Journal of Social Science Humanities and Literature, Vol.6, Issue 6, (Dec)
ISSN 2277-842X **2023**
DOI: 10.53469/jsshl.2023.06(06).39

A Comparative Study of Prosodic Cues to Turn-taking: Evidence from Perception

Wei Zhang

School of Foreign Languages, Shangrao Normal University, Shangrao, Jiangxi, China

Abstract: ***The objective of this study is to identify the disparities between individuals who speak English as their first language and those who are learning English as a foreign language, specifically in relation to the prosodic signals that indicate when it is appropriate to take turns in conversation. Data collection involved the implementation of two perceptual studies, one involving text and the other involving audio. The results indicate that there are notable distinctions between two groups of individuals in response to specific stimuli. Implications have been made about the instruction of English as a Foreign Language (EFL) learners.***

Keywords: prosodic cues, turn-taking, perception experiment, English teaching, comparative study.

1. INTRODUCTION

The study of turn-taking seeks to elucidate the process of initiating, maintaining, and concluding a discussion by relinquishing the floor to the listener. This aspect plays a pivotal part in the overall structure of a conversation (Sacks et al., 1974). The turn-taking process is associated with two models: the rule-based model and the signal-based model. The current study refers to the latter one, as proposed by Duncan (1972). The signal-based paradigm primarily emphasizes non-linguistic characteristics that are presented, such as prosody. In the present era, the progress in speech technology through machine learning allows for the modelling and integration of the prosodic part of turn-taking in IVR (Interactive Voice Response) systems. The majority of research on prosodic cues for turn-taking has been conducted through production studies (Caspers, 2001; Gravano & Hirschberg, 2011, 2012; Gravano et al., 2017;Koiso et al., 1998). The current study is inspired by research conducted by Wichmann and Caspers (2001), in which they devised two perceptual experiments to examine the melodic signals in turn-taking. The reported findings suggest that when an utterance is syntactically incomplete, individuals tend to interpret it as a signal to maintain their turn in the conversation, regardless of the melodic contour. However, if the statement is grammatically full, individuals often interpret it as a shift in conversation (giving the floor), except when the intonation is a high-level tone (H*%). In this case, maintaining control of the conversation (keeping the floor) is favored, regardless of grammatical completion.

The current study conducted a perceptual experiment to examine the utilization of prosodic cues between native English speakers and English as a Foreign Language (EFL) learners, with the objective of identifying the disparities between them.

2. 2.METHODS

2.1 Subjects

A total of 30 subjects, consisting of 15 males and 15 females, were selected at random from the student population of Shangrao Normal University. All of these students have successfully passed both the CET-6 (College English Test Band 6) and TEM4 (Test for English Majors Band 4). Furthermore, they have obtained a National Mandarin Test Certificate at level 2, achieving a grade of A or higher. This ensures that their spoken Mandarin is as standardized as possible, without any noticeable impact from their dialects. None of them possess the experience of residing or journeying abroad.

2.2 Stimuli

Participants received 20 declarative statements that were syntactically finished. These sentences were taken from dialogues read by a male Native English speaker. Each statement is duplicated, once in a turn-medial position and once in a turn-final position, in the original conversations. Thus, there are precisely 10 distinct utterances in terms of orthography.

The speech was delivered in its original audio format. The retrieved utterances were consolidated into a single audio file, with a 5-second delay between each pair.

2.3 Procedure

The perception experiment consists of two sequential stages: text judgment (as described in Appendix 1) and audio judgment. The participants were provided with a questionnaire containing ten concise statements. They were instructed to evaluate each statement on a scale ranging from 1 to 5, where 1 represents "definitely requires further elaboration," 2 represents "likely requires further elaboration," 3 represents "could be either ongoing or completed," 4 represents "likely completed," and 5 represents "definitely completed."

