<LRH>Talker Variability in Production Learning

<RRH>Talker Variability in Production Learning

<ArtType>EMPIRICAL STUDY

<AT>Talker and Acoustic Variability in Learning to Produce Second Language Sounds: Evidence from Articulatory Training

<AU>Natalia Kartushina[a, b] and Clara D. Martin[b, c]

<AF>[a]Department of Psychology, Faculty of Social Sciences, University of Oslo, [b]Basque Center on Cognition, Brain and Language, and [c]Ikerbasque Basque Foundation for Science

<AN>

The authors are grateful to Dr. Martijn Baart for his help with the preparation of the stimuli and to Dr. Alexis Hervais-Adelman for proofreading the manuscript. The authors would like to thank anonymous reviewers for their insightful comments and suggestions. This work was funded by a Swiss National Science Foundation Research Fellowship to NK (P2GEP1_165307). CM was supported by the grant PSI2017-82941-P from the Spanish government.

Correspondence concerning this article should be addressed to Natalia Kartushina, Department of Psychology, Faculty of Social Sciences, University of Oslo, Forskningsveien 3A, 0373 Oslo, Norway. E-mail: natalia.kartushina@psykologi.uio.no

<ABS>

Compared to low variability perception training, high variability training leads to better learning outcomes and supports generalization of learning. However, it is unclear whether the learning advantage is due to a presence of multiple talkers or to enhanced acoustic variability across target sounds. The current study addressed this issue in nonnative production learning. Spanish speakers were trained to produce the French /e/–/ɛ/ vowel contrast. The stimuli were recorded by five native French talkers for the multiple taker (MT) group or by one talker for the single-talker (ST) group, but acoustic dispersion of the vowels and context were matched between the two groups. Both training paradigms improved production accuracy, with slightly greater improvement in the ST group. However, only MT training enhanced the compactness of vowel categories and generalized to the production of sounds elicited by an unfamiliar speaker. This suggests that talker variability supports the establishment of abstract phonemic categories in production.

<KWG>Keywords production learning; talker variability; phonetic training; acoustic variability; second language; speech learning

<A>Introduction

<TXT>

Second language (L2) speakers can improve their perception and production of L2 speech sounds (vowels and consonants) by undergoing training whereby they receive feedback on their performance. Previous research has shown that perception training with highly variable stimuli (i.e., produced in multiple contexts by multiple talkers), as opposed to training with low variability stimuli (i.e., produced in a few contexts by a single talker), leads to greater improvements in perceptual identification performance (Logan, Lively, & Pisoni, 1991; Wang, Spence, Jongman, & Sereno, 1999) and supports generalization of learning to new speech tokens (Sadakata & McQueen, 2013) and speakers (Lively, Logan, & Pisoni, 1993). Therefore, in perception, high variability phonetic training supports the formation of robust abstract categories for L2 sounds. However, it remains unclear what drives the effect, mainly because the two main sources of variability—talker and context—are generally confounded in L2 research. Regarding speech production, to our knowledge, no study so far has addressed the role of stimulus variability in L2 learning. Yet a recent perception training study that assessed transfer of perceptual learning to production has shown that speakers had smaller improvements in the production of the trained sounds in the high than in the low variability group, suggesting that variability might not help to form accurate pronunciation patterns for L2 sounds (Evans & Martin-Alvarez, 2016). Given those unexplored aspects in L2 training, the current study had two aims: (a) to investigate, for the first time, the role of stimulus variability in L2 vowel production learning, and (b) to examine the contribution of talker variability while controlling for context because (thus far) these two main sources of variability have remained confounded in L2 research.

<A>Background Literature

<B>Variability in L2 Phonetic Learning

<TXT>

Speakers who acquire a L2 in adolescence or later often experience difficulties in the perception and production of L2 sounds. These difficulties can be minimized by phonetic training whereby L2 learners receive feedback on their perception and/or production of L2 sounds on a trial-by-trial basis. For instance, Iverson and Evans (2009) showed that German learners of English improved their perception of English vowels by 20% after undergoing five sessions of vowel identification training.

Training that contains highly variable stimuli has been shown to be more effective for the learning of L2 phonetic contrasts than low variability training (Brosseau-Lapré, Rvachew, Clayards, & Dickson, 2013; Lively et al., 1993; Logan et al., 1991; Sadakata & McQueen, 2013; Wang et al., 1999). For instance, in the study by Sadakata and McQueen, native Dutch speakers who were trained to identify the Japanese geminate–singleton fricative contrast /ss/–/s/ with a variable set of words recorded by multiple speakers (the high variability group) improved their identification of this contrast more than those who were trained in a low variability group, with a limited set of words recorded by a single speaker. Importantly, after training, only speakers in the high variability group were able to generalize learning to new sounds and speakers as attested by their better identification of (a) untrained stops and affricates and (b) trained fricatives produced by a new speaker (see also Lively et al., 1993, for similar results). There was no difference between the two groups in the amount of training-related gains in the discrimination of the two sounds. The authors concluded that input variability enhances categorical rather than (early or precategorical) acoustic processing of speech sounds and leads to the establishment of abstract representations for L2 sounds.

Other perception training studies, however, have shown that in some circumstances, variability can hinder learning. First, enhanced variability impairs learning of difficult L2 contrasts, for example, unfamiliar contrasts which depend on dimensions that do not exist in the learners' first language (L1) (Giannakopoulou, Brown, Clayards, & Wonnacott, 2017; Wade, Jongman, & Sereno, 2007; Wayland & Guion, 2004). For instance, Chinese speakers (who have linguistic experience with tonal contrasts) benefited from variability when learning to perceive a novel tonal contrast, whereas native English speakers did not (Wayland & Guion). In the same vein, Wade and colleagues showed that variability can diminish learning for highly confusable English vowels, compared to less confusable ones (i.e., those that were separated farther from each other in the acoustic space and overlapped less). These studies suggest that the effectiveness of high variability training depends on the nature of the categories to be learned, with unfamiliar (Wayland & Guion) or difficult L2 categories (Giannakopoulou et al.; Wade et al.) being learned less effectively in variable sets of stimuli.

Second, variability diminishes learning in novice learners (Chang & Bowles, 2015; Kingston, 2003). For instance, English speakers with no experience learning Mandarin acquired Mandarin tones better in monosyllabic words (characterized by less variation in individual tone contours) than in disyllabic words (characterized by more variation in tone contours) (Chang & Bowles, 2015). Interestingly, recent studies have shown that novice learners, can still benefit from variability if they have strong general perceptual abilities, such as the ability to detect a pitch contour (Antoniou & Wong, 2015; Perrachione, Lee, Ha, &

Wong, 2011; Sadakata & McQueen, 2014). The results of these studies suggest that variability interferes with initial learning of L2 contrasts for novice learners with poor perceptual abilities. L2 learners with high perceptual abilities, on the other hand, benefit from variability: They extract category-relevant information from phonetically variable input and establish phonologically constant categories. In addition, variability in stimuli seems to hinder acquisition in young learners (Evans & Martin-Alvarez, 2016; Giannakopoulou et al., 2017), perhaps because it places significant demands on attentional and/or cognitive resources which may present difficulties for children.

In sum, although high input variability is crucial for the generalization of perceptual learning, it represents a challenge for learners who lack experience with the L2 in general or with the target L2 phonetic property, and for individuals with weak initial perceptual abilities or limited attentional resources (as in the case of children). In addition, variability generally seems to hinder learning of difficult L2 sounds (i.e., those that are perceptually confusable with similar L1 sounds or with each other). One goal of the current study was therefore to investigate whether a hypothetically similar detrimental effect of variability extends to L2 production learning, by exploring the acquisition of a difficult L2 nonnative vowel contrast in novice learners.

<B>Sources of Variability in Perceptual Training

<TXT>

Traditionally, two sources of variability are manipulated in training studies—variability related to phonetic context and variability related to talker. Contextual variability refers to differences in phonetic environment—that is, surrounding sounds and prosodic positions—and results, in particular, in variability in sound-specific cues. For instance, some English vowels (Vs) have higher first and second formant (F1 and F2) values when produced in /hVt/ as compared to /bVt/ context (Steinlen, 2005). Talker variability refers to differences in the production of speech sounds between speakers and is due to variation in vocal tract morphology and physiology (Lammert, Proctor, & Narayanan, 2013; Ménard, Schwartz, Boë, & Aubin, 2007). For instance, men have larger vocal tracts than women; consequently, vowels produced by men have lower fundamental frequency (F0 or pitch), F1, and F2, than those produced by women (e.g., Peterson & Barney, 1952). Importantly, both contextual and talker-specific sources of variability create acoustic variability in sound-specific cues (e.g., F1 and F2 for vowels).[1] Talker variability, in addition, involves variability in talker-specific cues (e.g., pitch amplitude, pitch contour, speech rate).

There is currently no agreement as to how contextual and talker-specific sources of variability should be combined in order to create low versus high variability training conditions. Some studies have targeted variable sets of words produced by multiple talkers for the high variability group and a limited set of words produced by a single talker for the low variability group (Sadakata & McQueen, 2013). Others have maintained a constant context but manipulated the number of talkers/repetitions between the conditions (e.g., Barcroft & Sommers, 2005). These methodological differences across studies limit the generalizability of the results and obscure the origins of the variability advantage in perceptual learning.

Nevertheless, a few studies have contributed to a better understanding of the complex interaction between the two sources of variability. For instance, although the study by Chang and Bowles (2015) was not designed to contrast different sources of variability, it suggested that training with multiple phonetic contexts has more detrimental effects on learning than training with multiple talkers. In particular, introducing a new talker was shown to lead to fewer difficulties in tone identification than introducing a new context, suggesting that variability arising from differences in talkers is easier to process, presumably, because it is language universal (e.g., reflects gender-based differences in vocal production) and hence easier to predict and cope with. Acoustic differences driven by contextual (acoustic) variability, on the other hand, arise from language-specific allophonic and/or coarticulatory mechanisms, which are more difficult to predict on language-universal grounds (Chang & Bowles).

Brosseau-Lapré and colleagues (2013) have attempted to disentangle the effects of talker and acoustic variability on L2 phonetic learning. In this study, native English speakers were trained to identify the French /ə/–/ø/ vowel contrast. In order to create various training groups, the authors manipulated the distribution of the contrastive cues (F1) and the number of talkers. In the far condition, the within-category acoustic variability in F1 was 60 Hz, and the distance between the closest tokens of the two categories was 60 Hz; in the close condition, both of them were 40 Hz. Each of these training conditions was crossed with two talker conditions: multiple talker (*n* = 3) and single talker. The results showed that training was effective only in the two multiple talker conditions. The far multiple talker condition, however, led to greater improvements than the close multiple talker condition. The authors also tested the generalization of learning. There were no main effect of talker or acoustic variability, but the talker × acoustic variability interaction was significant. The multiple talker training was most effective in the far condition, whereas the single talker training was most effective in the close condition. Interestingly, the learning slopes in the identification were overall steeper (although not significantly different) in the single talker close condition, compared to all other conditions.

