Proceedings of DiSS-LPSS Joint Workshop 2010,
25-26 September 2010, Tokyo, Japan, pp. 39-42

Contextual effects in recognizing reduced words in spontaneous speech

Shu-Chuan Tseng and Tzu-Lun Lee

Institute of Linguistics, Academia Sinica, Taiwan
{tsengsc, tzulun}@gate.sinica.edu.tw

Abstract

This study investigates the effects of context on recognizing reduced word forms in spontaneous speech. Sixteen high-frequency disyllabic targets, eight disyllabic and eight combinations of monosyllabic words are presented to 48 subjects in a spoken word recognition experiment in three conditions: in their original context, in isolation, and embedded in a carrier sentence. Results show that context, degree of reduction, word unit type, gender, and age group all show an effect on the accuracy rates of recognizing the target items. Most interestingly, while a meaningful context helps recognize reduced word forms, a less meaningful context inhibits the recognition more than no context.

Index Terms: Spoken word recognition, context effect

1. Introduction

Spoken word recognition is a complex process involving integration of acoustic, lexical, and grammatical information. The process may be carried out differently due to the conditions of speech context. In an experimental design, in which the subjects only need to focus on fine, acoustic cues, presented in tiny frames, the working pattern certainly distinguishes itself from that in realistic speech communication. In other words, the task of matching sensory cues with lexical entries in the mental lexicon may be conducted in different manners. Words in a semantically coherent sentence can be more accurately recognized than those in isolation [1]. Studies have also shown that high-frequency words draw faster and more accurate responses than low-frequency words in experiment conditions [2]. At the same time, high-frequency words also tend to be more reduced acoustically than less frequently used words [3]. High-frequency words can be better recognized. But when they are highly reduced, that is, only incomplete (sometimes even incorrect), sensory cues are provided, it is also likely that they would result in a recognition barrier. How these effects interact with each other is a challenging topic. Autonomous and interactive models have been suggested to explain when the effect of the context starts to act in the process of recognizing spoken words [4]. Either kind of the models has to allow a space for the effect of the context to function at an early stage, if we take into account the fact from real speech communication. In spite of limited or misleading acoustic cues provided in spontaneous speech, reduced speech can still be correctly recognized with proper syntactic and semantic cues from the context [4]. Previous studies also conclude that words heard in the context are often recognized long before their full acoustic signal has been delivered [5, 6].

Many studies have demonstrated these effects by adopting on-line task designs such as phoneme monitoring, lexical decision, and gating. However, realistic speech data like informal reduced word forms have not been used to experiment stimuli, until recent studies have focused on the effects of natural speech on spoken word recognition [7]. Highly reduced words can be well recognizable only when presented in their original context, because acoustic cues along with semantic and syntactic information are provided through the context. And they cannot be properly recognized without context or with limited context information of adjacent syllables [7]. The main purpose of this study is to use well-defined reduced word forms extracted from the spontaneous speech corpora of Taiwan Mandarin embedded in three context conditions: in the original context, no context, and in a controlled context to experiment how well subjects can recognize the phonetically reduced target words.

2. Method and data

2.1 Reduced speech: syllable contraction

Before we proceed to introduce our stimuli for the experiment, we need to clarify the phenomenon of reduced speech in Mandarin: syllable contraction. Syllable contraction can sometimes result in word mergers, and mergers sometimes have influences on the writing system. Characters borrowed from newly invented or existing characters, may be used to represent the resulting merger syllables. Phonetically, the main characteristic of syllable contraction is either an omission of syllables or an omission of syllable boundaries. Extreme cases are deletions of syllables, where two syllables are merged into one. For instance, the word “therefore” *suoyi* is a disyllabic word. When produced as a single syllable [sue], it is a case of syllable contraction. Also, when none of the syllables is completely omitted, and no clear perceptual and acoustic cues for a syllable boundary can be obtained, it is also regarded as a syllable contraction. Take the frequently produced disyllabic word “if” *ruguo* as an example, *ruguo* is often reduced to *ruo* in spontaneous Mandarin speech. A segmental deletion leads to the change of syllable structure from CV+CV to CV+V. Because the boundary of the syllables cannot be perceptually identified, it is considered a case of syllable contraction, too. Syllable contraction can occur within a multi-syllabic word, for instance *suoyi* and *ruguo*. It can also occur in syllables across word boundary, for instance *youyi* (have one). In our stimuli, we selected two groups of syllable contraction. One group contains disyllabic words; another group contains combinations of monosyllabic words. The reason to have this distinction is to test whether different types of word units play a role in the perception of spoken word.

