This manuscript was a book chapter contribution.

Citation in APA style:


**Brain plasticity and phonetic training for English-as-a-Second-Language learners**

Yang Zhang¹, Bing Cheng²
Department of Speech-Language-Hearing Sciences & Center for Neurobehavioral Development, University of Minnesota, Minneapolis, MN 55455, USA
Department of English, School of International Studies, Xi’an Jiaotong University, Xi’an, Shaanxi, 710049, China.

Address for Correspondence:
Yang Zhang, Ph.D.
Department of Speech-Language-Hearing Sciences,
115 Shevlin Hall
University of Minnesota
Minneapolis, MN 55455
USA
E-mail: zhang470@umn.edu
Abstract

English as a second language (ESL) education has gained an increasingly important role in career development in science, business, and industry on the global stage. One great challenge for adult ESL learners is to reduce or eliminate “foreign accent” in their English pronunciation. Decades of behavioral and brain research have shown that language experience early in life leaves an indelible mark on speech perception and production. There is converging evidence for a self-reinforcing bonding process in the brain in the course of largely implicit learning, leading to a system that is neurally committed to the articulatory, acoustic, and perceptual properties of sounds and sound patterns in the first language. As a consequence, it is often difficult for adults to learn the speech sounds in a second language that do not conform to the phonology of the first language. This chapter examines the underlying causes for the foreign accent problem in second language learning and discusses the approaches to promote brain plasticity through phonetic training. The first section provides a summary of the main research findings to illustrate the role of phonetic knowledge in language learning, the neural mechanisms of speech perception, and the relationship between perception and production. The second section outlines a theoretical framework of brain plasticity for phonetic learning, featuring quantifiable measures to test relevant hypotheses about second language acquisition in terms of neural sensitivity, neural efficiency, neural specificity, and neural connectivity and their behavioral correlates. The third section introduces a synergistic Speech Assessment and Training (SAT) software program, to overcome first-language interference. The chapter concludes with a discussion of the implications for second-language education and future research.

Keywords: language acquisition; neural commitment; brain imaging; speech perception; speech production; speech training
0 Introduction

Adult ESL learners face many great challenges in improving their English pronunciation and fluency. Common pronunciation errors include nonnative substitutions of individual consonants and vowels, consonant cluster reduction, and epenthetic vowel insertion. Parallel to the “foreign accent” problem, there is a “foreign listening syndrome” characterized by reduction in accuracy and speed at the perceptual level (Mehler et al., 1994). For instance, many adult Japanese speakers cannot hear the distinction between /r/ and /l/ sounds in English despite years of school education or immigration to an English-speaking country (Takagi, 2002). In contrast, young children, whose linguistic knowledge and perceptual/cognitive systems are presumably not as developed as adults, can acquire native-like performance without sharing the same struggle.

Many researchers have proposed that inaccurate perceptual representations lead to non-native production in second language learners (Polivanov, 1931, Trubetzkoy, 1969, Strange, 1995, Guion et al., 2000). Cross-language perception research starting from the 1960s has provided solid evidence in support of this view (Strange, 1995). Nevertheless, not all the sounds that are correctly perceived can be correctly produced. In certain cases, accurate production of the L2 sounds may counter-intuitively precede their perception (Goto, 1971). Some bilingual studies have also shown that production of the less proficient language can be more accurate than its perception (Caramazza et al., 1973, Elman, Diehl and Buchwald, 1977).

A proper understanding of the foreign accented speech and foreign listening syndrome requires thorough investigation of how production and perception are linked in developing and mature minds in language acquisition. The perceptual advantage for the native language gains early ground during infancy. Developmental research has documented a clear perceptual transition from “language-general” to “language-specific” within the first year of life (Werker and Tees, 2005). Gains in L1 (first language) is accompanied with losses for L2, which can be behaviorally measured as early as 6 months of age for vowels and 10-12 months of age for consonants (Kuhl and Rivera-Gaxiola, 2008). Studies of bilingualism and perceptual training on adult ESL learners vividly illustrate the profound effects of first-language interference in speech perception. Despite high proficiency in the second language, there is strong evidence of first-language dominance in perception (Pallier, Bosch and Sebastian-Gallés, 1997, Bosch, Costa and Sebastián-Gallés, 2000). Similarly, perceptual training studies and other speech research have shown limited success in improving listeners’ perception and production of the difficult nonnative contrasts (Jamieson and Morosan, 1986, Akahane-Yamada et al., 1997, Tremblay et al., 1997, Bradlow et al., 1999b, Zhang et al., 2000, McCandliss et al., 2002, Wang et al., 2003, Iverson, Hazan and Bannister, 2005, Hazan et al., 2006, Pruitt, Jenkins and Strange, 2006, Zhang et al., 2009).

Advancement in brain imaging techniques has opened a new venue for studying the brain mechanisms in support of speech and language. The online non-invasive measures of spatial and temporal dynamics in the brain provide critical information for a better understanding of perceptual and motor processes that mediate speech perception and production. Cognitive brain research has shown accumulating evidence of “memory traces” for language-specific phonemes in adult listeners and its emergence in infants before the age of twelve months (Näätänen et al., 1997, Cheour et al., 1998). The results indicate strong “neural commitment” to the sounds and sound patterns of the native language (Zhang et al., 2005). The Native-Language-Neural-Commitment (NLNC) theory asserts that phonetic learning is an implicit self-reinforcing computational process, which promotes learning patterns that conform to the first language and interferes with those that do not (Kuhl et al., 2008). Furthermore, the behavioral and neural sensitivity measures for phonetic discrimination are good predictors of the development of higher-level language skills. However, neural commitment is by no means equivalent to irreversible hardware formatting of the computational neural machinery into the phonological structure of the first language. There are a number of issues that remain to be resolved in the NLNC theory. Research has demonstrated the human brain shows substantial neuroplasticity in emergent cortical connections or functional reorganization early in life as well as in adulthood (Neville and Bavelier, 1998, Golestani et al.,
A number of effective speech training methodologies have been identified with an aim to promote neuroplasticity for second language learning and intervention (Iverson, Hazan and Bannister, 2005, Zhang and Wang, 2007).

This chapter provides an integrated discussion of key empirical and theoretical issues for a proper characterization of neuroplasticity in speech learning as related to ESL. The first section provides a selective review of findings to illustrate the role of phonetic knowledge in language learning, the neural mechanisms of speech perception, and the relationship between perception and production. The second section outlines a theoretical framework of brain plasticity for phonetic learning, featuring quantifiable measures to test relevant hypotheses about second language acquisition in terms of neural sensitivity, neural efficiency, neural specificity, and neural connectivity and their behavioral correlates. The third section introduces a synergistic perceptual training software program to overcome first-language interference. The chapter concludes with a discussion of the limitations of the current approach and its implications for second-language education and future research.

1 Neural Commitment: The profound effects of language experience

Speech perception involves neural coding at the level of peripheral and central auditory processing. At a simple approximation, the peripheral auditory system extracts the acoustic information from speech sounds, and the cortical central system processes language-specific category information for higher-order sound-meaning mapping. Advances in neurophysiological methods have begun to refine and add to the knowledge of temporal dynamics and functional neuroanatomy for linguistic processing.

1.1 Two Basic Perceptual Phenomena

Behavioral research in the past six decades has demonstrated two basic phenomena in human listeners’ organization of phonetic categories. These phenomena characterize the listener’s ability to discriminate and categorize speech sounds, and how these abilities change with linguistic experience. The first is “categorical perception” (CP), a phenomenon found when using a synthetic stimulus continuum that embodies acoustic changes from one speech sound to another in equalized steps. In CP, subjects show enhanced discrimination for pairs equally distant in acoustic steps when the stimulus pair crosses the 50% phonetic identification boundary (Liberman, 1957). In other words, CP selectively shows better discrimination sensitivity for between-category stimuli when compared to within-category stimuli. The key component of CP was the limitation of discrimination by identification; listeners’ discrimination was “categorical.”

