Prosodic Cues in the Interpretation of Echo Questions in Chinese Spoken Dialogues

Aijun Li[1], Gan Huang[1,2], Zhiqiang Li[3]
[1]Institute of Linguistics of CASS and Graduate School of CASSU, China
E-mail: liaj@cass.org.cn
[2]School of Chinese Culture and Communication, Beijing International Studies University, China
E-mail: huanggan666@yeah.net
[3]Department of Modern and Classical Languages, University of San Francisco, USA
E-mail: zqli@usfca.edu

Abstract— **This study examines the effect of prosodic cues in the disambiguation of five discourse-pragmatic functions of echo questions and the corresponding statements in Chinese spoken dialogues. Data were collected in a "role-play" format to mimic different communicative functions of echo questions in real-life situations. Statistical analyses were performed on both global and local F0 variations associated with intonation patterns in echo questions and corresponding statements. Results showed that boundary tone features alone are not good predictors in distinguishing echo questions and statements; variations in intonation patterns are related to the different discourse-pragmatic functions that echo questions serve; echo questions and statements, as well as different discourse-pragmatic functions of echo questions, can be distinguished on the basis of global variations of prosodic features such as overall F0 slope and average F0, combined with local changes due to boundary tone features; and when information about morpho-syntactic structures and boundary tone features were included in the analysis, the accuracy of discriminant analysis was at 76.5%~94.1% for statements and echo questions, and at 57.6%~83.5% for different discourse-pragmatic functions. The accuracy dropped to 70.9% (2 groups) and 40.9% (6 groups) when morpho-syntactic structural information was not included, indicating that structural and contextual information contributed 30% and 60% respectively.**

I. INTRODUCTION

In spoken dialogue systems, encoding and decoding of interrogative information is an important aspect of intention understanding and generating. In human interactions, echo questions (EQ) can be used when one interlocutor did not hear properly or understand what was said and raised questions by repeating another interlocutor's utterance, in whole or in part, for clarification or confirmation. Echo questions are frequently used in services like ordering meals, booking tickets or making hotel reservations and occasionally used in daily talking like sharing an experience.

Interrogative information is relatively easy to decode in questions with syntactic markers in Chinese, but harder without. Previous research shows that contextual information plays a critical role in the decoding of unmarked echo questions [1]. When taken out of context, 93% of the echo questions were heard with a lower degree of interrogative mood, and about half could not be perceived as questions. In the context, all echo questions could be perceived as being in the interrogative mood, with 70% of them sounding more interrogative than indicative.

Prosody plays an important role in conveying interrogative information as well. In the autosegmental-metrical model of intonation [2-7], question intonation in American English is most frequently marked by a high boundary tone (H%), using a rising final F0 contour transcribed as L*L-H% in the ToBI system [8]. In a lexical tone language like Chinese, boundary tone effectively elevates the pitch register of the tonal contour associated with the pre-boundary stressed syllable in question intonation when the question marker such as the sentence-final particle "ma" is not used [9,10]. There are also other global and local F0 variations in different intonation patterns in Chinese [11].

Prosody has been shown to be provide information related to emotion, attitude and speaker-listener interactions at the discourse level. Roth et al. [12] indicate that prosody in teacher's utterances may result in students' cooperative or uncooperative behaviors, hence influencing the classroom dynamics. Simple words like "okay" and "uh-huh" could be employed to convey different meanings with disparate prosodies in a discourse [13]. In a perception study, Gravano et al. [14] found that contextual cues are stronger predictors of discourse-pragmatic functions than acoustic features in their study of "okay" in English, but word-final intonation seems to play a significant role. Therefore, it is not surprising that more attention has been paid to prosody, particularly the role of intonation, in conversation analysis [15-18]. For example, Ward and Tsukahara [19] identified a region of low pitch could be a good predictor of subsequent backchannel feedback in English and Japanese. What has been discovered in analyses of spontaneous dialogues shows that a response often does not occur the way as predicted by the logic of traditional grammar, but potentially relates to the prosody in which the previous conversation turn is uttered. Our study of

This research is supported by the Key NSSFC Granting (No. 15ZDB103), the National Key R&D Program of China (No. 2017YFE0111900), the Research Foundation of Beijing Municipal Science & Technology Commission (No. Z181100008918002).

echo questions aims to provide further evidence for the role of prosody in naturally occurring conversations by exploring the role of prosodic cues in conveying different discourse-pragmatic functions in spoken dialogues in Chinese.

