Interspeech 2025
17-21 August 2025, Rotterdam, The Netherlands

# The Development of Speech Rhythm in Putonghua-Learning Preschool Children in South Xinjiang Uyghur Autonomous Region of China

*Aijun Li*[*1,2], *Zhiwei Wang*[3], *Jun Gao*[1], *Xin Zhou*[4]

[1]Key Laboratory of Linguistics, Chinese Academy of Social Sciences , China
[2]Corpus and Computational Linguistics Center, Chinese Academy of Social Sciences, China
[3]University of Chinese Academy of Social Sciences, China,
[4]Beijing International Studies University, China,

liaj@cass.org.cn, wangzhiwei1@ucass.edu.cn, gao-jun@cass.org.cn,zhouxin@bisu.edu.cn

## Abstract

This study investigates Putonghua rhythm development among Uyghur preschoolers in South Xinjiang, China. We examined 40 children (ages 5-7) grouped by Putonghua learning time: mid-class (12 months, mean age 6;1) and senior-class (24 months, mean age 7;1). Speech samples were elicited through picture-guessing tasks and analyzed using rhythmic metrics. Our analyses reveal distinct developmental patterns compared to Putonghua-speaking children, with neither group achieving metrics comparable to age-matched Putonghua-speaking 6-year-olds. Consonant-related metrics emerged as increasingly salient markers in later developmental stages. The findings demonstrate that rhythm acquisition in this bilingual context is modulated by cross-linguistic influence from Uyghur. These results advance our understanding of L2 rhythm acquisition in bilingual settings and have implications for developing automated speech assessment tools and optimizing speech recognition systems for multilingual contexts.

**Index Terms**: speech rhythm, Putonghua acquisition, Uyghur children, preschool children

## 1. Introduction

Rhythm, a regular perceptual pattern of prominent elements in speech or poetry, is an important aspect of prosody in speech interaction. Speech rhythm varies by language or dialect, speaking style or genre, speaker's personal traits, communicative intentions, or emotions, and so on. It was previously believed to be mastered by children at a young age like other prosodic aspects such as intonational contours, phrasing, and rhythm. This belief stemmed from the observation that infants consistently adjust pitch, duration, and intensity in their prelingual vocalizations to effectively respond to adult speech patterns. Infants not only can detect the rhythmic patterns of their native language but also possess the capacity to perceive related rhythmic categories from an early age. This ability varies among infants exposed to different languages. By 9 months of age, Spanish-learning infants can recognize the stress patterns of their native language, particularly when listening to "wug" words. This demonstrates their sensitivity to contrastive lexical stress and their ability to detect changes in stress patterns even without lexical knowledge [1]. French-learning infants, on the other hand, show a preference for the rhythmic structure of their native language as early as 6 months, especially when processing stress information. At the same age, German-learning infants exhibit a language-specific iambic-trochaic bias, a pattern that is less evident in French-learning infants [2].

In contrast to the early acquisition of language-specific intonation, children do not produce fully adult-like rhythm patterns in terms of interval-based metrics until the age of 4 or 5 years for "syllable-timed" languages, and not until after the age of 5 years for "stress-timed" languages [3, 4]. English-speaking children reach the adult-like rhythmic target by 11 to 12 years [5]. Infants' rhythm perception and early speech production are not synchronized. Initially, rhythm patterns resembling those of adults are acquired on a word-by-word basis [6]. Children's speech production tends to be more "syllable-timed" [7]. As they grow, the stress-timed nature of their speech rhythm becomes increasingly evident [8, 9]. By the age of 2, when children can already produce familiar words, they approximate adult rhythm patterns by adjusting the pitch, intensity, and duration of these words [10, 11].

In our previous study [12], for Putonghua (commonly referred to as Mandarin Chinese)-speaking children, the rhythm development pattern would not reach maturity until age 6-7. A comparison of rhythm metrics in child-directed speech and children across age of 3-7 reveals significant differences related to variances in consonant-related metrics, indicating that consonants (initials) are produced with greater development in preschool children. This result also supports the investigation of 4000 Putonghua-speaking children's word production [13]. As children age, their vowel (final)-related metrics contribute more to rhythm. For Chinese, the effect of duration variability of finals on rhythm gradually increases.

