

More than Words: Modeling What Teachers Praise and How They Praise It

Jiseung Yoo Paiheng Xu Jing Liu

University of Maryland, College Park, MD 20742, USA

{jyoo20, paiheng, jliu28}@umd.edu

Abstract

Praise is one of the most frequent evaluative acts in teaching, but what a praise turn communicates depends less on its words than on how it is delivered. Text-based measures cannot see this channel, and absolute acoustic measures confound stable teacher style with event-level delivery. An expressive teacher sounds animated during routine instruction, while a reserved teacher marks praise with a small change in pitch or timing. We propose measuring delivery as *self-deviation*, comparing a praise turn against two references drawn from the speaker themselves, the teacher’s other praise turns (Global Reference) and the speech immediately surrounding the turn (Local Reference). We instantiate this in SERMA (Self-Reference Modeling of Affective Delivery), a low-resource model built over frozen SSL speech features.

1 Introduction

Teachers praise constantly. “Good job,” “Nice work,” and “Great thinking” recur throughout classroom and tutoring transcripts. Counting these expressions is straightforward, but determining what they communicate is not. A single praising utterance does several things at once. It carries information about what the teacher valued, evaluates the student against some standard of worth, and conveys the teacher’s affect (Brophy, 1981; Brummelman et al., 2016; Stout, 2020). Because these functions travel together, the meaning of praise is not fixed in its words. It is a situated act whose effect depends on timing, on how the student perceives it, and above all on how the teacher delivers it (Morton et al., 2022; Wolff and Brechmann, 2015). Praising effort rather than results is generally more motivating, yet older students can read the same praise as a signal of low ability (Amemiya and Wang, 2018).

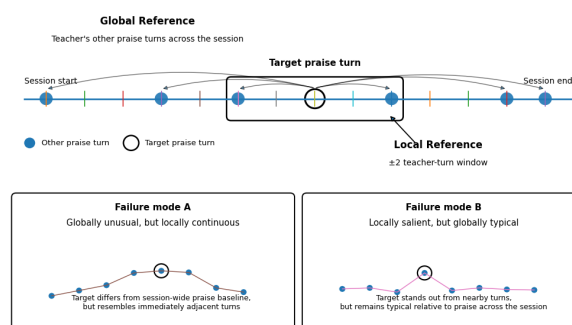


Figure 1: Conceptual scheme of global and local references.

Textually identical utterances differ sharply in felt authenticity (Burnett, 2002; Weinstein et al., 2020; Paulmann and Weinstein, 2025).

Measurement has lagged behind this understanding. Early work treated praise as reinforcement and counted frequencies (Becker et al., 1967). Attribution research turned to what is praised (Mueller and Dweck, 1998), and work on behavior-specific praise showed that naming the recognized action communicates more than general approval (Polick et al., 2012; Floress et al., 2018), while self-determination theory made quality rather than quantity the central concern (Ryan and Deci, 2000; Soenens and Vansteenkiste, 2020). The empirical literature nonetheless remains largely quantitative, counting praise events or coding wording as general versus behavior-specific (Chen et al., 2025). Text-based NLP can identify praise phrases at scale, but transcripts erase the acoustic channel through which teachers mark significance, warmth, and immediacy, and existing systems cannot distinguish a “good job” delivered warmly from the same words spoken flatly (Chen et al., 2025; Göllner et al., 2025).

The channel is known to be informative. Prosody carries warmth, energy, emphasis, and restraint

(Paulmann and Weinstein, 2025), and vocal information improves both human judgments and model performance beyond text alone (Onishi et al., 2020). What is missing is a way to turn that signal into a measure, because there is no universal acoustic definition of good praise. Autonomy-supportive delivery may be calm and restrained while other positive engagement is energetic. Absolute acoustic values therefore confound stable teacher style with event-specific delivery, which is precisely the quantity of interest.

Speech emotion recognition (SER) offers a route to separating the two. Norm theory holds that judgment compares an observed event against an available standard (Kahneman and Miller, 1986), and speaker-sensitive SER systems evaluate a turn against the speaker’s own distribution rather than a population threshold. The gain comes from what normalization removes, since a pitch or intensity value that signals arousal in one voice falls within the resting range of another, and only the residual after conditioning on the speaker is comparable across speakers. Conversational models extend the same logic to surrounding talk, showing that speaker identity and local context jointly determine how a turn is heard (Shi et al., 2025; Du et al., 2025; Zhang and Tan, 2025; Jeong and Nyamathi, 2026). Together these imply that enthusiastic delivery is not an absolute property of a voice but an event-level deviation from the speaker’s own pattern against the surrounding talk (Martinez-Lucas et al., 2026; Tamir et al., 2025).

