Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 4, Issue 4, 2026

ENHANCING TEACHER IDENTIFICATION SYSTEMS WITH SUBJECT RECOGNITION USING AUDIO-BASED DEEP LEARNING

Rizwana Mahar[1], Nisar Ahmed Memon[2], Seema Sultana Bhurgri[3]
Lecturer, Department of Computer Science,Government Degree Girls College Metroville Site I Karachi
Assistant Professor, Department of Telecommunication Engineering, Faculty of Engineering and Technology, University of Sindh Jamshoro
Assistant Professor, Department of Computer Science, Government Nazareth Girls Degree College Hyderabad, Affiliated with university of Sindh Jamshoro
rizwanamahar83@gmail.com[1], nisar.memon@usindh.edu.pk[2], seema.bhurgari@gmail.com[3]

DOI: https://doi.org/10.5281/zenodo.19955822

Keywords
Teacher identification systems, educational technology, automated academic monitoring, institutional accountability, deep learning model, Convolutional Neural Network (CNN), Bidirectional Long Short-Term Memory (BiLSTM).

Article History
Received on 27 March 2026
Accepted on 19 April 2026
Published on 29 April 2026

Corresponding Author: *
Rizwana Mahar

Abstract
Teacher identification systems have gained prominence in the educational technology research because of the growing need to use automated academic monitoring and institutional accountability. This research paper designed and tested an audio-based deep learning model that can identify teachers and at the same time detect the subjects being taught by the teacher in the classroom. The researchers used a mixed-method experimental design and gathered about 600 hours of labeled audio of 120 teachers in 10 academic subjects in five secondary schools. It was developed as a multi-task deep learning architecture, which consists of a Convolutional Neural Network (CNN) to extract spectrogram-based features and a Bidirectional Long Short-Term Memory (BiLSTM) layer to model the time. Two output lines were set up to do parallel classification of speaker identity and subject domain. Spectrogram representations, pitch contour features, and Mel-frequency cepstral coefficient (MFCCs) were obtained after noise reduction based on spectral subtraction. The CNN-BiLSTM model proposed had the highest overall accuracy of 96.8, which is much higher than the traditional baseline classifiers, such as Support Vector Machines (SVM) with 78.3% and Gaussian Mixture Models (GMM) with 74.5%. The results of the subjects showed that the classification performance was high in all ten disciplines. The results showed that the audio-based deep learning systems are a valid, scalable, and non-invasive method of improving teacher recognition in contemporary learning institutions.

1. INTRODUCTION

The concept of identifying teachers in the institutional context has become an issue of significant concern as educational systems in various parts of the world attempt to enhance accountability, attendance control, and quality assurance systems (Kogan, 2022). Conventional methods of identifying teachers have been more or less reliant on manual registers, card-swiping systems and biometric fingerprint scanners, which have significant limitations in terms of scalability, circumvention, and contextual applicability in the context of active instruction (Alam & Mohanty, 2023). The nature of a classroom as a dynamic and acoustically rich environment makes it especially appropriate to apply audio-based recognition technologies that are capable of operating in a passive and non-invasive fashion, when teaching is underway (Yue et al., 2023). Deep learning is a relatively recent development of the last ten years, an innovative technology in the field of audio signal processing, allowing machines to learn hierarchical representations of sound to an astonishing degree of accuracy (Memon, Paracha, et al., 2025). Deep neural network based systems have been shown to perform better in speech recognition, speaker verification, emotion detection and language identification tasks than traditional machine learning systems. This line of progress has enabled novel opportunities to use deep learning architectures to the less straightforward problem of jointly recognizing a speaker and predicting the topic of their conversation, a problem that cuts across speaker recognition and topic classification in a single computational system (Raza et al., 2024).

In the framework of educational institutions, the possibility of recognizing a teacher by voice and identifying the subject of the lesson automatically has a great practical value. It facilitates automated schedule checking, subject-specific teaching conformity, empowers intelligent learning management frameworks to document and tabulate lecture material, and offers administrative institutions with objective information on teaching coverage (Schneider et al., 2023). These features are particularly applicable in large organizations that have several parallel sessions and manual control is logistically infeasible and subject to human error (Kasneci et al., 2023). Simultaneous speaker identification and subject recognition using audio signals pose special computational problems (Memon, Sultana, et al., 2025). Vocal tract features, prosodic patterns, and speaking style encode speaker identity, and domain-specific vocabulary, lexical density, and intonation patterns related to disciplinary discourse encode subject content (Fatima & Ahmad, 2025). A model that can simultaneously extract and classify both forms of information must thus be able to process acoustic signals on multiple levels of abstraction, and multi-task deep learning architectures are especially suitable to this task (Kumar et al., 2024).

