



DOI: 10.22363/2313-1683-2026-23-1-75-87

EDN: IUVVNK

UDC 159.9.072:81'23

Brief research report

Natural and Synthesized Voices in Foreign-Language Listening: A Pilot Study of Text Comprehension and Semantic-Differential Ratings

Mikhail O. Matveev¹, Olga A. Ulyanina^{1,2}

¹ Moscow Institute of Physics and Technology (National Research University), *Dolgoprudny, Russian Federation*

² Moscow State University of Psychology & Education, *Moscow, Russian Federation*
mike-matveev@mail.ru

Abstract. Prosodic organization of speech is relevant to both the comprehension of the spoken text and subjective voice evaluation. This pilot study compared direct comprehension and subjective evaluation of the same English-language educational text voiced by a professional female narrator and a synthesized voice perceived as female. Data from 145 university students (females = 37.9%, males = 62.1%) with an upper-intermediate level of English or higher were analyzed. Each study group listened to only one recording (natural voice: $n = 77$; synthesized voice: $n = 68$). After listening, the participants answered five content-related questions and rated the recording on 12 five-point semantic-differential scales. Direct comprehension scores did not differ between the conditions. In contrast, the semantic-differential ratings varied: after applying the Holm correction, the synthesized recording received less positive ratings on the attractiveness, interest, informativeness, vividness, and scientific validity scales. In the natural-voice condition, factorization revealed components of general evaluation/attractiveness, cognitive accessibility/clarity, and activation. In the synthesized voice condition, the structure was less compact: cognitive-evaluative, activation/richness, and expressive-appeal components were identified, with the fourth eigenvalue also exceeding 1. The obtained data indicate that the two recordings provided similar direct factual comprehension but different subjective ratings. Because the results relate to two specific stimuli, it is impossible to separate the voice source from the speaker's identity and other delivery features.

Keywords: prosody, synthesized speech, listening comprehension, semantic differential, educational audio, foreign-language learning

Funding. The Ministry of Science and Higher Education of the Russian Federation supported the research under agreement No. 075-03-2026-305 of 16 January 2026, within the project Applied Research on the Implementation of Artificial Intelligence Technologies in Higher Education (project code FSMG-2025-0086).



For citation: Matveev, M.O., & Ulyanina, O.A. (2026). Natural and synthesized voices in foreign-language listening: A pilot study of text comprehension and semantic-differential ratings. *RUDN Journal of Psychology and Pedagogics*, 23(1), 75–87. <http://doi.org/10.22363/2313-1683-2026-23-1-75-87> EDN: IUUVNK

Introduction

Spoken language reaches the listener as a structured acoustic event rather than as a sequence of lexical units alone. Pitch changes, rhythm, stress, and pauses mark phrase boundaries, indicate emphasis, and help the listener organize incoming speech (Cole, 2015; Cruttenden, 1997; Cutler et al., 1997; Hirschberg, 2002). Prosodic phrasing can reduce structural ambiguity in sentence comprehension (LaCroix et al., 2019; Schafer et al., 2000). In this functional sense, intoning is part of meaning construction: the same verbal sequence can be perceived differently depending on how emphasis, tempo, and melodic contour are distributed throughout the utterance.

This point is especially relevant to foreign-language listening. Lexical access and syntactic parsing are less automated in a second language, so listeners may rely more on temporal and prosodic cues. Pitch stress can influence what learners remember from spoken discourse (Guan et al., 2022), while speech rate influences perceived comprehensibility and listener evaluation (Munro & Derwing, 1998, 2001). Prosody also conveys emotion and pragmatic stance, which means that the listener processes not only propositional content but also information about how that content is presented (Mauchand et al., 2021; Paulmann & Kotz, 2008).

The voice therefore becomes part of the educational stimulus. Short fragments are sufficient for listeners to form social impressions of the speaker, and those impressions are sensitive to voice pitch and tempo (Apple et al., 1979; Baus et al., 2019; McAleer et al., 2014). In multimedia learning, the speaker's voice can function as a social cue (Mayer et al., 2003). In classroom communication, vocal delivery changes how motivational messages are perceived (Paulmann & Weinstein, 2025). These findings make it possible to identify at least two outcomes of educational listening: whether the learner can reconstruct the factual content and how the learner evaluates the auditory presentation itself.

