Speech Perception in Chinese:

How Are the Different Levels of Ambiguity Resolved?

TSANG, Yiu Kei

A Thesis Submitted in Partial Fulfillment

of The Requirements for the Degree of

Doctor of Philosophy

in

Psychology

July 2009

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346

Thesis/ Assessment Committee

Professor John Xuexin Zhang (Chair)

Professor Hsuan-Chih Chen (Thesis Supervisor)

Professor Him Cheung (Committee Member)

Professor Anne Cutler (External Examiner)

Professor Ellick Kin-Fai Wong (External Examiner)

Acknowledgements

This work was supported in part by grants from the Research Grants Council of the Hong Kong Special Administrative Region, China (CUHK4142/04H and CUHK441008) to Prof Hsuan-Chih Chen and a dissertation grant from the Department of Psychology in The Chinese University of Hong Kong.

I would like to thank my thesis supervisor, Prof. Hsuan-Chih Chen, for his valuable comments on this work. Thanks are also given to my thesis/ assessment committee chair, Prof. John Xuexin Zhang, and my thesis committee member, Prof. Him Cheung, and the external examiners, Prof. Anne Cutler and Prof. Ellick Kin-Fai Wong for their constructive advices.

TABLE OF CONTENTS

List of Tables

List of Figures

Abstract

Three experiments were conducted to provide a better understanding about the fundamental processes involved in Chinese speech recognition. Specifically, we intended to answer three questions. First, are subsyllabic units like individual phonemes or whole syllables the basic encoding units in Chinese speech recognition? Second, does tone play a significant role in generating candidate words before correct identification? Third, how can the different meanings of homophones be resolved? In Experiment 1, we used the gating paradigm to explore the three issues. Results suggested that both subsyllabic (onset) and syllabic representations were important in recognizing Chinese monosyllables. Tonal constraints emerged only when context was available. And context also facilitated homophone recognition. In Experiment 2, the visual-world paradigm was used to verify the major findings in gating. While the salience of syllable and the absence of tonal constraints without context were replicated, the onset effect was greatly diminished. Further analyses suggested that acoustic similarity might also play a role in speech recognition. Experiment 3 also employed the visual-world paradigm. The resolution of Chinese homophones was found to be influenced by relative meaning frequency and context position. Based on these findings and those from related studies, we proposed a model of Chinese speech perception, in which initially, segmental and suprasegmental types of information were processed in separate but interacting pathways. Outputs from the two pathways were then combined at a later time point and jointly activated the corresponding morpheme. Implications of the model and its relations to previous findings are discussed.

摘要

本論文旨在探討中文當中語言識別的基本過程。我們特別關注的問題有三個：第一，中文語言識別的基本單位是音節還是音素？第二，聲調在言語識別中有何作用？第三，如何提取中文同音字的語意？實驗一採用閘門範式 (Gating Paradigm)對這三個問題進行初步研究，結果顯示，聲母和整個音節都對中文單音節詞語識別產生了重要作用，而聲調則只於存在語境的情況下才起作用，同時，語境亦能協助提取同音字的正確意思。實驗二採用視覺世界範式 (Visual-world Paradigm) 驗證實驗一的結果，雖然成功複製了整個意節的重要性和聲調的次要性，但聲母的效果卻大爲減弱。進一步的分析顯示：兩個音節在聲學上的相似度也會影響識別它們時的困難。實驗三同樣採用視覺世界範式探討中文同音字語意提取的問題，結果顯示特定語意的常用度和語境的位置都是重要因素。基於以上的發現以及相關研究的結果，我們提出了一個中文語言識別的模型，此模型表明，在語言識別的初期，音段訊息和超音段訊息循不同管道進行加工，這兩個管道獨立而又相互影響，在加工的較後階段，音節加工和聲調加工的結果則會統合起來激活相應的詞素。最後，我們會討論此模型對中文語言識別的含意，並它與前人結果的關聯。

Chapter 1

Introduction

While language is perhaps the "default" tool of communication for ordinary people all over the world, the form by which linguistic ability is realized differs substantially. For example, different languages have their own phonemic distinctions (e.g., /r/ and /l/ are treated the same in Japanese, but not in English), are governed by unique principles in word formation (e.g., no inflection exist in Chinese), and involve their own structural constraints (e.g., Japanese prefers a SOV structure than SVO). Whether differences in these surface characteristics are translatable into differences in the underlying principles in language processing is an interesting and important theoretical question. In the present dissertation, answers to part of this question will be explored from the perspective of ambiguity resolution in Chinese speech, which has a number of unique features distinct from the mostly-researched Indo-European languages such as English. Through such investigation, it is expected that a working model of Chinese speech comprehension can be constructed, and contrasted with existing theories based primarily on Indo-European languages.

This dissertation is divided into seven chapters. The first chapter aims at introducing general concepts about how ambiguities exist at different levels of the speech signal. Also, three important unique features of Chinese speech are described with reference to how they possibly create distinct processing demands in ambiguity resolution in Chinese speech, and they are related to the research goal of illuminating unexplored areas of speech processing in previous studies. The second chapter focuses on past research on speech processing and lexical ambiguity resolution conducted mainly in English. We will see that while theories developed based on these studies can explain various phenomena in processing English, they may not be

directly applicable to the case of Chinese speech due to its unique characteristics. Chapter Three outlines the paradigms and the hypotheses of three experiments aiming at investigating whether the three unique features reviewed in the introduction actually matter in Chinese speech processing. Chapters Four, Five, and Six report the three experiments conducted and discuss their results separately. Chapter Seven offers a general discussion of how the data from the three experiments contribute to the understanding of Chinese speech comprehension, and the more general issue of cross-language difference in processing dynamics. The ultimate goal is to provide empirical constraints for models of Chinese speech comprehension.

1.1 Multi-level ambiguities in language

In language processing research, presenting participants with stimuli of multiple interpretations and examining how they can come up with a proper understanding has always been an important technique. This procedure has provided valuable insights in studying the underlying cognitive mechanisms in various aspects of language processing. Representative examples adopting such approach include studies on lexical access (e.g., Borowsky, & Masson, 1996; Hino, Lupker, & Pexman, 2002; Rubenstein, Garfield, & Millikan, 1970), syntactic processing (e.g., Grodner, Gibson, & Tunstall, 2002), and discourse effects (e.g., Binder, 2003). In the present dissertation, the same approach will be taken to study the unique problems in Chinese speech processing.

1.1.1 Temporary perceptual ambiguity

Indeed, despite the apparent ease in comprehending utterances made by others, a closer inspection of the speech signal will reveal a high degree of ambiguities at

different levels, from the very fundamental uncertainties due to insufficient acoustic inputs (Dahan & Gaskell, 2007) up to higher level language uses such as metaphors and indirect requests (Coulson & van Petten, 2002). Specifically, unlike the relatively simultaneous availability of word features in the visual domain, acoustic signals of a spoken word only unfold gradually over time. This temporal stretching creates temporary acoustic under-specification in signal interpretation. For instance, simply perceiving the onset consonant /b/ will be insufficient for determining whether the syllable will be /bi:/ or /be/. Similarly, knowing that it is /bi:/ will not guarantee it is /bi:m/ but not /bi:t/. Therefore, to make sense of the inputs, one has to monitor the unfolding signal continuously and integrate later acoustic inputs incrementally with earlier ones until enough information is gathered. Actually, one fundamental finding in speech perception is that people appear to make hypothetical candidate words based on the acoustic information available thus far and use new incoming information to test those hypotheses. Candidates inconsistent with the bottom-up inputs would be rejected from further consideration (Marslen-Wilson & Welsh, 1978; see also Chapter 2.1).

Evidence for such acoustic level disambiguation comes mainly from studies related to the uniqueness point, which refers to the point at which the string of inputs is consistent with only a single interpretation (i.e., it is no longer ambiguous). These studies employed paradigms such as gating (Grosjean, 1980) and pause-detection (Mattys & Clark, 2002). Reaction time data showed that words with earlier uniqueness point, irrespective of word length, are recognized faster, presumably because the initial ambiguity inherent in these words requires less acoustic information to clarify. More importantly, the pattern of responses in gating tasks directly highlighted the presence of competition among partially active lexical

candidates that were consistent with the current acoustic inputs; participants provided various alternatives and showed low level of confidence about their responses until enough acoustic inputs had been received. This competition among lexical candidates and the corresponding disambiguation are captured by most contemporary theories of speech perception as an incremental integration of phonemes or phonetic features from onset to the uniqueness point (Gaskell & Marslen-Wilson, 1997; Marslen-Wilson & Welsh, 1978; McClelland and Elman, 1986). These theories, and their applicability to Chinese speech, will be reviewed in Chapter Two.

1.1.2 Morphemic ambiguity

Besides acoustic under-specification, ambiguity can also be characterized at the morphemic level when the same physical unit is linked to different meanings. For instance, the prefix "in-" of "invalid" and "inside" are quite different: The former means "not" while the latter denotes the opposite of "out". Similarly, the suffix "-er" in "heater" and "faster" also represents distinct meanings. In principle, this situation reflects ambiguity resolution on the smallest meaningful scale because the remaining portion of the word should have provided enough contextual information to retrieve the correct morpheme meaning. This allows the investigation of sensitivity to minimal ambiguity. Despite such theoretical potential, very limited works have been done on it, especially in the auditory modality. Bertram and his colleagues were perhaps the only group of researchers who had systematically investigated the issue (e.g., Bertram, Hyönä, & Laine, 2000a; Bertram, Laine, Baayen, Schreuder, & Hyönä, 2000b). They studied what they called "affixal homonymy" in Finnish, which is considered to be a morphologically rich language. Specifically, they tested whether the ambiguous suffix "-ja" (either as a deverbal subject noun marker or to denote

partitive plural) led Finnish readers to rely more on the morphological route of lexical access. Results of lexical decision on isolated words showed no evidence of morphological effect (Bertram et al., 2000a). On the other hand, employing an eye tracking and self-paced reading technique, it was found that reading times of the same stimuli in sentence context displayed a delayed morpheme (meaning) frequency effect, suggesting that morphemes are nevertheless a valid processing unit among Finnish readers (Bertram et al., 2000b). Bertram et al. (2000b) attributed the discrepancy between studies to the support of prior sentential context in resolving the ambiguous suffix "-ja", causing recognition through morpheme to be more efficient than whole word access. Two important conclusions can be drawn from their studies: First, morphemic ambiguity is a valid and testable phenomenon. Second, the prior sentence context is effective in affecting the relative availability of the meanings of the suffix.

However, the studies by Bertram et al. (2000a; 2000b) failed to answer the more critical question about the processing dynamics in resolving morphemic ambiguity. The only information available is the relatively late effect of morpheme frequency compared with word frequency. Without more details about when and how the correct meaning of a suffix is available, it is impossible to confidently interpret the null effect in lexical decision as evidence of less efficient disambiguation of suffix by base morpheme alone (Bertram et al., 2000a) compared with disambiguation by prior sentence (Bertram et al., 2000b). It is because the resolution processes may simply be too fast to be captured by traditional approaches like lexical decision. Indeed, a previous eye-tracking study of reading Chinese two-character words in sentences did reveal a rapid ambiguity resolution process once contextual information was encountered (Wong, 2000). Similar conclusion was arrived in Tsang (2006), who showed that a single morpheme is sufficient to trigger rapid meaning resolution in

Chinese disyllabic words. In two experiments, Tsang investigated the disambiguation of homonymic morphemes (i.e., those that share both orthography and phonology such as gaau3 [1]: 教, which means either education or religion). Results suggested that an initial ambiguous syllable can be resolved by a contextual syllable without any delay (Experiment 1). Moreover, a prior sentential context even allowed anticipation of the correct meaning before the ambiguous syllable is encountered (Experiment 2). The results thus converged with Bertram et al. (2000b) on the existence of morphemic ambiguity. Moreover, it further illustrated the power of a minimal context in resolving this ambiguity, at least in languages where words are easily decomposed into isolated, discrete morphemes, such as Chinese (Packard, 1999) or Finnish (Bertram et al., 2000a; 2000b)

Still, one aspect of morphemic ambiguity resolution was left unanswered. In Chen, Tsang, Chan, and Wong (Experiment 1; manuscript) and Tsang (2006, Experiment 2), a prior sentential context was available before the ambiguous morpheme. There were no conditions in which a single preceding morpheme served as the disambiguating context. Therefore, the results only provided partial evidence about the effect of minimal context in morphemic ambiguity resolution; while a succeeding morpheme is useful in Tsang's Experiment 1, it is unclear whether a preceding context morpheme is equally constraining. This issue will be elaborated in Chapter Three.

1.1.3 Lexical ambiguity

While the issue of morphemic ambiguity is new and not much is known about its underlying resolution processes, the investigation of lexical ambiguity resolution

[1] We follow the Cantonese Romantization Scheme proposed by the Linguistic Society of Hong Kong.

has a long history since the 1970s. Actually, besides validating the phenomenon of morphemic ambiguity, another equally important contribution in Tsang (2006) is the discovery that the mechanisms for resolving morphemic ambiguity parallel those for resolving lexical ambiguity. In his experiments, two variables, namely the position of disambiguating context and the relative frequency between the alternative meanings, were manipulated. Both variables exerted significant impact on the resolution processes. This pattern of results was consistent with the reordered access model developed to explain lexical ambiguity resolution (Duffy, Morris, & Rayner, 1988). Therefore, the results in Tsang provided the theoretical basis for researchers to borrow concepts and models established in lexical ambiguity to study the issue of morphemic ambiguity resolution.

Specifically, lexical ambiguity refers to the situation in which the same word is linked to different concepts. A classic example is the word "bank", which means either the "river bank" or the "financial institute". The investigation of the underlying principle by which the meanings of these ambiguous words are represented and retrieved can make a significant contribution to the more fundamental question about the architecture of the mental lexicons. Numerous studies have confirmed an advantage for such ambiguous words over unambiguous words (e.g., Kellas, Ferraro, & Simpson, 1988; Rubenstein et al., 1970). Specifically, Kellas et al. obtained faster reaction times and lower error rates for ambiguous words compared with control unambiguous words and pseudowords in two lexical decision experiments. Similarly effects were obtained in word naming (see Borowsky & Masson, 1996; Hino et al., 2002; Lichacz, Herdman, LeFevre, & Baird, 1999). This pattern of results was typically interpreted as supporting evidence for models assuming an interactive distributed lexical representation with extensive feedback connections from semantic

layer to orthographic and phonological layers (Borowsky & Masson, 1996; Hino et al., 2002; Kellas et al., 1988).

Another implication from the studies of lexical ambiguity concerns the mechanisms involved in semantic retrieval during comprehension. Interestingly, in contrast to observations of ambiguity advantages in lexical decision and naming, responses tend to be slower for ambiguous words when semantic access is required, such as semantic categorization (Hino et al., 2002), relatedness judgment (Chwilla & Kolk, 2003), and sentence comprehension (Rayner & Frazier, 1989). These studies provided important empirical constraints for the construction of models related to meaning resolution in lexical ambiguity. Details of these models will be discussed in Chapter Two.

1.1.4 Non-literal meanings

Ambiguities in comprehension penetrate through higher levels of language use such as indirect request, metaphor and irony. A good example is the utterance "It is very hot", which may act as an indirect request to turn on the air-conditioner, rather than simply describing a fact or event. To comprehend these non-literal meanings, one will require a combination of various contextual supports, some of which are perhaps verbal signals, and many others are conveyed nonverbally (Kelly, Barr, Church, & Lynch, 1999). For instance, Kelly et al. have demonstrated that pointing gestures can effectively facilitate the indirect request interpretation of utterances. Moreover, other studies suggested that factors like predictability, conventionality, and plausibility are important in determining the relative time course of activation of literal and non-literal meanings (Gibbs, 1983; Titone & Connine, 1994). Therefore, resolution at the pragmatic level appears to be no less simple. It is co-determined by a

number of factors related to the contextual bias and relative availability of different meanings, just as the ambiguity resolution at other levels.

In the present dissertation, focus will be on the more fundamental resolutions at the acoustic and morphemic levels. It should be noted that although traditionally, the investigation of basic speech unit is not a particular concern in the ambiguity resolution literature, this issue is actually highly relevant. It is because when the unit of processing is expanded into the whole phrase or even the whole sentence, there will no longer be any ambiguity as the processing unit is disambiguated internally. In other words, the exploration of processing unit is indeed fundamental to ambiguity resolution. This is less an issue in English because researchers generally agreed that phonemes and words are important encoding units. However, much more works about this have to been done in Chinese. It would be important to clarify the basic encoding unit of Chinese speech to ensure the description of ambiguity resolution at the appropriate levels. In particular, if acoustic inputs are mapped onto Chinese words directly, ambiguities at prelexical levels will be irrelevant to Chinese speech processing. Moreover, understanding the processes at different levels will be crucial in constructing a possible model of Chinese speech perception, which has a number of unique features that undermines the validity of generalization of models developed in Indo-European languages.

1.2 Unique characteristics of Chinese speech

Although the study of speech processing in Indo-European languages has a long history, the issue is relatively unexplored in Chinese. It is unclear whether the results obtained in Indo-European languages can be generalized to Chinese, which has many unique characteristics. Specifically, Chinese has a very simple morphosyllabic

structure such that each syllable has its own meaning (Chen, 1992; 2001; Chen & Yip, 2001). This increases the saliency of whole-syllables compared with individual phonemes. Moreover, Chinese speech can be categorized as a tonal language. In contrast to most Indo-European languages, in Chinese lexical distinction is made not only by segmental cues (i.e., phonemes) but also by pitch, a suprasegmental cue. The reliance of lexical tone in speech perception presents a fundamental challenge to modeling speech recognition because pitch level is correlated with many other features like gender and emotion. Only very limited information is known about how lexical tone is extracted and integrated to segmental information in spoken word recognition (e.g., Cutler & Chen, 1997; Sum, 2003; Ye & Connine, 1999). Finally, while lexical tone helps differentiate otherwise identical segments, homophony (identical syllable-plus-tone) is still serious in spoken Chinese. Given that each syllable-plus-tone unit is also a morpheme and has its own meaning, the homophonic morphemes will also create morphemic ambiguity similar to the homonymic morphemes in Tsang (2006). Whether the resolution of such ambiguity is similar to that observed by Tsang previously is an empirical question.

Given the special properties of Chinese speech, it is desirable to understand its underlying processing mechanisms and compare them with those observed in Indo-European languages in order to construct a universal model of speech perception. In the following, we will discuss in more details the uniqueness of Chinese speech, and how the special properties might produce distinctive processing demand.

1.2.1 *Syllable as the prominent unit*

There are at least three reasons to speculate that syllables, instead of phonemes, are the fundamental unit in Chinese speech perception. First, linguistically, each

spoken syllable in Chinese corresponds tightly to a written character and a meaningful morpheme (Chen, 2001). Moreover, syllables in Chinese are relatively simple (fewer possible combinations than other languages, see Chen, Chen, & Dell, 2002; Cheung, Chen, Lai, Wong, & Hills, 2001) and more "outstanding" acoustically (no "unstressed" syllables). In contrast, English syllables often involve complex phonemic combination such as consonant clusters. Also, only one syllable is stressed in multi-syllabic words. Therefore, compared to English speakers, Chinese speakers are potentially more capable to extract structures consistent with syllable boundary during speech comprehension.

Second, developmental data among Chinese-speaking children has consistently shown better performance in tasks assessing syllable awareness than those assessing phonemic awareness (Cheung et al., 2001; McBride-Chang, Bialystok, Chong, & Li, 2004). For instance, Cheung et al. (2001) delivered a sound-matching task to children aged four to eight. When children at different ages performed the matching based on whole-syllable, accuracy reached 90%, as compared to about 50% accuracy when matching onset, rime or coda. Even within a group of children who had received explicit Pinyin training that sensitized them to phonemes, accuracy in phoneme matching was still about 20% behind that of syllable matching. Converging evidence of syllable prominence in Chinese children was obtained in another study using syllable deletion and onset deletion tasks (McBride-Chang et al., 2004). Furthermore, in this study syllable awareness predicted Chinese reading better than phonemic awareness while the reverse was true for reading English. These results suggested that syllables are particularly relevant to Chinese processing and pre-school children have already been sensitized to this representation unit.

Third, results from Chinese speech production converged on the importance of syllables among adult Chinese speakers (Chen, 2000; Chen et al., 2002). In a speech error study, Chen (2000) found that Chinese speakers frequently commit errors involving movement of the whole syllable. For example, the word /qing1zhuo2du4/ ("清楚度"; "clarity") was spoken as /qing1du2du4/, a case of anticipation of the third syllable. In another well-controlled experimental study employing the implicit priming paradigm, Chen et al. (2002) discovered that while whole syllables facilitated responses significantly, onset alone failed to produce any effect. Assuming perception is the reverse of production, it is reasonable to expect similar syllable prominence during spoken word recognition. In other words, Chinese word perception will be less sensitive to phonemic contrasts during the course of activating and selecting items in the candidate set.

In summary, the research reviewed above supported taking syllables as the basic units in constructing Chinese speech perception model. This means that syllable is the first abstract phonological representation (i.e., having psychological validity) in the Chinese speech perception system. On the other hand, some studies suggested that among adult Chinese speakers subsyllabic units might recruit distinct neural mechanisms (Siok, Jin, Fletcher, & Tan, 2003) and prime auditory word recognition (Sum, 2003). Also, mismatches in subsyllabic units of target words appeared to create interference in sentence comprehension (Schirmer, Tang, Penney, Gunter, & Chen, 2005) as well. Yet, whether these results suggest that phonemes mediates normal speech perception or simply reflect Chinese adults' ability to decompose syllables into phonemes in specific tasks is still unclear. For instance, the study by Siok et al. (2003) required participants to perform phonemic judgment (see also Chen et al. (2002) for proposal of phonemic decomposition after syllable is available).

Therefore, a better understanding on this issue would be important to accurately describe the phonological unit(s) mediating Chinese speech perception.

1.2.2 *Importance of lexical tone*

Besides the prominence of syllable, Chinese has another interesting feature different from other Indo-European languages, namely the use of fundamental frequency (i.e., pitch) to distinguish lexical units. In other words, distinct Chinese lexical units can be constructed through assigning different tone values to the same syllable. For instance, the Mandarin (the major dialect of Chinese) syllable /ma/ can mean "mother" or "horse" when it is produced with a high rising pitch (tone 1) and a dipping pitch (tone 3) respectively. Therefore, lexical tone serves to clarify the high degree of meaning uncertainty that an isolated syllable carries alone. This kind of lexical tone differentiation is exclusive to tonal languages, although other suprasegmental cues like stress pattern in multi-syllabic words and intonation over the whole sentence are also present in non-tonal languages like English. To a certain extent, the employment of tonal distinction increases the information density carried in monosyllabic units and reduces the reliance on extra segments for disambiguation. For example, the syllable /yi/ has more than170 homophones (e.g., " ", "意", "疑", etc…), but this number is greatly reduced for the syllable-plus-tone unit /yi4/ (e.g., "意"; see Li & Yip, 1998). Therefore, in principle, tones should be a valuable cue for maximizing efficiency in spoken word identification in tonal languages. Indeed, empirical evidence supported the importance of tonal constraints in Chinese speech perception (e.g., Schirmer et al., 2005; Ye & Connine, 1999; Zhou, Qu, Shu, Gaskell, & Marslen-Wilson, 2004; but also see Cutler and Chen, 1997, for weak tone effects).

Acoustically, tone is carried by the pitch level (or fundamental frequency F0) in speech. When onset pitch level combines with another acoustic feature, the pitch contour, which refers to the variation in F0 over time, the identity of the tone being heard can be defined. Therefore, varying either or both of these tonal features over the same syllable leads to entirely different meanings. How tone is combined with the segmental information during Chinese speech comprehension has been the focus of previous research. Some studies found that during semantic access, tonal constraint was employed at least as quickly as the segmental one (Brown-Schmidt & Canseco-Gonzalez, 2004; Schirmer et al., 2005; Ye & Connine, 1999 Experiment 2; Zhou et al., 2004). Zhou et al. (2004) employed a cross-modal semantic priming procedure to study the availability of tonal constraints. They presented their participants with auditory disyllabic words before a visual target. The congruent prime was semantically related to the target while the tone-altered one was not. If tone did not exert early effect on meaning access, participants would confuse the tone-altered prime with the congruent one. Then lexical decision on the visual target should be facilitated in both congruent and tone-altered conditions. However, no facilitation was obtained in the tone-altered condition, suggesting that tonal constraint is immediate and would not allow semantic access of words sharing segments but differing in tonal information.

In another study, Ye and Connine (1999) found that when the target syllable was preceded by a strong context, tonal violations were detected immediately. Specifically, when the target syllable served to complete an idiom, recognition of the tone in this syllable became faster than that of vowel. Tone has also been found to exert early constraints on semantic access in the context of sentence comprehension. For instance, Brown-Schmidt and Canseco-Gonzalez (2004) studied the N400 ERP

component elicited when listeners encounter an anomalous word in a sentence that differed from the predicted one either by syllable or tone. Results indicated that the onset of N400 in the tone-mismatch condition is as early as, if not earlier than, the syllable-mismatch condition. Similar results were obtained in Schirmer et al. (2005), which contrast rime-mismatch condition with tone-mismatch condition. Therefore, empirical evidence suggested that tone is available as quickly as, or even more accessible than (as in Ye and Connine, 1999), segmental information in meaning retrieval in Chinese speech.

Although tone can provide early constraints in meaning retrieval of spoken Chinese, it may indeed be available late perceptually. Specifically, according to Cutler and Chen (1997) the acoustic features of tones "are primarily realized upon vowels" (p. 176), presumably because the vowel nucleus allows more F0 variations, which is necessary to reveal the shape of tone contour. On the other hand, phonemic contrasts are more localized as specific articulatory cues (e.g., voice onset time; place of articulation, etc…), Cutler and Chen hypothesized that this precision allows the onset and rime of a syllable to be available earlier than tone. This hypothesis received supports from an auditory lexical decision task. Participants made more false alarms (incorrectly accepted nonwords as real words) on disyllabic words when syllables mismatched on tone than on onset or vowel, suggesting that participants relied more on segmental information in lexical access. Similar difficulty was observed in a syllable-matching task (Experiment 2, Cutler & Chen) and a tone-vowel detection task (Ye & Connine, 1999). Furthermore, Luo et al. (2006) discovered that pre-attentively, tonal features are indeed first processed in the right hemisphere, which is also the processing center for music and intonation. However, computation of lexical tone relies primarily on the left hemisphere, just as other segmental information

(Gandour, et al., 2003). Therefore, tone has both linguistic and non-linguistic properties and the perception of fundamental frequency may take time to be transferred to the language system and converted to lexical tone. When semantic context is available, this transfer process may be speeded up and lead to the early constraints observed in other studies.

Conclusively, empirical evidence suggested an important role of tone during Chinese spoken word recognition. Its effect appears to be slower when its perceptual nature was emphasized, but the effect emerges much faster when the task concerns more with meaning retrieval. However, not much is known about the role of tone in candidate generation and elimination before the actual target is finally identified[2]. In other words, what we know is simply the product after stable identification. However, it is possible that tone can activate its own set of lexical candidates. Hints for such activation could be found in previous studies. First, Li and Yip (1998) used a gating paradigm to study the recognition of homophones. Interestingly, they found that, for example, when the target is /kwong3/, not only do segment-sharing candidates like /kwong4/ were reported, items sharing tone only, such as /gok3/, /gwok3/ and /kok3/ were also produced, suggesting that they had been included in the candidate set. Similarly, in Sum (2003), even when an auditory prime was altered such that only lexical tone remained the same (e.g., /baau1/ changed into /kwing1/), it could still facilitate subsequent auditory word recognition to the same degree as the original prime, suggesting that the original prime had been activated based on tone-sharing alone.

1.2.3 *Highly homophonic*

[2] Although Ye and Connine (1999) proposed a modified TRACE model that incorporated a "toneme" layer, they did not actually study the time-to-time changes of activation level due to tonal information.

Although tone greatly reduces ambiguities carried in a syllable, the syllable-plus-tone unit can still be ambiguous due to the existence of homophones. For instance, /gaau3/ in Cantonese can refer to morphemes meaning "education" or "religion" (as in the character "教" in "教師/teacher" and "教堂/church") and "comparison" (as in the character "較" in "較量/compete"). This one-to-many mapping between syllable and morpheme is nothing trivial when one realizes how frequent this occurs in Chinese. For instance, Li and Yip (1998) noted that in the four-tone Mandarin system, 80 percent of syllable-plus-tone units are ambiguous and 55 percent even have five or more meanings. In some extreme cases, such as in the six-tone Cantonese dialect, more than 50 meanings like lion, death, silk, thought, teacher, poem, etc., can share one single pronunciation /si1/ (Chinese character database, http://humanum.arts.cuhk.edu.hk/Lexis/lexi-can/). In this article, this phenomenon of one-to-many mapping between particular syllable and morpheme is called morphemic ambiguity, an issue that must be addressed in order to construct a comprehensive model of spoken Chinese comprehension.

The most straight-forward way to solve the problem of morphemic ambiguity is to bypass the morpheme layer and directly activate whole word lexical entries from acoustic features. As words are accessed holistically, any sublexical ambiguity will be irrelevant and will not require extra resources to compute (Packard, 1999). However, such position is inconsistent with the large body of empirical research confirming a morpheme decomposition route of lexical access (e.g., Alvarez, Carreiras, & Taft, 2001; Frost, Kugler, Deutsch, & Forster, 2005; Marslen-Wilson, 1999; Niswander-Klement & Pollatsek, 2006; Peng, Liu, & Wang, 1999; Taft, Liu, & Zhu, 1999; Zwitserlood, Bolwiender, & Drews, 2005). These studies generally supported a morpheme effect beyond that induced by mere form overlapping (Alvarez

et al., 2001; Frost et al., 2006), which is still present after controlling for whole-word factor (Niswander-Klement & Pollatsek, 2006; Taft, et al., 1999) and is observable in different languages (e.g., Chinese, English, and Spanish).

Yet, morphemic ambiguity may actually alter the availability of morphemes in lexical access. For instance, according to a study in Finnish (Bertram et al., 2000b), ambiguous affix might reduce the reliance of morphemes in lexical access. Previously, Tsang (2006) tested whether morphemes were bypassed in Cantonese spoken word recognition and the results he obtained indeed suggested robust morphemic involvement. In his study, a special case of homophonic morpheme, the homonymic morpheme, was studied. Homonymic morphemes share both orthography and phonology but have two or more meanings (e.g., "教" /gaau3/), one of which has a higher frequency of use (i.e., the dominant meaning). Testing the issue of morpheme involvement with these specific cases has the advantage of intrinsic control over confounding due to physical form because the visual and auditory forms are exactly identical across morpheme conditions. However, this also greatly limits the usable item pool, causing difficulties in generalizing the findings to other ambiguous syllables. Thus, it is unclear whether the conclusions in Tsang (2006) are equally applicable to the homophonic morphemes such as /si1/ (meaning "獅", lion; "師", teacher; and "絲", silk, etc...).

Chapter 2

Theories

In this chapter, important theories relevant to speech perception and ambiguity resolution will be reviewed. These theories were based mainly on research in Indo-European languages. Much fewer works have been done in other languages, such as the East-Asian ones. Therefore, it is unclear whether the theories can be generalized to Chinese, a language with many distinctive features. In the last part of each section, recent progress in modeling Chinese speech comprehension will also be reviewed.

2.1 Theories of speech comprehension

"The history of research on spoken word recognition is largely a history of word recognition models" (Jusczyk & Luce, 2002, p. 502). Although these models vary in the exact mechanism, such as whether top-down feedback or lateral inhibition is incorporated, most of them agree that speech recognition is characterized by cycles of activation and selection. In each cycle, the speech recognizer receives acoustic input and integrates it with inputs from previous cycles. Lexicons that are consistent with the inputs thus far will be activated. This "candidate set" of lexical items will then enters a stage of selection. When a particular item reaches certain activation threshold or remains the most active at a pre-specified processing deadline, this item will be "selected" (recognized). Within this general framework, researchers have worked for thirty years to answer three fundamental questions about the processing dynamics in speech recognition (Frauenfelder & Peeters, 1998):

1.) What items will be included in the candidate set?

2.) How do the candidates affect each other during the selection process?

3.) When is the word recognize (i.e., the definition of selection deadline)?

In this paper, three widely-discussed spoken word recognition models, including the cohort model, TRACE, and Neighborhood Activation Model (NAM) will be reviewed. As we will see, existing models answers these questions mainly through proposing specific phonemic processing mechanisms.

2.1.1 Cohort model

According to the cohort theory (Marslen-Wilson & Welsh, 1978), bottom-up phonemic input is solely responsible for determining what items will be included in the candidate set. Initially, all words that are consistent with the first phoneme heard will become active. This initial candidate set is called the "cohort". For instance, when the onset /b/ is perceived, onset-matched lexical items such as "beam", "beetle" and "beaker" will be activated. Uncertainties in this initial cohort are then disambiguated by subsequent incoming acoustic signals such that only candidates which still align completely with the input remains and any slight mismatches are "kicked out" from the cohort. The process continues until only one candidate in the set is consistent with all incoming signals received thus far (i.e., reaching the uniqueness point of that word). This candidate is consciously recognized. It should be noted that in this architecture, words in the cohort do not compete with each other directly. A word, even with high frequency and small number of competitors, will be recognized late when there is another word that diverges with it at a late acoustic point. In other words, the time required for successful word recognition is entirely dependent on the position of uniqueness point.

The cohort model is an important first step in modeling speech recognition. It successfully established several important processing principles, like the effect of uniqueness point, that need to be considered in understanding the processing

dynamics in spoken word recognition. It has been modified in light of new empirical evidence against some of its major assumptions, such as its over-reliance of word initial information (Marslen-Wilson, 1987). Gaskell and colleagues have also explicitly implemented the model in computer simulations using a distributed processing architecture (Gaskell & Marslen-Wilson, 1997; Gaskell, Hare, & Marslen-Wilson, 1995). Despite these variations, the model maintains its original claim that there are no direct competitions among the candidate items. This stands in sharp contrast to TRACE, which relies heavily on lateral inhibition among partially activated words in the selecting the "correct" candidate for stable recognition.

2.1.2 TRACE

Adopting an interactive-activation framework, McClelland and Elman (1986) proposed the TRACE model of spoken word recognition. The model incorporates three hierarchically organized layers, namely feature, phoneme and word. Between layers are extensive feedforward and feedback excitatory connections. More importantly, units within a layer are connected with inhibitory loops, leading to a winner-take-all phenomenon. Therefore, when a particular word is highly active because it matches well with the incoming features and phonemes, it will also exert strong inhibition on other words, leaving itself as the only active unit in this layer. In other words, unlike the cohort model, TRACE assumes that partially activated words compete with each other directly. When an absolute activation threshold is reached, or one item is more active compared with other partially active words to a certain degree, the best-fit item wins the competition and is recognized.

Another important distinctiveness of TRACE is related to its definition of candidate set from which the "correct" word is selected. The "cohort set" in the

traditional cohort model includes only onset-aligned words so any noise at word onset will lead to failures in word recognition because there are no mechanisms allowing acoustic features at later time points to recover early mismatches. This over-reliance on onset information is clearly inconsistent with our daily experiences of natural conversation, in which we can nevertheless recognize what others say despite a noisy environment or unclear word segmentation. The interactive-activation architecture of TRACE provides a natural solution to this problem: Acoustic features entering the system at any time point can lead to excitations at the phoneme layer, which in turn activates the corresponding matched words. Therefore, even when a word initial is not perceived properly, the word can still be activated if later acoustic cues converge on it. In short, at any time slice, some items in the candidate set will be inhibited due to mismatches with the input, while other items may receive activations from new inputs strong enough to overcome the lateral inhibition originally exerted on them. This greatly improves the "mobility" of items in the candidate set.

The inclusion of candidates that match the input at any position may dramatically increase the candidate set size. This seemingly unrealistic consequence is partly avoided by the lateral inhibition: Items that are only weakly consistent with the inputs will not enter the set because they are strongly "repelled" by current items that match better. Another approach to alleviate the problem is to assume an upper limit in the number of items that can concurrently enter the selection process. This approach is adopted in another connectionist model, namely the Shortlist model (Norris, 1994). In this model, the system first "shortlists" some highly activated items. These items are then fed into an interactive-activation network similar to TRACE. A word is "recognized" when certain threshold is reached.

2.1.3 Neighborhood Activation Model (NAM)

While the cohort model and TRACE model are explicit processing model that emphasize on the underlying mechanisms of recognition, proponents of NAM (Luce & Pisoni, 1998) are more interested in how the global similarity among words can affect the outcomes of complex internal processing. It has the most detailed specification of factors involved in determining successful recognition. For instance, the model proposes that recognition of a word will be less successful (leading to longer reaction time and higher error rates) when it is low in frequency, has many neighbors (particularly when these neighbors are comparatively high in frequency) and has many confusable phonemes.

The most important deviation of NAM with other models is the concept of neighbors, which characterizes the competitor set in this architecture. Neighbor in NAM is defined as a group of "similar sounding words". More specifically, neighbors of the word "cat" include words that differ from it by only one phoneme through addition (e.g., "cast"), deletion (e.g., "at") or substitution (e.g., "kit"). Activation level of each neighbor is directly proportional to the "degree of match with the input" (Jusczyk & Luce, 2002, p.504). The activation level of a particular neighbor is then translated to the probability that it is what the input actually conveys. This probability is in turn correlated with the time required for successful recognition. Although the model provides a detailed description of factors involved in computing the probability that a particular word will be recognized, it does not mention much about the temporal dynamics during such computation. In particular, it includes no descriptions of how the unfolding signals are mapped to the lexicons. Yet, its success in correctly predicting performances on lexical decisions and naming tasks supports the validity of certain model assumptions (e.g., neighborhood size effect). This

motivates Luce, Goldinger, Auer, and Vitevitch (2000) to revise NAM to accommodate the incremental nature of spoken word recognition to make it a real processing model that take into consideration processing dynamics like competition among neighbors.

2.1.4 Current trend in model development

The most recent development of speech perception models took into account the statistical regularities of speech signal in a corpus-based manner (Protopapas, 1999). These Bayesian models of speech perception complemented the traditional approach by showing how the acoustic variability could be processed (e.g., Gaskell & Marslen-Wilson, 2002; Norris & McQueen, 2008). Units such as phonemes and syllables become less meaningful in the probabilistic context. Moreover, these models also emphasized the abundance of information in the speech signals (Chater & Manning, 2006), thereby challenging the nativist view (e.g., Chomsky, 1965), which claimed that external inputs alone would be insufficient for language acquisition. This idea explicitly linked the acquisition and processing aspects of linguistic function together. However, a prerequisite of developing this type of models would be a lot of empirical works and corpus analyses, both of which are lacking in Chinese speech processing at present. Therefore, this study serves more as an initial step in gathering useful empirical data. Concepts and terms will follow more closely the traditional approach to facilitate a comparison across different languages.

