Journal of Uludag University Faculty of Education
Uludağ Üniversitesi Eğitim Fakültesi Dergisi
e-ISSN: 2667-6788

Year/Yıl 2026, **Volume/Cilt** 39, **Issue/Sayı** 1, 33-49
https://doi.org/10.19171/uefad.1728485

Comparing Human and AI Judgments of Turkish EFL Learners' Intelligibility

Canan DEVECİ[1]

Research Article

Article History

Received 30.06.2025,
Accepted 26.09.2025

Keywords

Intelligibility, Pronunciation, Turkish EFL learners, AI, ChatGPT

Abstract

This study investigates how the intelligibility of Turkish learners of English as a foreign language (EFL) is assessed by both human raters and artificial intelligence (AI), specifically ChatGPT-4, across two different speaking tasks: a controlled read-aloud passage and a spontaneous picture-description task. Drawing on intelligibility-focused pronunciation research, the study aims to explore how task type, rater type, and pronunciation features (segmental and suprasegmental) affect intelligibility ratings. 30 intermediate-level Turkish learners of English completed both tasks, and their recordings were evaluated by three native English-speaking human raters and an AI model. Quantitative results showed that spontaneous speech received higher intelligibility scores than the read-aloud task, despite including more segmental errors. Suprasegmental features such as rhythm, stress, and phrasing played a greater role in determining intelligibility across tasks. While AI ratings closely matched human judgments in most cases, discrepancies emerged, particularly in samples where prosodic nuances were critical. Qualitative analysis further revealed that both rater types frequently flagged vowel distortions, stress misplacement, and a lack of rhythmic cohesion as common intelligibility detractors. The findings underscore the importance of integrating suprasegmental instruction in EFL pronunciation pedagogy and highlight the potential role of AI tools in supporting intelligibility assessment. Nevertheless, when it comes to assessing natural, prosody-rich speech, human judgment is still crucial. By providing empirical evidence from a Turkish EFL environment and linking L2 pronunciation research with emerging technology, this study contributes to applied linguistics.

İngilizceyi Yabancı Dil Olarak Öğrenen Türk Öğrencilerin Konuşma Anlaşılırlığının İnsan ve Yapay Zekâ Değerlendirmeleriyle Karşılaştırılması

Özet

Araştırma Makalesi

Makale Geçmişi

Başvuru 30.06.2025,
Kabul 26.09.2025

Anahtar Kelimeler

Anlaşılırlık, Sesletim, İngilizceyi yabancı dil olarak öğrenen Türk öğrenciler, Yapay zeka, ChatGPT

Bu çalışma, İngilizceyi yabancı dil olarak öğrenen Türk öğrencilerin İngilizce konuşma anlaşılırlığının insan değerlendiriciler ve yapay zeka (ChatGPT-4) tarafından biri kontrollü okuma ve diğeri serbest resim betimleme olmak üzere iki farklı konuşma görevi üzerinden nasıl değerlendirildiğini incelemektedir. Anlaşılırlık odaklı sesletim araştırmalarından hareketle, çalışmada görev türü, değerlendirici türü ve sesletim özelliklerinin (parçalı ve parçalarüstü ses birim) anlaşılırlık puanlarını nasıl etkilediği araştırılmıştır. Orta düzeyde İngilizce yeterliliğe sahip 30 Türk öğrenci her iki görevi tamamlamış ve ses kayıtları anadili İngilizce olan üç değerlendirici ve bir yapay zeka sürümü olan ChatGPT-4 tarafından değerlendirilmiştir. Nicel bulgular, serbest resim betimleme görevinin parçalı sesbirim hatalarını daha fazla barındırmasına rağmen okuma görevine kıyasla daha yüksek anlaşılırlık puanları aldığını göstermiştir. Özellikle ritim, vurgu ve söz gruplaması gibi parçalarüstü ses birim özelliklerin anlaşılırlığın belirlenmesinde daha etkili olduğu görülmüştür. Yapay zeka puanlamaları çoğu durumda insan değerlendirmeleriyle yakınlık göstermiş, ancak özellikle prosodi ayrıntıların kritik olduğu örneklerde farklılıklar ortaya çıkmıştır. Nitel analiz, her iki değerlendirici türünün de sıkça ünlü ses bozulmaları, yanlış vurgu ve ritmik bütünlük eksikliğini anlaşılırlığı düşüren unsurlar olarak işaretlediğini ortaya koymuştur. Bulgular, İngilizce öğretiminde parçalarüstü ses birim özelliklerin daha güçlü biçimde ele alınması gerektiğini vurgulamakta ve anlaşılırlık değerlendirmesinde yapay zeka araçlarının potansiyel rolüne dikkat çekmektedir. Bununla birlikte, özellikle prosodi açısından daha zengin olan serbest konuşma değerlendirilmesinde insan yargısı vazgeçilmezdir. Bu çalışma, ikinci dil sesletim araştırmalarını gelişen teknolojilerle buluşturarak uygulamalı dilbilime katkı sağlamakta ve yabancı dil olarak İngilizce öğretimi bağlamından deneysel kanıtlar sunmaktadır.

[1] Post-doctoral Scholar, University of Utah, Department of Linguistics, u6060352@umail.utah.edu

1. Introduction

In the global landscape of English as a Foreign Language (EFL) education, achieving intelligibility in spoken communication has emerged as a central objective in pronunciation instruction. As English continues to serve as a global lingua franca, the pedagogical focus has shifted from aiming for native-like articulation to promoting understandability across a range of international interlocutors (Jenkins, 2000; Seidlhofer, 2013). Within this paradigm, intelligibility—the extent to which a listener actually understands an utterance—has been prioritized over accent reduction, which emphasizes practical communicative competence over phonetic precision (Derwing & Munro, 1995; Levis, 2005, 2020). This shift aligns with the understanding that intelligibility encompasses more than phonemic accuracy; it includes listener expectations, linguistic background, suprasegmental features, and communicative context (Munro & Derwing, 1995; Levis, 2020). Jenkins (2000) argues in her influential work on English as a Lingua Franca (ELF) that native-speaker norms are neither necessary nor realistic for successful international communication. Similarly, Seidlhofer (2013) calls for the reevaluation of pronunciation targets in EFL contexts, promoting intelligibility-oriented instruction tailored to diverse language users.

Despite the theoretical clarity of this approach, intelligibility remains difficult to assess objectively. Pronunciation assessment traditionally depends on subjective human judgment, which can vary due to raters' backgrounds, experience with accented speech, and task type (Kormos & Dénes, 2004). Furthermore, as Kang et al. (2010) and Saito (2018) note, the type of speaking task—controlled (e.g., read-aloud) versus spontaneous (e.g., picture description)—can significantly influence perceived intelligibility through differences in prosody, lexical selection, and fluency.

At the same time, the emergence of artificial intelligence (AI) in language education has prompted interest in its application to pronunciation evaluation. Although tools such as automatic speech recognition (ASR) and AI-based scoring systems are gaining popularity, little is known about their reliability in assessing complex constructs like intelligibility. As such, empirical investigations are needed to determine how AI-generated evaluations compare to native-speaker judgments, particularly in relation to task type and error type.

This study contributes to this growing body of research by comparing human and AI evaluations of intelligibility across read-aloud and picture-description tasks performed by Turkish EFL learners. Specifically, the study asks: (1) Does task type influence intelligibility ratings? and (2) To what extent do AI-generated evaluations align with human raters' judgments? By analyzing both segmental and suprasegmental error patterns, this study also seeks to better understand the phonological variables underlying intelligibility in different speech tasks and assess the viability of AI as a tool for pronunciation assessment in EFL contexts.

2. Literature Review

2.1. Defining Intelligibility

Intelligibility, comprehensibility, and accentedness have long been recognized as distinct yet interrelated constructs in L2 pronunciation research. *Intelligibility* refers to the extent to which an utterance is actually understood by a listener, whereas *comprehensibility* reflects the listener's subjective perception of how easy or difficult the utterance is to understand (Derwing & Munro, 1997, 2015; Munro & Derwing, 2020). *Accentedness*, on the other hand, pertains to the degree to which a speaker's pronunciation diverges from native norms (Munro & Derwing, 1995; Munro et al., 2006). These three constructs, although often correlated, are influenced by different factors and have unique pedagogical implications.

