 doi:10.20944/preprints202211.0037.v1

Article

Cognitive Factors in Nonnative Phonetic Learning: Impacts of Working Memory and Selective Attention on the Benefits and Costs of Talker Variability

Xiaojuan Zhang [a, b], Bing Cheng [a,*], Yang Zhang [b,*]

[a] English Department & Language and Cognitive Neuroscience Lab, School of Foreign Studies, Xi'an Jiaotong University, 710049, China

[b] Department of Speech-Language-Hearing Sciences & Center for Neurobehavioral Development, University of Minnesota, Minneapolis, MN 55455, USA

* Corresponding authors: Bing Cheng, PhD, Professor, bch@mail.xjtu.edu.cn (B. Cheng); Yang Zhang, PhD, Professor, zhanglab@umn.edu (Y. Zhang); Email addresses for co-authors: Xiaojuan Zhang, PhD candidate, xiaojuan.110577@stu.xjtu.edu.cn

Abstract: Talker variability has been reported to facilitate generalization and retention of speech learning, but is also shown to place demands on cognitive resources. Our recent study provided evidence that phonetically-irrelevant acoustic variability in single-talker (ST) speech is sufficient to induce equivalent amounts of learning to the use of multiple-talker (MT) training. This study is a follow-up contrasting MT versus ST training with varying degrees of temporal exaggeration to examine how cognitive measures of individual learners may influence the role of input variability in immediate learning and long-term retention. Native Chinese-speaking adults were trained on the English /i/-/ɪ/ contrast. We assessed the trainees' working memory and selective attention before training. Trained participants showed retention of more native-like cue weighting in both perception and production regardless of talker variability condition. The ST training group showed long-term benefit in word identification, whereas the MT training group did not retain the improvement. The results demonstrate the role of phonetically-irrelevant variability in robust speech learning and modulatory functions of nonlinguistic working memory and selective attention, highlighting the necessity to consider the interaction between input characteristics, task difficulty, and individual differences in cognitive abilities in assessing learning outcomes.

Keywords: non-native speech learning; talker variability; phonetically-irrelevant variability; long-term retention; cognitive abilities

Introduction

Learning to perceive and produce nonnative or second language (L2) speech sounds can be challenging, with problems persisting even after years of learning in an immersion environment (e.g., Flege et al., 1997; Flege & MacKay, 2004). One line of research has focused on determining laboratory training conditions that support and facilitate this learning process for adults with a variety of target sounds (e.g., Cheng et al., 2019; Iverson & Evans, 2009; Kondaurova & Francis, 2010, for vowels; Iverson et al., 2005; Logan et al., 1991, for liquids; Pruitt et al., 2006, for stops; Jamieson & Morosan, 1989; Sadakata & Mcqueen, 2013, for fricatives; Y. Wang et al., 1999, for suprasegmentals). Importantly, generalization and long-term retention of trained knowledge are of paramount concern for determining successes and failures of training. In this regard, one prime example is the high variability phonetic training (HVPT) protocol, which employs natural speech spoken by multiple talkers in various phonetic contexts. It has been reported to show the benefit of generalization to novel stimuli and talkers (Lively et al., 1993, 1994; Logan et al., 1991) and retention of learning in long-term memory (Bradlow et al., 1999; Lively et al., 1994) as well as transfer of perceptual learning to production (Bradlow et al., 1999).

 doi:10.20944/preprints202211.0037.v1

Despite the efficacy of HVPT, the role of input variability in L2 speech learning is still controversial and underappreciated. As most HVPT studies did not attempt to isolate or determine the specific source of variability which may lead to robust learning of L2 speech sounds, it remains an open question why some studies found an advantage of talker variability (Brosseau-Lapré et al., 2013; Deng et al., 2018; Hardison, 2003; Kartushina & Martin, 2019; Lively et al., 1993; Perrachione et al., 2011; Uchihara et al., 2022), while others did not (Brekelmans et al., 2022; Dong et al., 2019; Giannakopoulou et al., 2017; Wiener et al., 2020). Our recent training studies (Cheng et al., 2019; Zhang et al., 2021) provided tentative evidence that generalization can be induced by acoustic variability in phonetically irrelevant (or secondary) cues that may reside in but are not limited to talker variability. The question remains as to whether acoustic variability in phonetically irrelevant cues is conducive to long-term retention of learning.

One complication here is that increased variability may come at a processing cost (Antoniou & Wong, 2015; Fuhrmeister & Myers, 2017, 2020; Luthra et al., 2021; Sadakata & Mcqueen, 2014; Saltzman et al., 2021). Across domains of visual perception, auditory perception, motor learning, language, inductive reasoning, problem solving, and computational modeling, a general observation is that increased input variability may come at a cost of initially hindering learning but often show subsequent benefits in generalization (Raviv et al., 2022). Although a significant amount of work has been devoted to understanding the cognitive and neural mechanisms supporting speech learning (e.g., De Diego-Balaguer & Lopez-Barroso, 2010; Zhang et al., 2009), much less work has considered how individual learners' cognitive abilities may influence the efficacy of speech training in terms of perceptual generalization, transfer of learning to production and long-term retention, as perceptual learning does not solely depend on the nature of exposure, but also learner ability to cope with stimulus variability.

The Role of Input Variability in L2 Speech Learning

Variability in speech used to be considered as "noise", an unwanted property that obscures meaningful linguistic information. However, input variability in talker and phonetic context can facilitate speech categorization in terms of generalization and retention (Bradlow et al., 1999; Lively et al., 1993, 1994; Logan et al., 1991). In an early attempt, Strange and Dittmann (1984) used a continuum of synthesized speech to train Japanese learners of the English /l/-/r/ contrast. They found that the learners significantly improved in discrimination and identification of the contrast, but notably, the training effect did not generalize to natural speech. Of significance, by using variable natural speech as training input, Logan et al. (1991) found that the learners improved successfully on untrained items produced by an untrained talker. In their follow-up study, Lively et al. (1993) examined the effect of a more specific type of input variability by contrasting multiple-talker and single-talker conditions and found only the multiple-talker group successfully generalized to new phonetic contexts and a new talker. The researchers suggested that single-talker training led to stimulus-specific learning, whereas multiple-talker training facilitated robust categorization. Their subsequent studies further demonstrated that this learning effect could be retained several months after training (Bradlow et al., 1999; Lively et al., 1994).

If the benefit of talker variability to generalization and retention is robust and reliable, it has significant implications for implementing this training technique as a viable pedagogical tool. However, the essential role of talker variability in producing generalization is not always supported. First, only six participants were tested in both studies of Logan et al. (1991) and Lively et al. (1993). Second, the benefit to generalization was only described and not statistically tested as the two experiments were analyzed separately. Third, these two studies employed only one single male talker in the generalization test. Moreover, follow-up studies did not uniformly demonstrate the talker variability benefit. In some cases, studies demonstrated that only exposure to multiple-talker speech led to

 doi:10.20944/preprints202211.0037.v1

generalization outcomes (e.g., Kartushina & Martin, 2019). Other studies showed that single-talker training also engendered generalization, which could be further enhanced by greater talker variability (Perrachione et al., 2011; Wong, 2014). There have also been reports that single-talker versus multiple-talker training produced equivalent amounts of generalization (Brekelmans et al., 2022; Dong et al., 2019; Giannakopoulou et al., 2017). The researchers cautiously interpreted that these mixed results might be due to learner differences or other sources of variability introduced by training input or setting.

In comparison, long-term retention has been examined in relatively fewer HVPT studies. Studies showed that training improvement sustained for a time period varying from two weeks to six months (Bradlow et al., 1999; Carlet, 2017; Flege, 1995b; Iverson & Evans, 2009; Lively et al., 1994; Nishi & Kewley-Port, 2007; Thomson, 2012; Wang & Munro, 2004; Wang et al., 1999). However, very few studies have directly contrasted multiple-talker and single-talker training conditions, and the evidence in support of greater advantage for talker variability is inconclusive. One study was conducted by Macdonald (2012) who found that only multiple-talker training effects were retained one month later when training native English-speaking learners on French vowel contrasts (/u/ vs. /y/ and /ɑ̃/ vs. /ɔ̃/). Another study is Silpachai (2020) who trained English speakers who have limited tonal language experience to perceive Mandarin tones (Tones 1-4). The results showed that the multiple-talker group retained their learning of Tone 2, 3, and 4 six months after training, while the single-talker group also retained the learning of Tone 3 and 4.

Intriguingly, although the talker variability benefit has been reported in various language learning paradigms, the extant literature has interpreted this benefit somewhat differently. Researchers exploring adult L2 phonetic learning have suggested that talker variability is conducive to forming generalized representations that include only phonetically-relevant cues and exclude irrelevant talker identity cues (e.g., Lively et al., 1993, 1994; Logan et al., 1991). On the other hand, researchers focusing on adult L2 lexical learning have argued that talker-specific cues are also incorporated into representations to form more "associative hooks" and thus more robust representations for target words (e.g., Barcroft & Sommers, 2005, 2014). By contrast, developmental researchers have assumed that varying talker cues prevent consistent talker-specific cues from being associated with the object being learned at the cost of phonetically relevant cues (Apfelbaum & McMurray, 2011; Quam & Creel, 2021; Rost & McMurray, 2009, 2010). This view is compatible with a broader class of linguistic learning models in which linguistically relevant and irrelevant cues compete and generalization occurs via a discriminative process that dissociates the irrelevant features (e.g., Ramscar et al., 2010).

If dissociation of phonetically-irrelevant cues is the key advantage of talker variability in speech categorization, it is plausible that other sources of variability in phonetically-irrelevant cues (or secondary cues to the target contrast) in single-talker speech can also produce generalization and retention outcomes. This hypothesis may help account for the mixed results pertaining to the advantage of multiple-talker speech in the literature. Support comes from the work of lexical learning. For example, Barcroft and Sommers (2005) contrasted three conditions of talker variability and found that the learners improved systematically, from low- to moderate- and from moderate- to high-variability conditions. Crucially, the same pattern of results was observed when the talker was held constant and variability in speaking style was similarly manipulated. Their follow-up studies suggest that this variability benefit seems to be found only when the variability was phonetically-relevant (to learners' native language), that is, when there was variation in speaking rate which was phonetically-relevant for English speakers, while variability in a cue that was not phonetically-relevant, such as fundamental frequency (F0) for English speakers, did not lead to any learning effect; in contrast, speakers of a tonal language (where contrasts in F0 are lexically relevant) did show a variability effect for F0 (Barcroft & Sommers, 2014; Sommers & Barcroft, 2007, 2011). It bears noting that the term "phonetically-relevant"

 doi:10.20944/preprints202211.0037.v1

used in this series of studies refers to the relevance of the cue to word recognition in the learners' native language (L1) but not to the L2 target contrast as we refer to in this study.

The cumulative evidence suggests that at least some sources of acoustic variability (including but not specific to talker variability) appear to have a positive influence in facilitating contrast learning. In a developmental study that manipulated variability in the phonologically-noncontrastive cue (i.e., phonetically-irrelevant talker cue) versus the phonologically-contrastive cue (i.e., phonetically-relevant voice onset time), Rost and McMurray (2010) found that greater variability on irrelevant dimensions is what matters for infant word learning. In a subsequent study, Galle et al. (2015) reported infants' successful learning of words without the use of multiple talkers by increasing overall acoustic variability in the single-talker speech. Taken together, these findings imply that input variability that really matters for contrast learning may be the variability in the phonetically-irrelevant cue (or secondary cue to the L2 contrast) that may reside in but is not limited to talker variability. Similar results abound in early work on learning theory, wherein the variability of irrelevant cues appears to help attune focus on critical cues (Bourne & Restle, 1959; Bush & Mosteller, 2006; Restle, 1955).

In our recent study, Zhang et al. (2021) integrated talker variability with varying degrees of acoustic exaggeration along the secondary dimension of vowel duration to train adult Chinese learners on the English /i/-/ɪ/ contrast. The results demonstrated that the single-talker training engendered generalization comparable to the multiple-talker training by shifting attention from the secondary duration cue to the primary spectral cues. In comparison, the natural single-talker speech without varying temporal exaggeration did not show similar benefits compared to the natural multiple-talker speech stimuli. The results provided support for the hypothesis that acoustic variability along the secondary dimension is beneficial to generalization for adult L2 learners, at least in the case of non-native learning of the English /i/-/ɪ/ contrast (Kondaurova & Francis, 2010). The question remains as to whether this training protocol is helpful to long-term retention as well. If phonetically-irrelevant acoustic variability does play a role in robust L2 speech learning, there should be a subsequent retention benefit in the single-talker condition in our training protocol.

The Impact of Cognitive Capacities on Learning from Variability

High talker variability may place demands on learners' cognitive resources, which is intuitively sensible in the way that talker variability leads to more variations and additional processing in the way acoustic patterns map onto phonetic categories (Dorman et al., 1977; Peterson & Barney, 1952). This demand has been reported in the field of language processing for both native and non-native listeners (e.g., Antoniou et al., 2015; Heald & Nusbaum, 2014; Lee et al., 2009; Martin et al., 1989; Mullennix et al., 1989; Wiener et al., 2018). In fact, non-native listeners may experience more difficulties with multiple-talker input in cases where native listeners do not. For instance, Antoniou et al. (2015) compared native listeners to two groups of non-native listeners with different levels of previous language exposure in a word-monitoring task with either single- or multiple-talker sentences. The results showed that non-native listeners with less exposure were slower and less accurate than native listeners regardless of the input variability, whereas the non-native listeners with more exposure were only weaker than the native listeners in the multiple-talker condition.

Training research has witnessed this greater cognitive demand for processing high variability. Increased variability appears to impede learning unfamiliar or difficult nonnative contrasts (e.g., Wade et al., 2007; Wayland & Guion, 2004). For example, Wade et al. (2007) showed that high variability could diminish learning effects for highly confusable vowels (i.e., vowels with more overlap in the acoustic space) compared with less overlapping vowels. Additionally, high variability imposes a burden on perceptually weak or immature learners and novice learners with low perceptual aptitudes (Antoniou & Wong, 2015; Chang & Bowles, 2015; Fuhrmeister & Myers, 2020; Perrachione et al., 2011;

 doi:10.20944/preprints202211.0037.v1

Sadakata & Mcqueen, 2014; Sinkeviciute et al., 2019), suggesting a potential trade-off between the more demanding nature of processing multiple-talker speech and the benefit this variable input might have for generalization and retention. In this regard, individual L2 learners' working memory capacity may play a role in determining training outcomes.

Working memory is assumed to involve the short-term storage, processing, and manipulation of information (Baddeley, 1986; Baddeley & Hitch, 1974). Its role as a source of individual differences in L1 is well studied (e.g., Conway & Engle, 1996; Daneman & Green, 1986; Just & Carpenter, 1992). There is also mounting evidence for the role of working memory capacity as a potential constraint on L2 processes, including reading (e.g., Leeser, 2007; Walter, 2006), writing (e.g., Adams & Guillot, 2008), sentence processing (e.g., Felser & Roberts, 2007; Juffs, 2004), speech production (e.g., O'Brien et al., 2006; Weissheimer & Mota, 2009), speech perception (e.g., Isaacs & Trofimovich, 2011), vocabulary development (e.g., Cheung, 1996; Papagno & Vallar, 1995), and grammar learning (e.g., French & O'Brien, 2008; Williams & Lovatt, 2005). Studies have generally shown that individuals with a higher working memory capacity tend to outperform those with a lower capacity. When employing the HVPT protocol to train advanced L2 learners to perceive English monophthongs, Aliaga-García et al. (2011) used a serial non-word recognition task to measure the learners' phonological short-term memory (PSTM) capacity, a subcomponent of the working memory construct, and assigned the learners into two PSTM capacity groups through the median split. The results showed that the high PSTM group obtained higher accuracy scores and greater perceptual gains than the low PSTM group. Another study by McHaney et al. (2021) reported similar results that individuals with higher working memory learned faster and to a greater extent than those with lower working memory when training native English speakers to learn Mandarin tone categories. These findings suggest that higher nonlinguistic working memory capacity will lead to more successful L2 learning.

Another key cognitive factor is the attention mechanisms that allow the learner to selectively maintain focus and inhibit distracting information (Conway et al., 1999; Engle, 2002; Kane et al., 2001). Attentional control is thought to play an important role in L2 learning (Ellis, 2006; Francis et al., 2000; Goldstone, 1998). The attention-to-dimension models (A2D models) of speech perception characterize perceptual space as a multidimensional structure whereby both L1 and L2 speech learning can be understood as changes in the distribution of selective attention to certain dimensions, specifically, shifting attention to dimensions relevant for categorization and withdrawing attention from irrelevant dimensions (Francis et al., 2008; Francis & Nusbaum, 2002; Goldstone, 1993, 1994; Kuhl & Iverson, 1995; Nosofsky, 1986; Pisoni et al., 1994). Due to such experience-induced changes in the distribution of attention, the perceptual space of adult learners has already been ''warped'' due to their experience with L1, which may cause difficulties in perceiving L2 phonetic contrasts that are not distinguished in L1 (Iverson & Kuhl, 1995; Kuhl & Iverson, 1995; Pisoni et al., 1994). In this view, adult learners have to overcome interference from too much attention directed to acoustic cues that are not used in L2. To illustrate, the English /i/ and /ɪ/ contrast can be distinguished along the spectrum dimension (vowel quality, related mainly to the first and second formant frequencies, F1 and F2) and the duration dimension (vowel length). Native English speakers have been reported to rely primarily on spectral properties, with vowel duration playing only a secondary role (Hillenbrand et al., 2000; Mermelstein, 1978), whereas adult Chinese learners of English rely predominantly on vowel duration instead of spectral properties in both perception and production (Escudero & Boersma, 2004; Liu et al., 2014). According to the A2D models, successful acquisition of the English /i/ and /ɪ/ categories by native Chinese learners involves not only enhancement of attention to the under-attended spectral properties, but also a simultaneous withdrawal of attention from vowel duration. Of particular relevance to this study is that shifting attention to phonetically relevant cues and away from irrelevant cues is the assumed benefit of high variability training. Therefore, the incorporation of working memory and selective attention into our study represents a step in the

 doi:10.20944/preprints202211.0037.v1

right direction for understanding speech learning in terms of changes in the distribution of attention, a view that is pervasive in the literature. This approach may also offer potential insights into why some individual learners benefit from high variability training but not others.