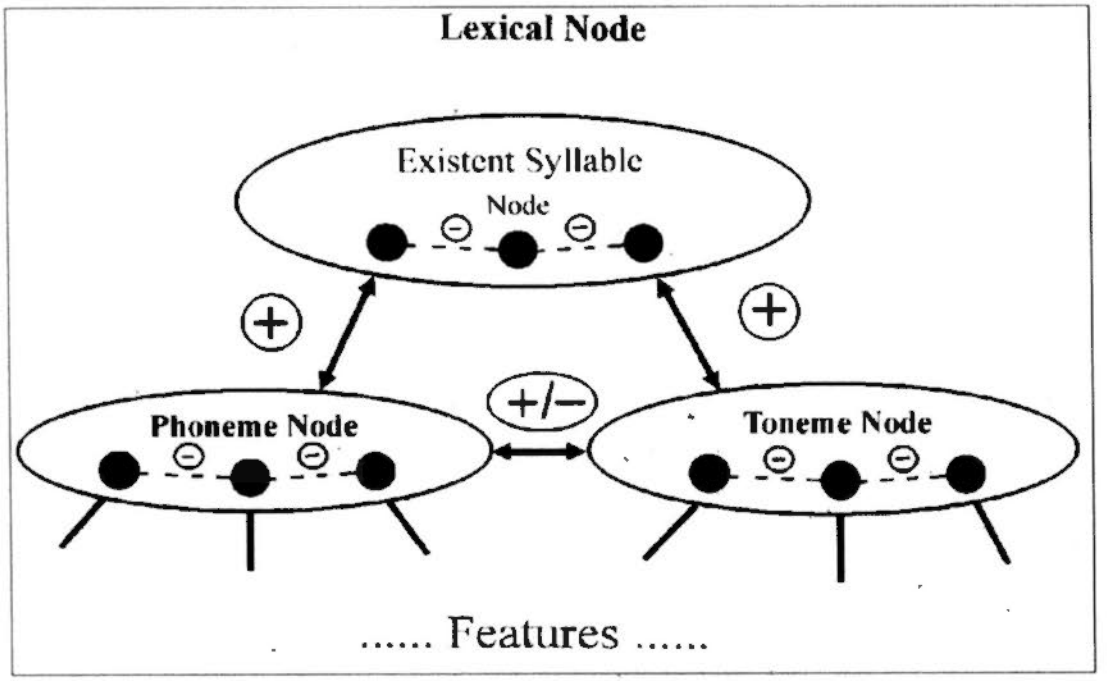
2.1.5 Chinese speech comprehension

As reviewed above, most existing theories consider phonemes or phonemic features as the fundamental input units in modeling the processing dynamics of

spoken word recognition. In contrast, the effect of suprasegmental cues on spoken word recognition is left unspecified. The over-emphasis on phonemes or phonemic features and the ignorant of suprasegmental cues in speech perception can probably be attributed to the fact that existing models of speech perception are constructed based on English and other Indo-European languages, in which suprasegmental information plays only a limited role in differentiating lexical items. Therefore, conducting research in other languages is important to extend the limitations in existing models.

Although not many works have been done on modeling Chinese speech perception, Ye and Connine's study (1999) provided an encouraging framework for further investigation. They were interested in studying how lexical tone affected Chinese speech perception. Results suggested that although Chinese speakers were in general more sensitive to segmental cues, tonal features were still important in lexical access, especially in highly constraining context. Based on these results, Ye and Connine modified the TRACE model (McClelland & Elman, 1986) to explain phenomena in Chinese speech perception. Specifically, they proposed a "toneme" layer to incorporate the unique reliance of lexical tone in distinguishing lexical items. They also replaced the word layer in the original model with a syllable layer to account for the prominence of syllable in Chinese speech. Activation in the toneme layer was graded, such that more similar tones would be confused more easily. Moreover, the toneme layer received both bottom-up feedforward inputs from acoustic features and top-down lexical feedbacks. Yet, because of the high rate of homophone in Chinese, feedback activation was in general weak unless a highly constraining context was provided. Other aspects, such as the existence of lateral inhibition, and the connection between syllable and phoneme layers, were identical to the original TRACE model. Figure 1 presents the modified model.

Figure 1 Modified TRACE model in Ye and Connine (1999)



Despite the important contributions by Ye and Connine (1999), there were several potential limitations. First, even though the importance of syllables was acknowledged, input features were still mapped onto phonemes. This fit the vowel sensitivity Ye and Connine observed, but the special status of syllable in Chinese led one to wonder whether features were also directly connected to syllables. Second, while tone is a suprasegmental cue very different from segmental ones, its treatment in the model as a toneme layer resembled phonemes and syllables. In contrast, other suprasegmental cues such as stress pattern were treated differently in the literature. For instance, stress was considered as part of the information in the "metrical frame", which was separated from processes of segmental retrieval in speech production (Levelt, Roelofs, & Meyer, 1999). Finally, although the model was intended to be a processing model of speech perception, the vowel/tone monitoring task in Ye and Connine did not provide information about candidate generation and elimination

before actual identification. It also failed to reveal the time course by which various speech units were available. This information would be essential for model evaluation and construction.

2.2 Theories of lexical ambiguity resolution

The discrepancy between the fact that many lexical items have multiple meanings and people's subjective unawareness of ambiguity has evoked considerable interest among psycholinguists. This is because, on the one hand, lexical ambiguity is a prominent phenomenon in many languages that models for language comprehension must account for. On the other hand, studying ambiguity resolution can provide important insights about the nature of human language processing system, such as whether semantic access should be considered as a modular or an interactive system (Duffy et al., 1988; Hogaboam & Perfetti, 1975; Kambe, Rayner, & Duffy, 2001; Simpson, 1981; Swinney, 1979; Tabossi, 1988; Vu & Kellas, 1999; Vu, Kellas, & Paul, 1998). More specifically, researchers usually conceptualize the extraction of the appropriate meaning in ambiguous words as involving two mechanisms. The first is a lexical-retrieval system which is sensitive to lexical properties such as meaning frequency. The second is a context-integration system that allows the selection of context-fit meanings. Although it is beyond doubt that the ultimate constraint for appropriate meaning selection is context, there are rigorous debates on when contextual effects can be seen. Researchers in the field can be divided into two groups along the modular-interactive dimension. Advocates of the modular view suggested that context only exerts a late effect after initial lexical access. In contrast, other researchers asserted that early meaning retrieval has already been constrained by word-context interaction.

2.2.1 Modular view of semantic access

Models adopting the modular view usually assume ambiguity resolution to go through a strict serial two-stage process, in which the initial stage of lexical access is automatic and encapsulated. Context can only exert effects on the product of initial access but not alter the access itself. A classical example of this type of model is the exhaustive access model (Swinney, 1979), which suggests that all meanings of an ambiguous word are activated upon initial encounter. Confirmatory evidence of this claim was obtained in studies using cross-modal priming, in which participants were presented with a spoken sentence that ended with a two-meaning ambiguous word. When a visual probe appeared immediately (0 ms SOA) after the auditory prime, facilitation was obtained for probes related to either meaning of the final ambiguous word, irrespective of the prior sentential context. Facilitation on probes related to the context-inappropriate meaning ceased when SOA is lengthened to 250 ms, a period required for context to exert effects.

Another model with a modular architecture is the ordered access model proposed by Hogaboam and Perfetti (1975). They noticed that when asked to provide definitions of words with multiple meanings, participants usually started from the most frequently used meaning (i.e., the dominant meaning). In other words, availabilities of the different meanings of an ambiguous word appeared to be ranked by their relative dominance. With this in mind, Hogaboam and Perfetti did not agree with the idea of simultaneous activation of all meanings. Instead, in their theory, only the most dominant meaning is activated during initial access. That meaning is then integrated with the preceding context. If it fits, the system will process the next word,

otherwise the system need to go back to retrieve the next most dominant meaning. The cycle continues until integration with context is successful.

2.2.2 Interactive view of semantic access

In contrast to the modular view, models with an interactive architecture do not dissect the comprehension process into two distinct stages. Rather, lexical and contextual influences are argued to operate together during comprehension. For instance, according to the selective access model (Simpson, 1981; Tabossi, 1988), context does not simply "select" the appropriate meaning out of the available candidates (as exhaustive access model implies) but can "determine" the entire meaning retrieval process such that only the context-fit meaning will ever get activated in the first place. Vu et al. (1998) further suggested that contextual effect would be strengthened if its constraint is "strong" enough, which means the context can form a coherent representation with one meaning of the ambiguous word.

Another interactive view of ambiguity resolution is the reordered access model proposed in Duffy et al. (1988). Whilst selective access model emphasizes the absolute influence of context, reordered access retains the role of lexical properties. Similar to the ordered access model (Hogaboam & Perfetti, 1975), Duffy et al. highlighted the importance of relative dominance among the different meanings an ambiguous word has. They proposed that upon encountering an ambiguous word, all meanings will be activated, with their activation levels being proportional to their relative frequency of use. In other words, the most dominant meaning will be the mostly activated one. However, a prior context can boost the activation level of the context-fit meaning so that with contextual support, a subordinate meaning can become as active as the dominant one. This implies a cost for retrieving the

subordinate meaning not because it is available late but because it is as active as the dominant meaning and thus requires extra resources to select among the equally available meanings. Supports for this subordinate bias effect came mainly from eye movement studies on reading comprehension (e.g., Duffy et al., 1988; Sereno, O'Donnell, & Rayner, 2006; Sereno, Pacht, & Rayner, 1992). The basic rationale for recording eye movements during reading is that any difficulty in meaning retrieval and integration will be directly translated into longer fixations times at the point difficulty is met (Rayner, 1998). Empirical evidence revealed lengthened reading times at the ambiguous word when previous context recruited its subordinate meaning, thereby confirming predictions made by the reordered access model.

2.2.3 Ambiguity resolution in Chinese speech

As reviewed previously, the morphosyllabic structure in Chinese, combined with its highly homophonic nature, leads to the prevalence of morphemic ambiguity in Chinese speech perception. While there are many studies on lexical ambiguity, virtually no published data concern ambiguity resolution at the morpheme level. However, as the smallest meaningful linguistic unit, meaning resolution of morpheme is non-trivial. For instance, the effective meaning resolution for the ambiguous prefix "in-" in "inside" and "invalid" or the ambiguous suffix "-er" in "beaker" and "quicker" is fundamental for successful comprehension of the whole word. Thus, theoretically, understanding its underlying mechanisms is crucial for constructing a valid model of word recognition. Moreover, this issue allows us to test whether a general mechanism for meaning resolution is applicable to both lexical and morpheme level resolutions, or there are distinct processes involved.

As a first attempt to investigate the issue, Tsang (2006) studied the recognition of Cantonese spoken disyllabic words with the visual world paradigm (Tanenhaus & Spivey-Knowlton, 1996; Tanenhaus, Spivey-Knowlton, Eberhard & Sedivy, 1995). The materials used in this study began with a homonymic syllable and ended with a disambiguating context morpheme. For example, the syllable "教" /gaau3/ is ambiguous, meaning "education" and "religion" in "教師" /gaau3si1/ (teacher) and "教堂" /gaau3tong2/ (church) respectively. All ambiguous materials used in the study were biased, meaning there was a highly frequent dominant meaning (such as "education" in "教"). In the first experiment, Tsang investigated the time course of activation of the different meanings "教" /gaau3/ contains and the temporal locus of disambiguation by the contextual morphemes "師" /si1/ and "堂" /tong2/. The results are interesting. First, the dominant meaning was always more active than the subordinate one during initial access of the ambiguous syllable. Second, after the contextual morpheme disambiguated the ambiguous syllable towards the subordinate meaning, the dominant meaning remained more active than an unrelated distracter for about 300 ms. Finally, albeit delayed and weakened compared to the dominant meaning, the subordinate meaning was always more active than a control baseline.

In the second experiment, a prior sentential context was constructed before the target disyllabic words. Prior context greatly increased the activation level of the context-appropriate meaning, especially when it is the dominant one. Therefore, the whole pattern observed in Tsang (2006) suggested that morphemes are important processing units during Chinese spoken word recognition, and the one-to-many syllable-morpheme mapping did lead to the issue of morphemic ambiguity, which should be considered in constructing models of comprehension. Most importantly, both the dominant and the subordinate meanings were always available, with a faster

and stronger activation for the dominant meaning. And the activation level of a particular meaning can be flexibly altered by a prior context. The results were overall consistent with the predictions made by the reordered access model (Duffy et al., 1988). In other words, morphemic ambiguity resolution is resolved using similar mechanisms as those employed in lexical ambiguity resolution.

However, as mentioned in the introduction, the use of a sentential context in the second experiment in Tsang (2006) precludes us from drawing a strong conclusion about morphemic level ambiguity resolution with preceding disambiguating information. It is because the constraints induced by a prior sentence and a prior morpheme may be qualitatively distinct. It is thus desirable to conduct an experiment in which the ambiguous syllable is preceded by a single contextual morpheme. Although such experiment is not possible in Tsang due to limitations in his homonymic materials, using homophonic morphemes as materials could solve this problem and allow more rigorous testing on the issue.

Chapter 3

Overview of Three Experiments

This chapter provides a general overview of the three experiments to be reported. Specifically, purposes of the experiments, backgrounds about the paradigms used, and the main hypotheses will be covered.

3.1 Aims

It is it is surprising to find so few studies on Chinese speech given the aforementioned uniqueness. Therefore, the purpose of this thesis is to construct a working model of Chinese speech perception through studying the mechanisms involved in resolving the multi-level ambiguities inherent to the speech signal. Three issues are of particular interests, including the fundamental encoding unit in Chinese speech, the processing dynamics of tone, and the resolution of homophonic morphemic ambiguity. Answers to these questions not only allow us to better understand the processing of Chinese speech but also illuminate on the unexplored areas in existing speech perception models that are based primarily on Indo-European languages. Furthermore, it provides the ideal opportunity to test the generalizability of processing principles discovered in other languages such as English or Dutch.

In particular, three experiments were conducted to study Chinese speech perception. The first two experiments aimed at offering general information about the more fundamental issues about the role of tone and other segmental units in the identification of Chinese monosyllabic words. Experiment 1 employed the gating paradigm (Grosjean, 1980), which has been a fruitful technique in exploring the basic units and mechanisms involved in speech perception (see Grosjean, 1996 for a brief review). In Experiment 2, preliminary results obtained in Experiment 1 were further

tested with the visual-world paradigm developed by Tanenhaus and his colleges (Tanenhaus & Spivey-Knowlton, 1996; Tanenhaus et al., 1995). Using the same visual-world paradigm, Experiment 3 tested how the correct meaning of homophonic morphemes, a frequent phenomenon in Chinese, could be retrieved. This experiment also served as an extension of the study of homonymic morphemes in Tsang (2006).

3.2 Paradigms

This section presents the basics for the gating paradigm and the visual-world paradigm, which were two important techniques in studying spoken word recognition.

3.2.1 Gating

The gating paradigm used in the present experiment was developed by Grosjean (1980). In a typical gating paradigm, participants will be presented with segments of a speech stimulus repeatedly. The first segment is usually very short but the presentation time (duration from onset) increases in successive pass (gate) until the entire stimulus is presented finally. After each gate, participants are asked to write down their guess to the stimulus identity, and judge how confident they are to their guess, based on the information they heard so far. Depending on the issues under investigation, procedure can be modified accordingly. For instance, the unit of segment (time, phoneme, syllable, etc.), segment size, presentation format, context availability, and response type can vary flexibly. In the present study, we followed the most traditional version including successive presentation of time segments with written responses.

The variables of interest in the gating paradigm include the number of gates that participants need to correctly recognize the target and establish high confidence.

Moreover, the tentative answers made by the participants and how their guesses changed as more acoustic inputs are gathered also provide valuable information about the trajectory of lexical retrieval of the targets. Actually, the pattern of guesses participants made before correct target identification constitutes the earliest evidence supporting the incremental nature of speech perception: Initially, participants produce diverged responses with low confidence. As they gathered more acoustic inputs, however, the diversity of guesses decreases until they finally converged on a single answer with high confidence. This pattern also fits perfectly with Cohort model's (Marslen-Wilson & Welsh, 1978) description of the candidate generation and elimination processes. Moreover, correct identification can usually be achieved before the final gate, indicating that partial information is sufficient for recognizing spoken words. All these findings provide important constraints in attempts to model the detailed dynamics of speech perception.

The gating paradigm has been validated by showing sensitivity to various well-established factors of word recognition, such as frequency and word length (Grosjean, 1980; Tyler, 1984). Robust context effect was also demonstrated with the paradigm. More importantly, the pattern of candidate generation has been argued to provide important insights on what hypothetical linguistic units are important in human speech perception. In particular, Tyler showed that a group of Dutch speakers, initial acoustic information was mapped onto consistent phonemes, suggesting that phoneme was a valid processing unit.

Given its relevance to the present research goal and the ease in administration, we employed the gating paradigm in Experiment 1 as our first step to understand the mechanisms in Chinese speech perception. It should be noted, however, that the paradigm has been criticized for involving unnatural strategies and/or contamination

from post-access processes. Nevertheless, results obtained with gating would still be valuable in guiding further rigorous test with the visual-world paradigm in Experiments 2 and 3.

3.2.2 Visual-world paradigm

The basic theoretical ground for the visual world paradigm can be traced back to Cooper's work (1974), in which participants simultaneously saw a visual display containing several objects and heard a prose passage that included the objects' names. Although participants were not explicitly requested to look at any object during listening to the speech, results indeed suggested that they spontaneously fixated on the object once it was supported by the acoustic signals. While Cooper's work did not produce an immediate impact on speech comprehension research, Tanenhaus and colleagues (Dahan & Tanenhaus, 2004; Tanenhaus & Spivey-Knowlton, 1996; Tanenhaus et al., 1995) employed this idea and developed the visual-world paradigm for online monitoring of speech processing with high temporal resolution. The typical procedure of their experiments included presenting participants with a display containing several concrete objects. At the same time the experimenter would deliver a target detection instruction (e.g., "click on the X"). Participants' eye-movements on the various objects were recorded as they performed the target detection task.

Usually, objects in the visual display are having particular relationship. For instance, in Allopenna, Magnuson and Tanenhaus (1998), objects with names sharing the same onset (e.g., "beaker" and "beetle") or rime (e.g., "beaker" and "speaker") were put together. Through a simple linking hypothesis, Allopenna et al. concluded that the fixation proportion on the various objects in the display is a direct function of the underlying lexical activation of the objects' names. Higher fixation proportion

reflects stronger activation. In other words, by tracing the changes in fixation proportion on different objects, we can obtain an effective index of relative lexical activation over time[3]. It is similar to a "continuous" version of gating, in which the availability of different candidate words in the participants' mind at different times is probed by the fixation proportions, rather than by stopping the signal and asking the participants to write down what they have in mind. In addition, later results confirmed the validity of the paradigm in reflecting the time course of activation of various candidate words given the inputs received so far. Such activation is a genuine lexical event involving conceptual information of the target, rather than simply matching a pre-activated phonological form with the incoming acoustic signals (Dahan & Tanenhaus, 2005; Huettig & Altmann, 2005).

Employing the visual-world paradigm, various issues about speech perception have been studied. For example, Dahan, Magnuson and Tanenhaus (2001) investigated the time course of the frequency effect in spoken word recognition. They found a higher activation level for words with higher frequency at the earliest possible moment, ruling out a postlexical decision bias interpretation. In another line of studies, researchers (Altmann & Kamide, 1999; Kamide, Altmann, & Haywood, 2003) were interested in the effects of sentential constraints. Results indicated that the context effect was so strong that even a single verb could affect the activation level of the subsequent noun. For example, activation of "cake" rose much more quickly following the verb "eat" than "move", presumably because typically people think of a cake as edible rather than movable. In short, the paradigm had been validated by showing sensitivity to both frequency information and contextual constraints. Moreover, results from Allopenna et al. (1998) also indicated that the paradigm could

[3] It was estimated that about 200 ms is necessary to program and execute a saccade. Therefore, at any time point the fixation is reflecting the activation level 200 ms ago.

be used to investigate the basic units of candidate generation and elimination in speech perception. When objects in the display shared an important processing unit, hearing the name of one object would lead to fixations on the other because both were receiving supporting bottom-up evidence and would be included in the candidate sets. For example, if the onset was an important unit in candidate generation, hearing "beaker" would also activate the onset-sharing word "beetle" and lead to more fixations on it. Otherwise, fixation proportions on "beetle" would be identical to an unrelated distracter (sharing nothing) such as "dolphin".

Using the visual-world paradigm to further investigate the results obtained in Experiment 1 has several advantages. First, the paradigm makes use of the natural human tendency to look at the objects in the visual field as they are mentioned in speech. As eye movements are continuous and fast, this provides high temporal resolution information necessary to reveal the changes in activation level of various candidate words. Second, interpretation is easy because the activation level of a specific lexical unit can be inferred directly from the fixation data (Allopenna et al., 1998). Finally, the paradigm allows continuous monitoring of the relative activation levels of different meanings over the whole time course of word recognition. This is much more natural than gating or other priming procedures.

3.3 Major hypotheses

Consistent with the findings in Indo-European languages, studies in Chinese speech perception (e.g., Li & Yip, 1998; Schirmer et al., 2005) have shown that it is a highly incremental process involving candidate generation and elimination. However, given the unique properties of Chinese speech compared with other Indo-European languages, the incremental nature might not be realized in the same way in different

languages. Specifically, we hypothesized that Chinese speakers would rely on lexical tone in generating hypothetical candidates before correct identification. In other words, we expected to observe participants writing down word candidates that shared only tone with targets in the gating paradigm. We also expected to see higher fixation proportion on objects sharing tone with targets than objects sharing nothing with targets in the visual-world paradigm.

Moreover, given the salience of syllable in Chinese speech, we hypothesized that the acoustic information was mapped directly onto whole syllable rather than individual phonemes in the course of speech perception. In this case, participants would rely less on subsyllabic information in candidate generation. They would not produce candidates that shared only phonemes but not the whole syllable in the gating task. In addition, participants would be fixated more on syllable-sharing competitors than on competitors sharing on subsyllabic units.

Finally, Tsang (2006) proposed that the resolution of homonymic morphemic ambiguity in Chinese speech followed closely the prediction of reordered access model. In particular, both relative meaning frequency and context exerted strong influence in retrieving the correct meaning. We expected the same pattern would be true in resolving homophonic morphemic ambiguity. Without a prior context, the dominant meaning would be more available than the subordinate meaning initially until the disambiguating morpheme helped activate the subordinate meaning. In contrast, when there was a preceding contextual morpheme, the subordinate meaning could be available immediately upon encountering the ambiguous homophone.

Chapter 4

Experiment 1 – Recognition of Chinese Syllables

Two gating experiments were conducted to provide a brief overview of Chinese speech perception. Specifically, they served to replicate in Chinese speech several fundamental properties found in perceiving Indo-European speech, such as its incremental nature, the recognition based on partial information, and its sensitivity to context. Furthermore, by analyzing the types of error made by our participants in the experiments, we could obtain a general idea of whether the unique characteristics of Chinese speech could produce significant impacts on speech perception.

4.1 Experiment 1A – recognizing isolated monosyllables

In Experiment 1A, we employ the gating paradigm (see Grosjean, 1980; 1996) in testing how Chinese speakers recognized isolated monosyllabic words as increasing amount of acoustic information was provided. First, based on the results from previous gating experiments in English, we expected our participants to generate various response candidates when only limited information was given. However, with more information, participants would converge on a single "correct" interpretation. Moreover, such convergence would occur before the whole syllable was available. Second, if Chinese speakers are still sensitive to subsyllabic units despite the prominence of syllables, they should produce a significant amount of onset-sharing and rime-sharing errors during the course of recognition. On the other hand, if syllables are the prominent processing units in Chinese speech, participants would be as likely to commit whole-syllable errors. Finally, given that lexical tone was found to be an important feature in Chinese speech processing in previous research (Li & Yip, 1998; Ye & Connine, 1999), we expected to observe candidate

generation based solely on tone-sharing. In other words, there would be a significant amount of errors with intact lexical tone.

4.1.1 Participants

Twenty undergraduates (12 males) in The Chinese University of Hong Kong participated in the experiment. All of them were native Cantonese speakers and none reported hearing deficits. They were paid $50 for participation. Informed consent was obtained and full debriefing was delivered after the experiment.

4.1.2 Materials and Design

Twenty-four Cantonese monosyllabic words were prepared (see Appendix A). To facilitate the direct comparison across tasks, identical materials were used in the present gating experiment and Experiment 2A. Therefore, the words were chosen on the basis of satisfying the constraints of the visual-world paradigm. In particular, each target monosyllabic word could be paired with an onset-shared competitor (O), a rime-shared competitor (R), a syllable-shared competitor (S), and a tone-shared competitor (T). For instance, the target /bou3/ (布; cloth) was paired with /bui1/ (杯; cup), /dou1/ (刀; knife), /bou1/ (煲; pot), and /coi3/ (菜; vegetable) for the four conditions respectively. Although we only tested participants' responses towards segments of the target syllables in the gating experiment, it should be noted that all items were picturable concrete nouns for the presentation in visual-world. Moreover, only the tone-shared competitor had identical tone to the target, while the other competitors shared another lexical tone. Finally, according to the Chinese Character Database developed by Research Centre for the Humanities Computing (http://humanum.arts.cuhk.edu.hk/Lexis/lexi-can/), items across conditions were

closely matched for log-frequency (log-frequency 2.84, 2.75, 2.82, 2.93, and 2.93 for the target, O-, R-, S-, and T-conditions respectively; $F_{(4, 92)}$ .261, *n.s.*).

The stimuli were recorded by a female native Cantonese speaker who was naïve to the details of the experiment. She had received training in phonetics and was teaching relevant courses in kindergartens. However, she was instructed to produce the target monosyllables naturally to resemble daily conversation. Recording was made on a DAT tape in a sound-proof chamber. Each word was recorded three times and the best token (in terms of clarity and naturalness) was selected for digitalization at 44.1kHz. All digitalized materials were manipulated with Cool-Edit Pro (Syntrillium Software). Peak amplitude was normalized to 70dB. Each selected token of the target monosyllables was then segmented into gates with 40ms increment size, until the last gate contained the whole syllable. For example, there were 13 gates for the syllable /bou3/, which was 494 ms long (40 ms, 80 ms, 120 ms, 160 ms, 200 ms, 240 ms, 280 ms, 320 ms, 360 ms, 400 ms, 440 ms, 480 ms, and 494 ms). The durations and the number of gates for each target word are listed in Appendix A.

4.1.3 Procedure

Participants were tested in groups of five in a sound-proof room. They were told that in each trial, they would hear segments of a Cantonese monosyllable, one at a time with increasing segment size (i.e., successive presentation). In particular, the first segment contained the initial 40ms of the target syllable. The second segment included the first gate plus an additional 40ms, and so on, until the whole syllable was presented. Participants were instructed to try their best to identify the target monosyllabic word based on the segment they heard, write down their answer, and rate their confidence level about their "guess" on a 10-point scale (10 = very

confident). They were encouraged to write down the first word that came to their mind, and if they were not sure about how to write the word in mind, they could replace it with a homophone or write down a contextual character for disambiguation. Presentation of materials was via speakers connected to a personal computer and was controlled by an experimenter. There were 412 segments in total. After each segment, participants were given about 10 seconds to respond. The whole experiment lasted for about an hour.

4.1.4 Results and Discussion

The target syllable /jyu5/ (雨; rain) was eliminated from further analyses because no participants could recognize it correctly[4]. Three variables were particularly relevant to the present research questions. First, we calculated the amount of information needed for identifying the target monosyllabic word (or its homophones) after excluding error trials. In this study, the "recognition point" was defined as 1.) the gate at which a correct identification was made with confidence rating of 8 or above; or 2.) the intermediate gate in which the correct syllable was written down for three consecutive segments regardless of confidence level. The second criterion was included because some participants were very persistence in their "guesses" despite their subjective feelings of uncertainty. The two criteria converged most of the time. If they conflicted, the criterion that led to an earlier estimation was taken[5]. The gate of recognition of a particular syllable was divided by its total number of gates to estimate the amount of information needed for recognition.

[4] Re-examination of the segments suggested that a wrong token was selected for this syllable in gating. This token was replaced with the better one in Experiment 2, which improved the recognition accuracy of the whole syllable to 97%

[5] In 1% of all trials, participants produced the correct response but conform to neither criteria (e.g., they only wrote down the same syllable for two consecutive trials with low confidence). These trials were coded as missing and were not analyzed further.

Consistent with studies in Indo-European languages (e.g., Grosjean, 1980; Marslen-Wilson, & Welsh, 1978), participants could recognize the Cantonese monosyllables when only partial information was given (66.39% of the whole syllable, range = 50.55 to 84.72; $t_{(22)}$ = 14.80, p < .01, one-sample t-test against 100).

Next, for trials in which participants failed to identify even when the whole syllable was presented (i.e., errors in final responses), we analyzed the type of errors they made. Table 1 shows the distribution of different types of error with examples.

Table 1. Distribution of different types of error in Experiment 1A.

Type of error	Example	% of error
No-sharing	faan6 → baak1	1.30
Onset-sharing	coeng4 → cat1	3.48
Rime-sharing	bou3 → dou1	2.39
Syllable-sharing	bou3 → bou1	20.00
Tone-sharing	fu3 → si3	1.52
Onset-plus-tone-sharing	leoi6 → lou6	2.83
Rime-plus-tone-sharing	tai4 → cai4	2.17

The total error rate was 33.70%, which indicated that the recognition of isolated Cantonese monosyllables was indeed difficult. Inspecting Table 1, it should be clear that participants confused syllables with different tones most easily ($F_{(6, 132)}$ = 8.99, MSE = 0.012 p < .01). Tonal error was much more prominent than any other types of error in protected LSD test (ts = 3.07 to 3.51, all ps < .01). Moreover, only a few (1.52%) errors were generated based on tonal-sharing only, arguing against our expectation that tone alone was enough to create lexical candidate activation. Table 2 displays the exact distribution of tonal errors when participants produced syllable-sharing responses. This is essentially a confusion matrix which provides details about which tones are more easily confused.

Table 2. Confusion matrix of tonal errors in Experiment 1A.

		Correct response					
		Tone 1	Tone 2	Tone 3	Tone 4	Tone 5	Tone 6
Actual response	Tone 1	-	0	24	2	0	6
	Tone 2	0	-	0	3	0	0
	Tone 3	2	0	-	18	0	15
	Tone 4	0	0	0	-	0	0
	Tone 5	0	7	0	7	-	0
	Tone 6	0	1	5	0	1	-

The confusion matrix clearly shows that there seems to be more errors for target monosyllables with Tone 3 and Tone 4. Some observations could be made. First, the three level tones (1, 3, and 6) seemed to be highly confusable. Second, the two rising tones (2 and 5) are also confusable. Third, the falling tone (4) was difficult to recognize and participants did not prefer to give Tone 4 responses.

The distribution of errors actually resembled the pattern observed in Cutler and Chen (1997). In one of their experiments, they asked participants to decide whether two successively presented Cantonese monosyllables were identical. In the mismatching trials, they found that when the syllables differed only in tone, participants' error rate was much higher than in conditions where the difference was on other dimensions. In other words, the present gating result converged with previous evidence and argued against the constraining power of lexical tone in perceiving spoken Chinese.

To provide further information about whether tonal constraint was utilized in recognizing the isolated monosyllables, we inspected the word candidates participants generated before they arrived at the correct interpretation. Results suggested that in 13.91% of trials, participants had generated at least one candidate word that shared with the actual target only on tone (e.g., /sat6/ as /faan6/; /ci2/ as /sing2/) during the

course of identification. However, given that there were six tones in Cantonese, this value was actually no better than chance ($t_{(22)}$ = .73, *n.s.*, one-sample *t*-test against 16.67%)[6]. This again suggested that lexical tone might not be particularly constraining in recognizing isolated monosyllables.

There were also cases in which candidates sharing rime were proposed (e.g., saan1 as 'faan6'; 'sau6' as 'zau6'). Yet, most of the times, participants generated onset-shared candidates during recognition. For instance, a typical response profile towards 'coeng4' (牆; wall) would include initial "guesses" such as /cyut3/, /ceot1/, coek3', and 'coeng3' before the correct interpretation arrived finally. This reliance of word onset in generating hypotheses about what was being heard supported the idea that Cantonese speakers were also sensitive to subsyllabic information. Moreover, it was consistent with studies in Indo-European languages showing the special role of word onsets in spoken word recognition (e.g., Allopenna et al., 1998; Marslen-Wilson, & Zwitserlood, 1989). Such onset-sensitivity made sense in gating because the initial portion of target words was all participants could base in responding during early gates. This issue would be covered again in Experiment 2A.

On the other hand, sensitivity to subsyllabic units should not be taken as evidence against the role of whole-syllable in Chinese speech perception. Actually, the prevalence of syllable-sharing errors suggested that syllables were also highly prominent. To test whether syllable representation existed over and above onset and rime, a two-way Analysis of Variance (ANOVA) with onset-sharing and rime-sharing as two within-item factors was conducted on the proportion of errors (Figure 2A)[7]. We obtained a significant interaction ($F_{(1, 22)}$ = 8.93, *MSE* = 0.015, $p < .01$) such that

---

[6] Obviously, it was just an estimation because not all tones were equally productive.

[7] We excluded the tone-sharing cases in order to observing the pure effect of segmental information. In other words, we only included syllable-sharing (tonal) errors, onset-sharing (rime-plus-tone) errors, rime-sharing (onset-plus-tone) errors, and no-sharing (complete) errors in the analysis.

there were more syllable-sharing errors ($t_{(22)}$ = 3.10, 3.35, and 3.51 when compared with onset-sharing, rime-sharing, and no-sharing conditions, all *p*s < .01). The significant interaction suggested that syllable was more than just the combination of onset and rime.

Figure 2A. Proportion of errors as a function of onset-sharing and rime-sharing.

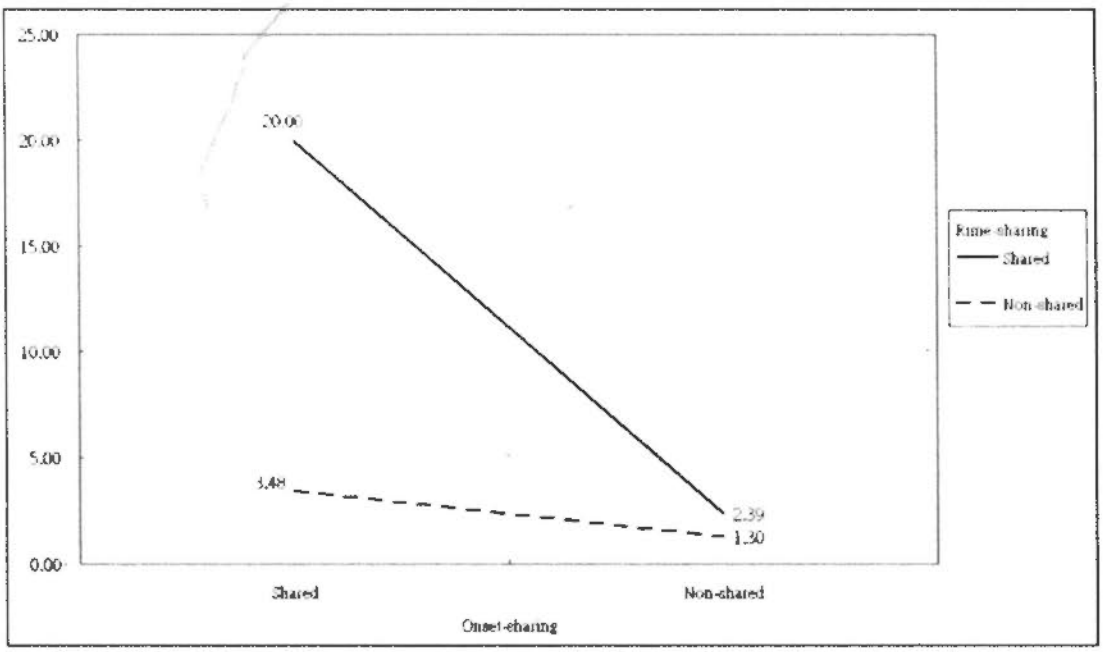


Figure 2B. Proportion of errors as a function of onset-sharing and tone-sharing.

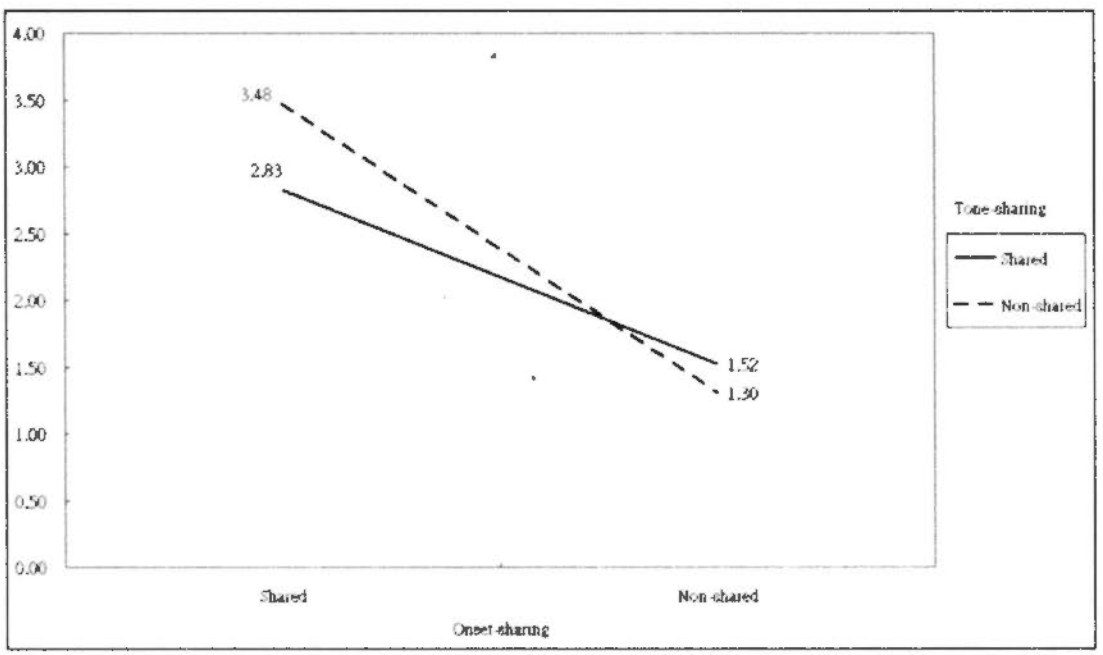
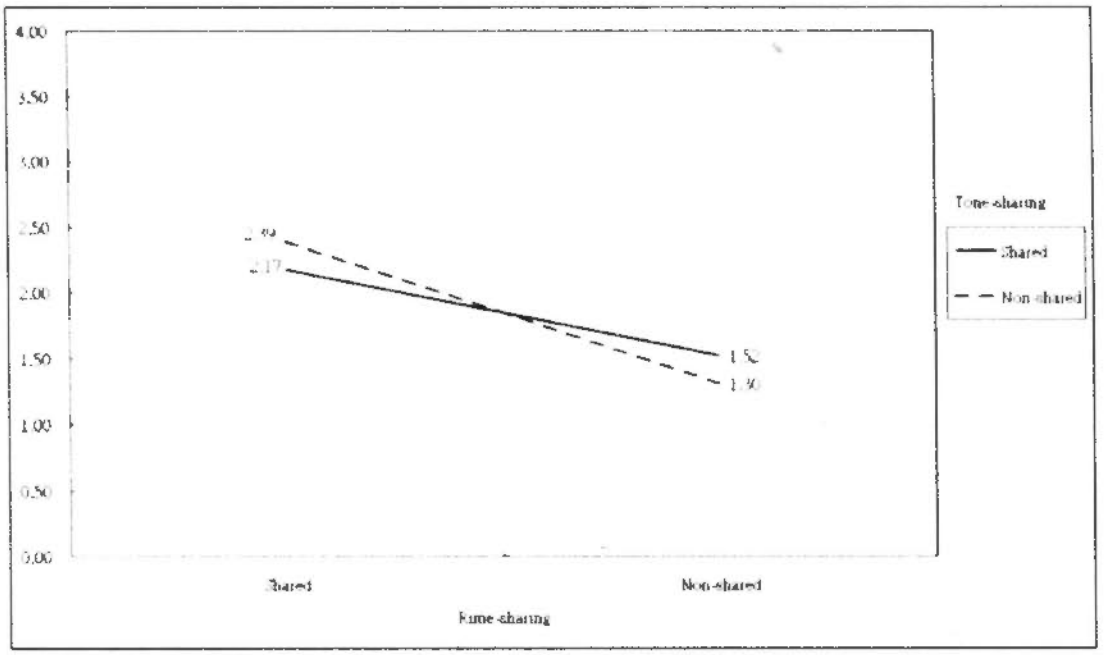


Figure 2C. Proportion of errors as a function of rime-sharing and tone-sharing.



We also conducted similar two-way ANOVAs crossing onset-sharing/rime-sharing with tone-sharing (Figure 2B to 2C). None of the effects approach significance ($Fs_{(1, 22)} < 1$).

