Tests of a dual-system model of speech category learning

W. TODD MADDOX and BHARATH CHANDRASEKARAN

Bilingualism: Language and Cognition / Volume 17 / Issue 04 / October 2014, pp 709 - 728
DOI: 10.1017/S1366728913000783, Published online: 17 January 2014

Link to this article: http://journals.cambridge.org/abstract_S1366728913000783

How to cite this article:

Request Permissions : Click here
Tests of a dual-system model of speech category learning

W. TODD MADDOX
BHARATH CHANDRASEKARAN
University of Texas, Austin

In the visual domain, more than two decades of work has argued for the existence of dual category learning systems. The reflective system uses working memory in an explicit fashion to develop and test rules for classifying. The reflexive system operates by implicitly associating perception with actions that lead to reinforcement. Dual-system models posit that in learning natural categories, learners initially use the reflective system and with practice, transfer control to the reflexive system. The role of reflective and reflexive systems in second language (L2) speech learning has not been systematically examined. In the study reported in this paper, monolingual native speakers of American English were trained to categorize Mandarin tones produced by multiple speakers. Our computational modeling approach demonstrates that learners use reflective and reflexive strategies during tone category learning. Successful learners use speaker-dependent, reflective analysis early in training and reflexive strategies by the end of training. Our results demonstrate that dual-learning systems are operative in L2 speech learning. Critically, learner strategies directly relate to individual differences in successful category learning.

Keywords: L2 acquisition, dual-learning systems, reflexive processing, reflective processing, individual differences

Introduction

A large body of neuropsychological, neuroimaging, and behavioral studies in vision has identified two separate systems that are operative during category learning: a reflective system, in which processing is under conscious control, and a reflexive system that is not under conscious control (Ashby & Maddox, 2011; Ashby & Spiering, 2004; Maddox & Ashby, 2004; Nomura, Maddox, Filoteo, Ing, Gitelman, Parrish, Mesulam & Reber, 2007; Nomura & Reber, 2008; Poldrack, Clark, Pare-Blagoev, Shohamy, Creso Moyano, Myers & Gluck, 2001; Poldrack & Foerde, 2008; Poldrack & Packard, 2003). The reflective learning system uses working memory (WM) and executive attention to develop and test verbalizable rules based on feedback (Ashby, Maddox & Bohil, 2002; Maddox & Ashby, 2004; Maddox, Filoteo, Lauritzen, Connaily & Hejl, 2005; Maddox, Love, Glass & Filoteo, 2008). In contrast, the reflexive learning system is neither consciously penetrable nor verbalizable, and operates by automatically associating perception with actions that lead to reward (Seger, 2008; Seger & Cincotta, 2005; Seger & Miller, 2010). This system is not dependent on WM and executive attention (DeCaro, Thomas & Beilock, 2008): it is implicit and procedural. Although there is anatomical evidence of extensive connectivity between auditory regions and the reflective and reflexive systems (Petrides & Pandya, 1998; Yeterian & Pandya, 1998), the dual-system framework has not been systematically applied in the auditory domain.

This paper applies the dual-system theoretical framework to speech category learning, using computational modeling to examine the learning of natural speech categories (Mandarin tones) by adult English speakers. Computational modeling provides insight into the specific learning operations (e.g. reflective vs. reflexive) that participants employ (Cleeremans & Dienes, 2008) and can account for individual variability in task performance. Although accuracy rates provide information about the level of performance, they tell us little about the specific strategy being used by a participant because there are a number of reflexive and reflective strategies that yield the same accuracy rate. The computational models used in the current study are constrained by our understanding of the neurobiology of the reflective and reflexive learning systems. Modeling therefore allows a more systematic examination of the computational strategies learners use while acquiring novel speech categories. We turn next to the neurobiology underlying the two learning systems.

Dual-learning systems: Neurobiology

The competition between verbal and implicit systems (COVIS) model captures the dual-system framework, and proposes a neural circuitry involved in visual category learning (Ashby & Maddox, 2011; Maddox & Ashby, 2004; Nomura & Reber, 2008).
In COVIS, processing in the reflective, hypothesis-testing system is available to conscious awareness and is mediated by a circuit primarily involving structures in the dorsolateral prefrontal cortex, anterior caudate nucleus, anterior cingulate, and medial temporal lobe. Processing in the reflexive, procedural-based learning system is not consciously penetrable and operates by associating perception with actions that lead to reinforcement. Learning in the procedural system is mediated primarily by the posterior caudate nucleus and putamen (Ashby & Maddox, 2005; Ashby & O’Brien, 2005). COVIS assumes that the two systems compete throughout learning, but that initial learning is dominated by the reflective, hypothesis-testing system. Learners will continue to use the hypothesis-testing system until the output of the procedural system is more accurate, at which point control is passed to the reflexive, procedural system.

Anatomical studies in animal models suggest that the primary and auditory association cortical regions are strongly connected to the reflective and reflexive systems. Retrograde anatomical labeling studies in primates demonstrate that the primary and association cortices are bi-directionally connected to the prefrontal cortex and form many-to-one projections to the caudate (Petrides & Pandya, 1988; Yeterian & Pandya, 1998). These studies lend neurobiological plausibility to the operation of a dual-system framework in the auditory domain. Note that extant dual-system models restrict themselves to the visual domain (Ashby & Ell, 2001; Ashby & Ennis, 2006; Ashby & Maddox, 2005, 2011), although the underlying systems are assumed to be domain-general.

**Dual-system framework: Category structures**

COVIS assumes that the reflective and reflexive systems are complementary and that each is dedicated to the learning of different types of category structures. Category structures that are learnable by the reflective system, like those shown in Figure 1a, are referred to as rule-based (RB) structures (Ashby & Maddox, 2011). The optimal verbal rule (denoted by the solid horizontal and vertical lines) is to “respond A to short, high frequency sounds, B to short, low frequency sounds, C to long, high frequency sounds, and D to long, low frequency sounds”. In contrast, the structure in Figure 1b is not learnable by the reflective system because the optimal rule (represented by the solid diagonal lines) is not verbalizable (because frequency and duration involve incommensurable units). However, such structures can still be learned and are referred to as information-integration (II) category structures. Participants who learn the II category structure are unable to articulate any verbal rules used to learn these categories, but likely integrate dimensions prior to decision-making.

These artificial category structures have been used to test the extent to which the two learning systems are separate. A large number of dissociations have been found between the two learning systems in the visual domain. One such critical difference is the reliance on WM (DeCaro, Carlson, Thomas & Beilock, 2009; DeCaro et al., 2008). The reflective learning system critically depends on WM, while the reflexive learning system does not. Individuals with high working memory capacity (WMC) learn rule-based categories faster than those with...
low WMC (DeCaro et al., 2008, 2009). Interestingly, this pattern does not hold true for II category structures. While experimenter-determined artificial category structures are a good test-bed on which to examine dissociations between the two systems, most naturally occurring categories cannot be cleanly demarcated as either RB or II categories.