Central to pronunciation instruction is the "intelligibility principle", which posits that learners should focus on being understandable rather than aiming for native-like pronunciation (Levis, 2005). According to this view, intelligibility is a more practical and attainable goal in international communication settings where English functions as a lingua franca (Jenkins, 2000; Seidlhofer, 2013). Levis (2020) emphasizes that intelligibility is not a fixed quality inherent to the speaker but rather a dynamic interaction between speaker output and listener perception. Factors such as segmental accuracy, prosodic features (stress, intonation, rhythm), lexical choice, and fluency all contribute to intelligibility outcomes (Pease, 2016). In line with this view, Pease (2016) also stresses that intelligibility is often judged more favorably when suprasegmental elements—such as sentence stress and phrasing—are appropriately used, even if minor segmental deviations are present.

Research has further demonstrated that intelligibility is task-dependent and listener-dependent (Lochland, 2020). For instance, controlled speech tasks like read-aloud passages may elicit greater segmental accuracy, while spontaneous speech can highlight fluency and suprasegmental control. However, both task types may yield different intelligibility profiles depending on listener familiarity, expectations, and language background. Lochland (2020) argues that

intelligibility is more appropriately assessed across a range of speaking tasks and listener types to account for the variability in real-world communication.

As intelligibility becomes a core objective in L2 pedagogy, assessment methods have begun to shift accordingly. While subjective human judgment remains common, researchers are increasingly calling for consistent rating criteria and exploring the potential of automated systems in evaluating speech intelligibility (Derwing & Munro, 2009). These shifts reflect a broader recognition of intelligibility as a multi-dimensional, interactional, and context-sensitive phenomenon. A more recent article by Levis (2018) argues for intelligibility-based instruction as a pedagogical objective in TESOL contexts, stressing the need for understanding intelligibility within a larger framework of communicative competence. There are clear consequences for pronunciation pedagogy in the work of Derwing and Munro (2015), who point out that segmental accuracy is not always more important for understanding than suprasegmental elements like rhythm, stress, and intonation.

2.2. Intelligibility in English Pronunciation Research

The field has witnessed a significant expansion in empirical studies on intelligibility within second language (L2) pronunciation. Research has increasingly highlighted the complex interplay between segmental and suprasegmental features in shaping listener perception. For instance, Kang et al. (2010) demonstrated that suprasegmental features such as word stress, pitch range, and intonation contour significantly predict intelligibility ratings by native listeners. Similarly, Saito (2018) found that rhythm and speech fluency contribute more to comprehensibility than precise phoneme articulation. From a developmental perspective, Aoyama and Guion (2007) examined the acquisition of English /r/ and /l/ by Japanese learners and emphasized the importance of both perceptual and production-based training for improving intelligibility, especially for L1 speakers with phonemic inventories divergent from English. This finding underscores the need for integrated phonological training that goes beyond isolated segmental correction. Foote et al. (2011) further supported this multidimensional view in their study on ESL pronunciation teaching practices in Canada. They reported that instructors increasingly prioritize intelligibility and comprehensibility over native-like pronunciation, which is in line with the communicative goals of modern language pedagogy. These findings echo the pedagogical stance advocated by Celce-Murcia et al. (2010), who proposed a communicative framework for pronunciation instruction that incorporates form-focused activities, communicative tasks, and feedback strategies—all aimed at improving functional intelligibility.

Lee et al. (2015) provided more evidence in a meta-analysis indicating, in ESL/EFL settings, suprasegmental-focused instruction had a greater impact on comprehensibility than segmental training. This aligns with the communication paradigm proposed by Celce-Murcia et al. (2010), which suggests that rhythm, stress, and intonation patterns should be prioritized in the classroom. Collectively, these studies establish suprasegmental instruction as an essential component of contemporary TESOL pedagogy and lay the theoretical groundwork for investigating prosody-driven intelligibility in the present investigation.

In terms of assessment, tools such as speech shadowing (Field, 2005) and keyword transcription tasks (Munro & Derwing, 1995) have been used to obtain more objective measures of intelligibility. However, studies like Saito (2012) caution that human raters, even with training, may vary in how they weigh segmental versus prosodic errors. These converging lines of evidence point to a paradigm shift in L2 pronunciation research—from accent reduction to intelligibility-focused instruction and assessment—where learner success is evaluated in terms of their ability to be understood rather than their approximation of native speaker norms.

2.3. Intelligibility Studies in the Turkish EFL Context

In Türkiye, English language learners frequently struggle with intelligibility due to structural differences between Turkish and English phonology. In Turkish EFL contexts, learners commonly display persistent pronunciation difficulties due to segmental mismatches between Turkish and English, the predominance of orthography-based instruction, and limited access to authentic spoken English (Demirezen, 2005; Dikilitaş & Geylanioğlu, 2019; Ercan, 2018). Segmental challenges include the articulation of interdental fricatives (/θ/, /ð/) and reduced vowel systems, leading to inaccuracies in producing English vowels (Demirezen, 2005; Yavaş, 2011). Suprasegmental difficulties, such as misplaced stress, incorrect rhythm, and flat intonation, also hinder communication effectiveness (Bayraktaroğlu, 2008; Uzun, 2022).

Hismanoğlu (2012) highlighted that while Turkish learners may achieve high grammatical proficiency, their intelligibility often lags due to limited pronunciation instruction. Similarly, Khalilzadeh (2014) emphasized the detrimental effects of orthographic interference in English pronunciation instruction in Türkiye, particularly in settings where English is taught with a heavy reliance on textual materials. Bayraktaroğlu (2008) also noted the lack of systematic pronunciation training in Turkish curricula, which leads to the fossilization of unintelligible speech patterns.

Recent empirical research by Uzun (2022) sheds light on the most salient pronunciation errors affecting intelligibility among Turkish learners of English. In his study, native English listeners and expert raters identified mispronounced segmental features—especially /ə/ and /θ/—as the most impactful on intelligibility. Additionally, stress placement errors in strong syllables were found to have a particularly detrimental effect on listener comprehension, which corroborates previous findings by Demirezen (2005).

Despite growing attention to intelligibility, few studies in Türkiye have employed both controlled and spontaneous speech tasks to assess L2 pronunciation. Most rely on scripted material, which limits the assessment of real-world communicative abilities. The current study addresses this gap by incorporating both read-aloud and picture-description tasks to capture a more holistic view of learner pronunciation performance and its perceived intelligibility by human and AI raters. The importance of this gap in TESOL-oriented pronunciation pedagogy is being more acknowledged (Levis, 2018; Lee et al., 2015), but it is especially noticeable in the Turkish EFL environment, where learners' communicative intelligibility is still constrained by inadequate suprasegmental instruction.

2.4. Intelligibility and Artificial Intelligence

The use of artificial intelligence in pronunciation training and assessment is an emerging frontier in applied linguistics. AI systems, particularly large language models (LLMs) such as ChatGPT, have increasingly been employed in language learning contexts for their capacity to provide scalable and immediate feedback (Babaeian, 2023; Zechner et al., 2009). These tools offer learners an accessible means to practice and refine pronunciation outside traditional classroom settings.

While promising, AI-based assessment tools still face challenges in replicating the complexity of human perceptual judgments. Zou et al. (2024) argue that although AI systems can reliably detect segmental errors—such as misarticulations of consonants or vowels—they are less adept at capturing suprasegmental elements like stress, rhythm, and intonation. Similarly, research by Zechner et al. (2009*)* and Somasundaran et al. (2015) have demonstrated that while AI-generated scores align well with human ratings in structured tasks such as read-alouds, notable discrepancies arise in spontaneous speech contexts due to the AI's limited sensitivity to prosodic variation and discourse-level features.

The application of artificial intelligence in pronunciation instruction is rapidly advancing, particularly with tools like ChatGPT offering immediate, scalable feedback. Mompean (2024) evaluated ChatGPT's performance in providing pronunciation feedback and found that it could accurately flag segmental errors (such as vowel substitutions and consonant omissions) and offer clear, pedagogical suggestions. However, its analysis of suprasegmental features—especially intonation and rhythm—was less reliable, which echoes earlier concerns by Zechner et al. (2009). These findings indicate that while AI tools are valuable for initial pronunciation support, they may not fully replace human evaluation in nuanced speech aspects, especially in spontaneous or discourse-rich contexts. Zechner et al. (2009) demonstrated that speech recognition technologies can predict intelligibility with moderate success but noted that these systems tend to prioritize lexical clarity over natural prosody. This limitation is especially relevant for EFL learners whose intelligibility is often shaped more by prosodic appropriateness than by phonemic accuracy (Derwing & Munro, 2015).