Echo questions in Chinese can be constructed in various ways, for example, by adding specific markers, repeating the previous turn in whole or in part, or manipulating the prosody [9, 20]. Based on the way they are constructed and the discourse-pragmatic functions that they express in question-response sequences in spoken dialogues [20-26], a functional classification system and annotation scheme was proposed for echo questions [27]. In this pilot study, we examine the effect of prosodic cues in the disambiguation of five discourse-pragmatic functions of echo questions without syntactic markers and the corresponding statements in Chinese spoken dialogues. The goal of the study is to understand how interlocutors encode echo questions to express different communicative functions via prosodic means in Chinese spoken dialogues.

II. Structural Properties and Functional Classification of Echo Questions

Following the framework of interactional linguistics [28], Huang et al. [27] developed a functional classification system and annotation scheme for echo questions. Echo questions occur in a conversational sequence in which an utterance in the form of a statement or question is followed by an echo question and the response, forming a chain of information flow, in which the echo question does not simply perform the task of "asking". Its function in the sequence is largely dependent on the response. In the example below, A2 repeats B1 and is uttered as an echo question, followed by an indirect response (B2), which not only confirms the answer, but also provides further details to the question in A2. Accordingly, the discourse-pragmatic function of A2 here is "Request for Details", labelled as "rdt".

A1:刚刚您点的什么 (What did you just order?)

B1:酸菜鸡丝 (Sliced chicken with pickled cabbage)

A2:酸菜鸡丝(rdt) (Sliced chicken with pickled cabbage?)

B2:少油少盐 (Less oil and less salt)

A3:噢 好 (Oh, OK)

Statistical analysis was performed on dialogues from a discourse speech corpus of six hours of recordings, created in Chinese Academy of Social Sciences (CASS) [29]. It was found that echo questions could appear as various interrogative sentences, expressing the intention of requesting confirmation and explication. Confirmation requests can be further divided into requests for affirmation, repetition, and supplement. Echo questions can also serve the functions of backchanneling or comprehension check. Explication requests can be requests for details and further explanation. In conversations, echo questions most often express the intention of requesting affirmation, which is always in the form of a yes-no question and tag question.

The structural and functions classifications of echo questions are given in Tables 1 and 2 respectively.

Table 1 Structural classification and tags.

Structures		Tag
Yes-no question (*e.g. ... ma0?*)		qy
Tag question	Yes-no tag (*e.g. ..., shi4ma0?*)	qt~tqy
	A-not-A tag (*e.g. ...,shi4bu4shi4?)*	qt~tqpn
A-not-A question		qpn
Wh-question		qw
Alternative question: *e.g.: ...hai2shi4...?*		qr
Yes-no plus disjunctive question *e.g.: ...ma0? hai2shi4...? (Is it ...? Or ...?)*		qrr
Statement question		dq

Table 2 Functional classification and tags.

Functions	Following Responses	Tag
Request Affirmation	Yes, no, right	raf
Request Repetition	Repeat what echo question aimed at	br
Request Supplement	Subsequent part (for long message only, such as address, number, code)	rsup
Request Details	Details, concrete info about echo question	rdt
Request Explanation	Reason, explanation	rex
Backchannel	No response or silence taken as positive answer	b
Comprehension Check	Continue speaking instead of waiting for a response	bu

III. Prosodic Features of Echo Questions

Echo questions occur frequently in spoken dialogues. Their interpretation can be dependent on context and prosody, as they lack clear syntactic markers in many cases. The main challenge is how speakers encode and listeners interpret variations in discourse-pragmatic functions. Understanding the way speakers and listeners negotiate meaning in spoken dialogues is important for human-machine systems. Acoustic features, including prosodic features, have been shown to play a role in the disambiguation of discourse-pragmatic functions in different languages. We will examine the role that prosodic features play in encoding discourse-pragmatic functions in echo questions in Chinese spoken dialogues.