Speaker normalization is not itself new (Gat et al., 2022; Shi et al., 2025). What the construct contributes is a principled answer to a question that normalization alone leaves open: what to normalize against. We argue that the reference set must be selected by the construct rather than by convenience. Because the question is whether this praise stands out among the ways this teacher normally praises, the Global Reference is the teacher’s other praise turns rather than their speech in general, which would fold the ordinary prosodic contrast between instruction and evaluation into the measure and report it as enthusiasm; work on conversational annotation likewise conditions judgments on the speaker’s own characteristics. The Local Reference addresses a complementary failure mode. A turn can be globally unusual yet locally continuous, as when the teacher has been animated for the past minute, or locally salient yet unremarkable for a routinely expressive teacher; speakers align their

prosody to the immediate talk rather than to a fixed personal baseline, so the surrounding discourse is the operative norm at that moment (Ostrand and Chodroff, 2021). This two-axis, construct-selected design is what separates the measure from a generic emotion classifier, which registers whether a voice is loud or high rather than whether a turn departs from the speaker’s own pattern.

Our contributions are threefold. (i) We present a measurement framework that operationalizes praise delivery through theory-informed reference-relative representations. Theory fixes the construct and its reference relation before any model is fit, the representation encodes that relation, and predictive accuracy is then used to test the representation rather than to select it. (ii) Because praise is both what is said and how it is said, we develop annotation rubrics and an identification pipeline for both the linguistic and the acoustic channel, and we use the linguistic channel to define the praise-relative reference set. Modeling the two channels jointly is the natural next step and is not attempted here. (iii) We develop SERMA, a low-resource method combining frozen self-supervised speech features, two reference channels, teacher-grouped evaluation, and strategies for limited annotation and severe class imbalance (Ueda et al., 2026; Naini et al., 2025; Gong et al., 2025). Freezing the backbone reduces overfitting and annotation cost, which matters because classroom corpora are hard to annotate; layer-wise analyses of frozen SSL features show that mid-layer prosodic content with statistical pooling, retaining the variance, carries the within-turn variability that enthusiasm lives in (Maji et al., 2024; Uniyal and Abrol, 2025), and naturalistic benchmarks establish that imbalance-aware objectives and uncertainty-aware label handling are needed when positive events are scarce (Omid and Hansen, 2026).

This paper reports an in-progress instrument. We present the construct, the annotation protocol, and a first evaluation of reference-relative representations against absolute-feature and zero-shot baselines under leave-one-teacher-out cross-validation. The ablation that zeroes both deviation channels, and the paired cluster-bootstrap intervals over per-fold differences, are specified below but not yet run, so the results here establish an ordering across a family of representations rather than a causal account of where the gain originates.

2 Data and Annotation

2.1 Data

We analyze 157 elementary tutoring sessions from five teachers in an online early-literacy program. All taught Grades 2 to 3 in groups of two or three. Sessions averaged around 40 minutes. Audio came from Zoom recordings with speaker-attributed, timestamped human transcripts, yielding 21,843 teacher turns from a total of 46,437 turns. The setting suits measurement development because it offers sustained interaction, repeated evaluative exchanges, and many observations per teacher. Praise is frequent. Teachers produced 0.84 praise turns per minute, and 24.7% of teacher turns contained a praise expression.

2.2 Annotation

We annotate two constructs, one over the words of a praise turn and one over its delivery. Table 1 summarizes both.

For **linguistic specificity**, four coders worked over a teacher-stratified random sample of 800 candidate turns. They first judged whether each candidate functioned as praise, and the turns judged to be praise were then coded for specificity. A turn counts as specific if it names an action or effort (C1), echoes the student’s contribution (C2), or personalizes recognition by name or pronoun (C3). Meeting any one criterion sufficed. Coders saw three preceding turns and one following turn. First-round agreement was Fleiss’ kappa 0.803 for the praise decision and 0.743 for specificity, and rose to 0.854 and 0.797 after a rubric revision round. Per-criterion agreement for both rounds appears in Table A1.

For **acoustic enthusiasm**, three coders rated 757 praise turns on a five-point scale anchored at no enthusiasm (1) and high enthusiasm (5), with the neutral midpoint (3) removed, leaving four options, so that coders committed to the lower or upper half of the range. Coders heard audio with synchronized transcripts. Turns were batched by teacher and session, so each coder heard a teacher’s full set before rating any single turn. First-stage agreement was moderate (Fleiss’ kappa 0.471). After a teacher-grouped pairwise calibration stage on 201 difficult turns, agreement over the full set reached 0.629. The final annotation gave 113 positive and 644 negative turns, a positive rate of 14.9%.

2.3 Label–construct alignment, and what it costs

Since coders heard a teacher’s full set before rating any single turn, the labels are themselves speaker-relative, which could advantage a speaker-relative model by construction. We treat this as a consequence of the construct: if enthusiasm is defined relative to a speaker, a protocol that induces that standard is the right operationalization, whereas shuffling turns across teachers would elicit judgments of absolute vocal arousal. The cost is that the present results cannot separate how much of SERMA’s advantage comes from the construct and how much from the protocol that produced the labels. We plan to re-rate a subset of turns under a shuffled protocol to estimate how far judgments move when the speaker-relative standard is removed.

