Using Whisper Embeddings for Audio-Only Latent Token Classification of Classroom Management Practices

Wesley Morris, Jessica Vitale, and Isabel Arvelo
Vanderbilt University

Abstract

In this study, we developed a textless NLP system using a fine-tuned Whisper encoder to identify classroom management practices from noisy classroom recordings. The model segments teacher speech from non-teacher speech and performs multi-label classification of classroom practices, achieving acceptable accuracy without requiring transcript generation.

1 Introduction

Positive and proactive classroom management establishes a foundation for equitable and inclusive environments where all students can learn. Research demonstrates that effective classroom management increases student engagement and academic achievement, particularly for students with learning and behavioral differences [1]. Despite identifying evidence-based classroom management practices, a significant implementation gap exists in their consistent classroom application [25]. Teachers often report feeling underprepared to support student behavior and express a need for ongoing, job-embedded professional development to implement practices effectively. Coaching and observational feedback improve teachers' classroom management practices and enhance their self-efficacy, reducing stress and mitigating burnout [19, 30]. However, these traditional approaches are resource-intensive and difficult to scale, particularly in historically marginalized communities. Advances in natural language processing (NLP) and machine learning present an innovative opportunity to address these challenges. Automated feedback tools can deliver frequent, timely, and actionable insights to teacher practice, bridging the gap between evidence-based practices and their real-world implementation, providing accessible professional development at scale.

Current automated feedback tools for teacher classroom practices rely solely on transcripts generated by Automatic Speech Recognition (ASR) tools. However, teacher affect, including tone and delivery, is critical in shaping positive student-teacher interactions, fostering social-emotional learning, and reinforcing classroom expectations [15]. Research indicates that transcription alone often fails to capture these suprasegmental speech features, resulting in losing vital information about prosody and intonation [26]. To address this limitation, we are developing a Multimodal Automatic System for the Classification of Teacher Classroom Practices (MASCoT-CP) to automatically detect classroom management practices using both audio and text-based data. This system aims to provide teachers with actionable insights into their practices, leveraging multi-modal inputs to enhance the feedback they receive. Unlike current automated feedback tools that rely exclusively on text-based transcript analysis, MASCoT-CP incorporates prosody, intonation, and affect, key elements of spoken language essential for understanding the nuances of classroom culture and teacher-student interactions.

This study presents findings from the audio-only component of the MASCoT-CP system. This component, designed as part of a larger, multi-modal system that will integrate audio and text transcripts, serves two purposes: diarizing classroom audio into teacher speech and non-teacher speech segments, and generating predictions about classroom management practices present within those segments. Future research will

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 1: Full Papers, pages 154–162

integrate the output of the audio-only model with a text classification model to create an ensemble system that enhances classification accuracy. This comprehensive approach will provide teachers with fine-grained feedback on their classroom practices, allowing them to focus on refining specific elements of their practices, thereby enhancing their students' learning experiences.

2 Background

2.1 Classroom Management Practices

Classroom management includes the strategies and practices teachers implement to establish and maintain structured, supportive learning environments. Research consistently demonstrates that effective classroom management is fundamental to maximizing instructional time, sustaining student engagement, and building positive student-teacher relationships. Systematic reviews identify several evidence-based practices that contribute to successful classroom management, particularly frequent opportunities for active student engagement and feedback for student behaviors [5, 8].

Central to effective classroom management are opportunities to respond (OTRs), questions or prompts that elicit student participation. Research shows that high rates of OTRs help sustain student engagement, increase on-task behaviors, and improve accuracy in student responses [7]. Complementing these engagement strategies, teacher feedback further shapes student behavior. Feedback typically falls into two categories within classroom management: reinforcing appropriate behavior through positive feedback (such as specific praise) and addressing inappropriate behavior through redirections or corrective responses. Evidence indicates that delivering specific praise and maintaining a positive ratio of positive to corrective interactions strengthens student-teacher relationships and increases students' on-task behaviors [2, 9]. Together, these practices create positive classroom environments that establish a foundation necessary for effective academic instruction.