The purpose of this stage is to ensure that the subjects' decisions are influenced by the auditory information rather than the textual information. An optimal outcome would entail an average rating of 3 for every sentence. However, it is inevitable that there will be a tiny inclination towards either the middle ratings or the extreme end of the spectrum. Hence, I anticipated that the average ratings would hover around 3, slightly surpassing or falling short of this value. Regarding the second stage, a total of 20 utterances were individually provided to the individuals in three groups. The utterances were randomly ordered and there was a 5-second interval between each presentation. If deemed essential, they have the option to initiate a fresh start. However, during the listening process, they are unable to interrupt the audio playback. They indicated their decision by marking either HOLD or CHANGE on their sheets.

3. RESULTS

3.1 Text judgement

All the sentences in this task are given as follows:

① *I couldn't get it done.*
② *I'm still not fluent in Chinese.*
③ *It was a surprise party.*
④ *I've ever had in all my life.*
⑤ *And they just ask questions.*
⑥ *It just threw me off guard.*
⑦ *There is mutual respect between students and teachers.*
⑧ *They also respect the teachers' rules.*
⑨ *It was nice to meet somebody from San Francisco.*
⑩ *We didn't visit friends.*

Table 1 below shows the descriptive statistics of the text judgement by the subjects in three groups (30 English native speakers participated in the perception experiment, but 6 sheets turned out to be invalid).

Table 1: Descriptive statistics of text judgement

Statistics

Level			T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
advanced	N	Valid	30	30	30	30	30	30	30	30	30	30
		Missing	0	0	0	0	0	0	0	0	0	0
	Mean		2.93	3.13	2.67	3.33	2.53	2.97	2.60	3.23	2.87	2.60
beginner	N	Valid	30	30	30	30	30	30	30	30	30	30
		Missing	0	0	0	0	0	0	0	0	0	0
	Mean		1.93	2.87	3.17	2.73	2.93	2.83	3.17	3.00	2.97	2.67
native	N	Valid	24	24	24	24	24	24	24	24	24	24
		Missing	0	0	0	0	0	0	0	0	0	0
	Mean		3.50	3.54	3.00	2.79	2.63	3.42	3.25	3.04	3.42	2.88

The means of the text judgement in the Advanced group range from 2.53 to 3.33, with the final mean rating of 2.886. The means in the Beginner group range from 1.93 to 3.17, with the final mean rating of 2.827. The means in the Native group range from 2.79 to 3.54, with a final rating of 3.147.

It was expected that the ideal result to be a mean rating of 3 for each sentence. But considering that there would be a slight bias towards either medial ratings or the final end of the scale. Thus, the final mean ratings obtained from the text judgment are acceptable. In this case, it may be confident to say that no intrinsic bias was confounding the effects of the auditory cues.

The one-way ANOVA results show that the factor of the group has a significant main effect on the text judgement, $F(2, 837) = 4.012$, $p = .018$. The LSD *Post hoc* tests show that English native speakers differ significantly from the two Chinese English learner groups.

3.2 Audio judgement

The one-way ANOVA results show that the factor of the group has no significant main effect on the audio judgement, $p > .05$. Therefore, another one-way ANOVA was conducted on each of the sentences in the audio stimuli. The results show that 7 out of 20 sentences display significant differences within groups.

Then the analysis is to be conducted on those significant ones in sequence.

3.3 Analysis of the sentences with a significant difference

(1) S5: *I'm still not fluent in Chinese*.

The descriptive statistics (Table 2) and pitch contour of S5 (Figure 2) are given below.

Table 2: Frequency of S5 across groups

S5

Level			Frequency	Percent	Valid Percent	Cumulative Percent
advanced	Valid	H	15	50.0	50.0	50.0
		C	15	50.0	50.0	100.0
		Total	30	100.0	100.0	
beginner	Valid	H	14	46.7	46.7	46.7
		C	16	53.3	53.3	100.0
		Total	30	100.0	100.0	
native	Valid	H	20	83.3	83.3	83.3
		C	4	16.7	16.7	100.0
		Total	24	100.0	100.0	

