EasyChair Preprint

№ 15289

Course Stage Recognition for Online Course Recordings Using Spoken Language Understanding

Yi-Ting Yuan, Ke-Ching Hong and Yu-Hsiang Wang

October 23, 2024

Course Stage Recognition for Online Course Recordings Using Spoken Language Understanding

Yi-Ting Yuan, Ke-Ching Hong
Chingshin Academy,
Taipei, Taiwan
{yiting9510,anniehung1030}@gmail.com

Yu-Hsiang Wang *
Department of Electrical Engineering,
National Taiwan University,
Taipei, Taiwan
r11921057@ntu.edu.tw

Abstract

This study investigates models for course stage recognition, a novel task in Spoken Language Understanding (SLU) aimed at segmenting classroom recordings into distinct instructional phases. Two approaches are evaluated: an end-to-end SLU model based on the WavLM base+ speech encoder, and a multistage SLU method integrating Whisper for Automatic Speech Recognition and ChatGPT 4o for Natural Language Understanding. The study compares the performance of these models to explore stage recognition without relying on intermediate text representations. Results indicate that the multistage approach excels in fine-grained classification across five stages—Opening, Lecture, Break, Conclusion, and Others—but is outperformed by the end-to-end model in distinguishing the Lecture stage. The findings suggest that a speech-language model capable of performing in-context learning directly on speech data could further enhance the accuracy of course stage recognition.

Keywords: Course Stage Recognition, Spoken Language Understanding, Speech Model, Large Language Model

1 Introduction

With the rapid advancement of artificial intelligence(AI), speech recognition and large language models (LLMs) have significantly reduced the cost of human-computer interaction, leading to widespread applications in various fields, including education. AI has been applied to assist children's learning (Okur et al., 2023), automatically assess students' attention (Parambil et al., 2022), and analyze classroom discourse to enhance teaching quality (Wang et al., 2024). Although deep learning models have been successfully used in these domains, AI applications for segmenting classroom activities based on content to support teaching organizations mostly rely on traditional machine learning models classroomdiscourse1,classroomdiscourse2,classroomdiscourse3.

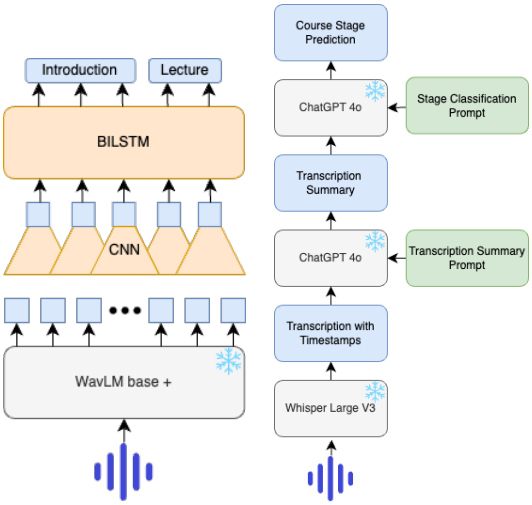
Classroom activity segmentation can be achieved with a multistage Spoken Language Understanding (SLU) model, using Whisper Large V3 (Radford et al., 2023) for ASR and ChatGPT, the gpt-4o version [1] for NLU. Whisper's noise resilience and multilingual support make it ideal for classroom recordings, while ChatGPT can often interpret correct meanings despite ASR errors. This combination reduces the impact of ASR errors on NLU and enables training-free classroom activity segmentation. However, the reliance on text limits its effectiveness for low-resource or unwritten languages.

To explore alternatives to the multistage SLU approach for segmenting classroom activities, we propose the task of course stage recognition. The goal is to segment long classroom recordings into five categories: Opening, Lecture, Break, Conclusion, and Others.