The first language of Uyghur children in South Xinjiang Uyghur Autonomous Region of China (south Xinjiang) is Uyghur. As a Turkic language, it belongs to the larger Altaic family. It's belieived that Uyghur rhythm is predominantly stress-timed, softened by its agglutinative structure and syllabic simplicity. This hybridity places it on a continuum between stress- and syllable-timed patterns, a characteristic typical of many Turkic languages. The acquisition of Putonghua among Uyghur children in South Xinjiang is relatively complex. Before entering kindergarten, they have been exposed to Putonghua, the national common language, and are not entirely unfamiliar second-language learners. Although existing research on south Xinjiang preschoolers has predominantly examined segment and tone [14, 15], suprasegmental features remain understudied, which are related to their verbal interaction ability and cognitive maturation. We attempt to address the following questions: What is the rhythmic pattern of Putonghua produced by Uyghur children in South Xinjiang? How do these patterns differ from those of Putonghua-speaking children of the same age?

*Corresponding author


10.21437/Interspeech.2025-1847

## 2. Participants and Materials

### 2.1. Participants

The recruited participants included 40 children (20 boys and 20 girls) from the mid and senior class of a kindergarten in south Xinjiang Uyghur Autonomous Region of China. None of the children had a history of hearing impairment or intellectual disabilities. The children started learning Putonghua (PTH) after entering kindergarten and have learned for about 12 or 24 months. The 40 children were divided into two age groups as seen in Table 1.

Table 1: *Demographic information of the children.*

| Group | Age of Children | Children | PTH Learning Time |
|---|---|---|---|
| G1 | 5;6-6;7, Mean: 6;1 | 20 (10F 10M) | 12m |
| G2 | 5;11-7;7, Mean: 7;1 | 20 (10F 10M) | 24m |
| Total | | 40 (20F 20M) | |

### 2.2. Date Collecting

The target sequence is a five-syllable SVO phrase expressed as Adjective 1 (A1) + Noun 1 (N1) + Verb (V) + Adjective 2 (A2) + Noun 2 (N2), for example, HēiA1 māoN1 qīnV xiǎoA2 yāN2 ("The black cat kisses the little duck."). One contrastive focus is on A1, V, A2, or N2 in one condition by the experimental design, i.e., the sentence-initial, -medial, or -final positions. The focus-bearing syllables at each position cover four lexical tones in Mandarin Chinese.

A picture-guessing game is designed to simulate the conversational interaction between children and the experimenter. Two pictures differed in only one element, that is, contrasting in one word in the target sequence were shown on the experimenter's screen. All the audio samples were pre-recorded by a native Mandarin-speaking adult female. The experimenter was asked to choose one of the pictures and play the corresponding of a question to guess if this was the picture on the child's screen, and the child was asked to answer whether the experimenter's guess was correct. Therefore, the participants might embed the target sequence in a yes/no question or declarative sentence. After that, it was the child's turn to guess and ask and the experimenter's turn to answer about the same elements. In this turn, the picture on the experimenter's screen was the same as one of the pictures on the child's screen.

Given the research design, the experimenter's guess (the sounds played by the experimenter) would always be wrong so as to prompt children's production of corrective foci. The child had a 50% chance of guessing correctly, and the experimenter would play a declarative sentence with a contrastive focus when the child's guess is wrong, and one with a broad/neutral focus when it is correct.

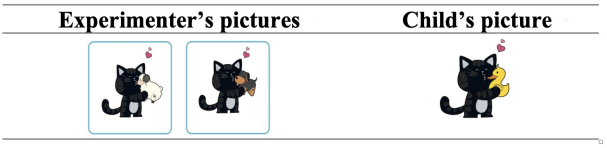


Figure 1: *Schematic diagram of the picture-guessing game.*

### 2.3. Data Analysis

The five-syllable target sequences were extracted from the child productions, yielding 2256 target utterances whose detailed information along with the participant's demographic information is given in Table 1. The number of valid pronunciation samples produced by middle-class children was 1124, whereas senior class children produced 1132. The implementation of this procedure ensured a uniform structure among participants, thereby facilitating the subsequent statistical analysis.