The results of Brosseau-Lapré and colleagues' study suggest that acoustic and talker variability contribute differently to L2 phonetic learning. Multiple talker training appears to be crucial for learning to discriminate ~~L2~~ L2 unfamiliar contrasts; more importantly, however, in the presence of acoustic variability, the benefits of multiple talker training are amplified. Acoustic variability, on the other hand, appears to aid the generalization of learning when speakers are trained with multiple talkers only; when trained with a single talker, such variability seems to hinder the establishment of abstract L2 categories. However, these findings should be taken with caution for two reasons. First, the target stimuli were not representative of natural variability found in speech production, because they contained only three examples of each vowel (repeated multiple times) in each condition. Second, the within-category acoustic variability and the acoustic distance between the vowels were confounded; therefore, better learning outcomes in the far condition could be attributed to greater acoustic distance between the vowel categories (hence less confusion) rather than greater acoustic variability across the vowels.

To summarize, the current state-of-the-art in L2 perception learning does not offer a clear understanding of the role that different sources of variability play in the process, nor does it examine potential origins of the variability advantage/disadvantage. Most importantly, there are no studies that have examined the acoustic variability associated with each source independently from each other and no studies exploring potential interactions between these sources of variability[2]. In particular, it remains unclear whether effects of variability arise from an increase in contextual, talker, or both sources of variability, or from an increase in the acoustic dispersion in the targeted speech sounds.

<B>Transfer of Perceptual Learning to Production

<TXT>

Although, to date, no research has addressed the role of variability in production learning, studies that have looked at the transfer of high variability perceptual training to production provide indirect evidence that perceptual learning gained in high variability conditions transfers to production (Bradlow, Pisoni, Akahane-Yamada, & Tohkura, 1997; Jügler, Zimmerer, Möbius, & Draxler, 2015; Wong, 2013; however, see Hwang & Lee, 2015, for contradictory results). For instance, Japanese learners of English improved their production of the English /r/–/l/ consonants by 8% after undergoing a three-week high variability identification training (Bradlow et al.). Wong has shown that high variability perception training, in combination with production training where participants saw videos of native speakers producing L2 vowels, was more effective (resulting in 66% accuracy) than high variability perception (36% accuracy) or production (29% accuracy) training implemented alone. However, it remains unclear, based on the results of these studies, whether the benefit to learning comes from the variability in stimuli or from a combination of production and perception training techniques.

To the best of our knowledge, only two studies have specifically compared the effects of high versus low variability perception training on production (Brosseau-Lapré et al., 2013; Evans & Martín-Alvarez, 2016). However, the results of these studies are inconclusive. For instance, Evans and Martín-Alvarez reported that perception training with low speaker variability, but not training with high speaker variability, improved the production of the English /i/–/ɪ/ vowels in young Spanish children and adults. Brosseau-Lapré and colleagues, on the other hand, did not find significant differences in the amount of improvement for the production of the French /ɔ/–/ø/ vowels between the English adults trained in multiple talker versus single talker conditions, nor between the groups trained with acoustically more versus less variable stimuli. These differences in study outcomes might stem from differences in participants tested (e.g., children vs. adults), amount of their L2 experience (some vs. none), their L1 (Spanish vs. English), training methods (picture- vs. word-identification), and/or amount of acoustic variability in the input. To summarize, the results of the above studies are contradictory in that they do not answer the question whether input variability during perception training is beneficial when learners attempt to produce L2 sounds.

Two dominant models of L2 phonological acquisition—the Perceptual Assimilation Model by Best (1995) and the Speech Learning Model by Flege (1995)—state that accurate perception of L2 sounds is a prerequisite for their accurate production. Indeed, multiple studies have shown that L2 production is related to L2 perception (Bradlow et al., 1997; Flege, Bohn, & Jang, 1997; Flege, MacKay, & Meador, 1999; Ingram & Park, 1997), and this relationship is stronger in advanced/proficient L2 learners (Flege, 1999; Flege & Schmidt, 1995). Therefore, similar to the results of perceptual training studies, one could expect that high variability production training should lead to greater learning outcomes than production training with low variability, unless, for example, L2 learners are inexperienced speakers and/or the targeted L2 sounds are difficult or confusable.

However, there are three reasons to believe that variability in production training might not offer the same benefits as variability in perception training. First, L2 sound repetition (imitation), which is often used in production training studies, has recently been shown to rely on different processes or skills than those involved in sound perception (Hao & de Jong, 2016). In this study, L2 learners performed better in imitation, compared to perceptual identification and reading aloud, suggesting that imitation bypasses some aspects of sound processing shared by the two latter tasks. Second, in production training paradigms, participants are encouraged to rely almost exclusively on visual articulatory feedback, that provides (among others) information about the position of the target in the acoustic space,, which might discourage participants from paying attention to perceptual input. Finally, a growing body of research showing no robust relationship between L2 perception and production, including weak relationships, absence of transfer of perception training to production (e.g., Lopez-Soto & Kewley-Port, 2009; Peperkamp & Bouchon, 2011), and disruptive effects of speech production on perceptual learning (Baese-Berk & Samuel, 2016), also suggests that the effects of variability found in perception might not be the same in production.

<A>The Current Study

<TXT>

The current study fills a gap in our understanding of the role of variability in L2 production learning by assessing the effects of two types of training input, both having the same acoustic (and contextual) variability, but differing with respect to the number of talkers (five talkers vs. one talker) used to record the stimuli. The question guiding this study was whether training with multiple talkers leads to greater learning outcomes and greater generalization of learning than training with a single talker.

In order to address the role of input variability in learning to produce L2 sounds, native Spanish speakers with no knowledge of French were trained to produce the French mid-open and mid-close front unrounded /ɛ/–/e/ vowels using the same articulatory training technique as in recent training studies (Kartushina, Hervais-Adelman, Frauenfelder, & Golestani, 2015, 2016). Adult Spanish learners of French have difficulties in the discrimination of the French /ɛ/–/e/ vowel height contrast which they assimilate perceptually to the Spanish /e/ vowel. For instance, Spanish learners misidentify the French /ɛ/ as the French /e/ 55% of the time, and they associate the French /e/ with the French /ɛ/ 35% of the time (Kartushina & Frauenfelder, 2014). Spanish speakers also confuse the French /ɛ/–/e/ vowels in production: The acoustic spaces for these vowels overlap largely in the F1/F2 space, and the F1 distance between the two vowels, which is the crucial parameter distinguishing them, is 10 times smaller in the production of these vowels by Spanish learners than by native French speakers. Even after four years of classroom experience with French, Spanish speakers still show persistent difficulty in the perception and production of the French /ɛ/–/e/ vowels, suggesting that this height contrast is difficult for them to acquire (Kartushina & Frauenfelder).

The present study made use of an articulatory training technique that implements a real-time analysis of the acoustic properties of vowels (F0, F1, and F2) produced by learners to provide them with immediate, trial-by-trial visual feedback about their articulation alongside the target vowel acoustic space derived from the productions of native French speakers (as in Kartushina et al., 2016). Previous production training studies have shown that one hour of training following this technique improves learners' pronunciation of nonnative vowels by 17–19% (Kartushina et al., 2015, 2016). In addition, articulatory training with feedback appears to enhance learners' perception of the trained sounds (Kartushina et al., 2015) and their articulatory compactness (which is the inverse of variability), with greater enhancements in compactness for those vowels that had more training-related improvements in production accuracy (Kartushina et al., 2016).

Compactness reflects the acoustic stability of articulatory production between different realizations (tokens) of the same sound (type). Compactness in the production of L2 categories is related to their accuracy. For instance, speakers who produce L2 vowels more accurately are those who produce them more compactly (Kartushina & Frauenfelder, 2014). In a L1, the compactness in the production of vowel categories is also related to perception accuracy: Speakers with more distinct (separated farther apart) and compactly distributed vowel categories in production discriminate vowels more accurately than those whose productions are less distinct and more overlapping (Franken, Acheson, McQueen, Eisner, & Hagoort, 2017; Perkell et al., 2004). These results are in line with the Directions Into Velocities of Articulators (DIVA) neural network model of speech acquisition and production (Guenther, 1994), which states that articulatory movements for speech sounds are primarily planned in the perceptual space. Hence, greater compactness in sound production stems from more precise, accurate perceptual targets. In sum, the results of L1 and L2 research suggest that an enhancement in vowel compactness can be considered as an indirect measure of improvement in vowel perception.

Two training conditions (multiple talker vs. single talker) were developed to test the role of talker variability in L2 production learning. The two conditions crucially differed in between-talker variability in F0 (pitch), which was greater in the multiple talker condition than in the single talker condition. All the other key parameters were matched between the two conditions: the total number of tokens, the number of contexts in which the stimuli were recorded (i.e., contextual variability) and, importantly, the acoustic dispersion of the target categories in F1/F2 space, that is, acoustic variability (see below for details). The effect of training on the production of the trained vowels was assessed in a vowel repetition task, performed before and after training. In this task, half of the stimuli were produced by a familiar voice used for training, the other half were spoken by an unfamiliar voice, which was used to test the generalization of learning to a new speaker. Two dependent measures were considered: (a) the acoustic distance from each participant's vowels to the target vowel space (i.e., accuracy of production relative to the acoustic values for the target sound) and (b) the compactness of participants' vowel categories in the acoustic space.

Assuming a relationship between L2 perception and production (Best, 1995; Flege, 1995), the following predictions, stemming from L2 perception research, were considered for the accuracy and compactness measures. For the accuracy measure, given that participants were novice learners and that the target contrast was confusable, it was expected that, for the familiar voice, training with multiple talkers would lead to less improvement in the production accuracy of the target vowels, compared to training with a single talker. However, for the unfamiliar voice, greater improvements were expected for training with multiple talkers, because variability was expected to support generalization of learning. For the compactness measure, in line with previous production training studies showing that improvements in production accuracy were accompanied by an enhancement in the acoustic compactness of the trained vowels (Kartushina et al., 2015, 2016), both groups were expected to show an enhancement in the acoustic compactness of the trained vowels. However, because higher compactness in production is related to greater perception accuracy (Franken et al., 2017), and the latter is achieved through high variability training (Lively et al., 1993), greater enhancements in compactness were expected for training with multiple talkers, compared to training with a single talker, particularly in response to the unfamiliar voice. Yet, as stated previously, there might be dissociations between L2 speech production and perception in (inexperienced) L2 learners (Flege & Schmidt, 1995), with higher perception benefits obtained in perception training involving multiple talkers but leading to no benefits for production (Brosseau-Lapré et al., 2013). Given that outcome, and considering previously mentioned specific aspects of the current design (i.e., use of repetition task and visual articulatory feedback in response to participants' production), no benefits of talker variability might emerge for either production measure.

