INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Lexically Guided Perceptual Learning in Mandarin Chinese

L. Ann Burchfield[1]*, San-hei Kenny Luk*[2], *Mark Antoniou*[1], *Anne Cutler*[1]

[1]The MARCS Institute, Western Sydney University, Australia
[2] Department of Linguistics & Languages, McMaster University, Canada
a.burchfield@westernsydney.edu.au

Abstract

Lexically guided perceptual learning refers to the use of lexical knowledge to retune speech categories and thereby adapt to a novel talker's pronunciation. This adaptation has been extensively documented, but primarily for segmental-based learning in English and Dutch. In languages with lexical tone, such as Mandarin Chinese, tonal categories can also be retuned in this way, but segmental category retuning had not been studied. We report two experiments in which Mandarin Chinese listeners were exposed to an ambiguous mixture of [f] and [s] in lexical contexts favoring an interpretation as either [f] or [s]. Listeners were subsequently more likely to identify sounds along a continuum between [f] and [s], and to interpret minimal word pairs, in a manner consistent with this exposure. Thus lexically guided perceptual learning of segmental categories had indeed taken place, consistent with suggestions that such learning may be a universally available adaptation process.

Index Terms: speech perception, adaptation, Mandarin, phonology

1. Introduction

Listeners encounter enormous variability in the sounds of everyday speech. Multiple talkers, multiple accents, variable speech rates, and background noise are all normal parts of speech recognition. Typically, listeners easily adapt to this variability and understand speech well regardless of it. Most listeners encounter on a daily basis, in person or via the media, talkers they have never heard before; rarely do they experience trouble understanding such talkers, even though variation across vocal tracts can be very large. With some practice, listeners are also able to adjust to foreign-accented speech [1] and even highly degraded speech, such as vocoded [2] or compressed [3] speech.

Previous work has found that listeners can adapt to newly encountered talkers by drawing on existing knowledge, such as knowledge of words, to make talker-specific adjustments to phoneme category boundaries. This has been demonstrated in experiments in which listeners are exposed to ambiguous sounds in a context that favors one interpretation over the other [4,5,6]. For example listeners might be exposed to an ambiguous sound halfway between [s] and [f]. If this sound occurs in contexts such as *gira?*, listeners tend to interpret the sound as [f] because *giraffe* is a word but *girasse* is not. Conversely, if the same sound occurs in contexts such as *hor?*, the sound tends to be interpreted as [s] because *horse* is a word but *horf* is not. After this kind of training, listeners continue to interpret the ambiguous sound if uttered by the same talker in a manner consistent with their exposure, even in contexts that do not favor one interpretation over the other. This process is termed lexically guided perceptual learning.

Lexically guided perceptual learning has been extensively studied. It has been found to be long-lasting [7,8] and to occur automatically, without explicit attention from the listener [9]. It has been observed for a variety of phoneme types including fricatives [4,6-9], stops [10], and vowels [11] and phonemes in both word-medial [10,11] and word-final [4-9] positions, and the learning generalizes across the vocabulary [12]. These findings suggest a robust and highly generalizable adaptation process. However, the large majority of these experiments have focused on English [7,10,11] and Dutch [4,6,8,9,12]. Studies of lexically guided perceptual learning outside this small group of languages have been much more limited.

In particular, it has not been studied extensively in tonal languages, such as Mandarin Chinese. Lexically guided retuning of tonal categories has been observed in Mandarin [13]; an ambiguous tone disambiguated by lexical knowledge can induce talker-specific adaptation, similar to that demonstrated for phonemic segments in English and Dutch experiments. But such retuning has not been studied for segments in tonal languages. The need to adjust to multiple talkers and other sources of variability should be present in all languages, so in principle we should expect this process to be universal. However, properties of specific languages could affect how listeners adjust to variable language input. Particular properties of Mandarin that could affect the process of lexically guided perceptual learning include not only its status as a tonal language but also its phonotactics.

Tone could influence perceptual learning for segments by affecting the timepoint at which adaptation occurs. Previous work has suggested that listeners do not process tonal information as quickly or accurately as phonemic information; for instance, Cantonese listeners more often accepted nonwords as real words when they differed from real words only in tone, and were also slower and less accurate in discriminating syllables that differed only in tone, as compared to those that differed in phonemes [14]. Similar results have also been observed for Mandarin Chinese [15]. Thus tonal distinctions may be processed more slowly than segmental distinctions.