Despite strong evidence supporting classroom management's impact on student outcomes, many teachers face challenges in consistently implementing these practices. Pre-service teacher preparation programs often provide limited training in classroom management [12], leading teachers to identify it as one of the most challenging aspects of their job and a primary factor contributing to teacher attrition [13, 28]. These implementation challenges underscore the need for effective professional development. Traditional approaches to supporting teacher development, such as coaching and observational feedback, have effectively improved practice implementation. However, scaling these support presents logistical and financial barriers due to time and resource constraints. Recent advances in NLP technologies offer promising solutions for addressing these scalability challenges. NLP tools capable of analyzing classroom discourse and generating automated feedback represent an emerging approach to supporting teaching practices at scale [10, 15].

Multiple research teams have developed text-based classification models using transformer architectures to analyze classroom transcripts. These studies demonstrate the feasibility of automated classroom discourse analysis across different instructional contexts and pedagogical practices. Alic et al. [1] fine-tuned a RoBERTa-based model with paired teacher-student utterances for binary classification of focusing questions, achieving an F1 score of 0.501. Suresh et al. [24, 25] trained a RoBERTa-base model to classify teacher utterances into one of ten math talk moves, incorporating surrounding transcript lines as context, and achieved an average F1 score of 0.79. Similarly, Jensen et al. [17] fine-tuned BERT to classify seven discourse-related teaching practices, obtaining an average area under the curve (AUC) of 0.84 across classifications.

2.2 Audio Classification

The studies mentioned above analyzed transcripts of teacher speech, rather than classifying directly from audio. Unlike text data, which consists of discrete words and subwords easily tokenized through dictionary lookup, audio data presents as a continuous information stream. While previous research has used feature engineering approaches to extract information from classroom audio [11, 16, 23], the current study uses a modified form of token classification approach that converts raw audio into latent token embeddings. Whisper [20], developed by OpenAI, is a sequence-to-sequence transformer model for automatic speech recognition (ASR). In the original architecture, the encoder's final hidden state feeds into a decoder

block that recursively generates text conditioned on both the encoder's final hidden state and previously generated tokens. The model was trained on 680,000 hours of speech with transcripts, including 117,000 hours in 96 non-English languages. As a result, Whisper achieves strong results in ASR and translation tasks [20].

Recent interest in textless NLP has focused on directly extracting semantic information from the audio without intermediate transcription [14]. Although designed for ASR and translation, multimodal sequence-to-sequence models show promise for audio classification tasks. Ma et al. [19] fine-tuned Whisper to generate label tokens, effectively performing zero-shot audio sound event classification. Classification can also be performed by separating the Whisper encoder block and using the final hidden state embeddings directly, as demonstrated in predicting speech disorders such as dysarthria [21] and stuttering [3]. In this audio classification approach, the encoder's final hidden states pass through a projection layer into a classification head that generates sequence predictions.

2.3 Current Study

In this study, we develop an audio-only tool that identifies classroom management practices in teacher speech segments. Our approach uses a three-state process using a modified Whisper architecture. First, we detach the Whisper encoder from the decoder and fine-tune it for latent token classification, similar to text-based NLP token classification, to predict the most probable teaching practice in each 0.02-second audio window. Second, we use these predictions to differentiate segments containing teacher speech from non-teacher speech segments. Finally, we use the predictions from the Whisper encoder to perform multi-label classification on teacher speech segments to identify which specific classroom management practices are present. The study addresses two primary research questions:

RQ1: Can an audio-only model accurately distinguish between teacher and non-teacher speech in elementary classroom recordings?

RQ2: Can an audio-only model accurately identify classroom management practices present within teacher speech segments from elementary classroom recordings?

3 Methods

3.1 Dataset

The dataset used to train the classification model included 29.91 hours of audio recordings from 131 classroom sessions. The recordings were collected from 28 teachers (15 general education, 13 special education) across kindergarten through 4th-grade classrooms. The sample included 6 male and 22 were female teachers. Teachers self-identified as White (n=16), Black (n=7), Latinx (n=4), and Biracial (n=1). Their average teaching experience was 11 years (range = 1-30). Each teacher contributed 4 to 5 recordings to the dataset. The recordings from special education teachers primarily consisted of small-group interventions, while general education teachers recorded themselves conducting whole-group instruction with an average of 21 students per class.

The audio recordings were annotated for 10 specific teaching practices and two non-teacher talk labels, organized into 6 broader categories related to classroom management. The six categories include instructional talk, social talk, positive teacher-student interactions (i.e., specific praise, general praise, and affirming correct student responses), negative teacher-student interactions (i.e., reprimands, redirections, and correcting incorrect student responses), opportunities to respond (OTRs) (i.e., academic and social demands and questions) and non-teacher speech (e.g., student talk and prolonged instances of silence).

