Production training in Second Language Acquisition: A comparison between objective measures and subjective judgments

Véronique Delvaux[1,2], Kathy Huet[1,2], Myriam Piccaluga[1,2], Bernard Harmegnies[1,2]

[1] Laboratoire de Phonétique, [2]Institut de Recherche en Sciences et Technologies du Langage, UMONS, Belgium

delvaux@umons.ac.be

Abstract

This paper reports on an exploratory study of the processes involved in the acquisition of new phonetic control regimes in L2 learning. We focus here on the acquisition of long VOT initial stops by French native speakers undergoing production training. Francophone speakers were asked to repeat /ta/ stimuli varying in VOT and burst intensity. The performances in production are assessed through a comparison between *objective* measures performed on the speech signals (VOT, burst intensity) and *subjective* measures from L2 listeners themselves (in terms of similarity between the model and the response) and from American English native listeners (in terms of similarity as well as L1 typicality). Results show that (i) the Francophone speakers reasonably matched in their responses the VOT and burst intensity variations of the stimuli; (ii) that the three subjective indices are highly correlated with each other, but that they only partially correlate with the acoustic parameters measured on the signals; (iii) that inter-individual variation is very large, among the speakers'productions as well as among the listeners'judgments.

Index Terms: SLA, production training, VOT

1. Introduction

1.1. State of the art

In Second Language Acquisition (SLA) studies, two main types of independent variables have been manipulated in the laboratory in order to get a better understanding of the processes involved in the acquisition of L2 phonetics and phonology, i.e. (i) the amount and the nature of L2 stimuli and (ii) the specificities of the training methods. Concerning the stimuli, it has been shown that providing high variability within the training stimuli set makes nonnative listeners better at processing novel, untrained stimuli [1,2,3,4,5,6,7,8,9,10].

Concerning the training, the vast majority of laboratory studies used perception rather than production tasks to train participants. Different perceptual training procedures have been implemented in a large number of SLA studies, most of them resorting either to discrimination or to identification tasks, both tasks allowing significant improvements in performances [1,2,3,4,5,6,8,10,11,12,13,14,15,16,17,18,19, 20,21,22,23,24,25,26,27]. However, only a minor part of these perceptual training studies are concerned with the potential transfer of the newly acquired knowledge into the production domain [5,6,15,18,21,22,27], and when so, the performances in L2 production (after perceptual training) are usually assessed through the perceptual judgments from native listeners. Direct measurements of the productions are rarely provided. As to L2 laboratory studies involving production training itself, they are far less frequent than those involving perceptual training, although they have proved quite successful [28,29,30,31,32,33].

It is now necessary to carry out a large number of laboratory studies testing a variety of production tasks, if one wants to assess the specific gains production training (vs. perceptual training) can potentially yield in L2 learning, as well as to investigate the fundamental mechanisms underlying the acquisition of new, appropriate, perceptuo-motor routines in L2. Direct measurements of production performances are also needed, to be compared with perceptual judgments from native listeners. Indeed, while native listeners' judgments constitute the endpoint of any process of L2 learning (if the learners' productions pass the native listeners' test, then the learning process has achieved its goal), by using perceptual tests only one considers speech production in L2 learning as a 'black box'. If the L2 speakers fail to make progress from the native listeners' point of view, it may be because they did not diverge from their original (L1 typical) productions, or because they did diverge from them but in a wrong direction, or because they proceeded in the good direction but did not go far enough (i.e. they did not reach a certain threshold that is constitutive of the native listeners' decision). The three terms of the alternative have very different implications for any specific SLA study, and more broadly, for SLA theory and its applications in foreign languages classrooms.

1.2. The present study

The present paper reports on a laboratory study focusing on production training in L2 learning. In the main task of the experimental paradigm, adult early beginners are exposed to a set of sound stimuli, most of which they are not familiar with in L1, and are asked to (re)produce them as faithfully as possible. Stimuli are semi-synthetic sounds which internal phonetic properties are carefully controlled. Performances in production are assessed through direct objective measurements on the speech signal, as well as subjective measurements from both L2 learners themselves (in terms of similarity between the model and the response), and native listeners (in terms of similarity, as well as L1 typicality). A short perceptual task, i.e. an AX discrimination task (not reported here), was carried out before production training itself, allowing the participants to acquaint with the stimuli and their unfamiliar phonetic properties.

