MASTER IN COGNITIVE SCIENCE AND LANGUAGE

MASTER THESIS

July, 2024

Acoustic study of non-interrogative usages of the interrogative pronoun in Mandarin

by Jiaqi Wang

Under the supervision of:

Wendy Elvira-García

and

Mireia Farrús

Abstract: In Mandarin Chinese, the interrogative pronoun “shen2me0” not only can be used to express interrogation, but also has multiple non-interrogative usages. In sentences with same syntactic structure, by applying different intonations, this wh-word can convey various meanings. However, due to the lack of grammatical markers, it could be a complex problem for Automatic Speech Recognition (ASR). Therefore, this study chose to investigate the acoustic features of these usages of “shen2me0”. Through two experiments, this present compared the acoustic features including pitch contours, sentence stress, duration, pitch range, boundary tone of sentence and the wh-word of the interrogative, empty reference, rhetorical, and referential substitution usages in same sentences under different contexts. Also, the use of modal particles at the end of the sentences was considered. The results showed that interrogative usage had moderate pitch fluctuation and the use of modal particle would influence how people pronounced the wh-word. In empty reference usage, “shen2me0” had a neutral nature as a placeholder rather than a focus, and thus the sentence showed a flat pitch curve. Rhetorical usage had dynamic pitch changes, especially at the word “hai2”, to express strong emotion. Referential substitution usage had a flatter pitch curve at the beginning, which rose higher at the end, with prolonged pronunciation of “me0”. Moreover, this study also discussed about meaning for ASR and the improvements in further study.

Keywords: Mandarin interrogative pronoun, non-interrogative usages, acoustic features, ASR

Acknowledgement

First, I would like to express my sincere gratitude to my supervisors, Wendy and Mireia, for your patience and guidance.

I am very grateful for my parents' unconditional love and support. You encourage me all the time, make me feel confident, and get me out of difficulties and anxiety no matter what happens.

Many thanks to my friends and colleagues for always being supportive and understanding. You always care about me and believe in me.

I also want to thank all the people who participated in the experiments and made this study possible.

Thanks to Fan Zhendong. Your spirit motivates me to move on. I feel happy to have witnessed the moment you finally won the gold medal in the Olympic Men's single and had the grand slam.

CONTENTS

1 Introduction

The interrogative pronoun is usually a pronoun used to express a question, and the object of it in an interrogative sentence is to be revealed or to be known. The researchers hold different opinions about the usage of interrogative pronouns in Mandarin. Some proposed that the basic usage of wh-words is questioning, and the non-interrogative usages are derivation. Some researchers also suggested that interrogation is a function of the whole sentence. The wh-words do not have interrogative markers per se, and there is equality between their interrogative usages and the non-interrogative ones, which have different functions depending on the contexts. In this study, the wh-word "shen2me0" is chosen to investigate its diverse usages, the phonetical form of which could be ambiguous.

1.1 Interrogative usage of "shen2me0"

When used for interrogation, syntactically, the interrogative usage of "shen2me0" mainly could be seen in specific reference questions, and a small part of it is used in declarative sentences of questioning nature. Sometimes modal particles expressing interrogation appear at the end of the sentence, such as "ne0". The position of "shen2 me0" in a sentence is not fixed. It can be used as the subject, the object, the determiner and other elements.

When used alone, "shen2 me0" can represent an object, but also can stand for action and behavior, or nature and state. When referring to place, it can be used as

“shen2me0 di4fang1” (which place). It can be used to ask “why” or “which reason” by forming a verb-object phrase as “wei4 shen2me0” before verbs, adjectives, or at the beginning of the sentence.

Semantically, “shen2me0” is mainly used to indicate doubt, and if there is an answer, it is the answer to the question to which the pronoun refers to. It is either hoping that someone else will answer, or by way of rhetoric question, raising a question and answering it oneself.

1.2 Non-interrogative usages of “shen2me0”

When “shen2me0” is used to express non-interrogative meanings, there are different views on the classification of its non-interrogative usages. This study took the previous studies into account, and generally categorized the non-interrogative usages into with-reference and no-reference, with eight kinds of usages as in Figure 1.

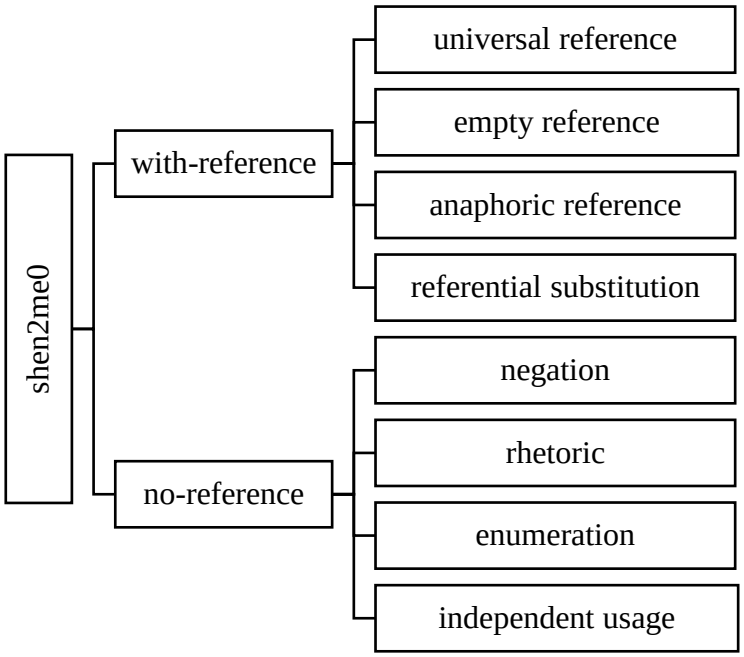

Figure 1 Classification of non-interrogative usages of "shen2me0"

Among these non-interrogative usages, four of them were selected to investigate their acoustic features, which could be used sharing the same grammatical structure, which could be problematic for Automatic Speech Recognition (ASR):

1.2.1 Empty reference

Empty reference indicates objects that is uncertain, i.e., something does exist, but is shown to be unknown or unspoken. Semantically, this function can be equivalent to "some" and "certain". The difference between empty reference and universal reference is that the former focuses more on individuals and small quantities, while the latter is groups and larger quantities.

Formally, "shen2me0" of empty reference is mostly used as object, determiner and complement in the sentences. According to Lv (1985), given that "shen2me0" has an

indefinite semantic feature of empty reference, it tends to appear in non-assertive context sentences such as negative sentences, hypothetical complex sentences, yes/no questions, affirmative-negative questions, and selective questions, or sentences with words that indicate speculation such as “xiang3bi4” (presumably), “yi1ding4” (must), “kong3pa4” (supposedly), “ye3xu3” (maybe, perhaps).

a. As illustrated in example (1), in broad-focus informative sentence, “shen2me0” can be used as a determiner modifying a noun to indicate an indefinite empty reference.

(1) 她 只要 讨得 一点 什么 便
Ta1 zhi3yao4 tao3de2 yi1dian3 shen2me0 bian4
都 献给 祖母 吃 自己 情愿
dou1 xian4gei3 zu2mu3 chi1 zi4ji3 qing2yuan4
饿肚子
e4du4zi2

Whenever she got something to eat, she gave to her grandma, even to starve herself.

b. “Shen2me0” in speculative declarative sentences also usually is of empty reference, in which usually contain words such as “fang3fu2”, “hao3xiang4”, “si4de0”, “si4hu1” (it seems that, it is like), as in example (2).

(2) 但是 他们 总 感到 没有
Dan4shi4 ta1men2 zong3 gan3dao4 mei2you3
得到 真正 的 满足 生活
de2dao4 zhen1zheng4 de4 man3zu2 sheng1huo2
中 好像 还 缺点 什么
zhong1 hao3xiang4 hai2 que1dian3 shen2me0

However, they always feel that they haven’t been truly fulfilled, and that there seems to be missing something in their lives.

c. In speculative sentences expressing empty reference, as example (3), "shen2me0" also appears with verbs expressing mental activities, such as "xi1wang4" (hope), "yi3wei2" (suppose), "yuan4yi4" (be willing, would like), "da3suan4" (prepare to), "xiang3" (want), "cai1" (guess), etc.

(3) 他们 并不 希望 放下 什么
Ta1men2 bing4bu4 xi1wang4 fang4xia4 shen2me0
They are not hoping to let go of anything.

d. "Shen2me0" used as empty reference with speculative adverbs such as "xiang3bi4" (presumably), "yi1ding4" (must), "kong3pa4" (supposedly), "ye3xu3" (maybe, perhaps), as shown in example (4).

(4) 她 以为 那是 什么 美味
Ta1 yi3wei2 na4shi4 shen2me0 mei3wei4
抓了 一块 放在 嘴里
zhua1le0 yi1kuai4 fang4zai4 zui3li3
Thinking it was some kind of delicacy, she grabbed a piece and put in her mouth.

Due to the presence of these speculative words, the whole sentence has a distinctly speculative tone, and thus, constrained by this kind of context, the questioning tone of "shen2me0" is weakened, and transformed into a tone of uncertainty and disbelief. Phonetically, when indicating empty reference, "shen2me0" do not need to be stressed.

1.2.2 Rhetoric

Although formally it indicates questioning, the speaker has already had a clear idea in mind, using affirmative expression to indicate negation and vice versa. “Shen2me0” in such sentence does not carry interrogative information, but to use to strengthen the tone, to reinforce the negative or affirmative content. “Shen2me0” normally need to be stressed to express an extra pragmatic meaning of interrogation, surprise, reprimand or emphasis. This meaning is produced with the strong emotional tone such as demonstrated in examples (5) to (7).

a. X + shen2me0

(5)

他们	又	不是	故意	的	你
Ta1men2	you4	bu4shi4	gu4yi4	de4	ni3
有	什么	必要	这样	对待	他们
you3	shen2me0	bi4yao4	zhe4yang4	dui4dai4	ta1men2
吗					
ma0					

They didn’t do it on purpose, so what’s the point of treating them like that?

b. X + shen2me0 + Y

(6)

本来	就	和	你	没	关系
Ben3lai2	jiu4	he2	ni3	mei2	guan1xi4
你	在	这里	认	什么	真
ni3	zai4	zhe4li3	ren4	shen2me0	zhen1
没	必要				
mei2	bi4yao4				

It had nothing to do with you in the first place, so why are you here being serious? There’s no need to do that.

c. You3 shen2me0 hao3 X de4

(7)

事情	就是	这样	发生	了	我
Shi4qing2	jiu4shi4	zhe4yang4	fa1sheng1	le0	wo3
有	什么	好	说	的	
you3	shen2me0	hao4	shuo1	de0	

It just happened. What do I have to say?

1.2.3 Referential substitution

Shao (1989) suggests that "shen2me0" can act as referential substitution, i.e., it is borrowed to replace a certain object temporarily. Formally speaking, "shen2me0" can replace a syllable, a word, a phrase, or even a sentence or a paragraph. He also divided the usage of this function into three types in terms of meaning:

a. To replace unknow information. Due to the lack of knowledge of the speaker, or be out of the mind for a moment, the current discourse is unable to be proceed smoothly, and needs to temporarily substitute the obstacle in the communication using "shen2me0" to make it continue as in example (8).

(8)

你	刚才	所说	的	就是
Ni3	gang1cai2	suo3shuo1	de4	jiu4shi4
说	我	和	华	华
shuo1	wo3	he2	hua2	hua2
什么	华泰	房地产	公司	这个
shen2me0	hua2tai4	fang2di4chan3	gong1si1	zhe4ge4
买卖				
mai3mai4				

What you just said is the business between me and the real estate company Hua, something, Huatai.

b. The speaker thinks that due to inconvenient or taboo, it is not possible or convenient to express directly, and therefore uses “shen2me0” to convey a kind of vague message. This kind of message may be vague seemingly. However, because of the shared knowledge background of both sides in the conversation, it does not affect the communication effect in practice. It is shown in example (9).

(9)

你	知道	那个	人	吗	他们俩
Ni3	zhi1dao4	na4ge4	ren2	me0	ta1men4lia3
之间	是不是	有	什么		
zhi1jian1	shi4bu2shi4	you3	shen2me0		

Do you know that people? Is there “something” between the two of them?

c. To replace unimportant information in the dialogue. For secondary information in a conversation, if the speaker does not feel the need to fully say it, it can be replaced using “shen2me0” to make the discourse more concise, for example, as shown in (10).