<A>Method

<B>Participants

<TXT>

Native Spanish female speakers[3] (N = 30; M_{age} = 24 ± 3 years) took part in the study. Prior to the experiment, participants completed a language background questionnaire. None had previous experience with French, Galician or Catalan (three languages frequently learnt in Spain that contain the target /e/–/ɛ/ vowels). Some participants had knowledge of other languages (German, English), but none were reported as being spoken proficiently. Participants reported no history of speech or hearing impairment. Informed consent was obtained from all participants, and they received financial compensation for their participation.

<B>Stimuli

<TXT>

Six female native monolingual French speakers were recorded reading two French sentences. Each sentence contained the target vowel in isolation and in five consonant–vowel (CV) phonetic contexts: /dV/, /gV/, /lV/, /sV/, /tV/. For example, for the target /ɛ/ vowel, the sentence was *Je dis* /ɛ/ *comme dans dès*, *gai*, *lait*, *sait*, *tait* ("I say /ɛ/ as in from, cheerful, milk, knows, stays, quiet"). The sentences were repeated three times. Recordings were carried out in a quiet room using a Marantz PMD670 portable recorder with a Shure Beta 58A microphone sampled at 22.05 kHz directly to 16-bit audio files.

Vowel tokens produced by all speakers (1 through 6) in words and in isolation (108 per target vowel) were extracted and trimmed to 250 milliseconds, relative to the midpoint, using PRAAT software (Boersma & Weenink, 2010). The spectrograms were inspected. Vowel onset and offset were marked at the first and last glottal striations showing formant structure. The midpoint of the vowel was located, and a segment of 250 milliseconds, centered at the midpoint, was extracted. A linear amplitude ramp of 20 milliseconds was then applied to the onsets and offsets of the trimmed vowels. The amplitude of all vowel tokens was equalized across speakers using an automatic procedure in PRAAT. All vowel tokens were resampled to 11 kHz.

<B>Training Stimuli

<C>*Multiple Talker Condition*

<TXT>

Vowel tokens from five speakers (1 through 5) were selected for the multiple talker training condition. In total, there were 90 training tokens for each vowel (5 speakers × 6 contexts × 3 repetitions).

<C>*Single Talker Condition*

<TXT>

Vowel tokens from one of the five speakers used for the multiple talker condition, (Speaker 3) were selected for the single talker condition. There were 18 tokens for each vowel produced by Speaker 3. In order to create stimuli for the single talker condition that had similar acoustic dispersion in F1/F2 space to the stimuli in the multiple talker condition but that differed with respect to the amount of interspeaker variability (i.e., variability in F0), a source-filter synthesis was performed using PRAAT. First, we extracted the glottal source from the 18 tokens produced by Speaker 3 using the inverse-filtering procedure detailed in the PRAAT manual, which yielded 18 glottal sources. Second, we extracted the filter from the 90 tokens (produced by the five speakers) using the linear prediction technique with the following parameters: prediction order of 11, window length of 0.01, time step of 0.005, and preemphasis frequency of 50 Hz. Third, we synthesized new vowel tokens using the 18 sources obtained from Speaker 3 and the 90 filters from the five speakers, with each source synthesized using five randomly chosen filters.[4] This procedure yielded 90 tokens with the source characteristics of Speaker 3, but with the formant values belonging to the productions of the five speakers used in the multiple talker condition. This procedure was performed separately for the /ɛ/ and /e/ vowels. Vowel spaces and individual tokens used for each of the training conditions are illustrated in Figure 1.

<COMP: Place Figure 1 near here>

<C>*Comparison of Training Stimuli*

<TXT>

All vowel tokens from both training conditions underwent acoustic analyses in MATLAB (summarized in Table 1) using a similar automatic procedure to that used by Kartushina et al. (2015). The formant values F1 and F2 were averaged over a 100 millisecond segment centered at the midpoint of the tokens. The F1 and F2 were computed by solving for the roots of the Linear Predictive Coding polynomial, using an adaptation of the scripts from the

COLEA software for speech analysis (Loizou, 1998) and setting the value for the Linear Predictive Coding order at 11. The F0 was analyzed using the cepstrum method and was averaged over the whole duration of the tokens.

<COMP: Place Table 1 near here>

Statistical analyses were used to examine differences in vowel formant frequencies between the two training conditions. Due to the specificity of the source-filter synthesis, the F1 and F2 of the resulting tokens differed slightly from those of the corresponding filters; however, the overall mean F1 and F2 did not differ significantly between the two settings for either of the target vowels. Crucially, there was no difference in the amount of acoustic variability in F1 or F2 between the two settings, as attested by the results of the Bartlett test of homogeneity of variances (see Table 1). The variability in F0, on the other hand, was higher in the multiple talker than in the single talker condition (as shown in Figure 2 and summarized in Table 1). The F1–F0 and F2–F0 differences of the vowel tokens were used to construct representative matching target vowel space composed of 90 vowel tokens for each vowel and setting.

<COMP: Place Figure 2 near here>

<C>Pre-/Posttraining Tests

<TXT>

The effects of training on the production of the target vowels were assessed in a vowel repetition task using familiar and unfamiliar voice sets. The familiar voice set included 36 vowel tokens (18 per vowel) produced by Speaker 3, to whom both groups were exposed to during training. The unfamiliar voice set included 36 vowel tokens (18 per vowel) produced by Speaker 6 (unfamiliar to both groups). The distribution of both vowels in the unfamiliar voice set was acoustically closer to the prototypical Spanish /e/ vowel (F1 = 531 Hz, F2 = 2,159 Hz for female speakers, based on Chládková, Escudero, & Boersma, 2011) than the distribution of vowel tokens in the familiar voice set (for an illustration, see Appendix S1 in the Supporting Information online). Thus, given that Spanish speakers assimilate the French /e/–/ɛ/ contrast to the Spanish /e/ vowel, we expected to find better production accuracy for the unfamiliar than for the familiar voice set.

<B>Procedure

<TXT>

Participants were trained to produce nonnative French /ɛ/ and /e/ vowels over three training sessions that were administered on alternating days: Monday, Wednesday, and Friday. Participants were assigned to one of the two training conditions pseudorandomly (*n* = 15 per condition). A vowel repetition task, performed during the first (before training) and last (after training) session, was used to assess training effects on vowel production.

The tasks were performed on a DELL computer, using Sennheiser PC-350 headphones fitted with a microphone. The pre- and posttraining tests were administered using the DMDX software (Forster & Forster, 2003). The training procedure was administered using MATLAB (MATLAB Release 2014b) and Psychophysics Toolbox extensions (Brainard, 1997; Pelli, 1997). Throughout the entire experiment (testing and training), stimuli were presented at a comfortable listening level, which was adjusted on a participant-by-participant basis.

<C>*Training*

<TXT>

The training procedure was similar to that used by Kartushina et al. (2016) and included three training sessions. The first and last sessions consisted of three training blocks, and the second session consisted of four training blocks. Each training block comprised two mini-blocks (one per vowel), the order of which was randomized. Within each mini-block, 90 condition-specific tokens of the given vowel were presented resulting in 900 productions of each vowel

(68 minutes in total). There were pauses between blocks, and the duration of these was controlled by the participants.

At the beginning of the first session, participants received instructions explaining the nature of the feedback and its correspondence to the articulators (i.e., tongue position with respect to the vertical and horizontal axes) during production. The visual feedback consisted of a two-dimensional visual display showing F1 (which corresponds to tongue height/mouth openness) along the y-axis, and F2 (which corresponds to the front-back position of the tongue) along the x-axis. Prior to the main experiment, participants were familiarized with this feedback while producing the Spanish /i/, /a/, and /u/ vowels (shown on the screen), six times each. The familiarization phase lasted 5 minutes and was immediately followed by the training.

Trials began with the appearance of a white screen for 500 milliseconds. Immediately after that, a vowel was presented over the headphones for 250 milliseconds. Then, participants saw a visually presented countdown signal “3, 2, 1” that lasted 1,050 milliseconds. This was followed by the “!” sign, at exactly which point a recording lasting 700 milliseconds began. Participants were instructed to repeat the vowel as accurately as possible while the “!” sign was on the screen. The produced vowels were recorded on a hard-disk as 16-bit .wav files, sampled at 11 kHz. An automatic procedure identified the onset and the offset of the recorded vowels and calculated vowel length and intensity. If the vowel was shorter than 50 milliseconds and/or had very low intensity, participants were asked to repeat the trial by producing the vowel longer and/or louder. Otherwise, the F0, F1, and F2, averaged over a 100 millisecond segment centered at the midpoint of the produced vowel, were estimated and visual feedback was presented on-screen for 1,500 milliseconds. We encouraged participants to rely primarily on visual feedback (and less on their own perception of production accuracy). To calculate the formants, we used the same automated procedure as the one used for the training stimuli. The trial ended with a cross that appeared at the center of the screen for 500 milliseconds. One trial lasted 4.5 seconds.

<C>*Visual Feedback During Training*

<TXT>

The articulatory feedback was based on an immediate, trial-by-trial acoustic analysis of the vowels produced by participants. It showed (a) the position of the participant's vowel production in F1–F0 (y-axis)/F2–F0 (x-axis) space, which corresponds to the degree of openness and to the back-to-front position of articulation, and (b) the target vowel acoustic space. Participants saw the target vowel space derived from the 90 productions by five speakers in the multiple talker conditions and from the same number of productions by a single speaker in the single talker condition. The target vowel space was represented as an ellipse having major and minor axes with a length of 0.5 standard deviations of the mean along the given axis (for examples of visual feedback, see Appendix S2 in the Supporting Information online). The visual feedback also included information about the position of the L2 vowel produced on the previous trial. Importantly, since the visual feedback regarding the F1 and F2 was adjusted for F0 (Ménard, Schwartz, Boë, Kandel, & Vallée, 2002), the target acoustic space was also adjusted for F0 (F0 was subtracted from F1 and from F2). The feedback image was presented on a 43 centimeter screen, in a window of 1,024 × 768 pixels. The range of values was 2,600–600 Hz on the x-axis, and 650–50 Hz on the y-axis.