The second is known as the “perceptual magnet effect” (PME), a phenomenon based on two findings concerning the structure and relationship of speech sounds within a single phonetic category. First, listeners can consistently judge the category goodness of individual speech tokens and their judgments show that certain sounds are particularly representative of one phonetic category as opposed others (Samuel, 1982, Grieser and Kuhl, 1989, Volaitis and Miller, 1992). Second, category goodness strongly influences human listeners’ sensitivities to within-category differences (Grieser and Kuhl, 1989, Kuhl, 1991b, Kuhl, 1992). In demonstrating the PME, subjects show reduced sensitivity to acoustic differences near the best instances (prototypes) and increased sensitivity in the vicinity of poor exemplars. Kuhl (1991) referred to the perceptual effects of speech prototypes as a “warping” of phonetic perception that was based on linguistic experience.

While both categorical perception and the perceptual magnet effect predict poor within-category discrimination and good between-category discrimination, there are fundamental theoretical differences. The two critical aspects of categorical perception are (a) the innate existence of phonetic boundaries and (b) the functional equivalence of sounds within a phonetic category produced by reliance on an innate articulatory representation of phonetic units that can be altered by learning experience. In contrast, the two critical aspects of the perceptual magnet effect are (a) a learned internal structure of phonetic categories based on statistical analyses of the input and (b) a “warping” of perception that results from
“neural commitment” to a particular kind of acoustic analysis. Categorical perception does not address the issue of internal organization of a phonetic category because it treats within-category discrimination as one unitary phenomenon as against cross-category discrimination. Computational modeling for CP and PME indicate that speech perception involves a statistical mapping process that weighs the acoustic, psychophysical, and linguistic properties of any input against speech sound target categories and contextual phonological and lexical targets. Data from nonhuman subjects and nonspeech stimuli suggest that the phoneme boundary effect and the perceptual magnet effect could coexist at different levels of phonetic organization, one stemming at birth from basic properties of the auditory system and the other from learning speech categories for language acquisition. For example, while CP was found in humans as well as birds, chinchillas, and monkeys, PME was found only in human infants and adults, not in monkeys. These two phenomena have generated a vast amount of research and controversy (Harnad, 1987, Iverson and Kuhl, 2000, Feldman, Griffiths and Morgan, 2009).

1.2 Effects of Linguistic Experience

Both CP and PME demonstrate that human listeners show discontinuous sensitivity for continuously varying acoustic parameters in speech. We are endowed with innate auditory sensitivities to discriminate speech sounds. Infants younger than 6 months of age (including newborns as young as 4 days old) are able to discriminate native as well as nonnative phonemes categorically (Eimas et al., 1971). Infants between 10-12 months old have been shown to exhibit adult-like perception with poor discrimination at nonnative contrasts, especially those whose shared features have close resemblance to a single native phonetic category (Werker and Tees, 1984, Best, 1994). Consistent with developmental data, numerous studies indicate that human adults show categorical perception for phonetic units in their native language but not for units from a nonnative language.

Phonemic relevance, phonetic familiarity, acoustic and psychophysical factors all contribute to perceptual changes that take place in the course of language learning (Polka, 1991; Werker & Tees, 1984). Behavioral data suggest that the language-specific pattern of perception appears to result from selective attentional mechanisms and not a change in basic auditory sensitivities. For example, Miyawaki et al. (1975) and Zhang et al. (2005) found that the American subjects showed typical CP for the /ra-la/ phonetic continuum whereas Japanese subjects did not. However, the two subject groups had nearly identical discrimination performance on the stimuli based on the acoustic component of third formant, which was critical to the /r-l/ distinction. Cross-language studies of the perceptual magnet effect provide further evidence that the PME is specific to native language categories (Iverson and Kuhl, 1996, Bosch et al., 2000). Multidimensional scaling analysis showed that the perception of English /r/ and /l/ in Japanese, German, and American adult listeners was strongly influenced by the native language magnets for the speakers of each language (Figure 1). In particular, Japanese adults showed high sensitivity to an acoustic cue, F2, which is irrelevant to the English /r/-/l/ categorization. German adults were similar to American adults, who were highly sensitive to the critical acoustic cue of F3 transition. Despite the similarity in the American and German adults, there were also noticeable differences in their perceptual warping maps for the /r/ and /l/ tokens.

In cognitive brain research, Näätänen et al. (1997) provided strong evidence for the existence of language-specific phoneme representations in the auditory cortex. The study used the well-known MMN (mismatch negativity) paradigm, in which repeated presentations of the same sound was occasionally replaced with a different sound. The MMN response peaks at 100-250 ms after the onset of a change in the stimulus, and the MMN magnitude provides an objective measure of neural sensitivity, independent of attentional processing, to perceived acoustic change with good correlation with psychophysical discrimination (Näätänen, 2007). Despite the fact that the acoustic difference in the nonnative contrast was larger than that of the native contrast, stronger language-specific MMN activity for native vowel discrimination was found in the left auditory cortex. Cheour et al. (1998) used exactly the same stimuli and paradigm on infants to determine the age at which the language-specific memory traces of the mother tongue emerge. These MMN results suggest that language-specific representations are developed between 6 and 12 months of age. Similar MMN results have been reported in many other adult studies (Dehaene-Lambertz, 1997, Winkler et al., 1999, Dehaene-Lambertz, Dupoux and Gout, 2000, Rivera-Gaxiola et al.,
2000, Sharma and Dorman, 2000, Zhang et al., 2005). The language samples of these studies covered a wide range, including English, Estonian, Finnish, French, Hindi, Hungarian, Japanese, and Spanish.

Does L2 learning produce permanent neural representations for nonnative speech in the brain? Winkler et al. (1999) found that MMN for the Finnish /e/-/æ/ contrast was elicited in Hungarians who could speak fluent Finnish, but not in monolingual Hungarians. Furthermore, the MMN response for the fluent Finnish-speaking Hungarians was almost identical to that shown in the Finnish subjects, suggesting that language learning may lead to long-term changes in neural representations of speech at the preattentive level. Two parallel processes contribute to the MMN response of speech discrimination (Näätänen et al., 1997). One is the process of acoustic change detection with bilateral activation, and the other is phoneme-specific processing, which is lateralized in the left auditory cortex. According to this theory, phonetic learning or training effects would be prominent in the left hemisphere as compared with the right hemisphere. Tremblay et al. (1997, 1998) were able to show more pronounced training effects in MMN over the left frontal cortex than the right. In two recent studies to train Japanese subjects to learn the /l-t/ contrast, Zhang et al. (2005, 2009) found a bilateral activation pattern of MMNm in the auditory cortex with enhanced activity in the left hemisphere relative to the right hemisphere after training. These results suggest that successful training can result in phonetic processing in the left hemisphere.

Taken together, linguistic experience may alter not only higher-level processes but also attentional and pre-attentional processes so that reduction in neural and behavioral sensitivities for non-native phonetic contrasts becomes difficult to reverse in adulthood. This difficulty arises because of changes in perceptual resolution that do not show the same level of sensitivity to critical acoustic variations important for L2 speech contrasts.

1.3 ESL Learning under the influence of Native-Language-Neural-Commitment

Cross-linguistic data strongly suggest that adults may become neurally committed to a particular network structure dedicated to L1 analysis. In this perspective, the foreign accented speech and foreign listening syndrome are attributable to self-reinforcing perceptual interference rather than to any age-related biological limitations. This hypothesis was confirmed in recent neuroimaging studies (Zhang et al., 2005; 2009). In particular, the superior temporal and inferior parietal regions in the brain were identified to show more focal and efficient processing of speech in native speakers than nonnative speakers.