The dual-system framework in second language acquisition (SLA)

While the dual-system framework has not been systematically examined in L2 speech category learning, there exists a large body of work that focuses on the role of explicit and implicit learning systems in SLA (see DeKeyser, 2008 for an extensive review). The mechanistic role of multiple learning systems has generated considerable interest as well as controversy, both in the empirical and practical (e.g. second language instruction) domain (Hulstijn, 2005; Krashen, 1982; Schmidt, 1995, 2012) and in the theoretical domain (e.g. models of SLA) (Paradis, 1985, 2004; Ullman, 2004, 2006). Early models argue for the existence of conscious and subconscious systems in language acquisition, with the subconscious system more critical to language learning. For example, Krashen’s Monitor model argues that instruction may not be optimal for all language learners, because complex language rules are mostly acquired subconsciously (Krashen, 1982). In contrast, other accounts (such as the Noticing hypothesis) suggest that conscious learning processes are critical to SLA (Schmidt, 1995, 2012). More recent dual-system models argue for different roles for explicit and implicit learning systems in mediating different components of language. In the DP (Declarative-Procedural) model (Ullman, 2004, 2006) (for a related model and application see McClelland, McNaughton & O’Reilly, 1995) a declarative system subserves vocabulary learning, and a more automatic, procedural system, mediated by the basal ganglia, subserves key aspects of grammar processing. Learning by the declarative system involves fast mapping, while learning by the procedural system is more gradual, requiring more extensive practice. In SLA, the DP model predicts that the fast-mapping declarative system initially dominates during grammar processing, making L2 processing effortful and less automatic (Hernandez & Li, 2007). With practice and training (such as in advanced L2 learners), the more gradual procedural learning system becomes increasingly dominant, allowing more automaticity in L2 processing. The neurolinguistic theory of bilingualism also hypothesizes the existence of declarative and procedural learning systems that mediate metalinguistic and implicit linguistic competence, respectively (Paradis, 2004). In this model, implicit linguistic competence is achieved via incidental learning. It proposes that L2 learners have maximum difficulty acquiring language components that are implicitly acquired. While much of the literature on implicit vs. explicit learning processes has focused on morphology, syntax, or vocabulary learning, much less is known about the role of the two systems in L2 phonetic acquisition. Here we extend a dual-system approach to understanding phonetic learning. Our approach differs from existing dual-system models in the language domain in several important ways. In SLA literature, the declarative system, mediated by the medial temporal lobe, is argued to play a critical role in explicit processing. The neural bases of the implicit system are less clearly delineated. For example, the DP model proposes the basal ganglia as a possible structure for procedural learning. Others suggest a non-specific locus for implicit processes (Reber, 2013), or view implicit processes as a form of computing statistics (Evans, Saffran & Robe-Torres, 2009; Hay, Pelucchi, Graf Estes & Saffran, 2011; Romberg & Saffran, 2010; Saffran, Aslin & Newport, 1996). In our dual-system theoretical framework, the prefrontal cortex (PFC) is the primary system involved in explicit, RB learning. The PFC, in coordination with the head of the caudate and the anterior cingulate, is important in generating, testing, and revising hypotheses, based on feedback. When rules are complex, the MTL-based declarative memory system keeps track of hypotheses that have been tested and those that have been rejected. The implicit reflexive learning system in our model is instantiated within anatomic regions in the striatum. Importantly, the computational modeling used in the current study is neurobiologically constrained, based on knowledge about the reflective/reflexive systems and their communication with the primary and secondary sensory regions.

Learning L2 speech categories as adults

A significant challenge in speech perception and learning is the mapping of highly variable signals, produced by multiple speakers, to relevant categories, which can be likened to a categorization problem (Holt & Lotto, 2008, 2010). Electrophysiological work argues for the existence of stored representations of speech categories that are language-specific (Cheour, Cepioniene, Lehtokoski, Luuk, Allik, Alho & Näätäinen, 1998; Näätänen, Lehtokoski, Lennes, Cheour, Huotilainen, Ivonen, Vainio, Alku, Ilmoniemi, Luuk, Allik, Sinkkonen & Alho, 1997). In existing neuroscientific models, these category representations are instantiated within auditory association regions in the posterior superior temporal gyrus (Hickok & Poeppel, 2004, 2007). In neuroimaging studies examining L2 speech learning, short-term auditory training has been shown to enhance
activation of the auditory association areas (Wang, Sereno, Jongman & Hirsch, 2003).

The mechanisms underlying speech category learning have been a focus of several studies. Previous research has put forward a number of reasons for difficulties in L2 speech learning, attributing it to various levels of speech processing, including interference caused by existing speech categories, and interference due to a “warping” of auditory-perceptual space by prior experience with native speech categories (Best, 1993; Best, Morrone-RObson, 1981; Best & Tyler, 2007; Feige, 1999; Francis, Ciocca, Ma & Fenn, 2008; Francis & Nusbaum, 2002). Laboratory training studies have typically used trial-by-trial feedback and high variability (multiple speakers) training to teach L2 speech categories (Bradlow, Akahane-Yamada, Pisoni & Tohkura, 1999; Lim & Holt, 2011; Lively, Pisoni, Yamada, Tohkura & Yamada, 1994; Tricipi, Delgado, McCandliss, McClelland & Fiez, 2006; Zhang, Kuhl, Imada, Iverson, Pruitt, Stevens, Kawakatsu, Tohkura & Nemoto, 2009). Feedback enhances learning by reducing errors, and multiple-speaker training results in learners refocusing their attention to cues that are relevant for distinguishing speech categories and/or reducing attention to irrelevant cues (Bradlow, 2008). Although unsupervised training results in some speech learning in adults, the addition of feedback results in substantially larger learning gains (Goudbeek, Cutler & Smits, 2008; McClelland, Fiez & McCandliss, 2002; Vallabha & McClelland, 2007). Studies have also examined the role of high-variability training in speech learning. Training with multiple speakers leads to better real-world generalization than single-speaker training (Bradlow, 2008; Lively, Logan & Pisoni, 1993; Lively et al., 1994). This suggests that high variability training may be a good way of creating robust representations of categories that are resistant to variation across speakers. However, the ability to successfully learn with high speaker variability during training may depend on the learner. For example, Perrachione, Lee, Ha and Wong (2011) showed that high variability training was beneficial to some individuals, but others performed better when exposed to less variability during training.

Several speech perceptual learning studies have examined the role of speaker normalization processes in speech category learning (Kraljic & Samuel, 2005, 2006; Samuel & Kraljic, 2009). These studies show that adults continually adjust their phonemic categories to incorporate phonetic variations produced by novel speakers. Speaker-dependent perceptual analysis during categorization is driven by the nature of the speech categories (e.g. fricatives vs. stops). For fricatives, the spectral mean is a critical acoustic cue differentiating categories (e.g. [sh] vs. [s]; [sh] has a lower spectral mean); this cue also correlates with sex differences (male speakers have a lower spectral mean than female speakers). Perceptual training resulted in sex-specific adjustment of [s] and [sh] categories (Kraljic & Samuel, 2007). In contrast, perceptual training leading to adjustment of stops (cued by voice-onset time, a timing cue that is not gender-specific) was not gender-specific.

Taken together, these studies suggest that feedback and speaker variability lead to significant L2 speech learning. While much of this research has focused on the mechanics of the perceptual system in speech learning, much less is known about the role of the dual category learning system, which previous studies suggest is critical to learning RB and II category structures. This leads us to an important point: are speech categories similar to RB category structures, or II category structures? Speech categories are typically difficult to verbalize, have multiple dimensions, and are highly variable. Generating and testing hypothesis for categories involving many dimensions is resource-intensive. Since the reflective system is dependent on WM and attention, generating rules/hypotheses for multiple dimensions may not be efficient. In addition, the redundancy and variability of cues available during speech perception prevents a simple one-to-one mapping of cues to categories. These suggest that reflexive learning may be optimal for speech categories. Does this mean there is no role for the reflective system during L2 speech learning? Speaker-dependent analyses engage WM and executive attention resources (Wong, Nusbaum & Small, 2004). Therefore, it is possible that resolving speaker variability in a multi-speaker training paradigm may significantly engage the reflective system. Our hypothesis is therefore that speech learning is reflexive-optimal, but may involve some reflective analysis. During natural visual category learning, the dual-system framework assumes that the reflective and reflexive learning systems compete throughout learning for control (Ashby & Maddox, 2011). Early learning is mostly reflective and involves actively testing hypotheses and using feedback to validate or invalidate rules. With practice, learners switch to the more automatic, reflexive learning if the output of the reflexive system is more accurate than the reflective system. In line with dual-system prediction, we propose that resolving speaker variability is at least partially a reflective process. In contrast, learning natural speech categories, we argue, is reflexive-optimal. We examine these hypotheses across two experiments involving Mandarin tone category learning by native English speakers with no prior experience with tone languages.

