CHAPTER 1

The Revised Speech Learning Model (SLM-r)

James Emil Flege and Ocke-Schwen Bohn[*]

Like its predecessor, the revised Speech Learning Model (SLM-r) focuses on the learning of L2 vowels and consonants (or "sounds," for short) across the life-span. To define the context in which the original SLM (Flege, 1995) developed, we begin by presenting some key studies carried out before 1995. After summarizing the SLM with clarification of some key points, we present the SLM-r.

The primary aim of the SLM-r differs from that of its predecessor, which was to "account for age-related limits on the ability to produce L2 vowels and consonants in a native-like fashion" (Flege, 1995, p. 237). The SLM focused on differences between groups of individuals who began learning an L2 before versus after the close of a supposed Critical Period (CP) for speech learning (Lenneberg, 1967). Closure of the CP was regarded as an undesired consequence of normal neurocognitive maturation that arose from diminished cerebral plasticity and a reduced ability to exploit L2 speech input. The SLM-r offers an account for differences between "early" and "late" learners, but its primary aim is to provide a better understanding of how the phonetic systems of individuals reorganize over the life-span in response to the phonetic input received during naturalistic L2 learning.

1.1 Work Prior to 1995

A phonemic level of analysis dominated early L2 research. Bloomfield (1933, p. 79) posited that because monolinguals learn to respond only to

* The work presented here was supported by grants from the National Institute of Deafness and Other Communicative Disorders. Susan Guion played an important role in this research, and we truly miss her. Special thanks are also due to Katsura Aoyama, Wieke Eefting, Anders Højen, Satomi Imai, Ian MacKay, Murray Munro, Thorsten Piske, Carlo Schirru, Anna Marie Schmidt, Naoyuki Takagi, Amanda Walley and Ratree Wayland. We thank Charles Chang, Olga Dmitrieva, Francois Grosjean, Nikola Eger, Natalia Katushina and Juan Carlo Mora Bonilla for comments of earlier versions of this chapter.

distinctive features, they can "ignore the rest of the gross acoustic mass that reaches [their] ears." Hockett (1958, p. 24) defined the phonological system of a language as "not so much a set of sounds as … a network of differences between sounds." Trubetskoy (1939) proposed that the phonological system of the native language (L1) acts like a "sieve" that passes only phonetic information in the production of L2 words that is needed to distinguish words found in the L1. This approach shifted attention away from the language-specific phonetic details of the L1 to which children attune slowly during infancy and childhood and it implied that such details might be inaccessible to individuals who learn the same language as an L2. One dissenting voice was that of Brière (1966), who maintained that the relative ease or difficulty of learning specific L2 sounds could only be predicted through "exhaustive" analyses of phonetic details (p. 795).

The aim of the contrastive analysis (CA) approach was to identify learning problems that would need to be addressed through instruction in the foreign-language classroom. Its general prediction was that L2 phonemes that do not have a counterpart in the L1 would be difficult to learn whereas those having an L1 counterpart would be relatively easy to learn. The CA approach assumed that pronunciation errors observed in L2 speech were the result of faulty articulation (i.e., production), not the results of incorrect targets resulting from faulty perception. Just as importantly, the CA approach ignored the fact that the "same" sound found in two languages may differ greatly at the phonetic level.

Another problem for the CA approach was that allophonic distributions of the "same" phonemes found in two languages often differ cross-linguistically, making point-by-point comparisons of phonemes difficult or meaningless (Kohler, 1981). The phonemes in a contrastive analysis were defined primarily in terms of a static articulatory description of a single canonical variant. This ignored the fact that an important part of L1 acquisition is the integration of conditioned variants of a phoneme (Gupta & Dell, 1999; Song, Shattuck-Hufnagel, & Demuth, 2015). In addition, the CA approach tacitly assumed that L2 learners make errors even after having received adequate input and that knowing how the L1 is learned is irrelevant for an understanding of L2 speech learning.

The one-time, one-size-fits-all CA approach soon fell from favor. As Lado had already noted in 1957, not all individual speakers of a single L1 make the same errors when speaking the same L2. Flege and Port (1981) showed that the distinctive features needed to distinguish L1 phonemes cannot be recombined freely to produce an L2 sound that is not present

in the L1. Most importantly, as noted in 1953 by Weinreich, the nature and extent of the mutual "phonic interference" between the sounds in a bilingual's two languages depends, in addition to phonological differences, on factors such as language dominance, demography (e.g., ethnicity, gender, age), years of L2 use, and the domains in which the L1 and L2 are used (pp. 83–110; see also Grosjean, 1998).