In the text-to-speech context, this distinction becomes increasingly important. Early work showed increased short-term memory loads when synthesized speech was used (Luce et al., 1983), while more recent studies provided a less consistent picture. Human speech is often preferable, but comprehension depends on the task, the listener, and the particular synthesis used (Curtin & Ayaz, 2017; Gong, 2023; Rodero, 2017). Dylman and colleagues (2025), for example, found that both intonation and voice type could influence listening comprehension in children. Therefore, the practical question is not simply whether synthesized speech is 'better' or 'worse' than natural speech. A synthesized recording can remain intelligible while eliciting different emotional or evaluative responses.

The semantic differential offers a compact way to describe such responses. Osgood and colleagues (1957) showed that bipolar adjective scales are often

organized around broad dimensions, commonly described as Evaluation, Potency, and Activity. However, these dimensions should not be viewed as invariant templates. A task-specific semantic differential is constructed around the rated object, and its factorial organization can vary depending on the stimulus and the group. Voronin (2018), for example, demonstrated that contrasting groups of respondents can construct different semantic structures around the same domain. This is relevant in this case, as the presented adjective pairs were selected to represent several aspects of listening comprehension—overall evaluation and attractiveness, cognitive accessibility and credibility, and activation—yet their empirical organization could have emerged from the ratings.

The present study combines a short comprehension test with a semantic-differential evaluation of two female-voiced realizations of the same English educational text. It is deliberately framed as a pilot study. The study design does not involve manipulation of individual acoustic variables and does not allow for the general effect of natural and synthesized speech to be determined. However, it can address a more specific question of whether the two recordings with identical verbal content produce similar direct comprehension and whether the students describe their listening experience in the same way. Factorization of the semantic-differential ratings is used as a research method to examine how the adjective pairs are organized within each voiced version.

Purpose and hypotheses

The study compared direct comprehension and semantic-differential ratings after the participants listened to the same foreign-language educational text delivered the natural or synthesized female voice.

Hypothesis 1: Direct comprehension will be higher after listening to the natural voice recording than after listening to the synthesized voice recording.

Hypothesis 2: The natural voice recording will receive more positive semantic-differential ratings than the synthesized voice recording.

Methods

Design and stimuli

The pilot analysis used two conditions: a natural female professional voice and a synthesized female voice.

The verbal material was a 789-word English informational text on the history of coffee, its social role, and its physiological effects. The professional recording was taken from a publicly available copy of the BBC video *What does coffee do to your body?*¹. The professional narrator was a woman. The first author transcribed the text and used an online text-to-speech service to generate a synthetic version based on the transcript. The selected synthesized voice sounded female. Both recordings contained the same spoken text.

¹ Learn English through Mass Media. (2024, May 16). *What does coffee do to your body?* BBC World Service.mp4 [Video]. VK Video. https://vkvideo.ru/video-200164765_456240346

The recordings were played in routine classroom settings. Before playback, the instructor asked the students to listen carefully to the recording and independently complete an online form after listening. Written informed consent was obtained before participation. The responses were anonymous.

Participants

The sample included 145 persons, 77 of whom listened to the professional natural voice and 68 listened to the synthesized voice. The participants were students at Moscow Institute of Physics and Technology (MIPT) or Moscow State Linguistic University (MSLU) and had an English proficiency level of upper-intermediate or above, as determined by university placement or achievement testing and program requirements. The stimuli were divided into groups, and each participant listened to only one recording. The main characteristics of the samples are presented in Table 1.

Table 1

Sample characteristics depending on voice conditions

Characteristic	Professional natural voice	Synthesized voice
<i>N</i>	77	68
Age, <i>M</i> (<i>SD</i>)	20.91 (1.70)	19.78 (1.42)
Age, median [IQR]	21 [20–22]	20 [19–21]
Women / men, <i>n</i>	31 / 46	24 / 44
MIPT / MSLU, <i>n</i>	51 / 26	45 / 23
Study year, median [IQR]	4 [2–4]	3 [1–3]

Note. The conditions were similar in gender ($p = 0.539$) and institution ($p = 0.994$), but differed in age ($p < 0.001$) and study year ($p < 0.001$).

Measurements

Five items were scored for direct comprehension. Three were multiple-choice questions and two required short answers (see Appendix). Each correct answer received one point; the total score ranged from 0 to 5. Spelling errors and minor spelling variations were ignored.