This study was designed as an exploratory study, in that several variables were manipulated in combination in the production training phase of the paradigm, in a first attempt to assess their potential role in supporting the efficiency of the production training. Three variables were investigated, one of them concerning the order of presentation of the stimuli, the other two relating to the way the stimuli set has been built, i.e. the use of a secondary cue (namely, Burst Intensity) potentially trading relations with VOT, and the resort to enhanced stimulation through extreme values of VOT and

Burst Intensity. The full experimental design is detailed below, whereas the present paper is less about the detailed examination of the factors affecting production performances than about the comparison between objective and subjective assessment of these performances.

2. Material and methods

2.1. Stimuli

The stimuli set is made of CV (pseudo-)words in which the phonetic properties of the initial voiceless consonant are manipulated along two dimensions: Voice Onset Time (VOT) and Burst Intensity (BI) [34].

All the stimuli originate from a natural speech signal consisting in the production of a */ta/ pseudo-word by a native speaker of American English. From this natural speech signal were retained in all the stimuli: the duration of the burst phase (20ms) and the level of aspiration noise, as well as the entire vowel (total duration: 210ms including a breathy onset of 30ms). The duration of the aspiration phase (and consequently the VOT) and the level of the burst have been manipulated.

These manipulations allowed us to build 36 stimuli varying in VOT and BI. VOT varies from 20ms to 100ms by incremental steps of 10ms (9 levels). BI can take 4 levels that have been computed by considering the overall level of the burst relatively to the level of the speech signal in the vowel, specifically in a 20ms-window centered on the vowel's peak amplitude. The original burst level (-15dB) has been multiplied, in instantaneous amplitude, by a factor of 1, 2, 4, and .5, resulting in stimuli's BI of respectively -15dB, -9dB, -3dB and -21 dB. Thus, both dimensions extend from a typical L1 value to an extreme value (infrequently observed in the world's languages), passing by a middle value which is typical of some languages having aspirated initial stops (e.g. English).

2.2. Experimental paradigm

The experimental paradigm consisted in a succession of 4 tasks. Task 1 was an AX discrimination task on pairs of stimuli differing in VOT. Task 2 was a production task. Stimuli were presented one at a time at fixed interval and the instructions were to "repeat as faithfully as possible, as if it was a word from a foreign language". Stimuli were presented in three successive blocks separated by short pauses. Each block is as follows. Consecutively for each BI level, the 9 stimuli were presented first in ascending order (i.e. from 20ms-VOT to 100ms-VOT), then in descending order (i.e. from 100ms-VOT to 20ms-VOT), then in pseudo-random order. After each order serie, the 20-ms VOT stimulus was presented six times in a row, in order to avoid contamination effects from the preceding serie. In total, each speaker produced 324 /ta/ (excluding the 20-ms VOT sequences).Task 3 was a similarity judgment task. Stimuli were presented by pairs and the instructions were to judge the similarity of the two members of the pair by means of a computerized Visual Analogue Scale (VAS). Each pair consisted in one production from task 2 (block 2 only) and the corresponding stimulus. A subset of the stimuli only was included in the similarity judgment test, i.e. the 60-ms stimuli of the 4 BI levels. Each of these four stimuli led to 42 productions during block 2 (14 speakers * 3 orders of presentations), for a total of 168 pairs to be judged. Task 4 was a typicality judgment task. The same 168 productions were presented one at a time and the instructions were to judge the typicality of the production just heard with respect to American English by VAS as in task 3.

2.3. Participants

A total of 30 subjects participated in the experiment: 14 French native speakers from Belgium (who participated in tasks 1 and 2, 6 of which also participated in task 3 one week later), and 16 American English native speakers from Louisiana who participated in tasks 3 and 4 in one single experimental session. None of the participants reported any hearing loss. All participants were administered a linguistic questionnaire in which they detailed their knowledge and experience with any relevant foreign language.

The Belgian French native speakers had a very limited experience of English and virtually no knowledge of any other language contrasting unaspirated vs. aspirated voiceless consonants. They were one male and nine female speakers aged 19 to 51. The American English native speakers had a basic knowledge of French through French classes and occasional practice with colleagues, family and friends.

