g. Bauer, Der & Herrmann 1996; Guenther et al. 1999). From a psychophysical point of view, this mode of perception relates to the so-called *just-noticeable difference* (jnd) (a.k.a. difference limen or differential threshold), which is known to be sensitive to testing conditions (e.g. ISI, volume level). In addition, according to the (inverted) magnification factor hypothesis, the jnd should be smaller in cortical areas with a high level of activated cell density, and larger in areas with a low level of cell density activation, since according to this theory, the degree of perceived acoustic detail is proportionally related to the density of cell activation. As discussed in chapter 2, categorical training affects the discrimination of stimuli at categorical centers, creating some kind of perceptual magnet effect. The inverted magnification factor was shown to account for this effect (e.g. Guenther & Bohland 2002). Hence, acoustic perception as posited by Werker and Logan (1985) appears to correspond to the ability to perceive a jnd along a given acoustic dimension, whether within or outside a speech category, and is affected by testing conditions and by the shape of the neural map (i.e. by whether the map has been subjected to a positive or inverted magnification factor). Hence, the way speech sounds are perceived in a discrimination task with a very short ISI is not representative of a level of speech processing *per se*, though it can be affected by changes in cell density resulting from exposure to speech.

The phonetic level, as posited by Werker and Logan (1985), is assumed to be activated at intermediate ISI values (500ms) and to represent listeners' ability to discriminate the acoustic differences *between* speech categories that exist in some languages. In their experiment, English speakers were sensitive to the contrast between

the Hindi dental and retroflex stops, even though this contrast is not distinctive in English. From a neural perspective, it is not entirely clear what the phonetic level posited by Werker and Logan (1985) represents, since there is no reason for listeners not to perceive non-native contrasts if the testing conditions are conducive to the processing of small differences between non-native sounds. That is, listeners *are* able to perceive differences between tokens of the same speech category given the proper testing conditions, as long as the acoustic difference between the tokens is larger than the jnd (see, for instance, studies conducted by McMurray & Aslin 2005; McMurray, Tanenhaus, and Aslin 2002; Miller & Eimas 1996; Pisoni & Tash 1974). Therefore, there is no need to posit a separate level of processing to account for this ability.

The last level posited by Werker and Logan (1985) is the phonemic level, which corresponds to the perceiver's ability to discriminate native speech categories when the ISI is longer (at least 1500ms). Within the BLIP approach, this level is argued to correspond to discrimination of sounds *between* neural maps (as opposed to *within* neural maps when the ISI is short (250ms) or intermediate (500ms)). Importantly, the phonemic level posited by Werker and Logan is *not* posited to correspond to the *phonological* level in the BLIP model. This is because the discrimination task does not specifically require lexical access; therefore, discrimination of sound pairs can be achieved simply by using the neural mapping level of speech processing (because meaning is not specifically relevant for this task).

Instead of using different testing conditions (e.g. different ISIs), Curtin, Goad and Pater (1998) used different tasks to assess English and French speakers' perception of non-native (Thai) stop contrasts: an ABX discrimination task versus a picture

identification task. The authors argue that the ABX task activates the surface level of processing, while the picture identification task activates the lexical level. In the ABX task, listeners are presented with three words. The first two words are different from each other, and the third word (word X) is the one that needs to be identified as the same as word A or B. In their experiment, this task appeared to enable English speakers to perceive Thai contrasts that are allophonic in English (a.k.a. phonetic categories), that is, contrasts that correspond acoustically to the aspirated versus non-aspirated /t/ in English. According to the BLIP model, since this contrast is context-bound, English speakers would have a neural map for each of these allophones. That is, one neural map would process the aspirated sound and another map the non-aspirated sound; and these maps would be context-bound. In short, the neural mapping level in the BLIP model can be said to be the neural-based equivalent of the surface level posited by Curtin, Goad and Pater (1998).

On the other hand, the same native English speakers in Curtin, Goad and Pater (1998) were unable to distinguish the allophonic (aspirated vs. non-aspirated) contrast when presented with a picture identification task, presumably because this task taps into contrasts which are meaningful for lexical distinction. Hence, the authors argued that this task activates the lexical level of processing. Within a neural approach such as the BLIP model, the lexical level posited by Curtin, Goad and Pater (1998) corresponds to contrasts that are meaningful for lexical distinction; and therefore, to contrasts between neural maps that serve the same purpose (or feature). In this case, the context-bound neural maps that capture the aspirated versus non-aspirated contrast at the neural mapping level are presumably associated with the same feature at the phonological level. Hence,

the lexical level posited by Curtin, Goad and Pater (1998) is roughly equivalent to the phonological level in the BLIP model.

The levels posited by Werker and Logan (1985) were meant to account for specific experimental results when changing the task conditions, while the levels posited by Curtin, Goad and Pater (1998) were meant to capture different results obtained when changing tasks. The levels (or planes) posited by the PRIMIR model (general perceptual, phonemic, and word form) proposed by Werker and Curtin (2005) were designed to reconcile the array of divergent experimental results in L1 studies obtained when changing either the task type, task conditions, or age of infants used in the experiments. The general perceptual plane in PRIMIR, as described by Werker and Curtin (2005), refers to "all the information that is in the signal. This plane includes those properties that are specifically phonetic" (2005, p. 213). In other words, like the surface level proposed by Curtin, Goad and Pater (1998), this level encodes phonetic categories but also captures acoustic details pertaining to indexical information. In this sense, the perceptual plane of PRIMIR and the neural mapping level of the BLIP model are also comparable, since although the neural maps encode phonetic categories, they also capture detailed acoustic information within each neural map. What differentiates the neural mapping level of the BLIP model from the general perceptual plane of PRIMIR is the BLIP model's explicit grounding in neural processing, and the fact that, unlike PRIMIR, it is not restricted to L1. Thus, the BLIP model can account for discrepancies in L2 experiments as well (see chapter 4). Another important difference between BLIP and PRIMIR lies in the use of terminology: in the BLIP model, allophonic contrasts are captured by different *neural maps* at the neural mapping level, whereas in PRIMIR, allophonic contrasts are referred

to as *General Perceptual (phonetic) features* that are "the bases of phonetic and indexical categories […] described as clusters of exemplar-like distributions (Werker & Curtin 2005: 214)." In other words, in PRIMIR, the concept of features relates to allophonic (a.k.a. phonetic) contrasts, rather than to phonological ones, as in BLIP.

The other level of speech processing posited by Werker and Curtin (2005) relevant to the current discussion is the phonemic plane.[48] The phonemic plane is presumed to emerge after the acquisition of phonetic categories (or "phonetic features" in PRIMIR) at the general perceptual level, a process that may be thought of as equivalent to the formation of neural maps in the BLIP model. The progressive overlapping of these "phonetic features" through lexical development results in the emergence of abstract phonemes in PRIMIR. It has been shown that infants at 14 months are generally able to distinguish well-known minimal pairs, but are unable to distinguish novel sets of minimal pairs (Fennell & Werker 2004; Swingley & Aslin 2002). However, this discrepancy appears partly related to lexical development, where infants with a larger vocabulary size were found to better discriminate novel minimal pairs than infants of the same age with a smaller vocabulary size (Werker et al. 2002), hence supporting the idea that phonemic development emerges in concert with lexical development.[49] Similarly, the BLIP model also posits that the phonology emerges from phonetic categories (i.e. neural maps), hand-

---

[48] I will not discuss here the word form plane posited by Werker & Curtin (2005) in PRIMIR for comparison with the BLIP model, since at this point, the BLIP model is not designed to account for lexical processing or retrieval. The word form plane proposed by Werker & Curtin was introduced briefly in chapter 2, section 2.2.

[49] Note that the minimal pairs used in these experiments are also distinguished minimally by only one feature. Hence, it is unclear whether a phonemic approach to this issue as proposed by PRIMIR provides a better account than an approach in terms of phonological features, as proposed by the BLIP model.

in-hand with lexical development,[50] since the phonological features in the BLIP model are related to meaning contrasts. However, unlike PRIMIR, which argues for the existence of abstract phonemes, the BLIP model posits the existence of abstract *phonological features* (the BLIP model remains agnostic at this point about whether phonemes are represented as separate entities by the neurology as discussed in 3.3). PRIMIR posits only the existence of *phonetic* features (e.g. VOT); it is unclear whether phonemes are actually decomposable into abstract *phonological* features (e.g. |voice|) in their model. Additionally, while the BLIP model explicitly accounts for the processing of suprasegmental elements such as lexical tones, stress, and accent, PRIMIR does not. Hence, although the phonemic plane proposed by PRIMIR is similar in many respects to the phonological level posited by the BLIP model, BLIP permits linguistic analyses in terms of both abstract phonological features and phonemes, and provides an account for the processing of suprasegmental elements, whereas PRIMIR does not. In addition, the BLIP model presents implications not only for the study of L1 development (since it provides a framework to study the processing of acoustic cues in different languages and at different stages of development with its bi-level approach), but also for the study of L2 processing and acquisition, as discussed thoroughly in the next chapter.

To sum up, different tasks, testing conditions, and infant vocabulary size (among other factors that play a role in L1 development) may activate different levels of speech processing, and consequently, yield seemingly contradictory results. However, the two

---

[50] Besides, the BLIP acknowledges the possible contribution of motor development and articulatory awareness to phonological development.

levels of speech processing posited by the BLIP model, the neural mapping and phonological levels, can capture these results by reconciling the different levels posited by the linguistic and psycholinguistics models described above.

3.5 The BLIP model in a nutshell

The Bi-Level Input Processing (BLIP) model builds on the assumptions about the formation of invariant neural parameters summarized in the previous chapter, and aims to address the questions raised at the end of chapter 2. To recapitulate, the BLIP model proposes two levels of speech processing: the neural mapping level and the phonological level. The neural mapping level consists of groups of neurons sensitive to acoustic details in the speech input, which are organized into neural maps that reflect the contrastive use of spectral and timing components in a given language or dialect. These maps capture phonetic contrasts, including context-bound allophonic variations. The phonological level is tentatively presumed to consist of neurons sensitive to abstract meaningful contrasts. These abstract, behaviorally relevant contrasts are derived from the neural maps that are established based on the input distribution. These maps are then associated with a contrastive feature by the phonology, based on information provided through lexical development, motor development, and articulatory awareness. Features may combine to yield phonemic, moraic, or syllabic representations, although it is not yet clear whether these units are encoded separately by the neurology.

Acoustic cues relevant for speech contrasts may be processed in three different ways by the neurology. First, they can be processed additively, where each cue is processed by a separate group of neurons or neural map and associated with different

features. Second, some cues may be processed connectively, where more than one acoustic component must be evaluated in conjunction with another by the same group of neurons, the result of this computation being associated with only one feature value. Third, some cues may be processed competitively, where each cue is processed by a separate group of neurons but contributes to the identification of only one feature value. The acoustic cues of context-bound allophones are also processed competitively, since different neurons are necessary to process the allophonic variations at the neural mapping level, yet the context-bound neural maps are associated with the same abstract, behaviorally relevant feature.

In the BLIP model, neural maps are designed to optimize the perception of a contrast rather than to strictly reflect the input distribution. Accordingly, the BLIP model posits that the categorical centers of neural maps along a given acoustic dimension are as far apart as permitted by the actual realization of the speech contrast. This optimized contrastive design has the potential advantage of providing greater latitude to the perceiver to cope with individual and dialectal differences in the realization of the acoustic contrast: the perceiver can adjust the boundary between two maps to fit the speech of each encountered individual. That said, speech cues are not necessarily processed in exhaustive detail, since activation of a few neurons may stimulate a chain reaction, activating all the remaining components typically activated in a learned pattern. The perceiver's expectations may also trigger this chain reaction prior to the activation of any neurons associated with a given pattern. That is, in the BLIP model the brain is *not* viewed as a passive receiver reacting mechanically to the acoustic input. On the contrary, the perceiver may exert some control over neural development through focus on, and

awareness of, the relevant acoustic contrasts. This last notion is particularly relevant for L2 acquisition, as explained in the last chapter.

In addition to providing a tenable account of first language development (as briefly introduced in this chapter), and accounting for cross-linguistic differences in the perception of acoustic contrasts, the BLIP model has crucial implications for a better understanding of second language perception and acquisition. These implications are discussed in the following chapter, which also reports the results of four behavioral experiments that support predictions derived from these implications.

Chapter Four: Implications of the BLIP Model for L2 perception

In the previous chapter, I described the assumptions and proposals of the BLIP model in relation to speech perception. This model proposes that context-bound allophones (a.k.a. phonetic categories), such as [i] and [ɪ] in Canadian (Québécois) French, are captured at the *neural mapping level* by separate context-bound neural maps along contrastive acoustic dimension(s). The neural maps are subsequently associated with a distinctive feature at the *phonological level*, where a combination of a set of features corresponds to a phonemic, moraic, or syllabic representation. In the case of the Canadian French vowels, the neural maps corresponding to the [i] and [ɪ] allophones are argued to be associated with the same feature |i|, which in turn represents the phoneme /i/. In this chapter, I demonstrate how these two levels of speech processing can serve to make predictions about the perception of non-native contrasts by adult L2 learners.

The BLIP model builds on important assumptions posited by Guenther and colleagues' neural-based perception model (1999, 2002, 2004), mainly the inverted magnification factor hypothesis as applied to L2 learning, and the concept of the overlapping map as a possible cause of L2 learners' difficulties with L2 contrasts. The BLIP model provides additional contributions: it further considers the possible role of neural maps and phonology for L2 perception and acquisition by making specific predictions.

This chapter summarizes the approach adopted by previous L2 models such as PAM and SLM, along with their shortcomings, in identifying the exact source of L2 learners' difficulties with non-native speech contrasts (4.1). The approach and predictions

of the BLIP model for L2 perception and acquisition are presented in section 4.2. Sections 4.3, 4.4, 4.5, and 4.6 discuss four perceptual experiments with native English, Japanese, and Canadian French speakers that support these predictions. Section 4.7 summarizes the predictions of the model along with the supporting experiments. Finally, section 4.8 wraps up this chapter by discussing the additional contribution of the BLIP model to the L1, L2 and neural-based models of speech processing introduced in this work.

4.1 The notion of cross-linguistic perceptual similarity

Despite the marketing fanfare[51] and the not uncommon belief that an L2 is best learned by replicating the conditions present during L1 development, there are critical differences between infant and adult language learners. Crucially, adult L2 learners, unlike infants learning their L1, already speak at least one language, and the way the L1 is processed may interfere with the perception and acquisition of novel speech contrasts. The unresolved issue, however, concerns exactly *how* the learners' L1 interferes with L2 acquisition, and whether the difficulties resulting from this interference can be overcome.

Various models have been proposed to account for L1 interference on the perception or acquisition of L2 segmental contrasts. To cite a few, Best and colleagues (Best 1993, 1994, 1995; Best & McRoberts 2003; Best, McRoberts, & Goodell 2001; Best & Strange 1992) put forward the Perceptual Assimilation Model (PAM); Flege

[51] The marketing slogan of one of the biggest companies specializing in L2 software is: "The key to the Rosetta Stone method is that it unlocks your natural ability to learn a language – the same way you learned your first."

(1992a, 1992b, 1993, 1995) proposed the Speech Learning Model (SLM); Major (Major & Kim 1999) articulated the Similarity Differential Rate Hypothesis (SDRH); and, more recently, Escudero (Escudero 2005; Escudero & Boersma 2003) presented the Second Language Linguistic Perception model (L2LP), derived from the principles of Optimality Theory. I am concerned here only with PAM and SLM, since these two models have been the most influential in L2 studies over the past two decades, and because they are representative of the shortcomings shared by most L2 models proposed to date, as discussed below.

PAM was designed to make predictions about the initial *perception* of L2 segmental contrasts, while SLM was designed to make predictions about the possible *acquisition* of these contrasts by L2 learners, especially in production. Hence, although many L2 researchers have attempted to evaluate which model better accounts for their experimental results, it is important to keep in mind that these models were meant to serve different objectives. Therefore, their predictions are not necessarily mutually exclusive.

According to PAM, which assumes a direct realist approach,[52] "similarity between non-native segments and native gestural constellations […] are predicted to determine listeners' perceptual assimilation of the non-native phones to native categories" (Best 1995: 194). Thus, this model predicts that perception of a non-native contrast is excellent if the two non-native segments are assimilated to different L1 categories, but

---

[52] See Best (1995) for a summary of the different approaches to speech perception and acquisition: psychoacoustic, direct realist, and motor theory. In contrast, the approach adopted by the BLIP model is neural-based.

poor if the non-native sounds are assimilated to the same L1 category. On the other hand, discrimination will vary from poor to very good if "both non-native sounds fall within phonetic space but outside of any particular native category […] depending upon their proximity to each other and to native categories within native phonological space" (Best 1995: 195). These are only a few of the predictions made by the PAM model, but they serve to illustrate an important cornerstone shared by PAM and all the other L2 models previously cited: the predictions of these models rely on the notion of cross-linguistic perceptual similarity. That is, their predictions are based on whether, and to what degree, the non-native sounds are perceived to resemble native speech categories, a concept referred to as *cross-linguistic perceptual similarity*.

The SLM also relies on the notion of perceptual similarity, but is concerned instead with adult L2 learners' ultimate attainment in the production of non-native segmental categories. In short, this model hypothesizes that the "difference in how new and similar sounds are treated perceptually leads to the prediction that new but not similar sounds in an L2 may be mastered eventually by adult L2 learners (Flege 1993: 1589)." While PAM, SLM and other models that rely on the notion of perceptual similarity generally formulate straightforward and testable predictions, the concept of *similarity* remains ambiguous and difficult to assess. While Flege (1991), who proposed the SLM model, admitted that "no satisfactory method now exists for determining whether an L2 vowel will be treated as new or similar" (p. 704), Ladefoged (1990) argued "it is not even technically possible to devise a measure of auditory distinctiveness among speech sounds

without becoming entangled in the problem of observer bias"[53] (p. 344). Bohn (2005) concurs that indirect measures (e.g., comparing phonetic or phonemic inventories or comparing acoustic or articulatory characteristics) are generally inadequate in evaluating the perceptual distance between L1 and L2 sounds. However, he argues that perceptual similarity can be evaluated by conducting perceptual assimilation identification experiments, where L2 learners are asked to identify the L1 category that is most similar to the L2 sound and to assess the goodness of fit with the L1 category using a grading scale (e.g. from 1-bad to 7-good). These experiments are understandably time-consuming and must be done for each and every L2 sound of interest, since the information provided about the perception of one speech category cannot readily be extended to other categories; under this approach, it is unclear what causes two sounds to be perceived as similar.

That being said, the predictions made by the PAM and SLM models appear reasonable in most cases, and are often borne out in experimental results. Hence, the issue at stake is not whether it is possible to predict which non-native speech contrasts are most problematic for a group of speakers sharing the same L1, but rather, whether it is possible to identify the exact source of the difficulty and to extend this knowledge to predict the perception of any non-native contrast in any other language. A better understanding of the causes of L2 learners' difficulties with non-native speech contrasts would, in addition to being theoretically relevant, have valuable implications for language education, especially for the development of training materials designed to

---

[53] Quotes cited by Bohn (2005).

optimize the time and effort necessary to enable L2 learners to perceive non-native contrasts. In the remainder of this chapter, I demonstrate that the approach provided by the BLIP model offers significant potential in this direction.

4.2 Predictions of the BLIP model for L2 perception

The idea that L1 interferes with the perception and acquisition of L2 contrasts is so entrenched in the field of L2 modeling that we may have neglected to focus on a question that is at least as important: how does the processing of L1 sound contrasts facilitate the perception and acquisition of a new speech contrast? Unlike infants, who are born with neural structures ready to be optimized to perceive any speech contrasts, adult L2 learners have already forged neural maps associated with a set of phonological features relevant to the process of contrasts in their L1. In this section, I demonstrate that the neural maps, and the linkages between these neural maps and phonological features in the learners' L1, might sometimes impinge on, and sometimes facilitate, the perception and acquisition of novel speech contrasts.

In chapter 3, cross-linguistic differences in the neural mapping of spectral and temporal cues were introduced; these differences are summarized in (3) below. In this chapter, I posit that these differences are the potential sources of L2 learners' difficulties in perceiving non-native speech contrasts. However, I also argue that L2 learners may be able to capitalize on their sensitivity to the neural mapping of speech contrasts along a given acoustic dimension to perceive non-native contrasts, even if those contrasts are generally neutralized at the phonological level in their L1.

(3) Potential source of difficulty for the perception of non-native speech contrasts:

a. The number of neural maps along a region of the contrastive acoustic dimension differs in the L1 and L2 (e.g. L1 and L2 use a different number of contrasts along the spectral peak dimension for fricatives);

b. The neural maps in the L1 are not linked with the same features in the L2 (e.g. vowel duration is associated with a vowel length contrast in Japanese but with a coda voicing contrast in English);

c. An acoustic cue is processed differently in L1 vs. L2 (e.g. F0 is processed by AM maps for stress identification in English, but by AM-AM maps for tone identification in Mandarin).

When L2 learners are faced with non-native speech contrasts, a number of things can "go wrong". Mainly, three potential hindrances may impinge on L2 learners' perception and acquisition of non-native contrasts. These are summarized in (3) above. More central to the current discussion, these impediments yield at least five specific predictions about the perception of non-native contrasts. These predictions are numbered from one to five in Figure 4–1 below, yielding a list of questions designed to help identify the exact cause of L2 speakers' difficulties in perceiving a non-native contrast, and to help predict the relative degree of difficulty involved in the perception and acquisition of this contrast.

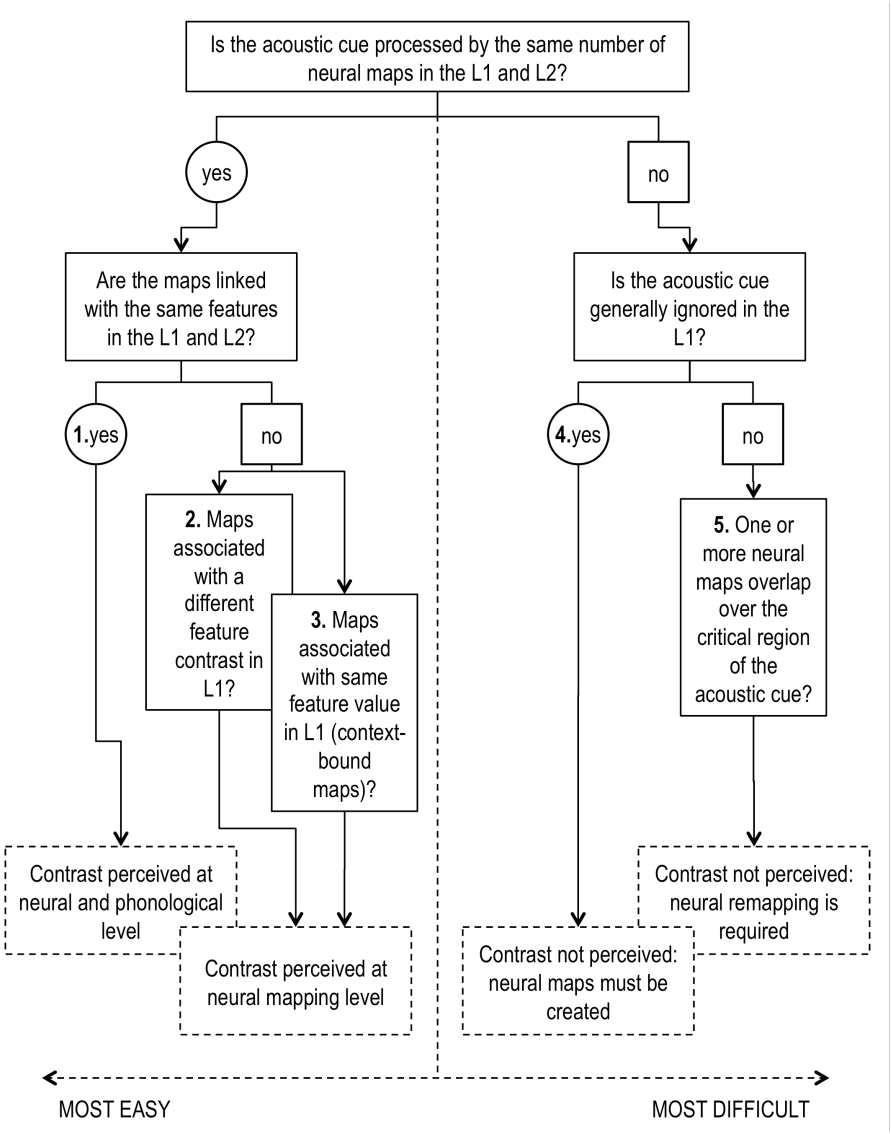

Figure 4–1 Predictions of the BLIP model for perception and acquisition of non-native speech contrasts.

To begin, the first important question is whether the acoustic contrast of interest is processed by the same number of neural maps in learners' L1 and L2. For instance, the processing of the high front vowel contrast in English, as in the words *beat* and *bit,* presumably requires two separate CF-CF neural maps in the acoustic F1-F2 space

occupied by these vowels. Although these vowels are not contrastive at the phonological level in Canadian French, under the current approach, speakers of this French dialect also make use of two (context-bound) neural maps to process these vowels: one to process the so-called tense vowel in open syllable context, the other to process its lax allophonic variant in closed-syllable context (see previous chapter for details). Japanese speakers, on the other hand, are presumed to make use of only one CF-CF neural map for all high front vowels, since they lack the /i/-/ɪ/ contrast. Hence, in the case of Canadian French speakers, the answer to the first question in Figure 4–1 (top) is "yes": Canadian French speakers do have the same number of neural maps along the F1-F2 acoustic dimension used to make the English contrast. Conversely, in the case of Japanese speakers, the answer is "no": Japanese speakers make use of a single neural map that most likely overlaps the two English vowel categories. The first question, therefore, relates to the processing of the acoustic cue at the neural mapping level. If answer to the first question is "yes," the L2 speakers should at least be able to perceive the non-native contrast at the neural mapping level. Conversely, if the answer to the first question is "no," the contrast cannot be perceived at any level, at least not categorically. Presumably, the acquisition of speech contrasts that requires neural organization not already in place may involve a complex restructuring of neural (or synaptic) connections, and is expected to be more difficult than the acquisition of contrasts for which appropriate neural maps are already in place. The vertical doted line in Figure 4–1 denotes this important distinction: contrasts that can be perceived at the neural mapping level are to the left of the vertical line, while contrasts that may require changes in neural organization are on the right side. However, this is not the end of the story: the phonology may also play a non-negligible role in the

perception of non-native speech contrasts (see predictions 1-3), as well as the exact neural mapping or lack of neural mapping of a given acoustic cue (see predictions 4-5). A grading scale from "most difficult" to "most easy" at the very bottom of the figure specifies the relative degree of difficulty in acquiring contrasts that fit the five predicted scenarios described and exemplified below.

