The effect of phonetic production training with visual feedback on the perception and production of foreign speech sounds[a)]

Natalia Kartushina,[1,b)] Alexis Hervais-Adelman,[2] Ulrich Hans Frauenfelder,[1] and Narly Golestani[2]

[1]*Laboratory of Experimental Psycholinguistics, Faculty of Psychology and Educational Sciences, University of Geneva, 42 bd du Pont d'Arve, 1205 Geneva, Switzerland*

[2]*Neuroscience Department, Brain and Language Lab, Faculty of Medicine, University of Geneva, Campus Biotech, 9 Chemin des Mines, 1211 Geneva, Switzerland*

(Received 25 July 2014; revised 13 June 2015; accepted 27 June 2015; published online 17 August 2015)

Second-language learners often experience major difficulties in producing non-native speech sounds. This paper introduces a training method that uses a real-time analysis of the acoustic properties of vowels produced by non-native speakers to provide them with immediate, trial-by-trial visual feedback about their articulation alongside that of the same vowels produced by native speakers. The Mahalanobis acoustic distance between non-native productions and target native acoustic spaces was used to assess *L*2 production accuracy. The experiment shows that 1 h of training per vowel improves the production of four non-native Danish vowels: the learners' productions were closer to the corresponding Danish target vowels after training. The production performance of a control group remained unchanged. Comparisons of pre- and post-training vowel discrimination performance in the experimental group showed improvements in perception. Correlational analyses of training-related changes in production and perception revealed no relationship. These results suggest, first, that this training method is effective in improving non-native vowel production. Second, training purely on production improves perception. Finally, it appears that improvements in production and perception do not systematically progress at equal rates within individuals. © *2015 Acoustical Society of America*.
[http://dx.doi.org/10.1121/1.4926561]

[LK] Pages: 817–832

I. INTRODUCTION

The attainment of native-like pronunciation in a second language (*L*2) is claimed to be beyond the grasp of learners who start acquiring the *L*2 after the age of 6 yrs, or even earlier (Piske *et al.*, 2001). These "late" *L*2 speakers tend to be easily identified by native speakers as having a foreign accent. The current work focuses on difficulties that *L*2 speakers experience in the production of non-native isolated phonological segments, in particular, vowels, and aims to evaluate the effect of production training with acoustic feedback regarding the articulation of vowel production in *L*2 learners.

A. Role of native language (*L*1) phonology and other factors in the production and perception of *L*2 sounds

Foreign accents in late *L*2 learners, and even in early bilinguals (i.e., when *L*2 is learned at or before the age of 3), are widely documented in the literature for different *L*1–*L*2 combinations (Ingram and Park, 1997; Piske *et al.*, 2001). For instance, Korean and Spanish learners of English have difficulties in producing the /i/-/ɪ/ contrast (e.g., sheep-ship), whereas Italian learners of English experience more difficulties with the /ɒ/-/ʌ/ contrast (e.g., bought-but). The dominant theoretical perspective attributes these production difficulties to a bias in *L*2 perception stemming from the *L*1 phonology and its relation to that of the *L*2 (Best, 1995; Flege, 1995).

According to the Full Transfer Model, the *L*1 system constitutes the initial state for *L*2 acquisition, in perception and in production. This means that at the beginning of *L*2 learning, when the *L*2 system is "empty," the *L*1 system (i.e., *L*1 abstract categories and perception grammar that "maps auditory events to phonological structures," Escudero and Boersma, 2004, p. 583) is used to process *L*2 sounds. At this stage it is assumed that the non-native (empty) and native languages form a single inter-language phonological space, where the native phonological system prevails, and filters foreign sounds to established *L*1 categories. *L*2 sounds are perceived through this filter as a function of their similarity and dissimilarity to the *L*1 sounds. Perceptually similar *L*2 sounds assimilate to native sounds (i.e., they are integrated into an existing *L*1 category), whereas dissimilar ones do not (Best's Perceptual Assimilation Model [PAM], 1995; Flege's Speech Learning Model [SLM], 1995). According to the SLM (Flege, 1995), *L*2 speakers will have the most difficulty in correctly discerning the phonetic differences between an *L*2 sound and a similar existing *L*1 category. According to the PAM (Best, 1995), *L*2 sounds are perceived based on their gestural similarity to native phonemes, and

[a)]Portions of this work were presented at International Workshop on Language Production, Geneva, July 2014.

[b)]Electronic mail: Natalia.Kartushina@unige.ch

can be either assimilated or uncategorized. Three patterns of assimilation are identified: (i) Two-Category assimilation, where two *L*2 sounds correspond to two different *L*1 categories, (ii) Category-Goodness (CG) assimilation, where both *L*2 sounds map onto one *L*1 category, but one *L*2 sound is a better exemplar of this *L*1 category than the other, and (iii) Single-Category (SC) assimilation, where both *L*2 sounds are approximately equally acceptable or equally deviant exemplars of the *L*1 category. The perception of the various phonetic contrasts of *L*2 is predicted to range from perfect to poor across these respective assimilation patterns. For example, Japanese *L*2 speakers tend to have difficulty discriminating English /r/ and /l/, which they assimilate to one Japanese phoneme in a SC manner (Bradlow *et al.*, 1997).

The above models assume that the acquisition of *L*2 production follows that of *L*2 perception. Once the *L*2 categories are established in perception, the same categories will be used to guide *L*2 production: "… without accurate perceptual" targets "to guide the sensorimotor learning of *L*2 sounds, production of the *L*2 sounds will be inaccurate…" (Flege, 1995, p. 238). For example, Korean speakers who fail to perceive the difference between the English /e/ and /æ/ vowels will also fail to differentiate these vowels in production. Indeed, acoustic analyses of their productions reveal that the formant spaces (first formant [*F*1] and second formant [*F*2]) of these two vowels largely overlap (Ingram and Park, 1997).

Other studies, however, challenge this "perception first" assumption by showing either no correlation (Peperkamp and Bouchon, 2011) or only moderate correlations (Flege *et al.*, 1999) between *L*2 perception and production accuracy. Moreover, some studies show inaccurate production despite successful perception of non-native phonetic differences (Golestani and Pallier, 2007), or accurate production despite poor perception (Sheldon and Strange, 1982). These dissociations imply that factors other than *L*2 perception also influence *L*2 production performance, and indeed, there is evidence for other such factors. First, there exist factors related to *L*2 production itself such as the articulatory difficulty of *L*2 sounds and their acoustic-phonetic environment. For example, the same *L*2 sound may be produced more or less accurately depending on the phonetic context that precedes or follows it (Saito and Lyster, 2012; Sheldon and Strange, 1982; Steinlen, 2005). Second, certain learner/talker characteristics such as sensory-motor control (Simmonds *et al.*, 2011) and the compactness (i.e., consistency/lack of variability) of *L*1 productions (Kartushina and Frauenfelder, 2014) may also play a role. Third, there exist factors related to individual differences outside the domain of language competence: individuals' *L*2 production depends on their motivation to learn and on their empathy levels (Hu *et al.*, 2013). Fourth, there are factors related to the *L*2 learning process such as *L*2 learning strategies and quality of *L*2 input (Flege, 2002), that influence the quality of *L*2 productions. Last, there are factors that are directly related to experience in *L*2, such as age of arrival and length of residence in an *L*2-speaking environment (Flege, 1995), age of *L*2 acquisition, experience with *L*2, and amount of *L*1 use (Piske *et al.*, 2001).

The difficulties that *L*2 speakers experience in the perception and production of non-native sounds can be partly overcome by training. Although both *L*2 perception and production can in principle be trained, training studies have more often focused on improving the former.

B. *L*2 training studies

1. Perception

*L*2 perception training studies usually aim to improve the perception of "difficult" *L*2 contrasts, i.e., ones that assimilate to one *L*1 category (SC or CG, according to Best's PAM, 1995) (e.g., for consonants: Bradlow *et al.*, 1997; Lively *et al.*, 1993; Lopez-Soto and Kewley-Port, 2009; for vowels: Wong, 2013; for tone contrasts: Perrachione *et al.*, 2011). The training procedures have typically involved one or more perception tasks (for example, discrimination and/or identification) where, for example, minimal pairs of words, pseudo words or segments are trained and tested (Bradlow *et al.*, 1997; Lopez-Soto and Kewley-Port, 2009). During training, trial-by-trial feedback on performance ("correct" vs "incorrect") is usually given. Perception training studies have shown relatively consistent results: *L*2 speakers' perception clearly benefits from perception training, with performance improving by 10%–20% (Bradlow *et al.*, 1997; Lopez-Soto and Kewley-Port, 2009; Wong, 2013). However, improvements in perception have been shown to only partially transfer to production (7% improvement in production of words in a repetition task, as judged by native speakers of English in Bradlow *et al.*, 1997), or not at all (Lopez-Soto and Kewley-Port, 2009).

2. Production

Techniques have also been developed that more directly aim at improving the accuracy of *L*2 sound production. Some of these studies adopted a combined approach where *L*2 perception was trained in conjunction with *L*2 production (e.g., Aliaga-Garcia and Mora, 2009; Delvaux *et al.*, 2013; Massaro *et al.*, 2008; Wong, 2013). These studies, with the exception of that by Aliaga-Garcia and Mora (2009), which tested multiple production tasks and computer-based visual feedback, have concluded that the approach of combining both pronunciation and perception training is effective in improving *L*2 production accuracy.

Although studies that exclusively train *L*2 production are still rare, some existing research provides encouraging results. Pure production training (where perception is not trained) improves the pronunciation of *L*2 sounds, as judged by acoustic measures (Dowd *et al.*, 1998; Leather, 1996; Öster, 1997; Pillot-Loiseau *et al.*, 2013), native speakers' ratings (Akahane-Yamada *et al.*, 1998; Dowd *et al.*, 1998; Wong, 2013) or phoneticians (Carey, 2004; Catford and Pisoni, 1970). Importantly, corrective feedback appears to be crucial in production training, since articulatory instruction alone has not been shown to improve *L*2 pronunciation as attested by acoustic analyses of production before and after training [Saito and Lyster (2012), but see Catford and Pisoni (1970) and Leather (1996), for positive effects of articulatory instruction on production of "exotic" sounds and of tones in naive listeners]. The impact of production training on

perception (the untrained modality) ranges from zero (Wong, 2013) to 3%–4% improvement (Akahane-Yamada *et al.*, 1998; Wik, 2004). These effects of production training on perception are similar in magnitude to those of perception training on production described above. The relatively weak or non-existent transfer effects suggest that training is, for the most part, modality-specific.