The automatic segmentation, manual annotation, and extraction of temporal metrics were conducted in Praat [16]. The rhythm metrics measured based on the extracted metrics are listed in Table 2, with citations to established research, where C and V respectively refer to the initials and finals of Chinese syllables.

Table 2: *Rhythm metrics measured in this study.*

| Rhythm metrics | Definition/calculation |
|---|---|
| Interval measures (IM) [5] | |
| PVI_V / PVI_C | Pairwise variability index (pvi) [5] |
| mean_V / mean_C | |
| percent_V / percent_C | Vowel (V) sequence duration ratio [17, 18] |
| delta_V / delta_C | Standard deviation of durations for vowel and consonant sequences [17, 18] |
| Varco_C / Varco_V | Percent standard deviation of consonantal and vocalic durations, each normalised by dividing by the mean [19] |
| nPVI | Normalized PVI [20] |
| rPVI | Non-normalized raw PVI [20] |
| speech rate (SR) | Syllable numbers /total duration of the utterance including silent pauses |
| articulatory rate (AR) | Syllable numbers/total duration of the utterance excluding silent pauses |

## 3. Results

### 3.1. Rhythm metrics distribution patterns

The Mann-Whitney U test was employed to analyze rhythm metrics between two groups of preschool children from South Xinjiang: Mid-Class (ages 5;6-6;7, mean: 6;1, Putonghua Learning Time: 12 m) and Senior-Class (ages 5;11-7;7, mean: 7;1, Putonghua Learning Time: 24 m), as illustrated in Figure 2.

For two groups of children from South Xinjiang, rhythm metrics related to vowel duration (e.g., sum_V, mean_V, delta_V, percent_V) and variation (e.g., varco_V, nPVI_V, rPVI_V) showed decreasing trends with increasing age and longer Putonghua learning. Similarly, consonant duration metrics (e.g., mean_C, delta_C) and variation metrics (e.g., varco_C, nPVI_C, rPVI_C), except for sum_C, also declined. Meanwhile, speech rate (SR) and articulation rate (AR), reflecting fluency, improved significantly in senior-class children with age and extended learning time.

From mid to senior class, we observed the following de-

velopmental patterns: vowel duration and variability decreased while rhythmic regularity increased, accompanied by reduced vowel dominance in speech rhythm. Consonant-related metrics showed increased duration and rhythmic complexity, yet decreased variability, resulting in more stable consonant timing and enhanced rhythmic regularity. Additionally, both SR and AR improved significantly, indicating enhanced fluency and rhythmic control.

To examine rhythmic production differences, we compared our data from South Xinjiang children with previously reported metrics from Putonghua-speaking children (ages 3-6) [12], as shown in Figure 2. Notably, while Putonghua-speaking children were grouped by chronological age, South Xinjiang children were categorized by their time of Putonghua learning in kindergarten.

Comparing South Xinjiang preschool children with their age-matched Putonghua-speaking peers (age 6, shown in Figure 2, bottom right), the latter demonstrated significantly lower mean_C and mean_V values, indicating that South Xinjiang children have not yet achieved comparable fluency levels. Variability metrics (varco_V, varco_C, and nPVI_C) revealed significant differences, with Putonghua-speaking children exhibiting more stable vowel durations and more consistent consonant pronunciation, while South Xinjiang preschool children showed greater variability. Speech and articulation rates were highest in six-year-old Putonghua speakers, followed by senior-class children, with mid-class children showing the lowest. The gap between senior-class and Putonghua-speaking children was less statistically significant.

Mid-class children displayed high vowel proportions, greater rhythm variability, and slower speech rates, resembling the rhythm metrics of 3-4-year-old Putonghua speakers. In contrast, senior-class children exhibited lower vowel proportions, reduced rhythm variability, increased regularity, and improved speech rates and rhythmic control, aligning with the metrics of 5-year-old Putonghua speakers. These developmental patterns likely reflect the influence of Putonghua learning time, language environment, and acquisition processes, highlighting age-specific characteristics in rhythm development among South Xinjiang preschoolers.

