PIAS: Transparent Multimodal Assessment of Presentation Delivery

Nina Hosseini-Kivanani
University of Luxembourg / RTL
Luxembourg
nina.hosseinikivanani@ext.uni.lu

Nafiseh Taghva
Shiraz University
Iran
taghvanafiseh@gmail.com

Oliver Niebuhr
University of Southern Denmark
Denmark
olni@sdu.dk

ABSTRACT

Oral presentation training requires feedback on delivery dimensions such as pacing, vocal variation, and gesture, yet many automated systems rely on opaque models that are difficult to align with pedagogical rubrics. We present the Prosodic Impact Analysis System (PIAS), an interpretable multimodal framework for formative presentation analytics. Its core component, the Prosodic Impact Index (PII), is a transparent 0–100 score derived from prosodic and coarse gesture features. We evaluate the system on 33 short presentations from 11 native Persian speakers recorded under three delivery conditions: seated (SIT), standing (STA), and gesture–encouraged (MOD). PII increases systematically across conditions and correlates strongly with expert ratings of overall presentation impact. A speaker–independent classifier further shows that combining audio and gesture cues improves condition discrimination over unimodal baselines. These findings position PIAS as a compact and auditable component for educational speaking–feedback systems.

Keywords

multimodal learning analytics, presentation skills, prosody, gesture, interpretable assessment

1. INTRODUCTION

Oral presentation skills are central in many educational settings, including classroom presentations, oral examinations, project defenses, and professional communication training. In such contexts, delivery quality is often assessed through rubric dimensions such as pacing, vocal variety, clarity, audience engagement, and physical expressiveness [14, 3]. These dimensions are pedagogically meaningful because they reflect observable behaviors that learners can practice and improve over time and can provide signals relevant to learner performance and coaching [15]. However, they are also difficult to assess consistently and at scale. Instructors may differ in how they interpret the same performance, and learners often receive feedback that is broad or impressionistic rather than analytically grounded. This creates a need for automated support tools that provide interpretable and pedagogically meaningful feedback rather than opaque end–to–end scores [7, 5].

This need is particularly salient in formative assessment settings, where the aim is not simply to assign a score, but to help learners understand which aspects of their performance are effective and which require improvement [15]. In presentation training, feedback is most useful when it can be connected to concrete delivery traits such as pacing, vocal variation, rhythm, and visible expressiveness [12]. Systems that output only a global score or a class label may be operationally efficient, but they provide limited support for coaching, reflection, and self–regulated learning [4]. For this reason, useful analytics should not only detect performance differences, but also preserve traceability between measured signals and feedback dimensions.

Prior work in automated speaking assessment, presentation–support systems, and multimodal behavior analysis has shown that vocal and visual cues can capture aspects of engagement, fluency, expressiveness, and delivery quality [14, 3, 17]. These studies demonstrate that acoustic and gestural signals carry useful information for computational support tools in presentation training. Prosodic features are especially relevant because they correspond closely to delivery traits that instructors frequently comment on directly. Measures related to pitch variation, speaking rate, rhythm, pause structure, and voice quality have long been associated with listener perceptions of engagement, emphasis, and communicative impact [2, 6]. Gesture contributes an additional visible channel, often reflecting confidence, emphasis, and bodily expressiveness [10, 8]. In educational and presentation settings, even relatively simple gesture cues can complement vocal analysis by capturing whether speakers remain physically constrained or adopt a more dynamic delivery style [3, 14].

