COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 35 (2011) 184–197
Copyright © 2010 Cognitive Science Society, Inc. All rights reserved.
ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/j.1551-6709.2010.01140.x

Phonological Abstraction in Processing Lexical-Tone Variation: Evidence From a Learning Paradigm

Holger Mitterer,[a] Yiya Chen,[b] Xiaolin Zhou[c]

[a]*Max Planck Institute for Psycholinguistics*
[b]*Leiden University Centre for Linguistics & Leiden Institute for Brain and Cognition*
[c]*Department of Psychology, Peking University*

Received 29 April 2009; received in revised form 2 July 2010; accepted 3 July 2010

Abstract

There is a growing consensus that the mental lexicon contains both abstract and word-specific acoustic information. To investigate their relative importance for word recognition, we tested to what extent perceptual learning is word specific or generalizable to other words. In an exposure phase, participants were divided into two groups; each group was semantically biased to interpret an ambiguous Mandarin tone contour as either tone1 or tone2. In a subsequent test phase, the perception of ambiguous contours was dependent on the exposure phase: Participants who heard ambiguous contours as tone1 during exposure were more likely to perceive ambiguous contours as tone1 than participants who heard ambiguous contours as tone2 during exposure. This learning effect was only slightly larger for previously encountered than for not previously encountered words. The results speak for an architecture with prelexical analysis of phonological categories to achieve both lexical access and episodic storage of exemplars.

Keywords: Speech perception; Lexical tone; Mandarin Chinese; Phonological abstraction; Episodic models

1. Introduction

The mental lexicon is a memory store for the words of the language(s) we speak. But in what format is this knowledge stored and, consequentially, how can it be accessed for online word recognition? In an influential paper, Klatt (1989) contrasted two strategies: lexical access from spectra versus lexical access from features. The first strategy assumes that the mental lexicon contains detailed acoustic representations of how a given word sounds in a variety of contexts and from a variety of different speakers. The lexicon then has to be

Correspondence should be sent to Holger Mitterer, Max Planck Institute for Psycholinguistics, P.O. Box 310, NL-6500 AH Nijmegen, The Netherlands. E-mail: holger.mitterer@mpi.nl accessed by ‘‘grainy spectrograms’’ with no intermediate representations. According to the second strategy, an intermediate stage mediates between the acoustic input and abstract representations in the mental lexicon, by generating an abstracted form of the input (in terms of phonetic features, phonemes, allophones, or demi-syllables) that are independent of speaker and context.

Going beyond this dichotomy, there is in fact evidence for both episodic and abstract representations operating in perception. Evidence for episodic storage comes from three sources. First, listeners are able to retain detailed acoustic information about individual tokens of spoken words (for a review, see Goldinger, 1998). Second, Connine, Ranbom, and Patterson (2008) showed that the efficiency of recognition of words with deleted schwas (e.g., *camera* [kæmərɑ] produced as [kæmrɑ]) depends on how often schwa deletion occurs in a given word. In a similar vein, Pitt (2009) showed that variant forms of newly learned words are only recognized if the variant form has been encountered before. Finally, in linguistics, it has been argued that ‘‘near-mergers’’ are best explained in terms of an exemplar model (Yu, 2007). A near-merger describes the situation that a distinction between two phonological categories is blurred so strongly that it is not functional anymore in perception, but still maintained probabilistically in production. Yu explained this phenomenon by proposing largely overlapping clouds of exemplars stored for the two categories.

Evidence for abstract representations comes from a study by McQueen, Cutler, and Norris (2006). They used a paradigm in which a phoneme representation is ‘‘recalibrated’’ on the basis of experience with a limited set of words. Dutch listeners were first exposed to words in which the natural [s] or [f] was replaced by an ambiguous fricative between [f] and [s] (henceforth referred to as [$^{s}{}_{f}$]). One group of listeners heard [$^{s}{}_{f}$] replacing the [f] in /f/-final words; the other group heard [$^{s}{}_{f}$] replacing the [s] in /s/-final words. Replicating the typical lexical bias (Ganong, 1980), listeners perceived the [$^{s}{}_{f}$] as [f] in /f/-final words and as [s] /s/-final words (as indicated by ‘‘yes’’ responses to words with ambiguous fricatives in a lexical decision task). McQueen et al. (2006) further showed that such an exposure had consequences for the perception of other /s/- and /f/-final words. Listeners exposed to the ambiguous fricative in /s/-final words interpreted the ambiguous sound in new words as /s/ during the test phase, whereas those exposed to the ambiguous fricative in /f/-final words interpreted the ambiguous sound in new words as /f/. This suggests that listeners not only learned that the speaker produced the words that happened to be presented in the exposure phase with an ambiguous fricative; they also learned that the speaker produces the abstract phoneme /s/ (or /f/) with an ambiguous fricative.