As in any language with lexical tone, Mandarin contains many minimal pairs distinguished by tone alone. Because of this, tone may influence whether listeners can access lexical contexts that bias their interpretation of an ambiguous sound. For example, replacing [f] in /fu3/ (府 'official residence, mansion') with [s] will make the non-word /su3/, but /su/ is a word in three other Mandarin tones (e.g., /su1/ 酥 'shortbread'; /su2/ 俗 'vulgar'; /su4/ 素 'vegetarian food'). If tonal information is not processed in the same timeframe as phonemic information, the lexically guided perceptual learning process could be delayed.

http://dx.doi.org/10.21437/Interspeech.2017-618

Mandarin phonotactics could affect perceptual learning by the constraints it places on segments. Previous perceptual learning studies [4-6] have most often placed the ambiguous sound in word-final position. At the end of a word, possible interpretations are highly constrained, meaning that listeners will often already know what phoneme to expect. For example, after hearing *gira-*, giraffe can be identified as the only possible lexical interpretation, even before the onset of the final phoneme. Mandarin, however, does not allow most consonants (including fricatives) to occur in the coda position. This means that in order to investigate perceptual learning of fricatives in Mandarin, the critical sounds must occur earlier in the word where possible interpretations are less constrained and more lexical competition is expected. Additionally, Mandarin's highly constrained syllable structure results in large numbers of homophones and dense lexical neighborhoods, which could have effects on the lexical competition process and the interpretation of ambiguous phonemes, effectively reducing the likelihood of rapid retuning such as is possible in the languages tested to date.

In this study we evaluate lexically guided perceptual retuning for the fricatives [f] and [s] in Mandarin Chinese. Learning is induced by a training phase in which participants hear an ambiguous mixture of [f] and [s] in lexical contexts favoring interpretation as either [f] or [s]. Retuning is tested at the phoneme level by a [f]-[s] categorization task as in [4] (Experiment 1) and at the word level by a cross-modal priming task with minimal pairs [12] (Experiment 2).

2. Experiment 1

2.1. Methods

2.1.1. Participants

16 participants took part in the pretest and 48 took part in the main experiment. All were students currently attending university in Hong Kong; all had been born in China and had acquired Mandarin Chinese as their primary home language. Some reported speaking Cantonese as a second language, but all were Mandarin-dominant. None reported any vision, hearing, or language impairments. Participants were paid for their participation.

2.1.2. Apparatus

The experiment was conducted in a quiet room, and all auditory stimuli were presented over headphones at a comfortable listening level. E-prime was used for the presentation of materials and recording of responses.

2.1.3. Pretest

A pretest was conducted in order to select an ambiguous fricative and a test continuum for the main experiment.

A female native speaker of Mandarin produced the syllables /fu/, /su/, and /θu/ with a high level tone (Mandarin tone 1). The fricative portions of the /fu/ and /su/ recordings were extracted and used to create a continuum following the procedure used in Norris et al. [4]. The [f] and [s] waveforms were mixed at different proportions that were equally spaced along a 41-step continuum so that the token at one end of the continuum was 100% [f] and 0% [s] and that at the other end was 0% [f] and 100% [s]. Each token was concatenated with a /u/ token that was excised from the same speaker's production of /θu/ (to avoid coarticulatory cues in vowels biasing listeners towards interpreting the ambiguous sounds as either [f] or [s]).

Fourteen steps were chosen from this [f]-[s] continuum as stimuli for the pretest. These were the two endpoints and 12 steps from the most ambiguous part of the continuum: 1 ([f]), 7, 9, 11, 13, 15, 17, 19, 21, 23, 25, 27, 29 and 41 ([s]).

The pretest consisted of 10 blocks, each containing the 14 steps in random order. Participants were instructed to indicate whether the stimulus was the word "夫" (/fu1/ 'husband'; an [f] response) or "苏" (/su1/ 'Su' [a surname]; an [s] response) by pressing "F" or "S" on a computer keyboard. A short break followed the fifth block of stimuli.

2.1.4. Pretest results

Figure 1 shows the proportion of [f] responses for each step in the [f]-[s] continuum in the pretest. Step 17 was the most ambiguous step, with 55% [f] responses. For this reason, it was selected as the source for the ambiguous fricative used in constructing the ambiguous stimuli in the main experiment. This allowed us to minimize bias toward either end of the continuum in the materials.