Prediction 1

For acoustic contrasts that are already mapped contrastively in the L1 (i.e. for which the answer to the first question is "yes"), there are three possible predictions. If the neural maps in L1 are associated with the same features in the L2 (prediction 1), L2 learners should encounter no major problems. An example that would fall under prediction 1 is French speakers' perception of the English alveolar (e.g. *sea* [si]) and palato-alveolar (e.g. *she* [ʃi]) fricative contrast: French has a comparable distinction (e.g. *sa* [sa] vs. *chat* [ʃa]), the perception of which presumably requires two neural maps based on spectral peak (CF maps) to be associated with equivalent abstract and behaviorally relevant features at the phonological level. That is, even though this contrast may not be acoustically realized in exactly the same way in French and English, French speakers are predicted to be able to perceive the English contrast at both the neural mapping and phonological levels. This prediction implies that French and English speakers should perceive this contrast in largely similar ways, irrespective of task type (e.g. auditory discrimination vs. word-object identification task) and task conditions (e.g. short vs. long ISI). At the same time, it is reasonable to suppose that if some important acoustic differences exist in the realization of those sounds in the two languages, the L2 speakers

(e.g. native French speakers) may need a minimum of exposure to the contrastive distribution used in the L2 to enable them to adjust their neural map boundary to correspond to the one used by native speakers of the L2 (e.g. native English speakers). Non-native contrasts that meet the criteria of prediction 1 should be the easiest to perceive and acquire by L2 learners. Note that I am only concerned here with perception of those contrasts, not their production.

Another example that falls under prediction 1 is the perception of the voicing contrast based on the presence or absence of a periodic signal. For example, since Japanese, English, and French speakers are all sensitive to the presence of a periodic signal to contrast voiced and voiceless fricatives, they should be able to use that same cue to contrast voiced and voiceless stops in word-final position. This latter prediction is tested in Experiments III and IV, presented in sections 4.5 and 4.6 below.

Prediction 2

A second possibility is that while a given acoustic cue in the L2 is processed by the same number of neural maps in L1 and L2, the maps are associated with different feature contrasts in the two languages (prediction 2). For instance, vowel duration is associated with a vowel length contrast in Japanese, but with a coda voicing contrast in English. Consequently, Japanese speakers should be able to perceive the vowel contrast in English at the neural mapping level, but not at the phonological level. In concrete terms, this difference means that perception of the L2 contrast may be dependent on the task and testing conditions: a listener should be able to perceive the non-native contrast if the task requires categorical discrimination of an acoustic cue (e.g. forced-choice

auditory identification task), but unable to perceive the contrast if the task requires processing at a behaviorally relevant level (word-object identification task or real speech conversation). The ability of Japanese speakers to use their sensitivity to vowel duration to perceive the English coda voicing contrast at the neural mapping level is tested in Experiment III, presented in section 4.5 below.

Prediction 3

The third prediction of the BLIP model is that a contrastive acoustic cue in the L2 that is processed by context-bound neural maps associated with the same feature value in the L1 should allow L2 learners to perceive L2 contrasts based on that acoustic component at the neural mapping level only. This would presumably be the case of the perception of the high front unrounded vowels (i.e. [i] and [ɪ]) by native Canadian French speakers. This contrast is context-bound in Canadian French, but contrastive at both the neural mapping and phonological levels in English. As in prediction 2, non-native contrasts that fall within this category are expected to be perceived at the neural mapping level only. However, these contrasts are expected to be somewhat more difficult to perceive and acquire than contrasts that fall under prediction 2 (i.e. they should be more difficult than contrasts that are already distinctive at the phonological level of the learners). The perception of a contrast that falls under prediction 3 is tested in Experiment II, described in section 4.4 below.

Prediction 4

In some cases, the answer to the first question in Figure 4–1,"Is the acoustic cue processed by the same number of maps in the L1 and L2?" is "no." L2 contrasts that fit within this category are expected to be the most problematic and difficult to acquire. In some cases, it is possible that a given acoustic cue is not used for the perception of any speech contrasts in the L1, and this situation corresponds to prediction 4. For instance, vowel duration is not used for any categorical contrasts in Canadian French. Hence, the neurons specialized to process duration contrasts in the auditory cortex of Canadian French speakers have presumably never been mapped into categories: their organization remains neutral, potentially similar to their initial organization at birth (which is still unknown). In this case, the neural maps must be *created*. The level of difficulty related to the creation of these maps depends on the complexity involved in the neural processing of the contrast, or on the general neurological predisposition to perceive that contrast. For instance, vowel duration is partly captured by the duration of tonic responses of CF-CF neurons, which are used to process vowel quality. Hence, speakers of any language should be able to perceive some differences in duration, but the processing may not be sufficiently efficient to allow for the categorical processing of speech contrasts. At this time, it is impossible to determine the degree of difficulty involved in the acquisition of the vowel length contrast (e.g. in Japanese or English) by speakers of languages that do not use durational contrasts (e.g. French speakers), as we do not know enough about the neural processing of vowel duration. However, the prediction that French speakers should be sensitive to changes in vowel duration, but that they may not use this cue as a primary

cue for speech contrasts, is tested in Experiment IV, the results of which are reported and discussed in section 4.6 below.

Prediction 5

Finally, it is possible that a given acoustic cue may not be ignored in an L2 learner's L1, but simply mapped differently. This prediction seems to apply to the case of vowel spectral contrasts; although the number of vowel contrasts may vary from one language to another, the entire vowel space appears to be somehow mapped, irrespective of the number of vowels used within this space. That is, if a vowel contrast is not used in a language, the neural mapping of the closest vowel(s) generally overlaps the "unused" categories. This is presumably the case for Japanese speakers, who are expected to classify the two high front English vowels into a single category based on the F1-F2 contrast (e.g. Morrison 2002). This prediction is tested in Experiment I and discussed in section 4.3 below.

Importantly, since according to the inverted magnification factor hypothesis, the categorical center of an overlapping map exhibits a decrease in cell density activation, reducing listeners' ability to discriminate acoustic details in this area, the task of *remapping* this area into separate categories (prediction 5) is hypothesized to be more difficult than the creation of a new neural map (prediction 4).

Summary

In sum, instead of relying on the notion of cross-linguistic perceptual similarity, the predictions of the BLIP model rely on comparisons of how spectral and temporal

components are mapped and associated with phonologically distinctive features in the L1 versus L2. In some cases, the mapping of a given acoustic cue in an L1 might impinge on the learning of a novel speech contrast (e.g. overlapping maps). In other cases, however, L2 learners may be able to capitalize on their sensitivity to various L1 contrasts to perceive L2 sounds, even if the neural maps that capture this contrast are not contrastive at the phonological level in the L1 (e.g. context-bound neural maps).

Within the BLIP model, there is no need to evaluate the perceptual similarity between L1 and L2 sounds. Rather, under this approach, what is crucial is L2 learners' ability to perceive *contrasts* along the relevant acoustic dimension(s). The BLIP model assumes that if a listener can perceive a sound contrast categorically, this perceiver should possess distinct neural maps to process this contrast. Moreover, these separate neural maps, whether they are associated with the same or different abstract (phonological) features, should facilitate the acquisition of those novel contrasts that are based on the same acoustic contrast (i.e. contrasts that employ the same neural maps).

A noticeable advantage of this approach is that once sensitivity to a given acoustic contrast is assessed, the result can be used to predict the perception of any other contrast in any other language that relies on the same contrastive cue. For instance, if a speaker is sensitive to a speech contrast based on the presence or absence of a periodic signal (whether for a contrast between voiced and voiceless fricatives, or between voiced and voiceless stops), this speaker should be able to perceive and acquire any contrast in any language that uses this cue contrastively. Sensitivity to a contrastive acoustic cue can be tested in two ways: by evaluating which cue, among possible alternatives, speakers generally attend to for their native contrasts (i.e. L1 experiment), or by evaluating which

cue (out of two or more) speakers attend to for categorization of non-native contrasts (i.e. L2 experiment). The following experiments use the second approach, as the purpose here is to evaluate the predictions of the BLIP model for the perception of non-native contrasts.

4.3 Experiment I

The first experiment was designed to reassess results of previous studies suggesting that Japanese speakers are highly sensitive to vowel duration when having to classify the English front vowels while generally ignoring spectral differences (i.e. Morrison 2002). The purpose here is twofold: 1) to provide support for prediction 5 of the BLIP model, that if Japanese speakers' (CF-CF) neural mapping of their high front vowel (/i/) overlaps over the English high front vowel categories (/i-ɪ/) these speakers should be unable to use spectral differences to categorize the English vowels, and 2) to confirm that Japanese speakers may apply their sensitivity to vowel duration to classify an L2 contrast, in this case English vowel categories. The latter objective constitutes a premise to Experiment III that evaluates Japanese speakers' ability to apply their sensitivity to vowel duration for identification of a different speech contrast, that is the English coda stop voicing contrast as discussed below.

The English vowel inventory includes a contrast between a high front tense vowel as in 'beat' (/i/) and a high front lax vowel as in 'bit' (/ɪ/), which is partly captured acoustically by variations in F1 and F2, with the tense vowel having a lower F1 and higher F2 than its lax counterpart (Ladefoged 2001), and the tense vowel may also have a tendency to be slightly longer in duration than its lax counterpart. Vowel duration is

generally contrastive in English, but is primarily used contrastively as a cue for stress and for the coda-voicing contrast (as discussed in Experiment III). By contrast, Japanese has a five-vowel system that includes only one cardinal /i/ vowel and a vowel length contrast (e.g. *chizu* 'map' versus *chiizu* 'cheese') (Akamatsu 1997; Vance 1987). According to the BLIP model, English speakers possess two distinct F1-F2 (CF-CF) neural maps to process the high front vowels, while the neural map used by Japanese speakers to process the vowel /i/ in their L1 is speculated, in the BLIP model (following the overlapping map hypothesis proposed by Guenther and Bohland 2002), to overlap the spectral space occupied by the two English vowels. This case is illustrated in Figure 4–2 below.

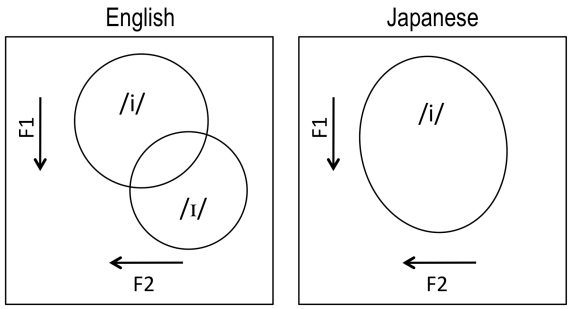

Figure 4–2 Neural mapping of high front vowels in English and Japanese.

Since Japanese speakers possess only one neural map to process an area of the F1-F2 acoustic space that is processed by two neural maps by English speakers, and assuming that the neural map in Japanese at least partly overlaps the two English vowel categories, Japanese speakers are expected to be unable to categorize the words 'beat' and 'bit' based on spectral information (i.e. changes in F1 and F2), as suggested by prediction 5 in Figure 4–1. Given Japanese speakers' sensitivity to vowel duration, it is expected that these speakers will instead use this cue to make a contrast between the two words,

provided that they know that these words are contrastive, as suggested by a previous study by Morrison (2002).

4.3.1 Methodology

Participants

Characteristics of the two groups of speakers who participated in this experiment are summarized in Table 4–1. Twenty-four native speakers of English (12 males, 12 females) were tested at the University of Victoria in Canada, and 24 native speakers of Japanese (12 males, 12 females) were tested at Waseda University in Tokyo. Participants in the native English group spoke only North American English, mostly Canadian English from the West coast. Participants in the native Japanese group were from the Kanto region around Tokyo, and spoke the so-called standard Japanese dialect. All participants had some university education, all reported having no known hearing impairments, and all received a monetary compensation for participating in this experiment ($10 CAD or 1000 YEN).

Table 4–1 Characteristics of the English and Japanese participants. Standard deviations are in parentheses.

Group	Gender	Age in years	Mean age started study	Means years studied	TOEIC score[a]
Japanese	12 males	19.5	12.3	7.2	600
	12 females	(0.9)	(0.9)	(1.2)	(110.8)
English	12 males	20.4	…	…	…
	12 females	(2.8)			

[a] TOEIC, Test of English for International Communication, total score out of 990.

English-speaking participants were aged between 18 and 30 years (mean 20.4) at the time of testing, whereas Japanese-speaking participants were aged between 18 and 21 years (mean 19.5). All the Japanese participants were from monolingual homes, and started receiving English language instruction in school between the age of 10 and 13 years (mean 12.3), for an average of 7.2 years of instruction. None of the Japanese participants reported having ever been abroad, with the exception of one female speaker who studied English in Australia for 10 months about three years before taking part in this experiment.[54] Their most recent reported scores on the TOEIC English proficiency test varied from 330 to 875 (mean 600, std. dev. 110) out of a possible score of 990.

Stimuli

Twenty-four 'beat' and 'bit' tokens were created by cross-splicing and editing portions of a natural speech sample using Praat (Boersma & Weenink 2007). A native Canadian English female speaker was recorded reading 'beat' and 'bit' tokens presented in isolation and within short sentences at three different speech rates (slow, normal, fast) using her natural voice level. The recordings were performed in a sound-attenuated booth with a high quality microphone directed at a 45^{o} angle a few inches from the woman's mouth. The samples were recorded at a sampling rate of 11 025Hz[55] directly to computer using Praat. The averaged formant values produced by the female speaker were compiled for

[54] The most recent TOEIC score of this participant was comparable to the other participants (i.e. 680).

[55] This sampling rate was chosen to avoid having to resample the token at 11,000Hz for manipulations in Praat.

each word featuring the contrast and were compared across speech rates and contexts to be used as guidelines for manipulation of the test tokens.

The test tokens were created by manipulating a single 'bit' token chosen for its sound quality and steady formants throughout the production of the vowel. To preserve the naturalness of the speech sample, the formant transitions in word-initial and word-final positions were not manipulated; nor were any of the formant bandwidths or pitch contours manipulated. The closure duration during the production of the final consonant was fixed to 100ms and the burst release to 130ms for all tokens.

To manipulate the steady portion of the vowel, a filter was extracted from the 'bit' sample source, and manipulated using the FormantGrid editing options in Praat. All points corresponding to the vowel formants were removed and replaced by one constant value per formant, except for those points corresponding to the initial and final formant transitions. For all 24 tokens used for the experiment, F3 was set to 3099Hz, F4 to 4115Hz, and F5 to 5000Hz. For the first test token, F1 and F2 were set at 468Hz and 2200Hz respectively, corresponding to an approximation of the 'bit' samples produced by the female speaker. The F1 was subsequently lowered and F2 increased in steps of 50 Mel (Stevens, Volkmann, & Newman 1937), yielding four spectrally different vowels: F1(468Hz)/F2(2200Hz), F1(417Hz)/F2(2332Hz), F1(368Hz)/F2(2469Hz), and F1(322Hz)/F2(2613Hz). The four manipulated filters were recombined with the initial source, yielding four tokens varying in spectral quality. Pilot testing with native English speakers confirmed that two of the manipulated tokens (those with the highest F1 and lowest F2) were, most of the time, perceived as 'bit,' while the other two tokens (those with lowest F1 and highest F2) were perceived as 'beat' by most speakers.

The duration of each of the four spectrally distinct vowels was lengthened to 210ms. Portions of the vowel were then removed from the center (i.e. without affecting the transitions) in equal steps of 30ms until four vowel continua were obtained, each containing six tokens varying in vowel duration from 60 to 210ms, as illustrated in Figure 4–3.

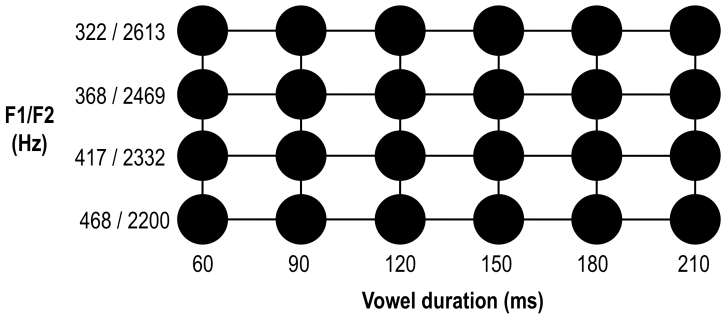

Figure 4–3 Tokens used for Experiment I, which vary in terms of vowel duration and values of F1 and F2 (vowel quality). F1 and F2 varied in equal steps of -50 and +50 Mel respectively (Stevens et al. 1937).

A sample of the test token is provided in Figure 4–4, with a detailed description of the stimulus provided in Table 4–2 below.

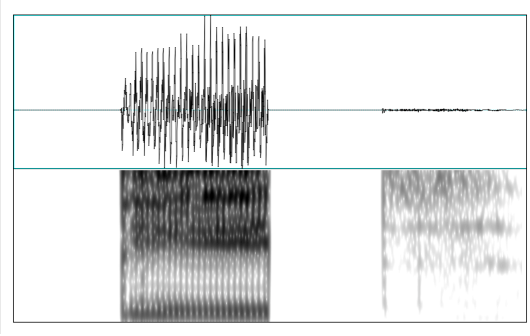

Figure 4–4 Example of a manipulated speech sample used for Experiment I, with a vowel duration of 120ms, F1 of 322Hz and F2 of 2613Hz.

Table 4–2 Acoustic description a test stimulus used for Experiment I with a 120 ms vowel, the lowest F1 and highest F2 frequencies.

Description	Values
Duration of the utterance	459 ms (includes 96 ms of silence prior to the initial noise burst)
Sampling frequency	11 025Hz
Number of formants	5
Duration of the onset noise burst	13 ms
Duration of vowel including transitions	120 ms
Closure duration of word-final consonant	100 ms
Duration of word-final burst release	130 ms (average intensity 45 dB)
Pitch (during vowel production)	Average 192 Hz (min. 184 to max. 208 Hz)
Intensity (during vowel production)	Average 79 dB (min. 77 to max. 80 dB)
Frequency of 1st formant	322 Hz
	Initial transition starting at 332 Hz
	Final transition ending at 365 Hz
Bandwidth of 1st formant[a]	80 Hz
Frequency of 2nd formant	2613 Hz
	Initial transition starting at 2122 Hz
	Final transition ending at 2589 Hz
Bandwidth of 2nd formant[a]	179 Hz
Frequency of 3rd formant	3099 Hz
	Initial transition starting at 3033 Hz
	Final transition ending at 3190 Hz
Bandwidth of 3rd formant[a]	329 Hz
Frequency of 4th formant	4115 Hz
	Initial transition starting at 3891 Hz
	Final transition ending at 4188 Hz
Bandwidth of 4th formant[a]	200 Hz
Frequency of 5th formant	5000 Hz
	Initial transition starting at 4894 Hz
	Final transition ending at 5108 Hz
Bandwidth of 5th formant[a]	124 Hz

[a] Values of the bandwidths were measured by Praat at mid-vowel, and may vary slightly at other locations.

Procedure

A forced-choice identification task was used for this experiment and the following ones, since this auditory task does not require lexical access,[56] and as a result is presumed to activate the neural mapping level of processing without necessitating processing at the phonological level. In the forced-choice identification task, participants listened to one word presented in isolation (the words presented were separated by a long ISI of at least 1500ms) and were required to select which word, out of two choices (in this experiment: 'beat' or 'bit'), the word they heard corresponded to.

The stimuli were presented to participants using SuperLab (computer software version 4.0.7b) installed on a MacBook Pro (2.4 GHz) in stereo through professional quality circumaural (full size) headphones, while participants sat alone in a quiet room. Sound level was adjusted for each participant to a comfortable level. Identification response and response time were entered and recorded using the same computer used to present the stimuli (the same computer was used for all participants).

Instructions to complete the forced-choice identification task were written on the computer screen in the participant's native language. Participants were aware that their response times were recorded and were specifically asked to enter their choice as quickly as possible. Prior to the task onset, the experimenter verified that the non-native participants were familiar with all the test words, in this case 'beat' and 'bit', and provided

---

[56] Conversely, a picture identification task, in which participants are asked to relate a word presented aurally to a picture, is expected to require lexical access, and consequently, activation of the phonological level (e.g. see Curtin, Goad & Pater 1998). On the opposite end, a discrimination task, where two words are presented one after another separated by a very short ISI, presumably taps into the perceiver's just noticeable difference, rather than the neural mapping level, as discussed in chapter 3, section 3.4.

them with a definition or translation of any unfamiliar word. However, the test words were not pronounced to participants before the experiment.

Presentation of each token was preceded by a quick visual prompt on the screen (a red plus sign in the middle of the screen) to direct participants' attention, following which the choices 'beat' and 'bit' appeared on the screen with their corresponding keys on the computer keyboard as the stimulus was presented through the headphones. The words disappeared when the participant successfully entered a choice on the keyboard. There was no time limit set for entering a choice, but after either one of the two possible answer keys was pressed, there was a delay of 1500ms before presentation of the next token (i.e. minimum ISI of 1500ms).

Each participant completed a practice block of 24 trials with each of the possible tokens presented once in a random order. After completing the practice block, the experimental session consisted of three blocks of the 24 tokens (72 trials) with the order of trials randomized within each block. Experiment I lasted about 10 minutes.

4.3.2 Results and discussion

To assess inter-individual variability in overall identification responses, the percentage of 'beat' judgments averaged across stimuli and trials were calculated for each participant. Histograms of the 'beat' identification percentages per language group, shown in Figure 4–5, indicate that English and Japanese participants labeled between 40% and 70% of the tokens as corresponding to the word 'beat', and incidentally between 30% and 60% of the tokens as 'bit'.

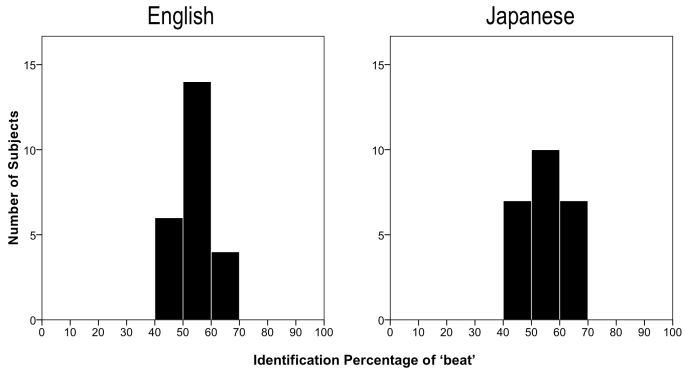

Figure 4–5 Histograms of the aggregated identification percentage (as 'beat') for individual subjects in each language group: English versus Japanese.

Although identification judgments were generally balanced between 'bit' and 'beat' for all participants, the two language groups differed significantly in their use of the acoustic cues manipulated to make their categorical decision. Figure 4–6 summarizes the averaged identification judgment for each test token across all English (top) and Japanese (bottom) participants. In this figure, a white circle corresponds to a token most frequently identified as 'beat' and a black circle to 'bit'. The identification percentage for this category is provided within each circle (with standard errors in parentheses). As can be seen in this figure, English speakers relied primarily on changes in formant values to make their decision; they generally identified tokens with the highest F1 and lowest F2 values as 'bit' and tokens with the lowest F1 and highest F2 values as 'beat'. In contrast, Japanese speakers relied on vowel duration instead; they identified the tokens with the shortest vowels as 'bit' and tokens with the longest vowels as 'beat', and that, irrespective of changes in the F1 and F2 values.

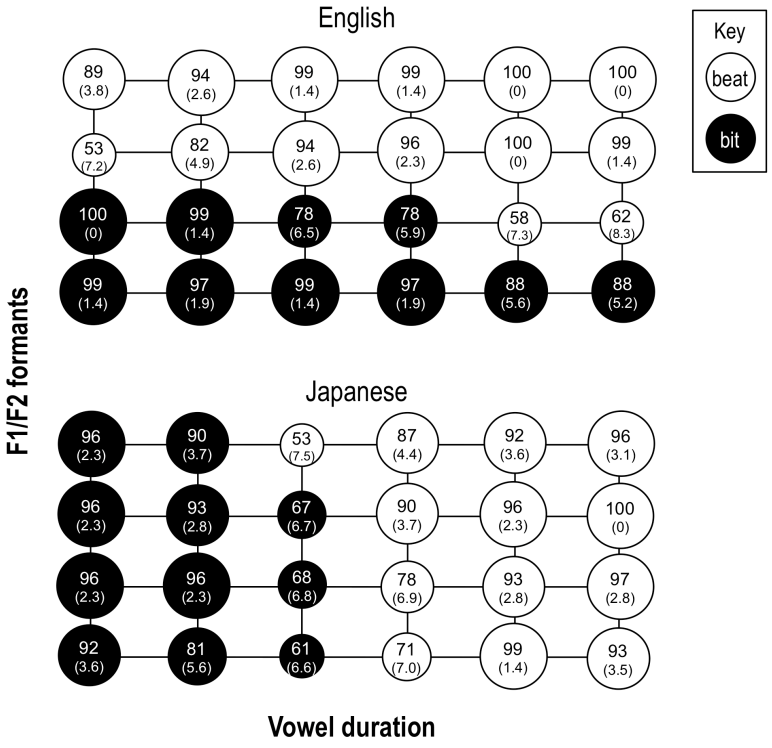

Figure 4–6 Averaged identification of tokens as either 'beat' or 'bit' across English (top) and Japanese (bottom) speakers. The size of each circle corresponds to its identification frequency: large circles indicate higher identification percentages. The shading of the circle indicates the most frequently identified category: white for 'beat' and black for 'bit'. The number within each circle indicates the identification percentage for the most frequently identified category with standard error in parentheses.

A repeated-measure ANOVA was used to evaluate the effect of native language (group) on the use of formants and vowel duration for identification of the manipulated tokens. Results show a significant interaction between group and formants, $F(2.24, 44) =$

212.14, $p < .001$, as well as between group and vowel duration, $F(3.02, 42) = 74.49$, $p < .001$, confirming that English and Japanese speakers used spectral and duration information in a significantly different way to categorize the 'beat'-'bit' contrast.