The Current Study

The present study extends our previous study on training adult Chinese speakers to distinguish the English /i/ and /ɪ/ categories, contrasting the multiple-talker (MT) versus single-talker (ST) training conditions with the joint use of varying degrees of temporal exaggeration and audio-visual exposure (Zhang et al., 2021). Specifically, our first research question was whether the phonetically-irrelevant acoustic variability benefit can be retained three months after training. If acoustic variability in phonetically-irrelevant cues plays a comparable role as talker variability does, we should expect the single-talker training group to show a benefit of retention in the delayed post-test. The second research question was whether training outcomes are related to nonlinguistic cognitive abilities. To this end, we measured individual learners' working memory capacity and selective attention before training. Our hypothesis was that, to the extent speech categorization may draw on domain-general skills such as working memory and selective attention, individual differences in these abilities would be reflected in training results with significant advantages in favor of individuals with higher cognitive abilities.

In order to provide a more nuanced assessment of training-induced changes in behavioral performance, we further examined the learners' attention to the primary spectral cues and the secondary durational cue in pre- and post-tests in addition to using a naturally-produced word identification task. Of relevance to our aims, researchers have characterized the mechanism of attention shifts as adjusting attentional weights assigned to certain dimensions of the contrast (Aha & Goldstone, 1990; Nosofsky, 1986). Therefore, we used a carefully-controlled grid of computer-synthesized phonemes in an identification task and adopted a logistic regression model to fit the participants' response data that could provide values of stimulus-tuned coefficients as measures of acoustic cue weights (Morrison, 2005, 2007; Morrison & Kondaurova, 2009). Our previous study (Zhang et al., 2021) has shown that for the trained groups, the primary spectral cues of F1 and F2 began to show increased weight and the secondary durational cue was significantly less weighted after training. It remains to be tested whether this training-induced cue reweighting would be retained months after training.

In addition to perception performance, we were also interested in the participants' changes in production as a result of our training protocol. The motor theory (e.g., Liberman et al., 1952, 1967) and the direct realist theory (e.g., Best, 1995; Fowler, 1986) both assume commonly-shared representations between speech perception and production. Furthermore, two influential models of L2 speech learning, the Perceptual Assimilation Model (PAM, Best, 1995) and the Speech Learning Model (SLM, Flege, 1995a), also assume that accurate perception is a prerequisite for accurate production. Hence, perceptual training that has been proved to change perceptual representations can provide a direct opportunity for investigating the relationship between speech perception and production in the learning process. Particularly, several studies have compared the effect of multiple- versus single-talker training on production (Brosseau-Lapré et al., 2013; Dong et al., 2019; Kartushina & Martin, 2019; Wiener et al., 2020), and the evidence is inconclusive. To our knowledge, previous training studies did not examine the transfer of learning by testing perceptual training effects on the use of primary vs. secondary cues in speech production, especially in terms of long-term benefit. Our recent study (Zhang et al., 2021) found that the trained participants significantly increased the use of the critical spectral cues and decreased using the secondary durational cue, crucially in line with the cue reweighting pattern shown in their perception data. In the current study, therefore, we further examined the possible retention of the training-induced changes in the use of spectral vs. durational cues in production as a result of our training protocol.

 doi:10.20944/preprints202211.0037.v1

Method

Participants

Sixty native speakers of Mandarin Chinese (30 females and 30 males, $Mean_{age}$ = 21.45, *SD* = 2.82) at Xi'an Jiaotong University were enrolled with written informed consent, following the ethical research approval of the Institutional Review Board at Xi'an Jiaotong University. All participants are right-handed, and none reported a history of speech, language, or hearing problems or disorders during screening. They had studied English in school for at least 9 years but did not use English regularly. None of the participants reported having spent over one month in an English-speaking country or community. The participants received monetary reimbursement for participation and were randomly assigned to three groups: 20 in the multiple-talker (MT) group, 20 in the single-talker (ST) group, and 20 in the control (CTRL) group that did not receive training.

Stimuli

Training Stimuli

Thirty different monosyllabic words (15 minimal pairs, see Appendix A in Supplementary Material) containing the target phonemes of English /i/ and /ɪ/ were used as the training input for both the MT and ST groups. The words were repeated four times in pseudo-random order in each of the seven sessions (840 trials in total: 30 words × 4 times × 7 sessions). The only difference in the MT versus ST condition was the number of native speakers of American English (four talkers versus one talker). Systematic lengthening (300, 208, 144, and 100%) along the duration dimension was equally applied to all the tokens used in the MT and ST conditions using PRAAT (Boersma & Weenink, 2016), which was irrelevant and uninformative to the category distinction. By contrast, the stimuli included variations along the spectral dimensions (i.e., F1 and F2) that were informative to the category membership (see Appendix A for the acoustic plot and measures of F0, F1, F2, and vowel duration of the target vowels contained in the training tokens). We statistically analyzed the acoustic values of the vowels using one-way ANOVA. In addition to the mean value and standard deviation, we calculated the relative variability of each acoustic feature, defined as the variance of the log-transformed values (Lewontin, 1966). Results showed that the overall mean of F0, F1, F2, and vowel duration did not differ significantly between the two training conditions (see Appendix A). While there was no significant difference in the amount of acoustic variability in F1, F2, and vowel duration, the variability in F0 did differ significantly between the two training conditions, with a greater amount in the MT condition than in the ST condition for both vowel categories.

In addition to the auditory stimuli, digital video recordings of the talkers were utilized as visual cues in the training (Zhang & Cheng, 2011). The video frames of all tokens were edited using Final Cut Studio (Apple Inc.) to match the four levels of temporal exaggeration. Five new minimal pairs recorded by two novel talkers (one female and one male, only audios) were used in the progress-monitoring quizzes that followed each of the seven training sessions.

Test Stimuli

The natural word stimuli in the identification task were produced by four native speakers of American English (2 males and 2 females) that were not employed in training, with a total of 160 natural tokens (10 minimal pairs × 4 talkers × 2 times). Four of the 10 minimal pairs were selected from the training stimuli, while the other six were untrained words (see Appendix A).

The stimuli employed in the synthetic phoneme identification task were made using the design of English /i/ to /ɪ/ continuum (Cheng & Zhang, 2013). The recordings of the /i/ and /ɪ/ sounds were made by a male native English speaker in the "hVd" context at a 44.1 kHz sampling rate. The sounds were then digitally cut out and processed in Sound Forge

 doi:10.20944/preprints202211.0037.v1

(SoundForge9, Sony Corporation, Japan) to have an equal duration of 170 ms and fade-in/fade-out time of 10 ms. The F1 and F2 frequencies of the endpoint /i/ sound are 353 and 2340 Hz. The F1 and F2 of the endpoint /ɪ/ sound are 437 and 2005 Hz. A seven-step continuum between these two endpoints was created by varying spectral properties utilizing the morphing technique in the STRAIGHT package (Kawahara et al., 1999) on the MATLAB platform (MathWorks Corporation, United States). For each sound in the continuum, the PSOLA technique in PRAAT was utilized to generate a seven-step continuum by varying vowel duration with a range from 250 to 100 ms (25 ms/step). The reason for the endpoints to be 250 ms and 100 ms and 7 steps for the duration continuum was to avoid listeners' possible preference on the dimension with the least variability (Bohn, 1995). Each step along the duration continuum was made approximately equivalent to one just noticeable difference (JND) for English listeners (Klatt, 1976), as was each step along the spectral continuum based on the F1 values 11–14 Hz following Kewley-Port and Watson (1994). The 49 synthetic stimuli (7 steps × 7 steps) were normalized to have the same average root mean square intensity (Fig. 1).

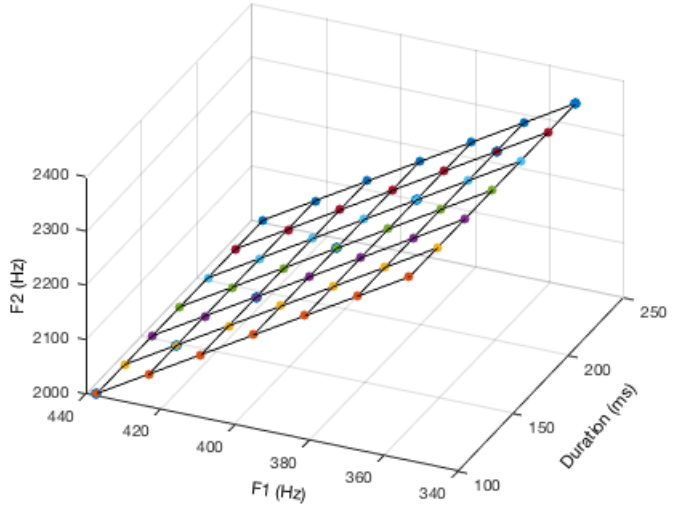


Fig. 1. Plot of the 7x7 stimulus grid along the F1, F2, and duration dimensions for the synthetic phoneme identification task. Each circle represents a stimulus.

Procedure

Cognitive Tests

Before training, we measured the participant's nonlinguistic working memory capacity and selective attention for the two training groups. We used the backward digit span task to assess the capacity of complex working memory (Colom et al., 2006). This task is used widely in behavioral and neuroscientific research (e.g., Oberauer et al., 2003) and is a subpart of the Wechsler IV intelligence test (Gathercole & Alloway, 2008), which minimizes the influence of language knowledge. In this task, a series of digits would flash up in the center of the computer screen and the participants were expected to recall this series of digits in reverse order. Each digit was presented for one second (in Arial, font size 100) against a black background, and with one second in between consecutive digits. The participants were first presented with two 3-digit trials for practice and then were tested on 2- up to 10-digit sequences (two trials for each sequence length, making up 18 trials in total). Performance was determined by calculating the proportion of correctly recalled digit sequences. The higher the score, the better working memory the participant has. The mean score of the participants was 0.84 (SD = 0.09). The mean working memory scores did not differ significantly between the MT and ST groups, $F(1,38) = 0.09$, $p = .77$, $\eta^2_p = 0.002$.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 November 2022 doi:10.20944/preprints202211.0037.v1

To measure selective attention, we used a computerized variant of the flanker task which is commonly considered a measure of inhibitory control (Eriksen & Eriksen, 1974). In this task, the participants saw a row of five white arrows (in Arial, font size 100) against a black background. The middle symbol in the row (the target) was either a leftward- or rightward-pointing arrowhead. The participants were required to report the direction of the middle arrow by pressing the corresponding keys on a CHRONOS keypad (Psychology Software Tools Inc., United States). The middle target was flanked on either side by two congruent or incongruent arrows or by neutral lines (e.g., for a leftward-pointing target < : < < < < < as the congruent condition; > > < > > as the incongruent condition; and – – < – – as the neutral condition). In the incongruent condition, the participant's performance provided a measure of inhibitory control in the context of selective attention. Each trial started with a fixation cross that remained on the screen for 250 ms. Following this fixation cross, the arrow row was presented for 1500 ms, with inter-trial time of 1000 ms. There were six different stimuli (2 pointing directions for the target × 3 flanker conditions). Each of the six different stimuli was presented 12 times in the test part in random order (72 trials in total). Six practice trials were presented before the test. The mean accuracy of the responses was 93.99% (*SD* = 9.23), and the mean response time was 417.08 ms (*SD* = 72.37) from visual presentation onset. The participant's performance was determined by the flanker score which is a composite of both response accuracy and time, following NIH toolbox guidelines (Zelazo et al., 2014). If the participant's accuracy level was less than or equal to 80%, the final flanker score for the participant was equal to the accuracy score on a scale from 0 to 5 (Eq. 1). If the accuracy level reached more than 80%, the participant's response time (RT) score was added to the accuracy score to obtain a final flanker score. The RT score was calculated based on each participant's median RT. If the participant's median RT fell outside of the range of 500-3000 ms, the median RT less than 500 ms was set equal to 500 ms, and the median RT more than 3000 ms was set equal to 3000 ms. The truncated median RTs were then rescaled to obtain the response time score (Eq. 2). The participants' mean flanker score was 9.32 (*SD* = 1.75). Higher flanker scores indicate higher levels of selective attention ability. The mean flanker scores did not differ significantly between the MT and ST groups, $F(1,38) = 0.12$, $p = .73$, $\eta^2_P = 0.003$.

$$\text{Accuracy score} = \frac{5}{72} \times Number\ of\ correct\ responses \quad \text{(Eq.1)}$$

$$\text{Response time score} = 5 - 5 \times \frac{\log RT - \log 500}{\log 3000 - \log 500} \quad \text{(Eq.2)}$$

Pre- and Post-tests

Identical tests were implemented one week before training (the pre-test), one week after training (the immediate post-test), and three months after training (the delayed post-test), including both perception and production tests. The perception test included a synthetic phoneme identification task and a natural word identification task. For the synthetic phoneme identification task, the vowel stimuli were presented on a DELL Desktop via headphones (Sennheiser, CX1) at about 70 dB SPL using E-PRIME Version 2.0 (Psychology Software Tools Inc., United States). The auditory stimulus was displayed with two possible answers on the computer screen (i.e., /i/ and /ɪ/, represented in International Phonetic Alphabet, and a written example word was given for each vowel). The participants had learned the IPA symbols since middle school. The experimenters verified that each participant correctly identified the /i/ and /ɪ/ symbols in association with a minimal pair ("beat" vs. "bit") prior to the experiment. The participants were asked to determine which of the two vowels they heard by pressing the corresponding buttons on a CHRONOS keypad. Although each trial was self-paced, the participants were instructed to respond as quickly as possible. Each sound in the 7 × 7 stimulus grid was repeated 10 times within one randomized block. The experiment began with a brief familiarization phase followed by the test phase. With a total of 490 trials, the average running time was about 12 minutes.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 November 2022 doi:10.20944/preprints202211.0037.v1

To test generalization to new talkers and new phonetic contexts, the natural word identification task included 8 trained minimal-pair words and 12 untrained minimal-pair words, recorded by four native American English speakers (2 males and 2 females) who were not present in the training program. The participants were required to determine whether the word they heard included /i/ or /ɪ/ by clicking the icons of the two vowels on the screen. With each word presented eight times, the average completion time for this task was about 8 minutes.

The identical set of 20 words in the perception test was utilized as production prompts in the production test. The visually-printed words were shown to the participants with phonetic transcriptions that they were familiar with. The participants were required to speak the words twice at a normal rate as in an example carrier "Say the word 'heat'". The recordings were made with the SHURE SM58 (SHURE, United States) microphone positioned in a sound-proofed booth approximately 20 cm in front and 45° to the right of the participants' lips, using PRAAT at 44.1 kHz, 16-bit quantization. To minimize the impact of exposure to the same words spoken by native speakers, the participants completed the production test before the perception test.

Training Program

The 60 participants were randomly assigned to three groups with 20 in each, including two training groups (MT and ST) and one control group (CTRL). To ensure that participants' perceptual performance did not differ significantly across groups before training, we conducted a one-way ANOVA with the between-group factor Group (MT, ST, CTRL) for the perception data obtained from the pre-test. There were no significant differences between the three groups in the relative weights of the three acoustic cues, $F(2,57) = 0.075$, $p = .928$, $\eta^2_p = 0.003$ for F1, $F(2,57) = 0.016$, $p = .984$, $\eta^2_p = 0.001$ for F2, $F(2,57) = 0.004$, $p = .996$, $\eta^2_p < 0.001$ for vowel duration, and in the word identification accuracy, $F(2,57) = 0.013$, $p = .987$, $\eta^2_p < 0.001$. The MT condition employed the minimal-pair words produced by four talkers, while the ST condition used the same words produced by a single talker. All other aspects were matched between the two conditions. Both training groups completed seven sessions during one lab visit, starting from the most exaggerated sounds with one single talker and gradually decreasing exaggeration level and adding talkers solely in the MT condition (Table 1).

Table 1 Talker number and temporal exaggeration setup for the seven training sessions of the MT (multiple talkers) and ST (single talker) training groups

Session	Stimuli	Talker number	
		MT	ST
Session 1	exaggerated 300%	1	1
Session 2	exaggerated 300%	2	1
Session 3	exaggerated 300%	3	1
Session 4	exaggerated 300%	4	1
Session 5	exaggerated 208%	4	1
Session 6	exaggerated 144%	4	1
Session 7	natural 100%	4	1

In each of the seven sessions, the participants were required to click either icon of vowels presented by the IPA on the computer screen after they heard a word containing the clicked vowels and saw a talker utter the word in the visual video. Each icon contained ready-to-click 60 words. After completing the 120 trials in a training session, the trainees were given a quiz of 10 untrained words produced by new talkers. The trainees could move on to the next session only when the percentage accuracy from the quiz was equal to or greater than 90%. If the trainees failed to meet the criterion, they had to receive the

 doi:10.20944/preprints202211.0037.v1

same training session again, followed by the same quiz. There was no such criterion after the repeat session for them to move on to the next session. In both the MT and ST groups, the participants completed 9 to 12 training sessions (including repeat sessions) which took them about 60 to 90 minutes, depending on each participant's number of repeat sessions and response pace.