The semantic differential contained 12 bipolar scales presented in Russian. The pairs were selected to cover three content domains relevant to auditory evaluation: general evaluation and attractiveness (Useful—Harmful, Attractive—Repulsive, Interesting—Boring, Bright—Dull, Captivating—Ordinary, Beautiful—Prosaic, Rich—Empty), cognitive accessibility and credibility (Simple—Complex, Informative—Confusing, Concrete—Abstract, Scientifically Valid—Dubious), and activation (Calming—Stimulating). This grouping was a content rationale for selecting adjectives, not a predetermined factor model. The procedure followed the general semantic-differential approach of Osgood and colleagues (1957) and the logic of task-specific semantic differentials (Voronin, 2018); it was not intended as a psychodiagnostic scale. Responses ranged from 1 for the left pole to 5 for the right pole. For the rating pairs, lower scores therefore indicated proximity to the more favourable left pole. The Calming—Stimulating pair was interpreted as an activation continuum rather than as a positive—negative pair. The order of the poles was not varied, which could have increased position-related response bias.

Data analysis

The two voice conditions were compared using the Mann—Whitney *U*-test. The Holm correction was applied to all the 12 scale-level comparisons, and the rank-biserial correlation (*rrb*) was used as the effect size. The semantic differential dimension was examined separately in the two conditions. Exploratory principal component analyses utilized the correlation matrix and Varimax rotation. The number of components was examined using the Kaiser criterion (eigenvalues > 1), visual inspection of scree plots, and comparison of two- and three-component solutions (Fabrigar et al., 1999; Voronin, 2018). Because the sample was small and the solutions were not identical across conditions, factorization was interpreted descriptively, and factor scores were not compared between the groups.

Results

Objective comprehension

The professional natural voice group obtained $M = 3.23$ ($SD = 0.89$; median = 3), and the synthesized voice group obtained $M = 3.10$ ($SD = 0.93$; median = 3). The distributions did not differ, $U = 2800.5$, $p = 0.445$, $rrb = 0.070$. Hypothesis 1 was not confirmed.

Semantic-differential ratings

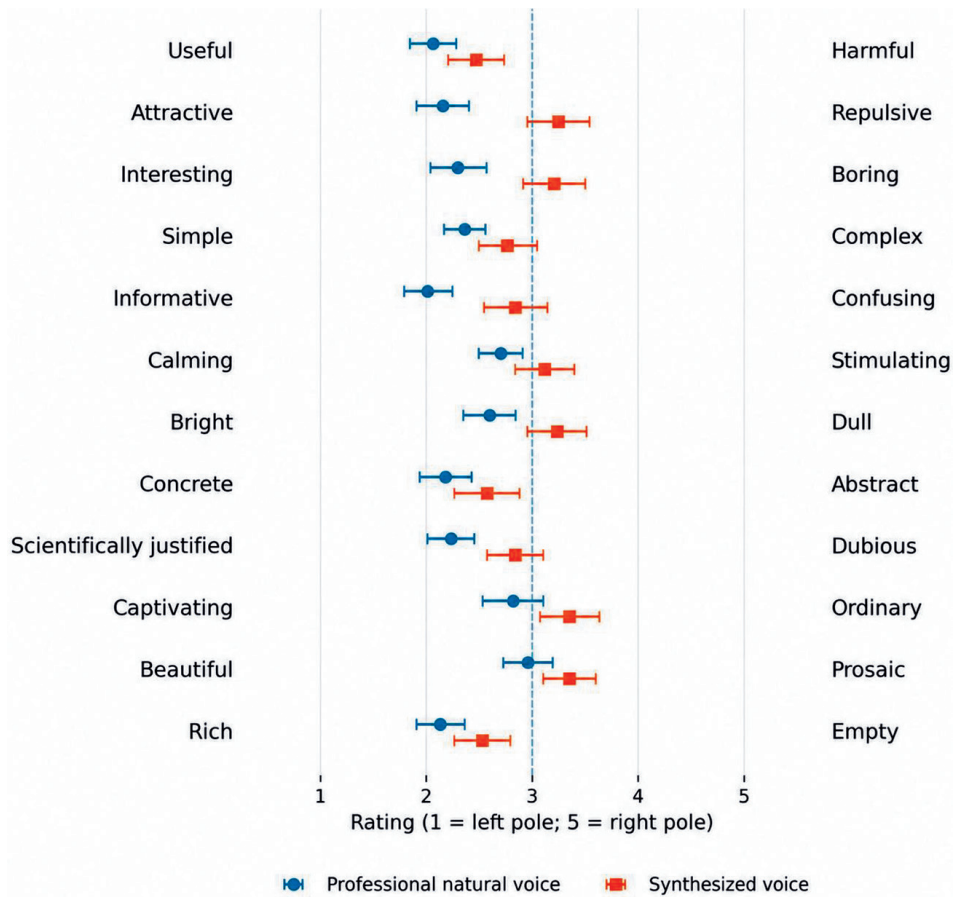
Figure presents mean ratings for all the 12 semantic pairs, and Table 2 provides the corresponding descriptive statistics, Mann—Whitney tests, effect sizes, and Holm-corrected *p* values. The left and right poles are shown directly in the figure. Scores ranged from 1 (left pole) to 5 (right pole); for the rating pairs, lower scores indicate a more favourable evaluation. The largest differences between the conditions concerned attractiveness, interest, informativeness, brightness, and scientific credibility.

