Proceedings of the 2nd International Conference on Interdisciplinary Humanities and Communication Studies
DOI: 10.54254/2753-7048/34/20231974

Analysis of Speech Prosody Characteristics of Teachers

Wei Deng[1,2,a,*], Meifeng Zhu[1,2,b], Min Ma[1,2,c], Yu Tian[1,2,d]

[1]Hubei Research Center for Educational Informationization, Central China Normal University, Luonan Street, Hongshan District, Wuhan, China
[2]Faculty of Artificial Intelligence in Education, Central China Normal University, Luonan Street, Hongshan District, Wuhan, Hubei, China
a. sdengwei@mail.ccnu.edu.cn, b. 2849737721@qq.com, c. 2794362403@qq.com, d. 2822706312@qq.com
**corresponding author*

Abstract: During the teaching process, the prosodic features of a teacher's speech can influence students' auditory perception of the teacher's speech and, subsequently, affect students' enthusiasm, concentration, and comprehension of the lesson content. This study aims to analyze the differences in prosodic features, including pitch, intensity, pauses, and speech rate, among speech samples from different categories of teacher instructional videos. The research algorithm is implemented in Python to calculate prosodic feature values for speech segments. Comparative data analysis revealed that, in comparison to negative teacher instructional video speech samples, positive teacher instructional video speech samples exhibit significantly higher average fundamental frequencies (pitch) and more noticeable variations in speech rate rhythm.

Keywords: prosodic features, pitch, intensity, speech rate, pauses

1. Introduction

The digital transformation of education has become an inevitable trend. The key to this transformation lies in the deep integration of digital technology into all aspects of the teaching and learning process, aiming for a comprehensive innovation centered on teaching and management. Numerous studies have employed digital technology to analyze the impact of teachers' behaviors and emotional expressions on students' knowledge construction during the teaching process. These studies extensively explore information related to teachers' behaviors and emotions, allowing for a comprehensive and precise analysis and diagnosis of the entire teaching process. The results of these analyses and diagnoses are then provided as feedback to teachers, enabling them to continuously adjust their teaching methods and styles to enhance teaching quality and better assist students in their learning process [1]. At the current research stage, most studies primarily focus on analyzing teachers' behaviors and emotions during the teaching process, with limited research on the prosodic features of teachers' speech.

Psychological acoustics suggests that within the audible range of the human ear, the prosodic features of speech are crucial factors influencing auditory perception, and a teacher's speech during the teaching process plays a vital role in the construction of students' cognitive structures. Therefore, the study of the prosodic features of teachers' speech is essential.

DOI: 10.54254/2753-7048/34/20231974

The purpose of this research is to analyze the prosodic features of teachers' speech during the teaching process. Using Python, we aim to analyze prosodic features in teachers' speech, such as pitch, intensity, pauses, and speech rate. The analysis results can provide teachers with information on the prosodic features of their speech, enabling them to reflect on their classroom teaching. By improving their teaching behaviors based on the analysis platform's results, teachers can enhance their practical skills in the classroom, promote the formation of a collaborative school culture, and support their professional development [2].

2. Research Foundation

Prosodic features are a phonological structure of language closely related to other linguistic structures such as syntax, discourse structure, and information structure. Prosodic features can be divided into three main aspects: intonation, temporal distribution, and stress, implemented through suprasegmental features. Suprasegmental features include pitch (the range of pitch in voice fundamental frequency), intensity (sound energy or amplitude), and temporal characteristics [3]. Linguistically, "prosody" generally refers to suprasegmental features of speech, including pitch, intensity, and variations in speech rate and pauses.

Psychophonetics indicates that the perception of sound by the human ear is highly nonlinear. Within the audible range of the human ear, there is a certain auditory range for perceiving sounds of different intensities and frequencies. This auditory range encompasses subjective perceptions of sound intensity, pitch, pauses, and speech rate as rhythmic features. Rhythmic features are also a crucial form of linguistic expression, playing a significant role in mutual understanding and expression in verbal communication. Proper utilization of rhythmic features can assist both communicators in constructing a cognitive context for interpreting speech.

Based on the theoretical analysis of prosodic features outlined above, it is evident that different prosodic features in the educational teaching process have a significant impact on students' knowledge understanding and cognitive structure. This paper, in alignment with the mentioned viewpoints, extracts and compares prosodic features of speech segments from 10 teachers in various classrooms, focusing on aspects such as pitch, intensity, pauses, and speech rate. The goal is to analyze the prosodic features of speech segments from different teachers.