The NLNC theory aims to provide an integrated account for research findings from infants and adults regarding the profound effects of language experience on speech perception (Kuhl et al., 2008). In addition to behavioral measures, innovative imaging techniques and analysis methods have been developed to investigate the experience-driven changes. Four basic principles of NLNC emerge from this body of research, which hold important implications for ESL learning.

(a) Early learning produces neural commitment to the abstract phonetic units and their statistical and combinatorial patterns in the native language (Zhang et al., 2000, Zhang et al., 2005, Kuhl et al., 2008). Specific temporal and parietal regions in the brain become specialized for native language processing with focal and efficient activation.

(b) The effects of NLNC are self-reinforcing and bidirectional – it enhances the detection of compatible higher-order linguistic patterns, while at the same time hindering the detection of non-conforming patterns contained in foreign languages, as shown behaviorally (Iverson et al., 2003) and neurally at the preattentive level (Zhang et al., 2006).

(c) Neural commitment is not irreversible – Enriched exposure can induce substantial plasticity in the adult brain for second language learning.

(d) Neural commitment involves the binding of perception and action, which depends on social learning experience early in life (Imada et al., 2006, Kuhl, 2007).
These principles are consistent with the developmental framework that views language acquisition as a computational process to extract the abstract speech categories and higher-order structures.

Admittedly, the NLNC theory remains underdeveloped on a number of important issues. The first issue is to address how talker variability facilitates language learning. Sources of talker variability include age, gender, affect, and dialectal influences, which can be acoustically measured in fundamental frequency (interchangeably referred to as “pitch” in the literature), speaking rates, specific features such as voice onset time, frication noise, formant structures of vowels and consonants (Creel, Aslin and Tanenhaus, 2008). Many behavioral studies have shown data in favor of exemplar-based representations for speech. Talker variability can affect speech intelligibility for children and adult listeners because talker-specific information is encoded and retained in memory (Nygaard and Pisoni, 1995, Bradlow, Torretta and Pisoni, 1996, Remez, Fellowes and Rubin, 1997, Goldinger, 1998, Nygaard and Pisoni, 1998, Pierrehumbert, 2003, Allen and Miller, 2004, Eisner and McQueen, 2005, Burk et al., 2006, Johnson, 2006, Rosenblum, Miller and Sanchez, 2007). Furthermore, sub-phonemic variations have significant effects in behavioral responses (Andruski, Blumstein and Burton, 1994, McMurray, Tanenhaus and Aslin, 2002). The NLNC model incorporates the statistical learning mechanism computationally grounded in variability (Maye, Werker and Gerken, 2002, Vallabha et al., 2007). However, such an implicit mechanism cannot fully explain the role of social cognitive skills and the potential negative effects of high variability input in language learning (Kuhl, Tsao and Liu, 2003). Vocal variables such as gender and emotion in speech have not been fully addressed in statistical learning models.

In second language acquisition, nonnative speakers may be less affected by speech variability due to their greater exposure to variability in L1 and L2 and the shared phonetic space for the L1 and L2 phonetic categories (Wade, Jongman and Sereno, 2007). At least in the case of difficult L2 vowel distinctions, training with low-variability stimuli may have better results than training with high-variability stimuli. These data suggest that the different sound categories may differentially benefit from the existence of variability in the language input. Thus the NLNC theory needs to take into account category-specific mechanisms concerning how the amount of variability in language input influences learning outcome.

The second issue related to variability is to specify how exaggerated speech facilitates language learning. Enriched exposure such as infant-directed speech (IDS) contains not only exaggerated spectral and temporal cues but also and listener-oriented pitch variations. IDS-based input manipulation can enhance second-language phonetic learning in adulthood (Zhang et al., 2009). This idea is consistent with other speech learning models (Flege, 1995, Pisoni and Lively, 1995, Best, McRoberts and Goodell, 2001, McClelland, 2001, Escudero and Boersma, 2004, Vallabha and McClelland, 2007) and computational neural network models (Guenther and Gjaja, 1996, McClelland, 2001, de Boer and Kuhl, 2003, Kirchhoff and Schimmel, 2005b). But not all aspects of IDS facilitate learning. For example, heightened pitch range and contour do not necessarily assist phonetic categorization (Trainor and Desjardins, 2002, Kitamura and Burnham, 2003). Robust statistical computation derived from IDS does not necessarily lead to more accurate or efficient categorization (Kirchhoff and Schimmel, 2005a). Comparisons between infant-directed, foreigner-directed, animal-directed speech (Burnham, Kitamura and Vollmer-Conna, 2002, Uther, Knoll and Burnham, 2007) and studies on clear vs. conversational speech (Bradlow and Bent, 2002, Ferguson and Kewley-Port, 2007) suggest that linguistic and nonlinguistic modifications contribute differently to speech intelligibility and phonetic learning. Further research is necessary to investigate mechanisms supporting the differential contributions of linguistic, paralinguistic, and nonlinguistic processing in speech perception and production.

The third issue concerns how “prototypical” speech is developed and coded in the brain. The NLNC theory asserts one pivotal concept – the effect of perceptual magnets or phonetic prototypes that easily assimilate within-category variations of speech sounds (Kuhl, 1991a). Non-prototypes such as the difficult speech sounds in a foreign language do not have the same assimilatory power (Iverson et al.,
However, other studies did not replicate the effect (Sussman and Lauckner -Morano, 1995, Lively and Pisoni, 1997, Lotto, Kluender and Holt, 1998, Frieda et al., 1999). ERP/MEG studies also did not show uniform support (Aaltonen et al., 1997, Sharma and Dorman, 1998, Zhang, 2002). The operational definition of phonetic prototype did not take into account paralinguistic variations. Paradoxically, one outstanding feature of infant-directed speech, upon which the phonetic prototypes are based for early learning, is prosodic exaggeration in pitch range and level (Fernald and Kuhl, 1987, Fernald, 1993). Prosodic cues not only provide important information for gender discrimination and emotion recognition, but also affect language development, as well as social and musical learning (Murry and Singh, 1980, Fernald, 1989, Trehub et al., 1997, Trainor, Austin and Desjardins, 2000, Reissland, Shepherd and Herrera, 2003, Fu, Chinchilla and Galvin, 2004, Weger et al., 2007, Fu, Chinchilla and Galvin, 2004, Weger et al., 2007, Dara and Pell, 2008). Yet little is known about how paralinguistic processing affects within-category phonetic organization in the brain. Assuming the pivotal role of prototypical speech representation for accurate speech perception and production, it also remains unclear how phonetic prototypes for a second language can be established irrespective of the paralinguistic variations.

The fourth issue is to account for how perceptual-motor binding supports L1 acquisition and phonetic interactions between L1 and L2. Facial and articulatory motion not only provides cues for speech categories but also conveys important social signals such as gender, emotion, social eye gaze, and joint attention. The well-known McGurk effect shows the automaticity of audiovisual fusion in speech (McGurk and MacDonald, 1976). This phenomenon has been the basis for a great many studies on multisensory integration in speech perception, language learning and modeling (e.g., Rosenblum, Yakel and Green, 2000, Movellan and McClelland, 2001, Massaro, 2004, Skipper, Nusbaum and Small, 2005, Brancazio, Best and Fowler, 2006, Davis et al., 2008, Teinonen et al., 2008, Wilson, Molnar-Szakacs and Iacoboni, 2008). While the NLNC model asserts “socially-gated” language learning (Kuhl, 2007), there is no specification on how social signals may affect audiovisual integration for speech perception. Bimodal speech perception theories have not addressed how multisensory integration in speech processing can be affected by social signal processing in spoken language. Much less is known about how the brain resources for automatic audiovisual processing are employed to support phonetic interactions between L1 and L2.