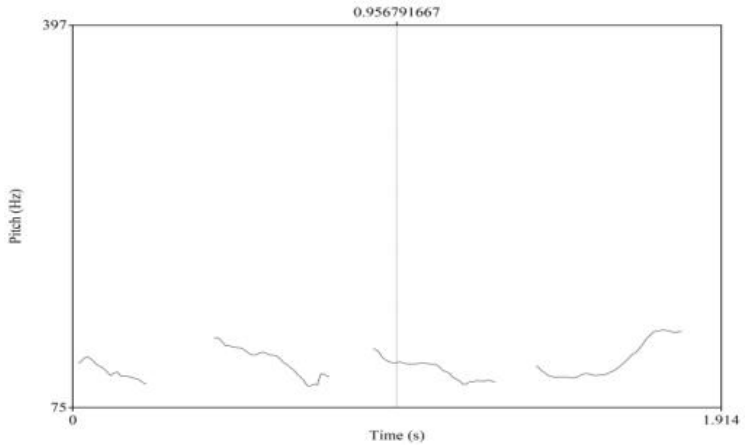

Figure 1: The pitch contour of S5

The native speakers' audio judgement of this sentence is in line with the one in text judgement, in which they gave it the highest scale as a turn-medial one. This judgement turns out to accord with the original dialogue. As is shown in the pitch contour in Figure 1, it is obvious that the sentence is a declarative one with a final rising tone, indicating the current speaker has something more to say. However, in judging this sentence, the Chinese learner groups display almost a balance between the choice of turn-medial or turn-final. That's to say, the EFL learners are not familiar with this tone: L-H% (low phrase tone with a high boundary tone). They have always been told that a rising tone goes with the interrogative sentence. However, this sentence is a declarative, thus making the ELF

learners confused. This can also be seen in their production, they seldom use this tone as a prosodic means to hold the turn.

(2) S6: *I've ever had in all my life*.

The descriptive statistics (Table 3) and pitch contour of S6 (Figure 2) are given below.

Table 3: Frequency of S6 across groups

S6

Level			Frequency	Percent	Valid Percent	Cumulative Percent
advanced	Valid	H	7	23.3	23.3	23.3
		C	23	76.7	76.7	100.0
		Total	30	100.0	100.0	
beginner	Valid	H	13	43.3	43.3	43.3
		C	17	56.7	56.7	100.0
		Total	30	100.0	100.0	
native	Valid	H	2	8.3	8.3	8.3
		C	22	91.7	91.7	100.0
		Total	24	100.0	100.0	

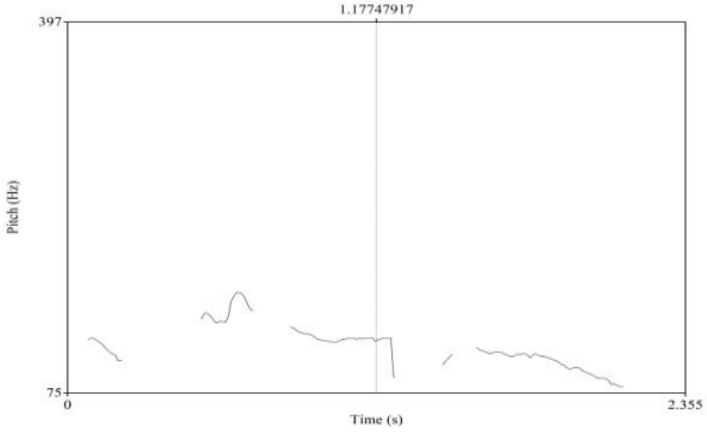

Figure 2: The pitch contour of S6

It can be seen from above Table 3 that native speakers show a significant difference: 91.7% vs. 8.3%, followed by the advanced learner group: 76.7% vs. 23.3%. In contrast, the beginner learner group kept almost a balance here. It can be seen in Figure 2 that an obvious falling tone (L-L%) is there, which is considered an indication of turn-yielding, i.e. the change of the turn. The results of the contrast show us that the beginner group are not clear about the use of this tone, although they display it often while speaking English. They are not aware a falling low tone indicates a turn change. One possible explanation is that they speak English as how they speak Chinese, without noticing any prosodic features.

(3) S7: *There is mutual respect between students and teachers*.

The descriptive statistics (Table 4) and the pitch contour of S7 (Figure 3) are given below.

