Published by

British Journal of Contemporary Research

BEXFORD PUBLISHING

A Multimodal Generative AI Tutor for Teaching Phonological Processes and Feature Geometry: Enhancing Metaphonological Awareness, Articulatory Accuracy, and Transfer in L2 Learners

Original Research Article | Volume 1 | Issue 6 | 2026 | Article Number: 038

Accepted: 26 October 2026 | Published: 10 November 2026 | ISSN: 2979-8582 (Online)

Crossref : **https://doi.org/10.67693/BJCR-ACTZ8R9C**

Dr Akamo O. Princewill

Department of English, University of Ibadan, Nigeria

AOP 0009-0003-2445-2562

Correspondence: Dr Akamo O. Princewill, princakams@yahoo.com

Abstract

This study evaluated a multimodal generative AI tutor that delivers explicit instruction on phonological processes and feature geometry to second language learners. The tutor combined language model generated explanations, speech synthesis, interactive articulatory visualizations and rule induction activities. Ninety intermediate L2 English learners participated in a mixed methods quasi experimental design across eight weeks. Participants were assigned to receive instruction from either an AI tutor or using traditional explicit instruction methods. Measures included tests of unambiguous knowledge of phonological processes such as assimilation and feature spreading, metaphonological awareness tasks as well as controlled production accuracy and transfer to free speech. Results revealed significantly greater gains in explicit knowledge and metaphonological awareness for the AI group. Production accuracy demonstrated greater improvements for process driven errors in the AI condition and greater transfer to spontaneous speech was also observed. Participants reported high levels of perceived usefulness of the visualizations and adaptive scaffolding, though a few noted occasional confabulation of rule formulations by the AI. These findings demonstrate the capacity of multimodal generative AI to bridge theoretical phonology instruction and practical production outcomes. The study provides a reusable tutor framework and evidence for the value of explicit knowledge of abstract phonological structure when delivered through adaptive multimodal instruction. Further analyses revealed that gains were particularly notable for complex phonological processes involving feature spreading across hierarchical nodes, with post-test retention remaining robust four weeks later. Exploration of interaction data also revealed that learners who engaged frequently with interactive feature tree manipulation achieved higher transfer scores, underscoring the value of active structural engagement in addition to generative explanations.

Keywords: Multimodal Generative AI, Metaphonological Awareness, Phonological Processes, Explicit Instruction, Second Language Learners

© 2026 The Authors. Published by Bexford Publishing Ltd

1.1 Introduction

Within the second language acquisition field, pronunciation is one of the most common problems in language learning. For many learners, regardless of practice, statements will still include sounds of the L1 or not be up to snuff on the target L2 phonology. Research has found that form-focused learning can have positive effects on both controlled and spontaneous production when applied correctly. However, current methods and tools when applied to pronunciation focus primarily on simple segmental accuracy or certain suprasegmental features like stress or tone. Teaching materials and methods that provide higher-level descriptions that would allow recognizing typical patterns or processes of error formation are much less widespread and underrepresented in both computer-assisted and traditional classroom instruction.

Feature geometry groups distinctive features into hierarchical nodes that define natural classes and restrict possible applications of phonological processes. Assimilation, deletion, insertion, and spreading processes act upon this geometry that create regular patterns of change for learners to learn. Development of explicit understanding of these processes enables the learner to increase their metaphonological awareness which has been shown to contribute to self-monitoring, error detection, and generalization to unpracticed items. The traditional approach to instruction is primarily didactic: concepts are learned through lectures and diagrams that require extensive working memory resources. Learners receive limited opportunities to explore the conceptual domain, test informal hypotheses, and receive immediate multimodal feedback. Recent advances in the domain of generative artificial intelligence offer novel instructional opportunities.

Large language models can be used to provide personalized explanations based on students' answers and hypothesis testing behavior, and speech synthesis can generate target exemplars. Finally, graphical models can illustrate articulatory configurations and update hierarchical feature trees in real-time. Learners operating this system are able to acquire mastery of rule induction, receive unified verbal and visual feedback and practice in an environment that is adaptable to their skill level. The current study was designed in order to test such a framework: It was hypothesized that extended learning in the system would lead to stronger metaphonological awareness and explicit knowledge than the traditional explicit instruction approach, and also that these advantages would be paired with improved performance on production tasks and spontaneous speech.