Each audio file was annotated by trained labelers using Audacity [4], where labelers listened to the complete recording and noted each segment's start and end times. This approach allowed us to establish ground-truth boundaries for each segment, enabling us to compare multiple diarization tools and align with methods used in systematic directional observation of classrooms [18, 29]. Since spoken language in classrooms does not follow traditional written sentence structures, annotators applied two stop rules to determine segment boundaries: a shift to a new practice category (e.g., a teacher transitioning from providing instructional talk to asking a question, signaling an opportunity to respond) or silence lasting at least two seconds (e.g., a teacher pausing mid-instructional talk to think). Table 1 displays the count of each classroom practice in the full dataset as well as aggregate statistics about their durations. To ensure reliability, each recording was

annotated by two independent labelers, followed by consensus coding meetings to resolve discrepancies. Inter-rater agreement (IRA) was calculated using the Multi-Option Observation System for Experimental Studies (MOOSES) [26], with agreement defined as both labelers identifying the same practice within a two-second window. The average IRA across the 131 recordings was 74%, with most disagreements occurring around segment start and end times rather than label assignment.

		duration (s.)	
	count	mean	std
Instructional Talk	5318	6.65	9.2
Social Talk	2477	3.52	3.8
Positive Interactions	2270	2.76	2.2
Corrective Interactions	981	3.25	2.7
Opportunity to Respond	5749	2.99	2.1
Non-teacher	9102	2.78	3.4

Table 1: Counts and durations of each classroom management practice category

3.2 Training

We used the encoder stack of Whisper base [20], as the foundation for a custom audio latent token classification model. Figure 1 illustrates the architecture of our modified version of the Whisper encoder stack with the shapes of embedding matrices listed on the bottom. The Whisper preprocessor first uses fast Fourier transforms that generate 80-channel log-mel spectrograms from 30 second segments of raw audio using 16 kHz sampling, 25ms window length, and 10ms stride. These spectrograms serve as input to two convolutional layers with a filter width of 3 and GELU activation function. The first layer maps the 80 spectrogram channels to embedding dimension d = 768. The second layer uses a stride of 2 to reduce the 3,000 windows to T = 1,500 latent token embeddings, each spanning 0.02 seconds. Sinusoidal position embeddings are then added to produce the final T×d dimensional hidden states $\mathbf{h}_{1,2...L} \in \mathbb{R}^{T\times d}$ that define each layer's embedding dimensionality in the encoder.

We first removed all audio files from nine (31%) of the teachers as a hold-out test set to ensure that the model generalizes to speakers outside of its training set. We then split each audio file into thirty-second clips with a fifteen-second overlap so that the model would be exposed to all audio twice

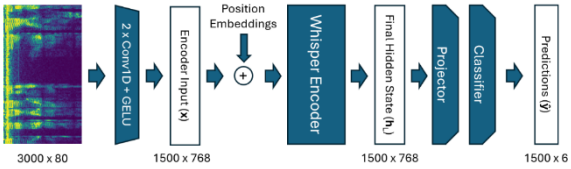

Figure 1: Architecture of the latent token classification model with dimensionality of matrices

per epoch except for the first and last fifteen seconds of the audio file. We included this overlap to ensure that each timestep had at least 15 seconds of previous context to inform the classification. Thus, the final hidden state of the Whisper encoder had a dimensionality of $\mathbf{h}_L \in \mathbb{R}^{T\times d}$ where T is the total number of time steps (i.e. 1,500) and d is the embedding dimensionality (i.e. 768). On top of the Whisper encoder block, we applied linear layers for token classification. The first, a projector, reduced the dimensionality from 768 to 256 and applied a ReLU activation function. Finally, our classification head further reduced the dimensionality to six, our number of labels k, with a sigmoid activation function. Therefore, the output of the model had a dimensionality of $\hat{\boldsymbol{y}} \in \mathbb{R}^{T\times k}$.

We used the Whisper encoder's output to create target labels for training. For each 30-second audio clip, we generated 1,500 target labels by mapping the original hand-annotated labels to each of the 1,500 timesteps t. At each timestep, we identified the predominant label from the annotations. The model's predictions were then compared against these labels using cross-entropy loss. We fine-tuned the model for six epochs using the AdamW optimizer. Following the specifications from the original Whisper training [20], we used a learning rate of 3.75e-05 and a weight decay of 0.1.