### 3.2. PCA Analysis of Rhythm Metrics by Age Group

PCA results (Table 3) reveal distinct developmental patterns in rhythm metrics among South Xinjiang preschoolers. At the mid-class stage, vowel-related indices dominate rhythm variation (delta_V = 0.71, rPVI_V = 0.71), while consonant-related metrics show moderate negative loadings (delta_C = -0.67, rPVI_C = -0.66). In contrast, senior-class data shows stronger consonant-related influences (delta_C = 0.80, rPVI_C = 0.80), indicating a developmental shift from vowel- to consonant-based temporal control.

We compared the PCA results of rhythm metrics between South Xinjiang preschoolers and previously studied Putonghua-speaking children (ages 3-6). Line chart in Figure 3 displays cosine distances of rhythm metric PCA results (based on the first three principal components). Consistent with earlier findings, South Xinjiang mid-class children's PCA results most closely match the 3-year-old Putonghua-speaking group, while senior-class results align with the 5-year-old group. For vowel-related metrics (delta_V, mean_V, nPVI_V, rPVI_V, varco_V), the 4-year-old Putonghua-speaking group shows closer proximity to both South Xinjiang groups—particularly the mid-class—compared to consonant-related metrics. Neither South

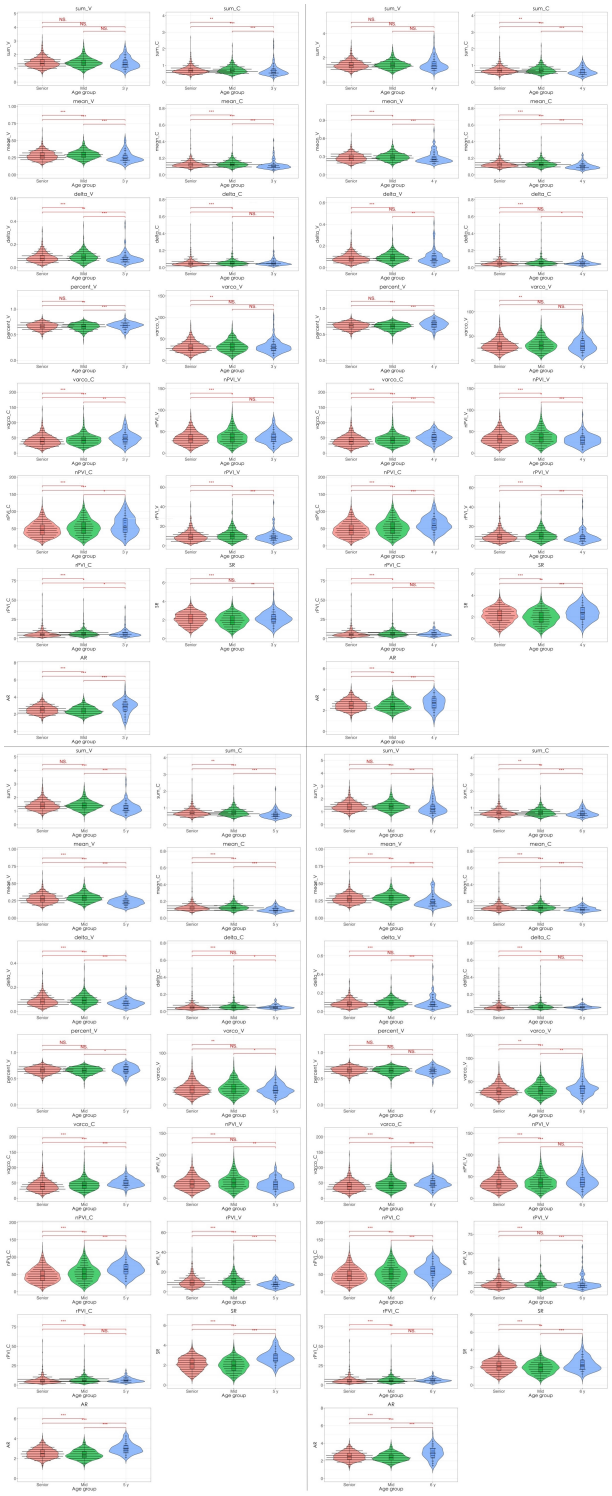

Figure 2: *Rhythm metric distributions across age groups: south Xinjiang preschoolers (red and green violins) compared with Putonghua-speaking children aged 3 (top left), 4 (top right), 5 (bottom left), and 6 years (bottom right).*

Xinjiang group achieved rhythm metric levels comparable to 6-year-old Putonghua speakers.