At the same time, multimodal learning analytics has increasingly emphasized the importance of explainability, pedagogical alignment, and learner agency [7, 5]. For educational deployment, it is not sufficient for a model to achieve strong predictive performance if its outputs cannot be related back to interpretable traits or meaningful feedback dimensions. Many existing speaking–analysis systems are designed primarily for prediction, ranking, or end–to–end scoring, and therefore provide limited insight into which observable de-

Nina Hosseini-Kivanani, Nafiseh Taghva, and Oliver Niebuhr. PIAS: Transparent Multimodal Assessment of Presentation Delivery. In Anthony Botelho, Maria Mercedes T. Rodrigo, Adish Singla, Hiroaki Ogata, Hyojeong So, and Young Hoan Cho (eds.) Proceedings of the 19th International Conference on Educational Data Mining, Seoul, Republic of Korea, June, 2026, pp. 721–726. International Educational Data Mining Society (2026).

https://doi.org/10.5281/zenodo.21039784

livery traits are driving the final output. This is where an important gap remains. While prior work establishes the value of prosodic and gestural information, relatively fewer approaches aim to produce compact multimodal indices that remain auditable at the component level and can be directly connected to formative feedback [3].

We address this need with the Prosodic Impact Analysis System (PIAS), a multimodal framework for interpretable presentation analytics. Its core component, the Prosodic Impact Index (PII), is a deterministic score that combines prosodic and coarse gesture cues while preserving feature–level traceability. We evaluate PIAS on 33 short presentations by 11 Persian speakers across three delivery conditions, seated (SIT), standing (STA), and gesture–encouraged (MOD).

This paper contributes a compact, interpretable multimodal framework for presentation–skills analytics that links auditable delivery traits to pedagogically meaningful feedback dimensions. The contribution is therefore not a new black–box modeling architecture, but a transparent scoring–and–validation layer that makes multimodal delivery analytics inspectable, lightweight, and easier to align with formative assessment practice. We focus here on the scoring layer and its initial validation, while leaving richer gesture modeling, uncertainty analysis, and generated feedback to future work.

We address two questions in this study: (I) whether an interpretable multimodal index can capture systematic differences in delivery across controlled presentation conditions, and (II) whether the resulting prosodic and gestural cues support speaker–independent discrimination and align with expert judgments of presentation impact.

2. METHOD

Figure 1 summarizes the overall PIAS workflow from multimodal recording to scoring and evaluation.

2.1 Data and procedure

The dataset contains 33 short presentations from 11 native Persian speakers, comprising 6 female and 5 male speakers, aged 25–45. All participants had prior experience with academic or professional presentation contexts. Each participant delivered the same short slide–based presentation under three conditions: seated (SIT), standing (STA), and gesture–encouraged (MOD), yielding one recording per condition. The order of the seated and standing conditions was randomized across speakers, while the gesture–encouraged condition was always presented last because it required an explicit gesture–stimulation manipulation. This repeated–measures design allows delivery style to vary while keeping the speaker and content constant across conditions.

The stimulus consisted of three content slides on the same topic, with short controlled text and visual prompts. Using a shared presentation script reduces semantic variability and allows the analysis to focus on delivery rather than content differences. A closing slide was included for naturalness but excluded from analysis. Recordings were conducted in a controlled setting with synchronized audio and video capture. Persian was selected as a controlled testbed because it provides a consistent single–language setting for initial validation, while its prosodic structure, commonly described as stress–accented and syllable–timed, is well documented in prior phonetic work [16, 13].

The three conditions were selected to induce different levels of expressiveness. SIT represents a relatively constrained delivery mode, STA introduces posture change without explicit movement prompting, and MOD encourages more dynamic visible expression. This design is particularly useful for an initial proof of concept because it creates structured within–speaker contrasts that can reveal whether the proposed index is sensitive to meaningful changes in delivery behavior.

2.2 Interpretable scoring

PIAS centers on the Prosodic Impact Index (PII), a transparent composite score scaled to 0–100. PII combines 12 interpretable components spanning prosody and coarse gesture activity. These components capture interpretable dimensions of delivery, including pitch dynamics, two speaking–rate measures, timing regularity, percentage of voiced material, spectral brightness, voice–quality proxies, and two coarse gesture indicators: gesture rate and gesture count. Components are normalized and combined using fixed predefined weights so that the final score remains auditable at the feature level.