Over the last decade, the existing evidence for both abstract and episodic representations has led to a growing consensus that the mental lexicon should be considered a hybrid which entails both episodic and abstract representations (Cutler & Weber, 2007; Goldinger, 2007). Such a consensus, however, has a serious drawback, because any hybrid model is obviously difficult to falsify. In order to constrain the possible architectures of a hybrid model, we here strove to explicitly pit episodic learning against phonological abstraction and numerically compare their effectiveness in the same paradigm.

This is possible in the exposure-test design used by McQueen et al. (2006) if some minimal pairs occur in the exposure phase, and in the test phase, learning is compared to old, that is, previously encountered words, with learning for new words. As this design requires many minimal pairs, we tested this with minimal tone pairs in Mandarin Chinese, using the tone contrast between the high-level tone1 and the mid-rising tone2 (e.g., *bo1* ''wave'' *bo2* ''thin''). Just as in McQueen et al. (2006), the experimental sessions contained an exposure and a test phase. During the exposure phase, participants heard syllables with an ambiguous contour between tone1 and tone2. In a between-subject manipulation, the perception of these ambiguous contour was influenced by sentence context, so that half of the participants learned to associate the ambiguous contour with tone1 and the other half with tone2.

In the test phase, listeners heard syllables from a tone1–tone2 continuum and had to decide whether they perceive the high-level tone1 or the rising tone2. Half of the target syllables in the test phase had already occurred in the exposure condition, and the other half were new. If there is a recalibration of the tone1–tone2 contrast, we expect the listeners who heard ambiguous contours in a tone2 biasing context in the exposure phase to label ambiguous contours in the test phase as tone2 more often then listeners who heard ambiguous contours in a tone1 biasing context during exposure. If phonological abstraction is more important, there should be strong learning effects for both old and new words. If episodic storage is more important, learning effects should be much weaker for new than for old words.

It is worth noting that previous investigations of similar recalibration effects (Eisner & McQueen, 2005, 2006; Kraljic & Samuel, 2005, 2006; McQueen et al., 2006; Norris, McQueen, & Cutler, 2003) have focused on segmental contrasts, in which case, the essential acoustic cues (i.e., voice over time in stop voicing and fricative spectra) are concentrated on a small stretch of the acoustic signal. Contrasts between lexical tones are quite different, as the acoustic cues are distributed over a much longer stretch of time (often the complete tone-carrying syllable). Previous research uncovered similar normalization processes in tone and segment perception: Listeners normalize in vowel and consonant perception according to the segmental context (Mann, 1980), speaker variation (Ladefoged & Broadbent, 1957), as well as sentence-level prosody (Kuzla, Ernestus, & Mitterer, 2010). The normalization for lexical tones shows similar inter- and intratalker context effects (Francis, Ciocca, Wong, Leung, & Chu, 2006; Leather, 1983; Moore & Jongman, 1997; Wong & Diehl, 2003; Xu, 1994). If we do find evidence that the recalibration of tone contrasts is analogous to that of segmental contrasts, we would then provide further converging evidence that tone contours in tone languages are processed similarly as segmental phonetic features (in both tone and other languages).

In addition to testing the possibly different recalibration effects for previously encountered versus new words, we also investigated to what extent recalibration effect is moderated by other variables. To this end, we manipulated the tone context in which the critical syllables appeared. It is well documented that the phonetic implementation of Mandarin tones depends on the neighboring tones with great carryover effect from the preceding tone (Chen & Gussenhoven, 2008; Xu, 1994). During the exposure phase, the critical words were preceded by a falling tone with a low offset. During the test phase, the critical words were either preceded by a falling tone with a low offset (like in the exposure test) or a high-level tone with a high offset (i.e., a new phonetic context).

2. Method

2.1. Participants

Eighty students at the Peking University participated in the experiment for pay. The participants, of whom 66 were female, were aged between 18 and 27. They were native speakers of Standard Chinese, who were born and grew up in Beijing.

Forty participants heard the critical words in the same tonal context during exposure and test. Half of these were assigned to the exposure condition with clear tone1 contours but ambiguous contours on tone2-bearing targets, the other half to the exposure condition with clear tone2 contours and the ambiguous contours on tone1-bearing targets.

The other 40 participants heard the critical words in a different tonal context during exposure and test. Twenty-one of these participants were assigned to the exposure condition with clear tone1 contours and the ambiguous contours on tone2-bearing targets, the other nineteen to the exposure condition with clear tone2 contours and the ambiguous contours on tone1-bearing targets.

2.2. Stimuli

All auditory stimuli were recorded by a male native speaker of Standard Chinese. The stimuli used in the exposure phase were recorded in the sentence frame *ta1 xie3 … zhe4 ge4 ci2* ''he wrote the word….'' The empty slot in this frame was filled by a member of one of the twenty tone1–tone2 minimal pairs (e.g., *bo1* ''wave'' *bo2* ''thin'') plus an accompanying syllable that disambiguated which member of the minimal pair was intended (e.g., *du3 bo2* ''to gamble'' and *duan3 bo1* ''short wave''). Appendix S1 lists all experimental items. We also recorded 160 filler sentences in which the empty slot was filled by two syllables carrying tone3 or tone4, so that there were no targets with tone1 or tone2 besides the minimal pairs.