Table 3: *PCA factor rotation matrix of rhythm metrics.*

| | | | |
|---|---|---|---|
| Mid (5;6-6;7, Mean: 6;1) | | | |
| | 32.23 | 30.68 | 15.01 |
| PC1 | delta_V:0.71 | rPVI_V:0.71 | rPVI_C:0.70 |
| PC2 | delta_C:-0.67 | rPVI_C:-0.66 | varco_C:-0.64 |
| PC3 | percent_V:0.81 | nPVI_C:0.58 | mean_C:-0.58 |
| Senior (5;11-7;7, Mean: 7;1) | | | |
| | 36.00 | 30.08 | 13.79 |
| PC1 | delta_C:0.80 | rPVI_C:0.80 | varco_C:0.66 |
| PC2 | delta_V:0.72 | rPVI_V:0.69 | varco_V:0.67 |
| PC3 | percent_V:0.78 | mean_V:0.71 | nPVI_V:0.41 |

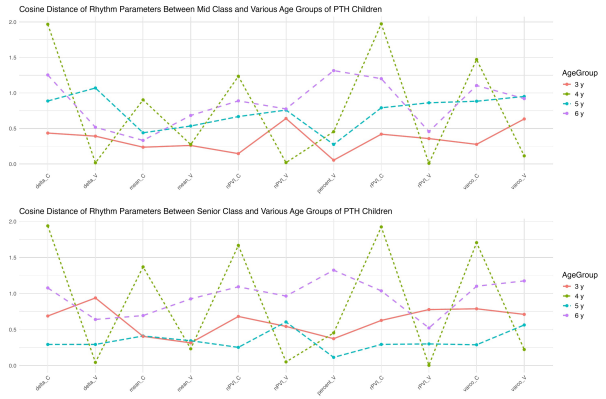


Figure 3: *Cosine distances of rhythm metric PCA results (first three principal components) between South Xinjiang preschoolers—mid-class (mean: 6;1) and senior-class (mean: 7;1)—and Putonghua-speaking children across age groups (3y-6y).*

# 4. Discussion and Conclusion

## 4.1. Putonghua Rhythm Pattern of South Xinjiang Preschoolers

Research on English-speaking children reveals that their vowel interval variability (Varco_V and nPVI_V) is lower than adults', reflecting reduced vowel duration variation. This pattern stems from children's limited syllable structure repertoire and fewer weak syllables [21]. Some scholars interpret this as a developmental progression from syllable-timed to stress-timed rhythm [22]. Conversely, children exhibit higher consonant interval variability (delta_C and rPVI_C) compared to adults, indicating less developed consonant articulatory control in early development [21].

Our study of preschoolers (ages 5-7) in the South Xinjiang region shows that vowel interval variability (Varco_V and nPVI_V) decreases with age grow and Putonghua learning time lengthen. This pattern may reflect cross-linguistic differences in syllable structure complexity and stress patterns. Uyghur—a pitch-accent language—marks stress primarily through pitch rather than intensity or loudness, typically placing stress on the final syllables of word. The bi/multilingual environment of South Xinjiang children, exposing them to multiple rhythm types, may yield vowel variability developmental patterns distinct from monolingual children.

Consonant interval variability (delta_C and rPVI_C) similarly decreases with age grow and Putonghua learning time lengthen, likely reflecting progressive mastery of complex consonant structures. Preschoolers in the South Xinjiang face the additional challenge of processing cross-linguistic consonant structures, potentially intensifying consonant control demands. Through sustained learning and practice, the children gradually develop precise consonant control. This extended developmental trajectory for consonant acquisition aligns with Allen & Hawkins's (1980) findings.