<C>*Vowel Repetition Task*

<TXT>

In the vowel repetition task, participants were asked to repeat, as accurately as possible, the 36 exemplars of the French /ɛ/ and /e/ vowels, two times each, for a total of 72 production trials per vowel (with no feedback). Vowel tokens produced by different talkers (familiar and unfamiliar voices) were mixed randomly for each participant. On each trial, a cross appeared on the screen for 1,000 milliseconds, then an auditory stimulus was presented for 250 milliseconds, and participants were prompted by an image of a microphone to repeat the vowel. The cue remained on the screen for 2,000 milliseconds, and responses were recorded during this period.

<A>Data Analysis

<TXT>

Recordings from the pre- and posttraining vowel repetition tasks were analyzed acoustically in order to compute two dependent measures: production accuracy (distance) and compactness. Before completing the analyses, the recordings were verified for their auditory quality in terms of intensity (i.e., sound should be clearly audible), length (i.e., minimum 50 milliseconds), and absence of noise (e.g., coughs, sneezes, sighs, etc.). Then, the remaining vowel tokens were analyzed acoustically in MATLAB using an automated four-step procedure. First, for each token (18 per vowel and condition) we identified its onset and offset; second, we computed the F0 and the F1 and F2 formant values, averaged over a 100 millisecond segment centered at the midpoint (the same procedure as the one used for the training stimuli), then obtained the F1–F0 and F2–F0 differences; these values were used to calculate the production accuracy (in step three) and vowel compactness (in step four).

Accuracy of the nonnative productions was measured using the Mahalanobis acoustic distance—henceforth, distance score (DS)—in the F1–F0/F2–F0 space between each token produced by the participant and the representative target vowel space (feedback) used for training (also adjusted for F0). The Mahalanobis distance was used in order to take into account the natural (unequal) variability in F1 and F2 in vowel production, as characterized by the target elliptic spaces. The DS is a scale-invariant, unitless measure of distance, in terms of standard deviations, from a given point to a distribution. By this means, we assessed whether participants learned the phonetic properties of the prototypical vowels used in training (as illustrated by visual feedback representing 0.5 standard deviations of the mean of the heard tokens along F1 and F2). Thus, a decrease in DS for both the familiar and unfamiliar voice sets would indicate that participants produced the vowels closer to the mean of the “representative” target space irrespectively of the familiarity with the voice. For each participant, 72 DSs (2 vowels × 2 sets × 18 tokens) were obtained per vowel before and after training.

The compactness of vowels in the acoustic space was represented by the compactness score (CS). The CS represents the size of vowel category in the acoustic space and reflects the acoustic stability of productions between realizations. CS was based on a joint analysis of the distribution of 18 tokens in the F1–F2 space (adjusted for F0). It was calculated as the area of an ellipse with major (a) and minor (b) axes, having a length of one standard deviation along the given axis, CS = π × a × b (Figure 3). For each participant, eight CSs (2 vowels × 2 sets × 2 sessions) were obtained.

<COMP: Place Figure 3 near here>

Statistical analyses were run on the DS and CS values using the R software (R Core Team, 2012). A two-step procedure was used: First, the dependent variable was fitted to a general linear mixed-effects model using the lme4 package (Bates, Mächler, Bolker, & Walker, 2015). The model was always composed of both a fixed and a random structure. The fixed structure included the main factors of interest and their interactions; the random structure included both by-subject and by-token random intercepts and slopes adjusted for the main effect of training, namely, session; the slopes included a correlation parameter with the fixed factor (Barr, Levy, Scheepers, & Tily, 2013). Second, the significance of the main effects and of the interactions was computed using the anova function, implemented in the lmerTest package (Kuznetsova, Brockhoff, & Christensen, 2017). The effect size estimate—the conditional coefficient for mixed-effect models (R^2)—was computed using the r.squaredGLMM function from the “MuMIn” package (Nakagawa & Schielzeth, 2013). If multiple paired analyses were necessary, the lsmeans function implemented in the lsmeans package was performed on the model (Lenth, 2016). In the latter analyses, the *p* value was adjusted for multiple comparisons using the default built-in Tukey method.

<A>Results

<B>Vowel Accuracy

<TXT>

All recordings from Participant 26 in the multiple talker condition were discarded due to technical issues. Based on initial quality checks, 35 additional tokens were discarded. Outliers and extreme values were detected in R using Q-Q plots and were removed; these represented 0.9% of the data. The remaining DS values ranged from 0.01 to 11.96, and had a mean of 2.06 ± 1.63.

A mixed-effect model examining the effects of session (pre vs. post), condition (multiple vs. single talker), set (unfamiliar vs. familiar voice), vowel (/e/ vs. /ɛ/), and their interactions on DS yielded a significant effect of session, $F(1, 27) = 15.4$, $p < .0001$, $R^2 = .30$, indicating that productions of the French vowels were more accurate (indicated by a decrease in DSs) at the posttraining test ($M = 2.46$) than at the pretraining test ($M = 3.14$). There was also a significant effect of set, $F(1, 3986) = 706.7$, $p < .001$, $R^2 = .41$, with vowels being produced more accurately in the unfamiliar voice set ($M = 2.18$) than in the familiar voice set ($M = 3.43$). The following interactions were significant: (a) condition × vowel, $F(1, 3986) = 19.11$, $p < .001$, $R^2 = .31$, with larger group differences in production accuracy for the /ɛ/ vowel than for the /e/ vowel; (b) session × set, $F(1, 3986) = 8.3$, $p = .004$, $R^2 = .41$, with larger effects of session for the vowels in the familiar voice set than in the unfamiliar voice set; and (c) set × vowel, $F(1, 3986) = 12.79$, $p < .001$, $R^2 = .41$, with larger differences in production accuracy between the two voice sets for the /e/ vowel than for the /ɛ/ vowel. More importantly, there was also a significant session × condition × set interaction, $F(1, 3986) = 11.32$, $p < .001$, $R^2 = .42$ (see Figure 4). All remaining effects were not significant ($p > .10$).

<COMP: Place Figure 4 near here>

<C>*Familiar Voice Set*

<TXT>

The significant three-way interaction was explored further through planned paired comparisons using the lsmeans function. For the familiar voice set, there was a significant effect of session both in the multiple talker condition, $\beta = 0.66$, $SE = 0.26$, $t = 2.61$, $p = .01$, $R^2 = .29$, and in the single talker condition, $\beta = 0.96$, $SE = 0.25$, $t = 3.8$, $p < .001$, $R^2 = .51$ (see Figure 4, left panel). As a result of training, for the familiar voice set, the multiple talker and the single talker groups gained, on average, 0.66 and 0.96 DS units, which corresponded to 18% and 24% of improvement in production accuracy, respectively. To explore whether the improvements in vowel production differed between the two groups, the amount of training-related improvement in vowel production for each participant and vowel was calculated. A one-tailed *t* test revealed a marginally significant effect of condition, $t(38) = –1.5$, $p = .06$, $d = 0.55$, indicating that participants in the single talker group showed a tendency to improve their production accuracy more than those in the multiple talker group.

<C>*Unfamiliar Voice Set*

<TXT>

For the unfamiliar voice set, similar analyses revealed a significant effect of session for the multiple talker group only, $\beta = 0.72$, $SE = 0.25$, $t = 2.80$, $p = .0026$, $R^2 = .36$ (see Figure 4, right panel). As a result of training, for the unfamiliar voice set, the multiple talker training group gained 0.72 DS units, which corresponded to 29% of improvement in production accuracy. The single talker training group gained 0.37 DS units (15.5%), but this improvement was not statistically significant (*p* = .084). Individual learning slopes for the multiple and single talker groups for each voice set are illustrated in Figure 5.

<COMP: Place Figure 5 near here>

<B>Vowel Compactness

<TXT>

A mixed-effect model examining the effects of session (pre vs. post), condition (multiple vs. single talker), set (unfamiliar vs. familiar voice), vowel[5] (/e/ vs. /ɛ/), and their interactions on CSs revealed a significant effect of session, $F(1, 189) = 10.20$, $p < .001$, $R^2 = .12$, with CSs decreasing between pretraining (M = 186 kHz) and posttraining (M = 142 kHz), and a significant effect of set, $F(1, 189) = 4.06$, $p = .04$, $R^2 = .09$, with the vowels produced more compactly in the familiar voice set (M = 150 kHz) than in the unfamiliar voice set (M = 178 kHz). There was also a marginal effect of vowel, $F(1, 189) = 3.31$, $p = .07$, $R^2 = .09$, with participants demonstrating a tendency to produce the /ɛ/ vowel (M = 151 kHz) more compactly than the /e/ vowel (M = 177 kHz). More importantly, there was a significant session × condition interaction, $F(1, 189) = 4.06$, $p = .04$, $R^2 = .14$. No other effects reached significance ($p > .20$). Paired comparisons further revealed that training improved the compactness of vowel categories in the acoustic space for the multiple talker group, $\beta = 72.63$, $SE = 20.05$, $t = 3.62$, $p < .001$, $R^2 = .19$, but not for the single talker group ($p = .39$), as illustrated in Figure 6. As a result of training, vowel categories produced by participants in the multiple talker training group gained, on average, 73 kHz in their compactness, with vowels becoming 38% more compact (less variable) after training (for individual gain scores, see Appendix S3 in the Supporting Information online).

<COMP: Place Figure 6 near here>

<A>Discussion

<TXT>

This study examined the role of talker variability in the learning to produce L2 sounds and in the generalization of learning to a new (unfamiliar) speaker. Native Spanish speakers with no knowledge of French were trained to produce the French /e/–/ɛ/ vowel contrast using an articulatory feedback technique. During training, participants heard one of the vowels, repeated it and received immediate visual feedback showing (in F1/F2 space) the position of their production along with the target vowel space. The multiple talker group received the stimuli spoken by five native French speakers, whereas the single talker group was exposed to the productions by one native speaker only. Most importantly, the acoustic dispersion of vowels in the F1/F2 space was matched between the two groups. To assess the effects of training on vowel production and the generalization of learning, participants performed a vowel repetition task using both familiar and unfamiliar voice sets, before and after training.

<B>Effectiveness of Training for Familiar Voices

<C>*Vowel Accuracy*

<TXT>

One hour of training with articulatory feedback generally improved the production of the trained vowels in both training groups. Participants in the multiple talker and the single talker groups produced the trained vowels acoustically closer to the target vowel spaces after training than before training. There were no differences in the amount of improvement between the trained vowels /e/ and /ɛ/. Both benefited from training in a similar way. Nevertheless, training-related improvements in vowel production were slightly (although marginally significantly) higher in the single talker group (24%) than in the multiple talker group (18%). Given our results, greater benefits of low variability training, previously reported for perception in novice speakers (Chang & Bowles, 2015; Kingston, 2003) and for difficult L2 sounds (Wade et al., 2007), could be similarly attributed to less variability in talker-specific cues rather than to a decrease in sound-specific acoustic variation. However, the results reported by Chang and Bowles suggest that a decrease in variability in sound-specific cues (e.g., pitch contour for tones) might also contribute to better perception learning. Yet because pitch is a talker-specific cue, speakers of nontonal languages might need more time and resources to learn to process it phonemically, which might explain their better performance with the less variable stimuli (in terms of sound-specific cues). More research is needed to understand the role of variability in talker- versus sound-specific cues on L2 learning, examining, in particular, whether the effect of variability depends on the type of a phonetic cue used to distinguish L2 sounds.