In the 1970s research began examining purely phonetic aspects of L2 segmental production and perception. Much of this early work focused on the voice-onset time (VOT) dimension in the production and perception of word-initial English stops by native Spanish speakers. For example, Elman, Diehl, and Buchwald (1977) examined the identification of naturally produced consonant-vowel syllables initiated by stops having VOT values in the "lead," "short-lag," or "long-lag" ranges. Spanish and English monolinguals labeled stops having short-lag VOT as /p/ and /b/, respectively. The Spanish-English bilinguals who participated were asked to label the same stimuli in two "language sets" intended to induce a Spanish or English perceptual processing mode. The effect of the language set manipulation was small for most participants, but five of the 31 bilingual participants, referred to as "strong" bilinguals, were far more likely to identify short-lag stops as /b/ in the English set than in the Spanish set. These five participants seem to have been early learners (R. Diehl, personal communication, June 3, 1990).

Many accepted the hypothesis by Lenneberg (1967) that a critical period (CP) exists for speech learning and that it closes at about the age of 13 years as the results of normal neurological maturation. Lenneberg (1967, p. 176) also suggested that following the close of a CP, L2 learners cannot make "automatic use" of L2 input from "mere exposure" to the input as children do when learning their L1.

To evaluate the "automatic use" hypothesis, Flege and Hammond (1982) recruited native English (NE) university students in Florida. All of them had been previously exposed to Spanish-accented English and, in addition, were taking a Spanish class taught by a native Spanish speaker who spoke English with a strong Spanish accent. The students were asked to read English sentences containing two variable test words (*The X is on the Y*) with a feigned Spanish accent. The amount of prior exposure to Spanish-accented English the students had received was estimated by counting the number of expected "Spanish accent substitutions" (e.g., [vel] or [veˈl] for *bail*, [big] for *big*) they produced in the English test words. VOT was measured in additional test words beginning in /t/.

Members of both the higher- and lower-exposure groups shortened VOT in the direction of values typical of Spanish, but only the higher-exposure group produced significantly shorter VOT values than did the members of a control group who read the sentences without special instruction. Importantly, the students who produced significantly shortened VOT values when speaking English with a feigned Spanish accent did not accomplish this by using a short-lag English /d/ to produce the /t/-initial test words.

Flege and Hammond (1982) concluded that monolingual adults are able to access cross-language phonetic differences through mere exposure to speech after the supposed closure of a CP for speech learning. The results indicated that NE monolinguals with substantial exposure to Spanish-accented English could not only detect phonetic differences between standard and Spanish-accented English, they could also store that information in long-term memory and use it to guide production (see also Reiterer, Hu, Sumathi, & Singh, 2013).

Other research examined production and perception of the VOT dimension in L2 learners. Williams (1977) found that the "phoneme boundary" between stops such as /b/ and /p/ occurred at significantly longer VOT values for adult English than Spanish monolinguals. Flege and Eefting (1986) reported that this also held true for monolingual children. They also reported that, within languages, phoneme boundaries occurred at longer average VOT values for adults than for eight- to nine-year-old children. Indeed, the phoneme boundaries of NE 17-year-olds occurred at significantly shorter VOT values than those of NE adults, suggesting that attunement to L1 phonetic-level details may continue into the late teenage years. Not surprisingly, Flege and Eefting (1986) observed cross-language production differences that mirrored the above-mentioned perception differences in phoneme boundaries. Both monolingual NE adults and children produced /p t k/ with longer VOT values than age-matched native Spanish (NS) monolinguals and, within both languages, adults produced longer VOT values than children did.

Research began to focus on providing an explanation for differences observed for early and late learners. Flege (1991) compared VOT in stops produced by groups of NS adults differing in age of arrival in the United States (means = 2 vs. 20 years). These early and late learners also differed in percentage English use (means = 82 vs. 66 percent). The early learners produced English stops with native-like VOT values, both individually and as a group. The average values obtained for late learners, on the other

hand, were intermediate to the values observed for Spanish and English monolinguals. This finding suggests that the speech learning ability of the late learners may have been partially compromised, perhaps due to the closing of a critical period.