4.2 Experiment 1B – recognizing monosyllables in context

Although researchers differed in their views about when contextual effects emerged, it is generally agreed that context must at some points exerts effects on lexical access. However, virtually all previous research on contextual effect concerned how a preceding sentence constrained recognition. It is still unclear whether contexts as short as a single morpheme can also influence succeeding meaning retrieval in speech perception. Actually, in the classical gating study, Grosjean (1980) demonstrated that short contexts facilitated recognition less than long contexts. On the other hand, Tsang's result (2006) suggested that a single morpheme was enough to disambiguate another homonymic morpheme. In Experiment 1B, we intended to verify Tsang's finding in homophonic morphemes, which is the more general case of the previously tested homonyms. Two factors were factorially

manipulated, namely the position of contextual morpheme and the relative dominance of the intended homophone. The main purpose was to see if a morphemic context could facilitate syllable identification in gating. Furthermore, we compared the recognition performance across conditions and contrasted the patterns observed with the processing dynamics predicted by various ambiguity resolution models.

4.1.1 Participants

Twenty-four undergraduates (12 males) in The Chinese University of Hong Kong participated in the experiment. All of them were native Cantonese speakers and none reported hearing deficits. None of them had participated in Experiment 1A. They were paid $50 for participation. Informed consent was obtained and full debriefing was delivered after the experiment.

4.2.2 Materials and Design

The present experiment adopted a 2 (context position: preceding or succeeding) X 2 (meaning frequency: dominant or subordinate) factorial design. Again, the materials in Experiment 1B were identical to those in Experiment 3 to facilitate comparison across tasks. Twenty pairs of homophonic syllable were selected such that each could form picturable disyllabic concrete nouns in the four conditions (SD: succeeded context-dominant meaning, SS: succeeded context-subordinate meaning, PD: preceded context-dominant meaning, PS: preceded context-subordinate meaning; see Appendix B).

Moreover, we conducted a series of pilot testing to ensure that the materials in the four conditions were properly matched. First, the dominant meaning of the syllable should be more frequent than the subordinate meaning. However, the written

character log-frequency (http://humanum.arts.cuhk.edu.hk/Lexis/lexi-can/) between the two conditions was significant only in one-way t-test ($t_{(19)} = 1.85$, $p < .05$). Therefore, to validate the meaning frequency manipulation, 20 students who did not participate in any main experiments were presented with the twenty syllables in isolation. They were instructed to write down the first monosyllabic word that came to their mind after hearing the syllables. If they were not sure, we presented the syllable again until it was clear to them. Results suggested that the dominant meaning (mean = 54%, S. D. = 25%, range = 10% to 86%) was always written down at least twice as frequently as the subordinate meaning (mean = 11%, S. D. = 12%, range = 0% to 24%). Paired-sample *t*-tests indicated a significant difference between frequencies of the two meanings ($t_{(19)} = 6.82$, $p < .01$).

Second, written character log-frequency for the contextual morphemes (http://humanum.arts.cuhk.edu.hk/Lexis/lexi-can/) was matched across conditions ($Fs_{(1,19)} = .51$ to 1.62, all *n.s.*). Finally, because there was no available word frequency corpus in Hong Kong undergraduates, another independent group of 20 participants rated their subjective familiarity to the words on a 6-point Likert scale (6 = highly familiar). Again, there were no significant differences observed (mean = 3.98, 3.77, 4.01, and 3.86, S. D. = 0.98, 0.90, 0.86, and 0.57, for SD, SS, PD, and PS conditions respectively, all $Fs_{(1,19)} < 1$). Any effects observed in the present experiment were thus irrelevant to unmatched word level familiarity.

Recordings were done by the same female speaker in Experiment 1A. We recorded both the isolated monosyllabic homophones (for the present gating) and the disyllabic words for the four conditions (for Experiment 3). Other details of recording and editing were identical to those in Experiment 1A. Appendix B also lists information about the durations of these stimuli.

4.2.3 Procedure

The procedure was identical to that in Experiment 1A, with a few exceptions. First, for every trial, a contextual morpheme was given in the answer sheet on which participants wrote their "guesses" towards the syllable segments they heard. Context position was indicated by leaving a space before or after the context for participants to write down their responses. Participants were told that the context "might help them to identify the speech segments" but they were not requested to write down words that had to fit the context. Second, each syllable was presented only once to each participant. Therefore, the eighty possible target words were divided into four lists such that within list, the same syllable occurred only in one condition and there were equal number of syllables in each condition; across list, the same syllable occurred in all conditions. Six participants completed each list as a group. Each of them received a total of 337 gates. They were given about 10 seconds to respond after each segment. The whole experiment took about an hour.

4.2.4 Results and Discussion

In 2.29% of all trials, participants wrote down the correct response but conform to neither criterion of correct recognition (criteria identical to those in Experiment 1A). These trials were excluded from further analyses. In 6.25% of trials (no difference across conditions, $Fs_{(1,19)} = .07$ to 1.49, all *n.s*), participants produced a correct homophonic response which was not consistent with the contextual morpheme. However, they indicated that they simply forgot how to write the intended character. Therefore, these responses were considered as correct. Table 3 presents the

proportion of gates needed for correct recognition and the distribution of error types across the four conditions for the remaining data.

Table 3. Mean proportion of gates for recognition and types of error in Experiment 1B.

	SD	PD	SS	PS
% of gate needed for recognition	44.77	43.82	44.90	47.39
No-sharing	0.21	0.00	0.00	0.21
Onset-sharing	0.42	0.42	0.83	0.42
Rime-sharing	0.42	0.00	0.00	0.21
Syllable-sharing	0.21	0.21	1.88	1.04
Tone-sharing	0.00	0.00	0.00	0.00
Onset-plus-tone-sharing	0.00	0.62	0.00	0.21
Rime-plus-tone-sharing	1.04	0.21	0.21	0.21

Note: SD = succeeded context-dominant meaning; SS = succeeded context-subordinate meaning, PD = preceded context-dominant meaning, PS = preceded context-subordinate meaning.

Similar to Experiment 1A, participants could identify the target syllable with only partial information. Although correct identification was apparently faster when context was available (mean proportion of gate for recognition = 66.38% and 44.69% for Experiments 1A and 1B respectively), we would not elaborate on this further because the two experiments employed different stimuli. Yet, it should be noted that facilitatory context effect on spoken word recognition has been documented in similar gating experiments (e.g., Grosejan, 1980).

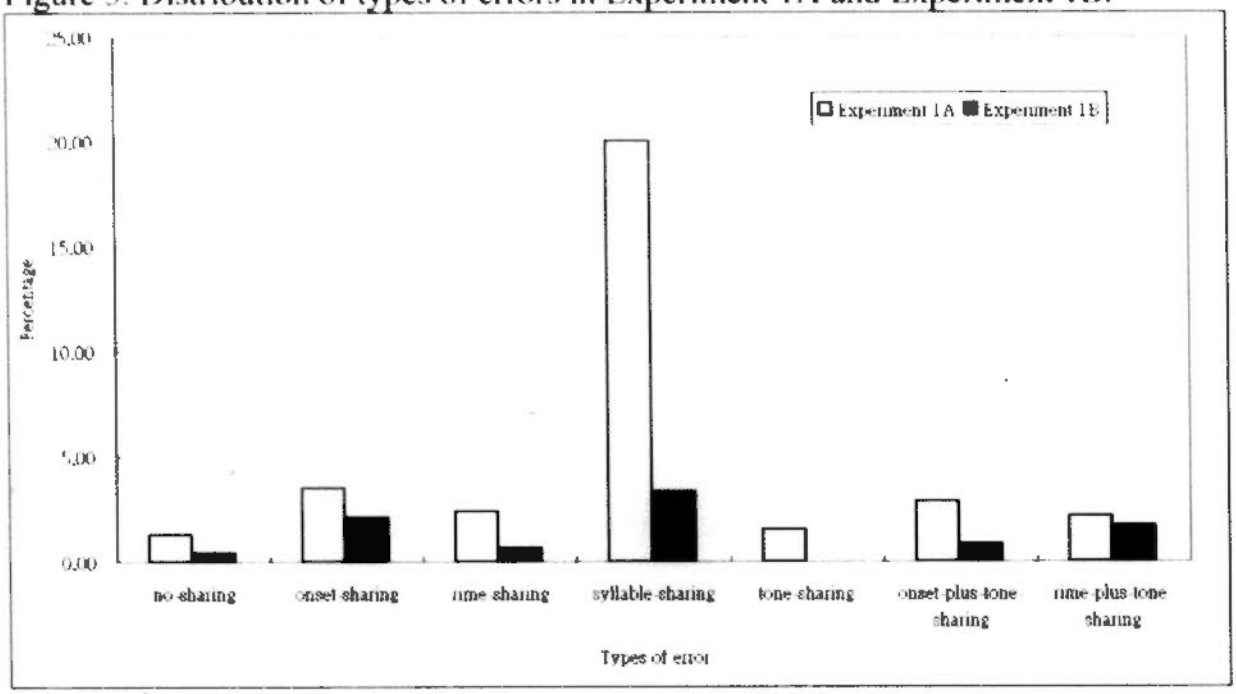
Furthermore, there were no significant differences among the four conditions ($Fs_{(1,19)} < 1$, all *n.s*). In other words, in the present gating experiment, the mere presence of a morphemic context could facilitate recognition, regardless of whether the context was available early or late, and whether the more frequent or the less

frequent meaning was to be retrieved. This pattern seems to fit well with the selective access model of ambiguity resolution (Simpson, 1981; Tabossi, 1988) which assumes an absolute influence by context. In contrast, it is inconsistent with models that include effects of meaning frequency such as ordered access model (Hogaboam & Perfetti, 1975) and reordered access model (Duffy et al., 1988; Sereno et al., 2006). However, the pattern might be attributed solely to the loose time constraint of responding in a gating task. We will return to this issue in Experiment 3.

The overall error rate of Experiment 1B (8.96%) was much lower than that in Experiment 1A (33.70%, $t_{(41)}$ = 4.16, p < .01). Also, there were no significant differences across conditions ($Fs_{(1, 19)}$ = .02 to 1.28, all *n.s*). These results again supported the mere presence of context was enough to facilitate syllable recognition in gating experiments (Grosjean, 1980; 1985). Furthermore, comparing the distributions of error types across Experiment 1A and 1B, it is obvious that the largest discrepancy lies on the amount of tonal errors (Figure 3). Indeed, the interaction between experiment and type of errors was highly significant ($F_{(6, 246)}$ = 5.26, *MSE* = 67.74, p < .01). Post-hoc pairwise comparison suggested that only the proportion of tonal errors differed across experiments ($t_{(41)}$ = 3.00, p < .001). Therefore, in contrast to Experiment 1A, tone was actually quite constraining in Experiment 1B. Interestingly, in a recent experiment, Schirmer et al. (2005) discovered that when prior context was available, lexical tone could be as constraining as segmental information (rime) in lexical activation. In their experiment, participants heard a series of sentence and completed a congruity judgment. In some sentences, a correct monosyllabic word (e.g., 'beng6; 病; illness) was replaced by a mismatched word. Results indicated that a tone-mismatch word (e.g., 'beng2; 餅; biscuit; when the actual target was) produced an N400 of comparable size and latency to the one

generated by a rime-mismatch word (e.g., /bou6/; 步; step) in sentence comprehension. Therefore, Chinese speakers appeared to be sensitive to lexical tone in context (Experiment 1B; Schirmer et al., 2005) but not when the stimuli were presented in isolation (Experiment 1A; Cutler & Chen, 1997).

Figure 3. Distribution of types of errors in Experiment 1A and Experiment 1B.



Inspecting the candidate words participants generated before they reached the final answers indicated that the faster convergence in Experiment 1B was not simply due to expectation of the target based on the contextual morpheme given. If they did, all their responses would have fit the context very well. This was not the case, however. Rather, resembling the pattern in Experiment 1A, participants produced many onset-shared candidates when limited information was given. For instance, participants produced candidates such as /luk6/ and /laap6/ before the correct identification /laam5/, even though they could not combine meaningfully with the context (see Tyler, 1984 for similar findings). However, participants needed less information for correct identification in Experiment 1A, showing that context might act to speed up candidate convergence. In other words, under the present gating

procedure, bottom-up acoustic signal cooperated with top-down contextual bias to constrain lexical activation (see Christiansen & Chater, 1999; Gaskell & Marslen-Wilson, 2001 for theoretical significance for such cooperation).

4.3 Summary of Experiment 1

In two gating experiments, we successfully replicated in our Chinese speakers major findings of speech recognition observed in Indo-European languages with similar methodologies. First, partial information (about 50 to 60% of total acoustic inputs) would be sufficient for correct recognition. In addition, before successful recognition, they generated candidates that fit the acoustic information available so far (i.e., onset-fit words). This illustrated that Chinese speakers also conform to the basic principle of "active" speech processing: People will not just sit and wait until all speech signals are available (e.g., Allopenna et al., 1998; Grosjean, 1980). Rather, they actively generate possible hypothetical words based on what they have and test these hypotheses against the incoming information. This continuous or incremental processing appears to be a universal characteristic of speech perception. After all, given the limited capacity of short term memory (e.g., Miller's magical number seven, 1956), it would be highly risky to hold a copy of the speech signal and delay immediate processing. Also, in the acoustic signal there is actually no clear syllable boundary. It is thus not obvious when listeners can stop waiting and start processing. A more active processing dynamics may therefore be the more efficient alternative to passive waiting.

Second, in both Experiments 1A and 1B, our participants did not commit whole-syllable errors significantly more. In contrast, analyses on the distribution of error types suggested that most of the time, the errors made would share either onset

or rime with the actual target. Moreover, the profile of candidate generation before correct recognition also indicated sensitivity to word onsets. These data converged on the significant role of subsyllabic unit in Chinese speech perception, despite the observed prominence of syllabic unit in Chinese reading acquisition (Cheung et al., 2001; McBride-Chang et al., 2004) and Chinese speech production (Chen, 2000; Chen et al., 2002). Actually, latest finding (Wong & Chen, 2008) also supported the role of subsyllabic units such as rime in Chinese speech production. It is unclear why such discrepancy exists; tasks, materials, and procedure probably jointly contribute. In any case, it seems that the role of subsyllabic unit should not be overlooked in studying Chinese speech processing.

On the other hand, it is also unlikely that syllable is completely irrelevant to Chinese speech perception because there were more errors sharing the whole syllable (onset-plus-rime; i.e., tonal errors) than the sum of errors sharing onset (i.e., rime-plus-tone errors) or rime alone (i.e., onset-plus-rime errors; see Experiment 1A). This interactive nature supported a role of syllabic unit over and above subsyllabic ones. In other words, both syllabic and subsyllabic representations exist in processing Chinese speech.

Finally, while lexical tone is a distinctive feature in Chinese speech that helps disambiguate between otherwise identical syllables (e.g., /bou1/ and /bou3/), its functional role seems to be qualitatively different from that of segmental units like onset and rime. Such a conclusion is inconsistent with Sum (2003), in which a primed auditory lexical decision task was used. In the present experiment, Cantonese speakers did not appear to use tone to generate possible candidates before correct identification like they did with onset. Moreover, without context, lexical tone could only play minimal role in constraining lexical access, whereas onset and rime were

still highly effective cues (Experiment 1A). Indeed, there were a large amount of tone errors. Also, the difference between proportions of onset-error and onset-plus-tone error, and that between rime-error and rime-plus-tone error, were not statistically significant.

Yet, we cannot completely dismiss the role of lexical tone in Chinese speech perception because when context was available, it became as important as segmental information (Experiment 1B; Brown-Schmidt & Canseco-Gonzalez, 2004; Schirmer et al., 2005; Ye & Connine, 1999). Given that fundamental frequency of speech (F0, the physical realization of lexical tone) varies according to gender, age, and/or emotion, generating candidates based on the perceived pitch level is likely to be misleading, especially when there are not enough acoustic context available for normalization (e.g., Moore & Jongman, 1997; Peng, 1997). Therefore, we speculate that lexical tone is playing less a generative role and more as a selective bias in the dynamics of Chinese speech perception: When enough contextual information is available, it acts as a secondary source of constraints to eliminate candidates that fit the segments but violate the tonal envelop required by context.

It should be noted that the successive presentation procedure adopted in the present gating experiments might have produced a conservative estimate of the recognition point (Walley, Michela, & Wood, 1995). Yet, this will simply reinforce our proposal that partial information is sufficient for successful recognition. Also, the conclusions based on error types and candidates generation are independent to the estimation of recognition point. Therefore, the particular presentation parameters used here are unlikely to contribute to the effects observed. On the other hand, given that participants were asked to pause and write down their responses after hearing each segment, the gating procedure itself appears to encourage post-access processing

and response strategies (but see Grosjean, 1996, for counter-arguments). Specifically, participants may be more sensitive to word onset because it has been presented repeatedly with increasing segment sizes. Also, the long lag between hearing the segment and deadline of responding could have allow participants to resolve ambiguity completely before writing down their responses, thus masking the more dynamical details of the resolution process. Given these concerns, it is desirable to replicate Experiment 1 with a more natural, sensitive, and online paradigm. In Experiments 2 and 3, we studied the same issues addressed in Experiment 1 with the recently developed visual-world paradigm (Tanenhaus & Spivey-Knowlton, 1996; Tanenhaus et al., 1995).

Chapter 5

Experiment 2 – Fundamental units in Chinese speech comprehension

Three experiments employing the visual-world paradigm were conducted to provide a detailed account about the fundamental processing units in spoken Chinese. In particular, we first tested the psychological significance of various hypothetical linguistic units such as onset, rime, tone, and syllable in Chinese speech perception (Experiment 2A). Next, given the special role of word onset as hypothesized in different models of speech processing, we focused more specifically on the role of onset in Experiment 2B and 2C. Data gathered in these experiments would create the basis on which formal models of Chinese speech comprehension was constructed.

5.1 Experiment 2A – the role of onset, rime, syllable, and tone

Experiment 2A served as a rigorous test for the relative importance of onset, rime, syllable, and tone in Chinese speech perception. Although in Experiment 1, we demonstrated that both onset and syllable were important units, while the role of rime and tone appeared to be unstable and secondary, the controversies over the validity of the gating paradigm precluded a strong conclusion on the issue. To gather additional data on the basic processing units in Chinese speech, we presented participants with the same monosyllabic words used in Experiment 1A and simultaneously tracked their eye-movements over several pictures. Actually, a previous study by Allopenna et al. (1998) showed that the visual-world paradigm could detect robust effects due to rime-sharing, while gating was only sensitive to word onset. Given that the acoustic details of rime and tone were available at a similar time point, visual-world might well be more sensitive to tone-sharing than the conventional gating procedure. Therefore, we might be able to observe candidate generations based not only on onset

and syllable, but also on rime and tone in Chinese speech perception with the more sensitive visual-world paradigm. In other words, as participants heard the target monosyllables and decided whether a concurrent visual display contained the presented word, we expected to see more fixations on the competitors that shared onset, rime, syllable, or tone with the target than distracters that shared nothing, which indicated that these competitor candidates were activated and considered during the course of spoken word recognition.

5.1.1 Participants

Thirty-two undergraduates (12 males) in The Chinese University of Hong Kong participated in the experiment. All of them were native Cantonese speakers with no known hearing-deficits and had normal or correct-to-normal vision. None of them had participated in the previous experiments and pilot tests. They were paid $50 for participation. Informed consent was obtained and full debriefing was delivered after the experiment.

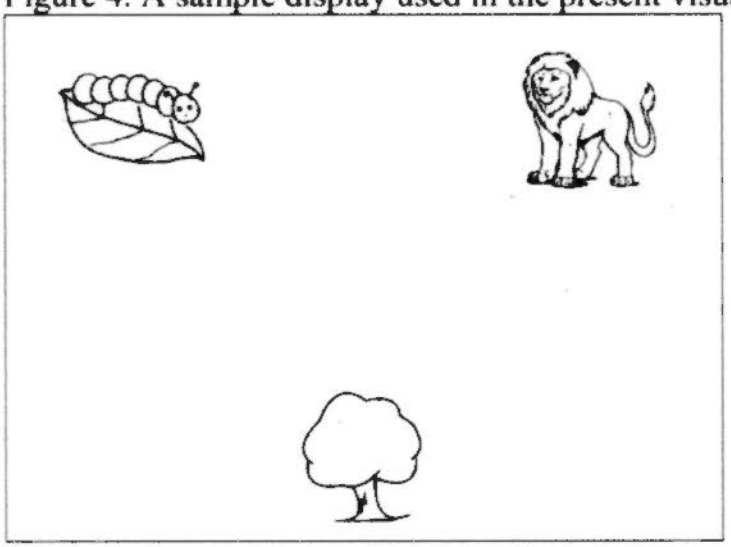
5.1.2 Materials and Design

Materials used in Experiment 2A were identical to those in Experiment 1A (see Appendix A). Specifically, each of the 24 target monosyllables (e.g., /bou3/) formed four conditions by pairing with an onset-shared competitor (O condition; e.g., /bui1/), a rime-shared competitor (R condition; e.g., /dou1/), a syllable-shared competitor (S condition; e.g., /bou1/), and a tone-shared competitor (T condition; e.g., /coi3/), creating a total of 96 experimental trials. Each trial contained a target, one of its competitors and an unrelated control distracter, which was actually taken from the competitors of other targets, shown on a visual display. For instance, one valid

experimental trial would include the target /bou3/, the onset-competitor /bui1/, and the unrelated distracter /dau2/ (which was the onset-shared distracter for the target /dou6/). Also, there were no orthographic or semantic relationships among the three items. Other triplets of items were created similarly so that each target formed trials with all competitors in its set. In addition, another set of 24 items was created with the same composition of conditions. However, the auditory target in these trials could not be found in the visual display (e.g., the syllable /ting4/ was presented for a visual display containing /min6/, /muk6/, and /wong4/, noticed that there was onset-sharing between two of the items). These trials served as fillers which required "no" responses in the target detection task.

Objects shown in the visual display were simple line drawings obtained from the picture database in our laboratory or by searching in the Google Image Database (2007). Each display contained the three objects in a triplet arranged in either v-shape or inverted v-shape (see Figure 4). The target object appeared equally often in each of the three positions.

Figure 4. A sample display used in the present visual-world experiments.



Because the same target was paired with four different conditions, to avoid target repetition, the 96 experimental trials were divided into four lists. Within a list,

each target and its corresponding competitors/distracters appeared only once. There were 6 trials for each condition in one list. Across lists, the same target was paired with all conditions and each competitor/distracter appeared twice. The same set of 24 fillers was added to the experimental items, resulting in 48 trials in each list. Finally, eight practice trials were prepared to acquaint participants with the procedure.

5.1.3 Procedure

Participants were tested individually in a sound-proof room. They were seated about 50cm from the computer screen. At this distance, each picture on the visual display was within the size of 9° X 9°, and the inter-center distance between pictures was about 20°. Participants' eye-movements were recorded by the Eyelink 1000 Desktop System (SR Research, Canada) with a sampling rate of 1000Hz. Viewing was binocular but we recorded fixation position of the right eye only. A chin-rest was used to minimize head movements. The standard nine-point calibration procedure was conducted to ensure the accuracy of fixation recording to be within 0.5°.

Each participant was assigned randomly to one experimental list (eight participants for each list). They were told that in each trial, a central fixation point would appear (for drift correction). After they established stable fixation on that point, they would be presented with a Cantonese monosyllabic word via speaker and a visual display containing three objects on the LCD monitor simultaneously. The visual display would remain on the screen until participants responded or after three seconds had passed (trial time-out). Participants were asked to decide whether the target word they heard was depicted by one of the objects on the screen by pressing corresponding keys on a gamepad connected to the Eyelink system. Their eye-movements and responses would be recorded as they searched through the display.

Before the experimental list, participants were given eight practice trials with materials not used in the actual experiment. Clarifications and additional practices were delivered upon request. After that, the 48 experimental trials were presented randomly. Stimuli presentation and recording were controlled by the Experiment Builder provided by SR Research. Participants' fixation accuracy was checked every trial when the fixation point was shown. Re-calibration was conducted if needed. The whole experiment took about 20 minutes.

5.1.4 Results and Discussion

Trials with incorrect responses or without responses within three seconds were coded as errors and eliminated from further analyses (11.33%). Trials in which responses were made without having fixated on the target were also discarded (1.04%) because it was unclear whether participants had really found the target with peripheral vision or they had misinterpreted the spoken syllable. Table 4 presents the mean reaction times (standard deviations) and error rates of target detection as a function of competitor conditions.

Table 4. Mean reaction times and error rates (standard deviations in parentheses) of target detection in Experiment 2A.

Condition	Reaction time (ms)	Error rate (%)
Onset-sharing	1516 (253)	13.2 (11.0)
Rime-sharing	1474 (196)	6.3 (11.8)
Syllable-sharing	1509 (246)	14.2 (13.4)
Tone-sharing	1389 (200)	12.2 (8.8)

From Table 4 it is obvious that participants could recognize the target much easier than in Experiment 1A. Identification seems to be particularly effortless in the

rime-sharing condition (low error rate) and tone-sharing condition (fast reaction time). To verify these observations, one-way Repeated Measure ANOVA was conducted treating subject (*F1*) and item (*F2*) as random factor. Results indicated that the four conditions differed significantly in reaction times ($F1_{(3, 93)} = 4.68$, $MSE = 23220$, $p < .01$; $F2_{(3, 69)} = 3.70$, $MSE = 36884$, $p < .05$). Further pairwise comparisons with LSD correction suggested that the difference could mainly be attributed to faster reaction times in the tone-sharing condition ($t1_{(31)} = 3.01, 2.71, 2.92$ compared with onset-sharing, rime-sharing, and syllable sharing conditions, all $ps < .05$; $t2_{(23)} = 3.22, 1.74, 2.30$ compared with onset-sharing, rime-sharing, and syllable sharing conditions, all $ps < .1$). There was also a significant difference in error rates across the four conditions in subject analysis ($F1_{(3, 93)} = 3.15$, $MSE = 0.013$, $p < .05$; $F2_{(3, 69)} = 1.37$, $MSE = 0.021$, *n.s.*). Rime-sharing condition had the lowest error rate compared with other conditions ($t1_{(31)} = 2.29$ to 2.55, all $ps < .05$).

Next, in each condition, we computed the mean fixation proportions towards the target, competitor, and distracter pictures over 100 ms time windows from 0ms to 1500ms after trial start. The results are presented in Figures 5A to 5D. Important similarities across conditions emerge when inspecting the four figures. First, well before 1500ms (roughly the grand mean reaction time), participants were looking at the target exclusively in all conditions. This conformed to the basic underlying assumption of visual-world paradigm that participants needed to look at the target before responding. Second, consistent with other visual-world studies, it took about 150 ms to 200 ms to plan and execute saccades to any pictures on the display. Third, there was a period of confusion (looking at all pictures) before being able to fixate on the target unambiguously.

Figure 5A. Fixation proportion over time for the onset-sharing competitor condition.

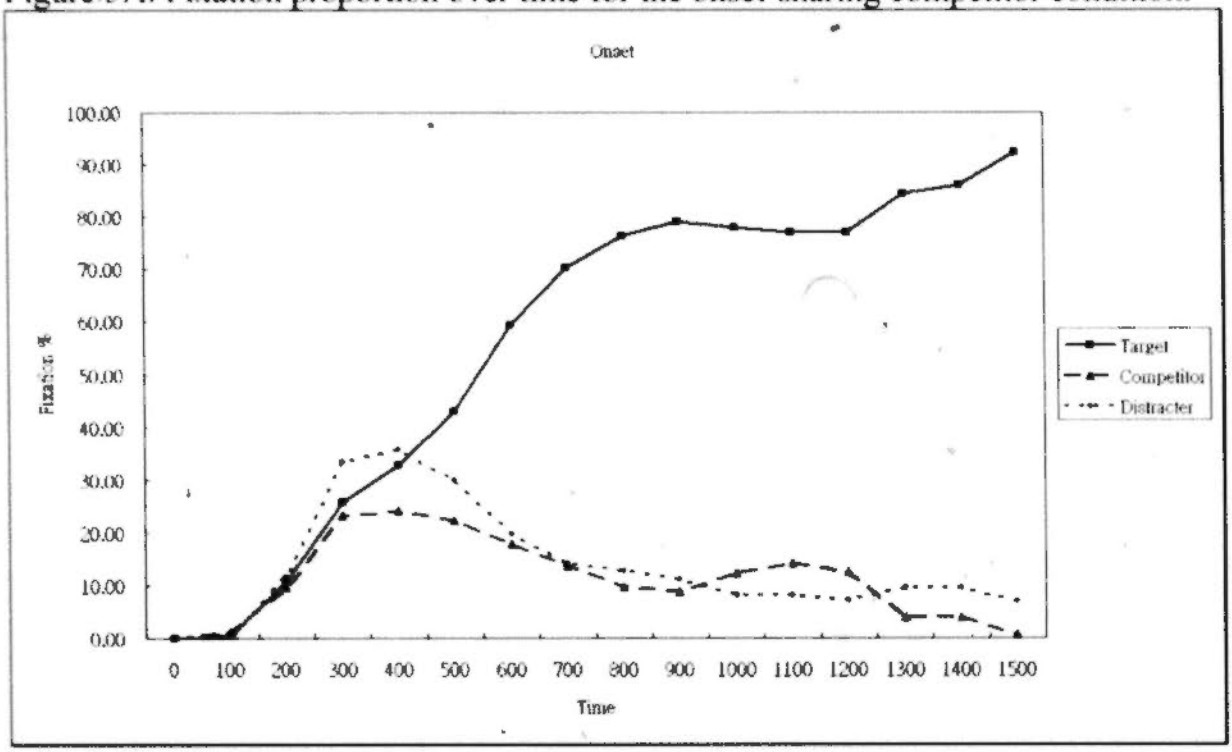


Figure 5B. Fixation proportion over time for the rime-sharing competitor condition.

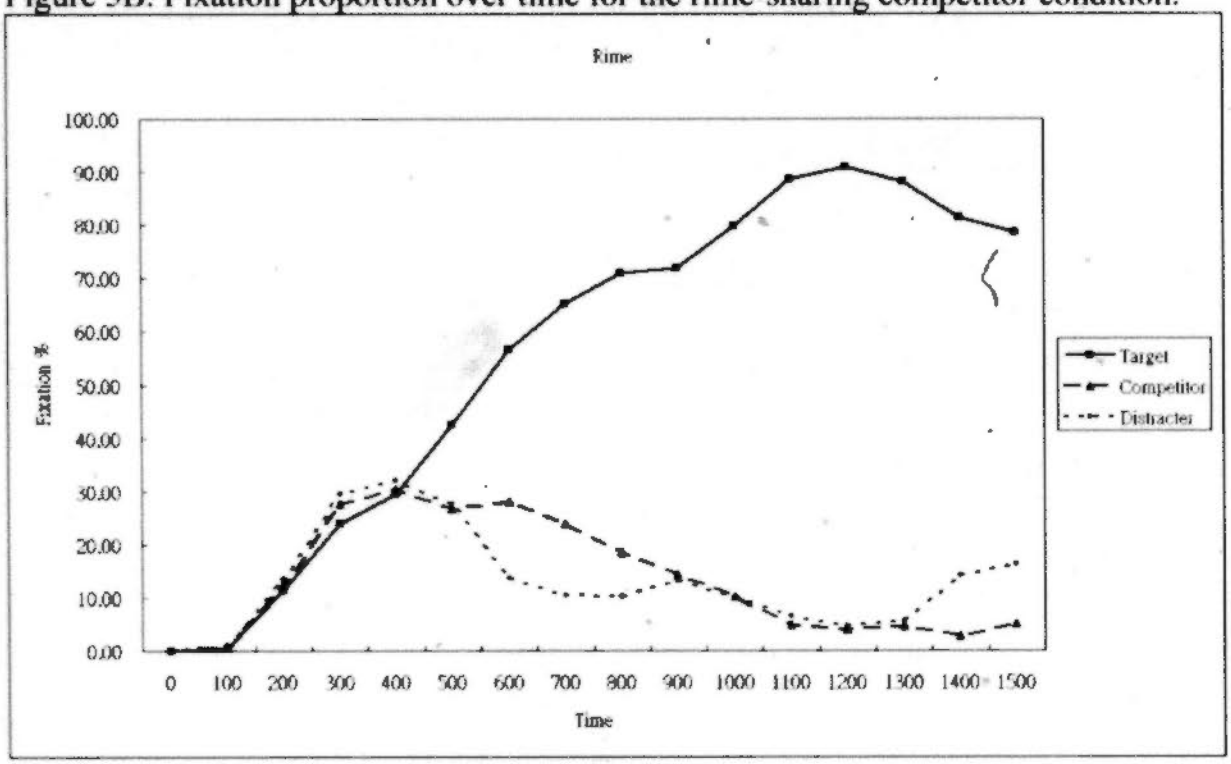


Figure 5C. Fixation proportion over time for the syllable-sharing competitor condition.

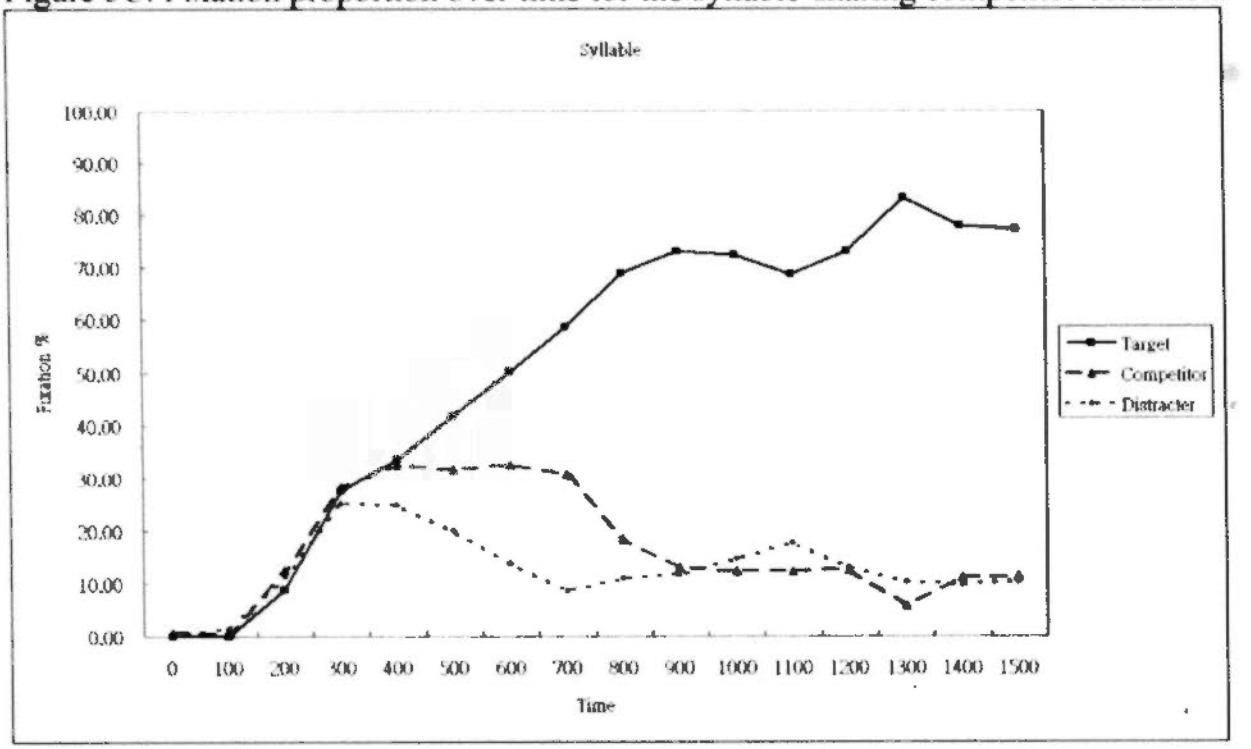
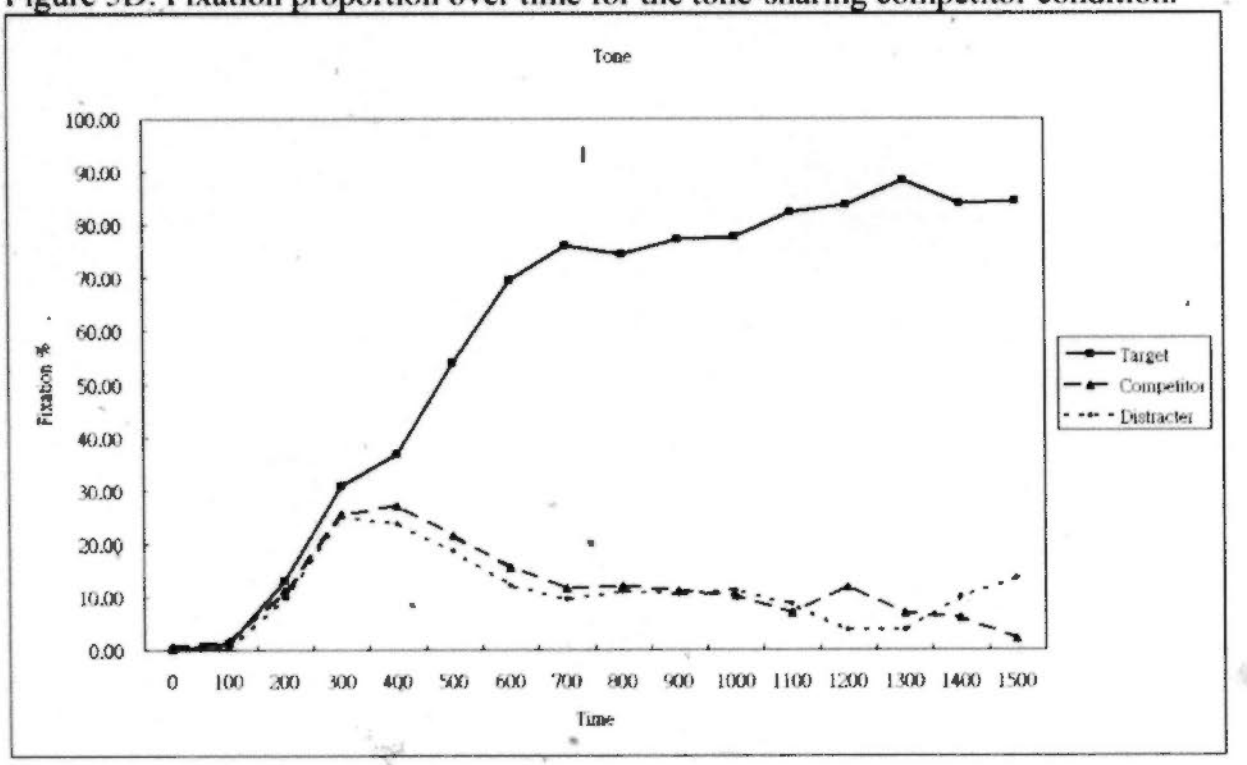


Figure 5D. Fixation proportion over time for the tone-sharing competitor condition.



A closer inspection in the course of target fixations reveals clear differences across conditions. For the tone-sharing competitor condition, the period of confusion

was very brief. Fixation proportion towards the actual target started rising monotonically at about 300 ms after word onset and reached a plateau at about 700 ms. In contrast, the other conditions had the target fixation increased at 400ms and the plateau reached at much later (800 ms to 1000 ms). Moreover, the rime-sharing competitors appeared to be fixated more within a time window of 500 ms to 900 ms. Similarly, the syllable-sharing competitors were also fixated more from 300 ms to 900 ms. It should also be noted that in the onset-sharing condition, the unrelated control distracters were receiving more fixations initially.

To verify the above mentioned observation with statistical analyses, we collapsed the fixation proportions on different pictures from 201 ms to 900 ms to represent the time window of target syllables (200 ms for saccade planning and execution; mean target syllable duration roughly equal to 700 ms). We also included a more fine-grained analysis by separating the whole syllable time window into two parts (first part: 201 ms – 500 ms; second part: 501 ms – 900 ms). The mean fixation proportions (standard deviations) of the four conditions in different time windows are presented in Table 5.

