Language, Cognition and Neuroscience, 2015
Vol. 30, No. 5, 491–505, http://dx.doi.org/10.1080/23273798.2014.942326

Routledge
Taylor & Francis Group

Multi-level processing of phonetic variants in speech production and visual word processing: evidence from Mandarin lexical tones

Jessie S. Nixon[a,b]*, Yiya Chen[a,b] and Niels O. Schiller[a,b]

[a]*Leiden Institute for Brain and Cognition (LIBC), Leiden University, 2300 RB Leiden, The Netherlands;* [b]*Leiden University Centre for Linguistics (LUCL), Leiden University, 2300 RB Leiden, The Netherlands*

(*Received 9 August 2013; accepted 19 June 2014*)

Two picture–word interference experiments provide new evidence on the nature of phonological processing in speech production and visual word processing. In both experiments, responses were significantly faster either when distractor and target matched in tone category, but had different overt realisations (toneme condition) or when target and distractor matched in overt realisation, but mismatched in tone category (contour condition). Tone 3 sandhi is an allophone of Beijing Mandarin Tone 3 (T3). Its contour is similar to another tone, Tone 2. In Experiment 1, sandhi picture naming was faster with contour (Tone 2) and toneme (low Tone 3) distractors, compared to control distractors. This indicates both category and context-specific representations are activated in sandhi word production. In Experiment 2, both contour (Tone 2) and toneme (low Tone 3) picture naming was facilitated by visually presented sandhi distractors, compared to controls, evidence that category and context-specific instantiated representations are automatically activated during processing of visually presented words. Combined, the results point to multi-level processing of phonology, whether words are overtly produced or processed visually. Interestingly, there were differences in the time course of effects.

Keywords: speech production; Mandarin Chinese; lexical tone; phonetic variation; sub-phonemic detail; phonological processing; picture-word interference

How are the sounds of language stored in memory and accessed during language production? Early accounts assumed phonology to be processed in terms of (optimally) functional units that distinguish between lexical items: phonemes. Phonemes were conceptualised as abstract, idealised representations of sound (Foss & Swinney, 1973; Meyer, 1990, 1991; Roelofs, 1999). In most experiments investigating phonology, phonological relatedness is measured in terms of phoneme overlap. In addition, some of the most influential models of language production (Dell, 1986, 1988; Indefrey & Levelt, 2004; Levelt, 2001; Levelt, Roelofs, & Meyer, 1999) posit lexical access to involve activation of sequences of phonemes.

Phonemes (*e.g.* /t/ or /k/) are the smallest units of sound that distinguish between words in a particular language (*e.g.* ‘top’ vs. ‘cop’ in English). In contrast, allophones vary with phonetic context, but do not affect word meaning. For example, word-initially, English /t/ is aspirated (has a puff of air, *e.g.* ‘top’), but is unaspirated (no puff of air) following /s/ (*e.g.* ‘stop’). Experimental evidence suggests that phoneme-like generalisation plays a role in *online* speech processing. For instance, in a perceptual learning experiment, McQueen, Cutler, and Norris (2006) had Dutch participants perform a training phase of auditory lexical decisions to words in which either the final /f/ or the final /s/ was replaced by an ambiguous (f-s) fricative sound. These words created a lexical bias to interpret the ambiguous sound as a particular phoneme. For example, participants in the ambiguous /f/ condition heard (witlɔ?), where *witlof* is a real Dutch word, but *witlos* is not, thereby creating a bias to interpret the ambiguous sound as an /f/. In the following test phase, participants made lexical decisions to visually presented minimal pair words (*e.g. doof* ‘deaf’; *doos* ‘box’) preceded by auditory primes containing the ambiguous sound (*e.g.* doo?). Facilitation depended on which ambiguous phoneme participants were trained with. Participants who heard the ambiguous sound in /f/-words during training were faster to identify visually presented /f/-words (*e.g. doof*), whereas participants who heard ambiguous /s/ were faster to name /s/-words (*e.g. doos*). Participants had adjusted (‘re-tuned’) their perceptual categories by matching the distorted sound to lexical items stored in memory. Importantly, since different sets of words were used in training and test, re-tuning was not restricted to specific words, but instead must have generalised to elements common to both training and test words; that is, to phoneme categories.

Similarly, McLennan, Luce, and Charles-Luce (2003) found evidence for category-level processing in production. In American English, word-medial /d/ and /t/ are

*Corresponding author. Email: jess.s.nixon@gmail.com often produced as a flap, making the two sounds ambiguous. In a repetition priming experiment, McLennan et al. (2003) had participants produce words containing /d/ and /t/, preceded by auditory primes that were either carefully articulated or flapped. Results showed that flapped and carefully produced forms primed each other, evidence for processing at the category level.

On the other hand, there is mounting evidence that processing of phonetic information goes far beyond distinguishing phonemes (Clayards, Tanenhaus, Aslin, & Jacobs, 2008; Ju & Luce, 2006; McMurray, Tanenhaus, & Aslin, 2009b; Mitterer, Chen, & Zhou, 2011; Newman, Clouse, & Burnham, 2001; Trude & Brown-Schmidt, 2012). Exquisite perception and memory for detail have also been shown in auditory (Agus, Thorpe, & Pressnitzer, 2010) and visual processing (Brady, Konkle, Alvarez, & Oliva, 2008). At the extreme, it has been proposed that lexical processing can be explained without any sub-lexical categories. For example, the memory model MINERVA 2 (Hintzman, 1986) takes phonological representations to be built up from episodic memory traces of whole lexical items. Goldinger (1998) found that MINERVA 2 correctly predicted both reaction times and speakers' spontaneous mimicking of voice onset time in perceived speech, which cannot be explained by purely abstractionist models. This has been taken as evidence that there are no abstract categories below the word level (but cf. Fowler, 2010; Mitterer, 2006).

Taken together, the above findings suggest that speech processing involves both phonemic and sub-phonemic representations. This conclusion is further supported by recent evidence for both abstraction and detailed information obtained within the same experiment (Mitterer et al., 2011; Nielsen, 2011). For instance, Mitterer et al. (2011) tested the extent to which abstract and detailed acoustic information influence perceptual learning of tones in Mandarin. Analogous to McQueen et al.'s (2006) perceptual learning study, listeners heard ambiguous tonal contours (a synthesised continuum between Tones 1 and 2) in phrases that biased interpretation to either Tone 1 or Tone 2. Results showed that participants who received the ambiguous contours in contexts that biased interpretation to Tone 1 in the exposure phase were more likely to perceive ambiguous tones as Tone 1 during test than those who received the ambiguous Tone 2 context. This generalised to words not in the exposure phase, suggesting sub-lexical abstraction of tone category. There was also a specific-word effect: perceptual learning was greater for exposure-phase words than new words, evidence that detailed acoustic information was retained in representations of individual words. In this paper, we extend the investigation of specificity in lexical prosodic representations to the realm of speech production. In addition, the study makes an important distinction between the level and the nature of phonological representations.

The present study

The aim of the present study is to determine two interrelated aspects of lexical tone processing in Beijing Mandarin. The first concerns whether the level of processing corresponds to the tone category or a context-specific sub-phonemic level. The second question examines whether the nature of the representations is purely abstract or involves an internal instantiation of an actual sound, that is, the tonal contour. In addition, Experiment 2 investigates whether sub-phonemic processing occurs with visually presented words.

Beijing Mandarin has four lexical tones (*tonemes)* represented schematically in Figure 1. Characters that have the same *segmental syllable* (sequence of phonemes) can be distinguished by this inherent pitch contour, such as bi2[1] (鼻, 'nose') versus bi3 (笔, 'pen'). In connected speech, Tone 3 (T3) has at least two variants (*allotones*), shown in Figure 2. The canonical realisation is the low contour,[2] but preceding another T3 syllable, T3 is realised with a rising contour. This allophonic variant of T3 is known as *third tone sandhi* (hereinafter, 'T3 sandhi'). Tone sandhi refers to the phenomenon whereby the acoustic realisation of a tone is influenced by a neighbouring tone in a particular environment. Importantly for the present study, the contour of T3 sandhi is very similar to another tone, Tone 2. Figure 3 shows the tonal contours for Tone 2, T3 sandhi and the canonical, low Tone 3. Detailed acoustic analyses have been able to detect subtle differences between the pitch contours of Tone 2 and the T3 sandhi (Yuan & Chen, 2014). However, listeners generally cannot consciously distinguish between them (Peng, 2000; Wang & Li, 1967). Peng (2000) had native Mandarin speakers produce minimal pairs of bisyllabic words that were sequences of either Tone 3 + Tone 3 (sandhi) or Tone 2 + Tone 3. Although subtle acoustic differences were detected, in a following identification task, a different group of native speakers performed only at chance level in distinguishing the two word types.