We developed an end-to-end SLU model using the self-supervised learning speech encoder WavLM base+ (Chen et al., 2022) to extract speech features, followed by Convolutional Neural Networks to reduce the length of the speech features sequence, and Bidirectional Long Short-Term Memory (Schuster and Paliwal, 1997) to predict the classroom stage category for each time frame. This study compares the performance of end-to-end

*Corresponding authors

[1] https://platform.openai.com/docs/models/gpt-4o



(a) (b)

Figure 1: The proposed end-to-end and multistage spoken language understanding (SLU) model. (a) The end-to-end model uses WavLM (Chen et al., 2022) for sequence labeling. (b) The multistage model uses Whisper Large V3 (Radford et al., 2023) for ASR and ChatGPT4-o[2] for NLU.

and multistage SLU models in the classroom stage recognition task. Figure 1 illustrates the model architecture of both SLU models. We summarize the contributions of this study in the following:

- We propose course stage recognition, a novel SLU task designed for long audio recordings.
- We analyze the advantages, limitations, and trade-offs between multistage and end-to-end approaches in the context of course stage recognition.
- To enhance transparency and promote further research, we publicly release the dataset and code used in this study at `https://github.com/yiting9510/Course-Stage-Recognition`.

2 Related Works

In previous studies aimed at improving teaching quality, researchers employed random forest models to classify three types of classroom activities: teacher lecturing, whole class discussion, and student group work, using audio data (Wang et al., 2014). Other studies applied Naïve Bayes classifiers on audio (Donnelly et al., 2016a) and multi-sensor (Donnelly et al., 2016b) data to recognize five instructional segments: question and answer, procedures and directions, supervised seatwork, small group work, and lecture. Our research leverages a more advanced deep learning model to analyze online course audio, categorizing it into five segments: Opening, Lecture, Break, Conclusion, and Others.

Course stage recognition is built upon Spoken Language Understanding (SLU), with two primary approaches: a multistage approach (Bastianelli et al., 2020) that separates Automatic Speech Recognition (ASR) and Natural Language Understanding (NLU), and an end-to-end approach (Wang et al., 2023; Arora et al., 2024). While the multistage method allows independent ASR and NLU training, it is vulnerable to ASR errors in noisy environments like classrooms (Schlotterbeck et al., 2022). The end-to-end approach, although potentially could have better performance, often faces limited training data, especially in specialized domains such as education.

SLU tasks have traditionally used Long Short-Term Memory (Schmidhuber et al., 1997; Schuster and Paliwal, 1997) (LSTM) and encoder-decoder architectures (Sutskever, 2014; Wu, 2016; Chiu et al., 2018), with recent advances in self-attention mechanisms (Vaswani, 2017) enhancing models' ability to capture dependencies within input sequences. These advancements led to the development of large language models (LLMs), such as GPT (Radford et al., 2018, 2019; Brown, 2020), which have strong capabilities in Natural Language Processing (NLP). Speech models built on transformer architectures, like Wav2Vec 2.0 (Baevski et al., 2020), HuBERT (Hsu et al., 2021), and WavLM (Chen et al., 2022), use self-supervised learning (SSL) to learn to extract robust speech features applicable to various speech tasks (wen Yang et al., 2021; Tsai et al., 2022). Whisper (Radford et al., 2023) is an encoder-decoder ASR transformer-based model trained on large annotated datasets, demonstrating strong ASR performance across languages and robust to noisy conditions.

Given the limited labeled data and potential noise in our online course audio, we leveraged

pre-trained models and off-the-shelf LLM services to enhance performance. Specifically, we compared two models: a multistage design using Whisper large v3 for ASR and with ChatGPT 4o for NLU (He and Garner, 2023), and an end-to-end model combining WavLM base with CNN and LSTM layers. We evaluated the performance of these models in classroom stage recognition and discussed potential areas for improvement.

3 Method

In this study, we propose a novel task of course stage recognition and have constructed a dataset for this purpose. We developed an end-to-end spoken language understanding (SLU) model based on WavLM and compared it to a multistage SLU model combining Whisper and ChatGPT. We analyzed the performance of these systems in the context of course stage classification. Below, we describe the dataset preparation process and then explain the design of the two SLU models.

3.1 Data Preparation

To create a dataset for course stage recognition, we focused on Mandarin online teaching courses in Taiwan, such as recordings of school lessons for junior high and high school students, as well as videos from tutoring centers. We searched for suitable course recordings on online video platforms and public course websites, extracted the audio, and manually analyzed the content to label different course stages.