The results of a speech imitation study (Flege & Eefting, 1988) suggested that NS early learners of English can form new long-lag phonetic categories for English /p t k/. This finding led Flege (1991) to suggest that the accurate production of VOT in English stops by NS early but not late learners arose from the inability by the late learners to form new phonetic categories. Had this been true it would have provided a solid empirical basis for the CP proposed by Lenneberg (1967). However, the interpretation suggested by Flege (1991) was problematic for two reasons. First, the VOT values produced by individual NS Late learners ranged from Spanish-like to English-like. If a CP exists, it should affect everyone in much the same way. Second, the results for the late learners may have reflected learning in progress rather than the performance that might have been evident had they received as much L2 phonetic input as monolingual NE children need to achieve an adult-like production of VOT.

For this chapter, we estimated years of full-time equivalent (FTE) English input that had been received by the NS participants in Flege (1991). These values were calculated by multiplying years of residence in the United States by proportion of English use (self-reported by each participant as a percentage). The mean estimated FTE years of English input was much higher for the early than late learners (means = 17.2 vs. 9.2 FTE years). Thus, if category formation is a slow process requiring input that gradually accumulates over many years of daily use, the early–late difference observed by Flege (1991) might simply have been the result of input differences, not the loss of capacity by the late learners to form new phonetic categories.

FTE years of L2 input may be a somewhat better estimate of quantity of L2 input than LOR alone, but it says nothing regarding the quality of L2 input. Early learners acculturate more rapidly following immigration than late learners do (Cheung, Chudek, & Heine, 2011; Jia & Aaronson, 2003). Acculturation involves the establishment of social contacts with native speakers of the target L2. This means that the NS late learners tested by Flege (1991) were likely to have been exposed more often to Spanish-accented English than the early learners were, and so they may have been exposed to shorter VOT values in English words overall than the early learners and NE monolinguals.

The effect of foreign-accented input was observed in research examining NS early learners who learned English in an environment where Spanish-accented English was the rule rather than the exception. The mean VOT values obtained by Flege and Eefting (1987) for early learners in Puerto Rico were intermediate in value, in both production and perception, to values obtained for English and Spanish monolinguals, and so were similar to the values obtained for NS late learners of English in Texas. The difference between the early learners tested in Puerto Rico and Texas suggested that the quality of L2 input may matter more than the age of first exposure to an L2.

Research in the period we are considering also showed that the magnitude of cross-language phonetic differences matters. Flege (1987) examined the production of French vowels by NE speakers who had lived in France for an average of 10 years. Unlike French, English has no /y/ and its /u/ differs acoustically from the /u/ of French. The three vowels of interest (French /y/ and /u/, English /u/) differ primarily in (F2) frequency, and NE speakers generally hear the French /y/ as their English /u/ (Levy, 2009a). Flege (1987) hypothesized that NE speakers would be able to produce the "new" French vowel, /y/, more accurately than the "similar" French /u/. In fact, the difference between the NE speakers and French monolinguals in terms of the crucial acoustic phonetic dimension, F2 frequency, was nonsignificant for /y/ but not /u/.

Flege (1992) further evaluated the new-similar distinction by examining the production of English vowels by native Dutch (ND) adults. The English vowel in *hit* (/ɪ/) was classified as "identical" to a Dutch vowel based on previously published acoustic data and on reports that the auditory differences between English /ɪ/ and the closest Dutch vowel are likely to go undetected by native Dutch-speaking listeners. The English vowel in *hat* (/æ/) was classified as "new" because it occupies a portion of vowel space not exploited by Dutch and because earlier research suggested that /æ/ is learnable. The vowels in *heat*, *hoot*, *hot* and *hut* (/i/, /u/, /ɑ/, /ʌ/) were each categorized as "similar" to a Dutch vowel.

The results obtained by Flege (1992) for the "new" vowel in *hat* supported the view that ND late learners can form new phonetic categories for certain L2 vowels. However, the results obtained for English vowels classified as "similar" to a Dutch vowel did not support the hypothesis that native versus nonnative differences persist for L2 vowels that are similar but not identical to an L1 vowel. Two Dutch vowels classified as similar were produced quite well but two others were produced poorly. Flege (1992, p. 162) concluded that no principled method existed

for distinguishing "new" from "similar" L2 sounds and so the trichotomy "new-similar-identical" was not included in the SLM (Flege, 1995).