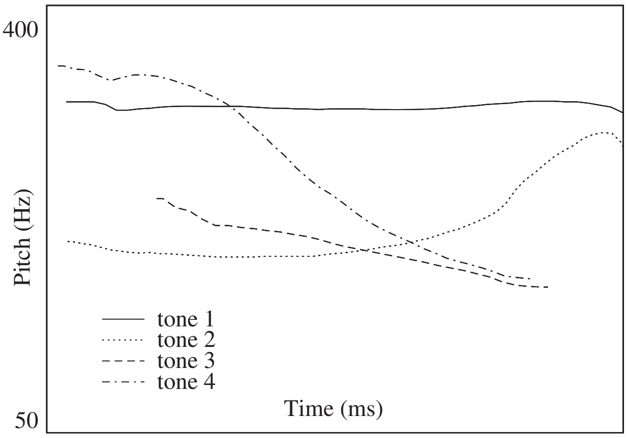

Figure 1. Pitch contours of the four tones of Beijing Mandarin.

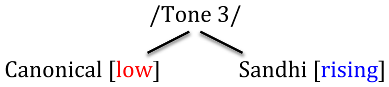

Figure 2. Schematic representation of the two variants of Beijing Mandarin Tone 3.

These aspects of the tonal system of Beijing Mandarin allowed us to manipulate two types of phonological relatedness: tone contour and tone category. In two picture–word interference (PWI) experiments (Damian & Martin, 1999; Lupker, 1982; Rosinski, Golinkoff, & Kukish, 1975; Schriefers, Meyer, & Levelt, 1990; Starreveld & La Heij, 1996), T3 sandhi and Tone 2 words share the same (rising) realisation contour, but belong to different tone categories (*contour* condition), or T3 sandhi and low Tone 3 words share the tone category (T3), but have different overt realisations (*toneme* condition). Experiment 1 investigates processing of speech category and sub-phonemic information during overtly produced T3 sandhi words. Facilitation in the toneme condition indicates processing at the tone category level; facilitation in the contour condition indicates sub-phonemic processing of tone. Experiment 2 tests whether these two levels of processing occur during visual processing of T3 sandhi words that are not overtly produced. In addition, both experiments used two stimulus onset asynchronies (SOAs) to investigate differences in the time course of processing between the contour and the tone category. Target picture and superimposed distractor word were presented either simultaneously (SOA = 0 ms) or with the distractor word delayed by 83 ms (SOA = 83 ms). Varying of SOAs has been used in this paradigm to investigate the time course of processing in speech production. Although the details of the stages and their time course are disputed, it is generally agreed across speech production models that producing speech involves access to at least two levels of information: a conceptual level and a form (phonological and/or orthographic) level. The distractor word (and the phonological, orthographic, or semantic information it contains) can be made available before, at the same time as or after the target picture is presented, with differential effects. For example, with visual presentation of distractor words, phonological facilitation from overlap of segmental phonemes has been found from 200 ms preceding up to 100 ms following target presentation, while semantic interference has been found only at simultaneous and positive SOAs, 0–200 ms (Damian & Martin, 1999). The decrease or disappearance of phonological facilitation between 100 ms and 200 ms presumably occurs because phonological processing has by 200 ms already reached a stage at which the speaker no longer benefits from the segmental overlap.

To date, very little is known about tone processing in speech production in general, or its time course in particular. However, Zhou and Zhuang (2000) found in a PWI experiment that tone processing is faster than segmental processing. While facilitation from segmental overlap occurred at both short and long SOAs, facilitation for tone was found only at the short SOA. We therefore selected a relatively short positive SOA in the present experiments to maximise the chances of obtaining facilitation effects.

If differences in the time course are found between the tonal category and the tonal contour, we see a number of possibilities for how this could manifest. Firstly, it is possible that each word is initially processed holistically so that each morpheme is processed in its context-specific form. This would then be followed by inductive activation of the context-general tone category. That is, in this scenario, the initial syllable of T3 sandhi words is processed as a rising tone first, followed by activation of the Tone 3 toneme. The second possibility is that activation begins at the general category level, followed by processing of the context-specific variant. If this is the case, we would expect an early effect in the toneme congruent conditions (i.e. at SOA = 0 ms), and late effects of the contour congruent conditions (at SOA = 83 ms). A third possibility is that both of these processes occur simultaneously, leading to simultaneous activation of both levels of processing. Finally, it is also possible that there are differences in the time course of activation of the two levels of processing, depending on the task. We might expect the actual contour of the context-specific variant to play a greater role in overt speech production than in silent processing of written words, while the reverse might be true for the tone category.

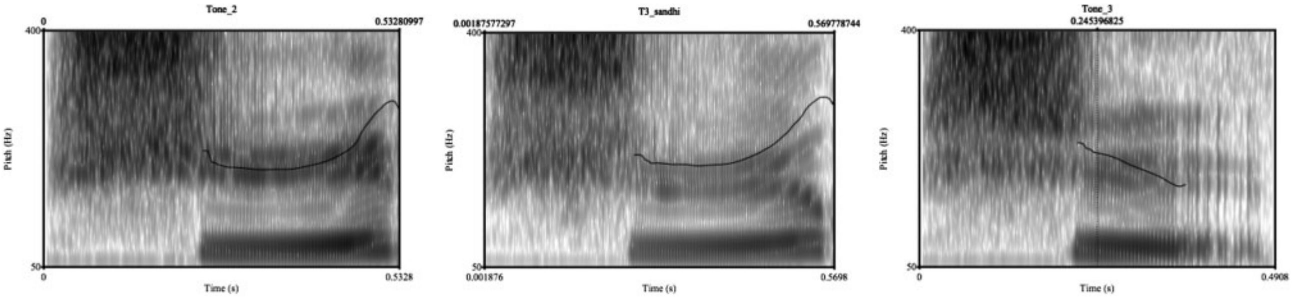

Figure 3. Pitch contours of Tone 2, Tone 3 sandhi and the canonical, low variant of Tone 3.

Experiment 1: Tone 3 sandhi picture naming with contour and toneme distractors

Method

In Experiment 1, participants named pictures of objects with T3 sandhi names. Recall that T3 sandhi words are made up of two Tone 3 characters. When both characters are Tone 3, the first character has a rising contour, instead of the canonical low contour. Crucially, this rising contour is similar to the contour of Tone 2. Therefore, Tone 2 distractors share the overt realisation – the rising contour – with T3 sandhi target pictures (contour condition). Low T3 distractors differ in overt contour realisation, but match in tone category (toneme condition).

If T3 sandhi picture naming is facilitated by toneme distractors, this demonstrates that there is activation at the tone category level, despite differences in the overt pronunciation. If contour distractors facilitate T3 sandhi picture naming, this indicates two things. Firstly, it is evidence for context-specific processing of the T3 sandhi allotone. Since, in most contexts, the contour of T3 is unrelated to T2, shorter latencies in the contour condition indicate a context-specific representation of the T3 sandhi allotone (rather than the general Tone 3 category). Secondly, even though T3 sandhi and T2 have similar realisations, if they are represented in a purely abstract form, they could still be processed as separate categories. Only through similarities in the actual pitch contour can facilitation from contour distractors occur. This suggests activation of an *instantiated* representation of the tonal contour.

Participants

Thirty native speakers of Beijing Mandarin (24 females; mean age: 21.5), students at universities within Haidian district in Beijing, were paid for their participation. All participants and their parents were born and raised in Beijing, except three participants who had one parent from the nearby Northern Mandarin-speaking province of Hebei, two participants for whom both parents were from Hebei, and one participant whose parents were from Shanghai.