3.3 Diarization

Our first goal was to correctly distinguish segments of audio where the teacher was speaking from segments of audio where the teacher was not speaking (e.g. student speech, silence). For inference, we first split the audio in the test set into thirty-second clips, overlapping with a step of fifteen seconds, as during training. We then used our model to generate logits for each 0.02-second window. Because of our method of splitting the audio files into overlapping clips, all audio in a file aside from the first and last 15 seconds is analyzed twice. We therefore calculate final logits for each

0.02-second window as the mean of the two predictions. Finally, we take the maximum logit for each window to determine the predicted class. If the class is predicted to be anything other than one of the classroom management practices, then we classify it as non-teacher speech. Any classroom management practice was classified as teacher speech. We evaluated our success using a modified diarization error rate (mDER), defined as:

$$mDER = \frac{Seconds\ of\ misclassified\ audio}{Total\ seconds}$$

Due to the "noisy" nature of elementary classroom environments, we noticed occasional very short segments. To address this, we implemented a minimum speaker turn length, merging segments shorter than a certain threshold with adjacent speech. We empirically determined the optimal threshold length, by assessing mDER at minimum length thresholds between 0.1 and 0.8. This threshold optimization was conducted exclusively on the training set to prevent information leak into the test set.

Finally, we tested our diarization method on the withheld test set, comparing our results against other open-source diarization tools including Pyannote [6] and SpeechBrain [22]. Unlike traditional diarization models that precisely mark the start and stop boundaries of speech and silence, our labeling scheme captures higher level speaker turns. For example, if a teacher pauses briefly during a classroom practice and then continues, our label extends across the entire segment rather than breaking it at the silence. This distinction is particularly important when comparing our approach to diarization tools designed to detect precise speech boundaries. These models segment speech with frequent breaks and allow for speaker overlap, which is not possible in our framework. Because our evaluation metric is based on non-overlapping, high-level speaker turns, other diarization models may be penalized under our modified DER, even when they have correctly identified what occurred in the audio. For our use case, where we aim to broadly classify whether a given segment of audio represents teacher speech or non-teacher speech, diarization serves primarily as a necessary preprocessing step rather than an end goal. Our segmentation approach is well-suited for our application because it reduces noise from minor pauses, interruptions, or overlapping speech that are not critical to our analysis.

3.4 Classification

After diarizing the audio into teacher speech and non-teacher speech, we used the logits computed by the classification tool to identify all teacher classroom practices present in each segment of audio. Each segment was assigned a vector $\hat{\mathbf{y}} \in \mathbb{R}^{K\times 1}$ where K is equal to the number of classes. If the model predicted the label for any of the 0.02-second windows within that segment, its value was predicted as 1, otherwise it was predicted as 0. Similarly, if a label k was present in a segment of the target dataset, $\mathbf{y}_k = 1$ otherwise 0. We evaluated success by calculating precision, recall, and f1 scores for each of the classes across all the segments.

4 Results

4.1 Diarization Results

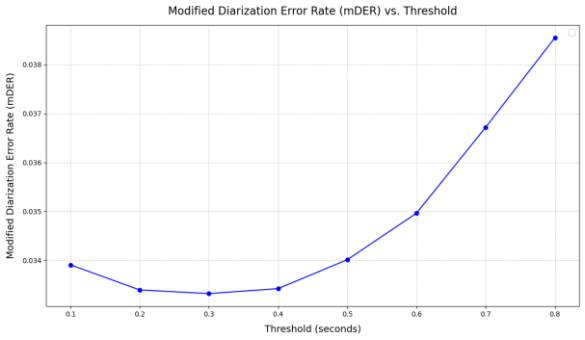

Figure 2: Identifying optimal maximum segment length for audio segmentation

Tool	Total mDER	Teacher mDER	Non-Teacher mDER
MASCoT-CP	0.086	0.06	0.149
Pyannote	0.264	0.29	0.196
SpeechBrain	0.324	0.28	0.438

Table 2: Modified Diarization Error Rate for MASCoT-CP vs. other diarization systems

We first attempted to determine the optimal minimum segment size. As Figure 2 shows, we found 0.3 seconds to be the optimal minimum segment length, and used this parameter for all further experiments. Using the minimum segment

length on our test set, we found an mDER of 0.086, indicating that 8.6% of all audio segments were misclassified. This outperformed other open source diarization tools such as Pyannote (mDER = 0.264) and SpeechBrain (mDER = 0.324). However, it should be noted that our study does not take other elements of diarization into account, such as voice overlap and speaker identification. Table 2 shows results from the three diarization tools.