Data Analysis

For perception data, we compared the percent accuracy of trained and untrained word identification from the pre- and post-tests to assess training effects. Additionally, we conducted a binary logistic regression analysis to fit each participant's /i/ responses to the synthetic vowel stimuli in order to examine changes in the weighting of the spectral and duration cues. The logistic regression model provided spectrally- and duration-tuned coefficients to describe the effect of a one-unit change in the predictors (F1, F2, and vowel duration) on the log odds (the probability of selecting /i/). Before the fit, the spectral and duration values were standardized by centering and scaling (Escudero et al., 2009). Instead of directly employing standardized regression coefficients as cue weights of the predictors, we used the relative weight analysis (RWA) as a supplement to the logistic regression and adopted the resulting coefficients as measures of cue weights (Tonidandel & LeBreton, 2015). The RWA addresses the problem of multicollinearity by transforming original predictors into a new set of orthogonal predictors (Tonidandel & LeBreton, 2010). As such, it quantifies a predictor's importance by incorporating both its direct effect and its joint effect with other predictors.

The production data were assessed acoustically in terms of vowel duration, and F1 and F2 frequencies. Specifically, the vowel onset was marked at the first positive peak in the periodic portion, and the offset was identified at the point of a considerable decrease in overall amplitude and waveform complexity. Based on a 14-pole linear predictive coding analysis, the F1 and F2 values were calculated as an unweighted mean of frequencies at five temporally equidistant positions corresponding to the 20–35–50–65–80%-point of the vowel, by centering a 25-ms Hanning window at each temporal location (Jacewicz et al., 2011). To reduce frequency variations between males and females, frequency values were transformed from Hertz to Barks (Traunmüller, 1990), and were then converted to *z*-scores for each participant. The duration data were also converted to *z*-scores to control differences in speaking rate. Following the perceptual data analysis, we adopted the RWA as a supplement to the logistic regression to measure each participant's use of F1, F2, and vowel duration in production.

For statistical analysis, we constructed linear mixed-effects models (LMM) for the participants' word identification accuracy and RWA coefficients of F1, F2, and duration in perception and production, using the lme4 package (Bates et al., 2015) in R (R Development Core Team, 2021). To answer the first research question concerning training effects, the model for the word identification accuracy was constructed with Group (MT, ST, CTRL), Test (pre-test, post-test1, post-test2), Word Type (trained vs. untrained word), and their interactions as the fixed effects. We also constructed the respective LMMs for the RWA coefficients of F1, F2, and duration measured in perception and production with the fixed effects of Group (MT, ST, CTRL), Test (pre-test, post-test1, post-test2), and their interactions. To answer the second research question that examined the influence of working memory capacity and selective attention on training outcomes, we added the fixed effects of working memory accuracy, flanker scores of selective attention, and their interactions together with the experiment design factors into the aforementioned LMMs for the trained participants. Before being added to the models, the working memory accuracy was log-transformed and centered due to its positively skewed distribution, and the flanker score was centered.

For model selection, we used the maximal random effects structure justified by the experiment design, including all possible by-item and by-participant random intercepts and slopes for the main effects in the fixed model (Barr et al., 2013). If the model failed to

 doi:10.20944/preprints202211.0037.v1

converge, we first considered adjusting the optimizer, that is, the method whereby the model finds an optimal solution, using the all_fit function in the afex package (Singmann et al., 2021). If none of the optimizers was helpful, we simplified the random effects structure by first removing random correlations, and then by removing random slopes that accounted for the least variance, until the model converged. We utilized the anova function in the lmerTest package (Kuznetsova et al., 2017) to examine the main effect for each included fixed effect, and used the emmeans function in the emmeans package (Lenth, 2021) to compare differences between factor levels based on adjusted Tukey tests. The Satterthwaite method was employed to approximate degrees of freedom for all models.

In order to better understand the relationship between word performance and cue weighting strategies, we used Pearson's product-moment correlation to examine the relationship between the word identification accuracy and cue weights of F1, F2, and duration before and after training. Moreover, Pearson's product-moment correlation was used to investigate the relationship between speech perception and production in terms of cue weights of F1, F2, and duration before and after training (Schertz et al., 2015; Shultz et al., 2012).

Results

Training Effects on Perception and Production

Perception Data

For word identification performance before and after training (Fig. 2), we modeled the percent correct accuracy and long-transformed response time of word identification respectively, with the fixed effects of Group (MT, ST, CTRL), Test (pre-test, post-test1, post-test2), Word Type (trained versus untrained word), and their interactions. For the sake of brevity, we only reported the significant results pertaining to the training effect (see Appendix B for the summary output for the final models). Table 2 summarizes the results of Type III analysis of variance for the main effects of each included fixed effect with the identification accuracy as the dependent variable. There was a significant interaction effect between Group and Test, $F(4,59.2) = 5.47$, $p < .001$ on the word identification accuracy. Results showed that the identification accuracy of the MT group significantly improved in the immediate post-test compared to the pre-test, Estimate = 0.16, $SE = 0.03$, $p < .001$, whereas there was no statistically significant difference between the performance of the delayed post-test and the pre-test, Estimate = 0.08, $SE = 0.04$, $p = .10$, suggesting that the training effect on word identification was not retained after three months for the MT group. By contrast, the ST group significantly improved their identification accuracy in the immediate post-test, Estimate = 0.17, $SE = 0.03$, $p < .001$, and the improvement was retained in the delayed post-test compared to the performance in the pre-test, Estimate = 0.14, $SE = 0.04$, $p = .002$. There was no significant difference between the performance of the pre-test and the two post-tests for the CTRL group (ps > .1). In the model for the response time, no significant effect of Test was found (see Appendix B).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 November 2022 doi:10.20944/preprints202211.0037.v1

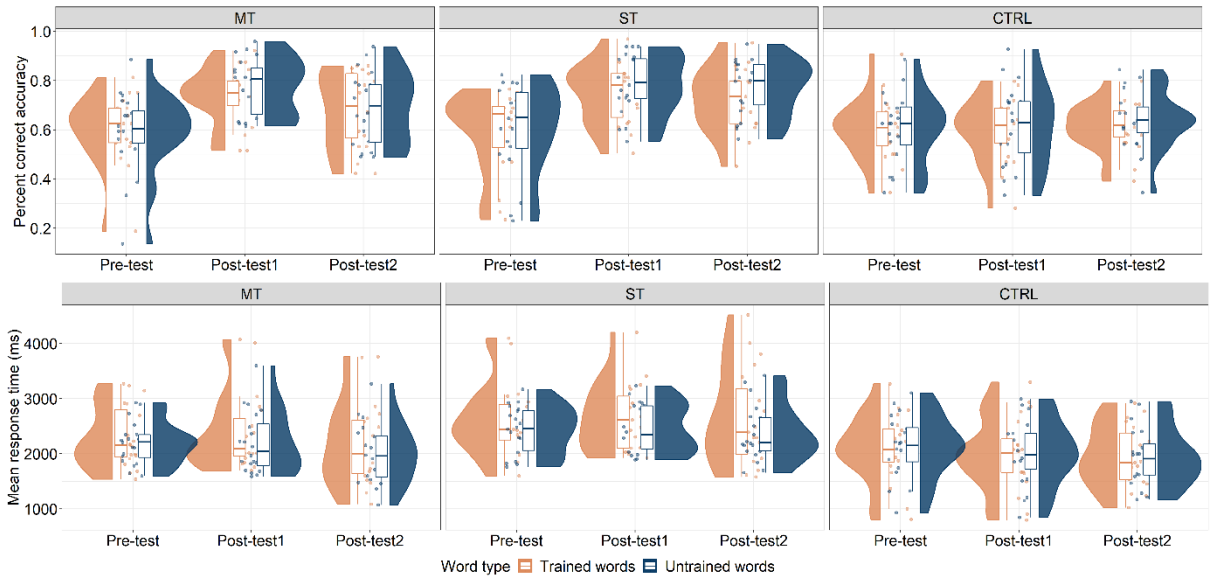


Fig. 1. Half-violin plots and boxplots visualizing percent correct accuracy (top panel) and response time (bottom panel) of identification performance of the trained and untrained words for each group (MT, ST, CTRL) in the pretest, post-test (Post-test1), and delayed post-test (Post-test2).

Table 2 Results of Type III analysis of variance on the word identification accuracy with Satterthwaite approximation for degrees of freedom

Fixed effect	*Sum Sq*	*Mean Sq*	*NumDF*	*DenDF*	*F* value	*p* value
Final Model: Accuracy ~ Group * Test * Type + (1 + Test | Subject) + (1 | Word)						
Group (MT, ST, CTRL)	0.57	0.28	2	57.5	4.50	.02
Test (pre-test, post-test1, post-test2)	1.97	0.98	2	59.2	15.65	< .001
Type (trained versus untrained)	0.02	0.02	1	18	0.26	.62
Group:Test	1.38	0.34	4	59.2	5.47	< .001
Group:Type	0.09	0.04	2	3393	0.71	.49
Test:Type	0.21	0.10	2	3393	1.64	.20
Group:Test:Type	0.06	0.01	4	3393	0.23	.92

Fig. 3 shows the participants' categorization pattern of the synthetic continua before and after training, illustrating the F1 and vowel duration dimensions for the sake of simplicity. Each cell indicates one stimulus in the 7 × 7 stimulus grid, with darker cells representing a larger proportion of /i/ response. Visual inspections suggested training-induced changes. Before training, longer vowel duration values elicited a fair amount of /i/ response in all of the three groups. In the immediate post-test, the /i/ response of the MT and ST groups clustered more in the space defined by lower F1 values, and the difference between shorter duration and longer duration was not so distinct, suggesting the two training groups began to rely more on F1 instead of vowel duration. In the delayed post-test, the two training groups appeared to retain a similar categorization pattern. But the two groups showed different trends of slight changes. For the MT group, higher F1 elicited more /i/ responses compared to those of the immediate post-test, whereas for the ST group longer duration elicited more /i/ responses compared to the immediate post-test.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 November 2022 doi:10.20944/preprints202211.0037.v1

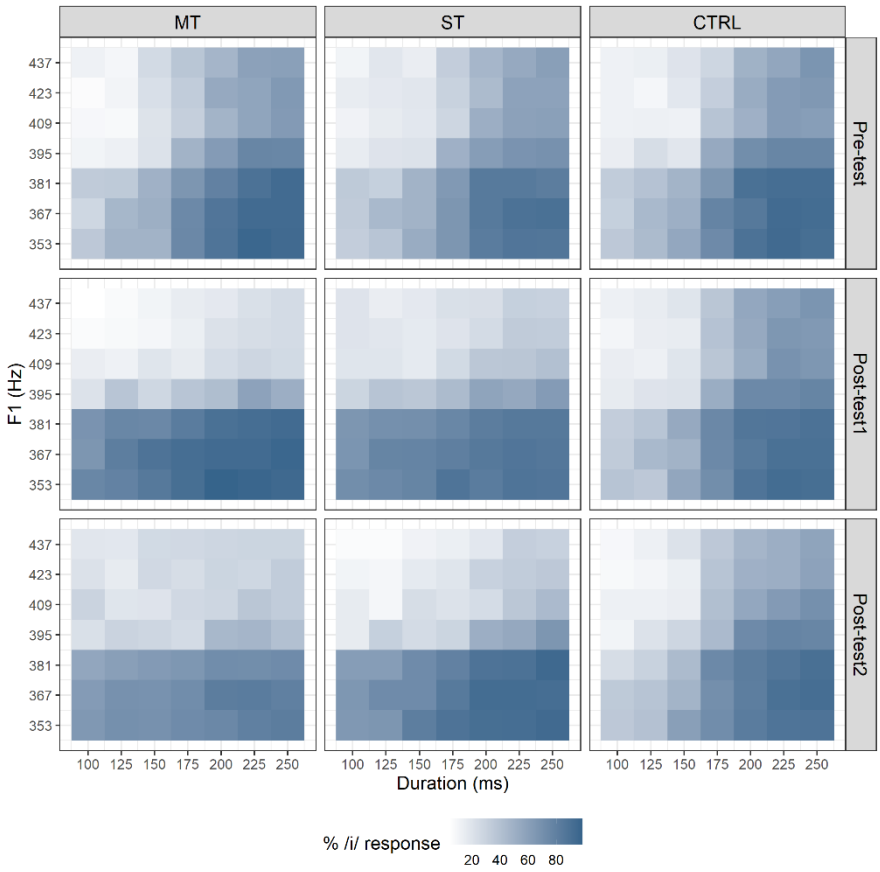


Fig. 2. Heat plots for categorization of the synthetic /i/-/ɪ/ grid along the F1 and duration dimensions in the pre-test, immediate post-test (Post-test1), and delayed post-test (Post-test2). Each cell indicates one stimulus in the 7x7 grid. The darkness of the cell represents the percentage response for /i/, with the darkest cells eliciting the highest portion of /i/ response and the white cells the lowest.

For statistical analysis, we used the spectrum- and duration-tuned RWA coefficients to estimate the relative cue weights used by each participant, with a greater value of the coefficient representing heavier weighting for the cue (see Fig. 4). We modeled the three outcomes of interest (i.e., the RWA coefficients of F1, F2, and duration) as the respective dependent variables, including the same fixed effects of Group (MT, ST, CTRL) and Test (pre-test, post-test1, post-test2). Table 3 summarizes the results of Type III analysis of variance for the main effects of each included fixed effect (see Appendix B for the summary output for the final models). For the F1-tuned RWA coefficients, a significant interaction effect for Group and Test was found, $F(4,114) = 6.71$, $p < .001$.

Further comparisons showed that the F1 weight of the MT group significantly increased in the immediate post-test, Estimate = 0.20, $SE = 0.04$, $p < .001$, as well as in the delayed post-test compared to the pre-test, Estimate = 0.17, $SE = 0.04$, $p < .001$. For the ST group, the F1 weight also significantly increased in the immediate post-test, Estimate = 0.21, $SE = 0.04$, $p < .001$, and in the delayed post-test relative to the pre-test, Estimate = 0.19, $SE = 0.04$, $p < .001$. The CTRL group did not change significantly before and after training ($ps > .1$). Likewise, for the F2-tuned RWA coefficients, there was a significant interaction effect for Group and Test, $F(4,114) = 6.40$, $p < .001$. Pairwise contrasts showed that the MT group significantly increased their use of F2 in the immediate post-test, Estimate = 0.20, $SE = 0.04$, $p < .001$, and in the delayed post-test compared to the weight used in the pre-test, Estimate = 0.19, $SE = 0.04$, $p < .001$. For the ST group, the weight of F2 also significantly increased in the immediate post-test, Estimate = 0.19, $SE = 0.04$, $p < .001$, and the increase retained in the delayed post-test, Estimate = 0.18, $SE = 0.04$, $p < .001$. Again, there was no

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 November 2022 doi:10.20944/preprints202211.0037.v1

significant change in the CTRL group (*p*s > .1). For the duration-tuned RWA coefficients, the results also indicated a significant interaction effect for Group and Test, $F(4,114) = 6.90$, $p < .001$. Comparisons showed that the duration weight of the MT group significantly decreased in the immediate post-test, Estimate = -0.34, *SE* = 0.06, $p < .001$, and the weight decreased to a greater extent in the delayed post-test compared to the pre-test, Estimate = -0.40, *SE* = 0.06, $p < .001$. For the ST group, the duration weight also significantly decreased in the immediate post-test, Estimate = -0.30, *SE* = 0.06, $p < .001$, and the decrease retained in the delayed post-test, Estimate = -0.29, *SE* = 0.06, $p < .001$. In contrast, no significant change was found for the CTRL group (*p*s > .1). These results suggest that both MT and ST groups significantly increased their weight of F1 and F2, and in the meantime significantly less weighted vowel duration, and the changes were retained three months after training.