Stimuli

The experimental conditions and sample stimuli are shown in Table 1. A complete list of stimuli for Experiment 1 is provided in Appendix 1. Critical targets were 27 pictures with two-character T3 sandhi names. Pictures were black-on-white line drawings selected from the MPI (10 pictures, two with modifications) and the Alario and Ferrand (1999) picture databases (three pictures), supplemented with pictures from the Internet (14 pictures). Distractors were contour (T2 characters), toneme (T3 characters) and control (T1 or T4 characters) one-character words with the same segmental syllable as the target initial syllable. Contour, toneme and control distractors were matched for word frequency and stroke number. Targets and distractors were semantically and orthographically unrelated. An additional 27 picture–distractor pairs were used as fillers to add variety and make the design less obvious to participants. None of the characters or initial syllables used in critical trials appeared in filler trials. Word and character frequencies were obtained from Subtlex-CH, a large (46.8 million characters, 33.5 million words) Chinese database based on film subtitles (Cai & Brysbaert, 2010).

Before going on to the experiment design, we make a brief note about the notion of 'word' in Chinese. The distinction between words and phrases is less clear-cut in Chinese than it is in alphabetic languages. Although lexicality could be said to be a gradient property in any language, for alphabetic language speakers, intuitions about what constitutes a word may be so deep-seated that we do not usually define it. Generally speaking, word boundaries are indicated by white spaces in the script. In Chinese script, spaces are instead inserted between characters. Characters correspond not to words, but to single syllables and (almost always) single morphemes. A word can consist of one or more characters. However, native Chinese speakers do not always agree on what constitutes a word versus a phrase. Therefore, in this paper, we have used bigram frequencies as a measure of lexicality. Sandhi stimuli had medium-to-high bigram frequencies, so they are expected to be processed more like 'words' than multi-word phrases. In Experiment 1, mean bigram frequency (measured in mutual information; MI) was 6.3 (SD = 3.7). MI is a measure of how likely two characters are to co-occur (see Da, 2004, for an explanation of the calculation method).

With medium-to-high bigram frequencies, one might expect the surface level to play a greater role. However, as described above, McLennan et al. (2003) failed to find evidence for surface-level processing in within-word

Table 1. Experiment design and sample stimuli for Experiment 1.

		Distractor conditions		
	Target pictures	Toneme	Contour	Control
Tone category	Tone 3 + Tone 3	Tone 3	Tone 2	Tone 1/4
Tonal contour	Rising	Low (dipping)	Rising	Other (high or falling)
Example	fu3dao3 辅导	fu3 斧	fu2 服	fu4 付

American English flap production. This has yet to be investigated in tone processing.

Design

Experiment 1 consisted of 324 trials, divided into six blocks of 54 trials, with breaks between the blocks. The experiment followed a 3 × 2 within-participant factorial design, with the factors distractor type (contour, toneme and control) and SOA (0 or 83). At SOA = 0 ms, the target picture and distractor word appeared simultaneously, while at SOA = 83 ms the distractor word was presented 83 ms after target picture onset. A relatively short delay was selected for the positive SOA because tonal effects have been found to be short-lived relative to segmental effects (Zhou & Zhuang, 2000). The SOA of 83 ms was calculated to match the screen refresh rate (60 Hz). There were 27 trials per condition. Three distractor word lists were constructed for each SOA, with distractor words divided equally between conditions. Each target word was presented six times (once in each distractor condition for each SOA). The script was programmed to counterbalance the order of presentation of the distractor word lists across participants. All lists were pseudo-randomised for each participant. Each block was preceded by three warm-up trials, which were excluded from analysis.

Procedure

Participants were tested individually in a quiet room at the Psychology Institute of the Chinese Academy of Sciences in Beijing. Stimulus presentation and data acquisition were conducted using the E-Prime 2.0 software package with the addition of a voice key. After being familiarised with target pictures and picture names, participants were seated approximately 60 cm from a 17-inch cathode ray tube computer monitor and given a practice session prior to the actual experiment.

Each experimental trial began with a fixation cross for 500 ms, followed by the target picture for a maximum of 2000 ms, or until the participant responded. Distractor words appeared superimposed on target pictures either simultaneously (SOA = 0 ms) or 83 ms after picture onset (SOA = 83 ms). An inter-stimulus interval of 500 ms preceded the next trial. Participants were instructed to ignore the words and name the pictures as quickly and accurately as possible. The experimenter coded response accuracy during the experiment. Response time was calculated from the time of target picture presentation until the voice key was triggered by the participant response.

Results

Data were analysed using linear mixed effects (LME) modelling, using the lmer function of the lme4 package (Bates, Maechler, & Bolker, 2013; see also Baayen, 2008; Baayen, Davidson, & Bates, 2008) in R (R Core Team, 2013). Analysis was conducted on the 4667 data points remaining after stutters, errors, false starts (3%) and null responses (0.8%) were removed. Since error rates were low, no further analyses were conducted on the errors. Inspection of response latency distributions revealed a skewed distribution, which was normalised by logarithmic transformation. Mean response times per distractor condition and SOA are shown in Figure 4.

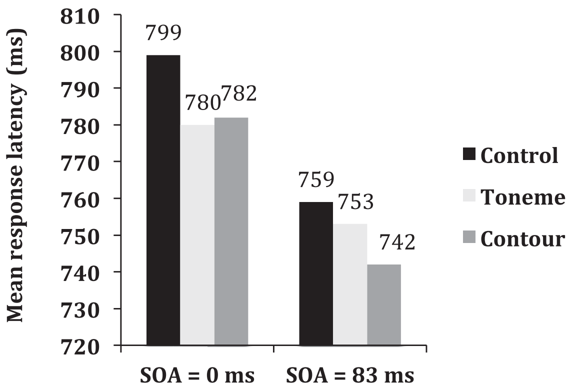

Figure 4. Mean reaction times (ms) per distractor type (toneme vs. contour vs. control) and SOA (0 ms vs. 83 ms) in Experiment 1.

There is currently debate in the literature concerning the appropriate method for constructing statistical models. Therefore, in this paper, both forward and backwards algorithms were used for model comparison (Baayen, 2008; Barr, Levy, Scheepers, & Tily, 2013). The two types of algorithm converged on the same final model. Firstly, a forward algorithm was conducted, which gradually built up complexity in the model. The baseline model was a regression line of log reaction times, with random intercepts for subjects and target pictures. Each fixed effect and interaction was individually added to the model and tested by comparing the log likelihood ratio to that of the simpler model. Trial was included as a control variable to investigate effects of learning or fatigue over the course of the experiment (Baayen, 2008). Only effects that significantly improved the fit were retained in the final model. Once the fixed effects were established, random effects structure was tested. A random slope was individually added and tested for each of the significant fixed effects. Only random slopes that improved model fit were retained.

It has been argued that random effects structure should be kept maximal, in that model testing should be conducted by starting first with maximal fixed effects, eliminating non-significant predictors, then entering maximal random slopes for all significant predictors of interest and eliminating only those that do not improve model fit (Barr et al., 2013). Therefore, in addition to the forward algorithm, a backwards elimination algorithm was also implemented here. After fixed effects were established, random intercepts and slopes for all significant fixed effects (trial, SOA and distractor type) were added to the model.

However, the maximum likelihood estimation of this model failed to reach convergence. This is a common problem with complex maximal models, particularly those with complex random effects structure (Barr et al., 2013). Barr and colleagues suggest that by removing correlation parameters, this problem can be solved while still meeting the objective of maximising the model. Therefore, a model was constructed containing separate by-subject random slopes for trial, SOA and distractor type without correlation parameters. With three random slopes, the likelihood estimation still failed to converge. With two random slopes, convergence was reached and model comparisons could be completed. No significant difference was found when trial or distractor random slopes were removed, but removing SOA significantly reduced log likelihood ratio. The backwards algorithm converged on the same final model as the forward algorithm.

The best-fit model (Table 2) included main effects of trial, distractor type and SOA, but no interactions, random intercepts for subjects and target pictures, and a by-subject random slope for SOA. Bigram frequency was tested, but did not improve the model as a main effect or interaction with other fixed effects, so was removed. In the model summary in Table 2, the control condition at SOA = 0 ms lies on the intercept (the baseline condition) and the estimates show the coefficients for each of the predictors. The trial coefficient indicates there was a small but significant increase in reaction times across participants over the course of the experiment. The main effect of SOA indicates that responses were faster when distractors were delayed (SOA = 83 ms) than with simultaneous presentation of stimuli (SOA = 0 ms). More importantly, naming latencies were significantly[3,4] shorter for both contour and toneme distractors, compared to controls. The effect appears to be slightly stronger in the contour condition than in the toneme condition.