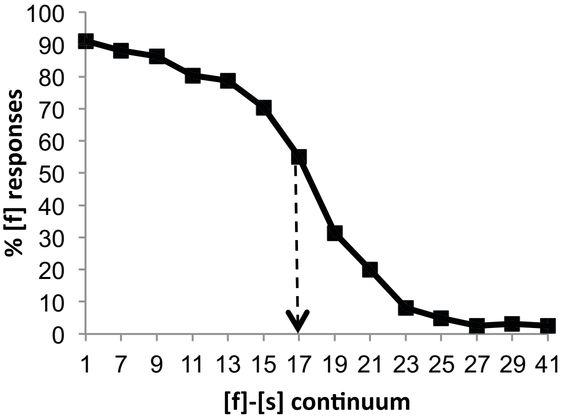

Figure 1: *A plot of the total proportion of [f] responses for each step in the pretest, averaged across participants*

2.1.5. Training materials

Training materials consisted of 100 Mandarin words and 100 non-words. All words and nonwords were disyllabic. Of the 100 words, 60 were filler items and 40 were training items. Of the 40 training items 20 contained [f] as the first phoneme of the second syllable (f-words) and 20 contained [s] as the first phoneme of the second syllable (s-words). These words were selected such that substituting [s] for [f] in the f-words would result in a nonword and vice-versa. These two sets of words were matched for tones and word frequency. The mean frequency was 3.68 per million for the f-words and 3.82 per million words for the s-words, as calculated from [16].

For each of the training items, two versions were selected for presentation: an unaltered version and a version in which the critical word-medial fricative was replaced by an ambiguous mixture of [f] and [s] extracted from step 17 of the /fu/-/su/ continuum. The original recordings were cut at zero-crossings near the onset and offset of frication and replaced by the ambiguous sound which will be henceforth referred to as [?]. This editing was carried out in Praat [17].

We constructed 100 nonwords by changing the tone of the second syllable in a real word (e.g., /ji1-dan4/ 鸡蛋 'egg' became /ji1-dan2/). No nonwords or filler words contained [f], [s], or any of the phonemes [ʂ], [ɕ], [ts], or [ts^h] (this, following the procedure of earlier studies, was in order to avoid any sounds too perceptually similar to the critical fricatives). All training materials were recorded by the same speaker who produced the recordings for the [f]-[s] continuum.

2.1.6. *Training phase: lexical decision task*

Participants were instructed to decide whether the item in each trial was a real word or a non-word in Mandarin and indicate their response with a button press. "Yes" responses were made with the dominant hand.

Four stimulus lists were constructed for this task. Each list contained the same 100 words and 100 non-words. The items were arranged in two different pseudo-random orders, such that no more than four words or four non-words were presented successively. For each of these two orders, two versions were created, one in which all instances of [f] were replaced with [?] and one in which all instances of [s] were replaced with [?]. The initial 12 trials, that included both words and non-words, were the same across all four lists and did not contain the ambiguous fricative. Half of the participants (the f-trained group) heard a list in which [f] had been replaced by [?] while the other half (the s-trained group) heard a list in which [s] was replaced by [?]. Both groups heard the same set of filler items and non-words.

2.1.7. *Test phase: categorization task*

The test phase consisted of a phonetic categorization task. Listeners were presented with recordings of steps 6, 12, 16, 20, and 26 of the /fu/-/su/ continuum and asked to categorize each item as either "夫" (/fu1/ 'husband') or "苏" (/su1/ 'Su [a surname]'). These five recordings were each heard 30 times (giving 150 trials per participant) in random order.