Based on the framework outlined by (Davis, 2009), courses can be divided into nine stages: Introduction, Opening, Lecture, Presentation, Break, Transition, Conclusion, Summary, and Others. We used this definition as a reference but merged several similar stages, ultimately classifying the course stages into the following five categories: **Opening** includes teacher greetings and pre-class reminders, which are either unrelated to the main content or provide recaps of previous lessons. **Lecture** covers all content related to the course topic, including guided exercises. **Break** refers to silent periods after the teacher announces a break, as well as off-topic conversations. **Conclusion** summarizes what was covered in the lesson, previews the next session, and includes farewells. **Others** encompasses any other periods, such as time before the class officially begins, after it ends, or interruptions due to technical issues.

Additionally, we defined a simplified classroom stage recognition task, in which courses are categorized into only two stages: Lecture and Others. In this setting, the original Opening, Lecture, and Conclusion stages are grouped under the broad category of Lecture, while Break and any other non-instructional periods are grouped into the Others category.

We manually collected and annotated the recordings according to these standards, labeling each time segment according to its corresponding classroom stage. We ensured that each time segment in the course had only one assigned stage. In total, we collected 65 course recordings and manually annotated the classroom stages. The statistical data of the dataset is shown in Table 1. Notably, the average course length was 15 minutes, which required the method to handle long audio files effectively. The proportion of Lecture stage recordings was 87.8%, significantly higher than the other four stages, necessitating methods capable of addressing extreme class imbalance.

Statistic Description	**Value**
Total Recordings	65
Total Duration (hrs)	90.67
Average Length (min)	15.15
Stage Duration Percentages	
Opening Stage %	2.50%
Lecture Stage %	87.80%
Break Stage %	2.80%
Conclusion Stage %	2.60%
Other Stage %	4.30%

Table 1: Course Recording Statistics

Our observations also revealed that not all five stages were present in every course, and transitions between stages did not always follow fixed patterns. This indicates that methods must be able to analyze contextual information to accurately perform classroom stage recognition.

For the testing dataset, a pre-selected set of 10 recordings was used to ensure coverage of

all stages. The remaining 55 recordings are split into training and validation sets at an approximate 8:2 ratio based on their total duration. Two different training-validation splits were randomly sampled, and the average score across these splits was used as the final performance metric.

3.2 Model Design

3.2.1 End-to-End SLU Model

The proposed architecture integrates a Transformer-based Self-Supervised Learning (SSL) speech encoder, Convolutional Neural Networks (CNN), and Bidirectional Long Short-Term Memory (BiLSTM) layers, addressing the task as a sequence labeling problem.

Figure 1a illustrates the end-to-end SLU model structure. The SSL speech encoder, specifically the WavLM Base+ model pretrained on large-scale speech data, is chosen due to its proven ability to extract robust speech features that generalize well across various speech processing tasks. These pretrained models have shown superior performances (wen Yang et al., 2021; Tsai et al., 2022; Feng et al., 2023), even when the encoder remains frozen, making them well-suited for the course stage recognition task where labeled data may be scarce. Since the repesentation extracted by the speech encoder is too long for efficient training in stage classification, a CNN is applied to further reduce the sequence length, followed by a BiLSTM layer that classifies each time frame into its corresponding course stage.

In this architecture, the SSL encoder remains frozen, and only the CNN and BiLSTM layers are trained. Experiments were conducted with 2-stage configurations: a detailed 5-stage classification and a simplified 2-stage classification. Due to the computational cost of Transformer-based models increasing rapidly with input length, the maximum input duration is limited to 30 seconds. For longer audio files, a sliding window approach is used to segment them into 30-second chunks with a 10-second overlap.