A. Data

a. Material design and recording

The material used in this experiment was designed based on the spoken dialogues in the CASS discourse corpus. We decided to use read speech instead of analyzing dialogues in the corpus directly for the following reasons. First, recorded speech data in the discourse corpus came from naturally occurring real-life conversations and were not well-controlled for good coverage of tonal combinations and discourse functions. For example, in order to study possible effects of boundary tone on signaling discourse functions, we would have to measure utterance-final syllables in 4 lexical tones, each having its own distinctive F0 shape. Second, echo questions in our study are all short utterances consisting of 3 to 4 syllables. Different morpho-syntactic structures (i.e. 2+1, 1+2 or 2+2) turned out to be relevant in distinguishing echo

questions from corresponding statements. Third, our use of read speech is the first step in an effort to gain sufficient insights into the prosodic cues in disambiguation of discourse-pragmatic functions in Chinese spoken dialogues. Third, the decision is also made on the assumption that listeners (i.e. participants in the experiment) have the ability to identify the intended functions correctly and reproduce the dialogues with their intended meaning.

The material was thus created with the above-mentioned considerations in mind and contained dialogues that were supposed to happen between a waiter/waitress and a customer in a restaurant setting. See Appendix 1 for the list of 6 dialogues. Dialogues 1 to 5 (D1-D5) are illustrative of 5 different discourse-pragmatic functions. For example, the function of the echo question in Dialogue 1 is "Request Affirmation", i.e. EQ1/D1=raf, EQ2/D2=rdt, EQ3/D3=rex, EQ4/D4=br, and EQ5/D5=b. The target word in D6 is a statement (SD) in declarative intonation. As a result, there are 6 functions in total.

The target words (echo questions and statements) are all dish names in Chinese, each consisting of 3 or 4 syllables. The full list is given in Table 3, grouped by the tone in the final syllable, i.e. before the intonation phrase boundary. Note that Mandarin Chinese has 4 lexical tones: T1 (H), T2 (LH), T3 (L) and T4 (HL). T3 is realized as a low tone in non-final position, but a low-rise tone in pre-boundary position. The penultimate syllable is set to be in T1 in all words. Morpho-syntactic structures of these words are provided too.

Table 3 Target words grouped by boundary tones.

Boundary Tone	Target echo questions	Structure
T1	suan1 cai4 ji1 si1 (stir fried chicken slices with pickled cabbage)	2+2
	bai2 qie1 ji1 (steamed chicken)	2+1
	ban4 ji1 si1 (mixed chicken slices)	1+2
T2	hong2 shao1 zhu1 ti2 (braised pork trotter)	2+2
	qiu1 dao1 yu2 (saury)	2+1
	chao3 xia1 ren2 (fried shrimp)	1+2
T3	suan4 xiang1 ji1 liu3 (garlic chicken)	2+2
	hua1 diao1 jiu3 (huadiao rice wine)	2+1
	qiang4 dong1 sun3 (stir fried bamboo shoots)	1+2
T4	huang2 men4 ji1 kuai4 (stewed chicken nuggets)	2+2
	dong1 po1 rou4 (dongpo pork)	2+1
	zheng1 hua1 xie4 (steamed crab)	1+2

Each echo question is inserted into one of the 5 dialogues, D1 to D5. It is also inserted into D6 as a statement. The whole set of recording material includes 72 dialogues (4 tones * 3 phrases for each tone * 6 functions): 4*3*5=60 echo questions and 12 statements.

b. Participants

16 students (9 males and 7 females) were recruited to participate in the experiment. They are from different universities in Beijing with an average age of 21. They all speak Standard Mandarin and reported no hearing problems.

c. Recording procedures

Before the recording, all participants were given the sample dialogues in Appendix 1 and listened to recordings of dialogues of ordering food over the telephone, selected from the CASS discourse corpus. They listened to the telephone recordings to understand the context of situation for each dialogue and the role that they would play in reading the dialogues in the material. They were reminded to speak as naturally and colloquially as possible. Minor changes to the reading material were allowed as long as the speakers read them naturally. Every two speakers were recorded as A and B, and then switched roles in the second recording.