Our study revealed a clear age-related trend in the rhythm production of Putonghua among preschoolers in South Xinjiang, highlighting that speech rhythm development is a continuous process shaped by both age and language experience. According to the "Input-Driven Hypothesis" [23], the quantity and quality of language input are critical for children's speech development. Increased learning time likely provides more language input and practice opportunities, facilitating speech rhythm development.

However, unlike previous findings on Putonghua-speaking children, the rhythm development of South Xinjiang preschoolers shows increasing significance of consonant-related metrics with age and extended learning time, likely due to their linguistic environment. Uyghur, a Turkic language, features vowel harmony, word-final consonant clusters, and phonetic influences from Indo-European languages such as Tokharian and Sogdian. These influences include vowel weakening in unstressed or open syllables, absent in other Turkic languages [24]. This bi/multilingual environment appears to generate rhythm development patterns distinct from monolingual Putonghua speakers.

## 4.2. Rhythm Contrasts: South Xinjiang versus Putonghua Preschoolers

Our research found that, in terms of rhythm metrics distribution, mid class children (mean: 6;1) from South Xinjiang are most similar to 3-4-year-old Putonghua-speaking children, while senior class children (mean: 7;1) are closest to 5-year-old Putonghua-speaking children. PCA results further support this observation. This phenomenon may be attributed to the following factors:

1. Input Rhythm Characteristics

   Children in South Xinjiang receive multi-language input, with diverse rhythmic patterns potentially influencing their Putonghua rhythm development. This linguistic diversity likely contributes to their rhythm parameter distributions diverging from age-matched Putonghua monolinguals.

2. Rhythm Acquisition Pattern

   The bi/multilingual environment may moderate rhythm development pace. Early-stage language switching demands could slow Putonghua rhythm acquisition, explaining the observed age-group displacement. While pre-kindergarten Putonghua exposure may accelerate rhythm acquisition, the reduced input intensity—particularly given that exposure predominantly comes from media sources like television rather than home environment—appears insufficient to eliminate developmental timing differences.

Our research demonstrates that rhythm development is shaped by multiple factors: age, linguistic environment, language input, and educational background. This complexity suggests the need for more sophisticated developmental models that can account for rhythmic diversity across different linguistic contexts, particularly in bi/multilingual environments like South Xinjiang.

## 5. Acknowledgements

This work was supported by the Independent Topic Selection Project of Cultural Experts and "Four Batches" of Talents awarded to Aijun Li, and also Key Laboratory of Linguistics, Chinese Academy of Social Sciences (2024SYZH001).

## 6. References

[1] S. Katrin et al, "Language-specific stress perception by 9-month-old French and Spanish infants," *Developmental science,* vol. 12,6. 2009.

[2] H. Barbara et al, "Language specific prosodic preferences during the first half year of life: evidence from German and French infants," *Infant behavior & development,* vol. 32, 3, pp. 262-74. 2009.

[3] E. Payne, B. Post, L. Astruc, P. Prieto, and M. M. Vanrell, "Measuring child rhythm," *Language and Speech*, vol. 55, no. 2, pp. 203–229, Jun. 2012.

[4] L. Polyanskaya and M. Ordin, "Acquisition of speech rhythm in first language," *Journal of the Acoustical Society of America*, vol. pp. 138, no. 3, pp. 199-204, Sep. 2015.

[5] L. White and S. L. Mattys, "Calibrating rhythm: First language and second language studies," *Journal of Phonetics*, vol. 35, no. 4, pp. 501–522, Oct. 2007.

[6] V. Marilyn, N. Satsuki and D. Rory, Getting the rhythm right: A cross-linguistic study of segmental duration in babbling and first words. New York: De Gruyter , pp. 341-366, 2006.

[7] G. D. Allen and S. Hawkins, "Chapter 12 - PHONOLOGICAL RHYTHM: DEFINITION AND DEVELOPMENT," *Child Phonology*, Pages 227-256, 1980.

[8] P. Elinor et al. "Measuring child rhythm," *Language and speech*, vol. 55. 2012.