**The current study**

In the current study we employ trial-by-trial feedback and high-speaker variability to examine the computational strategies that native English adult speakers employ while learning non-native Mandarin tone category structures.
Mandarin Chinese has four tone categories (ma¹ ‘mother’ [T1], ma² ‘hemp’ [T2], ma³ ‘horse’ [T3], ma⁴ ‘scold’ [T4]), described phonetically as high level, high rising, low dipping, and high falling, respectively (Figure 2a). Native English speakers find it particularly difficult to learn tone categories (Wang, Jongman & Sereno, 2003), and although training can enhance their tone identification and discrimination, such training paradigms have typically resulted in significant inter-individual differences in learning success (Perrachione, Lee, Ha & Wong, 2011).

Previous speech learning studies have typically relied on behavioral measures of accuracy to examine category learning. This is problematic, since the same accuracy rate can often be achieved by using qualitatively different strategies (e.g., reflective or reflexive). We apply reflective and reflexive computational models to a) determine the extent to which speech category learning is mediated by reflective and reflexive learning, and b) examine the source of individual differences in category learning success. Our working hypothesis, consistent with dual-system predictions, is that successful learners initially rely on the reflective learning system to perform speaker-dependent analyses, and switch to the reflexive learning system by the end of training. We will expand on our predictions in the next section.

**Tone category learning: Category structure**

A number of dimensions, such as pitch height or pitch direction, may serve as cues to tone identification. The perceptual saliency of these dimensions may be influenced by the presence of specific types of pitch patterns in a language’s tonal inventory (Gandour, 1978, 1983) as well as by the occurrence of abstract tonal rules in the listener's phonological system (Hume & Johnson, 2001). Multidimensional scaling analysis of dissimilarity judgments, reaction time measures, and event-related responses, converges on two primary dimensions that underlie the tone space: pitch height and pitch direction (Figure 2b).

Native speakers of Mandarin Chinese emphasize pitch direction more than pitch height. In contrast, English listeners place less emphasis on pitch direction in disambiguating tone categories. This is consistent with cue-weighting theories, which suggest that category learning difficulties in adults may be due to reduced emphasis on critical dimensions that are more invariant, and greater reliance on dimensions that are less invariant (Francis et al., 2008). This is consistent with reasons given for learner difficulties with other speech categories. For example, one reason given for difficulties in learning the /l/ vs. /r/ category difference in Japanese listeners is reduced emphasis on the third formant, a critical dimension that differentiates the two categories (Hattori & Iverson, 2009; Iverson, Kuhl, Akahane-Yamada, Diesch, Tohkura, Kettermann & Siebert, 2003).

In Figure 3a, we plot the 80 stimuli used in our experiments (five consonant-vowel segments x four speakers x four tones) along two dimensions (pitch height: average fundamental frequency (x-axis) and pitch direction: slope (y-axis)). A visual inspection of this space suggests that this category structure is most likely information-integration (compare with Figure 1b), and therefore most likely learned by the reflexive learning system. Of the two major acoustic dimensions, pitch height provides information about speaker identity and gender (a simple verbalizable strategy, e.g. male vs. female). Across languages, high pitch is typically aligned with female speakers, and a low pitch with male speakers. Therefore, in a high-variability training environment involving male and female speakers, the initial reflective strategy that learners may attempt is one
that creates gender-dependent perceptual spaces. This is consistent with the hypothesis of a previous study that showed gender-specific perceptual learning for category structures cued by dimensions that are also involved in processing sex differences across speakers (Samuel & Kraljic, 2009). In line with this, we suggest that separating the perceptual space on the basis of the sex of the speaker (Figures 3b and 3c) is a simple verbalizable strategy (“male” vs. “female”) that listeners use. As Figures 3b and 3c show, separating the perceptual space by sex of speaker allows for significantly less overlap between tone categories, leading to less category confusion. Also, within the male and female perceptual spaces, there is little to distinguish between speakers (i.e., male speaker 1 vs. male speaker 2). Our working hypothesis is that learners use a reflective strategy early in training to resolve speaker variability, and a reflexive strategy later in training to differentiate tonal categories based on feedback.

In Experiment 1, we examine the extent to which learners use reflective and reflexive strategies while learning tone categories. Our prediction based on the dual-system framework is that successful learners use a combination of reflective and reflexive strategies. Second, we predict that reflexive learning may be better for successful categorization of tone categories. Using separate perceptual spaces (male vs. female) may be a good learning strategy to reduce tone confusion, but likely requires more WM resource than a strategy that does not use separate perceptual space. In Experiment 2, we study the extent to which the two strategies (speaker separation vs. non-separation) are mediated by WM and executive attention resources. Specifically, we examine WMC differences between speaker separators and non-separators. We predict that individuals with lower WMC will be less effective at using separate perceptual spaces and therefore less effective at tone category learning. The logic here is that separating perceptual space using a verbalizable strategy (male vs. female) is a reflective strategy that therefore requires WM resources. We predict that individual differences in category learning success are related to differential strategy use between learners.

Experiment 1

Material and methods

In this section, we describe the stimulus space used to examine strategy in this study. We will then describe the model fitting approach in detail.

Stimuli

These consisted of natural native exemplars of the four Mandarin tones, tone 1 (T1), tone 2 (T2), tone 3 (T3), and tone 4 (T4). Monosyllabic Mandarin Chinese words
(bu, di, lu, ma, and mi) that are minimally contrasted by the four tone categories were used in the experiment. Since these syllables exist in the American English inventory, this avoids the need to learn phonetic structures as well as the tone distinction (Alexander, Wong & Bradlow, 2005). By using different segments and multiple speakers, our aim is to expose learners to the variability inherent in natural language. Each of these syllables was produced in citation form with the four Mandarin tones. The speakers in Experiment 1 were two male and two female native speakers of Mandarin Chinese originally from Beijing. Two of these speakers (one male and one female) were used in Experiment 2. The stimuli were RMS amplitude and duration normalized (70 dB, 0.4 s) using the software Praat. Duration and amplitude envelope may be useful cues for the disambiguation of lexical tones. However, behavioral studies (Howie, 1976) and multidimensional scaling (MDS) analyses have shown that dimensions related to pitch (especially pitch height and pitch direction) are used primarily to distinguish tone categories (Francis et al., 2008). In fact, phonetically, Mandarin tones 1–4 are described using these two dimensions as “high-level”, “low-rising”, “low-dipping”, and “high-falling” respectively. Five native speakers of Mandarin were asked to identify the tone categories (they were given four choices) and rate their quality and naturalness. High identification (> 95%) was achieved across all five native speakers. Speakers rated these stimuli as highly natural.