2.2 Data

The stimuli used for this experiment was extracted from two speech corpora of Taiwan Mandarin constructed at Academia Sinica: the Mandarin Conversational Dialogue Corpus (MCDC) and the Mandarin Map Task Corpus (MMTC) [8]. The MCDC corpus consists of eight orthographically transcribed free conversations, in total eight hours of data. The MMTC corpus comprises of 26 orthographically transcribed

tasked-oriented conversations. The transcripts were processed by the automatic word segmentation and POS tagging system developed for modern Mandarin by the CKIP at Academia Sinica [9]. One thing to note is that word segmentation criteria may vary according to different approaches of language processing or grammar. For processing our data, we followed the word segmentation criteria of the CKIP system. From the annotated occurrences of syllable contraction based on the abovementioned criteria, we selected 16 high-frequency disyllabic targets of syllable contraction spoken by a female speaker as listed in Table 1. In the abovementioned spoken corpora, 327,590 words were segmented, the percentage of the disyllabic words in disyllabic targets and the more frequent monosyllabic words in monosyllabic combination targets are given in the table, too.

Table 1. *Words used in the experiment.*

Disyllabic words	% in corpora	Combinations of monosyllabic words	% in corpora
Jiushi (is)	0.83	*Da-de* (big-structural particle)	2.45
Ranhou (then)	0.65	*Dou-shi* (all-is)	2.2
Juede (think)	0.49	*Ta-shi* (it-is)	2.2
Yinwei (because)	0.48	*Ni-you* (you-have)	1.29
Suoyi (so)	0.33	*Nei-yi* (that-one)	0.9
Qishi (in fact)	0.27	*You-yi* (have-one)	0.9
Zheyang (so, such)	0.25	*Ni-yao* (you-want)	0.51
Keyi (can)	0.2	*Ne-zhong* (that-kind)	0.27

They are equally distributed in two categories. The group of disyllabic words consists of syllable contraction which forms a lexical unit. The group of monosyllabic word combinations consists of syllable contraction across the word boundary. Adopting the criteria of [10], for each target two variants were selected with two degrees of acoustic information from the corpus: contracted and semi-contracted. The selected 32 speech stretches were cut from the original context and were presented to the subjects in three contextual conditions: 1) Full context: in the original context, 2) Isolation context: no context, and 3) Controlled context: with less meaningful context, i.e. embedded into a carrier sentence: *'I spoke the word _____'*. Table 2 shows examples of these three contextual conditions.

Table 2. *Examples of contextual conditions.*

Context	Example
Full	Rao ge liang san quan jiu **keyi** zhaodao na suan henhao Go around two three circle just **can** find that is pretty good *It was pretty good to find a parking space, just going around two or three times*
Isolation	**keyi** **Can**
Controlled	wo nian le **keyi** zhege ci I speak PAST-tense **can** this word *I spoke the word "can"*

In addition to the targets taken from the spontaneous speech corpora, we recorded a version of clearly read speech of these 16 targets. All of the sound files were tape-recorded by a DAT recorder (Sony TCD-D10 Pro II DAT), and then converted to WAV format with Steinberg Cubasis VST Convertor (US-428) at a sampling rate of 48 kHz. Adobe Audition 1.0 was used to extract the target forms from their original context and PRAAT was used for manipulating the targets with the carrier sentence [11]. Please also note that the data extracted from the speech corpora and the clearly read version of target words were all produced by the same female speaker, though recorded in different times.

In summary, 16 disyllabic target word forms (equally distributed in two categories: disyllabic words and combinations of monosyllabic words), with two degrees of reduction (Contracted and Semi-contracted) were presented in three contextual conditions (Full, Isolation, and Controlled). As a result, 112 trials were used for the experiment consisting of 96 (16*2*3) trials extracted from the spontaneous speech corpora and 16 trials which were additionally recorded clear speech.

2.3 Subjects

48 subjects sampled from the citizens of Taipei City based on age ranges 20-29, 30-39, 40-49, and 50-59, equally distributed in gender, participated in this experiment. All were native Mandarin speakers with no known hearing problems.