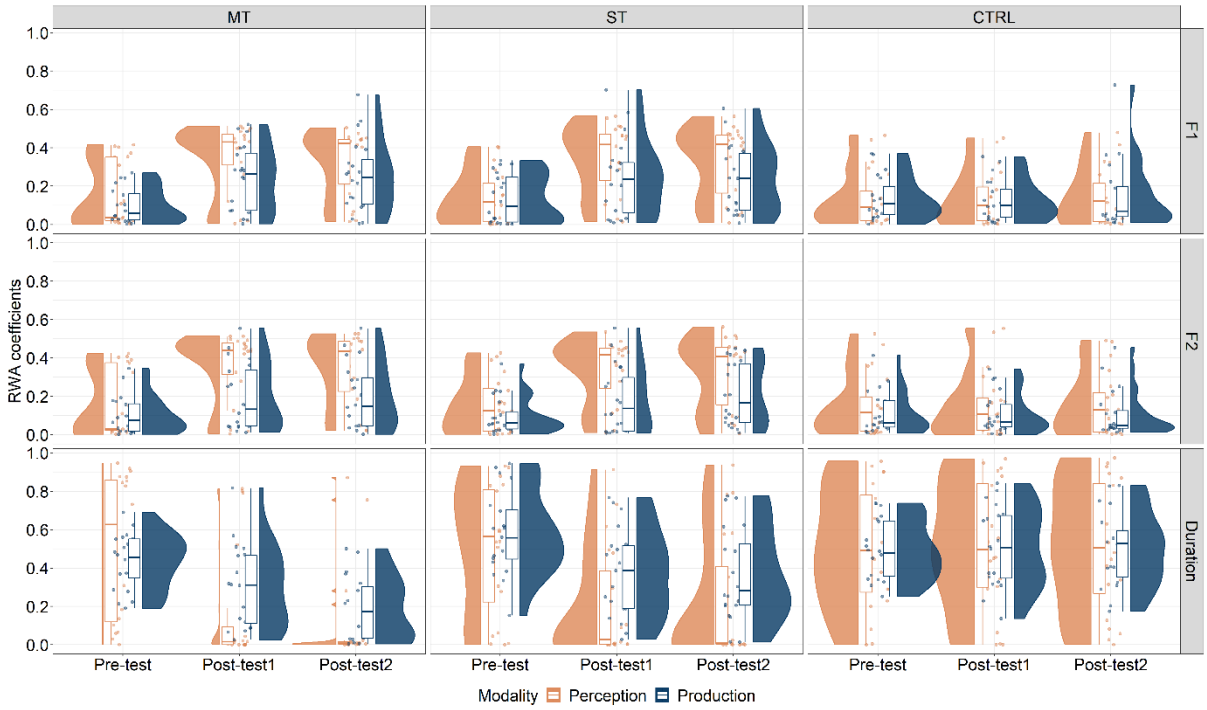


Fig. 3. Half-violin plots and boxplots visualizing the RWA coefficients of F1, F2, and duration measured in the perception and production data for each group in the pre-test, immediate post-test (Post-test1), and delayed post-test (Post-test2).

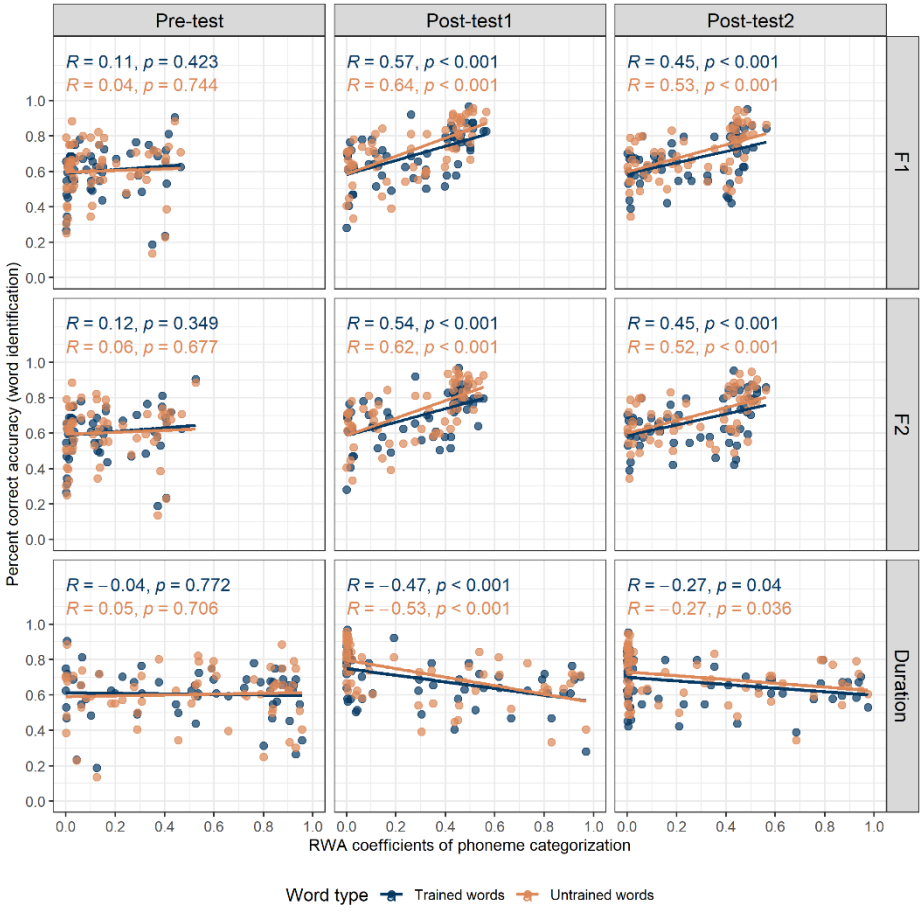
Table 3 Results of Type III analysis of variance on the perceptual RWA coefficients of F1, F2, and duration with Satterthwaite approximation for degrees of freedom

Fixed effect	*Sum Sq*	*Mean Sq*	*NumDF*	*DenDF*	*F* value	*p* value
Final Model: F1 ~ Group * Test + (1 | Subject)						
Group (MT, ST, CTRL)	0.17	0.08	2	57	6.85	.002
Test (pre-test, post-test1, post-test2)	0.69	0.34	2	114	27.83	< .001
Group:Test	0.33	0.08	4	114	6.71	< .001
Final Model: F2 ~ Group * Test + (1 | Subject)						
Group (MT, ST, CTRL)	0.13	0.07	2	57	5.32	0.01
Test (pre-test, post-test1, post-test2)	0.64	0.32	2	114	25.84	< .001
Group:Test	0.32	0.08	4	114	6.40	< .001

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 November 2022 doi:10.20944/preprints202211.0037.v1

Final Model: Duration ~ Group * Test + (1 | Subject)						
Group (MT, ST, CTRL)	0.34	0.17	2	57	4.33	0.02
Test (pre-test, post-test1, post-test2)	1.90	0.95	2	114	24.03	< .001
Group:Test	1.09	0.27	4	114	6.90	< .001

To explore relationships between word identification and cue use strategies, we used Pearson's product-moment correlation to examine correlations between the participants' word identification accuracy and RWA coefficients of F1, F2, and duration measured in the pre-test, immediate post-test, and delayed post-test (Fig. 5). The results showed that before training there was no significant correlation between the word identification accuracy and any of the dimensions. By contrast, in the immediate post-test, we found significant positive correlations for the F1 dimension (r = 0.57, p < .001 for the trained words, r = 0.64, p < .001 for the untrained words), and for the F2 dimension (r = 0.54, p < .001 for the trained words, r = 0.62, p < .001 for the untrained words). There were also significantly negative correlations between the word identification accuracy and the duration weight (r = -0.47, p < .001 for the trained words, r = -0.53, p < .001 for the untrained words). These correlations were retained in the delayed post-test, though to a less extent, with positive correlations for the F1 dimension (r = 0.45, p < .001 for the trained words, r = 0.53, p < .001 for the untrained words) and for the F2 dimension (r = 0.45, p < .001 for the trained words, r = 0.52, p < .001 for the untrained words), as well as negative correlations between the word identification accuracy and the duration weight (r = -0.27, p = .04 for the trained words, r = -0.27, p = .04 for the untrained words). It is noticeable that there were stronger correlations between the RWA weights and the untrained word identification accuracy compared to the trained words after training.



Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 November 2022 doi:10.20944/preprints202211.0037.v1

Fig. 4. Scatter plots for correlations between the word identification accuracy and the RWA coefficients of F1, F2, and duration measured in the pre-test, immediate post-test (Post-test1), and delayed post-test (Post-test2) on the trained versus untrained words.

Production Data

Fig. 6 shows the scatter plots for acoustic measures of the target vowels in the combination of F1 and vowel duration dimensions produced in the pre-test, immediate post-test, and delayed post-test. Before training, the two vowel categories showed much overlapping in the F1 dimension and were largely separated along the duration dimension for all the three groups. In the immediate post-test, the two categories for the MT and ST groups appeared to be more separated along the F1 dimension. In the delayed post-test, the pattern seemed to be retained but with more separating distributions of each category, especially for the MT group.

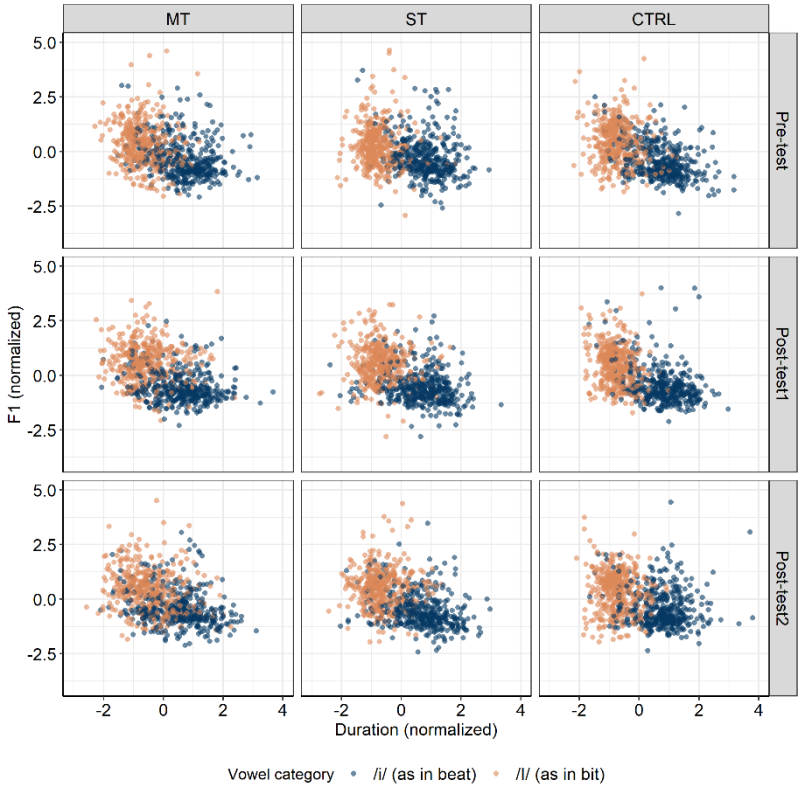


Fig. 6. Mean production values in the combination of F1 and vowel duration dimensions, normalized by the speaker (converted to *z*-scores along each dimension), measured in the pre-test, immediate post-test (Post-test1), and delayed post-test (Post-test2).

Following the perception data analysis, we adopted the spectrum- and duration-tuned RWA coefficients to gauge the relative cue weights used by each participant in production, with higher coefficients representing heavier weighting for the cue (see Fig. 4). The three outcomes of interest (i.e., the RWA coefficients of F1, F2, and duration) as the respective dependent variables were modeled, including the same fixed effects of Group (MT, ST, CTRL) and Test (pre-test, post-test1, post-test2). Table 4 summarizes the results of Type III analysis of variance for the main effects of the fixed effects (see Appendix B for the summary output for the final models). For the F1-tuned RWA coefficients, there was a significant interaction effect for Group and Test, $F(4,114) = 3.11$, $p = .02$.

Further comparisons showed that the F1 weight of the MT group significantly increased in the immediate post-test, Estimate = 0.14, $SE = 0.03$, $p < .001$, and in the delayed post-test compared to the pre-test, Estimate = 0.15, $SE = 0.03$, $p < .001$. For the ST group, the results also showed a significant increase in the use of F1 in the immediate post-test,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 November 2022 doi:10.20944/preprints202211.0037.v1

Estimate = 0.11, *SE* = 0.03, *p* = .01, and the increase was retained in the delayed post-test, Estimate = 0.10, *SE* = 0.03, *p* = .01. The CTRL group did not change significantly before and after training (*p*s > .1). For the F2-tuned RWA coefficients, we also observed a significant interaction effect for Group and Test, $F(4,114) = 3.22$, $p = .02$ (Figure 11). Specifically, the MT group significantly increased their use of F2 in the immediate post-test, Estimate = 0.09, *SE* = 0.03, *p* = .002, and in the delayed post-test compared to the pre-test, Estimate = 0.10, *SE* = 0.03, *p* < .001. For the ST group, the weight of F2 also significantly increased in the immediate post-test, Estimate = 0.09, *SE* = 0.03, *p* = .005, and in the delayed post-test, Estimate = 0.11, *SE* = 0.03, *p* < .001. By contrast, there was no significant change for the CTRL group (*p*s > .1). For the duration-tuned RWA coefficients, the results also showed a significant interaction effect for Group and Test, $F(4,114) = 5.49$, $p < .001$. Pairwise contrasts showed that for the MT group, there was a significant decrease in the use of duration in the immediate post-test, Estimate = -0.14, *SE* = 0.05, *p* = .007, and the weight decreased to a greater extent in the delayed post-test compared to the pre-test, Estimate = -0.25, *SE* = 0.05, *p* < .001. For the ST group, the weight of duration also significantly decreased in the immediate post-test, Estimate = -0.20, *SE* = 0.05, *p* < .001, and the decrease retained in the delayed post-test, Estimate = -0.21, *SE* = 0.05, *p* < .001. No significant change was found in the CTRL group (*p*s > .1). These results indicated that the MT and ST group similarly shifted their cue weighting in production with more reliance on F1 and F2 and less on vowel duration, and crucially, the cue reweighting was retained three months after training.

Table 4 Results of Type III analysis of variance on the RWA coefficients of F1, F2, and vowel duration measured in production with Satterthwaite approximation for degrees of freedom

Fixed effect	*Sum Sq*	*Mean Sq*	*NumDF*	*DenDF*	*F* value	*p* value
Final Model: F1 ~ Group * Test + (1 | Subject)						
Group (MT, ST, CTRL)	0.04	0.02	2	57	1.65	0.20
Test (pre-test, post-test1, post-test2)	0.30	0.15	2	114	12.20	< .001
Group:Test	0.15	0.04	4	114	3.11	0.02
Final Model: F2 ~ Group * Test + (1 | Subject)						
Group (MT, ST, CTRL)	0.02	0.01	2	57	1.38	0.26
Test (pre-test, post-test1, post-test2)	0.18	0.09	2	114	12.46	< .001
Group:Test	0.09	0.02	4	114	3.22	0.02
Final Model: Duration ~ Group * Test + (1 | Subject)						
Group (MT, ST, CTRL)	0.30	0.15	2	57	7.43	.001
Test (pre-test, post-test1, post-test2)	0.73	0.37	2	114	18.04	< .001
Group:Test	0.45	0.11	4	114	5.49	< .001

Correlations between Perception and Production

In order to better understand how speech perception and production are related, we used Pearson's product-moment correlation to examine correlations between the participants' cue weights of F1, F2, and duration measured in the perception and production data (Fig.7). The results showed that before training there was only a marginally significant positive correlation between perception and production on the duration dimension ($r = 0.25$, $p = .05$). In the immediate post-test, we found significant positive correlations between perception and production on all of the three dimensions ($r = 0.53$, $p < .001$ for F1,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 November 2022 doi:10.20944/preprints202211.0037.v1

$r = 0.48$, $p < .001$ for F2, $r = 0.41$, $p < .001$ for duration). These significant positive correlations were retained in the delayed post-test ($r = 0.37$, $p = .003$ for F1, $r = 0.52$, $p < .001$ for F2, $r = 0.52$, $p < .001$ for duration).

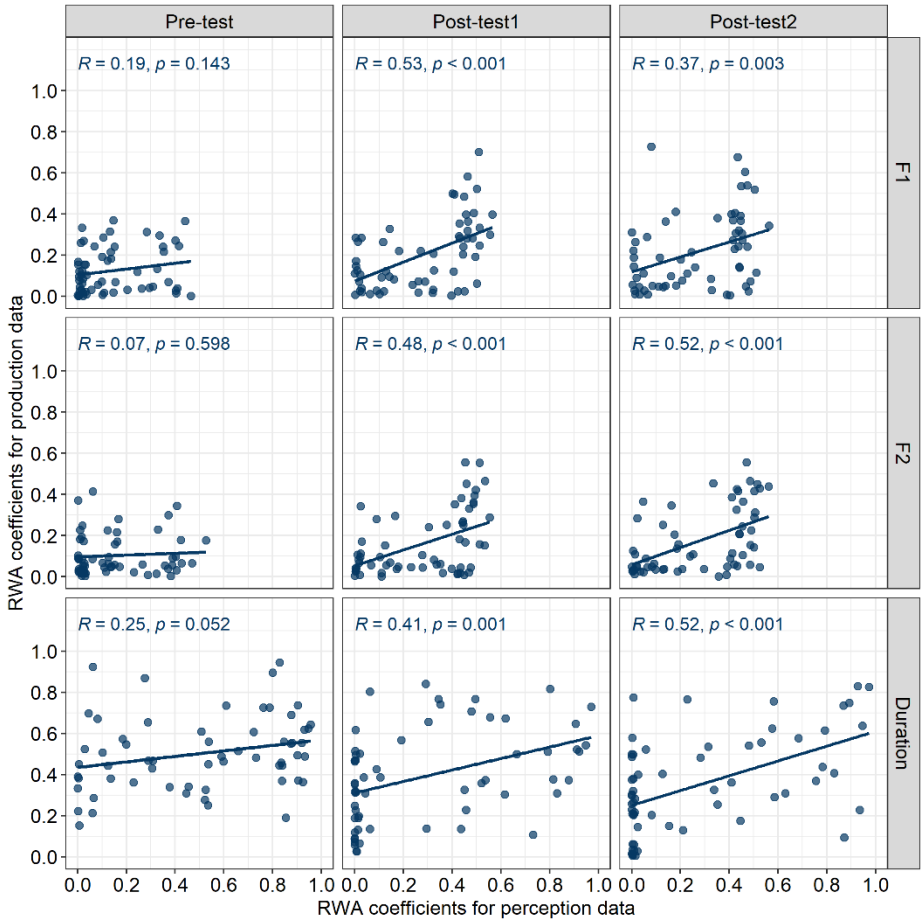


Fig. 7. Scatter plots for correlations between perception and production in terms of the RWA coefficients of F1, F2, and vowel duration measured in the pre-test, immediate post-test (Post-test1), and delayed post-test (Post-test2).