Multiple regressions using the forced entry method were subsequently performed on the English and Japanese data (i.e. separately) to evaluate the relative use of each independent variable—formant and vowel duration—for categorization of the vowel contrast. The effect of formants and vowel duration predicts 72% of the identification results for English speakers in this model ($R^2 = .723$). As summarized in Table 4–3 below, changes in both vowel duration and formants contributed significantly to the identification decision (each $p < .001$). However, the effect of formants ($\beta = .814$) is greater than the effect of vowel duration ($\beta = .247$). Hence, although English speakers are sensitive to vowel duration, their categorization decision for the high front vowel is based mostly on spectral differences.

Table 4–3 Regression results for English speakers (Experiment I)

	B	SE B	β
Constant	-.521	.032	
Formants	.333	.009	.814*
Vowel duration	.066	.006	.247*

Note: Model $R^2 = .723$, *$p < .001$, B = regression coefficient, SE B = standard error of B, β = standardized regression coefficient

The combined effect of vowel duration and formants also explains about 70% of the Japanese identification results ($R^2 = .698$), suggesting that Japanese speakers relied on the information provided by the manipulated cues to complete the task. However, unlike

English speakers, Japanese speakers did not make significant use of spectral information for identification of the English vowel contrast ($\beta = .025$, p = n.s.), but instead relied on vowel duration ($\beta = .835$, $p < .001$).

Table 4–4 Regression results for Japanese speakers (Experiment I)

	B	SE B	β
Constant	-.233	.032	
Formants	.010	.009	.025
Vowel duration	.215	.006	.835*

Note: Model $R^2 = .698$, *p < .001

The response time (RT) results demonstrate a very different pattern between English and Japanese participants, as illustrated in Figure 4–7 below, which includes the log-transformed[57] RT measures for each language group. A close look at Figure 4–7 suggests that formants and vowel duration are processed competitively by English speakers for the identification of the vowel contrast; the response time for these speakers generally *increases* as the duration of the vowel extends over 150ms for tokens with the two highest F1 and lowest F2 (identified generally as 'bit' by English speakers, henceforth the 'bit' series), whereas the response time gradually *decreases* as the vowel duration varies from 60ms to 210ms for tokens with the two lowest F1 and highest F2 in this experiment (henceforth the 'beat' series).

[57] Log transformations were applied to the response times to reduce the effect of a few outliers found in the data (i.e. response times that were more than 3 standard deviations from the mean).

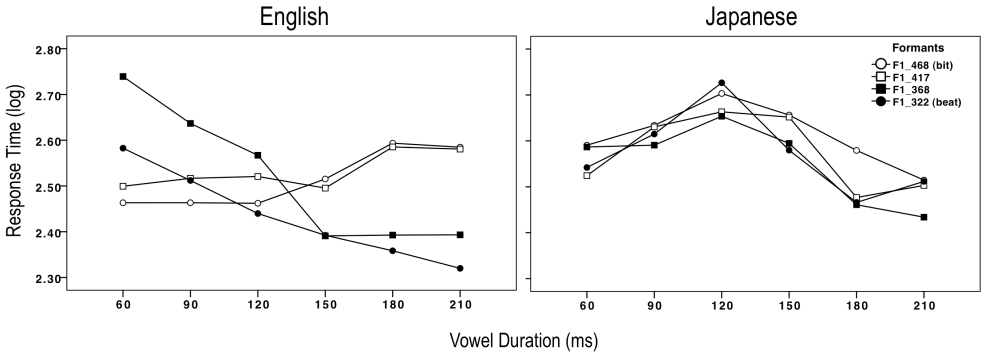

Figure 4–7 Average (log-transformed) response times for the English and Japanese group for each of the 24 tokens in Experiment I.

Segmented regression analyses performed on the split data into the 'bit' and 'beat' series confirm a significant positive effect of vowel duration on response time for the 'bit' series ($\beta = .204$, $p = .001$), and a significant negative effect of vowel duration on response time for the 'beat' series ($\beta = -.496$, $p < .001$), indicating that the effect of vowel duration is more important in the 'beat' series. These results are interpreted as providing evidence for the fact that although spectral information is used as the primary cue by English speakers to distinguish the high front vowel contrast, vowel duration and spectral information may sometimes be processed competitively in English: a relatively short vowel duration is generally associated with the lax vowel in 'bit,' whereas a long vowel is associated with the high tense front vowel in 'beat'. A possible explanation for this tendency may be that the tense vowel tends to occur more often in stressed syllable position—in which case the vowel is lengthened—while the lax vowel often occurs in unstressed, reduced position. An alternative explanation may be that the lax vowel tends to be shorter than its tense counterpart in English, irrespective of context. Either way, English speakers would have come to associate (through a learned auto-associated

pattern) a short duration with the vowel in 'bit,' and a long vowel with the vowel in 'beat,' even if the vowel contrast is based mainly on spectral differences.

In addition, formants were found to have a significant negative effect on response time, but only in the 'beat' series ($\beta = -.195$, $p < .001$); tokens with the lowest F1 (322 Hz) and highest F2 (2613Hz) in this experiment were generally processed more quickly than tokens with a higher F1 (368Hz) and lower F2 (2469Hz) that were also perceived as 'beat', indicating a possible boundary effect. According to the inverted magnification factor hypothesis, tokens at categorical boundaries should activate a higher level of cell density than at categorical centers, and consequently are expected to require a longer processing time. The presence of a boundary effect provides support for the BLIP model hypothesis that native English speakers possess two separate neural maps along the F1-F2 dimension. This finding also has important implications for theories of speech processing by showing that vowels, like consonants, may be perceived categorically, and that the perceptual magnet effect may simply be the result of this categorical perception (see chapter 2, section 2.3.2 for a complete discussion of this issue).

Japanese speakers, on the other hand, exhibit a very different pattern than English speakers, as shown in Figure 4–7. The RTs of Japanese speakers are affected by vowel duration, with a relatively short RT when the vowel duration is either very short (60 < 120ms) or very long (120 < 210ms), but a relatively long RT at the intermediate value (= 120ms). Hence, the words with a vowel duration of 120ms seem to be treated as the categorical boundary along the durational dimension by Japanese speakers in this experiment. Segmented regression analyses performed on the Japanese data split at 120ms confirms a significant positive effect of vowel duration on RT when the vowel

varies from 60 to 120ms (β = .219, p = .001) and a significant negative effect of vowel duration on RT when the vowel varies from 120 to 210ms (β = -.300, p = .001). In either case, no significant effect of formants on RT was found.

These results indicate that Japanese speakers systematically ignore spectral differences between the high front vowels in English, presumably because the F1-F2 neural map in Japanese overlaps the two English categories, which makes perception of variations in F1-F2 in the area occupied by the overlapping map very difficult for Japanese speakers. However, as discussed previously when looking at the English data, the vowel in 'beat' may generally be longer than the vowel in 'bit.' While this cue is not the most reliable one to distinguish the English vowels, Japanese speakers appear to have somehow picked up this distinction and applied the neural maps used to distinguish the short versus long vowel contrast in their L1 to distinguish the high front vowels in English. In this sense, the neural mapping for vowel duration in Japanese interferes with Japanese learners' acquisition of the English vowel contrast by preventing them from attending to spectral differences. Since they are able to make *a* contrast between the vowels, it may not appear necessary to take into consideration other possibly contrastive acoustic cues for the same contrast.

Taken together, these results are interpreted as supporting prediction 5 of the BLIP model by showing that Japanese speakers are unable to perceive the high front English vowel contrast at the neural mapping level in terms of spectral differences because the F1-F2 neural map for /i/ in their L1 overlaps the critical F1-F2 boundary that distinguishes /i/ from /ɪ/ in English. However, according to prediction 2 of the BLIP model, Japanese speakers should be able to use their categorical sensitivity to vowel

duration, as demonstrated in this experiment, to perceive a novel phonological contrast using this cue, such as the English coda voicing contrast. This hypothesis is tested as part of Experiment III reported later in this chapter.

4.4 Experiment II

In the previous experiment, it was shown that Japanese speakers, who do not distinguish phonologically between spectrally different vowels, as in the English words 'beat' and 'bit', systematically ignore spectral differences in order to classify the English high front vowels. Presumably, Japanese speakers process high front vowels using a single F1-F2 map that overlaps the two English categories, as posited by prediction 5 of the BLIP model. However, the BLIP model predicts that not all non-native contrasts absent from the learners' L1 phonology are treated equally or are equally problematic. According to prediction 3 of the BLIP model, if an L2 phonological contrast corresponds to context-bound allophones in the learners' L1, these learners should be able to draw on the distinct neural maps used to perceive these allophones in their L1 to perceive phonological contrasts based on the same acoustic cues in the L2. To test this prediction, perception of the English 'beat' and 'bit' contrast was tested with Canadian French speakers.

Canadian French speakers, as opposed to European French speakers, are known to have a context-bound contrast between the high front (unrounded) vowels, where the tense vowel [i] is used in open syllable context, as in *petit* 'small, masc.' [pətˢi], while its lax counterpart [ɪ] is used in closed syllable context, as in *petite* 'small, fem.' [pətˢɪt] (see Martin 1996 for additional examples). Provided that this context-bound variation exhibits a bimodal distribution along the F1-F2 dimension, as discussed in chapter 3, native

Canadian French speakers should have two neural maps to process these vowels. This scenario is illustrated in Figure 4–8 below, where the neural mapping of the same vowel by native English speakers is shown for comparison. However, since these allophones are not contrastive at the phonological level, their corresponding neural maps are presumably associated with the same feature in French, that is |i|. Accordingly, prediction 3 of the BLIP model implies that Canadian French speakers should be able to capitalize on their sensitivity to the vowel contrast at the neural mapping level to perceive the English phonological contrast, provided that the task and testing conditions allow them to use their context-bound neural maps. Experiment II uses the same forced-choice identification task as the one used in Experiment I, which should allow listeners to complete the task using their neural mapping level of processing (although the use of the phonology cannot be totally excluded).

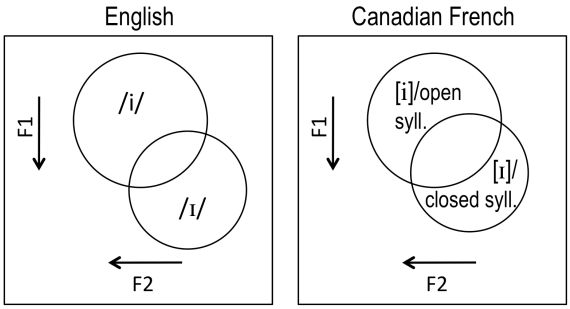

Figure 4–8 Neural mapping of high front vowels in English and Canadian French. The amount of overlapping between the categories in each language is arbitrary.

4.4.1 Methodology

Participants

Characteristics of the Canadian French speakers who participated in Experiment II are summarized in Table 4–5. The English and Japanese speakers used in this experiment to compare with the Canadian French speakers are the same speakers used in Experiment I. For the current experiment, 24 native Canadian French speakers (7 males, 17 females)[58] from the province of Québec were tested at Laval University, in Québec City. Two additional French-speaking participants were tested, but their data were excluded from analysis for the following reasons: one participant had systematically inverted the labels for the test words, while the other participant's English proficiency level was beyond the desired range for this experiment.[59] All participants had at least some university or college education; none reported having any known hearing impairment; and the Canadian French participants received the same monetary compensation for their participation as the English-speaking participants in Experiment I ($10 CAD).

---

[58] During the recruitment at the university level, it was very difficult to find Canadian French male participants who were not fluent in English. Therefore, more female participants were recruited instead to make the overall data as comparable as possible with the Japanese group in terms of L2 proficiency level. Many of the potential male candidates excluded from the experiment reported being engaged in English computer games from a very young age, a factor that may have contributed to their (self-reported) higher English proficiency than the female candidates encountered during the recruitment process.

[59] The experiment evaluates the perception of non-native sounds by L2 learners. Therefore, bilingual or near-bilingual speakers were excluded. Bilingual or near-bilingual proficiency was determined based on the participants' last English proficiency test result, whenever available, and on their answers to questions about their use of English and level of English exposure in their day-to-day life.

Table 4–5 Characteristics of the Canadian French participants. Standard deviations are in parentheses.

Group	Gender	Age in years	Mean age started study	Mean years studied
Canadian French	7 males	21.3	9.6	8.9
	17 females	(2.7)	(0.9)	(1.2)

All French-speaking participants grew up in monolingual homes, and none of them had stayed or traveled in an English-speaking environment, with the exception of one participant who spent five weeks in an English immersion program in Alberta about three years before taking part in this experiment. Participants in the French-speaking group were aged between 17 and 29 years old (mean 21.3) at the time of testing; started studying English at school between the age of 8 and 12 years old (mean 9.6); and had completed, on average, 8.9 years of education in the English language.

Stimuli

The stimuli were the same as in Experiment I.

Procedure

The procedure was the same as in Experiment I.

4.4.2 Results and discussion

To assess inter-individual variability in overall identification responses, the percentage of 'beat' judgments averaged across all stimuli and trials was calculated for each participant. The histogram of the 'beat' identification percentages, shown in Figure 4–9, indicates that Canadian French speakers identified between 30% and 60% of the tokens as

corresponding to the word 'beat'; and conversely, 40% to 70% of the tokens as 'bit'. Thus, on average, Canadian French speakers had a tendency to identify more tokens as corresponding to 'bit' in this experiment than did the English and Japanese speakers in the previous experiment (who identified 30% to 60% of the tokens as 'bit').

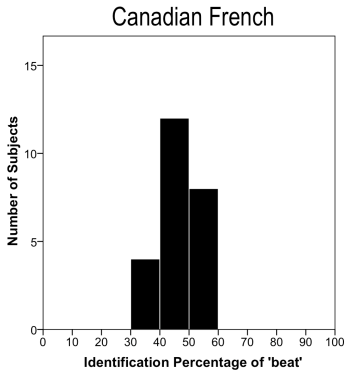

Figure 4–9 Histogram of the aggregated identification percentage (as 'beat') for individual Canadian French participants.

Figure 4–10 presents the averaged identification judgment for each test token across the Canadian French participants. These speakers' pattern of identification resembles that of the native English speakers reported in Experiment I: tokens with the highest F1 and lowest F2 (bottom rows) were overwhelmingly identified by Canadian French speakers as corresponding to the word 'bit', while tokens with the lowest F1 and highest F2 (top rows) were generally identified as corresponding to the word 'beat'. However, a repeated-measure ANOVA conducted on the English and Canadian French data indicates a significant effect of native language (group) on the use of formants $F(1.42, 44) = 17.96$, $p < .001$, and on the interaction between formants and duration (group x formants x duration) $F(8.98, 32) = 3.78$, $p < .001$, but no significant effect of native language on the use of duration alone $F(5, 42) = 0.94$, $p =$ n.s. These results

suggest that, on average, Canadian French participants used vowel duration, but not formants, in a comparable way to native English speakers. However, further analyses presented and discussed below reveal that the identification patterns vary considerably across Canadian French speakers. In fact, most French speakers appear to use formants in a way that is comparable to the native English-speaking participants. Before discussing the results of these analyses, the effect of each acoustic cue, as assessed by multiple regressions performed on the overall Canadian French data, is summarized in Table 4–6 below.

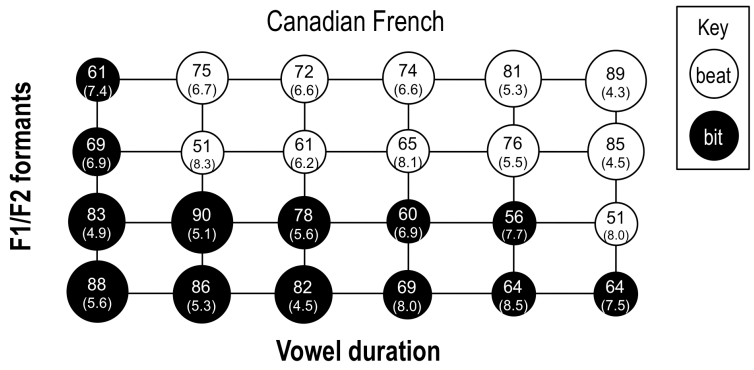

Figure 4–10 Averaged identification of tokens as either 'beat' or 'bit' across Canadian French speakers. The size of each circle corresponds to its identification frequency: large circles indicate higher identification percentages. The shading of the circle indicates the most frequently identified category: white for 'beat' and black for 'bit'. The number within each circle indicates the identification percentage for the most frequently identified category with the standard error in parentheses.

Results of the regression analyses indicate that the use of formants and vowel duration accounts for only 34% of the variance in the model (R^2 = .344), which is lower than for both the English and Japanese data (72% and 70% respectively). Nevertheless, changes in both vowel duration and formants contributed significantly to the identification judgments of Canadian French participants (each $p < .001$). Changes in formants had a greater effect (β = .479) than changes in vowel duration (β = .338). These results are consistent with the general hypothesis that native Canadian French speakers may possess two separate neural maps to process the high front (unrounded) allophonic contrast in their L1, and that these speakers are able to capitalize on these neural maps to perceive the English phonological vowel contrast.

Table 4–6 Regression results for Canadian French speakers (Experiment II)

	B	SE B	β
Constant	-.237	.043	
Formants	.172	.012	.479*
Vowel duration	.079	.008	.338*

Note: Model R^2 = .344, *p < .001

A visual inspection of the individual data reveals that the pattern of identification appears to differ considerably among the French speakers: some speakers relied primarily on formant changes, while other speakers exhibited a bias towards the use of vowel duration (accounting, perhaps, for the fact that all the shortest vowels were generally identified as 'bit' irrespective of formant changes in the overall results in Figure 4–10 above). To assess these differences, participants were divided into groups according to

their respective bias pattern, determined by compiling a ratio of 'bit' responses across formants and vowel duration categorical boundaries. To calculate this ratio, the total number of 'bit' responses for the tokens with the two highest F1 and lowest F2 (two bottom rows) were first compiled across trials and across all 12 tokens for each participant. The same procedure was applied to the tokens with the two lowest F1 and highest F2 (two top rows), and a ratio was compiled by taking the latter result divided by the former. A ratio of .5 indicates that the participant labeled at least twice as many tokens in the bottom rows as 'bit' than in the top two rows, suggesting that the participant had a bias toward using formant changes as a cue for categorical decisions. The ratio of .5 was chosen based on the results from Experiment I and on visual inspection of the individual ratios of English and Japanese participants compiled in the same manner. It has been established in the previous experiment that English speakers use formants as the primary cue for categorizing the tokens as 'beat' or 'bit.' The ratio of all English participants was lower than .5, meaning that the English participant labeled *at least* twice as many tokens in the bottom rows as 'bit' than in the top two rows. Conversely, Japanese speakers were found to ignore changes in formants (p = n.s.), and the ratio was above .5 for all individual Japanese speakers. Thus, a ratio of .5 was established as a sensible cut-off point to identify a possible bias toward the use of a given cue for each individual participant. The same procedure was used to establish a possible vowel duration bias, by dividing the set of tokens into shortest vowels (three leftmost columns) and longest vowels (three rightmost columns). Again, a ratio was compiled and a cut-off point of .5 was again used as the criterion for a possible bias. This procedure yields four bias scenarios, all of which were found in the Canadian French data but only two of which

were found in the English data: formant bias, duration bias, formant and duration bias, and no clear bias.

As illustrated in Figure 4–11, 12 Canadian French participants exhibited a bias toward using formants as a major cue for their categorical decisions; five relied heavily on vowel duration; three exhibited a bias towards using both cues; and four did not exhibit any clear bias toward using either cue.

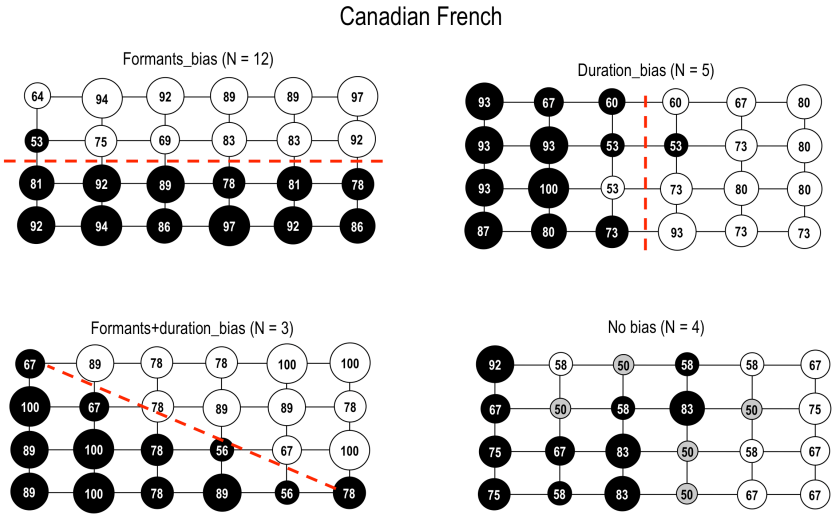

Figure 4–11 Averaged identification of tokens as either 'beat' (white circles) or 'bit' (black circles) across Canadian French speakers classified according to their pattern of response: formants bias, duration bias, formants + duration bias, or no bias.

For comparison, the same procedure was applied to the English and Japanese data collected in Experiment I. Using the ratio of .5 as the criterion for a clear bias, 23 of 24 Japanese participants exhibited a clear bias toward the use of vowel duration for their categorical judgments; and only one Japanese participant did not exhibit a clear bias

toward the use of either formants or vowel duration (though the ratio of this participant for the duration bias was .58, as compared to 1.11 for formants, indicating that this participant still used duration more than formants to categorize the test tokens). The results for the English participants, recompiled in Figure 4–12 according to their identification pattern, indicate that 19 had a clear bias toward the use of formants, while five are biased toward the use of vowel duration and formants.

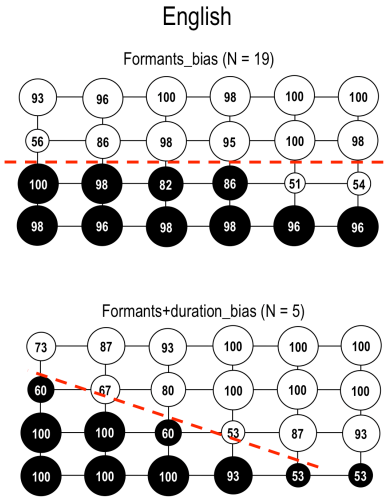

Figure 4–12 Averaged identification of tokens as either 'beat' (white circles) or 'bit' (black circles) across English speakers classified according to their pattern of response: formant bias, or formant + duration bias.

Since these differences in identification pattern may partly explain the relatively low fit of the general model of the Canadian French speakers reported in Table 4–6, additional regression analyses were performed using only the responses of the 12 participants showing a formant bias, combined with the three participants showing a formant + duration bias, since these patterns correspond to the two patterns observed in

the English data. The results of these regression analyses are reported in Table 4–7 below, and show a better fit to the model: For 15 of the 24 Canadian French participants, the effect of formants and vowel duration accounts for 60% of their results (an increase of 26% from the model including all Canadian French participants). In addition, these participants appear to use both formants and vowel duration significantly ($p < .001$) and to an extent comparable to the English-speaking participants (effect of formants: $\beta = .814$ for the English group vs. $\beta = .743$ for the French group; effect of vowel duration: $\beta = .247$ for the English group vs. $\beta = .220$ for the French group). Accordingly, it can be concluded that 15 of 24 (62.5%) Canadian French participants were able to categorize the English high front vowels in a way comparable to the English-speaking participants without training or instruction in the use of English acoustic cues for vowels. Accordingly, these results are interpreted as supporting the prediction of the BLIP model that Canadian French speakers use two separate neural maps to process their context-bound allophonic variants of /i/, and that these neural maps may help them to perceive and acquire a new phonological vowel contrast that uses an equivalent F1-F2 bimodal mapping, such as the English vowels in 'beat' and 'bit'.

Table 4–7 Regression results with Canadian French speakers showing a formant bias or formant + duration bias (N = 15)

	B	SE B	β
Constant	-.401	.044	
Formants	.276	.012	.743*
Vowel duration	.053	.008	.220*

Note: Model $R^2 = .601$, *$p < .001$

However, nine of 24 native Canadian French speakers either use vowel duration as a cue for their categorical judgments (N = 5) or appear to have used neither vowel duration nor formants (N = 4). It is important to keep in mind that although participants may have been able to complete the task without processing at the phonological level, the task does not necessarily prevent them from accessing this level. The four participants that did not show a clear bias toward either duration or formants may have considered the phonology and treated the two high front vowels as belonging to the same abstract category. In other words, without explicit instruction, L2 listeners may not know which cue they should attend to, and may choose to respond according to L1 phonological contrasts. However, the majority of the Canadian French speakers (N = 15) did take into consideration variations in formants and vowel duration, indicating that speakers of this French dialect should be able to use the formant contrast if properly guided or instructed.

The fact that five French speakers appear to have ignored changes in formants in favor of vowel duration reveals two additional important facts. First, without proper instruction, it is not instinctively obvious for L2 learners which cue is most important for an L2 contrast, especially when different cues are processed competitively (i.e. both may concur to identify the same speech contrast). Second, the ability of 15 of 24 Canadian French speakers to attend to variations in formants is not the result of their inability to perceive durational variations for vowel categorization.

Regression analyses on the overall response time of the Canadian French participants, shown in Figure 4–13 below, reveal a small but significant effect of vowel duration ($\beta = -.114$, $p < .01$), where the RT decreases as the duration of the vowel (and of the stimulus) increases, but reveal no significant effect of formants ($\beta = -.050$, p = n.s.)

Comparable results were also obtained on the data when including only the participants exhibiting a formant or formant + duration bias. These results may simply suggest that Canadian French speakers do not process vowel duration and spectral information as competitively as English speakers. Among English speakers, vowel duration had a positive effect on response time for tokens generally labeled as 'bit' and a negative effect on response time for tokens generally labeled as 'beat'.

Furthermore, no significant effect of formants was found on the segmented data into 'bit' and 'beat' series, as described in the previous experiment. A significant effect of formants on response time was found only in the 'beat' series in the English data, pointing to a possible boundary effect on processing time. The lack of a similar observable effect in the Canadian French data may simply indicate that no boundary effect was detectable in the RT results in this experiment for these speakers. The fact that the posited separate neural maps for processing the Canadian French allophones are context-bound and associated with the same phonological feature in that dialect, may obscure a possible boundary effect when the L2 learners are faced with the same contrast presented in the same context (in this case, both English high front vowels were presented in closed syllable context, where the lax allophone usually occurs in the Canadian French dialect, rather than its tense counterpart).