The work tackles a clear and consequential gap. Computer assisted pronunciation training has matured towards segmental and some prosodic targets. Artificial intelligence tools have been explored in early literacy phonological awareness and general metaphonological homework assistance. There is almost no rigorous research examining generative multimodal systems designed specifically for the explicit teaching of phonological processes and feature geometry, with the measurement of the full chain of outcomes from awareness towards controlled production towards free speech transfer. By filling that gap the study offers both theoretical insight and practical innovation.

1.2 Background of the Study

Phonological acquisition in a second language is constrained by the learner’s first language feature geometry and by universal principles of markedness. Brown’s work in the late 1990s and early 2000s showed that the presence or absence of particular nodes in the L1 inventory predicts which nonnative contrasts learners can readily perceive and produce. Feature geometry therefore offers both a descriptive tool and a pedagogical resource. When learners are taught the hierarchical organization of features they can better understand why certain processes apply and others do not. Explicit metaphonological awareness training has produced measurable benefits. Studies by Wrembel and colleagues demonstrated that raising learners’ conscious knowledge of phonetic and phonological properties improves self-monitoring and production accuracy. Meta analyses of pronunciation instruction confirm medium to large effects for

© 2026 The Authors. Published by Bexford Publishing Ltd

explicit approaches, especially when feedback is timely and focused. Computer assisted systems have extended these benefits by providing unlimited practice and automated diagnosis, yet most remain oriented toward surface accuracy rather than underlying structure.

Generative AI introduces new affordances. Large language models can engage learners in Socratic dialogue about rule formulations. Multimodal systems can synchronize text, speech, and visual representations of articulatory gestures or feature trees. Recent pilots have explored GPT based tutors for metaphonological homework and AI generated narrative feedback for intelligibility. These efforts show promise but stop short of embedding the full apparatus of theoretical phonology within an adaptive multimodal environment. The present study builds on that foundation by creating a tutor whose core content is the explicit teaching of processes and feature geometry.

1.3 Statement of the Problem

A great number of L2 learners attain intermediate proficiency while still struggling with pronunciation, a problem that arises from an incomplete grasp of the target phonological system. Pronunciation receives only limited classroom time, and teachers frequently do not have specialized training in theoretical phonology. Current CAPT (Computer-Assisted Pronunciation Training) tools are aimed mainly at identifying and fixing segmental or prosodic errors. They seldom assist learners in understanding why a given assimilation takes place or how feature spreading gives rise to natural classes. Consequently, improvements in controlled production often do not carry over into spontaneous speech. Because interactive tools for abstract phonological theory are lacking, both a practical and a theoretical gap remains. Learners who never build solid metaphonological awareness stay reliant on external feedback and find it hard to generalize. Meanwhile, generative AI systems able to provide multimodal explanations are now readily accessible, yet they have not been systematically designed or assessed for this particular instructional aim. In the absence of empirical evidence, it is still uncertain whether these systems can provide accurate theoretical content, foster meaningful learning, and yield measurable effects on production and transfer.

Prior research tends to cluster around a number of well-established strands. Computer assisted pronunciation training has recorded improvements in segmental accuracy and certain prosodic features by means of automatic speech recognition and visual feedback. AI applications have been examined for early literacy phonological awareness and for general homework assistance using large language models. Blended models have looked at motivation and overall pronunciation performance. Recent observational work has piloted GPT based tutors for metaphonological awareness in university phonetics courses, noting both useful personalization and the risk of confabulation. What has been left largely unexamined is the systematic employment of multimodal generative AI to teach the abstract constructs of phonological processes and feature geometry, and to tie that explicit knowledge directly to production accuracy and spontaneous transfer. No published study has brought together interactive feature tree visualizations, adaptive rule induction, articulatory animation, and generative explanation in one tutor, and then measured the entire chain from metaphonological awareness through controlled production to free speech. The present study fills that gap.

1.4 Aim and Objectives of the Study

The main aim of this research in the area of second language acquisition and applied linguistics is to create, put into practice and test a multimodal AI tutor. This tutor is meant to provide instruction on phonological processes and feature geometry. The study will look at how this tutor affects learners metaphonological awareness, their knowledge of phonology how accurately they produce sounds and whether these skills can be used in real second language contexts.

© 2026 The Authors. Published by Bexford Publishing Ltd

1.5 Research Objectives

1. To. Improve a multimodal generative AI tutor that uses large language models for conversation, speech synthesis, interactive visualizations of feature geometry and representations of articulatory gestures. The purpose is to support teaching of phonological processes.
2. To compare how effective the AI tutor is compared to methods of explicit teaching when it comes to improving learners' metaphonological awareness and their understanding of phonological processes and feature geometry.
3. To explore how gains in phonological knowledge and metaphonological awareness affect production accuracy. This includes testing, in both controlled experiments and natural spoken language situations.