2.4. Measures

The speech productions from task 2 were segmented semi-automatically. The segmentation process consisted in the positioning of 5 time labels on each production, respectively at: (1) burst onset; (2) aspiration phase onset; (3) onset of the vowel; (4) offset of the breathy phase of the vowel, if any; (5) end of the vowel.

Three measures have been computed on the speech productions based on these 5 labels: (i) VOT ('RespVOT', in ms), defined as the time interval between labels (1) and (3); (ii) vowel duration ('VowDur', in ms), corresponding to the time interval between labels (4) and (5); (iii) the relative-to-vowel burst intensity ('RespBI', in dB), computed as the ratio between RMS amplitude over the burst ((1)-(2) time interval) and RMS amplitude over a similar-duration window centered on the vowel's peak instantaneous amplitude. Note that the assessment of vowel duration does not directly stem from the experimental design (duration variability in uttered vowels was not expected, since vowel duration was kept constant across all stimuli), but from first informal listening of the participants' productions by the experimenters.

Raw data collected in tasks 3 and 4 by VAS have been transformed in 0-100 index values expressing the degree of similarity/typicality gathered for all (pairs of) stimuli. The data were transformed into a standardized unit, namely the *z*-score. *Z*-scores were computed for each listener separately.

3. Results

3.1. Task 2: production task

A MANOVA was carried out on the data from the productions of the Belgian French speakers (10 speakers S1 to S10 have been currently analysed). The dependent variables were the three measures: RespVOT, RespBI and VowDur, and the independent variables were the specifications of the stimuli (StimVOT; StimBI), the order of presentation of the stimuli (Order) and the Speaker variable (also considered as a fixed factor). The MANOVA revealed a variety of significant effects, from which we selected the most relevant ones for the description that follows.

First, the statistical analysis revealed several significant effects of the Speaker factor (alone and in interaction with other independent variables) on the three dependent variables, but mostly on RespVOT. This means that the speech productions collected during task 2 exhibit a large amount of inter-individual variation, most of which is modulated by the stimuli specifications. Fig.1 illustrates the significant interaction between Speaker and StimVOT on RespVOT ($F_{72, 2005}$=3.072; p<.001). Two speakers, S4 and S7, produced short VOT, typical of French, throughout the production training, whereas the remaining 8 speakers carried out the (re)production task with a certain amount of success, RespVOT increasing as StimVOT increased. However, only a few speakers could adjust their productions to the exact StimVOT values to be reproduced, some speakers producing extra-long VOT values, others exhibiting a large variability in their responses.

Second, RespVOT significantly varied as a function of StimVOT and StimBI (in interaction: $F_{24,2005}$=3.133; p<.001). Interestingly, the same was true of VowDur, although the StimVOT ($F_{8,2005}$=9.505; p<.001) and StimBI ($F_{3,2005}$=7.745; p<.001) had independent effects. As illustrated in Fig.2, the 10 Belgian French speakers tend to produce longer VOT and longer vowels when StimVOT is long and burst intensity is low.

3.2. Similarity and typicality judgments

A MANOVA with mean *z*-scores for the three subjective indices as dependent variables and StimBI and Order as independent variables revealed that StimBI has a significant effect on the three dependent variables (typicality: $F_{3,156}$= 14.998; p<.001; similarity (AE listeners): $F_{3,156}$= 29.221; p<.001 ; similarity (BF listeners): $F_{3,156}$= 28.679; p<.001), whereas neither Order nor the interaction of Order and StimBI does. As illustrated in Fig.3, the three subjective indices exhibit rather close values across stimuli although these indices are based on different judgment tasks and/or different listeners. The effect of StimBI essentially resides in the fact that the responses to -21dB stimuli are overall poorly rated.

Three separate repeated measures ANOVA were further performed, one per subjective index, with individual mean *z*-scores as the within-subjects variables and StimBI and Speaker as between-subjects factors. Results are notably similar across the judgment tasks and/or group of listeners. Concerning first the within-subjects variables, when considered alone they never reaches statistical significance, whereas all two-way and three-way interactions do. As to the between-subjects effects, StimBI, Speaker as well as the interaction between them all yield significant variations in the three types of perceptual ratings. In other words, any particular score, whether in a typicality task or in a similarity task, whether given by an American English listener or by a Belgian French listener, depends not only on the BI level of the original stimulus, but also on the listener who performed the task and on the speaker who pronounced the production to be rated. The interaction between the Speaker and StimBI effect is illustrated in Fig.4, in that some speakers are overall (reasonably) well rated for their pronunciations in response to -21dB-BI stimuli, whereas others are not.