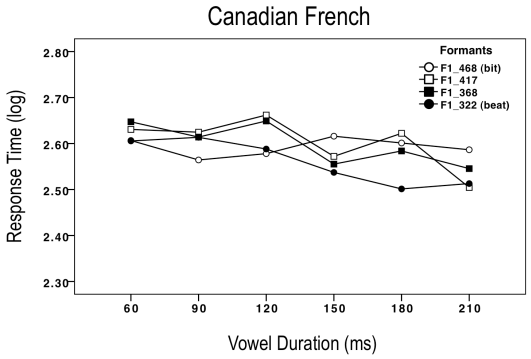

Figure 4–13 Average (log-transformed) response times for the Canadian French group for each of the 24 tokens in Experiment II.

At this point, it impossible to properly assess whether the different identification patterns correlate with participants' English proficiency level.[60] Although this is not impossible, the BLIP model suggests that if Canadian French speakers were aware of which cue is relevant to contrast the English vowels, and shown that these vowels are roughly equivalent to the allophonic contrast in their L1, all participants would be able to use this cue to contrast these vowels irrespective of their proficiency in English. This approach also predicts that European French speakers should be less sensitive than Canadian French speakers to variations in formants in identifying the high front English vowel contrast. This prediction has not yet been tested, but a project is currently underway to evaluate this prediction with Parisian French speakers.

---

[60] The level of English proficiency reported by the participants in the language profile designed to assess participants' past and daily amount of exposure to English and formal English education does not suggest a correlation between the pattern of responses of the French participants and their level of proficiency in English.

In summary, the results of Experiment II demonstrate that the majority of Canadian French speakers in this experiment —unlike Japanese speakers—are able to attend to variations in formants to identify the high front English vowel contrast, as in 'beat' and 'bit'; and that they are able to categorize the English vowels in a way comparable to native English speakers. These results support the context-bound neural map hypothesis, according to which context-bound allophones are processed by separate neural maps. The results also support the related hypothesis that L2 learners can capitalize on these separate L1 maps to perceive novel L2 contrasts, even if these contrasts are neutralized at the phonological level (in their L1), as suggested by prediction 3 of the BLIP model.

4.5 Experiment III

Japanese speakers are sensitive to vowel duration in distinguishing short versus long vowels in their L1 and may apply this ability to categorize non-native vowels as demonstrated in Experiment I. It is questionable, however, whether they can apply their sensitivity to vowel duration to categorize a different speech contrast, such as the coda voicing contrast in English. Experiment III evaluates the ability of native Japanese speakers to capitalize on their sensitivity to vowel duration and periodicity to categorize the English stop consonant in coda position in a way comparable to native English speakers. According to the predictions of the BLIP model, Japanese speakers should be able to use both periodicity (prediction 1) and vowel duration (prediction 2) to categorize English coda consonants in the words 'bit' versus 'bid', since these acoustic cues are used for other phonological contrasts in their own language.

In English, the coda stop voicing contrast is partly captured by the duration of the preceding vowel; a voiceless coda consonant (e.g. *bit* [bɪt̲]) is generally preceded by a shorter vowel than its voiced counterpart (e.g. *bid* [bɪd̲]) as exemplified in Figure 4–14 below. The coda stop voicing contrast exhibits additional acoustic differences: the voiced stop (bottom) may contain the presence of glottal pulses during the stop closure (i.e. periodicity may be present) as well as in the release burst. Further, the voiced consonant generally exhibits relatively shorter closure duration, shorter release burst, and lower F1 offset than its voiceless counterpart (top). Native English speakers are particularly sensitive to vowel duration for the categorization of stop voicing contrasts (Flege 1993) as well as to transitions into the final stops (Fischer & Ohde 1990). Although the presence of periodicity during the stop closure may not be a crucial cue for the stop voicing contrast in coda position (e.g. as suggested by the pilot testing reported in Flege 1993), prediction 1 of the BLIP model implies that English speakers should be sensitive to the presence or absence of periodicity for the categorization of any speech contrast, whether in an L2 or in their L1, since English speakers are presumably sensitive to this cue, among others, to distinguish their voiced from voiceless fricatives. Hence, native English speakers in this experiment are predicted to take into consideration both vowel duration and the presence of periodicity in making categorical judgments pertaining to the words 'bit' and 'bid'.

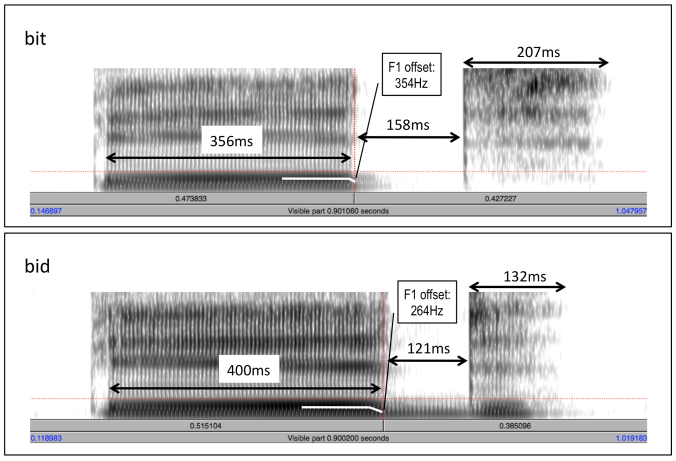

Figure 4–14 Spectrograms of the words 'bit' and 'bid' produced by a female native speaker of Canadian English showing possible acoustic differences between the voiced and voiceless stop in coda position: the voiced stop (bottom) is characterized by a longer preceding vowel, shorter closure duration, shorter release burst, and lower F1 offset than its voiceless counterpart (top), as well as by the possible presence of glottal pulses (i.e. periodicity) during the stop closure and during the release burst.

Japanese, on the other hand, has severe restrictions on the consonants that can occur in coda position. Only a nasal consonant is permitted in word-final position (e.g. *hon* [ho<u>n</u>] 'book'), while in word-medial position a coda may contain a nasal that is homorganic to a following onset obstruent (e.g. *onpu* [o<u>m</u>.pu] 'musical note') or the first part of a geminate consonant (e.g. *kappatsu* [ka<u>p</u>.pa.tsu] 'briskness') (Akamatsu 1997). Hence, Japanese does not exhibit a phonological coda voicing contrast, whether for stop or fricative consonants. However, Japanese does have a voicing contrast for both stops and fricatives in onset position. Of particular interest, the stop voicing contrast in

Japanese is characterized by a negative VOT for voiced stops and short-lag VOT for voiceless stops (Shimizu 1996). This means that Japanese speakers should be sensitive to the presence of periodicity during stop closure to make a voicing contrast. Accordingly, prediction 1 of the BLIP model implies that Japanese speakers should be able to capitalize on their sensitivity to periodicity to perceive the English stop coda consonants, since Japanese speakers already use this cue for other native contrasts. In addition, prediction 2 of the BLIP model posits that Japanese speakers should be able to apply their sensitivity to vowel duration to perceive the English coda voicing contrast, even though vowel duration is associated with a different phonological feature in Japanese than in English.[61]

The difference in neural mapping of vowel duration and periodicity between English speakers and Japanese speakers is illustrated in Figure 4–15 below. As shown in the figure, according to the BLIP model, vowel duration is processed by $CF_{onset\text{-}offset}$ neural maps in both languages. However, this map is presumably associated with a voicing contrast in English at the phonological level (and potentially also a stress contrast), whereas the same map is associated with a vowel length contrast in Japanese at the phonological level. Periodicity is processed by the AM neural map and associated with the same voicing contrast at the phonological level in both languages. In sum, the only difference in the processing of these cues by English and Japanese speakers occurs

[61] Note that the "easiest to most difficult" scale in relation to the predictions in Figure 4–1 does not imply that if two cues are processed competitively for identification of the same abstract feature, the cue towards the "easiest" end will be weighted more. At this point, the scale is simply intended to predict whether a cue can be perceived or not, and if so, what the relative difficulty in perceiving this cue contrastively is, in relation to a given speech contrast.

at the phonological level, not the neural mapping level. Therefore, in a task that does not specifically require lexical access and processing, Japanese speakers should be able to use both cues for categorization of speech sounds in a way comparable to native English speakers.

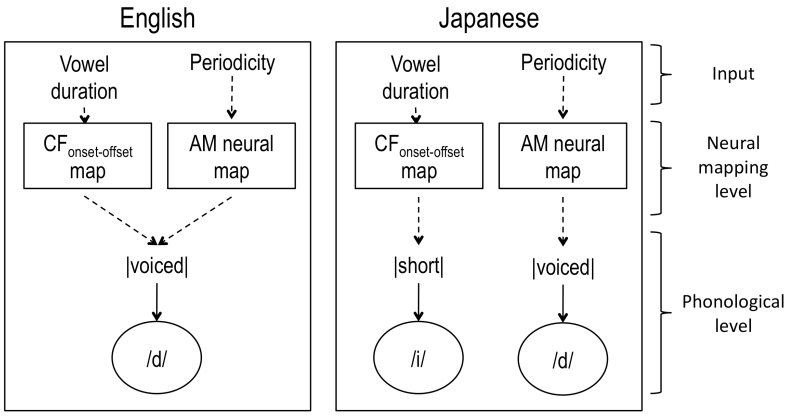

Figure 4–15 Processing of vowel duration and periodicity for speech contrasts in English versus in Japanese.

4.5.1 Methodology

Participants

The English and Japanese speakers who participated in Experiment III are the same as those who participated in Experiment I.

Stimuli

Twenty-four /bit/ and /bid/ tokens were created by cross-splicing and editing portions of a natural speech sample obtained from the same recording session with the native female

speaker of Canadian English described in Experiment I. The manipulations were performed in Praat (Boersma & Weenink 2007).

The test tokens were created by using a /bid/ sample as the starting point, which was chosen for its good sound quality and the presence of periodicity throughout the production of the coda stop closure. The vowel formants and duration were manipulated by following the same steps described in Experiment I, except that in this case only one vowel quality was created, and the vowel duration was varied in equal steps of 50 ms, from 60 ms to 310 ms. For all test tokens, the F1 was set to 415 Hz, F2 to 2163 Hz, F3 to 3027 Hz, F4 to 4130 Hz and F5 to 4846 Hz. The closure duration of the word-final stop consonant was fixed to 100 ms, and its following release burst to 35 ms. The release burst was taken from a /bit/ sample, and was considerably shortened. Its very short duration may suggest a voiced consonant, but the absence of any periodicity during its production may instead suggest a voiceless stop. Hence, this cue was mainly unreliable in deciding the voicing status of the final stop. Although optional in utterance-final position in English, a release burst was included to emphasize the presence of a word-final consonant, which may not be obvious for non-English speakers without the presence of any release burst. For similar reasons, the original formant transitions of the /bid/ token were used, since the offset of the formant transitions were lower (i.e. more salient) in the /bid/ samples than in the /bit/ samples. Pilot testing with a few native and non-native speakers of English suggested that the lower transitions were easier to perceive, especially for non-native speakers, and highlighted the presence of a word-final consonant.

Periodicity during closure duration of the natural /bid/ sample was first set to a duration of 60 ms starting at the vowel offset. Portions of the periodicity were gradually removed from the end of the 60 ms in segments of 20 ms, to create 4 continua varying in the duration of periodicity during stop closure, from 0 ms to 60 ms, as schematized in Figure 4–16.

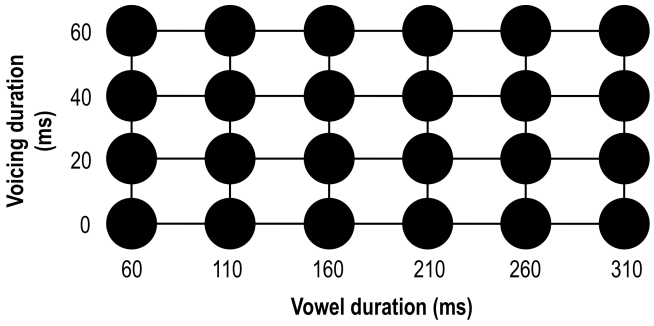

Figure 4–16 Tokens used for Experiment III, which vary in terms of vowel duration and duration of periodicity during word-final stop closure.

Importantly, this procedure ensures that the closure duration is never fully voiced, since the total closure duration exceeds the duration of periodicity in any of the continua, as illustrated in Figure 4–17 with a token containing 60 ms of voicing compared to a token containing no voicing. A full acoustic description of one of the test tokens with 60 ms of periodicity is provided in Table 4–8 below.

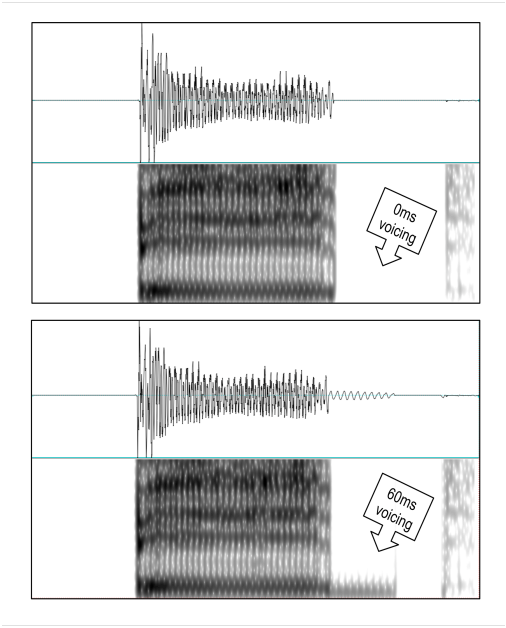

Figure 4–17 Example of a manipulated speech sample used for Experiment III: test token with 0ms of periodicity (top) and 60ms of periodicity (bottom).

To summarize, only the duration of periodicity during the final stop closure and vowel duration were manipulated for Experiment III testing English and Japanese speakers' perception of the English coda voicing contrast in /bit/ vs. /bid/ stimuli. For English speakers who generally associate vowel duration with a coda voicing contrast, the presence of periodicity in tokens with short vowel duration is contradictory, since a short vowel duration suggests a *voiceless* consonant, while the presence of periodicity suggests a *voiced* consonant. As a result, it is predicted that these cues may be processed competitively by native English speakers. The specific case just described should trigger an increase in response times.

Table 4–8 Acoustic description a test stimulus for Experiment III with a 160 ms vowel and 60 ms of periodicity during closure of the coda consonant.

Description	Values
Duration of the utterance	403 ms (includes 96 ms of silence prior to the initial noise burst)
Sampling frequency	11 025Hz
Number of formants	5
Duration of the onset noise burst	12 ms
Duration of vowel including transitions	160 ms
Closure duration of word-final consonant	100 ms
Duration of periodicity during coda closure	60 ms (average intensity 67 dB, starting from max. of 72 dB near the vowel and gradually decreasing to min. of 60 dB)
Duration of word-final burst release	35 ms (intensity was too low to be measured by Praat)
Pitch during vowel production	Average 189 Hz (min. 178 to max. 210 Hz)
Intensity during vowel production	Average 80 dB (min. 73 to max. 85 dB)
Frequency of 1st formant	415 Hz
	Initial transition starting at 434 Hz
	Final transition ending at 339 Hz
Bandwidth of 1st formant[a]	72 Hz
Frequency of 2nd formant	2163 Hz
	Initial transition starting at 1959 Hz
	Final transition ending at 1924 Hz
Bandwidth of 2nd formant[a]	163 Hz
Frequency of 3rd formant	3027 Hz
	Initial transition starting at 2777 Hz
	Final transition ending at 2901 Hz
Bandwidth of 3rd formant[a]	147 Hz
Frequency of 4th formant	4130 Hz
	Initial transition starting at 3693 Hz
	Final transition ending at 3919 Hz
Bandwidth of 4th formant[a]	207 Hz
Frequency of 5th formant	4846 Hz
	Initial transition starting at 4880 Hz
	Final transition ending at 4698 Hz
Bandwidth of 5th formant[a]	460 Hz

[a] Values of the bandwidths were measured by Praat at mid-vowel, and may vary slightly at other locations.

Procedure

The same procedure was used as in the previous experiments, except that in this experiment, participants were presented with a choice between the words 'bit' and 'bid' on the computer screen. Experiment I and III were conducted during the same session. Half the participants completed Experiment III before Experiment I, and the two experiments were separated by a mandatory two-minute pause. Experiment III, like Experiment I, lasted about 10 minutes.

4.5.2 Results and discussion

To get a general picture of inter-individual variability in overall identification responses, the percentage of 'bid' judgments averaged across all stimuli and trials was calculated for each participant and compiled into histograms, as shown in Figure 4–18. These results indicate that most English speakers labeled between 40% and 80% of the tokens as corresponding to the word 'bid, and between 20% and 60% of the tokens as 'bit'. That is, some English speakers had a tendency to identify more of the tokens as 'bid' than as 'bit'. By contrast, most Japanese speakers generally identified as many tokens as 'bid' (40% to 60%) as 'bit' (40% to 60%).

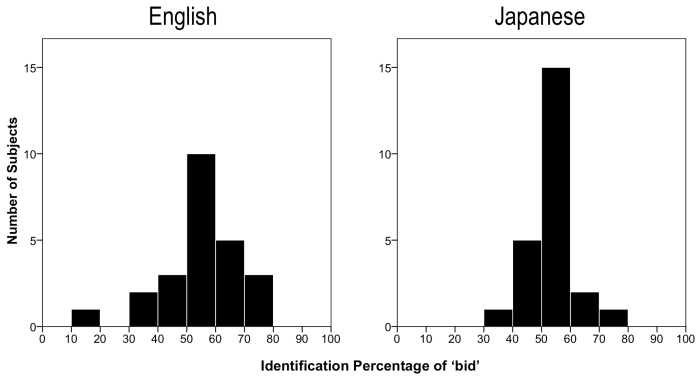

Figure 4–18 Histograms of the aggregated identification percentage (as 'bid') for individual subjects in each language group: English versus Japanese.

The averaged identification patterns for the English and Japanese groups, however, are very similar, as shown in Figure 4–19. In this figure, a white circle corresponds to a token identified mainly as 'bid' and a black circle to a token identified in most cases as 'bit'. As predicted, English and Japanese speakers appear to have taken into consideration vowel duration and duration of voicing during the coda stop closure to categorize the manipulated stimuli: tokens with a short vowel and short voicing duration are generally identified as 'bit' (left bottom rows), whereas tokens with a longer vowel and longer voicing duration (right columns and top rows) are generally identified as 'bid' by speakers of both language groups.

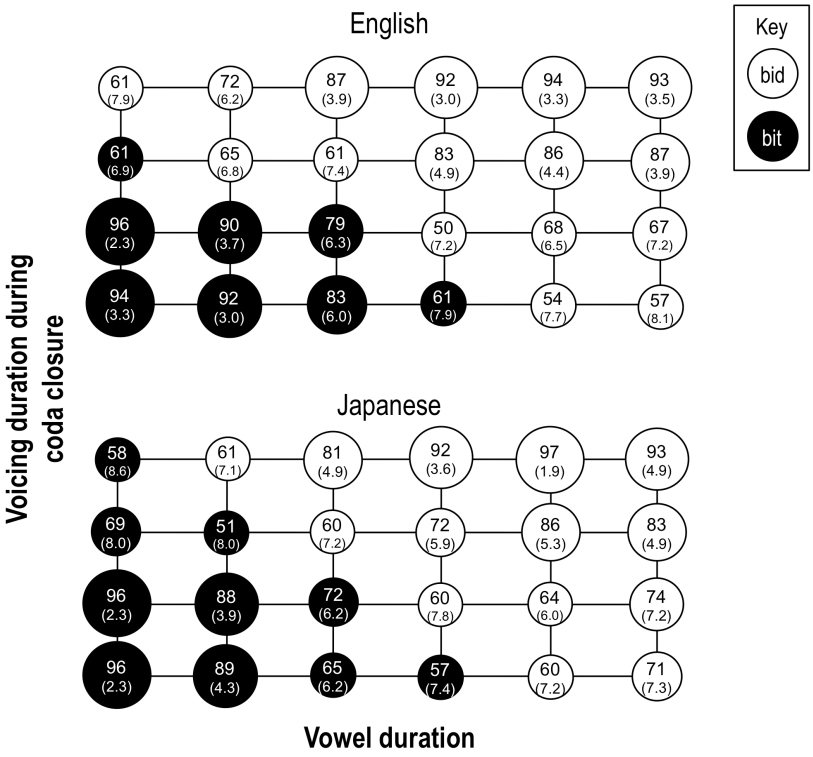

Figure 4–19 Averaged identification of tokens as either 'bit' or 'bid' across English (top) and Japanese (bottom) speakers. The size of each circle corresponds to its identification frequency: large circles indicate higher identification percentages. The shading of the circle indicates the most frequently identified category: white for 'bid' and black for 'bit'. The number within each circle indicates the identification percentage for the most frequently identified category, with standard error in parentheses.

A repeated-measure ANOVA evaluating the effect of native language (group) on the use of vowel duration and voicing duration reports a significant effect of group on the use of voicing duration, $F(2.09, 140) = 4.67$, $p < .01$. However, the interaction between group and vowel duration was not significant ($F(3.15, 138) = 1.75$, p = n.s.); nor was the

thee-way interaction between group, vowel duration, and voicing duration ($F(12.45, 128) = .938$, p = n.s.), indicating that Japanese speakers used vowel duration in a way comparable to native English speakers. This suggests that Japanese speakers are indeed able to apply their sensitivity to vowel duration to a novel phonological contrast, in this case the English voicing contrast in coda position. These results are interpreted as supporting prediction 2 of the BLIP model, according to which L2 learners can use the neural mapping of a given acoustic contrast to perceive a different phonological contrast using the same acoustic cue.

Multiple regressions were performed on the English and Japanese data (i.e. separately) to evaluate the relative use of each independent variable—voicing duration and vowel duration—for categorization of the coda contrast. The effect of voicing duration and vowel duration in this experiment predicts 48% of the identification responses for English speakers ($R^2 = .479$). As summarized in Table 4–9, the effect of voicing duration ($\beta = .530$, $p < .001$) was slightly greater in the English data than the effect of vowel duration ($\beta = .446$, $p < .001$). However, analyses of individual data using the ratio method described in Experiment II (again, using a cut-off point of .5) indicate that most English-speaking participants (N = 10) had a bias towards using both cues, while some had a bias towards using mainly changes in voicing duration (N = 7) or changes in vowel duration (N = 5). Two English-speaking participants did not demonstrate a clear bias towards either of these cues. These results contrast with those obtained in Experiment I, where only two patterns of identification emerged in the English group: a bias towards using formants (which was the pattern used by most English participants), or a bias towards using both formants and vowel duration for

identification of the 'beat' and 'bit' contrast. Thus, the results of Experiment III indicate that unlike in the first experiment, native English speakers can freely use both acoustic cues together or separately for identification of the coda voicing contrast. This finding suggests that these cues are more clearly processed competitively for identification of the same feature. The response time results, provided below, also concur with this conclusion.

Table 4–9 Regression results for English speakers (Experiment III)

	B	SE B	β
Constant	-.307	.039	
Voicing duration	.194	.011	.530*
Vowel duration	.107	.007	.446*

Note: Model $R^2 = .479$, *$p < .001$

The combined effect of voicing duration and vowel duration explains about 45% of the Japanese identification results ($R^2 = .446$). Unlike the participants in the English group, Japanese participants appear to use vowel duration ($\beta = .537$, $p < .001$) to a greater extent than voicing duration ($\beta = .397$, $p < .001$). An analysis of the individual Japanese data confirms that most Japanese participants used vowel duration as the most important cue to make their categorical judgments (N = 9). However, some listeners had a bias towards using voicing duration (N = 6) or towards using both cues (N = 5); and four of the Japanese participants did not demonstrate a clear bias towards the use of either cue. The response time results below further suggest that voicing duration and vowel duration

are not processed as competitively by most native Japanese speakers as they are by native English speakers.

Table 4–10 Regression results for Japanese speakers (Experiment III)

	B	SE B	β
Constant	-.261	.040	
Voicing duration	.144	.011	.397*
Vowel duration	.128	.007	.537*

Note: Model $R^2 = .446$, *$p < .001$

As can be seen in Figure 4–20 below, unlike the identification pattern described above, the response time pattern of English and Japanese speakers in Experiment III differs considerably. First, we can see that the response time of English speakers is relatively constant across changes in vowel duration when there is no voicing during the stop closure or when the voicing is very short (i.e. 20ms). Their response time is generally shorter when there is no voicing present (0ms), suggesting that this cue is sufficient and straightforward for English speakers to categorize coda consonants in terms of the voicing contrast. Second, the RT of English speakers is noticeably higher when a short vowel duration is combined with a voicing of 40 to 60ms (top left area on the English graph). However, the RT for the tokens with long voicing duration decreases as the vowel is lengthened. Segmented regression analyses confirm that vowel duration has a significant effect on RT for tokens with voicing duration between 40 to 60ms only ($\beta = -.305$, $p < .001$). In addition, segmented regressions on the split data into short and long vowels confirm a significant negative effect of voicing duration when the vowel is

relatively short (i.e. ≤ 160ms) (β = .252, p < .001). No other significant effects using segmented regression analyses were found in the English data. These results are interpreted as confirming that voicing duration and vowel duration for identification of the coda stop voicing contrast in English are both important cues that are processed *competitively* by native English speakers, as suggested by the analyses of individual data reported above, which indicated that many of the English participants in this study (i.e. 10) had a clear bias towards using both cues for their categorical judgments.

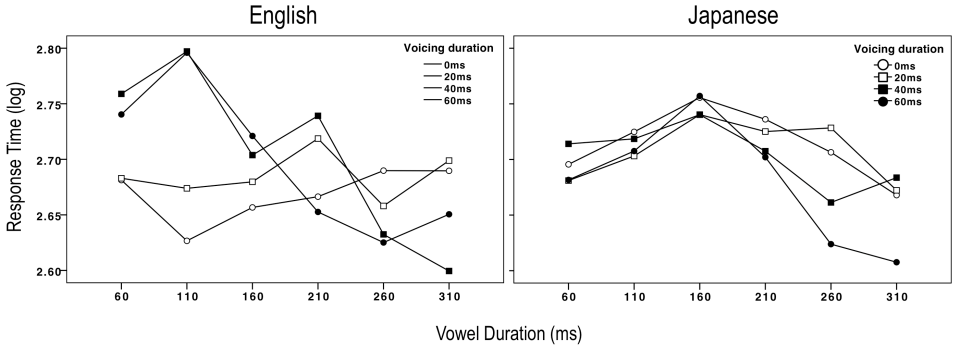

Figure 4–20 Average (log-transformed) response times for the English and Japanese group for each of the 24 tokens in Experiment III.