In summary, speech research has found convincing evidence for neural commitment to sound patterns in L1. The profound effects of language experience on speech perception and production for L2 learning are well established. There are also a number of unresolved issues that require further in-depth investigations. For example, speech research on adopted children and adults has presented a problematic case for the NLNC theory (Pallier, 2003, Ventureyra, Pallier and Yoo, 2004). Behavioral and fMRI data from the participants who were adopted after the age of three indicate that early neural commitment to one’s native language may not be a simple unidirectional path towards L1 dominance. Language use and input in childhood may rewire the neurally committed system and reallocate its resources to reflect changes in behavior. The challenges that adult ESL learners face in overcoming foreign accented speech and perceptual difficulties provide important questions for behavioral and brain imaging research on speech perception and production.

1.4 The role of phonetic learning in language acquisition

Theorists and researchers are well aware of the gaps between acoustical properties of linguistic input, measures of speech perception, and data on language acquisition beyond the realm of phonetics. Are there close links among the various levels of linguistic processing? Recent human and computer simulations studies have provided insightful answers to this question.

First, maternal speech clarity, as measured by the degree of acoustic expansion of the vowel space, is significantly correlated with infants’ phonetic discrimination ability (Liu, Kuhl and Tsao, 2003). Second, phonetic learning as early as 6 months of age is significantly correlated with their later (under the age of 3) language comprehension and production skills in terms of vocabulary and syntax (Tsao, Liu and
Kuhl, 2004, Kuhl et al., 2005, Kuhl et al., 2006). Strikingly, an early competitive pattern between native and nonnative phonetic perception can be seen even in infants who are raised in monolingual families. At a group level, infants who showed better native phonetic discrimination at 7 months old had better performance in their later language skills as assessed by using the MacArthur-Bates Development Communicative Inventory (Fenson et al., 1993). On the contrary, infants who had better nonnative phonetic discrimination at 7 months old demonstrated reduced later language abilities in their first language development during the first three years (Fig. 1).

Consistent results were found in MMN measures (Molfese and Molfese, 1985, Molfese and Molfese, 1997, Molfese, 2000, Kuhl et al., 2004), showing that early neural sensitivity to phonetic categories is a reliable predictor of later language skills. These results, though at a different level, are highly compatible with Li and colleague’s computational modeling of lexical development in the first and second languages (Li, Zhao and MacWhinney, 2007).

Different underlying mechanisms have been proposed to account for the association between phonetic perception and language learning. Some researchers believe that the association is primarily based on low-level domain-general auditory mechanisms rather than infants’ phonetic knowledge (Visto, Cranford and Scudder, 1996, Tallal et al., 1998, Benasich and Tallal, 2002, Ulla et al., 2004, Benasich et al., 2006). Their data show that processing abilities of rapid auditory/visual information for both speech and nonspeech stimuli in early infancy are highly predictive of later language development and literacy including language delay and impairment. More importantly, although individual differences in speech and non-speech discrimination in infancy are both correlated with later linguistic outcome, non-speech perceptual ability – the fine acoustic discrimination of rapid tones – appears to have a stronger predictive role than the discrimination of consonant-vowel pairs. Support for the domain-general auditory learning mechanisms has also been found in some adult speech perception studies (Poldrack et al., 2001, Diehl, Lotto and Holt, 2004, McNealy, Mazzotta and Dapretto, 2006). For example, fMRI activities in the left inferior and middle frontal cortex that index the implicit learning of word boundaries in continuous speech are positively correlated with listeners’ rapid auditory processing skills (McNealy et al., 2006).

However, other researchers argue that the early language mapping process is specifically based on statistical, spectro-temporal, phonetic and phonotactic properties of the native language (Saffran, 2003, Kuhl et al., 2005). In particular, the differences in language development associated with early native vs. nonnative speech discrimination abilities could not be explained by the auditory mechanisms – because low-level auditory processing should equally apply to any language. There is also behavioral and neural evidence showing that speech perception can be selectively affected by language experience without a parallel compromise in auditory processing of nonspeech stimuli that simulate the essential spectral properties of speech (Miyawaki et al., 1975, Zhang et al., 2005). Literature review on the rapid auditory processing deficits further suggests that not all people with specific language impairment and dyslexia exhibit auditory deficits. In many cases, there can be little or no relationship between the severity of auditory deficits and language performance (Rosen, 2003). Therefore, the existence of significant statistical correlations between speech and nonspeech deficits as well as those between phonetic perception and language development should not be interpreted as causative or exclusive of cognitive and maturational involvement.

Despite the theoretical debate, both behavioral and neurophysiological data consistently demonstrate the pivotal role of phonetic perception in language learning and the existence of large individual differences. The empirical data provide support for both longitudinal and cross-sectional associations between phonetic perception and language skills. More importantly, speech modifications and acoustic enhancements, which can be found in child-directed speech at the phonetic as well as syntactic and semantic levels, not only facilitates language learning in normal children but also achieves some remarkable success in training second-language learners and treating children with language disabilities.
1.5 Relationship between speech perception and production

L2 research has shown inconsistent results regarding the relationship between perception and production. Most studies indicate that accurate perception is a pre-requisite for good production. There is also some counter evidence that L2 production may surpass perception (Bohn and Flege, 1996, Kluge et al., 2007). Moderate correlations between perception and production have been reported between perception and production (Flege, 1999, Rauber et al., 2005, Cheng and Zhang, 2009). In the course of L2 learning, it could be the case that perceptual skills improve faster than productive skills or vice versa. A number of factors need to be considered to study the links between perception and production. One confounding factor is that it is difficult to assess measurements in the two domains in a balanced way. Methodological diversity could have led to the diverging patterns obtained in the literature.

It is well-known that certain speech contrasts in English are difficult for ESL learners to perceive and produce. Early bilinguals generally have milder foreign accented speech than late bilinguals. The critical period hypothesis assigns a pivotal role to maturational constraints – the ability to attain native-like perception and production for a second language starts to decline by age 6 and beyond the age of 12, it deteriorates further regardless of the learners’ motivational level or opportunities available to practice the L2 sounds (Johnson and Newport, 1989, Long, 1990). Researchers attribute the putative end of critical period at around 12 years of age to decrease in neuroplasticity as a result of neurological maturation (Long, 1990, Birdsong and Molis, 2001). However, Grosjean challenged the conventional notion of age advantages in L2 learning by emphasizing the psychosocial factors that drive L2 learning (Grosjean, 1982). Flege and colleagues demonstrate that while the phonetic systems in adults is still malleable and incremental progress is attainable in L2 perception and production regardless of age limit, L2 learning may proceed nonlinearly at a much faster rate in the initial learning period (Flege, 2007).

Links between age of acquisition and degree of foreign accent may result from phonetic interactions between L1 and L2 in addition to maturational constraints. The notions of common phonological/phonetic space, phonetic/phonological similarity, category assimilation and dissimilation across linguistic systems are crucial to L2 speech theories. Best’s Perceptual Assimilation Model focuses on phonological alignment between non-native sounds and native phonological system (Best, 1992). Flege’s Speech Learning Model (SLM) emphasizes the acoustic similarities and phonetic differences upon which L2 target categories can be acquired (Flege, 1995). SLM specifically predicts that accurate production of L2 sounds cannot occur unless there is accurate perception. According to SLM, the learning process is marked by gradual experience-dependent changes for second language phonological acquisition rather than a categorical shift known as the “critical period” biologically constrained by age. In this perspective, foreign accent for ESL learners naturally results from the development of the first language phonetic system, which creates a phonological filter that applies equivalence classification by equating similar sounds in L2 with sounds in L1.

Regardless of the theoretical differences, both length of residence/exposure and total language input for ESL learners need to be taken into account in predicting the improvement in terms of foreign accent performance. To better understand the links between perception and production, proper behavioral assessment is also needed to tease apart local improvements in individual speech sounds from global improvements that take place in the prosodic domain in terms of stress, duration, and pitch modulation. One way of addressing this issue directly in second language acquisition is to investigate the effects of training in one domain (either perception or production) on the other domain.