3. Overview of Prosodic Features and Extraction Methods

3.1. Pitch

Pitch represents the subjective perception of the highness or lowness of sound by the human ear. Objectively, the size of pitch primarily depends on the fundamental frequency of sound. The fundamental frequency effectively reflects the pitch characteristics of sound. Therefore, fundamental frequency is a significant parameter in studying the rhythmic feature of pitch in speech. A larger fundamental frequency value corresponds to a higher pitch, while a smaller fundamental frequency value corresponds to a lower pitch. Typically, male speakers have a lower fundamental frequency, mostly within the range of 70-200Hz, whereas female speakers and children have relatively higher fundamental frequencies, ranging from 200-450Hz.

Currently, methods for fundamental frequency detection fall into two main categories: event-based and non-event-based detection. Non-event-based methods include auto-correlation, cepstral analysis, average magnitude difference function, among others. Event-based methods involve techniques like wavelet transforms. In this study, the auto-correlation method implemented in Python is used for fundamental frequency calculation. The auto-correlation function is used to determine the fundamental frequency by comparing the similarity between the original signal and the signal shifted in time. This algorithm is a common method for fundamental frequency estimation, particularly

DOI: 10.54254/2753-7048/34/20231974

suitable for extracting fundamental frequency in noisy environments. The formula for auto-correlation is as follows:

$$R(m) = \sum_{n=-\infty}^{n=+\infty} x(n)x(n+m)$$

X represents the short-time signal, R is the auto-correlation function of the short-time signal, and it is used to obtain the fundamental frequency of the signal.

In this study, the average value of the fundamental frequency (F_{0avg}) is selected as a parameter for analysis, with the calculation formula as follows:

$$F_{0avg} = \frac{1}{N} \cdot \sum_{j=1}^{N} F_{0j}$$

3.2. Intensity

Intensity represents the subjective perception of sound strength when sound waves of a certain intensity act on the human auditory organs. Objectively, sound intensity can be reflected by the magnitude of short-term average energy. The greater the intensity, the larger the short-term average energy of the sound, and vice versa. In this study, the analysis of speech signal intensity is based on the short-term average energy. For signal x(m) at time n, the short-term energy is defined as E_n, and its calculation formula is as follows:

$$E_n = \sum_{m=n-(N-1)}^{n} [x(m)w(n-m)]^2$$

X represents the short-time signal, w is the window function. In the equation, N represents the window length. It's evident that short-term average energy is the weighted sum of squared values for a frame of sample points.

In this study, the short-time energy calculation uses a frame length of 256 sample points and employs the Hamming window function. The parameter for analysis is the average of short-term average energy (E_{navg}), with the calculation formula as follows:

$$E_{navg} = \frac{1}{k} \cdot \sum_{j=1}^{k} E_{nj}$$

3.3. Pauses

Pauses refer to the intervals occurring in speech flow, manifesting as intervals and breaks between words, sentences, or paragraphs. Different pauses in the same language can produce varying expressive effects. Correctly timed pauses are crucial for accurate comprehension of the entire speech segment. During speech communication, speakers consciously vary the rhythm of pauses to attract the listener's attention and stimulate their thoughts. Pauses in speech can be categorized into two types: grammatical pauses and logical pauses. Grammatical pauses reflect the grammatical structure within a sentence and are primarily indicated by punctuation marks. Logical pauses, on the other hand, serve the purpose of emphasizing specific elements, highlighting particular semantics, or conveying

DOI: 10.54254/2753-7048/34/20231974

certain emotions. These pauses may occur at locations where there are no punctuation marks in written text, and they may also be more significant pauses at locations with punctuation marks [4].

Currently, detecting speech pauses primarily involves using a threshold approach. It observes the short-term average energy of the current frame, determining speech when it exceeds the threshold and detecting pauses when it does not. Continuous frames below the threshold confirm the presence of a pause.

In this study, the energy threshold for pause detection is set at 20dB, and it is considered a pause when four or more consecutive frames are below this threshold.

3.4. Speech Rate

Speech rate is typically measured by the number of syllables per unit time. Research suggests that excessively fast speech impedes learners' comprehension, and once the speech rate slows to a certain point, further slowing does not significantly enhance comprehension. Very slow speech requires listeners to retain information for longer periods in their short-term memory, which can lead to memory decay [5].