3.3. Comparison between objective and subjective measurements

A series of two-tailed Pearson correlations were carried out in order to investigate the relationships between the three subjective indices (mean *z*-scores) on one hand, and between these indices and the acoustic properties as measured directly on the productions on the other hand (Table 1). Results show that the agreement between the subjective indices is moderate to high, but always significant. Moreover, Typicality ratings (from AE listeners only) are more closely related to Similarity ratings from the same group of listeners than are Similarity ratings from the two different groups of listeners.

Concerning the relationships between the three types of indices and the properties of the rated productions, there is only a moderate (but highly significant) correlation between RespBI and the judgments from the American listeners in both tasks. Notably, no subjective index is significantly correlated with the VOT of the productions to be rated (which could largely vary even if all these productions were recorded in Task 2 in response to 60ms-VOT stimuli). However, a comparison between Fig.1 and Fig.4 suggests that the mean values of the VOT produced by each speaker in response to 60ms-VOT stimuli could be a good predictor of his/her overall rating by the listeners. Indeed, the most poorly rated speakers are S4, S7 and S5, whose meanVOT in response to 60ms-VOT stimuli is respectively of 17ms, 18ms and 118ms. The best subjective ratings are attributed to S6 (62ms).

4. Discussion

In summary, results show that the adult L2 learners were quite successful when performing the imitation task. On average, acoustic measurements show that the speakers' productions varied as a function of the properties of the stimuli they were exposed to, and in the appropriate direction, although the data exhibit a large amount of intra-speaker as well as inter-speaker variation. Eight out of 10 speakers increased their VOT as StimVOT increased, the more so when long VOT were enhanced by a secondary phonetic cue, i.e. reduced Burst Intensity. These first results point to the usefulness of further investigating the potential of various modalities of production training in SLA.

Moreover, objective acoustic measurements proved valuable for the assessment of production performances. For example, they show that S5 is sensitive to VOT variations in the stimuli (Fig.1), although he produces far too long VOT in response, and is thus poorly rated by all listeners in the three subjective judgment tasks. They also allow to investigate how speakers behave when asked to modified their motor routines in order to achieve new targets (here a new type of inter-gestural timing between laryngeal and supralaryngeal gestures). Indeed, several Belgian French speakers happen to vary vowel duration in response to the (fixed vowel duration) stimuli they are told to reproduce instead of, or in complement with, varying VOT. This (compensatory?) strategy seems poorly rewarded by the listeners since VowDur is not correlated with any subjective index. However, overall it remains difficult to draw specific connections between variations in these subjective indices and variations in the acoustic properties of the judged productions. Since there is a fair amount of agreement among subjective judgments (in spite of high inter-listeners variability), further work is needed in order to establish which compound of acoustic-phonetic properties they are based on.

5. Acknowledgements

This research has been subsidized by the AUWB- 08/12-UMH 17 ARC convention from the Communauté française de Belgique and by the MCF/FRC project - FRFC 2.4644.09 from the Belgian Fonds National de la Recherche Scientifique.

Table1.*Results of the two-tailed Pearson correlations*

	Similarity (AE)	Similarity (BF)	Resp VOT	Resp BI	Vow Dur
Typicality	.85**	.5**	.09	-.3**	-.03
Similarity (AE)		.59**	-.12	-.24**	-.08
Similarity (BF)			-.02	.08	.03
RespVOT				.21*	.34**
RespBI					.34**