Typically, production-training studies involve providing *L*2 speakers with visual feedback comparing their production of *L*2 sounds to that of native *L*2 speakers (the "target"). Based on an analysis of methods that have been used in past studies, we note that although the actual feedback that is provided to participants differs considerably across studies, it can be divided into two main types: direct and indirect feedback. The former provides the participant with an immediate and dynamic view of the position and movements of their articulators during production. This can be achieved, for example, with ultrasound imaging of the articulators during production (Pillot-Loiseau *et al.*, 2013; Wilson and Gick, 2006) or electropalatography. Indirect feedback provides information about the articulation of *L*2 sounds, derived from acoustic analyses of the participants' *L*2 productions (Akahane-Yamada *et al.*, 1998; Dowd *et al.*, 1998; Öster, 1997). Although studies using direct feedback have been shown to improve *L*1 speech production in clinical populations such as patients with speech disorders (e.g., Wilson and Gick, 2006), their effectiveness in the context of *L*2 learning has not been systematically evaluated in large samples, principally because of their cost. A recent study by Pillot-Loiseau and colleagues (2013) has, however, demonstrated improvements in production of French /u/ and /y/ vowels in two Japanese learners who were provided with ultrasound images of their articulation.

Indirect feedback (i.e., via acoustic comparisons of the produced and target speech) is most frequently used in *L*2 training studies. It can also be divided into two sub-types. The first involves providing participants with a representation of the raw acoustic properties of *L*2 sounds, e.g., formants or resonance frequencies (Akahane-Yamada *et al.*, 1998; Dowd *et al.*, 1998). For instance, Akahane-Yamada and colleagues (1998) showed Japanese participants spectrograms depicting the first three formants (*F*1/*F*2/*F*3) of their productions of the English /r/ and /l/ phonemes, compared to those of native English speakers. The intelligibility of the *L*2 productions, as evaluated by native English listeners, was shown to improve by 22% after 4 h of training. In another study, native speakers of Australian English were trained to pronounce isolated French oral vowels using spectrograms of the resonance frequencies of the vocal tract (*F*1, *F*2) as feedback (Dowd *et al.*, 1998). Although training was effective for phonetic differences that were easily detected based on this feedback (e.g., *F*2 is noticeably different for the /y/-/u/ vowels), participants did not improve on vowels that were similar to one another along the dimensions used for feedback (e.g., both *F*1 and *F*2 are similar for /a/-/ɑ/ and for /e/-/ɛ/).

Indirect feedback can also be provided using simplified or more abstract graphic representations of acoustic information (i.e., instead of using raw values), where phonetic differences that are relevant to the trained contrast(s) are enhanced or highlighted (Carey, 2004; Öster 1997). For instance, Öster (1997) trained 13 participants with different native languages (individuals from Bosnia, Cuba, Peru, Saudi Arabia, and Russia) to pronounce voiced and voiceless contrasts (e.g., Swedish "buss" versus "puss") by showing them colored contrastive visual acoustic patterns representing the presence or absence of voicing. This study revealed that training was beneficial for *L*2 perception and production, but quantitative details relating to the improvement were not reported. Carey (2004) trained native Korean speakers to produce the English vowels /æ/, /ɜ:/ and /ɔ/ by providing them with both verbal instructions about articulation (supported by videos of native speakers producing vowels), and graphic representations, in the *F*1/*F*2 acoustic space of the trained English vowels and of similar Korean vowels /ɛ/ and /o/. This method was successful in improving the production of only one vowel (/æ/) out of the three trained ones.

In parallel with the development of *L*2 training methods that provide feedback about articulation via acoustic analyses of the produced speech, there also exists a very promising line of engineering research, which aims to create virtual three-dimensional (3D) space language tutors, or humanoid talking heads (Massaro *et al.*, 2008; Wik, 2004). These 3D tutors are shown using a dynamic frontal view or sometimes a dynamic mid-sagittal section of the articulatory gestures that are required for accurate *L*2 production. This approach aims to help *L*2 learners to improve their pronunciation (i) by making them aware of the required articulatory gestures, and (ii) by correcting their pronunciation. These 3D tutor methods must be improved before they can be used in *L*2 learning settings (cf. Wik, 2004). Moreover, it appears that use of these animated tutors results in a similar amount of improvement as the use of indirect articulatory feedback (e.g., spectrograms, as used by Wik, 2004).

3. Limitations of existing L2 production training studies

The number of production training studies available in *L*2 learning settings is limited. It appears that training with indirect, acoustically-based feedback offers a good balance between implementation difficulty and observed improvement, for both *L*2 learners and for experimenters. However, the results of the above-described *L*2 production training studies are neither conclusive, nor comparable, with one another, for a number of reasons. First, control groups have rarely been included, and if so, they have not been rigorously matched to the experimental ones (Akahane-Yamada *et al.*, 1998; Aliaga-Garcia and Mora, 2009; Carey, 2004; Dowd *et al.*, 1998; Öster, 1997; Wong, 2013). As a result, improvements observed in the experimental group may have arisen simply as a result of repeated attempts to produce the targets (i.e., articulatory practice), and not from the feedback itself. Second, most of the above-described studies can be criticized for the heterogeneous *L*2 proficiency levels of the participants. Third, the type of feedback that has been used is not always easy for *L*2 speakers to use rapidly and to interpret during training, and may even not be effective for

certain contrasts (e.g., studies showing no improvement on some *L*2 sounds: Aliaga-Garcia and Mora, 2009; Dowd *et al.*, 1998; Carey, 2004). It is, for example, important for the visual representation of the feedback to be logical and easily interpretable (Öster, 1997). The feedback should provide contrastive training (i.e., the crucial differences have to be represented), and it should be presented immediately after *L*2 production (Öster, 1997). Last, and critically, several previous studies have utilized subjective ratings by native listeners rather than objective acoustic analyses to assess the success of their training approach. Subjective ratings are prone to rater bias, and are not fully reproducible. As recently shown by Delvaux and colleagues (2013), objective measures of production accuracy (i.e., acoustic analyses) are a valuable and more sensitive tool for the assessment of training effects. Our study addresses all of the above considerations.

C. Current study: Overview and goals

The current study aims to evaluate the effect of simplified, indirect articulatory feedback training on the production of non-native Danish (DK) vowels by native French speakers who have no experience of the Danish language. To this end, four isolated DK vowels (/e/ and /ɛ/, and /y/ and /ø/, which form two height contrasts, see Fig. 1) were trained over five sessions, with 1 h of training per vowel in total. Their production was assessed before and after training, using repetition of isolated vowels. We also tested whether improvements in production were accompanied by improvements in perception; for this purpose, an *ABX* discrimination task (within the height-contrastive vowel pairs) was used before and after the training. An *ABX* discrimination task was used rather than an identification task in order to avoid the memory load associated with having to retain labels.

Our training method involves a real-time analysis of the acoustic properties of the vowels spoken by participants, and immediate, trial-by-trial visual feedback on two parameters that are deemed to be critically contrastive for these vowels within their pairs: tongue height (mouth openness), as reflected by *F*1, and tongue front–back position, as reflected by *F*2 (which can also reflect lip rounding). On each trial, participants repeated the DK target vowels, and immediately afterwards saw a visual display showing the location, in *F*1/*F*2 space, of their production, along with that of the DK vowel which they had just repeated. Participants were trained using tokens recorded by a speaker of the same sex as themselves.

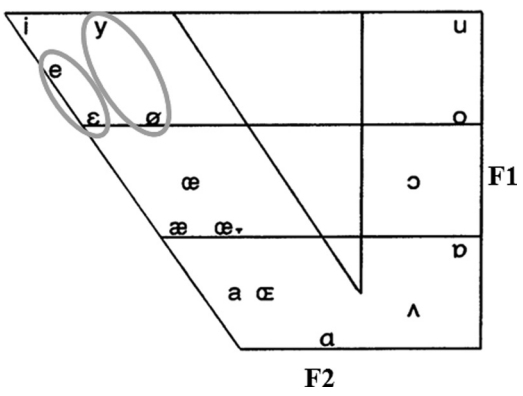

FIG. 1. Danish vowel space (adapted from Grønnum, 1997) indicating the stimuli used for training (circled); the *F*1 and *F*2 values increase from up to down on the *y* axis and from right to left on the *x* axis, respectively.

In order to properly assess the training effect associated with the visual articulatory feedback based on the participants' own productions, and to distinguish it from mere exposure to, and repetition of, the non-native vowels, a control group was included. The production and perception performance of this group was assessed before and after training under the same conditions as the experimental group. During their training, participants in the control group received as many exposures to the Danish vowels, saw the same visual *F*1 and *F*2 information about these vowels produced by native Danish speakers (but not by themselves) on every trial, and repeated them as often as the experiment group. In order to keep them engaged and motivated in performing the task, the control group received aggregate feedback based on the mean proximity of their productions in *F*1 but not *F*2, presented at the end of each block of 21 trials. This impoverished feedback served to help to match the two groups with respect to the motivation that receiving feedback regarding performance might confer.

The effectiveness of the experimental training and of the control training was assessed by computing the Mahalanobis distances (calculated as described in Sec. II D 2 below), between the vowels produced during pre- and post-training tests and the native DK target spaces derived from recordings of native Danish speakers. The Mahalanobis distance is a unitless, scale-invariant measure of the distance between a point and a distribution that takes into account the distance, measured in standard deviations, along the principal component axes of the distribution. It has been used in techniques of pattern recognition, data clustering and classification, and speaker identification, where an unknown speech sample is assigned to a speaker on the basis of the minimum distance between a test speech sample and the reference samples. It has previously been used to calculate *L*2 production accuracy (Kartushina and Frauenfelder, 2014). This metric has the advantage over simple Euclidian distances in that it allows the natural variability of native speech production to be taken into consideration when assessing the accuracy of production.

We hypothesized that articulatory feedback training would improve production accuracy of non-native DK vowels, but that there would be no change in performance in the control group, because repetition and articulatory instruction alone do not lead to improvements in *L*2 production (e.g., Saito and Lyster, 2012). Based on the above-mentioned studies showing (1) dissociations between *L*2 perception and production (Peperkamp and Bouchon, 2011), (2) no correlation between improvements across these modalities (Bradlow *et al.*, 1997), and (3) weak transfer of training effects between the two modalities (Akahane-Yamada *et al.*, 1998; Bradlow *et al.*, 1997; Lopez-Soto and Kewley-Port, 2009; Wik, 2004; Wong, 2013), we further predicted that improvements in production would partially transfer to perception,

but that some participants may show no relationship between improvements in performance across modalities.