Table 4: Frequency of S7 across groups

S7

Level			Frequency	Percent	Valid Percent	Cumulative Percent
advanced	Valid	H	16	53.3	53.3	53.3
		C	14	46.7	46.7	100.0
		Total	30	100.0	100.0	
beginner	Valid	H	18	60.0	60.0	60.0
		C	12	40.0	40.0	100.0
		Total	30	100.0	100.0	
native	Valid	H	1	4.2	4.2	4.2
		C	23	95.8	95.8	100.0
		Total	24	100.0	100.0	

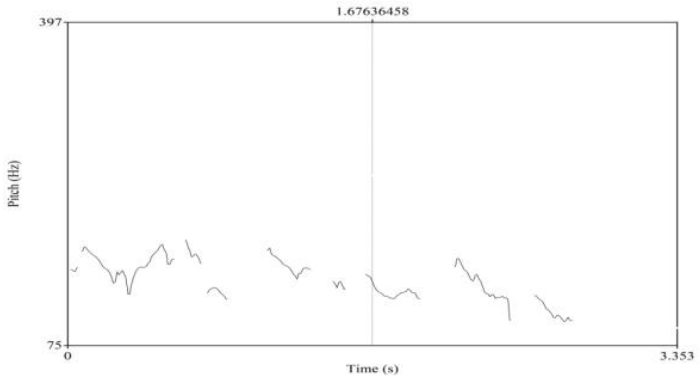

Figure 3: The pitch contour of S7

It can be seen from above Table 4 that native speakers show a significant difference: 95.8% vs. 4.2%. As is shown in Figure 3 this is a relatively long utterance compared with other ones, which ends with a low tone (L-L%). A low tone is always associated with a turn change. However, both the learner groups tend to take it as a turn-medial one, especially the beginner group. It may have something to do with the length of the utterance, besides their unfamiliarity with this tone used in turn taking, which leads to the confusion of the EFL learners.

(4) S10: *We didn't visit friends*.

The descriptive statistics (Table 5) and the pitch contour (Figure 4) of S10 are given as follows.

Table 5: Frequency of S10 across groups

S10

Level			Frequency	Percent	Valid Percent	Cumulative Percent
advanced	Valid	H	14	46.7	46.7	46.7
		C	16	53.3	53.3	100.0
		Total	30	100.0	100.0	
beginner	Valid	H	15	50.0	50.0	50.0
		C	15	50.0	50.0	100.0
		Total	30	100.0	100.0	
native	Valid	H	23	95.8	95.8	95.8
		C	1	4.2	4.2	100.0
		Total	24	100.0	100.0	

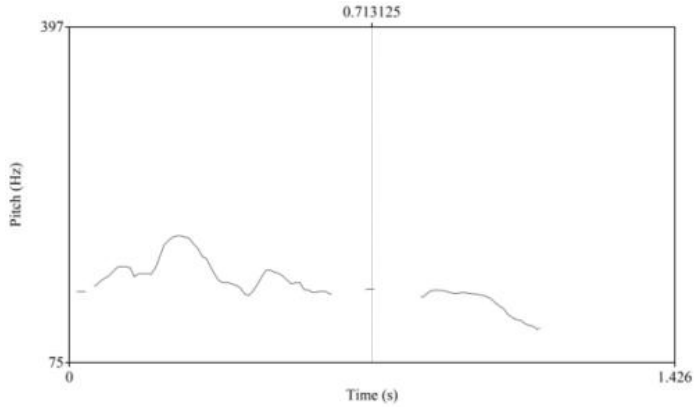

Figure 4: The pitch contour of S10

It can be seen from the above table that native speakers show a significant difference: 95.8% vs. 4.2%. In contrast, the two Chinese learner groups show roughly the same case, especially the beginner group, their choices show an exact balance. As is shown in Figure 4, a downward-sloping pitch contour subsides before reaching the bottom of the speaker's range. This tone is classified as partial fall (pL-L%) in this study. Based on the results of production by EFL learners, it is quite rare in their speech. Thus, there is no doubt they have difficulty in making a choice between Hold and Change based on the auditory cues.