The trend for greater improvements shown by the single talker group with the familiar voice set may be attributed to a higher cognitive load needed to simultaneously learn two confusable vowels in the multiple talker context. Recently, Antoniou and Wong (2015) showed that resolving talker variability requires the allocation of additional cognitive resources. Under high cognitive load conditions, low aptitude perceivers showed a significant deterioration in their accuracy and speed in the identification of nonnative (variable) pitch-contours, whereas high aptitude perceivers showed no sign of deterioration. Considering this result, and given that the target vowels used here (French /e/–/ɛ/) are very difficult to discriminate for native Spanish speakers (Kartushina & Frauenfelder, 2014), learning to produce acoustically close L2 sounds through multiple talker training might place major demands on learners’ cognitive resources. Consistent with this view, Brosseau-Lapré and colleagues (2013) showed that perception training with multiple talkers was most effective when the trained L2 categories were acoustically far from each other. A detailed look at individual data in our study also supports this hypothesis. In fact, the three participants who had the highest improvements in the production accuracy for the /ɛ/ vowel were the same who had the lowest benefits for the /e/ vowel; indeed, their production accuracy decreased for the /e/ vowel after training (see the data for Participants 6, 8, and 17 in Appendix S3). These observations suggest that these participants may not have been able to learn two confusable vowels at the same time through training with multiple talkers. Instead, they may have concentrated on the production of the more difficult vowel (which was /ɛ/),[6] showing an improvement in their production accuracy. More research is needed to understand the role of individual factors and the acoustic proximity between L2 sounds in production learning.

Two further alternative explanations for greater improvement demonstrated by the single talker group with the familiar voice set could also be considered. First, this trend could be attributed to a greater exposure (5 times more) to the stimuli produced by the same (familiar) speaker in the single talker group than in the multiple talker group.[7] However, if frequency of exposure mattered, then participants in the single talker group should also have shown a reduction in variability of their realizations of individual vowel tokens, which would indicate their better knowledge of the vowel distribution in the single talker's speech. However, this was not the case. Alternatively, group differences in the amount of improvement could be due to a relatively larger pretraining distance to the target in the single talker group (M = 4.02 DS units), compared to the multiple talker group (M = 3.64 DS units, as shown in Figure 4), suggesting that participants in the single talker group had more room for improvement (Kartushina et al., 2015). However, additional analyses revealed that these pretraining differences were not statistically significant (p = .40); therefore, it is more likely that the trend for greater improvements in the single talker group was due to higher cognitive resources needed to learn two confusable L2 sounds under high talker variability.

<C>*Vowel Compactness*

<TXT>

The analyses of compactness of vowel productions for the familiar voice set showed training-related improvements only for the multiple talker group. There were no changes in vowel compactness in the single talker group. The lack of improvements in vowel compactness in the single talker group, despite considerable gains in production accuracy, was not anticipated, and it suggests that the benefits of training for the two measures are independent. Given that previous training studies showing a reduction in vowel compactness also included talker variability (three talkers) in the stimuli, the current results suggest that the variability in acoustic cues (e.g., pitch strength, pitch contour, timbre) arising from a single talker does not suffice to stabilize L2 production. Moreover, considering that single talker training does not improve categorical perception (Sadakata & McQueen, 2013) and that categorical perception is related to an individual's vowel compactness (Franken et al., 2017; Perkell et al., 2004), the lack of improvement in compactness observed in the single talker group might be due to the lack of improvement in the categorical perception of the trained sounds. However, data on participants' perception of the target sounds are needed to fully support this interpretation.

Taken together, the results for both vowel accuracy and compactness measures suggest that the single talker group enhanced the acoustic (precategorical) processing of speech sounds, in that trained vowels were produced better after training through participants' approximating the respective targets in the acoustic space. However, single talker training had no impact on the stability of these vowel productions. Although, on average, the acoustic values for the vowels spoken by single talker group were closer to the targets after training, the vowel categories remained widely distributed and overlapping (for an illustration, see Appendix S4 in the Supporting Information online).

In contrast, the multiple talker training group had two significant outcomes: an enhanced acoustic processing of L2 sounds accompanied by more stable realizations of those sounds, as attested by an increase in accuracy and compactness of vowel categories after training. Although, after training, Spanish speakers' acoustic spaces for the trained vowels (M = 118 kHz) did not reach the compactness values for the vowels produced by native French speakers (M = 79 kHz), training-related gains in compactness for the multiple talker group sufficed to make the categories more distinguishable and not as overlapping in the acoustic space as they were at pretraining (see Appendix S4). In sum, training with acoustically variable stimuli alone (without speaker variability) likely led to better learning outcomes in terms of production accuracy, but did not suffice to create more stable realizations. To do so would require both acoustic and speaker variability. Assuming a relationship between L2 perception and production (Best, 1995; Flege, 1995), the current results suggest that the benefits of training reported in previous perception studies are due, analogously, to variability in talker-relevant cues.

<B>Generalization of Learning to an Unfamiliar Speaker

<C>*Vowel Accuracy and Compactness*

<TXT>

The results for the vowel accuracy and compactness measures showed that only multiple talker training led to the generalization of learning to an unfamiliar speaker. Participants who were trained with multiple talkers produced the /e/–/ɛ/ vowels in the unfamiliar voice set more accurately (29% improvement) and less variably after training. Single talker training, on the other hand, did not support the generalization of learning. Participants in the single talker group did not gain in the compactness of the trained vowels, nor did they improve their production accuracy (although their productions were 15% closer to the target after training, this improvement was not significant).[8] These results are in line with previous perception studies showing that variability is required for the generalization of learning, be it new speakers, new contexts, or new sounds (Lively et al., 1993; Sadakata & McQueen, 2013, 2014; Wang et al., 1999). More importantly, our results demonstrated that the generalization of learning in the multiple talker group arises from between-talker variability in the training materials and not from the acoustic or contextual variability, because contextual variability was strictly matched between the two training conditions. These results are consistent with those of a recent study by Brosseau-Lapré and colleagues (2013) showing that acoustic variability helps to generalize learning only when speakers are trained with multiple talkers; single talker training hinders the establishment of abstract sound categories.

Given our dependent measures, significant training effects in the multiple talker group revealed that participants' productions were acoustically closer to the mean of the trained vowel category and distributed tighter around this mean after the training. In other words, this might suggest that vowel categories became more prototypical or categorical. Graphic representations of the participants' vowel productions in the acoustic space (see Appendix S4) support this interpretation by showing that, after training, vowel categories in the multiple talker group became more distinguishable from each other and closer to the target spaces. No study, to date, has specifically addressed the mechanism underlying the relationship between speech sound production accuracy and its variability. Nevertheless, the results of previous studies showing (a) high correlations between production accuracy and variability in L2 speakers (Kartushina & Frauenfelder, 2014) and novice learners (Kartushina et al., 2015), (b) a decrease in variability as a result of improvements in production and perception accuracy (Kartushina et al., 2015, 2016), and (c) a relationship between these two production measures and categorical perception in the L1 (Franken et al., 2017; Perkell et al., 2004) suggest that gains in accuracy and compactness in the multiple talker group might be mediated by more accurate sound perception.

The interpretation that multiple talker production training enhances categorical perception of the trained sounds is also supported by theories of L1 sound production learning and control, which assume a relationship between production and perception. For instance, the DIVA model claims that learners first establish auditory targets for sounds and then tune their production (mainly through auditory feedback) to reach these auditory targets (Guenther, 1994). Therefore, this model assumes that the establishment of sound categories in perception precedes their stabilization in production. Previous perception studies have shown that multiple talker training encourages categorical processing of sounds (Brosseau-Lapré et al., 2013; Lively et al., 1993; Sadakata & McQueen, 2013). Such processing might lead to an establishment of more accurate auditory targets, which, in turn, might tune the production, leading to more compact/stable articulatory realizations. Other production training studies have shown that training-related improvements in production were accompanied by a reduction in production variability and also led to slight improvements in the perception of the trained sounds (Kartushina et al., 2015). The results of the current study suggest that hearing multiple talkers during production training encourages participants to disregard irrelevant acoustic information (here, variability in pitch or between-talker variability) and to focus on the relevant phonetic features (here, F1 and F2), which allows for the two trained categories to be distinguished. Such focused processing of sounds may have boosted the establishment of (prototypical) vowel representations (in the F1/F2 space), which, in turn, tuned L2 vowel production so it became more accurate and stable, in line with the predictions of the DIVA model for L1 sound learning (see Guenther, 1994). However, as stated previously, data on participants' perception are needed to fully support this interpretation.

Recently, Simmonds (2015) proposed that increased variability in sound realization during L2 learning (i.e., when learners try out and explore different ways to produce L2 sounds before stabilizing on the "best" exemplar) might support accurate L2 pronunciation. Based on this account, we performed additional analyses on vowel compactness during training. Those analyses revealed no differences in vowel compactness during training between the two trained groups, suggesting that both groups showed similar amounts of variability in their articulatory behavior, independent of the between-talker input variability. Therefore, our results showing more stable production in the multiple talker group cannot be attributed to greater variability in articulatory output during training (as proposed by Simmonds), but rather to the presence of multiple talker variability in (perceptual) input (i.e., variability in mean F0, including changes in F0 and timber), which likely encouraged abstract, phonemic processing of sounds (Sadakata & McQueen, 2013).

<A>Limitations and Future Directions

<TXT>

Here we propose some improvements for studying L2 production training, both in terms of the tasks and measures used to evaluate its accuracy and suggest several specific future directions. In this study, we used a relatively new measure of stability of articulatory realizations—the compactness of vowels in the acoustic space. Most previous training studies have measured the effects of training by subtracting the mean production accuracy before training from the production accuracy after training; thus, the results show whether, on average, participants produced targeted more accurately after the training. The measure of compactness, on the other hand, provides a valuable additional piece of information about production learning because it shows whether participants have stabilized their articulatory realizations and have constructed more compact vowel categories after training. However, a more direct measure could be used in future studies, for example, sonographic imaging, which can provide more precise and dynamic information on the training-related changes in tongue height and shape.