(10)

他们	谈起	小学	的	同学
Ta1men2	tan2qi3	xiao3xue2	de4	tong2xue2
某某	现在	在	什么	城市
mou3mou3	xian4zai4	zai4	shenm2me0	cheng2shi4
在	搞	什么	工作	
zai4	gao3	shen2me0	gong1zuo4	

They talked about their primary school classmates, someone is in some city, doing some job.

1.3 Previous acoustic research on wh-words

From the phonetic point of view, previous studies performed acoustic analyses of wh-words, specially, on the sentence level. The researchers chose to focus on features like the stress and focus, intonation, the boundary tone, etc. and they mainly paid attention to the basic usage of wh-words, i.e., the interrogative usage.

1.3.1 Wh-word, emphasis and focus

Emphasis, often referred to in Chinese literature as stress, is a highlighting phenomenon when speaking (Lin & Wang, 2013). Speakers usually choose stressing to show emphasizing in order to gain attention of hearers. And focus is normally viewed as a way to reflect new information (Halliday, 1967). The sentence elements that have focus are normally notional words (Zhao, Yang & Lv, 2013). It is generally believed that there are close relationships between stress and focus. Although most focus would be stressed, their degree of stressing vary widely. Only around half of the focuses are strongly stressed, but the possibility of the focus at the end of the sentence gaining stress could reach 88% (Zhao, Yang, Yang, et al., 2012). Qi (2012) explained their relations from semantic perspective. He suggested that the speaker offer more energy to the focus while speaking, which is an encoding process from focus to stress, and the listeners would pay their attention on the words have higher pitch and longer duration while understanding the information, which is a decoding process from stress to the semantics. In interrogative sentences, some researchers, for example, Lin (1985) and Tang & Shi

(2009) believed that the interrogative pronoun and focus markers could have similarities in their function of expression. The interrogative pronoun marks the unknow information of a sentence, while the focus marks the most important new information in it. However, researchers such as Lambrecht & Michaelis (1998), Hedberg & Sosa (2002) held the opinions that in wh-questions the wh-word gains its focus by structure (form and location) rather than intonation, and therefore the sentence stress would fall on other sentence element instead of the wh-word. While Haan-van Ditzhuysen (2001) and Chen (2006) proved that in Dutch, wh-words are not only the focus of sentence, but also where the sentence stress is.

As for sentences with wh-words of non-interrogative usages, in these cases, the querying function of wh-word decreases, i.e., instead of conveying questioning, it is only for expressing the narration of opinion of the speaker. Therefore, in these sentences, the wh-words lose their stress and focus. Zhao (1979) pointed out that wh-words of empty reference normally should be pronounced using neutral tone, while those of universal reference not. Lv (1982) also indicated that if there is modal particle “ma0” at the end of wh-question sentences, the “ma0” would move the questioning point, which makes the wh-question into yes/no question, and wh-words into indefinite referents. Therefore, the interrogative pronoun would not be stressed.

1.3.2 Declarative tone vs. interrogative tone

In most languages, the declarative sentences have falling or low tone. While the interrogation requires to consider the types of question: Yes/no question is mainly with rising tone, and wh-question is normally expressed in falling tone. Lee (2005) and for English, Cruttenden (1997) suggested that there are differences between the falling tone of declaration and wh-question: the former is a gradually falling process, while the latter shows a high or rising tone, and followed by a sharp fall, until reaching the bottom of the range. The same is true for many Romance languages where statement intonation and wh-question intonation can be quite distinct (Frota & Prieto, 2015). As for Chinese, De Francis (1963), Yuan, Shih & Kochanski (2002), Wu, Tao & Lu (2006) suggested that the difference between declaration and interrogation lies in the overall pitch contour of the former is higher than the latter. Shen (1990), Shiamizadeh, Caspers & Schiller (2015) suggested that their differences appear at the front part of sentences: pitch contour of words before the wh-word of wh-question is higher than the corresponding declarative sentence. Wang (2009) and Lin (2006, 2012) believed that only the boundary tone could play the role of distinguishing the declarative and interrogative tone. Liu & Xu (2005) thought that their discrepancy starts from the focus in sentence: before the focus, the differences in pitch are not significant; after the focus, the pitch curve of interrogative tone is higher than the corresponding declarative tone.

1.4 Automatic Speech Recognition of wh-words

In ASR of sentences with interrogative words, the first feature to be detected is the lexical-syntactic feature, including the words appearing in specific sentence structures, the order of words and the location of words in sentences. Besides, the information conveyed by the speaker could also be recognized from the context.

However, the ASR of sentences with wh-words in Mandarin is facing problems due to these reasons: From semantic and pragmatic level, there exists non-interrogative usages of wh-words that express different meanings. From syntactic level, some of the non-interrogative usages do not have specific or unique sentence structure that can be easily detected, and the positions of wh-words in the sentence could be various. Therefore, it is necessary to seek for other features, such as intonation, to improve the ASR of wh-words in sentences expressing different tones. And it requires finding acoustic features that are essential in the recognition. Jiang & Cai (2003) employed Fisher discriminant analysis to examine average frequency, slope of fundamental frequency (f0) and energy curve after linear fitting, duration and other acoustic features to distinguish the two tones, and discovered that the combination of f0, duration and energy was the most effective way to improve the accuracy of classification. Liu, Surendran & Xu (2006) considered the influences of f0, intonation and focus on the recognition of tones. Yuan & Jurasfsky (2005) chose to extract pitch, spectrum characteristics ad duration for the investigation.

In this study, in order solve this problem, we especially focused on the different usages of wh-words in the same grammatical structure. And through the analysis of acoustic features of these sentences and the wh-words, we look for improvement of ASR in Mandarin Chinese.

In the following sections, we first show the methodology adopted in this study, including the research method, the experiment design, the data collection and processing, and the data analysis. Furthermore, the results of the study are demonstrated. Then we discussed the results, and the limitation of this present. Finally, the conclusions of the study are drawn.

2 Methodology

2.1 Research method

This study employs a mixed-methods approach, integrating both quantitative and qualitative analyses to comprehensively investigate the acoustic features of non-interrogative usages of wh-words in Mandarin. Two experiments are designed to reflect that when using the same sentence pattern, in different contexts, due to the difference between the interrogative and non-interrogative pragmatic functions as well as the use of modal particles, the meanings of expression are diverse. And after the collection of speech data, the local and global acoustic features of Mandarin speakers' voice are analyzed.

2.2 Data collection

2.2.1 Materials

This study contains two experiments. Based on the factors that may affect the results of Mandarin wh-words' ASR, a set of sentences with the same syntactic structure, but with different functions of the wh-word "shen2me0" in different contexts are designed. Experiment Part I aims to investigate the acoustic performance of the wh-word "shen2me0" in its (1) interrogative usage and three non-interrogative usages, which are (2) empty reference, (3) rhetoric and (4) referential substitution. In experiment Part II, modal particles "me0", "ba0", and "a0" are added to the end of sentences in the first

three usages, in order to study the effect of these modal particles on the acoustic features and ASR of wh-word “shen2me0” in Mandarin.

2.2.2 Procedure

The collection was conducted via WeChat, where subjects first filled in their basic background information, read and fully understood the contexts, then played the role in the dialogues, and sent their speech through WeChat voice message. And in Experiment II, they were also required to first select and add the appropriate modal particle before speaking. The subjects were asked to complete the recording in silent environment and the use of earphones are preferred.

2.2.3 Participants

Thirty native Mandarin speakers were recruited and selected by the researcher, making sure the participants in this study with a balanced gender and age range of 18-30 years old, as well as excluding subjects with symptoms affecting their voice.

2.2.4 Ethical considerations

This study adhered to strict ethical guidelines to protect the privacy of participants and data integrity. Informed consent was obtained from all participants clearly outlining the purpose of the study and their rights.

2.3 Data analysis

2.3.1 Transcription

All sound files were transcribed after collection and WAV file conversion. The transcribing was performed in Praat. First, we adopted Montreal Forced Aligner for automated annotation, which is a tool based on Kaldi ASR toolkit, for time speech-text aligning. Four tiers of textgrids were annotated in Praat: 1. Sentence (sentence), 2. Phone (phones), 3. Word (words), and 4. Character (CHs), as shown in Figure 2.

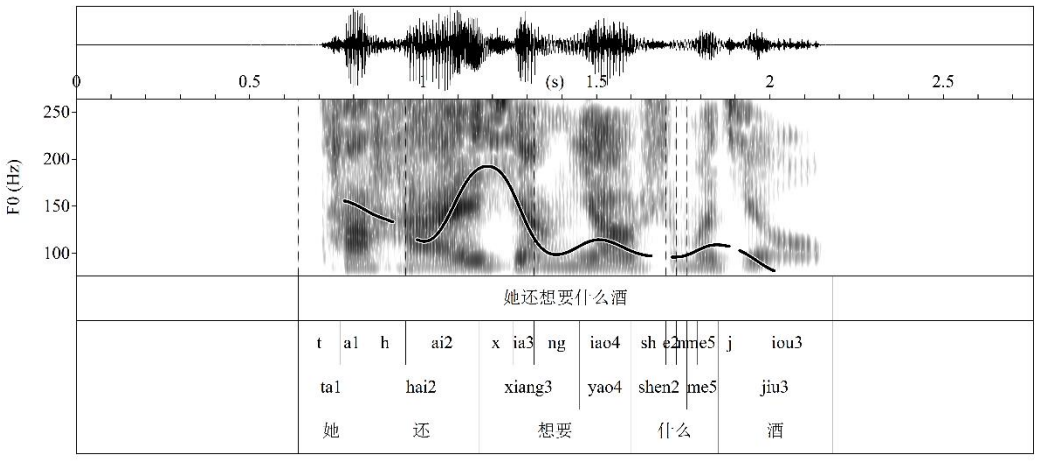

Figure 2 Annotation in Praat (*in annotation system, the tone 0 was represented using tone 5)

Then, every textgrid file was manually adjusted for better accuracy.

2.3.2 Feature extraction

The feature extraction was conducted using Praat. The target acoustic features are divided into local features and global features. This study collected the original acoustic data of these features using Praat script, which extracts the duration and pitch value of 10 averagely taken points, and finally stores the data in a text file. The "name" row data is determined by the tier and its corresponding intervals. For example, when the "words" tier is selected, the results are as in Figure 3.

fileName	name	duration	Pitch1	Pitch2	Pitch3	Pitch4	Pitch5	Pitch6	Pitch7	Pitch8	Pitch9	Pitch10
01 1 1 1	1.TextGrid	ta1	0.105	114	114	114	114	113	112	110	109			
01 1 1 1	1.TextGrid	hai2	0.141	109	108	108	108	109	110	111	113	115	117	
01 1 1 1	1.TextGrid	xiang3	0.210	117	116	113	109	102	93	83	78	78	83	
01 1 1 1	1.TextGrid	yao4	0.164	83	88	94	98	99	98	94	90	86	84	
01 1 1 1	1.TextGrid	shen2	0.166	84	83	82	82	82	81	81	82	84	88	
01 1 1 1	1.TextGrid	me5	0.113	88	92	95	99	102	104	105	105	106	107	

Figure 3 Example of extracted data in text file

The target acoustic features investigated in this study are:

A. Local features

a. Duration of "shen2me0"

It is directly extracted using Praat script.

b. Pitch of "shen2me0" (range, highest and lowest point)

First extracted ten points' pitch values of each "shen2me0", and then selected the highest and the lowest values, and used the highest minus the lowest to get the range.

c. Boundary tone of “shen2me0”

The data were normalized using Python by calculating the z-score of the last point’s pitch value of each “shen2me0” to see how many standard deviations this value was from the mean ten-points pitch of “shen2me0”.