As for the Japanese RT data, we can see that unlike English speakers, Japanese speakers do not seem to process vowel duration and voicing duration competitively, since their response time is relatively short when a short vowel is combined with a long voicing duration (left area of the Japanese graph). While segmented regression analyses on the data split into short and long vowels reveal an effect of *voicing duration* in the English data, the same procedure applied to the Japanese data reveal a small significant effect of *vowel duration* (β = .125, p < .05). Hence, Japanese speakers appear to use either one of the manipulated cues to make their categorical judgments, rather than processing these

cues competitively like English speakers. This is again consistent with the analyses of individual data reported above that indicated that only five Japanese participants demonstrated a clear bias towards using both cues. There was also a small significant effect of vowel duration on tokens with voicing duration between 40 and 60ms ($\beta = -.134$, $p < .05$). No other results of the segmented analyses were significant. These results, combined with the identification results described previously, suggest that the primary cue used by native Japanese speakers in this experiment was vowel duration. Nevertheless, the identification results did show that Japanese speakers could also use voicing duration to make their judgments about the voicing status of the coda stop consonants in English.

To sum up, native English speakers use both vowel duration and voicing duration to perceive the coda voicing contrast, as in 'bit' vs. 'bid'. Moreover, these cues are processed competitively by native English speakers; voicing duration appears to have a slightly stronger effect on their categorical decision than vowel duration, at least in this experiment. Conversely, although Japanese speakers are able to carry their sensitivity to periodicity for voicing contrast in onset position to a non-native coda voicing contrast, as suggested by prediction 1 of the BLIP model, they generally favor the use of vowel duration. Crucially, the fact that Japanese speakers are able to use their sensitivity to vowel duration—associated with a vowel contrast in their L1—to perceive a non-native contrast associated with a different phonological contrast in the L2 (i.e. a coda voicing contrast), provides support for prediction 2 of the BLIP model. Thus, these results are interpreted as demonstrating that L2 learners are able to capitalize on the neural mapping in their L1 to perceive novel L2 (phonological) contrasts at the neural mapping level.

4.6 Experiment IV

In the previous experiment, it was shown that Japanese speakers could use both vowel duration and periodicity for perception of the coda voicing contrast in English, presumably because both cues are used in their L1 for other phonological contrasts. Experiment IV evaluates the perception of the 'bit' vs. 'bid' contrast by native Canadian French speakers to verify that the ability to use these cues, particularly vowel duration, is indeed language-specific. According to prediction 4 of the BLIP model, if vowel duration is not used contrastively in the L1 of the speakers for any phonological or context-bound allophonic contrasts, these speakers should generally be unable to use this cue to categorize L2 contrasts. This hypothesis is tested with Canadian French speakers, in whose L1 vowel duration is generally ignored for phonemic contrasts. Importantly, the results of Experiment II revealed that Canadian French speakers are able to perceive differences in vowel duration. According to prediction 4 of the BLIP model, acoustic cues ignored in the L1 of the L2 learners may be perceived by these speakers more easily than acoustic contrasts processed by overlapping neural maps (e.g. Japanese speakers are generally unable to use spectral differences to distinguish the high front English vowels as demonstrated in Experiment I). Hence, the BLIP model does not predict that Canadian French speakers are totally unable to perceive any change in vowel duration, but simply that they may rely on more familiar cues for categorizing non-native speech contrasts if such cues are available.

The cues manipulated in Experiment IV for the 'bit' versus 'bid' contrast are vowel duration and duration of periodicity during the coda stop closure. French contrasts coda

consonants in terms of voicing, distinguishing a phrase like *il vente* [ɪl vɑ̃t] 'it's windy' from *ils vendent* [ɪl vɑ̃d] 'they sell'. However, French speakers are not known to use vowel duration for this contrast, presumably using other cues, such as the presence of periodicity and difference in formant transitions instead. Hence, this experiment verifies that Canadian French speakers will use periodicity as their primary cue for identification of the coda voicing contrast in English, rather than vowel duration.

4.6.1 Methodology

Participants

The Canadian French participants for Experiment IV were the same as for Experiment II. The English and Japanese data used for comparison are those reported in Experiment III above.

Stimuli

The stimuli were the same as in Experiment III.

Procedure

The procedure was the same as in Experiment III.

4.6.2 Results and discussion

The overall identification percentages averaged across stimuli and trials for each Canadian French participant, reported in Figure 4–21, indicate that Canadian French speakers generally labeled between 40% and 70% of the tokens as 'bid' and between 30% and 60% as 'bit'. These identification percentages are similar to those reported in

Experiment III for English speakers (who identified between 40% and 80% of the tokens as 'bid').

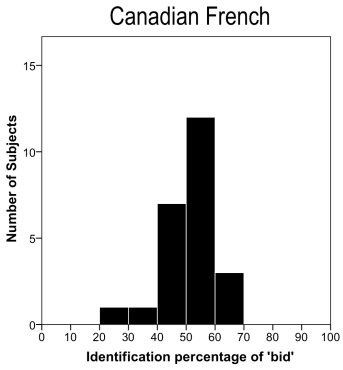

Figure 4–21 Histogram of the aggregated identification percentage (as 'bid') for individual Canadian French subjects.

The averaged identification pattern of Canadian French speakers, however, differs significantly from that of the English speakers reported in the previous experiment, as shown in Figure 4–22 below. Unlike English speakers, who generally use both voicing duration and vowel duration as cues to categorize the 'bit' and 'bid' tokens, Canadian French speakers appear to rely overwhelmingly on changes in voicing duration, and generally ignore changes in vowel duration.

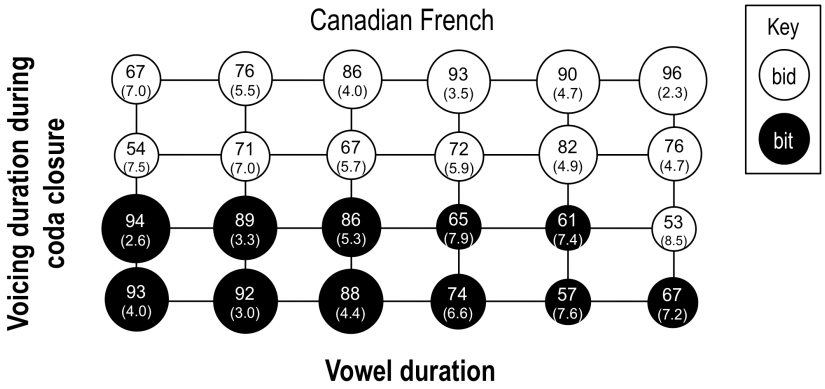

Figure 4–22 Averaged identification of tokens as either 'bit' or 'bid' across Canadian French speakers. The size of each circle corresponds to its identification frequency: large circles indicate higher identification percentages. The shading of the circle indicates the most frequently identified category: white for 'bid' and black for 'bit'. The number within each circle indicates the identification percentage for the most frequently identified category with standard error in parentheses.

A repeated-measure ANOVA confirms a significant effect of native language (group) on the use of vowel duration, $F(3.70, 138) = 4.82$, $p < .001$, as well as on the use of voicing duration, $F(2.07, 140) = 3.58$, $p < .05$. No effect was found on the three-way interaction between group, vowel duration, and voicing duration, $F(12.85, 128) = 1.03$, p= n.s.

Multiple regressions performed on the Canadian data alone to evaluate the use of each independent variable for categorization of the coda voicing contrast indicate that the effect of vowel duration and voicing duration predicts 49% of the results for the Canadian French speakers ($R^2 = .487$), which is roughly equivalent to the overall effect of these

cues reported for the English (48%) and Japanese (45%) group. However, as summarized in Table 4–11, the effect of voicing duration was greater in the Canadian French group ($\beta = .639$, $p < .001$) than in the English ($\beta = .530$, $p < .001$) and Japanese ($\beta = .397$, $p < .001$) groups. By constrast, the effect of vowel duration was lower in the Canadian French group ($\beta = .280$, $p < .001$) than in the English ($\beta = .446$, $p < .001$) or Japanese ($\beta = .537$, $p < .001$) groups. Hence, as anticipated by prediction 4 of the BLIP model, Canadian French speakers appear to use voicing duration as their primary cue for identification of the English coda voicing contrast rather than vowel duration, presumably because vowel duration is not used for any speech contrasts in their L1. Analyses of individual data using the bias ratio method described in Experiment II confirms that most French speakers in this experiment (N = 15) had a clear bias towards using only voicing duration for their identification judgments. Only a few French speakers had a bias towards using only vowel duration (N = 3), while some of them had a bias towards using both cues (N = 4), and two Canadian French participants did not demonstrate a clear bias towards using either cue. Interestingly, five of the seven participants who used vowel duration in Experiment IV also demonstrated a bias towards using vowel duration in Experiment II. Hence, the Canadian French speakers who exhibited a bias towards the use of vowel duration may have figured out that vowel duration is an important cue in English, and used this cue to perceive both vowel and coda consonant contrasts in Experiments II and IV. Conversely, none of 12 Canadian French participants who exhibited a bias towards using only spectral changes in Experiment II for the identification of the vowel contrast had a bias towards using vowel duration in Experiment IV. These results provide support for prediction 4 of the BLIP model by

confirming that most Canadian French speakers are generally insensitive to changes in vowel duration in perceiving L2 speech contrasts because French does not use vowel duration for any phonological or context-bound allophonic contrasts.

Table 4–11 Regression results for Canadian French speakers (Experiment IV)

	B	SE B	β
Constant	-.310	.039	
Voicing duration	.233	.011	.639*
Vowel duration	.067	.007	.280*

Note: Model $R^2 = .487$, *p < .001

The log-transformed RT results of Canadian French speakers are shown in Figure 4–23 below. Segmented regressions performed on the Canadian French log RT data reveal a small negative effect of vowel duration on the tokens, with voicing duration equal to or above 40ms ($\beta = -.177$, $p < .01$), similar to the effect found in the Japanese data ($\beta = -.134$, $p < .05$), but weaker than that found in the English data ($\beta = -.305$, $p < .001$). Additionally, there was a small effect of voicing duration on tokens with vowel duration equal to or above 210ms (right side of the graph) ($\beta = -.143$, $p < .05$). This result means that Canadian French speakers found it easier to categorize tokens with long vowels combined with a long voicing duration during the final stop closure than tokens with short vowels, with or without voicing during the coda closure. These results contrast with that of the English speakers, who clearly had a harder time processing tokens with short vowels when these tokens also had a long voicing duration, presumably because native English speakers generally process these cues competitively. Hence, taken

together, the Canadian French and Japanese results suggest that L2 learners may have difficulties in processing many cues at the same time (i.e. in processing cues competitively in the L2) and may instead use only one cue at a time to make their categorical judgments.

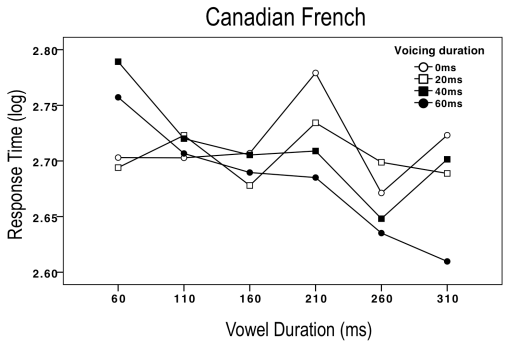

Figure 4–23 Average (log-transformed) response times for the Canadian French group for each of the 24 tokens in Experiment IV.

In sum, in this experiment, Canadian French speakers overwhelmingly favor the use of changes in voicing duration, rather than changes in vowel duration, for categorization of the English coda voicing contrast in the words 'bid' versus 'bit'. Results from Experiments I through IV appear to confirm that L2 learners' ability to use an acoustic cue for perception of an L2 contrast is language-specific, and directly dependent on whether this cue is used contrastively in the L1, either for a phonological or context-bound allophonic contrast.

4.7 Summary of the predictions of the BLIP model and supporting experiments

The BLIP model predicts that the neural mapping of acoustic components in the L1 can either facilitate (predictions 1, 2 & 3) or impede (prediction 4 & 5) the perception and

acquisition of non-native speech contrasts. The five predictions of the BLIP model were presented in Figure 4–1, but are repeated below in Figure 4–24 for convenience.

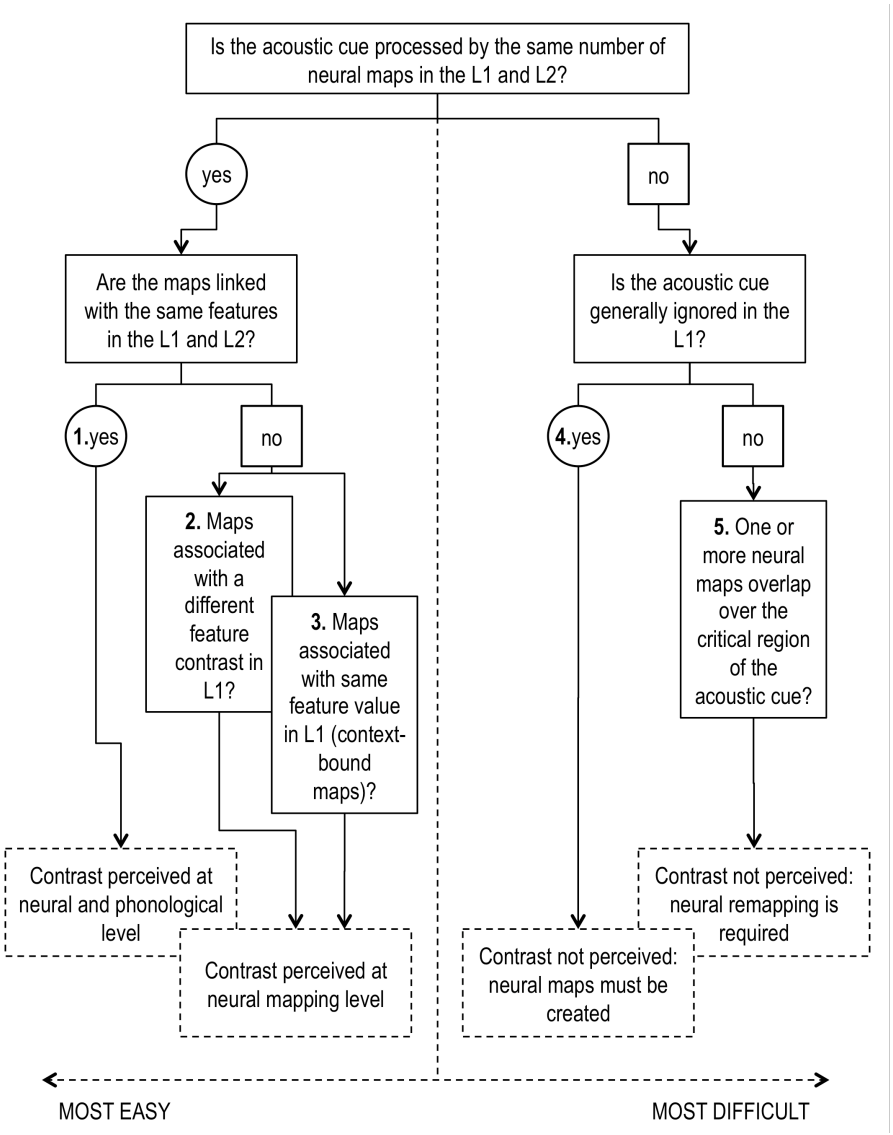

Figure 4–24 Predictions of the BLIP model for perception and acquisition of non-native speech contrasts.

Prediction 1 stipulates that acoustic cues that are processed by the same number of neural maps in the L1 and L2, and which are also associated with the same phonological feature at the phonological level in both languages, should allow L2

learners to perceive any non-native contrasts based on these cues. Experiments III and IV provided evidence for this prediction. Experiment III demonstrated that Japanese speakers could perceive the coda voicing contrast in English based on the presence or absence of periodicity during the final stop closure—even though Japanese lacks a voicing contrast in coda position—presumably because Japanese speakers are sensitive to speech contrasts involving the activation of the AM map (which fires to the presence of a periodic waveform) in processing voiced versus voiceless stops and fricatives in onset position. Experiment IV demonstrated that Canadian French speakers who are sensitive to the voicing contrast, involving the presence or absence of periodicity in both onset and coda positions in their L1, could use this cue to contrast the coda voicing contrast in English.

Prediction 2 of the BLIP model posits that if an L2 contrast is processed by the same number of neural maps in the L1 as in the L2, but that these maps are associated with a different feature at the phonological level in the two languages, L2 learners should still be able to perceive this contrast at least at the neural mapping level—provided, that is, that the task type and testing conditions do not specifically require lexical access. Experiment III provided support for this prediction by showing that Japanese speakers could apply their sensitivity to vowel duration—generally associated with a vowel length phonological contrast in their L1—to perceive the coda voicing contrast in English—associated with a voicing contrast in the L2.

Prediction 3 of the BLIP model speculates that if an L2 contrast is processed by the same number of neural maps in the L1 as in the L2, but that the L1 maps are not contrastive at the phonological level because they process context-bound allophones in

the L1, L2 learners should still be able to perceive the L2 contrast based on these maps, provided that the type of task and testing conditions only tap into the neural mapping level of processing. Experiment II provided support for this prediction by demonstrating that most Canadian French speakers were able, without any instruction about the relevant acoustic cue distinguishing the English high front vowels, to use spectral changes (i.e. changes in F1 and F2) to contrast the English vowel contrast. This result is presumably due to Canadian French speakers' having context-bound allophones, corresponding roughly to the English phonemic vowel categories. According to the BLIP model, these allophones are processed by separate neural maps, even though these maps are associated with the same phonologically relevant vowel quality (i.e. a feature relevant for meaning contrast) at the phonological level in French.

Prediction 4 of the BLIP model posits that if an L2 contrast is based on an acoustic cue that is not used in the listeners' L1 to contrast any native speech sounds, the L2 acoustic contrast may generally be ignored even at the neural mapping level, especially if the L2 learners are unaware that this cue is important for the L2 contrasts and if more familiar cues are available. The results of Experiment IV are consistent with this hypothesis: most Canadian French speakers ignored vowel duration for categorization of the coda voicing contrast in English and relied instead on the presence or absence of periodicity, presumably because vowel duration is not used for any segmental contrasts in French. This finding does not suggest that Canadian French speakers are totally insensitive to changes in vowel duration, but simply that these speakers have not developed specific neural maps to process this cue categorically. Hence, neural organization for the processing of vowel duration by speakers of languages

that do not use this cue contrastively is expected to remain generally neutral, that is, to correspond to their initial organization at birth. This means that neural maps based on this cue must be created for efficient processing of speech contrasts based on this cue.

Prediction 5 of the BLIP model speculates that if an L2 contrast is based on an acoustic cue that is used in the learner's L1, but that two or more categories in the L2 are mapped by a single overlapping map in their L1, this contrast will be most difficult to perceive. Results of Experiment I provide support for this prediction, by showing that none of the native Japanese speakers in this experiment were able to use spectral changes (i.e. information pertaining to vowel quality) to contrast the high front English vowel contrast, but relied instead on the use of vowel duration.

Hence, predictions 1, 2 and 3 of the BLIP model suggest that non-native speech contrasts that fall under one of these categories should be relatively easy to perceive and acquire by non-native speakers, since the neural maps necessary to distinguish the relevant L2 categories are already in place and used for native contrasts. Conversely, predictions 4 and 5 of the BLIP model suggest that L2 acoustic contrasts that fall under these latter categories may not be perceived and acquired as easily by non-native speakers, and may require specific instruction and intensive training, since the neural maps necessary to perceive the relevant L2 acoustic contrasts are not already in place. That is, the processing of L1 contrasts does not always interfere with the perception and acquisition of L2 contrasts, but it may, rather, facilitate their acquisition in many cases.

An important conclusion following from this theoretical approach and from the experimental results reported above, is that speech categories first emerge at the neural mapping level. Incidentally, the most problematic difficulties encountered by L2 learners

with non-native speech contrasts are predicted to be with acoustic contrasts that are not mapped contrastively in the learners' L1 for any phonemic or context-bound allophonic contrasts. At this stage in the development of the BLIP model, I have not evaluated the degree of difficulty related to the creation of novel phonological contrasts (i.e. with the creation of features absent from the learners' L1). Although the phonological model proposed by Brown (1997; 1998; 2000) predicts that L2 contrasts based on features present in the learner's L1 can be perceived, whereas contrasts based on features absent from the learner's L1 cannot be perceived, these predictions were not completely borne out in a previous investigation by Grenon (2008), who evaluated these hypotheses with a wide range of L2 contrasts. Here, this discrepancy is argued to stem from the fact that phonologically based models do not take into consideration the neural mapping level of processing, where the speech categories first emerge.

Another difference between Brown's model and the BLIP model—where both models are designed to account for *perception* of speech contrasts and their potential acquisition—is that Brown's model takes the strong stance that speech contrasts based on features that are not exploited in the learner's native language are impossible to acquire. The BLIP model does not support this position. On the contrary, since synaptic connections can be altered throughout the lifespan, the BLIP model suggests that there is no reason for L2 learners to be unable to acquire new speech categories by creating new contrastive maps. To which extent the newly created maps can be used with the same efficiency as those developed by native speakers early in life—that is, when the infant's brain has more synaptic connections available—is a matter yet to be investigated. Nevertheless, a study reported by Grenon (2006) suggests that at least in some

experimental settings, non-native speakers can perform as well as native speakers on a hitherto novel speech contrast: Most Japanese speakers in her experiment were able to discriminate the /s-θ/ and /z-ð/ contrasts in English in a way comparable to the native English controls, despite the fact that these contrasts presumably employ a phonological feature ([distributed]) that is absent from Japanese phonology (as well as neural maps that are not contrastive in Japanese).

The next section departs from the implications of the BLIP model for the study of L2 perception and discusses the additional contributions of the model. For this purpose, the general discussion presents a succinct comparison of the BLIP model with the models previously presented in this thesis.

4.8 General discussion

In this work, I introduced a variety of neural-based, psycholinguistic, and L2 models of speech processing. Neural-based models suggest that speech categories are embedded in the neurology through the formation of neural maps designed to capture acoustic contrasts. Psycholinguistic models have been proposed to account for behavioral discrepancies in the perception of speech contrasts depending on the task type and testing conditions used in experimental settings. Finally, L2 models were generally designed as a tool to study the perception and possible acquisition of L2 contrasts. All these models serve their own purpose, and their contribution to that purpose is not disputed here.

However, it seems a worthwhile endeavor to merge the various models to create a unified model of speech processing that can capture psycholinguistic behavior related to the perception of linguistic units, while taking into consideration the possible constraints

imposed by the neurology in processing acoustic speech stimuli. This multidisciplinary approach may have implications for the study of both first and second language perception. Since each field (neurology, psychology, linguistics, and phonetics) looks at speech processing from a different angle and uses a different jargon, building a unified model is not a small task. This dissertation makes such an effort, the result of which is the BLIP model. In this section, I review how the BLIP model builds on or integrates concepts from the various models described throughout this work. At the end of the next chapter, I exemplify how the BLIP model can potentially serve as a useful framework for psycholinguistic and neurolinguistic research on L1 and L2 by discussing the future directions intended with this model.

The neuronal model of vowel normalization proposed by Sussman (1986), the neural-based model of locus equations suggested by Sussman and colleagues (Sussman 1989, 1999, 2002; Sussman et al. 1991), and the neural-based model of speech perception put forward by Guenther and colleagues (Guenther & Bohland 2002; Guenther et al. 1999, 2004), build on the assumption that speech categories are embedded into neural organization through the formation of neural maps. This proposal is applied to the categorical processing of vowels in Sussman's normalization model; to the processing of stop place of articulation in Sussman and colleagues' locus equation model; and to the processing of /r/ and /l/ in English, as compared to the processing of the Japanese flap, in Guenther and Bohland's model. The BLIP model incorporates the general hypotheses related to the processing of stop place of articulation and vowel quality as suggested by the above-mentioned neural-based models. However, the BLIP model further extends these hypotheses to model the processing of voicing contrasts (for stops and fricatives),

fricative place of articulation contrasts, vowel duration contrasts, lexical stress, and lexical tone contrasts.

The locus equations model proposed by Sussman and colleagues (Sussman 1989, 1999, 2002; Sussman et al. 1991) speculates that noise burst, F2 onset, and F2 value at mid-vowel may be processed in stages by the neurology for identification of stop place of articulation. Neurons at the different stages project to the possible outcomes, and only the outcome that receives most "support" or projections is analyzed as the activated abstract category. This process may be seen as a kind of "competitive" (or alternatively, "complementary") processing, where many cues that contribute to identification of the same abstract characteristic or linguistic unit (i.e. phonological feature, segment, mora, or syllable) are processed by different types of neurons for identification of the proper abstract feature. Based on this general idea, the BLIP model proposes three possible ways in which two acoustic cues may be processed by the neurology for identification of a linguistic unit: as discussed at length in the previous chapter, two cues may be processed competitively, additively, or connectively. Additionally, within the BLIP approach to speech perception, it is argued that the way acoustic cues are processed may impact the processing of non-native contrasts that feature one or more of these cues. For instance, it is argued that the difficulties encountered by English speakers with the perception of lexical tone contrasts in Chinese do not stem from English speakers' inability to perceive pitch variations, but from the different way pitch is processed by speakers of each language (see section 3.2.4 for more detail).

Sussman and colleagues' model of locus equations also implies that the brain goes through steps or stages during speech processing. Psycholinguistic research further

documents that the task type or testing conditions used in experimental settings appear to tap into one or more of these stages. Accordingly, psycholinguistic models, such as PRIMIR (Werker & Curtin 2005), generally propose two distinct stages or levels of processing (besides lexical encoding) to account for behavioral results. One stage involves the processing of categorical acoustic information (referred to in PRIMIR as the general perceptual plane), and another stage the processing of abstract phonemic contrasts (referred to in PRIMIR as the phonemic plane). The BLIP model proposes two levels of processing as well. These levels are *de facto* similar to the levels proposed in the PRIMIR model, in that they capture the same behavioral facts. Unlike any comparable psycholinguistic models, however, the levels posited by the BLIP model, particularly the neural mapping level (which captures categorical acoustic information), are informed by neurological processing, as documented in neural (non-human animal) experiments. For instance, the BLIP model posits (in-line with Sussman's model of vowel normalization) that abstract vowel categories are derived from the combinatory processing of at least the first two formants at the neural mapping level by combination-sensitive neurons. As a result, in Sussman's model and in the BLIP model, the only possible abstract representation of vowels in the neurology based on the processing of F1 and F2 corresponds to a vowel quality, rather than to an articulatory-based characteristic such as tongue height, frontness, or backness.[62] In contrast, PRIMIR is not explicit about how the

[62] Although, as mentioned in the previous chapter, this does not prevent the listener from being sensitive to variations in F1 and F2, which generally correspond to variations in tongue height, frontness, and backness. The implication here, though arguable, is that phonological processes (such as vowel harmony) are primarily production phenomena, not perceptual ones, since these rules are not, under the current approach, specifically encoded at the neural mapping level (phonological rules may, however, be somehow encoded

processing of speech contrasts is achieved at the general perceptual plane. That is, it does not provide a framework to make specific predictions about the (neural) processing of acoustic contrasts, and remains vague about how this processing may be related to phonemic representations. In this sense, the BLIP model is compatible with PRIMIR, but offers the advantage of being grounded in what we know about neural processing. As a consequence, it is able to make specific predictions about the relative ease of processing of different acoustic cues, based on how they are (speculated to be) processed by the neurology.