The fixed weights were specified a priori to reflect delivery dimensions emphasized in prior phonetic and presentation–analysis literature, rather than optimized for predictive performance [14, 11]. In this sense, PII is intended as an interpretable summary of delivery behavior rather than as a black–box prediction target. To assess sensitivity to weighting choices, we sampled 5,000 random weight vectors from a Dirichlet distribution. In 99.9% of cases, the monotone SIT $<$ STA $<$ MOD ordering was preserved. Equal weights yielded the same Kendall's W as the original specification. These results suggest that the main condition–ordering finding is robust to reasonable variation in the weighting scheme.

The prosodic component groups were chosen to reflect dimensions that are both measurable and educationally meaningful. Pitch dynamics capture vocal variation. Speaking rate and timing regularity reflect pacing and rhythmic control. Spectral brightness and voice–related measures provide coarse proxies for vocal energy and quality. The gesture component is intentionally simple. Rather than modeling full–body kinematics or fine–grained gesture taxonomy, we use communicative gesture count and gesture rate as minimal visible indicators of bodily expressiveness [10, 8].

Gestures were annotated as communicative hand or arm strokes overlapping with speech, while non-communicative movements such as self-adaptors, scratching, or touching one's hair were excluded. Movements that did not partly or fully coincide with the speech signal were also excluded. Gesture annotation and counting were conducted in ELAN using condition-specific annotation tiers for SIT, STA, and MOD. All annotations and counts were checked by two annotators, yielding substantial inter-annotator agreement (Cohen's $\kappa = 0.83$); disagreement cases were discussed until a final consensus count was reached.

For the speech channel, utterance–level timing was obtained

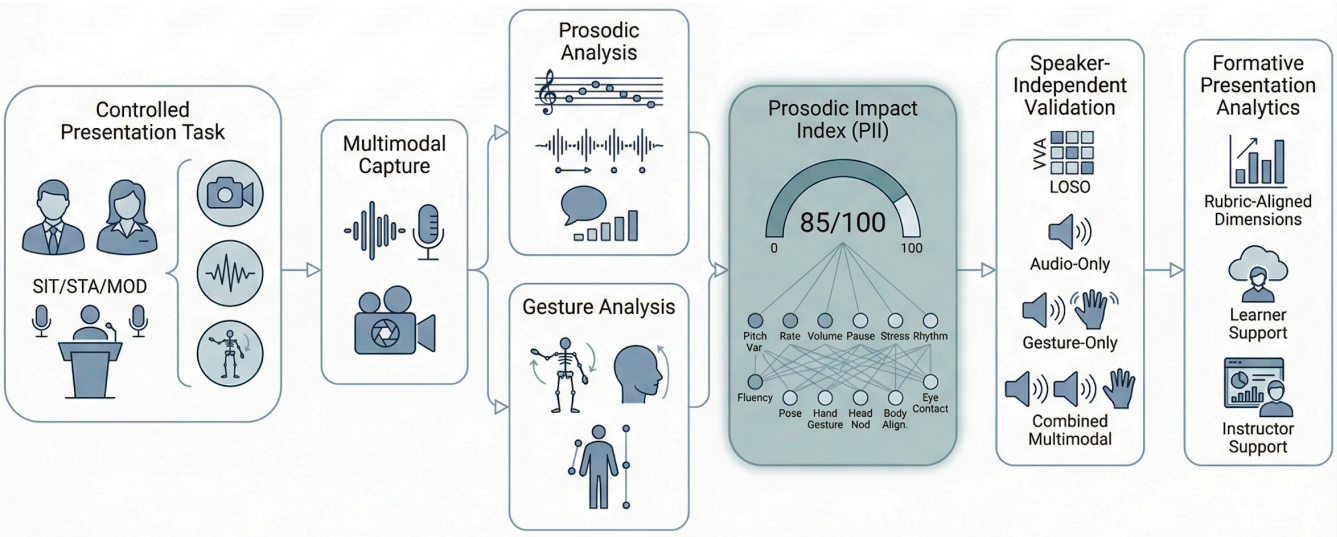

Figure 1: Overview of PIAS. Audio and video recordings are processed into prosodic and coarse gesture measures, combined into the PII, and evaluated through condition comparison, expert validation, and speaker–independent classification.