(5) S12: *It just threw me off guar*d.

The descriptive statistics (Table 6) and the pitch contour (Figure 5) of S12 are given as follows.

Table 6: Frequency of S12 across groups

S12

Level			Frequency	Percent	Valid Percent	Cumulative Percent
advanced	Valid	H	19	63.3	63.3	63.3
		C	11	36.7	36.7	100.0
		Total	30	100.0	100.0	
beginner	Valid	H	10	33.3	33.3	33.3
		C	20	66.7	66.7	100.0
		Total	30	100.0	100.0	
native	Valid	H	17	70.8	70.8	70.8
		C	7	29.2	29.2	100.0
		Total	24	100.0	100.0	

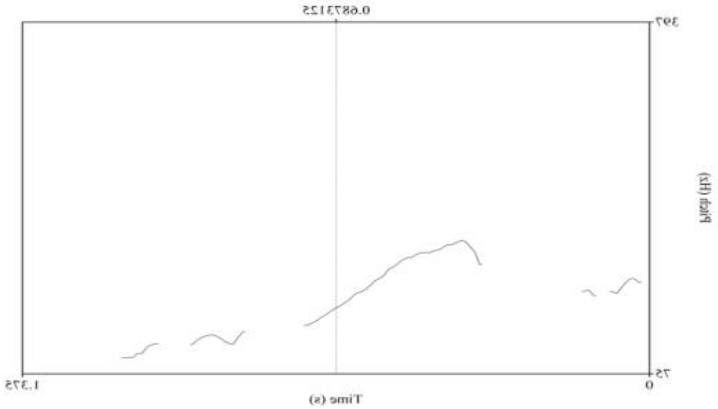

Figure 5: The pitch contour of S12

It can be seen from above Table 6 that the native speakers are significantly more inclined to take it as a turn-medial one, which in fact belongs to the turn-final. Besides, another striking result lies in that the two learner groups show the opposite results. By listening to the audio file, we can hear the speaker speaking quite fast, with a falling tone, yet not quite clear enough. However, as is seen in Figure 5 above, there is a clear indication of downward sloping pitch contour. Therefore, the speaking rate may have an influence on the listeners' perception. And that might be the reason that leads to a sharp contrast between the two learner groups. Speaking rate tends to influence the perception. Furthermore, with the pitch contour presented, the tone is clearly seen, so such visual aids should contribute to the perception with regard to the prosodic features. Therefore, the continuous imitation of native speakers' speech with the visual aids of pitch or intonation contours is strongly suggested.

(6) S14: *There is mutual respect between students and teachers.*

The descriptive statistics (Table 7) and the pitch contour of S14 (Figure 6) are given as follows.

Table 7: Frequency of S14 across groups

S14

Level			Frequency	Percent	Valid Percent	Cumulative Percent
advanced	Valid	H	8	26.7	26.7	26.7
		C	22	73.3	73.3	100.0
		Total	30	100.0	100.0	
beginner	Valid	H	16	53.3	53.3	53.3
		C	14	46.7	46.7	100.0
		Total	30	100.0	100.0	
native	Valid	H	17	70.8	70.8	70.8
		C	7	29.2	29.2	100.0
		Total	24	100.0	100.0	