Figure 1: *Production data: RespVOT as a function of StimVOT and Speaker*

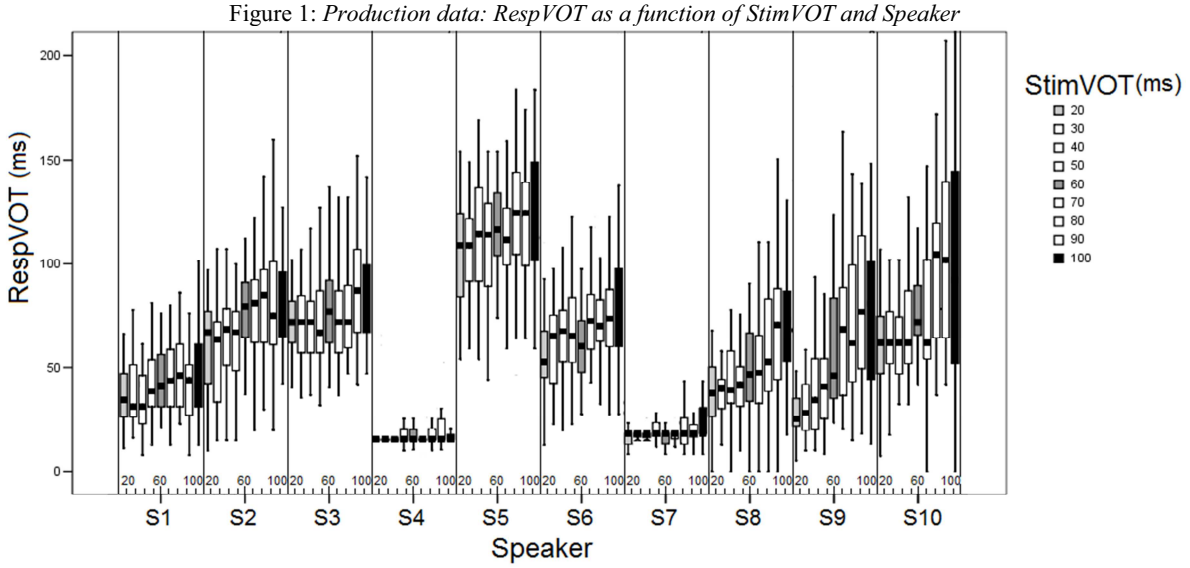


Figure 2: *Production data: RespVOT (left) and VowDur (right) as a function of StimVOT and StimBI (10 speakers pooled)*

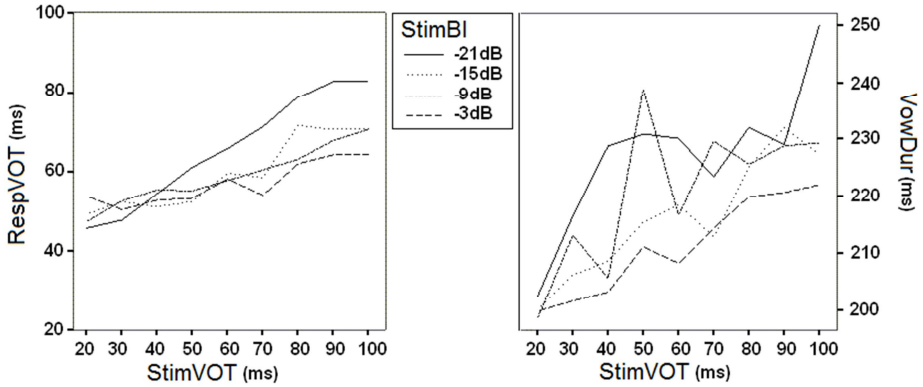


Figure 3. *Subjective indices (mean z-scores) as a function of Burst Intensity*

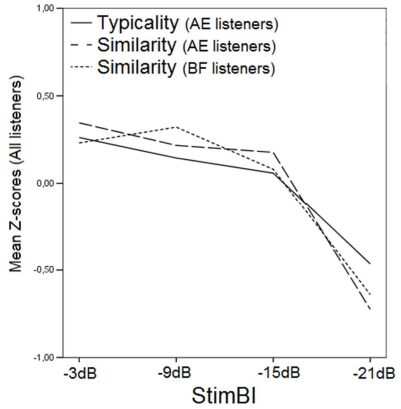
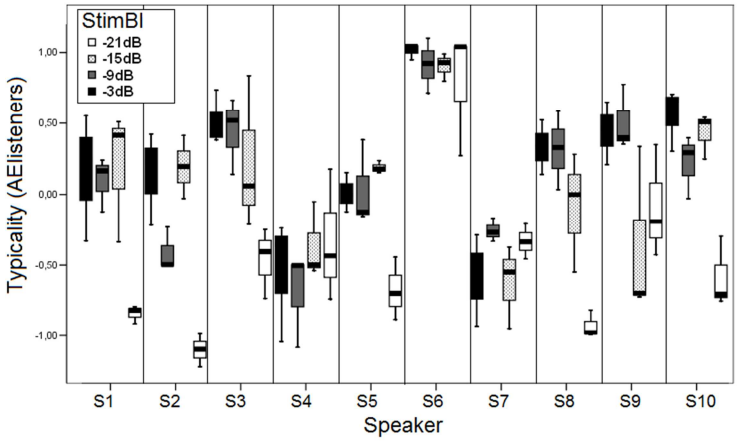