In this study, we assessed the effectiveness of training immediately after it took place using a vowel repetition task. However, it is important to address the longevity of the learning outcomes and their generalization to the production of more complex (untrained) structures (e.g., words). Previous perception studies have shown that high variability training encourages long-term retention of new phonemic categories. For instance, six months after training, participants showed no significant decrease in their perception accuracy for the trained vowels produced by familiar and unfamiliar talkers, compared to posttraining performance (Lively, Pisoni, Yamada, Tohkura, & Yamada, 1994; Wang et al., 1999). Other perception studies have shown that only training incorporating talker variability supports generalization to untrained sounds, such as timing-based discrimination of fricatives transfers to stops and affricates (Sadakata & McQueen, 2013). Future research needs to explore the longevity of learning outcomes in sound production training and the generalization of learning to untrained complex structures, including the use of the trained sound in words.

In this study, there were no overall differences in the production accuracy between the two vowels. Given that Spanish /e/ is acoustically intermediate between French /e/ and /ɛ/, this result suggests, in line with other studies (e.g., Kartushina & Frauenfelder, 2014), that Spanish speakers used their native /e/ category to produce both French vowels. There were, however, differences between conditions, with both vowels being produced better in the unfamiliar than the familiar voice sets.[9] These differences could arise due to the distribution of acoustic realizations of the two vowels relative to the prototypical Spanish /e/ vowel (see Appendix S4), such that French /e/ and /ɛ/ could become acoustically closer to the Spanish /e/ in the productions by an unfamiliar speaker than by a familiar one. These potential differences in the vowel position in the acoustic space between familiar and unfamiliar speakers likely led to lower distance scores for the vowels produced in the unfamiliar than in the familiar voice sets. Future studies need to address, specifically, the role of the acoustic distance between the target vowels and learners' realizations of these vowels as well as the role of vowel compactness in learners' output in the development of L2 speech production. Also, in the same vein, it would be important to compare training conditions which are similar in the number of speakers used for training but differ in the amount of acoustic variability (compactness) between them. Indeed, between-vowel distance (and vowel compactness) might interact with the talker variability effect (Brosseau-Lapré et al., 2013).

In the current study, we opted for maximal control over talker-relevant acoustic variation between the two training groups, which resulted in two times less acoustic variability in pitch in the single talker group than in the multiple talker group (see Table 1), with length and pitch contours being similar. However, research on word learning in young infants, for instance, involving recognition of similar sounding objects (e.g., /buk/-/puk/), has shown that single talker settings can be as effective as multiple talker settings, as long as single talker stimuli present high talker-relevant acoustic variation, for example, variation in pitch amplitude, pitch contour, and length (Galle, Apfelbaum, & McMurray, 2015). Future research thus needs to address whether single talker training that incorporates high acoustic variability (in talker-relevant dimensions) leads to better learning outcomes in terms of production stability and the generalization of learning to a new speaker. Finally, the results of this study suggest that L2 learning is more efficient (regarding the quality of sound production) if learners are exposed to multiple talkers. Such a favorable learning environment could be easily implemented in classrooms, with L2 teachers using teaching material (recordings, videos, etc.) with substantial talker variability. This assumption, however, should be validated through research in L2 classrooms.

<A>Conclusion

<TXT>

The current findings revealed that training with a stimulus set featuring a single speaker leads to improvements in production accuracy by enhancing acoustic processing of nonnative sounds, with learners' productions becoming, on average, acoustically closer to the target (native speaker) acoustic values. However, single speaker training did not seem to promote stability in the newly learned articulatory realizations of vowels in the sense that vowel categories did not become more compact after training. Furthermore, single speaker training (i.e., training which is low in between-talker variability) did not appear to contribute to the development of abstract vowel categories of nonnative sounds, because learners exposed to single talker training failed to generalize learning to an unfamiliar speaker. Training with multiple speakers, on the other hand, likely contributed to the establishment of accurate and stable articulatory representations for nonnative sounds. These representations also appeared to be sufficiently abstract to allow for successful generalization of learning to novel speakers.

Final revised version accepted 5 June 2018

<A>Notes

[1] Acoustic variability can also refer to variability in amplitude (e.g., loud, whisper), speech style, and register (childlike, excited), among others. Although these sources of variability are not examined here, they have been shown to have no effect on lexical acquisition in L2 learning, likely because these sources of variability do not affect relevant phonetic properties of speech sounds (Barcroft, 2001). In the current study, the term acoustic variability refers to the acoustic dispersion of vowels in the F1/F2 space. Talker variability refers to differences in pitch (F0).

[2] Yet it is possible that the amount of acoustic variability for a given sound produced by one speaker (e.g., in a case of a “sloppy” speaker) is larger than that produced by multiple speakers.

[3] In order to avoid gender differences between auditory input (we recorded native French female speakers) and articulatory output, only female participants were recruited to take part in the experiment.

[4] We carefully controlled the perceptual quality of the synthesized stimuli: If a synthesized sound had audible unnatural tone or clicks at the end, we replaced this sound and resynthesized the source with another filter until we were satisfied with the quality.

[6] The decrease in the production of /e/ cannot be attributed to its production difficulty. At pretraining, each of these three participants produced /e/ better than the group did, on average; however, all of them produced /ɛ/ considerably less accurately than /e/.

[7] Related to this point, some idiosyncratic features of the familiar voice could potentially account for some performance differences in the experiment, with this voice being less intelligible/more difficult to learn. Additional mixed-effects regression analyses of training data in the multiple talker group with the fixed structure including block, talker, and their interaction, and the random structure including by-subject random slopes adjusted to the effects of talker, revealed a significant effect of block, $F(9, 24293) = 90.00$, $p < .00001$, but no interaction between block and talker, $F(36, 24293) = 1.17$, $p = .21$. These results suggest that participants improved their production gradually from block to block irrespective of the heard talker. Also, the results revealed no differences in production accuracy during training between stimuli produced after familiar talker (T3) and other talkers (T1–T2 and T4–T5), all $p > .70$ (see Appendix S5 in the Supporting Information online for an illustration of these results).

[8] The lack of generalization to an unfamiliar speaker in the single talker group cannot be attributed to larger variance in participants' performance, compared to the performance of the multiple talker group (i.e., Levene's test comparing the measure of gains was not significant). In addition, the number of participants who did not benefit from training (negative values) was similar in both training groups (for /e/, $n = 3$ in both groups; for /ɛ/, $n = 4$ in the single talker group and $n = 3$ in the multiple talker group; see Appendix S3). This observation suggests that both training methods helped participants to generalize learning to unfamiliar speakers. However, after training, each individual gained less in production accuracy in the single talker than in the multiple talker group.

9 Greater improvements in the production of vowels for the unfamiliar than for the familiar voice set, shown in the multiple talker group (Figure 4), were not expected and suggest that the phonetic properties in the speech by an unfamiliar speaker, somehow, allowed participants to produce the trained vowels, after training, even closer to the respective (trained) mean acoustic values. A post hoc comparison revealed that the distribution of 18 tokens (within each vowel) used in unfamiliar voice set overlapped considerably with the respective vowel target spaces used in visual feedback (ellipses showing 0.5 standard deviations from the mean). In the familiar voice set, on the other hand, vowel (formant) spaces for /e/ and /ɛ/ overlapped only moderately with those used for visual feedback in training. This observation, which must be addressed in future research, suggests that participants demonstrated greater benefits of training when talker's vowels matched acoustically the "prototypical" target vowel spaces shown in training.

<A>References

<REF>Antoniou, M., & Wong, P. C. M. (2015). Poor phonetic perceivers are affected by cognitive load when resolving talker variability. *The Journal of the Acoustical Society of America*, *138*, 571–574. https://doi.org/10.1121/1.4923362

<REF>Baese-Berk, M. M., & Samuel, A. G. (2016). Listeners beware: Speech production may be bad for learning speech sounds. *Journal of Memory and Language, 89,* 23–36. https://doi.org/10.1016/j.jml.2015.10.008

<REF>Barcroft, J. (2001). Acoustic Variation and Lexical Acquisition. *Language Learning, 51*, 563–590. https://doi.org/10.1111/0023-8333.00168

<REF>Barcroft, J., & Sommers, M. S. (2005). Effects of acoustic variability on second language vocabulary learning. *Studies in Second Language Acquisition*, *27*, 387–414. https://doi.org/10.1017/S0272263105050175

<**REF**>Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*, 255–278. https://doi.org/10.1016/j.jml.2012.11.001

<**REF**>Bates, D. M., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*, 1–40.

<**REF**>Best, C. T. (1995). A Direct Realist View of Cross-Language Speech Perception. In W. Strange (Ed.), *Speech perception and linguistic experience: Theoretical and methodological issues* (pp. 171–204). Baltimore: York Press.

<**REF**>Boersma, P., & Weenink, D. (2010). *{P}raat: doing phonetics by computer [Computer program]*. *Version 5.2*. Retrieved from http://www.praat.org

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

<**REF**>Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision*, *10*, 433–436.

<**REF**>Brosseau-Lapré, F., Rvachew, S., Clayards, M., & Dickson, D. (2013). Stimulus variability and perceptual learning of nonnative vowel categories. *Applied Psycholinguistics*, *34*, 419–441. https://doi.org/10.1017/S0142716411000750

<**REF**>Chang, C. B., & Bowles, A. R. (2015). Context effects on second-language learning of tonal contrasts. *The Journal of the Acoustical Society of America*, *138*, 3703–3716. https://doi.org/10.1121/1.4937612

<**REF**>Chládková, K., Escudero, P., & Boersma, P. (2011). Context-specific acoustic differences between Peruvian and Iberian Spanish vowels. *The Journal of the Acoustical Society of America*, *130*, 416–428. https://doi.org/10.1121/1.3592242

<REF>Evans, B. G., & Martin-Alvarez, L. (2016). Age-related differences in second-language learning? A comparison of high and low variability perceptual training for the acquisition of English /i/-/I/ by Spanish adults and children. Presented at the International Symposium on the Acquisition of Second Language Speech, New Sounds, Aarhus, Denmark.

<REF>Flege, J. E. (1995). Second language speech learning Theory, findings, and problems. In Strange, Winifred (Ed.), *Speech Perception and Linguistic Experience: Issues in Cross-Language Research* (pp. 233–277). Timonium, MD: York Press.

<REF>Flege, J. E. (1999). Age of learning and second-Ianguage speech. In D. Birdsong (Ed.), *Second Language Acquisition and the Critical Period Hypothesis* (pp. 101–132). Hillsdale, NJ: Lawrence Erlbaum.