In the Turkish context, the integration of AI in pronunciation assessment remains at a nascent stage. There is limited empirical research exploring how well AI-generated intelligibility ratings align with human perceptions among Turkish EFL learners. The current study addresses this gap by comparing AI and human judgments across two task types—read-aloud and picture description. In doing so, it seeks to evaluate not only the reliability of AI assessments but also their sensitivity to the suprasegmental features that critically impact intelligibility.

3. Method

3.1. Research Design

This study employed a within-subjects comparative design (Dörnyei, 2007) to examine the intelligibility of Turkish EFL learners' speech across two speaking tasks: a read-aloud and a picture-description task. Each participant completed both tasks, and their recordings were evaluated by three native English-speaking human raters and one AI-based rater (ChatGPT-4). Quantitative comparisons (mean scores, correlations) and qualitative analyses (rater comments) were used to explore differences across task types and rating sources.

3.2. Participants

This study recruited 30 adult Turkish-speaking learners of English, all enrolled in a university-level English preparatory program in Türkiye. The participants, aged between 18 and 24, had completed at least one academic term of English

instruction and were placed at CEFR A2–B1 levels according to institutional placement tests. None reported any history of speech or hearing impairments. Informed consent was obtained from all participants in line with ethical guidelines. Although the sample size was sufficient for capturing initial patterns, the study is positioned as an exploratory investigation into human–AI rating alignment rather than a definitive large-scale validation.

3.3. Data Collection Instruments

Participants completed two distinct speech production tasks designed to elicit both controlled and spontaneous speech:

Reading-Aloud Task: Participants read aloud the Rainbow Passage (Fairbanks, 1960; see Appendix A), a widely used text in speech research for its rich segmental and suprasegmental features.

Spontaneous Picture Description Task: Participants described a picture depicting people, objects, and actions for approximately 30–60 seconds (see Appendix B). The image was selected to elicit spontaneous speech containing a variety of grammatical structures and vocabulary items.

Each participant produced one recording for each task, resulting in a total of 60 speech samples.

3.4. Data Collection Procedure

Data were collected in June 2025. All recordings were performed individually in a quiet room using high-quality digital audio recorders. The speech samples were saved, and each participant completed both tasks in a single uninterrupted recording session. Prior to participation, informed consent was obtained from all individuals, and they were assured that their data would be anonymized and used solely for research purposes in accordance with institutional ethical guidelines. This study was approved by the Social and Human Sciences Ethics Committee of Atatürk University (Date: 20.06.2025; Decision No: E.88656144-000-2500199482).

3.4.1. Human Intelligibility Ratings

Three native speakers of English served as raters. Their demographic and professional details are outlined in Table 1.

Table 1. Demographic Data of Human Raters

Raters	Nationality	Age	Sex	Degree	Experience	Stay in Türkiye
1	US	25	F	BA	3 years	no
2	US	26	F	MA	3 years	3 days
3	US	29	M	PHD	5 years	1 week

All raters were from the United States; two were female (ages 25 and 26) and one was male (age 29). All raters held at least a bachelor's degree in language-related fields with at least 3 years of English teaching experience. Raters with relatively short stays in Türkiye were deliberately selected, as extended residence in the country could lead to increased familiarity with local English accents and potentially influence their judgments of intelligibility. Each rater independently evaluated the 60 anonymized audio recordings using the Washington University 7-point intelligibility scale, which is a tool originally developed for clinical assessment of speech intelligibility in motor speech disorders. The scale has demonstrated strong inter-rater reliability and construct validity in both clinical and applied linguistics contexts (Yorkston et al., 1984), which makes it suitable for capturing perceptual variation in L2 speech intelligibility. It provides a graded continuum of judgments that enables distinctions between clearly intelligible, moderately distorted, and unintelligible speech. Below, the scale is presented:

1. No noticeable differences from normal.

2. Intelligible though some differences occasionally noticeable.

3. Intelligible although noticeably different.

4. Intelligible with careful listening although some words unintelligible

5. Speech is difficult to understand with many words unintelligible

6. Usually is unintelligible.

7. Unintelligible.

Raters received no explicit training, calibration session, or predefined error criteria to mirror their authentic real-world listening experiences. They were instructed to evaluate recordings based solely on how easily a native English speaker

would understand the speaker. Raters were also asked to provide qualitative comments to explain their judgments. The inter-rater reliability was strong (Intraclass Correlation Coefficient, ICC = 0.81) indicating a high level of consistency among the human raters' intelligibility scores. For qualitative comments, a second researcher independently reviewed and coded a random 30% of the rater comments to establish inter-coder reliability. An agreement reached 87% and this supports the reliability of the qualitative analysis.

3.4.2. Intelligibility Ratings by Artificial Intelligence

The same 60 recordings were evaluated by ChatGPT-4, a generative AI model developed by OpenAI. Each audio file was analyzed individually, with the model instructed to simulate a native speaker of English rating intelligibility on the Washington University 7-point Intelligibility Scale (Yorkston et al., 1984), the same instrument used by the human raters. To parallel human rater instructions, ChatGPT-4 was given a structured rubric and specific prompts covering segmental and suprasegmental features as well as fluency. The prompt included the following:

"Rate how intelligible the speech sample is on the given scale from 1 (completely unintelligible) to 7 (fully intelligible). Focus only on how easily an average native speaker would understand it, not on grammar or content. Consider whether vowels and consonants are pronounced clearly and distinctly. Note substitutions, deletions, or distortions that reduce intelligibility. Consider stress placement, rhythm, and intonation. Evaluate whether these features help or hinder intelligibility. Consider fluency and pausing. Too many pauses, hesitations, or irregular speech rate may reduce intelligibility even if individual words are correct. Provide one overall score (1–7) reflecting intelligibility and a short qualitative comment highlighting the main features that influenced the score."

ChatGPT-4 produced both a scalar rating and a short qualitative comment for each speech sample. The model had no access to participant demographic information or to human rater scores, ensuring independent evaluation. This procedure resulted in a set of AI-generated intelligibility scores and diagnostic comments directly comparable to those produced by human raters.

3.5. Data Analysis

Descriptive statistics were used to analyze intelligibility scores across tasks and rater types. Mean scores and standard deviations were calculated for both human and AI ratings. Correlation analyses were conducted to determine inter-rater reliability and the extent of alignment between AI and human judgments. Thematic frequency data on error types were analyzed to highlight task-specific trends in segmental and suprasegmental features.

To better understand patterns of unintelligibility, qualitative comments provided by both human raters and ChatGPT were subjected to an inductive thematic analysis. Importantly, neither group was provided with a rubric or predefined error categories. Instead, raters commented based on their overall perceptions of intelligibility. The comments were coded and grouped, including the recurring error types, into three broad categories as given below:

Segmental Errors: Mispronunciations at the phoneme level (e.g., substitution of /θ/ with /t/, or vowel distortions).

Suprasegmental Errors: Problems related to stress, rhythm, intonation, and pausing (e.g., misplaced stress, monotone delivery).

Lexical and Fluency-Related Issues: Hesitations, unnatural pauses, false starts, and lexical retrieval problems.