2.2. Results and Discussion

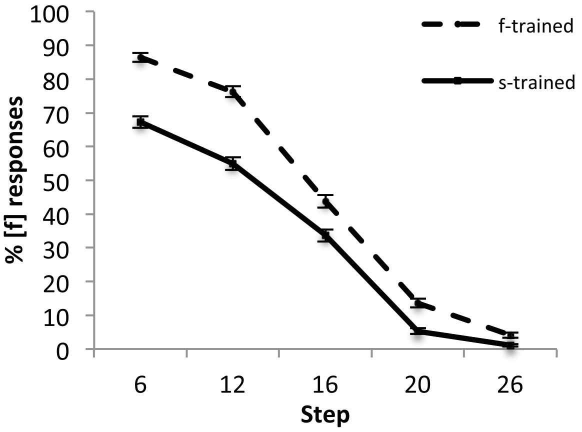

Figure 2: *A plot of the total proportion of [f] responses for the two training groups in the categorization task. Error bars represent standard errors.*

Figure 2 shows the proportion of [f] responses for the f-trained and s-trained groups at each step of the /fu/-/su/ continuum. Results were subjected to an ANOVA with training as a between-subjects factor and step as a within-subjects factor. Because only one version of each step on the continuum was presented, a by-items analysis was not possible.

There was a significant main effect of training condition, $F(1, 46) = 9.438$, $p < .01$, $\eta^2_p = .170$. This indicates that, as predicted, the f-trained group categorized the ambiguous sounds as [f] significantly more often than the s-trained group. This result suggests that perceptual learning did take place, as participants tended to identify the ambiguous fricatives in a manner consistent with their training.

There was also a significant main effect of step, $F(4, 43) = 71.715$, $p < .001$, $\eta^2_p = .870$. This indicates that participants categorized some sounds as [f] significantly more often than others. From Figure 2, we can see that the proportion of [f] responses decreases from step 6 (the most [f]-like of the step) to step 26 (the most [s]-like) of the steps.

Although the difference between the f-trained and s-trained groups appears to decrease from step 6 to step 26, the interaction of step and training group did not reach significance, $F(4, 43) = 1.827$, $p = .141$, $\eta^2_p = .145$.

3. Experiment 2

3.1. Methods

3.1.1. *Participants and Apparatus*

A subset of 23 participants from Experiment 1 also took part in Experiment 2. The apparatus was as in Experiment 1. The procedure was based on that of [12].

3.1.2. *Materials*

Testing materials consisted of 80 Mandarin words and 80 nonwords with legal Mandarin phonotactics. The critical test items consisted of 20 minimal pairs (40 words) that differed in whether they contained [f] or [s]/ (e.g. 家父 (jia1fu4)/加速 (jia1su4), father/accelerate). For each minimal pair, the speaker recorded a version with [θ] in place of [f] or /s/ (e.g. jia1θu4 for jia1fu4/jia1su4). The /θ/ was excised from each of these recordings and replaced with [?]. Again, this was done to avoid coarticulatory cues in the vowel influencing listeners' interpretation of the ambiguous sound. Of the remaining 40 words, 10 contained /f/, 10 contained /s/ and 20 contained neither /s/ nor /f/.

3.1.3. *Cross-modal priming task*

Participants heard auditory stimuli, then saw written stimuli in Simplified Chinese characters; they were instructed to indicate with a button press whether the written stimuli were real Mandarin words. Response time was measured from the onset of the written stimulus. In critical trials, visually presented members of [f]-[s] minimal pairs (e.g. 家父 (jia1fu4)/加速 (jia1su4), father/accelerate) followed either an unrelated prime (unrelated condition) or an ambiguous prime with [?] in place of the fricative (related condition). For example, 家父 (jia1fu4) would follow *jia1?u4* in the related condition and *wu4li3* (physics) in the unrelated condition.

Cross-modal identity priming refers to the well-known finding (e.g., [18,12]) that listeners are able to identify a written word more quickly if it is immediately preceded by an auditory version of the same word than by an unrelated word. If participants have learned to identify [?] as either [f] or [s], we therefore expect that cross-modal identity priming will occur for items containing [?] in place of the phoneme consistent with their training. Thus, for f-trained participants, we would expect a faster response for *jia1?u4* ->家父 (jia1fu4) than for *wu4li3* ->家父 (jia1fu4). However, we would not expect such an effect for *jia1?u4* ->加速 (jia1su4) compared to *wu4li3* ->加速 (jia1su4).

3.2. Results

Response times were analyzed with an ANOVA with training as a between-subjects factor and target type (f or s) and prime type (related or unrelated) as within-subjects factors. The main effect of prime type was significant in both the by-subjects ($F(1, 21) = 8.792$, $p < .01$, $\eta^2_p = .295$) and by-items ($F(1, 38) = 6.358$, $p < .05$, $\eta^2_p = .143$) analyses. This indicates that in trials with the related primes, containing the ambiguous fricative, responses were faster than in the trials with unrelated primes. The by-items analysis also revealed a main effect of training group ($F(1, 38) = 10.960$, $p < .01$, $\eta^2_p = .224$), indicating faster overall reaction times for the f-trained group.