3 Method

3.1 Linguistic pipeline

A two-stage LLM pipeline, prompted with the rubric and few-shot examples, identifies praise and its specificity. For each target utterance we build a window of three preceding turns, the target, and one following turn, matching the coders’ context. Stage 1 scans for evaluative language and extracts candidate phrases with temporal boundaries. Stage 2 decides whether each candidate functions as praise, then labels validated praise as specific or general by C1 to C3. Table 2 reports accuracy against the 800 human-coded candidate turns, and all downstream experiments use the praise decisions of the selected system there. Because these decisions are made without reference to enthusiasm, they do not leak the target construct.

3.2 Acoustic pipeline

3.2.1 Utterance encoder

Each turn is encoded from frozen SSL speech frames, WavLM layers 0 to 8, $X_i \in \mathbb{R}^{T_i \times D}$ ($D = 1024$). We restrict to lower layers because they retain more prosodic and less lexical information, and restricting rather than letting the mixture discover this is a regularization choice given 757 labels. A learned softmax over per-layer logits mixes retained layers, so layer selection is optimized with the downstream model while SSL weights stay fixed. Frames are projected to $h = 64$ and passed to a one-layer unidirectional GRU, conditioning

Table 1: Annotation taxonomy.

Channel	Component	What it marks	Example
Linguistic (what is praised)	Action verb (C1)	Names the specific action or effort, supporting autonomy by making the recognized behavior explicit	“You solved the problem perfectly” (Kamana et al., 2022; Stevens et al., 2011)
	Echo (C2)	Repeats or rephrases the student’s contribution, supporting competence by reflecting the work back	Teacher restates the student’s answer with affirmation (Brophy, 1981; Granger et al., 2024)
	Personal (C3)	Names the student or addresses them directly, supporting relatedness	“Great job explaining that, Sarah” (Kamana et al., 2022; Rila et al., 2024)
Acoustic (how it is delivered)	Expressive delivery	Variation in pitch, volume, and tempo, with rise–fall contours and sustained intensity at the peak	A wider pitch range than the speaker’s baseline, roughly 72 to 147 Hz (Weyman and Sy, 2018; Wolff and Brechmann, 2015)
	Contrast (global and local)	Departure from how this teacher usually praises and from the speech immediately surrounding the turn, giving selective signaling and the perceptual pop effect	Moderate to high intensity shift against the speaker’s own praise baseline and against the adjacent turns (Paulmann and Weinstein, 2025)

each frame on preceding acoustics. This frame-level module and the turn-level reference encoder described below are the two components varied across the model family, and SERMA-Transformer replaces both with transformer encoders. The hidden sequence is pooled by a masked-softmax scorer that weights valid frames, and the weighted mean and standard deviation are concatenated and projected to $z_i \in \mathbb{R}^h$. Retaining the second moment preserves within-turn variability, which mean pooling discards and which carries much of the prosodic dynamics of enthusiasm. Crucially, z_i is purely acoustic, and no lexical information enters, so the deviations below compare prosody against prosodic baselines rather than semantic drift. Reference turns are encoded by this same frame encoder with shared weights, so reference embeddings occupy the same space as the target. Sequence-length handling, dropout, and remaining dimensions are in Appendix B.

3.2.2 Reference sets

The two reference sets use deliberately different scopes because they answer different questions. The **Global Reference** set comprises the teacher’s own other praise turns, uniformly subsampled to at most $K_g = 24$ and excluding the target. The construct asks whether this praise stands out among the ways this teacher normally praises, so the baseline

must be teacher praise rather than teacher speech in general. The **Local Reference** set is a symmetric window of $K_l = 4$ teacher turns, the two preceding and the two following, capturing whether the turn stands out against surrounding speech at that moment. Both sizes were fixed before evaluation and never tuned on a held-out teacher. Neither set uses enthusiasm labels. Local membership derives from timestamps and speaker metadata alone, and global membership additionally uses the automatic praise decision, which is made independently of the target construct.

3.2.3 Deviations and availability gating

Reference embeddings are averaged into z_l and z_g , and the deviations $d_l = z - z_l$ and $d_g = z - z_g$ are each encoded as $[d, |d|]$ by a linear-GELU layer. The signed and absolute parts are both retained because direction and magnitude of departure are separately informative. Encoded deviations are then rescaled by availability gates computed from a descriptor of reference-set existence and size (Appendix B), so a deviation resting on few references is attenuated rather than consumed as a fabricated zero. Gating is load-bearing here, because a praise-filtered global set is sparse far more often than one drawn from all of a teacher’s speech. Writing l and g for the gated encoded local and global deviations, the two deviations recombine with the target

through a residual projection $c = z + W[z, l, g]$.