Guenther and colleagues' (Guenther & Bohland 2002; Guenther et al. 1999, 2004) neural-based perception model is founded on the inverted magnification factor hypothesis originally proposed by Bauer, Der and Herrmann (1996), according to which cell density activation decreases at categorical centers along the relevant acoustic dimension of a speech contrast. If few cells are activated during the perception of a given type of stimulus, the perceiver is presumably unable to discriminate that stimulus in detail or with high accuracy from stimuli around the same region of the input space. Guenther and colleagues argue that this hypothesis has crucial implications for L2 perception and acquisition: they demonstrate that Japanese speakers' difficulties in perceiving the English /r-l/ contrast can be explained as a decrease in cell density activation within the neural map used to process the Japanese flap along the F3 dimension, which roughly coincides with or overlaps the F3 boundary critical to the English /r/ and /l/ contrast. This

---

by or derived from lexical encoding at higher/subsequent levels of processing not yet considered in the current version of the BLIP model).

hypothesis was incorporated into the BLIP model and combined with the two levels of speech processing posited (neural mapping and phonological levels), to yield a set of five specific predictions about the perception of non-native contrasts. Each of these predictions was tested and supported in four behavioral experiments reported and discussed in this chapter.

In addition to general models of speech processing, a number of models have been proposed dealing specifically with L2 speech perception or acquisition, such as PAM (Best 1993, 1994, 1995; Best & McRoberts 2003; Best, McRoberts, & Goodell 2001; Best & Strange 1992) and SLM (Flege 1992a, 1992b, 1993, 1995). These models also make predictions about the perception or possible acquisition of L2 contrasts and may yield similar predictions as the BLIP model. The predictions of PAM and SLM have been tested numerous times with varying degrees of success, whereas the predictions of the BLIP model still await further testing. Nevertheless, the BLIP model holds three possible advantages over previous L2 models. First, the BLIP model does not require the evaluation of cross-linguistic perceptual similarity to make predictions (which is required by previous models, see 4.1 for details), and consequently removes a time-consuming step for L2 researchers. Second, the BLIP model's predictions are not based on absolute acoustic values for the comparison of L1 and L2 contrasts, but rather, on the number of contrasts (i.e. neural maps) within the region of interest along a given acoustic dimension. This feature is expected to yield more accurate predictions (since our sense of perception is relative, rather than absolute). Third, the BLIP model provides a framework to evaluate and identify the source of learners' difficulties with a given L2 contrast, which can in turn be extrapolated to other non-native contrasts in the same or other languages.

In sum, the multidisciplinary approach adopted by the BLIP model draws on a number of different fields of research and represents an innovative approach to the study of L2 perception and acquisition, offering some significant advantages over previously posited L2 models. While the experiments in this work have focused on L2 perception, the BLIP model is not restricted to L2 research, since it can just as easily be used for cross-linguistic comparisons of the processing of acoustic contrasts, or used to evaluate the development of acoustic and phonemic categories during L1 development.

Chapter Five: Conclusion

5.1 Summary of the model and its contribution to the field

The focus of the current work was the articulation of a unified model of speech processing that builds on previous neural-based, psycholinguistic and L2 models. The proposed model is founded upon the assumptions of neural-based models, according to which speech categories are encoded as neural maps by the neurology. The model uses this framework to account for documented psycholinguistic behavior related to the perception of speech contrasts. This unified model of speech processing may have applications to the study of language perception, as exemplified by a set of specific predictions about L2 perception derived from the model, which were supported by the results of four behavioral experiments. Hence, the proposed model was meant to serve as a basis for linguistic analyses of speech sound processing by bridging the gap between neural processing and conventional psycholinguistic descriptions. This model is called the Bi-Level Input Processing model (or BLIP) to emphasize the fact that human speech sound processing is best captured by positing two levels of speech processing.

Based on previous neuroethology experiments (e.g. Suga 2006) reported in chapter 2, it was shown that there appears to be a direct correlation between animals' (human and non-human species) perception of sound contrasts and their sensitivity to acoustic components and neural responses to these components. Accordingly, the BLIP model posits that the perception of speech sounds by humans corresponds to the processing of a limited number of acoustic components by neural maps tuned to these components, where each map corresponds to a contrastive speech category along the

relevant acoustic dimension in the listener's native language. One of the most innovative and valuable aspects of the BLIP model is its neural grounding of specific linguistic concepts, mainly features and allophones. This characteristic empowers the model to serve as a practical and sensible framework for the study of speech perception and acquisition, providing researchers with a concrete and testable way of identifying the source of L2 learners' difficulties with non-native contrasts. This approach prevents the reliance on the concept of perceptual similarity, which cannot be unambiguously defined.

Specifically, as discussed in chapters 2 and 3, the BLIP model speculates that the development of relevant speech categories in an L1, whether corresponding to phonemic or context-bound allophonic contrasts, first emerges at the neural mapping level, where each distinct category is processed by a separate neural map tuned to a specific acoustic component (e.g. CF-constant frequency components such as formants; FM-frequency-modulated components such as formant transitions; NB-burst noise components; AM-amplitude-modulated components; or relevant combinations of the foregoing). In L1 development, the neural maps emerge based on the statistical distribution of these components in the language to which an infant is exposed. In addition, it is speculated that each neural map is associated with a phonological feature used to contrast meaning in the language. In most cases, more than one neural map is associated with the same feature, since more than one acoustic cue can be used to identify a given speech contrast. Accordingly, the BLIP model posits that these cues may be processed in three different ways: additively, connectively, or competitively. Acoustic cues processed *additively* are processed by different types of neurons (hence, by different neural maps, such as an AM map and an FM map) and associated with different features (e.g. |voice| and |bilabial|).

Acoustic cues processed *connectively* are processed by the same group of neurons (hence, by only one neural map, such as a CF-CF map that captures F1 and F2 values at the same time) and associated with only one feature (e.g. |i|). Finally, acoustic cues processed *competitively* are processed by different neural maps (e.g. a NB map processes information pertaining to the noise burst and a FM map processes information related to the formant transition), while contributing to the identification of the same feature (e.g. |bilabial|). The number of neural maps used to process spectral and temporal cues, as well as the type of processing used to identify the speech categories associated with these cues, may differ from one language to another, accounting, incidentally, for cross-linguistic variation in language processing. The neural account of speech processing proposed by the BLIP model was exemplified in chapter 3 with the processing of fricative, vowel, stop, and suprasegmental contrasts in English, French, Japanese, and Mandarin.

In addition to accounting for cross-linguistic differences in L1 speech processing, the BLIP model has crucial implications for the study of L2 perception and acquisition. In Chapter 4, I described and discussed the five predictions for L2 perception derived from the model, along with four perceptual experiments conducted with native (North American) English, Japanese, and Canadian (Québécois) French speakers that support these predictions. To summarize, the BLIP model predicts that non-native acoustic contrasts that are not mapped in the learners' L1 for any native contrasts should be the most difficult to perceive and acquire, whereas acoustic contrasts that are already mapped in the learners' L1 should be more easily perceived and acquired. That is, if, as assumed in the BLIP model, there is a correspondence between speech contrasts and neural maps,

language learners should be able to capitalize on their sensitivity to acoustic contrasts (whether phonemic or context-bound allophones) to perceive novel speech sounds, even if these sounds are generally neutralized at the phonological level in the learners' L1. The BLIP also posits that in cases where L2 learners lack the proper neural maps to perceive the novel speech categories, neural mapping or remapping is possible with the proper training paradigm (this paradigm is described in the next section). That is, the BLIP model argues that there is no critical period for the acquisition of novel speech categories in perception, a position supported by the author's previous research (Grenon, 2006).

In short, the BLIP model was designed to fill the gap between neural processing and language processing. Moreover, while this dissertation focuses primarily on the application of the BLIP model for the study of L2 perception, the model is argued to have significant implications for the study of both L1 and L2 perception and acquisition. The model is still under development; further experimentation is necessary to provide additional support for its assumptions and proposals, and to refine the hypotheses put forward in this work, as discussed further in section 5.3. In any case, it is my hope that this work will contribute to assisting speech scientists in envisioning speech processing from a wider and integrative perspective, and that it will serve as a convenient framework to conduct phonetic, psycholinguistic, and neurolinguistic experiments designed to deepen our understanding of language processing. Further, I believe that the results of such experiments will have important implications in the domains of education (e.g. L1 development or L2 acquisition), health care (e.g. speech pathology), and speech technology (e.g. voice recognition systems).

5.2 Implications for second language education

Throughout the different chapters of this thesis, I have reported experimental studies pointing to the effect of different types of training (e.g. categorical versus discrimination training) on neural organization; to the possible effects of listeners' expectations, level of attention, and awareness about the L2 contrasts on their ability to acquire a novel contrast; and to the role of auto-associated patterns for speech perception in fluent conversation. In this section, I discuss how these factors may affect the perception and acquisition of non-native speech categories; and accordingly, how these factors may be used to optimize the time and effort involved in teaching or learning these categories.

Predictions 4 and 5 of the BLIP model posit that L2 categories that are not mapped contrastively in learners' L1 are difficult to perceive and acquire because the neural maps necessary to distinguish the L2 categories are not available. I argue here that these contrasts are still acquirable, provided that the type of training and the training conditions favor the creation of novel neural maps (prediction 4) or facilitate the reorganization of pre-existing maps (prediction 5). Based on the assumptions of the BLIP model and on previous results of training experiments, it is possible to speculate on the optimal training conditions for the creation or reorganization of neural maps by adult L2 learners, as discussed in the remainder of this section.[63]

To begin, it is worth highlighting that under the current approach, exposure to natural L2 speech is unlikely to be sufficient to trigger changes in adult L2 learners'

---

[63] Learners are also known to have different learning styles or perceptual preferences, mainly visual, aural, reading/writing or kinesthetic (Leite, Svinicki & Shi 2010). It remains unclear how these learning styles may impact specifically on neural reorganization for the perception of new sound contrasts. The assessment of this idea certainly deserves consideration, but is beyond the scope of the current work.

neural organization (i.e. for L2 contrasts that fall under predictions 4 and 5 only), especially if the learner is unaware that the L2 sounds are contrastive, or does not know which cue is crucial to discriminate the L2 sounds. Such exposure is likely inadequate for two reasons: (1) natural speech may not always contain an ideal statistical distribution or reliable cues that allow neural reorganization; and (2) the listeners' expectations may direct them to focus on the wrong contrastive cue, if they are able to pick up any cue at all (see Experiment I for an example of this phenomenon). As explained in the previous chapters, even though the speech input contains enough invariance to enable newborns to forge the speech categories required in the ambient language, adults have already constructed neural maps to process the categories relevant to their native language. Thus, unlike infants, adults are able to process another language by assimilating L2 categories into native categories. In cases where two L2 categories are assimilated to the same L1 category, this process may simply yield an increased number of perceived homophones in the L2 that must be inferred from contextual information. In practice, this process may be a fair compromise that allows adult L2 learners to understand the L2 to a workable extent, provided that: (a) their knowledge of the lexicon is sufficiently extensive (i.e. for advanced learners); and (b) there is a reasonably good one-to-one correspondence between L1 and L2 speech categories (i.e. not too many L2 categories are assimilated into fewer L1 categories). However, if some speech categories are not contrasted at the neurological level, some undesirable outcomes may result. First, the lack of contrastive neural organization may impact on production of these contrasts (assuming that production of distinctive feature contrasts is at least partly based on their encoding at the neural mapping and phonological levels). Second, learners may perceive more words as

homophones than exist in the language. That is, the lack of proper contrastive neural maps may slow down L2 processing and development by increasing the difficulty in understanding the language and by impeding L2 production. Hence, in some conditions (i.e. for L2 sounds under prediction 4 or 5 of the BLIP model), targeted training may be most appropriate and beneficial for adult L2 learners. The question is: what training paradigm optimizes the efficiency and time required for neural reorganization? Previous experiments point to a training paradigm that fits most of the requirements for optimal and efficient neural remapping, as discussed below.

Given that the training task may impact on neural organization by inducing either an increase (i.e. magnification factor) or a decrease (i.e. inverted magnification factor) in cell density activation around categorical centers (as suggested by Guenther et al. 1999), it appears essential to use a task that clearly compels L2 learners to perform categorical decisions about the target L2 speech contrast, such as a forced-choice identification task. For the most difficult contrasts (i.e. those that require neural remapping)—for instance, the distinction between the high front English vowels in terms of formant changes by native Japanese speakers—it may be appropriate to start with some kind of discrimination task (such as AX discrimination task). This strategy would enable L2 learners to perceive *a* difference along the relevant acoustic dimension by capitalizing on their ability to perceive small changes (i.e. just-noticeable-difference) between stimuli *within* the overlapping neural map used in their L1. The use of real words (i.e. minimal pairs or near-minimal pairs) has also been shown to provide better results than the use of nonsense words (Hayes-Harb 2007), presumably because it increases the learners'

awareness that the target speech contrast induces differences in meaning—that is, they are contrastive at the phonological level.

Various studies have tested the effectiveness of a high-variability forced-choice identification paradigm in improving perception of a difficult L2 contrast. For example, in this paradigm, non-native English speakers would listen to a series of pre-recorded words featuring the target sounds, such as /r/ and /l/, in various contexts (syllable onset, word-medial position, within a cluster, etc.), as pronounced by different English speakers. Participants might be presented with the word *light*, for instance, and they would then be asked to determine whether they heard the word *light* or *right*. This paradigm was shown to improve the perception—and whenever tested, also the production—of difficult non-native contrasts by L2 learners, for instance, of the English /r/-/l/ contrast by native Japanese speakers (Bradlow et al. 1997, 1999; Iverson, Hazan & Bannister 2005; Lively, Logan & Pisoni 1993; Logan, Lively & Pisoni 1991), of English vowels by Mandarin and Cantonese speakers (Wang & Munro 2004), and of Mandarin tones by American speakers (Wang et al. 1999). Under the current approach, we can assume that the use of this kind of identification task induces L2 learners to discriminate the target speech contrast by creating separate neural maps, since the task is categorical and uses real (meaningful) words. In addition, the use of multiple voices may help listeners to discard idiosyncratic acoustic differences and enable them to identify the most robust contrastive cue(s) across individuals, as previously suggested by Lively, Logan & Pisoni (1993).

Although the conclusions presented here are not new, they have never been shown to be consistent with what is known to date about neurological processing. The BLIP model fills this gap, providing a way to further improve the training paradigm by

identifying the exact source of difficulty of a given (L1) language group with specific L2 contrasts. For instance, the training paradigms tested by Iverson et al. (2005) to help native Japanese speakers to perceive the English /r/-/l/ contrast have shown that the manipulation of F2 (which was either set as constant, variable, or progressively variable across the training stimuli) did not significantly enhance Japanese speakers' perception of the English contrast relative to the simple high-variability paradigm used in previous experiments. The researchers had expected that manipulation of the F2 would force L2 learners to focus on variations in F3 (which was shown in a previous research by Iverson and Kuhl 1996 to be a critical cue for English speakers), but these expectations were not borne out by their experiments. However, according to the BLIP model, and consistent with the (overlapping map) hypothesis described by Guenther & Bohland (2002), the F3 (instead of F2 or in addition to F2) should be manipulated to facilitate acquisition of this contrast by native Japanese speakers, because the neural remapping must be done along the F3 dimension. That is, to optimize the neural reorganization of an overlapping map into two separate maps, the training paradigm should first present tokens that are far apart along the most reliable acoustic dimension (in this case, F3) to enable L2 learners to perceive this contrast more easily so they can more confidently build two categories based on this acoustic dimension. As the training progresses and neural reorganization starts to take place, the tokens can be manipulated, for instance, by progressively reducing the acoustic differences between them. In such a design, it seems important that the training paradigm be adaptive, to follow the listeners' progressive development of the separate neural maps, starting with the remapping of the map at its opposite extremities.

A training paradigm using manipulated cues has been shown to be particularly successful in the training of vowel contrasts. Mandarin speakers, for whom neither vowel duration nor vowel quality of the corresponding English high front vowels are contrastive, were able to improve their perception of the vowel contrast after training with natural and synthesized tokens of the vowels when duration was controlled to ensure these learners built the L2 categories based on formant changes rather than vowel duration (Wang & Munro 2004).

Importantly, intensive training, such as that described above, may also help to create auto-associated patterns of acoustic cues, which are necessary to easily perceive fluent speech (refer to section 3.3.3 for a discussion). These patterns help L2 learners to forge neural maps for many different cues and for the most reliable cues in the identification of a given word, enabling L2 learners to perceive some L2 sounds even when some of the cues are missing or unreliable (since they can rely on more than one cue).

Finally, the use of a high-variability training paradigm using manipulated cues (whenever appropriate), may not only be beneficial for the acquisition of segmental categories, but also for the acquisition of suprasegmental elements, such as lexical tones and speech rhythm. For instance, the rhythm pattern of native Japanese speakers of English was found to differ from that of native English speakers in their stressed-unstressed syllable ratio; Japanese speakers did not reduce unstressed syllables (i.e. as captured by vowel duration) to the same extent as native Canadian English speakers (Grenon & White 2008). Although Japanese speakers are sensitive to vowel duration, they may either neglect to apply this contrast to stressed versus unstressed syllables, or

they have their categorical boundary for the vowel contrast set at a duration longer than that typically used by native English speakers for this particular contrast. Hence, because suprasegmental elements such as rhythm are sometimes partly captured by segmental, moraic, or syllabic contrasts, a similar training paradigm as that described in this section could be used to teach some suprasegmental features. The use of the high-variability paradigm has already been tested and shown to be beneficial for training American speakers to perceive Mandarin tones (Wang et al. 1999).

To conclude this section, the BLIP model provides a framework to help identify the exact source of difficulty of L2 learners with non-native speech contrasts and the relative degree of difficulty related to the acquisition of these contrasts. In the case of speech categories falling under predictions 1, 2 and 3 of the BLIP model, L2 learners already possess the neural maps necessary to perceive the novel contrasts. Hence, in these cases, simple awareness about which acoustic cue is most relevant to perceive the L2 contrast may be sufficient to enable perception of these contrasts, as suggested by the results of Experiments II and III presented in chapter 4. For speech categories falling under predictions 4 and 5 of the BLIP model, L2 learners are presumed to lack the appropriate neural maps to efficiently perceive the L2 categories using the relevant acoustic cue employed by L1 speakers, as suggested by the results of Experiments I and IV reported previously. In these cases, computer-based training may be particularly useful, especially if the training paradigm involves a categorical task, such as a forced-choice identification task, which features real minimal pairs produced by various speakers and that contrasts the target sounds in different contexts. In some cases, it may be necessary or desirable to manipulate some acoustic cues to help L2 learners to focus

on the most reliable cue(s) that distinguish the L2 contrast, or to enable them to progressively reorganize a pre-existent neural map. Adding an adaptive component to the training paradigm is also expected to facilitate the rate and robustness of acquisition of the novel L2 categories by adult learners. The acquisition of appropriate neural maps to perceive novel L2 categories is expected to help not only with perception of these contrasts, but also with their production, as suggested by previous perceptual training experiments (e.g. Bradlow et al. 1997, 1999; Logan, Lively & Pisoni 1991).

5.3 Future directions

The BLIP model is sufficiently general and at the same time adequately specific to offer a convenient framework that could be applied to the study of L1 development, L2 perception and acquisition, or potentially to research in speech-language pathology. However, the model still has some limitations and shortcomings. For instance, the phonological level of processing in the model needs to be further developed and more firmly grounded; or contrasted with previous phonological literature and research. Also, although the model is informed by neurological processing, as documented in previous neuroethology and neurolinguistic experiments, the assumptions and mechanisms in the BLIP model have not yet been tested with neurolinguistic experiments. Finally, the predictions of the model for L2 perception and acquisition still need to be tested with a wider variety of speech contrasts using different language groups. In this section I briefly outline my plans to further develop the model in the near future, although I am hoping that other researchers will take the model in other directions as well.

One of the main assumptions upon which the BLIP model is built is the inverted magnification factor hypothesis, which posits that cell density activation should decrease during the processing of stimuli close to the categorical center of speech categories (and increase at categorical boundaries). Guenther and Bohland (2002) confirmed this hypothesis with an fMRI experiment evaluating the perception of a spectral contrast. In general, however, an increase—not a decrease—in brain activation is associated with better performance on a given task. For instance, typically reading children usually exhibit more brain activation during a reading task than children with dyslexia (Gabrieli 2009). Although this may be counterintuitive from a neurological point of view, in the case of speech processing, the converse may be true. Testing this hypothesis is critical for a better understanding of categorical processing as applied to speech and other modalities, as well as to confirm a central assumption of the BLIP model. Accordingly, an fMRI experiment is currently underway, which aims to: (1) document possible changes in cell density activation for processing durational contrasts; (2) evaluate the potential malleability of neural activation in the mature adult brain; and (3) compare cross-linguistic data about neural activation used to perceive spectral and durational contrasts.

With the possible exception of the inverted magnification factor hypothesis, to the best of my knowledge, it remains impossible to directly prove the mechanisms posited by the BLIP model with the current technology available for human experiments, especially the two levels of processing claimed to be performed by the neurology. It is possible, however, to confirm the psychological and behavioral reality of these levels by

conducting behavioral experiments, particularly by evaluating the predictions of the BLIP model for L2 perception and acquisition presented in chapter 4. An experiment is being planned to test the perception of the English high front vowel contrast, as in the words 'beat' and 'bit', with speakers of European French. For these speakers, unlike the Canadian French speakers tested in experiment II, these vowels are not context-bound allophones. If the predictions of the BLIP model are accurate, Canadian French speakers possess two context-bound neural maps to process this vowel contrast, while European French speakers do not. Accordingly, European French speakers should perform worse than their Canadian French counterparts on the same L2 vowel contrast. Testing of the predictions of the BLIP model with other language groups and other speech contrasts is also being considered, but not yet in motion.

Although the BLIP model was designed to capture the categorical processing of *acoustic* contrasts, the model could be extended to account for the processing of other cues, such as visual cues for instance, which may be used to process phonological categories in sign languages. The framework provided by the BLIP model technically allows for the processing of any type of cue, but the neural basis to capture the processing of non acoustic contrasts still needs to be investigated. Also in need of investigation and modeling, is the processing of speech contrasts not discussed in this dissertation (e.g. nasals, liquids, affricates, etc.) In sum, it is my hope that future research will be able to confirm some of the ideas put forward in this dissertation, and that the model will prove useful to other researchers interested in speech processing.

References

Abdelli-Beruh, N.-B. (2004). The stop voicing contrast in French sentences: Contextual sensitivity of vowel duration, closure duration, aspiration duration and closure voicing. *Phonetica, 61* (4), 201–219.

Akamatsu, T. (1997). *Japanese phonetics: Theory and practice.* Newcastle: Lincom Europa.

Allport, D. A. (1985). Distributed memory, modular subsystems and dysphasia. In S. Newman & R. Epstein (Eds.), *Current perspectives in dysphasia* (pp. 32-60). New York, NY: Churchill Livingstone.

Anderson, J. L., Morgan, J. L., & White, K. S. (2003). A Statistical Basis for Speech Sound Discrimination. *Language and Speech, 46* (2-3), 155-182.

Angelo, R., & Moller, A.R. (1990). Response from the inferior colliculus in the rat to tone bursts and amplitude-modulated continuous tones. *Audiology, 29*, 336-346.

Ashby, F. G., & Maddox, W. T. (1993). Relations between prototype, exemplar, and decision-bound models of categorization. *Journal of Mathematical Psychology, 37*, 372-400.

Aslin, R. N., Pisoni, D. B., Hennessy, B. L., & Perey, A. J. (1981). Discrimination of voice onset time by human infants: new findings and implications for the effects of early experience. *Child Development, 52*, 1135-1145.

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of Conditional Probability Statistics by 8-Month-Old Infants. *Psychological Science, 9* (4), 321-324.

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1999). Statistical learning in linguistic and nonlinguistic domains. In B. MacWhinney (Ed.) *The emergence of language* (pp. 359-380). Mahwah, NJ: Lawrence Erlbaum Associates.

Barrett, S. (1999). The perceptual magnet effect is not specific to speech prototypes: New evidence from music categories. *Speech*, *11*, 1-16.

Bartlett, E. L., & Wang, X. (2005). Long-lasting modulation by stimulus context in primate auditory cortex. *Journal of Neurophysiology, 94*, 83-104.

Bauer, H.-U., Der, R., & Herrmann, M. (1996) Controlling the magnification factor of self-organizing feature maps. *Neural Computation 8*, 757–771.

Behrens, S. J., & Blumstein, S. E. (1988). Acoustic characteristics of English voiceless fricatives: A descriptive analysis. *Journal of Phonetics 16*, 295 – 298.

Bendor, D., & Wang, X. (2005). The neuronal representation of pitch in primate auditory cortex. *Nature, 436*, 1161-1165.

Benkí, J. R. (1998). Evidence for phonological categories from speech perception. Doctoral dissertation. University of Massachusetts Amherst.

Benkí, J. R. (2001). Place of articulation and first formant transition pattern both affect perception of voicing in English. *Journal of Phonetics, 29* (1), 1-22.

Benner, A., Grenon, I., & Esling, J. (2007). Infants' phonetic acquisition of voice quality parameters in the first year of life. In *Proceedings of the XVIth International Congress of Phonetic Sciences (ICPhS)*, 2073-2076.

Best, C. T. (1993). Emergence of language-specific constraints in perception of non-native speech: A window on early phonological development. In B. de Boysson-Bardies, S. de Schonen. P. Jusczyk, P. McNeilage, & J. Morton (Eds.), *Developmental neurocognition: Speech and face processing in the first year of life* (pp. 289-304) Dordrecht, The Netherlands: Kluwer Academic Publishers.

Best, C. T. (1994). The emergence of native-language phonological influences in infants: A perceptual assimilation model. In J. Goodman & H. C. Nusbaum (Eds). *The development of speech perception: the transition from speech sounds to spoken words* (pp. 167-224) Cambridge, MA: MIT Press.