To address the issue of data imbalance, particularly the overrepresentation of the "Lecture" stage, several data augmentation techniques were applied. In addition to downsampling the "Lecture" segments to match the second most frequent class and upsampling non-Lecture segments until their total number reached approximately one-third of the total "Lecture" segments, we also employed augmentation techniques such as TimeStretch and Gaussian noise. TimeStretch was used to slightly alter the speed of the audio without affecting its pitch, while Gaussian noise was added to enhance robustness against noise in the input data. These augmentations helped improve model generalization and performance, especially in cases where training data was limited.

For model training, cross-entropy loss was used as the loss function. To evaluate model performance, the F1 score was chosen as the primary metric due to its ability to balance precision and recall, especially in imbalanced datasets. The F1 score was computed for each class and then aggregated using either macro-averaging, where all classes are treated equally, or micro-averaging, which in this case is equivalent to accuracy. This provided a comprehensive measure of the model's classification performance.

$$\text{Precision}_c = \frac{TP_c}{TP_c + FP_c} \tag{1}$$

$$\text{Recall}_c = \frac{TP_c}{TP_c + FN_c} \tag{2}$$

$$F1_c = \frac{2 \cdot \text{Precision}_c \cdot \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c} \tag{3}$$

$$\text{Macro-F1} = \frac{1}{C} \sum_{c=1}^{C} F1_c \tag{4}$$

During testing, the model's predictions are converted into time-based segments, with start time, end time, and the corresponding stage. Both predictions and ground truth are then converted into one-dimensional arrays with 1,000 frames per second, allowing precise alignment. The F1 score is calculated based on these frame-level vectors to evaluate the model's performance.

3.2.2 Multistage SLU Model

We developed a multistage SLU model combining Whisper large v3 and ChatGPT 4o for classroom stage classification, as illustrated in

Model	Macro-F1 score	Micro-F1 score
End-to-End	0.24559/0.88203	0.82984/0.97111
End-to-End w/ augmentation	0.33477/0.85778	0.84194/0.96629
End-to-End w/ augmentation & undersampling	0.30493/0.88434	0.84299/0.97160
End-to-End w/ augmentation & oversampling	0.39199/**0.91195**	0.85136/**0.97736**
Multistage	**0.49196**/0.74734	0.**85295**/0.94295

Table 2: Comparison of End-to-End model(WavLM base+ with CNN and BiLSTM) and Multistage model(Whisper Large v3 and ChatGPT 4o) performance for class stage recognition performance, where performance under 5 stage (left) and 2 stage (right) settings are shown.

Figure 1b. The model operates in two stages: first, Whisper large v3 performs automatic speech recognition (ASR), converting audio into text with timestamps. We use WhisperX (Bain et al., 2023) for more accurate long-form transcription and to reduce hallucination.

In the second stage, ChatGPT 4o processes the transcribed text for natural language understanding (NLU). It analyzes the transcription to infer classroom stages. Given the length of classroom sessions, ChatGPT 4o uses a two-pass process: first, summarizing chunks of transcription (up to 30 minutes each) to reduce the length, then analyzing the summaries to classify each time segment into its respective classroom stage.

The key benefit of this approach is the use of powerful pre-trained models that require no additional training. With appropriate prompts, the system can accurately predict stages on test data, even when ASR errors are present in the transcription. Performance is evaluated by comparing the predicted stage start/end times with the ground truth using the F1 score.