The elicited production was digitally recorded with Cool Edit pro 2.0 at a sampling rate of 44.1 kHZ with a 16-bit resolution in a sound-proof recording booth in the Phonetics and Speech Science Laboratory of the Institute of Languages at Chinese Academy of Social Sciences. The co-authors monitored the whole reading procedure and ensured the naturalness of the conversions produced. In total, 16*72*2=2304 dialogues were recorded with 2304 target sentences obtained.

d. Annotation and data extraction

The targeted sentences were segmentally and prosodically annotated using Praat [30]. F0 was extracted and manually checked for spurious cycles. According to the segmental annotations, duration and HNR (Harmony to Noise Rate) of each segment (syllable initials and finals), F0 values were calculated at 10 points with equal time interval for each syllable final. In this paper, we report findings based on the data from 5 speakers, which include 600 echo questions and 419 declarative sentences.

B. Analysis of Prosodic Features

a. Intonation patterns

We provide a few examples illustrating variations in intonation patterns of echo questions in 5 different discourse-pragmatic functions and the corresponding declarative sentences before echo questions in the dialogues. Fig. 1 and 2 present data for the 4-syllable words in 2+2 structure, ending in T4 and T1. Mean F0 curves corresponding to 6 different functions are plotted, i.e. EQ1 to EQ5 and SD. Similarly, intonation patterns for 3-syllable words in 1+2 structure and ending in T1 and T2 are plotted in Fig. 3 and 4.

Several observations can be made here. Overall pitch

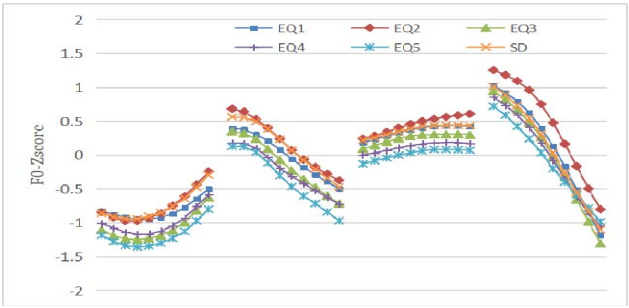

Fig. 1 Intonation patterns for "huang2men4ji1kuai(r)4" (2+2) in 6 functions

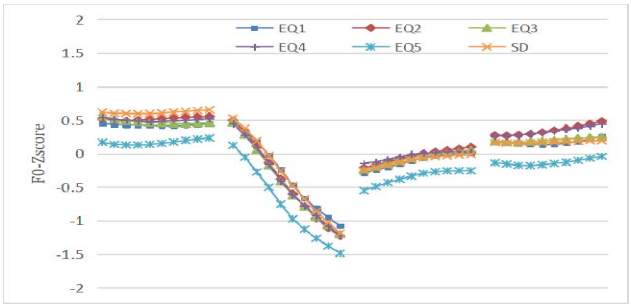

Fig. 2 Intonation patterns for "suan1cai4ji1si1" (2+2) in 6 functions.

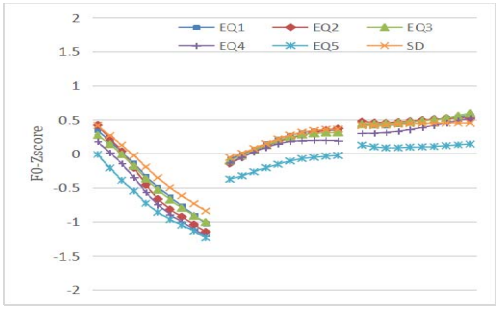

Fig. 3 Intonation patterns for "ban4ji1si1" (1+2) in 6 functions.