2 Neuroplasticity: The ability for the brain to reinvent itself in language learning

Neuroplasticity refers to the brain’s ability to change throughout life. In addition to genetic factors, the social environment and personal learning experience play an important role. Plasticity applies to processes operative at many levels of our neurocognitive system, an intrinsic property that persists throughout our lives (DeFelipe, 2006, Mahncke et al., 2006). Adaptive changes and reorganizations in structure and function can reflect changes in the neurochemical systems, cell assemblies and connections,
brain activation patterns and behavioral performance, which have important implications for psychological development, L1 intervention, L2 education, and neurological rehabilitation. Some changes in the brain are known to be genetically determined and “experience-independent” whereas others are either “experience-expectant” or “experience-dependent”, which require the reception of certain input from the external environment (Greenough et al., 1999). There are two basic facts in studying neuroplasticity associated with speech acquisition and language learning. First, the brain is an intricate and highly specialized neural network with excitatory as well as inhibitory interconnections. Second, language is arguably the most complex human behavior. Insightful discoveries have been found in phonetic training that shed light on the nature of brain plasticity.

2.1 Phonetic training: Methods, findings and remaining issues

It is an intuitive belief that “practice makes perfect” – the performance of a given task will improve with repetition and training. While perceptual as well as motor learning has been the subject of psychological studies for over a century (Gibson, 1969, Karni, 1996), the study of brain plasticity in speech perception has been a fairly recent phenomenon. Research has shown that adult listeners do not show good discrimination sensitivity for many nonnative speech contrasts. In the ontogeny of development, adults are beyond the “critical period” early in life during which the brain is most sensitive to acquire a language (Grimshaw et al., 1998). Training studies bring up the issue of brain plasticity in adulthood by highlighting the role of experience in changing a listener’s speech perception. Successful training is evidenced by the transfer of learning to untrained stimuli and long-term retention of the improvement. Assuming that behavioral gains reflect experience-dependent changes in the brain, the successes and failures of various training methods may provide us with a better understanding of the underlying perceptual mechanisms and the nature of neural plasticity in the acquisition of new phonetic categories.

A long deliberated issue in cognitive neuroscience of language learning is the extent to which human brain has the capacity to change resulting from learning. Current theories posit that language acquisition patterns are influenced by linguistic experience rather than biological or maturational constraints (Best, 1995, Flege, 1995, Kuhl, 2000, Hernandez, Li and MacWhinney, 2005). Research has consistently shown that language-related cortical responses differ as a function of linguistic experience (Yetkin et al., 1996, Binder, 1999, Gaillard et al., 2000). Bilinguals’ second language (L2) may share the same neural system with the first language (L1) or involve specialized neurocircuitry, depending on such factors as age of L2 learning and proficiency levels (Kim et al., 1997, Perani et al., 1998, Xue et al., 2004, Hernandez and Li, 2007).

Training studies provide the opportunity to examine the interrelationship between perception and production. Many perceptual training studies have shown positive transfer of learning from perception to production (Bradlow et al., 1997, Callan et al., 2003b). Similarly, some studies using production-based training have also reported significant carry-over effects to perceptual skills (Mathews, 1997, Hazan and Sennema, 2007). Mutual facilitation between perception and production is also found in cross-modal analysis of training effects (Leather, 1990, Gómez Lacabex, 2009). But correlational results do not mean causal effects in terms of interactions between the two domains.

It has been widely accepted that the phonetic categories of the native language acquired early in life, represented in the neural substrate of the brain for speech perception, function like an “assimilator” for perceiving speech sounds in general. By this account, the degree to which listeners fail to detect a foreign-language contrast is closely tied to the mapping relationship between the native and nonnative phonological systems. It remains a question how the phonetic similarity and category goodness of native and nonnative phonemes can be predictive of the learnability of nonnative speech sounds in training (Best, 1994; Flege, 1995). What is not clear is the extent to which measures of auditory processing and phonetic interaction would show equivalence in listeners across cultures for assimilation and dissimilation processes. That is, at preattentive levels, such as that measured by the MMN, do training effects show up at the level of early auditory analysis or
do they simply heighten the listeners’ ability to attend to the critical acoustic cues for higher-level phonetic processing?

Neurophysiological data in training provide not only important information of the brain regions specialized in speech perception but also a different perspective regarding anatomical changes and the level of processing that is altered by language experience (Golestani and Zatorre, 2004b). Behavioral studies suggest that the language-specific pattern of perception results from selective attentional mechanisms and not a change in basic auditory sensitivity. As the neurophysiological data are recorded while the subjects are required to attend to a distraction task such as reading a book or watching a movie, it seems reasonable to argue that language experience alters not only higher-level categorization but also lower-level perceptual processes.

Training results have been mixed, and most training studies failed to meet the criterion of excellent generalization to novel contexts and talkers. While it is relatively easier to train listeners to perceive the nonnative contrasts cued by spectral differences (Pisoni et al., 1982, Jamieson and Morosan, 1986; Jamieson and Morosan, 1989, Merzenich et al., 1996), spectrally cued nonnative contrasts present much more difficulty (Pruitt, 1995). The /l-r/ distinction for Japanese listeners is one such example cued by the spectral difference in the third-formant transition. Various training studies of Japanese adults have showed large improvement, small but statistically significant improvement, or little change (Yamada and Tohkura, 1992, Lively, Logan and Pisoni, 1993, Lively et al., 1994, Bradlow et al., 1999, Takagi, 2002). In all the training studies, the Japanese trainees did not reach the level of perceptual performance of native English speakers.

The most successful training procedure in /r/-/l/ training studies for Japanese adults has used high-variability stimulus sets of natural speech tokens from multi-talkers (e.g., Logan, Lively and Pisoni, 1991). While smaller stimulus sets are easier to learn, the efficacy of the training on adult listeners does not readily extend to novel stimuli (Strange and Dittmann, 1984). Variability in speech tokens is thought to be helpful because it highlights the context in which critical acoustic parameters for nonnative contrasts are embedded and trains the listeners to extract the key acoustic cues and ignore irrelevant variation. Moreover, it has been found that acoustic modification and exaggeration in training tokens can be particularly useful in successful training (e.g., McCandliss et al., 2002; Merzenich et al., 1996; Iverson et al., 2005; Zhang et al., 2009). The exaggerated form of speech, as originally found in infant-directed speech (Kuhl et al., 1997), is helpful because it offers greater separation of the phonetic categories while highlighting the salient acoustic features that must be attended to. Japanese adults, for example, require greater separation between /r/ and /l/ tokens to escape the effects of the perceptual magnet effect around Japanese /r/.

According to the NLNC account, Japanese adult trainees have tremendous difficulty and limited success in /l-r/ training probably because exposure to the natural speech alone is unable to circumvent the strong preattentive filtering that hinders full access to or proper use of the critical acoustic dimension(s) for accurate phonetic categorization in the second language. It may require special enriched exposure analogous to "motherese" to circumvent the perceptual interference of NLNC in order for adults to learn non-native phonetic categories, especially those difficult contrasts that map onto single native categories. The signal enhancement approach was previously found to be effective in treating children with language disabilities in improving their phonetic discrimination and language skills (Tallal et al., 1998). This idea has been tested by developing a training software program that incorporated signal enhancement, visible articulation cues, a large stimulus set with high variability, and self-directed adaptive training (Zhang et al., 2000; Zhang et al., 2001). Approximately 12 hours of training showed over 20% identification improvement with excellent generalization to untrained voices. Training also resulted in a notable enhancement in neural sensitivity to the /r-l/ distinction in the left hemisphere, as measured by preattentive MMN, and increased neural efficiency in both hemispheres, as measured by a reduction in the amount of activation spread and duration of activation. The training-induced changes in behavioral discrimination were significantly correlated with changes in both neural sensitivity and efficiency.
measures. The results suggest that both neural sensitivity and neural efficiency appear to be good predictors of phonetic learning.