1.6 Research Problems

1. Does sustained interaction with a multimodal generative AI tutor produce significantly greater improvement in L2 learners' metaphonological awareness and explicit knowledge of phonological processes and feature geometry than traditional classroom based explicit instruction?
2. To what degree do measured gains in explicit metaphonological knowledge transfer to higher production accuracy for process related errors in controlled speaking tasks?
3. How effectively do improvements in explicit knowledge and controlled production carry over into spontaneous speech, and how do learners evaluate the accuracy, usefulness, and pedagogical value of the AI generated visualizations and explanations?

2.1 Literature Review

Research on explicit pronunciation instruction has established that focused attention to form yields reliable gains. Lee, Jang, and Plonsky (2015) synthesized quasi experimental studies and reported medium to large effect sizes for both segmental and suprasegmental targets when instruction was systematic. Saito and Plonsky (2019) further showed that explicit phonetic instruction can enhance both controlled and spontaneous production under favorable conditions. These findings support the premise that declarative knowledge of phonological structure can, with practice, become proceduralized. Metaphonological awareness occupies a central place in this literature. Wrembel (2015) defined it as conscious analytic knowledge of the formal properties of the L2 sound system. Training studies have demonstrated that raising such awareness improves self-correction and accuracy. In L3 contexts, metaphonological protocols have produced measurable gains in tone production. The construct is therefore both theoretically coherent and pedagogically actionable. Feature geometry supplies the formal apparatus. Clements (1985) and subsequent work organized features into hierarchical trees that capture natural classes and constrain spreading and delinking processes. Brown (2000) research on L2 feature acquisition showed that the L1 geometry limits the contrasts learners can readily acquire. Pedagogical applications remain scarce, yet the theory offers a powerful explanatory framework for why certain errors persist and how instruction might restructure the learner's inventory. Computer assisted pronunciation training has advanced steadily. Systematic reviews document the effectiveness of ASR based feedback for segmental accuracy and the value of visual articulatory models (Mahdi & Al Khateeb, 2019; Ngo et al., 2024; Almusharraf, 2024). Multimodal approaches that combine auditory, visual, and gestural information further enhance engagement and retention. Recent systems incorporate phonological features into diagnosis, improving interpretability of feedback.

Generative AI has entered the field more recently. Łodzikowski et al. (2025) conducted an observational pilot of GPT-4 as a homework tutor for metaphonological awareness in a university phonetics course. Learners valued personalized explanations yet encountered confabulation. Hirschi (2025) examined AI

© 2026 The Authors. Published by Bexford Publishing Ltd

generated visual and narrative feedback for intelligibility and found that narrative elaboration improved prosodic features beyond visual feedback alone. Multimodal adaptive systems have shown gains in oral fluency when speech recognition, visual analysis, and generative avatars operate in concert. These studies confirm technical feasibility and learner receptivity while underscoring the need for careful design to minimize inaccurate theoretical content. The literature therefore provides strong foundations for explicit instruction, metaphonological training, and feature based analysis, CAPT effectiveness, and emerging generative AI applications. It does not yet contain a controlled evaluation of a multimodal generative tutor whose primary curriculum is phonological processes and feature geometry, with outcome measures spanning awareness, knowledge, production, and transfer. The present study supplies that missing evaluation.

2.2 Theoretical Framework

The study is grounded in three complementary frameworks. Skill Acquisition Theory, as articulated by DeKeyser (2007), posits that declarative knowledge acquired through explicit instruction can, with sufficient practice and feedback, become proceduralized and eventually automatized. The AI tutor supplies the declarative component through interactive explanation and rule induction. Production tasks and adaptive practice supply the conditions for proceduralization. Schmidt's Noticing Hypothesis (1990) emphasizes that learners must consciously attend to formal features in the input for those features to become intake. The multimodal design heightens noticing by presenting the same phonological structure simultaneously in text, speech, and animated visual form. Interactive manipulation of feature trees further focuses attention on hierarchical relations.