B. Global features

a. Sentence stress

For the sentence stress of Chinese, researchers hold different opinions on its determination. Zhao (1968) believed that it requires first the expanding of pitch range and duration of syllables, and then increasing the airflow. Lin, Yan & Sun (1984) suggested that the most important feature of Chinese stress is the increasing of duration, rather than the function of intensity. Shen (1994) indicated that when recognizing the stress, the role of pitch is important, while the function of duration is not obvious. There are studies on the auto-annotation of stress of Mandarin, but not widely applied (Ni, Liu & Xu, 2012). In experimental studies, some researchers like Liu (2016) chose to annotate the stress manually. Taking into the previous studies of Chinese sentence stress into consideration, in this present study the sentence stress was determined by pitch and duration, using a self-designed calculating method:

(1) Extraction of duration and ten-points’ pitch value of each word

(2) Calculation

$$mean\ pitch\ of\ sentence = \frac{sum\ of\ all\ point's\ pitch\ values}{sum\ of\ pitch\ points\ (the\ word\ number \times 10)}$$

$$mean\ pitch\ of\ word = \frac{sum\ of\ ten\ points' pitch\ value\ of\ the\ word}{10}$$

$$mean\ duration\ of\ word = \frac{sum\ of\ duration\ values\ of\ each\ word}{the\ word\ number}$$

(3) Comparison of the mean pitch of word with mean pitch of sentence, and each word's duration with the mean duration of word. It is shown in Figure 4.

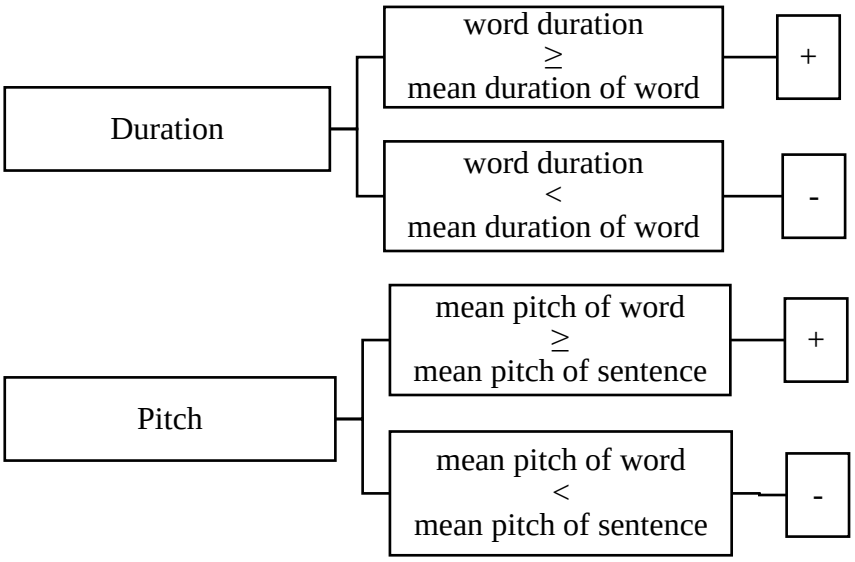

Figure 4 Comparison method of duration and pitch

(4) Determination the stress levels. It is shown in Figure 5.

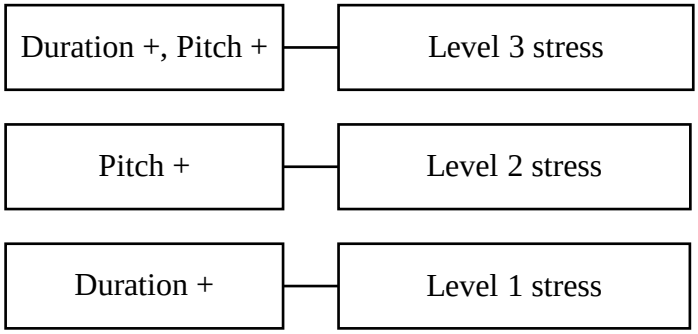

Figure 5 Determination method of stress levels

(5) Calculation of the total number of level 3, level 2 and level1 stress, and the final stress score:

$$Final\ stress\ score = level\ 3\ number \times 50\% + level\ 2\ number \times 30\% + level\ 1\ number \times 20\%$$

b. Sentence duration

It was directly extracted using Praat script.

c. Sentence pitch range

Highest and lowest value are obtained after comparing all words' ten-points-pitch value, and the range value is the highest value minus the lowest value.

d. Sentence boundary tone

The data were normalized using Python by calculating the z-score of the last point's pitch value of the sentence to see how many standard deviations this value was from the mean sentence pitch.

2.3.3 Pitch contour visualization

In order to visualize the pitch contour of different context, first the "word" tier data were stored and converted into excel files. And normalization was performed to the pitch data using Python, which aimed to eliminate the voice variations across participants and recording conditions, making it comparable across different recordings. For this study, the pitch data were normalized by calculating z-score to show the distance of each point's pitch value to the mean pitch of each sentence.

For model fitting, this study adopted Generalized Additive Mixed Models (GAMMs) using R. GAMMs provided a robust statistical framework to analyze the complex patterns in the pitch data. It extended the traditional linear mixed-effects models by allowing the inclusion of smooth functions for predictors, thus providing more flexibility to capture non-linear relations between the tones and different context. For this study, the pitch, expressed as z-score, was modeled as a function of the pitch point thin plate regression spline. The choice of spline basis and the number of knots were optimized to capture the underlying pitch contour without overfitting.

Finally, each word's pitch contour was shown in plots. They were aligned by their order in the sentences, which were also classified according to Parts (two parts of the experiment) Scenarios (four usages of wh-word in part I and three usages with modal particles in part II) and Tones (nouns with four Mandarin tones).

2.3.4 Statistical analysis

In order to further investigate the target acoustic features of different usages of wh-word “shen2me0”, a series of statistical analysis were employed using SPSS to analyze the sentence duration, sentence pitch range, sentence boundary tone, wh-word duration, wh-word pitch range and wh-word boundary tone. But for different comparison groups, two kinds of tests were used: mixed model analysis and independent samples t-tests.

To assess the differences in acoustic features across different wh-word usages scenarios, the mixed linear model analysis was applied. This statistical approach is particularly suitable for data with hierarchical structure. The primary focus of the analysis was the fixed effect of “scenario”, which represents the four distinct usages of “shen2me0”. It was designed as the fixed effects in the tests. The experiment part and tones of nouns are included in the random effects. The significance of the fixed effects was tested using F-tests. The model fit was evaluated using criteria such as the Akaike Information Criterion (AIC) and The Bayesian Information Criterion (BIC). Post-hoc pairwise comparisons were conducted to further investigate the differences between scenarios. If the p-value < 0.05, the data was considered significant.

Independent samples t-tests were conducted to compare the acoustic features between the two parts of the experiment, i.e., the influence of modal particles. It is suitable for comparing the means of two independent groups. Before performing the t-tests, Levene’s test for equality of variances was conducted to check if the assumption

of equal variances holds. If the variances were not equal, the results of the t-test assuming unequal variances were reported. If the p-value < 0.05, the result was significant.

3 Results

In this part, the pitch contour and statistical analysis results of the target acoustic features will be demonstrated and further analyzed.

3.1 Pitch contour analysis

The pitch contours are shown in Figure 6-13. In Figure 6-9, there are four curves and in Figure 10-13, there are three curves. S1, S2, S3 and S4 represent Interrogative, Empty reference, Rhetoric and Referential substitution usages of “shen2me0” in the sentences accordingly.