4 Experimental Results

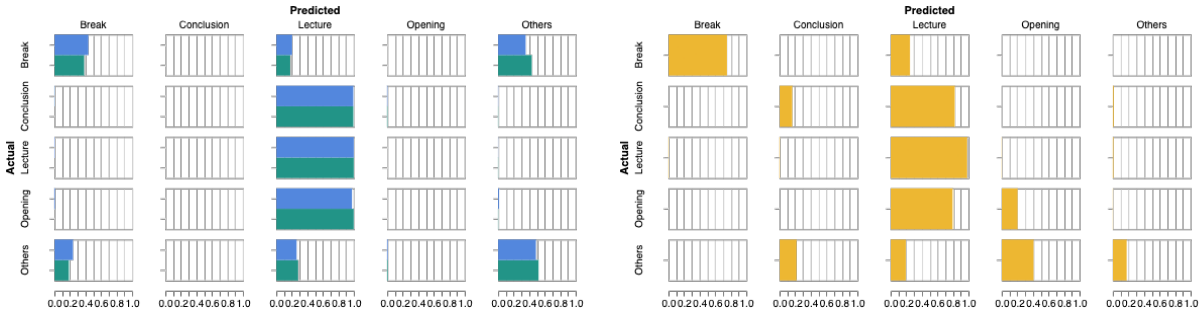
4.1 Model Performance Evaluation

As shown in Table 2, for the 5-stage course classification, the multistage model achieves the highest Macro-F1 score (0.49196) and slightly outperforms the end-to-end models in Micro-F1 score (0.85295 vs. 0.85136). This indicates the multistage model handles class imbalance better, particularly for less frequent categories, resulting in a superior Macro-F1 score. Among the end-to-end models, the one using augmentation and oversampling performs best, with a Macro-F1 score of 0.39199 and Micro-F1 score of 0.85136, though it still lags behind the multistage model in handling imbalanced data.

In the 2-stage classification, both models improve significantly. The end-to-end model with augmentation and oversampling achieves the highest Micro-F1 (0.97736) and Macro-F1 (0.91195) scores, outperforming the multistage model. This result reflects the simpler task's reduced complexity, where the end-to-end model excels by focusing on the two main categories, "Lecture" and "Others." While the multistage model performs decently with a Micro-F1 of 0.94295, it shows a larger gap in Macro-F1 (0.74734), highlighting its less effective handling of the simplified task.

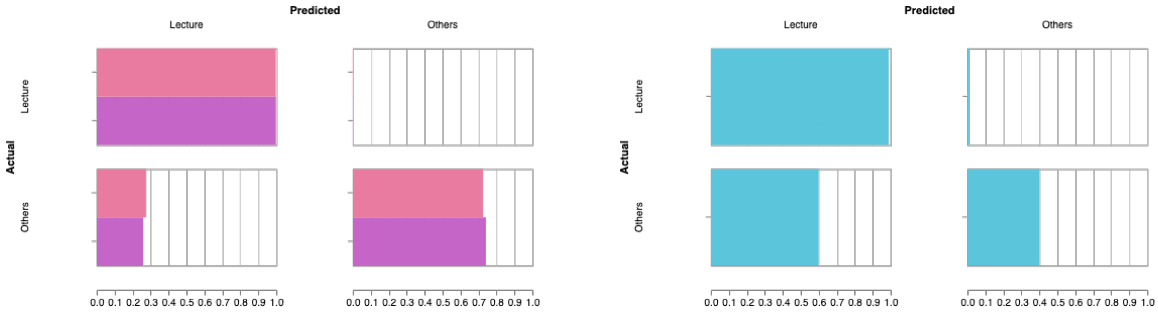
Confusion matrices (Figures 2a and 2b) reveal that in the 5-stage classification, the end-to-end model tends to overpredict the "Lecture" category, leading to an imbalanced performance, while the multistage model distributes predictions more evenly, contributing to its higher Macro-F1 score. However, in the 2-stage classification, the end-to-end model performs better, reducing prediction imbalance (Figures 3a and 3b).

In summary, each model demonstrates distinct advantages. The end-to-end model performs exceptionally well in the 2-stage task, achieving the highest Micro-F1 and Macro-F1 scores. On the other hand, the multistage model shows superior performance in the more complex 5-stage task, particularly in handling class imbalances. However, both models exhibit limitations: the end-to-end model faces challenges in distinguishing between stages with similar characteristics, while the multistage model risks losing important information, which could impair its ability to accurately recognize certain categories.



(a) (b)

Figure 2: 5-stage recognition Confusion matrix of (a) End-to-end model with data augmentation and oversampling (b) multistage model



(a) (b)

Figure 3: 2-stage recognition Confusion matrix of (a) End-to-end model with data augmentation and oversampling (b) multistage model

4.2 Applications and Impacts

The rise of online educational resources has led to a lack of proper segmentation in video and audio materials, mainly due to the high cost of manual annotation. Course stage recognition addresses this by helping students quickly find relevant content, improving their learning experience. For educators, it offers insights into course structure, enabling better content organization. Educational platforms benefit from more accessible and streamlined content. Our research also advances textless spoken language understanding (SLU), particularly for low-resource and unwritten languages, promoting broader access to educational resources for underrepresented language communities.