Statistical analyses confirmed the observation. Repeated Measure ANOVAs treating competitor condition and picture of fixation as two independent variables revealed robust interactions between the two factors in 201 – 900 ms ($F1_{(6, 186)} = 9.79$, $MSE = 77.96$, $p < .01$; $F2_{(6, 138)} = 4.14$, $MSE = 126.97$, $p < .01$), in 201 – 500 ms ($F1_{(6, 186)} = 7.08$, $MSE = 164.24$, $p < .01$; $F2_{(6, 138)} = 1.18$, $MSE = 124.76$, *n.s.*), and 501 – 900ms ($F1_{(6, 186)} = 8.71$, $MSE = 151.49$, $p < .01$; $F2_{(6, 138)} = 4.15$, $MSE = 269.50$, $p < .01$). A series of one-way ANOVAs was then conducted compare the fixation proportions on the different pictures in each condition and each time window.

Table 5. Mean fixation proportions (%) and standard deviations (in parentheses) across conditions in Experiment 2A.

	Onset-sharing competitor condition			Rime-sharing competitor condition			Syllable-sharing competitor condition			Tone-sharing competitor condition		
	T	C	D	T	C	D	T	C	D	T	C	D
201 – 500 ms	23.04 (11.73)	19.02 (10.58)	26.96 (11.72)	21.55 (9.63)	23.41 (12.20)	25.00 (11.69)	34.22 (16.12)	30.75 (10.99)	23.28 (12.81)	40.60 (14.02)	24.77 (10.35)	22.38 (13.13)
501 – 900 ms	62.27 (14.05)	15.93 (9.46)	19.14 (11.13)	58.79 (9.43)	24.31 (9.38)	15.53 (7.88)	54.91 (16.20)	28.19 (12.30)	13.27 (9.86)	68.59 (13.30)	15.20 (10.07)	12.74 (9.23)
201 – 900 ms	45.46 (11.34)	17.25 (7.41)	22.49 (7.12)	42.83 (6.77)	23.92 (7.69)	19.59 (7.05)	41.33 (12.29)	26.56 (7.70)	16.34 (7.95)	50.72 (10.31)	17.81 (6.46)	15.64 (8.08)

Note: T = Target picture; C = Competitor picture; D = Unrelated control distracter picture.

Onset-sharing competitor condition. One-way ANOVAs revealed a significant picture type main effect within 201 – 500 ms ($F1_{(2, 62)} = 3.73$, $MSE = 135.45$, $p < .05$; $F2_{(2, 46)} < 1$). Contrary to our expectation, however, participants fixated on the unrelated distracter more often than the onset-sharing competitor ($t1_{(31)} = 2.56$, $p < .05$); other differences were not significant. Within 501 – 900 ms, there was again a picture type main effect ($F1_{(2, 32)} = 106.91$, $MSE = 200.42$, $p < .001$; $F2_{(2, 46)} = 17.68$, $MSE = 747.54$, $p < .01$). There were more fixations on the target picture than both the competitor and distracter (all $ps < .01$). For the whole time window 201 – 900 ms, again the picture effect was significant ($F1_{(2, 32)} = 72.73$, $MSE = 99.01$, $p < .001$; $F2_{(2, 46)} = 7.35$, $MSE = 570.75$, $p < .01$). Pairwise comparisons indicated that there were more fixations on target than competitor and distracter (all $ps < .01$). There were also more fixations on distracter than competitor in subject analysis ($t1_{(31)} = 3.10$, $p < .01$).

Rime-sharing competitor condition. Analyses revealed no significant picture type effect within 201 – 500 ms ($F1_{(2, 62)} = 0.69$; $F2_{(2, 46)} = 0.10$). In contrast, there was a robust picture type effect within 501 – 900 ms ($F1_{(2, 32)} = 142.11$, $MSE = 117.75$, $p < .001$; $F2_{(2, 46)} = 35.67$, $MSE = 354.07$, $p < .01$). Pairwise comparisons indicated that the fixation proportion on targets higher than that on competitors (all $ps < .01$) which was in turn higher than that on distracters ($t1_{(31)} = 3.50$, $p < .01$; $t2_{(23)} = 2.09$ $p < .05$). Finally, the picture type main effect was also significant for the overall window 201 – 900 ms ($F1_{(2, 32)} = 74.22$, $MSE = 65.86$, $p < .001$; $F2_{(2, 46)} = 7.83$, $MSE = 453.80$, $p < .01$). Participants fixated more on target pictures than both the competitor and distracter pictures (all $ps < .05$).

Syllable-sharing competitor condition. The fixation proportion within 201 – 500 ms showed a significant picture type effect ($F1_{(2, 62)} = 4.86$, $MSE = 205.69$, $p < .05$; $F2_{(2, 46)} < 1$). In this early time window, both the target and the competitor had already been fixated significantly more than the unrelated control ($t1$s = 2.54 and 2.36 respectively, both ps < .05). The picture type effect was further amplified in the later time window 501 – 900 ms ($F1_{(2, 32)} = 58.76$, $MSE = 242.36$, $p < .001$; $F2_{(2, 46)} = 17.61$, $MSE = 491.45$, $p < .01$). Target pictures were fixated significantly more than competitors (all ps < .05), which were in turn fixated more than distracters ($t1_{(31)} = 4.71$, $p < .01$; $t2_{(23)} = 2.53$, $p < .05$). Finally, within the whole time window 201 – 900 ms, a significant picture type effect was observed ($F1_{(2, 32)} = 47.05$, $MSE = 107.35$, $p < .001$; $F2_{(2, 46)} = 7.32$, $MSE = 407.58$, $p < .01$). Participants fixated more on the targets than on the competitors ($t1_{(31)} = 8.98$, $p < .01$; $t2_{(23)} = 3.98$, $p < .01$). However, the difference between competitor and distracter fixations was only significant in the subject analysis ($t1_{(31)} = 4.62$, $p < .01$; $t2_{(23)} = 1.60$, *n.s.*).

Tone-sharing competitor condition. There was a significant picture type main effect within 201 – 500 ms in the subject analysis ($F1_{(2, 62)} = 19.49$, $MSE = 160.98$, $p < .01$). The target was fixated significantly more than both competitors and distracters (all ps < .05), which did not differ by themselves. For the later time window 501 – 900 ms, the picture type effect was robust ($F1_{(2, 32)} = 187.86$, $MSE = 169.66$, $p < .001$; $F2_{(2, 46)} = 58.19$, $MSE = 412.11$, $p < .01$). Again, the target was fixated significantly more than both competitors and distracters (all ps < .05). Exactly the same pattern could be observed when the whole 201 – 900 ms time window was considered: Overall picture type effect was significant ($F1_{(2, 32)} = 163.88$, $MSE = 74.43$, $p < .001$; $F2_{(2, 46)} = 22.37$,

MSE = 388.73, $p < .01$), with the difference attributable to higher fixation proportion on the target pictures than the other two pictures (all $ps < .01$).

Table 6 presents a brief summary of the fixation proportions on targets, competitors, and distracters within different time windows for each condition. The most notable finding is that in the tone-competitor condition, participants could start fixating on the actual target in the early time window of 201 – 500ms. In contrast, in all other conditions, there was a relatively long period of uncertainty in which participants looked at various pictures before making a strong commitment over any object. Such discrepancy seems to be consistent with the results in Experiment 1A in suggesting that lexical tone is not involved in candidate generation. In other words, the tone-competitor was not activated due to its tone-sharing with the target. As its activation level remained low, it would not be confused with the target and interfere with target fixations. Furthermore, the fast convergence of eye fixations on the target in this condition supported the incremental nature of speech perception in Cantonese. Participants could rapidly integrate the acoustic signals received and came up with hypothetical candidates. As long as a specific candidate met other response constraints (the visual display in the present study), it could create a response bias (recognition based on partial information) accordingly.

Table 6. Summary of differences in fixation proportion across conditions.

Condition	201 – 500 ms	501 – 900 ms	201 – 900 ms
Onset-sharing	D = T > C *	T > C = D	T > D > C
Rime-sharing	T = C = D	T > C > D	T > C = D
Syllable-sharing	T = C > D	T > C > D	T > C > D
Tone-sharing	T > C = D	T > C = D	T > C = D

Note: T = target fixation proportion; C = competitor fixation proportion; D = unrelated distracter fixation proportion; * = only the difference between distracter and competitor reached significance.

For the rime-sharing condition, initially participants were not sure about which picture would be the target before they finally gather enough information to recognize the actual target, a pattern that fit quite well with the results in gating. More importantly, however, in this condition there were more competitor fixations than distracter fixations in the late time window (501 900 ms). Participants were sensitive to the later portion of the target words and used it to activate rime-matched candidates. The high activation of competitor words was then translated into a stronger bias of eye gaze towards them than towards the corresponding distracters. Result of this condition not only supported the greater sensitivity to rime in the visual-world paradigm than in gating (Allopenna et al., 1998), it also suggested that, given appropriate experimental procedure, Cantonese speakers could respond based on subsyllabic information.

Again, the ability to utilize the subsyllabic rime did not imply that syllable played no role in Chinese speech perception. It is clear from Table 6 that unlike the rime-sharing condition, which showed more competitor fixations only in the later time window, competitor fixation proportions were in general higher than the unrelated distracters (201 900 ms) in the syllable-sharing condition. Even in the early time window (201 500 ms), the competitors were already fixated significantly more than the distracters. Actually, they were as active as the real targets. In other words, contrary to the complete initial confusion in rime-sharing, participants could very quickly isolate the correct syllable and separate it from other totally mismatched words. However, since there was a total segmental overlap between targets and competitors, participants needed more time to gather subtle acoustic information (probably including lexical tone) before arriving at the correct interpretation finally. The general pattern of high competitor fixations in this condition was thus consistent

with Experiment 1A in supporting the existence of independent syllabic effects beyond the subsyllabic ones[8].

On the other hand, not all results in the present experiment were equally interpretable. The most unexpected finding was the observation of high distracter fixations in the onset-sharing condition. In both the early (201 – 500ms) and the overall (201 – 900ms) time window, there were significantly more fixations on the unrelated distracters than on the onset-sharing competitors. We also tested whether the slight elevation of fixation proportion on onset competitor within 901 – 1200ms was significant (see Figure 5A). Unfortunately, it was not (both *p*s > .1). Following the same linking hypothesis as before, we were forced to conclude that distracters were more highly activated than the corresponding competitors![9] This was odd because distracters should be total mismatches to the speech signal. Furthermore, these results seems particularly unusual given the strong onset effect in candidate generation in Experiment 1A and in previous studies using Indo-European languages (e.g., Allopenna et al., 1998; Marslen-Wilson & Zwitserlood, 1989). They were also inconsistent with the rapid convergence on targets over distracters in the syllable-sharing and tone-sharing conditions in the present experiment.

One possibility for such unexpected finding is that the onset-sharing was too brief in most of our materials. Its facilitation on competitor activation was therefore very small and could be easily overridden by quick negative feedback from later phonemes. Support for this proposal came from the visual-world study by Allopenna et al. (1998), who actually defined onset-sharing as overlapping of the whole initial

---

[8] One possible concern here is that our participants were mistaking the competitor as the actual target. Two observations argue against this possibility. First, target fixation was still the highest, indicating that participants could recognize the target. Second, we would expect much lower miss (error) rate if competitors were mistaken as target. This was not the case, however, as Table 3 shows.

[9] It should be noted that the high fixation on distracters was irrelevant to uncontrolled picture properties because the same distracter was the competitor in another trial. Its "attractiveness" was thus balanced across conditions.

syllable (e.g., "beaker" and "beetle" share the entire first syllable). In their study, strong onset-competitor activation could be observed over a 600ms time window. In order to provide more information about how onset-sharing affect Chinese speech perception through candidate generation and selection, we conducted Experiment 2B with materials that should maximize the chance of observing onset-sharing effects.

5.2 Experiment 2B – a deeper look to the role of word onset

The primary goal of Experiment 2B was to investigate more deeply the role of onset in Chinese speech perception. To be specific, we tested whether the absence of competitor activation over distracter in the onset-sharing condition (Experiment 2A) could be attributed to the limited phonemic overlapping (a single initial phoneme) between the target and competitor. To achieve this goal, we explicitly manipulated the degree of initial phoneme overlapping between the targets and its corresponding competitors. In the present experiment, we compared the activation of three types of competitors, namely the onset-sharing condition, the onset-plus-sharing condition, and the embedded word condition. While the onset condition again shared only a single phoneme with the target, the latter two conditions had larger overlapping with the target. For instance, 'caai4 (柴; firewood; onset-plus) and /caa4/ (茶; tea; embedded word) shared two phonemes with the target /caau4/ (巢; nest). If the degree of phoneme overlapping really matters, we would observe more fixations on competitors than distracters in the onset-plus and embedded word conditions.

Experiment 2B also served the secondary goal in providing empirical data on how fully embedded words were recognized in Chinese speech. Answer to this question is not quite straight forward (Grosjean, 1985; Salverda, Dahan, & McQueen, 2003). In principle, the isolation point of an embedded word should locate after its

word offset because by definition, there is always another word that shares all phonemic content with it (e.g., /caa4/ was totally nested within /caau4/). With this in mind, a reasonable expectation is that recognition of embedded words should be delayed. Yet, some previous studies (e.g., Davis, Marslen-Wilson, & Gaskell, 2002; Salverda et al., 2003) showed the contrary: An embedded word could indeed be recognized very quickly. Research on this topic has the important potential of informing how participants could successfully segment a particular word from the continuous speech stream. Virtually nothing about this was done in Chinese despite the presence of heavy word nesting (e.g., 煙; smoke was embedded in 煙灰; ash, which was in turn embedded in 煙灰缸; ashtray). The present experiment intended to provide preliminary data on this issue.

5.2.1 Participants

Twenty undergraduates (eight males) in The Chinese University of Hong Kong were recruited for this experiment. They were paid $50 for participation. All of them were native Cantonese speakers, reported no hearing deficits and had normal or corrected-to-normal vision. None of them had participated in the previous experiments. Informed consent was obtained and full debriefing was delivered after experiment.

5.2.2 Materials and Design

Eighteen target monosyllables were prepared for Experiment 2B. Although we tried to pair each target syllable with competitors from all three conditions (onset, onset-plus, embedded), we failed because of the constraints in visual-world paradigm that all items should be concrete nouns. Therefore, each target could only be paired

with two conditions. For example, the target /caau4/ could only be paired with the onset-plus competitor /caai4/ and the embedded competitor /caa4/, while the target /baau1/ could be paired with the onset competitor /bou1/ and embedded competitor /baa1/. Items in a set also shared tone so any effect could only be attributed to the degree of segmental sharing. In total, there were 12 items for each condition (see Appendix C). In addition, we prepared 12 filler trials which had the target sharing rime or tone with the competitor to decrease participants' awareness to manipulation of the initial phonemes. Furthermore, there were 30 trials requiring "no" responses in the target detection task. Procedure for syllable recording and editing was identical to previous experiments. Appendix C also shows the duration of target syllables for experimental items.

The experimental trials were divided into two lists such that each target appeared only once in each list. Within a list, there were six items for each condition. And across lists, all target-competitor pairs were presented. The same set of fillers (12 items) and "no" trials (30 items) were added to each list, resulting 60 items in total. Visual displays were prepared in the same way as Experiment 2A.

5.2.3 Procedure

Participants were randomly assigned to experimental list until there were 10 persons in each list. Procedure was identical to that in Experiment 2A. The whole experiment lasted for about 20 minutes.

5.2.4 Results and Discussion

All responses were made after fixating at the target. Trials with incorrect responses or without responses within three seconds were coded as errors and

eliminated from further analyses (16.67%). Table 7 presents the mean reaction times (standard deviations) and error rates of target detection as a function of competitor conditions. The overall values were comparable to those in Experiment 2A. Despite the apparent longer reaction time and higher error rate in the onset-sharing condition, none of the effect approached statistical significance (all *p*s > .1). In this sense, difficulty in target detection was closely matched across conditions.

Table 7. Mean reaction times and error rates (standard deviations in parentheses) of target detection in Experiment 2B.

Condition	Reaction time (ms)	Error rate (%)
Onset-sharing	1542 (216)	19.2 (14.5)
Onset-plus-sharing	1518 (246)	12.6 (11.9)
Embedded word	1442 (226)	18.4 (14.1)

Again, we plotted the fixation proportion over time curves for each condition from 0ms to 1500 ms. We also collapsed the three competitor conditions to produce an "onset (all)" curve which resembled the unspecified onset-sharing condition in Experiment 2A. Figures 6A to 6D display these curves. The shape of the "average" curve indeed fits quite well with Figure 5A. So we have successfully replicated the unexpected pattern in the previous experiment with a different set of materials.

Next, we examined the data with reference to our hypotheses. Contrary to our expectation, sharing more word initial phonemes did not help much in boosting the activation level of competitors. Actually, in all time windows, there were no interactions between competitor condition and picture of fixation ($F1_{(4, 76)}$ = .32 to 1.14; $F2_{(4, 66)}$ = .07 to .37)[10]. Yet, a closer inspection on the curves suggests some potential differences in competitor activation across conditions: In the late time

[10] In the item analyses of Experiment 2B, competitor condition was treated as a between-item factor because target-competitor pairing was incomplete.

window (501 – 900 ms), participants appeared to fixate more on the embedded word competitors than on the unrelated distracters. This competitor fixation was much weaker in the other two conditions. We performed a series of one-way ANOVAs to further test the fixation proportion as a function of picture type in each condition and time window. Table 8 presents the mean fixation proportions (standard deviations) in each condition.

Figure 6A. Fixation proportion over time for the onset-sharing condition.

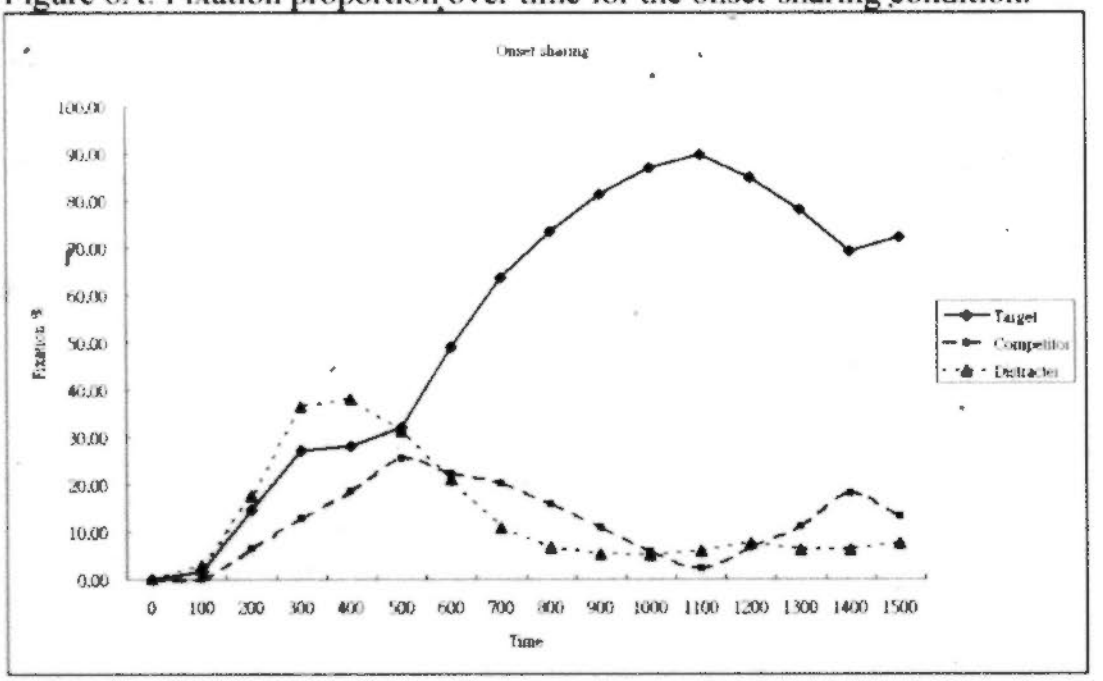


Figure 6B. Fixation proportion over time for the onset-plus-sharing condition.

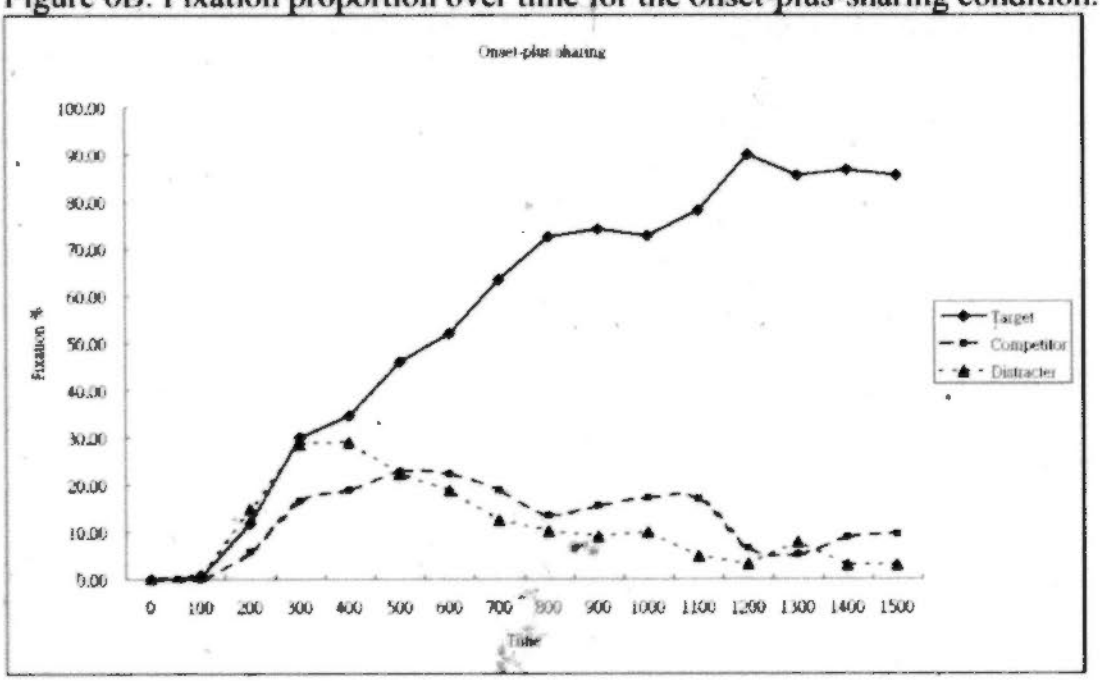


Figure 6C. Fixation proportion over time for the embedded condition.

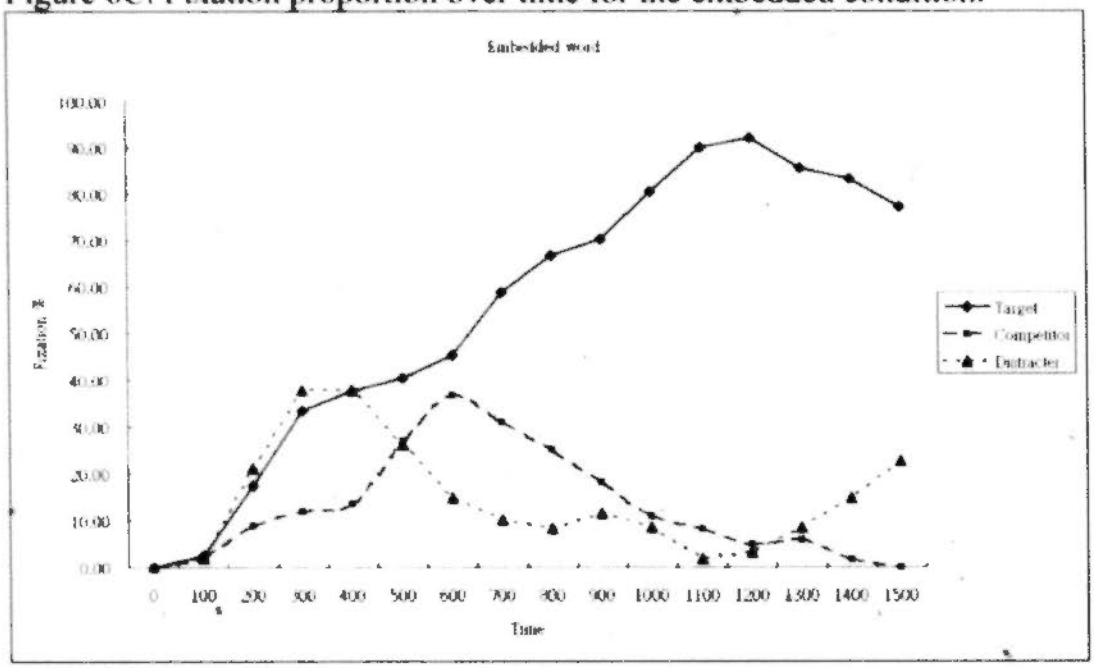


Figure 6D. Fixation proportion over time after collapsing the three conditions.

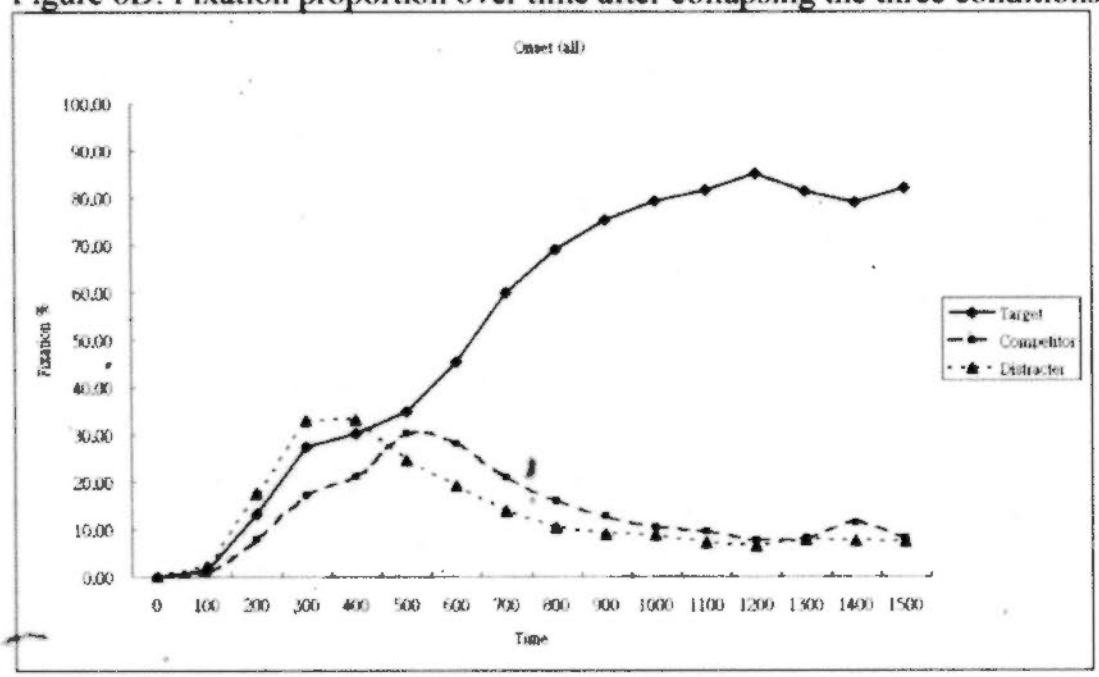


Table 8. Mean fixation proportions (%) and standard deviations (in parentheses) across conditions in Experiment 2B.

	Onset-sharing competitor condition			Onset-plus-sharing competitor condition			Embedded word competitor condition		
	T	C	D	T	C	D	T	C	D
201 – 500 ms	23.34 (23.47)	12.60 (15.37)	30.79 (19.51)	25.54 (24.58)	13.68 (12.03)	24.23 (21.48)	29.51 (10.95)	11.39 (9.46)	32.33 (11.95)
501 – 900 ms	54.52 (23.21)	21.01 (12.05)	17.52 (15.32)	58.60 (22.11)	19.39 (12.41)	16.03 (10.08)	52.86 (13.50)	29.94 (15.92)	14.91 (9.14)
201 – 900 ms	41.15 (21.55)	17.41 (11.78)	23.21 (15.73)	44.43 (21.01)	16.94 (10.31)	19.54 (12.80)	42.85 (8.21)	21.99 (9.87)	22.38 (6.97)

Note: T = Target picture; C = Competitor picture; D = Unrelated control distracter picture.

Onset-sharing competitor condition. Within the time window 201 – 500 ms, one-way ANOVAs revealed a significant picture type main effect in subject analysis ($F1_{(2, 38)}$ = 3.42, MSE = 489.34, p < .05; $F2_{(2, 22)}$ < 1). Pairwise comparisons suggested that distracters were fixated more than competitors ($t1_{(19)}$ = 3.76, p < .01); other differences were not significant. Within 501 – 900 ms, there was again a picture effect ($F1_{(2, 38)}$ = 19.40, MSE = 430.34, p < .01; $F2_{(2, 22)}$ = 6.04, MSE = 592.43, p < .01), with more fixations on targets than on distracters ($t1_{(19)}$ = 4.70, p < .01; $t2_{(11)}$ = 3.14, p < .01) and on competitors ($t1_{(19)}$ = 4.62, p < .01; $t2_{(11)}$ = 2.15, p = .054). For the whole time window 201 – 900ms, participants fixated on targets ($F1_{(2, 38)}$ = 7.97, MSE = 384.42, p < .01; $F2_{(2, 22)}$ = 2.22, *n.s.*) significantly more than on competitors ($t1_{(19)}$ = 3.45, p < .01) and marginally more than on distracters ($t1_{(19)}$ = 2.45, p < .05).

Onset-plus-sharing competitor condition. There were no significant differences in 201 – 500 ms ($F1_{(2, 38)}$ = 1.62; $F2_{(2, 22)}$ = 0.26). The superior target fixations emerged in 501 – 900 ms ($F1_{(2, 38)}$ = 31.35, MSE = 357.26, p < .01; $F2_{(2, 22)}$ = 11.79, MSE = 470.80, p < .01). Pairwise comparisons showed that targets were fixated more than both the competitors and distracters (all ps < .05), which did not differ by themselves. The same pattern holds for the whole time window 201 – 900 ms ($F1_{(2, 38)}$ = 14.04, MSE = 328.01, p < .01; $F2_{(2, 22)}$ = 5.35, MSE = 400.44, p < .05). The higher fixation proportion on targets over competitors and distracters was also robust (all ps < .05).

Embedded word competitor condition. The overall fixation pattern (201 – 900 ms) was similar to the other two conditions ($F1_{(2, 38)}$ = 30.01, MSE = 94.91, p < .01; $F2_{(2, 22)}$ = 1.83, *n.s.*) in showing significantly more target fixations (ps < .01). Yet, the detailed time course in the embedded condition appeared to be quite unique: Even

though in the early time window participants again had lower fixation proportions ($F1_{2,\ 88}$ = 19.04, MSE = 135.66, p < .01; $F2_{2,\ 22}$ = 0.74, *n.s*) on competitors than on targets or distracters (ps < .01), there was a rapid boost in competitor activation within 501 – 900 ms ($F1_{2,\ 88}$ = 28.49, MSE = 256.27, p < .01; $F2_{2,\ 22}$ = 5.62, MSE = 668.93, p < .05). In this time window, fixation proportion on targets was higher than that on competitors ($t1_{(44)}$ = 3.67, p < .01), which in turn was higher than that on unrelated distracters ($t1_{(44)}$ = 3.17, p < .01; $t2_{(11)}$ = 1.92, p = .081). In other words, perhaps being delayed, but the embedded words were really activated and considered in the candidate set during the course of recognizing the corresponding target monosyllables.

Table 9 presents a brief summary of the fixation proportion on targets, competitors, and distracters within different time windows for each condition. We replicated the lack of competitor activation with onset sharing in Experiment 2A. Onset alone is therefore not a particularly strong determinant for candidate generation in online Chinese speech perception. Moreover, in general participants performed quite similarly across the three conditions (201 – 900 ms): They could reach the target and fixated on it significantly more than other pictures before responding. This might account for the lack of interaction between condition and picture in the initial analysis, indicating that more phoneme overlapping did not necessarily lead to overall stronger competitor activation.

Table 9. Summary of differences in fixation proportion across conditions.

Condition	201 – 500 ms	501 – 900 ms	201 – 900 ms
Onset-sharing	D = T > C *	T > C = D	T > C = D
Onset-plus-sharing	T = C = D	T > C = D	T > C = D
Embedded word	D = T > C	T > C > D	T > C = D

Note: T = target fixation proportion; C = competitor fixation proportion; D = unrelated distracter fixation proportion; * = only the difference between distracter and competitor reached significance.

However, looking more closely into the temporal evolution of competitor fixations could reveal effects due to larger phoneme overlapping. Although usually competitors had lower, or at best, equal, fixation proportions to the distracters, within 501 – 900 ms, embedded words were actually more highly active than the unrelated control distracters. This finding appeared to be consistent with previous studies showing confusion between embedded and carrier words (e.g., Vroomen & de Gelder, 1997). It should be noted that the embedded words did not have the same rime as the actual target (e.g., /aa/ in /caa4/ but /aau/ in /caau4/), so activation could not be reduced to the rime-overlapping demonstrated in Experiment 2B. Moreover, the fact that competitors were not fixated more than distracters in the onset-plus condition suggested that merely having more phoneme overlapping was insufficient to be included in the candidate set. Rather, the complete embedding allowed facilitation to accumulate without inhibition from mismatched phonemes (as in the onset-plus cases). As a result, perhaps fully embedded words would have a better chance to be included into the candidate set than the onset or onset-plus competitors. This also suggested that Chinese speakers might actually be sensitive to phoneme level processing in speech perception.

Yet, we still have to explain why the activation of embedded word in Chinese speech recognition was delayed rather than immediate (Salverda et al., 2003). There are no readily available solutions to this question. After all, given the incremental nature of human speech perception, existing models usually assume auditory inputs to exert immediate influence. When searching through the literature, we noticed that there was a recent interest in studying how fine acoustic details influence speech comprehension. Results generally supported a strong role of subphonemic features;

participants were highly sensitive to subtle acoustic variations in vowel length (Davis et al., 2002; Salverda et al., 2003), voice-onset time (McMurray, Aslin, Tanenhaus, Spivey, & Subik, 2008), and phoneme identity (Marslen-Wilson & Warren, 1994). It is also well-established that co-articulation could alter the acoustic realization of a specific phoneme (e.g., /d/ in /du/ vs. /di/). Therefore, we speculated that the acoustic forms of the targets and competitors used were actually rather different. In this case, perhaps the low activation level of competitors could be attributable to our participants' sensitivity to acoustic details.

To test whether the acoustic realization of the target was different from that of the competitor, we obtained typical tokens of the syllables we used from an online database (http://humanum.arts.cuhk.edu.hk/Lexis/lexi-can/) and use Praat (Boersma & Weenink, 2009) to generate the spectrograms for the initial 300ms of each syllable. Then we presented side-by-side the two spectrograms of each target-competitor pair to 23 naïve participants, who rated the similarity between the two items on a 6-point Likert scale (6 = very similar). Two representative sets of spectrograms are shown in Figures 7A and 7B. The similarity ratings differed across the three conditions ($F2_{(2, 44)}$ = 13.29, $p < .01$). The similarity of onset-sharing condition (mean = 1.78, S.D. = 0.59) was significantly lower than that of onset-plus condition (mean = 3.33, S.D. = 0.92, $p < .01$) and that of embedded word condition (mean = 3.50, S.D. = 1.11, $p < .01$)[11].

More importantly, we divided the target-competitor pairs into acoustically similar and acoustically dissimilar items (cutoff similarity = 4) and plotted the fixation proportion over time curves separately. Only seven items were in the similar group. We expected the competitors would be activated more strongly when they

[11] I acknowledged that this procedure was not optimal in comparing acoustic similarity. A better procedure would be directly comparing the values of formants and formant transitions, etc. However, this was the best I could do under limited resources. And as we will see, meaningful data emerged even with this sub-optimal procedure.

were sharing acoustic features with the targets. The curves are presented in Figures 8A to 8B.

Figure 7A. An item used in the spectrogram rating task (similarity rating > 4).

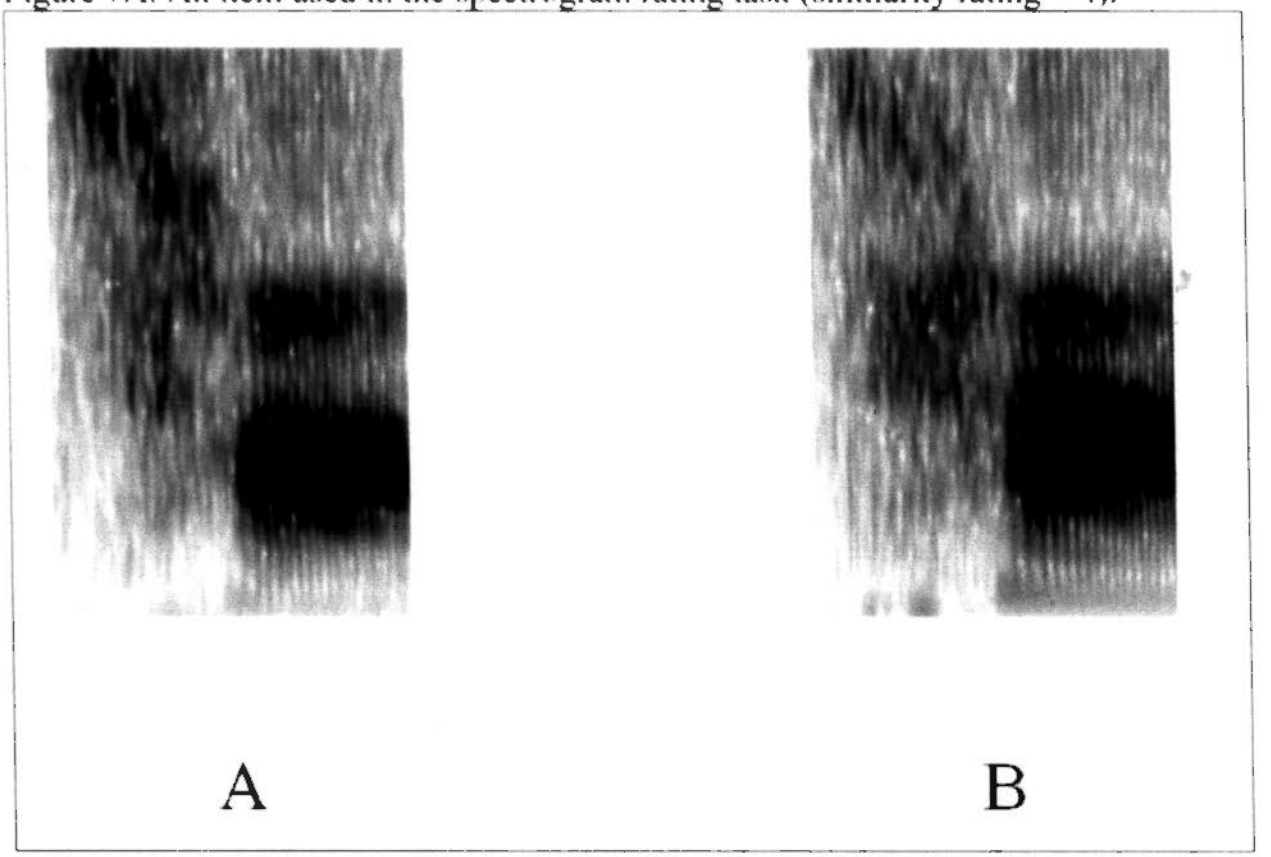


Figure 7B. An item used in the spectrogram rating task (similarity rating < 4).