Table 2

Semantic-differential ratings depending on the voice conditions

Scale	<i>M</i> (<i>SD</i>)		<i>U</i>	<i>rrb</i>	Holm <i>p</i>
	Professional voice	Synthesized voice			
Useful—Harmful	2.06 (0.99)	2.47 (1.14)	2084.0	–0.204	0.166
Attractive—Repulsive	2.16 (1.10)	3.25 (1.24)	1369.0	–0.477	< 0.001
Interesting—Boring	2.30 (1.18)	3.21 (1.23)	1554.5	–0.406	< 0.001
Simple—Complex	2.36 (0.90)	2.76 (1.13)	2138.5	–0.183	0.166
Informative—Confusing	2.01 (1.04)	2.84 (1.28)	1647.0	–0.371	< 0.001
Calming—Stimulating	2.70 (0.95)	3.12 (1.17)	2111.5	–0.193	0.166
Bright—Dull	2.60 (1.09)	3.24 (1.19)	1848.0	–0.294	0.013
Concrete—Abstract	2.18 (1.10)	2.57 (1.33)	2209.0	–0.156	0.166
Scientifically justified—Dubious	2.23 (1.00)	2.84 (1.11)	1779.0	–0.320	0.005
Captivating—Ordinary	2.82 (1.27)	3.35 (1.21)	1980.5	–0.244	0.068
Beautiful—Prosaic	2.96 (1.09)	3.35 (1.06)	2105.0	–0.196	0.166
Rich—Empty	2.13 (1.04)	2.53 (1.14)	2085.0	–0.204	0.166

Note. Lower values indicate proximity to the adjective on the left. Negative *rrb* values indicate lower ratings in the professional natural voice condition.



Mean semantic-differential ratings depending on the voice conditions with 95% bootstrap confidence intervals

Note. Scores range from 1 (left pole) to 5 (right pole). For evaluative pairs, lower values indicate a more favorable rating; the Calming–Stimulating pair represents activation rather than valence.

Source: created by Mikhail O. Matveev and Olga A. Ulyanina using Python 3.13.5 and Matplotlib 3.10.8.

After the Holm correction, significant differences remained for the pairs: Attractive—Repulsive, Interesting—Boring, Informative—Confusing, Bright—Dull, and Scientifically Valid—Dubious. The synthesized recording received less favorable mean across all the significant comparisons.

Factor structure

In the professional natural voice condition, the Kaiser criterion retained three components (eigenvalues = 5.00, 1.46, and 1.11); together, they explained 63.1% of the variance. In the synthesized voice condition, four eigenvalues exceeded 1 (4.69, 1.34, 1.28, and 1.18). The first three explained 60.9% of the variance. Scree plots showed a dominant first component followed by smaller components in both conditions. Two-component solutions explained 53.8 and 50.3% of the variance and left the Calming–Stimulating scale weakly represented. Therefore, three-component Varimax solutions were inspected as a parsimonious descriptive representation (Tables 3 and 4). The fourth synthesized component was viewed as

evidence that the structure in that condition could be more differentiated than a three-component summary might suggests.

Table 3

Three-component Varimax solution in the professional natural voice condition

Scale	Component 1	Component 2	Component 3
Useful—Harmful	0.627	0.102	0.374
Attractive—Repulsive	0.800	0.266	0.104
Interesting—Boring	0.829	0.155	–0.091
Simple—Complex	–0.070	0.801	0.099
Informative—Confusing	0.485	0.642	–0.134
Calming—Stimulating	–0.029	–0.021	0.903
Bright—Dull	0.726	0.044	–0.253
Concrete—Abstract	0.135	0.684	–0.103
Scientifically justified—Dubious	0.612	0.401	–0.005
Captivating—Ordinary	0.773	–0.053	0.026
Beautiful—Prosaic	0.636	0.024	–0.210
Rich—Empty	0.879	0.040	0.010
Variance explained, %	38.8	15.0	9.3

Note. Loadings with |.40| or greater are shown in bold.