2.4 Procedure

The experiment was carried out with a working panel shown on a 14 inch screen of a laptop computer (IBM T43 2665-02V) with the Windows XP Pro operating system. High-quality monitor headphones (Audio Technica ATH-M50) were used for listening to the stimuli. For Full and Controlled context conditions, the panel shows the context in Chinese characters on the screen, leaving a blank space for the subject to type in the recognized content of the stimuli. Subjects were seated in a soundproof booth with the laptop computer in front of them. 112 trials were presented to the subjects in a random order. To familiarize the subjects with the procedure, four training trials with three conditions and the read version were set up. Because the subjects were not linguists, it was not possible to use the IPA symbols to write down what they heard from the stimuli. Thus, they were instructed to key in their answers according to the following principles:

(1) Type the Chinese characters they think the stimuli are;

(2) Type the most likely Chinese characters they think it is, if they are not sure about what exactly the stimuli are,

(3) Type the phonetic symbols of the Mandarin Phonetic Symbols I to write down the most likely pronunciation of the stimuli, if they cannot find the suitable characters. The Mandarin Phonetic Symbols I is a transcription convention used in Taiwan, also called Zhuyin Fuhao or Bopomofo.

If the subjects can neither recognize which characters nor which sounds the stimuli are, they can leave the answer open and proceed. Subjects can listen to the sound files as many times as they wanted with no time limit for the experiment. Volume adjustment was also permitted to suit the hearing ability of the subjects.

2.5 Factors

An answer is only considered correct when the exactly matched Chinese characters or Mandarin phonetic symbols were answered. Answers left blank were considered as incorrect. We studied the distribution of correctly recognized tokens in terms of context, degree of reduction, word unit type, gender, and age group. Stimuli presented without context should be harder to recognize than with context; more reduced stimuli harder than less reduced, and two-words units harder than one-word units. As for subject-related factors, the gender and age, we did not have any prior prediction. But triggered by the fact that the male speakers tend to produce more syllable contraction than the female, approximately 4:3, we selected gender- and age equally distributed subjects to test

whether gender and age have also an effect in the task of spoken word recognition.

3. Results

3.1 Overall results

4,608 (96 trials *48 subjects) reduced target tokens extracted from spontaneous speech and 768 (16*48) clearly read target tokens were tested in the experiment. Table 3 shows the results classified in different contexts. The clearly read trials were mainly used to test whether subjects can identify the target words without syntactic and semantic context, when the sensory cues are complete. The subjects achieved a 92.97% correction rate. All 54 errors were monosyllabic word combinations. Without context, some of the monosyllabic word combinations were not recognized by the subjects. The rest of the targets were recognized correctly, suggesting that sensory cues are sufficient for the task of recognizing these target words (at least the disyllabic word) without the aid of contextual information. In the reduced speech, although also presented without context, only 46.61% of the targets in the Isolation context were correctly recognized, which is clearly different from the clearly read speech. This result shows that reduced speech does result in difficulties in recognizing spoken word forms. With full context, the rate of correctly identified tokens is much higher than the other two context conditions. However, it is still a bit lower than the clearly read speech forms. Interestingly, we also found that the overall performance in the context of carrier sentence is worse than without any contextual information, suggesting that inappropriate semantic information could result in an intervening effect in recognizing reduced speech.

Table 3. *Overall results in different contexts.*

Context	Correct # (%)	Incorrect # (%)	Total # (%)
Full	1,396 (90.89)	140 (9.11)	1,536 (100)
Isolation	716 (46.61)	820 (53.39)	1,536 (100)
Controlled	456 (29.69)	1,080 (70.31)	1,536 (100)
Total	**2,568 (55.73)**	**2,040 (44.27)**	**4,608 (100)**
Clearly read (isolation)	714 (92.97)	54 (7.03)	768 (100)

3.2 Effects of factors

Table 3 identified a clear effect of the context. We furthermore illustrated the percentage of correct answers which were arranged in degree of reduction, word unit type, gender and age group. As illustrated in Figure 1, the stimuli-related factors (degree of reduction and word unit) show greater differences in terms of the contextual conditions than the subject-related factors. For statistical analysis, we adopted the logistic regression likelihood ratio test to investigate whether this is a statistically significant effect of each individual factor and the cross-interaction between context factor and the two stimuli-related factors: degree of reduction and word unit type. The results are summarized in Table 4. All factors: context, degree of reduction, word unit type, gender, and age group show an effect in the accuracy rate of the spoken word recognition experiment. Cross-effect of context and degree of reduction also shows an effect influencing the accuracy rate. However, although we noticed a difference of word unit type in terms of the contextual conditions, no statistically significant effect was found.