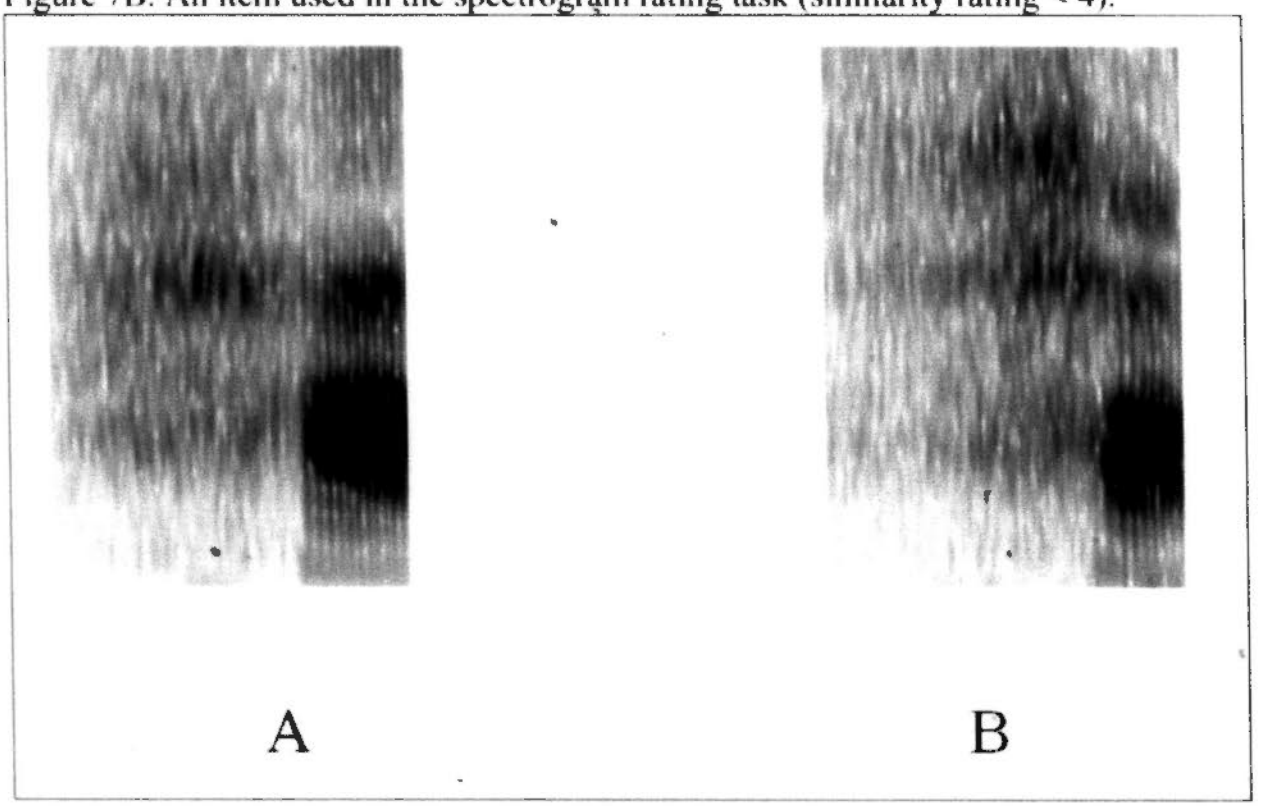


Figure 8A. Fixation proportion over time for the acoustically similar items.

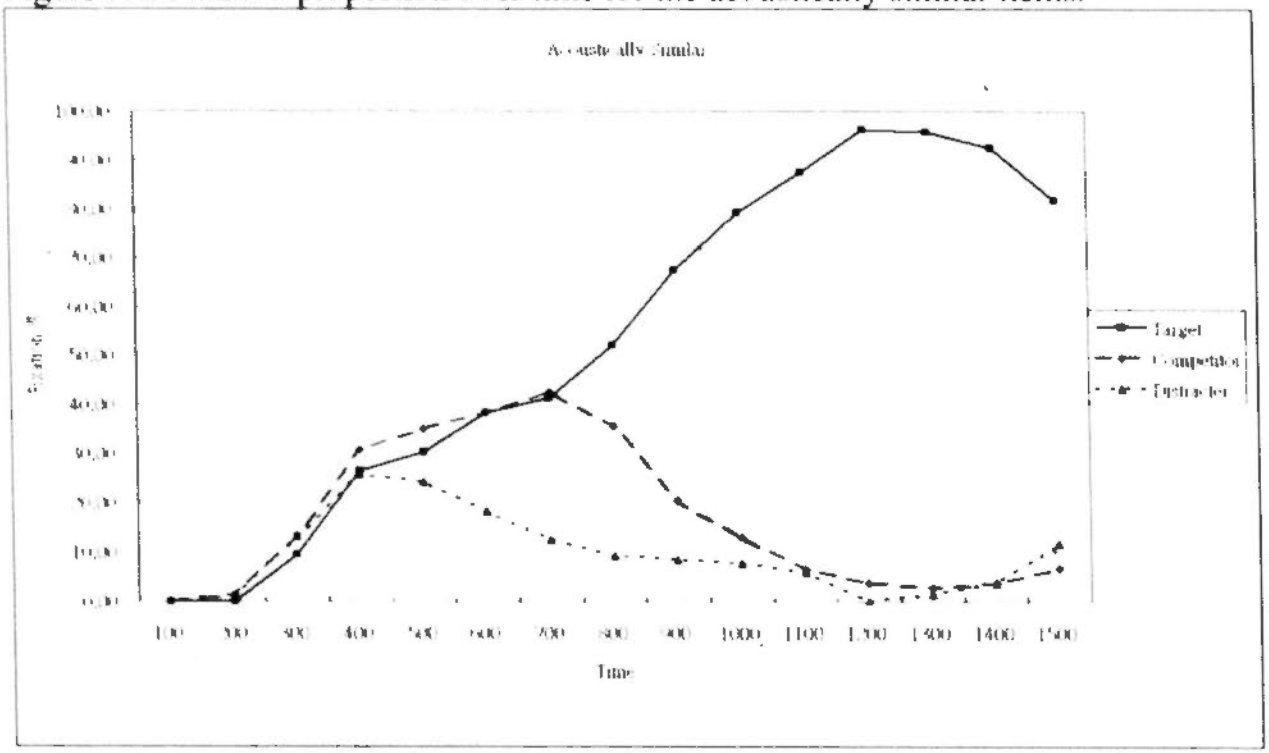
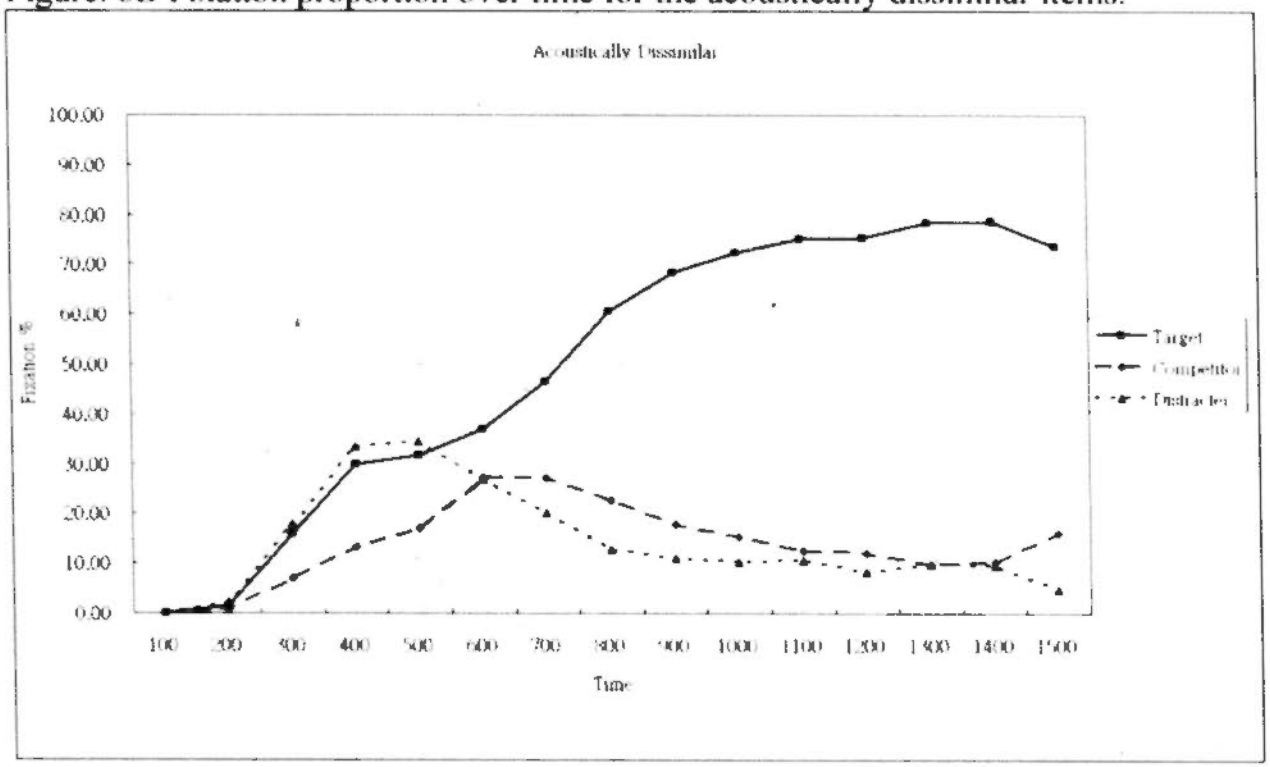


Figure. 8B Fixation proportion over time for the acoustically dissimilar items.



As expected, for the acoustically similar pairs, there was a long period in which the competitors were indiscriminable from the targets. In contrast, the activation level for competitors was much lower than the targets throughout the whole time window. Table 10 presents the mean fixation proportions (standard deviations) on each type of picture in different time windows separated by acoustic similarity.

Table 10. Mean fixation proportions (%) and standard deviations (in parentheses) separated by acoustic similarity.

	Acoustically similar			Acoustically dissimilar		
	T	C	D	T	C	D
201 – 500 ms	22.18 (19.93)	26.52 (34.00)	21.10 (26.49)	25.74 (30.51)	12.43 (20.71)	28.56 (28.96)
501 – 900 ms	49.80 (23.77)	34.24 (28.35)	12.18 (10.30)	53.11 (22.69)	23.69 (18.52)	17.79 (14.18)
201 – 900 ms	37.96 (19.34)	30.93 (24.99)	16.00 (14.51)	41.38 (22.96)	18.86 (14.45)	22.40 (17.58)

Note: T = Target picture; C = Competitor picture; D = Unrelated control distracter picture.

Despite obvious difference between the two curves, the interactions between similarity and picture type on fixation proportion were not significant (all *p*s > .1). To be consistent with previous analyses, we also looked into the details for each individual curve. However, interpretations should be cautious because this post-hoc analysis was done on unbalanced items.

Acoustically similar condition. There was no early difference (201 – 500ms) in fixation proportions across the three types of pictures ($F2_{(2, 12)} < 1$), supporting the early confusion seen from the curve. In the late time window (501 – 900ms), the picture type effect was only marginally significant ($F2_{(2, 33)} = 3.42$, $MSE = 731.00$, $p = .067$). The target pictures were fixated significantly more than the unrelated distracters ($t2_{(6)} = 3.89$, $p < .05$). However, the 22% different between competitors

and distracters was just approaching significance ($t2_{(6)} = 1.75$, $p = 0.13$). Finally, no significant results were found in the overall time window ($F2_{(2, 12)} = 1.49$).

Acoustically dissimilar condition. The apparent differences in the early window turned out to be non-significant ($F2_{(2, 12)} = 2.03$, $p > .1$). In the later time window (501 900ms), the picture type effect was significant ($F2_{(2, 33)} = 19.82$, $MSE = 523.78$, $p < .01$). Targets were having higher fixation proportions than competitors and distracters ($ps < .01$). The same pattern was true for the overall time window ($F2_{(2, 33)} = 8.31$, $MSE = 511.82$, $p < .01$). Targets were again fixated more than competitors and distracters ($ps < .05$).

Table 11 presents the summary of the fixation proportions for each curve. Generally, there was a considerable period of confusion between targets and the corresponding competitors when they shared acoustic features. In contrast, the competitors could be differentiated from targets much more easily when they were dissimilar acoustically. In other words, consistent with recent demonstrations of subphonemic sensitivity in the perception of Indo-European speech, Chinese speakers also appeared to utilize fine acoustic details to constrain spoken word recognition.

Table 11. Summary of differences in fixation proportion across conditions.

Condition	201 – 500 ms	501 – 900 ms	201 – 900 ms
Acoustically similar	T = C = D	T = C > D *	T = C = D
Acoustically dissimilar	T = C = D	T > C = D	T > C = D

Note: T = target fixation proportion; C = competitor fixation proportion; D = unrelated distracter fixation proportion; * = only the difference between targets and distracters reached significance.

Specifically, Chinese speakers showed sensitivity to how well the incoming acoustic features fit with different onset-sharing candidates. In other words, acoustic inputs could vary in typicality towards a given phonological form in different words. For better fits, the corresponding phonemic representation could be activated more quickly, resulting in stronger competitor activations. For instance, perhaps due to co-articulation, the 'g' in 'gun2' was highly similar to the /g/ in /gu2/, but not the /g/ in 'geng2'. Participants would therefore found that /gu2/ was also supported by the acoustic inputs when 'gun2' was heard. As both /gun2/ and /gu2/ were activated, they competed for being recognized. The mutual inhibition might temporarily lead to a reduction of activation level for both candidates, compared with the mismatched distracters. And the closely matched pair would also require a longer time for disambiguation, producing the observable confusion in the target detection task. In contrast, /geng2/ would not be highly activated because of its acoustic mismatch with the typical /g/ present in /gun2/. Actually, the more highly activated /gun2/ might exert strong interference on /geng2/ at both the phonemic level and lexical level via lateral inhibition (e.g., McClelland & Elman, 1986). The strong lateral inhibition might explain why onset-sharing competitors usually were fixated even less than the unrelated distracters.

The acoustic constraints observed in the present experiments appeared to be even stronger than previous studies. For example, Salverda et al. (2003) discovered that syllable length could help disambiguate embedded words from carrier words, yet the effect "was modest and the disfavored competitor remained active for a substantial amount of time after the disambiguating information was available" (p.81). They concluded that acoustic cues operated more like recognition bias favoring a particular candidate without being actively involved in the elimination of alternative candidates.

We would defer the discussion of such discrepancy to the General Discussion. At the moment, it should be noted that in all onset-sharing condition, there was a delayed elevation in competitor fixations at some points in the curves. In other words, perhaps the acoustic constraint was just temporary in Chinese speech. Obviously, more works have to be done to illuminate the importance of subtle acoustic details in Chinese speech perception.

In any case, the effect of acoustic similarity in the present experiment suggested that embedded words posed challenges to Chinese spoken word recognition through sharing acoustic features with the carrier words. Competitors that were not fully embedded in the targets (i.e., the onset-plus condition) could still be highly activated as long as they received acoustic support from the physical inputs. The role of acoustic inputs echoed with speech perception models assuming phonetic features as the basic inputs into the system. Before discussing in more details about how the present findings could be related to these models, we conducted an extra experiment testing whether the availability of a concurrent visual display during spoken word recognition would be responsible for the patterns observed.

5.3 Experiment 2C – the role of concurrent visual display

As McMurray et al. (2008) noted, there were arguments around whether the presence of a concurrent visual display in the visual-world paradigm would encourage participants to adopt specific strategies, such as implicit naming, that were irrelevant to normal speech processing. Although there were direct evidence against such view (e.g., Dahan & Tanenhaus, 2005; Huettig & Altmann, 2005; Magnuson, Dixon, Tanenhaus, & Aslin, 2007), we conducted Experiment 2C to explore how the visual display might be related to the pattern observed in Experiment 2B. Specifically, in

contrast to the simultaneous presentation, we introduced a 300ms SOA between syllables and visual displays. We expected that, the delayed appearance of visual displays would discourage participants from implicitly naming the objects in the display and then matching the names with the auditory inputs. If such implicit naming could actually account for the low competitor fixations observed previously, we should be able to see higher competitor fixations in the present experiment.

5.2.1 Participants

Fourteen undergraduates (five males) in The Chinese University of Hong Kong were recruited for this experiment. They were paid $50 for participation. All of them were native Cantonese speakers, reported no hearing deficits and had normal or corrected-to-normal vision. None of them had participated in the previous experiments. Informed consent was obtained and full debriefing was delivered after experiment.

5.2.2 Materials and Design

The same set of materials in Experiment 2B was used.

5.2.3 Procedure

Procedure was identical to Experiment 2B, except that the visual displays appeared 300ms after the onset of the spoken syllables.

5.2.4 Results and Discussion

Results were analyzed in the same way as in previous experiments. In 1.2% of trials, participants responded before having fixated on the target. These trials were

excluded from further analyses. The overall error rate was 13.89%, which was comparable to that in Experiment 2B (16.67%). Table 12 presents the mean reactions times (standard deviations) and error rates of target detection as a function of competitor conditions. The differences among conditions turned out to be statistically non-significant (all *ps* > .1).

Table 12. Mean reaction times and error rates (standard deviations in parentheses) of target detection in Experiment 2C.

Condition	Reaction time (ms)	Error rate (%)
Onset-sharing	1316 (229)	10.7 (15.5)
Onset-plus-sharing	1360 (244)	17.9 (21.1)
Embedded word	1296 (220)	13.1 (14.9)

Figures 9A to 9C shows the fixation proportion over time for each condition from 0 ms to 1500 ms. It is clear from the figures that we "successfully" replicated the low competitor fixations for onset-sharing and onset-plus-sharing competitors in previous experiments. On the other hand, the delayed activation of embedded competitors was also weaker than that seen in Experiment 2B, probably because when eye movements started to be influenced by the activation of embedded competitors[12], participants had already gathered some acoustic information that unambiguously pointed towards the real targets. To quantify the observed pattern, we again calculated the mean fixation proportion on each picture for different conditions. However, given the 300 ms delay of the visual displays, our time windows of interest also shifted 300 ms accordingly. In other words, the overall time window now spanned from 501 to 1200 ms, with the early and late windows from 501 to 800ms and 801 to 1200 ms respectively.

[12] It should be noted that competitors started attracting eye fixations at around 500ms in Experiment 2B, which was consistent with the time when competitors were fixated more than distracters in Experiment 2C (at 500+300 = 800ms).

Figure 9A. Fixation proportion over time for the onset-sharing condition.

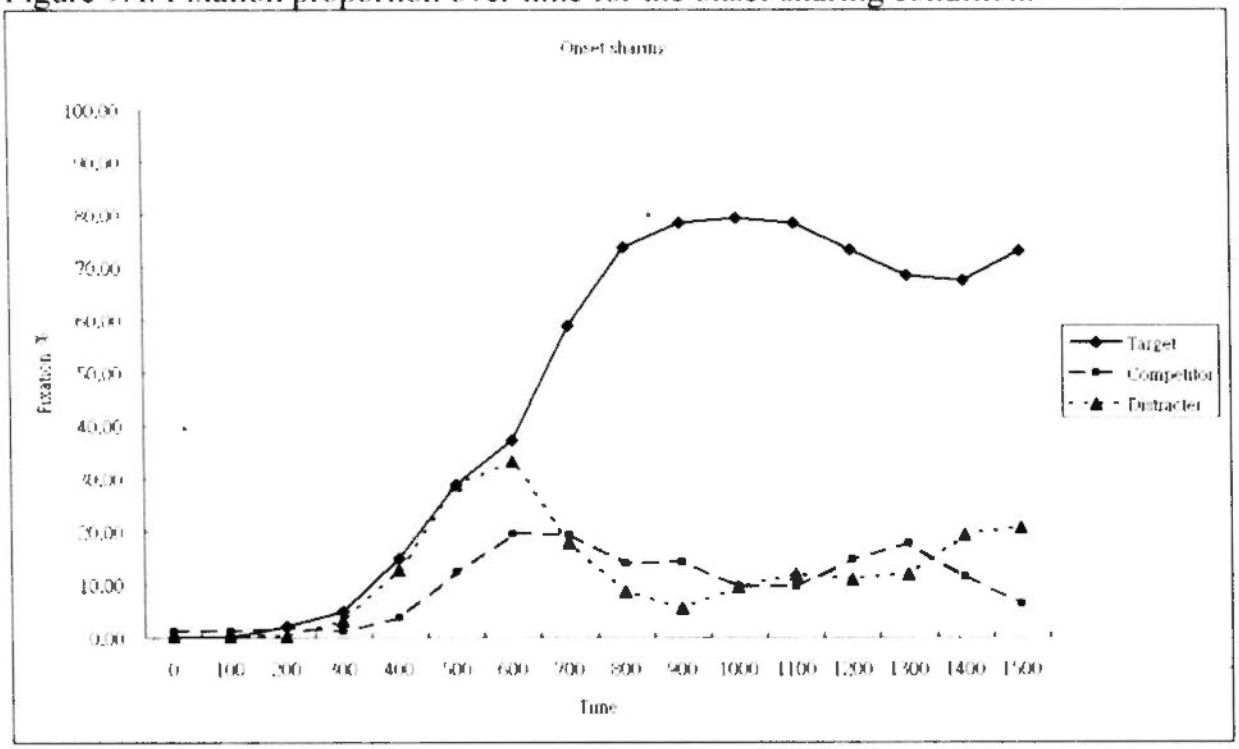


Figure 9B. Fixation proportion over time for the onset-plus-sharing condition.

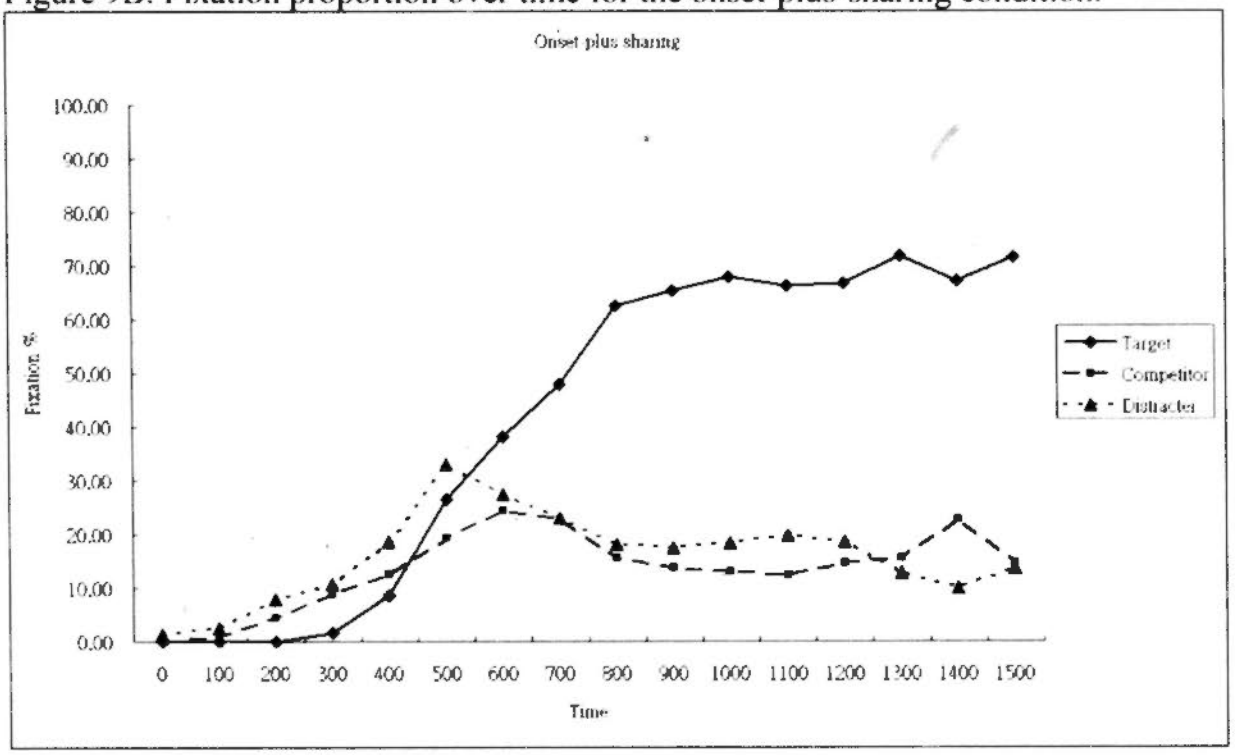


Figure 9C. Fixation proportion over time for the embedded condition.

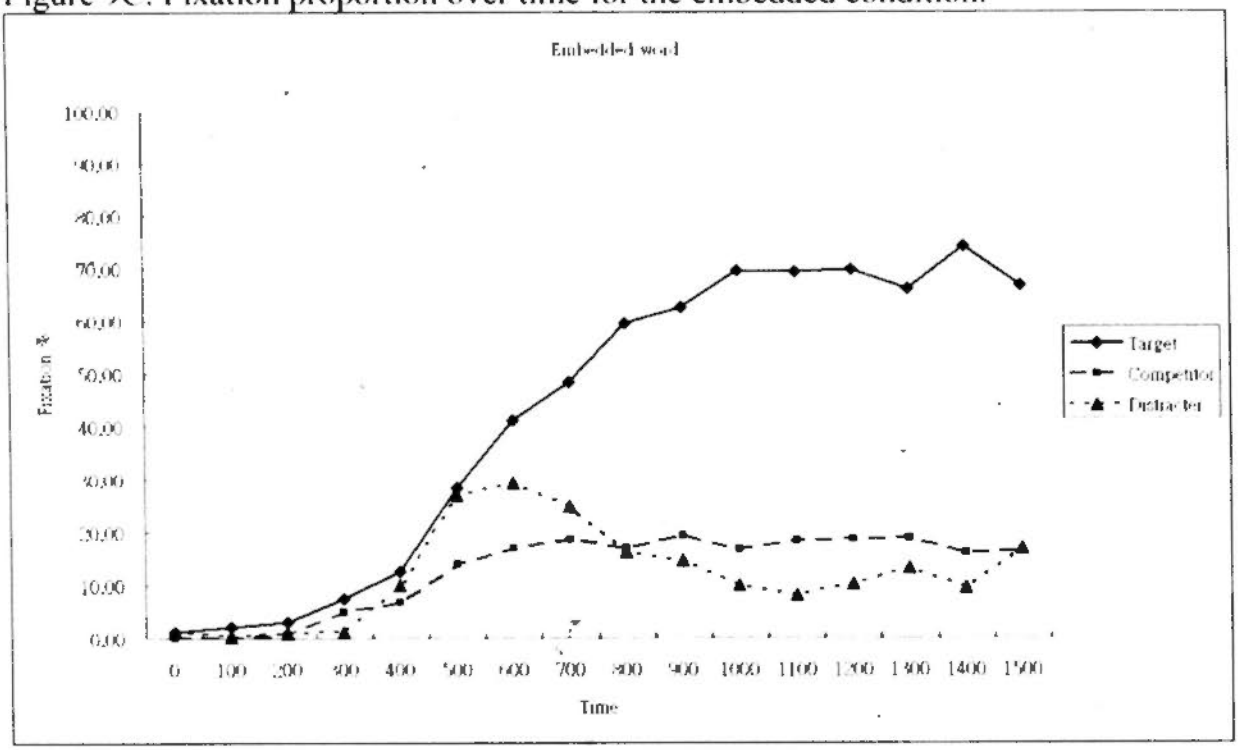


Table 13 displays the mean fixation proportions (standard deviations) on each picture in different conditions. The most important finding was the non-significant three-way interactions in the Experiment (2B vs. 2C) X Condition (onset vs. onset-plus vs. embedded word) X Picture type (target vs. competitor vs. distracter) mixed ANOVA (*ps* > .1 in all time windows)[13]. This supported our previous observation that having a 300ms-delay between speech onset and visual display did not significantly alter the relationship between degree of phoneme sharing and the pattern of competitor fixations. On the other hand, there were strong two-way interactions between Experiment and Picture type ($F1_{(2, 64)}$ = 8.29 to 16.57; $F2_{(2, 66)}$ =9.76 to 22.91, all *ps* < .01). This interaction obviously emerged due to the higher target fixation proportion in Experiment 2C than 2B. To provide more information, and to facilitate comparisons with previous experiments, we again conducted a series of analyses separately for each condition.

[13] Interaction was also non-significant when time windows were defined with reference to word onset.

Table 13. Mean fixation proportions (%) and standard deviations (in parentheses) across conditions in Experiment 2C.

	Onset-sharing competitor condition			Onset-plus-sharing competitor condition			Embedded word competitor condition		
	T	C	D	T	C	D	T	C	D
501 – 800ms	39.28 (15.87)	16.55 (7.86)	27.11 (13.95)	41.71 (24.74)	17.00 (14.70)	26.73 (11.49)	37.48 (27.59)	22.18 (15.23)	27.79 (20.38)
801 – 1200ms	65.17 (17.42)	17.88 (14.29)	12.28 (9.45)	77.48 (12.74)	11.80 (8.91)	8.97 (6.58)	65.49 (17.36)	13.65 (8.98)	18.38 (13.00)
501 – 1200ms	54.08 (13.91)	17.31 (9.08)	18.63 (7.57)	62.15 (13.422)	14.03 (8.84)	16.58 (6.05)	53.49 (17.20)	17.31 (7.91)	22.41 (12.76)

Note: T = Target picture; C = Competitor picture; D = Unrelated control distracter picture.

Onset-sharing competitor condition. Within the time window 501 800 ms, the picture type main effect was significant by subject ($F1_{(2,26)}$ 5.21, *MSE* 416.40, p .05; $F2_{(2,22)}$ 1.60, *n.s.*). Pairwise comparisons with LSD correction revealed lower fixations on competitor than on targets and distracters ($t1_{(13)}$ 2.72 and 2.47 respectively, ps .05). In the later time window (801 1200 ms), however, picture type effect was robust ($F1_{(2,26)}$ 152.44, *MSE* 138.01, p .001; $F2_{(2,22)}$ 79.09, *MSE* 236.33, p .01). Targets were significantly fixated more than both the competitors and distracters (all ps .01). Overall (501 1200 ms), participants also fixated more on targets ($F1_{(2,26)}$ 80.25, *MSE* 127.88, p .001; $F2_{(2,22)}$ 28.05, *MSE* 322.83, p .01) than on competitors (ps .01) or distracters (ps .01). In other words, generally the targets could be reached quite easily and there was no competitor activation over distracter. In contrast to Experiment 2B, although distracters also appeared to be fixated more frequently than competitors, the difference was only marginal. This could be attributed to the fact that, even in the earliest time window concerned, participants had already received some inputs necessary to recognize the actual targets.

Onset-plus-sharing competitor condition. The small difference in fixation proportions among different pictures was not statistically significant ($F1_{(2,26)}$ 1.32; $F2_{(2,22)}$ 0.62) in the early time window. In contrast, the effect was robust in the later time window ($F1_{(2,26)}$ 42.42, *MSE* 271.16, p .01; $F2_{(2,22)}$ 16.18, *MSE* 605.66, p .01). Again, targets were fixated more than both the competitors and distracters (all ps .05). In the overall time-window, the picture type effect was significant ($F1_{(2,26)}$ 22.44, *MSE* 239.25, p .01; $F2_{(2,22)}$ 8.70, *MSE* 527.41, p .01). Targets were fixated more than competitors (ps .01) and distracters (ps .05). This pattern is

consistent with the one observed in Experiment 2B and the onset-sharing condition in Experiment 2C.

Embedded word competitor condition. Similar to Experiment 2B, in the first time window, participants fixated less on the embedded word competitors ($F1_{(2,26)}$ 12.00, MSE 150.85, p .01; $F2_{(2,22)}$ 1.90, *n.s.*) than both the targets and distracters (*ps* .05). In contrast, the significant picture type effect ($F1_{(2,26)}$ 43.94, MSE 269.02, p .01; $F2_{(2,22)}$ 31.22, MSE 380.75, p .01) in the late time window (801 1200 ms) could be fully attributed to the higher fixation proportions for targets than competitors or unrelated distracters (all *ps* .01). Unlike Experiment 2B, however, the higher competitor fixation over distracter did not approach significant. Finally, similar to the other two conditions, fixation proportions differed among the three pictures in the overall time window ($F1_{(2,26)}$ 53.99, MSE 112.76, p .01; $F2_{(2,22)}$ 16.51, MSE 374.75, p .01). Fixation on targets was significantly more than that on competitors or distracters (all *ps* .01).

Table 14 presents a brief summary of the fixation proportions on different pictures in each condition and time window. An alternative account to the low competitor fixations in Experiment 2B rested on the arguments that performance in the visual-world paradigm involved strategic processing. According to this view participants would implicitly name the objects in the visual display and matched the object names with incoming acoustic inputs. As a result, participants would be highly sensitive to even the slightest acoustic mismatches. The competitors were supported less and received stronger lateral inhibition from the partially supported targets[14]. If

[14] It should be noted that an increased sensitivity to acoustic mismatch of competitors due to implicit naming was not equivalent to an increased sensitivity to acoustic match of target. The latter would need another mechanism to explain why target fixation could not emerge immediately. It also failed to explain why the target fixation was indeed higher with a delayed visual display.

such mechanism could indeed explain the low onset-related competitor fixation in the previous experiments, we should be able to obtain more fixations on competitor when implicit naming was prohibited by the delayed visual display. This prediction was not supported by the results in Experiment 2C. In general, the pattern observed in this experiment is highly similar to that in Experiment 2B. Interactions involving Experiment and Condition were not significant. If anything, the only difference was that embedded word competitors were fixated even less than the previous experiment, which was directly contradictory to the prediction by strategic processing.

Table 14. Summary of differences in fixation proportion across conditions.

Condition	501 – 800ms	801 – 1200ms	501 – 1200ms
Onset-sharing	T = D > C	T > C = D	T > C > D
Onset-plus-sharing	T > C > D	T > C = D	T > C = D
Embedded word	T > D > C	T > C > D	T > C > D

Note: T = target fixation proportion; C = competitor fixation proportion; D = unrelated distracter fixation proportion.

On the other hand, the lower fixation proportion on embedded word competitors in Experiment 2C than in 2B could be explained easily by assuming that participants had gathered sufficient acoustic information to identify the targets unambiguously (or at least confidently) at the time they could start searching from the visual display given the 300ms delay. The target was thus less ambiguous and could attract eye fixations earlier and more strongly. This account could also accommodate the overall faster reaction time and higher target fixation rates in Experiment 2C.

Finally, in order to provide a rigorous test to the implicit naming hypothesis, we divided the experimental materials into acoustically similar and acoustically dissimilar items as in the experiment 2B. If implicit naming was responsible for the

acoustic sensitivity in our participants, we should obtain identical response pattern in acoustically similar or dissimilar items. The corresponding curves of fixation proportion over time are plotted in Figures 10A and 10B. Table 15 shows the mean fixation proportions (standard deviations) for each type of items.

Figure 10A. Fixation proportion over time for the acoustically similar items.

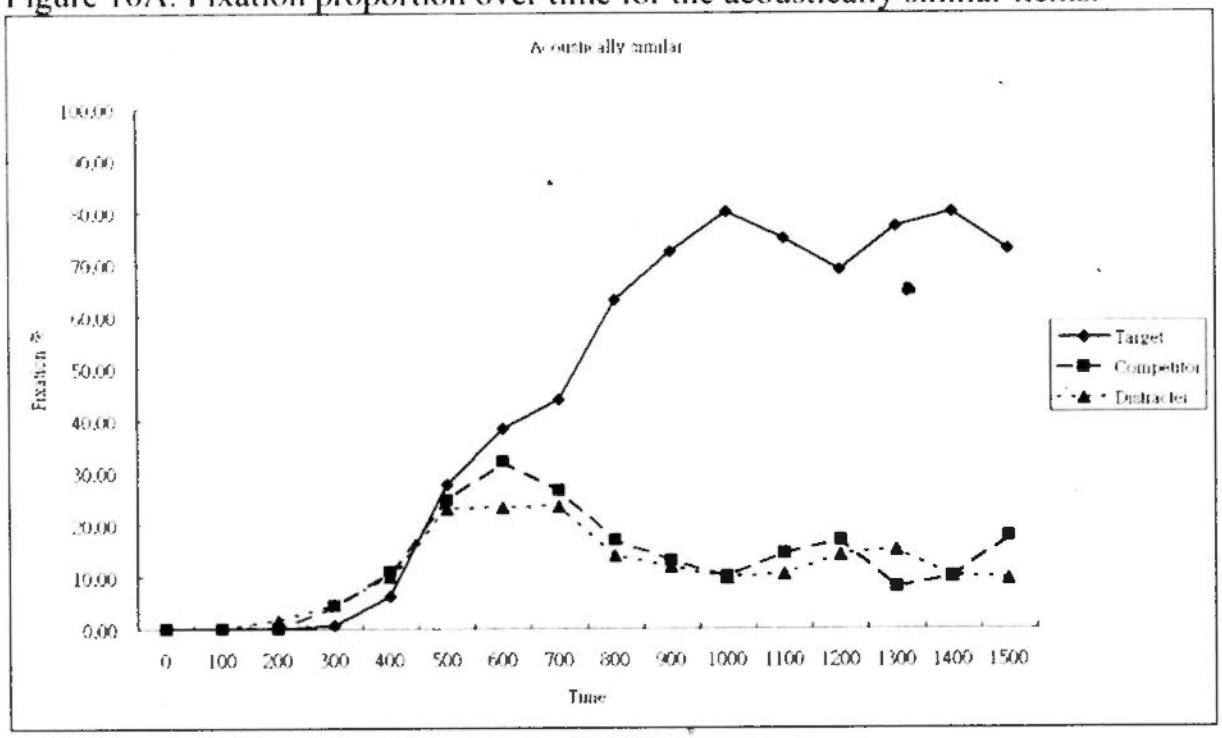


Figure 10B. Fixation proportion over time for the acoustically dissimilar items.

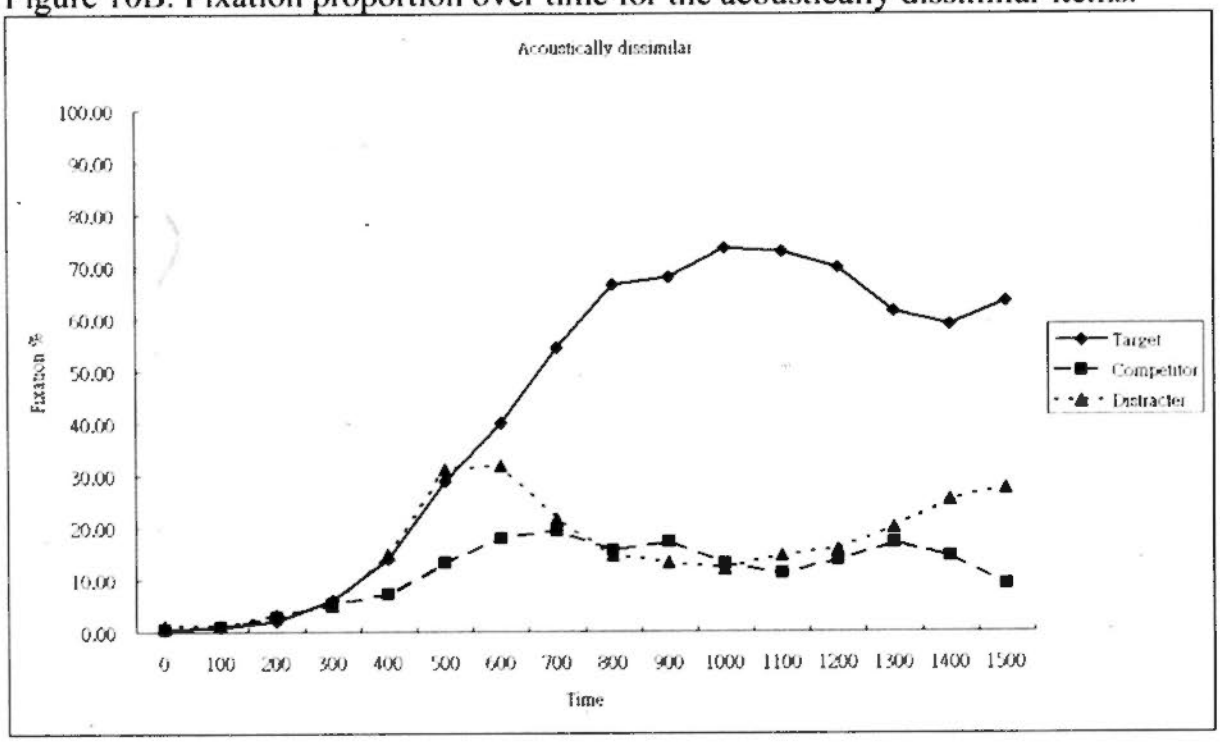


Table 15. Mean fixation proportions (%) and standard deviations (in parentheses) separated by acoustic similarity.