3.2.4 Context, fusion, and the two heads

A turn-level encoder, a GRU in SERMA and a transformer encoder in SERMA-Transformer, reads the ordered Global Reference sequence terminated by the target, and its output at the target position is projected and added to c . This encoder supplies how the teacher’s praise delivery drifts across a session, rather than local discourse context, which the local deviation already carries. Where transcripts exist, a linear projection of a 774-d per-turn RoBERTa embedding is added at this same point, after the acoustic deviations, keeping the reference pathway audio-only and the lexical contribution separable, with the consequence that the target head stays audio-only even in the audio+text condition. Every text and audio+text row reported below uses this same text encoder. Two linear heads then train jointly, a **target head** reading z alone and a **context head** reading c . The loss is the mean of the two (BCE, or class-weighted cross-entropy with inverse-frequency weights fit on the training fold), and at inference the two head probabilities are averaged to give the score that is calibrated and thresholded as described below. The target head is not redundant. Keeping a path from acoustics to prediction that carries no reference information means the model is never *forced* to define enthusiasm as unusualness for this teacher, and if absolute delivery is what annotators responded to, that path can express it.

4 Experimental Setup

4.1 Protocol

We use leave-one-teacher-out cross-validation. Each fold trains on four teachers and tests on the fifth, so no teacher appears in both splits. This matches the construct, since a model measuring deviation from a teacher’s own baseline should transfer to a new teacher with no labeled data from them. Within each fold, 20% of training turns (approximately 120) form a calibration set used only for calibration and threshold selection. The systems differ in how a score is produced. For the fixed-feature representations (Praat, WavLM-mean, FixedRef) the classifier is L2-regularized logistic regression (balanced weights, $C = 1.0$, in-fold standardization), chosen for stability with high-dimensional features and few labels. The SERMA variants are trained end-to-end with the two prediction heads

and joint loss described above, and their averaged head probability is the score. Given the 14.9% positive rate, we apply sigmoid (Platt) calibration, since isotonic overfits a calibration set this small; we then sweep the threshold from 0.1 to 0.9 in steps of 0.05 and apply the best value once to the held-out teacher. The zero-shot scores are thresholded on the same calibration folds under this same procedure, using no held-out labels. With roughly 18 positive turns in a calibration set, this sweep is itself a source of fold-to-fold variance, and a fixed rule that sets the threshold at the training base rate is an alternative we have not yet compared against.

4.2 Compared systems

Praat uses hand-crafted descriptors (Appendix B). *WavLM-mean* uses mean-pooled frozen WavLM frames with no sequence model. *FixedRef* implements the same two reference relations in a fixed basis. Frames are projected by an unsupervised PCA to 64 dimensions, global and local deviations are taken in that fixed space and retained at full dimensionality, KL-divergence statistics over the two reference distributions and their interaction are appended alongside prosody-shape descriptors and an absolute content channel, and the resulting stack is classified by the same logistic regression used for the other fixed-feature representations. *FixedRef* is therefore the control for how the relations are encoded rather than for whether they matter at all, which is the role of the remove-both ablation. *SERMA-Transformer* replaces both recurrent modules, the frame-level encoder and the turn-level reference encoder, with transformer encoders (Appendix B).

Three zero-shot systems run on the same held-out teachers. **ZS-Point** prompts GPT-5.4 to score a single turn from Praat z -scores rendered as prosody descriptions. **ZS-Pair-Audio** gives Gemini-3.5-flash-lite two praise turns from the same teacher, with audio and transcripts, and asks which is more enthusiastic, aggregating preferences by regularized Bradley–Terry. **ZS-Pair-Praat** repeats that protocol with Praat z -scores as text in place of audio. ZS-Pair-Audio versus ZS-Pair-Praat isolates audio perception from reasoning within one model. ZS-Point versus ZS-Pair-Praat compares pointwise with pairwise judgment, though these differ in vendor as well, so the comparison is not fully isolated.

4.3 Metrics

The primary metric is positive-class F1 at the selected threshold, with PR-AUC as a threshold-independent complement. Macro-F1 appears for comparability with prior work but never drives threshold selection, since negative-class F1 is high by default at this base rate and inflates the average even when enthusiastic praise is missed. Metrics are computed per held-out teacher and averaged across folds rather than pooled, so each teacher counts equally. All systems see identical folds, which permits paired per-fold differences with cluster-bootstrap intervals over teachers. We report point estimates here and leave the paired intervals to the next revision, since marginal intervals on five folds are wide enough to be uninformative about contrasts.

5 Results

5.1 Linguistic pipeline result

All three models detect praise at comparable F1 (0.878 to 0.887), so detection is not where they separate. They differ on specificity, where GPT-5.6-luna reaches 0.825 F1 and is the most even across the three criteria. Qwen-3.7-28B is close overall (0.816) but weaker on naming an action (C1, 0.707 against 0.754), while GPT-5.4 loses most of its accuracy on both C1 (0.622) and personalized recognition (C3, 0.628). We therefore use GPT-5.6-luna’s praise decisions throughout, and note that it also has the highest detection precision (0.938 against 0.918 and 0.888), which is the quantity that governs how much non-praise speech contaminates the Global Reference set. Full results are in Table 2.