by forced alignment with the Persian WebMAUS configuration, and prosodic and voice–related measures were then extracted with standard phonetic analysis tools, including Praat–based routines [2, 9, 6]. The resulting index is therefore transparent at two levels. First, each component corresponds to a recognizable delivery dimension. Second, the score is deterministic once the component values are extracted.

In parallel to PII, we extracted a broader feature set for speaker–independent classification. This auxiliary representation includes acoustic features related to rhythm, timing, pitch, and voice quality, together with two gesture–based indicators, gesture count and gesture rate. The classifier was used only as a complementary analysis and was not trained to predict PII from its own components. This separation avoids circularity and allows us to treat PII as a transparent scoring layer, while using classification only to test whether multimodal information generalizes across unseen speakers.

2.3 Evaluation

We evaluated PIAS in three ways. **First**, we tested whether PII differentiates the three delivery conditions using repeated–measures nonparametric statistics. Since each speaker appears in all three conditions, this analysis directly assesses whether the score behaves in the expected direction under controlled within–speaker contrasts.

Second, we examined convergent validity with expert judgments. Twenty trained instructors rated the presentations on a Likert scale using a study–specific protocol focused on overall presentation impact, and the aggregated ratings were correlated with PII. To reduce order effects in the listening experiment, presentation stimuli were shown to raters in a counterbalanced order. This step is important because a delivery index may be statistically sensitive to condition differences without necessarily aligning with human perceptions of presentation quality. The human–rating analysis, therefore, serves as a pedagogically relevant validation layer.

Third, we tested speaker–independent discrimination using a Random Forest classifier under LOSO cross–validation with three feature configurations, audio–only, gesture–only, and combined, to determine whether multimodal information improves condition recognition across unseen speakers. The classification task is not intended as an end use in itself, but as a complementary test of whether multimodal delivery cues differentiate controlled presentation states across unseen speakers. LOSO is appropriate here because it prevents the model from being tested on the same speaker it was trained on. The primary classification metric was balanced accuracy, which is suitable for multi–class comparison under a small sample setting.

Taken together, these evaluation steps map directly onto the two research questions. Condition comparison and expert correlation assess whether PII behaves as an interpretable and educationally meaningful index of delivery, while the LOSO classification analysis tests whether the broader multimodal cue set captures controlled delivery differences in a speaker–independent manner.

3. RESULTS

3.1 PII differentiates delivery conditions

PII differed substantially across delivery conditions. Mean scores were lowest for SIT (13.5 ± 13.1), higher for STA (29.5 ± 18.5), and highest for MOD (76.6 ± 15.2). A Friedman test confirmed a significant condition effect, $\chi^2(2) = 14.36$, $p = .0008$, indicating that the index captures systematic variation from constrained to more expressive delivery. Figure 2 visualizes this progression.

This pattern is important for two reasons. **First**, it shows that the index is responsive to experimentally induced delivery differences rather than remaining flat across conditions. **Second**, the ordering is pedagogically sensible. The seated condition corresponds to the most constrained speaking mode, while the gesture–encouraged condition supports