Table 4

Three-component Varimax solution in the synthesized voice condition

Scale	Component 1	Component 2	Component 3
Useful—Harmful	0.788	0.032	0.325
Attractive—Repulsive	0.478	–0.092	0.637
Interesting—Boring	0.650	–0.117	0.497
Simple—Complex	0.123	0.441	0.471
Informative—Confusing	0.793	0.186	0.276
Calming—Stimulating	0.261	0.739	–0.123
Bright—Dull	0.158	–0.172	0.828
Concrete—Abstract	0.675	–0.078	–0.027
Scientifically justified—Dubious	0.750	–0.019	0.140
Captivating—Ordinary	0.181	0.140	0.794
Beautiful—Prosaic	0.374	–0.335	0.387
Rich—Empty	0.471	–0.609	0.000
Variance explained, %	28.5	11.1	21.3

Note. Loadings with |.40| or greater are shown in bold.

In the professional natural voice condition, Component 1 was dominated by general evaluation/attractiveness, Component 2 was dominated by cognitive accessibility/clarity, and Component 3 was dominated by activation. In the

synthesized voice condition, the three-component solution separated cognitive-evaluative descriptors, activation/richness, and expressive attractiveness less cleanly; the fourth eigenvalue above 1 indicates additional residual structure.

Discussion

Both recordings produced similar scores on the five items related to direct comprehension. In this pilot sample, the results do not support the superiority of the professional natural voice for factual comprehension. This finding is compatible with previous research showing that synthesized speech can remain intelligible even when it differs from natural speech in terms of processing requirements or listener evaluation (Curtin & Ayaz, 2017; Rodero, 2017). The result should not be generalized to other voices or longer-term learning outcomes.

The semantic differential revealed a different pattern. After correcting for multiple testing, the students rated the synthesized recording as less attractive, less interesting, less informative, more boring, and less scientifically valid. This result is consistent with previous findings that human and synthesized voices can differ in listener evaluations even when both remain intelligible (Curtin & Ayaz, 2017; Gong, 2023; Luce et al., 1983; Rodero, 2017). Thus, Hypothesis 2 was partially supported: the expected direction was observed across several rating scales, but not on every pair.

The factorization is also informative. The 12 adjective pairs were selected to represent several content domains, but this selection did not predetermine the statistical structure. In the professional natural voice condition, the three-component solution identified a broad evaluation/attractiveness component, a cognitive-accessibility component, and an activation component dominated by Calming–Stimulating. In the synthesized voice condition, the organization was less compact: the Kaiser criterion indicated four components, while the three-component solution only approximately separated cognitive-evaluative descriptors, activation/richness, and expressive attractiveness. Thus, Simple–Complex and Calming–Stimulating do not behave as conventional indicators of the same general evaluative dimension. The difference between the conditions is consistent with the psychosemantic view that the structure of the task-specific semantic differential depends on the rated object and the context of evaluation (Osgood et al., 1957; Voronin, 2018). Given the small group sizes, these factorial solutions should be considered preliminary.

The practical implications are intentionally limited. A brief comprehension assessment revealed no differences between the two recordings, whereas several subjective ratings did. For developers and teachers using synthesized audio in foreign language educational materials, this suggests that a pilot evaluation with a target group of learners could reveal differences in perceived attractiveness, informativeness, brightness, or credibility that a brief factual test might miss. This study does not determine the acoustic characteristic responsible for these differences, nor does it establish a general advantage of natural speech.

Limitations. The study compared one professional natural voice recording with one synthesized voice recording. The voice source was not separated from the speaker's identity, vocal quality, duration, or tempo of speech. Therefore, the results obtained are applicable only to these stimuli.

The semantic-differential poles were presented in a fixed order, which could have led to position-related response bias. The comprehension test contained only five items and assessed direct fact retrieval. The factor analyses were conducted on relatively small subgroups and should therefore be considered preliminary; replication with a larger sample and a balanced pole order is needed.

Conclusion

In this pilot study, a professional natural female voice and synthesized female voice produced similar direct comprehension of the same English educational text but elicited different subjective evaluations. The natural voice recording was rated more favourably on attractiveness, interest, informativeness, brightness, and scientific credibility. Exploratory factor analysis also revealed that the semantic-differential ratings were organized differently in the two conditions. These findings apply to the two recordings used here and cannot be attributed solely to the voice source. However, they support the feasibility of analyzing comprehension and listener evaluation as separate outcomes in foreign-language educational listening.