II. MATERIALS AND METHODS

A. Participants

Twenty-seven monolingual native French speakers from the student body of the University of Geneva participated in the study (20 female and 7 male, mean age = 24 yrs, 8 months). Some participants had some knowledge of other languages (German, English, Spanish, and Italian), but none of these languages was reported as being spoken proficiently. They reported no history of speech or hearing impairment. None reported any experience with Scandinavian languages (Danish, Norwegian, or Swedish). Participants were paid for their participation. The experiment was conducted in accordance with the declaration of Helsinki. Participants gave informed consent and were free to withdraw from the experiment at any time.

B. Stimuli

We trained participants to produce four isolated DK vowels that form two height contrasts: /e/ and /ɛ/, and /y/ and /ø/. Typically, height contrasts differ in terms of mouth closure during articulation, and are well characterized by information along the $F1$ dimension (see below). Figure 1 shows the positions in vowel space of the DK stimuli used in this experiment.

The Danish vowel inventory is larger than that of French. Danish contains 16 monophthongs (unevenly distributed across the vowel space with most vowels placed in the upper third of the vowel space), most of which can occur in both short and long forms (Grønnum, 1997), whereas French contains 10 evenly distributed oral (and 3 nasal) monophthongs (Georgeton *et al.*, 2012). The larger repertoire of DK vowels and the specificity of their distribution allowed us to select DK phonemes which differ phonetically from French. Second, because Danish is not a language that is frequently spoken or taught in Switzerland or in France, we were able to ensure that neither group had any previous experience with Danish.

Though the International Phonetic Alphabet (IPA) symbols representing certain Danish phonemes are also used to represent certain French vowels, their realizations differ across the two languages (Steinlen, 2005; Georgeton *et al.*, 2012). For instance, the DK front unrounded vowels represented by the IPA symbols /e/ and /ɛ/ are described as raised close-mid and close-mid vowels, respectively (Basbøl, 2005), whereas the French (FR) corresponding /e/ and /ɛ/ vowels are described as close-mid and open-mid vowels, respectively (Georgeton *et al.*, 2012). Danish /e/-/ɛ/ vowels are therefore much closer to one another, especially in Copenhagen Danish, which is spoken by our target native speakers. These articulatory differences are reflected in acoustic values. The relatively more closed DK /e/ vowel has a lower $F1$ compared to the FR /e/, and, therefore it is acoustically intermediate between the FR /i/ and /e/ vowel categories. The DK /ɛ/ vowel, in turn, is acoustically closer

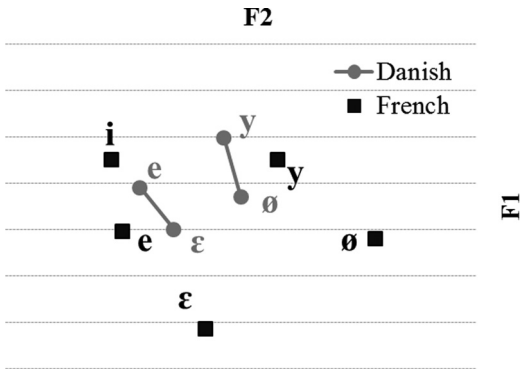

FIG. 2. Schematic representation of selected Danish and French vowels (based on a review of the literature, i.e., Steinlen, 2005; Georgeton *et al.*, 2012; Basbøl, 2005, and citations therein).

to the FR /e/ vowel than to the FR /ɛ/ vowel. The DK front rounded vowels represented by the IPA symbols /y/ and /ø/ are described as close and close-mid vowels, respectively (Basbøl, 2005). The same description is given to the FR /y/ and /ø/ vowels (Georgeton *et al.*, 2012). However, when we compare the acoustic values for these vowels, we note that the DK /y/ and /ø/ vowels have lower $F1$ and higher $F2$ values than the FR /y/ and /ø/, respectively, making the DK /ø/ vowel, for example, acoustically more similar to the FR /y/ than to the FR /ø/ vowel. Figure 2 schematically illustrates the reported characteristics of the DK vowels that we selected, compared to their French counterparts, represented by the corresponding IPA symbols. In order to make predictions for the performance of our participants on the discrimination task, ten native French speakers who did not participate in the main experiment took part in a categorization task. They heard the Danish vowels used for training (see Sec. II) and were asked to transcribe them, using French vowels. Six examples of each Danish vowel were presented. The results revealed that the DK /y/ and /ɛ/ vowels were consistently categorized as the FR /y/ and /e/ vowels, respectively, by all participants. The DK /ø/ vowel was perceived as being intermediate between the FR /y/ and /ø/ vowels by eight participants (the participants were permitted to choose two vowels if they felt it was similar to both or intermediate to them), one participant perceived it as a FR /y/ and one as a FR /ø/. The DK /e/ vowel was perceived as being intermediate between the FR /e/ and /i/ vowels by all participants. Based on the results of the above reported analyses, we hypothesized, in accordance with the SLM, that the DK /y/ and /ɛ/ vowels would be perceived more accurately before training, due to their assimilation to the perceptually similar FR /y/ and /e/ vowels. The DK /e/ and /ø/ vowels, on the other hand, were predicted to be perceived less accurately (and therefore confused more with similar DK vowels) due to their less consistent similarity to corresponding FR vowels. With regard to discrimination performance, in accordance with the PAM (Best, 1995), we predicted that the DK /e/-/ɛ/ and /y/-/ø/ vowel contrasts would assimilate to the French /e/ and /y/ categories, respectively. We also expected better performance on the /y/-/ø/ than on the /e/-/ɛ/ contrast, since the acoustic values for the DK /y/ vowel are markedly

different from those for the FR /y/ vowel. This difference should make it easier for participants to distinguish between the DK /y/ (similar to FR /y/) and DK /ø/ (similar to FR /y/ and /ø/) vowels.

C. Stimulus recording

To create the DK stimuli, one female and one male native speaker of Danish were recorded. The speakers were of similar age (32 and 36 yrs old, respectively), and were both from the Copenhagen region in order to minimize the effect of variability arising from dialect and age (Grønnum, 1997). They had lived in Switzerland for 2 yrs (the male speaker) and 5 months (the female speaker) at the time of the recordings. They had only very basic knowledge of and limited exposure to French. At their respective jobs, both at Danish organizations, they were constantly exposed to Danish and English (as they would be in Denmark), and they spoke Danish at home. They were not learning nor were they planning on learning French during their stay here, and both left Switzerland 2 months after the study was completed.

Recordings were carried out in a quiet room, using a Marantz PMD670 (MARANTZ EUROPE B.V., Eindhoven, Netherlands) portable recorder and a Shure Beta 58A microphone (Mexico, Mexique) [frequency response 50 Hz–16 kHz, sensitivity −38 dB V at 94 dB Sound Pressure Level (SPL)], sampled at 22.05 kHz directly to 16-bit mono.wav files.

Speakers read a list of four DK sentences, one per target vowel. Each comprised the opening context "Jeg siger" ("I say"), followed by a three-element sequence, consisting of a bisyllabic DK real word, a pseudo-word always of the form hVde [hVde] (Danish real words with this frame do not exist for all tested vowels) and another bisyllabic DK real word. The hVd [hVd] context is considered to allow for neutral production of vowels, with minimal impact of co-articulation from the phonemic context. The words and pseudo-words each contained the long target vowel in their first syllable, e.g., for the /e/ vowel, "Jeg siger mele, *hede, sene" ("I say mile, *hede, late"), where * denotes a pseudo-word. The list of four sentences was repeated five times. The vowels that were produced in pseudo-word contexts were extracted, analyzed, and trimmed to 350 ms from beginning and end using MATLAB (MATLAB Release 2011a, The MathWorks, Inc., Natick, Massachusetts). The beginning of a vowel started at the offset of the /h/ noise and the end was determined at the $F2$ movement toward /d/. Then the middle of the vowel was determined and the segment of 350 ms (175 ms from the middle in both directions) was trimmed. The vowels were trimmed for two reasons: first, to remove those portions that were affected by preceding and following consonants; and second, to make vowel duration constant across all trained vowels. A 20 ms linear ramp was applied to the onset and offset of the trimmed vowels. The first two formant frequencies ($F1$ and $F2$) of the produced token, averaged over the 350 ms time period, were computed by solving for the roots of the Linear Predictive Coding (LPC) polynomial. For this purpose, we adapted the scripts from the COLEA for speech analysis software (COLEA is a suite of tools that are a subset of the COchLEA Implants Toolbox; Loizou, 1998). An LPC order of 24 was used based on the rule of thumb that the LPC order should be equal to 2+ (the sampling frequency/1000). The formant tracks were then visualized using the Praat software package for acoustic analysis (Boersma and Weenink, 2010) and assessed for formant stability by eye. We retained the three exemplars of each target vowel extracted from the pseudo-words with the most stable formant tracks for each speaker. For validation purposes, we asked four native Danish speakers from Denmark (not the recorded talkers) to identify the extracted vowels. Their identification performance showed that the stimuli were recognized as members of the categories from which they had been drawn. These 24 vowels (12 for each speaker), hereafter called "DK target vowels," were used for training.

Vowels from the real words were also extracted and analyzed using the above-mentioned procedure, for a total of ten further exemplars of each of the four vowels per speaker. The $F1$ and $F2$ of these ten exemplars and of the three tokens extracted from pseudo-words were analyzed in MATLAB, and used to construct a representative acoustic space for each vowel and speaker, hereafter referred to as the "DK target space." These spaces enabled us to represent some of the natural variability in vowel realizations, and were used in assessing production performance (see Sec. II D 2). Note that these DK target spaces were constructed using a larger selection of tokens than was subsequently used for the training.

As mentioned above, participants were trained using tokens recorded by a speaker of the same sex as themselves. This was intended to minimize differences between the vocal-tract characteristics of the experimental participants and the native Danish speakers.