		Acoustically similar			Acoustically dissimilar		
		T	C	D	T	C	D
501	800 ms	36.67 (27.04)	27.77 (23.18)	23.16 (28.96)	41.16 (31.49)	16.76 (19.72)	28.24 (27.09)
801	1200 ms	72.56 (13.95)	13.61 (17.93)	11.54 (6.29)	70.40 (20.83)	14.25 (13.68)	13.41 (17.27)
501	1200 ms	57.18 (17.90)	19.68 (12.16)	16.52 (12.50)	57.88 (19.75)	15.29 (13.54)	19.76 (16.97)

Note: T = Target picture; C = Competitor picture; D = Unrelated control distracter picture.

The three-way interactions among Experiment (2B vs. 2C), Picture type (target vs. competitor vs. distracter), and Acoustic similarity (similar vs. dissimilar) were not significant (all $F2s_{(2, 68)} < 1$). This suggested that the 300ms delay of visual display did not significantly change the relationship between acoustic similarity and picture fixation observed in Experiment 2B. In contrast, as in the previous analysis, there were more target fixations in Experiment 2C than in 2B, as revealed by the significant Picture X Acoustic similarity interaction ($F2s_{(2, 68)} = 5.76$ to 16.48, all $ps < .01$). Further comparisons were done separately for each level of acoustic similarity.

Acoustically similar condition. There was no significant difference of picture type effect ($F2_{(2, 12)} < 1$) in the early time window (501 – 800 ms). In the late time window (801 – 1200 ms), there were more target fixations ($F2_{(2, 12)} = 30.39$, $MSE = 276.49$, $p < .01$) than competitors and distracters (all $ps < .01$). The difference between competitor and distracter was not significant. Targets were also fixated more ($F2_{(2, 12)} = 11.55$, $MSE = 310.15$, $p < .01$) than both competitors and distracters (all $ps < .05$) in the overall time window (501 – 1200 ms).

Acoustically dissimilar condition. Unlike the similar condition, there was a significant picture type effect even in the early time window ($F2_{(2, 56)} = 4.12$, MSE 1047.87, $p < .05$), which was attributed solely to the higher fixation proportion of target over competitor ($t2_{(28)} = 3.03$, $p < .01$). In the late time window, picture type effect was robust ($F2_{(2, 12)} = 67.56$, $MSE = 458.04$, $p < .01$). Fixation proportion on target was higher than that on competitors or distracters (all $ps < .01$). In the overall time window, again target was fixated more ($F2_{(2, 12)} = 37.14$, $MSE = 427.78$, $p < .01$) than competitors and distracters (all $ps < .01$).

Table 16 presents the summary of fixation proportions separated by acoustic similarity. The two patterns were similar, but there was a crucial difference: For the acoustically dissimilar items, there were more target fixations than competitor fixations in the early time window. However, it was not the case in acoustically similar items, in which participants were completely confused among targets, competitors and distracters. This result supported the prior analyses that strategic processing (such as implicit naming) alone could not account for the patterns observed in previous experiments. On the other hand, the present analysis and that in Experiment 2B (Table 11) also diverged in several ways. The primary difference was the stronger (acoustically similar items) and earlier (acoustically dissimilar items) target fixations. Another notable difference was the greatly attenuated difference between competitor and distracter fixations for acoustically similar items in the late time window (22ms vs. 2ms). The discrepancy across experiments could again be attributed to the lower ambiguity in target detection when the visual display was presented.

Table 16. Summary of differences in fixation proportion across conditions.

Condition	501 – 800ms	801 – 1200ms	501 – 1200ms
Acoustically similar	T = C = D	T > C = D	T > C = D
Acoustically dissimilar	T > C = D *	T > C = D	T > C = D

Note: T = target fixation proportion; C = competitor fixation proportion; D = unrelated distracter fixation proportion; * = only the difference between targets and competitors reached significance

5.4 Summary of Experiment 2

Employing the visual-world paradigm, three experiments were conducted to verify and extend the patterns observed in Experiment 1. Consistent with the results in gating, participants in Experiment 2A did not show any candidate activation based on tonal information alone. In contrast, they were highly sensitive to syllables such that syllable-sharing competitors were readily available and attracted eye fixations during target detection. On the other hand, syllabic-sensitivity did not preclude the influence from subsyllabic units. In the time window when rime was acoustically available (501 – 900 ms), it could activate the rime-sharing competitor immediately (see Allopenna et al., 1998 for identical results in English), suggesting that rime was also a valid unit in candidate generation and elimination during Chinese speech perception.

The most unusual finding in Experiment 2A was the lack of candidate generation based on onset. Actually, fixation proportion on onset-sharing competitors was even lower than that on unrelated distracters. Given that speech onset was physically available first in the spoken word, it should receive zero inhibition from other phonemes and could fully activate its corresponding candidates. Therefore, typical models of speech perception (Marslen-Wilson & Welsh, 1978; McClelland & Elman, 1986) would assume a stronger role of onset than rime (Tyler, 1984). More

importantly, in Experiment 1A we actually obtained results confirming such claim. The absence of onset-effect in Experiment 2A was thus inconsistent with the results in both previous studies in English and in Experiment 1A.

Experiments 2B and 2C were conducted to examine in more details the lack of onset effect. We speculated that a single phoneme sharing in Experiment 2A was too brief to allow facilitation to accumulate. Rather, as information starting from the second phoneme was inconsistent with the onset-competitor, the inhibition resulted would outweigh the initial facilitation. If this speculation was correct, we should obtain stronger competitor activation when the degree of phoneme overlapping was higher. Results in Experiment 2B partially supported this hypothesis. Fixation proportion on embedded competitors was higher than that on distracters in the late time window (501 – 900 ms). However, the same pattern was not observed in the early time window or in the onset-plus conditions. The fact that onset-plus-sharing competitors were not fixated more despite sharing the same number of phonemes with the actual targets as the embedded words suggest that fully embedded words were indeed special candidates in the course of speech perception. Perhaps the extra phoneme in the onset-plus competitors had created inhibition, which offset the facilitation by the shared phonemes. In this sense, Chinese speakers might be sensitive to individual phonemes as well, even though syllables were important and prominent units in Chinese speech (Chen, 2000; Chen et al., 2002; Cheung et al., 2001; McBride-Chang et al., 2004)

Interestingly, when we split the competitors used in Experiment 2B by their acoustic similarity to the targets, we observed competitor activation over distracter in the acoustically similar items (it should be noted that, however, strictly speaking, the difference failed to attain statistical significance). Although far from being conclusive,

this result appeared to suggest that Chinese speakers might be sensitive not only to subsyllabic units, but also to the more fine-grained subphonemic acoustic details in the speech signal. Indeed, such acoustic sensitivity fit well with recent evidence that speech perception was highly influenced by subphonemic features (e.g., Davis et al., 2002; Marslen-Wilson & Warren, 1994; McMurray et al., 2008; Salverda et al., 2003). Moreover, it provided a simple explanation for why the activation of embedded competitors appeared to be delayed: Early acoustic mismatch inhibited competitor activation, while later phonemic match reintroduced it back to the candidate set. In other words, acoustic and phonemic processes were two distinct but interacting systems. Initially, acoustic process dominated candidate generation and elimination but the abstract phonemic representations slowly took the control. Similar mechanisms have been proposed to explain results in orthographic processing. Specifically, it was well-established that words which looked similar (physically) would inhibit each others, unless they also shared abstract orthographic representation. For example, Ding, Peng, and Taft (2004) showed that in Chinese, primes which shared strokes with targets were inhibitory, while primes which shared radicals with targets were facilitatory.

Indeed, inspecting the fixation proportion over time curves for each onset related condition more closely, one would discover small (statistically non-significant) but consistent elevations of competitor fixation in a later time window. Presumably, this could also be explained by assuming that effects due to phonemic match lagged behind the acoustic effects. Such lag in phonemic influence also naturally accounted for the discrepancy between Experiments 1 and 2: Given the loosened time limit in responding in a gating paradigm, participants' responses had a higher chance to be

affected by the slowly emerging phonemes. They could thus produce responses that shared onset with the actual targets in their guesses.

Finally, Experiment 2C was conducted to ensure the effects observed in Experiment 2B could not be attributed to strategic processing in the visual-world paradigm. In this experiment, the visual display was presented 300 ms after speech onset. This minimized the chance of implicit naming. Yet, virtually identical results were obtained, except that participants were in general faster in arriving at the target. This pattern argued against the strategic account and validated the conclusions drawn from the previous experiments.

The results obtained in Experiment 2 had important implications on modeling Chinese speech perception. On the one hand, they provided empirical support to the modified TRACE model proposed by Ye and Connine (1999), but on the other hand, they also suggested that the model needed further revision to account for all findings. In particular, the model emphasized the role of syllables Chinese speech and its interaction with lexical tone: The constraining power of lexical tone would remain weak until enough syllabic feedback was received. Such emphasis was justified by the observation that syllable-sharing competitors in Experiment 2A were activated more early and more strongly than would be predicted by the addictive effect of onsets and rimes. Only at a later time point (501 – 900 ms) would tone be employed to differentiate the otherwise identical syllables.

However, the assumption that "tone information is a separate level of representation" (Ye & Connine, 1999, p. 619) similar to the phoneme level might be misleading. Candidate generation and elimination have been a standard component in speech perception models. And most existing models considered phonemes as the primary contributor of these processes. Treating lexical tones like phonemes would

imply similar tonal involvement in candidate generation and elimination. Although some earlier studies had reported positive evidence (e.g., Li & Yip, 1998), in Experiment 2A we failed to find any traces of candidate activation based on pure tone overlapping. Rather, as mentioned in the previous paragraph, lexical tone appeared to be more relevant in eliminating tone-mismatch syllable candidates. Therefore, it might be more appropriate to consider lexical tone as something qualitatively different from other segmental representations. We will go back to this issue in the General Discussion.

Consistent with models in Indo-European languages, Ye and Connine (1999) also assumed that acoustic features were mapped onto phonemes, which in turn activated the corresponding lexical representations. Although the differential activation patterns between onset-plus competitors and embedded word competitors (Experiment 2B) also seemed to support the role of phoneme, given the salience of syllables, one might need to consider the possibility that features could be directly mapped onto whole-syllable representations in Chinese speech.

A related issue concerned the bottom-up acoustic inputs. While no one would deny acoustic signals as the fundamental components in speech, most researchers implicitly assumed that mental operations worked on phonemes instead of acoustic features. Subphonemic acoustic variations were simply regarded as noises to the recognition system. For instance, in TRACE model (McClelland & Elman, 1986) acoustic variations due to coarticulation had to be "recovered" by lexical feedback for successful word recognition. This tradition of considering acoustic variations as noises probably rooted in the findings of categorical perception of phoneme, and gained nutrients from successful demonstration of phoneme effects in various experiments. However, actually more than 30 years ago there had been reports on

acoustic effect in gating paradigm (Ellis, Derbyshire, & Joseph, 1971, see also Warren & Marslen-Wilson, 1987). Yet, interests in conducting empirical (e.g., Davis et al., 2002; Marslen-Wilson & Warren, 1994; McMurray et al., 2008; Salverda et al., 2003) and theoretical (e.g., Gaskell & Marslen-Wilson, 1997; Norris, McQueen, & Cutler, 2000) works on how subphonemic features affected speech perception did not emerge until more recently. In these works, acoustic variations were not treated as noises, but as useful information for activating the appropriate words during recognition. For example, Salverda et al. (2003) proposed that a longer syllable duration helped recognition of embedded words from their carrier words.

The modified TRACE model in Chinese speech perception (Ye & Connine, 1999) also lacked specification about the role of acoustic details. Although far from being conclusive, results from Experiments 2B and 2C appeared to suggest that Chinese speakers may also be sensitive to acoustic match between targets and competitors. Therefore, a comprehensive model of Chinese speech perception would have to include a more detailed description of the roles acoustic information played. A possible model will be proposed in the General Discussion

Chapter 6

Experiment 3 – Morphemic ambiguity in Chinese homophones

An experiment using the visual-world paradigm was conducted to investigate meaning retrieval of homophones in spoken Chinese. Specifically, we tested the effects of two factors, namely relative meaning frequency and context position, on ambiguity resolution of the homophonic morphemes. This study served as the replication and extension of a previous study on homonymic morphemes (Tsang, 2006). Results obtained in this experiment would provide important insights on how fluent comprehension was possible in Chinese speech regardless of the frequent ambiguities encountered.

6.1 Experiment 3 – the role of meaning frequency and context

In this experiment, we tested the resolution of homophonic morpheme in Chinese disyllabic words with the visual-world paradigm. Because multiple objects could be presented in the visual display, it allowed us to monitor the activation levels of different meanings simultaneously. This provides a direct test for the multiple meaning activation in exhaustive access (Swinney, 1979) and reordered access (Duffy et al., 1988). Moreover, having a high temporal resolution, the paradigm is also useful in tracking the changes in activations level of each meaning during ambiguity resolution. Employing the same technique, Tsang (2006) found that both meaning frequency and the presence of prior context could affect the ambiguity resolution of homonymic morphemes during target detection. Specifically, when no prior context was available, both dominant and subordinate meanings were always activated, but the dominant one could be activated earlier and to a higher level. On the other hand, when biasing context preceded the ambiguous morpheme, the subordinate meaning

could be as available as the dominant one. In short, the overall pattern fits well with the reordered access model, which was originally proposed to explain lexical ambiguity resolution. The similarity in resolving morphemic and lexical ambiguity suggested that perhaps a common system governed the correct meaning retrieval under uncertainty. Following this, we expected the resolution of homophonic morphemes would follow the same principles. In other words, we predicted that the dominant meaning would again be more available, and the presence of prior context could boost the activation level of the subordinate meaning.

6.1.1 Participants

Twenty-four undergraduates (eleven males) in The Chinese University of Hong Kong were recruited. They were paid $50 for participation. All of them were native Cantonese speakers, reported no hearing deficits and had normal or corrected-to-normal vision. None of them had participated in the previous experiments. Informed consent was obtained and full debriefing was delivered after experiment.

6.1.2 Materials and Design

Design and materials were identical to those used in Experiment 1B (see also Appendix B). As described in Experiment 1B, properties of the experimental items were matched across different conditions (SD: succeeded context-dominant meaning, SS: succeeded context-subordinate meaning, PD: preceded context-dominant meaning, PS: preceded context-subordinate meaning). Each trial was composed of a target (dominant or subordinate), a competitor (the alternative meaning), and a distracter (items in other sets). For example, a valid trial in the SD condition contained the dominant meaning target "風箏" (/fung1 zaang1/; kite), the subordinate meaning

competitor "蜂巢" (fung1caau4 ; honeycomb), and the unrelated distracter "貝殼" (bui3hok3 ; shell), which was actually the items in other sets.

The twenty sets of experimental items were arranged into four blocks such that within a block, each item and the target syllable only appeared once, and there were five trials for each condition. Across blocks, all items appeared three times (as target, competitor, and distracter) and each target syllable was paired with contextual syllables from all conditions. Eighty filler items that required "no" responses in target detection were also prepared. Each object in the filler trials also appeared three times to resemble the composition of experimental trials. These filler trials were split into four and added to each block. The forty items in a block were presented in random order. Eight practice trials were prepared. The preparation of spoken disyllabic words and visual displays followed the same procedure in previous experiments.

6.1.3 Procedure

Each participant performed target detection for all four blocks of items. Each block began with calibration of eye-tracker. Between blocks, there was a short break. Block order was counterbalanced across participants. Other aspects of the procedure were identical to those in Experiment 2A. The whole experiment lasted for an hour.

6.1.4 Results and Discussion

In 0.56% of all trials, participants responded without having any fixations on the targets. These trials were discarded. Trials with incorrect responses or without responses within three seconds were coded as errors and excluded from further analyses (9.74%). Table 17 presents the mean reaction times (standard deviations) and error rates of target detection for the remaining data in each condition. Although

the mean duration of the disyllabic words (about 1200ms) was longer than the monosyllabic words in previous experiments, target detection times were comparable. There were significant interactions between meaning frequency and context position in both reaction times ($F1_{(1,24)}$ 5.37, MSE 14928.22, p .05; $F2_{(1,19)}$ 1) and error rates ($F1_{(1,24)}$ 13.35, MSE 0.002, p .01; $F2_{(1,19)}$ 1.54, *n.s.*). Pairwise comparisons suggested that dominant meaning produced faster responses than subordinate meaning only when no prior context was available ($t1_{(24)}$ 2.04, p .053). In contrast, dominant meaning produced less errors than subordinate meaning only when prior context was available ($t1_{(24)}$ 4.05, p .01). Collectively, these results suggested that the dominant meaning of the homophonic morphemes was more available in target detection.

Table 17. Mean reaction times and error rates (standard deviations in parentheses) of target detection in Experiment 3.

Condition	Reaction time (ms)	Error rate (%)
Succeeded context-Dominant meaning	1463 (180)	12.29 (6.91)
Succeeded context-Subordinate meaning	1555 (181)	11.25 (8.37)
Preceded context-Dominant meaning	1551 (170)	4.58 (4.87)
Preceded context-Subordinate meaning	1527 (149)	10.83 (7.76)

Next, following Experiment 2, we plotted the fixation proportion over time for each condition from 0ms to 1800 ms after spoken word onset (Figures 11A to 11D). Striking differences can be observed when comparing the fixation patterns between conditions with (PD and PS) and without prior context (SD and SS). Within each context position, the dominant meaning also appeared to be more activated earlier

when it was the target (PD vs. PS) and also remain active for a longer period when it was the competitor (SD vs. SS).

Figure 11A. Fixation proportion over time for the SD condition.

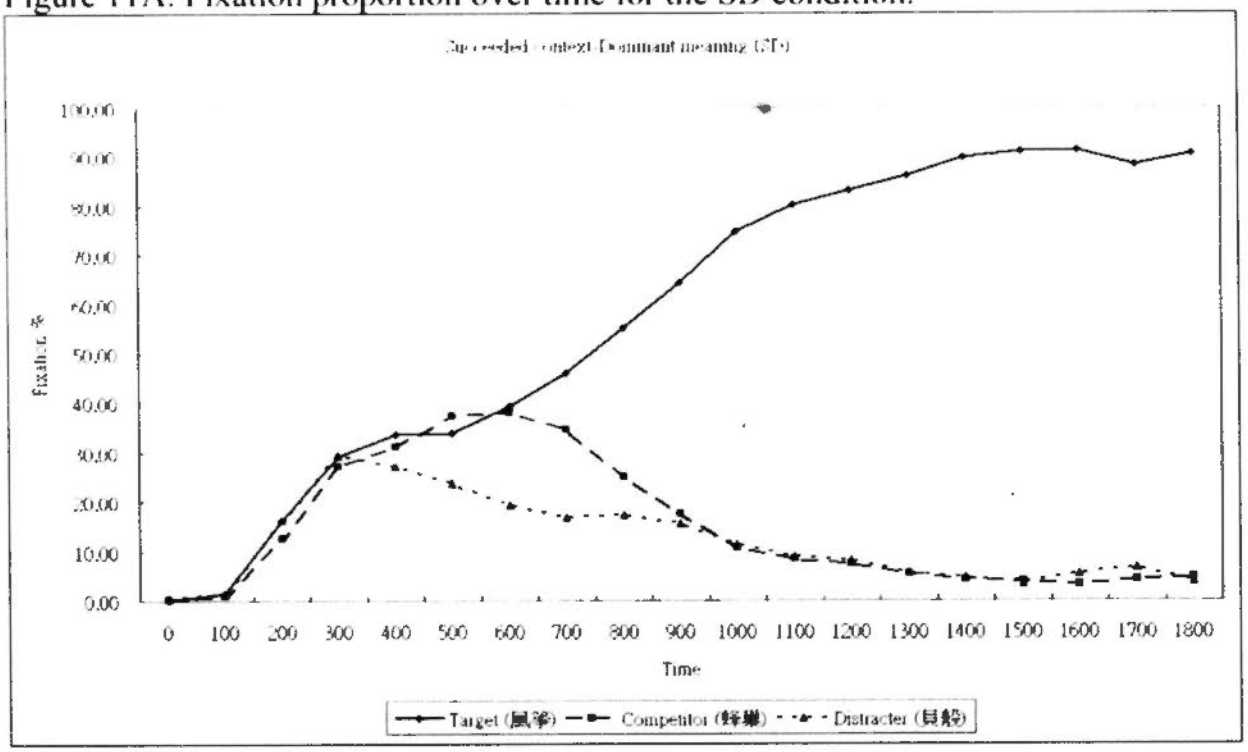


Figure 11B. Fixation proportion over time for the SS condition.

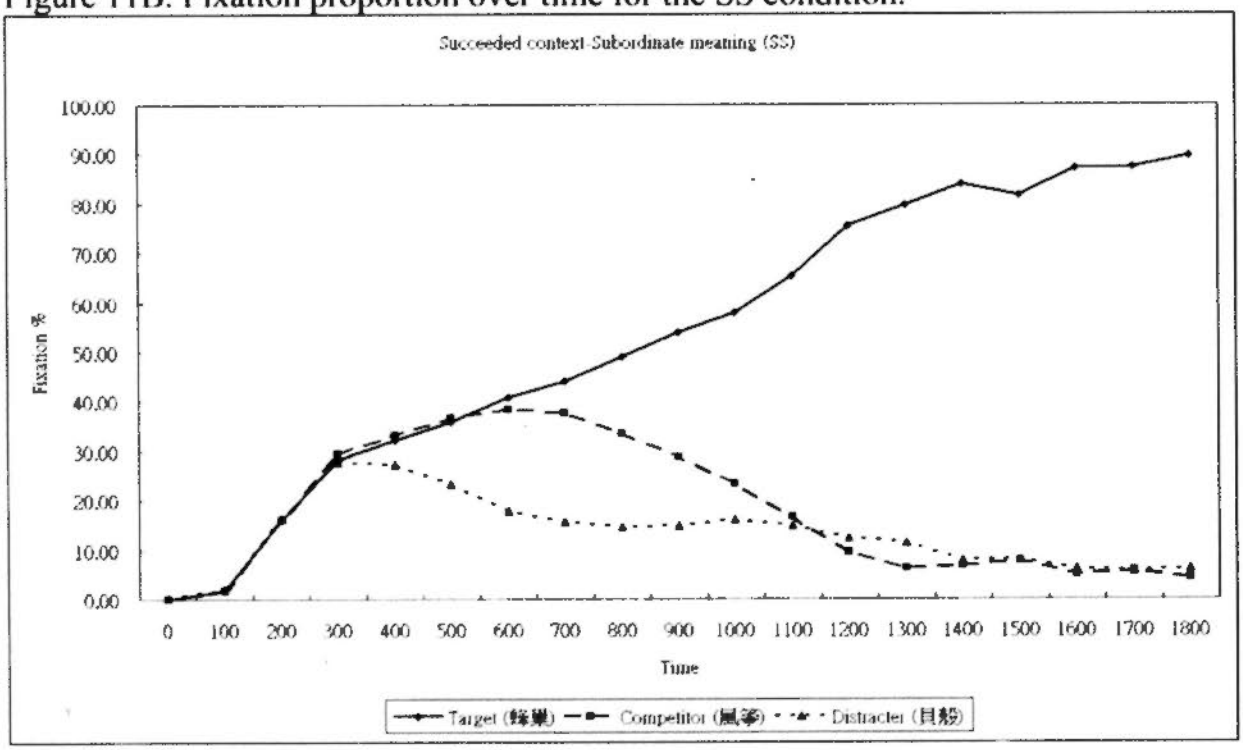


Figure 11C. Fixation proportion over time for the PD condition.

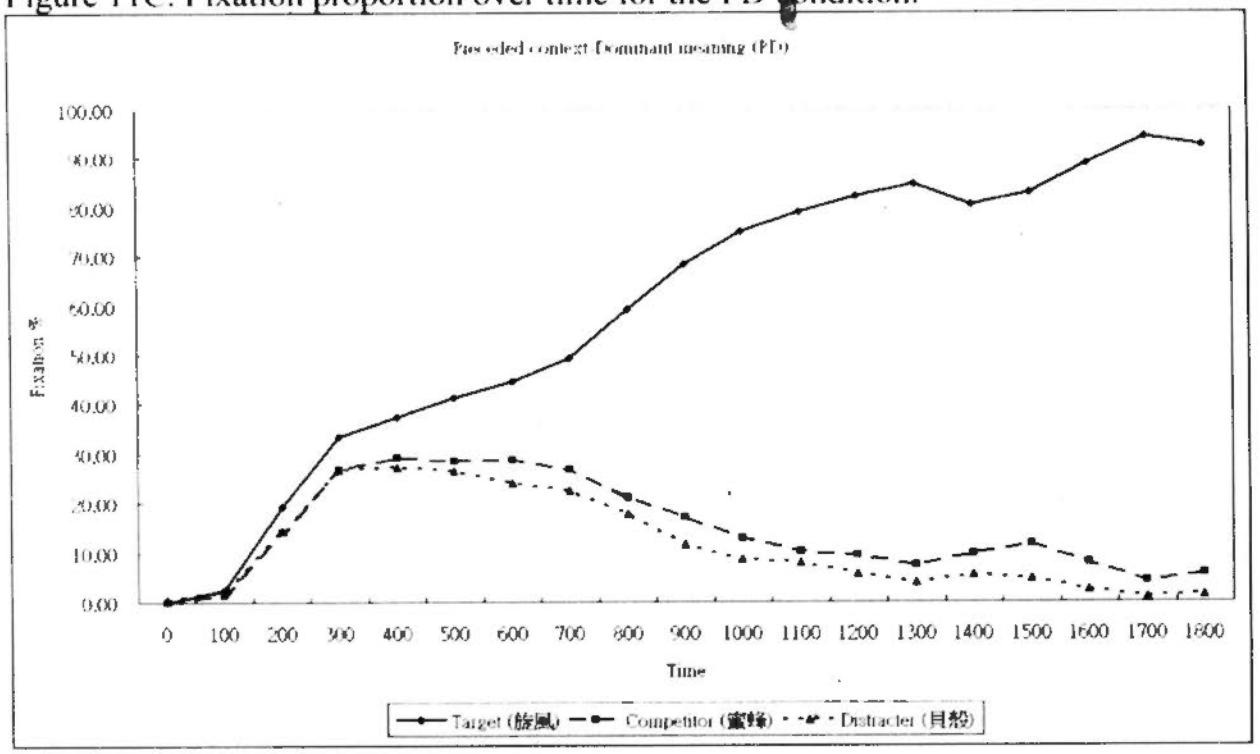
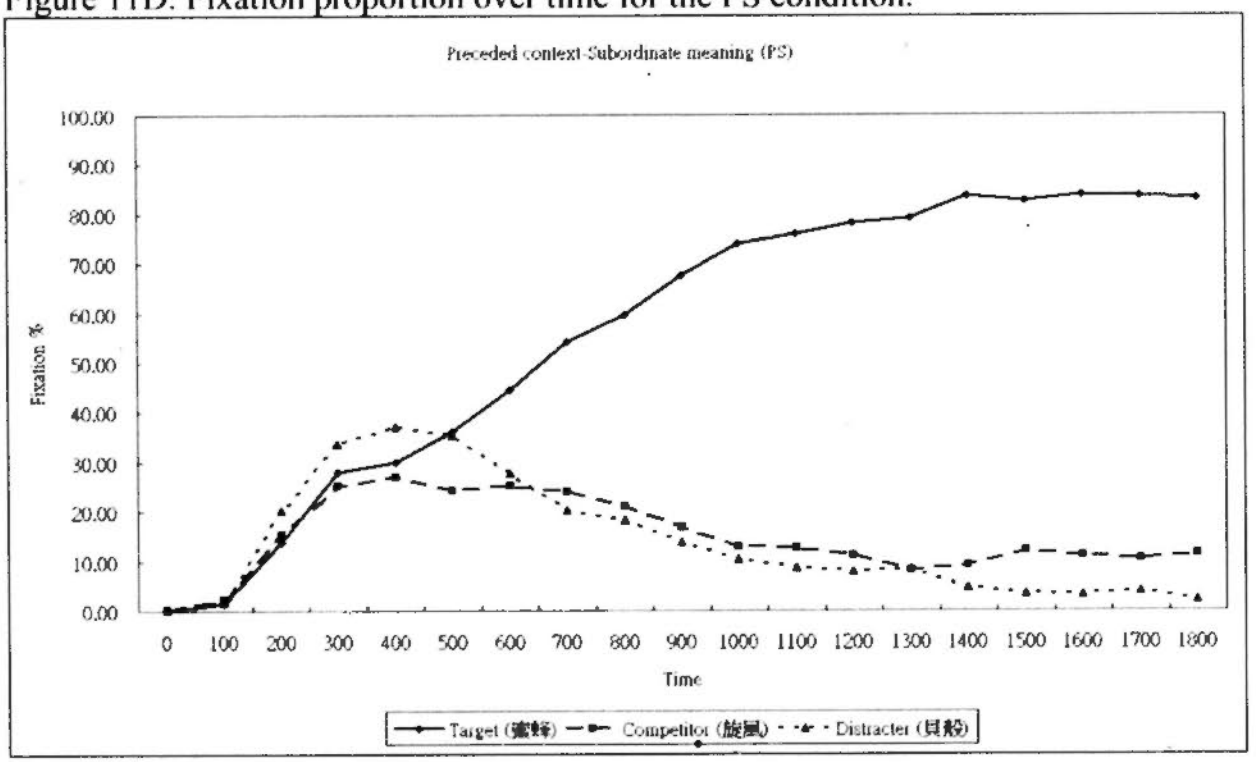


Figure 11D. Fixation proportion over time for the PS condition.



To verify the observation with statistical analyses, we again collapsed the fixation proportions on different pictures in each condition. Following Tsang (2006), we extracted the time windows of interest by syllable durations. Given the typical

200 ms delay for saccade planning and execution, the first and second syllable time windows were defined as 201 800 ms and 801 1400 ms respectively. We further included a time window of 201 400 ms to look at the early differences in the PD and the PS conditions. The mean fixation proportions (standard deviations) on each picture in each condition are presented in Table 18. The three-way interaction among Relative Meaning Frequency (dominant vs. subordinate), Context Position (succeeded vs. preceded), and Picture Type (target vs. competitor vs. distracter) was significant in the first syllable time window ($F1_{(2,46)}$ 4.90, MSE 37.90, p < .05; $F2_{(2,38)}$ < 1) and in the second syllable time window ($F1_{(2,46)}$ 6.20, MSE = 51.23, p < .01; $F2_{(2,38)}$ 2.76; p .076). We conducted further comparisons in each condition to study the origin of interactions.

Succeeded context-Dominant meaning. In the first syllable time window, there was a significant picture type main effect ($F1_{(2,46)}$ = 20.26, MSE = 37.75, p < .01; $F2_{(2,38)}$ < 1). While the dominant target and the subordinate competitor did not differ in fixation proportions, both of them were fixated more than the unrelated distracter (ps < .01). In the second syllable time window, the activation of the dominant target continued to rise ($F1_{(2,46)}$ = 523.17, MSE = 59.10, p < .001; $F2_{(2,38)}$ =391.36, MSE = 66.17, p < .001) over both competitors and distracters (ps < .01). On the other hand, activation of the subordinate competitor dropped back to the baseline distracter level (ps > .1). In brief, without a prior context, both dominant and subordinate meanings were equally available upon encountering the homophonic morpheme. However, once context was available, it exerted strong constraints so that only the context-fit. dominant meaning remained active.

Succeeded context-Subordinate meaning. Again, there was a significant picture type effect in the first syllable time window ($F1_{(2,46)}$ = 32.97, MSE = 29.74, p < .01; $F2_{(2,38)}$ = 1.25, *n.s.*). Similar to the dominant meaning condition, when no prior context was available, both dominant and subordinate meanings were equally available, both of which were fixated more than the unrelated distracters (*p*s < .01). In the second syllable time window, targets were robustly fixated more ($F1_{(2,46)}$ = 349.57, MSE = 50.83, p < .001; $F2_{(2,38)}$ = 70.30, MSE = 224.59, p < .01) than competitors (*p*s < .01) and distracters (*p*s < .01), indicating that context effect was strong and immediate even when the subordinate meaning was requested. More importantly, however, was the significantly higher competitor fixation than distracter fixation ($t1_{(23)}$ = 3.79, p < .01; $t2_{(19)}$ = 1.61, p > .1). Such prolonged activation of dominant meaning competitors stood in sharp contrast to the rapid deactivation of subordinate competitors shown in the previous paragraph. Actually, this was consistent with the hypothesis that the dominant meaning was more available even when context biased towards the subordinate reading.

Table 18. Mean fixation proportions (%) and standard deviations (in parentheses) across conditions in Experiment 3.

	succeeded context-dominant meaning			succeeded context-subordinate meaning			preceded context-dominant meaning			preceded context-subordinate meaning		
	T	C	D	T	C	D	T	C	D	T	C	D
201 – 400 ms							26.32 (7.50)	20.51 (6.53)	20.63 (6.81)	21.01 (7.25)	20.32 (7.06)	26.91 (8.92)
201 – 800 ms	33.05 (6.24)	30.15 (6.32)	22.15 (4.35)	32.87 (5.25)	31.90 (5.77)	21.34 (4.78)	37.49 (6.16)	25.66 (5.11)	23.61 (6.61)	34.49 (5.94)	23.50 (5.41)	29.04 (6.23)
801 – 1400 ms	73.99 (8.59)	12.44 (4.71)	11.22 (5.18)	63.63 (8.33)	19.53 (5.10)	13.97 (4.41)	74.68 (8.70)	13.00 (3.59)	9.26 (3.96)	72.44 (10.50)	13.77 (7.02)	11.28 (5.91)

Note: T = Target picture; C = Competitor picture; D = Unrelated control distracter picture.

Preceded context-Dominant meaning. In contrast to the previous two "late context" conditions, participants converged on the target much quicker in this condition. We first analyzed the fixation proportions in the 201 400 ms time window. Participants had already fixated more on targets ($F1_{(2,46)}$ = 10.82, MSE = 24.46, p < .01; $F2_{(2,38)}$ < 1) than both the competitors and distracters ($t1_{(23)}$ = 4.40 and 3.70 respectively, ps < .01). Similarly, in the first syllable time window, participants fixated on the targets ($F1_{(2,46)}$ = 35.79, MSE = 37.65, p < .01; $F2_{(2,38)}$ 2.51, MSE 507.27, p = .095) significantly more than both the competitors and distracters ($t1_{(23)}$ = 7.96 and 6.79 respectively, ps < .01; $t2_{(19)}$ = 1.61 and 1.96, ps > .05). The superior target fixation was maintained in the second syllable time window ($F1_{(2,46)}$ = 692.38, MSE = 46.78, p < .001; $F2_{(2,38)}$ = 168.42, MSE 162.01, p < .01). More interestingly, while targets were fixated more the competitors (ps < .01), competitors were also fixated more than the control distracters ($t1_{(23)}$ = 3.38, p < .01; $t2_{(19)}$ = 1.20, p > .1). In other words, even with a prior context that biased towards the dominant meaning, and even when the dominant meaning was indeed retrieved very strongly and quickly, the subordinate meaning was not completely eliminated. This provided the clearest evidence against the assumption of absolute contextual constraint: Meaning consistent with context would be facilitated. But the alternative meaning would still be available, at least temporarily. In short, there was multi-meaning activation as proposed in exhaustive access (Swinney, 1979) and reordered access (Duffy et al., 1988).

Preceded context-Subordinate meaning. When a prior context was biased towards retrieving the subordinate meaning, there was a brief period (201 – 400 ms) of complete confusion, in which distracters ($F1_{(2,46)}$ = 5.72, MSE = 55.18, p < .01; $F2_{(2,38)}$ < 1) were fixated more than the targets and competitors ($t1_{(23)}$ = 2.91 and 2.84

respectively, $ps < .01$). However, participants could go back to the target very quickly in the first syllable time window ($F1_{(2,46)} = 17.15$, $MSE = 42.10$, $p < .01$; $F2_{(2,38)} < 1$). Yet, unlike the previous condition, participants not only fixated more on targets than competitors and distracters ($t1_{(24)} = 5.93$ and 3.31, $ps < .01$), they actually also fixated more on distracters than competitors ($t1_{(24)} = 2.63$, $p < .05$). This pattern was unexpected because in the succeeded context conditions, the dominant meaning was always more readily available. Perhaps when a prior context requested for the subordinate reading, it worked not only by boosting the subordinate meaning but also by inhibiting the dominant meaning. Finally, as in other conditions, there was a robust picture type effect in the late time window ($F1_{(2,46)} = 302.32$, $MSE = 95.11$, $p < .001$; $F2_{(2,38)} = 238.85$, $MSE = 98.04$, $p < .001$). Targets were fixated more than both the competitors and the distracters (all $ps < .01$).

Table 19. Summary of differences in fixation proportion across conditions.

Condition	201 – 400 ms*	201 – 800 ms	801 – 1200 ms
Succeeded context-Dominant meaning		T = C > D	T > C = D
Succeeded context-Subordinate meaning		T = C > D	T > C > D
Preceded context-Dominant meaning	T > C = D	T > C = D	T > C > D
Preceded context-Subordinate meaning	D > T – C	T > D > C	T > C – D

Note: T = target fixation proportion; C = competitor fixation proportion; D = unrelated distracter fixation proportion; * = only the PD and PS conditions included analyses in this time window because there were obviously no differences in the SD and SS conditions.

Table 19 summarizes the fixation patterns in each condition and time window. First, there was a clear dominance effect. The dominant competitors remained fixated more than unrelated distracters in the SS condition, while fixation proportion of the subordinate competitors dropped more quickly in the SD condition. Similarly, the

dominant target could be reached faster (higher target fixation in 201 – 400 ms) in the PD condition than the subordinate target in the PS condition. Both pieces of evidence suggested that the dominant meaning was more available than the subordinate meaning.

Moreover, a robust context position effect was also observed: Without prior context, both meanings were activated initially until disambiguating information was received. In contrast, when context preceded the ambiguous homophonic morpheme, context-fit meaning would be more available once the homophone was perceived. Such context sensitivity in ambiguity resolution is inconsistent with modular models like exhaustive access (Swinney, 1979) or ordered access (Hogaboam & Perfetti, 1975), but fits perfectly with interactive models such as selective access (Simpson, 1981) and reordered access (Duffy et al., 1988). Still, simple selective access is unable to account for the activation of context-inconsistent meanings in the SS and PD conditions. In both conditions, the alternative reading was activated even when the context had decided what should be retrieved. Actually, similar multi-meaning activation could be observed in all conditions except in the PS condition.

Collectively, the pattern appeared to be most consistent with the reordered access model proposed by Duffy et al. (1988): All meanings of an ambiguous unit would be activated, with the level of activation ranked by relative meaning frequency. On the other hand, this "default" activation ranking could be rearranged by contextual information so that context fit meaning could receive a boost. Yet, the reordered access model in its original form would have difficulty explaining why the dominant meaning competitor in the PS condition was not more available than the unrelated distracter. This is particularly odd because the subordinate competitor was activated in the PD condition, indicating that indeed the constraint of prior context was not all-

or-none but graded. According to the original model, contextual bias worked by increasing the activation of context-fit meanings, while the alternative meanings were unaffected. The consequence was slower subordinate meaning retrieval because an extra selection process was needed to choose from the equally active dominant (more frequent) and subordinate (boosted by context) meanings.