Table 2. Results summary Experiment 1: coefficient estimates, standard errors (SE) and *t* values for all significant predictors in the log-transformed naming latencies for pictures with Tone 3 sandhi names.

Predictor	Coefficient estimate	SE	*t*
(Intercept)	6.6108	0.0282	234.75
Trial	0.0014	0.0002	8.15
Distractor type contour	−0.0207	0.0065	−3.20
Distractor type toneme	−0.0164	0.0065	−2.53
SOA 83	−0.0479	0.0191	−2.51

Table 3. Results summary Experiment 1 SOA 0: coefficient estimates, standard errors (SE) and *t* values for all significant predictors in the log-transformed naming latencies for Tone 3 sandhi pictures.

Predictor	Coefficient estimate	SE	*t*
(Intercept)	6.6099	0.0282	234.80
Trial	0.0015	0.0002	6.53
Distractor type contour	−0.0200	0.0089	−2.25
Distractor type toneme	−0.0232	0.0089	−2.61

Although the log likelihood ratio showed no significant improvement in model fit by adding an interaction between distractor type and SOA (p > .23), the mean reaction times (Figure 4) suggest differences in effects between the SOAs. Since our primary interest was to investigate the effects of different distractors on target picture naming, we split the dataset by SOA and ran separate models for each. The model summary for SOA = 0 ms (Table 3) shows that the predictors for the SOA = 0 ms model are similar to that of the full dataset. Model fit was improved by main effects of trial (p = 0) and distractor (p <.02), but not their interaction (p =.8). Random slopes did not improve the model. The model summary shows that for both contour and toneme distractors, response times are significantly faster than with the control distractor.

The model for the SOA = 83 ms dataset is shown in Table 4. There was a main effect of trial (p = 0), but distractor type only approached significance (p > .07). It may be that there was insufficient power in the experiment to yield a significant improvement in model fit for the three-level factor at the later SOA. However, since the predictor approached significance, we include it in the model here for comparison with SOA = 0 ms. The interaction with trial was not significant (p > .3), but there was a significant random slope for trial (p < .01). With delayed presentation of the distractor (SOA = 83 ms), the *t* values of this model suggest that while T3 sandhi naming seems to be faster in the contour condition than the control condition, there is no longer facilitation from toneme distractors.

Table 4. Results summary Experiment 1 SOA 83: coefficient estimates, standard errors (SE) and *t* values for all significant predictors in the log-transformed naming latencies for Tone 3 sandhi pictures.

Predictor	Coefficient estimate	SE	*t*
(Intercept)	6.5651	0.0275	238.66
Trial	0.0013	0.0004	3.69
Distractor type contour	−0.0233	0.0955	−2.44
Distractor type toneme	−0.0076	0.0952	−0.80

Discussion

The purpose of Experiment 1 was to investigate whether production of T3 sandhi words activates the Tone 3 toneme, the context-specific rising contour, or both. The results show that both levels of activation occur. Main effects of SOA and trial indicate that responses were faster when distractors were delayed and that there was a slight increase in reaction times over the course of the experiment. More importantly, there was also a main effect of distractor type, such that responses were faster when the distractor and target matched in contour, but mismatched in toneme (contour distractors), or when they matched in toneme, but mismatched in contour (toneme distractors), compared to unrelated controls.

Although there were no significant interactions, the numerical means (Figure 4) indicate a difference in facilitation effects between SOAs. This was investigated in a separate model for each SOA. Similar to the full dataset model, with simultaneous presentation of target picture and distractor word (SOA = 0 ms), faster naming latencies were found when distractors matched either the realisation (contour distractors) or the Tone 3 category (toneme distractors).

At the later SOA, including distractor type in the model resulted in only a marginal improvement of model fit. However, the model summary suggested faster naming with contour distractors, but not toneme distractors, compared to controls. This suggests that the congruent context-specific rising contour continues to facilitate production even with delayed presentation. This would suggest that while activation of the tone category may be fleeting, similarity in the actual acoustic-phonetic contour continues to facilitate production at later stages during overt production. Presumably, this reflects activation of an acoustic and/or articulatory target in preparation for speech.

In summary, the finding of both contour and toneme priming effects in Experiment 1 indicates activation of multiple levels of representation during T3 sandhi word production. This raises the question of whether the results are due to automatic, lexical processes or to articulation preparation. It is possible that lexical processing of T3 sandhi words involves only their abstract form, but that the context-specific instantiation is only generated for overt speech. If this is the case, visually presented T3 sandhi words that are not overtly produced should lead to activation of the Tone 3 category only. Experiment 2 addresses this question.

Experiment 2: naming of contour and toneme pictures with visually presented Tone 3 sandhi distractors

Method

Experiment 2 reversed the distractor and target conditions of Experiment 1, such that distractors were T3 sandhi words or controls and targets were pictures with contour (Tone 2 initial syllable) and toneme (Tone 3 initial syllable) bisyllabic names. If context-specific representations are activated only during speech preparation, then toneme targets, but not contour targets, should see facilitation from T3 sandhi compared to control distractors. If contour picture naming is quicker with T3 sandhi compared to control distractors, this suggests automatic activation of an instantiated representation of the context-specific T3 sandhi allotone, even when it is not overtly produced.

Participants

Thirty native Beijing Mandarin speakers (24 females; mean age: 22.7), students at universities within Haidian district of Beijing, were paid for their participation. None of them had participated in Experiment 1. All participants and their parents were born and raised in Beijing, except for four participants who had one parent from another Northern Mandarin-speaking province, and two participants whose parents were each from (different) Northern Mandarin-speaking provinces.

Stimuli

The experiment design and sample stimuli are shown in Table 5. Appendix 2 lists the stimuli used in Experiment 2. Targets were 48 bisyllabic pictures; 24 with initial Tone 2 syllable (contour condition) and 24 with initial Tone 3 names (toneme condition). Distractors were bisyllabic T3 sandhi or control (Tones 1 or 4) words that shared the same initial segmental syllable. T3 sandhi and control distractors were matched for word frequency, first character frequency, second character frequency, whole word stroke number, first character stroke number and second character stroke number. Mean bigram frequency of sandhi stimuli was 6.44 MI (SD = 3.7). There was no orthographic overlap or semantic relatedness between prime and target. An additional 48 picture–distractor pairs were used as filler trials.

Design

Experiment 2 consisted of a factorial 2 × 2 × 2 within-participants design. Experimental factors were target type (contour vs. toneme), distractor type (T3 sandhi versus control) and SOA (0 ms vs. 83 ms). The experiment consisted of 384 trials, divided into six blocks of 64 trials, with breaks between the blocks. Each target word was presented four times (once in each distractor condition for each SOA). Other aspects of the design were the same as Experiment 1.

Procedure

The procedure for Experiment 2 was identical to Experiment 1.

Table 5. Experiment design and sample stimuli for Experiment 2.

	Target conditions		Distractor conditions	
	Contour	Toneme	Sandhi	Control
Lexical tone	Tone 2 + Tone X	Tone 3 + Tone X	Tone 3 + Tone 3	Tone 1/4 + Tone X
Tonal contour	Rising	Low	Rising	Other (high or falling)
Example	bi2kong3	bi3ji4	bi3shou3 匕首	bi4zhi4 币值

Results

Analysis was conducted on the 5589 data points remaining after removal of stutters, errors, false starts (2%) and voice key errors (0.9%). Mean reaction times are shown in Figure 5. Forward and backwards algorithms were used to establish fixed and random effects structure, and converged on the same final model. In the backwards algorithm, the maximal model with random slopes for distractor type, SOA and trial reached convergence, but the log likelihood ratio revealed that the random slope for distractor type was not significant, so it was removed.