To generate ambiguous syllables for the exposure phase, the critical syllables from the tone1–tone2 minimal pairs were excised from both the exposure sentence with a tone1 bias and those with the tone2 bias. The excised syllables were zero-padded with 25 ms of silence to allow valid pitch estimation at the start and the end of the utterance. The pitch curves of both members of a pair were estimated using PRAAT (Boersma, 2001). Tone1–tone2 continua were then generated by interpolating (in 10 equal steps on a semitone scale) between the two pitch curves. This gives rise to 11 steps (100% tone1-0% tone2, 90% tone1-10% tone2, … , 0% tone1-100% tone2). Fig. 1 shows the result of this interpolation for the syllable *qin*. Using the Pitch-Synchronous-Overlap-and-Add (PSOLA) method, the syllables with an original tone1 contour were then resynthesized after replacing the original contour with the 11 altered f0 contours. The changed syllables were then reinserted into the original sentence frames after removal of the 25 ms of silence at the beginning and end. For the exposure items, a pretest (see Appendix S2 and Figure 2) was used to determine an ambiguous version of the syllable, which was strongly susceptible to contextual effect. Note that

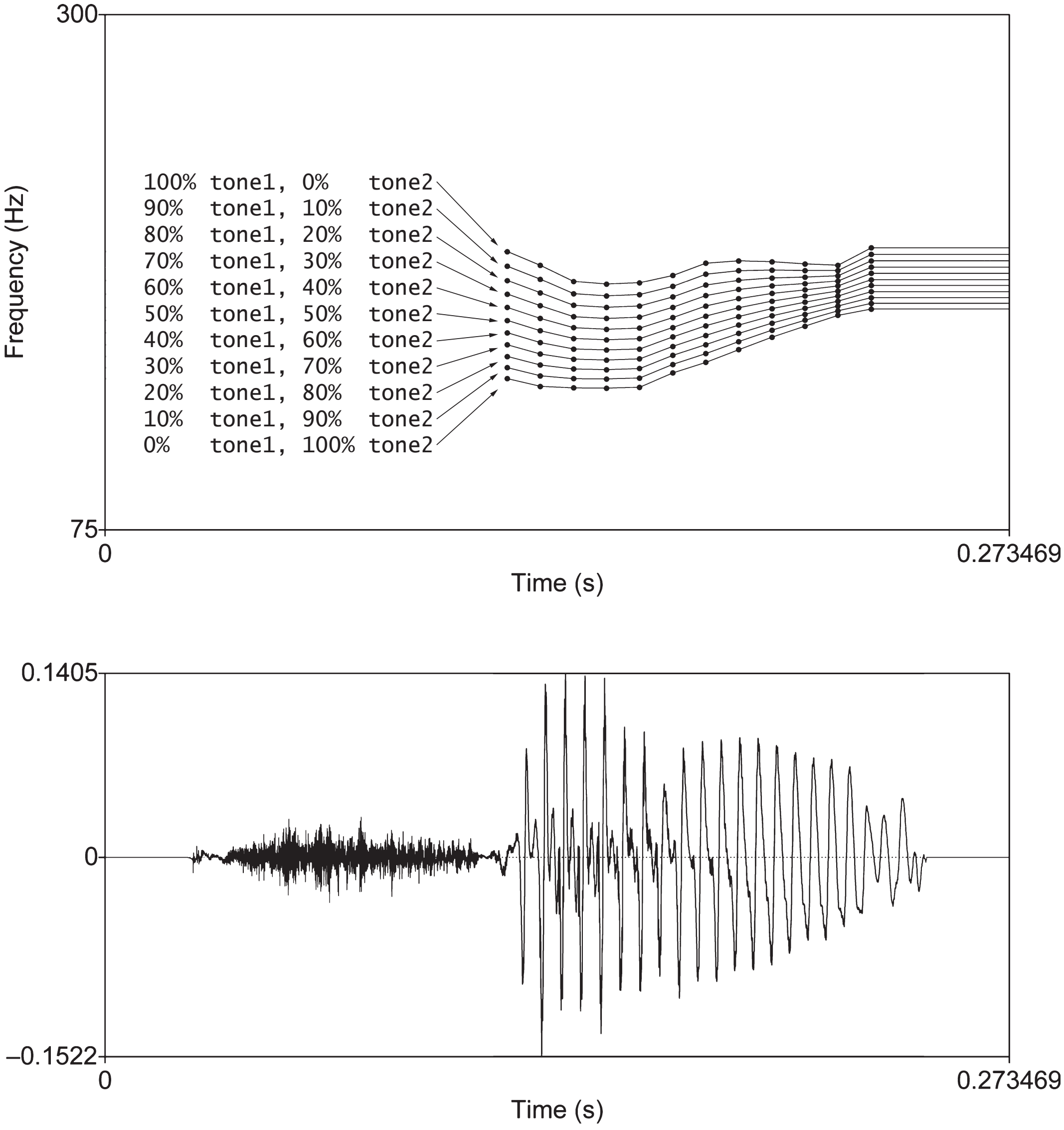

Fig. 1. Pitch curve (upper panel) for the qin1 (uppermost line) and qin2 (lowest line) synchronized with the wave form (lower panel). The intermediate lines show the different steps of the tone1–tone2 continuum for this syllable.

this was not necessary for the items used at test, in which a range of different stimuli are presented.