5.2 Acoustic pipeline result

SERMA is the strongest system at 0.542 F1 and 0.729 macro-F1 (Table 3). The ordering across the supervised family is stable and interpretable. Hand-crafted descriptors and mean pooling fail (0.303, 0.353), representations that encode the two reference relations reach 0.475 and above, and encoding those relations in a learned, gated space rather than a fixed PCA basis adds a further 0.067. Because FixedRef computes the same global and local deviations, this margin is attributable to how the relations are encoded rather than to their presence, and the complementary question of what the relations contribute at all is left to two ablations that are not yet run, one zeroing both deviation channels and one replacing the availability gates

with a constant. One caveat governs how all of these margins should be read. The fold-to-fold standard deviations, 0.144 for SERMA and 0.180 for FixedRef, are of the same order as the differences between systems, so the three-tier ordering is consistent across folds while individual pairwise margins remain provisional.

Two further observations. Text alone is weak everywhere (0.272 to 0.329), which is the premise of the paper, since what was said is nearly uninformative about how enthusiastically it was said, though SERMA’s text-only PR-AUC of 0.402 sits close to its audio-only 0.446, which suggests that much of the text-only collapse is thresholding fragility on roughly 18 calibration positives rather than an absence of signal. Audio alone recovers most of the combined performance, trailing the audio+text rows by 0.06 to 0.11 among the reference-relative systems and by at most 0.03 for the two absolute-feature systems. The zero-shot comparison is also instructive about where the difficulty lies. ZS-Pair-Audio at 0.471 matches the strongest supervised baseline without task-specific training, while ZS-Pair-Praat, which gives the same model the same protocol over textual z-score descriptions, collapses to 0.281. Listening, not reasoning, is what carries the judgment, and the pairwise framing is itself doing work, since it forces the model into the same within-teacher comparison that our reference relations encode.

6 Discussion and Future Directions

This study builds and evaluates an instrument for measuring praise delivery. We ask whether a theory-specified representation aligns with human judgments, not whether it maximizes prediction accuracy, and with five teachers the results show feasibility rather than generalizability.

Three results stand out. Representations that encode self-deviation reach 0.475 positive-class F1 and above, against 0.303 for hand-crafted prosodic descriptors and 0.353 for mean-pooled WavLM, and the ordering holds on all five held-out teachers. SERMA is the strongest at 0.542, but its 0.067 margin over the same relations computed in a fixed unsupervised basis is smaller than the fold-to-fold standard deviations of 0.144 and 0.180, and it differs from FixedRef in several ways at once, so the result establishes an ordering while the size of the gain remains open. The signal is also acoustic: audio alone recovers most of the audio+text result,

Table 2: Performance of praise detection and specificity models.

Model	Task	F1	Precision	Recall
GPT-5.4	Detection	0.878	0.888	0.868
	Specific	0.779	0.723	0.844
	Verb (C1)	0.622	0.491	0.849
	Echo (C2)	0.740	0.675	0.818
	Personal (C3)	0.628	0.696	0.571
Qwen-3.7-28B	Detection	0.887	0.918	0.859
	Specific	0.816	0.710	0.959
	Verb (C1)	0.707	0.592	0.879
	Echo (C2)	0.755	0.638	0.923
	Personal (C3)	0.746	0.759	0.733
GPT-5.6-luna	Detection	0.880	0.938	0.829
	Specific	0.825	0.773	0.885
	Verb (C1)	0.754	0.722	0.788
	Echo (C2)	0.763	0.726	0.803
	Personal (C3)	0.764	0.778	0.750

Table 3: Performance of enthusiastic praise detection models.

System	Modality	F1	$\pm$	Macro-F1	PR-AUC
Praat	audio+text	0.303	0.030	0.435	0.257
	audio	0.273	0.081	0.486	0.250
	text	0.285	0.058	0.468	0.308
WavLM-mean	audio+text	0.353	0.057	0.561	0.267
	audio	0.340	0.084	0.560	0.329
	text	0.272	0.051	0.410	0.217
FixedRef	audio+text	0.475	0.180	0.640	0.430
	audio	0.367	0.099	0.590	0.381
	text	0.306	0.128	0.427	0.218
SERMA	audio+text	0.542	0.144	0.729	0.460
	audio	0.481	0.157	0.686	0.446
	text	0.324	0.119	0.469	0.402
SERMA-Transformer	audio+text	0.519	0.152	0.689	0.433
	audio	0.453	0.197	0.630	0.402
	text	0.329	0.098	0.527	0.328
ZS-Point	Praat-as-text	0.352	n/a	0.541	0.244
ZS-Pair-Audio	audio+text	0.471	n/a	0.661	0.368
ZS-Pair-Praat	Praat-as-text	0.281	n/a	0.327	0.201

text alone is weak in every system, and a zero-shot model given audio reaches 0.471 against 0.281 for the same model given the same pairwise protocol over Praat z-scores as text. None of these comparisons isolates the reference relations, since every system that encodes them also differs in how it encodes them.