The summary of results for the LME model for Experiment 2 is shown in Table 6. The control distractor condition at SOA = 0 ms lies on the intercept. The model was improved by a main effect of trial, reflecting an overall increase in response time over the course of the experiment. Responses were faster with a delayed distractor (SOA = 83 ms) than with simultaneous presentation, replicating the findings in Experiment 1. More interestingly for the present study, the model reveals a main effect of distractor type, with log naming latencies significantly shorter for T3 sandhi distractors, compared to control distractors. No improvement of the model was found with either a main effect of target type ($p = .12$), nor its interaction with distractor type ($p = .28$). The main effect of bigram frequency was not significant, and there were no significant interactions, so it was not included in the model. With a random slope for SOA, the fixed effect for SOA was no longer significant, but since the random effect term was significant, the fixed effect was retained in the model.

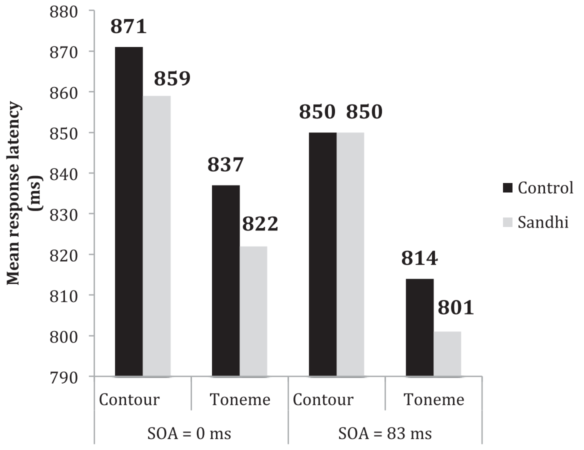

Figure 5. Mean reaction times (ms) per target type (contour vs. toneme), distractor type (sandhi vs. control) and SOA (0 ms vs. 83 ms) in Experiment 2.

The main effect of distractor type in absence of an interaction between distractor type and target type or SOA suggests that the T3 sandhi distractors facilitated both target types at both SOAs. However, as with Experiment 1, between-condition differences in the numerical means suggest differential effects for the two target types, particularly at the late SOA. In order to further investigate these numerical differences, we split the data and modelled each target type separately. The model summary for toneme targets is shown in Table 7. The model contained very similar predictors as in the combined dataset. Model fit was improved by main effects of trial ($p = 0$), SOA ($p = 0$) and distractor type ($p < .01$), but no interactions. With a random slope for SOA, the t value for the fixed effect was only marginally significant (Table 7), but since SOA significantly improved model fit and because the random slope term was significant, the fixed effect was retained in the model. The model confirms for toneme targets the findings of the full model, namely that presentation of T3 sandhi distractor words facilitates production of picture names with the same tone category.

The model summary for contour targets is shown in Table 8. The only significant fixed effects were trial ($p < .001$) and SOA ($p < .02$), which did not interact ($p = .42$). Distractor type did not significantly improve the model. Model fit was further improved by a random slope for SOA ($p = 0$).

Table 6. Results summary Experiment 2: coefficient estimates, standard errors (SE) and t values for all significant predictors in the log-transformed picture naming latencies with Tone 3 sandhi versus control distractors.

Predictor	Coefficient estimate	SE	t
(Intercept)	6.6951	0.0230	291.49
Distractor type sandhi	−0.0127	0.0048	−2.67
SOA 83	−0.0221	0.0151	−1.47
Trial	0.0006	0.0001	5.17

Table 7. Results summary Experiment 2 Toneme targets: coefficient estimates, standard errors (SE) and t-values for all significant predictors in the log-transformed picture naming latencies with Tone 3 sandhi versus control distractors.

Predictor	Coefficient estimate	SE	*t*
(Intercept)	6.6854	0.0286	233.99
Distractor Type Sandhi	−0.0224	0.0068	−3.29
SOA 83	−0.0293	0.0154	−1.90
Trial	0.0005	0.0001	4.41

Although the model shows no effect of distractor type for the contour target only data with both SOAs included, there was a substantial difference in mean response times between T3 sandhi and control distractors at SOA = 0 ms (Figure 5). Because neither the interaction with SOA nor the interaction with target type was significant, it is unclear whether this numerical difference is due to facilitation in processing from contour overlap or simply due to random variation. In order to further investigate this issue, we ran a Bayesian analysis on the SOA = 0 ms data. If it is found that both target types are facilitated by T3 sandhi, compared to controls (i.e. there is no interaction between distractor type and target type), this would provide evidence that the contour is activated in visual word processing. Each of the predictors included in the LME model were individually tested in the Bayesian model. Substantial support was found for including a predictor of trial, compared to the baseline intercept (BF = 10.5).[5] Adding a predictor for distractor type further improved the model (BF = 42.6), but neither target type (bf = .08) nor a target type:distractor type interaction (BF = .08) were supported. The absence of an interaction indicates that both target types were facilitated by T3 sandhi, compared to control distractors, providing further evidence for activation of the context-specific rising contour in visual processing of T3 sandhi words.

Discussion

Experiment 2 investigated whether the findings from Experiment 1, namely that overt production of speech variants involves multi-level phonological processing, can be extended to visual processing of written speech variants. There was a main effect of trial, reflecting an increase in reaction times over the course of the experiment. More importantly, there was a main effect of distractor type, indicating faster picture naming with T3 sandhi distractors, compared to control distractors. Although the interaction between distractor type and target type was not significant, mean reaction times (Figure 5) suggested differences in the amount of facilitation for toneme and contour targets. We therefore split the data and analysed each target type separately. A robust effect of distractor type remained for the toneme targets, indicating activation of the tone category during visual presentation of T3 sandhi words.

Table 8. Results summary Experiment 2 contour targets: coefficient estimates, standard errors (SE) and *t* values for all significant predictors in the log-transformed picture naming latencies with Tone 3 sandhi versus control distractors.

Predictor	Coefficient estimate	SE	*t*
(Intercept)	6.6073	0.0276	242.86
SOA 83	−0.0163	0.0160	−1.02
Trial	0.0006	0.0001	5.30

For the contour targets, distractor type did not improve model fit with both SOAs in the model. However, a number of factors pointed to a facilitatory effect of contour. Firstly, there was a substantial difference in mean response times at SOA = 0 ms (Figure 5). In addition, in the full model containing both target types, the effect of distractor type was significant, and there was no statistical support for an interaction with target type, suggesting that distractor type plays a role for both toneme and contour targets. Therefore, a Bayesian analysis was run in order to investigate whether the main effect of distractor type in the full model and the shorter reaction times at SOA = 0 ms could indeed be attributed to contour facilitation. The model showed a preference for including distractor type, but not target type or the interaction, confirming that the context-specific contour is activated during visual processing of T3 sandhi words.

General discussion

Two PWI experiments provide new evidence on the nature of the phonological processing during speech production and visual processing of words that are not overtly produced. In particular, they address the question of whether allophonic variation is processed at the higher level of the phonemic category or at the lower, sub-phonemic level of the context-specific variant. In Experiment 1, during overt production of T3 sandhi picture names, significantly shorter naming latencies were found when distractor and target picture matched in tone category, but had different overt realisations (toneme condition), and when target and distractor matched in overt realisation, but mismatched in tone category (contour condition). This demonstrates that production of allophonic variants of Mandarin tones involves multi-level phonological processing: both the tone category and the context-specific variant are activated. The time course of activation was further investigated by splitting the data by SOA. When target and distractor were presented simultaneously (SOA = 0 ms), there was facilitation from both contour and toneme distractors, compared to controls, indicating early activation of both the tone category and the context-specific contour. With delayed presentation of the distractor (SOA = 83 ms), only the contour distractor showed significant effects on overt production; the toneme distractor no longer facilitated naming latencies. This can be explained if the overt realisation contour remains activated for longer than the tone category. An alternative explanation is that, while both the contour and the category remain activated, as the task shifts from lexical retrieval to articulation preparation, only the articulatory/acoustic congruency benefits production. The present results do not allow us to tease apart these two possibilities.

In Experiment 2, target and distractor conditions were reversed to investigate whether the multi-level processing of allophonic variants found in Experiment 1 could be extended to visual processing of ignored distractor words. The model revealed a significant effect of distractor type that did not interact with target type, suggesting that both the tone category and the context-specific variant were activated. Since mean response times differed between target types, the data were split and analysed separately to verify whether the effect held for both contour and toneme targets. The model for toneme targets confirmed the results from the full model: T3 sandhi distractor words facilitated naming of toneme (low Tone 3) pictures, compared to control distractors. This demonstrates that automatic visual processing of allophonic variants in ignored distractor words involves processing of the tone category.