4.2 Classification Results

Practice Category	n	prec.	recall	F1
Instructional Talk	2,011	0.627	0.509	0.562
Social Talk	1,222	0.334	0.548	0.415
Positive Interactions	757	0.556	0.528	0.542
Corrective Interactions	512	0.179	0.221	0.198
Opportunity to Respond	2,194	0.637	0.528	0.577
Non-teacher	4,360	0.875	0.55	0.675
mean	**1,843.7**	**0.535**	**0.481**	**0.495**

Table 3: Counts in test set and metrics for each classroom practice

Once we had separated each audio file into teacher-speech and non-teacher-speech segments, we generated labels for each segment according to whether a classroom practice was predicted in each segment. Table 3 shows precision, recall, and f1 scores for each classroom practice, as well as the number of occurrences of each classroom practice in the test set. The model's classification F1 scores were above 0.4 for all classroom practices aside from corrective, which may be a result of the low prevalence of this practice. However, while praise had a similarly low prevalence, the model was much more likely to identify this classroom practice correctly (F1 = 0.542).

5 Discussion

In this study, we developed an audio-only tool which uses a fine-tuned version of the Whisper base model's encoder stack to segment and classify teacher speech for the classroom management practices. We fine-tuned the model on a dataset of almost 30 hours of classroom audio annotated by expert raters for the start and end times of classroom management practices. Finally, we process the output of the model to identify segments of teacher speech and classify the classroom management practices in those segments. This study demonstrates that models can be trained to identify classroom practices with reasonably performance levels without access to text transcripts.

Our model effectively distinguishes between teacher speech and non-teacher speech, achieving a low misclassification rate of 8.6% - a significant improvement over other open-source diarization models. However, it is important to note that other diarization models are not specifically tuned for this task or classroom contexts. Our approach differs from traditional diarization methods, which precisely segment speech boundaries and capture overlapping speakers. Regardless, our results suggest that our model is well-suited for automatic identification of teacher speech in classroom recordings without requiring prior training on individual teacher voices, making it a practical alternative to traditional diarization tools when the goal is classification of classroom discourse rather than precise speaker diarization.

For classification performance, our tool attained F1 values between 0.4 and 0.7 for all but one teaching practice. The lowest F1 score of 0.2 occurred for corrective interactions, likely due to the limited representation of this class in the training dataset. With only 981 instances (3.8% of the training dataset), correctives were the least frequent classroom practice we labeled, potentially limiting the model's ability to learn robust patterns for this category. While our classification accuracy was lower than that of previous studies, reporting F1 scores between 0.79 and 0.84 for multi-class classification of teacher discourse moves [17, 25], it is important to note that prior work relied on hand-transcribed textual data. In contrast, our study uses raw, noisy, audio-only data.

Our study was principally limited by the relatively small sample size of only 30 hours from 28 teachers. We need to train our model on a larger and more diverse labeled dataset to develop a tool that generalizes effectively across diverse linguistic environments. Additionally, while our results are promising, given that they are derived directly from

audio in naturally noisy classroom recordings, they lag behind studies using clean text transcripts. One potential solution is to integrate this audio model into a larger multi-modal ensemble model that leverages audio and transcripts to achieve higher accuracy in identifying classroom practices.

6 Conclusion

In this study, we trained the encoder block of the Whisper to predict classroom management practices in small time windows of teacher speech. We then used these predictions for two purposes: segmenting audio into teacher speech and non-teacher speech segments with high accuracy and predicting which classroom management practices were present in the segments with reasonably high performance. These results demonstrate that it is possible to classify classroom management practices using textless NLP methods, even in noisy classroom recordings.

While observation and feedback are established methods for supporting teacher development, their implementation is resource-constrained, particularly in under-resourced educational settings. Automatically identifying teaching practices from authentically noisy audio recordings can allow teachers to reflect and improve their use of effective classroom management practices. This can have significant downstream effects on students' educational experiences, particularly those with learning differences and those in under-resourced settings.