Figure 4. *Typicality values (mean z-scores) as a function of Speaker and StimBI*



6. References

[1] Logan, J. S., Lively, S. E., & Pisoni, D. B., "Training Japanese listeners to identify English /r/ and /l/: A first report", Journal of the Acoustical Society of America, 89: 874-886, 1991.

[2] Lively, S. E., Logan, J. S., & Pisoni, D. B., "Training Japanese listeners to identify English /r/ and /l/: II. The role of phonetic environment and talker variability in learning new perceptual categories" Journal of the Acoustical Society of America, 94: 1242-1255, 1993.

[3] Lively, S. E., Pisoni, D. B., Yamada, R., Tohkura, Y., & Yamada, T., "Training Japanese listeners to identify English /r/ and /l/. III. Long-term retention of new phonetic categories", Journal of the Acoustical Society of America, 96:2076-2087, 1994.

[4] Magnuson, J. S., Yamada, R. A., Tohkura, Y., & Bradlow, A. R., "Testing the importance of talker variability in non-native speech contrast training", Journal of the Acoustical Society of America 97(5), Pt. 2: 3417, 1995.

[5] Bradlow, A. R., Pisoni, D. B., Yamada, R. A., & Tohkura, Y., "Training Japanese listeners to identify English /r/ and /l/: IV. Some effects of perceptual learning on speech production", Journal of the Acoustical Society of America, 101: 2299–2310, 1997.

[6] Bradlow, A. R., Akahane-Yamada, R., Pisoni, D. B., & Tohkura, Y., "Training Japanese listeners to identify English /r/ and /l/: Long-term retention of learning in perception and production", Perception and Psychophysics, 61: 977–985, 1999.

[7] Wang, Y., Jongman, & A. Sereno, J.A., "Acoustic and perceptual evaluation of Mandarin tone productions before and after perceptual training", Journal of the Acoustical Society of America, 113 (2): 1033-1043, 2003.

[8] Pruitt, J. S., "Perceptual training on Hindi dental and retroflex consonants by native English and Japanese speakers", Journal of the Acoustical Society of America, 97: 3417, 1995.

[9] Kingston J., "Learning foreign vowels", Language and Speech, 46: 295–349, 2003.

[10] Iverson, P., Hazan, V., & Bannister K., "Phonetic training with acoustic cue manipulations: A comparison of methods for teaching English /r/–/l/ to Japanese adults", Journal of the Acoustical Society of America, 118: 3267–3278, 2005.

[11] Flege, J., "Second-language speech learning : theory, findings, and problems", in W. Strange [Ed], Speech Perception and Linguistic Experience. Issues in Cross-Language Research, 233-277, York Press, 1995.

[12] Flege, J., & Wang, C., "Native language phonotactic constraints affect how well Chinese subjects perceive the word-final English /t/-/d/ contrast", Journal of Phonetics, 17: 299–315, 1990.

[13] Hirata, Y., "Training native English speakers to perceive Japanese length contrast in word versus sentence contexts", Journal of the Acoustical Society of America, 116: 2384–2394, 2004.

[14] Jamieson, D. G. &Morosan, D. E., "Training new, nonnative speech contrasts: A comparison of the protoTypicalitye and perceptual fading techniques", Canadian Journal of Psychology, 43: 88–96, 1989.

[15] Lambacher, S, Martens, W. L., Kakehi, K., Marasinghe, C. A., Molholt, G., "The Effects of Identification Training on the Perception and Production of American English Vowels by Native Speakers of Japanese", Applied Psycholinguistics, 26 (2): 227-247, 2005.