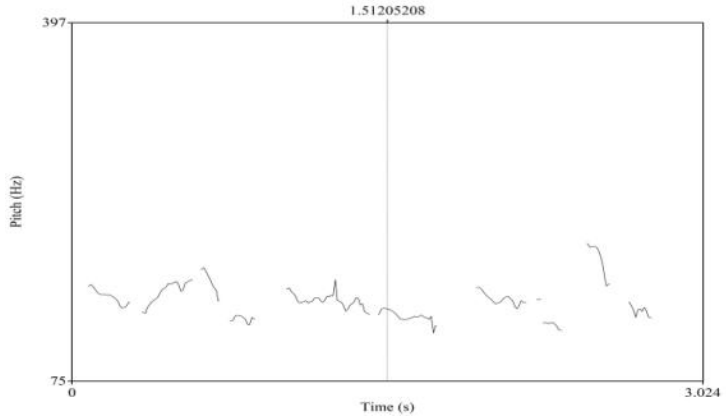

Figure 6: The pitch contour of S14

It can be seen that a sharp contrast can be found between the advanced and the native groups. These two groups show completely different choices. Although the utterance is the same as that in S7 literally, this is a different case with respect to prosodic features. As can be seen in Figure 6, the pitch stays flat at a fairly high level, through the end of the word, forming a "plateau". Plateau, rather than an obvious rising or falling, which indicates the speaker has more to say. However, as for the learner groups, quite different choices have been made. This might be attributed to the function of the plateau tone which the learners are not familiar with.

(7) S15: *I couldn't get it done.*

The descriptive statistics (Table 8) and the pitch contour (Figure 7) of S15 are given as follows.

Table 8: Frequency of S15 across groups

S15

Level			Frequency	Percent	Valid Percent	Cumulative Percent
advanced	Valid	H	13	43.3	43.3	43.3
		C	17	56.7	56.7	100.0
		Total	30	100.0	100.0	
beginner	Valid	H	13	43.3	43.3	43.3
		C	17	56.7	56.7	100.0
		Total	30	100.0	100.0	
native	Valid	H	19	79.2	79.2	79.2
		C	5	20.8	20.8	100.0
		Total	24	100.0	100.0	

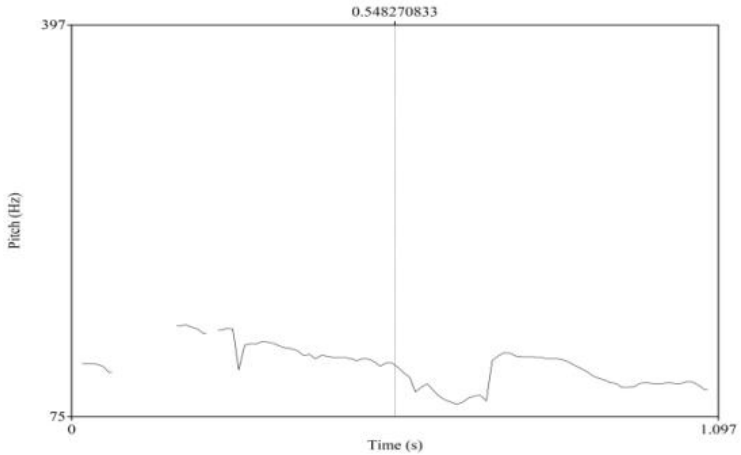

Figure 7: The pitch contour of S15

It is interesting to find that two learner groups made the same choices. They tend to take it as a turn-medial one. In contrast, the native speakers' choices are significantly different. As is shown in Figure 7, the pitch contour shows a tendency downward, without approaching the bottom of the speaker's pitch range, which as mentioned in the

former section, is a partial Fall (pL-L%). It is a tone indicating turn hold. Such kind of tone may not be so noticeable for learners, thus leading to difficulties for the learners in making their choices about whether the speaker is finished or not.

4. DISCUSSION

4.1 Major findings

The results from the perception experiment indicate that native speakers do not exhibit substantial variations in their audio judgments. The feedback from native English speakers has suggested such outcomes. The majority of individuals encountered challenges when making their decisions. The response is heavily influenced by the transient nature of intonation, making it challenging to grasp, even for native speakers, let alone second language learners.

Selected sentences of utmost significance have been chosen for further examination. The choices made by the native group differ significantly from those made by the two learner groups. However, it is worth noting that the learner groups sometimes make completely different choices on certain sentences, suggesting that they are confused about the use of prosodic features in turn-taking. Regarding the indigenous community, people primarily rely on their instincts when making decisions. For EFL learners who are not aware of L2 intonation, it is difficult for them to understand the implied meaning of intonation without sufficient training.