For the test phase, there were two different sentence frames, one in which the critical word was preceded by a falling tone, just as in the exposure (*ta1 xie3 … zhe4 ge4 ci2* ''he wrote the word…''), and one in which the critical word was preceded by a high-level tone (''*ta1 shuo1 … zhe4ge4ci2*'' ''he said the word…''). Eighty sentences for each frame were

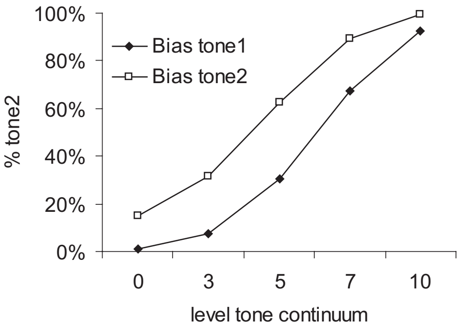

Fig. 2. Mean percentages of tone2 identifications in the pretest depending on semantic bias and tone continuum averaged over all syllables.

generated by filling the empty slot with a member of one of forty tone1–tone2 minimal pairs. Twenty of these pairs had been already used in the exposure test and the other twenty were new minimal pairs. Just as for the exposure phase, we generated an 11-step continuum by excising the members of the minimal pair from the sentence context and interpolating the pitch contours. The stimuli with an interpolated contour with 80% tone1–20% tone2, 60% tone1–40% tone2, 50% tone1–50% tone2, 40% tone1–60% tone2, and 20% tone1–80% tone2 were used in the test phase. With this selection, participants could still make phonetically based decisions given the phonetic variation. As our main goal was to test the perception of ambiguous contours, we excluded the unambiguous continuum endpoints (see Norris et al., 2003, for a similar stimulus selection for the test phase).

For both the monitoring task in the exposure phase and the 2AFC task in the test phase, we generated bitmaps with Chinese characters. The characters had an approximate size of 50 × 50 pixels and were presented on a white rectangle of 189 × 113 pixels.

2.3. Apparatus and procedure

The experiment was run on a standard PC. Stimulus presentation was controlled by MATLAB using the Psychophysics Toolbox (version 2.54) (Brainard, 1997). Participants were seated in front of the monitor and reacted to the stimuli with the mouse.

The participants first responded to 200 trials on which they had to monitor whether one visually depicted character (henceforth: target) was present in the sentence they heard. Their task was to click with any mouse button if the sentence contained the target. The sentence started 600 ms after the onset of the visual target. Participants could react until 1.5 s after sentence offset. After a reaction or a time-out, there was a 400-ms interval before the next trial started.

Table 1 indicates the different types of trials of the exposure phase. The visual target character referred 45 times to a tone1 word or a tone2 word, and 55 times to a tone3 or tone4 word. For the 90 trials with either a tone1 or a tone2 target, there were 40 trials in which the target was present in the sentence either in a clear or an ambiguous form. These were the critical exposure trials. Half of the participants heard the tone1 syllables with an unaltered

Table 1
Number of trials for each tone category in the exposure conditions for the two different groups, respectively

Tone of Visual Target Character	Group 1 (Tone1 Bias) Target Present			Group 2 (Tone2 Bias) Target Present		
	Yes	No	Ambiguous	Yes	No	Ambiguous
Tone1	–	25	20	20	25	–
Tone2	20	25	–	–	25	20
Tone3	30	25	–	30	25	–
Tone4	30	25	–	30	25	–
Total	80	100	20	80	100	20

Note. The visual target character refers to words presented in the auditory sentence.

tone1 and the tone2 syllables with an ambiguous contour. The other half of the participants heard the same ambiguous contour in the tone1 words and natural versions of the tone2 words. In the remaining 110 trials with tone3- or tone4-bearing targets, the target was present in the sentence on 60 trials so that overall there was a balanced number of ''target present'' and ''target absent'' trials. The 200 trials of the exposure phase were presented in a different random order for each participant, with no critical items in the first 10 trials.

In the test phase, participants heard the syllables of tone1–tone2 minimal pairs in one of the two neutral sentence frames (''*ta1 xie3/shuo1 … zhe4ge4ci2*'' ''he wrote/said the word…''), which varied between subjects. On each trial, two characters were presented on the screen, referring to the tone1 and tone2 interpretation of the target syllable on that trial. That is, if the syllable *bo* was presented acoustically, the characters for *bo1* and *bo2* were presented on the screen. The sentence started 0.2 s after the onset of the two visual characters. The participants then had to move the mouse over one of the characters and click on it. They had 5 s time from the onset of the sentence for this task. Clicks on other parts of the screen were ignored. During the test phase, each of the 40 minimal pair syllables (20 old and 20 new) was presented five times, each time with one of the five f0 contours used for testing (levels 3, 5, 6, 7, and 9 of the tone1–tone2 continuum; see also the Stimuli section). A different random order was used for each participant.