Model fitting approach

We fit each model at the individual participant level because of problems with interpreting fits to aggregate data (Ashby, Maddox & Lee, 1994; Estes, 1956; Maddox, 1999). We fit each model to a block of 80 trials, assuming a fixed decision strategy on each trial within the block. We fit three classes of models with the multiple instantiations of a fixed decision strategy on each trial within the block. We use different segments and multiple speakers, our aim is to expose learners to the variability inherent in natural language. Each of these syllables was produced in citation form with the four Mandarin tones. The speakers in Experiment 1 were two male and two female native speakers of Mandarin Chinese originally from Beijing. Two of these speakers (one male and one female) were used in Experiment 2. The stimuli were RMS amplitude and duration normalized (70 dB, 0.4 s) using the software Praat. Duration and amplitude envelope may be useful cues for the disambiguation of lexical tones. However, behavioral studies (Howie, 1976) and multidimensional scaling (MDS) analyses have shown that dimensions related to pitch (especially pitch height and pitch direction) are used primarily to distinguish tone categories (Francis et al., 2008). In fact, phonetically, Mandarin tones 1–4 are described using these two dimensions as “high-level”, “low-rising”, “low-dipping”, and “high-falling” respectively. Five native speakers of Mandarin were asked to identify the tone categories (they were given four choices) and rate their quality and naturalness. High identification (> 95%) was achieved across all five native speakers. Speakers rated these stimuli as highly natural.

The model parameters were estimated using maximum likelihood procedures (Ashby, 1992; Wickens, 1982) and the models compared using Akaike weights (Wagenmakers & Farrell, 2004), as described in detail in the Results section. The details of each model are outlined in the next section.

Striatal Pattern Classifier

The SPC assumes that stimuli are represented perceptually in higher level auditory areas, such as the superior temporal gyrus. Because of the massive many-to-one (approximately 10,000-to-1) convergence of afferents from the primary and secondary auditory cortices to the striatum (Ashby & Ennis, 2006; Wilson, 1995), a low-resolution map of perceptual space is represented among the striatal units. Within the auditory domain it is well known that there are direct projections from secondary auditory areas such as the superior temporal gyrus and supratemporal plane to the caudate (Arnauld, Jeantet, Arsaut & Desmotes-Mainard, 1996; Hikosaka, Sakamoto & Usui, 1989; Yeterian & Pandya, 1998). During learning the striatal units become associated with one of the category labels, so that after learning is complete, a category response label is associated with each of a number of different regions of perceptual space. In effect, the striatum learns to associate a response with clumps of cells in the auditory cortex. The SPC assumes that there is one striatal “unit” in the pitch height-pitch direction space for each category, yielding a total of four striatal units. Because the location of one of the units can be fixed, and since a uniform expansion or contraction of the space will not affect the location of the resulting response region partitions, the SPC contains six free parameters: five that determine the location of the units, and one that represents the noise associated with the placement of the striatal units. Figure 4a displays a scatterplot of the responses and response regions for the four tone categories in Figure 3a generated from a version of the SPC. It is worth mentioning that versions of the SPC have already been applied in the auditory domain. Specifically, Maddox, Ing and Lauritzen (2006) applied the model to data from an artificial auditory category learning task, and Maddox, Molis and Diehl (2002) to data from an auditory vowel categorization task.

1 It is important to be clear that the SPC is a computational model inspired by what is known about the neurobiology of the striatum. The striatal “units” are therefore hypothetical and could be interpreted in terms of other computational models (e.g., as “prototypes” in a multiple prototype model such as SUSTAIN: Love, Medin & Gureckis, 2004). In addition, we do not model learning in the SPC, in the sense that we do not update association weights between units and category labels. Learning models have been proposed (Ashby & Maddox, 2011), but are not used here, because of their complexity.
Figure 4. (Colour online) Scatterplot of the responses along with the decision boundaries that separate response regions from versions of the (a) striatal pattern classifier, (b) conjunctive rule-based, (c) uni-dimensional_height, and (d) uni-dimensional_direction models as applied to the stimuli from Figure 3a.

**Conjunctive rule-based model**
A conjunctive RB model that assumes that the participant sets two criteria along the pitch height dimension and a third along the pitch direction dimension was also applied to the data. The model assumes that the two criteria along the pitch height dimension are used to separate the stimuli into low, medium or high. Low pitch height items are classified as tone category 3 (T3) and high pitch height items as tone category 1 (T1). If an item is classified as medium pitch height, the pitch direction dimension is examined. The single criterion along the pitch direction dimension is used to separate the stimuli into low and high pitch direction. Stimuli that have medium pitch height and low pitch direction (negative slope) are classified as tone category 4 (T4) and medium pitch height items of high pitch direction as tone category 2 (T2). Figure 4b displays a scatterplot of the responses and response regions for the four tone categories in Figure 3a generated from a version of the Conjunctive model. This model contains four free parameters: three criteria and one noise parameter.

**Uni-dimensional rule-based models**
A uni-dimensional_height RB model that assumes that the participant sets three criteria along the pitch height dimension was also applied to the data. This model assumes that the three criteria along the pitch height dimension are used to separate the stimuli into low, medium-low, medium-high or high, each of these being associated with one of the tone categories. Notice that this model completely ignores the pitch direction. Although, given four category labels, 24 versions of the model are possible, some are highly unrealistic (e.g., a model that assumes that tone category 1 (T1) was the lowest in pitch height). We examined the eight most reasonable variants of the model.

A uni-dimensional_direction RB model that assumes that the participant sets three criteria along the pitch direction dimension was also applied to the data. This model assumes that the three criteria along the pitch direction dimension are used to separate the stimuli into low, medium-low, medium-high or high, each of
these being associated with one of the tone categories. Notice that this model completely ignores the pitch height dimension. Although, given four category labels, 24 versions of the model are possible, many are highly unrealistic. We examined the two most reasonable variants of the model. Figure 4c displays a scatterplot of the responses and response regions for the four tone categories in Figure 3a generated from a version of the uni-dimensional_height model, and Figure 4d displays the corresponding scatterplot generated from a version of the uni-dimensional_direction model. The uni-dimensional models each contain four free parameters: three criteria and one noise parameter.

Random Responder Model

This model assumes a fixed probability of responding tone 1, tone 2, tone 3, and tone 4 but allows for response biases. The model has three free parameters to reflect the predicted probability of responding “1,” “2,” or “3”, the probability of responding “4” being equal to one minus the sum of the other three.

Modeling speaker separation

An a priori prediction was that learners use a verbalizable strategy (speaker sex) to separate male and female perceptual spaces. We predicted that using separate perceptual spaces reduces overlap between categories and increases successful learning of the tone categories. The model procedure assumes that each model applied to a block of 80 trials using the 80 stimuli. Figure 3a shows a modeling procedure that assumes no speaker separation (a non-separation model).

To model the presence of speaker separation (a separation model), we assume that the participant converted the 80 stimuli perceptual space in Figure 3a into two separate perceptual spaces, one that characterizes the 40 stimuli spoken by male speakers and one that characterizes the 40 stimuli spoken by female speakers. A scatterplot of the stimuli associated with the male and female sub-perceptual spaces are displayed in Figures 3b and 3c.

We fit each of the models outlined above (SPC, conjunctive, uni-dimensional_height, uni-dimensional_direction, and random responder) separately to the 40 trials with a female speaker and the 40 trials with the male speaker and estimated separate parameters for each of the relevant perceptual spaces (male or female). For example, the conjunctive model required four parameters when no speaker separation was assumed, but eight when speaker separation was assumed.