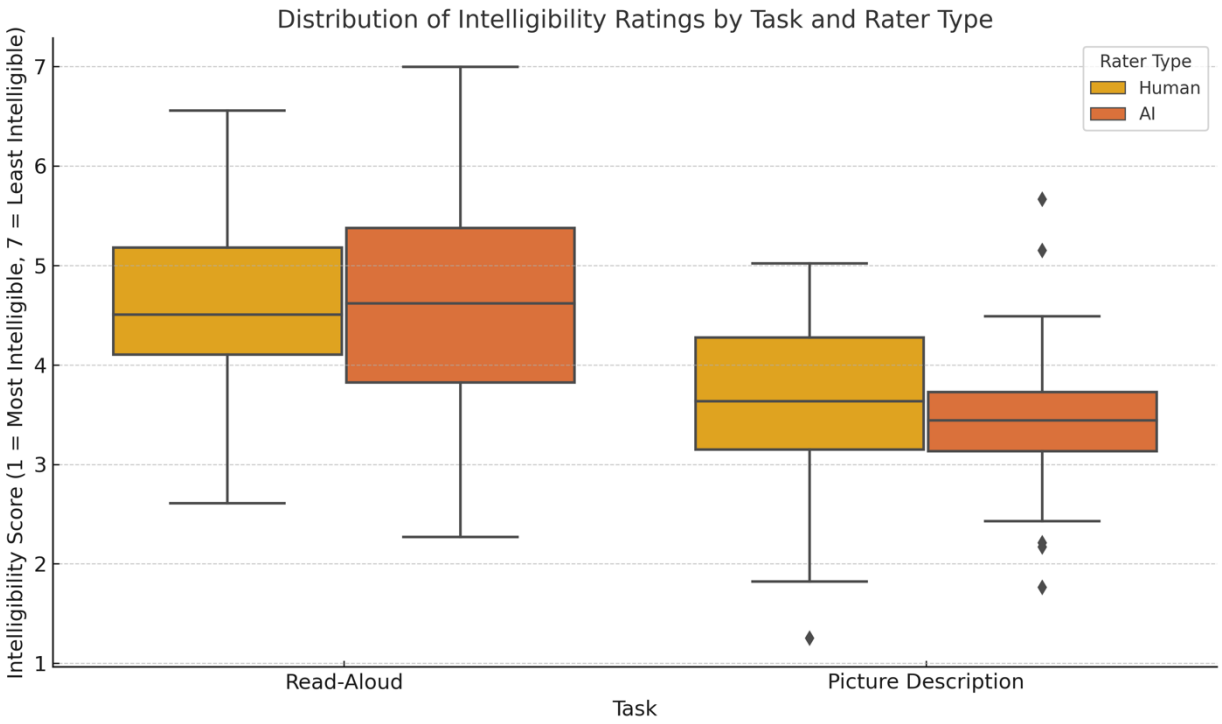
4. Findings

This section presents a comprehensive analysis of intelligibility ratings assigned by three native English-speaking human raters and an artificial intelligence (AI) system (ChatGPT-4) across two speaking tasks: a controlled read-aloud and a spontaneous picture-description task. Findings are reported through descriptive statistics, rater agreement patterns, error typologies, and illustrative examples highlighting the nature and frequency of specific pronunciation issues.

4.1. Descriptive Statistics of Intelligibility Ratings

Figure 1 displays the distribution of intelligibility scores assigned by human raters and ChatGPT across both speaking tasks in boxplots. Ratings were based on a 7-point scale where 1 represents "no noticeable difference from native speech" and 7 indicates "unintelligible".

Figure 1. Intelligibility Ratings by Task and Rater Type



Both human and AI ratings showed higher intelligibility for the picture-description task (M=3.61 from human raters, M=3.47 from AI) compared to the read-aloud task (M=4.77 from human raters, M=4.70 from AI). This trend indicates that spontaneous speech, despite its potential for grammatical or lexical errors, may afford greater fluency and listener comprehensibility through more natural prosody and discourse patterns.

Boxplots in Figure 1 demonstrated that AI ratings closely tracked human scores, particularly at the lower and higher ends of the intelligibility spectrum. Median scores were lower (better) for the picture-description task across both groups. The spread of scores was noticeably wider for human raters in both tasks, suggesting more variability in subjective judgments, whereas AI ratings were more compactly distributed, reflecting higher internal consistency. Outliers appeared in the picture-description condition, particularly for human raters, which indicates that a few learners' performances were perceived as substantially less intelligible despite the general trend. Human raters demonstrated slightly greater variability in their assessments, especially in mid-range scores (3–5), whereas AI produced more tightly clustered ratings. Notably, Rater 3 exhibited a slightly more severe rating trend overall. Strong positive correlations were observed between human and AI intelligibility ratings (r = .76 for the read-aloud task; r = .83 for the picture-description task), suggesting that the AI system demonstrates considerable alignment with human judgments and holds promise as a reliable supplementary assessment tool.

Overall, human-human agreement was strongest for extreme scores—very intelligible or unintelligible speech. Moderate variation appeared for mid-range scores in the read-aloud task, where raters showed some inconsistency in interpreting prosodic variation. AI ratings were less variable than human ratings, particularly in the picture-description task, suggesting greater internal consistency**.** However, they showed slightly more leniency, especially for segmental errors that did not impede meaning. Taken together, the figure highlights that while AI tends to compress scores toward the middle and reduce variability, human raters capture more nuance at both ends of the scale, especially when prosodic features are salient. The discrepant patterns are given below:

In 6 cases, AI scores were more generous by 1–1.5 points.

In 3 cases, AI was stricter—mostly when disfluencies disrupted lexical cohesion.

Qualitative comments revealed that human raters emphasized prosody, while AI focused more on word recognition and syntax.

4.2. Error Typologies and Frequencies

To better understand the basis of intelligibility judgments, the qualitative comments from both human and AI raters were systematically coded for recurring pronunciation issues. These were categorized as segmental (phoneme-level

errors) or suprasegmental (rhythm, stress, prosody) and quantified separately by task. While both rater types identified similar issues, notable differences emerged in their sensitivity and prioritization of these features. To provide a clear overview of the distribution of error types, Table 2 summarizes the frequencies of segmental and suprasegmental features identified across the two speaking tasks by both human raters and the AI system.

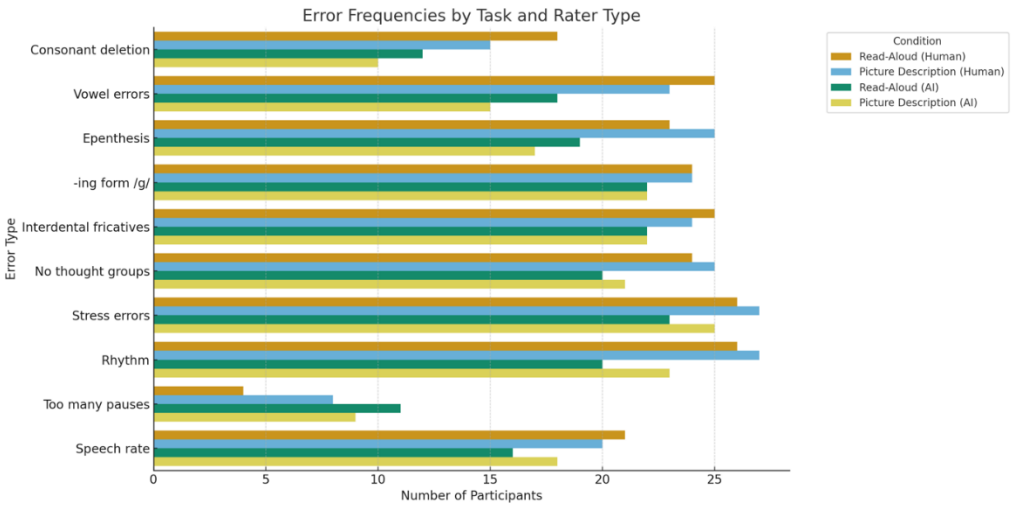
Table 2. Frequencies of Errors by Task and Rater Type

Error Type	Read-aloud (Human)	Picture Description (Human)	Read-aloud (AI)	Picture Description (AI)
Consonant deletion	18	15	12	10
Vowel errors	25	23	18	15
Epenthesis	23	25	19	17
-ing form /g/	24	24	22	22
Interdental fricatives	25	24	22	22
No thought groups	24	25	20	21
Stress errors	26	27	23	25
Rhythm	26	27	20	23
Too many pauses	4	8	11	9
Speech rate	21	20	16	18

As the table shows, human raters consistently reported higher frequencies of vowel errors, stress errors, and rhythm issues compared to the AI model, whereas the AI tended to detect more instances of excessive pauses. Both rater types identified stress and rhythm as the most frequent suprasegmental problems, particularly in the picture-description task, which underscores the central role of prosodic features in shaping intelligibility.