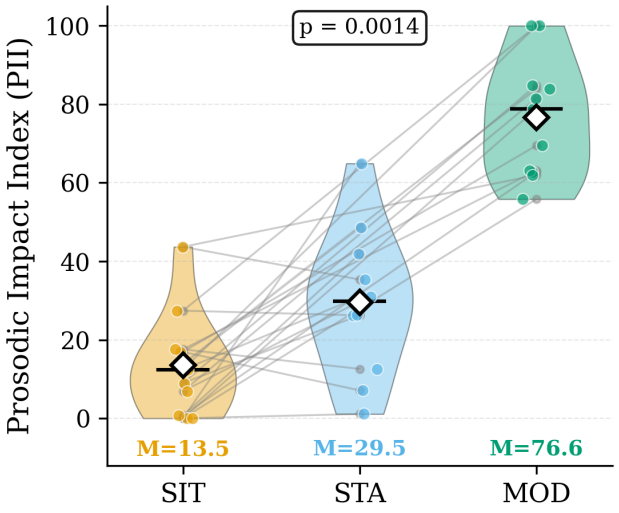

Figure 2: PII distributions across seated, standing, and gesture–encouraged conditions.

the greatest degree of visible and vocal expressiveness. The strong increase in PII from SIT to MOD therefore supports the intended interpretation of the score as a marker of delivery impact rather than merely a descriptive summary of unrelated features.

3.2 PII aligns with expert ratings

To assess convergent validity, trained instructors evaluated each presentation for overall impact. Automated PII scores correlated strongly with human ratings (Spearman $\rho = 0.84$, $p < .001$), indicating that the index captures delivery dimensions that expert evaluators also perceive as impactful. This human alignment is important for educational use because it suggests that the score reflects recognizable presentation traits rather than only dataset–specific patterns.

The strength of this relationship also supports the educational relevance of the component design. PII is intentionally constructed from observable dimensions such as pacing, vocal variation, and bodily expressiveness. The high correlation with expert judgments suggests that these dimensions are not only measurable, but also meaningful within human evaluation practice. In other words, the score appears to summarize aspects of delivery that trained raters already use, implicitly or explicitly, when forming judgments about presentation impact.

3.3 Multimodal cues improve speaker–independent discrimination

As a complementary speaker–independent analysis, we trained Random Forest classifiers under LOSO cross–validation to distinguish the three delivery conditions. The combined multimodal model achieved the best balanced accuracy, 63.6%, outperforming audio–only, 51.5%, and gesture–only, 45.5%, baselines. Although this performance is not sufficient for high–stakes classification, it shows that prosodic and gesture cues provide complementary information that generalizes across speakers.

Table 1 provides a compact summary of the condition effect, expert alignment, and speaker–independent classification results. Taken together, these findings support both research questions. Panel A shows that PII is sensitive to controlled differences in delivery and aligns strongly with expert judgments, while Panel B indicates that the underlying multimodal cues retain speaker–independent discriminative value beyond the handcrafted score itself.

Table 1: Core results for the proposed framework.

Panel A. PII validity	
SIT mean ± SD	13.5 ± 13.1
STA mean ± SD	29.5 ± 18.5
MOD mean ± SD	76.6 ± 15.2
Friedman test	$\chi^2(2) = 14.36,\ p = .0008$
PII vs. expert ratings	$\rho = 0.84,\ p < .001$
Panel B. Speaker–independent classification	
Audio–only	51.5%
Gesture–only	45.5%
Combined	**63.6%**

Note. Panel B reports balanced accuracy under LOSO evaluation.

The ablation pattern is informative. Audio–only features capture useful prosodic differences across all three conditions, while gesture–only features are more limited when visible movement is sparse or constrained. Their combination produces the strongest speaker–independent result, suggesting that the two modalities contribute different but complementary evidence. For the purposes of this paper, the classifier is not the main contribution. Rather, it provides supporting evidence that multimodal signals carry generalizable information beyond a single handcrafted score.

4. DISCUSSION AND LIMITATIONS

As summarized in Table 1, PIAS differentiated constrained and expressive delivery conditions, aligned strongly with expert judgments, and showed moderate speaker–independent classification performance. Together, these findings suggest that transparent prosodic and gestural cues can support formative assessment of presentation delivery, even in a compact controlled setting.