Results: Overall accuracy

Figure 5a displays the average accuracy along with standard error bars. We also include average accuracy for participants best fit by a separation or non-separation model (discussed below). We conducted a two-way ANOVA (block x tone category) and found significant learning across blocks (F (5, 95) = 23.98, p < 0.001, partial $\eta^2 = .56$), with average accuracy increasing from 42% in block 1 (b1) to 71% in block 6 (b6). We also found a main effect of tone category (F (3, 57) = 4.51, p = 0.007, partial $\eta^2 = .19$). Average accuracy was greatest for T3 (69%); pairwise comparison showed significantly better learning for T3 stimuli than for T2 (p = 0.004) or T4.

Results: Learning across blocks

We fit the 24 versions of the model procedure assumes that each model applied to a block of 80 trials using the 80 stimuli. Figure 3a shows a modeling procedure that assumes no speaker separation (a non-separation model).

To model the presence of speaker separation (a separation model), we assume that the participant converted the 80 stimuli perceptual space in Figure 3a into two separate perceptual spaces, one that characterizes the 40 stimuli spoken by male speakers and one that characterizes the 40 stimuli spoken by female speakers. A scatterplot of the stimuli associated with the male and female sub-perceptual spaces are displayed in Figures 3b and 3c.

We fit each of the models outlined above (SPC, conjunctive, uni-dimensional_height, uni-dimensional_direction, and random responder) separately to the 40 trials with a female speaker and the 40 trials with the male speaker and estimated separate parameters for each of the relevant perceptual spaces (male or female). For example, the conjunctive model required four parameters when no speaker separation was assumed, but eight when speaker separation was assumed.

Random Responder Model

This model assumes a fixed probability of responding tone 1, tone 2, tone 3, and tone 4 but allows for response biases. The model has three free parameters to reflect the predicted probability of responding “1,” “2,” or “3”, the probability of responding “4” being equal to one minus the sum of the other three.

Modeling speaker separation

An a priori prediction was that learners use a verbalizable strategy (speaker sex) to separate male and female perceptual spaces. We predicted that using separate perceptual spaces reduces overlap between categories and increases successful learning of the tone categories. The model procedure assumes that each model applied to a block of 80 trials using the 80 stimuli. Figure 3a shows a modeling procedure that assumes no speaker separation (a non-separation model).

To model the presence of speaker separation (a separation model), we assume that the participant converted the 80 stimuli perceptual space in Figure 3a into two separate perceptual spaces, one that characterizes the 40 stimuli spoken by male speakers and one that characterizes the 40 stimuli spoken by female speakers. A scatterplot of the stimuli associated with the male and female sub-perceptual spaces are displayed in Figures 3b and 3c.

We fit each of the models outlined above (SPC, conjunctive, uni-dimensional_height, uni-dimensional_direction, and random responder) separately to the 40 trials with a female speaker and the 40 trials with the male speaker and estimated separate parameters for each of the relevant perceptual spaces (male or female). For example, the conjunctive model required four parameters when no speaker separation was assumed, but eight when speaker separation was assumed.

Participants

Twenty-four participants aged 18–35 were recruited from the undergraduate population at the University of Texas at Austin; they were paid $10 per hour for their participation. The participants were monolingual and raised in monolingual English households, as reported in detailed background questionnaires. Participants with significant exposure to another language before the age of 12 were not included. Participants reported no history of neurological, visual, or hearing deficits. A detailed language history questionnaire was used to ensure that participants had no previous exposure to a tone language. All participants provided informed consent and were debriefed following the conclusion of the experiment, in accordance with UT IRB requirements. Participants (n = 24) with poorer performance in the final block than the first were excluded, leaving data from 22 participants for the statistical analyses. Average accuracy in the final block for the excluded participants was 29%, which is close to chance-level performance (25%).

Tone category training procedure

In each trial, participants were presented with a single example from one of four Mandarin tone categories (T1, T2, T3, or T4) and instructed to place it into one of four categories. The participants were told that high accuracy levels are possible. They were given feedback on each trial and exposed to multiple speakers throughout the training program. They listened to 80 stimuli per block (four tone categories x five syllables x four speakers). Within each block, the speakers were randomized. Each participant completed six 80-trial blocks of training. Participants responded by pressing one of four buttons on a response box labeled “1,” “2,” “3,” or “4”. Corrective feedback was displayed for one second on the screen immediately following the button press and consisted of the word “Correct” or “Error”, followed by the correct label for the tone. For example, on a correct tone 2 (T2) trial the feedback display was as follows: “Correct, that was a 2”. On an incorrect response trial, if tone 3 was the correct response, the feedback display was: “Error, that was a 3”. A one-second inter-trial interval followed the feedback.
Figure 5. Overall proportion correct for final blocks separators versus non-separators in Experiment 1 (A) and Experiment 2 (B).

Results: Modeling

Model fitting and model comparison

As outlined above, each model was fit to the data from each participant on a block-by-block basis. The models were fit to the Mandarin tone category learning data from each trial by maximizing negative log-likelihood. We used Akaike weights to compare the relative fit of each model (Akaike, 1974; Wagenmakers & Farrell, 2004). Akaike weights are derived from Akaike’s Information Criterion (AIC), which is used to compare models with different numbers of free parameters (AIC penalizes models with more free parameters). For each model \( i \), AIC is defined as:

\[
AIC_i = -2\log L_i + 2V_i
\]

(1)

where \( L_i \) is the maximum likelihood for model \( i \) and \( V_i \) is the number of free parameters in the model. Smaller AIC values indicate a better fit to the data. We first computed AIC values for each model and for each participant’s data in each block. Akaike weights were then calculated to obtain a continuous measure of degree-of-fit. A difference score is computed by subtracting the AIC of the best fitting model for each data set from the AIC of each model for the same data set:

\[
\Delta_i (AIC) = AIC_i - \min AIC
\]

(2)

From the differences in AIC we then computed the relative likelihood \( L \) of each model \( i \) with the transform:

\[
L(M_i|data) \propto \exp \left\{ -\frac{1}{2} \Delta_i(AIC) \right\}
\]

(3)

Finally, the relative model likelihoods were normalized by dividing the likelihood for each model by the sum of the likelihoods for all models. This yields Akaike weights:

\[
w_i(AIC) = \frac{\exp \left\{ -\frac{1}{2} \Delta_i(AIC) \right\}}{\sum_{k} \exp \left\{ -\frac{1}{2} \Delta_k(AIC) \right\}}
\]

(4)

These weights can be interpreted as the probability that the model is the best model given the data set and the set of candidate models (Wagenmakers & Farrell, 2004). Akaike weights range from 0 to 1.0, an Akaike weight of 0 implying that a given model is the best model with probability 0, and an Akaike weight of 1 implying that a given model is the best model with probability 1.0. Equivocal evidence in support of a given model is associated with an Akaike weight of \( 1/n \) where \( n \) denotes the number of models being compared (e.g., with two models, an Akaike weight of 0.5 implies equivocal support for the given model).

Best fitting model vs. random responder model

We began by comparing the Akaike weights from the best fitting uni-dimensional, conjunctive or SPC model that assumed non-separation or separation with the best fitting random responder model. This comparison allowed us to determine whether the best fitting model is capturing noise or meaningful strategic responding. The results were clear. The resulting Akaike weights were .964, .990, .946, .956, .991, and .991 in blocks 1–6, respectively. In each case these values were significantly above 0.5 based on a one-sample t-test (all p’s < .0001), which indicates that the best fitting models are effectively fitting the data.