While the table allows for a detailed inspection of error frequencies, the patterns of divergence between human and AI judgments become more apparent when presented visually. Figure 2 illustrates these differences showing how human raters and the AI model diverge in sensitivity to particular error types. For example, human raters highlighted stress and rhythmic cohesion as major contributors to intelligibility, whereas AI judgments clustered more around segmental accuracy and pause distribution. The visual representation thus complements the tabular summary by highlighting the relative weight placed on different error categories by each rater type. Figure 2 illustrates the overall comparison of error types, task types, and rater types.

Figure 2. Frequency of Errors by Rater and Task Type



A number of key observations emerged from the comparative analysis of human and AI rater feedback, summarized in Table 2 and Figure 2. Both rater types consistently identified /ŋ/ reduction, interdental substitutions (e.g., /θ/ realized as [t], /ð/ as [d]), and vowel distortion as the most frequent segmental pronunciation issues. These error patterns are in line with established findings on L1 Turkish interference in English pronunciation (Yavaş, 2011). Human raters, however, demonstrated heightened sensitivity to suprasegmental elements such as stress placement and intonation.

In fact, at least two of the three human judges flagged 28 out of the 30 read-aloud samples for misplaced primary stress. Qualitative feedback frequently included observations like “stress not aligned with sentence meaning” and “monotonic delivery obscures key content”. In the spontaneous picture-description task, human raters emphasized long, unnatural pauses between phrases as the most salient suprasegmental issue affecting intelligibility.

The AI rater, in contrast, focused more on lexical-level clarity and fluency. It commonly highlighted problems such as “disjointed timing”, “lack of word boundary clarity”, and “reduced lexical transparency”. Although the AI system did identify some prosodic concerns—including “unnatural phrasing” and “uneven timing between ideas”—it appeared to assign less importance to rhythm and stress unless these features directly compromised the overall message. Notably, in the spontaneous task, the AI provided more detailed commentary on semantic coherence and word choice, occasionally overlooking stress-related issues that were consistently flagged by human raters.

Qualitative feedback from human and AI raters revealed overlapping concerns but highlighted differences in rater focus. Human raters frequently noted issues related to prosody and stress. One rater commented, “No rising or falling pitch—completely monotone”, drawing attention to the absence of natural intonation contours. Another noted, “Frequent stress on function words confuses meaning”, emphasizing how inappropriate prosodic emphasis disrupted listener comprehension. Segmental problems were also highlighted, with comments such as “Clear consonant errors with /θ/, /ð/, and /ŋ/ across multiple words”, which points to recurring difficulties influenced by L1 Turkish phonology.

The AI rater (ChatGPT-4) similarly flagged intelligibility issues but framed its feedback with a focus on fluency and lexical clarity. One comment stated, “The speech lacks rhythmic grouping, affecting listener comprehension”, which indicates some recognition of prosodic deficiency. Another remark, “Vowel articulation is imprecise in multisyllabic words”, emphasized segmental accuracy. A third comment, “Fluency is affected by pauses between content words”, revealed the AI’s sensitivity to temporal delivery and pacing, especially in spontaneous speech.

To summarize the results, a comparison of error dimensions revealed consistent patterns across rater types with some divergence in focus. Both human raters and the AI system were sensitive to segmental accuracy, particularly substitutions involving interdental fricatives (/θ/ and /ð/) and the nasal /ŋ/. However, human raters were more attentive to suprasegmental features, frequently commenting on stress, rhythm, and overall prosodic delivery. While the AI did acknowledge prosodic issues, it appeared less likely to penalize them unless they directly hindered message clarity. Lexical clarity emerged as a greater priority for the AI, which regularly flagged unclear or imprecise vocabulary. In contrast, human raters only occasionally referenced vocabulary, focusing more on naturalness and communicative effort. Both rater types identified pausing and fluency as important, especially in the picture-description task. Overall, human judgments tended to emphasize delivery, listener effort, and the naturalness of speech, while AI leaned toward evaluating the structural clarity and coherence of the spoken message.

4.3. Task-based Comparison of Intelligibility Patterns

Table 3 presents the task-based comparison of intelligibility ratings by human and AI raters.

Table 3. Human and AI Intelligibility Ratings by Task

Task	Human Mean	AI Mean	Correlation (r)
Read-aloud	4.77	4.70	.76
Picture description	3.61	3.47	.83

As shown in Table 3, both human raters and the AI assigned higher intelligibility to the picture description task than to the read-aloud task. Despite more segmental errors in the read-aloud task, it appeared more structured, which helped some learners maintain rhythm and pacing. However, 12 participants received human intelligibility ratings of 6 or higher on this task indicating near-unintelligibility. For example, Participant 19’s read-aloud sample was labeled as “too slow, very unnatural” and “difficult to follow”.

Conversely, spontaneous speech in the picture-description task often lacked precision but benefited from more natural prosody and discourse markers. Learners appeared more communicatively focused and employed strategies such as self-repair and gesturing (inferred from context) that aided intelligibility. Participant 4, for instance, received human scores averaging 2.0 and an AI score of 2.1 with comments like “excellent use of rhythm and timing” and “minor mispronunciations did not affect understanding”.

Building on the quantitative trends and rater comments discussed earlier, two learner profiles further illustrate how individual variation intersects with task type and rater perception. Participant 22, whose performance on the read-aloud task received a human average score of 6.0 and an AI score of 6.2, exhibited relatively strong segmental accuracy

but lacked prosodic control. Human raters noted “mispronunciations, no attempt to make key words prominent”, while the AI echoed similar concerns flagging “absent stress” and “phonetic distortion”. In contrast, Participant 4, evaluated during the picture-description task, scored 2.0 from human raters and 2.1 from the AI. However, despite minor segmental issues, both rater types highlighted the participant’s use of effective prosodic strategies. Human feedback emphasized “clear rhythm” and “adequate stress”, while the AI described the sample as “mostly intelligible” and credited stress patterns for aiding comprehension.

These contrasting profiles reinforce key findings from the broader dataset: intelligibility is not solely a function of phoneme-level accuracy but emerges from a dynamic interplay of prosodic fluency, task context, and individual learner strategies. They also substantiate the broader trend of AI-human convergence in rating patterns while simultaneously underscoring that AI tools still require human interpretive support when assessing nuanced suprasegmental features.

5. Discussion, Conclusion and Implications

This study aimed to explore (1) how intelligibility ratings of Turkish EFL learners differ across controlled and spontaneous speaking tasks and (2) to what extent AI-generated evaluations align with human judgments. The discussion below integrates the key findings with relevant literature and outlines pedagogical and technological implications.

One major finding was that learners’ speech in the picture-description (spontaneous) task received higher intelligibility ratings than in the read-aloud (controlled) task—despite containing more segmental errors. Quantitative analyses provided clear evidence for task-related differences in intelligibility. As presented in Table 3, mean ratings for the picture-description task were significantly lower (Human M = 3.61, AI M = 3.47) than for the read-aloud task (Human M = 4.77, AI M = 4.70), with strong correlations between human and AI raters (r = .76–.83). These numeric results establish the basis for the subsequent discussion highlighting that task type and rater perspective systematically shaped intelligibility judgments. This aligns with previous research suggesting that prosodic fluency often carries more communicative weight than phonetic precision (Hahn, 2004; Kang et al., 2010; Saito, 2012). Human raters emphasized rhythm, stress placement, and natural pausing as key contributors to intelligibility, especially in less scripted tasks. These findings echo Jenkins’ (2000) Lingua Franca Core model, which underscores suprasegmental features as crucial to mutual intelligibility in global contexts. In contrast, the read-aloud task elicited choppy delivery and reduced comprehension, a trend also observed by Lochland (2020) and Foote et al. (2011), who noted that overly controlled reading suppresses natural speech rhythm and communicative intent. Particularly for Turkish EFL learners—whose L1 prosodic patterns differ markedly from English—natural speech seems to allow more room to deploy adaptive strategies that compensate for phonemic deviations (Bayraktaroğlu, 2008; Uzun, 2022).