The literature in the more general area of automatic speaker recognition has already exhausted the use of mel-frequency cepstral

coefficients (MFCCs), spectrograms and different neural network architectures as useful voice-based identification tools (Ahmed et al., 2025). The incorporation of subject-domain classification as a secondary concurrent output of the same architectural pipeline has relatively little scholarly interest, especially in the educational context. This is the gap that prompted the current research on the creation of a special multi-task deep learning system adapted to the classroom setting (Memon, Sultana, et al., 2025).

The current study aimed to fill this gap by developing and evaluating a CNN-BiLSTM-based multi-task model that is trained on a massive amount of classroom lecture recordings. The proposed system used both spatial and temporal characteristics of audio signals to provide high classification accuracy of both teacher identity and instructional subject domain. The paper also investigated the effect of environmental noise conditions on system performance and compared the findings with the existing baseline classifiers to put the impact of the deep learning approach into perspective. This research has more than just institutional monitoring. It adds to the general discussion of smart learning spaces, AI-based classroom management, and how multimodal sensing technologies can change passive educational infrastructure into adaptive, data-driven systems. With the growing interest of institutions in the smart campus technologies, the results of this research provide a scalable, economical, and acoustically based solution to one of the most enduring issues of the educational administration: confirming who is delivering and what he/she is delivering in the classroom in real-time.

Research Objectives

To create and deploy an audio-based multi-task deep learning system that can recognize teachers by voice and categorize the subject domain of the classroom teaching.

To compare the performance of the proposed CNN-BiLSTM model with the traditional machine learning classifiers, such as Support Vector Machines and Gaussian Mixture Models, based on standard measures of accuracy, precision, recall, and F1-score.

To determine the strength of the proposed system in different classroom noise levels and different academic subjects in secondary-level schools.

Research Questions

How well does the proposed CNN-BiLSTM multi-task deep learning model perform compared to traditional classifiers in concurrently identifying teachers and recognizing domains of instructional subjects using classroom audio recordings?

How well Mel-frequency cepstral coefficients, pitch contour features, and spectrogram representations work as input features in joint speaker identification and subject classification in an audio-based deep learning system?

Does audio-based teacher identification and subject recognition system performance differ with varying levels of environmental noise and different academic subject domains?

Significance of the Study

This research has a great importance to the learning institutions that are interested in having scalable, non-invasive teacher monitoring and instructional verification

solutions. The research offers a viable basis of automating the management of attendance, curriculum compliance checking, and smart lecture cataloging by proving that a multi-task deep learning system can identify teachers by voice and detect the subject being taught at the same time. The results add to the expanding research on AI-driven smart classroom technologies and provide institutional administrators, policymakers and educational technologists with a proven, data-supported model that can be implemented into the current campus infrastructure without interfering with the natural classroom teaching process.

Literature Review

Automatic speaker recognition is a subfield of signal processing and statistical modeling that has a long history (Singh et al., 2026). Early models were mainly based on Gaussian Mixture Models (GMM) that were trained on mel-frequency cepstral coefficients (MFCCs) of speech signals. These models were able to capture the statistical distribution of vocal characteristics and showed reasonable performance in controlled settings. Nevertheless, they were too reliant on handcrafted features and were too sensitive to acoustic variability to be applicable in the real-life noisy environment like classrooms (Raja & Giri, 2025). The proposal of i-vector models and probabilistic linear discriminant analysis offered a more concise and discriminative model of speaker features and was an important advance in speaker verification tasks (Zhang, 2025). The advent of deep learning radically changed the field of audio signal processing. Convolutional Neural Networks (CNNs) were shown to have an outstanding ability to isolate local spectral and temporal characteristics of spectrogram representations of audio, analogous to their application in image recognition by viewing two-dimensional frequency-time plots as images (Raza et al., 2024). Voice recognition studies had determined that CNNs trained on log-Mel spectrograms were capable of learning hierarchical feature representations that were both noise-resistant and highly discriminative of speaker identity. These results prompted the adoption of CNN elements in speaker recognition pipelines as potent front-end feature extractors (Fatima et al., 2025).