The findings of the perception experiment indicate that participants consistently opt for either maintaining their current option or altering it when a clear auditory signal is present, regardless of whether it is ascending, descending, or steady. The greater the distinctiveness of the tone, the more pronounced the disparity in choice will be. This is particularly true for individuals who are native speakers. Regarding them, the intonation has become instinctive and inherent to the message. Although individuals may not always be conscious of it, they are able to perceive something when it is sufficiently noticeable or when they direct their attention towards it. However, when it comes to English learners who are studying it as a foreign language, the focus has primarily been on aspects of language other than intonation. Despite their awareness of the noticeable intonation, individuals may nonetheless be confusion over its pragmatic significance. Despite the recognition of intonation as a crucial aspect of language acquisition and communication in recent decades, it remains one of the most overlooked areas of linguistics in classroom instruction (Verdugo, 2006).

Like many English as a Foreign Language (EFL) learners, they lack the opportunity to engage in English conversation on a daily basis due to their limited language context. Typically, English is the sole language used in classes. In other words, the people they are speaking with are also Chinese. Having a shared cultural background facilitates mutual understanding between them. The issue, therefore, occurs when they encounter individuals who are fluent in English from birth. The obstacles extend beyond mere vocabulary or grammar. Even proficient learners may encounter difficulties while communicating with native English speakers. Frequently, EFL learners express dissatisfaction with the speaking speed of native speakers. They often perceive those factors as the underlying reasons for miscommunication or communication breakdown. EFL learners exhibit a notable characteristic of speaking English with a Chinese accent. In contrast to English, suprasegmental cues such as pitch and loudness have a significant impact on word recognition. However, when non-native English speakers employ this approach in their English conversations, it often perplexes or offends native English speakers, resulting in communication breakdowns or failures.

4.2 Pedagogical implications:

(1) Raise the awareness

According to Wells (2006), it is recommended to be understanding of mistakes in individual sounds while communicating with non-native English speakers. However, errors in intonation should not be overlooked, as stated by Jiang (2011). Native speakers prioritize the correct utilization of intonation for efficient communication. Hence, the initial step is to assist learners in recognizing the significance of prosody and intonation. As an illustration, I can construct a statement using distinct prosodic characteristics, such as diverse pitch, ascending, descending, or steady tone, elongating the duration of the utterance, varying the speaking speed, etc. Subsequently, I can prompt learners to discern the distinctions in addition to the intended significance. Subsequently, elucidates the concept of prosodic traits and their potential impacts on everyday communication, particularly in the context of

cross-cultural interactions. Simply said, learners are unlikely to detect these prosodic traits unless they are specifically made aware of them.

Prosodic qualities are frequently conveyed implicitly in foreign language instruction. Typically, the prosodic qualities are neither categorized nor are their meanings elucidated. The inadequate familiarity of a foreign language teacher with the pertinent prosodic characteristics could be a significant factor contributing to the lack of success.

(2) From CALL (Computer-Assisted Language Teaching) to EVF (Electronic Visual Feedback)

The practical applications of CALL are growing at such a rapid pace that it is almost impossible for a classroom teacher to keep up with the field (Brown, 2007). Nowadays, classrooms are almost all equipped with multimedia. But will it lead to happy and successful learning? The answer I think remains uncertain. A good and updated use of such aids will be of great help. As for the teaching of prosody, I suggest learning visually and watching people speak. Traditionally, we ask learners to listen carefully for the content of what the speaker says, whether it is in an audio file or a video file. Learning visually as well as watching people speak is far more than getting the meaning of what people say. Visual feedback for teaching tone and intonation began more than forty years ago (Anderson, 1992, 1994; Chun, 1998; Verdugo, 2006), but with the development of speech technology and the popularity of personal computers, it has become accessible both in the classroom and at home. PRAAT proposed by Boersma and Weenink (2001) is popular and easily available online, so my suggestion is that both teachers and learners can use it to provide auditory and visual displays of pitch contours as feedback. By means of acoustic analysis with the software, we can compare the features of learners' speech with those of native speakers. With the visual differences, then we can give learners corresponding training. We always say practice makes perfect, but with proper training, we can make the practice more effective and efficient.