In the present experiment, delay in subordinate meaning retrieval was absent or at best minimal (in the 201 – 400ms window). Perhaps the inconsistent meanings sometimes would receive inhibition from the context as well (see, for example, Martin, Vu, Kellas, & Metcalf, 1999; Vu & Kellas, 1999; Vu et al., 1998). In the PD condition, because the dominant meaning was supported by both dominance and contextual bias, rapid meaning retrieval and integration could succeed without inhibition on the subordinate meaning. In contrast, when the subordinate sense was required in the PS condition, both facilitation on the subordinate meaning and inhibition on the dominant meaning were necessary so that integration was fast enough to keep up with the speed of incremental speech processing. In short, although ambiguity resolutions in different situations (e.g., at the morphemic level and at the lexical level; in the auditory modality and in the visual modality) may follow identical basic principles, details might differ to catch up with the specific processing demands in different modalities.

The pattern observed in Experiment 3 with homophonic morphemes is in general consistent with that in Tsang (2006) about homonymic morpheme resolution. First, potential confounding factors such as familiarity were matched for whole words and contextual morphemes in both experiments. This allowed attribution of any effects observed to properties of the homophonic or homonymic morphemes. The observed meaning frequency effect thus supported the role of morphemes in Chinese

disyllabic word recognition (e.g., Zhou, Marslen-Wilson, Taft, & Shu, 1999), even in conditions where whole word access might be more efficient because of the morpheme level ambiguity (Baayen, Dijkstra, & Schreuder, 1997; Bertram et al., 2000a). Therefore, although some linguists believed that morpheme level integration was "unwieldy and costly" (Packard, 1999, p. 91), morphological-mediation was nevertheless the routine in word recognition, at least in languages like Chinese, where words can be easily decomposed into individual and isolated morphemes. Actually, meaning integration is an essential component for comprehension. After all, online integration is necessary at the lexical and sentence levels. Logically speaking, it is unclear why a system that includes an online integration of morpheme would be especially disruptive.

Second, both the results in Experiment 3 and Tsang (2006) could best be explained by the reordered access model of ambiguity resolution (Duffy et al., 1988). Specifically, relative meaning frequency interact with the presence of a prior biasing context to determine the course of meaning retrieval. Recently, Chen and Boland (2008) also employed the visual-world paradigm to investigate meaning resolution of homophones. Similar to our conclusion, they also discovered both meaning frequency and context effects. The primary difference between our study and theirs lay in materials: We used homophones that mapped onto morphemes while they used homophonic words. The fact that Chen and Boland arrived at a similar conclusion as ours using homophonic lexical materials further strengthened our confidence that ambiguity resolution at different levels follows identical principles.

This experiment also extended the results in Chen et al. (manuscript). In Chen et al., participants' eye movements during the processing of written two-character Chinese words were monitored. Results also supported the reordered access model of

ambiguity resolution (Duffy et al., 1988). The fact that similar findings could be found in processing the simultaneously shown written words and the serially available auditory words is a strong evidence supporting a general mechanism in ambiguity resolution.

There are subtle differences between the present experiment and Tsang (2006), however. Specifically, in the previous study, Tsang always obtained higher dominant than subordinate meaning activation in the first syllable time window when no prior context was available (i.e., the SD and SS conditions). In contrast, both dominant and subordinate meanings were equally available initially in the present study. The stronger activation of the dominant meaning was only revealed as a more prolonged activation in the SS condition (Figure 11B). There were no readily available answers for this discrepancy. Given that materials in the two experiments were not matched, factors such as the strength of dominance bias, distinctiveness between meaning, whole word familiarity, etc... might all contributed to the difference. For instance, homophone density[15] (the number of characters sharing the same pronunciation) is higher in the present experiment than in Tsang (mean = 1.32 and 1.10, S.D. 0.25 and 0.33, respectively; $t_{(34)} = 2.28$, $p < .05$). In other words, upon hearing the first syllable, activation would spread across more morphemes in the present experiment, which might have masked the activation in the dominant meanings. Also, in the present experiment the average (log) character frequency of the intended target is lower than that in Tsang (mean = 3.10 and 3.83 respectively, $t_{(34)} = 4.76$, $p < .01$). Given recent demonstrations of orthographic effects in spoken word recognition (e.g., Slowiaczek, Soltano, Wieting, & Bishop, 2003; Ziegler, Ferrand, & Montant, 2004), the lower written frequency might also contribute to the slower emergence of meaning

[15] Log (number of homophones) was used to compute the difference because the distribution of homophone density is skewed.

frequency effect because there were other more frequent characters available. Further research would be needed to test the influence of these factors and to verify whether homonymic and homophonic morphemes elicited subtly different resolution mechanisms. In any case, the general principle of interactivity between meaning frequency and context is valid across different materials.

To summarize, Experiment 3 revealed robust frequency effect and context effect in the resolution of homophonic morphemes. The two factors interacted and constrained the retrieval of correct meaning. Such interaction appeared to be critical in catching up with the demand of rapid speech perception so that comprehension could remain fluent. Actually, the high speed of online language processing might have contributed to the apparent universality of ambiguity resolution mechanisms in different linguistic units, across modalities, and various languages. In the General Discussion, we will see how these mechanisms constrain Chinese speech perception.

Chapter 7

General Discussion

Despite the fact that Chinese is one of the most widely-used languages in the world, relatively few empirical works have been done to understand the cognitive mechanisms underlying its production and comprehension. There are virtually no established theories of Chinese speech processing (except the simple description in Ye & Connine, 1999). Therefore, researchers have to rely on models developed in Indo-European languages in guiding their research direction. However, given the many unique characteristics of Chinese speech, such as the existence on lexical tone (Cutler & Chen, 1997; Schirmer et al., 2005; Ye & Connine, 1999) and the salience of syllabic units (Chen, 2000; Chen et al., 2002; Cheung et al., 2001; McBride-Chang et al., 2004), the validity in directly employing ideas generated from other languages is left in doubt. In other words, a better understanding about the fundamentals of Chinese language processing is in urgent need. The purpose of this thesis is to gather empirical information on Chinese speech perception. In three experiments, we have studied how Chinese speakers deal with the ambiguities intrinsic to the speech signals. Specifically, we first ask what the basic segmental unit is in Chinese speech perception. Is it phoneme, just as in English or Dutch (Protopapas, 1999 for review), or is it syllable? Second, we investigate the role of tone in speech perception. By definition, a tonal language such as Mandarin Chinese or Cantonese Chinese employs tonal contrast to differentiate lexical entries. Then, does tone behave similarly as segmental cues? Is it also involved in the candidate generation and elimination central to speech perception? Finally, we test how Chinese speakers retrieved the correct meaning in case of homophony.

Briefly stated, in the present experiments we found out that while syllables were highly salient in Chinese speech (Experiments 1A and 2A), our participants were nevertheless sensitive to subsyllabic information during candidate generation and elimination (onset in Experiment 1A, rime in Experiment 2A). Yet, phonemic effects appeared to be weak, and lagged behind the influence by whole-syllables (onset effect in Experiment 1A vs. 2A). More interestingly, subphonemic details also strongly constrained the initial activation of hypothetical word candidates: Only words that fit the incoming acoustic features would be included in the candidate set (Experiment 2B). The sensitivity to fine acoustic characteristics in Chinese is consistent with similar findings in other Indo-European languages (e.g., Davis et al., 2002; Marslen-Wilson & Warren, 1994; Salverda et al., 2003). It highlighted the needs to go beyond the traditional phoneme-based approach in speech perception.

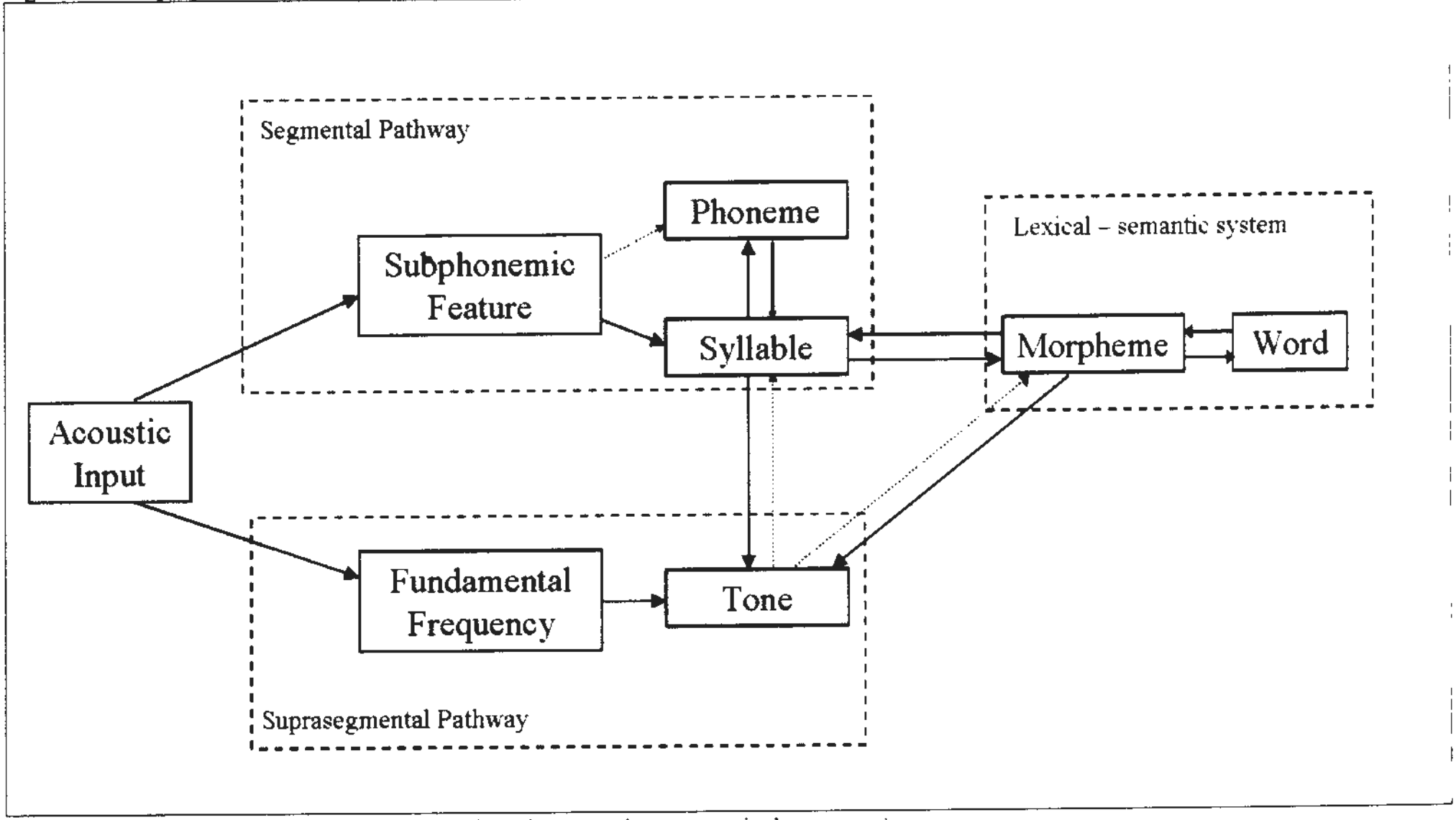
On the other hand, even though lexical tone is crucial for the recognition of Chinese spoken words, it does not appear to have a unique role in candidate generation. In contrast to the preliminary report in Li and Yip (1998), or Sum (2003), we did not observe word activation based solely on tone-sharing without segment-sharing (Experiments 1A and 2A). Rather, lexical tone was used to eliminate mismatched candidates when context was available (Experiment 1B) or after sufficient segmental features had been received (Experiment 2A). This pattern was consistent with the perceptual disadvantage of tone (Cutler & Chen, 1997) and the tonal constraint on semantic access in idioms or sentences (Schirmer et al., 2005; Ye & Connine, 1999). However, our results were inconsistent with Sum (2003), who proposed that any differences between segmental and suprasegmental effects were negligible. The results also did not justify the phoneme-metaphor of lexical tone proposed by Ye and Connine (1999). Suprasegmental information like tone and

segmental information like phonemes or syllables should better be treated as something qualitatively different.

Although tone might help meaning retrieval by disambiguating otherwise identical syllables, as reviewed in the introduction, syllabic homophony is still prevalent in Chinese speech. Yet, most of the time, ambiguities are left unnoticed in natural conversation. This unawareness about signal uncertainty cannot be attributed to a simple delayed meaning access after enough inputs are gathered. Actually, meaning retrieval is immediate and governed by the relative dominance among various meanings such that dominant meaning was activated more strongly than subordinate ones as shown in Experiment 3. By definition the dominant meaning is the more frequently used one. Therefore, a stronger activation of this meaning would guarantee successful comprehension in most cases. Also, the ambiguity resolution system is sensitive to the bias of prior context such that context-fit meaning could be activated more quickly. Overall, the results are consistent with the reordered access model of ambiguity resolution (Duffy et al., 1988). Moreover, these mechanisms could ensure efficient and correct comprehension most of the time.

In this General Discussion, we will first outline a model of Chinese speech recognition based on the aforementioned findings. For instance, to account for the unique characteristics of lexical tone compared with other segmental units like phonemes and syllables, the model incorporates two independent yet interacting pathways, one suprasegmental and the other segmental. We will go into details the mechanisms in each pathway one-by-one. Then, the mechanisms by which outputs from the two pathways combine to activate the corresponding morphemes and words will be discussed, with emphasis on resolution of ambiguous homophonic morpheme. Finally, possible directions for future research will be proposed.

Figure 12. The proposed model of Chinese spoken word recognition.



Note: Solid and dotted lines indicate strong and weak connections respectively (see text).

7.1 A possible model of Chinese spoken word recognition

Given the empirical constraints revealed in the present experiments (as well as findings in previous research), we propose a possible model of Chinese spoken word recognition here. The general workings of the model are presented in Figure 11.

Figure 11 shows the flow of information among different systems and units. The most noticeable difference between this proposed model and the one in Ye and Connine (1999) is the separation of acoustic information into two types from the very beginning of speech perception. While the acoustic correlates of phonemes or syllables[16] are processed in the segmental pathway, tonal features (i.e., fundamental frequency F0) are processed in the suprasegmental pathway. This separated processing of segmental and suprasegmental actually received neurological support. Employing a mismatch-negativity (MMN) paradigm, Luo et al. (2006) found that in a group of Mandarin Chinese speakers, the acoustic cues of lexical tone elicited stronger right hemisphere responses than the left hemisphere. In contrast, consonants produced the reversed pattern such that higher activation was found in the left hemisphere. Luo et al. concluded from this double-dissociation that different aspects of the speech inputs were processed separately based on their acoustic properties very early in the pre-attentive stage. In other words, segmental and suprasegmental features are processed in the left and right hemispheres respectively. Only at a later time point will pitch information be transferred to the left hemisphere to fulfill its role as lexical tone in differentiating otherwise identical syllables. Such two-stage processing of tonal information received partial support from an fMRI study by Gandour et al (2003). In their study, participants were asked to judge whether two

[16] It should be noted that there are still many uncertainties about the acoustic correlates underlying a specific phoneme. However, consensus has been reached on a number of possible important dimensions, such as voice-onset time (VOT), direction of formant transition, etc. (Raphael, 2005).

successively presented three-syllable phrases were identical. In critical trials, the final syllables of the two phrases differed in lexical tone. Gandour et al. found that activation was stronger in the left hemisphere while performing lexical tone differentiation. Given that fMRI measured brain activity with low temporal resolution, the results appeared to reflect left hemisphere involvement when the pitch information was processed lexically (i.e., in the second stage).

The second important difference of the present model compared with the modified TRACE model (Ye & Connine, 1999) is that we allow the segmental cues to map directly onto syllables. Actually, we also propose that the acoustic-syllabic linkage is stronger than the acoustic-phonemic one to account for the salience of syllable in the present Experiments 1 and 2A (as indicated by the larger arrow pointing towards the syllable box). On the other hand, we retain the phoneme representation in the model because when given appropriate instruction, native Chinese speakers could still demonstrate phoneme level awareness (e.g., McBride-Chang et al., 2004; Ye & Connine, 1999). Moreover, the existence of a phoneme layer also helps explaining why our participants sometimes produced candidate words which shared only the onset phoneme with the actual target. However, to be consistent with the observation that such onset-sharing activation only occurred when the time limit of responding was loose (Experiments 1A and 1B), or after a relatively long time delay (Experiments 2A and 2B), we hypothesize that the acoustic-phoneme connection is weak. It will thus require a longer time before phoneme level representation accumulates enough activation to influence candidate word generation.

We also include the conventional bidirectional linkage between phoneme and syllable, just as Ye and Connine (1999) did. The reason is simple: phonemes add together to form syllables, and syllables can be decomposed into constituent

phonemes. Although in the proposed model, this linkage is symmetrical, it should be noted that the influence of syllables on phonemes may be stronger than phonemes on syllables because the acoustic-syllable connection is tight. Syllable representations can reach recognition threshold very quickly (Experiment 2A), allowing less room for facilitation by phonemes. Syllabic feedback on phonemes is thus relatively strong, while the reversed link will be useful only under special circumstances such as the extraction of embedded words, where ambiguities in the correct target word persist long enough for phonemes to accumulate effects on syllables (Experiment 2B).

On the other hand, the bidirectional connections between syllable and tone, and between morpheme and tone, are asymmetrical. This proposal is necessary to account for the effect of contexts in improving tonal constraints (Experiment 1A vs. 1B; also see Brown-Schmidt & Canseco-Gonzalez, 2004; Schirmer et al., 2005; Ye & Connine, 1999). Contextual information creates feedback on tones, strengthening the activation of the matching tone, which in turn constrains the activation of "correct" syllable and the corresponding morpheme. Without a readily available context, tonal effects are weak and need a long time to accumulate before being able to affect candidate activation.

A further difference from Ye and Connine's model (1999) is the lack of a direct phoneme-tone linkage in the present proposal. This is because, in Chinese speech, tone can seldom be identified on individual phonemes. Indeed, Gandour (1983) proposed that the change of fundamental frequency over time (i.e., tone contour or direction) is an important tonal feature in Chinese speech. A longer steady state of sound energy is required to realize the pitch movement. Therefore, lexical tone should better be conceptualized as being connected with the entire syllables rather than with phonemes. On the other hand, there are also proposals that rime and

tone are actually a unified representation because the voiced part (i.e., the vowel) is the primary constituent of the steady state (see Howie, 1976 and Vance, 1977 for discussion). However, the independent effects of rime and tone observed in Experiments 1A and 1B (contrary to the interaction between onset and rime) suggested that they are indeed separate units in Chinese speech perception. Specifically, the rates for rime-sharing, tone-sharing, and rime-plus-tone-sharing errors were identical statistically in gating. Moreover, the activation for rime-sharing competitors, but not for tone-sharing competitors in Experiment 2A, also supported that rime alone is independent from tone. As a result, we did not propose a specific linkage between rime and tone.

Finally, in the present model, we explicitly propose that the direct output of phonological processing is morpheme. Word level representations are activated via morphemes. In other words, Chinese speakers recognize spoken words through their constituent morphemes. This proposal is contradictory to some linguists' belief that a morphemic route of word recognition is improbable (Packard, 1999). However, it fits perfectly with the rapid morpheme activation and integration demonstrated in Tsang (2006) and the present Experiment 3. It is also consistent with the large body of evidence supporting the word recognition models that assume morpheme-mediation (e.g., Taft & Forster, 1975; Taft & Zhu, 1995). We also proposed feedback linkage from morpheme to the syllable and phoneme representations to account for the context morpheme effect observed in Experiments 1B and 3.

Following this overview of the proposed model, we will cover in more details the processes in the segmental and suprasegmental pathways, and how their outputs combine to allow correct recognition of target morpheme. Special emphasis will be

placed on the implications of the proposed mechanisms and how they can be related to previous studies.

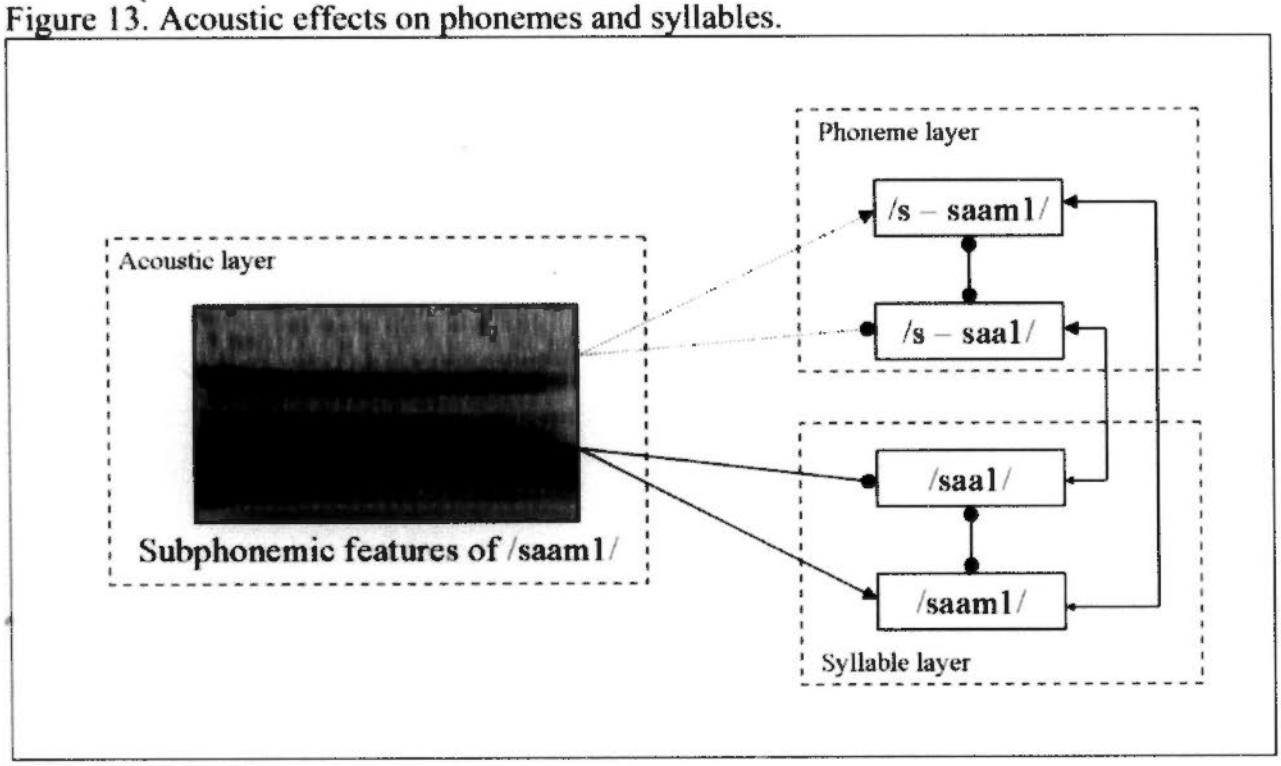
7.2 The segmental pathway

The segmental pathway starts from the extraction of subphonemic features from the speech stream. Although there are still uncertainties about the exact nature of the acoustic correlates of segmental units (Raphael, 2005), empirically speaking it has been demonstrated repeatedly that people are sensitive to these signals in spoken word recognition (Davis et al., 2002; Marslen-Wilson & Warren, 1994; McMurray et al., 2008; Salverda et al., 2003; Warren & Marslen-Wilson, 1987). Consistent with this, in the present study, fine acoustic details were captured online by our participants in their generation of hypothetical word candidates before successful target recognition (Experiments 2B and 2C): Candidate words are activated only when the inputs match well with the typical acoustic realization of these words. For instance, even though /haai4/ and /haa4/ are having extensive phonological overlapping, the embedded competitor /haa4/ was not activated in Experiment 2B, when participants heard /haai4/, because the typical acoustic realization of the initial portion of these two syllables differs drastically. Actually, as demonstrated in Experiments 2B and 2C, for competitors sharing initial phoneme(s) but with the typical acoustic realization being different from that of the presented target, inhibition occurred[17]. Moreover, the inhibition between the differentially realized identical phonemes (e.g., the /h/ in /haai4/ and /haa4/) was so strong that below-baseline fixation proportion was resulted.

[17] I acknowledged that calling this effect as "inhibition" might not be accurate because the below-baseline fixation proportion did not occur in all competitor conditions. And there was no condition in which the different competitors were directly compared against a common baseline. So the "inhibition" could simply be a "lower activation". However, in any case, we still need to explain what causes the "lower activation" and the mechanisms proposed here would equally apply.

The inhibition on acoustically different phonemes can be incorporated into the acoustic-phoneme/syllable linkage and/or realized as lateral inhibition among phonemes or syllables. In the former case, while acoustically-fit phoneme/syllable receives excitation from the supporting features, acoustically-misfit ones actually receive interference. In the latter case, identical phonemes/syllables that are having different acoustic realizations are connected with inhibitory links, such that when the acoustically-fit phoneme/syllable is activated, it will interfere with other misfit items. In both cases, however, the same phoneme or phoneme-cluster appears to be represented more than once. This proposal is consistent with recent suggestions of episodic or semi-episodic view of phonological representation (Mitterer & McQueen, 2009), which assumes that concrete instances of the spoken words are also stored in the mental lexicon. A schematic overview of these two hypotheses is illustrated in Figure 13.

Figure 13. Acoustic effects on phonemes and syllables.



We prefer the lateral inhibition hypothesis of acoustic similarity effect because it can also explain why the unrelated distracters were not inhibited to the same degree as the differentially realized identical phonemes despite both of them were different from the targets in terms of initial acoustic features. According to this hypothesis, the unbalanced inhibition on competitors and distracters might be attributed to the lack of direct inhibitory links among different phonemes (e.g., /h/ and /p/ are not connected). As acoustic inputs just activate the matching phonemes (e.g., the /h/ in /haai4/ but not haa4), the phonemes of unrelated distracters actually receive neither supportive nor counteractive acoustic evidence. In contrast, if inhibition is incorporated into the linkage between acoustic features and higher representations, the phonemes of unrelated distracters should receive equal inhibition as those of competitors.

A related issue is that the acoustic effects observed in the present experiments appears to be much stronger than those in previous reports (e.g., McMurray et al., 2008; Salverda et al., 2003), in which words containing acoustically dissimilar phonemes were still more highly active than the unrelated baseline. In contrast, we obtained below-baseline activation for the acoustically dissimilar competitors. No explanation for this discrepancy is readily available because many factors (e.g., task, language and phonemic structure) may have contributed. Yet, we speculate that the type of acoustic information studied may be particularly important. In previous studies employing the visual-world paradigm comparable to the present Experiment 2B, usually the temporal aspect of subphonemic feature was manipulated. For example, McMurray et al. (2008) varied voice-onset time along a continuum and Salverda et al. (2003) manipulated syllable duration. In contrast, the acoustic similarity effects in Experiment 2B reflected differences in spectral properties. In most dissimilar pairs, it is the formant frequency and/or shape of formant transition

that differ. Previous neurological evidence (e.g., Zatorre & Belin, 2001) has shown that temporal and spectral properties of acoustic signals are dissociable and processed separately by different hemispheres. Briefly stated, spectral information is processed in the right hemisphere while rapid temporal changes are processed in the left hemisphere.

This speculation received partial support from an eye-monitoring study by Dahan, Magnuson, Tanenhaus, and Hogan (2001). In their study, Dahan et al. created monosyllabic target words (W1) by cross-splicing two tokens of the same word (W1W1), two different words (W2W1), or a nonword and a real word (N3W1). Such cross-splicing is likely to produce extensive changes in spectral properties of the target (W1) for the latter two conditions. Results suggested that for trials with subphonemic mismatch (W2W1 and N3W1), target fixation was slower than in trials without mismatch (W1W1). More importantly, target fixation was also slower for W2W1 condition than N3W1, because participants had "wrongly" activated W2 based on the initial portion of the items (Experiment 2 in Dahan et al.). This also inhibited activation of the actual target W1. Moreover, the initial inhibition by acoustic mismatch was later overridden by the phonemic match such that target detection was still successful in W2W1 and N3W1. In other words, the subtle subphonemic features played a role in temporarily activating the acoustically-matched words and inhibiting those having a different acoustic realization. This fits perfectly with the competitor activation in the acoustically similar condition observed in Experiment 2B. Therefore, the inhibitory linkages among identical phonemes with differential acoustic realization may be particularly strong when they differ on spectral properties, leading to higher mutual inhibitions with spectral than temporal differences. However, one major problem exists in Dahan et al.: It is unclear whether

cross-splicing would have increased participants' sensitivity to subphonemic mismatches. It should also be noted that the acoustic effect in Dahan et al. (see Figure 3) was not as immediate as in Salverda et al. (2003) or in the present experiments. Certainly, more works have to be done in studying the role of fine acoustic details in speech processing, especially in Chinese speech perception.

From the acoustic layer there are two parallel route of activation spreading, namely the phoneme route and the syllable route. In the present model, we suggest that the phonemic route is just secondary to Chinese speech recognition because it is only weakly connected with the features. Rather, Chinese speakers mainly rely on the syllable route in spoken word recognition. This idea not only provides explanation for the robust syllable effects observed in Experiments 1A and 2A, it also conforms to linguistic hypothesis, developmental data (Cheung et al., 2001; McBride-Chang et al., 2004), and evidence from Chinese speech production (Chen, 2000; Chen et al., 2002). This "marginalization" of phoneme is perhaps the most unique aspect of Chinese speech processing compared with Indo-European languages. Although phoneme is just a linguistic construction, its psychological validity has been largely unquestioned since the demonstration of phoneme level categorical perception. Phoneme has been incorporated in virtually all speech perception and production models as an important unit that intervenes between the abstract lexical representation and its acoustic realization. Even in more recent connectionist models of speech perception that allowed direct feature-lexicon connection (Gaskell & Marslen-Wilson, 1997), phonemes were still considered to be the basic unit in the phonological system.

Actually, from the perspective of language acquisition, the salience of syllable in Chinese makes perfect sense because phonetics is not included in the syllabus of regular language education (at least in Hong Kong). Children grow up without having

learned explicitly how to decompose spoken words into phonemes. The logographic nature of written Chinese also precludes Hong Kong children from developing grapheme-phoneme conversion spontaneously. In contrast, because each written Chinese character corresponds to a monosyllable in speech, it is a common practice to learn Chinese through the pronunciation of the whole syllable. This further interferes with the robustness of phoneme level representation in Chinese. As a result, individual phonemes cannot develop into routine processing units in Chinese speech among average undergraduates in Hong Kong. Its effect will lag behind that of syllable and emerge only when sufficient time is allowed. Therefore, when /haai4/ is heard, initially the phoneme-sharing /ho4/ or /haa4/ will not be highly active. However, after the phoneme nodes have accumulated enough activation, they can still be more active compared with the control baseline.

While we assume that syllable is the actual phonological representation that maps onto morphemes in the semantic system, we do not intend to propose that Chinese speech perception is delayed until whole syllables are available. Actually, in Experiment 2, we observed clear target fixations before syllable offset, which strongly suggested that Chinese speech perception is also immediate and active. It involves the typical candidate generation and elimination found in other languages. The syllable-based hypothesis simply suggests that these processes operate primarily on syllables. In other words, when /haai4/ is heard, /haai5/ will be more highly active than /haa4/ and /haau4/, even though according to the model in Ye and Connine (1999), /haai5/ and /haa4/ and /haau4/ all differ from /haai4/ by only one distinctive feature and thus should be activated to a similar degree.

To summarize, we propose that acoustic features are important basic processing units in Chinese speech perception. Feature-level contrasts are mapped

directly to both phonemes and syllables, but the latter one plays the primary role in candidate generation and elimination before target identification. This pattern may be inconsistent with the general assumption of phoneme-based recognition in Indo-European languages, but it fits well with the results in the present study and the pattern of language acquisition among our Chinese participants.

7.3 The suprasegmental pathway

In contrast to a poor understanding of the acoustic correlates of segmental unit, it is generally agreed that lexical tone is primarily realized as the fundamental frequency in the speech signal, while factors like durations and intensity play less a role (Gandour, 1983; Khouw & Ciocca, 2007; Vance, 1976). As mentioned previously, the acoustic processing of fundamental frequency appears to be dissociable from segmental processing (Luo et al., 2006). This provides justification for proposing an independent route of tonal processing. Moreover, we assume that the suprasegmental route is slow when no contextual feedback from is available. In other words, tonal identity is available late after enough information about fundamental frequency has been received. This naturally explains the perceptual disadvantage of tone (Cutler & Chen, 1997) and the present findings that it is not particularly useful in generating hypothetical word candidates. After all, by the time the whole tonal contour is identified, there has already been a great deal of segmental constraints that inhibit other tone-sharing candidates. Therefore, perhaps lexical tone participated in Chinese speech perception only by differentiating otherwise identical syllables such as /haai4/ and /haai5/ (Experiment 2A).

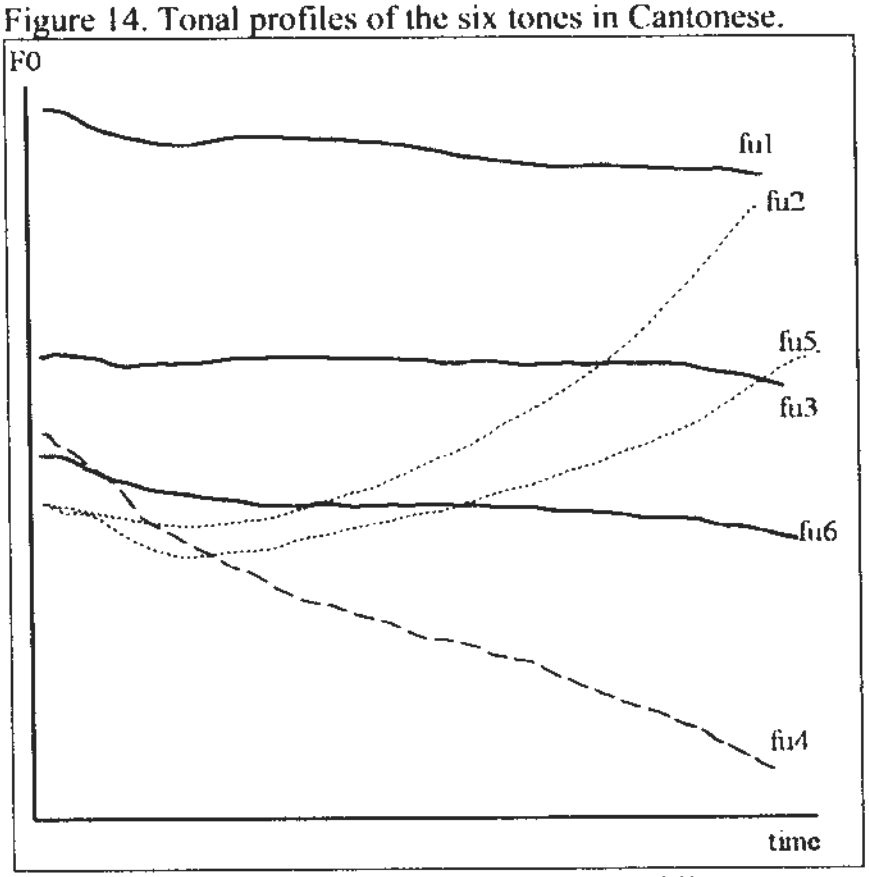
Given that the acoustic correlates of lexical tone (i.e., fundamental frequency) are also related to speaker variability (e.g., gender, age), emotion, and speech rate, it

is still unclear how abstract the representation of lexical tone should be. While some researchers (Moore & Jongman, 1997; Peng, 1997) believed that a normalization process would help adjusting each individual tone token according to an abstract "template" tone, it is also possible that each realization of spoken Chinese words is stored separately in the mental lexicon (i.e., the episodic view, Goldinger, 1998; also see Connine, 2004, for a semi-episodic view). According to the latter hypothesis, details of each tone token are stored. Tone perception is therefore sensitive to fine acoustic features, just as segmental units. At present, very limited works have been done on testing the two possibilities and results have been equivocal. For example, Yu (2007) showed that subtle acoustic difference in exemplars of the same tone was meaningful. However, he argued that the acoustic effects were restricted to modification of the tonal representation and did not reach consciousness. On the other hand, Francis, Ciocca, and Ng (2003) observed a flat discrimination curve of tonal contrasts, which indicated that acoustic details could reach conscious awareness and led to non-categorical discrimination of lexical tone. However, in the same study Francis et al. also discovered a steep identification slope, which indicated categorical perception. In the present thesis, we did not include any manipulation of acoustic details. Therefore, we cannot estimate how sensitive our participants were towards the subtle difference in tonal features. Also, all materials were recorded by the same female speakers, normalization should be easy. Given these, we refrain from drawing strong conclusions about the abstractness of tone representation. More research is clearly needed. It should be noted, however, that our assumption of a relatively slow tonal effect will still hold regardless of answer to this question.

A related issue concerns how the different tones are connected by themselves. In Cantonese, there are six tones. Figure 13 shows the typical tonal profiles of the six

tones on the syllable /fu/. As can be seen from the figure, there are similarities and differences among the six tones. For example, the shapes of tones 1, 3 and 6 are similar, and the starting F0 values of tones 2, 4, 5 and 6 are also similar. Actually, previous studies showed that Cantonese tones can be summarized by two dimensions (Gandour, 1983; Khouw & Ciocca, 2007). The first dimension is pitch height, which refers to the average fundamental frequency over the whole syllable and is useful to differentiate the three level tones (high, mid, and low; i.e., tones 1, 3, and 6 or tones 2 and 4). The second dimension is pitch contour, which is the change in pitch level over time. Pitch contour (especially the changes over the later part) is important to separate flat (tones 1, 3, and 6), rising (tones 2 and 4) and falling tones (tone 5). In other words, subtle differences in fundamental frequency may really be necessary to differentiate the six tones.

Figure 14. Tonal profiles of the six tones in Cantonese.



Note: Solid lines represent level tones; Dotted lines represent rising tones; Dashed line represent falling tone.

From Figure 14, one may come up with the hypothesis that tones 1 and 3 should be differentiable from the other tones at the earliest moment because the pitch heights of these two tones are distinct. On the other hand, tones 2, 4, 5 and 6 are distinguishable only after more information about the contour is also received. In Experiment 2A, we did not manipulate the tonal similarity between competitors and the target tone. Rather, the tone-sharing competitors we used shared maximum similarity with the target (i.e., the tones were identical). Given that no candidate generation had been obtained in such optimal condition, we speculated that candidate words which shared only pitch height or pitch contour would also remain inactive. Yet, it is still unclear whether the later constraint in eliminating tone-mismatched candidates would be dependent on tonal similarity. Actually, Cutler and Chen (1997) demonstrated that while tonal contrasts were in general more difficult to detect than phonemic contrasts in a matching task, differentiation of similar tones (tones 4 and 5) were particularly more difficult than dissimilar tones (tones 1 and 2). Unfortunately, the matching task in Cutler and Chen did not provide details about the time course of tonal influence. It would be important to replicate their results with the visual-world paradigm, which provides more information about activation changes over time.

The connection between morpheme and tone explains the stronger tonal constraints when context is present (Experiment 1B). Pre-existing morphemic context exerts facilitatory feedback on the lexical tone, increasing its activation level. As a result, the tonal representation reaches activation threshold faster and exerts feedback back to the syllable level earlier than when no morphemic context is available.

Finally, while we propose a suprasegmental route of lexical tone processing and assume that this route lags behind the segmental route (syllable), we are not suggesting that processing of tonal information only starts after syllables are available.

In contrast, we believe the segmental and suprasegmental routes run in parallel. Different aspects of the acoustic inputs are extracted by different brain regions (Luo et al., 2006), which then feed into the corresponding systems for further segment and tone processing. The two routes continuously interact with one another until they cooperate to activate the correct morpheme for recognition. Such proposal is indeed consistent with the results obtained when building speech recognition machine with statistical modeling in the engineering community. Specifically, Demmechai and Mäkeläinen (2001) proposed the linked detection mechanisms of tonal syllable recognition. In their model, acoustic inputs are separately analyzed to segmental and tonal features, which then feed into the syllable and tone recognizers respectively. The two recognizers influence each other at every time step before successful identification is established. Comparing with other possible models, linked detection yielded the best outcomes, suggesting the proposal of independent, yet interacting pathways of segmental and suprasegmental processing may also have computational validity. Yet, it should be noted that in the present model, the linkage between tone and syllable is asymmetric (to account for the context effects), contrary to the assumption of symmetry in Deemchai and Mäkeläinen.