4.3 Limitation and Future Work

The end-to-end SLU model in this study faces the challenge of having too short a context window, making it difficult to capture long-term dependencies, which results in an inability to differentiate between similar course stages. On the other hand, multistage models, constrained by their modular design, are prone to losing information, making certain stages harder to recognize. Additionally, relying on ChatGPT for NLU in multistage SLU raises privacy concerns for certain applications.

Future work includes developing an end-to-end SLU model capable of in-context learning. This could be achieved by incorporating a Speech Language Model (SLM), as suggested in recent work (Hsu et al., 2023). The goal would be to use trainable prompts, enabling the SLM (Lakhotia et al., 2021; Kharitonov et al., 2021) to perform SLU on the entire classroom recording, while preserving rich speech information. However, the main challenge is the high computational cost of handling long input sequences.

5 Conclusion

This paper introduces course stage recognition, a novel SLU task aimed at segmenting course content using audio. We propose two models: an end-to-end model based on WavLM and a multistage SLU model using Whisper for transcription and ChatGPT for text under-

standing. Experimental results demonstrate that both approaches show promising capabilities but have limitations. The end-to-end model can recognize some of the rarer stages but struggles with distinguishing other similar stages, while the multistage model effectively differentiates stages through text analysis but performs worse than the end-to-end model in identifying some of the rarer stages. These results highlight the challenges of course stage recognition. Future work includes developing a speech-language model with in-context learning on speech data to improve performance. We have made our dataset and code publicly available to encourage further research.

Declaration of the Use of Generative AI and AI-assisted Technologies in Writing

During the preparation of this paper, the author(s) used ChatGPT for writing improvement. After using these tools, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

References

Siddhant Arora, Hayato Futami, Jee-weon Jung, Yifan Peng, Roshan Sharma, Yosuke Kashiwagi, Emiru Tsunoo, Karen Livescu, and Shinji Watanabe. 2024. Universlu: Universal spoken language understanding for diverse tasks with natural language instructions. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2754–2774.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *NeurIPS*.

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. Whisperx: Time-accurate speech transcription of long-form audio. *arXiv preprint arXiv:2303.00747*.

Emanuele Bastianelli, Andrea Vanzo, Pawel Swietojanski, and Verena Rieser. 2020. Slurp: A spoken language understanding resource package. *arXiv preprint arXiv:2011.13205*.

Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE J. Sel. Top. Signal Process.*, 16(6):1505–1518.

Chung-Cheng Chiu, Tara N Sainath, Yonghui Wu, Rohit Prabhavalkar, Patrick Nguyen, Zhifeng Chen, Anjuli Kannan, Ron J Weiss, Kanishka Rao, Ekaterina Gonina, et al. 2018. State-of-the-art speech recognition with sequence-to-sequence models. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4774–4778. IEEE.

Barbara Gross Davis. 2009. *Tools for teaching*. Jossey-Bass.

Patrick J Donnelly, Nathan Blanchard, Borhan Samei, Andrew M Olney, Xiaoyi Sun, Brooke Ward, Sean Kelly, Martin Nystran, and Sidney K D'Mello. 2016a. Automatic teacher modeling from live classroom audio. In *Proceedings of the 2016 conference on user modeling adaptation and personalization*, pages 45–53.

Patrick J Donnelly, Nathaniel Blanchard, Borhan Samei, Andrew M Olney, Xiaoyi Sun, Brooke Ward, Sean Kelly, Martin Nystrand, and Sidney K D'Mello. 2016b. Multi-sensor modeling of teacher instructional segments in live classrooms. In *Proceedings of the 18th ACM international conference on multimodal interaction*, pages 177–184.