One of the benefits of EVF is that it provides the students with an accurate visual representation of suprasegmentals in real-time paired with the normal auditory feedback that occurs during speech. Students can thus more easily replicate native suprasegmental targets using both the target form and the visual feedback from their own speech to guide them (Chun, 1998). Although the technology and emerging software are continually and rapidly improving, some aspects are still limited when it comes to analyzing natural speech. Especially, in authentic, spontaneous conversation, the prosody is also not easy to get fully understood. However, one definite advantage of the application of EVF is that it can facilitate the L2 prosody acquisition.

5. CONCLUSION

By analyzing the seven most prominent sentences selected from a pool of 20 stimuli used in the audio judgment, it is evident that native speakers consistently choose between "Hold" and "Change" when there is a clear distinction in tone, whether it is rising, declining, or plateauing. Chinese EFL learners, regardless of their proficiency level, frequently exhibit uncertainty when it comes to making the appropriate selection. The neglect of suprasegmentals in English teaching may be attributed to the predominant emphasis on segmental aspects. Consequently, recommendations have been provided for instructional methods based on the findings of the research. These findings and their consequences are expected to be beneficial for effective communication.

ACKNOWLEDGEMENT

This article is a phased result of the 2022 project of Jiangxi Province's "14th Five-Year Plan" for Educational Science "Research on the Fluency of Spoken English Production from the Perspective of Prosody" (No.: 22YB203) and a phased result of Jiangxi Province's College and Universities Humanities and Social Sciences Research, 2022 annual Planning Project "Research on the Development of Academic Discourse Ability of Undergraduates Majoring in English Majors in Normal Colleges" (No.: YY22116).

REFERENCES

[1] Anderson, H. J. (1992). Using electronic visual feedback to teach suprasegmentals. System, 20 (1), 51–62.
[2] Anderson, H. J. (1994). Interpreting visual feedback on suprasegmentals in computer-assisted pronunciation instruction. CALICO Journal, 11 (4), 5–22.
[3] Boersma. P., & Weenink, D., 2001. Praat: Doing phonetics by computer, http://www.fon.hum.uva.nl/praat/

[4] Brown, H. D. (2007). Teaching by principles: an interactive approach to language pedagogy (3rd Edition). London: Pearson Longman.
[5] Brusco, P., Perez, J., & Gravano, A. (2017). Cross-linguistic study of the production of turn-taking cues in American English and Argentine Spanish. Interspeech.
[6] Chun, D. M. (1998). Signal analysis software for teaching discourse intonation.Language Learning and Technology, (2), 61–77.
[7] Duncan, S. (1972). Some Signals and Rules for Taking Speaking Turns in Conversations. Journal of personality and Social Psychology, 23, 283-292.
[8] Gravano, A., & Hirschberg, J. (2011). Turn-taking cues in task-oriented dialogue.Computer Speech and Language, 25, 601-634.
[9] Gravano, A., & Hirschberg, J. (2012). A corpus-based study of interruptions in spoken dialogue. INTERSPEECH 2012, 854-857.
[10] Gravano, A., Brusco, P., & Benus, S. (2016). Who do you think will speak next? Perception of turn-taking cues in Slovak and Argentine Spanish. INTERSPEECH 2016, 1265-1269.
[11] Jiang, H. (2011). Gender difference in English intonation. International Congress of Phonetic Sciences, 974–977.
[12] Koiso, H., Horiuchi, Y., Tutiya, S., Ichikawa, A., & Den, Y., (1998). An analysis of turn-taking and backchannels based on prosodic and syntactic features in Japanese Map Task dialogs. Language and Speech, 41, 295–321.
[13] Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). Linguistic society of America. Language, 50(1), 696-735.
[14] Verdugo, D. R. (2006). A study of intonation awareness and learning in non-native speakers of English. Language Awareness, 15 (3), 141-159.
[15] Wells, J. C. (2006). English intonation PB and Audio CD: An introduction. Cambridge University Press.
[16] Wichmann, A., & Caspers, J. (2001). Melodic cues to turn-taking in English: evidence from perception. Proceedings of the Second SIGDIAL Workshop on Discourse and Dialogue.