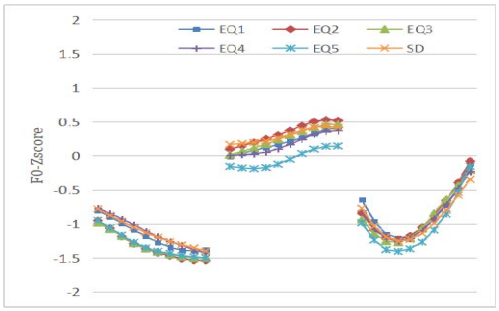

Fig. 4 Intonation patterns for "chao3xia1ren(r)2" (1+2) in 6 functions.

registers corresponding to the declarative intonation and EQ2 requesting for details (rdt) are higher than other EQs. Among them, EQ1 requesting for affirmation (raf) has slightly lower overall pitch register and EQ5 backchannelling has the lowest overall pitch register. Variations in F0 are also noticeable at the left and right boundaries of the intonational phrase between declarative sentences (SD) and echo questions. These observations seem to suggest that information about intonation does not solely reside in the boundary tone [9]. Rather global variations such as overall F0 trendline and average F0, combined with local changes due to boundary tone features contribute more to disambiguation of echo questions in different discourse-pragmatic functions, and differentiation of statements from echo questions in general.

b. Linear discriminant analysis (LDA)

In order to investigate the relationship between various acoustic features and the discourse-pragmatic functions of echo questions, we performed linear discriminant analysis (LDA) on various acoustic features. The features used for the analysis include global features and local features of the target utterances [11], as given in Table 4 below.

Table 4 Global and local acoustic measures: S=whole sentence, F, I, RO2, RO3=final, initial, penultimate and antepenult syllables in the test words (durps=duration per syllable).

Acoustic Measures	Meaning
S_durps, F_dur, I_dur, RO2_dur, RO3_dur	syllable duration
S_LR F_slope, I_slope, RO2_slope, RO3_slope	F0 slope of the syllable or of the linear regression of the sentence
S_F0max F_F0max, I_F0max, RO2_F0max F_F0max	maximum F0
S_F0min, F_F0min, F_F0min, RO2_F0min, RO3_F0min	minimum F0
S_F0mean, F_F0mean, I_F0mean, RO2_F0mean, RO3_F0mean	mean F0
S_F0range F_F0range, I_F0range, RO2_F0range, RO3_F0range	F0 range
F_smHNR, I_smHNR, RO2_smHNR, RO3_smHNR	HNR of syllable initial
F_ymHNR, I_ymHNR, RO2_ymHNR, RO3_ymHNR	HNR of syllable final

(1) We first conducted the LDA between echo questions and declarative sentences based on boundary tones and morpho-syntactic structures, 2+2, 2+1 or 1+2. The results showed that the accuracy of discriminant analysis on the acoustic features used is at 76.5%~94.1% (average 85.0%). The lowest rate of discrimination occurred with the 2+1 structure when the right boundary tone was T1 while the highest rate occurred with the 2+2 structure when the boundary tone was T4. Global F0 variations (F0 slope and average F0) contributed the most to disambiguation. Local F0 features, such as phrase-initial and phrase-final boundary tones, played a significant role in differentiating statements from echo questions. When the boundary tones were T2 and T3, the F0 value and duration of the penultimate syllable also played a role. An example of the F0 variations in a "2+1" structure is given in Fig. 5, where the overall pitch register of declarative sentences is scaled much higher than echo questions.

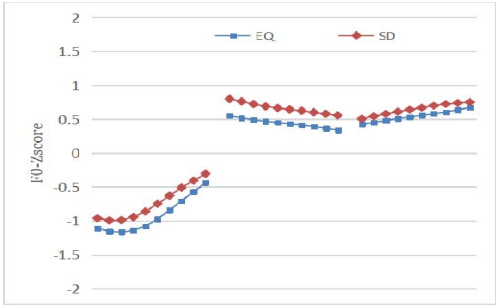

Fig. 5 Mean F0 curves for "bai2qie1ji1" (2+1) in echo questions and declarative sentences.