We plan four next steps. An ablation that zeroes both deviation channels, with the frame encoder, the turn-level encoder, and the prediction heads held fixed, would separate what the reference relations contribute from what the encoder carrying them contributes; a companion ablation replaces the availability gates with a constant to test whether attenuating deviations from sparse

reference sets matters. A sensitivity check varies the global and local set sizes and substitutes another teacher’s praise for the speaker’s own, testing whether the measure tracks the baseline it is given. Paired per-fold differences with cluster-bootstrap intervals over teachers would remove the shared fold difficulty behind most of the spread in Table 3 and support comparisons between systems. Re-rating a subset of turns under a teacher-shuffled protocol (Section 2.3) would estimate how far judgments move once the speaker-relative standard is removed. More teachers are needed before any gain can be sized, joint modeling of specificity and delivery would test an interaction the construct predicts, and single-pass inference would make the

method usable for feedback during instruction.

References

- Jamie Amemiya and Ming-Te Wang. 2018. Why effort praise can backfire in adolescence. *Child Development Perspectives*, 12(3):199–203.
- Wesley C. Becker, Madsen, Charles H., Jr., Carole R. Arnold, and Don R. Thomas. 1967. The contingent use of teacher attention and praise in reducing classroom behavior problems. *The Journal of Special Education*, 1(3):287–307.
- Jere Brophy. 1981. Teacher praise: A functional analysis. *Review of Educational Research*, 51(1):5–32.
- Eddie Brummelman, Jennifer Crocker, and Brad J. Bushman. 2016. The praise paradox: When and why praise backfires in children with low self-esteem. *Child Development Perspectives*, 10(2):111–115.
- Paul C. Burnett. 2002. Teacher praise and feedback and students’ perceptions of the classroom environment. *Educational Psychology*, 22(1):5–16.
- Zhongpu Chen, Decheng Ren, Zhongying Wang, Liyue Gao, Ying Shi, Haiying Liu, and Tao Yang. 2025. Enhancing student academic performance in middle school classrooms by fostering engagement motivation through intelligent assessment of teacher praise emotional intensity. *Learning and Instruction*, 99:102185.
- Hao Du, Shiyin Lu, Guangyu Zhou, and Ye Gao. 2025. EAA: Emotion-aware audio large language models with dual cross-attention and context-aware instruction tuning. In *Proceedings of Interspeech 2025*, pages 5433–5437. ISCA.
- Margaret T. Floress, Lyndsay N. Jenkins, Wendy M. Reinke, and Laura McKown. 2018. General education teachers’ natural rates of praise: A preliminary investigation. *Behavioral Disorders*, 43(4):411–422.
- Itai Gat, Hagai Aronowitz, Weizhong Zhu, Edmilson Morais, and Ron Hoory. 2022. [Speaker normalization for self-supervised speech emotion recognition](#). In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7342–7346. IEEE.
- Richard Göllner, Rebecca Lazarides, and Patrick Stark. 2025. Revealing teaching quality through lesson semantics: A GPT-assisted analysis of transcripts. *British Journal of Educational Psychology*, 95:S300–S315.
- Z. Gong, P. Shi, K. Donbekci, L. Ai, R. Chen, D. Sasu, Z. Wu, and J. Hirschberg. 2025. [Learning more with less: Self-supervised approaches for low-resource speech emotion recognition](#). *arXiv preprint arXiv:2506.02059*. Preprint.
- Kristen L. Granger, Melissa Washington-Nortey, Jason C. Chow, Michael D. Broda, Alexandra Montesion, Kevin S. Sutherland, and Maureen A. Conroy. 2024. [Child gender and challenging behavior influences early childhood teachers’ use of behavior specific praise](#). *Psychology in the Schools*, 61:4688–4704.
- Chanhee Jeong and Aditya Nyamathi. 2026. [Causal emotion recognition in conversation: Context saturation and discourse-marker evidence](#). *Transactions on Machine Learning Research*.
- Daniel Kahneman and Dale T. Miller. 1986. [Norm theory: Comparing reality to its alternatives](#). *Psychological Review*, 93(2):136–153.
- Bethany U. Kamana, Claudia L. Dozier, Elizabeth A. Foley, and April N. Hubbs. 2022. [Descriptive versus general praise with and without edibles for acquisition in young children](#). *Behavioral Interventions*, 37(2):222–238.
- Bidisha Maji, Rishabh Guha, Aurobinda Routray, Shabnam Nasreen, and Debasish Majumdar. 2024. Investigation of layer-wise speech representations in self-supervised learning models: A cross-lingual study in detecting depression. In *Proceedings of Interspeech 2024*. ISCA.
- Lucas Martinez-Lucas, Prashanth Mote, Amir R. Naini, Mohammed Abdelwahab, and Carlos Busso. 2026. MSP-Conversation: A corpus for naturalistic, time-continuous emotion recognition. *arXiv preprint arXiv:2603.22536*. Preprint.
- Jenna S. Morton, Moïra Mikolajczak, and Olivier Luminet. 2022. New perspectives on the praise literature: Towards a conceptual model of compliment. *Current Psychology*, 41(9):6038–6050.
- Claudia M. Mueller and Carol S. Dweck. 1998. Praise for intelligence can undermine children’s motivation and performance. *Journal of Personality and Social Psychology*, 75(1):33–52.
- Amir R. Naini, Luis Goncalves, A. N. Salman, Prashanth Mote, Ilker R. Ülgen, Thomas Thebaud, Laureano Moro-Velazquez, L. Paola Garcia, Najim Dehak, Berrak Sisman, and Carlos Busso. 2025. The interspeech 2025 challenge on speech emotion recognition in naturalistic conditions. In *Proceedings of Interspeech 2025*. ISCA.
- Zahra Omid and John H. L. Hansen. 2026. Learning from annotation uncertainty: Entropy-aware curriculum for speech emotion recognition. *arXiv preprint arXiv:2606.27536*. Preprint.
- Takamura Onishi, Akihiro Yamauchi, Ryo Ishii, Yuki Aono, and Atsushi Miyata. 2020. Analyzing nonverbal behaviors along with praising. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, pages 609–613. ACM.