For the tone contour targets, a Bayesian analysis was run in order to further investigate effects that could not be determined by the LME model. When the LME model was run with data from only the contour targets, but with both SOAs included, the model did not show a significant effect of distractor type. The picture was complicated by the fact that at SOA = 83 ms, mean reaction times (Figure 5) were identical in the two distractor conditions, but with simultaneous presentation of target and distractor (SOA = 0 ms), naming latencies were substantially shorter with congruent T3 sandhi distractors, compared to control distractors. We speculated that there may be an early facilitation effect from contour overlap, but that the present experiment had insufficient power to capture it in the reduced dataset, due to the absence of facilitation at SOA = 83 ms. We investigated this with a Bayesian model, which showed support for facilitation of T3 sandhi distractors on both target types, indicating activation of both the tone category and the tone contour at SOA = 0 ms during visual processing of ignored distractor words.

Overall, the present results point to a different pattern of effects for overt production compared to the ignored distractor words. When target and distractor were presented simultaneously, there was facilitation in both the contour and toneme conditions, regardless of whether the allophonic variants were overtly produced or processed visually. However, when presentation of the distractor was delayed, there was an effect of contour congruency only with overt production (Experiment 1) and an effect of the toneme only with visual processing (Experiment 2).

Taken together, the present results provide evidence for automatic activation of both categorical and context-specific, instantiated representations during both overt production and visual processing of T3 sandhi words. This is consistent with previous studies that have found both abstract and fine-grained processing of segments (McLennan et al., 2003, McLennan, Luce, & Charles-Luce, 2005) and tone (Mitterer et al., 2011) in speech perception, as well as segmental processing in speech production (McLennan et al., 2003, 2005; Nielsen, 2011). The present study extends the evidence for multi-level processing of speech variants to lexical tone production (see also Chen, Shen, & Schiller, 2011) and visual processing of Chinese characters. The results also provide evidence of differences in the time course of processing of the two levels during overt production, compared to when words are visually presented and not overtly produced.

The toneme and contour priming effects seem to reflect two separate processes corresponding to different levels of representation. The toneme effect informs the question of whether processing of allophonic variants activates a category-level representation. Facilitation in the toneme conditions must have occurred at the tone category level because the actual realisation of the T3 sandhi targets and distractors (rising contour) is unrelated to the toneme targets and distractors (low contour). During overt production of the T3 sandhi variants (Experiment 1), *t* values for the SOA = 83 ms model suggest continued facilitation from contour distractors, but not toneme distractors. In Experiment 2, the reverse pattern seems to emerge. The *t* values for SOA = 83 ms indicate that the toneme targets are still facilitated by T3 sandhi distractors, while the contour targets are not.

The present results have interesting implications for the debate about the role of statistical distributions in speech category acquisition and processing. Although a growing body of research demonstrates that statistical information about acoustic cues plays an important role in first and second language acquisition and speech perception (Escudero, Benders, & Wanrooij, 2011; Gulian, Escudero, & Boersma, 2007; Maye & Gerken, 2000; Maye, Werker, & Gerken, 2002; Wanrooij, Escudero, & Raijmakers, 2013), there is also substantial between-category overlap in distributions. In the case of T3 sandhi and Tone 2, the overlap is almost total. Accounts based purely on statistical distributions of acoustic cues would have difficulty explaining the formation of separate categories in such cases where acoustic information from two speech categories is very similar. Recent evidence from computational models also suggests that acoustic distributional information alone may not be sufficient for phonetic category acquisition (Feldman, Myers, White, Griffiths, & Morgan, 2011; McMurray, Aslin, & Toscano, 2009a). Feldman et al. (2011) suggest that phonetic category development occurs as part of extracting meaning from language, through association of phonetic distributions with lexical items. Association with the meanings (and orthography) of the respective characters can explain how different speech categories, such as T2 and T3 sandhi, can form separate representations at the category level, despite very similar acoustic distributions.

The contour priming effect, on the other hand, seems to reflect a different sort of representation. Contour facilitation must have occurred due to acoustic and/or articulatory similarity. This entails representations that are both context-specific and instantiated. They are context-specific because the variant occurs only in the particular phonetic environment when two or more Tone 3 characters occur directly one after the other. They are instantiated because, since target and distractor are unrelated at the category level, the effect must occur due to similar physical properties, that is, acoustic and/or motor-movement similarity. The idea of an instantiated internal representation is consistent with models that posit involvement of the sensori-motor system in speech production and studies showing that auditory and somatosensory feedback are utilised in guiding and adjusting speech production (Davis & Johnsrude, 2007; Guenther, Ghosh, & Tourville, 2006; Guenther & Vladusich, 2009; Houde & Jordan 1998; Jones & Munhall, 2002; Liberman & Whalen, 2000; Purcell & Munhall, 2006).

The contour effect in Experiment 1 opens up new questions about the processing of the contour itself. For example, is it stored lexically? In fact, T3 sandhi occurs not only in words but also across word boundaries. In the present study, there was no effect of bigram frequency. However, since we were interested in tone processing in sandhi *words*, the bigram frequency range in the selected stimuli was restricted. Recall that bigram frequency is a measure of how often two characters occur together, and is used here as a measure of lexicality, since word boundaries are not explicit in Chinese. If similar effects were found even with very low bigram frequency, this would rule out the possibility that the context-specific contour is stored only in lexical items. Two further possibilities are that the contour is stored as part of the morpheme and, alternatively, that it is stored as part of the tone category. Future research could disentangle these possibilities. If the sandhi contour is processed as part of a purely abstract Tone 3 category, effects should be equal for all morphemes. However, if the contour is processed by exemplar, morphemes which rarely occur in sandhi contexts should see significant attenuation of the contour effect relative to morphemes that frequently occur in sandhi contexts.

Conclusion

In conclusion, the present study provides new insights into phonological processing during speech production and visual word processing. Experiments 1 and 2 showed toneme and contour priming effects, indicating multiple levels of representation in production and visual processing of Mandarin tones. The toneme effect indicates activation of the tone category representation, which may be formed through processing of regularities in input data distributions. The contour effect suggests a context-specific and instantiated representation of the actual pitch contour. This can be explained in terms of a somatomotor/auditory target by which speakers gauge production accuracy.

Acknowledgements

We would like to express our gratitude to members of the Psychology Institute of the Chinese Academy of Science for their kind assistance while the first author was testing in Beijing. In particular, thanks to Yuan Yuan, Wang Cheng and Li Xiaoqing for help with participant recruitment and lab set-up and to Xia Jinyan for running additional participants. We also wish two thank two anonymous reviewers for their helpful and constructive suggestions on previous versions of this article. This research was conducted while the first author was at Leiden University. She is currently based at the Quantitative Linguistics Group, University of Tuebingen. This work was supported by the European Research Council under the ERC Starting [Grant 206198] to YC.

Notes

1. Mandarin tones are referred to using a number system (tones 1–4). Here, the numeral following the syllable represents the tone number, in this case, Tone 2.
2. The third tone is sometimes described as ‘falling-rising’. However, the rising part of the contour is optional and does not usually occur when there is a following syllable. In addition, the gradient of the fall is very shallow. For these reasons and for simplicity, we refer to the contour as ‘low’.
3. Significance is reported at the 95% confidence level, unless otherwise specified.
4. In LME, it is unclear what the appropriate degrees of freedom should be. Therefore, in the lme4 package, p values are not provided in the output. However, the t value provides confidence intervals for sufficiently large datasets (1000 data points or more). T values below −2 or above 2 can be taken as significant at the 95% confidence level (see, e.g. Baayen, 2008; Baayen, Davidson, & Bates, 2008; Baayen & Milin, 2010 for full discussion).
5. A Bayesian factor (BF) of 3 or more shows a substantial preference for the model over the alternative. A BF of less than 1 indicates lack of evidence or a preference for the alternative model.