[9] P. Leona and O. Mikhail, "Acquisition of speech rhythm in first language," *The Journal of the Acoustical Society of America*, 138. EL199-204, 2015.

[10] M. Kehoe et al. "Acoustic correlates of stress in young children's speech," *Journal of speech and hearing research*, vol. 38,2, pp. 338-50, 1995.

[11] N.P.V. Nair, N. Hariharasubramanian and C. Pilapil, "Circadian rhythm of plasma melatonin in endogenous depression," *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, Volume 8, Issues 4–6, Pages 715-718, 1984.

[12]A.Li, Z. Wang, S. Zhang, J. Gao, and X. Zhou, "The Development of Speech Rhythm in Mandarin-Speaking Children," *2024 IEEE 14th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. pp. 274 - 278, 2024.

[13] J. Gao, *The Manual of the Articulation Test for 1.5- to 6-year-old Putonghua-speaking Children in Beijing*. Beijing: China Social Sciences Press, 2024.

[14] Y. Jia, and L. Pan, A Research of Preschool Teachers' Production Characteristics and Teaching Strategies of Vowels in Standard Chinese in South Xinjiang Ethnic Areas. Kashgar: Journal of Kashi University,45(02), pp.56-63, 2024.

[15] A. Li, J. Gao and Z. Wang, "Effect of Complex Boundary Tones on Tone Identification: An Experimental Study with Mandarin-speaking Preschool Children." Proc. *Interspeech 2024*, pp. 4204-4208, 2024.

[16] P. Boersma and D. Weenink, *Praat: Doing Phonetics by Computer*, (version 6.4.12), [Computer program]. Retrieved: May 4, 2024. Available: http://www.praat.org/.

[17] F. Ramus, E. Dupoux, and J. Mehler, J, "The psychological reality of rhythm classes: Perceptual studies," in *Proceedings of the 15th International Congress of Phonetic Sciences*, Barcelona, pp. 337–342, Aug. 2003.

[18] F. Ramus, M. Nespor, and J. Mehler, "Correlates of linguistic rhythm in the speech signal," *Cognition*, vol. 73, no. 3, pp. 265–292, Dec. 1999.

[19] V. Dellwo and P. Wagner, "Relations between language rhythm and speech rate," in *Proceedings of the 15th International Congress of Phonetic Sciences*, Barcelona, pp. 471–474, Aug. 2003.

[20] E. L. Low, E. Grabe, and F. Nolan, "Quantitative characterisations of speech rhythm: 'Syllable-timing' in Singapore English," *Language and Speech*, vol. 43, no.4, pp. 377–401, Oct. 2000.

[21] Allen, G., & Hawkins, S. (1980). Phonological rhythm:Definition and development. In G. Yeni-Komshian, J. Kavanagh & C. Ferguson (eds.), Child phonology (vol. 1): Production, pp. 227–256. New York: Academic Press.

[22] M. Ordin and L. Polyanskaya, "Acquisition of English speech rhythm by monolingual children," *INTERSPEECH*. pp. 3120-3124, 2015.

[23] M. Harrington and S. Dennis, S, "Input-driven language learning," *Studies in Second Language Acquisition*, 24(2), pp. 261–268, 2002.

[24] L. Johanson, "The classification of the Turkic languages," in Martine Robbeets, and Alexander Savelyev (eds), *The Oxford Guide to the Transeurasian Languages*, Oxford Academic, 17 Sept. 2020.

[25] A. Li and P. Brechtje, "L2 ACQUISITION OF PROSODIC PROPERTIES OF SPEECH RHYTHM: Evidence from L1 Mandarin and German Learners of English," *Studies in Second Language Acquisition* 36.2, pp. 223–255, Web. 2014.

[26] K. de Bot and D. Larsen-Freeman, "Researching second language development from a Dynamic Systems Theory perspective," In Verspoor, M. H., de Bot, K., & Lowie, W. (Eds.), *A dynamic approach to second language development*, Amsterdam: Benjamins. 2011.

[27] K. de Bot, W. Lowie and M. Verspoor, "A Dynamic Systems Theory approach to second language acquisition," *Bilingualism: Language and Cognition*, 10, 7–21, 2007.