Although learning-induced enhancement in neural sensitivity has been consistently found in many other studies, the construct of neural efficiency as a neural signature of learning appears controversial. Intuitively, higher ability should translate into more efficient use of brain resources. Theoretically, learning can be conceived as an increased differentiation of the activation pattern, so that when performance is more specialized and highly efficient, only mechanisms absolutely necessary for the occurrence of performance are activated (Zhang et al., 2005). More focal activation in efficient neural processing does not necessarily equal smaller and shorter activation. An equally plausible model was based on the principle of Hebbian learning (McCandliss et al., 2002). This model predicts not only training-induced neural sensitivity but also increment in activation as a function of increased neural specificity (Patel et al., 1998, Johnsrude, Penhune and Zatorre, 2000) and connectivity (He et al., 2003, Horwitz and Braun, 2004) in the course of learning. The Hebbian model and the NLNC theory differs in their predictions about whether neural efficiency would lead to stronger or weaker activation in term of activation level and longer or shorter activation in terms of activation duration.

Speech perception involves brain regions for acoustic-phonetic as well as auditory-articulatory mappings (Callan et al., 2003a, Imada et al., 2006). Learning-induced plasticity in speech perception could also be associated with decreases, increases and shifts in brain activation to facilitate the behavioral improvement. At the cortical level, available data on phonetic training suggest that improved behavioral performance does not necessarily involve reduction of brain activities. Rather, reallocation in hemispheric resources (relative dominance of left and right hemispheres, for instance), recruitment of additional brain regions, strengthened anatomical (increased white-matter density) and functional connections (increased coherence among regions) in neural pathways, and increases or decreases in brain activation can all take place in the course of phonetic learning (Golestani, Paus and Zatorre, 2002, Callan et al., 2003a, Wang et al., 2003, Golestani and Zatorre, 2004a). For example, Golestani and Zattore (2004) found that the degree of successful phonetic training was correlated with more efficient processing (faster against slower learner) in temporal-parietal and inferior frontal activation, notably in the left hemisphere. Voxel-based morphometry analysis of MRI data also indicated that faster phonetic learners had more white matter in parietal regions for more efficient processing, especially in the left hemisphere (Golestani et al., 2002).

The phonetic training studies raise the possibility of progressive cortical changes with increased proficiency, suggesting that cortical representations may be continuously shaped with learning throughout life (van Turennout, Ellmore and Martin, 2000). The results support the view that language learning is not a strictly timed developmental process with rigid cut-off periods (Flege, 1995, Bongaerts et al., 1997, Hakuta, Bialystok and Wiley, 2003). While native speakers have acquired linguistic expertise by automatically focusing at a more abstract (linguistic) level of processing and thus freeing attentional resources, second language learners may rely more on bottom-up processing with varying demands on attention in the course of learning depending on such factors as age and proficiency. In fact, changes in brain activation may reflect not only interference from prior learning but also continuously updated cognitive and attentional strategies. Therefore, the learning trajectory does not necessarily entail a monotonic increase in neural efficiency (Wang et al., 2003, Werker, Hall and Fais, 2004, Werker and Tees, 2005, Zhang et al., 2005). One plausible brain activation trajectory in learning would be inverted U-shaped activation as a function of time. Initially, greater attention is required to attend to the task, leading to increased activation in specific brain regions; as the learning process moves toward expertise continues, processing may become more and more automatic so that the activation level gets reduced at certain point. In this perspective, there would be no contradiction between the observations of increased and decreased activations to characterize neural plasticity. Increased activation takes place at an earlier stage, accompanied by improved neural sensitivity and enhanced neural specificity at the cost of increased bottom-up processing. Decreased activation takes over at a later stage of expertise attainment, requiring the minimal level of activation for the ease of processing under the preemptive dominance of top-down
processing. Certainly, this learning trajectory hypothesis needs to be tested against a number of variables, including not only attentional demands but also age, handedness, relative language proficiency levels for the first and second languages, and motivational factors.

2.2 Imaging and neural markers of plasticity

In order to integrate brain research with theories built upon behavioral data and evaluate training-induced changes in the brain, one needs to understand the advantages and challenges of the various techniques. The value of cognitive neuroscience research lies in its ability to determine and validate non-invasive brain-based spatial and temporal markers for understanding human intelligence and complex behavior (Gazzaniga, 2000, Nelson and Luciana, 2001). Among the most commonly used non-invasive techniques are EEG (electroencephalography), MEG (magnetoencephalography), and fMRI (functional Magnetic Resonance Imaging). The EEG and MEG systems measure the variations in scalp voltages or magnetic fields generated by the postsynaptic neuronal activities in the order of milliseconds. Because the critical acoustic features for different speech sounds are only tens of milliseconds long in time or less, a detailed measurement of speech and voice processing requires excellent time resolution. In event-related paradigms, responses time-locked to the presentation of the stimuli are recorded. The event-related potentials (ERPs) or event-related field (ERF) can be taken from subjects when no behavioral response is required. To exploit the neuronal dynamics within specific brain regions, it is important to assess the localizing ability of EEG and MEG inverse solutions. EEG localization is limited by the volume conduction of currents through the tissues of the head and its primary sensitivity to radially oriented neural currents in the gyri, and MEG sensors are limited by their sensitivity to tangential neural currents primarily in the sulci (Hämäläinen et al., 1993, Baillet, Mosher and Leahy, 2001). Modeling and empirical data suggest that a combination of MEG and EEG can take advantage of their complementary sensitivities for more accurate results (Lopes da Silva, Wieringa and Peters, 1991, Liu, Dale and Belliveau, 2002, Pfieger, Greenblatt and Kirkish, 2004, Sharon et al., 2007).

One limitation with EEG/MEG source localization is the poor signal to noise ratio for deep sources due to the distance between the sensors and the deep brain structures that support important functions of attention, memory and emotion. This is not a problem with the fMRI technique, which is known for its high spatial resolution of hemodynamic changes in specific brain regions associated with experimental tasks regardless the depth of the source. But hemodynamic changes take place on a much slower temporal course in the order of seconds. A critical and practical challenge to the integration of EEG/MEG with fMRI lies in the spatial and temporal mismatches between relatively slow fMRI activations and instantaneous electrical source activities. Advances in MRI- and fMRI-constrained EEG/MEG localization with depth weighting have shown a promising venue for exploring the complementary capability of the methods (Dale and Halgren, 2001, Im, Jung and Fujimaki, 2005, Lin et al., 2006a, Lin et al., 2006b, Liu and He, 2008, Wibral et al., 2008).

The EEG, MEG and fMRI techniques can be employed to characterize neuroplasticity associated with language learning using the spatial and temporal markers in four categories of measurements: (a) neural sensitivity, (b) neural specificity, (c) neural connectivity, and (d) neural efficiency (Zhang and Wang, 2007). Sensitivity can be shown by the latency and amplitude of a change-detection measure called mismatch negativity (MMN). The MMN peaks typically at 100-300 ms after the onset of a change in the stimulus and is considered an “endogenous” component generated by the automatic detection of a stimulus change in auditory sensory memory (Näätänen et al., 2007). The MMN amplitude and latency measures demonstrate good correlations with psychophysical discrimination for various parameters of speech and nonspeech stimuli irrespective of the listener’s attention. Specificity is assessed by the degree to which specific parts of brain region(s) or neural pathway are dedicated to the processed information. Connectivity is structurally measured in grey-matter and white-matter density in regions of interest and functionally in the degree of coherence, synchronization between activated regions or measuring sensors and trial-to-trial analysis of phase locking (Gootjes et al., 2006). Finally, neural efficiency can be considered a derivative of the experience-driven changes in all the three measures, which presumably
leads to more focal, faster and shorter activation associated with learned categories as against new or unfamiliar ones (Zhang et al., 2005). Collectively, the operational definitions for these measures need to be tested for a thorough examination of neural commitment and neuroplasticity in spatiotemporal activation patterns. The fMRI data are particularly informative about subcortical involvement, which can be integrated into fMRI-constrained source localization analysis of the ERP and MEG data.