Table 4. *Logistic Regression Likelihood Ratio Test. (Significance level at 0.01)*

	LR Chisq	Df	Pr(>Chisq)
Context	1428.79	2	< .001*
Degree of reduction	37.30	1	< .001*
Word unit type	49.27	1	< .001*
Gender	6.80	1	< .01*
Age group	16.23	3	< .01*
Context:Degree of reduction	10.84	2	< .01*
Context:Word unit	5.59	2	< .1

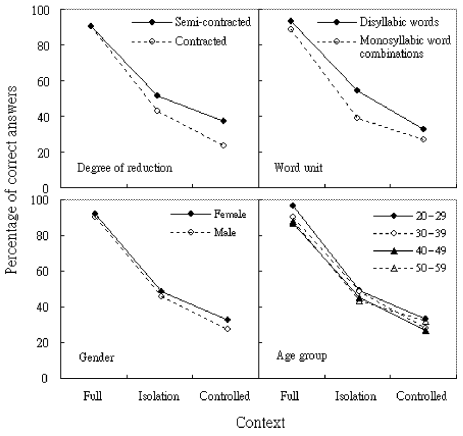

Figure 1. Percentage of correct answers.

To study each factor more closely, we used logistic regression measures. Compared with the Controlled context, the Isolated context [*Exp(β)=2.08, p<.001*] and the Full context [*Exp(β)=30.51, p<.001*] show different degrees of effect. The Full context facilitates spoken word recognition in a much larger extent than the Isolated context. With respect to degree of reduction, the semi-contracted targets [*Exp(β)=1.97, p<.001*] were more easily recognized than the contracted ones. One-word units can be better identified than two-words units, although they are all high-frequency words. The analysis shows that it is easier to identify targets which are disyllabic word units [*Exp(β)=1.33, p<.05*] than targets which are monosyllabic combinations. Interestingly, we found that female subjects did better than male subjects [*Exp(β)=0.83, p<.01*] and the youngest age group of subjects did better than the other age groups [*30-39yrs: Exp(β)=0.79, p<.05; 40-49yrs: Exp(β)=0.68, p<.001; 50-59yrs: Exp(β)=0.74, p<.01*].

Table 5. *Logistic Regression on semantic accuracy.* (*Significance level at 0.05*)

	β	Std. Error	z value	Pr(>|z|)	Exp(β)
(Intercept)	-1.06	0.13	-8.41	< .001*	0.35
Context (baseline: Controlled)					
Isolated	0.73	0.14	5.23	< .001*	2.08
Full	3.42	0.18	19.12	< .001*	30.51
Degree of reduction (baseline: Contracted)					
Semi-contracted	0.68	0.11	5.91	< .001*	1.97
Word unit (baseline: Monosyllabic word combinations)					
Disyllabic words	0.28	0.11	2.49	< .05*	1.33
Gender (baseline: Female)					
Male	-0.18	0.07	-2.61	< .01*	0.83
Age group (baseline: 20-29)					
30-39	-0.23	0.10	-2.34	< .05*	0.79
40-49	-0.38	0.10	-3.82	< .001*	0.68
50-59	-0.30	0.10	-2.98	< .01*	0.74

4. Discussion

This study focuses on the influence of contextual information on the recognition accuracy of reduced spoken words. In spontaneous speech, high-frequency words tend to be reduced, i.e. acoustic cues (sensory cues) provided to the listeners are incomplete or inadequate, although due to frequency effect, high-frequency words should facilitate understanding and identification [3]. Studies have shown that it was difficult for listeners to understand highly reduced words in isolation. They need to use acoustic and semantic/syntactic cues from the context to decode reduced word [7]. In this study, we are mainly concerned with the issue of context effect. But we took into account of more factors in our experiment: degree of reduction, word unit types, gender, and age group. We have shown that reduced word forms in their original sentential context are easier to recognize than those in an irrelevant context such as a less meaningful carrier sentence. The result suggests that even if words which are heavily reduced may be recognized with the syntactic or semantic information provided by a meaningful context. As mentioned in previous studies, poorly articulated words may not be well recognized until more contextual information of that word has been provided [1]. Moreover, the recognition performance turned out to be worse in the context of carrier sentence than in isolation, indicating inappropriate context prohibits word recognition more than no context.