PIAS contributes not only by detecting condition differences, but by doing so through auditable components that correspond to recognizable feedback dimensions. The score is constructed from traits such as pacing, pitch variation, timing regularity, and visible expressiveness, all of which relate closely to how presentation delivery is commonly discussed in instructional practice. This makes the framework better suited to formative use than to summative or high–stakes evaluation. In learner–facing settings, such an approach could support rehearsal tools, reflective dashboards, or instructor summaries that point to specific aspects of delivery rather than returning only a global score, consistent with current work on explainable and pedagogically aligned multimodal learning analytics [7, 5].

The results also illustrate the value of separating interpretable scoring from complementary predictive modeling. In PIAS, PII functions as the primary analytic layer, designed to remain transparent and pedagogically traceable. The speaker–

independent classifier plays a secondary role by testing whether multimodal information generalizes across unseen speakers. The fact that the combined model outperformed the unimodal baselines suggests that prosodic and gestural cues provide complementary evidence, while the strong correlation between PII and expert ratings suggests that the handcrafted score captures dimensions that human evaluators also perceive as relevant. This combination of interpretability and multimodal support is especially useful in educational contexts, where the goal is often not merely to detect performance differences, but to explain them in ways that can inform practice [14, 3, 17].

Several limitations should be noted. The study is based on a small sample and a single language, and the gesture representation is intentionally coarse. The participant group also consisted of speakers with prior academic or professional presentation experience. As a result, the current findings may not transfer directly to novice students, who may show different anxiety patterns, pacing problems, gesture behavior, or content-delivery trade-offs during classroom presentations. These constraints are acceptable for an initial proof of concept, but they mean that the current results should not be interpreted as a complete model of presentation quality across contexts. It also remains unclear to what extent the current component configuration transfers unchanged across languages, presentation genres, and cultural speaking norms [11, 1].

PIAS should not be interpreted as a complete assessment of presentation quality. It measures delivery–related expressiveness and vocal–gestural impact, but it does not evaluate content accuracy, argument structure, slide design, audience understanding, or rhetorical coherence. A speaker could therefore receive a high PII score while still giving a weak presentation in terms of content or organization. For educational deployment, PIAS should be used as a formative delivery–support layer rather than as a standalone judgment of whether a talk is good or bad.

A further direction concerns the connection between interpretable scoring and feedback generation. The present paper focuses on the measurement layer. A natural next step is to transform trait profiles into short rubric–aligned comments that learners can act on, while preserving the transparency of the underlying analytics. For example, a learner with low pitch variation and irregular pacing could receive feedback such as: "Try adding stronger pitch movement around key points and reduce long pauses between slide transitions." Similarly, a low gesture–rate profile could be mapped to feedback encouraging deliberate hand gestures during emphasis points. These examples illustrate the intended formative use of PIAS, although the present study does not yet evaluate whether such feedback improves later presentations.

In summary, PIAS offers a compact and auditable framework for multimodal presentation analytics. By combining transparent scoring with initial human and speaker–independent validation, it provides a promising foundation for educational speaking–feedback systems that prioritize interpretability, pedagogical alignment, and formative use.

5. CONCLUSION AND FUTURE WORK

This paper introduced PIAS as a compact and interpretable framework for multimodal presentation analysis. Across controlled delivery conditions, PII showed clear condition sensitivity and strong alignment with expert judgments, while complementary LOSO classification indicated that the underlying multimodal cues retain speaker–independent discriminative value. Future work should evaluate PIAS on broader and less controlled presentation datasets, incorporate richer gesture descriptors, and test whether trait–based feedback supports measurable improvement in presentation practice over time.

6. ACKNOWLEDGMENTS

This research was conducted in the context of the LuxVoice project, funded by the Luxembourg National Research Fund (FNR) under grant agreement (project reference 19205922). LuxVoice aims to advance Luxembourgish language technologies and support the development of robust multilingual AI resources. The authors gratefully acknowledge the support of RTL Lëtzebuerg and the University of Luxembourg. This work contributes to broader efforts toward inclusive, reliable, and linguistically grounded AI systems for Luxembourgish and other low–resource language settings.