One distinctive advantage of brain imaging is that brain activation provides direct information of relative hemispheric involvement and cortical localization. Although there have been many adult studies reporting left-hemisphere dominance for native speech perception, other data tend to show large individual variability and a bilateral activation pattern even for native speech processing (Zhang et al., 2005). Developmental data do not show a clear maturational pattern of left hemispheric involvement for native speech processing as against nonnative or general acoustic processing. Some imaging data suggest that the left dominant areas subserving speech processing are similar in infants and in adults (Dehaene-Lambertz, Dehaene and Hertz-Pannier, 2002, Pena et al., 2003). However, the MMNs in infant often show a bilateral temporal maximum with no laterality effect (Čeponienė, Rinne and Näätänen, 2002). The fact that different speech sounds are cued by different spectral and temporal parameters further complicates the interpretation. Some PET and fMRI data show that the left auditory cortex is specialized for processing rapidly changing broad-band stimuli including speech, whereas the right auditory cortex may be good at processing slower narrow-band stimuli such as tonal patterns (Zatorre and Belin, 2001). However, this theory does not provide specific predictions regarding the cortical basis of native vs. nonnative language processing.

Caution must be taken when interpreting results across different speech stimuli in different studies. Various speech sounds are characterized by markedly different temporal and spectral cues. The large differences in temporal and spectral features correspond to differences in voicing, place of articulation, and manner of articulation for consonants and tongue height, tongue backness, and lip rounding for vowels. Due to the time-based nature of EEG and MEG responses, a neural coding pattern for a particular phonetic feature may not be generalizable to other features. It is also important to note that evoked neural responses vary as a function of manipulations of experimental conditions and acoustic properties of the stimuli. For example, the MMN response is sensitive to minor acoustic differences in both intra- and inter-category domains. When the acoustic difference is large between the standard and deviant, the phonetic factor may be subdued by the acoustic factor.

A full account of brain mechanisms of neuroplasticity in language acquisition needs to specify how learning shapes automatic and attentional processes for efficient and accurate perception and production. The neural markers of speech learning need to specify how automatic and attentional processing differentially contributes to the quantifiable measures. Automaticity and attentional control mechanisms are core constructs underpinning theoretical accounts of human intelligence and behavior across research domains (Deacon et al., 1999, Pardo et al., 2000, Fan et al., 2005, Fang and He, 2005, Pulvermüller and Shtyrov, 2006, Jiang, Zhou and He, 2007, MacDonald, 2008). Automaticity is characterized by fast and accurate informational processing involving little or no conscious awareness, which is inherently bound with experience and learning. Automatic processing and attentional mechanisms can be systematically examined by the use of different paradigms that directly manipulate attentional resources. Specifically, the MMN response for automatic processing of speech can be modified to require different amounts of attention to the distracting modality. Selective listening and priming paradigms can be employed to examine the dissociable neural systems for acoustic, phonetic, phonological, and lexical factors that influence the acquisition of L2 phonology and their interactions. Future studies along these lines will contribute to a better integration of the findings in the literature for more accurate characterization of the temporal markers and neural systems involved in L1 and L2 acquisition.
3 From theory to practice: Software showcase towards optimization of phonetic training

3.1 The /r/-l/ training software

To address the underlying mechanisms of brain plasticity in speech learning, Zhang and colleagues designed a training software program to test its success on adult Japanese ESL learners (Zhang et al., 2000, 2009). The program incorporated features that were motivated by studies of infant-directed speech (IDS) or “motherese” (Fernald and Kuhl, 1987, Kuhl et al., 1997, Burnham et al., 2002, Liu et al., 2003), including adaptive signal enhancement, visible articulation cues, a large stimulus set with high variability, and self-initiated selection. The results indicate that substantial behavioral improvement in second language phonetic learning can be achieved in adulthood and simultaneously reflected by the spatiotemporal markers of neural sensitivity and neural efficiency, resulting in native-like perception and native-like brain activation patterns for learning the difficult speech contrasts in a second language. Furthermore, IDS-motivated training program can help circumvent interference from neural networks that have been shaped by native language experience, yielding significant brain-behavior correlations in both domains of sensitivity and efficiency.

The training software program (Figure 3) incorporated the following key features: 1. **Self-directed listening.** Trainees selected the sounds by clicking on iconic buttons that indicated talkers and vowel/syllabic contexts. 2. **Visible articulation cues.** Photographic facial images and visual animation effects of each talker articulating /r/ or /l/ were provided for each sound presentation. 3. **Large stimulus sets with high variability.** A total of 120 different tokens were used for training. A message would prompt the participant to select a different icon if one icon had been clicked on for 20 times. 4. **Adaptive scaffolding.** There were 12 training sessions, starting from the most exaggeration sounds with one speaker only. Each session consisted of 10 listening blocks with 50 tokens in each block. Each listening block was followed by an identification quiz of 10 sounds to monitor progress. Difficulty level was increased when the quiz score was 90% or above. The scaffolding system worked by first adding talkers (up to five talkers) and then reducing exaggeration level (down to Level Zero). 5. **Feedback outside of listening blocks for training.** Each correct answer in the quiz accumulated monetary reward of two cents in US currency for the participant. Incorrect answers were prompted with a one-time playback of the sound. No feedback was given during the listening blocks in training or pre-post tests.

The training stimuli were digitally synthesized based on recordings from eight native speakers of American English (4 males, 4 females) producing five vowels (/a/, /i/, /u/, /e/, /o/) in the Consonant-Vowel (CV) and Vowel-Consonant-Vowel (VCV) contexts. The training sessions used audiovisual /r/-l/ stimuli from five talkers (3 males, 2 females) and three of the five vowels (/a/, /e/, /u/) in the two syllabic contexts (CV and VCV). Adaptive training was implemented by using acoustic modification on the training stimuli with four levels of exaggeration on three parameters of the critical F3 transition for the /r/-l/ distinction, namely, F3 separation in frequency, F3 bandwidth, and F3 transition duration (Zhang et al., 2000). In the synthesis process, the recorded sounds were submitted to an LPC (linear predicative coding) analysis-resynthesis procedure to exaggerate the formant frequency differences between pairs of /r/-l/ tokens, and to reduce the bandwidth of F3. The LPC technique analyzed the speech signal by estimating the formants, removing their effects from the speech signal by inverse filtering, and estimating the intensity and frequency of the residual signal. Temporal exaggeration of the /r/-l/ stimuli was made using a time warping technique – pitch synchronous overlap and add (Moulines and Charpentier, 1990).

3.2 SAT (Speech Assessment and Training Program): New software development for ESL learners

The original training program (Zhang et al., 2000) was implemented using Macromedia Authorware on an Apple PowerPC. Zhang et al. (2009) made minor revisions to improve user interface on computer screen for the training sessions. But there are three major limitations. First, the software has very limited use as it was designed specifically for /r/-l/ training on adult Japanese speakers. Second, the audiovisual feature in the training program was based on simple animation of two static shots for the articulation of the stimuli at the beginning and at the end. This limitation reflected technical difficulty to
create the video clips that included the different amounts of temporal exaggeration for the training stimuli. Third, data collection generated a separate output file for pretest, posttest, and each training session for each subject without implementing a systematic database structure for all the data points. The multiple output files made it difficult for data analysis and online monitoring of progress across training sessions for individual trainees. Fourth, the software did not collect reaction time data for individual button responses.