Flege, Munro, and Skelton (1992) evaluated the effect of L2 experience by recruiting two groups each of native Mandarin (NM) and Spanish (NS) speakers. All had begun learning English as adults, but the same-language groups differed according to LOR in the United States (Mandarin means = 0.9 vs. 5.5 years; Spanish means = 0.4 vs. 9.0 years). The study focused on the production of word-final English /t/ and /d/ because these stops are not found in the final position of Mandarin or Spanish words. The authors hypothesized that the nonnatives with a relatively long residence in the United States would treat the word-final stops as "new" sounds and so produce them accurately.

NE-speaking listeners were more successful overall in identifying the nonnative speakers' productions of /t/ than /d/ (means = 82 vs. 65 percent correct). Acoustic analyses revealed that the NM and NS speakers produced smaller acoustic phonetic differences between /t/ and /d/ (longer vowels before /d/, higher F1 offset frequency for /d/, more closure voicing in /d/, longer closure for /t/) than the NE speakers did. Stops produced by both "experienced" and "inexperienced" nonnatives were significantly less intelligible (means = 68 percent for both groups) than stops produced by the NE speakers. Within languages, the LOR-defined groups did not differ significantly. Of the 40 NM and NS speakers tested, just six produced word-final stops that were as intelligible as the stops produced by the NE speakers.

One possible explanation for the frequent errors in nonnative speakers' final stop productions identified by Flege et al. (1992) is that adult learners of an L2 lack the capacity to learn new forms of speech. An alternative explanation is that the errors may have been the result of inadequate input. Monolingual NE children need approximately five years of full-time English input in order to produce /t/ and /d/ accurately in word-final position (e.g., Smith, 1979). The nonnative speakers designated as "experienced" had an average of just 4.2 FTE (full-time equivalent) years of English input and were likely to have often heard other nonnatives produce the word-final English stops inaccurately.

The same two explanations might be applied to the findings of Flege and Davidian (1984), who used a picture-naming task to elicit the production of /p t k/ and /b d g / in the final position of English words. Among the participants tested were immigrants from China and Mexico (12 each) who had all learned English as adults and had lived in Chicago for 4.2 years on average (range = 0.2 to 7.5 years). Unlike members of the

NE comparison group (n = 12), these late learners omitted (means = 2.3 vs. 3.4 percent), devoiced (means = 29.5 vs. 43.0 percent) and spirantized (means = 0.8 vs. 19.3 percent) the word-final English stops. The differing frequency of error types observed for the two L1 groups was readily understandable with reference to the inventory of word-final obstruents found in their L1s, but overall, they produced only about half of the stops without error. All 24 participants were enrolled in English as a second language classes at a local community college where they certainly heard one another, and other immigrants outside the classroom, producing final English stops with the same errors. At least some of them may have learned to accurately produce the wrong phonetic "models."

In summary, L2 speech research carried out prior to 1995 gradually began to focus on a phonetic rather than a phonemic level of analysis. Language-specific phonetic differences between the L1 and L2 became the focus of speech production and perception research. The existing research made clear that (1) the L1 phonetic system "interferes" with L2 speech learning; (2) some L2 sounds are learned better than others; (3) L2 sounds without an L1 counterpart might be learned more effectively than those with an L1 counterpart; and (4) the quantity and quality of L2 input that L2 learners receive may exert an important influence on phonetic-level learning. It appeared possible that early learners generally produce and perceive L2 sounds more effectively than late learners do because they, but not late learners, might be able to form new phonetic categories for L2 sounds. This inference was at odds, however, with evidence that late learners can gain access to L1–L2 phonetic differences, store the detected differences in long-term memory, and then use the stored perceptual representations to guide articulation.

1.2 The Speech Learning Model (SLM)

Flege (1995) observed that at a time when "children's sensorimotor abilities are generally improving, they seem to lose their ability to learn the vowels and consonants of an L2" (p. 234). We now know that earlier is generally better than later for those learning an L2, but only in the long run. Adults outperform children in the early stages of naturalistic L2 acquisition, but adult-child differences tend to recede over time until early learners outperform late learners (e.g., Jia, Strange, Wu, Collado, & Guan, 2006; Snow & Hoefnagel-Höhle, 1979).