Segmental errors—especially with interdental fricatives (/θ/, /ð/), nasal codas like /ŋ/, and unstressed vowels—were consistent across tasks and aligned with past Turkish EFL studies (Demirezen, 2005; Hismanoğlu, 2012; Yavaş, 2011). The error frequency distributions presented in Table 2 and Figure 2 corroborate these observations. Human raters consistently reported higher incidences of vowel errors, stress errors, and rhythm problems, whereas AI ratings were more conservative in these categories but more sensitive to excessive pauses. This pattern reinforces the interpretation that suprasegmental features played a decisive role in human judgments, while AI evaluations prioritized fluency disruptions and lexical clarity. However, these errors were more tolerated in spontaneous speech when prosodic features were intact. The suprasegmental domain emerged as more influential. Consistent with Saito (2012), listeners penalized monotone delivery, misplaced stress, and lack of speech chunking more harshly than individual phoneme errors. These results support research by Celce-Murcia et al. (2010) and Derwing and Munro (1995), who found that suprasegmental training had a stronger impact on perceived comprehensibility than segmental drilling.

Correlations between human and AI ratings were strong overall (r = 0.76 for read-aloud; r = 0.83 for picture-description), which confirms that AI systems like ChatGPT can replicate human intuitions under many conditions (Geng et al., 2025; Zechner et al., 2009). The alignment was especially close for clearly intelligible or unintelligible utterances. However, key differences emerged. Human raters focused more on prosodic nuance—intonation, stress timing, and rhythm—while AI ratings leaned toward lexical clarity, syntactic completeness, and semantic transparency (Mompean, 2024; Somasundaran et al., 2015). A closer inspection of divergence cases further illustrates this point. Out of 60 recordings, AI scores were more generous than human ratings in six cases (by approximately 1–1.5 points), particularly when segmental errors did not obscure meaning. On the other side, there were three instances where AI ratings were more stringent, in particular when disfluencies caused problems with lexical cohesiveness. These disparities show that, even though there was good alignment generally, there were small variances in how raters viewed the relative importance of prosodic and segmental dimensions in different samples. This divergence aligns with recent research in automated pronunciation assessment, such as the comprehensive review by Kheir et al. (2023) and the meta-analysis by Vančová (2023), both of which emphasize that although AI tools (e.g., CAPT systems) have made significant progress

in identifying segmental errors, they still often fall short in detecting nuanced prosodic features such as stress, rhythm, and intonation. These results also suggest that while AI tools can assist in large-scale pronunciation evaluation, they cannot yet substitute expert human judgments in nuanced assessments—especially in languages like Turkish where suprasegmental interference is pronounced (Uzun, 2022).

This study reinforces the shift from nativeness-based models to intelligibility-based pedagogy (Levis, 2005; Seidlhofer, 2013). Turkish EFL instructors should prioritize activities that promote rhythm, stress, and natural phrasing over segmental correction alone. As Pease (2016) notes, intelligibility-focused training enhances communicative effectiveness without demoralizing learners. Moreover, AI can be used as a supplementary rater in resource-limited contexts, enabling teachers to offer immediate feedback and conduct high-frequency assessments. However, as Mompean (2024) argues, teacher mediation is crucial to contextualize AI-generated feedback, especially when addressing prosodic shortcomings. Ultimately, just as Levis (2020) and Chau et al. (2022) suggest, intelligibility instruction in Turkish EFL settings should evolve toward an ecological and communicative model—prioritizing listener-oriented fluency over phonological perfection and integrating technology to scale and personalize pronunciation support.

While this study contributes to intelligibility literature by combining human and AI evaluations, it had limitations. The relatively small sample size and reliance on a single AI system constrain generalizability. Because of the relatively small sample, the findings are best understood as exploratory in nature. Nonetheless, this pioneering analysis provides empirical evidence of AI–human convergence in intelligibility assessment and establishes directions for future large-scale validation studies. Further studies should include longitudinal data to track intelligibility development, expand across diverse L1 groups to explore universal vs. language-specific patterns, and investigate how AI training models can be fine-tuned for better prosodic perception. Another limitation of this study is related to the way the AI rater was instructed. Although ChatGPT-4 was guided with the same Washington University 7-point intelligibility scale as the human raters, the absence of AI-specific rubrics designed exclusively for suprasegmental features remains a methodological gap. Future research should therefore develop and test AI-tailored rubrics to ensure even greater validity in human–AI intelligibility comparisons.

Overall, the findings confirm that intelligibility judgments in this study were firmly data-driven. Statistical evidence from task comparisons, error-type frequencies, and divergence analyses collectively support the conclusion that intelligibility is shaped by the interaction of prosodic fluency, segmental accuracy, and task context. This empirical grounding strengthens the interpretive claims that follow regarding pedagogy and technology. This study displays that intelligibility is shaped by a constellation of interacting factors—segmental precision, prosodic fluency, task type, and rater lens. Spontaneous speech, despite more phoneme-level errors, often yielded higher ratings due to better prosody and communicative flow. The substantial agreement between AI and human raters affirms the potential role of AI in pronunciation assessment, especially when used to complement—not replace—human evaluations. Future tools must improve in capturing prosodic cues to ensure holistic intelligibility assessment. Future research should aim to validate these findings with larger and more diverse learner populations across different L1 backgrounds. Comparative studies of multiple AI tools would further clarify whether the convergence patterns observed here are unique to ChatGPT-4 or generalizable across systems. Importantly, future work should also examine the pedagogical implications of AI–human rating differences, particularly how discrepancies in weighting segmental versus suprasegmental features may influence classroom feedback and instructional priorities. By pursuing these trajectories, researchers can both advance the empirical basis for AI-assisted intelligibility assessment and deepen its integration into TESOL-oriented pronunciation pedagogy.

References

Aoyama, K., & Guion, S. G. (2007). Prosody in second language acquisition: Acoustic analyses of duration and F0 range. In O.-S. Bohn & M. J. Munro (Eds.), *Language experience in second-language speech learning* (pp. 282–297). John Benjamins.

Babaeian, A. (2023). Pronunciation assessment: Traditional vs modern modes. *Journal of Education For Sustainable Innovation*, *1*(1), 61-68. https://doi.org/10.56916/jesi.v1i1.530

Bayraktaroğlu, S. (2008). Orthographic interference and the teaching of British pronunciation to Turkish learners. *Journal of Language and Linguistic Studies, 4*(2), 1-36.

Celce-Murcia, M., Brinton, D. M., Goodwin, J. M., & Griner, B. (2010). *Teaching pronunciation: A course book and reference guide* (2nd ed.). Cambridge University Press.

Chau, T., Huensch, A., Hoang, Y. K., & Chau, H. T. (2022). The effects of L2 pronunciation instruction on EFL learners' intelligibility and fluency in spontaneous speech. *TESL-EJ, 25*(4), n4.

Demirezen, M. (2005). Rehabilitating a fossilized pronunciation error: The/v/and/w/contrast by using the audio-articulation method in teacher training in Türkiye. *Journal of Language and Linguistic Studies*, *1*(2), 183-192.

Derwing, T. M., & Munro, M. J. (1995). Foreign accent, comprehensibility, and intelligibility in L2 learner speech. *Language Learning, 45*(1), 73–97. https://doi.org/10.1111/j.1467-1770.1995.tb00963.x

Derwing, T. M., & Munro, M. J. (1997). Accent, intelligibility, and comprehensibility: Evidence from four L1s. *Studies in second language acquisition*, *19*(1), 1-16. https://doi.org/10.1017/S0272263197001010

Derwing, T. M., & Munro, M. J. (2009). Putting accent in its place: Rethinking obstacles to communication. *Language teaching*, *42*(4), 476-490. https://doi.org/10.1017/S026144480800551X

Derwing, T. M., & Munro, M. J. (2015). *Pronunciation fundamentals: Evidence-based perspectives for L2 teaching and research.* John Benjamins.

Dörnyei, Z. (2007). *Research methods in applied linguistics*. Oxford university press.

Dikilitaş, K., & Geylanioğlu, S. (2019). Pronunciation Errors of Turkish Learners of English: Conceptualization Theory as a Teaching Method. *Journal of Language Teaching and Learning*, *2*(2), 38-50. Retrieved from https://www.jltl.com.tr/index.php/jltl/article/view/101

Ercan, H. (2018). Pronunciation Problems of Turkish EFL Learners in Northern Cyprus. *International Online Journal of Education and Teaching*, *5*(4), 877-893. *International Online Journal of Education and Teaching*, v5 n4 p877-893 2018

Fairbanks, G. (1960). *Voice and articulation drillbook* (2nd ed.). Harper & Row.