(2) When we pooled together declarative sentences and echo questions in 5 different functions and performed LDA on all 6 functions based on boundary tones and morpho-syntactic structures, the accuracy of discriminant analysis on the acoustic features used is at 57.6%~83.5% (average 68.6%). The lowest percentage occurred with the 2+1 structure when the boundary tone was T2 and the 1+2 structure when the boundary tone was T4. The highest percentage occurred with

the 2+2 structure when the boundary tones were T2, T3 and T4. Both global F0 variations and local changes due to boundary tone features played a significant role in disambiguation of different functions. When the boundary tones were T2 and T3, F0 variations in the middle syllables seemed to be relevant too.

(3) If boundary tone types and the number of syllables in each utterance were treated as covariates, without referencing morpho-syntactic structural information, LDA on all 6 functions only yielded the accuracy of discriminant analysis at only 40.9%. In this situation, overall F0 slope, initial boundary tone and F0 averages seemed to play a significant role in the discrimination. However, LDA on EQ as one group and SD produced the accuracy of discriminant analysis at 70.9%. Global features such as F0 slope figured most prominently in distinguishing echo questions from statements, with local prosodic cues such as F0 averages in the initial, penultimate and antepenult syllables as significant contributing factors too.

IV. CONCLUSION

In this study, we examined the effect of prosodic cues in the interpretation of echo questions and analyzed prosodic features in statements as well as echo questions to express five different discourse-pragmatic functions. LDA results showed that when morpho-syntactic structures and boundary tone features are considered, prosody would account for the correct discrimination of at least 85% of the statements and echo questions, and at least 68.6% of the statements and the echo questions in five different functions. When information about the morpho-syntactic structure of the target words is not available, prosodic cues alone could only discriminate 40.9% of all six discourse-pragmatic functions including statements, and 70.9% of statements and echo questions. Since people can discriminate functions of all sentences, the significant drop in the rate of the discrimination indicates that morpho-syntactic and contextual information likely contributes nearly 60% of the cues to discrimination of 6 different functions, nearly 30% of the cues to the discrimination of statements and echo questions in general.

Our findings suggested that participants in the spoken dialogues made use of information from a multi-dimensional source such as morpho-syntax, prosody and context of the discourse to encode communicative intent. While not a single factor played the decisive role, the morpho-syntactic structure and the context of the discourse did contribute significantly. The perception experiment reported in our previous work [1] showed that in the identification of echo questions and corresponding statements, the contextual information alone could discriminate 43% of the target utterances, similar to the results of the LDA analysis conducted in the current study. The context played an even bigger role in the interpretation of different discourse-pragmatic functions of echo questions.

In terms of prosodic features, global changes of F0 such as F0 slope was believed to have the most impact, followed by the F0 scaling and duration of the syllable at the boundary. The successive addition boundary [31] tone was also identified in the study and will be discussed in the follow-up reports.

We did not analyze the dialogues in the CASS discourse speech corpus directly. As mentioned in section III, the material used in this experiment was designed based on the spoken dialogues in the CASS corpus. The results produced in this report will provide guidance for the further study to be carried out on the spoken dialogues in the CASS corpus. After designating the functional categories of echo questions in spoken dialogues in the CASS corpus based on machine learning algorithms, a model will be constructed that integrates morpho-syntactic, prosodic and contextual features and information. In particular, three kinds of prosodic features will be utilized: global intonation features (F0 slope and average F0, top and bottom F0), local F0 features (boundary tones and F0 variations of the syllables at the boundaries) and microscopic features (the successive addition boundary tones).

REFERENCES

[1] A. Li, J. Yuan, et. al, "Decoding of echo question in Chinese spoken dialogue," in *Proc. of the International Conference on Phonetics of the Languages in China* (ICPLC-2013).

[2] J. Pierrehumbert, *The Phonology and Phonetics of English Intonation*, Ph.D. Dissertation, MIT, 1980.