References

[1] F. Biadsy, A. Rosenberg, R. Carlson, J. Hirschberg, and E. Strangert. A cross-cultural comparison of american, palestinian, and swedish perception of charismatic speech. In *Speech prosody*, volume 37, pages 579–82, 2008.

[2] P. Boersma. Praat, a system for doing phonetics by computer. *Glot. Int.*, 5(9):341–345, 2001.

[3] L. Chen, G. Feng, C. W. Leong, J. Joe, C. Kitchen, and C. M. Lee. Designing an automated assessment of public speaking skills using multimodal cues. *Journal of Learning Analytics*, 3(2):261–281, 2016.

[4] N. Fourati, A. Barkar, M. Dragée, L. Danthon-Lefebvre, and M. Chollet. Probing experts' perspectives on ai-assisted public speaking training. *arXiv preprint arXiv:2507.07930*, 2025.

[5] G. Ghane, S. Ghiyasvandian, A. M. Chekeni, and R. Karimi. Revolutionizing nursing education and care: the role of artificial intelligence in nursing. *Nurse Author & Editor*, 34(1):e12057, 2024.

[6] M. Gordon and P. Ladefoged. Phonation types: a cross-linguistic overview. *Journal of phonetics*, 29(4):383–406, 2001.

[7] J. D. Guerrero-Sosa, F. P. Romero, V. H. Menéndez-Domínguez, J. Serrano-Guerrero, A. Montoro-Montarroso, and J. A. Olivas. A comprehensive review of multimodal analysis in education. *Applied Sciences*, 15(11):5896, 2025.

[8] T. Jenkins and W. Pouw. Gesture–speech coupling in persons with aphasia: A kinematic-acoustic analysis. *Journal of Experimental Psychology: General*, 152(5):1469, 2023.

[9] T. Kisler, U. Reichel, and F. Schiel. Multilingual processing of speech via web services. *Computer Speech & Language*, 45:326–347, 2017.

[10] J. P. Momsen and S. Coulson. Decoding prosodic information from motion capture data: the gravity of cospeech gestures. *Open Mind*, 9:652–664, 2025.

[11] O. Niebuhr, J. Neitsch, and J. Michalsky. Akustisches charisma profiling: Auf dem weg zur digitalen rhetorik. *DEGA Akustik Journal*, 20(2):7–22, 2020.

[12] X. Ochoa and H. Zhao. Openopaf: An open-source multimodal system for automated feedback for oral presentations. *Journal of Learning Analytics*, 11(3):224–248, 2024.

[13] V. Sadeghi. A phonetic study of vowel reduction in persian. *Language Related Research*, 6(3):165–187, 2015.

[14] J. Schneider, D. Börner, P. Van Rosmalen, and M. Specht. Presentation trainer: what experts and computers can tell about your nonverbal communication. *Journal of Computer Assisted Learning*, 33(2):164–177, 2017.

[15] Ö. Sümer, C. Beyan, F. Ruth, O. Kramer, U. Trautwein, and E. Kasneci. Estimating presentation competence using multimodal nonverbal behavioral cues. *arXiv preprint arXiv:2105.02636*, 2021.

[16] N. Taghva, A. Moloodi, and V. Abolhasani Zadeh. Acoustic correlations of speech rhythms in persian based on variability of between-speakers characteristics. *Journal of Researches in Linguistics*, 12(2):27–50, 2020.

[17] T. Wörtwein, M. Chollet, B. Schauerte, L.-P. Morency, R. Stiefelhagen, and S. Scherer. Multimodal public speaking performance assessment. In *Proceedings of the 2015 acm on International Conference on Multimodal Interaction*, pages 43–50, 2015.