Tzu-hsun Feng, Annie Dong, Ching-Feng Yeh, Shuwen Yang, Tzu-Quan Lin, Jiatong Shi, Kai-Wei Chang, Zili Huang, Haibin Wu, Xuankai Chang, et al. 2023. Superb@ slt 2022: Challenge on generalization and efficiency of self-supervised speech representation learning. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 1096–1103. IEEE.

Mutian He and Philip N Garner. 2023. Can chatgpt detect intent? evaluating large language models for spoken language understanding. *arXiv preprint arXiv:2305.13512*.

Ming-Hao Hsu, Kai-Wei Chang, Shang-Wen Li, and Hung-yi Lee. 2023. An exploration of in-context learning for speech language model. *arXiv preprint arXiv:2310.12477*.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE ACM Trans. Audio Speech Lang. Process.*, 29:3451–3460.

Eugene Kharitonov, Ann Lee, Adam Polyak, Yossi Adi, Jade Copet, Kushal Lakhotia, Tu-Anh Nguyen, Morgane Rivière, Abdelrahman Mohamed, Emmanuel Dupoux, et al. 2021. Text-free prosody-aware generative spoken language modeling. *arXiv preprint arXiv:2109.03264*.

Kushal Lakhotia, Eugene Kharitonov, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte, Tu-Anh Nguyen, Jade Copet, Alexei Baevski, Abdelrahman Mohamed, et al. 2021. On generative spoken language modeling from raw audio. *Transactions of the Association for Computational Linguistics*, 9:1336–1354.

Eda Okur, Roddy Fuentes Alba, Saurav Sahay, and Lama Nachman. 2023. Inspecting spoken language understanding from kids for basic math learning at home. *arXiv preprint arXiv:2306.00482*.

Medha Mohan Ambali Parambil, Luqman Ali, Fady Alnajjar, and Munkhjargal Gochoo. 2022. Smart classroom: A deep learning approach towards attention assessment through class behavior detection. In *2022 Advances in Science and Engineering Technology International Conferences (ASET)*, pages 1–6. IEEE.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. In *OpenAI*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Danner Schlotterbeck, Abelino Jiménez, Roberto Araya, Daniela Caballero, Pablo Uribe, and Johan Van der Molen Moris. 2022. "teacher, can you say it again?" improving automatic speech recognition performance over classroom environments with limited data. In *International Conference on Artificial Intelligence in Education*, pages 269–280. Springer.

Jürgen Schmidhuber, Sepp Hochreiter, et al. 1997. Long short-term memory. *Neural Comput*, 9(8):1735–1780.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681.

I Sutskever. 2014. Sequence to sequence learning with neural networks. *arXiv preprint arXiv:1409.3215*.

Hsiang-Sheng Tsai, Heng-Jui Chang, Wen-Chin Huang, Zili Huang, Kushal Lakhotia, Shu-Wen Yang, Shuyan Dong, Andy T. Liu, Cheng-I Lai, Jiatong Shi, Xuankai Chang, Phil Hall, Hsuan-Jui Chen, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung-yi Lee. 2022. SUPERB-SG: enhanced speech processing universal performance benchmark for semantic and generative capabilities. In *ACL (1)*, pages 8479–8492. Association for Computational Linguistics.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Deliang Wang, Yang Tao, and Gaowei Chen. 2024. Artificial intelligence in classroom discourse: A systematic review of the past decade. *International Journal of Educational Research*, 123:102275.

Minghan Wang, Yinglu Li, Jiaxin Guo, Xiaosong Qiao, Zongyao Li, Hengchao Shang, Daimeng Wei, Shimin Tao, Min Zhang, and Hao Yang. 2023. Whislu: End-to-end spoken language understanding with whisper. In *Proc. Interspeech*, volume 2023, pages 770–774.

Zuowei Wang, Xingyu Pan, Kevin F Miller, and Kai S Cortina. 2014. Automatic classification of activities in classroom discourse. *Computers & Education*, 78:115–123.

Yonghui Wu. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Shu wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu-Hsien Huang, Wei-Cheng Tseng, Ko tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung yi Lee. 2021. SUPERB: Speech Processing Universal PERformance Benchmark. In *Proc. Interspeech 2021*, pages 1194–1198.