Field, J. (2005). Intelligibility and the listener: The role of lexical stress. *TESOL Quarterly, 39*(3), 399–423. https://doi.org/10.2307/3588487

Foote, J. A., Holtby, A. K., & Derwing, T. M. (2011). Survey of the teaching of pronunciation in adult ESL programs in Canada, 2010. *TESL Canada journal*, 1-22. https://doi.org/10.18806/tesl.v29i1.1086

Geng, H., Saito, D., & Minematsu, N. (2025). A perception-based L2 speech intelligibility Indicator: Leveraging a rater's shadowing and sequence-to-sequence voice conversion. *arXiv preprint arXiv:2505.24304*. https://doi.org/10.48550/arXiv.2505.24304

Hahn, L. (2004). Primary stress and intelligibility: Research to motivate the teaching of suprasegmentals. *TESOL Quarterly, 38*(2), 201–223. https://doi.org/10.2307/3588378

Hismanoglu, M. (2012). Teaching word stress to Turkish EFL learners through Internet-based video lessons. *US–China Education Review A, 1*(26), 26–40.

Jenkins, J. (2000). *The phonology of English as an international language*. Oxford University Press.

Kang, O., Rubin, D., & Pickering, L. (2010). Suprasegmental measures of accentedness and judgments of 1107 language learner proficiency in oral English. *Modern Language Journal*, *94*(4), 554–566. https://doi.org/10.1111/j.1540-4781.2010.01091.x

Khalilzadeh, A. (2014). Phonetic and non-phonetic languages: A contrastive study of English and Turkish phonology focusing on the orthography-induced pronunciation problems of Turkish learners of English as a foreign language (Turkish EFL learners). *International Journal of Languages' Education and Teaching*, *2*(1), 1-16.

Kheir, Y. E., Ali, A., & Chowdhury, S. A. (2023). Automatic pronunciation assessment--a review. *arXiv preprint arXiv:2310.13974*. https://doi.org/10.48550/arXiv.2310.13974

Kormos, J., & Dénes, M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System, 32*(2), 145–164. https://doi.org/10.1016/j.system.2004.01.001

Lee, J., Jang, J., & Plonsky, L. (2015). The effectiveness of second language pronunciation instruction: A meta-analysis. *Applied Linguistics, 36*(3), 345–366. https://doi.org/10.1093/applin/amu040

Levis, J. M. (2005). Changing contexts and shifting paradigms in pronunciation teaching. *TESOL Quarterly, 39*(3), 369–377. https://doi.org/10.2307/3588485

Levis, J. M. (2018). *Intelligibility, oral communication, and the teaching of pronunciation.* Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108241564

Levis, J. (2020). *Revisiting the intelligibility and nativeness principles*. *Journal of Second Language Pronunciation, 6*(3), 310–328. https://doi.org/10.1075/jslp.20050.lev

Lochland, P. (2020). Intelligibility of L2 Speech in ELF. *Australian Journal of Applied Linguistics*, *3*(3), 196-212.

Mompean, J. A. (2024). ChatGPT for L2 pronunciation teaching and learning. *ELT Journal*, *78*(4), 423-434. https://doi.org/10.1093/elt/ccae050

Munro, M. J., & Derwing, T. M. (1995). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning, 45*(1), 73–97. https://doi.org/10.1111/j.1467-1770.1995.tb00963.x

Munro, M. J., & Derwing, T. M. (2020). Foreign accent, comprehensibility and intelligibility, redux. *Journal of Second Language Pronunciation*, *6*(3), 283-309. https://doi.org/10.1075/jslp.20038.mun

Munro, M. J., Derwing, T. M., & Morton, S. L. (2006). The mutual intelligibility of L2 speech. *Studies in second language acquisition*, *28*(1), 111-131. https://doi.org/10.1017/S0272263106060049

Pease, C. (2016). *Accentedness, comprehensibility and intelligibility of L2 speech: A replication and extended study* (Doctoral dissertation, University of York).

Saito, K. (2012). Effects of instruction on L2 pronunciation development: A synthesis of 15 quasi-experimental intervention studies. *TESOL Quarterly, 46*(4), 842–854. https://doi.org/10.1002/tesq.67

Saito, K. (2018). Advanced second language segmental and suprasegmental acquisition. *The handbook of advanced proficiency in second language acquisition*, 282-303. https://doi.org/10.1002/9781119261650.ch15

Seidlhofer, B. (2013). *Understanding English as a lingua franca*. Oxford University Press.

Somasundaran, S., Chen, L., Cheng, X., & Zechner, K. (2015). Exploring content and discourse features for automated speech scoring. *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, 12–21. https://doi.org/10.3115/v1/W15-0602

Uzun, T. (2022). The salient pronunciation errors and intelligibility of Turkish speakers in English. *MEXTESOL Journal, 46*(1), 1–15. https://doi.org/10.61871/mj.v46n1-9

Vančová, H. (2023). AI and AI-powered tools for pronunciation training. *Journal of Language and Cultural Education*, *11*(3), 12-24.

Yavaş, M. (2011). *Applied english phonology* (2nd ed.). Wiley-Blackwell.

Yorkston, K. M., Beukelman, D. R., & Traynor, C. (1984). *Assessment of intelligibility of dysarthric speech*. Austin, TX: Pro-ed.

Zechner, K., Higgins, D., Xi, X., & Williamson, D. M. (2009). Automatic scoring of non-native spontaneous speech in tests of spoken English. *Speech communication*, *51*(10), 883-895. https://doi.org/10.1016/j.specom.2009.04.009

Zou, B., Liviero, S., Ma, Q., Zhang, W., Du, Y., & Xing, P. (2024). Exploring EFL learners' perceived promise and limitations of using an artificial intelligence speech evaluation system for speaking practice. *System*, *126*, 103497. https://doi.org/10.1016/j.specom.2009.04.011

Appendices

Appendix A

The Rainbow Passage

When the sunlight strikes raindrops in the air, they act as a prism and form a rainbow. The rainbow is a division of white light into many beautiful colors. These take the shape of a long round arch, with its path high above, and its two ends apparently beyond the horizon. There is, according to legend, a boiling pot of gold at one end. People look, but no one ever finds it. When a man looks for something beyond his reach, his friends say he is looking for the pot of gold at the end of the rainbow. Throughout the centuries people have explained the rainbow in various ways. Some have accepted it as a miracle without physical explanation. To the Hebrews it was a token that there would be no more universal floods. The Greeks used to imagine that it was a sign from the gods to foretell war or heavy rain. The Norsemen considered the rainbow as a bridge over which the gods passed from earth to their home in the sky. Others have tried to explain the phenomenon physically. Aristotle thought that the rainbow was caused by reflection of the sun's rays by the rain. Since then physicists have found that it is not reflection, but refraction by the raindrops which causes the rainbows. Many complicated ideas about the rainbow have been formed. The difference in the rainbow depends considerably upon the size of the drops, and the width of the colored band increases as the size of the drops increases. The actual primary rainbow observed is said to be the effect of super-imposition of a number of bows. If the red of the second bow falls upon the green of the first, the result is to give a bow with an abnormally wide yellow band, since red and green light when mixed form yellow. This is a very common type of bow, one showing mainly red and yellow, with little or no green or blue.

Appendix B

Picture for the spontaneous speech task

Geniş Özet

Giriş

İngilizcenin küresel iletişim dili olarak artan rolü, İngilizce öğretiminde sesletim becerilerine yönelik yaklaşımların yeniden değerlendirilmesini gerekli kılmıştır. Özellikle anlaşılırlık (intelligibility) kavramı, geleneksel yerli aksana ulaşma hedefinin yerine, anlaşılabilirliğe odaklanmayı savunan pedagojik yaklaşımların merkezine yerleşmiştir (Derwing & Munro, 1995; Jenkins, 2000; Levis, 2005). Bu değişim, özellikle anadili İngilizce olmayan ülkelerde, yabancı dil olarak İngilizce (EFL) öğrenen bireyler için önemli çıkarımlar taşımaktadır. Türkiye'de İngilizce öğretimi genellikle yazılı odaklı bir biçimde ilerlemekte, sesletim eğitimi ise yeterince sistematik ve işlevsel olarak ele alınmamaktadır (Demirezen, 2005; Bayraktaroğlu, 2008). Bu durum, Türk öğrencilerin anlaşılırlığını hem segmental (parçalı ses birimleri) hem de suprasegmental (parçalarüstü ses birimleri) düzeyde olumsuz etkileyebilmektedir. Bu çalışmada, Türk EFL öğrencilerinin İngilizce konuşmalarının anlaşılırlığı hem insan hem de yapay zeka değerlendiricileri tarafından değerlendirilmiş ve iki farklı konuşma görevi (okuma ve resim betimleme) üzerinden analiz edilmiştir.