D. Procedure

Participants were trained to produce four isolated DK vowels over five 45-min sessions (30 additional minutes of the first and last sessions were used to perform pre-/post-training perception and production tests), which were administered on separate, and if possible, alternating days. We alternated the training days when possible due to known learning benefit related to sleep consolidation (Davis *et al.*, 2009). Before and after training, the perception and production of the DK vowels was evaluated. Fifteen participants were randomly assigned to the experimental group, and were trained on the selected DK vowels and given feedback on their accuracy (see Sec. II D 3). The remaining 12 participants were assigned to the control group. All tasks were performed on a DELL computer, using Sennheiser PC-350 (Sennheiser, Germany) headphones fitted with a microphone (frequency response 50 Hz–16 kHz, sensitivity −38 dB V at 94 dB SPL). The production and perception tests were administered using E-prime 2 (Psychology Software Tools, Pittsburgh, PA), and training (including recording the participants' productions, analyzing them, and displaying feedback on the screen) was administered using MATLAB and Psychophysics

Toolbox extensions. For all parts of the experiment (testing and training), stimuli were presented at a comfortable listening level, which was adjusted on a participant-by-participant basis.

1. Evaluation of perception

Participants performed an *ABX* cross-sex discrimination task on six (three male and three female) different exemplars of the four vowels (two vowel pairs, i.e., /e/-/ɛ/, and /y/-/ø/) before and after the training. Participants were asked to indicate whether the third stimulus was more similar to the first or to the second one. *A* and *B* were always each of the two vowels in a pair (see Fig. 1, where vowel contrasts are displayed). On each trial, *A* and *B* were produced by one speaker, and *X* was produced by a speaker of the opposite sex. This between-sex approach was used in order to ensure that the task could only be correctly carried out using phonological rather than acoustical information. The composition of the *ABX* triplets was counterbalanced along the dimensions of speaker gender and vowel sequence, and presentation order was randomized. There were a total of 96 trials, with 24 per vowel category.

Each trial began with the appearance of a cross at the center of the screen. The *A*, *B*, and *X* stimuli were then presented with an Interstimulus interval (ISI) of 1000 ms, accompanied by a number (1, 2, or 3) presented visually, to provide a label for each stimulus. Participants were prompted to respond after the offset of the *X* by a visual instruction, displayed for 4000 ms. Participants were instructed to answer as accurately as possible within these 4000 ms.

2. Evaluation of production

In order to assess production before and after training, participants were asked to repeat each of the 3 target tokens of /e/, /ɛ/, /y/, and /ø/ (produced by the talker of their own sex) that had been retained for training 5 times (a total of 15 trials per vowel). On each trial, a token was presented, and participants were prompted, by a visual cue, to repeat the vowel they had heard as accurately as possible. The cue remained on the screen for 2000 ms, and responses were recorded during this period. Participants' productions were recorded to a hard-disk as 16-bit.wav files sampled at 22.05 kHz and analyzed in MATLAb. The *F*1 and *F*2 values were estimated using scripts adapted from the COLEA for speech analysis software (Loizou, 1998). The *F*1 and *F*2 were computed by solving for the roots of the LPC polynomial.

Accuracy of the vowel productions was determined by calculating the Mahalanobis distance between participants' productions and vowel target spaces in *F*1/*F*2 space. Rather than computing the Euclidean distance between produced tokens and targets in *F*1/*F*2 space, this metric was used in order to take account of the natural variability in speech production, as characterized by the target spaces derived from the recordings of the native speakers. This distance measure involves computing the number of standard deviations from participant's token to the mean of the target space, and this along each principal component axis of the target spaces. The distance is "0" if a token is at the mean of the target space, and it increases as token moves away from this mean. For each participant, 15 distance scores (DSs) were calculated per vowel before and after training.

3. Training

On the first session, after completing the pre-training tests, all participants received basic instructions explaining the nature of the feedback, and its correspondence to the position of the articulators (i.e., mouth and tongue position) during production. They were also familiarized with this feedback using a task involving reading FR vowels (five vowels, three times each). The familiarization phase lasted 5 min and was immediately followed by training.

Each of the five 45-min training sessions was composed of 5 blocks, with pauses between them. The length of the inter-block pauses was controlled by the participants. The training blocks were composed of five mini-blocks. The order of the mini-blocks was randomized. Within each mini-block, each of the three same-sex tokens of a given vowel was presented 7 times, resulting in 21 presentations of each vowel per mini-block. The same amount of training was administered per vowel; participants produced each trained vowel 525 times.

Trials began with the appearance of a fixation cross in the center of the screen for 500 ms. At the offset of the cross, a vowel was presented over the headphones. Participants were immediately prompted by an on-screen message to repeat the vowel, and a 500 ms recording was initiated. Feedback was then presented on-screen for 2000 ms.

a. Visual feedback during training. The articulatory feedback provided was based on an immediate, trial-by-trial acoustic analysis of the vowels produced by participants compared to that of the respective DK target vowel previously recorded by native DK speakers, which participants were required to repeat. The aim of this feedback was to provide participants with visual information, on each trial, regarding the acoustics (specifically *F*1 and *F*2) of their produced token, together with that of the DK target token that they had just heard. Since the selected vowels essentially differ from one another based on degree of openness and frontness (see Sec. II B), we assumed that information about *F*1 and *F*2 would be sufficient to help participants improve their production of the DK vowels which corresponds to tongue height. The *F*1 reflects the degree of openness of vowels: a vowel that is produced with a more open mouth tongue height (e.g., /a/) has a higher *F*1 (e.g., Ménard *et al*., 2002). The *F*2 reflects the back-to-front position of vowel articulation; i.e., it reflects whether the sound is produced more at the back (e.g., /u/) or at the front (e.g., /i/) of the mouth. Relatively more back vowels have relatively lower *F*2 values.

The produced vowels were recorded to a hard-disk as 16-bit.wav files sampled at 22.05 kHz. *F*1 and *F*2 values were estimated in MATLAB using scripts adapted from the COLEA for speech analysis software (Loizou, 1998). On each trial, the formant values were averaged over the duration of the produced token. These averaged formant values of the produced tokens were displayed, along with the time-

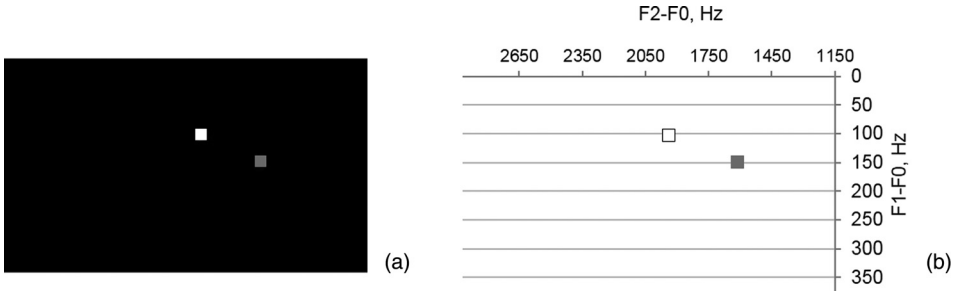

FIG. 3. (a) Example of visual feedback provided to participants on their vowel production during training. The white square corresponds to the participant's production and the gray square corresponds to the DK target vowel (the control group only saw the gray squares); (b) A schematic representation depicting the units and ranges used for calculating this feedback. Note that the axes are reversed, such that movement upwards along the ordinate indicates decreasing $F1$ and corresponds to a more closed mouth tongue height, and such that movement from left to right along the abscissa indicates decreasing $F2$ and corresponds to a more retracted tongue position. The rectangle represents the space of all four vowels.

averaged formant values of the target vowel that the participants had just heard. Vowels for which the recordings were shorter than 150 ms were discarded because on these trials the utterance was probably produced late, and an error message was displayed. The $F1$ and $F2$ were computed by solving for the roots of the LPC polynomial. For each trial and for each subject, the fundamental frequency ($F0$) was analyzed using a cepstral method. $F0$ was used to adjust the feedback on the $F1$ and $F2$ dimensions, since it is known that $F0$ plays a role in the perceptual normalization of vowels and in the disambiguation of vowels with similar $F1$ and $F2$ values (Ménard *et al.*, 2002). Moreover, the distance between $F0$ and $F1$ has been shown to be a good predictor of perceived vowel height (close–open dimension) (Ménard *et al.*, 2002). The feedback presented to the participants was adjusted by subtracting $F0$ from $F1$ and $F2$ (see Fig. 3). For the displayed feedback, the y axis showed $F1$–$F0$ (in Hz), and the x axis showed $F2$–$F0$ (also in Hz). The x and y axes ranged from 1150 to 2920 and 0 to 375 Hz, respectively. The axes were projected onto a screen area of 590 by 375 pixels, resulting in a mapping of 1 pixel to 3 Hz on the x axis and 1 pixel to 1 Hz on the y axis. Figure 3(a) shows an example of the feedback display provided to participants in the experimental group.

On each trial, the control group, like the experimental group, saw the $F1$ and $F2$ values of the target DK tokens that they had repeated, but they received no feedback on the acoustic properties of their own performance. Instead, at the end of each block, they were presented with an aggregate score (expressed as a percentage) indicating their average distance to the target space in terms only of $F1$. If the average difference between the participant's production and the target values was within 20 Hz, the participant was provided with an estimated accuracy of "90% correct"; this score was chosen arbitrarily. For each additional 10 Hz of excursion from the target vowel, the displayed estimation of correct production was reduced by 20%. In sum, control participants listened to and repeated as many training trials as the experimental participants, and received the same visual information about the position of the target vowels in $F1$–$F2$ space on screen. However, they received no visual feedback on their own productions, but in order to help maintain their motivation, they received some feedback about changes in the quality of their production on $F1$.

III. RESULTS

A. Production performance

Recordings from the pre- and post-training tests were verified for intensity and absence of noise (e.g., coughs, sneezes, sighs). Ten vowel productions were removed on this basis. All recordings from participants 18, 23, 25 (control group), and 21 (experimental group) were discarded due to technical issues (they spoke very quietly during the pre-training test, and it therefore was impossible to reliably analyze their productions). For each recorded token, the silent portions preceding and following the sound were removed using Praat, and the Mahalanobis distance (based on $F1$ and $F2$) between the token and the target space was calculated. Outliers and extreme values were detected using Quantile-Quantile plots, and were removed. They represented 1.19% and 0.77% of the data for the experimental and control group, respectively. The remaining DSs ranged from 0.045 to 14.86, and had a standard deviation of 2.46 and a mean of 3.75. Statistical analyses were run using the R software package.

Analysis was executed using general linear mixed-effects models. These were chosen over Analysis of (co)-variances because they are able to: (i) account for within and between speaker variability in non-native vowel production; (ii) account for variability in speakers' sensitivity to the training effects; (iii) include data with occasional missing points (e.g., coughing, sneezing); (iv) simultaneously model crossed random-speaker and random-vowel effects (i.e., hierarchical modeling); and finally, by virtue of the above four points (v) better generalize the findings.