This study contributes to advancements in textless NLP and automated measurement of classroom practices. Future research will build on the audio-only model by integrating it with a text-based classification approach using ASR-derived transcripts, forming a multi-modal automatic system for classifying classroom management practices (MASCoT-CP). By combining transcript analysis with prosodic and intonational features from the audio-only model, we anticipate improved accuracy in predicting teaching practices. This potentially enhanced measurement capability could be a foundation for developing automated feedback tools that provide teachers with data-driven insights into their classroom management strengths and areas for reflection and growth.

References

[1]Alic, S., Demszky, D., Mancenido, Z., Liu, J., Hill, H. and Jurafsky, D. 2022. Computationally Identifying Funneling and Focusing Questions in Classroom Discourse. arXiv.

[2]Allday, R.A., Hinkson-Lee, K., Hudson, T., Neilsen-Gatti, S., Kleinke, A. and Russel, C.S. 2012. Training General Educators to Increase Behavior-Specific Praise: Effects on Students with EBD. *Behavioral Disorders*. 37, 2 (Feb. 2012), 87–98. DOI:https://doi.org/10.1177/019874291203700203.

[3]Ameer, H., Latif, S., Latif, R. and Mukhtar, S. 2023. Whisper in Focus: Enhancing Stuttered Speech Classification with Encoder Layer Optimization. arXiv.

[4]Audacity Team 2014. Audacity(R): Free Audio Editor and Recorder.

[5]Brandi Simonsen, Sarah Fairbanks, Amy Briesch, Diane Myers, and George Sugai 2008. Evidence-based Practices in Classroom Management: Considerations for Research to Practice. *Education and Treatment of Children*. 31, 1 (2008), 351–380. DOI:https://doi.org/10.1353/etc.0.0007.

[6]Bredin, H., Yin, R., Coria, J.M., Gelly, G., Korshunov, P., Lavechin, M., Fustes, D., Titeux, H., Bouaziz, W. and Gill, M.-P. 2020. Pyannote.Audio: Neural Building Blocks for Speaker Diarization. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Barcelona, Spain, May 2020), 7124–7128.

[7]Cavanaugh, B. 2013. Performance Feedback and Teachers' Use of Praise and Opportunities to Respond: A Review of the Literature. *Education and Treatment of Children*. 36, 1 (2013), 111–136. DOI:https://doi.org/10.1353/etc.2013.0001.

[8]Conroy, J., Hulme, M. and Menter, I. 2013. Developing a 'clinical' model for teacher education. *Journal of Education for Teaching*. 39, 5 (Dec. 2013), 557–573. DOI:https://doi.org/10.1080/02607476.2013.836339.

[9]Conroy, M., Sutherland, K., Snyder, A., Al-Hendawi, M. and Vo, A. 2009. Creating a Positive Classroom Atmosphere: Teachers' Use of Effective Praise and Feedback. *Beyond Behavior*. 18, 2 (2009), 18–26.

[10]Demszky, D., Liu, J., Hill, H.C., Sanghi, S. and Chung, A. 2025. Automated feedback improves teachers' questioning quality in brick-and-mortar classrooms: Opportunities for further enhancement. *Computers & Education*. 227, (Apr. 2025), 105183. DOI:https://doi.org/10.1016/j.compedu.2024.105183.

[11]Donnelly, P.J., Blanchard, N., Olney, A.M., Kelly, S., Nystrand, M. and D'Mello, S.K. 2017. Words

matter: automatic detection of teacher questions in live classroom discourse using linguistics, acoustics, and context. *Proceedings of the Seventh International Learning Analytics & Knowledge Conference* (Vancouver British Columbia Canada, Mar. 2017), 218–227.

[12]Freeman, J., Simonsen, B., Briere, D.E. and MacSuga-Gage, A.S. 2014. Pre-Service Teacher Training in Classroom Management: A Review of State Accreditation Policy and Teacher Preparation Programs. *Teacher Education and Special Education: The Journal of the Teacher Education Division of the Council for Exceptional Children*. 37, 2 (May 2014), 106–120. DOI:https://doi.org/10.1177/0888406413507002.

[13]Gable, R.A., Tonelson, S.W., Sheth, M., Wilson, C. and Park, K.L. 2012. Importance, Usage, and Preparedness to Implement Evidence-based Practices for Students with Emotional Disabilities: A Comparison of Knowledge and Skills of Special Education and General Education Teachers. *Education and Treatment of Children*. 35, 4 (2012), 499–520. DOI:https://doi.org/10.1353/etc.2012.0030.