Impacts of Cognitive factors on Training

To examine the influences of individual learners' working memory capacity and selective attention on the training outcomes, the model for word identification accuracy included the fixed effects of Group, Test, Word Type, Working Memory Accuracy, Flanker Score, and their interactions. The results of Type III analysis of variance for main effects are summarized in Table 5 (see Appendix C for the summary output for the final model). We found a significant interaction effect for Group, Test, and Flanker Score, $F(2,2312) = 7.89$, $p < .001$, and a significant interaction effect for Group, Test, and Working Memory Accuracy, $F(2,2312) = 4.02$, $p = .02$. For the reason of brevity, we only reported the results of comparisons pertaining to the second research question.

Table 5 Results of Type III analysis of variance on the word identification accuracy with Satterthwaite approximation for degrees of freedom, with cognitive factors added as the fixed effects (Flanker_c represents centered flanker score, WM_c represents log-transformed and centered working memory accuracy)

Fixed effect	*Sum Sq*	*Mean Sq*	*NumDF*	*DenDF*	*F* value	*p* value
Final Model: Accuracy ~ Group * Test * Type * (Flanker_c+WM_c) + (1 | Subject) + (1 | Word)						

 doi:10.20944/preprints202211.0037.v1

Group (MT vs. ST)	0.05	0.05	1	34.35	0.72	0.40
Test (pre-test, post-test1, post-test2)	10.26	5.13	2	2312	79.67	< .001
Type (trained vs. untrained)	0.03	0.03	1	18.03	0.43	0.52
Flanker_c	0.27	0.27	1	34.35	4.24	0.05
WM_c	0.14	0.14	1	34.35	2.17	0.15
Group:Test	0.45	0.23	2	2312	3.50	0.03
Group:Type	0.09	0.09	1	2312	1.46	0.23
Test:Type	0.20	0.10	2	2312	1.59	0.20
Group:Flanker_c	< 0.01	< 0.01	1	34.35	0.02	0.88
Group:WM_c	0.05	0.05	1	34.35	0.74	0.40
Test:Flanker_c	0.80	0.40	2	2312	6.21	.002
Test: WM_c	1.85	0.92	2	2312	14.33	< .001
Type:Flanker_c	0.01	0.01	1	2312	0.10	0.76
Type: WM_c	< 0.01	< 0.01	1	2312	< 0.01	0.99
Group:Test:type	0.03	0.02	2	2312	0.26	0.77
Group:Test:Flanker_c	**1.02**	**0.51**	**2**	**2312**	**7.89**	**< .001**
Group:Test: WM_c	**0.52**	**0.26**	**2**	**2312**	**4.02**	**0.02**
Group:Type:Flanker_c	0.01	0.01	1	2312	0.09	0.76
Group:Type: WM_c	0.02	0.02	1	2312	0.28	0.60
Test:Type:Flanker_c	0.14	0.07	2	2312	1.11	0.33
Test:Type: WM_c	0.01	< 0.01	2	2312	0.04	0.96
Group:Test:Type:Flanker_c	0.04	0.02	2	2312	0.30	0.74
Group:Test:Type: WM_c	0.04	0.02	2	2312	0.28	0.75

To help interpret the three-way interaction effects, we divided the participants into three sub-groups of the low-, medium-, and high-level working memory capacity for the MT and ST groups respectively, with relatively equal sample sizes according to their working memory accuracy. For the MT group, the participants with relatively low-level working memory did not show significant improvement in the immediate post-test, Estimate = 0.05, *SE* = 0.03, $p = 1$, and in the delayed post-test, Estimate = -0.01, *SE* = 0.03, $p = 1$. In contrast, the medium-level participants improved significantly in the immediate post-test, Estimate = 0.13, *SE* = 0.03, $p < .001$, and a marginally significant difference between the delayed post-test and the pretest was found for these medium-level participants, Estimate = 0.09, *SE* = 0.03, $p = .05$. The high-level learners improved significantly in the immediate post-test, Estimate = 0.29, *SE* = 0.03, $p < .001$, and furthermore a significant improvement in the delayed post-test relative to the pre-test was found for these high-level participants, Estimate = 0.16, *SE* = 0.03, $p < .001$. But note that there was still a significant difference between the performance in the immediate post-test and the delayed post-test for these high-level learners in the MT group, Estimate = -0.13, *SE* = 0.03, $p = .004$. For the ST group, different from the MT group, the participants with relatively low-level working memory improved significantly in the immediate post-test, Estimate = 0.16, *SE* = 0.03, $p < .001$, and the improvement was retained in the delayed post-test, Estimate = 0.15, *SE* = 0.03, $p < .001$. The medium-level learners, surprisingly, did not improve significantly in

 doi:10.20944/preprints202211.0037.v1

the immediate post-test, Estimate = 0.11, *SE* = 0.04, *p* = .22, and in the delayed post-test compared to the pre-test, Estimate = 0.10, *SE* = 0.04, *p* = .34. The high-level learners improved significantly in the immediate post-test, Estimate = 0.22, *SE* = 0.03, *p* < .001, and the improvement was retained in the delayed post-test, Estimate = 0.17, *SE* = 0.03, *p* < .001.

In the same vein, we divided the participants in each training group into three levels of selective attention ability according to their flanker scores. For the MT group, the participants with relatively low-level selective attention improved significantly in the immediate post-test, Estimate = 0.18, *SE* = 0.03, *p* < .001, but the improvement was not retained in the delayed post-test, Estimate = 0.06, *SE* = 0.03, *p* = 1. It bears noting here that a significant difference between performance in the immediate and delayed post-test was found for the low-level learners, Estimate = -0.12, *SE* = 0.03, *p* = .02. The medium-level learners also improved significantly in the immediate post-test, Estimate = 0.12, *SE* = 0.03, *p* = .002, and similarly the improvement was not retained in the delayed post-test, Estimate = 0.07, *SE* = 0.03, *p* = .62. But different from those low-level learners, there was no significant difference between performance in the immediate post-test and the delayed post-test, Estimate = -0.05, *SE* = 0.03, *p* = .1. Similar pattern was found in the high-level learners in the MT group, in which case they benefited from training in the immediate post-test, Estimate = 0.15, *SE* = 0.03, *p* < .001, and the benefit was not retained in the delayed post-test, Estimate = 0.10, *SE* = 0.03, *p* = .13. There was also no significant difference between performance in the immediate post-test and the delayed post-test, Estimate = -0.05, *SE* = 0.03, *p* = .1. These results suggest that the low-level participants were poorer at retaining training effects compared to those with medium- and high-level selective attention abilities under the MT training condition. Different from the mixed pattern found in the MT group, for the ST group, the participants with different levels of selective attention all benefited from training, Estimate = 0.13, *SE* = 0.03, *p* < .001 for low-level learners, Estimate = 0.17, *SE* = 0.03, *p* < .001 for medium-level learners, Estimate = 0.24, *SE* = 0.04, *p* < .001 for high-level learners. And the benefit was retained in the delayed post-test, Estimate = 0.11, *SE* = 0.03, *p* = .003 for low-level learners, Estimate = 0.15, *SE* = 0.03, *p* < .001 for medium-level learners, Estimate = 0.20, *SE* = 0.04, *p* < .001 for high-level learners.

Discussion

This study examined the role of input variability and cognitive abilities in learning to perceive and produce the English /i/-/ɪ/ contrast by adult native speakers of Mandarin Chinese. We contrasted the multiple-talker (four talkers) versus single-talker (one talker) training conditions together with a control group who did not receive training. Meanwhile, we systematically lengthened the duration of the training stimuli used in both training conditions, which provided greater variability with irrelevant information to the vowel category distinction. By contrast, the training stimuli included variations along the spectral dimensions (i.e., F1 and F2) that were informative to the category membership. In order to assess training effects, we measured the participants' performance in terms of naturally-spoken word identification and cue weighting of three dimensions (F1, F2, and vowel duration) in perception and production at three time points: one week before training (pre-test), one week after training (immediate post-test), and three months after training (delayed post-test). For nonlinguistic cognitive abilities, we measured the trainees' working memory capacity and selective attention prior to training. The results demonstrated that the trained participants exhibited retention of more native-like cue weighting in both perception and production, regardless of talker variability condition. Regarding word identification performance, intriguingly, we found that the single-talker training showed a long-term benefit, whereas the multiple-talker training did not show a comparable retention effect. Our results suggested that the disappearance of high talker variability benefit could be explained by the interaction between the nature of training input, task difficulty, and the trainees' cognitive abilities.

 doi:10.20944/preprints202211.0037.v1

The success of the single-talker training in long-term retention of L2 speech learning in this study suggests an important role that phonetically-irrelevant acoustic variability plays beyond multiple talkers. This finding replicated the result of our previous study (Zhang et al., 2021) that introducing variability along the secondary dimension is sufficient to generalize trained knowledge to new phonetic contexts and novel talkers with a long-term benefit. This long-term benefit is consistent with previous HVPT research demonstrating that knowledge learned during training could result in lasting memory traces even after a period of no exposure (Bradlow et al., 1999; Lively et al., 1994; Wang et al., 1999). Our measures of cue weighting pre- and post-training showed that the training indeed directed the learners' attention away from the irrelevant and uninformative dimension of vowel duration and towards the relevant and consistent dimensions of the spectrum, suggesting that the phonetically-irrelevant acoustic variability is conducive for listeners to figuring out what cues are important for the speech contrast and which are not. The findings support the hypothesis that L2 phonetic learning can be understood as the operation of enhancement of attention increasing attentional weight placed on relevant dimensions (e.g., spectrum) and withdrawal of attention decreasing attentional weight to unimportant dimensions (e.g., vowel duration) (Francis & Nusbaum, 2002; Goudbeek et al., 2008; Iverson et al., 2005; Iverson & Kuhl, 1995; Kuhl & Iverson, 1995). Our findings of significant positive correlations between the word identification accuracy and the primary cue weight and negative correlations between the word identification accuracy and the secondary cue weight provided further evidence for the importance of shifting attentional cue weights. Notably, these correlations were stronger for the untrained words compared to the trained words. These results suggest that given appropriate exposure, learners' attentional weights would be shifted to accommodate the contrast being learned, which then facilitates generalizing knowledge of a trained set to a novel set of speech tokens. Our findings also add to a growing body of evidence for the benefit of variable input during learning in a variety of domains (e.g., Adwan-Mansour & Bitan, 2017; Apfelbaum et al., 2013; Bulgarelli & Weiss, 2021; Fuhrmeister & Myers, 2020; Helsdingen et al., 2011; Potter & Saffran, 2017; Rost & McMurray, 2010). These results converge to suggest that input variability that is irrelevant to the ultimate learning outcome may be a general principle of learning which is applicable to many, if not most, domains. Therefore, a critical learning problem is likely to involve sorting out which cues are relevant to what is being learned and which are not, the process of which can be achieved using principles like variability.

The sensitivity to irrelevant acoustic variability may be the reason for the talker variability benefit that has been generally reported in tasks of generalization and retention. In theory, the talker variability benefit has been interpreted differently. Researchers supporting the abstractionist view assume that listeners normalize talker variability and obtain the generalized information excluding indexical properties, which can then be matched to standardized representation in long-term memory (e.g., Ladefoged & Broadbent, 1957; Liberman, 1973). Alternatively, researchers adopting an exemplar view of speech perception (e.g., Johnson, 1994; Pierrehumbert, 2002) and a non-analytic episodic view (e.g., Goldinger, 1998; Pisoni, 1997) posit that talker-specific information is encoded together with linguistic contents and stored in long-term memory. Nonetheless, these views are not necessarily incompatible. Talker-specific properties may be encoded into representations, which thus form more associative hooks (Barcroft & Sommers, 2005, 2014) and aids subsequent speech processing, evidenced by better identifying tokens produced by encountered talkers (e.g., Nygaard et al., 1994; Palmeri et al., 1993). On the other hand, clusters of consistent information (e.g., phonetic units) may emerge naturally from these encoded and stored idiosyncratic talker-related attributes. Different clusterings could be focused on by directing attention to particular acoustic attributes, which thus helps generalize knowledge to new phonetic contexts and new talkers (Nosofsky, 1989; Werker & Curtin, 2005). Indeed, evidence indicates that attending to phonemic information may entail attending to talker-specific acoustic attributes, even if talker identity is

 doi:10.20944/preprints202211.0037.v1

not attended to, suggesting that phonetic information and talker-specific attributes may share common representations (e.g., Eimas et al., 1978; Green et al., 1997; Mullennix & Pisoni, 1990; Myers & Theodore, 2017; Tremblay et al., 2021). Therefore, it is likely that variation in idiosyncratic dimensions, such as talker-related dimensions, would make those consistent and informative dimensions stand out for learners easily to attend to.

In our view, however, these idiosyncratic variations for highlighting consistent and informative dimensions for robust categorization do not have to be talker-specific. If dissociation of phonetically-irrelevant dimensions underpins the effect of generalization and retention, directly introducing variability along phonetically-irrelevant dimensions would be more precise and helpful than the manipulation of talker variability, due to the fact that specific talker-varying acoustic properties vary across talkers. That is, the assumed phonetically-irrelevant feature that resides in multiple-talker speech may vary in specific groups of talkers. There is admittedly an inconsistent picture of cues that can consistently distinguish talkers (Van Lancker et al., 1985). For example, the most reliably consistent cue in talker attributes is F0, while studies have also reported that formant frequencies (Baumann & Belin, 2010; Murry & Singh, 1980), hoarseness (Singh & Murry, 1978), vowel duration (Murry & Singh, 1980; Singh & Murry, 1978), and shimmer (Kreiman et al., 1992) can be consistent in talkers, at least a fraction of talkers. Therefore, for example, for one group of talkers, vowel duration is consistently varied, while for another group of talkers, vowel duration displays inconsistent variations. In this case, the positive effect of talker variability may be mitigated by different types of acoustic variability that reside in multiple-talker speech. That is, talker variability can be beneficial when it provides acoustic variability in phonetically-irrelevant cues to the contrast being learned. Otherwise, talker variability appears to be not informative for identifying relevant cues. This is even complicated by the fact that some properties which vary idiosyncratically by talker in one language vary phonemically in others (Gordon & Ladefoged, 2001). Therefore, while multiple talkers may provide phonetically-irrelevant acoustic variability, the present findings do not support the notion that multiple talkers are necessary for successful generalization and retention. This premise is helpful to explain why some studies showed a talker variability effect, whereas others reported null effects. It is plausible that different sources of acoustic variability that are not specifically identified in multiple-talker or single-talker speech would lead to generalization and retention outcomes (Johnson et al., 1993; Peterson & Barney, 1952).

The Interaction of Task Complexity and Cognitive Abilities

Despite the fact that a large body of research has observed a positive effect for input variability (For a meta-analysis, see Zhang, Cheng, & Zhang, 2021), our results suggest there may be boundaries to the benefit of input variability. We did not see an additional benefit of multiple-talker speech relative to single-talker speech in our modified HVPT protocol. Particularly, the multiple-talker group did not show comparable retention of naturally-spoken word identification performance compared to the single-talker group, though both of the two trained groups retained their heavier weighting of the primary spectral cues and less weighting of the secondary durational cue. This null result could not be due to our multiple-talker speech material, given that the multiple-talker group who received the canonical HVPT without the additional enhancement features in our previous study confirmed significant benefits for our multiple-talker speech materials (Zhang et al., 2021). These findings appear to contradict earlier training studies that demonstrated the advantage of high talker variability for phonetic learning (e.g., Bradlow et al., 1997, 1999; Lively et al., 1994; Logan et al., 1991). However, our study has some key differences that may explain this inconsistency. First, we increased the amount of acoustic variability in the multiple-talker training by introducing varying degrees of acoustic exaggeration along the secondary dimension of the vowel contrast. It seems that conflating various sources of acoustic variability may interfere with learning, which may be partic-

 doi:10.20944/preprints202211.0037.v1

ularly true when the task at hand is more challenging. The acoustic exaggerations on irrelevant dimensions for L2 may impede this process due to the "Native Language Neural Commitment" that prioritizes the allocation of perceptual attention and processing resources to optimize efficient speech categorization in service of L1 phonology instead of L2 (Zhang et al., 2005, 2009). Given that we did find a long-term effect of our multiple-talker training on the synthetic phoneme identification task, it is possible that the identification of naturally-produced words was more difficult because it might have incurred an additional load for processing lexical information (Escudero et al., 2008) and semantic content (Guion & Pederson, 2007). When exposed to naturally-produced words, listeners may draw upon rich variations of information present in the tokens to make perceptual judgments. When exposed to stimuli drawn from a synthetic continuum, they may direct their perceptual processing to precisely those dimensions along which the stimuli vary. Evidence indeed shows that task difficulty may interact with input variability (Fuhrmeister & Myers, 2020; Goldinger et al., 1991). For example, Goldinger et al. (1991) found that single-talker lists produced better word recall than multiple-talker lists at short inter-word intervals, whereas this effect was reversed for longer inter-word intervals, suggesting that even for adults remembering L1 words, the nature of the task may place a burden on processing high talker variability.