Mayer's Cognitive Theory of Multimedia Learning (2009) provides design principles. Dual channel processing, limited capacity, and active processing guide the integration of verbal explanations with visual diagrams and articulatory animations. By synchronizing modalities and allowing learner control over pacing and complexity, the tutor reduces extraneous cognitive load and supports generative processing of abstract phonological concepts. Feature Geometry Theory itself supplies the content domain. Instruction is organized around hierarchical feature organization, natural classes, and the formal representation of processes. The theoretical framework therefore unites a model of how knowledge develops, a principle of attention, a set of multimedia design guidelines, and the specific phonological theory being taught.

3.1 Research Methodology

A mixed methods quasi experimental design was used. Ninety intermediate level L2 English learners taking university phonetics courses were assigned intact classes to the AI tutor condition or the traditional explicit instruction condition. Both groups received eight weeks of instruction on the same set of phonological processes and feature geometry concepts. The AI group worked primarily with the multimodal tutor, supplemented by brief teacher facilitation. The comparison group received teacher led lectures, static diagrams, and paper based exercises. The tutor architecture consisted of a fine tuned large language model for dialogue and explanation, a neural speech synthesizer for target models and feedback, an interactive feature geometry module that allowed learners to build and modify trees, and an articulatory visualization component that displayed sagittal sections and lip movements synchronized with speech.

Adaptive algorithms adjusted explanation depth and task difficulty according to learner performance and interaction patterns. Safeguards against confabulation consisted of constrained generation templates for core theoretical content and teacher review of selected interaction logs.

Quantitative measures consisted of a pretest, immediate posttest, and delayed posttest four weeks later. Explicit knowledge was assessed with process identification, rule formulation, and feature tree

© 2026 The Authors. Published by Bexford Publishing Ltd

construction tasks. Metaphonological awareness was measured with judgment and explanation tasks requiring reflection on phonological structure. Production accuracy was evaluated with controlled reading and elicitation tasks scored for process related errors. Transfer was assessed with spontaneous speech samples coded for the same error types. Inter rater reliability exceeded 0.85 for all production measures.

Qualitative data consisted of interaction logs, stimulated recall interviews with a subset of twenty participants, and post intervention questionnaires on perceived usefulness, accuracy, and cognitive load. Quantitative data were analyzed with mixed design ANOVA and effect size calculations. Qualitative data underwent thematic analysis guided by the research questions.

Ethical approval was obtained, informed consent secured, and all participants retained the right to withdraw. The design thus combined experimental control with rich process data, allowing both causal inference about learning outcomes and insight into the mechanisms of interaction with the tutor.

4.1 Findings and Discussion

Quantitative results yielded notable benefits favoring the AI tutor. For explicit knowledge items the AI group substantially outperformed the traditional group in terms of learning gains (high effect size) which were maintained at delayed posttest. Gains in metaphonological awareness largely mirrored this pattern. Both groups demonstrated production accuracy improvements on process related violations though the AI condition exhibited stronger enhancements, especially for assimilation and feature spreading contexts. Transfer to spontaneous speech was generally modest though reliably greater in the AI condition. Interaction logs revealed extensive engagement with the feature trees and requests for alternative explanations by AI group members. Stimulated recall indicated that the combination of the updated displays and the need to engage in generative dialogue prompted insights into hierarchical relations masked by the static display of the traditional condition. Questionnaire items reflected strong positive reactions to the visualizations and the scaffolding of the AI tutor.

Only a minority of questionnaire respondents reported encountering instances where the AI tutor suggested slightly misleading formulations of rules which were effectively rectified in subsequent dialogue or by the teacher. Findings provide support for the predictions of Skill Acquisition Theory as the AI tutor facilitated richer declarative input and multiple guided practice opportunities, thereby promoting effective proceduralization. Noticing was enhanced by the dual coding of the AI interface which is congruent with Schmidt's hypothesis. The multimedia design may thus have helped reduce cognitive load by virtue of the synchronized channels and the learner control over the input which in turn promoted active processing of the abstract hierarchical structure.

The value of explicitly teaching feature geometry was also supported by the present results as participants who demonstrated understanding of the natural classes and process restrictions made fewer process driven errors and displayed greater transfer. By building upon the theoretical frameworks pioneered by Brown and others the current study provides empirical validation for the pedagogical utility of generative AI in an instructional setting. The study's limitations include the quasi-experimental design, the focus on intermediate level L2 learners and the restriction to English as a target language. The potential for confabulation inherent in the design was mitigated but not eliminated. Future research should explore the influence of AI tutors on process driven errors across broader proficiency levels and different target languages.