A two-step analysis procedure was used. First, the DSs were fitted to a general linear mixed-effects model using the R software package. Here, the effects of Group (control vs experimental) and Session (pre vs post) (with "control" and "pre-training" as reference levels for Group and Session, respectively), and their interactions, were included as fixed factors. The "maximal" random structure with correlation parameters between the critical factors and random slopes was used: it included by-subject and by-vowel random slopes adjusted for Session, and Session and Group, respectively (Barr *et al.*, 2013). This structure allows for best generalization of the findings (Barr *et al.*, 2013). Second, the significance of the main effects and of the interactions was

computed using Markov Chain Monte Carlo (MCMC) sampling (10 000 simulations), implemented in the "LanguageR" package of the R software. The coefficient estimate (β), standard error (SE), t-value (t), and p-value (based on MCMC simulations) are used to report the predictor parameters.

There was a significant two-way Group-by-Session interaction ($\beta = -0.598$, SE $= 0.27$, $t = -2.182$, $p < 0.001$), and no significant effect of Group or Session at the reference level of session and group, respectively (see Fig. 4), indicating that there was no difference between groups before training and that there was no effect of training for the control group. Separate by-group linear mixed-effects analyses with Session (with pre-training as a reference level) as a fixed factor and by-subject and by-vowel random slopes as the random-effects structure revealed a significant effect of training for the experimental group ($\beta = -0.4654$, SE $= 0.18$, $t = -2.541$, $p < 0.001$), but not for the control group ($p > 0.1$). Importantly, since by-participant random slopes were adjusted for Session (to account for the heteroscedasticity by participants on session effects), the model included interactions between the fixed effect of session and the random effects of participant and vowel (these factors showed a non-negligible correlation of rho $= -0.37$). The relationship between these factors suggested that participants whose DSs were larger at the pre-training test session (i.e., those whose vowel production was worse) benefited more from the training than those whose production was already closer to the DK target vowel before training. There was measurable heterogeneity across participants (see Fig. 6 below). The group performance on each trained vowel before and after the training is presented in Table I.

For each trained vowel and for each subject of the experimental group, the average DSs for the pre- and post-training test sessions were used to calculate improvement in production performance relative to pre-training levels. The mean improvement was 17%, with improvements of 18%, 20%, 13%, and 18% for the /e/, /ɛ/, /y/, and /ø/ vowels, respectively. Across subjects, no significant correlations in improvement were found among vowels.

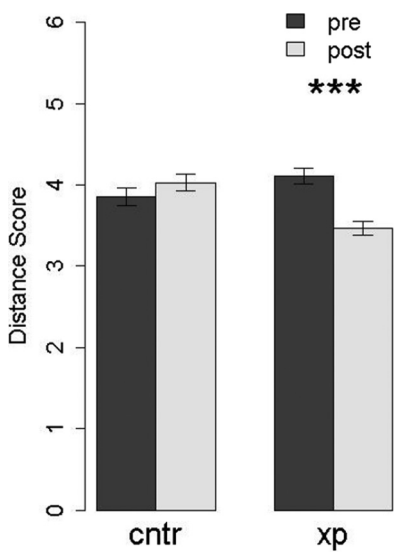

FIG. 4. Effect of training on production accuracy for the experimental (xp) and control (cntr) groups for trained and untrained vowels. Mean DSs are shown, and error bars represent ±1 SE of the mean. Three asterisks correspond to $p < 0.001$.

B. Perception performance

In order to test whether production training resulted in improvements in the perception of the trained vowels, we compared pre- and post-training performance on the *ABX* vowel discrimination task (i.e., perception of vowels within height-contrastive vowel pairs) for the control and experimental groups. Missed trials (0.22% of the data) were removed before performing these analyses.

Accuracy (a binary measure of 1 for correct and 0 for incorrect responses) was fitted to a mixed-effects logistic model that is traditionally used to analyze binomially distributed data. The Session, Group, and Session-by-Group interaction with pre-training and control as reference levels for Session and Group, respectively, were included as fixed factors and the maximal random-effects structure was again used (Barr *et al.*, 2013); it included random by-subject and by-vowel slopes, adjusted for correlational parameters between Session, and Session and Group, respectively. The analysis revealed a significant effect of Group at the reference level of session ($\beta = 0.66$, SE $= 0.26$, $z = 2.50$, $p = 0.012$), with the experimental group having more accurate performance before training, a marginally-significant Session-by-Group interaction ($\beta = -0.25$, SE $= 0.14$, $z = -1.74$, $p = 0.08$) and no effect of Session at the reference level of group ($p > 0.1$). As can be seen from Fig. 5 and as confirmed by the above analyses, the experimental group performed better on the perception task before training. It was surprising to find these differences given that participants were randomly assigned. In order to test whether the pre-training perception performance was related to the amount of improvement in production, additional correlational analyses comparing by-vowel pre-training perception accuracy and improvements in production were run. The results suggest that perception performance before the training was not related to the improvement in production across individuals ($p > 0.1$).

Due to *a priori* predictions of at least some transfer of production training to perception, we performed planned tests to further explore the effects of production training on perception. Separate logistic mixed-effects analyses with Session (with pre-training as a reference level) as a fixed factor and random by-subject and by-vowel slopes were run on the two groups separately. There was a significant effect of training for the experimental group ($\beta = -0.23$, SE $= 0.10$, $z = -2.199$, $p = 0.02$), and not for the control group ($p > 0.05$). As shown in Fig. 5, the perception of the trained vowels improved in the experimental group, but remained stable in the control group. The group performance on each trained vowel before and after the training is presented in Table II. In order to test whether pre-training perception accuracy was related to production accuracy, we ran an additional correlational analysis between pre-training perception

TABLE I. Pre- and post-training production accuracy for each of the four trained vowels for the experimental group. Mean DSs and SEs of the mean (in brackets) are presented.

Trained Danish vowel	Pre-training performance, DS	Post-training performance, DS
/e/	3.09 (0.33)	2.53 (0.28)
/ɛ/	5.06 (0.51)	4.07 (0.41)
/y/	4.94 (0.44)	4.32 (0.56)
/ø/	3.43 (0.86)	2.82 (0.59)

and production performance. There was no correlation between them ($p > 0.1$).

Improvement in perception performance in the experimental group was calculated as a difference of post-training relative to pre-training performance. This revealed improvements of 4.56% on average. There was substantial heterogeneity across participants, whose change from pre- to post-test ranged from −7% to +17%.

C. Relationship between changes in production and perception of trained vowels

In order to determine whether there was a correlation between changes in production and in perception performance in the experimental group, we performed a Spearman rank-correlation analysis on the changes in each participant's pre- versus post-training test performance on each task, averaged over all trained vowels. There was no significant correlation between changes in perception and production ($S = 358.8$, $p = 0.23$, rho = 0.21, one-tailed test), see Fig. 6(a). One participant (subject 12) exhibited remarkably greater improvement in production performance than the others (an improvement of 3.7 DS units, more than 2 standard deviations of mean DS), whereas his perception performance remained stable (within 2 standard deviations). For exploratory purposes, we therefore performed another Spearman rank-correlation analysis on the changes in production and perception, this time excluding this participant. This analysis revealed a correlation of rho = 0.39 ($p = 0.08$) [see Fig. 6(b)]. Separate Spearman rank-correlation tests were also run on the changes in production and perception for each trained vowel separately (subject 12 was included in the analyses). None of the correlations was significant ($p > 0.1$).

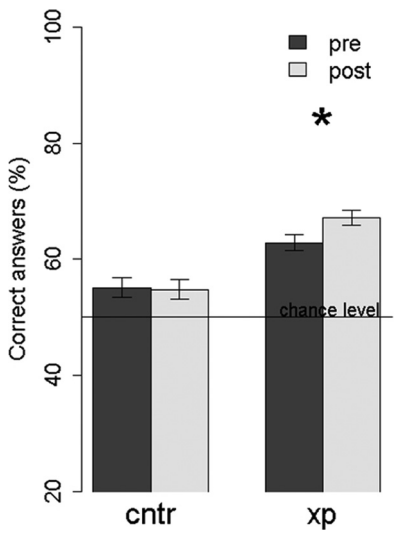

FIG. 5. Effect of training on discrimination accuracy for the experimental (xp) and control (cntr) groups for trained vowels. Mean percent of correct answers and their SEs of the mean are shown. One asterisk corresponds to $p < 0.05$.

D. Training-related changes in the stability of vowel production

In order to further explore the effect of production training, we examined training-related changes in the acoustic stability (or conversely, variability) of vowel productions for the trained vowels; we refer to this measure as "compactness." Compactness can be thought of as representing the consistency of the participants' phonological-motor mapping. It has been previously shown that the compactness of *L*2 productions is highly correlated with their accuracy: speakers whose productions are more compact are those who are more accurate (Kartushina and Frauenfelder, 2014). A compactness score (CS), based on an analysis of the *F*1–*F*0 and *F*2–*F*0 of the produced vowels, was estimated as follows: the distribution of the productions in *F*1–*F*0/*F*2–*F*0 space was assumed to be elliptical, and the angles of the major and minor axes of an ellipse centered on the mean of the productions were estimated. The CS was then calculated as the area of an ellipse having principal axes with a length of one standard deviation of the mean along the given axis. In order to make the CSs more meaningful, they were scaled as a proportion of the area of the native target space.

Two-tailed paired *t*-tests were run on the scaled CSs before and after the training in the experimental and control groups separately. The results revealed a significant effect of training on the CS in the experimental group [$t_{(55)} = 2.28$, $p = 0.026$], with more compact productions after training, but not in the control group ($p > 0.1$), see Fig. 7.

In order to explore the relationship between the compactness of vowel productions and their accuracy, Spearman's rank correlations were calculated between these two measures for the productions before and after the training, in the experimental group only. The results revealed significant positive correlations between vowel compactness and vowel accuracy before (rho = 0.58, $p < 0.001$) as well as after (rho = 0.51, $p < 0.001$) the training.