DeKeyser and Larson-Hall (2005) attributed the age-performance "cross-over" to age-related cognitive changes. If applied to L2 speech

learning, their hypothesis would mean that children learn L1 speech implicitly through massive exposure to the sounds making up the L1 phonetic inventory. Also by hypothesis, the efficacy of implicit learning mechanisms would be reduced following the close of a critical period because it would cause L2 learners to lose the ability to make "automatic" use of input from "mere exposure" to the sounds making up the L2 phonetic inventory (Lenneberg, 1967, p. 176).

The ability to make effective use of ambient language phonetic input is the acknowledged prerequisite for L1 speech acquisition (e.g., Kuhl, 2000). According to a "cognitive change" hypothesis (DeKeyser & Larson-Hall, 2005), late learners fare well in early stages of L2 learning through the use of explicit learning mechanisms, but such mechanisms are not well suited for the slow process of attunement to the language-specific details defining L2 sounds and their differences from L1 sounds. Early learners, on the other hand, learn L2 phonetic details well but slowly via implicit learning mechanisms.

The SLM provided a way to understand the cross-over paradox without positing a loss of neural plasticity or a change in the cognitive mechanisms needed for speech learning. As mentioned earlier, research has shown (e.g., Flege & Hammond, 1982; see also Reiterer et al., 2013) that even late learners can gain access to the language-specific details defining L2 sounds. The SLM proposed that L2 phonetic input is accessible and that L2 learners of all ages exploit the same mechanisms and processes they used earlier for L1 speech learning, including the ability to create new phonetic categories for certain L2 sounds based on the experienced distributions of tokens defining those L2 sounds.

The SLM focused on the development of language-specific phonetic categories and the phonetic realization rules used to implement those categories motorically. The model assumed a generic three-level perception-production framework, illustrated in Figure 1.1, that envisages a flow of information from a sensory motor level to a phonetic category level to lexico-phonological representations (see, e.g., Evans & Davis, 2015).

A precategorical, auditory level of processing is evident only in specific perceptual testing conditions and is imperceptible to listeners (e.g., Werker & Logan, 1985), whereas the distinction between the phonetic category and lexico-phonological levels is more readily evident. For example, listeners can "hear" (i.e., perceive) a sound in the speech stream even when the sound has been replaced by silence or noise, thereby removing any phonetic-level information (e.g., Samuel, 1981). Evidence

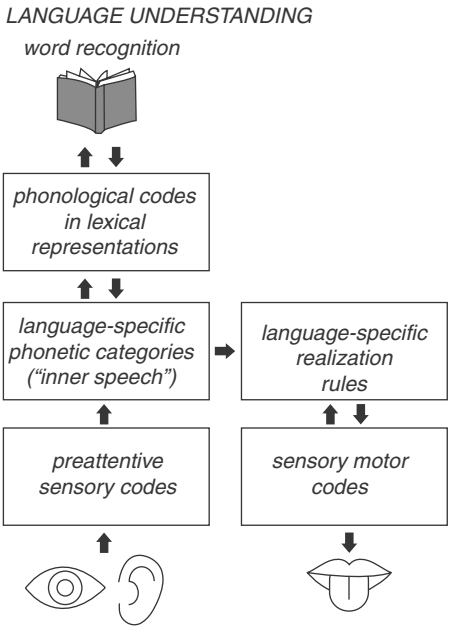

Figure 1.1 The generic three-level production–perception model assumed by the Speech Learning Model.

that sounds are categorized at a phonetic level is provided by the fact that monolingual listeners can recognize unfamiliar names heard for the first time.

Phonetic categories have two important functions. They define the articulatory goals used by language-specific phonetic realization "rules" in producing speech (but see Best, 1995, for a different perspective). More specifically, the realization rules "specify the amplitude and duration of muscular contractions that position the speech articulators in space and time" (Flege, 1992, p. 165). Second, phonetic categories are used to access segment-sized units of speech that, in turn, are used to activate word candidates during lexical access.

Listeners are usually not consciously aware of phonetic categories as they process speech because phonetic-level changes do not change meaning. However, language-specific phonetic categories are sufficiently rich in detail that they permit the detection of a fluent speaker as nonnative in as little as 30 ms (Flege, 1984). Moreover, phonetic-level differences can be detected when listeners focus attention on such differences (Best & Tyler, 2007; Pisoni, Aslin, Perey, & Hennessy, 1982).