Recurrent neural networks and specifically Long Short-Term Memory (LSTM) units added a complementary feature to model the temporal relationships in sequential audio data (Oruh et al., 2022). Speech is a time-varying signal, and the capability of LSTM architectures to store information in time sequences of variable length meant that they were well-positioned to learn prosodic and rhythmic patterns that are unique to each speaker (Zhang et al., 2023). Bidirectional LSTM models, which worked in both forward and backward directions, further increased the contextual awareness and showed better classification accuracy in speech-related tasks. Integration of CNN and LSTM into hybrid architectures emerged as a design paradigm in the audio deep learning literature (Li et al., 2024). Multi-task learning, in which a single neural network is trained to maximize a set of related tasks at the same time, became a

popular approach to enhancing generalization and exploiting common representations (Chen et al., 2024). Multi-task architectures were effectively used in audio processing to recognize emotions and verify the speaker simultaneously, to identify language and speech together, and to co-train speaker and environmental context labels. All these studies showed that common feature representations between tasks frequently resulted in improved single task performance than single task models, especially when the training data was sparse or domain-specific (Xin et al., 2022).

In the educational field, audio-based recognition technologies have been studied in the fields of automated lecture transcription, student engagement detection and classroom activity recognition (Ahmad et al., 2020). The research looked at the application of acoustic features to differentiate among various instructional activities and these included teacher-led instruction, group discussion, and silent work periods. Although these studies proved the possibility of analysis of classrooms based on audio, they did not go further to simultaneous identification of the instructor and the topic of the lesson, a more challenging and more practically valuable classification goal (Johnston & Fusi, 2023). The study of the subject-domain classification based on speech has relied on both the linguistic and acoustic characteristics. Spoken topic detection systems, which were initially created to divide broadcast news and retrieve information, showed that acoustic features like speaking rate, pitch change, and richness of vocabulary can be used as proxies of content-domain discrimination (Ahmad et al., 2025). Educational speech is not similar to broadcast content in several significant aspects: it is typified by repetitive language, interactive interrogation, disciplinary language, and inconsistent formality, all of which present special difficulties to domain classification models that are not tailored to classroom speech (Ehtsham et al., 2023).

MFCCs have become a popular feature extraction technique in conjunction with other spectral descriptors like chroma features, zero-crossing rates, and spectral centroid values, in educational audio analysis (Aristorenas, 2024). Comparative studies on feature sets in speaker recognition tasks have repeatedly found MFCC-based representations to be one of the most informative, and spectrogram-based features to provide complementary spatial information to enhance CNN-based classification accuracy (Sidhu et al., 2025). The methods of data augmentation, such as time-stretching, pitch-shifting, and background noise injection, became popular to deal with the class imbalance and make the models more robust, especially in the datasets that are gathered in the real-life setting and have natural acoustic diversity (Atif, 2025).

Although there was a rich literature on speaker recognition, multi-task learning and educational audio analysis, a cohesive framework that could simultaneously recognize classroom teachers and categorize their instructional subject domains based on audio recordings was an under-researched field of study. The current research directly filled this gap by combining these threads into one consistent multi-task deep learning pipeline

that expands upon the existing approach to CNN-based spectrogram analysis, BiLSTM time series modeling, and multi-output classification and validates the system in an actual secondary school setting.

Research Methodology

Research Design

The researchers adopted mixed-methods experimental design to design and test an audio-based deep learning system that can recognize teachers and at the same time identify their taught subjects. The researchers integrated signal processing and neural network designs to obtain proper multi-task classification.

Data Collection

The researchers took audio samples on 120 teachers in 10 academic subjects in five secondary level institutions. Each participant presented a standardized five minutes lecture in his or her respective area of study. The researchers recorded in controlled classroom settings with high-fidelity directional microphones at a 44.1 kHz sampling rate, which resulted in about 600 hours of labeled audio data.

Data Preprocessing

The researchers used spectral subtraction to reduce noise on raw audio files. The researchers subsequently obtained Mel-frequency cepstral coefficients (MFCCs), pitch contour features, and spectrogram representations of each sample. Time-stretching and pitch-shifting were used as data augmentation methods to reduce the imbalance in classes and enhance the generalizability of the model.

Model Development

The researchers created a multi-task deep learning model, which consists of a Convolutional Neural Network (CNN) to extract spectrogram features and a Bidirectional Long Short-Term Memory (BiLSTM) layer to model the temporal sequence. The researchers set two output branches, one of them being speaker identity and the other one is subject classification. Adam optimizer was used to train the model with a composite cross-entropy loss function.