[3] D.R. Ladd, *Intonational Phonology*, Cambridge: Cambridge University Press, 1996.

[4] D.R. Ladd, *Intonational Phonology* (2nd ed.), Cambridge: Cambridge University Press, 2008.

[5] S.A. Jun (ed.), *Prosodic Typology: The Phonology of Intonation and Phrasing*, Oxford: Oxford University Press, 2005.

[6] S.A. Jun (ed.), *Prosodic Typology II: The Phonology and Intonation of Phrasing*, Oxford: Oxford University Press, 2014.

[7] C. Gussenhoven, *The Phonology of Tone and Intonation*, Cambridge: Cambridge University Press, 2004.

[8] M. E. Beckman and J. Hirschberg, "*The ToBI annotation conventions*," Ohio State University, 1994.

[9] M. Lin, Z. Li, "Focus and Boundary in Chinese Intonation," in *the 17th International Congress of Phonetic Sciences* (ICPhs 2011), August 17-21, Hong Kong.

[10] M. Lin, *The Experimental Study of Intonation in Mandarin Chinese* (in Chinese), Beijing: Chinese Academy of Social Sciences Press, 2012.

[11] X. Liu, A. Li, & Y. Jia, "How does prosody distinguish Wh-statement from Wh-question? A case study of Standard Chinese," In *Proceedings of Speech Prosody*, pp.1076-1080, 2016

[12] W.M. Roth, K. Tobin, "Solidarity and conflict: aligned and misaligned prosody as a transactional resource in intra- and intercultural communication involving power differences," *Cultural Studies of Science Education,* vol. 5, pp 807-847, 2010.

[13] B. A. Hockey, "Prosody and the role of okay and uh-huh in discourse," *Proceedings of the Eastern States Conference on Linguistics*, pp. 128-136, 1992.

[14] A. Gravano, S. Benus, H. Chávez, J. Hirschberg, & L. Wilcox, "On the role of context and prosody in the interpretation of 'okay'," In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pp. 800-807, 2007.

[15] D. Wilson, T. Wharton, "Relevance and prosody," *Journal of Pragmatics*, vol. 38, pp. 1559-1579, 2006.

[16] T. Wharton, *Pragmatics and Nonverbal Communication*, Cambridge University Press, 2009.
[17] A. Wichmann, "Prosody and pragmatic effects," *Pragmatics of Society*, De Gruyter Mouton, 2011, pp. 181-214.
[18] K. Scott, "Prosody, procedures and pragmatics," In *Semantics and pragmatics: Drawing a Line*, I. Depraetere, and R. Salkie, Eds. Springer International Publishing, 2017, pp. 323-341.
[19] N. Ward and W. Tsukahara, "Prosodic features which cue back-channel responses in English and Japanese," *Journal of Pragmatics*, vol. 23, pp.1177-1207, 2000.
[20] J. Shao, *A Study of Questions in Modern Chinese*, China Commerce and Trade Press, 2014.
[21] T. Stivers and M. Hayashi, "Transformative answers: One way to resist a question's constraints", *Language in Society*, vol. 39, no. 1, pp. 1-25, 2010.
[22] J. Steensig and P. Drew, "Introduction: questioning and affiliation/disaffiliation in interaction," *Discourse Studies*, vol. 10, no. 1, pp. 5-15, 2008.
[23] T. Stivers, "An overview of the question-response system in American English conversation," *Journal of Pragmatics*, vol. 42, no. 10, pp. 2772-2781, 2010.
[24] T. Stivers and N. Enfield, "A coding scheme for question-response sequences in conversation," *Journal of Pragmatics*, vol. 42, no. 10, pp. 2620-2626, 2010.
[25] T. Stivers and F. Rossano, "Mobilizing response," *Research on Language and Social Interaction*, vol. 43, no. 1, pp. 3-31, 2010.
[26] N. Enfield, T. Stivers, S.C. Levinson, "Question-response sequences in conversation across ten languages: An introduction," *Journal of Pragmatics*, vol. 42, pp. 2615-2619, 2010.
[27] G. Huang, L. Zhu and A. Li, "Syntactic structure and communicative function of echo questions in Chinese dialogues," in *Proceedings of ISCSLP2018*, 2018.
[28] M. Selting, & E. Couper-Kuhlen (eds.), *Studies in Interactional Linguistics*, Benjamins, 2001.
[29] J. Yuan, and A. Li, "A linguistic annotation scheme of Chinese discourse structures and study of prosodic interactions," in *Proceedings of ISCSLP2016*, 2016.
[30] P. Boersma and D. Weenink, "Praat: Doing phonetics by computer," http://www.praat.org, 2001.
[31] Y.R. Chao, "Tone and Intonation in Chinese," *Bulletin of the Institute of History and Philology* 4, 121-134,1933.