IV. DISCUSSION

A. Pre-training perception and production performance

In accordance with the PAM (Best, 1995), poor discrimination performance on the /e/-/ɛ/ pair (55.9% accuracy) in the *ABX* task suggests that these vowels perceptually assimilated (Best, 1995) to the French /e/ category (see Table II). Weaker performance on the Danish /e/ compared to the /ɛ/ vowel is in line with the results of the categorization task showing that the Danish /ɛ/ vowel is perceived more consistently than the Danish /e/ vowel: French speakers perceived the Danish /ɛ/ as being perceptually close to the French /e/

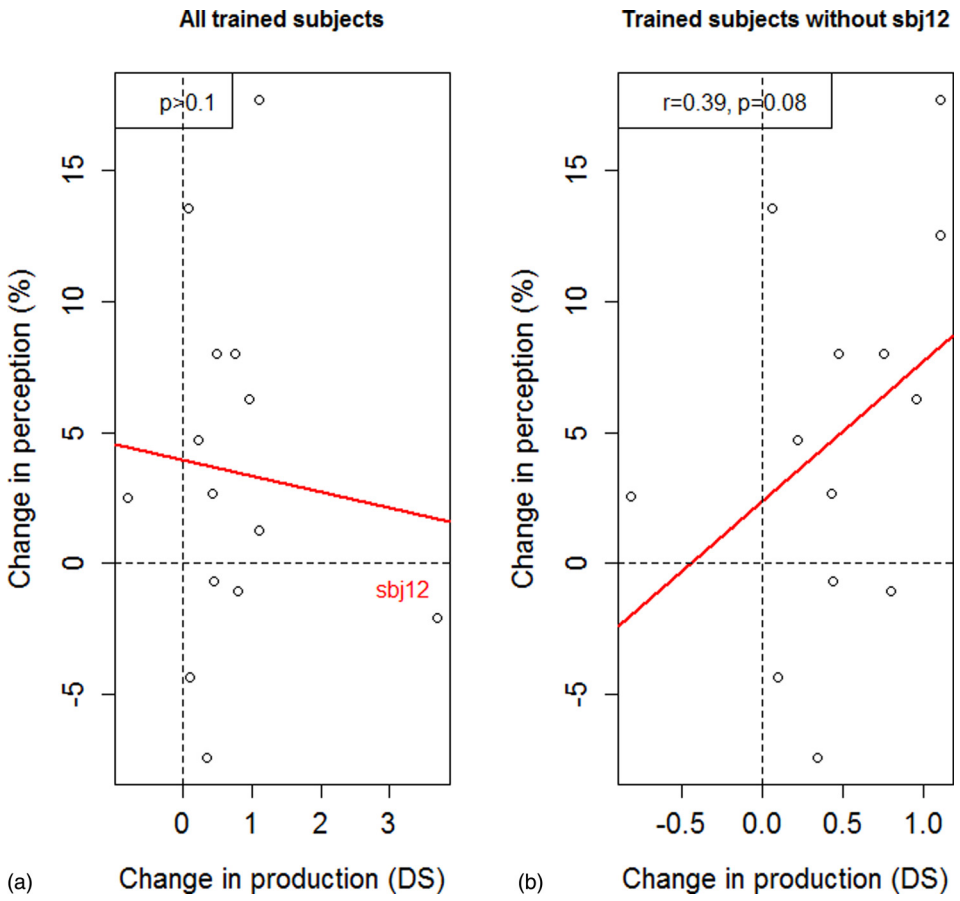

FIG. 6. (Color online) Scatterplot of the changes in production and perception and their correlation slope for all four trained vowels in participants of the experimental group (a); scatterplot of the changes in production and perception and their correlation slope when the data point from one atypical subject was excluded (b).

vowel, whereas the Danish /e/ vowel was perceived as being intermediate between the French /e/ and /i/ sounds. Intermediate performance on the DK /y/-/ø/ pair (69.5% accuracy) suggests that these vowels perceptually assimilated to the FR /y/ category. Better performance on the DK /y/ compared to the /ø/ vowel is likely due to its relatively further position in acoustic space from its French phonological counterpart /y/ as compared to the /ø/ vowel. This result suggests that the Danish /y/ vowel was perceived as being a poorer exemplar of the French /y/. Better performance on the /y/-/ø/ compared to the /e/-/ɛ/ contrast suggests that the former vowels assimilated in a CG manner to the French /y/ vowel, whereas that the latter vowels assimilated in a SC manner to the French /e/ vowel.

Pre-training production results showed that French speakers experienced greater difficulty with the Danish /ɛ/ and /y/ vowels than with the Danish /e/ and /ø/ vowels. Curiously, the latter two vowels are those which were perceived the worst before training. These results suggest that before training, there was no relationship between the perception and production of the Danish vowels in French speakers with no experience with Danish, i.e., those vowels that were poorly perceived were not those that were poorly produced.

TABLE II. Pre- and post-training perception accuracy for each of the four trained vowels for the experimental group. Mean % of correct discrimination and SEs of the mean (in brackets) are presented.

Trained Danish vowel	Pre-training performance	Post-training performance
/e/	43 (5.3)	51 (7.2)
/ɛ/	69 (2.8)	69 (3.6)
/y/	86 (2.8)	88 (4.5)
/ø/	53 (2.8)	61 (3.9)

B. Effect of production training on production

We have developed and extended an articulatory feedback training method, similar to that used by Carey (2004), and tested its effect on the production and perception of

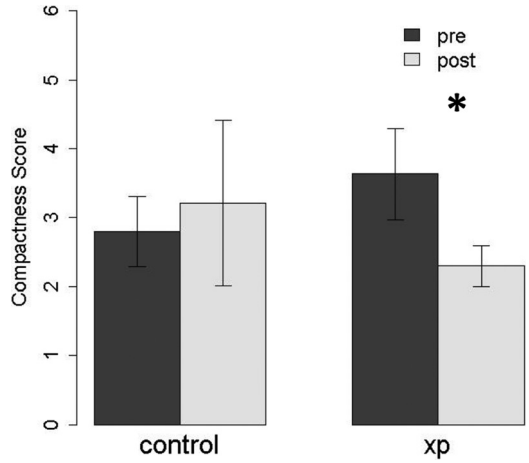

FIG. 7. Effect of training on compactness for the experimental (xp) and control groups for the trained vowels. Mean CSs are shown, and error bars represent ±1 SE of the mean. One asterisk corresponds to $p < 0.05$.

foreign speech sounds. The method involves providing participants with immediate visual feedback that represents the *F*1 and *F*2 values of their production displayed together with the same formant information about the Danish target vowels produced by native speakers. However, unlike Carey's (2004) study in which participants of the control group were not exposed to foreign sounds between pre- and post-training tests, our investigation compares the effect of the training method to an appropriate control group and, therefore, allows us to distinguish training effects from mere exposure to and repetition of the non-native vowels. Moreover, our method has the dual advantages of adjusting for *F*0 and using a more simple graphical representation of the feedback than Carey's study, in which all non-native vowels (and not only the trained one) were displayed.

The study shows that this training method improves the production of four non-native Danish vowels in the experimental group, whereas no changes were observed in the control group. Further, no differences in the production accuracy of these two groups were observed before training. Thus, our results indicate that the improvements observed in the experimental group are specifically due to the trial-by-trial feedback received during training. The results of our study converge with results of second-language learning studies, which have shown that descriptive articulatory instruction alone does not suffice to improve pronunciation of *L*2 phonemes, and that trial-by-trial corrective feedback is crucial (Saito and Lyster, 2012).

The results revealed that production training significantly decreased the variability of vowel productions in the experimental group only. This suggests that our articulatory feedback training method not only improves the production of the trained vowels but that it also makes them more stable. In other words, the vowel categories become more compact acoustically, suggesting a training-related enhancement in the stability of phonological-motor mapping in the trained group. It is interesting to note that more generally, before and after the training, there is a relatively strong relationship between vowel production accuracy and vowel production stability. This suggests that speakers whose productions were realized more accurately were also those whose productions were more stable. The relationship between non-native production accuracy and stability merits further study.

The results of our experiment showed an average improvement of 17% in production performance. Improvements by vowel were /e/: 18%, /ɛ/: 20%, /y/: 13%, and /ø/: 18%. The explanation for these qualitative differences in improvement across vowels may partly lie in the difficulty of these vowels before training: vowels that were produced less accurately before training (such as, for example, the /ɛ/ vowel), appeared to benefit more from the training than did vowels that were produced more accurately at pre-training (such as, for example, the /e/ and /ø/ vowels, see Table I). In other words, this could be the result of a ceiling effect. This explanation, however, does not hold for the /y/ vowel: its production accuracy before training was about as poor as that of /ɛ/, and yet, production improved by 7 percentage points fewer. Nevertheless, /ɛ/, which benefited the most from training (improvement of 20%) was the one that was pronounced least accurately prior to training (see Table I). A statistical analysis that included by-participant random slopes adjusted for session confirms this same pattern across individuals: participants whose DSs were larger prior to training (i.e., those whose vowel production was worse) benefited more from training than those whose DSs were already closer to the DK target vowel. Previous studies have similarly shown that better pre-training production is associated with less improvement (Bradlow *et al.*, 1997). This phenomenon might reflect an asymptotic learning pattern, whereby relatively more training would be required to further improve performance after it reaches a certain, higher level of performance.

The post-training production results suggest that the production of the high Danish vowel (/y/) is relatively more difficult compared to the other trained Danish vowels for French speakers, despite the fact that perception of this vowel was good (86% accuracy in the pre- training perception test, see Table II). Although training improves the production of the /y/ vowel, the DSs achieved by French speakers after training were very similar to those obtained for the other non-native vowels even before training. It is likely that due to assimilation mechanisms of the Danish /y/ to the French /y/ vowel, as predicted by the SLM (Flege, 1995) and by the categorization data, French speakers use the French phonological counterpart /y/ that is phonetically different from the Danish /y/. This mechanism prevents the formation of a new category for the Danish /y/. Note that the Danish /y/ is relatively higher and more front compared to French (it falls outside the French vowel space), as shown by acoustic values (see Fig. 2). It remains to be tested whether more training would encourage the formation of a distinct/new category for this difficult Danish /y/ vowel. The fact that this vowel was not well produced despite relatively accurate perception (see Table II) suggests that there is not a clear and consistent relation between *L*2 vowel perception and production, and that accurate *L*2 speech perception is not sufficient for accurate *L*2 speech production. Our results suggest that factors other than perceptual difficulty likely also contribute to *L*2 production, such as the articulatory difficulty of *L*2 vowels *per se* or speakers' sensorimotor control abilities (Simmonds *et al.*, 2011).