References

Agus, T. R., Thorpe, S. J., & Pressnitzer, D. (2010). Rapid formation of robust auditory memories: Insights from noise. *Neuron*, *66*, 610–618. doi:10.1016/j.neuron.2010.04.014

Alario, F. X., & Ferrand, L. (1999). A set of 400 pictures standardized for French: Norms for name agreement, image agreement, familiarity, visual complexity, image variability, and age of acquisition. *Behavior Research Methods, Instruments, & Computers*, *31*, 531–552. doi:10.3758/BF03200732

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modelling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*, 390–412. doi:10.1016/j.jml.2007.12.005

Baayen, R. H., & Milin, P. (2010). Analyzing reaction times. *International Journal of Psychological Research*, *3*(2), 12–28.

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

Bates, D., Maechler, M., & Bolker, B. (2013). *lme4: Linear mixed-effects models using S4 classes*. R package version 0.999999-2. Retrieved from http://CRAN.R-project.org/package=lme4

Brady, T. F., Konkle, T., Alvarez, G. A., & Oliva, A. (2008). Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences*, *105*, 14325–14329. doi:10.1073/pnas.0803390105

Cai, Q., & Brysbaert, M. (2010). SUBTLEX-CH: Chinese word and character frequencies based on film subtitles. *PLoS ONE*, *5*(6), 8. doi:10.1371/journal.pone.0010729

Clayards, M., Tanenhaus, M. K., Aslin, R. N., & Jacobs, R. A. (2008). Perception of speech reflects optimal use of probabilistic speech cues. *Cognition*, *108*, 804–809. doi:10.1016/j.cognition.2008.04.004

Chen, Y., Shen, R., & Schiller, N. O. (2011). Representation of allophonic tone sandhi variants. *Proceedings of Psycholinguistics Representation of Tone. Satellite Workshop to ICPhS*, Hongkong, 38–41.

Da, J. (2004). A corpus-based study of character and bigram frequencies in Chinese e-texts and its implications for Chinese language instruction: Studies on the theory and methodology of digitalized Chinese teaching to foreigners. In Z. Pu, T. Xie, & J. Xu (Eds.), *Proceedings of the Fourth International Conference on New Technologies in Teaching and Learning Chinese* (pp. 501–511). Beijing: Tsinghua University Press.

Damian, M. F., & Martin, R. C. (1999). Semantic and phonological codes interact in single word production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*, 345–361. doi:10.1037/0278-7393.25.2.345

Davis, M. H., & Johnsrude, I. S. (2007). Hearing speech sounds: Top-down influences on the interface between audition and speech perception. *Hearing Research*, *229*(1–2), 132–147. doi:10.1016/j.heares.2007.01.014

Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review*, *93*, 283–321. doi:10.1037/0033-295X.93.3.283

Dell, G. S. (1988). The retrieval of phonological forms in production: Tests of predictions from a connectionist model. *Journal of Memory and Language*, *27*, 124–142. doi:10.1016/0749-596X(88)90070-8

Escudero, P., Benders, T., & Wanrooij, K. (2011). Enhanced bimodal distributions facilitate the learning of second language vowels. *The Journal of the Acoustical Society of America*, *130*, EL206–EL212. doi:10.1121/1.3629144

Feldman, N., Myers, E., White, K., Griffiths, T., & Morgan, J. (2011). Learners use word-level statistics in phonetic category acquisition. In N. Danis, K. Mesh, & H. Sung (Eds.), *Proceedings of the 35th annual Boston University Conference on Language Development* (pp. 197–209). Somerville, MA: Cascadilla Press.

Foss, D. J., & Swinney, D. A. (1973). On the psychological reality of the phoneme: Perception, identification, and consciousness. *Journal of Verbal Learning and Verbal Behavior*, *12*, 246–257. doi:10.1016/S0022-5371(73)80069-6

Fowler, C. A. (2010). The reality of phonological forms: A reply to Port. *Language Sciences*, *32*(1), 56–59. doi:10.1016/j.langsci.2009.10.015

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

Guenther, F. H., Ghosh, S. S., & Tourville, J. A. (2006). Neural modeling and imaging of the cortical interactions underlying syllable production. *Brain and language*, *96*, 280–301. doi:10.1016/j.bandl.2005.06.001

Guenther, F. H., & Vladusich, T. (2009). A neural theory of speech acquisition and production. *Journal of Neurolinguistics*, *25*, 408–422. doi:10.1016/j.jneuroling.2009.08.006

Gulian, M., Escudero, P., & Boersma, P. (2007). Supervision hampers distributional learning of vowel contrasts. In J. Trouvain & W. J. Barry (Eds.), *Proceedings of the 16th International Congress of Phonetic Sciences* (pp. 1893–1896). Saarbrücken, Germany: Saarland University.

Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review*, *93*, 411–428. doi:10.3758/MC.38.1.102

Houde, J. F., & Jordan, M. I. (1998). Sensorimotor adaptation in speech production. *Science*, *279*, 1213–1216. doi:10.1126/science.279.5354.1213

Indefrey, P., & Levelt, W. J. (2004). The spatial and temporal signatures of word production components. *Cognition*, *92*(1–2), 101–144. doi:10.1016/j.cognition.2002.06.001

Jones, J. A., & Munhall, K. G. (2002). The role of auditory feedback during phonation: Studies of Mandarin tone production. *Journal of Phonetics*, *30*, 303–320. doi:10.1006/jpho.2001.0160

Ju, M., & Luce, P. A. (2006). Representational specificity of within-category phonetic variation in the long-term mental lexicon. *Journal of Experimental Psychology: Human Perception and Performance*, *32*(1), 120–138. doi:10.1037/0096-1523.32.1.120

Levelt, W. J. M. (2001). Spoken word production: A theory of lexical access. *Proceedings of the National Academy of Sciences*, *98*, 13464–13471. doi:10.1073/pnas.231459498

Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, *22*, 1–75. Retrieval from http://journals.cambridge.org/article_S0140525X99001776

Liberman, A., & Whalen, D. (2000). On the relation of speech to language. *Trends in Cognitive Sciences*, *4*(5), 187–196. doi:10.1016/S1364-6613(00)01471-6

Lupker, S. J. (1982). The role of phonetic and orthographic similarity in picture–word interference. *Canadian Journal of Psychology*, *36*, 349–367. doi:10.1037/h0080652

Maye, J., & Gerken, L. (2000). Learning phonemes without minimal pairs. In S. C. Howell, S. A. Fish, & T. Keith-Lucas (Eds.), *Proceedings of the 24th Annual Boston University Conference on Language Development* (pp. 522–533). Somerville, MA: Cascadilla Press.

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, *82*(3), B101–B111. doi:10.1016/S0010-0277(01)00157-3

McLennan, C. T., Luce, P. A., & Charles-Luce, J. (2003). Representation of lexical form. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*, 539–553. doi:10.1037/0278-7393.29.4.539

McLennan, C. T., Luce, P. A., & Charles-Luce, J. (2005). Representation of lexical form: Evidence from studies of sublexical ambiguity. *Journal of Experimental Psychology: Human Perception and Performance*, *31*, 1308–1314. doi:10.1037/0096-1523.31.6.1308

McMurray, B., Aslin, R. N., & Toscano, J. C. (2009a). Statistical learning of phonetic categories: Insights from a computational approach. *Developmental Science*, *12*, 369–378. doi:10.1111/j.1467-7687.2009.00822.x

McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2009b). Within-category VOT affects recovery from "lexical" garden-paths: Evidence against phoneme-level inhibition. *Journal of Memory and Language*, *60*(1), 65–91. doi:10.1016/j.jml.2008.07.002

McQueen, J. M., Cutler, A., & Norris, D. (2006). Phonological abstraction in the mental lexicon. *Cognitive Science*, *30*, 1113–1126. doi:10.1207/s15516709cog0000_79

Meyer, A. S. (1990). The time course of phonological encoding in language production: The encoding of successive syllables of a word. *Journal of Memory and Language*, *29*, 524–545. doi:10.1016/0749-596X(90)90050-A

Meyer, A. S. (1991). The time course of phonological encoding in language production: Phonological encoding inside a syllable. *Journal of Memory and Language*, *30*(1), 69–89. doi:10.1016/0749-596X(91)90011-8

Mitterer, H. (2006). Is vowel normalization independent of lexical processing? *Phonetica*, *63*, 209–229. doi:10.1159/000097306

Mitterer, H., Chen, Y., & Zhou, X. (2011). Phonological abstraction in processing lexical-tone variation: Evidence from a learning paradigm. *Cognitive Science*, *35*(1), 184–197. doi:10.1111/j.1551-6709.2010.01140.x

Newman, R. S., Clouse, S. A., & Burnham, J. L. (2001). The perceptual consequences of within-talker variability in fricative production. *The Journal of the Acoustical Society of America*, *109*, 1181–1196. doi:10.1121/1.1348009

Nielsen, K. (2011). Specificity and abstractness of VOT imitation. *Journal of Phonetics*, *39*(2), 132–142. doi:10.1016/j.wocn.2010.12.007

Peng, S.-H. (2000). Lexical versus 'phonological' representations of Mandarin sandhi tones. In M. B. Broe & J. B. Pierrehumbert (Eds.), *Acquisition and the lexicon: Papers in Laboratory Phonology V* (pp. 152–167). Cambridge: Cambridge University Press.