Best fitting non-separation model vs. best fitting separation model

Next we compared the Akaike weights from the best fitting non-separation model against the best fitting non-separation model.
A dual-system model of speech category learning

0.2
0.4
0.6
0.8
1
12345
0
0.2
0.4
0.6
0.8
1
0
0.2
0.4
0.6
0.8
1
12345

Figure 6. Proportion of participants whose data was best fit by a non-separation or separation model as a function of block in Experiment 1 (A) and Experiment 2 (B).

model. This comparison allows us to determine whether the best fitting model is truly capturing additional strategic responding or just more noise. Again, the results were clear. When a separation model provided the best account of the data, the Akaike weights ranged from .880 to .982 and in every block were significantly above 0.5 based on a one-sample t-test (all p's < .001). When a non-separation model provided the best account of the data, the Akaike weights ranged from .893 to .982 and in every block were significantly above 0.5 based on a one-sample t-test (all p's < .001). These findings suggest that the best fitting model (separation or non-separation) is capturing meaningful strategic variance in the data, not just random noise.

Distribution of best fitting non-separation and separation models

Because it was hypothesized that speaker separation improves performance, we predicted that as the participants gained experience they would increasingly fit one of the separation models rather than one of the non-separation models. Figure 6 displays the proportion of participants whose data was best fit by a separation or non-separation model in each block. As a formal test of our hypothesis, we compared the number of separators and non-separators across the first and final blocks. A χ² test suggested that the number of separators increased while the number of non-separators decreased from the first to the final block of trials (χ² (1, N = 22) = 5.94, p < .005).

Separation strategy distribution for final block separators and final block non-separators

Another way to examine changes in the use of separation strategies across blocks is to compare the number of blocks of trials best fit by the separation model for participants whose final block of trials is best fit by a separation (hereafter referred to as final block separators) or non-separation (hereafter referred to as final block non-separators) model. We hypothesized that participants whose data is best fit by a separation model in the final block of trials will also be more likely to use separation strategies earlier in learning. The results supported our prediction (see Figure 7a), with significantly more blocks of trials best fit by a separation model for final block separators (4.9 blocks) than for final block non-separators (1.0 blocks) (F(1, 20) = 15.449, p < .001, partial η² = .436).

We also examined the first block of trials for which a separation model provided the best fit to the data for final block separators versus final block non-separators. We hypothesized that final block separators would begin to speaker-normalize sooner. The results supported our prediction (Figure 7b), with final block separators speaker-normalizing earlier (1.65 blocks) than non-separators (4.50 blocks) (F(1, 20) = 7.20, p < .005, partial η² = .265).

Learning curves for final block separators and final block non-separators

Figure 5a displays the learning curves for final block separators and final block non-separators. A 2-model strategy x 6-block mixed ANOVA was conducted on these data. We observed a main effect of model strategy (F(1, 20) = 6.70, p < .05, partial η² = .251), with performance significantly better for separators (.65) than for non-separators (.27). We also observed a main effect of block (F(5, 100) = 3.83, p < .01, partial η² = .161), suggesting significant learning. Finally, we observed a significant model strategy by block interaction (F(5, 100) = 2.59, p < .05, partial η² = .151). The interaction is characterized by significant learning in the separator group (F(5, 95) = 28.77, p < .001, partial η² = .602), and non-significant learning in the non-separators (F(5, 5) < 1.0).

Reflective and reflexive strategies and accuracy rates for final block separators. Here we examined performance for the reflective and reflexive strategies used by final block separators. Of the 20 final block separators, ten were best fit by the SPC, two by the conjunctive RB model, and eight by the uni-dimensional_height model.
Because the optimal strategy requires application of a reflexive strategy, we predicted that reflexive participants would outperform reflective participants. To test this, we compared overall accuracy across the ten reflexive separators and the ten reflective separators. The effect of strategy was significant ($F(1, 18) = 9.44, p < .01, \eta^2 = .344$), with participants using a reflexive strategy ($.762$) outperforming those using a reflective strategy ($.532$).

**Discussion**

Experiment 1 examined Mandarin tone category learning in native English speakers. A series of computational models derived from a dual-system model of visual category learning were applied. These models capture two aspects of learning that are hypothesized as critical to a complete understanding of Mandarin tone category learning. First, the models capture the distinction between reflective category learning strategies that are available to conscious awareness and require WM and executive attention, and reflexive category learning strategies that are not available to conscious awareness and operate without relying on WM or executive attention (Ashby et al., 1998; Ashby & Maddox, 2011). This distinction is modeled by placing constraints on the nature of the participant’s decision process (see Figure 4). Second, the models capture speaker-dependent strategies. The lack of speaker dependency is modeled by assuming that the participant generates categorization responses from a perceptual space that makes no distinction between speakers (see Figure 3a). Speaker separation, on the other hand, is modeled by assuming that the participant first determines whether the speaker is male or female and then generates categorization responses from the relevant male or female pitch height-pitch direction perceptual space (see Figures 3b and 3c).

Several results emerged from Experiment 1. Behaviorally, significant learning occurred within one session of training. Importantly, learning across blocks was similar for all four tonal categories and did not differ between male and female speakers. We found a significant main effect of tone, driven by the fact that T3 was easier to learn than T2 or T4. This is an interesting
finding because T3 is by far the most distinct category for native English speakers. T2 and T4 can be mapped on to existing intonational categories (rising and falling pitch are important cues in intonational processing, indicating, for example, the difference between a question and a statement). This finding is consistent with predictions made by the Speech Learning Model (SLM) (Flege, 1995, 1999). Specifically, SLM predicts that a native category that is similar to a non-native category, but not identical to it, may interfere with processing and learning of the non-native category. By this account, T2 and T4 categories suffer interference from existing intonational categories; in contrast, T3 is learned more easily because there is no interference from existing category representation.

The basic computational modeling approach received strong validation from the data. First, Akaike weights, which can be interpreted as the probability that a particular model is the best model given the data set and the set of candidate models, were very large for the best fitting model. When the best fitting model was compared with a random responder model, all Akaike weights were larger than .946. In addition, when the best fitting model assumed speaker separation, the Akaike weights for this model compared with the best fitting non-separation model were all larger than .880. When the best fitting model assumed no speaker separation, the Akaike weights for this model compared with the best fitting separation model were all larger than .893. Second, we found that the number of participants whose data was best fit by a separation model increased across blocks, suggesting that the prevalence of speaker separation increased with experience. Third, we found that participants whose final block of data was best fit by a separation model increased across blocks, suggesting that the prevalence of speaker separation increased with experience. Third, we found that participants whose final block of data was best fit by a separation model were more likely to use a speaker-dependent strategy in other blocks and showed speaker separation earlier in learning than participants whose final block of data was best fit by a non-separation model. Finally, amongst final block separators, we found that those who used a reflexive strategy were more accurate than those who used a reflective strategy. Taken together, these results provide strong validation of our modeling approach and demonstrate its usefulness.

Experiment 2

In Experiment 2, we conducted a large-scale replication of Experiment 1 and extended the study by collecting measures of WM capacity. First, we hypothesized that individuals using separate perceptual spaces (separators) would be more effective at learning tone categories. Second, we predicted that those learners would show higher WMC than those who do not use separate perceptual spaces (non-separators). This prediction is very similar to that proposed by Tagarelli and colleagues, who showed that WMC correlated with performance on an explicit, but not an implicit, artificial language task (Tagarelli, Borges-Mota & Rebuschat, 2011), as well as earlier work by Reber and colleagues, who showed that aptitude measures influence explicit, but not implicit, learning processes (Reber, Walkenfeld & Hernstadt, 1991). Similarly, a recent neuroimaging study showed that in an artificial grammar learning task, individual differences in WM predicted the performance of participants who learned the task explicitly, but not those who learned implicitly (Yang & Li, 2012).