<REF>Flege, J. E., Bohn, O.-S., & Jang, S. (1997). Effects of experience on non-native speakers' production and perception of English vowels. *Journal of Phonetics*, *25*, 437–470. https://doi.org/doi.org/10.1006/jpho.1997.0052

<REF>Flege, J. E., MacKay, I. R., & Meador, D. (1999). Native Italian speakers' perception and production of English vowels. *The Journal of the Acoustical Society of America*, *106*, 2973–2987. https://doi.org/doi.org/10.1121/1.428116

<REF>Flege, J. E., & Schmidt, A. M. (1995). Native speakers of Spanish show rate-dependent processing of English stop consonants. *Phonetica*, *52*, 90–111. https://doi.org/10.1159/000262062

<REF>Forster, K., & Forster, J. (2003). DMDX: A Windows Display Program with Millisecond Accuracy. *Behavior Research Methods Instruments and Computers*, *35*, 116–124. https://doi.org/doi.org/10.3758/BF03195503

<REF>Franken, M. K., Acheson, D. J., McQueen, J. M., Eisner, F., & Hagoort, P. (2017). Individual variability as a window on production-perception interactions in speech motor control. *The Journal of the Acoustical Society of America*, *142*, 2007–2018. https://doi.org/10.1121/1.5006899

<REF>Galle, M. E., Apfelbaum, K. S., & McMurray, B. (2015). The Role of Single Talker Acoustic Variation in Early Word Learning. *Language Learning and Development : The Official Journal of the Society for Language Development*, *11*, 66–79. https://doi.org/10.1080/15475441.2014.895249

<REF>Giannakopoulou, A., Brown, H., Clayards, M., & Wonnacott, E. (2017). High or low? Comparing high and low-variability phonetic training in adult and child second language learners. *PeerJ*, *5*. https://doi.org/10.7717/peerj.3209

<REF>Guenther, F. H. (1994). A neural network model of speech acquisition and motor equivalent speech production. *Biological Cybernetics*, *72*, 43–53. https://doi.org/10.1007/BF00206237

<REF>Hao, Y.-C., & de Jong, K. (2016). Imitation of second language sounds in relation to L2 perception and production. *Journal of Phonetics*, *54*, 151–168. https://doi.org/10.1016/j.wocn.2015.10.003

<REF>Hwang, H., & Lee, H.-Y. (2015). The effect of high variability phonetic training on the production of English vowels and consonants. In *Proceedings of the 18th International Congress of Phonetic Sciences, Glasgow, UK, edited by The Scottish Consortium for ICPhS* (pp. 1041–1045). Retrieved from https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2015/Papers/ICPHS0466.pdf

<REF>Ingram, J. C., & Park, S.-G. (1997). Cross-language vowel perception and production by Japanese and Korean learners of English. *Journal of Phonetics*, *25*, 343–370. https://doi.org/10.1006/jpho.1997.0048

<REF>Iverson, P., & Evans, B. G. (2009). Learning English vowels with different first-language vowel systems II: Auditory training for native Spanish and German speakers. *The Journal of the Acoustical Society of America*, *126*, 866–877. https://doi.org/10.1121/1.3148196

<REF>Jügler, J., Zimmerer, F., Möbius, B., & Draxler, C. (2015). The effect of high-variability training on the perception and production of French stops by German native speakers. In *INTERSPEECH* (pp. 806–810). Retrieved from http://www.coli.uni-saarland.de/~juegler/Publications/juegler_etal_is2015.pdf

<REF>Kartushina, N., & Frauenfelder, U. H. (2014). On the effects of L2 perception and of individual differences in L1 production on L2 pronunciation. *Frontiers in Psychology*, *5*, 1246. https://doi.org/10.3389/fpsyg.2014.01246

<REF>Kartushina, N., Hervais-Adelman, A., Frauenfelder, U. H., & Golestani, N. (2015). The effect of phonetic production training with visual feedback on the perception and production of foreign speech sounds. *The Journal of the Acoustical Society of America*, *138*, 817–832. https://doi.org/10.1121/1.4926561

<REF>Kartushina, N., Hervais-Adelman, A., Frauenfelder, U. H., & Golestani, N. (2016). Mutual influences between native and non-native vowels in production: Evidence from short-term visual articulatory feedback training. *Journal of Phonetics*, *57*, 21–39. https://doi.org/10.1016/j.wocn.2016.05.001

<REF>Kingston, J. (2003). Learning foreign vowels. *Language and Speech*, *46*, 295–348. https://doi.org/10.1177/00238309030460020201

<REF>Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*, 1–26. https://doi.org/10.18637/jss.v082.i13

<REF>Lammert, A., Proctor, M., & Narayanan, S. (2013). Interspeaker variability in hard palate morphology and vowel production. *Journal of Speech, Language, and Hearing Research: JSLHR*, *56*, S1924-1933. https://doi.org/10.1044/1092-4388(2013/12-0211)

<REF>Lenth, R., V. (2016). Least-Squares Means: The R Package lsmeans. *Journal of Statistical Software*, *69*, 1–33. https://doi.org/doi:10.18637/jss.v069.i01

<REF>Lively, S. E., Logan, J. S., & Pisoni, D. B. (1993). Training Japanese listeners to identify English /r/ and /l/. II: The role of phonetic environment and talker variability in learning new perceptual categories. *The Journal of the Acoustical Society of America*, *94*, 1242–1255. https://doi.org/10.1121/1.408177

<REF>Lively, S. E., Pisoni, D. B., Yamada, R. A., Tohkura, Y., & Yamada, T. (1994). Training Japanese listeners to identify English/r/and/l/. III. Long-term retention of new phonetic categories. *The Journal of the Acoustical Society of America*, *96*, 2076–2087. https://doi.org/10.1121/1.410149

<REF>Logan, J. S., Lively, S. E., & Pisoni, D. B. (1991). Training Japanese listeners to identify English /r/ and /l/: a first report. *The Journal of the Acoustical Society of America*, *89*, 874–886.

<REF>Loizou, P. (1998). *COLEA: A MATLAB software tool for speech analysis*. Dallas, TX. Retrieved from http://ecs.utdallas.edu/loizou/speech/colea.htm

MATLAB Release 2014b, The MathWorks, Inc., Natick, Massachusetts, United States

<REF>Ménard, L., Schwartz, J.-L., Boë, L.-J., & Aubin, J. (2007). Articulatory–acoustic relationships during vocal tract growth for French vowels: Analysis of real data and simulations with an articulatory model. *Journal of Phonetics*, *35*, 1–19.

<REF>Ménard, L., Schwartz, J.-L., Boë, L.-J., Kandel, S., & Vallée, N. (2002). Auditory normalization of French vowels synthesized by an articulatory model simulating growth from birth to adulthood. *The Journal of the Acoustical Society of America*, *111*, 1892–1905. https://doi.org/10.1121/1.145946

<**REF**>Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, *4*, 133–142. https://doi.org/10.1111/j.2041-210x.2012.00261.x

<**REF**>Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: transforming numbers into movies. *Spatial Vision, 10,* 437–442.

<**REF**>Perkell, J. S., Guenther, F. H., Lane, H., Matthies, M. L., Stockmann, E., Tiede, M., & Zandipour, M. (2004). The distinctness of speakers' productions of vowel contrasts is related to their discrimination of the contrasts. *The Journal of the Acoustical Society of America*, *116*, 2338. https://doi.org/10.1121/1.1787524

<**REF**>Perrachione, T. K., Lee, J., Ha, L. Y., & Wong, P. C. (2011). Learning a novel phonological contrast depends on interactions between individual differences and training paradigm design. *The Journal of the Acoustical Society of America*, *130,* 461–472. https://doi.org/10.1121/1.3593366

<**REF**>Peterson, G. E., & Barney, H. L. (1952). Control Methods Used in a Study of the Vowels. *The Journal of the Acoustical Society of America*, *24*, 175. https://doi.org/10.1121/1.1906875

<**REF**>Sadakata, M., & McQueen, J. M. (2013). High stimulus variability in nonnative speech learning supports formation of abstract categories: Evidence from Japanese geminates. *The Journal of the Acoustical Society of America*, *134*, 1324–1335. https://doi.org/10.1121/1.4812767

<**REF**>Sadakata, M., & McQueen, J. M. (2014). Individual aptitude in Mandarin lexical tone perception predicts effectiveness of high-variability training. *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.01318

<REF>Steinlen, A. K. (2005). *The Influence of Consonants on Native and Non-native Vowel Production: A Cross-linguistic Study*. Gunter Narr Verlag.

<REF>Wade, T., Jongman, A., & Sereno, J. (2007). Effects of Acoustic Variability in the Perceptual Learning of Non-Native-Accented Speech Sounds. *Phonetica*, *64*, 122–144. https://doi.org/10.1159/000107913

<REF>Wang, Y., Spence, M. M., Jongman, A., & Sereno, J. A. (1999). Training American listeners to perceive Mandarin tones. *The Journal of the Acoustical Society of America*, *106*, 3649–3658. https://doi.org/10.1121/1.428217

<REF>Wayland, R. P., & Guion, S. G. (2004). Training English and Chinese Listeners to Perceive Thai Tones: A Preliminary Report. *Language Learning*, *54*, 681–712. https://doi.org/10.1111/j.1467-9922.2004.00283.x

<REF>Wong, J. W. S. (2013). The effects of perceptual and/or productive training on the perception and production of English vowels/ɪ/and/iː/by Cantonese ESL learners. In *Proceedings of International Congress of Speech Science* (pp. 2113–2117). Lyon, France.

<A>Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher’s website:

Appendix S1. Visual Illustration of French /e/ and /ɛ/ Vowels.

Appendix S2. Vowel Spaces Used in Feedback.

Appendix S3. Individual Training-Related Gains in Production Accuracy.

Appendix S4. Vowel Spaces for French /e/ and /ɛ/ Vowels Before and After Training.

Appendix S5. Production Accuracy During Training.