- Rachel Ostrand and Eleanor Chodroff. 2021. [It's alignment all the way down, but not all the way up: Speakers align on some features but not others within a dialogue](#). *Journal of Phonetics*, 88:101074.
- Silke Paulmann and Netta Weinstein. 2025. Motivating tones to enhance education: The effects of vocal awareness on teachers' voices. *British Journal of Educational Psychology*, 95(2):551–564.
- Amy S. Polick, James E. Carr, and Nicole M. Hanney. 2012. A comparison of general and descriptive praise in teaching intraverbal behavior to children with autism. *Journal of Applied Behavior Analysis*, 45(3):593–599.
- Ashley Rila, Allison L. Bruhn, and Lu Wang. 2024. [The disaggregated effects of visual performance feedback on teachers' use of behavior specific-praise and reprimands](#). *Journal of Behavioral Education*, 33:53–80.
- Richard M. Ryan and Edward L. Deci. 2000. Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary Educational Psychology*, 25:54–67.
- Xiaoran Shi, Ximin Li, and Tomoki Toda. 2025. Speaker-aware multi-task learning for speech emotion recognition. In *Proceedings of Interspeech 2025*. ISCA.
- Bart Soenens and Maarten Vansteenkiste. 2020. Understanding the complexity of praise through the lens of self-determination. In Eddie Brummelman, editor, *Psychological Perspectives on Praise*. Routledge.
- Colleen Stevens, Tina M. Sidener, Sharon A. Reeve, and David W. Sidener. 2011. [Effects of behavior-specific and general praise on acquisition of tacts in children with pervasive developmental disorders](#). *Research in Autism Spectrum Disorders*, 5(1):666–669.
- Nathan Stout. 2020. On the significance of praise. *American Philosophical Quarterly*, 57(3):215–226.
- Zachary Tamir, Thomas Thebaud, Jesus Villalba, Najim Dehak, and Oren Kurland. 2025. Multimodal emotion diarization: Frame-wise integration of text and audio representations. In *Proceedings of Interspeech 2025*. ISCA.
- Lucas H. Ueda, João G. T. Lima, and Paulo D. P. Costa. 2026. Crab: Multi layer contrastive supervision to improve speech emotion recognition under both acted and natural speech condition. *arXiv preprint arXiv:2603.23673*. Preprint.
- Deepak Uniyal and Vinay Abrol. 2025. From pretraining to performance: Benchmarking self-supervised speech models for interspeech-25 SER challenge. In *Proceedings of Interspeech 2025*. ISCA.
- Netta Weinstein, Maarten Vansteenkiste, and Silke Paulmann. 2020. Don't you say it that way! experimental evidence that controlling voices elicit defiance. *Journal of Experimental Social Psychology*, 88:103949.
- Jessica R. Weyman and Josephine R. Sy. 2018. [Effects of neutral and enthusiastic praise on the rate of discrimination acquisition](#). *Journal of Applied Behavior Analysis*, 51(2):335–344.
- Susanne Wolff and Andre Brechmann. 2015. Carrot and stick 2.0: The benefits of natural and motivational prosody in computer-assisted learning. *Computers in Human Behavior*, 43:76–84.
- Tianhao Zhang and Zhixing Tan. 2025. [ECERC: Evidence-cause attention network for multi-modal emotion recognition in conversation](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. ACL.