2.4. Design and analysis

For the exposure phase, the design entailed two factors, one within and one between subjects. The within-subjects factor was whether the target to be monitored for is underlyingly a tone1 word or a tone2 word. The between-subjects variable was Exposure condition, that is, whether ambiguous contours were presented on tone1 or tone2 words.

For the test phase, there were two between-subject variables: Exposure condition (tone1 ambiguous vs. tone2 ambiguous) and Tone Context (same vs. different as in the exposure). The within-subject variables were (a) the continuum steps (as a continuous predictor) and (b) old versus new word. The critical dependent variable was whether the test syllable was heard as tone1 or tone2. We expected that participants exposed to ambiguous tone1 words in the exposure phase would be more likely to interpret ambiguous contours in the test phase as tone1. To test whether this effect was moderated by other factors, we added the following interactions to the model: Exposure Group × Tone Context (same vs. different in Exposure and Test), Exposure Group by Old/New Word, and Exposure Group by Trial Number, which tests whether effects change over the course of the experiment. The statistical significance of the associations between dependent and independent variables was tested with linear mixed effect models in R with items and subject as random factors, using the lme4 package (Baayen, Davidson, & Bates, 2008). For categorical outcome variables, such as acceptance (yes/no), the binomial linking function was used (Jaeger, 2008).

3. Results

3.1. Exposure

The overall performance was quite accurate with 99% correct responses in the filler trials. In the critical trials, acceptance rates were somewhat lower (see Table 2) but still close to ceiling. There is a slightly higher acceptance rate for tone2 targets. Table 2 also shows the reaction times (RT) measured from word offset, excluding 78 trials on which a response was given before target offset. An analysis of the RT data revealed an interaction of Exposure Condition by Word Tone ($b_{\text{condition by Word Tone}} = -39$, $p < .05$, responses being faster if words carried a natural and not an ambiguous contour) and a significant decrease of RT over trials ($b_{\text{trial}} = -0.66$, $p < .001$). No other effects were significant.

3.2. Test

Fig. 3 shows how often the tokens in the test phase were categorized as tone2, with the results for old words (i.e., words used in the exposure phase) in the left panels and those for new words in the right panels. The top panels show data for the congruent condition in which the tone context was the same Low tone in Exposure and Test, whereas the lower panels show data for the incongruent condition (where in the Test, the target words were preceded by a High tone). In all panels, there is a clear effect of exposure: Participants exposed to ambiguous contours on tone2 targets during exposure give more tone2 responses

Table 2
Acceptance rates and estimated marginal means for natural clear versus synthesized ambiguous contours in the exposure phase

	Exposure Condition	
%Present RT	Tone1 Clear Tone2 Ambiguous	Tone2 Clear Tone1 Ambiguous
Tone1 targets	97.5% 389 ms	95.1% 405 ms
Tone2 targets	98.4% 427 ms	97.5% 404 ms