Sesletim araştırmalarında üç temel kavram öne çıkar: anlaşılırlık, anlaşılabilirlik ve aksanlılık (Munro & Derwing, 1995). Anlaşılırlık, dinleyicinin konuşulanı gerçekten anlayıp anlamadığını ifade ederken; anlaşılabilirlik, anlayışın zorluğu veya kolaylığına ilişkin algıyı belirtir. Aksanlılık ise, konuşmacının telaffuzunun yerli aksan normlarına ne ölçüde yakın olduğunu yansıtır. Son yıllarda, sesletim öğretiminde anlaşılırlığı artırmaya yönelik yaklaşımlar ön plana çıkmıştır (Celce-Murcia et al., 2010). Bu çerçevede, parçalı ses birim doğruluğundan ziyade, vurgu, ezgi ve ritim gibi parçalarüstü ses birim öğelerin iletişim başarısında daha belirleyici olduğu birçok çalışma tarafından ortaya konmuştur (Hahn, 2004; Kang, 2010; Saito, 2018). Ayrıca, Levis (2020) anlaşılırlığın sabit bir özellik olmadığını, konuşmacı ve dinleyici arasındaki etkileşimle belirlendiğini vurgular.

Yapay zeka destekli sistemler de, sesletim değerlendirmelerinde insan değerlendirmelerine alternatif veya tamamlayıcı araçlar olarak giderek daha fazla kullanılmaktadır (Zechner et al., 2009; Mompean, 2024). Bununla birlikte, yapay zeka sistemlerinin parçalarüstü ses birim unsurlarını tanıma ve değerlendirme yetilerinin sınırlı olduğu da vurgulanmaktadır (Zou et al., 2024).

Bu çalışma, Türk EFL öğrencilerinin İngilizce konuşmalarının anlaşılırlığını iki farklı değerlendirme biçimi (insan ve yapay zeka) ve iki farklı görev türü (okuma ve betimleme) üzerinden karşılaştırarak analiz etmeyi amaçlamaktadır. Çalışma şu sorulara yanıt aramıştır: (1) Türk EFL öğrencilerinin İngilizce konuşmalarında anlaşılırlık düzeyleri görev türüne göre (okuma ve resim betimleme) değişiklik göstermekte midir? (2) Yapay zeka ve insan değerlendirmeleri arasında hangi bağlamlarda uyum ya da uyumsuzluk gözlenmektedir?

Yöntem

Araştırma 30 Türk EFL öğrencisiyle gerçekleştirilmiştir. Katılımcılar, biri kontrollü (Rainbow Passage okuma) ve diğeri özgür üretim (resim betimleme) olmak üzere iki farklı konuşma görevi tamamlamıştır. Her görev için ses kayıtları alınmış ve bu kayıtlar üç anadili İngilizce olan değerlendirici ile ChatGPT (GPT-4) tarafından 7'li Likert ölçeğiyle değerlendirilmiştir. Değerlendirmeler, istatistiksel analizler (tanımlayıcı istatistikler, ortalamalar, varyanslar) ve nitel hata analizi yoluyla karşılaştırılmıştır.

Bulgular

Çalışmanın bulguları beş başlık altında sunulmuştur:

1. Resim betimleme görevleri, beklenenin aksine daha fazla parçalı ses birim hataları içerse de daha yüksek anlaşılırlık puanları almıştır. Bu durum, doğal konuşmanın ritim ve vurgu gibi parçalarüstü ses birim öğeleri açısından daha etkili olduğunu göstermektedir.

2. Değerlendirici Uyumları: Yapay zeka ve insan değerlendiriciler arasında genel olarak yüksek düzeyde bir uyum görülmüştür. Ancak sınırda kalan ("orta düzeyde" anlaşılan) örneklerde, yapay zekanın vurgu ve ritim gibi unsurlara duyarsız kaldığı ve insanlarla farklı değerlendirmeler yaptığı belirlenmiştir.

3. Hata Tipolojisi: Parçalı ses birim hataları arasında diş arası ünsüzlerin (/θ/, /ð/) yerine [t]/[d] olarak telaffuzu, -ing ekinde /ŋ/ yerine /n/ kullanımı ve ünlü uzunluk hataları öne çıkmıştır. parçalarüstü ses birim hataları ise düzensiz duraklamalar, yanlış vurgu yerleştirme ve düz intonasyon olarak kodlanmıştır.

4. Yapay Zekanın Güçlü ve Zayıf Yönleri: ChatGPT özellikle kelime seçimi, dilbilgisel tamlık ve semantik bütünlük üzerinden değerlendirme yaparken; insan değerlendiriciler, konuşma doğallığı ve işitsel akışkanlık gibi unsurlara dikkat çekmiştir.

5. Genel Değerlendirme: Yapay zeka ve insan puanları arasındaki ortalama fark çok küçük olmakla birlikte, yapay zekanın parçalar üstü ses birim hatalarını gözden kaçırdığı durumlar mevcuttur. Ancak genel eğilimler ve anlaşılırlık düzeyleri açısından uyumluluk sağlanmıştır.

Tartışma, Sonuç ve Öneriler

Bu çalışma, resim betimleme gibi daha serbest konuşma görevlerinin, parçaı ses birim doğruluğu açısından daha zayıf olmasına rağmen, parçalarüstü ses birim bütünlüğü sayesinde daha yüksek anlaşılırlık sağladığını göstermiştir. Bu bulgu, Hahn (2004), Jenkins (2000) ve Kang et al. (2010) gibi araştırmacıların vurgu ve ritmin anlaşılırlık üzerindeki önemine dikkat çeken çalışmalarını desteklemektedir. Ayrıca, ChatGPT gibi yapay zeka sistemlerinin, insan değerlendirmelerine büyük ölçüde yakın sonuçlar vermesi, bu teknolojilerin öğretim süreçlerinde tamamlayıcı araçlar olarak kullanılabileceğini göstermektedir. Ancak özellikle konuşma akışı ve vurgu gibi prosodik unsurların değerlendirilmesinde halen insan yargısına ihtiyaç duyulmaktadır. Çalışmanın sonuçlarına dayanarak yapay zeka uygulamaları öğretim sürecine entegre edilmesi ancak bu geri bildirimler öğretmen tarafından yorumlanarak öğrenciye açıklanması ve Türkiye'de daha fazla görev-temelli, çok değerlendiricili ve doğal konuşma örnekleri içeren anlaşılırlık çalışmaları yapılması önerilmiştir.

Yayın Etiği Beyanı

Bu araştırmanın, Atatürk Üniversitesi Sosyal ve Beşeri Bilimler Etik Kurulu Başkanlığı kurumu tarafından 20.06.2025 tarihinde E.88656144-000-2500199482 sayılı kararıyla verilen etik kurul izni bulunmaktadır. Bu araştırmanın planlanmasından, uygulanmasına, verilerin toplanmasından verilerin analizine kadar olan tüm süreçte "Yükseköğretim Kurumları Bilimsel Araştırma ve Yayın Etiği Yönergesi" kapsamında uyulması belirtilen tüm kurallara uyulmuştur. Yönergenin ikinci bölümü olan "Bilimsel Araştırma ve Yayın Etiğine Aykırı Eylemler" başlığı altında belirtilen eylemlerden hiçbiri gerçekleştirilmemiştir. Bu araştırmanın yazım sürecinde bilimsel, etik ve alıntı kurallarına uyulmuş; toplanan veriler üzerinde herhangi bir tahrifat yapılmamıştır. Bu çalışma herhangi başka bir akademik yayın ortamına değerlendirme için gönderilmemiştir.

Çatışma Beyanı

Araştırmanın yazarı olarak herhangi bir çıkar/çatışma beyanım olmadığını ifade ederim.