Currently, there are two methods for calculating speech rate: one includes silent pauses (hence, it accounts for pauses in speech), and the other does not (known as articulation rate). This study employs the second method, excluding silent pauses from the calculation, by measuring the number of syllables in a speech segment divided by the duration of pure speech.

4. Research Design

4.1. Research Questions

Through questionnaire-based research, it was found that the prosodic features of teachers' speech in the classroom affect students' enthusiasm for learning. Teachers with higher voice frequencies are more favorable for students' auditory perception and attention. A resonant and lively teacher's voice aids students in quickly engaging in the learning process. Variations in the teacher's speech rate and pause rhythm can contribute to a better teaching rhythm. Based on this investigation, this study selects 15 teaching video examples and analyzes the prosodic features of teachers' speech in these examples. The specific research questions are as follows:

What are the differences in pitch, intensity, pauses, and speech rate among prosodic features in the speech samples of positive and negative teaching video examples?

4.2. Study Subjects

In this study, a total of fifteen instructional video lessons delivered by male instructors from publicly available online courses were selected as the source of speech data. The duration of each instructional video lesson varied from 30 to 50 minutes. After categorization and screening through a questionnaire survey, five speech samples from positively perceived instructional video lessons and five from negatively perceived instructional video lessons were chosen. Subsequently, 50 small instructional video lesson speech segments were extracted from each instructional video lesson sample. These small segments varied in duration from 40 to 65 seconds. In total, there were 300 instructional video lesson speech segments. All samples were recorded in the WAV format and consisted of monaural signals.

5. Data Analysis

The study designed and implemented techniques to extract prosodic features, including intensity, pitch, pauses, and speech rate, from the speech segments.

DOI: 10.54254/2753-7048/34/20231974

Based on the prosody data, it can be calculated that the overall average fundamental frequency and average short-time average energy for the five positively perceived instructional video lesson speech samples are 176.07 Hz and 28.93 dB, respectively. For the five negatively perceived instructional video lesson speech samples, the corresponding prosody data yields values of 151.93 Hz for the average fundamental frequency and 28.78 dB for the average short-time average energy. Analyzing the pitch and intensity data (as shown in Figure 1) from the five positively perceived teacher speech samples and the five negatively perceived teacher speech samples, it is evident that, excluding certain outliers, the average fundamental frequency of the negatively perceived male teacher speech samples is lower than that of the positively perceived male teacher speech samples. However, the average short-time average energy of the negatively perceived male teacher speech samples is approximately equal to that of the positively perceived male teacher speech samples, with no significant difference.

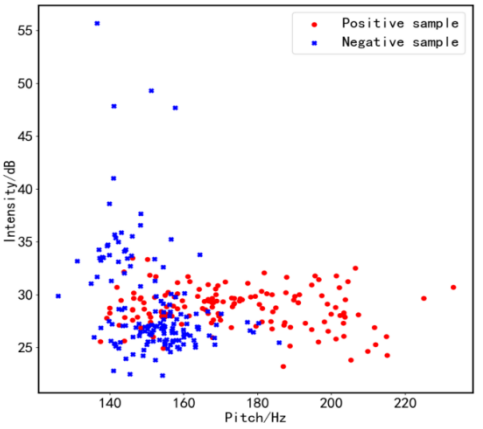

Figure 1: Comparison of Pitch and Intensity Data

Figure 2 displays box plots for the speech rate of five negative teacher speech samples and five positive teacher speech samples. M_NT1 represents the first negative teacher's teaching video example speech sample, while M_PT1 represents the first positive teacher's teaching video example speech sample. After removing outlier data, it is evident that the individual box plots of the speech samples from the five positive male teachers are relatively tall. This indicates significant variations in speech rate between each small sample within the same teacher's teaching video example speech sample. The individual teaching video example speech samples from each teacher show notable variations in speech rate and rhythm. In contrast, the individual box plots for the five negative male teachers' speech samples are relatively short, suggesting that within the same teacher's teaching video example speech sample, there is less variation in speech rate. The variations in speech rate and rhythm within the teaching video example speech samples of individual teachers are less pronounced.