Appendix

Table A1: Inter-annotator agreement on the linguistic rubric, before and after the rubric revision round.

Construct	Round 1	Round 2
Praise detection	0.803	0.854
Specificity (any criterion)	0.743	0.797
Verb (C1)	0.652	0.758
Echo (C2)	0.749	0.914
Personal (C3)	0.711	0.841

Table B1: Optimization and evaluation.

Setting	Value
WavLM checkpoint	WavLM-Large, layers 0 to 8 of 24, frozen
Praat	13 hand-crafted prosodic features per praise segment (librosa and parselmouth)
RoBERTa	774-d per turn, 768 from roberta-base plus 6 structural features
Frame encoder	1-layer unidirectional GRU
Turn-level encoder	1-layer GRU, or 1-layer Transformer (4 heads, FF 128) in SERMA-Transformer
Dropout	0.1
Optimizer	AdamW, weight decay 1e-4
Learning rate	1e-3
Batch size	16
Epochs	40, early stopping patience 5
Loss	mean of the two head losses, class-weighted CE with inverse-frequency weights fit on the training fold (BCE for binary tasks)
Seeds	base 42, offset by fold index, one run per fold

Prompt

Praise detection prompt

What is praise?

Praise is a positive evaluative statement — the teacher signals "I approve of what you did", not merely "I heard you" or "that is correct."

Decision — apply in order, stop at the first match

1. Contains a recognized praise expression → PRAISE

Single words: awesome, excellent, fantastic, wonderful, beautiful, brilliant, amazing, outstanding, terrific, bravo, phenomenal, impressive, incredible, marvelous, fabulous, perfect (evaluating work).

Phrases: good/great/nice + job/work/answer/thinking/try/reading/effort, well done, way to go, keep it up, I'm proud of you, I'm impressed, love it, that's + [any single-word above].

2. Explicitly validates what the student did or how they did it → PRAISE

e.g. "You read that correctly", "I can tell you practiced", "That's a clever way to think about it".

3. A short acknowledgment/confirmation/politeness word combined with any positive evaluative word → PRAISE

e.g. "OK, perfect", "That's right, good", "Thank you, nice job".

4. Otherwise → NO PRAISE. In particular, these alone are NOT praise:

- Bare acknowledgments: good, great, nice, right, yes, yeah, ok, all right, sure, mm-hmm, come on.
- Bare confirmations: that's right, correct, exactly, you got it right.
- Politeness / procedural: thank you, please, your turn, turn to page X, let's move on.
- Clearly sarcastic uses of positive words.

Boundary cases

- "Good." alone → NO PRAISE. "Good job." → PRAISE.

- "That's right." alone → NO PRAISE. "That's right, good." → PRAISE (Rule 3).

- "Perfect." after a student's answer → PRAISE. "Perfect, now sit down." → still PRAISE (the evaluative use came first).

- Praise + a following instruction still counts as praise ("Nice job. Turn to page 5." → PRAISE).

- "You read the sentence." → NO PRAISE (neutral). "You read the sentence correctly." → PRAISE (evaluative).

Praise classification prompt

Task

The teacher utterance below already contains praise. Decide which of the three specificity criteria are satisfied. Any one criterion is enough to make the praise SPECIFIC.

Return a list of the satisfied criterion numbers (e.g. [1, 3], or [] if none).

The three criteria

Criterion 1 — Descriptive expression.

The praise clause contains a word or phrase (verb, gerund, noun, or noun phrase) that names what the student did or produced.

✓ "Nice reading." · "Good job listening." · "Great answer." · "Beautiful handwriting." · "That's a creative solution."

× "Good job." · "Awesome." · "Perfect." (generic praise words alone)

Criterion 2 — Echo or paraphrase.

The teacher repeats, restates, or paraphrases a content word/phrase from the student's prior turn. Requires a prior student turn to check.

✓ Student: "gravity" → Teacher: "Gravity — exactly right!"

✓ Student: "I added 5 and 3" → Teacher: "You combined them — perfect!"

× Teacher: "That's great." with no matching content word from the student.

Criterion 3 — Student name attached to the praise.

The student's name is directly connected to the praise clause by comma (vocative). Pronouns like you/your do NOT count.

✓ "Good job, Sarah." · "Sarah, that was excellent!"

× "Good job. Sarah, read this." (name belongs to the following instruction, not the praise)

Key boundary rule

A descriptive word or name only counts if it is inside the praise clause, not inside a following instruction or question.

- "Nice job explaining that." → C1 ✓ (attached)

- "Nice job. Could you explain this?" → C1 × ("explain" is in a new request)

- "Good job, Sarah." → C3 ✓

- "Good job. Sarah, read this." → C3 ×