To summarize, although lexical tone is critical for the ultimate differentiation of otherwise identical syllables, its effect only emerges slowly. Indeed, it is more useful for eliminating mismatched words than for generating matched candidates. On the other hand, while our understanding about the nature of tonal representation is still shallow, we are pretty confident that the activation of tonal representations relies on operations independent to those dealing with segmental information. Yet, continuous interaction between the two systems seems necessary for efficient Chinese spoken word recognition.

7.4 Morpheme-mediation of spoken word recognition

In the proposed model, the outputs of segmental and suprasegmental pathways are combined as they jointly access the morpheme level representation. However, the mapping between phonological unit and morphemic unit is not one-to-one. Rather, extensive homophony exists in spoken Chinese such that the same tonal syllable may be linked to more than one morpheme. We term this phenomenon "morphemic ambiguity" because it resembles the one-to-many mapping between word and meaning in lexical ambiguity. It turns out that the morphemic ambiguity resolution also follows mechanism in lexical ambiguity. Both context and meaning frequency matter. Therefore, in the model we assume that the syllable and tone will first jointly activate the most frequent meaning before the subordinate ones. Moreover, there will be feedback from morpheme to syllable and tone to account for the context effect.

Investigation on homophonic morpheme resolution presents a strong test against the role of morpheme in spoken word recognition because ambiguities in morpheme may lead to more reliance on whole-word access (Bertram et al., 2000b). Contrary to the beliefs that morpheme simply does not exist (Packard, 1999) or it affects spoken word recognition only post-lexically (Greber & Frauenfelder, 1999), we hypothesize an obligatory morpheme-decomposition route in Chinese spoken word recognition. Linguistically, Chinese speech has a tight morphosyllabic structure such that each individual syllable clearly represents a single morpheme. Given the salience of syllables in Chinese as reviewed before, the assumption of morpheme-based meaning access seems reasonable and natural. Furthermore, this hypothesis is also grounded empirically from the results in Tsang (2006) and in the present thesis (Experiment 3). In both studies, we employed the visual-world paradigm which is

sensitive enough to reveal the online speech processing (e.g., Allopenna et al., 1998; Tanenhaus & Spivey-Knowlton, 1996). Results showed that morpheme frequency affected fixation proportion well before word offset. Given that word level characteristics were closely matched across conditions, the early morpheme effect could only be explained by morpheme activation prior to whole-word access. The observation of morpheme mediation in Chinese spoken word recognition is consistent with the findings in other languages (e.g., Taft, Hambly, & Kinoshita, 1986; Wurm, 1997) and in written word recognition (e.g., Zhou et al., 1999).

Besides the morpheme frequency effect, we also observed robust prior context effects in recognizing the correct morpheme. More importantly, previous studies on the issue typically employed sentential context (e.g., Tsang, 2006; Zwitserlood, et al., 2005), so it is unclear whether context effects also operate at the minimum meaningful scale of morpheme. In Experiment 3, we successfully demonstrated that a single prior morpheme could provide sufficient contextual constraints to resolve the ambiguities in a homophonic morpheme (PD vs. SD conditions; PS vs. SS conditions). Therefore, we incorporate a feedback link from the morpheme layer to syllable and tone layers to account for the contextual bias created by an identified morpheme. Overall, this context sensitivity and the effect of morpheme frequency jointly support the reordered access model (Duffy et al., 1988) of morphemic ambiguity resolution. This strongly supports the possibility of a general ambiguity resolution system responsible for handling the one-to-many mapping between form and meaning at different levels of language processing. This also suggests that we should be cautious when inspecting previous demonstrations of whole-word access of ambiguous morphemes (Bertram et al., 2000b). Given that a single morpheme is constraining enough for retrieving the correct morpheme, the apparent lack of morphemic effects

may simply be attributed to the low sensitivity of the dependent measures. For instance, while robust morpheme involvement could be seen in the eye fixation pattern, we actually failed to obtain significant effects in the reaction times and error rates of target detection. Presumably, when participants responded, all ambiguities had been resolved completely. The temporary difference in meaning availability was thus masked.

Moreover, the existence of morphemic ambiguity resolution revealed in Experiment 3 was particularly important because it represented a crucial piece of evidence supporting the presence of a separate morpheme layer beyond whole-word representation. An alternative explanation of the morphemic effect was to reduce it to form or meaning sharing at the lexical level (see Feldman, 2000, for discussion). According to this hypothesis, lexicons with similar forms and meanings would cluster together in the representation space. The apparent morphemic effect was simply an emergent property of such clustering. In other words, a separate morpheme layer intervening between features and whole words was not necessary. This proposal, however, could not account for the pattern observed in resolving morphemic ambiguity. If morphemic effect was purely due to semantic clustering, it was unclear why the distinctive (i.e., low clustering) meanings of the homophonic morphemes would be activated together. On the other hand, attributing morphemic effects solely to form sharing would also encounter difficulty because it could not explain why relative frequency of usage affected meaning availability. In short, the mere existence of morphemic ambiguity resolution supports Feldman's finding (2000) that morphological effect "cannot be described simply as the 'sum' of a semantic and an orthographic effect" (p. 1441).

It should be noted, however, that in Chinese compounding is the primary way of combining morphemes. There are in principle no inflections. So all materials we used in Experiment 3 are compound words. It is unclear whether morphemes play a stronger role in processing compound words because the constituent morphemes contribute to the whole-word meanings. For example, Taft and Kougious (2004) showed that morpheme-like processing occurred in reading monomorphemic words such as "virus" and "viral" because their form and meaning are correlated. On the other hand, results from studying inflected or derived words are equivocal. While some studies demonstrated robust morpheme-decomposition (e.g., Taft et al., 1986; Wurm, 1997, there were also reports of whole-word access (e.g., Schriefers, Zwitserlood, & Roelofs, 1991; Tyler, Marslen-Wilson, Rentoul, & Hanney, 1988). Therefore, it would be important to extend the present findings to other types of morphemes.

To summarize, although the activation of dominant meaning appeared to be weaker and delayed (see Figure 11B) compared to Tsang (2006), we in general successfully replicated the interaction between meaning frequency and context position in homonymic morphemes using homophonic morphemes in Experiment 3. Moreover, the exact mechanisms for resolution follow closely the pattern of reordered access model. Finally, the immediate morpheme frequency effect is interpreted as supporting a morpheme-decomposition view of word recognition and the existence of genuine morpheme representations.

7.5 Conclusion and Future directions

The present thesis aims at gathering empirical data about the fundamentals of Chinese speech. In three experiments, we showed that acoustic features, phonemes,

and syllables are all valid phonological units in Chinese speech. However, while the effects of acoustic features and syllables are salient, the role of phoneme seems to be secondary. We also obtained evidence of morpheme-mediation in recognizing Chinese disyllabic words. To summarize these findings, a possible model of Chinese speech perception is proposed in Figure 11. On the other hand, a good model not only serves to explain empirical findings, it also helps generating new hypothesis. We conclude the present thesis by suggesting several possible research directions.

First, the role of fine acoustic features in spoken word recognition has to be specified. Although we observe a clear difference in activation level between the acoustically similar and dissimilar competitors, we did not manipulate this as an experimental factor. Stronger evidence about acoustic effect can be obtained by directly manipulating the acoustic features and measuring its influences on lexical activation (e.g., McMurray et al., 2008; Salverda et al., 2003). Moreover, further studies are needed to compare whether the type of acoustic contrasts manipulated (temporal or spectral) affects the size of acoustic effect.

Second, the weak phonemic effect in Chinese speech perception may be related to the emphasis of syllable-character linkage during reading acquisition. If this hypothesis is true, we should obtain stronger phoneme effect when children receive training in phonetics. Actually, McBride-Chang et al. (2004) showed that children who have received PinYin (a phonetic system in Mandarin Chinese) trainings scored higher in phonemic awareness. It would be interesting to see if they also demonstrate stronger phoneme-sharing competitor activation in online spoken word recognition.

Third, as mentioned previously, the six tones in Cantonese can be organized by two dimensions, namely pitch height and pitch contour. Some tones are similar in

pitch height but differ in the contour. In the present experiments, we did not manipulate the degree of similarity between competitor and target tones. Given that previous studies in Cantonese and Mandarin suggested significant effect of tone similarity (e.g., Cutler & Chen, 1997; Ye & Connine, 1999), one may expect to see similar effects in Cantonese. On the other hand, the weak activation in identical-tone competitors may argue for the contrary. Further investigations will be needed to verify which of these positions is correct.

Fourth, while in speech perception, the relative meaning dominance of the homophonic morphemes significantly affects the pattern of meaning retrieval, it is unclear whether it plays the same role in speech production. For instance, Dohmes, Zwitserlood, and Bölte (2004) revealed identical priming effects on picture naming by transparent and opaque primes, suggesting that in speech production, morpheme might be coded only at the form-level without any linkage to meaning.

Fifth, methodologically speaking, one limitation of the visual-world paradigm is that only concrete nouns could be used as materials because an easily recognizable visual display must be constructed along the auditory target. However, recently, McQueen and Viebahn (2007) successfully replicated the competitor activations during spoken word recognition (Allopenna et al., 1998) when the visual display was constructed by printed words. This greatly reduced the limitation in materials choosing and allowed further testing of similar issues in Chinese speech.

Finally, the model outlined in this thesis is comprehensive, yet, it is still a descriptive model that lacks prediction about its detailed temporal characteristics. As speech perception is highly incremental and dynamic, it is desirable to elaborate the model further with computer simulation such that the outputs of simulations can be compared with the results in the visual-world experiments directly (see Allopenna et

al., 1998). However, a thorough understanding of Chinese speech perception is necessary to implement a successful computer simulation. At the present stage, perhaps the most important job is to accumulate more empirical knowledge on the issue. We believe the present thesis has contributed in achieving a better understanding of Chinese speech.

References:

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory & Language, 38*, 419-439.

Altmann, G. T. M., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition, 73*, 247-264.

Alvarez, C. J., Carreiras, M., & Taft, M. (2001). Syllables and morphemes: Contrasting frequency effects in Spanish. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*, 545-555.

Baayen, R. H., Dijkstra, T., & Schreuder, R. (1997). Singulars and plurals in Dutch: Evidence for a parallel dual-route model. *Journal of Memory and Language, 37*, 94-117.

Bertram, R., Hyönä, J., Laine, M. (2000a). The role of context in morphological processing: Evidence from Finnish. *Language and Cognitive Processes, 15*, 367-388.

Bertram, R., Laine, M., Baayen, R. H., Schreuder, R., & Hyönä, J. (2000b). Affixal homonymy triggers full-form storage, even for inflected words, even in a morphologically rich language. *Cognition, 74*, B13-B25.

Binder, K. S. (2003). Sentential and discourse topic effects on lexical ambiguity processing: An eye movement examination. *Memory and Cognition, 31*, 690-702.

Brown-Schmidt, S., & Canseco-Gonzalez, E. (2004). Who do you love, your mother or your horse? An event-related brain potential analysis of tone processing in Mandarin Chinese. *Journal of Psycholinguistic Research, 33*, 103-135.

Borowsky, R., Masson, M. E. J. (1996). Semantic ambiguity effects in word identification. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*, 63-85.

Boersma, P. & Weenink, D. (2009). Praat: doing phonetics by computer (Version 5.1.04) [Computer program]. Retrieved April 4, 2009, from http://www.praat.org

Chater, N., & Manning, C. (2006). Probabilistic models of language processing and acquisition. *Trends in Cognitive Sciences, 10*, 335-344.

Chen, H. C. (1992). Reading comprehension in Chinese: Implications from character reading times. In H. C. Chen & O. J. L. Tzeng (Eds*). Language Processing in Chinese* (pp. 175-205). Amsterdam, Netherlands: North-Holland.

Chen, H. C. (2001). Speech processing in Chinese: An introduction. *Journal of Psychology in Chinese Societies, 2*, 155-158.

Chen, H. C., Tsang, Y. K., Chan, N. Y., & Wong, E. K. F. (manuscript). Morphemic ambiguity resolution in Chinese: Evidence from eye movements studies.

Chen, H. C., & Yip, M. C. W. (2001). Processing syllabic and sub-syllabic information in Cantonese. *Journal of Psychology in Chinese Societies, 2*, 199-210.

Chen, J.-Y. (2000). Syllable errors from naturalistic slips of the tongue in Mandarin Chinese. *Psychologia, 43*, 15-26.

Chen, J.-Y., Chen, T.-M., & Dell, G. S. (2002). Word-form encoding in Mandarin Chinese as assessed b the implicit priming task. *Journal of Memory and Language, 46*, 751-781.

Chen, L., & Boland, J. E. (2008). Dominance and context effects on activation of alternative homophone meanings. *Memory and Cognition, 36*, 1306-1323.

Cheung, H., Chen, H.-C., Lai, C. Y., Wong, O. C., & Hills, M. (2001). The development of phonological awareness: Effects of spoken language experience and orthography. *Cognition, 81*, 227-241.

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

Christiansen, M. H., & Chater, N. (1999) Connectionist natural language processing: The state of the art. *Cognitive Science, 23*, 417-437.

Chwilla, D. J., & Kolk, H. (2003). Event-related potential and reaction time evidence for inhibition between alternative meanings of ambiguous words. *Brain and Language, 86*, 167-192.

Connine, C. M. (2004). It's not what you hear but how often you hear it: On the neglected role of phonological variant frequency in auditory word recognition. *Psychonomic Bulletin and Review, 11*, 1084-1089.

Cooper, R. M. (1974). The control of eye fixation by the meaning of spoken language: A new methodology for the real-time investigation of speech perception, memory and language processing. *Cognitive Psychology, 6*, 84-107.

Coulson, S., & van Petten, C. (2002). Conceptual integration and metaphor: An event-related potential study. *Memory and Cognition, 30*, 958-968.

Cutler, A., & Chen, H.-C. (1997). Lexical tone in Cantonese spoken-word processing. *Perception and Psychophysics, 59*, 165-179.

Dahan, D., & Gaskell, M. G. (2007). The temporal dynamics of ambiguity resolution: Evidence from spoken-word recognition. *Journal of Memory and Language, 57*, 483-501.

Dahan, D., Magnuson, J. S., & Tanenhaus, M. K. (2001). Time course of frequency effects in spoken-word recognition: Evidence from eye movements. *Cognitive Psychology, 42*, 317-367.

Dahan, D., Magnuson, J. S., Tanenhaus, M. K., & Hogan, E. M. (2001). Subcategorical mismatches and the time course of lexical access: Evidence for lexical competition. *Language and Cognitive Processes, 16*, 507-534.

Dahan, D., & Tanenhaus, M. (2005). Looking at the rope when looking for the snake: Conceptually mediated eye movements during spoken-word recognition. *Psychonomic Bulletin and Review, 12*, 453-459.

Davis, M. H., Marslen-Wilson W. D., Gaskell, M. G. (2002). Leading up the lexical garden path: Segmentation and ambiguity in spoken word recognition. *Journal of Experimental Psychology: Human Perception and Performance, 28*, 218-244.

Demmechai, T., & Mäkeläinen K. (2001). Recognition of syllables in a tone language. *Speech Communication, 33*, 241-254.

Ding, G. S., Peng, D., & Taft, M. (2004). The nature of the mental representation of radicals in Chinese: A priming study. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*, 530-539.

Dohmes, P., Zwitserlood, P., & Bölte, J. (2004). The impact of semantic transparency of morphologically complex words on picture naming. *Brain and Language, 90*, 203-212.

Duffy, S. A., Morris, R. K., & Rayner , K. (1988). Lexical ambiguity and fixation times in reading. *Journal of Memory & Language, 27*,429-446.

Ellis, L., Derbyshire, A. J., & Joseph, M. E. (1971). Perception of electronically gated speech. *Language and Speech, 14*, 229-240.

Feldman, L. B. (2000). Are morphological effects distinguishable from the effects of shared meaning and shared form? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*, 1431-1444.

Francis, A. L., Ciocca, V., & Ng, B. K. C. (2003). On the (non)categorical perception of lexical tones. *Perception and Psychophysics, 65*, 1029-1044.

Frauenfelder, U. H., & Peeters G. (1998). Simulating the time course of spoken word recognition: An analysis of lexical competition in TRACE. In J., Grainger, & A. M., Jacobs (Eds.), *Localist Connectionist Approaches to Human Cognition*, pp. 101-146. New Jersey: Lawrence Erlbaum Associates.

Frost, R., Kugler, T., Deutsch, A., & Forster, K. I. (2005). Orthographic structure versus morphological structure: Principles of lexical organization in a given language. *Journal of Experimental Psychology, Learning, Memory, and Cognition, 31*, 1293-1326.

Frost, R., Grainger, J., & Rastler, K. (2005). Current issues in morphological processing: An introduction. *Language and Cognitive Processes, 20*, 1-5.

Gandour, J. (1983). Tone perception in Far Eastern language. *Journal of Phonetics, 11*, 149-175.

Gandour, J., Dzemidzic, M., Wong, D., Lowe, M., Tong, Y., Hsieh, L., Satthamnuwong, N., & Luirto, J. (2003). Temporal integration of speech prosody is shaped by language experience: An fMRI study. *Brain and Language, 84*, 318-336.

Gaskell, M. G., & Marslen-Wilson, W. D. (1997). Integrating form and meaning: A distributed model of speech perception. *Language and Cognitive Processes, 12*, 613-656.

Gaskell, M. G., & Marslen-Wilson, W. D. (2001). Lexical ambiguity resolution and spoken word recognition: Bridging the gap. *Journal of Memory and Language, 44*, 325-349.

Gibbs, R. W. (1983). Do people always process the literal meaning of indirect requests? *Journal of Experimental Psychology: Learning, Memory and Cognition, 9*, 524-533.

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

Greber, C., & Frauenfelder, U. H. (1999). On the locus of morphological effects in spoken-word recognition: Before or after lexical identification. *Brain and Language, 68*, 46-53.

Grodner, D., Gibson, E., Tunstall, S. (2002). Syntactic complexity in ambiguity resolution. *Journal of Memory and Language, 46*, 267-295.

Grosjean, F. (1980). Spoken word recognition processes and the gating paradigm. *Perception and Psychophysics, 28*, 267-283.

Grosjean, F. (1985). The recognition of words after their acoustic offset: Evidence and implications. *Perception and Psychophysics, 38*, 299-310.

Grosjean, F. (1996). Gating. *Language and Cognitive Processes, 11*, 597-604.

Hino, Y., Lupker, S. J., & Pexman, P. M. (2002). Ambiguity and synonymy effects in lexical decision, naming, and semantic categorization tasks: Interactions between orthography, phonology, and semantics. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*, 686-713.

Hogaboam, T. W., & Perfetti, C. A. (1975). Lexical ambiguity and sentence comprehension. *Journal of Verbal Learning and Verbal Behavior, 14*, 256-274.

Howie, J. M. (1976). *Acoustic Studies of Mandarin Vowels and Tones*. Cambridge: Cambridge University Press.

Huettig, F., & Altmann, G. T. M. (2005). Word meaning and the control of eye fixation: Semantic competitor effects and the visual world paradigm. *Cognition,*

96, B23-B32.

Jusczyk, P. W., & Luce, P. A. (2002). Speech perception. In H., Pashler, & S., Yantis (Eds.), *Steven's Handbook of Experimental Psychology (Vol. 1): Sensation and Perception*, pp. 493-536. New York: John Wiley & Sons Inc.

Kamber, G., Rayner, K., & Duffy, S. A. (2001). Global context effects on processing lexically ambiguous words: Evidence from eye fixations. *Memory and Cognition, 29*, 363-372.

Kamide, Y., Altmann, G. T. M., & Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and Language, 49*, 133-156.

Kellas, G., Ferraro, F. R., & Simpson, G. B. (1988). Lexical ambiguity and the timecourse of attentional allocation in word recognition. *Journal of Experimental Psychology: Human Perception and Performance, 14*, 601-609.

Kelly, S. D., Barr, D. J., Church, R. B., & Lynch, K. (1999). Offering a hand to pragmatic understanding: The role of speech and gesture in comprehension and memory. *Journal of Memory and Language, 40*, 577-592.

Khouw, E., & Ciocca, V. (2007). Perceptual correlates of Cantonese tones. *Journal of Phonetics, 35*, 104-117.

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.

Li, P., & Yip, M. C. (1998). Context effects and the processing of spoken homophones. *Reading and Writing: An interdisciplinary Journal, 10*, 223-243.

Lichacz, F. M., Herdman, C. M., LeFevre, J., & Baird, B. (1999). Polysemy effects in naming. *Canadian Journal of Experimental Psychology, 53*, 189-193.

Luce, P. A., Goldinger, S. D., Auer, E. T., & Vitevitch, M. S. (2000). Phonetic priming, neighborhood activation, and PARSYN. *Perception and Psychophysics, 62*, 615-625.

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The Neighborhood Activation Model. *Ear and Hearing, 19*, 1-36.

Luo, H., Ni, J., Li, Z., Li, X., Zhang, D., Zeng, F., & Chen, L. (2006). Opposite patterns of hemisphere dominance for early auditory processing of lexical tones and consonants. *Proceedings of the national Academy of Sciences of the United States of America, 103*, 19558-19563.

Magnuson, J. S., Dixon, J. A., Tanenhaus, M. K., & Aslin, R. N. (2007). The dynamics of lexical competition during spoken word recognition. *Cognitive Sciences, 31*, 133-156.

Marslen-Wilson, W., & Warren, P. (1994). Levels of perceptual representation and process in lexical access: Words, phonemes, and features. *Psychological Review, 101*, 653-675.

Marslen-Wilson, W., & Welsh, A. (1978). Processing interactions during word-recognition in continuous speech. *Cognitive Psychology, 10*, 29-63.

Marslen-Wilson, W., & Zwitserlood, P. (1989). Accessing spoken words: The importance of word onsets. *Journal of Experimental Psychology: Human Perception and Performance, 15*, 576-585.

Matin, E., Shao, K. C., & Boff, K. R. (1993). Saccadic overhead: Information-processing time with and without saccades. *Perception and Psychophysics, 53*, 372-380.

Mattys, S. L., & Clark, J. H. (2002). Lexical activity in speech processing: Evidence from pause detection. *Journal of Memory and Language, 47*, 343-359.

McBride-Chang, C., Bialystok, E., Chong, K. K. Y., & Li, Y. P. (2004). Levels of phonological awareness in three cultures. *Journal of Experimental Child Psychology, 89*, 93-111.

McClelland, J. L. & Elman, J. (1986). The TRACE model of speech perception. *Cognitive Psychology, 23*, 1-44.

McMurray, B., Aslin, R. N., Tanenhaus, M. K., Spivey, M. J., & Subik, D. (2008). Gradient sensitivity to within-category variation in words and syllables. *Journal of Experimental Psychology: Human Perception and Performance, 34*, 1609-1631.

McQueen, J. M., & Viebahn, M. C. (2007). Tracking recognition of spoken words by tracking looks to printed words. *The Quarterly Journal of Experimental Psychology, 60*, 661-671.

Miller, George A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63*, 81-97.

Mitterer, H., & McQueen, J. M. (2009). Processing reduced word-forms in speech perception using probabilistic knowledge about speech production. *Journal of Experimental Psychology: Human Perception and Performance, 35*, 244-263.

Moore, C. B., Jongman, A. (1997). Speaker normalization in the perception of Mandarin Chinese tones. *The Journal of the Acoustic Society of America, 102*, 1864-1877.

Niswander-Klement, E., & Pollatsek, A. (2006). The effects of root frequency, word frequency, and length on the processing of prefixed English words during reading. *Memory and Cognition, 34*, 685-702.

Norris, D, McQueen, J. M., & Cutler, A. (2000). Merging information in speech recognition: Feedback is never necessary. *Behavioral and Brain Sciences, 23*, 299-370.

Parkard, J. L. (1999). Lexical access in Chinese speech comprehension and production. *Brain and Language, 68*, 89-94.

Peng, D. L., Liu, Y., & Wang, C. M. (1999). How is access representation organized? The relation of polymorphemic words and their morphemes in Chinese. In J. Wang, A. W. Inhoff, & H. C. Chen (Eds). *Reading Chinese Script: A Cognitive Analysis* (pp. 65-89). Mahwah, New Jersey: Lawrence Erlbaum.

Peng, S.-H. (1997). Production and perception of Taiwanese tones in different tonal and prosodic contexts. *Journal of Phonetics, 25*, 371-400.

Pollatsek, A., & Hyönä, J. (2005). The role of semantic transparency in the processing of Finnish compound words. *Language and Cognitive Processes, 20*, 261-291.

Protopapas A. (1999). Connectionist modeling of speech perception. *Psychological Bulletin, 125*, 410-436.

Raphael, L.J. (2005). Acoustic cues to the perception of segmental phonemes. In D.B. Pisoni & R.E. Remez (Eds.), *The handbook of speech perception*, pp. 182-206. Blackwell Publishing Ltd: Oxford, UK

Rayner , K. (1998).Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin, 124*, 372-422.

Rayner, K., & Frazier, L. (1989). Selection mechanisms in reading lexically ambiguous words. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15*, 779-790.

Rubenstein, H., Garfield, L., & Millikan, J. A. (1970). Homographic entries in the internal lexicon. *Journal of Verbal Learning and Verbal Behavior, 9*, 487-492.

Salverda, A. P., Dahan, D., & McQueen, J. M. (2003). The role of prosodic boundaries in the resolution of lexical embedding in speech comprehension. *Cognition, 90*, 51-89.

Schirmer, A., Tang, S., Penney, T. B., Gunter, T. C., & Chen, H. (2005). Brain response to segmentally and tonally induced semantic violations in Cantonese. *Journal of Cognitive Neuroscience, 17*, 1-12.

Schriefers, H., Zwitserlood, P., & Roelofs, A. (1991). The identification of morphologically complex spoken words: Continuous processing or decomposition? *Journal of Memory and Language, 30*, 26-47.

Sereno, S. C., O'Donnell, P. J., & Rayner, K. (2006). Eye movements and lexical ambiguity resolution: Investigating the subordinate-bias effect. *Journal of Experimental Psychology: Human Perception and Performance, 32*, 335-350.

Sereno, S. C., Pacht J. M., & Rayner, K. (1992). The effect of meaning frequency on processing lexically ambiguous words: Evidence from eye fixations. *Psychological Science, 14*, 328-333.

Simpson, G. B. (1981). Meaning dominance and semantic context in the processing of lexical ambiguity. *Journal of Verbal Learning & Verbal Behavior, 20*, 120-136.

Siok, W. T., Jin, Z., Fletcher, P., & Tan, L. H. (2003). Distinct brain regions associated with syllable and phoneme. *Human Brain Mapping, 18*, 201-207.

Slowiaczek, L. M., Soltano, E. G., Wieting, S. J., & Bishop, K. L. (2003). An investigation of phonology and orthography in spoken-word recognition. *The Quarterly Journal of Experimental Psychology, 56A*, 233-262.

Sum, K.-W. (2003). Spoken word recognition in Cantonese: Significance of onset, rime and tone in monosyllabic words. *Unpublished Master Thesis*.

Swinney, D. A. (1979). Lexical access during sentence comprehension:

(Re)consideration of context effects. *Journal of Verbal Learning & Verbal Behavior, 18*, 645-659.

Tabossi, P. (1988). Accessing lexical ambiguity in different types of sentential contexts. *Journal of Memory & Language, 27*, 324-340.

Taft, M., & Forster, K. (1975). Lexical storage and retrieval of prefixed words. *Journal of Verbal Learning and Verbal Behavior, 14*, 639-647.

Taft, M., Hambly, G. & Kinoshita, S. (1986). Visual and auditory recognition of prefixed words. *Quarterly Journal of Experimental Psychology, 38A*, 351-366.

Taft, M., & Kougious, K. (2004). The processing of morpheme-like units in monomorphemic words. *Brain and Language, 90*, 9-16.

Taft, M., Liu, Y., Zhu, X. P. (1999). Morphemic processing in reading Chinese. In J. Wang, A. W. Inhoff, & H. C. Chen (Eds). *Reading Chinese Script: A Cognitive Analysis* (pp. 65-89). Mahwah, New Jersey: Lawrence Erlbaum.

Taft, M., & Zhu, X. P. (1995). The representation of bound morphemes in the lexicon: A Chinese study. In L. Feldman (Ed.), *Morphological aspects of language processing*, pp. 293-319. Hillsdale, NJ: Erlbaum.

Tanenhaus, M. K., Spivey-Knowlton, M. J. (1996). Eye-tracking. *Language and Cognitive Processes, 11*, 583-588.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science, 268(5217)*, 1632-1634.

Titone, D. (1998). Hemispheric differences in context sensitivity during lexical ambiguity resolution. *Brain and Language, 65*, 361-394.

Tsang (2006). Semantic access in Chinese speech comprehension: The role of morpheme frequency and context. *Unpublished Master Thesis*.

Tyler, L. K. (1984). The structure of initial cohort: Evidence from gating. *Perception and Psychophysics, 36*, 417-427.

Tyler, L. K., Marslen-Wilson, W., Rentoul, J., & Hanney, P. (1988). Continuous and discontinuous access in spoken word-recognition: The role of derivational prefixes. *Journal of Memory and Language, 27*, 368-381.

Vance, T. J. (1976). An experimental investigation of tone and intonation in Cantonese. *Phonetica, 33*, 368-392.

Vroomen, J., & de Gelder, B. (1997). Activation of embedded words in spoken word recognition. *Journal of Experimental Psychology: Human Perception and Performance, 23*, 710-720.

Vu, H., & Kellas, G. (1999). Contextual strength modulates the subordinate bias effect: Reply to Rayner, Binder, and Duffy. *The Quarterly Journal of Experimental Psychology, 52A*, 853-855.

Vu, H., Kellas, G., & Paul, S. T. (1998). Sources of sentence constraint on lexical ambiguity resolution. *Memory and Cognition, 26*, 979-1001.

Walley, A. C., Michela, V., & Wood, D. R. (1995). The gating paradigm: Effects of presentation format on spoken word recognition by children and adults. *Perception and Psychophysics, 57*, 343-351.

Warren, P., & Marslen-Wilson, W. D. (1987). Continuous uptake of acoustic cues in spoken word recognition. *Perception and Psychophysics, 41*, 262-275.

Wong, A. W. K., & Chen, H.-C. (2008). Processing segmental and prosodic information in Cantonese word production. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 34*, 1172-1190.

Wong, E. K. F. (2000). The time course of semantic activation in reading Chinese two-character words. *Unpublished Doctor Thesis*.

Wurm, L. H. (1997). Auditory processing of prefixed English words is both continuous and decompositional. *Journal of Memory and Language, 37*, 438-461.

Ye, Y., & Connine, C. M. (1999). Processing spoken Chinese: The role of tone information. *Language and Cognitive Processes, 14*, 609-630.

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

Zatorre, R. J., & Belin, P. (2001). Spectral and temporal processing in human auditory cortex. *Cerebral Cortex, 11*, 946-953.

Ziegler, J. C., Ferrand, L., & Montant, M. (2004). Visual phonology: The effects of orthographic consistency on different auditory word recognition tasks. *Memory and Cognition, 32*, 732-741.

Zhou, X. L., Marslen-Wilson, W., Taft, M., & Shu, H. (1999). Morphology, orthography, and phonology in reading Chinese compound words. *Language and Cognitive Processes, 14*, 526-565.

Zhou, X. L., Qu, Y. X., Shu, H., Gaskell, G., & Marslen-Wilson (2004). Constraints of lexical tone on semantic activation in Chinese spoken word recognition. *Acta Psychology Sinica, 36*, 379-392.

Zwitserlood, P., Bolwiender, A., & Drews, E. (2005). Priming morphologically complex verbs by sentence contexts: Effects of semantic transparency and ambiguity. *Language and Cognitive Processes, 20*, 395-415.

Appendix A: Materials in Experiments 1A and 2A

	Target			Target duration	No. of gates	Onset-sharing		Rime-sharing		Syllable-sharing		Tone-sharing	
1	bou3	布	Cloth	494	13	bui1	杯	dou1	刀	bou1	煲	coi3	菜
2	ngau1	勾	Hook	647	17	ngaai4	崖	kau4	球	ngau4	牛	ziu1	魚
3	caa4	茶	Tea	633	16	ce1	中	faa1	花	caa1	叉	kiu4	橋
4	coeng4	牆	Wall	676	18	cim1	簽	soeng1	箱	coeng1	窗	wan4	雲
5	cong2	廠	Factory	722	19	cung4	蟲	long4	狼	cong4	床	daan2	蛋
6	dang3	凳	Chair	512	13	deng1	釘	gang1	鼻	dang1	燈	taap3	塔
7	dou6	稻	Rice	584	15	dau2	豆	bou2	饅	dou2	島	bei6	鼻
8	ci4	池	Pool	689	18	co3	鈸	zi3	鼠	ci3	刺	lou4	爐
9	faan6	飯	Meal	716	18	fan4	墳	lann4	欄	faan4	帆	jip6	葉
10	fu3	褲	Trousers	772	20	fo2	火	gu2	鼓	fu2	虎	tou3	兔
11	geng3	鏡	Mirror	590	15	gun2	管	beng2	餅	geng2	頸	gim3	劍
12	gwai2	鬼	Ghost	605	16	gwaa1	瓜	mai1	咪	gwai1	龜	cou2	草
13	haai5	蟹	Crab	693	18	ho4	河	naai4	奶	haai4	鞋	ji5	耳
14	hau2	口	Mouth	611	16	haa4	蝦	tau4	頭	hau4	猴	so2	鎖
15	jin2	硯	Inkstone	649	17	jing1	鷹	bin1	鞭	jin1	煙	mat6	襪
16	jyu5	雨	Rain	650	16	jau4	油	syu4	署	jyu4	魚	ngai5	蟻
17	leoi6	淚	Tears	765	20	lo4	籮	ceoi4	錘	leoi4	雷	gwai6	櫃
18	maau1	貓	Cat	740	19	mei4	眉	caau4	巢	maau4	矛	biu1	錶
19	sing2	繩	Rope	716	18	saam1	衫	bing1	冰	sing1	星	zeng2	井
20	syu6	樹	Tree	939	24	si1	獅	zyu1	豬	syu1	書	lou6	路
21	tai4	蹄	Hoof	603	16	taai1	呔	gai1	雞	tai1	梯	ngaa4	牙
22	zau6	袖	Sleeve	705	18	zeoi2	嘴	gau2	狗	zau2	酒	wai6	胃
23	zin1	氈	Blanket	649	17	zyun3	鑽	sin3	扇	zin3	箭	bat1	筆
24	zung2	粽	Zong	589	15	zam1	針	gung1	弓	zung1	鐘	wun2	碗

Appendix B: Materials in Experiments 1B and 3

	Syllable	Duration	No. of gates	Dom		Sub		SD		SS		PD		PS	
1	fung1	886	23	風	Wind	蜂	Bee	風箏	fung1zaang1	蜂巢	fung1caau4	旋風	syun4fung1	蜜蜂	mat6fung1
2	bui3	671	17	背	Back	貝	Shell	背囊	bui3nong4	貝殼	bui3hok3	駝背	to4bui3	扇貝	sin3bui3
3	zin3	706	18	箭	Arrow	墊	Cushion	箭靶	zin3baa2	墊褥	zin3juk6	弓箭	gung1zin3	坐墊	zo6zin3
4	zuk1	240	6	竹	Bamboo	燭	Candle	竹筍	zuk1seon2	燭台	zuk1toi4	炮竹	paau3zuk1	蠟燭	laap6zuk1
5	zyu1	724	19	豬	Pig	珠	Pearl	豬肉	zyu1juk6	珠寶	zyu1bou2	箭豬	zin3zyu1	珍珠	zan1zyu1
6	bou2	675	17	寶	Treasure	堡	Castle	寶石	bou2sek6	堡壘	bou2leoi5	元寶	jyun4bou2	城堡	sing4bou2
7	sin1	742	19	鮮	Fresh	仙	Fairy	鮮奶	sin1naai5	仙人	sin1jan4	海鮮	hoi2sin1	神仙	san4sin1
8	lung4	740	19	龍	Dragon	籠	Cage	龍蝦	lung4haa1	籠子	lung4zi2	恐龍	hung2lung4	鳥籠	niu5lung4
9	bing1	583	15	冰	Ice	兵	Solider	冰窖	bing1uk1	兵器	bing1hei3	溜冰	lau4bing1	士兵	si6bing1
10	ji1	808	21	衣	Clothes	醫	Doctor	衣架	ji1gaa3	醫生	ji1saang1	毛衣	mou4ji1	獸醫	sau3ji1
11	hung4	676	17	熊	Bear	洪	Flood	熊貓	hung4maau1	洪水	hung4seoi2	樹熊	syu6hung4	山洪	saan1hung4
12	bou3	494	13	布	Cloth	報	Report	布匹	bou3pat1	報紙	bou3zi2	紗布	saa1bou3	海報	hoi2bou3
13	laam5	650	17	纜	Rope	艦	Vessel	纜車	laam5ce1	艦艇	laam5teng5	繩纜	sing4laam5	戰艦	zin3laam5
14	ziu1	818	21	蕉	Banana	招	Attract	蕉葉	ziu1jip6	招財	ziu1coi4	香蕉	hoeng1ziu1	街招	gaai1ziu1
15	wun2	794	20	碗	Bowl	腕	Wrist	碗盤	wun2pun4	腕錶	wun2biu1	湯碗	tong1wun2	手腕	sau2wun2
16	coeng1	679	17	窗	Window	槍	Pistol	窗簾	coeng1lim4	槍械	coeng1haai6	門窗	mun4coeng1	手槍	sau2coeng1
17	haai4	660	17	鞋	Shoes	骸	Bones	鞋帶	haai4daai3	骸骨	haai4gwat1	皮鞋	pei4haai4	疑骸	si1haai4
18	ci4	618	16	池	Pool	匙	Spoon	池塘	ci4tong4	匙羹	ci4gang1	水池	seoi2ci4	湯匙	tong1ci4
19	zi2	676	17	紙	Paper	指	Finger	紙幣	zi2bai6	指紋	zi2man4	報紙	bou3zi2	戒指	gaai3zi2
20	gaap3	304	8	甲	Armor	鴿	Pigeon	甲蟲	gaap3cung4	鴿子	gaap3zi2	盔甲	kwai1gaap3	白鴿	baak6gaap3

Note: Dom = dominant meaning; Sub = subordinate meaning; SD = succeeded context-dominant meaning; SS = succeeded context-subordinate meaning; PD = preceded context-dominant meaning; PS = preceded context-subordinate meaning.

Appendix C: Materials in Experiments 2B and 2C

	Target			Duration	Onset-sharing		Onset-plus-sharing		Embedded word	
1	caau4	巢	Nest	764			caai4	柴	caa4	茶
2	saam1	衫	Clothes	850			saan1	山	saa1	沙
3	sing1	星	Star	791			siu1	燒	si1	獅
4	lou4	爐	Oven	737			long4	狼	lo4	籮
5	haai4	鞋	Shoes	749			hau4	猴	haa4	蝦
6	ngaai4	崖	Cliff	658			ngau4	牛	ngaa4	牙
7	baau1	包	Bread	724	bou1	煲			baa1	爸
8	coi3	菜	Vegetable	879	cek3	尺			co3	鈡
9	gun2	管	Pipe	708	geng2	頸			gu2	鼓
10	zyun1	磚	Brick	862	zung1	鍾			zyu1	豬
11	goek3	腳	Foot	401	gim3	劍			goe3	靴
12	sou1	鬚	Beard	1017	sat1	蝨			so1	梳
13	bin1	鞭	Whip	664	bat1	筆	biu1	錶		
14	zin1	氈	Blanket	748	zuk1	竹	ziu1	燒		
15	cou2	杯	Glass	775	caang2	橙	cong2	磡		
16	kam4	琴	Piano	718	kei4	旗	kau4	球		
17	tau4	頭	Head	689	tin4	田	tai4	睇		
18	jau4	油	Oil	673	joeng4	羊	jan4	人		

Note: Each target was paired with only two competitor conditions (see text)