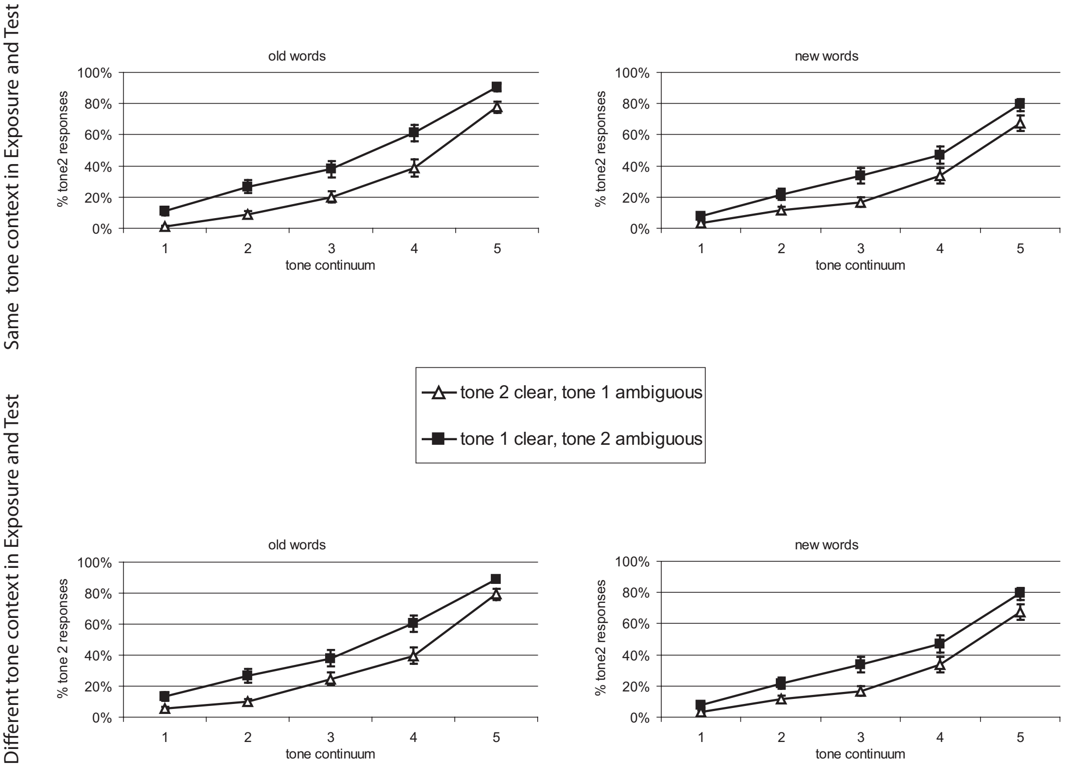

Fig. 3. Mean percentage of tone2 responses in the test phase depending on learning condition, match of tone context in Exposure and Test phase, and old versus new word for all steps of the tone1–tone2 continuum. Confidence intervals are based on the average standard error in the statistical model, which have been transformed back from the logistic space employed in the statistical analysis to the probabilities in the raw data.

during the test phase than participants exposed to ambiguous contours on tone1 targets. For ease of comparison, Fig. 4 summarizes the learning effects over the four panels of Fig. 3. It shows the overall difference in tone2 judgments made in the test phase between the group that heard ambiguous contours in the exposure phase as tone2 and the group that heard the ambiguous contours in the exposure phase as tone1. A positive value hence indicates recalibration.

The data analysis revealed the following significant effects (with $p < .01$ unless otherwise noted). The proportion of tone2 percepts increased over the continuum, so that there were (unsurprisingly) more tone2 responses the more tone2-like the contour was. Tone2 percepts were also more likely as the experiment progressed, with a significant effect of trial number, leading to 36.0% tone2 responses in the first half and 38.7% in the second half of the experiment. Importantly, there was also an effect of exposure condition. Participants in the group who heard the ambiguous contours in the exposure as tone2 gave more tone2 responses in the test phase (44.2%) than the group (31.1%) who heard the ambiguous contours as tone1 during exposure. The learning effect got significantly smaller as the test phase progressed

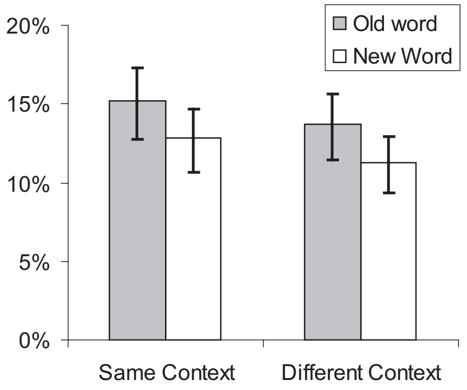

Fig. 4. Learning effects for old and new words in a different or the same phonetic context as in the exposure phase. Effects are expressed as the overall difference in tone2 percepts between the two learning groups. Confidence intervals are based on the standard error of the (significant) regression weight for Condition × Old/New Word in the statistical model, which have been transformed back from the logistic space employed in the statistical analysis to the probabilities in the raw data.

(15.0% in the first half vs. 10.8% in the second half, a condition by trial interaction). The learning effect was also significantly smaller with new words (11.8%) than with old words (14.4%, a condition by old vs. new interaction $p < .05$). There was no significant interaction of Tone Context with the Exposure Condition ($p > .2$).

4. Discussion

The current experiment tested whether lexically guided recalibration occurs for tone contrasts, how strongly learning generalizes to novel words, and to what extent learning is moderated by the phonetic context. As in previous experiments (Eisner & McQueen, 2005, 2006; Kraljic & Samuel, 2005, 2006; van der Linden & Vroomen, 2007; McQueen et al., 2006; Norris et al., 2003), we adopted a two-phase test procedure. Exposure to ambiguous tone contours in a biasing context led to recalibration of the tone contrast. Furthermore, this recalibration was effective for both old and new syllables in the test phase but was slightly stronger for old than for new words. The recalibration was not moderated by the phonetic context. These results show that such a form of recalibration is not restricted to acoustically ‘‘local’’ segmental contrast used in previous studies (for a review, see Samuel & Kraljic, 2009) but also applies to acoustically more distributed tone contrasts, therefore extending the generality of perceptual learning in speech perception.

Our main rationale was to directly compare the strength of episodic versus phonological abstract learning. We found evidence for both. Learning generalized from old to new words, which speaks for abstract phonological learning. The learning effect was, however, slightly larger for old than for new words, suggesting additional effects of episodic learning on a lexical level. This empirically strengthens the evolving consensus that models of word recognition need to incorporate both abstract phonological mechanisms as well as listeners’ ability to store individual exemplars of words. Furthermore, our results indicate that phonological learning is much more potent than episodic learning.