Further evidence for this processing difficulty comes from our findings that the learners with higher working memory capacities and selective attention were more likely to benefit from the multiple-talker training in terms of word identification performance, compared to those with relatively lower capacities. Critically, we did not see this interaction in cue weighting measured in the synthetic phoneme identification task for the multiple-talker group or in any measures for the single-talker group (the failure to benefit from the ST training was more likely due to the high accuracy before training for the participants with medium-level of working memory). These results suggest that the potential benefit of increased variability in the input appears to be outweighed by the processing cost, especially in terms of retention benefit, which may place major demands on learners' cognitive resources. More specifically, the participants with low-level working memory capacity were poorer at both the immediate improvement and the retention of improvement compared to those medium- and high-level participants, while the participants with low-level selective attention were poorer at retaining training improvement compared to those medium- and high-level participants. That is, working memory appears to influence both immediate improvement and long-term retention, while selective attention is more likely to play a role in long-term benefit. These results come as no surprise that a larger working memory capacity is suggested to enhance the quality of input (Darcy et al., 2015; Segalowitz et al., 2009). That is, larger working memory may allow the learners more time to process and learn from the input by maintaining longer access to it, and better storage quality might promote more accurate perception and learning (Goldstone, 1998). Furthermore, studies have shown that individuals with lower working memory exhibit a relatively reduced ability to attend to task-relevant dimensions relative to those with higher working memory (D'Esposito & Postle, 2015; Unsworth & Robison, 2017; Wöstmann & Obleser, 2016), suggesting the potential overlap in measuring the construct of working memory and selective attention.

Our findings regarding the impact of nonlinguistic cognitive abilities complement previous training studies showing that successful learning depends on an interaction between individual differences in perceptual abilities and the training protocol (e.g., Antoniou & Wong, 2015; Fuhrmeister & Myers, 2020; Perrachione et al., 2011; Sadakata & Mcqueen, 2014). For example, Fuhrmeister et al. (2020) found that the participants with higher pre-training discrimination abilities went on to learn the contrast more robustly, and crucially related to our findings, this relationship between aptitude and identification performance was stronger for the interleaved training group than the blocked training group, indicating that the higher-aptitude participants were better to take advantage of the training condition with greater variability. Antoniou and Wong (2015) attributed this

 doi:10.20944/preprints202211.0037.v1

learner by training interaction to the reduced availability of cognitive resources in low-aptitude participants in the mixed-talker training condition. Our results extend these previous findings by a clear interaction of input variability in the training protocol and individual differences in nonlinguistic cognitive abilities, which may manifest to a greater extent in more challenging tasks. Actually, evidence has shown a significantly high correlation between language aptitude and performance measured by the Backward Digit Span Task (which we adopted in this study) to measure working memory capacity (Kormos & Sáfár, 2008), suggesting that working memory capacity may be considered as a learner trait to influence outcome just like the construct of language aptitude.

When viewed across the scientific literature beyond phonetic learning, our results also agree very well with a broader literature on how a learner's cognitive abilities might constrain their ability to benefit from input variability (Raviv et al., 2022). For example, there is both empirical and computational evidence that encountering grammatical morphemes across a broader range of vocabulary, that is, high lexical variability, promotes generalization (Bybee, 1995; Plunkett & Marchman, 1991; Wonnacott et al., 2012). Particularly, Brooks et al. (2006) further examined this effect in English-speaking adult learners who were exposed to an unfamiliar L2 and showed that greater variability facilitated generalization only in learners with above-median scores on an intelligence test which is shown to correlate highly with working memory capacity (Colom et al., 2006). Taken together, it appears that striking a delicate balance between input variability and an appropriate level of task difficulty for individuals with certain abilities would be critical to engendering effective generalization and retention of perceptual learning.

Perceptual Transfer to Production

Our acoustic measures of production data showed that both the multiple-talker and single-talker group significantly increased their use of the critical spectral cues and decreased the use of the secondary durational cue, which was retained three months after training, aligning with the cue reweighting pattern found in their perception data. This perceptual transfer to production is consistent with previous studies showing that perceptual training can lead to more native-like changes in speech production (Bradlow et al., 1997; Rochet, 1995; Wang et al., 2003). In addition to the training-induced changes in production, we also assessed correlations between perception and production in terms of cue weighting. Prior to training, there were no significant correlations on any of the three dimensions, whereas significant positive correlations between perception and production on all of the three dimensions were found in the immediate post-test as well as the delayed post-test, suggesting that the relationship between perception and production changes as a result of the phonetic training. Our findings are consistent with previous studies showing that the perception-production link is time-varying (e.g., Jia et al., 2006; Rallo Fabra & Romero, 2012). For example, examining Mandarin speakers' perception and production of English vowels, Jia et al. (2006) compared three groups: foreign language learners in China, second language learners who had lived in the US for less than 2 years, and another group of second language learners who had lived in the US for 3-5 years. While the correlations for the foreign language learners and past arrivals were of similar magnitude (r = 0.42 and r = 0.46, respectively), the perception-production link was far weaker for recent arrivals (r = 0.25). Although the goal of Jia et al. (2006) was to examine the age and experience-related changes in perception and production of L2 sounds, the between-group differences in the strength of correlations could be interpreted as evidence of a change in the perception-production link. In fact, converging evidence shows that these correlations appear to be limited to proficient speakers of the language (Flege, 1999).

With respect to the perception-production link in terms of cue weighting, previous work actually has failed to detect correlations between the two modalities (Bohn & Flege, 1997; Schertz et al., 2015; Shultz et al., 2012). For example, Schertz et al. (2015) examined Koreans' production and perception of L1 Korean stop contrast and L2 English stop contrast and found that there was no correlation between individuals' use of any of the three

 doi:10.20944/preprints202211.0037.v1

cues across production and perception in either L1 and L2. Crucially, there appeared to be much more variability in native Korean listeners' perceptual cue weights for the L2 English stop contrast than there was in their productions of the English contrast. From a dynamic developmental view of the relationship, our findings are not contradictory to the previous null results. As hypothesized by Nagle and Baese-Berk (2021), during a period when perception and production improve relatively quickly, a large correlation between perception and production measures might be observed, with the correlation increasing in strength over time. Once one of the two modalities begins to stabilize, entering a developmental plateau, the perception-production link might also stabilize such that no strong correlation would be observed. This hypothesis converges with evidence suggesting that cross-sectional studies may over- or underestimate the perception-production link depending on the precise stage at which L2 learners are measured, which may result in the truncated observation of the link. The current state of perception-production research has been informed by a large number of cross-sectional studies, which are not well-suited to capture the time-varying nature of the link (Nagle, 2021). Future research on the perception-production link then may require a more developmental/longitudinal approach that takes into consideration the stage, strength, and duration of the relationship.

Limitations and Future Directions

The current study has several limitations. The primary limitation relates to the brief period of the training program which provided the trainees with only 60-90 minutes of seven training sessions. Other studies of L2 phonetic learning have involved more intense training, typically 10 to 15 hours of training (e.g., Lively et al., 1994; Nishi & Kewley-Port, 2007; Zhang et al., 2009). The limited benefit of multiple-talker speech observed in our study thus may reflect the brevity of our training program rather than the stimulus conditions *per se*, given recent evidence suggesting that learners may require more exposure to multiple talkers before talker-specific learning can be observed (Luthra et al., 2021). A second limitation concerns the measure of cognitive abilities. It is clear from the literature that working memory is not a unitary construct, while we used only one single working memory score per participant. Since previous studies suggest that the subcomponent, the Phonological Loop or more specifically phonological short-term memory, may be more related to language learning (e.g., Baddeley et al., 1998; Speciale et al., 2004), the use of more sophisticated or a wider range of measures may provide a much more nuanced understanding of the role of working memory capacity in L2 learning. A third limitation is that our participants exhibited a relatively high level of working memory and selective attention. Future studies can be conducted to take into account a wider range of individual L2 cognitive abilities to avoid influences from potential ceiling effects.

Even with these caveats, our data unequivocally confirmed the importance of input variability for generalization and retention of perceptual learning, and provided evidence that the benefit of input variability interacts with other variables such as task difficulty and individual learner's cognitive abilities. The ultimate goal of this line of research is to pull together and optimize resources on training conditions that are most conducive to yielding robust learning outcomes. Our results provide insights into the specific hypotheses that can be tested in future studies. For example, how much acoustic variability is necessary or sufficient for successful learning is still an open question. Particularly, the use of talker numbers as a measure of the amount of talker variability may prevent a precise exploration of why some studies show a talker variability effect while others do not. That said, knowing how many talkers are in the input does not necessarily indicate how many distinct talkers a listener actually hears, and it does not inform the acoustic variability provided by these talkers (Bulgarelli et al., 2021). It is thus of significance to specify this operationalization of talker variability in acoustic and perceptual spaces in future research. Moreover, given that acoustic variability can be characterized by relative variations in primary and secondary cues to the contrast being learned, exploring what prop-

erties underlie acoustic variability has great implications for ways to systematically manipulate this variability in cases where it may be beneficial to learning. It will also be necessary to distinguish between different roles that various sources of acoustic variability may play, together or separately, for the learner with different language experiences.

Conclusions

Taken together, we found evidence that phonetically-irrelevant acoustic variability in single-talker speech is conducive to L2 speech learning in terms of generalization and retention of trained knowledge. We did not observe an additional benefit of multiple-talker speech in this modified HVPT program, and crucially, we did not find the long-term benefit to natural word identification under the multiple-talker training condition. Given the brevity of our training protocol, we attribute the null results to the additional processing cost incurred by the enhanced variability in the multiple-talker training stimuli as the trainees with varying levels of cognitive abilities benefited to a different extent from the multiple-talker training measured in terms of their gains in natural word identification. Importantly, our findings reveal a clear interaction between learners' cognitive abilities, task difficulty, and the amount of input variability, highlighting the need for careful considerations of how specific combinations of input type and amount, task difficulty, and learner cognitive profile may lead to efficient and robust learning outcomes.

Acknowledgements

The research was supported by a grant from the National Social Science Fund of China (15BYY005). YZ additionally received support from University of Minnesota's Grand Challenges Exploratory Research Grant and Brain Imaging Grant to work on the project.

Data availability statement

All raw data files and analysis codes in this study are publicly available via Open Science Framework at **https://osf.io/e8sn6/**.

CRediT authorship contribution statement

Xiaojuan Zhang: Formal Analysis, Software, Investigation, Visualization, Writing – original draft. Bing Cheng: Conceptualization, Data Curation, Resources, Writing – review & editing, Supervision, Project administration, Funding acquisition. Yang Zhang: Conceptualization, Software, Visualization, Writing – review & editing, Supervision, Funding Acquisition.

References

Adams, A. M., & Guillot, K. (2008). Working memory and writing in bilingual students. *International Journal of Applied Linguistics*, *156*, 13–28. https://doi.org/10.2143/ITL.156.0.2034417

Adwan-Mansour, J., & Bitan, T. (2017). The effect of stimulus variability on learning and generalization of reading in a novel script. *Journal of Speech, Language, and Hearing Research*, *60*(10), 2840–2851. https://doi.org/10.1044/2017_JSLHR-L-16-0293

Aha, D. W., & Goldstone, R. L. (1990, July). *Learning attribute relevance in context in instance-based learning algorithms*. Proceedings of the Twelfth Annual Conference of the Cognitive Science Society (pp. 141-148). Hillsdale, New Jersey.

Aliaga-García, C., Mora, J. C., & Cerviño-Povedano, E. (2011). L2 speech learning in adulthood and phonological short-term memory. *Poznań Studies in Contemporary Linguistics*, *47*(1), 1–14. https://doi.org/10.2478/psicl-2011-0002

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*(2), 571–5744. https://doi.org/10.1121/1.4923362

Antoniou, M., Wong, P. C. M., & Wang, S. (2015). The effect of intensified language exposure on accommodating talker variability. *Journal of Speech, Language, and Hearing Research*, *58*(3), 722–727. PubMed. https://doi.org/10.1044/2015_JSLHR-S-14-0259

Apfelbaum, K. S., Hazeltine, E., & McMurray, B. (2013). Statistical learning in reading: Variability in irrelevant letters helps children learn phonics skills. *Developmental Psychology*, *49*(7), 1348–1365. https://doi.org/10.1037/a0029839

Apfelbaum, K. S., & McMurray, B. (2011). Using variability to guide dimensional weighting: Associative mechanisms in early word learning. *Cognitive Science*, *35*(6), 1105–1138. PubMed. https://doi.org/10.1111/j.1551-6709.2011.01181.x

Baddeley, A. D. (1986). *Working Memory*. Oxford University Press.

Baddeley, A. D., & Hitch, G. (1974). Working memory. In G. A. Bower (Ed.), *The Psychology of Learning and Motivation* (pp. 47–89). Academic Press.

Baddeley, A., Gathercole, S., & Papagno, C. (1998). The phonological loop as a language learning device. *Psychological Review*, *105*(1), 158–173. PubMed. https://doi.org/10.1037/0033-295x.105.1.158

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

Barcroft, J., & Sommers, M. S. (2014). Effects of variability in fundamental frequency on L2 vocabulary learning: A comparison between learners who do and do not speak a tone language. *Studies in Second Language Acquisition*, *36*(3), 423–449. https://doi.org/10.1017/S0272263113000582

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*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Baumann, O., & Belin, P. (2010). Perceptual scaling of voice identity: Common dimensions for different vowels and speakers. *Psychological Research*, *74*(1), 110–120. https://doi.org/10.1007/s00426-008-0185-z

Best, C. T. (1995). A direct realist view of cross-language speech perception. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 171–204). York Press.

Boersma, P., & Weenink, D. (2016). *Praat: Doing phonetics by computer* (6.0.22). http://www.praat.org/

Bohn, O.-S. (1995). Cross-language speech perception in adults: First language transfer doesn't tell it all. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 279–304). York Press.

Bohn, O.-S., & Flege, J. E. (1997). Perception and production of a new vowel category by adult second language learners. In A. James & J. Leather (Eds.), *Second-Language Speech: Structure and Process* (pp. 53–74). de Gruyter.

Bourne, L. E., & Restle, F. (1959). Mathematical theory of concept identification. *Psychological Review*, *66*(5), 278–296. https://doi.org/10.1037/h0041365

Bradlow, A. R., Akahane-Yamada, R., Pisoni, D. B., & Tohkura, Y. (1999). Training Japanese listeners to identify English /r/ and /l/: Long-term retention of learning in perception and production. *Perception & Psychophysics*, *61*(5), 977–985. https://doi.org/10.3758/BF03206911

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*(4), 2299–2310. https://doi.org/10.1121/1.418276

Brekelmans, G., Lavan, N., Saito, H., Clayards, M., & Wonnacott, E. (2022). Does high variability training improve the learning of non-native phoneme contrasts over low variability training? A replication. *Journal of Memory and Language*, *126*, 104352. https://doi.org/10.1016/j.jml.2022.104352

Brooks, P. J., Kempe, V., & Sionov, A. (2006). The role of learner and input variables in learning inflectional morphology. *Applied Psycholinguistics*, *27*(2), 185–209. https://doi.org/10.1017/S0142716406060243

Brosseau-Lapré, F., Rvachew, S., Clayards, M., & Dickson, D. (2013). Stimulus variability and perceptual learning of nonnative vowel categories. *Applied Psycholinguistics*, *34*(3), 419–441. https://doi.org/10.1017/S0142716411000750

Bulgarelli, F., Mielke, J., & Bergelson, E. (2021). Quantifying talker variability in North-American infants' daily input. *Cognitive Science*, *46*(1), e13075. https://doi.org/10.1111/cogs.13075

Bulgarelli, F., & Weiss, D. J. (2021). Desirable difficulties in language learning? How talker variability impacts artificial grammar learning. *Language Learning*, *71*(4), 1085–1121. https://doi.org/10.1111/lang.12464

Bush, R. R., & Mosteller, F. (2006). A Model for Stimulus Generalization and Discrimination. In S. E. Fienberg & D. C. Hoaglin (Eds.), *Selected Papers of Frederick Mosteller* (pp. 235–250). Springer New York. https://doi.org/10.1007/978-0-387-44956-2_13

Bybee, J. (1995). Regular morphology and the lexicon. *Language and Cognitive Processes*, *10*(5), 425–455. https://doi.org/10.1080/01690969508407111

Carlet, A. (2017). *L2 Perception and Production of English Consonants and Vowels by Catalan Speakers: The Effects of Attention and Training Task in A Cross-training Study* [Doctoral Dissertation]. Universitat Autonoma de Barcelona.