Participants

Ninety-eight monolingual participants aged 18 to 35 were recruited from the undergraduate population at the University of Texas; they were paid $10 per hour for their participation. No participant reported prior exposure to a tone language, or any history of neurological or hearing deficits. All participants provided informed consent and were debriefed following the conclusion of the experiment, in accordance with UT IRB requirements. Participants (n = 16) with poorer performance in the final block than the first were excluded, leaving data from 82 participants for the statistical analyses. Average accuracy in the final block for the excluded participants was 30%, which is close to chance (25%).

Stimuli for tone category training

The stimuli were identical to those in Experiment 1, with the exception that only one male and one female native speaker of Mandarin Chinese were included. This resulted in a shorter training experiment that allowed a WM measure to be included.

Procedure

The procedure was identical to that from Experiment 1 except that participants listened to 40 stimuli per block (four tone categories x five syllables x two speakers) across five blocks of training. In addition, following completion of the tone category learning task, each participant completed the immediate recall portion of the logical memory test in the Wechsler Memory Scale, third edition (WMS-III) (Wechsler, 1997). In this test, two stories were read at conversational rate. One story was read only once, the other twice. Participants were told to listen to the stories and repeat as much as they could recall immediately after each story was read, for a total of three story recall opportunities. The total number of key phrases or words correctly recalled across stories served as their raw score for the task.
Results: Overall accuracy

Figure 5b displays the average accuracy along with standard error bars. We also include average accuracy for participants best fit by a separation or non-separation model (discussed below). Participants showed significant learning across blocks (F(4, 324) = 133.59, p < .0001, partial $\eta^2 = 623$), with accuracy increasing from 40% in block 1 to 74% in block 5. A similar pattern held for tones spoken by female and male speakers: average accuracy increased from 40% in block 1 to 74% in block 5 for tones spoken by female and male speakers: average accuracy increased from 40% in block 1 to 74% in block 5. A similar pattern held for tones spoken by female and male speakers: average accuracy increased from 40% in block 1 to 74% in block 5.

Separation strategy distribution for final block separators and final block non-separators

Here we examined changes in the use of separation strategies across blocks by comparing the number of blocks of trials best fit by the separation model for final block separators and final block non-separators. As predicted (see Figure 7c), final block separators (3.84 blocks) used separation strategies in significantly more blocks than final block non-separators (1.58 blocks) (F(1, 80) = 56.90, p < .001, partial $\eta^2 = .416$).

We also examined the first block of trials for which a separation model provided the best fit to the data for final block separators and non-separators. As predicted (see Figure 7d), final block separators began using separate perceptual spaces earlier (1.95 blocks) than final block non-separators (2.68 blocks) (F(1, 80) = 5.09, p < .05, partial $\eta^2 = .060$).

Learning curves for final block separators and final block non-separators

Figure 5b shows the learning curves for final block separators and final block non-separators. A 2-model strategy x 5-block mixed ANOVA was conducted on these data. We observed a main effect of model strategy (F(1, 80) = 19.34, p < .001, partial $\eta^2 = .195$), with performance significantly better for separators (.67) than for non-separators (.44). We observed a main effect of block (F(4, 320) = 78.28, p < .001, partial $\eta^2 = .495$), suggesting significant learning. We also observed a significant model strategy by block interaction (F(4, 320) = 4.33, p < .005, partial $\eta^2 = .051$). Post hoc analyses suggested that performance of the separators was superior to that of the non-separators in every block (all p's < .05). Both groups showed significant learning, but a comparison of performance in block 1 with that in block 5 suggested that separators showed greater learning (an increase of .37) than non-separators (an increase of .22) (F(1, 80) = 8.64, p < .005, partial $\eta^2 = .097$).

Results: Modeling

The Experiment 1 model fitting and model comparison approach was used.

Best fitting model vs. random responder model

First we compared the Akaike weights from the best fitting model and compared that with the fit of the random responder model to determine whether the best fitting model was capturing noise or meaningful strategic responding. The resulting average Akaike weights were .944, .975, .990, .994, and .993 in blocks 1 to 5, respectively. In every case these values were significantly above 0.5 based on a one-sample t-test (all p's < .0001).

Best fitting non-separation model vs. best fitting separation model

Next we compared the Akaike weights from the best fitting separation model against the best fitting non-separation model. When a separation model provided the best account of the data, the Akaike weights ranged from .908 to .944 and in every block were significantly above 0.5 (all p's < .001). When a non-separation model provided the best account of the data, the Akaike weights ranged from .775 to .853 and in every block were significantly above 0.5 based on a one-sample t-test (all p's < .01). These findings suggest that the models are capturing meaningful strategic variance in the data.

Distribution of best fitting non-separation and separation model

Here we test the hypothesis that speaker separation leads to improved performance and thus that the number of participants best fit by one of the separation models will increase with experience. Figure 6b displays the proportion of participants whose data was best fit by a separation or non-separation model in each block. In support of our hypothesis, a $\chi^2$ test suggested that the number of separators increased while the number of non-separators decreased from the first to the final block of trials ($\chi^2(1, N = 82) = 22.61, p < .00001$).

2 It is worth noting that performance was very similar across the two experiments, despite the fact that four speakers were included (with 80 stimuli per training block) in Experiment 1 and only two (with 40 stimuli per training block) in Experiment 2. In Experiment 1, two male and two female speakers were included (see Figure 3) whereas in Experiment 2, one male and one female speaker were included. Importantly, the perceptual features (pitch height and pitch direction) were very similar across the two male and across the two female speakers. This likely explains why performance was similar in the two studies.
Working memory capacity and speaker separation
As outlined above, we expected that individuals who used speaker separation strategies would be more likely to have high WMC. As a test of this hypothesis we compared the WMS-III scores for final block separators and final block non-separators. As predicted, final block separators remembered significantly more story items (mean = 43.22; standard error = 1.21) than non-separators (mean = 37.95; standard error = 1.79), as measured by the WMS-III (F(1, 79) = 4.80, p < .05, partial η² = .057).

Discussion
In Experiment 1, we examined the extent to which reflective and reflexive strategies were used in tone category learning across different blocks. We showed that learners who used a reflexive strategy at the end of training were more accurate at categorization than those who used reflective strategies. Furthermore, our data demonstrated that a majority of learners use speaker-dependent strategies to deal with the multi-speaker variability in the training paradigm. The extent to which this strategy is beneficial in category learning success, and whether it is indeed dependent on WM (and therefore, reflective) was not addressed in Experiment 1. In Experiment 2 we used a larger sample, and examined WM ability to address both these issues. Our results showed that participants who used a speaker separation strategy showed superior learning to those who did not. Moreover, participants who did not use a speaker separation strategy showed lower WMC than participants who did. Experiment 2 thus shows that reflective strategy use is an important determiner of success in the category learning task.

General discussion
In this study we have provided a computational account of strategies that adults use while learning non-native speech sound categories using a dual-system framework. We used two experiments to study the learning of Mandarin tone categories by native English speakers with no prior experience of tone languages. Our results demonstrate significant individual differences in the computational strategies participants use during category learning. Importantly, these differences in strategies have a direct bearing on category learning success.

We hypothesized that tone category learning is reflexive-optimal, but that learners may use reflective, speaker-dependent strategies early in training. Consistent with these predictions, we demonstrated in Experiment 1 that learners who used reflexive strategies by the end of training were more accurate in category learning than those who used reflective learning strategies. We showed that successful learners initially separate perceptual spaces based on a reflective strategy (“male/female”) to perform category-related computations. Successful learners then shift to a reflexive computational strategy (SPC) by the end of the training session. According to the dual-system model, WMC is a critical component of the reflective learning system. Reflective learning is resource-intensive, requiring WM and executive attention to develop, test, and update verbal rules. We predicted that WMC may therefore be an important determiner of individual differences in category learning success. Indeed, in Experiment 2 we demonstrated that individuals who use the verbalizable speaker-dependent learning strategy also show enhanced learning and have higher WMC, as measured by standardized tests of WM.