Appendix 1: Recording material of 6 dialogues, where "蒸花蟹（Steamed Crab)" is selected as target word

D1	B1:喂 你好 我要订两个菜 wei4 ni3 hao3 wo3 yao4 ding4 liang3 ge4 cai4 (Hello, I'd like to order two dishes) A1:您说 nin2 shuo1 (Sure) B2:要蒸花蟹吧 yao4 zheng1 hua1 xie4 ba0 I'd like a Steamed Crab. A2:**蒸花蟹(raf)** **zheng1hua1xie4(Steamed Crab?)** B3:对 dui4 (Yes) A3:**还有** hai2 you3 (And?)
D2	A1:刚刚您点的什么 gang1 gang1 nin2 dian3 de0 shen2 me0 (What have you just ordered) B1:蒸花蟹 zheng1 hua1 xie4 (Steamed Crab) A2:**蒸花蟹(rdt) zheng1 hua1 xie4 (Steamed Crab?)** B2:多放孜然 duo1 fang4 zi1 ran2 (Put more cumin) A3:噢 好 o1 hao3 (Oh, OK)
D3	B1：你好 我们想换个菜 ni3 hao3 wo3 men0 xiang3 huan4 ge4 cai4 (Hello, we want to change a dish) A1：换掉什么 huan4 diao4 shen2 me0 (What do you want to change?) B2：蒸花蟹 zheng1 hua1 xie4 (Steamed Crab) A2：**蒸花蟹(rex)** **zheng1 hua1 xie4 (Steamed Crab?)** B3：因为有人不吃海鲜了 yin1 wei4 you3 ren0 bu4 chi1 hai3 xian1 le0 (Because someone don't want to eat seafood) A3：噢 那换成什么呢 o1 na4 huan4 cheng2 shen2 me0 ne0 (Oh, what do you want to change to?)
D4	B1:点个菜 dian3 ge4 cai4（I'd like to order dishes now) A1:您点啥 nin2 dian3 sha2 (What do you want?) B2:我点个蒸花蟹 wo3 dian3 ge4 zheng1 hua1 xie4 (I want a Steamed Crab) A2:**蒸花蟹(br)** **zheng1 hua1 xie4 (Steamed Crab)** B3:蒸花蟹zheng1 hua1 xie4 (Steamed Crab?) A4:好的 hao3 de0 (Sure, OK)
D5	B1:我点个麻婆豆腐 wo3 dian3 ge4 ma2 po2 dou4 fu0 (I want a Mapo Tofu) A1:麻婆豆腐(b) ma2 po2 dou4 fu0 (Mapo Tofu) (silence, longer break) B2:蒸花蟹zheng1 hua1 xie4 (Steamed Crab) A2:**蒸花蟹 (b)** **zheng1 hua1 xie4 (Steamed Crab?)** (silence, longer break) A3:**还有** hai2 you3 (And?)
D6	B1:他点了什么荤**菜** ta1 dian3 le0 shen2 me0 hun1 cai4 (What meat dish did he order?) A1:**蒸花蟹(SD)** **zheng1 hua1 xie4 (Steamed Crab)** B2:就一个啊 jiu4 yi2 ge4 a0 (Only one?) A2:嗯 就一个 en4 jiu4 yi2 ge4 (Yes, just one)