How to model the effects found in this experiment? Both extreme abstractionist models, such as Shortlist (Norris, 1994), and purely episodic models, such as Minerva (Goldinger, 1998), have trouble explaining the current data straightforwardly. Specifically, purely abstractionist models fail to explain the word-specific effects, and purely episodic models are challenged by the generalization of learning to new words. It is clear that a hybrid model is needed to explain the current results; furthermore, the architecture of such a hybrid model also needs to be constrained to explain the asymmetrical effects of the abstract phonological versus episodic learning.

One may conceive of models in which phonological categories are functional in lexical access or are epiphenomena of lexical access. In the first case, the input activates prelexical units, generating an abstraction of the acoustic signal with which the lexicon is addressed. In the latter case, the signal, often conceptualized as a grainy spectrogram (Pierrehumbert, 2002), directly accesses the mental lexicon, which would still consist of episodic traces without abstraction (Johnson, 1997). Abstraction only occurs after lexical access when the lexical entries in turn activate phonological categories. Phonological categories are then defined by all the words in which they occur. It is, however, difficult to see how such models can account for the current finding. If tone categories are defined by all the relevant words, the empirical basis for generalization is rather small. Participants are only exposed to 20 ambiguous tone contours. A lexical database for Modern Mandarin (http://lingua.mtsu.edu/chinese-computing/) lists 300 different tone1 syllables and 242 tone2 syllables. Therefore, in the experiment, listeners were exposed to <10% of the lexicon ([20 words in the exposure]/>[200 words in the lexicon]). Given that the episodic representations of about 90% of the words within the category (or cluster of exemplars) are not influenced by the experimental exposure, we expect that a postlexical tone category should be altered only slightly, if at all. Accordingly, the word-specific effect should be stronger than the generalization effect. This prediction is not borne out by our results, because the phonological learning effect is five times stronger than the word-specific learning effect.

A model that fits well with the relative strength of the episodic and categorical effects is the production model proposed by Pierrehumbert (2002). In this model, word-specific effects are second-order effects while categorical processing provides the backbone on which speech production operates. Without such categorical machinery, it cannot explain the relative strengths of abstract and episodic learning found here, nor can it explain more complex speaker normalization effects (Mitterer, 2006). Such an intervening level explains generalization of the current and other types of phonetic learning (Davis, Johnsrude, Hervais-Adelman, Taylor, & McGettigan, 2005; Maye, Aslin, & Tanenhaus, 2008).

While arguing for abstraction, the current results also underline the importance of storing episodic traces at different levels of processing. First of all, recalibration effects can be speaker specific (Eisner & McQueen, 2005) and may not be automatically driven by the acoustic input, but modulated by previous or a concurrent visual experience of the speaker (Kraljic, Samuel, & Brennan, 2008). These highly dynamic effects can only be explained if one assumes that episodic information, which encodes how particular speakers produce particular categories, is retained at a prelexical level of processing. Additionally, the results also indicate that listeners retain information about how speakers produce particular words.

It is worth noting that the emphasis of previous research has been to show that episodic information is encoded at all (Goldinger, 1996); the current results, however, indicate that such memory traces are in fact functional in word recognition. In other words, hearing a word in a phonetically ambiguous form not only creates an episodic memory but also influences the recognition of new tokens of this word with similar forms. Such lexical episodic storage, however, contributes less to spoken-word recognition than prelexical abstraction. The main burden of word recognition thus seems to lie on the prelexical abstraction (from the speech signal input) with which lexical access can be achieved efficiently. Because prelexical and lexical codes must be commensurable, it then follows that lexical representations should be abstract as well. Such an architecture facilitates fast adaptation to inter- and intraspeaker variation, which is necessary for effective and efficient speech communication and would otherwise be difficult to achieve.

Acknowledgments

We would like to thank Yingyi Luo for running the experiment. Support from the Netherlands Organization for Scientific Research (NWO-VIDI 016084338) and the European Research Council (ERC-Starting Independent Researcher Grant 206198) are gratefully acknowledged.

References

Baayen, H. R., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*, 390–412.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International*, *5*, 341–345.

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

Chen, Y., & Gussenhoven, C. (2008). Emphasis and tonal implementation in Mandarin Chinese. *Journal of Phonetics*, *36*, 724–746.

Connine, C. M., Ranbom, L. J., & Patterson, D. J. (2008). Processing variant forms in spoken word recognition: The role of variant frequency. *Perception & Psychophysics*, *70*, 403–411.

Cutler, A., & Weber, A. (2007). Listening experience and phonetic-to-lexical mapping in L2. In J. Trouvain & W. J. Barry (Eds.), *Proceedings of the 16th International Congress of Phonetic Sciences* (pp. 43–48). Dudweiler, Germany: Pirrot.

Davis, M. H., Johnsrude, I. S., Hervais-Adelman, A., Taylor, K., & McGettigan, C. (2005). Lexical information drives perceptual learning of distorted speech: Evidence from the comprehension of noise-vocoded sentences. *Journal of Experimental Psychology: General*, *134*, 222–241.