Finally, it is interesting to note that no relationship was observed between improvements in production accuracy across the different vowels. These results reveal considerable heterogeneity across participants in terms of their pattern of improvement for the different vowels—participants who improved on one vowel were not necessarily those who improved on the others. The compactness, or stability, and position of individual's production of native vowels in the acoustic space have been shown to predict *L*2 production accuracy (Kartushina and Frauenfelder, 2014). Speakers whose native individual vowels are close to similar non-native ones, and/or whose native vowels are more compact, produce non-native vowels better. In the current study, it is possible that French speakers whose native vowels were closer to the Danish ones, and were more compact, learned faster than those whose native vowels were more distant and more variably distributed. Within speakers, learning can be either

facilitated or impeded by native vowel categories depending on their position and compactness in acoustic space.

C. Effect of production training on perception

We tested for improvements in perception that may have resulted from the production training, and found an overall improvement of 4.56% in vowel perception, in the experimental group only. Similar results have been reported by Catford and Pisoni (1970): after training, speakers who had been trained with auditory instruction (i.e., listening to the stimuli and repeating them, as in our impoverished control group) performed less well in a perception task than speakers who underwent articulatory-instruction training.

The finding of a transfer of production training to improved perception suggests that the learning of new articulatory patterns led to a "tuning" of the corresponding perceptual representations. This conclusion is consistent with the Direct-Realism theory, which states that the basic perceptual unit is an articulatory gesture (Best, 1995). Similarly, in the Motor Theory of speech perception, it has been proposed that phonetic information is perceived in a cognitive module that is dedicated to the detection of the intended articulatory gestures of the speaker (Liberman and Mattingly, 1985). These two theories are compatible with our results, since they would predict that changes in the articulation of foreign speech sounds will have implications for perception. However, due to the lack of a robust relationship between improvements in the production and perception of Danish vowels, our results provide only modest support to the above-mentioned theories. A recent study by Lametti *et al*. (2014) shows that speech motor learning involving adaptation to altered auditory feedback changes the perceptual categorization of speech sounds. Interestingly, their experimental design allowed them to distinguish the effects of the sensory inputs during learning from those associated with the motor commands. The authors conclude that the perceptual changes that accompany speech motor learning are due to the motor processes of speech production.

Reduced training-related benefits in perception compared to production (4.56% versus 17%) are consistent with other *L*2 training studies showing that training is relatively specific to the modality being trained: it significantly improves the trained modality, and the untrained one benefits little, if at all, from it (Akahane-Yamada *et al*., 1998; Bradlow *et al*., 1997; Lopez-Soto and Kewley-Port, 2009).

Reduced training-related benefits in perception reported in our study can partly be attributed to the task used to assess the perception accuracy. Cross-sex discrimination tests participants' ability to discriminate non-native vowels in height-contrastive pairs across different voices, i.e., male and female. In order to successfully perform this task, participants have to have established abstract, speaker-independent, representations for non-native vowels. However, during the training, they were exposed to tokens produced by one single speaker of their own sex. As shown in some perception studies (e.g., Lively *et al*., 1993), such a small amount of variability in input does not suffice to generalize to unfamiliar voices, which was crucial in our discrimination task.

D. Relationship between changes in perception and production

To evaluate the relationship between the perception and production of the trained vowels, correlational analyses comparing global changes in perception and production were run across all vowels and participants. They revealed no reliable relationship between changes in perception and production, suggesting that the individuals whose production improved were not necessarily the ones whose perception improved. Nevertheless, when one outlying participant was excluded from the analysis, we found a trend for a relationship between improvements in production and perception. The lack of a robust relationship between improvements in production and perception is consistent with recent *L*2 studies showing no correlation between *L*2 phonological production and perception (Peperkamp and Bouchon, 2011), and with one *L*2 perception training study that also did not find a relationship between measures of improvement in these two modalities (Bradlow *et al*., 1997).

Other studies that trained naive listeners to produce exotic and tone contrasts have showed that after training on either perception or production, there was a correlation between the trained and the untrained modalities (Catford and Pisoni, 1970; Leather, 1996). These studies, however, do not report participants' performance before training; it is therefore unclear whether this correlation was due to transfer from the improved trained modality to the untrained one, or whether these were due to pre-existing (i.e., before training) relationships between the two modalities.

In sum, our correlational results and those of Bradlow and colleagues (1997) suggest that improvements in production and perception do not systematically progress at equal rates within individuals. However, the potential link revealed by the trend merits further examination in future studies.

E. Limitations of this study and directions for future research

The articulatory feedback training method that we developed for this study was effective at improving native monolingual French speakers' production of four different non-native vowels. However, given that we trained the production of isolated speech segments (cf. Dowd *et al*., 1998), we do not know whether this phonetic learning generalizes to the production of these vowels in the context of syllables and/or words. In *L*2 perception research, some studies suggest that the acquisition of robust *L*2 phonetic perceptual categories is a pre-requisite for the acquisition of lexical representations containing these phonemes (Pallier *et al*., 2001). The articulatory patterns learned during the training might also potentially allow learners to create more precise lexical representations for *L*2 production. There are other studies, however, that challenge this "phonetic first" hypothesis, by showing that good phonetic discrimination is not required for the creation of contrastive lexical representations (Darcy *et al*., 2012). Some studies of *L*2 production

have shown that late *L*2 speakers tend to share the same syllabic representations (CV and CVC structures) for similar *L*1 and *L*2 sounds (Alario *et al.*, 2010). It remains to be tested whether newly learned *L*2 vowels preserve their improved pronunciation in the context of the formation of new *L*2 syllables, or whether they would instead assimilate in production to similar *L*1 syllables.

In evaluating the production accuracy of our French participants, we opted for an objective measure of production accuracy, the Mahalanobis distance, rather than subjective ratings since the former minimizes rater bias and is fully reproducible. Despite these advantages of objective over subjective measures (Delvaux *et al.*, 2013), the former are limited in that they provide no information on native speakers' perception of the quality of the vowels. In particular, observed improvements in production accuracy may not be reflected in native listeners' judgements of non-native productions of accentedness. Ideally, both objective and subjective measures should be combined since they provide complementary approaches to the assessment of *L*2 production accuracy. Only few *L*2 production-training studies have used both approaches (Dowd *et al.*, 1998; Akahane-Yamada *et al.*, 1998). For example, some automated methods such as Hidden Markov Models (HMM) used by Akahane-Yamada *et al.* (1998) show a high correlation with human evaluations of *L*2 sound production accuracy.

Our results reveal considerable variability across participants and across vowels. First, production performance improved overall for the majority of the participants, except one, who showed deteriorated production performance (see Fig. 6). Second, participants showed vowel-specific effects: some achieved native-like production (i.e., their DSs were less than 1, indicating that their productions were situated within the one standard deviation from the DK mean) only on one vowel but not on the others, and some did not achieve native-like performance on any of the trained vowels. These results are consistent with other *L*2 training studies showing large individual differences in the amount of phonetic learning across *L*2 speakers (Bradlow *et al.*, 1997; Dowd *et al.*, 1998). Further studies are therefore needed to address the roles of other factors that may influence *L*2 production skills, including individual differences in pronunciation and imitation talent, sensorimotor control, and stability of *L*1 productions.

Although participants exhibited significantly improved production of the trained vowels, mean performance (averaged over the group) did not reach the performance levels of native Danish speakers (their mean post-training articulatory DS was approximately 3). There are several possible reasons for this. First, the amount of training may have been insufficient, being limited to 1 h per vowel. Second, we used low-variability stimuli (three tokens per vowel produced by only one sex-matched speaker). It is known from *L*2 perception training studies that high variability stimuli (here, with various phonetic contexts produced by multiple speakers) boost phonological (as opposed to acoustic) learning of speech sounds compared to low variability stimuli (Bradlow *et al.*, 1997; Lively *et al.*, 1993; Wong, 2013), and lead to greater transfer effects of improved perception to production (Wong, 2013). Other studies have nevertheless obtained training effects similar to ours (i.e., in terms of percentage of improvement) when using anywhere from three (Akahane-Yamada *et al.*, 1998) to seven (Dowd *et al.*, 1998) native speakers to record the target sounds. A recent study by Perrachione and his colleagues (2011) suggests that only speakers with high perceptive abilities benefit from perception training with high-variability stimuli. Finally, our participants may have not achieved native performance levels due to the nature of the feedback given about the target vowel. On each trial, participants received feedback in *F*1/*F*2 space on their production, alongside such information about individual DK target vowels. By providing participants with visual information reflecting the distribution of the DK target vowel category rather than of individual tokens, we might have observed the creation of more abstract, token-independent phonetic categories for the non-native sounds. Assessment of French speakers' productions using both objective and subjective measures (i.e., by additionally having native Danish speakers judge the productions for their accentedness or prototypicality) will be particularly relevant in this context. These directions will be pursued in our future studies.

V. CONCLUSIONS

Our experiment has shown that as little as 1 h of training with visual articulatory feedback containing acoustic information about tongue position and mouth openness is effective in improving the production accuracy of non-native vowels. This improvement was observed on all four trained vowels in terms of decreases in the Mahalanobis distance between the participants' non-native productions and target vowel spaces constructed from the productions of native speakers. The absence of training effects in the control feedback group demonstrates that the improvement in production accuracy of the experimental group was specifically due to the feedback and not to exposure to or repetition of the sounds. However, participants were not equally sensitive to the training, indicating that individual differences also influence performance. Finally, we observed some transfer of production training to perception; however, correlational analyses revealed the absence of a robust correlation between improvements in production and perception, suggesting no reliable relationship between learning in these two modalities across individuals.

Our findings on the transfer of improved production to perception suggest that learning to improve the articulation of foreign vowels results in a tuning of perceptual representations for those very vowels. This is compatible with several theories of speech perception, which claim that the basic perceptual unit is an articulatory (produced or intended) gesture (Best, 1995; Liberman and Mattingly, 1985). Other proposals (e.g., DIVA model in Guenther, 1994) claim that articulatory movements are planned in terms of the acoustic goals (i.e., that production is guided by perception). Training methods like the one we have presented here can help learners to overcome limitations that perception might impose on production, since by providing people with trial-by-trial visual representations of the acoustic features that they cannot easily perceive, we provide them with information that can serve as scaffolding for improved foreign speech sound

production. Our results with one of the more difficult Danish vowels, /e/, lend support to this; production of this vowel improved with training, even though pre-training perception of this vowel was very poor (43% accuracy in the pre-training discrimination test). More work is needed to further validate the efficacy of our new training method, both in the context of non-native speech sounds for which poor perception may impose a bottleneck for correct production, and in the context of non-native speech sounds that are easily distinguished perceptually from native ones but that require completely novel articulatory patterns.