In order to increase the usability of the software for ESL learners, Zhang took a further step in developing a new software program (Speech Assessment and Training, SAT) to overcome the limitations (Figure 4). The essential features for synthesizing the training stimuli are retained in new SAT software. Instead of robotic animations based on two static shots, realistic video clips for the synthetic stimuli are generated using Final Cut Pro 7 (Apple Inc.) for the different versions of slow motion movies synchronized with temporally stretched audio track for all the different speech tokens used in the program. A fully functional database structure is implemented using Microsoft ACCESS program to manage data access, retrieval, report, and analysis for a large number of participants with data from the pre-post tests and all the training sessions. To make it appealing and relevant to ESL learners, the training stimuli used real words in English, and the on-screen selective listening used IPA symbols (International Phonetic Alphabet). In addition to /r-l/ contrast, three more difficult speech contrasts have been added, including a nasal contrast and two vowel contrasts. Each contrast consists of 30 minimal pairs of words. The program also features an interface that includes login, password, and trainee IDs that allows secure remote training online for large scale training and study.

The prototype of the program has been implemented using the cross-platform Java language and tested on 38 adult ESL learners in Xi’an Jiaotong University, China (Cheng and Zhang, 2009). The initial goal was to conduct general assessment of Chinese ESL learners’ speech perception and production from two aspects: (a) Identification of four patterns of relationship between speech perception and production (good perception – good production; poor perception – poor production; good perception – poor production; poor perception – good production.). (b) Investigation of the relationships between phonetic learning and higher levels language proficiency in ESL learners. Intensive phonetic training was conducted on four pairs of most difficult contrasting segments: /l/ - /r/, /s/ - /z/, /n/ - /ŋ/, /i/ - /ɪ/. There were seven training sessions for each speech contrast, each lasting up to four hours over a two-week span. The initial results showed highly positive training effects in short-term improvement of perception and production in all four difficult contrasts. Although immediate posttest results did not show significant advantage of multi-talker training over single-talker training, the retention results after 6 months of training showed benefits of multi-talker training. Moreover, retention effects were significant for improved perception but not for production. Large individual differences were observed among the trainees, and significant differences in training effects were also found for the different speech contrasts.

The SAT software program allows more research options and practical applications for ESL learners. The vision in software design is to make it customizable and available not only as a research tool but also for online speech training. Control options are built in to allow flexibility in the use of the major features. For instance, training can be done in contrastive manners to tease apart how various features contribute to the learning processes: multi-talkers versus single talker, auditory versus audiovisual, formant exaggeration versus no exaggeration, temporal exaggeration versus no exaggeration, and feedback versus no feedback for short quizzes after each training session. The program can also be easily modified to include new speech contrasts for research or teaching purposes. More rigorous research and educational plans are currently being developed with research collaborators and English teachers to improve the software design. The new research studies aim to test neuroplasticity by including EEG and fMRI measures in the pre- and post- tests. The ultimate goal is to find optimal training procedures for different ESL learners by providing objective and comprehensive assessment of their skills and customized training programs by integrating the various features that have been found to promote brain plasticity for ESL learning in adulthood.
4 Implications and future directions for research and ESL education

With all the empirical evidence showing the effects of early language experience and neuroplasticity for L2 learning, some basic questions remain to be further explored. As discussed in the previous three sections, a number of important issues remain to be resolved for the NLNC theory. Language learning is not equivalent over the lifespan; there are at least four perspectives in studying neuroplasticity for language: the initial state of the infant, developmental changes, the mature state of the adult listener, and factors that facilitate speech perception and production training in infancy, childhood and adulthood.

Research has shown strong associations between the characteristics of language input and the development of speech perception in parallel with the strong associations between phonetic perception and other language skills. The implicit learning mechanisms that operate on the probabilistic transitions and statistical distributions of the language input are fundamental to language acquisition early in life (Kuhl et al., 2001, Saffran, 2003), second language acquisition (Zhang et al., 2000, Mueller et al., 2005) and artificial grammar learning (Lieberman et al., 2004, McNealy et al., 2006). Details of phonetic interactions between L1 and L2 and links between perception and production and transfer of learning across the two domains await further research for various speech sound categories as a function of age of acquisition, input content, and length of exposure. Future imaging studies will continue to refine the operational definitions of neural sensitivity, neural efficiency, neural specificity and neural connectivity and how these measures are affected by automatic and attentional processes, multimodal speech processing, and experimental manipulations on subject, stimulus, and task characteristics.

Phonetic training studies have shown promising approaches to promote neuroplasticity and overcome the NLNC constraints for L2 acquisition. Manipulations to the language input and delivery mechanisms, as shown in the SAT training software illustrated in this chapter, can effectively enhance L2 processing at the pre-attentive level and thus influence the efficiency of the developing brain in social, cognitive and linguistic domains. Such procedures have been previously shown to be effective and helpful in acquiring novel phonetic categories from brief exposures and improving literacy skills in children with reading problems. Further research and development using computer-assisted training technology holds the promise of delivering optimal tools for ESL learners.

Acknowledgements

Manuscript preparation and software development were supported by a Single Semester Leave and Brain Imaging Research Project Award from the College of Liberal Arts, University of Minnesota. Software development received additional support from Xi’an Jiaotong University’s Presidential Research Fund to Bing Cheng and Yang Zhang. The authors acknowledge assistance from David Akbari, Xue Bai, Xiao Dong, Tess Koerner, Sharon Miller, Sierra Novak, James Stevens and Liz Tusler for the development with the SAT software. Thanks are also due to Professors Bing Cheng (Xi’an Jiaotong University), Yue Wang (Simon Fraser University), Hua Shu (Beijing Normal University), Ping Li (Pennsylvania State University), and Patricia Kuhl (University of Washington) for technical and theoretical discussions.
References


Figure legends

Figure 1. Effects of linguistic experience shown in the Perceptual Magnet Effect (Iverson et al., 2002). Goodness, identification, and MDS (multidimensional scaling) solutions for Japanese, German, and American listeners. In the goodness and identification graphs, the size of the circle indicates the average goodness rating (larger circle for higher goodness), and the shading indicates the predominant phonetic category (black for the respective /r/ sounds in Japanese, German, and English; white for /l/ sounds in German and English, and /w/ in Japanese). The numbers within the circles list the average goodness ratings and the identification percentages for the predominant phonetic category. The MDS solutions are geometric representations of the average similarity ratings for these stimuli. The lines between stimuli reflect their spacing and the length of the lines reflect perceptual sensitivities for these acoustic differences (perceptually similar stimuli are placed close together; perceptually dissimilar stimuli are placed far apart).

Figure 2. Sample MEG waveforms for the mismatch negativity (MMN) response and the basic event-related positive (P) and negative (N) current peaks. The waveforms show the deviant and standard stimuli from a single subject at one recording channel with a 100 ms baseline. The evoked response components of P1m, N1m, P2m, and MMNm are marked on the waveforms; the subscript ‘m’ in the naming stands for magnetic response.

Figure 3. Screenshot of training program during one session for the /l-r/ training program. The mouse-on button shows a female talker articulating /ru/. The upper left corner registers the difficulty level, and the upper right corner the cumulative reward earnings for correct answers in the small quizzes after each listening block. The lower left corner shows the number of sounds played, and the lower right corner the current listening block in a training session.

Figure 4. Sample screenshots of training sessions in the Speech Assessment and Training (SAT) software program developed by Yang Zhang at the University of Minnesota. (a) Single talker training for the /i- i/ contrast. The trainee can click on the phonetic symbols to watch the video clip of articulation synchronized with different versions of exaggerated speech in the audio track. The four corners on the screen show training session number, the number of words heard, the talker icon, and the exaggerated level. (b) Multiple-talker training. The trainee can activate a talker by clicking on the targeted talker icon at the lower left corner of the screen.
Figure 1

Goodness and Identification

Japanese listeners

Multidimensional Scaling

German listeners

American listeners (Iverson & Kuhl, 1996)
Figure 2

Scale bars: Vertical = 100 fT/cm, Horizontal = 200 ms