Best, C. T. (1995). A direct realist view of cross-language speech perception. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp.171-204). Timonium, MD: York Press.

Best, C. T., & McRoberts, G. W. (2003). Infant perception of non-native consonant contrasts that adults assimilate in different ways. Language and Speech, 46 (2-3), 183-216.

Best, C. T., McRoberts, G. W., & Goodell, E. (2001). Discrimination of non-native consonant contrasts varying in perceptual assimilation to the listener's native phonological system. *The Journal of the Acoustical Society of America, 109* (2), 775-794.

Best, C. T., & Strange, W. (1992). Effects of phonological and phonetic factors on cross-language perception of approximants. *Journal of Phonetics, 20*, 305-330.

BharrathSingh, K. (2001). Prototypes and the perceptual magnet effect in the perception of vowels. Doctoral Dissertation. Carleton University.

Bitterman, Y., Mukamel, R., Malach, R., Fried, I., & Nelken, I. (2008). Ultra-fine frequency tuning revealed in single neurons of human auditory cortex. *Nature*, *451*, 197-201.

Boersma, P., & Weenink, D. (2007) *Praat: Doing phonetics by computer* (version 4.6.38) [computer program]. Retrieved 2007 from <http://www.praat.org/>.

Bohn, O.-S. (2005). Establishing cross-language perceptual similarity of speech sounds. Talk presented at the *Workshop on Models of L1 and L2 Phonetics/Phonology*. Utrecht, The Netherlands.

Bohn, O.-S. & Steinlen, A.K. (2003). Consonantal context affects cross-language perception of vowels. Proceedings of the 15th *International Congress of Phonetic Sciences*, 2289-2292

Bradlow, A. R., Akahane-Yamada, R., Pisoni, D., & Tohkura, Y. (1999). Training Japanese listeners to identify English /r/ and /l/: Long-term retention of learning in perception and production. *Perception & Psychophysics, 61* (5), 977-985.

Bradlow, A. R., Pisoni, D., Akahane-Yamada, R., & Tohkura, Y. (1997). Training Japanese listeners to identify English /r/ and /l/: Some effects of perceptual learning on speech production. *Journal of Acoustical Society of America*, *101* (4), 2299-2310.

Brown, C. (1997) Acquisition of segmental structure: Consequences for speech perception and second language acquisition. Doctoral dissertation. McGill University. Montréal, Québec.

Brown, C. A. (1998). The role of the L1 grammar in the L2 acquisition of segmental structure. *Second Language Research*, *14* (2), 136-193.

Brown, C. A. (2000). The interrelation between speech perception and phonological acquisition from infant to adult. In J. Archibald (Ed.), *Second language acquisition and linguistic theory* (pp. 4-63). Malden, MA & Oxford: Blackwell Publishers.

Brunelle, M. (2009). Tone perception in Northern and Southern Vietnamese. *Journal of Phonetics, 37* (1), 79-96.

Buckner, R. L., & Logan, J. M. (2001). Functional neuroimaging methods: PET and fMRI. In R. Cabeza & A. Kingstone (Eds.), *Handbook of functional neuroimaging of cognition* (pp. 27-48). Cambridge, MA: MIT Press.

Bybee, J. (2000). The phonology of the lexicon: evidence from lexical diffusion. In M. Broe & J. Pierrehumbert (Eds.) *Papers in Laboratory Phonology V: Acquisition and the Lexicon*. Cambridge, UK: Cambridge University Press.

Caramazza, A., & Yeni-Komshian, G. H. (1974). Voice onset time in two French dialects. *Journal of* Phonetics, 2, 239–245.

Caramazza, A., Yeni-Komshian, G. H., Zurif, E. B., & Carbone, E. (1973). The acquisition of a new phonological contrast: The case of stop consonants in French-English bilinguals. *Journal of the Acoustical Society of America, 54*, 421–428.

Catherwood, D., Crassini, B., & Freiberg, K. (1989). Infant response to stimuli of similar hue and dissimilar shape: Tracing the origins of the categorization of objects by hue. *Child Development, 60,* 752-762.

Chang, E.F., & Merzenich, M.M. (2003). Environmental noise retards auditory cortical development. *Science, 300*, 498–502.

Clarey, J. C., Paolini, A. G., Grayden, D. B., Burkitt, A. N., & Clark, G. M. (2004). Ventral cochlear nucleus coding of voice onset time in naturally spoken syllables. *Hearing Research, 190*, 37-59.

Clements, G. N., & Hume, E. V. (1995) The internal organization of speech sounds. In J. A. Goldsmith (Ed.), *The handbook of phonological theory* (pp. 245-306). Cambridge: Blackwell Publishers.

Cooper, F. S., Delattre, P. C., Liberman, A. M., Borst, J. M., & Gerstman, L. J. (1952). Some experiments on the perception of synthetic speech sounds. *Journal of Acoustical Society of America*, *24*, (6), 597-606.

Curtin, S., Goad, H., & Pater, J. V. (1998) Phonological transfer and levels of representation: the perceptual acquisition of Thai voice and aspiration by English and French speakers. *Second Language Research, 14* (4), 389-405.

Dear, S. P., Simmons, J. A., & Fritz, J. (1993). A possible neural basis for representation of acoustic scenes in auditory cortex of the big brown bat. *Nature*, *364*, 620-623.

Dear, S. P., & Suga, N. (1995). Delay-tuned neurons in the midbrain of the big brown bat. *Journal of Neurophysiology, 73*, 1084-1100.

Delattre, P. C., Liberman, A. M., & Cooper, F. S. (1951). Voyelles synthetiques a deux formants et voyelles cardinales. Le Maître Phonétique, 96, 30-36.

Diesch, E., & Luce, T. (2000). Topographic and temporal indices of vowel spectral envelope extraction in the human auditory cortex. *Journal of cognitive neuroscience, 12* (5), 878-893.

Drachman, D. (2005). "Do we have brain to spare?" *Neurology, 64* (12), 2004–2005.

Edamatsu, H., Kawasaki, M., & Suga, N. (1989). Distribution of combination-sensitive neurons in the ventral fringe area of the auditory cortex of the mustached bat. *Journal of Neurophysiology, 61*, 202-207.

Edamatsu, H., & Suga, N. (1993). Differences in responses properties of neurons between two delay-tuned areas in the auditory cortex of the mustached bat. *Journal of Neurophysiology, 69*, 1700-1712.

Eggermont, J. J. (2001). Between sound and perception: reviewing the search for a neural code. *Hearing Research, 157*, 1-42.

Eimas, P. D. (1975). Auditory and phonetic coding of the cues for speech: Discrimination of the /r-l/ distinction by young infants. *Perceptual Psychophysics 18,* 341-347.

Eimas, P. D., & Miller, J. L. (1992). Organization in the perception of speech by young infants. *Psychological Science, 3* (6), 340-345.

Eimas, P. D., Siqueland, E. R., Jusczyk, P., & Vigorito, J. (1971). Speech perception in infants. *Science, 171*, 303-306.

Escudero, P. (2005). Linguistic perception and second language acquisition. Doctoral dissertation. Utrecht: LOT.

Escudero, P., & Boersma, P. (2003). Modelling the perceptual development of phonological contrasts with Optimality Theory and the Gradual Learning Algorithm. *University of Pennsylvania Working Papers in Linguistics*, 8 (1), 71-85.

Evers, V., Reetz, H., & Lahiri, A. (1998). Crosslinguistic acoustic categorization of sibilants independent of phonological status. *Journal of Phonetics 26*, 345 – 370.

Fant, G. (1960). *Acoustic theory of speech production*. Mouton, The Hague.

Feng, A. S., Simmons, J. A., & Kick, S. A. (1978). Echo detection and target-ranging neurons in the auditory system of the bat *Eptesicus fuscus*. *Science, 202*, 645-648.

Fennell, C. T. & Werker, J. F. (2004). Infant attention to phonetic detail: Knowledge and familiarity effects. In A. Brugos, L. Micciulla, & C. E. Smith (Eds.), *Proceedings of the 28th Annual Boston University Conference on Language Development* (Vol. 1, pp. 165–176). Somerville, MA: Cascadilla.

Fischer, R., & Ohde, R. (1990). Spectral and duration properties of front vowels as cues to final stop voicing. *Journal of the Acoustical Society of America, 88*, 1250-1259.

Fitch, W. T., & Fritz, J. B. (2006). Rhesus macaques spontaneously perceive formants in conspecific vocalizations. *Journal of Acoustical Society of America*, *120*, (4), 2132-2141.

Flege, J. E. (1991). The interlingual identification of Spanish and English vowels: Orthographic evidence. *The Quarterly Journal of Experimental Psychology,* 43A, *3*, 701-731.

Flege, J. E. (1992a). Speech learning in a second language. In C. Ferguson, L. Menn, & C. Stoel-Gammon (Eds.) *Phonological development: Models, Research, and Application* (pp. 565-604). Timonium, MD: York.

Flege, J. E. (1992b). The intelligibility of English vowels spoken by British and Dutch talkers. In R. Kent (Ed.) *Intelligibility in speech disorders: Theory, measurement, and management* (pp. 157-232). Amsterdam: Benjamins.

Flege, J. E. (1993). Production and perception of a novel, second-language phonetic contrast. *Journal of the Acoustical Society of America, 93* (3), 1589-1608.

Flege, J. E. (1995). Second language speech learning theory, findings, and problems. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 233-277). Timonium, MD: York Press

Forrest, K., Weismer, G., Milenkovic, P., & Dougall, R. N. (1988). Statistical analysis of word-initial voiceless obstruents: Preliminary data. *Journal of the Acoustical Society of America, 84*, 115 – 124.

Fowler, C. A. (1994). Invariants, specifiers, cues: An investigation of locus equations as information for place of articulation. *Perception and Psychophysics, 55*, 597 – 611.

Franklin, A., & Davies, I. R. L. (2004). New evidence for infant colour categories. *British Journal of Developmental Psychology, 22,* 349-377.

Franklin, A., Pilling, M., & Davies, I. (2005). The nature of infant color categorization: Evidence from eye movements on a target detection task. *Journal of Experimental Child Psychology, 91,* 227-248.

Fruchter, D., & Sussman, H. M. (1997). The perceptual relevance of locus equations. *Journal of Acoustical Society of America*, *102*, (5), 2997-3008.

Fry, D. B. (1955). Duration and intensity as physical correlates of linguistic stress. *Journal of the Acoustical Society of America, 27*, 765-768.

Fry, D. B. (1958). Experiments in the perception of stress. *Language and Speech, 1*, 126-152.

Fuzessery, Z. M., & Feng, A. S. (1982). Frequency selectivity in the anuran auditory midbrain: Single unit responses to single and multiple tone stimulation. *Journal of Comparative Physiology A, 146*, 471-484.

Fuzessery, Z. M., & Feng, A. S. (1983). Mating call selectivity in the thalamus and midbrain of the leopard frog (*Rana p. pipiens*): Single and multiunit analysis. *Journal of Comparative Physiology A, 150*, 333-344.

Gabrieli, J. D. E. (2009). Dyslexia: A new synergy between education and cognitive neuroscience. *Science, 325*, 280-283.

Gallen C.C., Hirschkoff E.C., & Buchanan D.S. (1995). Magnetoencephalography and magnetic source imaging. Capabilities and limitations. *Neuroimaging Clinics of North America, 5* (2), 227-249.

Gehr, D. D., Komiya, H., & Eggermont, J. J. (2000). Neuronal responses in cat primary auditory cortex to natural and altered species-specific calls. *Hearing Research, 150*, 27-42.

Gerken, L., Wilson, R., & Lewis, W. (2005). Infants Can Use Distributional Cues to Form Syntactic Categories. *Journal of Child Language, 32*(2), 249-268.

Goldinger, S. D. (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*, 1166–1183.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review, 105*, 251–279.

Goldinger, S. D., Kleider, H. M., & Shelley, E. (1999). The marriage of perception and memory: Creating two-way illusions with words and voices. *Memory & Cognition, 27* (2), 328-338.

Goldstein, L., Nam, H., Kulthreshtha, M., Root, L., & Best, C. (2008). Distribution of tongue tip articulations in Hindi versus English and the acquisition of stop place categories. Paper presented at *Laboratory Phonology 11*, Wellington, New Zealand, 30 June-2 July 2008.

Greenberg, S. (2006). A multi-tier framework for understanding spoken language. In S. Greenberg, & W. A. Ainsworth (Eds.), *Listening to speech: An auditory perspective* (pp. 411-433). Mahwah, NJ: Lawrence Erlbaum Associates.

Grenon, I. (2006). Adults still have direct access to UG: Evidence from the perception of a non-native feature contrast. In M. G. O'Brien, C. Shea, & J. Archibald (Eds.) *Proceedings of the 8th Generative Approaches to Second Language Acquisition Conference* (GASLA) (pp. 51-62). Somerville, MA: Cascadilla Proceedings Project. www.lingref.com, document #1487.

Grenon, I. (2008) The acquisition of English soun[dz] by native Japanese speakers: A perceptual study. Saarbrücken, Germany: VDM Verlag.

Grenon, I. & White, L. (2008). Acquiring rhythm: A comparison of L1 and L2 speakers of Canadian English and Japanese. In *Proceedings of the 32nd Boston University Conference on Language Development (BUCLD)*, 155-166.

Grenon, I., Benner, A., & Esling, J. (2007). Language-specific phonetic production patterns in the first year of life. In *Proceedings of the XVIth International Congress of Phonetic Sciences (ICPhS)*, 1561-1564.

Guenther, F. H. (1995). Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech. *Psychological Review, 102* (3), 594-621.

Guenther, F. H. (2006). Cortical interactions underlying the production of speech sounds. *Journal of Communication Disorders, 39* (5), 350-365.

Guenther, F. H., & Bohland, J. W. (2002). Learning sound categories: A Neural Model and supporting experiments. *Acoustical Science and Technology, 23* (4), 213-221.

Guenther, F. H., Ghosh, S. S., & Tourville, J. A. (2006). Neural modeling and imaging of the cortical interactions underlying syllable production. *Brain & Language, 96* (3), 280-301.

Guenther, F. H., & Gjaja, M. N. (1996). The perceptual magnet effect as an emergent property of neural map formation. *Journal of the Acoustical Society of America, 100*, 1111–1121.

Guenther, F. H., Husain, F. T., Cohen, M. A., & Shinn-Cunningham, B. G. (1999). Effects of categorization and discrimination training on auditory perceptual space. *Journal of the Acoustical Society of America, 106* (5), 2900-2912.

Guenther, F. H., Nieto-Castanon, A., Ghosh, S. S., & Tourville, J. A., (2004). Representation of sound categories in auditory cortical maps. *Journal of Speech, Language, and Hearing Research, 47*, 46–57.

Guenther, F. H., Nieto-Castanon, A., Tourville, J. A., & Ghosh, S. S. (2001). Effects of categorization training on auditory perception and cortical representations. In *Proceedings of the Speech Recognition as Pattern Classification (SPRAAC) Workshop*, Nijmegen, The Netherlands, July 11-13.

Hayes-Harb, R. (2007). Lexical and statistical evidence in the acquisition of second language phonemes. *Second Language Research, 23* (1), 65-94.

Hebb, D.O. (1949). *The organization of behavior*. New York: Wiley.

Heinz, J. M., & Stevens, K. N. (1961). On the properties of voiceless fricative consonants. *Journal of the Acoustical Society of America, 33*, 589 – 596.

Hillenbrand, J. M., Clark, M. J., & Nearey, T. M. (2001). Effects of consonant environment on vowel formant patterns. *Journal of the Acoustical Society of America, 109* (2), 748–763

Hillenbrand, J. M., Houde, R. A., & Gayvert, R. T. (2006). Speech perception based on spectral peaks versus spectral shape. *Journal of Acoustical Society of America*, *119*, (6), 4041-4054.

Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of America, 97*, 3099–3111.

Horikawa, J., & Suga, N. (1986). Biosonar signals and cerebellar auditory neurons of the mustached bat. *Journal of Neurophysiology, 55*, 1247-1267.

Hose, B., Langner, G., & Scheich, H. (1987). Topographic representation of periodicities in the forebrain of the mynah bird: One map for pitch and rhythm? *Brain Research 422* (2), 367-373.

Hughes, G. W., & Halle, M. (1956). Spectral properties of fricative consonants. *Journal of the Acoustical Society of America 28*, 303 – 310.

International Phonetic Association. (1999). Handbook of the International Phonetic Association Cambridge: Cambridge University Press.

Iverson, P., & Kuhl, P. K. (1995). Mapping the perceptual magnet effect for speech using signal detection theory and multidimensional scaling. *Journal of the Acoustical Society of America, 97* (1), 553-562.

Iverson, P., & Kuhl, P. K. (1996) Influences of phonetic identification and category goodness on American listeners' perception of /r/ and /l/. *Journal of the Acoustical Society of America 99*, 1130-1140.

Iverson, P., Hazan, V., & Bannister, K. (2005). Phonetic training with acoustic cue manipulations: A comparison of methods for teaching English /r/-/l/ to Japanese adults. *Journal of the Acoustical Society of America 118* (5), 3267-3278.

Iverson, P., Kuhl, P. K., Akahane-Yamada, R., Diesch, E., Tohkura, Y., Kettermann, A., & Siebert, C. (2003). A perceptual interference account of acquisition difficulties for non-native phonemes. *Cognition, 87*, B47-B57.

Jiang, J., Chen, M., & Alwan, A. (2006). On the perception of voicing in syllable-initial plosives in noise. *Journal of the Acoustical Society of America, 119* (2), 1092-1105.

Johnson, D. (1980). The relationship between spike rate and synchrony in responses of auditory nerve fibers to single tones. *Journal of Acoustical Society of America*, *68*, 10157-10170.

Johnson, K. (1997). Speech perception without speaker normalization. In K. Johnson, & J. Mullennix (Eds.), *Talker variability in speech processing* (pp. 145–166). San Diego: Academic Press.

Johnson, K. (2005). Speaker Normalization in speech perception. In D. B. Pisoni, & R. Remez (Eds.) *The handbook of speech perception* (pp. 363-389). Oxford: Blackwell Publishers.

Jones, C. (2005). Effects of vocalic duration and first formant offset on final voicing judgments by children and adults (L). *Journal of the Acoustical Society of America, 117* (6), 3385-3388.

Jongman, A., Wayland, R., & Wong, S. (2000). Acoustic characteristics of English fricatives. *Journal of Acoustical Society of America*, *108*, (3), 1252-1263.

Jusczyk, P. W. (1987). Implications from speech studies on the unit of perception. In M. E. H. Schouten (Ed.), *The psychophysics of speech perception* (pp. 433-443). Dordrecht, The Netherlands: Nijhoff.

Jusczyk, P. W. (1993). From general to language-specific capacities: the WRAPSA Model of how speech perception develops. *Journal of Phonetics, 21*, 3-28.

Jusczyk, P. W., Rosner, B. S., Cutting, J. E., Foard, C. F., & Smith, L. B. (1977). Categorical perception of nonspeech sounds by 2-month-old infants. *Perception & Psychophysics, 21* (1), 50-54.

Keating, P. (1984). Phonetic and phonological representation of stop consonant voicing. *Language, 60*, 286–319.

Kenstowicz, M. (1994). Phonology in generative grammar. Malden, MA: Blackwell.

Kingston, J., & Diehl, R. L. (1994). Phonetic knowledge. *Journal of the Linguistic Society of America, 70*(3), 419-454.

Kiriloff, C. (1969). On the auditory discrimination of tones in Mandarin. *Phonetica, 20*, 63-67.

Kirmse, U., Ylinen, S., Tervaniemi, M., Vainio, M., Schröger, E., & Jacobsen, T. (2008). Modulation of the mismatch negativity (MMN) to vowel duration changes in native speakers of Finnish and German as a result of language experience. *International Journal of Psychophysiology, 67*, 131-143.

Kluender, K. R., Diehl, R. L., & Killeen, P. R. (1987). Japanese quail can learn phonetic categories. *Science, 237*, 1195-1197.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics, 43*, 59-69.

Kohonen, T. (2001). Self-organizing maps. Third, extended edition. Springer.

Krebs, B., Lesica, N. A., & Grothe, B. (2008). The representation of amplitude modulations in the mammalian auditory midbrain. *Journal of Neurophysiology, 100*, 1602-1609.

Krishna, B.S., & Semple, M.N. (2000). Auditory temporal processing: responses to sinusoidally amplitude-modulated tones in the inferior colliculus. *Journal of Neurophysiology, 84*, 255-273.

Krull, D. (1989). Second formant locus patterns and consonant-vowel coarticulation in spontaneous speech. Phonetic Experimental Research at the Institute of Linguistics, University of Stockholm, *PERILUS X*, 87-108.

Kuhl, P. K. (1981). Discrimination of speech by nonhuman animals: Basic sensitivities conducive to the perception of speech-sound categories. *Journal of the Acoustical Society of America, 70*, 340-349.

Kuhl, P. K. (1983). Perception of auditory equivalence classes for speech in early infancy. Infant Behavior and Development, 6 (3), 263-285.

Kuhl, P. K. (1991). Human adults and human infants show a "Perceptual Magnet Effect" for the prototypes of speech categories, monkeys do not. *Perception & Psychophysics, 50* (2), 93-107.

Kuhl, P. K. (1993a). Infant speech perception: A window on psycholinguistic development. *International Journal of Psycholinguistics, 9* (1), 33-56.

Kuhl, P. K. (1993b). Early linguistic experience and phonetic perception: Implications for theories of developmental speech perception. *Journal of Phonetics, 21* (1-2), 125-139.

Kuhl, P. K. (2007). Cracking the speech code: How infants learn language. *Acoustical Science and Technology, 28* (2), 71-83.

Kuhl, P. K., & Iverson, P. (1995). Linguistic Experience and the "Perceptual Magnet Effect". In W. Strange (Ed.) Speech perception and linguistic experience: Issues in cross-language research (pp. 121-154). Timonium, MD: York Press.

Kuhl, P. K., & Miller, J. D. (1975a). Speech perception by the chinchilla: Voiced-voiceless distinction in alveolar plosive consonants. *Science, 190*, 69-72.

Kuhl, P. K., & Miller, J. D. (1975b). Speech perception in early infancy: Discrimination of speech-sound categories. *Journal of the Acoustical Society of America, 58,* S56.

Kuhl, P. K., & Miller, J. D. (1978). Speech perception by the chinchilla: Identification functions for synthetic VOT stimuli. *Journal of the Acoustical Society of America, 63*, 905-917.

Kuhl, P. K., & Padden, D. M. (1982). Enhanced discriminability at the phonetic boundaries for the voicing feature in macaques. *Perception & Psychophysics, 32* (6), 542-550.

Kuhl, P. K., & Padden, D. M. (1983). Enhanced discriminability at the phonetic boundaries for the place feature in macaques. *Journal of the Acoustical Society of America*, *73*, 1003-1010.

Kuhl, P. K., Stevens, E., Hayashi, A., Deguchi, T., Kiritani, S., & Iverson, P. (2006). Infants show facilitation for native language phonetic perception between 6 and 12 months. *Developmental Science, 9,* 13-21.

Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., & Lindblom, B. (1992). Linguistic experience alters phonetic perception in Infants by 6 months of age. *Science, 255*, 606-608.

Kuo, Y., Rosen, S., & Faulkner, A. (2008). Acoustic cues to tonal contrasts in Mandarin: Implications for cochlear implants. *The Journal of the Acoustical Society of America, 123* (5), 2815-2824.

Lacerda, F. (1995). The perceptual-magnet effect: An emergent consequence of exemplar-based phonetic memory. In K. Elenius, & P. Branderud (Eds.), *Proceedings of the XIIIth International Congress of Phonetic Sciences*, vol. 2 (pp. 140-147). Stockholm: KTH and Stockholm University.

LaCharité, D., & Prévost, P. (1999). The role of L1 and of teaching in the acquisition of English sounds by Francophones. In Annabel Greenhill, Heather Littlefield and Cheryl Tano (Eds), *Proceedings of the 23rd annual Boston University Conference on Language Development* (pp. 373-385), Somerville: Cascadilla Press.

Ladefoged, P. (1990). Some reflections on the IPA. *Journal of Phonetics, 18*, 335-346.

Ladefoged, P. (2001). *Vowels and consonants*. Malden, MA: Blackwell.

Ladefoged, P., & Broadbent, D. E. (1957). Information conveyed by vowels. Journal of the Acoustical *Society of America, 29* (1), 98-104.

Laeufer, C. (1996). The acquisition of a complex phonological contrast: Voice timing patterns of English initial stops by native French speakers. *Phonetica, 53*, 86–110.

Langner, G. (1983). Evidence for neuronal periodicity detection in the auditory system of the guinea fowl: implications for pitch analysis in the time domain. *Experimental Brain Research, 52*, 333-355.

Langner, G. (1992). Periodicity coding in the auditory system. *Hearing Research, 60*, 115-142.

Langner, G., & Schreiner, C.E. (1988). Periodicity coding in the inferior colliculus of the cat. I. Neuronal mechanisms. *Journal of Neurophysiology*, *60*, 1799-1822.

Langner, G., Albert, M., & Briede, T. (2002). Temporal and spatial coding of periodicity information in the inferior colliculus of awake chinchilla (*Chinchilla laniger*). *Hearing Research, 168*, 110-130.

Lehiste, I. (1970). *Suprasegmentals*. Cambridge, MA: M.I.T. Press.

Leite, W. L., Svinicki, M. & Shi, Y. (2010). Attempted validation of the scores of the VARK: Learning styles inventory with multitrait-multimethod confirmatory factor analysis models. *Educational and Psychological Measurement, 70*, 323-339.

Liberman, A. M., Harris, K. S., Hoffman, H., & B. Griffith. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, *54*, 358-368.

Liberman, A. M., Harris, K. S., Kinney, J. A., & Lane, H. (1961). The discrimination of relative onset-time of the components of certain speech and non-speech patterns. *Journal of Experimental Psychology*, *61*, 379-388.

Lieberman, P. (1960). Some acoustic correlates of word stress in American English. *Journal of the Acoustical Society of America, 32*, 451-454.

Lieberman, P., & Blumstein, S. E. (1988). *Speech physiology, speech perception, and acoustic phonetics.* Cambridge University Press, Cambridge, UK.

Lindblom, B. (1963). On vowel reduction. Rep. No. 29, The Royal Institute of Technology, Speech Transmission Laboratory, Stockholm, Sweden.

Lisker, L. (1975). Is it VOT or a first-formant transition detector? *Journal of the Acoustical Society of America 57*, 1547–1551.