Table 1 Mean and standard deviation values of formant frequencies (F0, F1, F2) for vowels used in the multiple talker and single talker training conditions

	Multiple talkers		Single talker		Comparison	
	M	*SD*	*M*	*SD*	*t* test	Bartlett test
Vowel /ɛ/						
F0 (Hz)	196	36	168	11	*p* < .001	*p* < .0001
F1 (Hz)	607	63	606	65	*p* > .10	*p* > .10
F2 (Hz)	2129	82	2140	72	*p* > .10	*p* > .10
Vowel /e/						
F0 (Hz)	203	37	192	19	*p* < .01	*p* < .0001
F1 (Hz)	440	33	432	34	*p* > .10	*p* > .10
F2 (Hz)	2357	121	2368	117	*p* > .10	*p* > .10

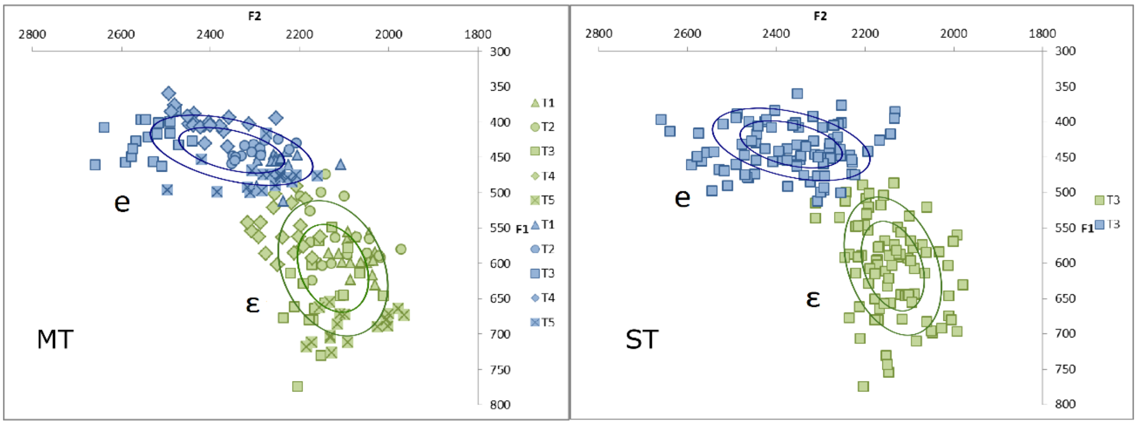

Figure 1 The French /e/ (mid-close) and /ɛ/ (mid-open) vowel tokens used in the multiple talker (MT) and single talker (ST) training conditions, plotted in the F1/F2 space (90 tokens per condition). Different markers designate talkers (T1–T5), and ellipses demarcate areas within one (inner) and two (outer) standard deviations from the mean formant values.

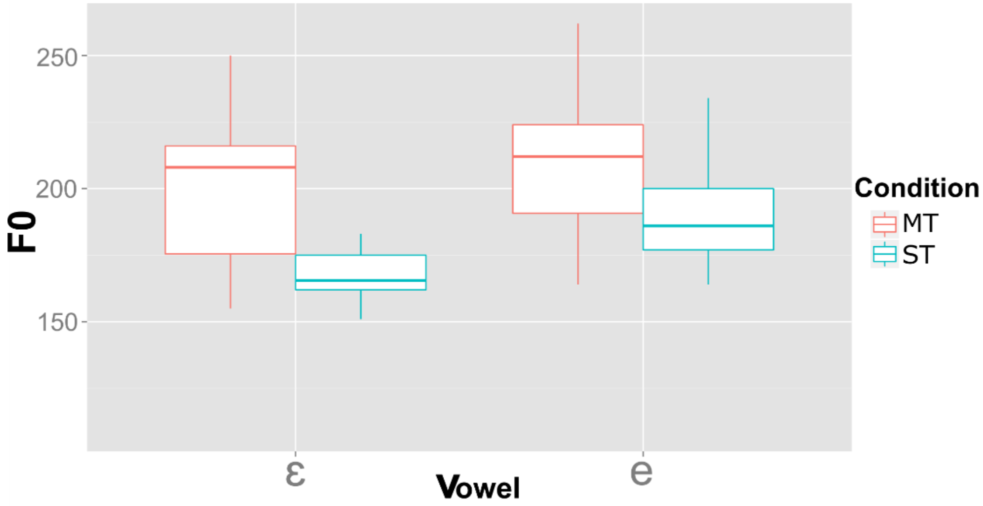

Figure 2 Boxplot of F0 (Hz) values for each target vowel in the multiple talker (MT) and single talker (ST) conditions.

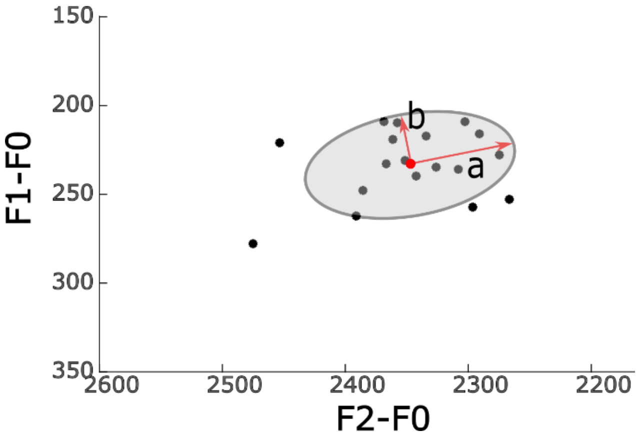

Figure 3 A schematic illustration for the computation of the vowel compactness score (CS). The CS is the area within the grey ellipse corresponding to one standard deviation from the mean F1 and F2 formant values (in Hz, adjusted for F0), with individual dots representing 18 vowel tokens and the red dot designating the mean F1–F0/F2–F0 value.

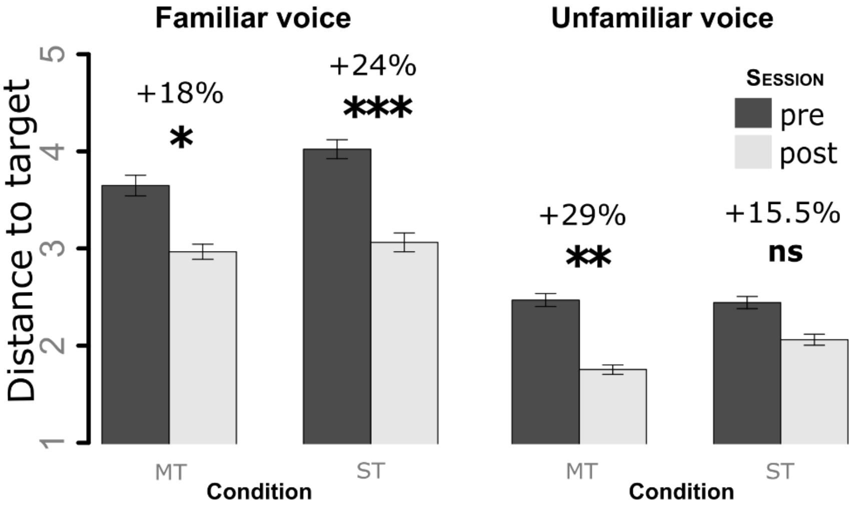

Figure 4 Mean production accuracy (distance score to target) for multiple talker (MT) and single talker (ST) conditions before and after training for both trained vowels (combined) in familiar and unfamiliar voice sets. Low distance scores correspond to higher production accuracy (i.e., closer to the target). Error bars represent ±1 SE of the mean, with extent of improvement indicated in percent values (**p* < .05, ***p* < .01, ****p* < .001).

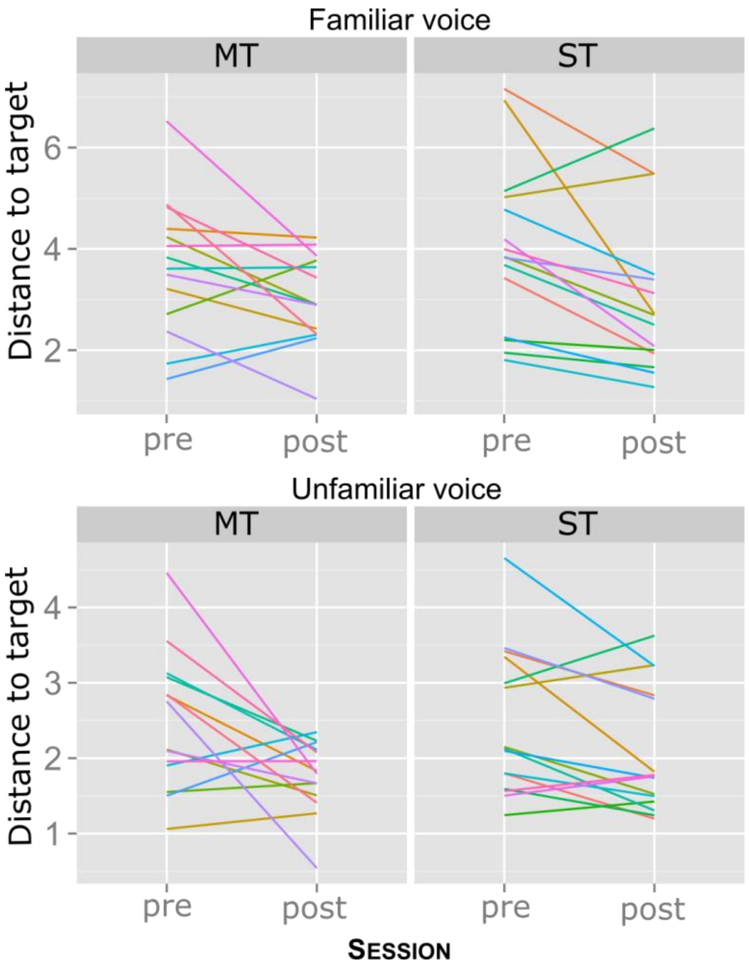

Figure 5 Individual learning slopes in multiple talker (MT) and single talker (ST) training conditions for familiar and unfamiliar voices before and after training. Low distance scores correspond to higher production accuracy (i.e., closer to the target). Colored lines designate individual participants.

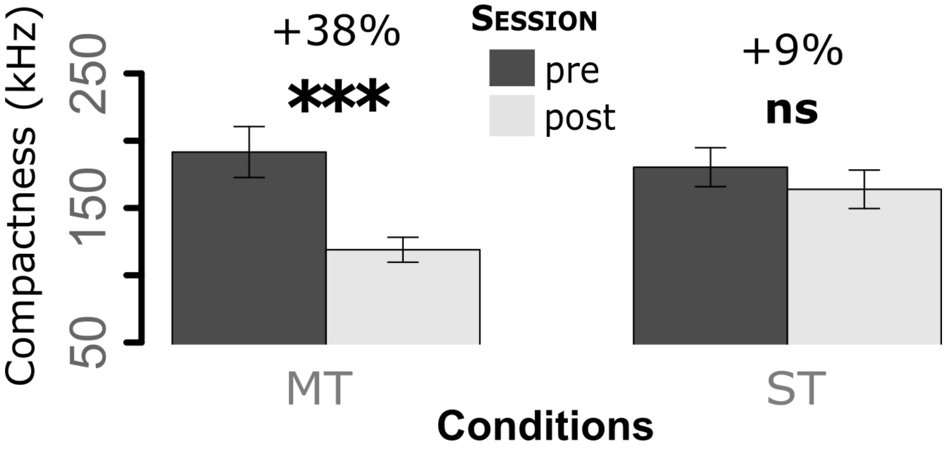

Figure 6 Mean compactness score for multiple talker (MT) and singe talker (ST) training conditions at the pre- and posttraining tests for both trained vowels (combined). Error bars represent ±1 SE of the mean, with extent of improvement indicated in percent values (*p* < .001).