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*(6), 3703–3716. https://doi.org/10.1121/1.4937612

Cheng, B., Zhang, X., Fan, S., & Zhang, Y. (2019). The role of temporal acoustic exaggeration in high variability phonetic Training: A behavioral and ERP study. *Frontiers in Psychology*, *10*, 1178. https://doi.org/10.3389/fpsyg.2019.01178

Cheng, B., & Zhang, Y. (2013). Neural plasticity in phonetic training of the /i-I/ contrast for adult Chinese speakers. *The Journal of the Acoustical Society of America*, *134*(5), 4245. https://doi.org/10.1121/1.4831610

Cheung, H. (1996). Non-word span as a unique predictor of second language vocabulary learning. *Developmental Psychology*, *32*, 867–873. https://doi.org/10.1037/0012-1649.32.5.867

Colom, R., Rebollo, I., Abad, F. J., & Shih, P. C. (2006). Complex span tasks, simple span tasks, and cognitive abilities: A reanalysis of key studies. *Memory & Cognition*, *34*(1), 158–171. https://doi.org/10.3758/BF03193395

Colom, R., Shih, P. C., Flores-Mendoza, C., & Quiroga, M. Á. (2006). The real relationship between short-term memory and working memory. *Memory*, *14*(7), 804–813. https://doi.org/10.1080/09658210600680020

Conway, A. R. A., & Engle, R. W. (1996). Individual differences in working memory capacity: More evidence for a general capacity theory. *Memory*, *4*(6), 577–590. https://doi.org/10.1080/741940997

Conway, A. R. A., Tuholski, S. W., Shisler, R. J., & Engle, R. W. (1999). The effect of memory load on negative priming: An individual differences investigation. *Memory & Cognition*, *27*(6), 1042–1050. https://doi.org/10.3758/BF03201233

Daneman, M., & Green, I. (1986). Individual differences in comprehending and producing words in context. *Journal of Memory and Language*, *25*(1), 1–18. https://doi.org/10.1016/0749-596X(86)90018-5

Darcy, I., Park, H., & Yang, C.-L. (2015). Individual differences in L2 acquisition of English phonology: The relation between cognitive abilities and phonological processing. *Learning and Individual Differences*, *40*, 63–72. https://doi.org/10.1016/j.lindif.2015.04.005

De Diego-Balaguer, R., & Lopez-Barroso, D. (2010). Cognitive and neural mechanisms sustaining rule learning from speech. *Language Learning*, *60*(s2), 151–187. https://doi.org/10.1111/j.1467-9922.2010.00605.x

Deng, Z., Chandrasekaran, B., Wang, S., & Wong, P. C. M. (2018). Training-induced brain activation and functional connectivity differentiate multi-talker and single-talker speech training. *Neurobiology of Learning & Memory*, *151*, 1–9. https://doi.org/10.1016/j.nlm.2018.03.009

D'Esposito, M., & Postle, B. (2015). The cognitive neuroscience of working memory. *Annual Review of Psychology*, *66*(1), 115–142. https://doi.org/10.1146/annurev-psych-010814-015031

Dong, H., Clayards, M., Brown, H., & Wonnacott, E. (2019). The effects of high versus low talker variability and individual aptitude on phonetic training of Mandarin lexical tones. *PeerJ*, *7*(6), e7191. https://doi.org/10.7717/peerj.7191

Dorman, M. F., Studdert-Kennedy, M., & Raphael, L. J. (1977). Stop-consonant recognition: Release bursts and formant transitions as functionally equivalent, context-dependent cues. *Perception & Psychophysics*, *22*(2), 109–122. https://doi.org/10.1121/1.2003596

Eimas, P. D., Tartter, V. C., Miller, J. L., & Keuthen, N. J. (1978). Asymmetric dependencies in processing phonetic features. *Perception & Psychophysics*, *23*(1), 12–20. https://doi.org/10.3758/BF03214289

Ellis, N. C. (2006). Selective attention and transfer phenomena in L2 acquisition: Contingency, cue competition, salience, interference, overshadowing, blocking, and perceptual learning. *Applied Linguistics*, *27*(2), 164–194. https://doi.org/10.1093/applin/aml015

Engle, R. W. (2002). Working memory capacity as executive attention. *Current Directions in Psychological Science*, *11*(1), 19–23. https://doi.org/10.1111/1467-8721.00160

Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & Psychophysics*, *16*(1), 143–149. https://doi.org/10.3758/BF03203267

Escudero, P., Benders, T., & Lipski, S. C. (2009). Native, non-native and L2 perceptual cue weighting for Dutch vowels: The case of Dutch, German, and Spanish listeners. *Journal of Phonetics*, *37*(4), 452–465. https://doi.org/10.1016/j.wocn.2009.07.006

Escudero, P., & Boersma, P. (2004). Bridging the gap between L2 speech perception research and phonological theory. *Studies in Second Language Acquisition*, *26*(4), 551–585. https://doi.org/10.1017/S0272226310400021

Escudero, P., Hayes-Harb, R., & Mitterer, H. (2008). Novel second-language words and asymmetric lexical access. *Journal of Phonetics*, *36*(2), 345–360. https://doi.org/10.1016/j.wocn.2007.11.002

Felser, C., & Roberts, L. (2007). Processing wh-dependencies in a second language: A cross-modal priming study. *Second Language Research*, *23*(1), 9–36. https://doi.org/10.1177/0267658307071600

Flege, J. E. (1995a). Second language speech learning: Theory, findings, and problems. In W. Strange (Ed.), *Speech perception and language experience: Issues in cross-language research* (pp. 233–277). York Press.

Flege, J. E. (1995b). Two procedures for training a novel second language phonetic contrast. *Applied Psycholinguistics*, *16*. https://doi.org/10.1017/S0142716400066029

Flege, J. E. (1999). Age of learning and second-language speech. In D. P. Birdsong (Ed.), *Second Language Acquisition and the Critical Period Hypothesis* (pp. 101–131). Erlbaum.

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*(4), 437–470. https://doi.org/10.1006/jpho.1997.0052

Flege, J. E., & MacKay, I. R. A. (2004). Perceiving vowels in a second language. *Studies in Second Language Acquisition*, *26*(1), 1–34. https://doi.org/10.1017/S0272263104026117

Fowler, C. A. (1986). An event approach to the study of speech perception from a direct–realist perspective. *Journal of Phonetics*, *14*(1), 3–28. https://doi.org/10.1016/S0095-4470(19)30607-2

Francis, A. L., Baldwin, K., & Nusbaum, H. C. (2000). Effects of training on attention to acoustic cues. *Perception & Psychophysics*, *62*(8), 1668–1680. https://doi.org/10.3758/BF03212164

Francis, A. L., Kaganovich, N., & Driscoll-Huber, C. (2008). Cue-specific effects of categorization training on the relative weighting of acoustic cues to consonant voicing in English. *The Journal of the Acoustical Society of America*, *124*(2), 1234–1251. https://doi.org/10.1121/1.2945161

Francis, A. L., & Nusbaum, H. C. (2002). Selective attention and the acquisition of new phonetic categories. *Journal of Experimental Psychology: Human Perception and Performance*, *28*(2), 349. https://doi.org/10.1037/0096-1523.28.2.349

French, L. M., & O'Brien, I. (2008). Phonological memory and children's second language grammar learning. *Applied Psycholinguistics*, *29*, 463–487. https://doi.org/10.1017/S0142716408080211

Fuhrmeister, P., & Myers, E. B. (2017). Non-native phonetic learning is destabilized by exposure to phonological variability before and after training. *The Journal of the Acoustical Society of America*, *142*(5), EL448–EL454. https://doi.org/10.1121/1.5009688

Fuhrmeister, P., & Myers, E. B. (2020). Desirable and undesirable difficulties: Influences of variability, training schedule, and aptitude on nonnative phonetic learning. *Attention, Perception, & Psychophysics*, *82*(4), 2049–2065. https://doi.org/10.3758/s13414-019-01925-y

Galle, M. E., Apfelbaum, K. S., & McMurray, B. (2015). The role of single talker acoustic variation in early word learning. *Language Learning and Development*, *11*(1), 66–79. https://doi.org/10.1080/15475441.2014.895249

Gathercole, S. E., & Alloway, T. P. (2008). *Working Memory and Learning: A Practical Guide for Teachers*. Sage.

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*(1), e3209. https://doi.org/10.7717/peerj.3209

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, *105*, 251–279. https://doi.org/10.1037/0033-295x.105.2.251

Goldinger, S. D., Pisoni, D. B., & Logan, J. S. (1991). On the nature of talker variability effects on recall of spoken word lists. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *17*(1), 152–162. PubMed. https://doi.org/10.1037//0278-7393.17.1.152

Goldstone, R. L. (1993). Feature distribution and biased estimation of visual displays. *Journal of Experimental Psychology: Human Perception and Performance*, *19*(3), 564–579. https://doi.org/10.1037/0096-1523.19.3.564

Goldstone, R. L. (1994). Influences of categorization on perceptual discrimination. *Journal of Experimental Psychology: General*, *123*(2), 178–200. https://doi.org/10.1037//0096-3445.123.2.178

Goldstone, R. L. (1998). Perceptual learning. *Annual Review of Psychology*, *49*(1), 585–612. https://doi.org/10.1007/978-1-4419-1428-6_147

Gordon, M., & Ladefoged, P. (2001). Phonation types: A cross-linguistic overview. *Journal of Phonetics*, *29*(4), 383–406. https://doi.org/10.1006/jpho.2001.0147

Goudbeek, M., Cutler, A., & Smits, R. (2008). Supervised and unsupervised learning of multidimensionally varying non-native speech categories. *Speech Communication*, *50*(2), 109–125. https://doi.org/10.1016/j.specom.2007.07.003

Green, K. P., Tomiak, G. R., & Kuhl, P. K. (1997). The encoding of rate and talker information during phonetic perception. *Perception & Psychophysics*, *59*(5), 675–692. https://doi.org/10.3758/BF03206015

Guion, S. G., & Pederson, E. (2007). Investigating the role of attention in phonetic learning. In O.-S. Bohn & M. Munro (Eds.), *Language Experience in Second Language Speech Learning* (pp. 57–77). John Benjamins.

Hardison, D. M. (2003). Acquisition of second-language speech: Effects of visual cues, context, and talker variability. *Applied Psycholinguistics*, *24*(4), 495–522. https://doi.org/10.1017/S0142716403000250

Heald, S., & Nusbaum, H. (2014). Talker variability in audiovisual speech perception. *Frontiers in Psychology*, *5*, 698. https://doi.org/10.3389/fpsyg.2014.00698

Helsdingen, A. S., van Gog, T., & van Merriënboer, J. J. G. (2011). The effects of practice schedule on learning a complex judgment task. *Learning and Instruction*, *21*(1), 126–136. https://doi.org/10.1016/j.learninstruc.2009.12.001

Hillenbrand, J. M., Clark, M. J., & Houde, R. A. (2000). Some effects of duration on vowel recognition. *The Journal of the Acoustical Society of America*, *108*(6), 3013–3022. https://doi.org/10.1121/1.1323463

Isaacs, T., & Trofimovich, P. (2011). Phonological memory, attention control, and musical ability: Effects of individual differences on rater judgments of second language speech. *Applied Psycholinguistics*, *32*(1), 113–140. https://doi.org/10.1017/S0142716410000317

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*(2), 866–877. https://doi.org/10.1121/1.3148196

Iverson, P., Hazan, V., & Bannister, K. (2005). Phonetic training with acoustic cue manipulations: A comparison of methods for teaching English /r/-/l/ to Japanese adults. *The Journal of the Acoustical Society of America*, *118*(5), 3267–3278. https://doi.org/10.1121/1.2062307

Iverson, P., & Kuhl, P. K. (1995). Mapping the perceptual magnet effect for speech using signal detection theory and multidimensional scaling. *The Journal of the Acoustical Society of America*, *97*(1), 553–562. https://doi.org/10.1121/1.412280

Jacewicz, E., Fox, R. A., & Salmons, J. (2011). Vowel change across three age groups of speakers in three regional varieties of American English. *Journal of Phonetics*, *39*(4), 683–693. https://doi.org/10.1016/j.wocn.2011.07.003

Jamieson, D. G., & Morosan, D. E. (1989). Training new, nonnative speech contrasts: A comparison of the prototype and perceptual fading techniques. *Canadian Journal of Psychology*, *43*(1), 88–96. https://doi.org/10.1037/h0084209

Jia, G., Strange, W., Wu, Y., Collado, J., & Guan, Q. (2006). Perception and production of English vowels by Mandarin speakers: Age-related differences vary with amount of L2 exposure. *The Journal of the Acoustical Society of America*, *119*(2), 1118–1130. https://doi.org/10.1121/1.2151806

Johnson, K. (1994). Memory for vowel exemplars. *The Journal of the Acoustical Society of America*, *95*(5), 2977–2977. https://doi.org/10.1121/1.408940

Johnson, K., Ladefoged, P., & Lindau, M. (1993). Individual differences in vowel production. *The Journal of the Acoustical Society of America*, *94*(2), 701–714. https://doi.org/10.1121/1.406887

Juffs, A. (2004). Representation, processing and working memory in a second language. *Transactions of the Philological Society*, *102*(2), 199–225. https://doi.org/10.1111/j.0079-1636.2004.00135.x

Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review*, *99*(1), 122–149. https://doi.org/10.1037/0033-295X.99.1.122

Kane, M. J., Bleckley, M. K., Conway, A. R. A., & Engle, R. W. (2001). A controlled-attention view of working-memory capacity. *Journal of Experimental Psychology. General*, *130*(2), 169–183. https://doi.org/10.1037//0096-3445.130.2.169

Kartushina, N., & Martin, C. (2019). Talker and acoustic variability in learning to produce nonnative sounds: Evidence from articulatory training. *Language Learning*, *69*(1), 71–105. https://doi.org/10.1111/lang.12315

Kawahara, H., Masuda-Katsuse, I., & De Cheveigne, A. (1999). Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds[1]. *Speech Communication*, *27*(3–4), 187–207. https://doi.org/10.1016/S0167-6393(98)00085-5

Kewley-Port, D., & Watson, C. S. (1994). Formant-frequency discrimination for isolated English vowels. *The Journal of the Acoustical Society of America*, *95*(1), 485–496. https://doi.org/10.1121/1.410024

Klatt, D. H. (1976). Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *The Journal of the Acoustical Society of America*, *59*(5), 1208–1221. https://doi.org/10.1121/1.380986

Kondaurova, M. V., & Francis, A. L. (2010). The role of selective attention in the acquisition of English tense and lax vowels by native Spanish listeners: Comparison of three training methods. *Journal of Phonetics*, *38*(4), 569–587. https://doi.org/10.1016/j.wocn.2010.08.003

Kormos, J., & Sáfár, A. (2008). Phonological short-term memory, working memory and foreign language performance in intensive language learning. *Bilingualism: Language and Cognition*, *11*(2), 261–271. https://doi.org/10.1017/S1366728908003416

Kreiman, J., Gerratt, B. R., Precoda, K., & Berke, G. S. (1992). Individual differences in voice quality perception. *Journal of Speech, Language, and Hearing Research*, *35*(3), 512–520. https://doi.org/10.1044/jshr.3503.512

Kuhl, P. K., & Iverson, P. (1995). Linguistic experience and the "perceptual magnet effect." In W. Strange (Ed.), *Speech Perception and Linguistic Experience: Issues in Cross-language Research* (pp. 121–154). York Press.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Ladefoged, P., & Broadbent, D. E. (1957). Information conveyed by vowels. *The Journal of the Acoustical Society of America*, *29*(1), 98–104. https://doi.org/10.1121/1.1908694

Lee, C.-Y., Tao, L., & Bond, Z. S. (2009). Speaker variability and context in the identification of fragmented Mandarin tones by native and non-native listeners. *Journal of Phonetics*, *37*(1), 1–15. https://doi.org/10.1016/j.wocn.2008.08.001

Leeser, M. J. (2007). Learner-based factors in L2 reading comprehension and processing grammatical form: Topic familiarity and working memory. *Language Learning*, *57*(2), 229–270. https://doi.org/10.1111/j.1467-9922.2007.00408.x

Lenth, R. V. (2021). *emmeans: Estimated Marginal Means, aka Least-Squares Means* (R package version 1.7.1-1). https://CRAN.R-project.org/package=emmeans

Lewontin, R. C. (1966). On the measurement of relative variability. *Systematic Biology*, *15*(2), 141–142. https://doi.org/10.2307/sysbio/15.2.141

Liberman, A. M., Cooper, F. S., Shankweiler, D., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, *74*(6), 431–461. https://doi.org/10.1037/h0020279

Liberman, A. M., Delattre, P., & Cooper, F. (1952). The role of selected stimulus-variables in the perception of the unvoiced stop consonants. *The American Journal of Psychology*, *65*(4), 497–516. https://doi.org/10.2307/1418032

Liberman, P. (1973). On the evolution of language: A unified view. *Cognition*, *2*(1), 59–94. https://doi.org/10.1016/0010-0277(72)90030-3

Liu, C., Jin, S. -h, & Chen, C. -t. (2014). Durations of American English vowels by native and non-native Speakers: Acoustic analyses and perceptual effects. *Language and Speech*, *57*(2), 238–253. https://doi.org/10.1177/0023830913507692

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*(3), 1242–1255. https://doi.org/10.1121/1.408177

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*(4), 2076–2087. https://doi.org/10.1121/1.410149

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*(2), 874–886. https://doi.org/10.1121/1.1894649

Luthra, S., Mechtenberg, H., & Myers, E. B. (2021). Perceptual learning of multiple talkers requires additional exposure. *Attention, Perception, & Psychophysics*, *83*(5), 2217–2228. https://doi.org/10.3758/s13414-021-02261-w

Macdonald, R. M. M. (2012). *Counteracting Age Related Effects in L2 Acquisition: Training to Distinguish between French Vowels* [Doctoral Dissertation]. The University of Edinburgh.

Martin, C. S., Mullennix, J. W., Pisoni, D. B., & Summers, W. V. (1989). Effects of talker variability on recall of spoken word lists. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *15*(4), 676–684. PubMed. https://doi.org/10.1037//0278-7393.15.4.676

McHaney, J. R., Tessmer, R., Roark, C. L., & Chandrasekaran, B. (2021). Working memory relates to individual differences in speech category learning: Insights from computational modeling and pupillometry. *Brain and Language*, *222*, 105010. https://doi.org/10.1016/j.bandl.2021.105010

Mermelstein, P. (1978). Difference limens for formant frequencies of steady-state and consonant-bound vowels. *The Journal of the Acoustical Society of America*, *63*(2), 572–580. https://doi.org/10.1121/1.381756

Morrison, G. S. (2005). An appropriate metric for cue weighting in L2 speech perception: Response to Escudero and Boersma (2004). *Studies in Second Language Acquisition*, *27*(4), 597–606. https://doi.org/10.1017/S0272263105050266

Morrison, G. S. (2007). Logistic regression modeling for first- and second-language perception data. In M. G. Sole, P. Prieto, & J. Mascaro (Eds.), *Segmental and Prosodic Issues in Romance Phonology* (pp. 219–236). John Benjamins.