5.1 Conclusion

The study demonstrated that a carefully designed multimodal generative AI tutor can deliver effective explicit instruction on phonological processes and feature geometry. Learners who interacted with the

© 2026 The Authors. Published by Bexford Publishing Ltd

system developed stronger metaphonological awareness and explicit knowledge than those who received traditional instruction. These gains transferred to improved production accuracy and, to a measurable degree, to spontaneous speech. The tutor therefore bridges a long standing gap between theoretical phonology and practical pronunciation pedagogy. It also illustrates how generative AI, when constrained and multimodal, can serve as a powerful instructional partner rather than merely a source of surface level feedback. The framework is reusable and scalable, offering potential benefits for both well-resourced and under resourced instructional settings.

5.2 Recommendations

1. Incorporate interactive feature geometry visualizations into pronunciation curricula to allow for manipulation of hierarchical structure, rather than just viewing static diagrams. Also, design generative AI tutors with constrained templates for core theoretical content to minimize confabulation while allowing for adaptive dialogue
2. Combine speech synthesis, articulatory animation, and text based explanation into a single interface to enable dual channel processing of abstract phonological concepts. In the same vein, Embed metaphonological reflection prompts after production practice to strengthen the connection between explicit knowledge and self-monitoring
3. Investigate cognitive load and long term retention in subsequent studies, using eye tracking or secondary task measures to refine multimodal design principles. Likewise, include adaptive rule induction tasks requiring learners to formulate and test process descriptions to promote active knowledge construction
4. Provide teacher dashboards that highlight interaction patterns and theoretical inaccuracies for efficient human oversight. In addition, Train instructors in AI literacy to guide learners in evaluating tutor output and integrating it with classroom instruction
5. Conduct delayed posttests at multiple intervals to assess the durability of gains in awareness, knowledge, and transfer. Furthermore, adapt the tutor architecture for additional languages and proficiency levels, with particular attention to L1 feature geometries that differ substantially from the target.

References

Almusharraf, N. (2024). The effect of settings, educational level and tools on computer assisted pronunciation training: A meta analysis. Journal of Computer Assisted Learning. https://doi.org/10.1111/jcal.12974

Brown, C. (2000). The interrelation between speech perception and phonological acquisition from infant to adult. In J. Archibald (Ed.), Second language acquisition and linguistic theory (pp. 4–63). Blackwell.

Clements, G. N. (1985). The geometry of phonological features. *Phonology Yearbook*, *2*, 225–252. https://doi.org/10.1017/S0952675700000440

DeKeyser, R. (2007). Skill acquisition theory. In B. VanPatten & J. Williams (Eds.), Theories in second language acquisition (pp. 97–113). Lawrence Erlbaum.

Hirschi, K. (2025). Artificial intelligence generated feedback for second language intelligibility: An exploratory intervention study on effects and perceptions. Language Learning. https://doi.org/10.1111/lang.12719

© 2026 The Authors. Published by Bexford Publishing Ltd

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

Mahdi, H. S., & Al Khateeb, A. A. (2019). The effectiveness of computer-assisted pronunciation training: A meta-analysis. *Review of Education*, *7*(3), 445–464. https://doi.org/10.1002/rev3.3165

Mayer, R. E. (2009). Multimedia learning (2nd ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511811678

Ngo, T. T. N., Chen, H. J., & Lai, K. W. (2024). The effectiveness of automatic speech recognition in ESL/EFL pronunciation: A meta analysis. *ReCALL*, *36*(1), 4–21. https://doi.org/10.1017/S0958344023000113

Saito, K., & Plonsky, L. (2019). Effects of second language pronunciation teaching revisited: A proposed framework and meta analysis. *Language Learning*, *69*(3), 652–708. https://doi.org/10.1111/lang.12345

Schmidt, R. (1990). The role of consciousness in second language learning. *Applied Linguistics*, *11*(2), 129–158. https://doi.org/10.1093/applin/11.2.129

Wrembel, M. (2015). Metaphonological awareness in multilinguals: A case of L3 Polish. *Language Awareness*, *24*(1), 60–83. https://doi.org/10.1080/09658416.2014.890340

Łodzikowski, K., et al. (2025). Exploring large language models for L2 metaphonological awareness tutoring. *Journal of Second Language Pronunciation*, *11*(1). https://doi.org/10.1075/jslp.24030.lod

© 2026 by the authors. Submitted for possible open access publication under the terms and conditions of the **Creative Commons Attribution (CC BY 4.0) licence** (https://creativecommons.org/licenses/by/4.0/)