[16] Polka, L. "Cross-language speech perception in adults: Phonemic, phonetic, and acoustic contributions", Journal of Acoustical Society of America, 89: 2961–2977, 1991.

[17] Pruitt, J., Strange, W., Polka, L., & Aquilar, M., "Effects of category knowledge and syllable truncation during auditory training on American's discrimination of Hindi retroflex-dental contrast", Journal of the Acoustical Society of America, 87: S72(A), 1990.

[18] Rochet, B., "Perception and production of second language speech sounds by adults", in W. Strange [Ed], Speech Perception and Linguistic Experience : Issues in Cross-Language Research, 379-410, York Press, 1995.

[19] Strange, W., & Dittman, S., "Effects of discrimination training on the perception of /r-l/ by Japanese adults learning English", Perception and Psychophysics, 36:131-145, 1984.

[20] Wang, Y., Spence, M. M., Jongman, A., & Sereno, J. A., "Training American listeners to perceive Mandarin tone", Journal of the Acoustical Society of America, 106:3649–3658, 1999.

[21] Wang, Y., Jongman, & A. Sereno, J.A., "Acoustic and perceptual evaluation of Mandarin tone productions before and after perceptual training", Journal of the Acoustical Society of America, 113 (2): 1033-1043, 2003.

[22] Wang, X., "Perceptual Training for Learning English Vowels: Perception, Production, and Long-term Retention", VDM Verlag Dr. Müller, 2008.

[23] Wayland, R., & Guion, S., "Perceptual discrimination of Thai tones by naı¨ve and experienced learners of Thai", Applied Psycholinguistics, 24: 113–129, 2003.

[24] Wayland, R., & Guion, S., "Training native English and native Chinese speakers to perceive Thai tones", Language Learning, 54(4): 681–712, 2004.

[25] Wayland, R. & Li, B., "Effects of two training procedures in cross-language perception of tones", Journal of Phonetics, 36:250–267, 2008.

[26] Akahane-Yamada, R., "Effects of extended training on /r/ and /l/ identification by native speakers of Japanese", Journal of the Acoustical Society of America, 93 (4, Pt. 2): 2391, 1993.

[27] Akahane-Yamada, R., Strange, W., Downs-Pruitt, J.C. and Masuda, Y., "Modification of L2 vowel production by perception training as evaluated by acoustic analysis and native speakers", Journal of the Acoustical Society of America, 103 (5):3089-3089, 1998.

[28] Akahane-Yamada, R., Adachi, T., & Kawahara, H., "Second language production training using spectrographic representations as feedback", Journal of the Acoustical Society of Japan, 18: 341–343, 1997.

[29] Dalby, J., & Kewly-Port, D., "Explicit pronunciation training using automatic speech recognition", in M. Holland [Ed], Tutors that listen: Speech recognition for language training, Special issue of the Journal of the Computer Assisted Language Learning Consortium (CALICO), 16(5): 425–445, 1999.

[30] Gómez Lacabex, E., "Perception and production of vowel reduction in second language acquisition", Unpublished doctoral dissertation, University of the Basque Country, Spain, 2009.

[31] Gomez Lacabex, E. & Garcia Lecumberri, "Over/generalization effects in the production of English weak and full vowels in unstressed syllables after vowel reduction training: evidence for L2 sound learning in progress", Proceedings EUROSLA 22, 2012.

[32] Hanulíková, A., Dediu, D., Fang, Z., Bašnaková, J. and Huettig, F., "Individual differences in the acquisition of a complex L2 phonology: a training study", special issue of Language Learning. In press.

[33] Aliaga-Garcia, C. and Mora, J.C. , "Assessing the effects of phonetic training on L2 sound perception and production", in M.A. Watkins, A.S. Rauber, B.O. Baptista [Ed], Recent reseach in second language phonetics/phonology, perception and production, 2-31, 2009.

[34] Piccaluga, M., Clairet, S., Delvaux, V., Huet, K., Harmegnies, B., "Guidage perceptuel de la production en L2: Tendances générales et variabilité individuelle", Recherches Anglaises et Nord-Américaines, Special Issue, 25-44, 2011.