[14]Haghani, P., Narayanan, A., Bacchiani, M., Chuang, G., Gaur, N., Moreno, P., Prabhavalkar, R., Qu, Z. and Waters, A. 2018. From Audio to Semantics: Approaches to End-to-End Spoken Language Understanding. *2018 IEEE Spoken Language Technology Workshop (SLT)* (Athens, Greece, Dec. 2018), 720–726.

[15]Jacobs, J., Scornavacco, K., Clevenger, C., Suresh, A. and Sumner, T. 2024. Automated feedback on discourse moves: teachers' perceived utility of a professional learning tool. *Educational technology research and development*. 72, 3 (Jun. 2024), 1307–1329. DOI:https://doi.org/10.1007/s11423-023-10338-6.

[16]James, A., Kashyap, M., Victoria Chua, Y.H., Maszczyk, T., Nunez, A.M., Bull, R. and Dauwels, J. 2018. Inferring the Climate in Classrooms from Audio and Video Recordings: A Machine Learning Approach. *2018 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE)* (Wollongong, NSW, Dec. 2018), 983–988.

[17]Jensen, E., L. Pugh, S. and K. D'Mello, S. 2021. A Deep Transfer Learning Approach to Modeling Teacher Discourse in the Classroom. *LAK21: 11th International Learning Analytics and Knowledge Conference* (Irvine CA USA, Apr. 2021), 302–312.

[18]Ledford, J.R. and Gast, D.L. 2024. *Single Case Research Methodology: Applications in Special Education and Behavioral Sciences*. Routledge.

[19]Ma, R., Liusie, A., Gales, M.J.F. and Knill, K.M. 2023. Investigating the Emergent Audio Classification Ability of ASR Foundation Models. arXiv.

[20]Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C. and Sutskever, I. 2022. Robust Speech Recognition via Large-Scale Weak Supervision. *arXiv preprint arXiv:2212.04356*. (2022), 28.

[21]Rathod, S., Charola, M. and Patil, H.A. 2023. Noise Robust Whisper Features for Dysarthric Severity-Level Classification. *Pattern Recognition and Machine Intelligence*. P. Maji, T. Huang, N.R. Pal, S. Chaudhury, and R.K. De, eds. Springer Nature Switzerland. 708–715.

[22]Ravanelli, M. et al. 2021. SpeechBrain: A General-Purpose Speech Toolkit. arXiv.

[23]Schlotterbeck, D., Uribe, P., Araya, R., Jimenez, A. and Caballero, D. 2021. What Classroom Audio Tells About Teaching: A Cost-effective Approach for Detection of Teaching Practices Using Spectral Audio Features. *LAK21: 11th International Learning Analytics and Knowledge Conference* (Irvine CA USA, Apr. 2021), 132–140.

[24]Suresh, A., Jacobs, J., Clevenger, C., Lai, V., Tan, C., Martin, J.H. and Sumner, T. 2021. Using AI to Promote Equitable Classroom Discussions: The TalkMoves Application. *Artificial Intelligence in Education*. I. Roll, D. McNamara, S. Sosnovsky, R. Luckin, and V. Dimitrova, eds. Springer International Publishing. 344–348.

[25]Suresh, A., Jacobs, J., Perkoff, M., Martin, J.H. and Sumner, T. 2022. Fine-tuning Transformers with Additional Context to Classify Discursive Moves in Mathematics Classrooms. *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)* (Seattle, Washington, 2022), 71–81.

[26]Tapp, J., Wehby, J. and Ellis, D. 1995. A multiple option observation system for experimental studies: MOOSES. *Behavior Research Methods, Instruments, & Computers*. 27, 1 (Mar. 1995), 25–31. DOI:https://doi.org/10.3758/BF03203616.

[27]Wallace, T., Anderson, A.R., Bartholomay, T. and Hupp, S. 2002. An Ecobehavioral Examination of High School Classrooms That Include Students with Disabilities. *Exceptional Children*. 68, 3 (Apr. 2002), 345–359. DOI:https://doi.org/10.1177/001440290206800304.

[28]Wei, R.C., Darling-Hammond, L. and Adamson, F. 2010. *Professional Development in the United States: Trends and Challenges. Phase II of a Three-Phase Study. Executive Summary*. National Staff Development Council.

[29]Yoder, P.J. and Symons, F.J. 2010. *Observational measurement of behavior*. Springer Pub. Co.