The extent to which L2 learning in adulthood is explicit versus implicit has been a significant topic of inquiry in the SLA literature (DeKeyser, 2008). Much of this work has focused on grammar, morphology, and vocabulary learning. In contrast, much less is known about the role of explicit/implicit processes during phonetic learning. Speech sounds are typically difficult to verbalize and involve many redundant dimensions. These features argue for a dominance of the procedure-based reflexive learning system. Indeed, our results show that participants who use reflexive strategies at the end of training learn more accurately than those who use reflective strategies. However, some reflective strategies are advantageous in certain situations. For example, in multi-speaker training paradigms (Lively et al., 1994) speaker variability has been shown to be important for generalizing categories to novel speakers. Indeed, for speech learning to be effective in the real world, categorization performance must transfer to novel speakers. However, a previous tone category learning study showed that not all learners benefit from high speaker variability (Perrachione et al., 2011), but that some may find high variability training too resource-intensive, and may therefore be less successful. Why some learners have difficulty in such a high variability environment is unclear. Our results show large individual differences in computational strategies that learners use to deal with speaker variability. Tone categories are cued by speaker-dependent and speaker-independent dimensions. We predicted that an initial reflective strategy is to create speaker-dependent perceptual spaces. Our computational modeling results show that reduced effectiveness in using this strategy results in substantially lower category learning success. While a majority of participants use separate perceptual spaces for male and female speakers, some operate with a single perceptual space containing both male and female speakers. Across both experiments, our results demonstrate that learning is enhanced when participants use separate perceptual spaces. However, this is a reflective strategy, and therefore requires WM resources, so individuals with low WMC may have difficulty maintaining separate perceptual spaces for
different speakers. Individuals with high WMC, on the other hand, have an advantage in learning, a finding that addresses a source of individual differences in category learning success. Our results suggest a trade-off in category learning. Multiple-speaker training benefits real-world generalization, but requires WM resources, and thus may lead to individual differences in learning success.

Previous studies have focused on perceptual differences as a source of individual variability in speech learning success. For example, neuroanatomical and neuropsychological studies show that pre-training differences in auditory regions are a significant predictor of individual differences in successful speech learning (Golestani, Molko, Dehaene, LeBihan & Pallier, 2007; Wong, Chandrasekaran, Garibaldi & Wong, 2011; Wong, Perrachione & Parrish, 2007). Individual differences in phonetic acquisition under natural learning situations were found to relate to neural processing of speech sounds, as evidenced by preattentive change-detection electrophysiological responses, but not basic psychoacoustic differences between learners (Diaz, Baus, Escera, Costa & Sebastián-Galles, 2008). It is interesting that functional connectivity between frontal and parietal brain regions at rest also relates to individual differences in speech learning success (Ventura-Campos, Sanjuán, González, Palomar-García, Rodriguez-Pujadas, Sebastián-Gallés, Deco & Ávila, 2013). Our results show that individual differences in computational strategies may further contribute to variability in speech learning success.

Previous studies have viewed speech learning from the perspective of unsupervised learning (McClelland et al., 2002; Vallabha & McClelland, 2007) or interactive (lexically-mediated), supervised learning mechanisms (McClelland, Mirman & Holt, 2006; Mirman, McClelland & Holt, 2006). The role of feedback-based multiple learning systems, known to play an important role in visual category learning, is unclear. Our theoretical approach argues for two competing learning systems that operate when feedback is provided: a reflective learning system that uses feedback to develop rules, and an implicit, reflexive learning system that unconsciously associates feedback with stimulus and reward. Our computational approach demonstrates that one predictor of learning success is the type of computational strategy that participants employ while learning category differences. Thus, our framework views L2 speech learning as a category learning problem (Lotto, 2000).

We are aware that our computational approach is just a start to understanding dual-learning system influences on L2 speech learning. First, speech categories are extensively multidimensional (Diehl, Lotto & Holt, 2004), a factor that distinguishes speech learning from visual category learning (Lotto, 2000). The two-dimensional perceptual space (pitch height/direction) that is derived likely underestimates the inherent complexity in listeners’ processing of linguistic tone. However, our model fits suggest that this approach is a good way to start examining strategy use during speech learning. Future experiments should build on complexity to allow a better understanding of real-world speech processing. Second, we use Mandarin tone categories to evaluate speech learning. Previous research suggests several similarities between segmental (vowels and consonants) and suprasegmental speech learning. From a neural perspective, the extent to which lexical tones behave like segmental information has been extensively studied (Gandour & Dardarananda, 1983; Gandour, Wong, Hsieh, Weinzapfel, Van Lancker & Hutchins, 2000; Gandour, Dzemidzic, Wong, Lowe, Tong, Hsieh, Satthamnuwong & Lurito 2003; Gandour, Wong & Hutchins, 1998; Klein, Zatorre, Milner & Zhao, 2001; Xu, Gandour & Francis, 2006). For the purpose of this paper, three main points are particularly relevant. First, native speakers use a left hemisphere-dominant network to process linguistic tones that is indistinguishable from the network used to process other speech sound categories (Gandour et al., 1998; Gandour & Dardarananda, 1983; Wong et al., 2004b). Second, when non-native participants learn lexical tone categories, there is increased activity in the left-hemisphere anterior and posterior language regions (Wang, Sereno, Jongman & Hirsch, 2003b). Third, native listeners process Mandarin tones categorically, in a manner similar to consonants (Xi, Zhang, Shu, Zhang & Li, 2010). While this suggests a certain degree of consistency in the processing of linguistically “tainted” stimuli, the extent to which our dual-system approach can be applied to segmental learning is unclear. The SPC model has been successfully applied to the examination of vowel categorization (Maddox et al., 2002), suggesting a possible generalization to segmental learning. The applicability of our approach to learning segmental information is an important direction for future research. Finally, our data suggests that successful learners use separate perceptual spaces for male and female speakers, to deal with a high degree of variability across speakers. We argue that this strategy is an explicit process, because participants who use this strategy have higher WM ability than those who do not. However, many studies have shown that normalizing for speaker differences is an automatic process (Holt, 2006; Huang & Holt, 2012; Laing, Liu, Lotto & Holt, 2012), reliant on general auditory processes such as neural adaptation to the long-term average spectrum of speech rather than cognitive effort. We reconcile these differences with the fact that processing speaker information while categorizing L1 speech sounds may be fundamentally different from the role of speaker information during the learning of L2 speech categories. Initial learning of categories in a high speaker variability environment may be an effortful process requiring speaker-dependent analysis.
Indeed, a number of studies demonstrate a processing cost in multi-speaker paradigms, as well as showing more cognitive effort in mixed-speaker presentations than blocked-speaker presentations (Wong et al., 2004). However, with practice, and a switch to reflexive analysis within the perceptual space, such effortful analysis may not be necessary. Although the current experiments were not designed to distinguish between various models of speaker normalization, we believe that the computational modeling approach developed here could contribute to this research topic.

In conclusion, our computational modeling results demonstrate that learners use a variety of reflective and reflexive strategies to learn new categories. Successful learners use a combination of both types of strategy and are likely to become more reflexive with practice. The computational modeling approach developed here provides a foundation for future work on L2 speech learning and perception.

References


Tagarelli, K., Borges-Mota, M., & Rebuschat, P. (2011). The role of working memory in implicit and explicit language...