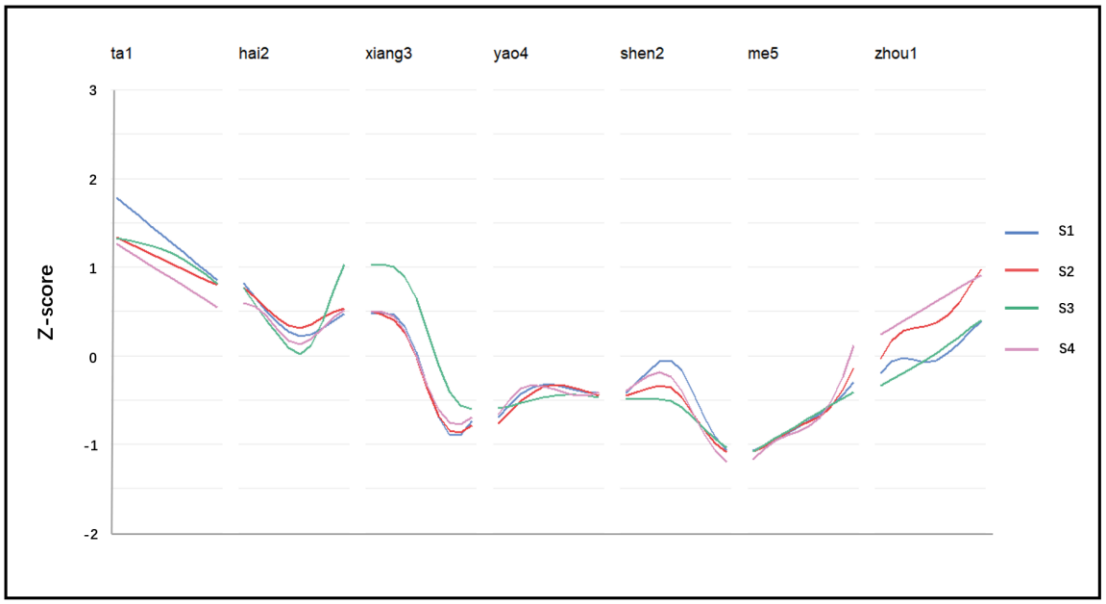

Figure 6 Pitch contour of data in experiment Part I and with nouns of tone 1

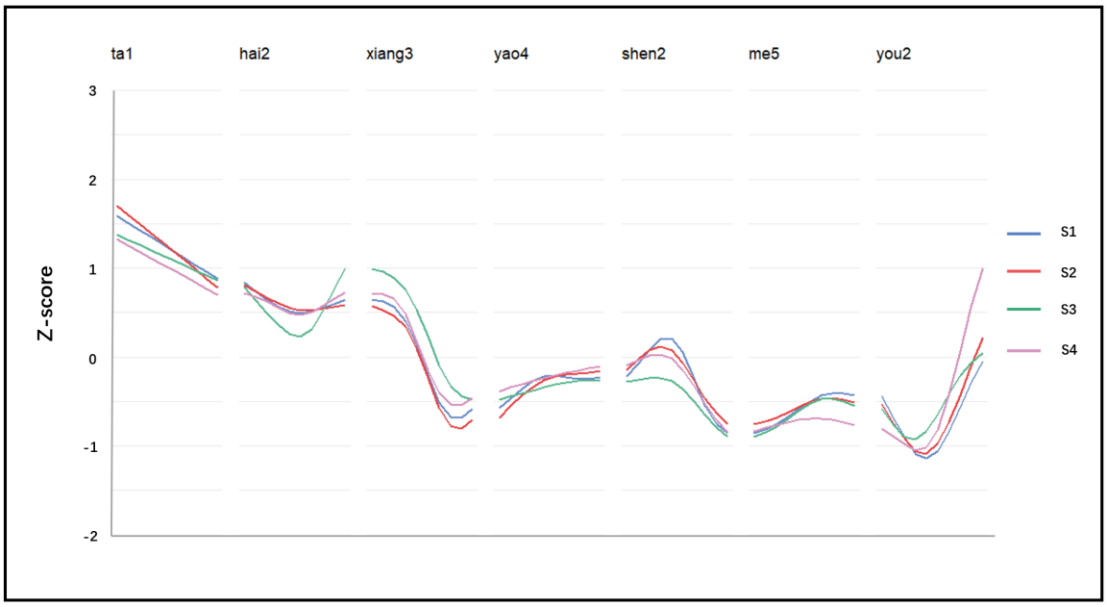

Figure 7 Pitch contour of data in experiment Part I and with nouns of tone 2

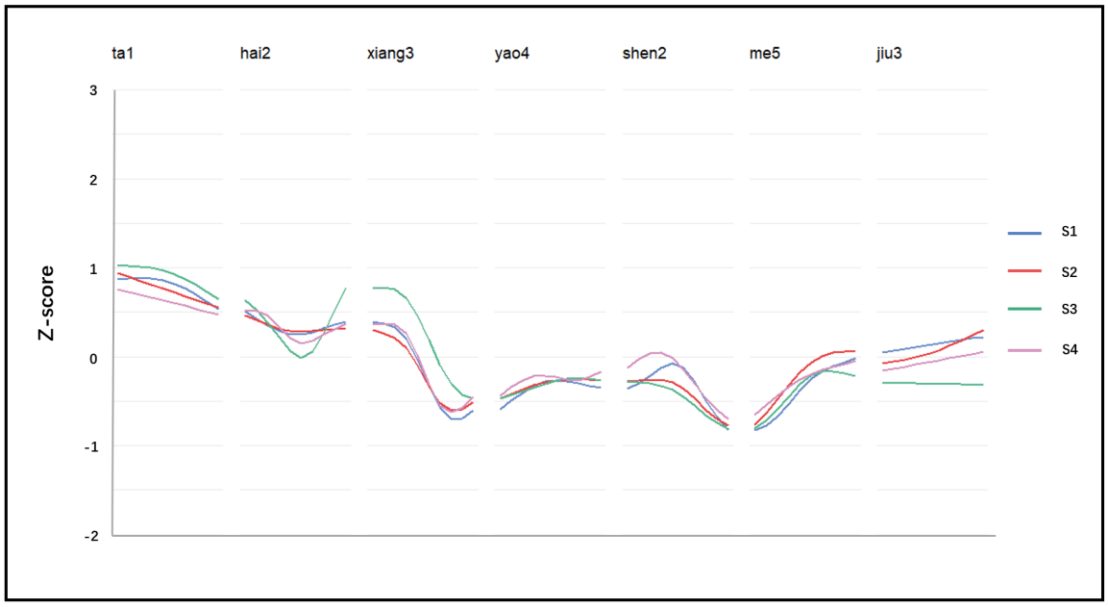

Figure 8 Pitch contour of data in experiment Part I and with nouns of tone 3

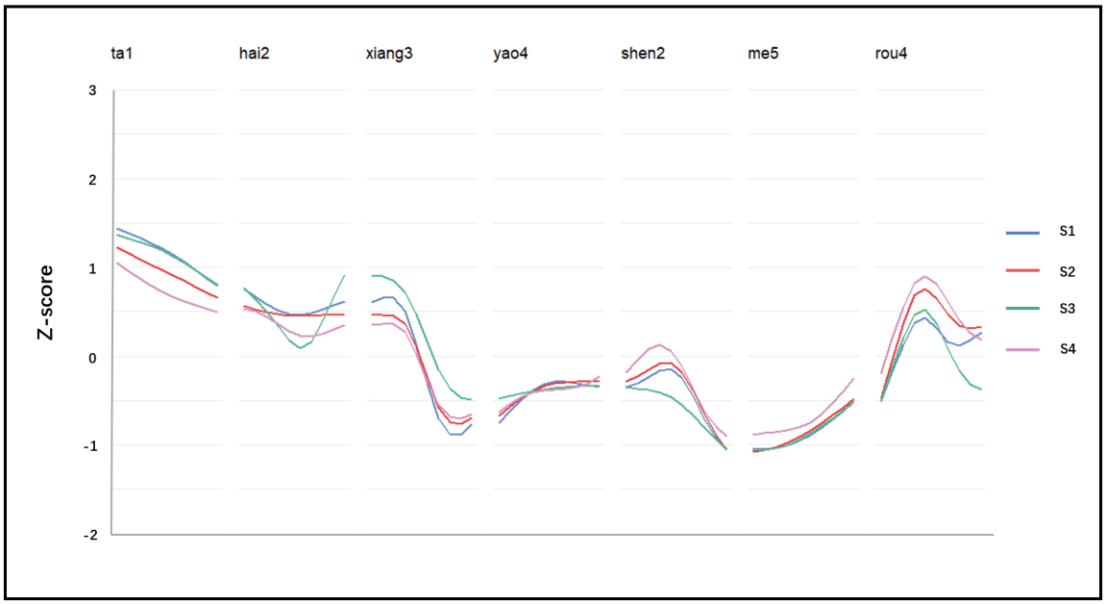

Figure 9 Pitch contour of data in experiment Part I and with nouns of tone 4

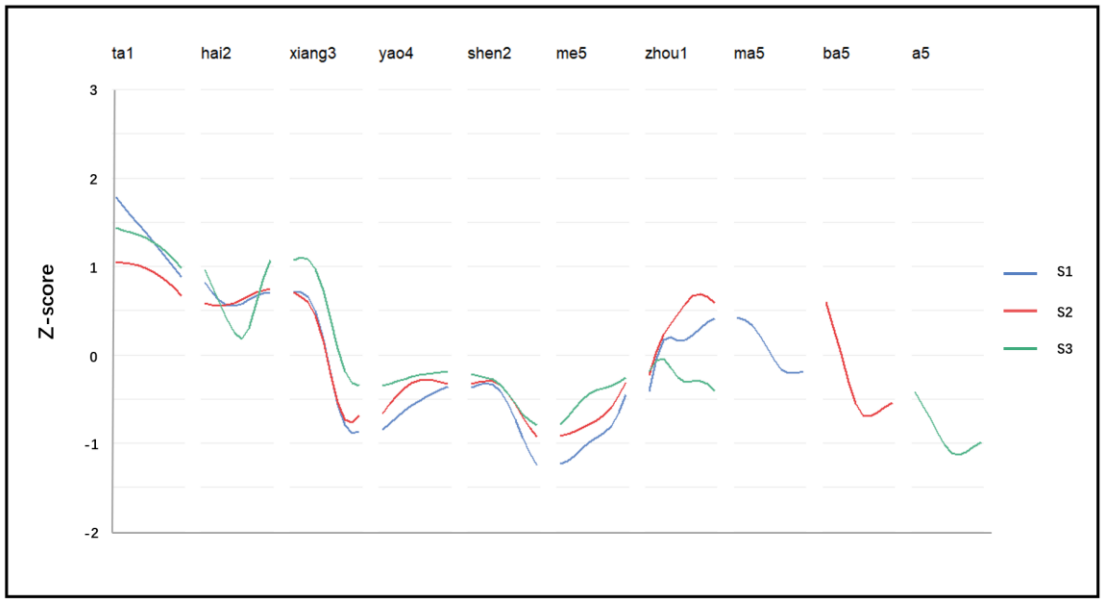

Figure 10 Pitch contour of data in experiment Part II and with nouns of tone 1

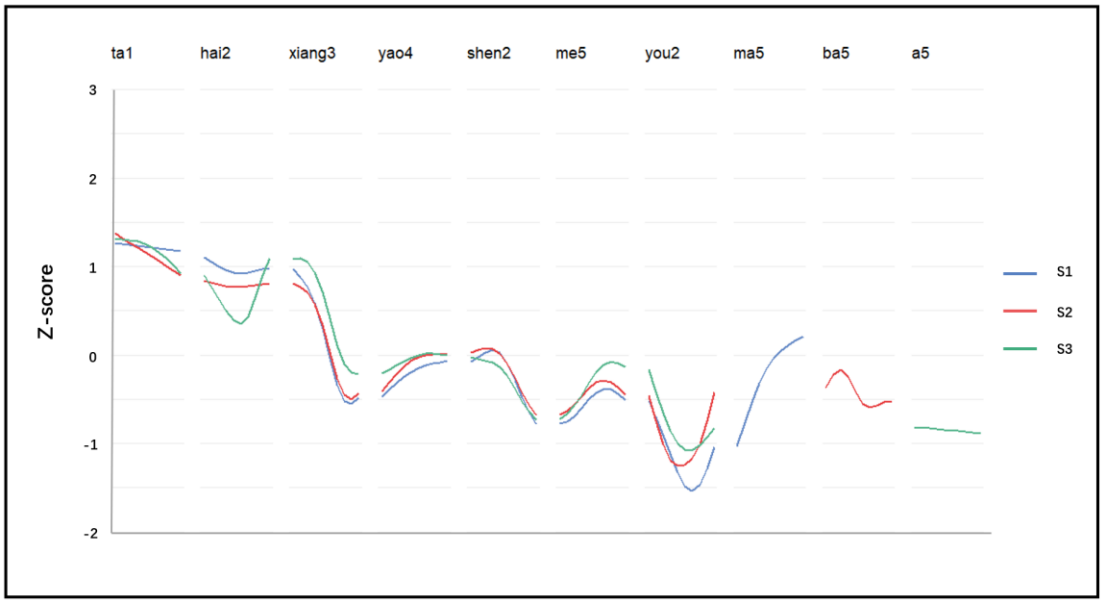

Figure 11 Pitch contour of data in experiment Part II and with nouns of tone 2

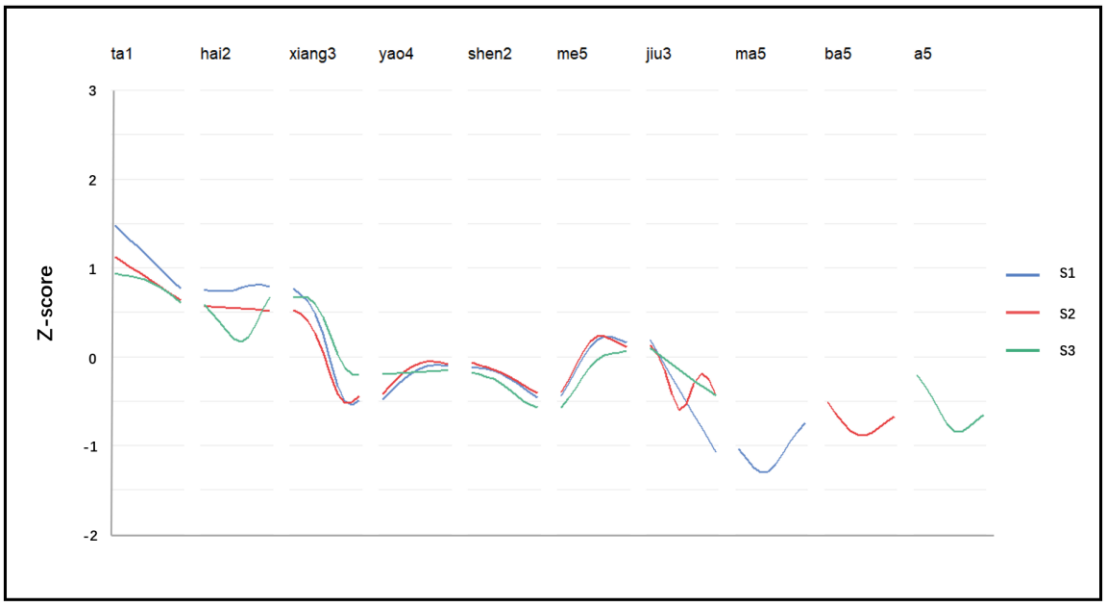

Figure 12 Pitch contour of data in experiment Part II and with nouns of tone 3

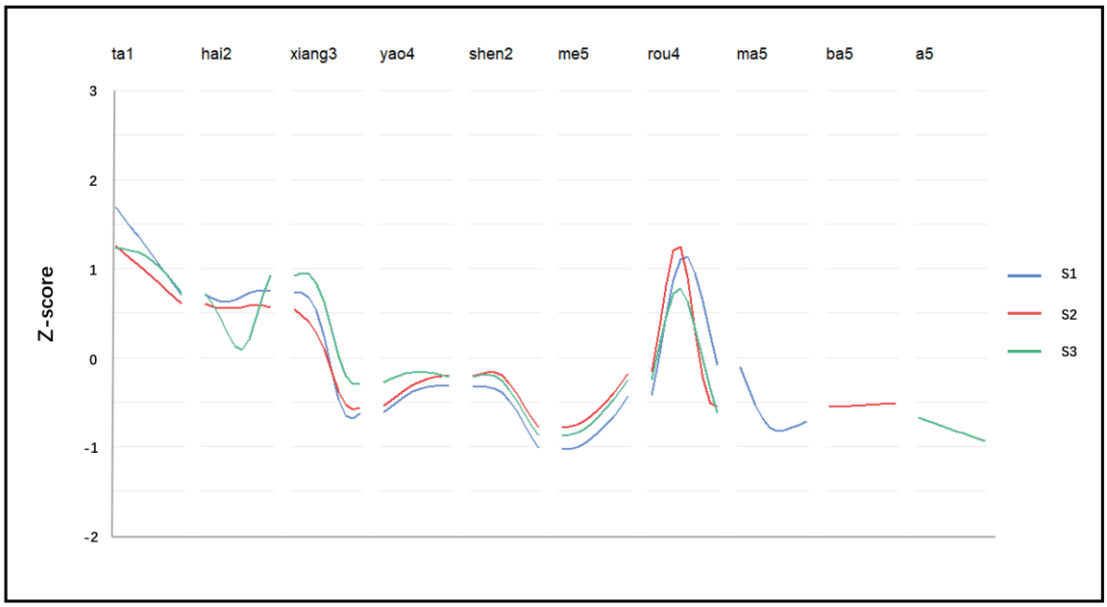

Figure 13 Pitch contour of data in experiment Part II and with nouns of tone 4

From the pitch contour figures, an overview of the pitch patterns throughout the sentences with different usages of "shen2me0" could be obtained. All four usages' sentences start from a relatively high pitch point on the word "ta1", and then go down. At "hai2" and "xiang3", it can be seen that rhetoric usage exhibit dynamic changes, while other usages are relatively flat. At the interrogative pronoun "shen2me0", it is the rhetoric usage that is generally flatter compared to other usages, with minor rises and falls. And as for the interrogation, it shows relatively sharper fluctuations at the word "shen2". Comparing the four figures representing different tones of nouns in the sentences, it is quite clear that the pitch contour of "me0" is influenced by the tone of nouns, which may be one of the factors that affect the acoustic performance of the wh-word in different usages. At the end of the sentences, the interrogative usage holds slight rising towards the final noun. Generally speaking, the interrogation usage shows moderate fluctuations throughout the sentence, and generally exhibits a rising tendency

at the end of the wh-word and the sentence. The empty reference and referential substitution usages are flatter compared to other usages with minor fluctuations, but the latter rises higher at the end of sentences. And the rhetoric usage shows overall distinct pitch changes, particularly at the front part of the sentence.