References

- Apple, W., Streeter, L.A., & Krauss, R.M. (1979). Effects of pitch and speech rate on personal attributions. *Journal of Personality and Social Psychology*, 37(5), 715–727. <https://doi.org/10.1037/0022-3514.37.5.715>
- Baus, C., McAleer, P., Marcoux, K., Belin, P., & Costa, A. (2019). Forming social impressions from voices in native and foreign languages. *Scientific Reports*, 9(1), 414. <https://doi.org/10.1038/s41598-018-36518-6>
- Cole, J. (2015). Prosody in context: A review. *Language, Cognition and Neuroscience*, 30(1-2), 1–31. <https://doi.org/10.1080/23273798.2014.963130>
- Cruttenden, A. (1997). *Intonation* (2nd ed.). Cambridge: Cambridge University Press. <https://doi.org/10.1017/cbo9781139166973>
- Curtin, A., & Ayaz, H. (2017). Cognitive considerations in auditory user interfaces: Neuroergonomic evaluation of synthetic speech comprehension. In D. Harris (Eds.). *Engineering Psychology and Cognitive Ergonomics: Performance, Emotion and Situation Awareness. EPCE 2017. Lecture Notes in Computer Science*, vol. 10275 (pp. 106–116). Cham: Springer. https://doi.org/10.1007/978-3-319-58472-0_9
- Cutler, A., Dahan, D., & van Donselaar, W. (1997). Prosody in the comprehension of spoken language: A literature review. *Language and Speech*, 40(2), 141–201. <https://doi.org/10.1177/002383099704000203>
- Dylman, A.S., Champoux-Larsson, M.-F., & Frances, C. (2025). Prosody! When intonation helps and there is an effect... on listening comprehension in children. *Educational Psychology*, 45(1), 1–17. <https://doi.org/10.1080/01443410.2024.2446778>

- Fabrigar, L.R., Wegener, D.T., MacCallum, R.C., & Strahan, E.J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, 4(3), 272–299. <https://doi.org/10.1037/1082-989x.4.3.272>
- Gong, C. (2023). AI voices reduce cognitive activity? A psychophysiological study of the media effect of AI and human newscasts in Chinese journalism. *Frontiers in Psychology*, 14, 1243078. <https://doi.org/10.3389/fpsyg.2023.1243078>
- Guan, C.Q., Meng, W., Morett, L.M., & Fraundorf, S.H. (2022). Mapping pitch accents to memory representations in spoken discourse among Chinese learners of English: Effects of L2 proficiency and working memory. *Frontiers in Psychology*, 13, 870152. <https://doi.org/10.3389/fpsyg.2022.870152>
- Hirschberg, J. (2002). Communication and prosody: Functional aspects of prosody. *Speech Communication*, 36(1-2), 31–43. [https://doi.org/10.1016/s0167-6393\(01\)00024-3](https://doi.org/10.1016/s0167-6393(01)00024-3)
- LaCroix, A.N., Blumenstein, N., Houlihan, C., & Rogalsky, C. (2019). The effects of prosody on sentence comprehension: Evidence from a neurotypical control group and seven cases of chronic stroke. *Neurocase*, 25(3-4), 106–117. <https://doi.org/10.1080/13554794.2019.1630447>
- Luce, P.A., Feustel, T.C., & Pisoni, D.B. (1983). Capacity demands in short-term memory for synthetic and natural speech. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 25(1), 17–32. <https://doi.org/10.1177/001872088302500102>
- Mauchand, M., Caballero, J.A., Jiang, X., & Pell, M.D. (2021). Immediate online use of prosody reveals the ironic intentions of a speaker: Neurophysiological evidence. *Cognitive, Affective, & Behavioral Neuroscience*, 21(1), 74–92. <https://doi.org/10.3758/s13415-020-00849-7>
- Mayer, R.E., Sobko, K., & Mautone, P.D. (2003). Social cues in multimedia learning: Role of speaker's voice. *Journal of Educational Psychology*, 95(2), 419–425. <https://doi.org/10.1037/0022-0663.95.2.419>
- McAleer, P., Todorov, A., & Belin, P. (2014). How do you say 'Hello'? Personality impressions from brief novel voices. *PLoS ONE*, 9(3), e90779. <https://doi.org/10.1371/journal.pone.0090779>
- Munro, M.J., & Derwing, T.M. (1998). The effects of speaking rate on listener evaluations of native and foreign-accented speech. *Language Learning*, 48(2), 159–182. <https://doi.org/10.1111/1467-9922.00038>
- Munro, M.J., & Derwing, T.M. (2001). Modeling perceptions of the accentedness and comprehensibility of L2 speech: The role of speaking rate. *Studies in Second Language Acquisition*, 23(4), 451–468. <https://doi.org/10.1017/s0272263101004016>
- Osgood, C.E., Suci, G.J., & Tannenbaum, P.H. (1957). *The measurement of meaning*. Urbana, IL: University of Illinois Press.
- Paulmann, S., & Kotz, S.A. (2008). Early emotional prosody perception based on different speaker voices. *NeuroReport*, 19(2), 209–213. <https://doi.org/10.1097/wnr.0b013e3282f454db>
- Paulmann, S., & Weinstein, N. (2025). Motivating tones to enhance education: The effects of vocal awareness on teachers' voices. *British Journal of Educational Psychology*, 95(2), 551–564. <https://doi.org/10.1111/bjep.12737>
- Rodero, E. (2017). Effectiveness, attention, and recall of human and artificial voices in an advertising story. Prosody influence and functions of voices. *Computers in Human Behavior*, 77, 336–346. <https://doi.org/10.1016/j.chb.2017.08.044>
- Schafer, A.J., Speer, S.R., Warren, P., & White, S.D. (2000). Intonational disambiguation in sentence production and comprehension. *Journal of Psycholinguistic Research*, 29(2), 169–182. <https://doi.org/10.1023/a:1005192911512>