ACKNOWLEDGMENTS

We are very grateful to Dr. Julien Mayor and Dr. Audrey Bürki for their assistance in statistical analyses. N.G. and A.H.-A. are supported by the Swiss National Science Foundation (PP00P3_133701).

Akahane-Yamada, R., McDermott, E., Adachi, T., Kawahara, H., and Pruitt, J.-S. (**1998**). "Computer-based second language production training by using spectrographic representation and HMM-based speech recognition scores," Proc. Interspeech **5**, 1–4.

Alario, F. X., Goslin, J., Michel, V., and Laganaro, M. (**2010**). "The functional origin of the foreign accent: Evidence from the syllable-frequency effect in bilingual speakers," Psychol. Sci. **21**(1), 15–20.

Aliaga-García, C., and Mora, J. C. (**2009**). "Assessing the effects of phonetic training on L2 sound perception and production," in *Recent Research in Second Language Phonetics/Phonology: Perception and Production*, edited by M. A. Watkins, A. S. Rauber, and B. O. Baptista (Cambridge Scholars Publishing, Newcastle upon Tyne, UK), pp. 2–31.

Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. (**2013**). "Random effects structure for confirmatory hypothesis testing: Keep it maximal," J. Memory Lang. **68**(3), 255–278.

Basbøl, H. (**2005**). *The Phonology of Danish* (Oxford University Press, New York), pp. 86–92.

Best, C. T. (**1995**). "A direct cross-realist view of cross-language speech perception," in *Speech Perception and Linguistic Experience: Theoretical and Methodological Issues*, edited by W. Strange (York Press, Baltimore, MD), pp. 171–204).

Boersma, P., and Weenink, D. (**2010**). "Praat: Doing phonetics by computer," [Computer program]. Version 5.2, http://www.praat.org/ (Last viewed October 19, 2010).

Bradlow, A. R., Pisoni, D. B., Akahane-Yamada, R., and Tohkura, Y. (**1997**). "Training Japanese listeners to identify English /r/ and /l/: IV. Some effects of perceptual learning on speech production," J. Acoust. Soc. Am. **101**(4), 2299–2310.

Carey, M. (**2004**). "CALL visual feedback for pronunciation of vowels: Kay Sona-Match," CALICO J. **21**(3), 571–601.

Catford, J. C., and Pisoni, D. B. (**1970**). "Auditory vs. articulatory training in exotic sounds," Modern Lang. J. **54**(7), 477–481.

Darcy, I., Dekydtspotter, L., Sprouse, R. A., Glover, J., Kaden, C., McGuire, M., and Scott, J. H. (**2012**). "Direct mapping of acoustics to phonology: On the lexical encoding of front rounded vowels in L1 English-L2 French acquisition," Second Lang. Res. **28**(1), 5–40.

Davis, M. H., Di Betta, A. M., Macdonald, M. J., and Gaskell, M. G. (**2009**). "Learning and consolidation of novel spoken words," J. Cogn. Neurosci. **21**(4), 803–820.

Delvaux, V., Huet, K., Piccaluga, M., and Harmegnies, B. (**2013**). "Production training in Second Language Acquisition: A comparison between objective measures and subjective judgments," Proc. Interspeech **14**, 2375–2379.

Dowd, A., Smith, J., and Wolfe, J. (**1998**). "Learning to pronounce vowel sounds in a foreign language using acoustic measurements of the vocal tract as feedback in real time," Lang. Speech **41**(1), 1–20.

Escudero, P., and Boersma, P. (**2004**). "Bridging the gap between L2 speech perception research and phonological theory," Stud. Second Lang. Acquisit. **26**(04), 551–585.

Flege, J. E. (**1995**). "Second language speech learning theory, findings, and problems," in *Speech Perception and Linguistic Experience: Issues in Cross-Language Research*, edited by W. Strange (York Press, Timonium, MD), pp. 233–277.

Flege, J. E. (**2002**). "Interactions between the native and second-language phonetic systems," in *An Integrated View of Language Development: Papers in Honor of Henning Wode*, edited by P. Burmeister, T. Piske, and A. Rohde (Wissenschaftlicher Verlag Trier, Trier), pp. 217–243.

Flege, J. E., MacKay, I. R., and Meador, D. (**1999**). "Native Italian speakers' perception and production of English vowels," J. Acoust. Soc. Am. **106**(5), 2973–2987.

Georgeton, L., Paillereau, N., Landron, S., Gao, J., and Kamiyama, T. (**2012**). "Analyse formantique des voyelles orales du français en contexte isolé: à la recherche d'une référence pour les apprenants de FLE (Formant analysis of the French oral vowels in isolated context: In a quest of a reference for French learners)," in *Proceedings of JEP-TALN-RECITAL*, pp. 145–152.

Golestani, N., and Pallier, C. (**2007**). "Anatomical correlates of foreign speech sound production," Cerebral Cortex **17**(4), 929–934.

Grønnum, N. (**1997**). "Danish vowels: The psychological reality of a morphophonemic representation," Proc. Journée dáEtudes Linguistiques [A day of Linguistic Studies] pp. 91–97.

Guenther, F. H. (**1994**). "A neural network model of speech acquisition and motor equivalent speech production," Biol. Cybern **72**(1), 43–53.

Hu, X., Ackermann, H., Martin, J. A., Erb, M., Winkler, S., and Reiterer, S. M. (**2013**). "Language aptitude for pronunciation in advanced second language (L2) learners: Behavioural predictors and neural substrates," Brain Lang. **127**(3), 366–376.

Ingram, J. C., and Park, S.-G. (**1997**). "Cross-language vowel perception and production by Japanese and Korean learners of English," J. Phonetics **25**(3), 343–370.

Kartushina, N., and Frauenfelder, U. H. (**2014**). "On the effects of L2 perception and of individual differences in L1 production on L2 pronunciation," Frontiers Psychol. **5**(1246), 1–17.

Lametti, D. R., Rochet-Capellan, A., Neufeld, E., Shiller, D. M., and Ostry, D. J. (**2014**). "Plasticity in the human speech motor system drives changes in speech perception," J. Neurosci. **34**(31), 10339–10346.

Leather, J. (**1996**). "Interrelation of perceptual and productive learning in the initial acquisition of second-language tone," in *Second-Language Speech: Structure and Process*, edited by A. James and J. Leather (Mouton de Gruyter, Berlin), pp. 75–101.

Liberman, A. M., and Mattingly, I. G. (**1985**). "The motor theory of speech perception revised," Cognition **21**(1), 1–36.

Lively, S. E., Logan, J. S., and Pisoni, D. B. (**1993**). "Training Japanese listeners to identify English/r/and/l/. II: The role of phonetic environment and talker variability in learning new perceptual categories," J. Acoust. Soc. Am. **94**(3 Pt 1), 1242–1255.

Loizou, P. (**1998**). "COLEA: A MATLAB software tool for speech analysis," Dallas, TX. Available at http://ecs.utdallas.edu/loizou/speech/colea.htm (Last viewed October 4, 2011).

Lopez-Soto, T., and Kewley-Port, D. (**2009**). "Relation of perception training to production of codas in English as a Second Language," J. Acoust. Soc. Am. **125**, 2756.

Massaro, D. W., Bigler, S., Chen, T. H., Perlman, M., and Ouni, S. (**2008**). "Pronunciation training: The role of eye and ear," Proc. Interspeech **9**, 2623–2626.

Ménard, L., Schwartz, J.-L., Boë, L.-J., Kandel, S., and Vallée, N. (**2002**). "Auditory normalization of French vowels synthesized by an articulatory model simulating growth from birth to adulthood," J. Acoust. Soc. Am. **111**(4), 1892–1905.

Öster, A.-M. (**1997**). "Auditory and visual feedback in spoken L2 teaching," in Reports from the Department of Phonetics, Umeå University, PHONUM **4**, 145–148.

Pallier, C., Colomé, A., and Sebastián-Gallés, N. (**2001**). "The influence of native-language phonology on lexical access: Exemplar-based versus abstract lexical entries," Psychol. Sci. **12**(6), 445–449.

Peperkamp, S., and Bouchon, C. (**2011**). "The relation between perception and production in L2 phonological processing," Proc. Interspeech **12**, 161–164.

Perrachione, T. K., Lee, J., Ha, L. Y. Y., and Wong, P. C. M. (**2011**). "Learning a novel phonological contrast depends on interactions between individual differences and training paradigm design," J. Acoust. Soc. Am. **130**(1), 461–472.

Pillot-Loiseau, C., Antolík Kocjančič, T., and Kamiyama, T. (**2013**). "Contribution of ultrasound visualisation to improving the production of the French /y/-/u/ contrast by four Japanese learners," in *Proceedings of the PPLC13: Phonetics, Phonology, Languages in Contact. Contact Varieties, Multilingualism, Second Language Learning*, pp. 86–89.

Piske, T., MacKay, I. R., and Flege, J. E. (**2001**). "Factors affecting degree of foreign accent in an L2: A review," J. Phon. **29**(2), 191–215.
Saito, K., and Lyster, R. (**2012**). "Effects of form-focused instruction and corrective feedback on L2 pronunciation development of /ɹ/ by Japanese Learners of English," Lang. Learn. **62**(2), 595–633.
Sheldon, A., and Strange, W. (**1982**). "The acquisition of /r/ and /l/ by Japanese learners of English: Evidence that speech production can precede speech perception," Appl. Psycholinguist. **3**(3), 243–261.
Simmonds, A. J., Wise, R. J. S., Dhanjal, N. S., and Leech, R. (**2011**). "A comparison of sensory-motor activity during speech in first and second languages," J. Neurophys. **106**(1), 470–478.
Steinlen, A. K. (**2005**). *The Influence of Consonants on Native and Non-native Vowel Production: A Cross-Linguistic Study* (Gunter Narr Verlag, Tübingen), pp. 71–117.
Wik, P. (**2004**). "Designing a virtual language tutor," in *Proceedings of the XVIIth Swedish Phonetics Conference*, Fonetik, pp. 136–139.
Wilson, S. M., and Gick, B. (**2006**). "Ultrasound technology and second language acquisition research," Proc. Generative Approach. Second Lang. Acquisit. **8**, 148–152.
Wong, J. W. S. (**2013**). "The effects of perceptual and or productive training on the perception and production of English vowels /I/ and /i:/ by Cantonese ESL learners," Proc. Interspeech **14**, 2113–2117.