Additionally, a three-way interaction of target type, training type, and prime type was marginally significant in both the by-items ($F(1, 38) = 3.216$, $p =.081$, $\eta^2_p = .078$) and by-subjects analyses ($F(1, 21) = 3.715$, $p =.068$, $\eta^2_p = .150$). This interaction was due to reaction times in the related versus the unrelated condition differing to a greater extent for prime-target pairs consistent with a listener's training (e.g. a greater priming effect for f-words in f-trained subjects; thus, as predicted, f-trained participants were more likely to interpret [?] as [f] in the minimal pair words while s-trained listeners were more likely to interpret the same sound as [s]). However, the pattern is more pronounced for the f-trained than the s-trained listeners. While such marginally significant results should be interpreted with caution, stronger effects with [f] than with [s] have also been observed in perceptual learning studies in other languages (e.g., [4, 6]).

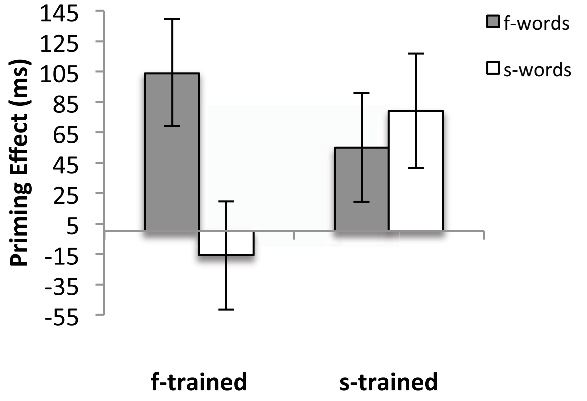

Figure 3: *Priming effect, measured as the difference in reaction time between the related and unrelated conditions, for f-words and s-words in each training group. Error bars represent standard errors.*

4. General Discussion

Mandarin listeners showed retuning of phonemic category boundaries for [f] and [s] after a brief training, comparable to that received by listeners in previous experiments with English and Dutch. The results we have observed were also in line with those of the predecessor studies. Overall, the results therefore demonstrate that lexically guided perceptual learning for adaptation to newly encountered talkers occurs in Mandarin Chinese, and moreover occurs under similar conditions and on a similar time scale to that shown in experiments in European languages.

The experiments extend the perceptual learning literature in more ways than by adding a language that was previously unexamined in this respect. The phonotactic requirements of Mandarin rendered word-medial ambiguous phonemes a necessary location for the critical phonetic ambiguity, which in turn has demonstrated that the lexical knowledge allowing interpretation of that ambiguity need not consist of the whole of the rest of the lexical item containing it.

Furthermore, we have demonstrated perceptual learning both with an explicit categorization task at test (as used in [4] and many later studies) and with a priming task (as first used in [12]). The latter finding shows that the retuning generalizes to untrained lexical items, in which the critical sound occurs in phonetic contexts not available in training, thus ruling out any possible exemplar-based account of the learning effect [19].

5. Conclusion

Previous work has extensively documented lexically guided perceptual learning for phonemes in English and Dutch [4-12]. Similar learning has also been observed for lexical tone in Mandarin [13] and even for written letters in English [20]. Taken together with this previous work, our findings are consistent with the view that perceptual retuning based on lexical knowledge is the application in speech of a powerful and general cognitive strategy for adaptation to variable input.

Listeners of any language encounter multiple talkers and need to adjust to the resulting variability in speech. Our findings show that lexically guided perceptual learning is useful for this, in languages with very different lexical and phonological structures. Consistent with findings that segmental structure is processed faster than tonal structure in Chinese word recognition [14,15], the results suggest that adaptation of segmental categories has a high priority in Chinese listeners' adaptation to new talkers. As suggested in [20], recourse to existing knowledge to resolve temporary ambiguity and hence retune perceptual categories, as a general cognitive strategy, is likely to be language-universal.

6. Acknowledgements

This research was supported by Australian Research Council Discovery Grant DP140104389. We would like to thank Mona Faris and Kay Hoi Yi Wong for assisting in data collection.