Purcell, D. W., & Munhall, K. G. (2006). Compensation following real-time manipulation of formants in isolated vowels. *The Journal of the Acoustical Society of America*, *119*, 2288–2297. doi:10.1121/1.2173514

R Core Team (2013). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Retrieval from URL http://www.R-project.org/.

Roelofs, A. (1999). Phonological segments and features as planning units in speech production. *Language & Cognitive Processes*, *14*(2), 173–200. doi:10.1080/016909699386338

Rosinski, R. R., Golinkoff, R. M., & Kukish, K. S. (1975). Automatic semantic processing in a picture–word interference task. *Child Development*, *46*, 247–253. doi:10.2307/1128859

Schriefers, H., Meyer, A. S., & Levelt, W. J. (1990). Exploring the time course of lexical access in language production: Picture-word interference studies. *Journal of Memory and Language*, *29*(1), 86–102. doi:10.1016/0749-596X(90)90011-N

Starreveld, P. A., & La Heij, W. (1996). Time-course analysis of semantic and orthographic context effects in picture naming. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 869–918. doi:10.1037/0278-7393.22.4.896

Trude, A. M., & Brown-Schmidt, S. (2012). Talker-specific perceptual adaptation during online speech perception. *Language and Cognitive Processes*, *27*, 979–1001. doi:10.1080/ 01690965.2011.597153

Wang, W. S-Y., & Li, K.-P. (1967). Tone 3 in Pekinese. *Journal of Speech and Hearing Research 10*, 629–636.

Wanrooij, K., Escudero, P., & Raijmakers, M. E. (2013). What do listeners learn from exposure to a vowel distribution? An analysis of listening strategies in distributional learning. *Journal of Phonetics*, *41*, 307–319. doi:10.1016/j.wocn.2013.03.005

Yuan, J. H., & Chen, Y. (2014). 3RD tone sandhi in Standard Chinese: A corpus approach. *Journal of Chinese Linguistics*, *42*, 218–234.

Zhou, X. L., & Zhuang, J. (2000). Lexical tone in the speech production of Chinese words. *Stroke* 9.8.8, 9-5. Retrieved from http://www.isca-speech.org/archive/icslp_2000/i00_2051.html

Appendix 1. Target picture names and distractor words used in Experiment 1.

Syllable	Target picture	Target picture (English)	Tone 2 distractor	Pinyin	Tone 3 distractor	Pinyin	Control distractor	Pinyin
Ye	野草	Grass	爷	ye2	也	ye3	业	ye4
Meng	蒙古	Mongolia	盟	meng2	猛	meng3	梦	meng4
Du	赌本	Gambling money	读	du2	笃	du3	督	du1
Li	礼品	Present	离	li2	里	li3	立	li4
Miao	秒表	Stopwatch	描	miao2	藐	miao3	庙	miao1
Wang	网孔	Net	王	wang2	往	wang3	望	wang4
Xi	喜酒	Wedding liquor	席	xi2	洗	xi3	系	xi4
Yan	鼹鼠	Mole	严	yan2	演	yan3	烟	yan1
Fu	辅导	Tutor	服	fu2	斧	fu3	付	fu4
Lv	旅馆	Hotel	驴	lv2	履	lv3	绿	lv4
Qi	起点	Starting line	其	qi2	启	qi3	期	qi1
Wu	舞女	(Female) Dancer	无	wu2	午	wu3	乌	wu1
Yang	仰泳	Backstroke	羊	yang2	养	yang3	样	yang4
Zhi	指骨	Finger bone	直	zhi2	止	zhi3	知	zhi1
Bi	匕首	Dagger	鼻	bi2	比	bi3	逼	bi1
Chi	尺码	Clothing size	持	chi2	齿	chi3	痴	chi1
Ji	脊髓	Spine	集	ji2	挤	ji3	机	ji1
Qian	浅海	Shallows	前	qian2	遣	qian3	铅	qian1
Wu	武警	Armed police	无	wu2	午	wu3	污	wu1
Yan	眼角	Corner of eye	言	yan2	掩	yan3	宴	yan4
Yu	雨伞	Umbrella	鱼	yu2	与	yu3	玉	yu4
Zhi	纸板	Cardboard	值	zhi2	趾	zhi3	至	zhi4
Shi	始祖	Ancestor	时	shi2	使	shi3	侍	shi4
Bao	宝塔	Pagoda	雹	bao2	保	bao3	胞	bao1
Zhu	主管	Leader	逐	zhu2	煮	zhu3	筑	zhu4
Chang	场景	Scene	常	chang2	厂	chang3	昌	chang1
Chan	产品	Merchandise	馋	chan2	谄	chan3	掺	chan1

Appendix 2. Target picture names and distractor words used in Experiment 2.

Syllable	T2 Target picture	T2 Target (English)	T3 Target picture	T3 Target (English)	Sandhi distractor	Sandhi distractor pinyin	Control distractor	Control distractor pinyin
bao	雹子	Hail	宝贝	Baby	堡垒	bao3lei3	报纸	bao4zhi3
bi	鼻孔	Nose	笔迹	Writing	彼此	bi3ci3	避免	bi4mian3
chang	长凳	Bench	厂房	Factory	场景	chang3jing3	唱片	chang4pian4
chi	池塘	Pool	齿轮	Gear	尺码	chi3ma3	吃惊	chi1jing1
du	毒蛇	Snake	肚皮	Stomach	赌场	du3chang3	度过	du4guo4
fu	浮标	Buoy	斧头	Axe	辅导	fu3dao3	夫妇	fu1fu4
ji	吉普	Jeep	济南	Ji'nan province	脊髓	ji3sui3	饥荒	ji1huang1
jia	夹克	Jacket	甲虫	Beetle	假想	jia3xiang3	佳人	jia1ren2
jie	洁具	Wash tubs	姐妹	Sisters	解体	jie3ti3	借用	jie4yong4
ju	橘子	Mandarin	矩形	Rectangle	举手	ju3shou3	居民	ju1min2
li	梨子	Pear	礼物	Present	理想	li3xiang3	立柜	li4gui4
qi	骑士	Cavalry	企鹅	Penguin	起点	qi3dian3	期刊	qi1kan1
wan	玩具	Toy	碗柜	Cupboard	婉转	wan3zhuan3	豌豆	wan1dou4
wang	王后	Queen	网球	Tennis	往返	wang3fan3	忘掉	wang4diao4
wei	围巾	Scarf	苇丛	Reeds	猥琐	wei3suo3	卫星	wei4xing1
wu	蜈蚣	Centipede	武器	Weapons	舞蹈	wu3dao3	巫婆	wu1po2
yan	盐巴	Salt	眼镜	Glasses	演讲	yan3jiang3	宴会	yan4hui4
yang	阳台	Balcony	氧气	Oxygen	养老	yang3lao3	样式	yang4shi4
ye	爷爷	Grandpa	野猪	Boar	也好	ye3hao3	叶子	ye4zi
yi	遗址	Ruins	椅子	Chair	以免	yi3mian3	抑郁	yi4yu4
yin	银行	Bank	饮料	Drink	引起	yin3qi3	阴暗	yin1an4
yu	鱼缸	Fish tank	羽毛	Feather	雨伞	yu3san3	玉米	yu4mi3
zao	凿子	Chisel	枣子	Date	早点	zao3dian3	噪音	zao4yin1
zhi	植物	Plant	指环	Ring	只好	zhi3hao3	支票	zhi1piao4