With the participation of modal particles at the end of the sentences, the pitch curves have some changes, which are generally shown at the rear part of the sentences. When there is the modal particle “ma5” to express questioning, the interrogative usage then does not show the rising and falling of the same degree as without the “ma5”. And the tone of nouns could also have influence on the pitch of modal particles.

3.2 Stress of sentence

Based on the complex theories of Mandarin stress, the pitch contour and the self-designed stress determination method were adopted together for the analysis of sentence stress of different usages of “shen2me0”.

It can be found from the stress analysis results that, in interrogation sentences, when there is no modal particle, the sentence stress is mainly on “ta1”, “hai2”, “xiang3” and the nouns. And when added the modal particle, it generally does not change the stress, but only with a not evident stress falling on the modal particle “ma0” that expresses the query tone at the end of the sentences.

For empty reference usage, in Part I, the four tones' sentences are mainly stressed on "ta1", "hai2" and the nouns. The stress on "xiang3" is not very obvious, which is mainly the increase in duration rather than the increase in pitch that affects its stress. In part II, they are also stressed on "xiang3". The prominence of modal particle is mainly due to a longer duration to express a speculative tone.

When expressing rhetoric meaning, the sentence stress still falls on the first three word and the nouns. However, in this case the difference is that the stress of "hai2" is especially obvious and prominent. The same is true for sentences with modal particles. And the modal particle "a0" mainly has a longer duration to express the rhetoric tone.

When the wh-word is used as referential substitution, the stress falls on "ta1", "hai2" and especially the nouns. But it is worth noticed that in such context the "me0" of the wh-word has particularly longer duration, which is caused by the speakers dragging out this word when their thinking and speaking got stuck.

3.3 Acoustic analysis of interrogative pronoun "shen2me0" with different usages

In this section the analysis will be separately performed on the target global and local features, which will be mainly based on the statistical tests results in order to show the difference between experiment groups. The statistical analysis was conducted using mixed linear model to investigate the differences between usages of "shen2me0", and to control the variables, each test was performed in the same experiment part and tone

of noun. Only significant results in mixed linear model are demonstrated, with all corresponding pairwise comparisons details.

3.3.1 Sentence duration

Here are the statistical analysis results of sentence duration in Table 1.1 and Table 1.2.

	Mixed Linear Model		**Pairwise Comparisons**		
Group	**F (3, 356)**	**P-value**	**Scenario**	**Mean Difference (s)**	**P-value**
Part I Tone 1	84.197	< 0.001	1 vs 2	-0.078	0.189
			1 vs 3	0.075	0.210
			1 vs 4	-0.763	< 0.001
			2 vs 3	0.153	0.011
			2 vs 4	-0.685	< 0.001
			3 vs 4	-0.838	< 0.001
Part I Tone 2	73.285	< 0.001	1 vs 2	-0.040	0.526
			1 vs 3	0.099	0.115
			1 vs 4	-0.728	< 0.001
			2 vs 3	0.138	0.027
			2 vs 4	-0.689	< 0.001
			3 vs 4	-0.827	< 0.001
Part I Tone 3	90.505	< 0.001	1 vs 2	-0.066	0.267
			1 vs 3	0.037	0.535
			1 vs 4	-0.809	< 0.001
			2 vs 3	0.103	0.084
			2 vs 4	-0.742	< 0.001
			3 vs 4	-0.846	< 0.001
Part I Tone 4	95.725	< 0.001	1 vs 2	-0.059	0.302
			1 vs 3	0.017	0.770
			1 vs 4	-0.807	< 0.001
			2 vs 3	0.076	0.185
			2 vs 4	-0.748	< 0.001
			3 vs 4	-0.824	< 0.001

Table 1.1 Mixed linear model results of sentence duration in Part I

	Mixed Linear Model		Pairwise Comparisons		
Group	**F (2, 267)**	**P-value**	**Scenario**	**Mean Difference (s)**	**P-value**
Part II Tone 1	3.756	0.025	1 vs 2	-0.091	0.007
			1 vs 3	-0.041	0.214
			2 vs 3	0.050	0.137
Part II Tone 2	8.424	< 0.001	1 vs 2	-0.118	< 0.001
			1 vs 3	-0.017	0.589
			2 vs 3	0.101	< 0.001
Part II Tone 3	8.074	< 0.001	1 vs 2	-0.120	< 0.001
			1 vs 3	-0.014	0.666
			2 vs 3	0.120	0.001
Part II Tone 4	7.940	< 0.001	1 vs 2	-0.125	< 0.001
			1 vs 3	-0.068	0.030
			2 vs 3	0.056	0.074

Table 1.2 Mixed linear model results of sentence duration in Part II

In Part I, Scenario 4 is significantly different. Referring to the results of wh-duration, it could be seen that it is due to the longer duration of the wh-word “shen2me0” when used as referential substitution. Speakers tend to express their thinking process for trying to capture the object they were going to refer to by extending their pronouncing of “shen2me0”, especially the “me0”. From the mean difference data, it can be specified that in Scenario 4, the duration of sentence is around 0.7 to 0.8 seconds longer than other usages, while among other usages, the differences of sentence duration are not outstanding.

In Part II, due to the participant of modal particles, the sentence duration of interrogation, empty reference and rhetoric question are significantly different. When the nouns are in Tone 1 and Tone 4, the sentence duration when expressing empty

reference and rhetoric question are tend to be longer. Also, combining the results of wh-word duration, it could be discovered that in this case, with the adding of modal particles, the duration differences are no longer caused by wh-word. For example, by examining the stress results, it can be found that in the rhetoric case, the duration of "hai2" is longer than other words.

3.3.2 Sentence pitch range

Statistical results of sentence pitch range are in Table 2.

	Mixed Linear Model		**Pairwise Comparisons**		
Group	**F (2, 267)**	**P-value**	**Scenario**	**Mean Difference (Hz)**	**P-value**
Part II Tone 2	4.830	0.009	1 vs 2	-39.344	0.007
			1 vs 3	-39.100	0.008
			2 vs 3	0.244	0.987
Part II Tone 3	9.291	< 0.001	1 vs 2	-35.478	0.038
			1 vs 3	-73.200	< 0.001
			2 vs 3	-37.722	0.027
Part II Tone 4	3.244	0.041	1 vs 2	-41.244	0.013
			1 vs 3	-27.889	0.093
			2 vs 3	13.356	0.420

Table 2 Mixed linear model results of sentence pitch range

In Part I, there is no significant difference of sentence pitch range. But in Part II, there exist significant difference when the tones of noun are 2, 3 and 4. More specifically speaking, the sentence pitch range of empty reference usage is higher than

interrogation in all four tones of noun. And for tone 2 and 3 cases, the overall pitch range of rhetoric question is also higher than interrogation.

3.3.3 Sentence boundary tone

The differences of sentence boundary tone across usages of "shen2me0" can be checked in Table 3.

	Mixed Linear Model		Pairwise Comparisons		
Group	**F (3, 356)**	**P-value**	**Scenario**	**Mean Difference**	**P-value**
Part I Tone 2	5.890	0.001	1 vs 2	-0.166	0.297
			1 vs 3	0.035	0.824
			1 vs 4	-0.560	< 0.001
			2 vs 3	0.201	0.206
			2 vs 4	-0.394	0.014
			3 vs 4	-0.595	< 0.001
Group	**F (2, 267)**	**P-value**	**Scenario**	**Mean Difference**	**P-value**
Part II Tone 2	30.237	< 0.001	1 vs 2	1.086	< 0.001
			1 vs 3	0.757	< 0.001
			2 vs 3	-0.329	0.022
Part II Tone 3	18.040	< 0.001	1 vs 2	0.819	< 0.001
			1 vs 3	0.755	< 0.001
			2 vs 3	-0.063	0.676

Table 3 Mixed linear model results of sentence boundary tone

From the statistical results, it can be concluded that when the noun is of tone 2, no matter there is a modal particle at the end of the sentence or not, the sentence boundary tone shows significant differences. Especially the boundary tone of referential substitution cases, whose pitch are relatively higher for the mean pitch of the ending nouns. When there is a modal particle and with the O being a noun with tone 2 and 3,

it can be seen that now the interrogation sentences have higher boundary tone. It might be influenced by the tone 2 and 3's rising that leads the modal particle to go even higher. Also, it is possible that when added the "ma0", the sentences have a semantic implicature of yes-no question for the speakers involuntarily, which caused the rising of sentence boundary tone to express the interrogative tone in some cases. While when there is no modal particle, which means the interrogation is expressed by a wh-question with a wh-word "shen2me0", there are no significant rising in boundary tone.

3.3.4 Wh-word duration

The details of differences of wh-word duration are demonstrated in Table 4.

	Mixed Linear Model		**Pairwise Comparisons**		
Group	**F (3, 356)**	**P-value**	**Scenario**	**Mean Difference (s)**	**P-value**
Part I Tone 1	193.682	< 0.001	1 vs 2	0.017	0.531
			1 vs 3	0.062	0.020
			1 vs 4	-0.495	< 0.001
			2 vs 3	0.045	0.088
			2 vs 4	-0.512	< 0.001
			3 vs 4	-0.557	< 0.001
Part I Tone 2	160.447	< 0.001	1 vs 2	0.024	0.363
			1 vs 3	0.054	0.043
			1 vs 4	-0.444	< 0.001
			2 vs 3	0.030	0.262
			2 vs 4	-0.468	< 0.001
			3 vs 4	-0.498	< 0.001
Part I Tone 3	163.509	< 0.001	1 vs 2	0.018	0.496
			1 vs 3	0.050	0.067
			1 vs 4	-0.465	< 0.001
			2 vs 3	0.031	0.249

Mixed Linear Model			Pairwise Comparisons		
Group	**F (3, 356)**	**P-value**	**Scenario**	**Mean Difference (s)**	**P-value**
Part I Tone 4	144.415	< 0.001	2 vs 4	-0.483	< 0.001
			3 vs 4	-0.514	< 0.001
			1 vs 2	0.013	0.639
			1 vs 3	0.039	0.168
			1 vs 4	-0.461	< 0.001
			2 vs 3	0.026	0.362
			2 vs 4	-0.475	< 0.001
			3 vs 4	-0.500	< 0.001

Table 4 Mixed linear model results of wh-word duration

Due to the referential substitution cases in Part I, the wh-word duration in all tone groups demonstrate significance for Scenario 4, which have a wh-word duration averagely 0.5 seconds longer than other context. It can also be confirmed by the results of sentence duration.