Voronin, A.N. (2018). Modifications of semantic space of interpersonal perception of bilingualism at different levels of discursive abilities. *RUDN Journal of Psychology and Pedagogics*, 15(4), 410–430. (In Russ.). <https://doi.org/10.22363/2313-1683-2018-15-4-410-430>

APPENDIX

Comprehension items and scoring key

No.	Item	Correct answer
1	How did coffee houses contribute to social change during the Enlightenment?	They provided venues for business, politics, and sharing ideas
2	What is the main active ingredient in coffee?	Caffeine
3	How does caffeine affect the nervous system?	It blocks adenosine receptors to boost alertness
4	Coffee has been consumed for at least _____ years, shaping global culture.	1,500
5	Drinking too much coffee may cause side effects such as _____ and nervousness.	Insomnia

Authors' contribution:

Mikhail O. Matveev — study design, stimulus preparation, data collection, statistical analysis, data visualization, interpretation of results, initial text preparation, and manuscript editing. *Olga A. Ulyanina* — study design, data collection, interpretation of results, project coordination, scientific supervision, and manuscript writing and editing. All authors have read and approved the final version of the article.

Conflict of interests:

The authors declare that there is no conflict of interest.

Bio notes:

Mikhail O. Matveev, PhD in Philology, Associate Professor, Department of Foreign Languages, Moscow Institute of Physics and Technology (National Research University) (9 Institutsky Lane, Dolgoprudny, Moscow Region, 141701, Russian Federation). ORCID: [0000-0003-3629-1332](https://orcid.org/0000-0003-3629-1332); Scopus Author ID: 57211157583; ResearcherID: AAS-7819-2020; eLibrary SPIN-code: 4248-4335. E-mail: mike-matveev@mail.ru

Olga A. Ulyanina, Dr. Sci. in Psychology, Associate Professor, Head of the Federal Coordination Center for the Development of Psychological and Pedagogical Assistance in the Education System of the Russian Federation, Moscow State University of Psychology & Education (29 Sretenka St, Moscow, 127051, Russian Federation); Chief Research Fellow, Center for Applied Linguistic Research and Testing ISTOK, Moscow Institute of Physics and Technology (National Research University) (9 Institutsky Lane, Dolgoprudny, Moscow Region, 141701, Russian Federation). ORCID: [0000-0001-9300-4825](https://orcid.org/0000-0001-9300-4825); 9283-7824; Scopus Author ID: 57207950411; ResearcherID: AAF-2050-2020; eLibrary SPIN-code. E-mail: ulyaninaoa@mgppu.ru

Article history:

Received 18 January, 2026

Revised 25 January, 2026

Accepted 25 January, 2026

DOI: 10.22363/2313-1683-2026-23-1-75-87

EDN: IUUVNK

УДК 159.9.072:81'23

Краткое сообщение

Естественный и синтезированный голос при аудировании на иностранном языке: пилотажное исследование понимания текста и оценок по семантическому дифференциалу

М.О. Матвеев¹ , О.А. Ульянина^{1, 2} 

¹ Московский физико-технический институт (Национальный исследовательский университет), Долгопрудный, Российская Федерация

² Московский государственный психолого-педагогический университет, Москва, Российская Федерация
mike-matveev@mail.ru