Morrison, G. S., & Kondaurova, M. V. (2009). Analysis of categorical response data: Use logistic regression rather than endpoint-difference scores or discriminant analysis. *The Journal of the Acoustical Society of America*, *126*(5), 2159–2162. https://doi.org/10.1121/1.3216917

Mullennix, J. W., & Pisoni, D. B. (1990). Stimulus variability and processing dependencies in speech perception. *Perception & Psychophysics*, *47*(4), 379–390. https://doi.org/10.3758/BF03210878

Mullennix, J. W., Pisoni, D. B., & Martin, C. S. (1989). Some effects of talker variability on spoken word recognition. *The Journal of the Acoustical Society of America*, *85*(1), 365–378. PubMed. https://doi.org/10.1121/1.397688

Murry, T., & Singh, S. (1980). Multidimensional analysis of male and female voices. *The Journal of the Acoustical Society of America*, *68*(5), 1294–1300. https://doi.org/10.1121/1.385122

Myers, E. B., & Theodore, R. M. (2017). Voice-sensitive brain networks encode talker-specific phonetic detail. *Brain and Language*, *165*, 33–44. https://doi.org/10.1016/j.bandl.2016.11.001

Nagle, C., & Baese-Berk, M. (2021). Advancing the state of the art in L2 speech perception-production research: Revisiting theoretical assumptions and methodological practices. *Studies in Second Language Acquisition*, 1–26. https://doi.org/10.1017/S0272263121000371

Nagle, C. L. (2021). Revisiting perception-production relationships: Exploring a new approach to investigate perception as a time-varying predictor. *Language Learning*, *71*(1), 243–279. https://doi.org/10.1111/lang.12431

Nishi, K., & Kewley-Port, D. (2007). Training Japanese listeners to perceive American English vowels: Influence of training sets. *Journal of Speech, Language, and Hearing Research*, *50*(6), 1496–1509. https://doi.org/10.1044/1092-4388(2007/103)

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39–57. https://doi.org/10.1037/0096-3445.115.1.39

Nosofsky, R. M. (1989). Further tests of an exemplar-similarity approach to relating identification and categorization. *Perception & Psychophysics*, *45*(4), 279–290. https://doi.org/10.3758/BF03204942

Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1994). Speech perception as a talker-contingent process. *Psychological Science*, *5*(1), 42–46. https://doi.org/10.1111/j.1467-9280.1994.tb00612.x

Oberauer, K., Süß, H.-M., Wilhelm, O., & Wittman, W. W. (2003). The multiple faces of working memory: Storage, processing, supervision, and coordination. *Intelligence*, *31*(2), 167–193. https://doi.org/10.1016/S0160-2896(02)00115-0

O'Brien, I., Segalowitz, N., Collentine, J. O. E., & Freed, B. (2006). Phonological memory and lexical, narrative, and grammatical skills in second language oral production by adult learners. *Applied Psycholinguistics*, *27*(3), 377–402. https://doi.org/10.1017/s0142716406060322

Palmeri, T. J., Goldinger, S. D., & Pisoni, D. B. (1993). Episodic encoding of voice attributes and recognition memory for spoken words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*, 309–328. https://doi.org/10.1037//0278-7393.19.2.309

Papagno, C., & Vallar, G. (1995). Verbal short term memory and vocabulary learning in polyglots. *The Quarterly Journal of Experimental Psychology Section A*, *48*(1), 98–107. https://doi.org/10.1080/14640749508401378

Perrachione, T. K., Jiyeon, L., Ha, L. Y. Y., & Wong, P. C. M. (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*(1), 461–472. https://doi.org/10.1121/1.3593366

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*(2), 175–184. https://doi.org/10.1121/1.1906875

Pierrehumbert, J. B. (2002). Word-specific phonetics. In C. Gussenhoven & N. Warner (Eds.), *Laboratory Phonology VII* (pp. 101–140). Mouton de Gruyter.

Pisoni, D. B. (1997). Some thoughts on "normalization" in speech perception. In J. Mullennix & K. A. Johnson (Eds.), *Talker Variability in Speech Processing* (pp. 9–32). Academic Press.

Pisoni, D. B., Lively, S. E., & Logan, J. S. (1994). Perceptual learning of nonnative speech contrasts: Implications for theories of speech perception. In H. C. Nusbaum & J. Goodman (Eds.), *The Development of Speech Perception: The Transition from Speech Sounds to Spoken Words* (pp. 121–166). MIT Press.

Plunkett, K., & Marchman, V. (1991). U-shaped learning and frequency effects in a multi-layered perception: Implications for child language acquisition. *Cognition*, *38*(1), 43–102. https://doi.org/10.1016/0010-0277(91)90022-V

Potter, C. E., & Saffran, J. R. (2017). Exposure to multiple accents supports infants' understanding of novel accents. *Cognition*, *166*, 67–72. https://doi.org/10.1016/j.cognition.2017.05.031

Pruitt, J. S., Jenkins, J. J., & Strange, W. (2006). Training the perception of Hindi dental and retroflex stops by native speakers of American English and Japanese. *The Journal of the Acoustical Society of America*, *119*(3), 1684–1696. https://doi.org/10.1121/1.2161427

Quam, C., & Creel, S. C. (2021). Impacts of acoustic-phonetic variability on perceptual development for spoken language: A review. *WIREs Cognitive Science*, *12*(5), e1558. https://doi.org/10.1002/wcs.1558

Rallo Fabra, L., & Romero, J. (2012). Native Catalan learners' perception and production of English vowels. *Journal of Phonetics*, *40*(3), 491–508. https://doi.org/10.1016/j.wocn.2012.01.001

Ramscar, M., Yarlett, D., Dye, M., Denny, K., & Thorpe, K. (2010). The effects of feature-label-order and their implications for symbolic learning. *Cognitive Science*, *34*, 909–957. https://doi.org/10.1111/j.1551-6709.2009.01092.x

Raviv, L., Lupyan, G., & Green, S. C. (2022). How variability shapes learning and generalization. *Trends in Cognitive Sciences*, *26*(6), 462–483. https://doi.org/10.1016/j.tics.2022.03.007

Restle, F. (1955). A theory of discrimination learning. *Psychological Review*, *62*(1), 11. https://doi.org/10.1037/h0046642

Rochet, B. L. (1995). Perception and production of second-language speech sounds by adults. In W. Strange (Ed.), *Speech Perception and Linguistic Experience: Issues in Cross-language Research* (pp. 379–410). York Press.

Rost, G. C., & McMurray, B. (2009). Speaker variability augments phonological processing in early word learning. *Developmental Science*, *12*(2), 339–349. https://doi.org/10.1111/j.1467-7687.2008.00786.x

Rost, G. C., & McMurray, B. (2010). Finding the signal by adding noise: The role of noncontrastive phonetic variability in early word learning. *Infancy*, *15*(6), 608–635. https://doi.org/10.1111/j.1532-7078.2010.00033.x

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*(2), 1324–1335. https://doi.org/10.1121/1.4812767

Sadakata, M., & Mcqueen, J. M. (2014). Individual aptitude in Mandarin lexical tone perception predicts effectiveness of high-variability training. *Frontiers in Psychology*, *5*(5), 1–15. https://doi.org/10.3389/fpsyg.2014.01318

Saltzman, D., Luthra, S., Myers, E. B., & Magnuson, J. S. (2021). Attention, task demands, and multitalker processing costs in speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, *47*(12), 1673–1680. https://doi.org/10.1037/xhp0000963

Schertz, J., Cho, T., Lotto, A., & Warner, N. (2015). Individual differences in phonetic cue use in production and perception of a non-native sound contrast. *Journal of Phonetics*, *52*, 183–204. https://doi.org/10.1016/j.wocn.2015.07.003

Segalowitz, N., Gatbonton, E., & Trofimovich, P. (2009). Links between ethnolinguistic affiliation, self-related motivation and second language fluency: Are they mediated by psycholinguistic variables? In Z. Dörnyei & E. Ushioda (Eds.), *Motivation, Language Identity and the L2 Self* (pp. 172–192). Multilingual Matters.

Shultz, A. A., Francis, A. L., & Fernando, L. (2012). Differential cue weighting in perception and production of consonant voicing. *The Journal of the Acoustical Society of America*, *132*(2), EL95-101. https://doi.org/10.1121/1.4736711

Silpachai, A. (2020). The role of talker variability in the perceptual learning of Mandarin tones by American English listeners. *Journal of Second Language Pronunciation*, *6*(2), 209–235. https://doi.org/10.1075/jslp.19010.sil

Singh, S., & Murry, T. (1978). Multidimensional classification of normal voice qualities. *The Journal of the Acoustical Society of America*, *64*(1), 81–87. https://doi.org/10.1121/1.381958

Singmann, H., Bolker, B., Westfall, J., Aust, F., & Ben-Shachar, M. S. (2021). *afex: Analysis of Factorial Experiments* (R package version 1.0-1). https://CRAN.R-project.org/package=afex

Sinkeviciute, R., Brown, H., Brekelmans, G., & Wonnacott, E. (2019). The role of input variability and learner age in second language vocabulary learning. *Studies in Second Language Acquisition*, *41*, 1–26. https://doi.org/10.1017/S0272263119000263

Sommers, M. S., & Barcroft, J. (2007). An integrated account of the effects of acoustic variability in first language and second language: Evidence from amplitude, fundamental frequency, and speaking rate variability. *Applied Psycholinguistics*, *28*, 231–249. https://doi.org/10.1017/S0142716407070129

Sommers, M. S., & Barcroft, J. (2011). Indexical information, encoding difficulty, and second language vocabulary learning. *Applied Psycholinguistics*, *32*(2), 417–434. Cambridge Core. https://doi.org/10.1017/S0142716410000469

Speciale, G., Ellis, N. C., & Bywater, T. (2004). Phonological sequence learning and short-term store capacity determine second language vocabulary acquisition. *Applied Psycholinguistics*, *25*(2), 293–321. Cambridge Core. https://doi.org/10.1017/S0142716404001146

Strange, W., & Dittmann, S. (1984). Effects of discrimination training on the perception of /r-l/ by Japanese adults learning English. *Perception & Psychophysics*, *36*(2), 131–145. https://doi.org/10.3758/BF03202673

Thomson, R. I. (2012). Improving L2 Listeners' Perception of English Vowels: A Computer-Mediated Approach. *Language Learning*, *62*(4), 1231–1258. https://doi.org/10.1111/j.1467-9922.2012.00724.x

Tonidandel, S., & LeBreton, J. M. (2010). Determining the relative importance of predictors in logistic regression: An extension of relative weight analysis. *Organizational Research Methods*, *13*(4), 767–781. https://doi.org/10.1177/1094428109341993

Tonidandel, S., & LeBreton, J. M. (2015). RWA web: A free, comprehensive, web-based, and user-friendly tool for relative weight analyses. *Journal of Business and Psychology*, *30*(2), 207–216. https://doi.org/10.1007/s10869-014-9351-z

Traunmüller, H. (1990). Analytical expressions for the tonotopic sensory scale. *The Journal of the Acoustical Society of America*, *88*, 97–100. https://doi.org/10.1121/1.399849

Tremblay, P., Brisson, V., & Deschamps, I. (2021). Brain aging and speech perception: Effects of background noise and talker variability. *NeuroImage*, *227*, 117675. https://doi.org/10.1016/j.neuroimage.2020.117675

Uchihara, T., Webb, S., Saito, K., & Trofimovich, P. (2022). The effects of talker variability and frequency of exposure on the acquisition of spoken word knowledge. *Studies in Second Language Acquisition*, *44*(2), 357–380. https://doi.org/10.1017/S0272263121000218

Unsworth, N., & Robison, M. K. (2017). The importance of arousal for variation in working memory capacity and attention control: A latent variable pupillometry study. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(12), 1962–1987. https://doi.org/10.1037/xlm0000421

Van Lancker, D., Kreiman, J., & Emmorey, K. (1985). Familiar voice recognition: Patterns and parameters Part I: Recognition of backward voices. *Journal of Phonetics*, *13*(1), 19–38. https://doi.org/10.1016/S0095-4470(19)30723-5

Wade, T., Jongman, A., & Sereno, J. (2007). Effects of acoustic variability in the perceptual learning of non-native-accented speech sounds. *Phonetica*, *64*(2–3), 122–144. https://doi.org/10.1159/000107913

Walter, C. (2006). Transfer of reading comprehension skills to L2 is linked to mental representations of text and to L2 working memory. *Applied Linguistics*, *25*(3), 315–339. https://doi.org/10.1093/applin/25.3.315

Wang, X., & Munro, M. J. (2004). Computer-based training for learning English vowel contrasts. *Incorporating Multimedia Capability in the Reporting of Applied Linguistics Research*, *32*(4), 539–552. https://doi.org/10.1016/j.system.2004.09.011

Wang, Y., Jongman, A., & Sereno, J. A. (2003). Acoustic and perceptual evaluation of Mandarin tone productions before and after perceptual training. *The Journal of the Acoustical Society of America*, *113*(2), 1033–1043. https://doi.org/10.1121/1.1531176

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*(6), 3649–3658. https://doi.org/10.1121/1.428217

Wayland, R. P., & Guion, S. G. (2004). Training English and Chinese listeners to perceive Thai tones: A preliminary report. *Language Learning*, *54*(4), 681–712. https://doi.org/10.1111/j.1467-9922.2004.00283.x

Weissheimer, J., & Mota, M. (2009). Working memory capacity and the development of L2 speech production. *Issues in Applied Linguistics*, *17*, 93–112. https://doi.org/10.5070/L4172005115

Werker, J. F., & Curtin, S. (2005). PRIMIR: A Developmental Framework of Infant Speech Processing. *Language Learning and Development*, *1*(2), 197–234. https://doi.org/10.1080/15475441.2005.9684216

Wiener, S., Chan, M., & Ito, K. (2020). Do explicit instruction and high variability phonetic training improve nonnative speakers' Mandarin tone productions? *The Modern Language Journal*, *104*(1), 152–168. https://doi.org/10.1111/modl.12619

Wiener, S., Ito, K., & Speer, S. R. (2018). Early L2 spoken word recognition combines input-based and knowledge-based processing. *Language and Speech*, *61*(4), 632–656. https://doi.org/10.1177/0023830918761762

Williams, J. N., & Lovatt, P. (2005). Phonological memory and rule learning. *Language Learning*, *55*(S1), 177–233. https://doi.org/10.1111/j.0023-8333.2005.00298.x

Wong, W. S. (2014, September). *The effects of high and low variability phonetic training on the perception and production of English vowels /e/-/æ/ by Cantonese ESL learners with high and low L2 proficiency levels*. Proceedings of the 15th Annual Conference of the International Speech Communication Association (Interspeech 2014) (pp. 524-528), Singapore.

Wonnacott, E., Boyd, J. K., Thomson, J., & Goldberg, A. E. (2012). Input effects on the acquisition of a novel phrasal construction in 5 year olds. *Journal of Memory and Language*, *66*(3), 458–478. https://doi.org/10.1016/j.jml.2011.11.004

Wöstmann, M., & Obleser, J. (2016). Acoustic detail but not predictability of task-irrelevant speech disrupts working memory. *Frontiers in Human Neuroscience*, *10*, 538. https://doi.org/10.3389/fnhum.2016.00538

Zelazo, P. D., Anderson, J. E., Richler, J., Wallner-Allen, K., Beaumont, J. L., Conway, K. P., Gershon, R., & Weintraub, S. (2014). NIH Toolbox Cognition Battery (CB): Validation of executive function measures in adults. *Journal of the International Neuropsychological Society*, *20*(6), 620–629. PubMed. https://doi.org/10.1017/S1355617714000472

Zhang, X., Cheng, B., Qin, D., & Zhang, Y. (2021). Is talker variability a critical component of effective phonetic training for nonnative speech? *Journal of Phonetics*, *87*, 101071. https://doi.org/10.1016/j.wocn.2021.101071

Zhang, X., Cheng, B., & Zhang, Y. (2021). The role of talker variability in nonnative phonetic learning: A systematic review and meta-analysis. *Journal of Speech, Language, and Hearing Research*, *64*(12), 4802–4825. https://doi.org/10.1044/2021_JSLHR-21-00181

Zhang, Y., & Cheng, B. (2011). Brain plasticity and phonetic training for English-as-a-second-language learners. In D. J. Alonso (Ed.), *English as a Second Language* (pp. 1–50). Nova Science Publishers.

Zhang, Y., Kuhl, P. K., Imada, T., Iverson, P., Pruitt, J., Stevens, E. B., Kawakatsu, M., Tohkura, Y., & Nemoto, I. (2009). Neural signatures of phonetic learning in adulthood: A magnetoencephalography study. *NeuroImage*, *46*(1), 226–240. https://doi.org/10.1016/j.neuroimage.2009.01.028

Zhang, Y., Kuhl, P. K., Imada, T., Kotani, M., & Tohkura, Y. (2005). Effects of language experience: Neural commitment to language-specific auditory patterns. *NeuroImage*, *26*(3), 703–720. https://doi.org/10.1016/j.neuroimage.2005.02.040