3.3.5 Wh-word pitch range

Here are the statistical test results of wh-word pitch range in Table 5.1 and Table 5.2.

Mixed Linear Model			Pairwise Comparisons		
Group	**F (3, 356)**	**P-value**	**Scenario**	**Mean Difference (Hz)**	**P-value**
Part I Tone 1	3.537	0.015	1 vs 2	18.967	0.132
			1 vs 3	30.678	0.015
			1 vs 4	-5.278	0.675
			2 vs 3	11.711	0.352
			2 vs 4	-24.244	0.054
			3 vs 4	-35.956	0.004
Part I Tone 2	3.401	0.018	1 vs 2	20.044	0.146
			1 vs 3	32.300	0.020

Mixed Linear Model			Pairwise Comparisons		
Group	**F (3, 356)**	**P-value**	**Scenario**	**Mean Difference (Hz)**	**P-value**
Part I Tone 3	5.618	0.001	1 vs 4	-6.656	0.629
			2 vs 3	12.256	0.374
			2 vs 4	-26.700	0.053
			3 vs 4	-38.956	0.005
			1 vs 2	3.944	0.768
			1 vs 3	25.656	0.056
			1 vs 4	-28.933	0.031
			2 vs 3	21.711	0.106
			2 vs 4	-32.878	0.015
			3 vs 4	-54.589	< 0.001
Part I Tone 4	7.220	< 0.001	1 vs 2	-4.100	0.743
			1 vs 3	14.378	0.250
			1 vs 4	-41.267	0.001
			2 vs 3	18.478	0.140
			2 vs 4	-37.167	0.003
			3 vs 4	-55.644	< 0.001

Table 5.1 Mixed linear model results of wh-word pitch range in Part I

Mixed Linear Model			Pairwise Comparisons		
Group	**F (2, 267)**	**P-value**	**Scenario**	**Mean Difference (Hz)**	**P-value**
Part II Tone 4	3.902	0.021	1 vs 2	-18.189	0.025
			1 vs 3	-20.533	0.011
			2 vs 3	-2.344	0.771

Table 5.2 Mixed linear model results of wh-word pitch range in Part II

In experiment Part I, the wh-word pitch range of all four tone groups are significantly different. For Tone 1 and Tone 2, "shen2me0' in the rhetoric question cases have around 30 to 35 Hz smaller pitch range than interrogation and referential substitution. For Tone 3 and Tone 4, the wh-word pitch range of referential substitution are larger than other cases. To be more specific, it is 20 to 30 Hz larger than interrogation and empty reference, and about 55 Hz larger than rhetoric question.

However, when there are modal particles at the end, generally existing rules could not be found.

3.3.6 Wh-word boundary tone

In following Table 6.1 and Table 6.2, the results of wh-word boundary tone can be checked.

	Mixed Linear Model		Pairwise Comparisons		
Group	**F (3, 356)**	**P-value**	**Scenario**	**Mean Difference**	**P-value**
Part I Tone 1	5.951	0.001	1 vs 2	-0.183	0.193
			1 vs 3	0.306	0.030
			1 vs 4	-0.226	0.108
			2 vs 3	0.489	0.001
			2 vs 4	-0.043	0.759
			3 vs 4	-0.532	< 0.001
Part I Tone 2	5.233	0.002	1 vs 2	0.322	0.067
			1 vs 3	0.205	0.242
			1 vs 4	0.675	< 0.001
			2 vs 3	-0.117	0.504
			2 vs 4	0.353	0.044
			3 vs 4	0.470	0.008
Part I Tone 3	6.378	< 0.001	1 vs 2	0.256	0.094
			1 vs 3	-0.496	0.001
			1 vs 4	0.617	< 0.001
			2 vs 3	0.239	0.119
			2 vs 4	0.361	0.019
			3 vs 4	0.122	0.427

Table 6.1 Mixed linear model results of wh-word boundary tone in Part I

	Mixed Linear Model		Pairwise Comparisons		
Group	**F (2, 267)**	**P-value**	**Scenario**	**Mean Difference**	**P-value**
Part II Tone 1	12.324	< 0.001	1 vs 2	0.137	0.362
			1 vs 3	0.703	< 0.001
			2 vs 3	0.566	< 0.001
Part II Tone 3	6.511	0.002	1 vs 2	0.103	0.461
			1 vs 3	-0.374	0.008
			2 vs 3	-0.477	0.001

Table 6.2 Mixed linear model results of wh-word boundary tone in Part II

There are significant differences appearing in most groups, among which the interrogation and empty reference mainly have more evident rising at the ending boundary of “shen2me0”. It might suggest that for interrogation of wh-question, the strategy is to express the questioning tone by rise the tone of wh-word, especially at the end of it. While for rhetoric question, when there is no modal particle, due to the focus being mainly at the front part of sentence, such as the adverb “hai2”, some speakers choose to pass the wh-word quickly, without modifying its tone. But when the sentences have the participation “a0” at the end, which forms part of the sentence stress, it can also be observed from the pitch curves that “me0” has a boundary tone going up that could be a way to express emotion and to join the stressed word after it.

3.4 Analysis of modal particles’ influence on acoustic features of interrogative pronoun “shen2me0” with different usages

In this part independent samples t-tests were conducted in order to quantify the influences of modal particles on sentence boundary tone, wh-word duration, wh-word

pitch range and wh-word boundary tone. Only significant results in the tests are shown in tables.

3.4.1 Sentence boundary tone

In Table 7 are the independent sample t-test results of differences of sentence boundary tone between experiment Part I and Part II.

Part I vs Part II		**Levene's Test for Equality of Variances**		**T-test for Equality of Means**		
Group		**F**	**Sig.**	**t**	**Sig. (2-tailed)**	**Mean Difference**
Scenario 1	Tone 1	0.019	0.890	5.455	< 0.001	0.833
	Tone 3	8.761	0.003	-4.211	< 0.001	-0.663
Scenario 2	Tone 1	9.038	0.003	6.606	< 0.001	0.979
	Tone 2	1.950	0.164	7.583	< 0.001	1.208
Scenario 3	Tone 1	1.455	0.229	3.825	< 0.001	0.621
	Tone 2	1.154	0.284	3.908	< 0.001	0.678

Table 7 Independent sample t-test results of sentence boundary tone

In Tone 1 group, the z-score of sentence boundary tone of Part I is significantly higher than Part II. In Tone 2 group, the z-score of Part I is higher than the other Part in empty reference and rhetoric cases. And in Tone 3, when express query tone using modal particle "ma0", the z-score of sentence boundary will be higher.

3.4.2 Wh-word duration

The influences of modal particles on wh-word duration can be found in Table 8.

Part I vs Part II		Levene's Test for Equality of Variances		T-test for Equality of Means		
Group		F	Sig.	t	Sig. (2-tailed)	Mean Difference (s)
Scenario 1	Tone 1	15.624	< 0.001	4.726	< 0.001	0.059
	Tone 2	8.013	0.005	5.276	< 0.001	0.057
	Tone 4	10.247	0.002	4.183	< 0.001	0.041
Scenario 2	Tone 1	14.494	< 0.001	3.276	0.001	0.039
	Tone 2	10.794	0.001	2.736	0.007	0.030
Scenario 3	Tone 3	1.964	0.163	-2.264	0.025	-0.019

Table 8 Independent samples t-test results of wh-word duration

Excluded the referential substitution cases (Part II does not have this group), comparing the impact of the presence or absence of modal particles on the duration of wh-word, it can be found that in the cases of nouns with tone 1 and 2, the duration of interrogative word without modal particles in the sentences is significantly different when expressing interrogation and empty reference. Among tone 3 groups, the duration of wh-word with modal particles when expressing rhetorical meanings is longer than when without modal particles. And in tone 4 groups, the wh-word duration is longer only when expressing interrogation and when there is no modal particle.

3.4.3 Wh-word pitch range

The results of wh-word pitch range are shown in Table 9.

Part I vs Part II		Levene's Test for Equality of Variances		T-test for Equality of Means		
Group		F	Sig.	t	Sig. (2-tailed)	Mean Difference (Hz)
Scenario 1	Tone 1	22.834	< 0.001	2.250	0.026	29.444
	Tone 2	15.239	< 0.001	2.084	0.039	30.244
	Tone 3	16.945	< 0.001	2.828	0.005	29.733
	Tone 4	21.516	< 0.001	2.948	0.004	25.644
Scenario 2	Tone 3	26.006	< 0.001	2.901	0.004	27.678

Table 9 Independent samples t-test results of wh-word pitch range

Generally speaking, in interrogation context, the sentences with modal particles have 25 to 30 Hz higher wh-word pitch range than those without modal particles.

3.4.4 Wh-word boundary tone

The results of wh-word boundary tone can be found in Table 10.

		Levene's Test for Equality of Variances		T-test for Equality of Means		
Group		F	Sig.	t	Sig. (2-tailed)	Mean Difference
Scenario 1	Tone 2	0.007	0.935	2.779	0.006	0.463
	Tone 3	6.898	0.009	4.195	< 0.001	0.552
Scenario 2	Tone 3	1.096	0.297	2.753	0.007	0.399
Scenario 3	Tone 3	11.954	0.001	-2.152	0.033	-0.318

Table 10 Independent samples t-test results of wh-word boundary tone

Overall, the significant differences in the boundary tone at the end of wh-word are concentrated in the case of noun of tone 3. Specifically, when expressing questioning and empty reference, the z-score of the boundary tone at the end of "shen2me0" without a modal particle in the sentences is higher than that with a modal particle. It could because that it is necessary to express interrogation by the boundary tone of wh-word when there is no help of the modal particle. While in rhetoric cases, the z-score of wh-word's boundary tone of sentences with modal particles is higher than those without them, which may due to the reason that the modal particle "a0" plays the role of strengthening the disapproval and dissatisfactory tone by rising the pitch.

4 Conclusions and discussions

In this study the acoustic features of various usages of the Mandarin interrogative pronoun “shen2me0” – interrogation, empty reference, rhetoric and referential substitution were analyzed, not only highlighting their acoustic distinctions but also reflect the underlying pragmatic functions. The influence of modal particles and the tone of nouns appearing between the interrogative pronoun and the modal particles were also considered into account, which resulted having certain influence on the acoustic features, but not general in all cases.

Interrogative usage of “shen2me0” is characterized by moderate pitch fluctuations, especially with sharper variations at the word “shen2”. Without modal particle, in which case the sentences are more of wh-question, the speakers choose to rise the tone at the end of the sentences to express the questioning intention. The presence of querying modal particle “ma0” at the end of the sentences would turn the sentences into a more yes/no question, which makes the speakers not raise their tone that much as in the former cases, but its presence still adds a slight stress, maintaining the query tone and enhancing the clarity of the question.

In contrast, the empty reference usage exhibits a relatively flat pitch contour with minor fluctuations both at sentence level and at wh-word level, reflecting its neutral or non-specific nature. This flatness in pitch and stress distribution aligns with this function of “shen2me0” as more of a placeholder rather than a focal point for eliciting

information. It accords with the intention of speculating but not expecting any specific answers.

Rhetoric usage, on the other hand, displays dynamic pitch changes, particularly at “hai2”, which shows significant dips followed by sharp rises. This word is particularly stressed in this usage with significantly higher pitch and longer duration. As for the wh-word “shen2me0”, it shows relatively flatter contour, even comparing to the empty reference cases. It could suggest that the interrogative pronoun plays less important role of expressing strongly negative emotion than the adverb “hai2” when speaking. Fewer speaker chooses to show their dissatisfaction or intention of persuasion using interrogative pronoun when there is an adverb that can be used to strengthen the mood.

Referential substitution usage has flatter pitch contour at the front of the sentences, but rises higher than other cases at the end of the sentences. And it is general that speakers choose to last their pronouncing of “me0” in “shen2me0”, which also extended the overall duration of sentence. These all points to thinking process in searching of the wanted word of the speaker themselves.