Аннотация. Просодическая организация речи связана как с пониманием звучащего текста, так и с субъективной оценкой голоса. В пилотажном исследовании сопоставлялись непосредственное понимание и субъективная оценка одного и того же англоязычного учебного текста, озвученного профессиональным диктором-женщиной и синтезированным голосом, воспринимаемым как женский. Проанализированы данные 145 студентов (37,9 % женщин, 62,1 % мужчин) с уровнем английского языка не ниже Upper-Intermediate. Каждая учебная группа прослушивала только одну запись (естественный голос: $n = 77$; синтезированный голос: $n = 68$). После аудирования участники отвечали на пять вопросов по содержанию и оценивали запись по 12 пятибалльным шкалам семантического дифференциала. Показатели понимания между условиями не различались ($U = 2800,5$; $p = 0,445$; $r_{rb} = 0,070$). Оценки по семантическому дифференциалу, напротив, различались: после поправки Холма синтезированная запись получила менее положительные оценки по шкалам привлекательности, интереса, информативности, яркости и научной обоснованности. Для естественного голоса факторизация выявила компоненты общей оценки/привлекательности, когнитивной доступности/ясности и активации. Для синтезированного голоса структура оказалась менее компактной: были выделены когнитивно-оценочный компонент, компонент активации/насыщенности и компонент выразительной привлекательности, при этом четвертое собственное значение также превышало 1. Полученные данные показывают, что две записи обеспечили сходное непосредственное понимание фактической информации, но получили разные субъективные оценки. Результаты относятся к двум конкретным стимулам и не позволяют отделить источник голоса от личности говорящего и других особенностей подачи.

Ключевые слова: просодия, синтез речи, аудирование, семантический дифференциал, образовательное аудио, обучение иностранному языку

Финансирование. Министерство науки и высшего образования Российской Федерации поддержало исследование в рамках соглашения № 075-03-2026-305 от 16 января 2026 г. в рамках проекта «Прикладные исследования по внедрению технологий искусственного интеллекта в высшее образование» (шифр научной темы: FSMG-2025-0086).

Для цитирования: Matveev M.O., Ulyanina O.A. Natural and synthesized voices in foreign-language listening: A pilot study of text comprehension and semantic-differential

ratings // Вестник Российского университета дружбы народов. Серия: Психология и педагогика. 2026. Т. 23. № 1. С. 75–87. <http://doi.org/10.22363/2313-1683-2026-23-1-75-87>
EDN: IUUVNK

Вклад авторов:

М.О. Матвеев — планирование исследования, подготовка стимульного материала, сбор данных, статистический анализ, визуализация данных, интерпретация результатов, подготовка первоначального текста и редактирование рукописи. *О.А. Ульянина* — планирование исследования, сбор данных, интерпретация результатов, координация и научное руководство проектом, редактирование рукописи. Все авторы ознакомлены с окончательной версией статьи и одобрили ее.

Заявление о конфликте интересов:

Авторы заявляют об отсутствии конфликта интересов.

Сведения об авторах:

Матвеев Михаил Олегович, кандидат филологических наук, доцент департамента иностранных языков, Московский физико-технический институт (Национальный исследовательский университет) (Российская Федерация, 141701, Московская область, Долгопрудный, Институтский пер., д. 9). ORCID: [0000-0003-3629-1332](https://orcid.org/0000-0003-3629-1332); Scopus Author ID: 57211157583; ResearcherID: AAS-7819-2020; eLibrary SPIN-код: 4248-4335. E-mail: mike-matveev@mail.ru

Ульянина Ольга Александровна, доктор психологических наук, доцент, руководитель Федерального координационного центра по обеспечению развития психолого-педагогической помощи в системе образования Российской Федерации, Московский государственный психолого-педагогический университет (Российская Федерация, 127051, Москва, ул. Сретенка, д. 29); главный научный сотрудник Центра прикладных лингвистических исследований и тестирования «ИСТОК», Московский физико-технический институт (Национальный исследовательский университет) (Российская Федерация, 141701, Московская область, Долгопрудный, Институтский пер., д. 9). ORCID: [0000-0001-9300-4825](https://orcid.org/0000-0001-9300-4825); Scopus Author ID: 57207950411; ResearcherID: AAF-2050-2020; eLibrary SPIN-код: 9283-7824. E-mail: ulyaninaoa@mgppu.ru

История статьи:

Поступила в редакцию 18 января 2026 г.

Доработана после рецензирования 25 января 2026 г.

Принята к печати 25 января 2026 г.