Lisker, L., & Abramson, A. S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word, 20*, 384–422.

Lisker, L., & Abramson, A. S. (1967). Some effects of context on voice onset time in English stops. *Language and Speech, 10*, 1–28.

Lively, S. E., Logan, J. S., & Pisoni, D. B. (1993). Training Japanese listeners to identify English /r/ and /l/ II: The role of phonetic environment and talker variability in learning new perceptual categories. *Journal of the Acoustical Society of America, 94* (3), 1242-1255.

Logan, J. S., Lively, S. E., & Pisoni, D. B. (1991). Training Japanese listeners to identify English /r/ and /l/: A first report. *Journal of Acoustical Society of America, 89* (2), 874-886.

Lotto, A. J. (2000). Reply to "An analytical error invalidates the 'depolarization' of the perceptual magnet effect" [J. Acoust. Soc. Am. 107, 3576-3577 (2000)]. *The Journal of the Acoustical Society of America*, *107* (6), 3578-3580.

Lotto, A. J., Kluender, K. R., & Holt, L. L. (1998). Depolarizing the perceptual magnet effect. *The Journal of the Acoustical Society of America*, *103* (6), 3648-3655.

Maekawa, M., Wong, D., & Paschal, W. G. (1992). Spectral selectivity of FM-FM neurons in the auditory cortex of the echolocating bat, *Myotis lucifugus*. *Journal of Comparative Physiology A, 171,* 513-522.

Major & Kim (1999). The similarity differential rate hypothesis. *Language Learning, 49*, supplement 1, 151-183.

Margoliash, D. (1983). Acoustic parameters underlying the responses of song-specific neurons in the white-crowned sparrow. *Journal of Neuroscience, 3*, 1039-1057.

Margoliash, D., & Fortune, E. S. (1992). Temporal and harmonic combination-sensitive neurons in the zebra finch's HVc. *Journal of Neuroscience, 12*, 4309-4326.

Martin, P. (1996). *Éléments de phonétique avec application au français*. Sainte-Foy: Les Presses de l'Université Laval.

Martin, P. (2004). Dévoisement vocalique en français. *La Linguistique 40* (2), 3-21.

Mather, G. (2006). *Foundations of perception.* Hove, East Sussex and New York, NY: Psychology Press.

Maye, J. C. (2000). The Acquisition of Speech Sound Categories on the Basis of Distributional Information. Doctoral Dissertation. University of Arizona.

Maye, J., & Gerken, L. (2000). Learning phonemes without minimal pairs. *Proceedings of the Annual Boston University Conference on Language Development, 24* (2), 522-533.

Maye, J., & Gerken, L. (2001). Learning phonemes: How far can the input take us? *Proceedings of the Annual Boston University Conference on Language Development, 25* (2), 480-490.

Maye, J., & Weiss, D. (2003). Statistical cues facilitate infants' discrimination of difficult phonetic contrasts. *Proceedings of the Annual Boston University Conference on Language Development, 27* (2), 508-518.

Maye, J., Weiss, D. J., & Aslin, R. N. (2008). Statistical phonetic learning in infants: facilitation and feature generalization. *Developmental Science, 11* (1), 122-134.

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition, 82* (3), B101-B111.

McMurray, B., & Aslin, R. N. (2005). Infants are sensitive to within-category variation in speech perception. *Cognition, 95*, B15–B26.

McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2002). Gradient effects of within-category phoneticvariation on lexical access. *Cognition, 86*, B33–B42.

McQueen, J. M., Norris, D., & Cutler, A. (2006) The dynamic nature of speech perception. *Language and Speech*, *49*, 1, 101-112.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review, 5*, 207-238.

Middlebrooks, J. C. (2008). Auditory cortex phase locking to amplitude-modulated cochlear implant pulse trains. *Journal of Neurophysiology, 100*, 76-91.

Miller, J. L., & Eimas, P. D. (1996). Internal structure of voicing categories in early infancy. *Perception and Psychophysics, 58* (8), 1157–1167.

Mittman, D. H., & Wenstrup, J. J. (1995). Combination-sensitive neurons in the inferior colliculus. *Hearing Research, 90*, 185-191.

Mol, H., & Uhlenbeck, E. M. (1965). The linguistic relevance of intensity in speech. *Lingua, 5*, 205-213.

Morrison, G. S. (2002). Effects of L1 duration experience on Japanese and Spanish listeners' perception of English high front vowels. MA thesis. Simon Fraser University, Canada.

Morton, J., & Jassem, W. (1965). Acoustic correlates of stress. *Language and Speech, 8*, (3) 159-181.

Mudry, K. M., Constantin-Paton, M., & Capranica, R. R. (1977). Auditory sensitivity of the diencephalon of the leopard frog *Rana p. pipiens*. *Journal of Comparative Physiology, 114*, 1-13.

Myers, E. B. (2007). Dissociable effects of phonetic competition and category typicality in a phonetic categorization task: An fMRI investigation. *Neuropsychologia, 45*, 1463-1473.

Nearey, T. M., and Shammass, S. E. (1987). Formant transitions as partly distinctive invariant properties in the identification of voiced stops. *Canadian Acoustics, 15* (4), 17-24.

Nelken, I. (2008). Processing of complex sounds in the auditory system. *Current Opinion in Neurobiology, 18*, 413-417.

Nittrouer, S., Studdert-Kennedy, M., & McGowan, R. S. (1989). The emergence of phonetic segments: evidence from the spectral structure of fricative-vowel syllables spoken by children and adults. *Journal of Speech Language and Hearing Research, 32*, 120 – 132.

Noreña, A.J., Gourévitch, B., Aizawa, N., & Eggermont, J.J. (2006). Spectrally enhanced acoustic environment disrupts frequency representation in cat auditory cortex. *Nature Neuroscience, 9*, 932–939.

Norris, D., McQueen, J. M., & Cutler, A. (2003) Perceptual learning in speech. *Cognitive Psychology, 47*, 204-238.

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*, 104-114.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General, 115*, 39-57.

Nosofsky, R. M., & Zaki, S. R. (2002). Exemplar and prototype models revisited: Response strategies, selective attention, and stimulus generalization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*(5), 924-940.

O'Neill, W. E. (1995). The bat auditory cortex. In A. N. Popper & R. R. Fay (Eds.), *Hearing by bats* (pp. 416-480). New York: Springer Verlag.

O'Neill, W. E., & Suga, N. (1979). Target range-sensitive neurons in the auditory cortex of the mustached bat. *Science, 203*, 69-73.

O'Neill, W. E., & Suga, N. (1982). Encoding of target-range information and its representation in the auditory cortex of the mustached bat. *Journal of Neuroscience, 47*, 225-255.

Ohl, F. W., & Scheich, H. (1997). Orderly cortical representation of vowels based on formant interaction. *Proceedings of the National Academy of Sciences of the United States of America, 94* (17), 9440-9444.

Olsen, J. F. (1994). Medial geniculate neurons in the squirrel monkey sensitive to inter-component delays that categorize species-typical calls. *Midwinter Meeting of the Association for Research in Otolaryngology*, p. 21(A).

Olsen, J. F., & Suga, N. (1991a). Combination-sensitive neurons in the medial geniculate body of the mustached bat: Encoding of relative velocity information. *Journal of Neurophysiology, 65*, 1254-1274.

Olsen, J. F., & Suga, N. (1991b). Combination-sensitive neurons in the medial geniculate body of the mustached bat: Encoding of target range information. *Journal of Neurophysiology, 65*, 1275-1296.

Palombi, P.S., Backoff, P.M., & Caspary, D.M. (2001). Responses of young and aged rat inferior colliculus neurons to sinusoidally amplitude modulated stimuli. *Hearing Research*, *153*, 174-180.

Pegg, J. E., & Werker, J. F. (1997). Adult and infant perception of two English phones. *Journal of the Acoustical Society of America, 102*, 3742-3753.

Phan, M. L., & Recanzone, G. H. (2007). Single-neuron responses to rapidly presented temporal sequences in the primary auditory cortex of the awake macaque monkey. *Journal of Neurophysiology, 97*, 1726-1737.

Pienkowski, M., & Eggermont, J. J. (2009). Long-term, partially-reversible reorganization of frequency tuning in mature cat primary auditory cortex can be induced by passive exposure to moderate-level sounds. *Hearing Research, 257*, 24-40.

Pierrehumbert, J. B. (2001). Exemplar dynamics: Word frequency, lenition, and contrast. In J. Bybee, & P. Hopper (Eds.) *Frequency effects and the emergence of linguistic structure* (pp. 137-157). Amsterdam: John Benjamins.

Pierrehumbert, J. B. (2002). Word-specific phonetics. In C. Gussenhoven, & N. Warner (Eds.), *Laboratory phonology*, Vol. VII (pp. 101–139). Berlin: Mouton de Gruyter.

Pierrehumbert, J. B. (2006). The next toolkit. Journal of Phonetics, 34, 516-530.

Pisoni, D. B., & Tash, J. (1974). Reaction times to comparisons within and across phonetic categories. *Perception & Psychophysics, 15* (2), 285–290.

Qin, L., Chimoto, S., Sakai, M., & Sato, Y. (2004). Spectral-shape preference of primary auditory cortex neurons in awake cats. *Brain Research, 1024*, 167-175.

Quinn, P. C. (2004). Spatial representation by young infants: Categorization of spatial relations or sensitivity to a crossing primitive? *Memory & Cognition, 32* (5), 852-861.

Randall, D., Burggren, W., & French, K. (1997). *Eckert animal physiology: mechanisms and adaptations* (4th ed.). New York: W. H. Freeman and Company.

Rauschecker, J. P., & Singer, W. (1981). The effects of early visual experience on the cat's visual cortex and their possible explanation by Hebb synapses. *Journal of Physiology*, *310*, 215-239.

Rauschecker, J. P., Tian, B., & Hauser, M. (1995). Processing of complex sounds in the macaque nonprimary auditory cortex. *Science, 268*, 111-114.

Razak, K. A., & Fuzessery, Z. M. (2006). Neural mechanisms underlying selectivity for the rate and direction of frequency-modulated sweeps in the auditory cortex of the pallid bat. *Journal of Neurophysiology, 96*, 1303-1319.

Rees, A., & Moller, A.R. (1983). Responses of neurons in the inferior colliculus of the rat to AM and FM tones. *Hearing Research*, *10*, 301-330.

Rose, S. P. R. (2008). Memory beyond the synapse. In R. Douglas Fields (Ed.) *Beyond the synapse: cell-cell signaling in synaptic plasticity*. Chapter 1. Cambridge: Cambridge University Press.

Rossing, T. D., & Houtsma, A. J. (1986). Effects of signal envelope on the pitch of short sinusoidal tones. *Journal of Acoustical Society of America*, *79*, 1926-1933.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-olds. *Science*, *274*, 1926–1928.

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word segmentation: The role of distributional cues. *Journal of Memory and Language, 35* (4), 606-621.

Sagey, E. (1986). *The representation of features and relations in nonlinear phonology*. Doctoral dissertation, MIT. New York: Garland Press, 1991.

Scheich, H. (1991). Auditory cortex: comparative aspects of maps and plasticity. *Curr Opin Neurobiology, 1*, 236-247.

Schreiner, C. E., & Urbas, J. V. (1986). Representation of amplitude modulation in the auditory cortex of the cat: I. The anterior auditory field (AAF). *Hearing Research, 21*, 227-241.

Schuller, G., O'Neill, W. E., & Radthe-Schuller, S. (1991). Facilitation and delay sensitivity of auditory cortex in CF-FM bats, *Rhinolophus vouxi* and *Pteronotus p. parnellii*. *European Journal of Neuroscience, 3*, 1165-1181.

Scott, D. J., Stohler, C. S., Egnatuk, C. M., Wang, H., Koeppe, R. A., & Zubieta, J. (2007). Individual differences in reward responding explain placebo-induced expectations and effects. *Neuron, 55* (2), 325-336.

Seung, H.S., & Sompolinsky, H. (1993). Simple models for reading neuronal population codes. Proceedings of the National Academy of Sciences. USA, 90, 10749-10753.

Shadle, C. H. (1990). Articulatory-acoustic relationships in fricative consonants. In W. Hardcastle & A. Marchal (Eds.) *Speech production and speech modeling* (pp. 187 – 209). Kluwer: Dordrecht.

Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science, 270*, 303-304.

Shimizu, K. (1996). A cross-language study of voicing contrasts of stop consonants in Asian languages. Tokyo: Seibido Publishing.

Silva, A. J., Zhou, Y., Rogerson, T., Shobe, J., & Balaji, J. (2009). Molecular and cellular approaches to memory allocation in neural circuits. *Science, 326*, 391-395.

Simos, P. G., Diehl, R. L., Breier, J. I., Molis, M. R., Zouridakis, G., & Papanicolaou, A. C. (1998). MEG correlates of categorical perception of a voice onset time continuum in humans. *Cognitive Brain Research, 7*, 215-219.

Sinnott, J. M. & Adams, F. S. (1987). Differences in human and monkey sensitivity to acoustic cues underlying voicing contrast. *Journal of the Acoustical Society of America, 82,* 1539-1547.

Sinnott, J. M., & Brown, C. H. (1997). Perception of the English liquid /ra-la/ contrast by humans and monkeys. *Journal of the Acoustical Society of America, 102*, 588-602.

Sinnott, J., Brown, C., & Borneman, M. (1997). Effects of syllable duration on stop-glide identification in syllable-initial and syllable-final position by humans and monkeys. *Perception & Psychophysics, 60*, 1032-1043.

Srivatsun, S., & Wang, X. (2009). Nonlinear spectrotemporal interactions underlying selectivity for complex sounds in auditory cortex. *Journal of Neuroscience, 29* (36), 11192-11202.

Steinlen, A. K. (2002). A cross-linguistic comparison of the effects of consonantal contexts on vowels produced by native and non-native speakers. Unpublished doctoral dissertation, Aarhus University, Denmark.

Steinschneider, M., Schroeder, C. E., Arezzo, J. C., & Vaughan Jr., H. G. (1995). Physiologic correlates of the voice onset time boundary in primary auditory cortex (A1) of the awake monkey: temporal response patterns. *Brain and Language, 48*, 326-340.

Stevens, K. N. (1985). Evidence for the role of acoustic boundaries in the perception of speech sounds. In V. A. Fromkin (Ed.) Phonetic Linguistics: Essays in Honor of Peter Ladefoged, (pp. 243-255). Orlando: Academic Press.

Stevens, K. N., & House, A. S. (1963). 'Perturbation of vowel articulations by consonantal context: An acoustical study. *Journal of Speech and Hearing Research, 6*, 111 – 128.

Stevens, S. S., Volkmann, J., & Newman, E. B. (1937) A scale for the measurement of the psychological magnitude pitch. *Journal of the Acoustical Society of America, 8*, 185-190.

Strange, W., Bohn, O.-S., Nishi, K., & Trent, S. (2005). Contextual variation in the acoustic and perceptual similarity of North German and American English vowels. *Journal of the Acoustical Society of America, 118*, 1751-1762.

Strevens, P. (1960). Spectra of fricative noise in human speech. *Language and Speech 3*, 32 – 49.

Suga, N. (1969). Classification of inferior collicular neurons of bats in terms of responses to pure tones, FM sounds and noise bursts. *Journal of Physiology, 200*, 555-574.

Suga, N. (1973). Feature extraction in the auditory system of bats. In A. Møller (Ed.), *Basic mechanisms in hearing* (pp. 675-744). New York: Academic.

Suga, N. (1978). Specialization of the auditory system for reception and processing of species-specific sounds. *Proceedings of the Federation of the American Society of Experimental Biology, 37*, 2342-2354.

Suga, N. (1982). Functional organization of the auditory cortex representation beyond tonotopy in the bat. In C. N. Woolsey (Ed.), *Cortical sensory organization* (Vol 3) (pp. 157-218). Totowa, NJ: Humana.

Suga, N. (1984). The extent to which biosonar information is represented in the bat auditory cortex. In G. M. Edelman, W. E. Gall & W. M. Cowan (Eds.) *Dynamic aspects of neocortical function* (pp. 315-373). New York: Wiley.

Suga, N. (1988). Auditory neuroethology and speech processing: Complex sound processing by combination-sensitive neurons. In G. M. Edelman, W. E. Gall, & W. M. Cowan (Eds.) *Functions of the auditory system* (pp. 679-720). New York: Wiley.

Suga, N. (2006). Basic acoustic patterns and neural mechanisms shared by humans and animals for auditory perception. In S. Greenberg & W. A. Ainsworth (Eds.) *Listening to speech: An auditory perspective* (pp. 159-181). Mahwah, NJ: Lawrence Erlbaum.

Suga, N., & Horikawa, J. (1986). Multiple time axes for representation of echo delays in the auditory cortex of the mustached bat. *Journal of Neurophysiology, 55*, 776-805.

Suga, N., & O'Neill, W. E. (1979). Neural axis representation target range in the auditory cortex of the mustached bat. *Science, 206*, 351-353.

Suga, N., O'Neill, W. E., Kujirai, K., & Manabe, T. (1983). Specificity of combination-sensitive neurons for processing of complex biosonar signals in the auditory cortex of the mustached bat. *Journal of Neurophysiology, 49*, 1573-1626.

Suga, N., O'Neill, W. E., & Manabe, T. (1978). Cortical neurons sensitive to combinations of information-bearing elements of biosonar signals in the mustached bat. *Science, 200*, 778-781.

Suga, N., O'Neill, W. E., & Manabe, T. (1979). Harmonic-sensitive neurons in the auditory cortex of the mustached bat. *Science, 203*, 270-274.

Summerfield, Q., & Haggard, M. (1977). On the dissociation of spectral and temporal cues to the voicing distinction in initial stop consonants. *Journal of the Acoustical Society of America 62*, 435–448.

Sundara, M. (2005). Acoustic-phonetics of coronal stops: A cross-language study of Canadian English and Canadian French. *Journal of the Acoustical Society of America, 118* (2), 1026-1037.

Sussman, H. M. (1986). A neuronal model of vowel normalization and representation. *Brain and Language, 28*, 12-23.

Sussman, H. M. (1989). Neural coding of relational invariance in speech: Human language analogs to the barn owl. *Psychological Review, 96*, 631 – 642.

Sussman, H. M. (1994). The phonological reality of locus equations across manner class distinctions: Preliminary observations. *Phonetica, 51*, 119 –131.

Sussman, H. M. (1999). A neural mapping hypothesis to explain why velar stops have an allophonic split. *Brain and Language, 70*, 294-304.

Sussman, H. M. (2002). Representation of phonological categories: A functional role for auditory columns. *Brain and Language, 80*, 1-13.

Sussman, H. M., & Shore, J. (1996). Locus equations as phonetic descriptors of consonantal place of articulation. *Perception and Psychophysics, 58,* 936 – 946.

Sussman, H. M., Fruchter, D., & Cable, A. (1995). Locus equations derived from compensatory articulation. *Journal of the Acoustical Society of America, 97* (5), 3112-3124.

Sussman, H. M., Hoemeke, K. A., & Ahmed, F. S. (1993). A cross-linguistic investigation of locus equations as a phonetic descriptor for place of articulation. *Journal of Acoustical Society of America*, *94*, (3), 1256-1268.

Sussman, H. M., McCaffrey, H. A., & Matthews, S. A. (1991). An investigation of locus equations as a source of relational invariance for stop place categorization. *Journal of Acoustical Society of America*, *90*, (3), 1309-1325.

Sussman, J. E., & Lauckner-Morano, V. J. (1995). Further tests of the "perceptual magnet effect" in the perception of [i]: Identification and change/no-change discrimination. *Journal of the Acoustical Society of America, 97* (1), 539-552.

Sussman, J., & Gekas, B. (1997). Phonetic category structure of [I]: extent, best exemplars, and organization. *Journal of Speech, Language & Hearing Research*, *40* (6), 1406-1424.

Swingley, D., & Aslin, R. N. (2002). Lexical neighborhoods and the word-form representations of 14-month-olds. *Psychological Science, 13* (5), 480–484.

Tanaka, H., Wong, D., & Taniguchi, I. (1992). The influence of stimulus duration on the delay tuning of cortical neurons in the FM bat, *Myotis lucifugus*. *Journal of Comparative Physiology A, 171,* 29-40.

Thyer, N., Hickson, L., & Dodd, B. (2000). The perceptual magnet effect in Australian English vowels. *Perception & Psychophysics*, *62* (1), 1-20.

Titze, I. R. (1994). *Principles of voice production*. Prentice Hall, Englewood Cliffs, NJ.

Tomiak, G. R. (1990). An acoustic and perceptual analysis of the spectral moments invariant with voiceless fricative obstruents. Doctoral dissertation, SUNY Buffalo.

Toro, J. M., Sinnett, S., & Soto-Faraco, S. (2005). Speech segmentation by statistical learning depends on attention. Cognition, 97 (2), B25-B34.

Tremblay, K. L., Shahin, A. J., Picton, T., & Ross, B. (2009). Auditory training alters the physiological detection of stimulus-specific cues in humans. *Clinical Neurophysiology, 120*, 128-135.

Tsao, F.-M. (2001). The effects of language experience on the perception of affricate and fricative consonants in English-speaking and Mandarin-speaking adults and young infants. Doctoral dissertation. University of Washington.

Tsao, F., Liu, H., & Kuhl, P. K. (2006). Perception of native and non-native affricate-fricative contrasts: Cross-language tests on adults and infants. *Journal of the Acoustical Society of America, 120* (4), 2285-2294.

Tsushima, T., Takizawa, O., Sasaki, M., Shiraki, S., Nishi, K., Kohno, M., Menyuk, P., & Best, C. (1994). Discrimination of English /r–l/ and /w–y/ by Japanese infants at 6–12 months: Language-specific developmental changes in speech perception abilities. *International Conference on Spoken Language,Vol. S28F-1*, Yokohama, Japan, 1695–1698.

Tsuzuki, K., & Suga, N. (1988). Combination-sensitive neurons in the ventro-anterior area of the auditory cortex of the mustached bat. *Journal of Neurophysiology, 60*, 1908-1923.

Tucker, B. V., & Warner, N. (2007). Inhibition of processing due to reduction of the American English flap. In *Proceedings of the XVIth International Conference of Phonetic Science*, Saarbrücken 6-10 August 2007, 1949-1952.

Tulving, E. (2002). Episodic memory: From mind to brain. *Annual Review of Psychology, 53*, 1-25.

Vaissière, J. (2006). *Que sais-je? La phonétique*. Paris: Presses Universitaires de France.

Vance, T. J. (1987) *An introduction to Japanese phonology*. Albany: State University of New York.

Vitevich, M., & Luce, P. (1998). When words compete: Levels of processing in perception of spoken words. *Psychological Science, 9*, 325–329.

Vitevich, M., Luce, P., Pisoni, D., & Auer, E. T. (1999). Phonotactics, neighborhood activation, and lexical access for spoken words. *Brain and Language, 68*, 306–311.

Wagner, H., Takahashi, T. & Konishi, M. (1987). Representation of interaural time difference in the central nucleus of the barn owl's inferior colliculus. *Journal of Neuroscience, 7*, 3105-3116.

Wallace, M. N., Shackleton, T. M., & Palmer, A. R. (2002). Phase-locked responses to pure tones in the primary auditory cortex. *Hearing Research, 163*, 1-12.

Wang, Q. (2008). *Perception of English stress by Madarin Chinese learners of English: An acoustic study*. Ph.D. Dissertation. University of Victoria, Victoria, Canada. http://hdl.handle.net/1828/1282

Wang, X. & Munro, M. J. (2004). Computer-based training for learning English vowel contrasts. *System, 32*, 539-552.

Wang, Y., Spence, M. M., Jongman, A., & Sereno, J. A. (1999). Training American listeners to perceive Mandarin tones. *Journal of the Acoustical Society of America, 106* (6), 1063-1074.

Warner, N., & Tucker, B.V. (2007). Categorical and Gradient Variability in Intervocalic Stops. Talk given at the annual meeting of the *Linguistic Society of America*.

Webster, M. A, Kaping, D., Mizokami, Y., & Duhamel, P. (2004). Adaptation to natural facial categories. *Nature, 428*, 557–561.

Werker, J. F. (1995). Exploring developmental changes in cross-language speech perception. In L. R. Gleitman, & M. Liberman (Eds.) *Language: An invitation to cognitive science, Part 1* (2nd ed., pp. 87-106). Cambridge, MA: MIT Press.

Werker, J. F., & Curtin, S. (2005). PRIMIR: A developmental framework of infant speech processing. *Language Learning and Development, 1* (2), 197-234.

Werker, J. F., Fennell, C. T., Corcoran, K. M., & Stager, C. L. (2002). Infants' ability to learn phonetically similar words: Effects of age and vocabulary size. *Infancy, 3* (1), 1–30.

Werker, J. F., & Lalonde, C. E. (1988). Cross-language speech perception: Initial capabilities and developmental change. *Development Psychology, 24* (5), 672-683.

Werker, J. F., & Logan, J. S. (1985). Cross-language evidence for three factors in speech perception. *Perception & Psychophysics, 37* (1), 35-44.

Werker, J. F., & Tees, R. C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development, 7* (1), 49-63.

White, K. S., Peperkamp, S., Kirk, C., & Morgan, J. L. (2008). Rapid acquisition of phonological alternations by infants. *Cognition, 107* (1), 238-265.

Wilde, L. (1993). Inferring articulatory movements from acoustic properties at fricative vowel boundaries. *Journal of the Acoustical Society of America, 94*, 1881.

Wong, D., Maekawa, M., & Tanaka, H. (1992). The effect of pulse repetition rate on the delay sensitivity of neurons in the auditory cortex of the FM bat *Myotis lucifugus. Journal of Comparative Physiology A, 170,* 393-402.

Woolsey, C. N., & Walzl, E. M. (1942). Topical projection of nerve fibers from local regions of cochlea to cerebral cortex of the cat. *Bulletin of the Johns Hopkins Hospital, 71*, 315-344.

Xu, Y., Gandour, J. T., & Francis, A. L. (2006). Effects of language experience and stimulus complexity on the categorical perception of pitch direction. *Journal of the Acoustical Society of America, 120* (2), 3649-3658.

Yan, J., & Suga, N. (1996). Corticofugal modulation of time-domain processing of biosonar information in bats. *Science, 273*, 1100-1103.

Yeou, M. (1997). Locus equations and the degree of coarticulation of Arabic consonants. *Phonetica 54*, 187-202.

Zlatin, M. A. (1974). Voicing contrast: perceptual and productive voice onset time characteristics of adults. *Journal of the Acoustical Society of America, 56*, 981-994.