Admittedly, there exist limitations in this study. First, due to the limited sample size, the fitting results might not be as comprehensive as desired to represent general Mandarin speaking population. The experimental design also presents certain limitations. There is a possibility that some speakers did not fully understand the context or were unable to immerse themselves adequately in the given scenarios, which might have influenced their speech patterns. In future research, it could employ more

immersive methods to create more realistic contexts, such as having conversations with speakers, using videos, or even applying AI technologies. Another significant limitation lies in the recording instruments used in this study. Due to the constraint, participants recorded their voices using different mobile phones and under varying environmental conditions, although they had instructions of recording and were requested to manage avoiding the noise. This variation likely introduced background noise and inconsistencies in the recording quality, which could have affected the accuracy of the acoustic data. Moreover, future research should explore better methods to quantify the acoustic data and to capture all relevant features.

This study could provide some suggestions for the ASR of non-interrogative usages of wh-word in Mandarin Chinese from the perspective of acoustic analysis in order to solve the problem of recognition and understanding of these types of special usages in sentences sharing the same syntactic structure.

In recent years, the ASR has seen great advancements due to the deep learning technology. The deep learning-based ASR models usually are trained on large datasets of labeled audio data, and future studies could enrich the training data with specific additional features such as the ones investigated in this study in order to have better efficiency and performance. And the models convert raw audio signals into features that can be processed by them, such as Mel-Frequency Cepstral Coefficients (MFCCs), Mel-spectrogram, etc., and process them using feature extraction layer in their structures. Researchers could also integrate the acoustic features analyzed in this

present to this layer along with MFCCs and other commonly used features in these models. Features such as MFCCs are indeed widely used in deep learning-based ASR models due to various reasons. They show good performance in capturing features that align well with human perceptions, such as in Speech Emotion Recognition (Nancy et al., 2018, Liu et al. 2023). And MFCCs have benefits such as reducing the dimensionality (Errity & McKenna, 2007), which is the limitation of this present study. Despite these positive aspects of MFCCs, integrating the pitch, duration, and other basic acoustic features in the feature set along with the standard features in deep learning-based ASR could still be beneficial. For example, Gupta et al. (2020) proposed pitch-synchronous single frequency filtering spectrogram to improve the recognition of speech emotion. Nevertheless, MFCCs also need to face the problem of processing signals with background noises (Khan et al., 2019). Moreover, modern ASR models require specific contextual information, and in recent studies, researchers have proposed ASR model based on audio conditioned large language models (AcLLM) that makes use of the LLM through the input of continuous speech representations and contextual information (Bai et al. 2024). The results of this study may also help such models evaluating the contexts.

References

Bai, Y., Chen, J., Chen, J., Chen, W., Chen, Z., Ding, C., ... & Zou, M. (2024). Seed-ASR: Understanding Diverse Speech and Contexts with LLM-based Speech Recognition. *arXiv preprint arXiv:2407.04675*.

Chen, A. (2006). Interface between information structure and intonation in Dutch wh-questions. In R. Hoffmann, & H. Mixdorff (Eds.), *Speech Prosody 2006*.

Cruttenden, A. (1997). *Intonation*. Cambridge University Press.

DeFrancis, J. (1976). *Beginning Chinese*. Yale University Press.

Errity, A., & McKenna, J. (2007). A comparative study of linear and nonlinear dimensionality reduction for speaker identification. *2007 15th International Conference on Digital Signal Processing*. https://doi.org/10.1109/icdsp.2007.4288650

Frota, S., & Prieto, P. (Eds.). (2015). *Intonation in romance*. OUP Oxford.

Gupta, S., Fahad, Md. S., & Deepak, A. (2020). Pitch-synchronous single frequency filtering spectrogram for speech emotion recognition. *Multimedia Tools and Applications*, *79*(31–32), 23347–23365. https://doi.org/10.1007/s11042-020-09068-1

Haan-van Ditzhuysen, J. J. M. (2001). *Speaking of questions: An exploration of Dutch question intonation*. Netherlands Graduate School of Linguistics.

Halliday, M. A. K. (1967). Notes on transitivity and theme in English: Part 2. *Journal of Linguistics, 3*(2), 199–244. http://www.jstor.org/stable/4174965

Hedberg, Nancy & Sosa, Juan. (2002). The prosody of questions in natural discourse. *Proceeding of Speech Prosody.*

Jiang, D. N., & Cai, L. H. (2003). Hanyu Yiwen Yuqi De Shengxue Tezheng Yanjiu [Study of acoustic features of Chinese interrogative tone]. *Proceedings of the 6th National Modern Phonetics Conference.*

Khan, U., Sarim, M., Bin Ahmad, M., & Shafiq, F. (2019). Feature extraction and modeling techniques in speech recognition: A Review. *2019 4th International Conference on Information Systems Engineering (ICISE)*. https://doi.org/10.1109/icise.2019.00020

Lambrecht, K., & Michaelis, L. A. (1998). Sentence accent in information questions: Default and projection. *Linguistics and Philosophy, 21*(5), 477–544.

Lee, O.J. (2005). The prosody of questions in Beijing Mandarin. Ohio State University.

Lin, M. C., Yan, J. Z., & Sun, G. H. (1984). Beijinghua Liangzizu Zhengchang Zhongyin De Chubu Shiyan [Primary experiment of the normal stress of two words combination in Beijinghua]. *Dialect, 1*, 57-73.

Lin, M. C. (2006). Yiwen He Chenshu Yuqi Yu Bianjiediao [Interrogative and declarative tone with boundary tone]. *Chinese Language Study, 4*, 364-376.

Lin, M. C. (2012). *Hanyu Yudiao Shiyan Yanjiu [Experimental study on Chinese intonation]*, Chinese Social Sciences Press.

Lin, T., & Wang, L. (2013). *Yuyingxue Jiaocheng (Zengding Ban)* [*Phonetics (Revised Edition)*]. Peking University Press.

Lin, Y. W. (1985). Tan Yiwen Ju [A discussion about interrogative sentence]. *Chinese Language Study, 2*, 91.

Liu, F., Surendran, D., & Xu, Y. (2006). Classification of statement and question intonations in Mandarin. *Proc. 3rd speech prosody*, 603-606.

Liu, F., & Xu, Y. (2005). Parallel encoding of focus and interrogative meaning in Mandarin intonation. *Phonetica, 62*(2-4), 70–87. https://doi.org/10.1159/000090090

Liu, X. F. (2016). Yiwen Daici Ju De Yuyin Yu Jufa Jiekou Yanjiu [Study of Interfaces between Phonetics and Syntax of interrogative pronoun sentence]. Graduate School of Chinese Academy of Social Sciences.

Liu, Z. T., Han, M. T., Wu, B. H., & Rehman, A. (2023). Speech emotion recognition based on convolutional neural network with attention-based bidirectional long short-term memory network and multi-task learning. *Applied Acoustics*, *202*, 109178.

Lv, S. X. (1982). *Zhongguo Wenfa Yaolue [Essentials of Chinese grammar]*. The Commercial Press.

Lv, S. X. (1985). Yiwen. Fouding. Kending [Interrogation. Negation. Affirmation]. *Chinese Language Study, 5*, 241-250.

Nancy, A. M., Kumar, G. S., Doshi, P., & Shaw, S. (2018). Audio based emotion recognition using mel frequency cepstral coefficient and support vector machine. *Journal of Computational and Theoretical Nanoscience*, *15*(6-7), 2255-2258.

Ni, C. J., Liu, W. J., & Xu, B. (2012). Hanyu He Yingyu Yingao Zhongyin Zidong Biaozhu Fangfa De Duibi Yu Fenxi [The comparison and analysis between Chinese Putonghua and English stress automatic annotation]. *Acta Acustica, 37*(5), 553-560.

Qi, F. (2012). *Xiandai Hanyu Jiaodian Yanjiu [Focus study of modern Chinese]*. Fudan University.

Shen, J. & Hoek, J. H. v. d. (1994). Hanyu Yushi Zhongyin De Yinli (Jianyao Baogao) [The acoustic theory of Chinese energy stress (a brief report)]. *Chinese Language Study, 3*, 10-15.

Shen, X. N. S. (1990). *The prosody of Mandarin Chinese*. University of California Press.

Shiamizadeh, Z., Caspers, J., & Schiller, N. O. (2015). Acoustic correlates of Persian in-situ-wh-questions. In *ICPhS*.

Tang, Y. L., Shi, Y. Z. (2009). Yiwen He Jiaodian Zhi Guanxi [Relationship between question and focus]. *Journal of Foreign Languages, 32*(1), 51-57.

Wang, M. L. (2009). Yinxixue Zhongyin Lilun Jianshu [A survey of theory of stress in Phonology]. *Foreign Language Learning Theory and Practice, 3*, 83-87.

Wu, Y. H., Tao, J. H., & Lu, J. L. (2006). Hanyu Yiwen Yudiao De Yunlv Fenxi [Intonational analysis of Chinese interrogative tone]. In 7th *Chinese Phonetics Conference and International Forum on Frontier Issues in Phonetics*.

Yuan, J., Shih, C., & Kochanski, G. (2002). Comparison of declarative and interrogative intonation in Chinese. In *Speech Prosody 2002*.

Yuan, J., & Jurafsky, D. (2005, November). Detection of questions in Chinese conversational speech. In *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2005. (pp. 47-52). IEEE.

Zhao, J. J., Yang, Y. F., & Lv, S. N. (2013). Jiyu Tongji De Jixuwen Yuju Jiaodian De Fenbu Tedian Yanjiu [Statistical analysis of the sentence focus distribution in the narrative discourse]. *Journal of Chinese Information Processing, 27*(1), 81-85.

Zhao, J. J., Yang, X. H., Yang, Y. F., & Lv, S. N. (2012). Hanyu Zhong Jiaodian Yu Zhongyin De Duiying Guanxi – Jiyu Yuliaoku De Chubu Yanjiu [The relationship between focus and accent in Mandarin: an exploratory study based on corpus]. *Studies in Language and Linguistics, 32*(4), 55-59.

Zhao, Y. R. (1968). *Yuyan Wenti [Issues on language]*. The Commercial Press of Taiwan.

Zhao, Y. R. (1979). *Hanyu Kouyu Yufa [A grammar of spoken Chinese]*. The Commercial Press.

APPENDIX A

Sentences stress score results

Part I – Scenario I – Tone 1

	ta1	hai2	xiang3	yao4	shen2	me5	zhou1
1	3	7	50	9	22	14	44
2	46	44	9	5	8	1	4
3	38	24	17	1	4	2	37
score	33.4	26.6	21.2	3.8	8.8	4.1	28.5

Part I – Scenario I – Tone 2

	ta1	hai2	xiang3	yao4	shen2	me5	you2
1	1	1	41	9	11	7	76
2	52	59	19	15	14	5	1
3	34	23	22	1	7	0	8
score	32.8	29.4	24.9	6.8	9.9	2.9	19.5

Part I – Scenario I – Tone 3

	ta1	hai2	xiang3	yao4	shen2	me5	jiu3
1	5	9	50	9	21	17	44
2	46	43	12	11	6	8	9
3	31	23	11	0	4	3	30
score	30.3	26.2	19.1	5.1	8	7.3	26.5

Part I – Scenario I – Tone 4

	ta1	hai2	xiang3	yao4	shen2	me5	rou4
1	0	8	42	8	30	4	47
2	53	53	11	13	6	0	3
3	33	24	24	1	2	2	38

score	32.4	29.5	23.7	6	8.8	1.8	29.3

Part I – Scenario II – Tone 1

	ta1	**hai2**	**xiang3**	**yao4**	**shen2**	**me5**	**zhou1**
1	2	4	29	9	8	11	29
2	24	52	21	8	3	7	6
3	63	23	8	0	2	3	50
score	39.1	27.9	16.1	4.2	3.5	5.8	32.6

Part I – Scenario II – Tone 2

	ta1	**hai2**	**xiang3**	**yao4**	**shen2**	**me5**	**you2**
1	0	4	35	6	6	7	76
2	27	56	22	17	10	8	1
3	61	24	11	3	6	0	8
score	38.6	29.6	19.1	7.8	7.2	3.8	19.5

Part I – Scenario II – Tone 3

	ta1	**hai2**	**xiang3**	**yao4**	**shen2**	**me5**	**jiu3**
1	4	4	31	7	9	12	44
2	22	46	20	15	7	15	9
3	52	12	9	2	3	4	32
score	33.4	20.6	16.7	6.9	5.4	8.9	27.5