Furthermore, we also found that context effects are sensitive to the degree of acoustic properties of reduced words. It was more difficult to correctly recognize word forms which were more reduced than less reduced. We have experimented on this condition by selecting stimuli from realistic spontaneous speech corpora. As shown in Figure 1, in the Isolation context, semi-contracted targets were more correctly identified than the contracted ones. That is, the extent of sensory cues does help recognize reduced word forms, when no context is provided. In consideration of word unit type, the result suggests that semantically coherent disyllabic words facilitate listeners to associate meaning and form of the reduced words more than combinations of monosyllabic words which are semantically ambiguous without context. This effect is especially apparent in the isolation condition.

In the design of this experiment, the visual information provided to the subjects with the carrier sentence shown on the screen should not affect the recognition result, as the original sentential context was also printed on the screen in the same layout. We would try to make the experiment solely verbal, i.e. no visual information is involved and to balance the stimuli with additional fillers.

In addition, we used the CKIP word segmentation criteria for preparing our stimuli, because there is no general rule to Chinese word segmentation. The CKIP system was designed to cut word units as small as possible. If we are considering the effect of semantic cues provided by the context, we need to be more careful about the definition of words in the experiment. A more semantically-defined word list would probably be more suitable.

With regard to the listener-related factors, we found that the effect of gender and age group had only a marginal influence on the recognition of reduced word forms in our experiment. Female subjects did slightly better than male subjects. The youngest subject group outperformed the other three groups. This seems to reflect the differences of gender and age in the production data. More analysis on the relationship between defined qualities of speech production and the perception performance can be done to study this issue more closely.

5. Conclusion

Our study has confirmed the effect of context in recognizing reduced speech forms extracted from spontaneous speech data. Effects of degree of reduction, word unit, gender, and age group were also clearly identified in the results. In presenting the results, we regarded only the exactly matched characters or phonetic symbols as correct answers. Some of the stimuli were extremely reduced, so that recognition of the exactly matched answers could be difficult. Some of the answers should be further analyzed in terms of the actual phonetic representation of the stimuli. More groups of answers should be sub-classified to study the relationship of the sensory input of the reduced forms, the perceived phonetic form, and the associated semantic interpretation with or without contextual information. The answers typed by the subjects were largely diverse. Currently, we are re-processing their phonetic forms for further analysis.

6. Acknowledgements

This study was financially supported by the National Science Council, under grant NSC 96-2411-H-001-067-MY2 and Academia Sinica. We would like to thank Alvin Chen for his help with the statistical analysis.

7. References

[1] Grant, K.W. and P.F. Seitz, "The recognition of isolated words and words in sentences: Individual variability in the use of sentence context", The Journal of the Acoustical Society of America, 107: 1000-1011, 2000.

[2] Benki, J.R., "Quantitative evaluation of lexical status, word frequency, and neighborhood density as context effects in spoken word recognition", The Journal of the Acoustical Society of America, 113: 1689-1705, 2003.

[3] Jurafsky, D., et al., "Probabilistic relations between words: evidence from reduction in lexical production", Frequency and the emergence of linguistic structure, ed. by Joan Bybee and Paul Hopper, 229-54. 2001, Amsterdam: John Benjamins.

[4] Frauenfelder, U.H. and L.K. Tyler, "The process of spoken word recognition: An introduction", Cognition, 25(1-2): 1-20, 1987.

[5] Grosjean, F., "Spoken word recognition processes and the gating paradigm", Perception & Psychophysics, 28(4): 267-283, 1980.

[6] Marslen-Wilson, W.D., "Functional parallelism in spoken word-recognition", Cognition, 25(1-2): 71-102, 1987.

[7] Ernestus, M., H. Baayen, and R. Schreuder, "The Recognition of Reduced Word Forms", Brain and Language, 81(1-3): 162-173, 2002.

[8] Tseng, S.C., "Processing spoken Mandarin corpora", Traitement automatique des langues. Special issue: Spoken corpus processing 45(2): 89-108, 2004.

[9] Chen, K.J., et al., "Sinica corpus: Design methodology for balanced corpora", PACLIC 11, 167-176, 1996.

[10] Cheng, C. and Y. Xu, "When and how disyllables are contracted into monosyllables in Taiwan Mandarin", The Journal of Acoustical Society of America, 123: 3864, 2008.

[11] Boersma, P. and D. Weenink, "Praat: doing phonetics by computer [Computer program]", Version 5.1.03 from http://www.praat.org/, retrieved 21 March 2009.