Eisner, F., & McQueen, J. M. (2005). The specificity of perceptual learning in speech processing. *Perception & Psychophysics*, *67*, 224–238.

Eisner, F., & McQueen, J. M. (2006). Perceptual learning in speech: Stability over time. *Journal of the Acoustical Society of America*, *119*, 1950–1953.

Francis, A. L., Ciocca, V., Wong, N. K. U., Leung, W. H. Y., & Chu, P. C. Y. (2006). Extrinsic context affects perceptual normalization of lexical tone. *Journal of the Acoustical Society of America*, *119*, 1712–1726.

Ganong, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance*, *6*, 110–125.

Goldinger, S. D. (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 1166–1183.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, *105*, 251–279.

Goldinger, S. D. (2007). A complementary-systems approach to abstract and episodic speech perception. In J. Trouvain & W. J. Barry (Eds.), *Proceedings of the 16th International Congress of Phonetic Sciences* (pp. 49–54). Dudweiler, Germany: Pirrot.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*, 434–446.

Johnson, K. (1997). Speech perception without speaker normalization: An exemplar model. In K. Johnson & J. Mullennix (Eds.), *Talker variability in speech processing* (pp. 145–165). San Diego, CA: Academic Press.

Klatt, D. (1989). Review of selected models of speech perception. In W. D. Marslen- Wilson (Ed.), *Lexical representation and process* (pp. 169–226). Cambridge, MA: MIT Press.

Kraljic, T., & Samuel, A. G. (2005). Perceptual learning for speech: Is there a return to normal? *Cognitive Psychology*, *51*, 141–178.

Kraljic, T., & Samuel, A. G. (2006). Generalization in perceptual learning for speech. *Psychonomic Bulletin and Review*, *13*, 262–268.

Kraljic, T., Samuel, A. G., & Brennan, S. (2008). First impressions and last resorts: How listeners adjust to speaker variability. *Psychological Science*, *19*, 332–338.

Kuzla, C., Ernestus, M., & Mitterer, H. (2010). Compensation for assimilatory devoicing and prosodic structure in german fricative perception. In C. Fougeron & M. D'Imperio (Eds.), *Laboratory phonology 10* (pp. 731–758). Berlin: Mouton.

Ladefoged, P., & Broadbent, D. E. (1957). Information conveyed by vowels. *Journal of the Acoustical Society of America*, *27*, 98–104.

Leather, J. (1983). Speaker normalization in perception of lexical tone. *Journal of Phonetics*, *11*, 373–382.

van der Linden, S., & Vroomen, J. (2007). Recalibration of phonetic categories by lipread speech versus lexical information. *Journal of Experimental Psychology: Human Perception and Performance*, *33*, 1483–1494.

Mann, V. A. (1980). Influence of preceding liquid on stop-consonant perception. *Perception & Psychophysics*, *28*, 407–412.

Maye, J., Aslin, R. N., & Tanenhaus, M. K. (2008). The weckud wetch of the wast: Lexical adaptation to a novel accent. *Cognitive Science*, *32*, 543–562.

McQueen, J. M., Cutler, A., & Norris, D. (2006). Phonological abstraction in the mental lexicon. *Cognitive Science*, *30*, 1113–1126.

Mitterer, H. (2006). Is vowel normalization independent of lexical processing? *Phonetica*, *63*, 209–229.

Moore, C. B., & Jongman, A. (1997). Speaker normalization in the perception of Mandarin Chinese tones. *Journal of the Acoustical Society of America*, *102* (3), 1864–1877.

Norris, D. (1994). Shortlist: A connectionist model of continuous speech recognition. *Cognition*, *52*, 189–234.

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology*, *47*, 204–238.

Pierrehumbert, J. (2002). Word-specific phonetics. In C. Gussenhoven & N. Warner (Eds.), *Laboratory phonology VII* (pp. 101–139). Berlin: Mouton de Gruyter.

Pitt, M. A. (2009). How are pronunciation variants of spoken words recognized? A test of generalization to newly learned words. *Journal of Memory and Language*, *61*, 19–36.

Samuel, A. G., & Kraljic, T. (2009). Perceptual learning for speech. *Attention, Perception, & Psychophysics*, *71*, 1207–1218.

Wong, P. C. M., & Diehl, R. L. (2003). Perceptual normalization for inter- and intratalker variation in Cantonese level tones. *Journal of Speech, Language, and Hearing Research*, *46*, 413–421.

Xu, Y. (1994). Production and perception of coarticulated tones. *Journal of the Acoustical Society of America*, *95*, 2240–2253.

Yu, A. C. L. (2007). Understanding near mergers: The case of morphological tone in Cantonese. *Phonology*, *24*, 187–214.

Supporting Information

Additional Supporting Information may be found in the online version of this article on Wiley Online Library:

Appendix S1: Experimental items.

Appendix S2: Pretest for exposure items.

Please note: Wiley-Blackwell is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.