Part I – Scenario II – Tone 4

	ta1	**hai2**	**xiang3**	**yao4**	**shen2**	**me5**	**rou4**
1	4	11	37	10	11	8	36
2	27	48	17	13	10	4	6
3	56	18	17	2	2	0	46
score	36.9	25.6	21	6.9	6.2	2.8	32

Part I – Scenario III – Tone 1

	ta1	hai2	xiang3	yao4	shen2	me5	zhou1
1	2	17	13	2	9	11	44
2	46	32	37	5	1	1	4
3	37	39	25	0	0	1	37
score	32.7	32.5	26.2	1.9	2.1	3	28.5

Part I – Scenario III – Tone 2

	ta1	hai2	xiang3	yao4	shen2	me5	you2
1	1	8	14	1	14	9	71
2	56	34	41	13	4	4	0
3	29	46	25	0	0	2	18
score	31.5	34.8	27.6	4.1	4	4	23.2

Part I – Scenario III – Tone 3

	ta1	hai2	xiang3	yao4	shen2	me5	jiu3
1	1	16	13	0	9	11	55
2	52	33	32	11	3	13	1
3	29	35	24	0	1	2	29
score	30.3	30.6	24.2	3.3	3.2	7.1	25.8

Part I – Scenario III – Tone 4

	ta1	hai2	xiang3	yao4	shen2	me5	rou4
1	0	18	20	2	20	7	47
2	52	34	33	17	3	1	0
3	33	38	32	0	0	0	39
score	32.1	32.8	29.9	5.5	4.9	1.7	28.9

Part I – Scenario IV – Tone 1

	ta1	hai2	xiang3	yao4	shen2	me5	zhou1
1	4	3	8	18	13	70	14
2	62	64	22	6	9	2	20

3	21	4	3	2	1	5	48
score	29.9	21.8	9.7	6.4	5.8	17.1	32.8

Part I – Scenario IV – Tone 2

	ta1	**hai2**	**xiang3**	**yao4**	**shen2**	**me5**	**you2**
1	0	1	13	12	9	71	65
2	64	74	42	20	12	2	1
3	20	6	8	2	5	3	15
score	29.2	25.4	19.2	9.4	7.9	16.3	20.8

Part I – Scenario IV – Tone 3

	ta1	**hai2**	**xiang3**	**yao4**	**shen2**	**me5**	**jiu3**
1	2	5	16	7	15	57	43
2	53	65	26	16	18	10	1
3	19	2	6	1	1	11	36
score	25.8	21.5	14	6.7	8.9	19.9	26.9

Part I – Scenario IV – Tone 4

	ta1	**hai2**	**xiang3**	**yao4**	**shen2**	**me5**	**rou4**
1	8	5	21	13	16	65	24
2	52	54	21	3	14	3	23
3	21	10	4	1	0	5	37
score	27.7	22.2	12.5	4	7.4	16.4	30.2

Part II – Scenario I – Tone 1

	ta1	**hai2**	**xiang3**	**yao4**	**shen2**	**me5**	**zhou1**	**ma5**
1	0	4	35	9	14	18	41	40
2	55	53	19	10	6	0	3	6
3	35	21	23	2	0	0	39	38
score	34	27.2	24.2	5.8	4.6	3.6	28.6	28.8

Part II – Scenario I – Tone 2

	ta1	hai2	xiang3	yao4	shen2	me5	you2	ma5
1	3	0	25	3	9	13	83	53
2	57	61	30	24	11	9	0	2
3	25	26	31	3	1	0	0	28
score	30.2	31.3	29.5	9.3	5.6	5.3	16.6	25.2

Part II – Scenario I – Tone 3

	ta1	hai2	xiang3	yao4	shen2	me5	jiu3	ma5
1	2	0	15	0	9	18	71	82
2	63	71	26	26	14	29	0	0
3	21	14	33	0	2	10	14	1
score	29.8	28.3	27.3	7.8	7	17.3	21.2	16.9

Part II – Scenario I – Tone 4

	ta1	hai2	xiang3	yao4	shen2	me5	rou4	ma5
1	0	2	29	5	10	5	23	76
2	57	70	14	11	4	2	0	1
3	27	14	31	2	1	0	67	7
score	30.6	28.4	25.5	5.3	3.7	1.6	38.1	19

Part II – Scenario II – Tone 1

	ta1	hai2	xiang3	yao4	shen2	me5	zhou1	ba5
1	1	5	32	2	4	26	27	37
2	39	58	27	18	8	8	2	19
3	46	15	19	0	1	3	61	7
score	34.9	25.9	24	5.8	3.7	9.1	36.5	16.6

Part II – Scenario II – Tone 2

	ta1	hai2	xiang3	yao4	shen2	me5	you2	ba5
1	4	1	24	1	5	6	82	43
2	41	61	35	28	18	15	0	12

3	43	20	21	1	0	3	7	6
score	34.6	28.5	25.8	9.1	6.4	7.2	19.9	15.2

Part II – Scenario II – Tone 3

	ta1	**hai2**	**xiang3**	**yao4**	**shen2**	**me5**	**jiu3**	**ba5**
1	9	5	19	0	1	15	59	46
2	43	62	18	28	17	28	0	5
3	36	10	21	3	0	15	30	9
score	32.7	24.6	19.7	9.9	5.3	18.9	26.8	15.2

Part II – Scenario II – Tone 4

	ta1	**hai2**	**xiang3**	**yao4**	**shen2**	**me5**	**rou4**	**ba5**
1	6	2	33	4	5	6	16	29
2	37	60	16	20	9	14	3	14
3	42	16	20	2	0	2	70	10
score	33.3	26.4	21.4	7.8	3.7	6.4	39.1	15

Part II – Scenario III – Tone 1

	ta1	**hai2**	**xiang3**	**yao4**	**shen2**	**me5**	**zhou1**	**a5**
1	0	7	9	1	16	16	62	38
2	60	29	39	24	5	9	1	3
3	28	50	36	0	0	3	24	5
score	32	35.1	31.5	7.4	4.7	7.4	24.7	11

Part II – Scenario III – Tone 2

	ta1	**hai2**	**xiang3**	**yao4**	**shen2**	**me5**	**you2**	**a5**
1	1	7	6	1	20	8	74	48
2	59	34	41	42	10	14	0	5
3	26	47	38	0	1	3	8	5
score	30.9	35.1	32.5	12.8	7.5	7.3	18.8	13.6

Part II – Scenario III – Tone 3

	ta1	hai2	xiang3	yao4	shen2	me5	jiu3	a5
1	3	13	17	7	15	16	55	63
2	52	27	39	26	18	24	5	9
3	20	36	23	1	0	6	29	12
score	26.2	28.7	26.6	9.7	8.4	13.4	27	21.3

Part II – Scenario III – Tone 4

	ta1	hai2	xiang3	yao4	shen2	me5	rou4	a5
1	0	15	14	1	16	5	30	42
2	54	29	30	25	2	7	0	7
3	29	44	38	2	3	0	57	7
score	30.7	33.7	30.8	8.7	5.3	3.1	34.5	14

APPENDIX B

Experiments

Experiment I

Background: You are roommates with Luo and Wan. Today, you and Wan go for shopping in the supermarket. Luo can't go with you because she is busy, and she asked you two to bring something for her.

Scenario I

You and Wan have taken the stuff that Luo needs. At this time Luo texts Wan that she needs something more.

1.1

小万：小罗说她还要买粥。(Wan: Luo said she also wants some porridge.)

你：她还想要什么粥？(You: What kind of porridge does she also want?)

小万：燕麦粥。(Wan: Oat porridge.)

1.2

小万：小罗说她还要买油。(Wan: Luo said she also wants some oil.)

你：她还想要什么油？(You: What kind of oil does she also want?)

小万：橄榄油。(Wan: Olive oil).

1.3

小万：小罗说她还要买酒。(Wan: Luo said she also wants some wine.)

你：她还想要什么酒？(You: What kind of wine does she also want?)

小万：红酒。(Wan: Red wine.)

1.4

小万：小罗说她还要买肉。(Wan: Luo said she also wants some meat.)

你：她还想要什么肉？(You: What kind of meat does she also want?)

小万：牛肉。(Wan: Beef.)

Scenario II

Before you and Wan left home, Luo asked you to contact her when you are in the supermarket because she wanted to have a choose before deciding. Now you and Wan have finished your shopping, and are ready to buy the stuff that Luo needs.

2.1

小万：小罗要什么？(Wan: What does Luo want?)

你：她还想要什么粥。我问问。(You: She wants some porridge. Let me ask her.)

你：她说要燕麦粥。(You: She said she wants oat porridge.)

2.2

小万：小罗要什么？(Wan: What does Luo want?)

你：她还想要什么油。我问问。(You: She wants some oil. Let me ask her.)

你：她说要橄榄油。(You: She said she wants olive oil.)

2.3

小万：小罗要什么？(Wan: What does Luo want?)

你：她还想要什么酒。我问问。(You: She wants some wine. Let me ask her.)

你：她说要红酒。(You: She said she wants red wine.)

2.4

小万：小罗要什么？(Wan: What does Luo want?)

你：她还想要什么肉。我问问。(You: She wants some meat. Let me ask her.)

你：她说要牛肉。(You: She said she wants beef.)

Scenario III

Luo asked you to buy too much same food for her, and You and Wan think she shouldn't do that, because she won't be able to eat them all.

3.1 Luo said she wanted five bowls of oat porridge, and texts you now that she wants five more.

小万：小罗还要买五份燕麦粥！(Wan: Luo want five bowls of oat porridge more!)

你：她还想要什么粥？吃得完吗？(You: What porridge? Can she eat them all?)

3.2 Luo said she wanted five bottles of olive oil, and texts you now that she wants five more.

小万：小罗还要再买五瓶橄榄油！(Wan: Luo want five bottles of olive oil more!)

你：她还想要什么油？吃得完吗？(You: What oil? Can she eat them all?)

3.3 Luo said she wanted five bottles of red wine, and texts you now that she wants five more.

小万：小罗还要再买五瓶红酒！(Wan: Luo want five bottles of red wine more!)

你：她还想要什么酒？喝得完吗？(You: What wine? Can she eat them all?)

3.4 Luo said she wanted five pieces of beef, and texts you now that she wants five more.

小万：小罗还要再买五份牛肉！(Wan: Luo want five pieces of beef more!)

你：她还想要什么肉？吃得完吗？(You: What meat? Can she eat them all?)

Scenario IV

You and Wan have finished your shopping, and are ready to buy the stuff Luo wants. Before leaving home, Luo told you that she wanted porridge/oil/wine/meat, but asked you to contact her when in the supermarket because she wanted to choose before deciding. Now you are texting her asking what she wants while talking with Wan.

4.1

小万：小罗要什么？(What does Luo want?)

你：（你一边给小罗发消息，一边说）她还想要什么粥… (You: (texting) She says she wants some porridge…)

你：（小罗回复了）她说要燕麦粥。(You: (Luo replied) She says she wants oat porridge.)

4.2

小万：小罗要什么？(What does Luo want?)

你：（你一边给小罗发消息，一边说）她还想要什么油…(You: (texting) She says she wants some oil…)

你：（小罗回复了）她说要橄榄油。(You: (Luo replied) She says she wants olive oil.)

4.3

小万：小罗要什么？(What does Luo want?)

你：（你一边给小罗发消息，一边说）她还想要什么酒…(You: (texting) She says she wants some wine…)

你：（小罗回复了）她说要红酒。(You: (Luo replied) She says she wants red wine.)

4.4

小万：小罗要什么？(What does Luo want?)

你：（你一边给小罗发消息，一边说）她还想要什么肉…(You: (texting) She says she wants some meat…)

你：（小罗回复了）她说要牛肉。(You: (Luo replied) She says she wants beef.)

Experiment II

The background, the scenarios I to III and the conversations are exactly the same as Experiment I. The only differences are the sentences “You” says have modal particles “ma0” in Scenario I, “ba0” in Scenario II, and “a0” in Scenario III, but they do not change the meaning nor the context of the sentences:

Scenario I

你：她还想要什么粥/油/酒/肉吗？

Scenario II

你：她还想要什么粥/油/酒/肉吧？

Scenario III

你：她还想要什么粥/油/酒/肉啊？