7. References

[1] A. Bradlow and T. Bent, "Perceptual adaptation to non-native speech", *Cognition*, vol. 106, no. 2, pp. 707-729, 2008.

[2] E. Dupoux and K. Green, "Perceptual adjustment to highly compressed speech: Effects of talker and rate changes.", *Journal of Experimental Psychology: Human Perception and Performance*, vol. 23, no. 3, pp. 914-927, 1997.

[3] M. Davis, I. Johnsrude, A. Hervais-Adelman, K. Taylor and C. McGettigan, "Lexical information drives perceptual learning of distorted speech: Evidence from the comprehension of noise-vocoded sentences.", *Journal of Experimental Psychology: General*, vol. 134, no. 2, pp. 222-241, 2005.

[4] D. Norris, J. McQueen and A. Cutler "Perceptual learning in speech", *Cognitive Psychology*, vol. 47, no. 2, pp. 204-238, 2003.

[5] A. Samuel and T. Kraljic, "Perceptual learning for speech", *Attention, Perception, & Psychophysics*, vol. 71, no. 6, pp. 1207-1218, 2009.

[6] M. Sjerps and J. McQueen, "The bounds on flexibility in speech perception.", *Journal of Experimental Psychology: Human Perception and Performance*, vol. 36, no. 1, pp. 195-211, 2010.

[7] T. Kraljic and A. Samuel, "Perceptual learning for speech: Is there a return to normal?", *Cognitive Psychology*, vol. 51, no. 2, pp. 141-178, 2005.

[8] F. Eisner and J. McQueen, "Perceptual learning in speech: Stability over time", *The Journal of the Acoustical Society of America*, vol. 119, no. 4, pp. 1950-1953, 2006.

[9] J. McQueen, D. Norris and A. Cutler, "The dynamic nature of speech perception", *Language and Speech*, vol. 49, no. 1, pp. 101-112, 2006.

[10] T. Kraljic and A. Samuel, "Generalization in perceptual learning for speech", *Psychonomic Bulletin & Review*, vol. 13, no. 2, pp. 262-268, 2006.

[11] J. Maye, R. Aslin and M. Tanenhaus, "The weckud wetch of the wast: Lexical adaptation to a novel accent", *Cognitive Science: A Multidisciplinary Journal*, vol. 32, no. 3, pp. 543-562, 2008.

[12] J. McQueen, A. Cutler and D. Norris, "Phonological abstraction in the mental lexicon", *Cognitive Science*, vol. 30, no. 6, pp. 1113-1126, 2006.

[13] H. Mitterer, Y. Chen and X. Zhou, "Phonological abstraction in processing lexical-tone variation: evidence from a learning paradigm", *Cognitive Science*, vol. 35, no. 1, pp. 184-197, 2010.

[14] A. Cutler and H. Chen, "Lexical tone in Cantonese spoken-word processing", *Perception & Psychophysics*, vol. 59, no. 2, pp. 165-179, 1997.

[15] Y. Ye and C. Connine, "Processing spoken Chinese: The role of tone information", *Language and Cognitive Processes*, vol. 14, no. 5-6, pp. 609-630, 1999.

[16] Center for Chinese Linguistics PKU. (2017, March 19). *CCL语料库检索系统（网络版)* [Online]. Available: http://ccl.pku.edu.cn:8080/ccl_corpus/

[17] P. Boersma and D. Weenink. (2017, March 19) *Praat: doing phonetics by computer* (Version 6.0.27) [Computer program]. Available: http://www.praat.org/

[18] P. Zwitserlood, "Form priming", *Language and Cognitive Processes*, vol. 11, no. 6, pp. 589-596, 1996.

[19] A. Cutler, F. Eisner, J. McQueen and D. Norris, “How abstract phonemic categories are necessary for coping with speaker-related variation”, in *Papers in Laboratory Phonology*, vol. 10, C. Fougeron, B. Kühnert, M.P. d'Imperio and N. Vallée, Eds. Berlin: Mouton de Gruyter, 2010, pp. 91-111.

[20] D. Norris, S. Butterfield, J. McQueen and A. Cutler, "Lexically guided retuning of letter perception", *The Quarterly Journal of Experimental Psychology*, vol. 59, no. 9, pp. 1505-1515, 2006.