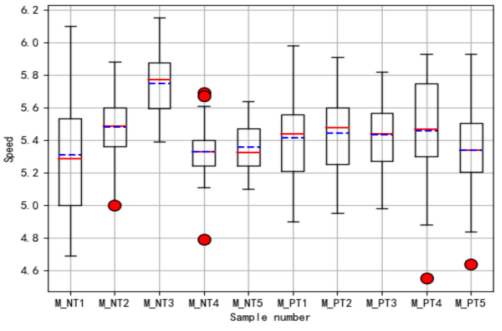

Figure 2: Box Plots for Speech Rate

DOI: 10.54254/2753-7048/34/20231974

Figure 3 presents box plots for pauses in the speech samples of five negative teachers and five positive teachers. M_NT1 represents the first negative teacher's teaching video example speech sample, and M_PT1 represents the first positive teacher's teaching video example speech sample. After eliminating outlier data, it becomes evident that the individual box plots of the speech samples from the five negative male teachers are relatively tall. This suggests significant variations in pauses within the speech samples of the same teacher's teaching video example. On the other hand, the individual box plots for the five positive male teachers' speech samples are relatively short, indicating that there is less variation in pauses within the speech samples of the same teacher's teaching video example. The variations in pauses within the teaching video example speech samples of individual teachers are less pronounced.

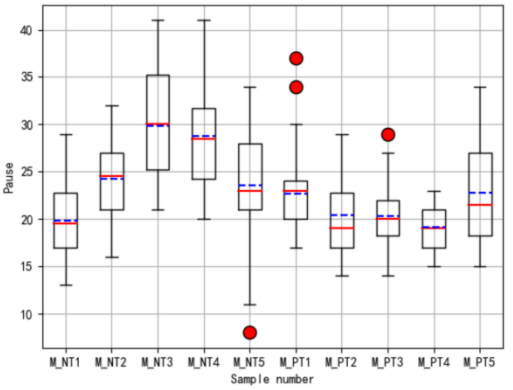

Figure 3: Box Plots for Pauses

6. Conclusion

This study has made significant progress in the analysis of the rhythmic features of teachers' instructional speech in the classroom. The rhythmic features, including pitch, intensity, pauses, and speech rate, in teachers' speech can influence students' perception of the teacher's speech, subsequently impacting students' enthusiasm, attention, and comprehension of the class content. Based on the data analysis, the following conclusions can be drawn: The average fundamental frequency of positive teacher speech samples in teaching videos is higher than that of negative lead teachers; There is no significant difference in the short-term average energy between positive teacher speech samples in teaching videos and those of negative lead teachers; The speech rate within individual speech segments in the teaching video samples of positive teachers exhibits more significant variations compared to negative teachers; Pauses within individual speech segments in the teaching video samples of positive teachers are smaller in variation compared to those of negative teachers. Upon analysis, it was observed that many negative samples in this study were from teaching examples of novice teachers with limited teaching experience. In certain teaching segments, these novice teachers exhibited less fluency in their speech, which were mistakenly identified as pauses. Consequently, this led to the presentation of smaller variations in pauses within the individual speech segments of the teaching video samples of positive teachers compared to negative teachers. This limitation needs to be considered when interpreting the results of this study.

Given the current progress of the research, future research work will mainly focus on the following aspects: 1. Further refining the data and exploring other differences in rhythmic features among different categories of teacher teaching video speech samples. 2. Conducting empirical research to investigate the impact of pitch as an independent variable on student learning.

DOI: 10.54254/2753-7048/34/20231974

References

[1] *Ling, L. (2011). Research on Rhythmic Features at the Turn of Conversation at Home and Abroad. *Journal of Taiyuan Urban Vocational and Technical College*, 2011(02), 205-206. DOI: [10.16227/j.cnki.tycs.2011.02.047](https://doi.org/10.16227/j.cnki.tycs.2011.02.047).*

[2] *Zhou, P. X., Deng, W., Guo, P. Y., et al. (2018). Research on Intelligent Recognition of S-T Behavior in Classroom Teaching Videos. *Modern Educational Technology*, 28(06), 54-59.*

[3] *Yang, Y. F., Huang, X. J., & Gao, L. (2006). Research on Rhythmic Features. *Advances in Psychological Science*, 2006(04), 546-550.*

[4] *Deng, H. H. (2006). On the Phonological Role of Pauses. *Journal of Hunan University of Science and Technology*, 2006(03), 203-204.*

[5] *Fan, Y. (2016). The Influence of Speech Rate, Video Type, and Vocabulary Knowledge on Foreign Language Audiovisual Comprehension. *Foreign Language Teaching*, 37(01), 58-62. DOI: [10.16362/j.cnki.cn61-1023/h.2016.01.013](https://doi.org/10.16362/j.cnki.cn61-1023/h.2016.01.013).*