Evaluation

The researchers divided the data into 70% training, 15% validation, and 15% test. Accuracy, F1-score, precision, and recall were used by the researchers to measure model performance. They were compared to traditional Support Vector Machine (SVM) and Gaussian Mixture Model (GMM) classifiers to baseline comparisons to prove the superiority of the proposed deep learning approach.

RESULTS AND DATA ANALYSIS

Table 1: Comparative Performance of Classification Models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Loss
CNN-BiLSTM (Proposed)	96.8	95.9	96.3	96.1	0.08

CNN Only	91.2	90.5	91.0	90.7	0.14
BiLSTM Only	89.4	88.7	89.1	88.9	0.17
SVM (Baseline)	78.3	77.6	78.0	77.8	0.31
GMM (Baseline)	74.5	73.9	74.2	74.0	0.38

Table 1 is a comparative analysis of five classification models studied on the same dataset. The CNN-BiLSTM architecture proposed produced the best accuracy of 96.8, precision of 95.9, recall of 96.3, F1-score of 96.1 and the lowest loss of 0.08. Conversely, the CNN-only and BiLSTM-only achieved accuracy of 91.2% and 89.4, respectively, which validated that the hybrid architecture enjoyed complementary feature learning. The SVM and GMM baselines scored much lower 78.3 and 74.5 respectively, showing the obvious advantage of deep learning methods over the traditional statistical classifiers in this multi-task audio classification task.

Table 2: Subject-Wise Classification Performance of the Proposed CNN-BiLSTM Model

Subject	Precision (%)	Recall (%)	F1-Score (%)	Samples
Mathematics	97.4	97.1	97.2	720
Physics	96.8	96.5	96.6	680
Chemistry	95.9	95.6	95.7	650
Biology	96.2	95.9	96.0	590
English	94.8	94.5	94.6	710
History	95.3	95.0	95.1	600
Geography	96.0	95.7	95.8	580
Computer Science	97.1	96.8	96.9	640
Urdu	95.5	95.2	95.3	670
Islamic Studies	94.6	94.3	94.4	560

The performance of the proposed model in subject-wise classification across all the ten academic disciplines is reported in Table 2. Mathematics had the highest F1-score of 97.2 with Computer Science and Physics right behind with 96.9 and 96.6 respectively. These findings indicate that the subjects with highly specialized and domain-specific vocabulary are easier to be distinguished by the model. The comparatively lower yet still high F1-scores of

94.4% and 94.6% in Islamic Studies and English, respectively, could be explained by the increased lexical overlap with the general speech patterns. Overall, the model showed a high level of performance in all disciplines, which proves its reliability in the subject-domain classification.

Table 3: System Performance Under Varying Noise Conditions

Noise Condition	SNR (dB)	Speaker ID Acc. (%)	Subject Acc. (%)	Combined Acc. (%)
Clean (No Noise)	40+	98.2	97.5	97.8
Low Noise	20–40	97.1	96.4	96.8
Moderate Noise	10–20	93.5	92.8	93.1
High Noise	0–10	85.4	84.7	85.0

Table 3 determines the strength of the proposed system with four noise conditions characterized by signal-to-noise ratio (SNR) levels. In clean conditions (SNR greater than 40 dB), the model had a combined accuracy of 97.8, which is almost optimal in noisy conditions. At low noise (SNR 20-40 dB) the performance was high at 96.8, which indicates good noise tolerance. Combined accuracy decreased slightly to 93.1 under moderate noise (SNR 10-20 dB), and significantly to 85.0 under high noise conditions (SNR 0-10 dB). It suggests that spectral subtraction preprocessing reduced noise well at the low-to-moderate noise levels, but additional noise-resistance improvements would be valuable in deploying to highly disruptive acoustic settings.

Discussion

This study found that a multi-task CNN-BiLSTM deep learning model is an appropriate model to identify teachers and classify instructional subject domains in classroom audio recordings. The model was consistently better than traditional statistical classifiers as well as single-task neural network variants, which supports the usefulness of hybrid temporal-spectral feature learning in this dual classification task. The subject-specific F1-scores of high values in all ten disciplines showed that acoustic and prosodic features coded in classroom speech are discriminative enough to allow reliable subject recognition. The resultant performance deterioration in the presence of high-noise conditions highlights the necessity of preprocessing quality and the necessity of employing adaptive noise suppression techniques in practical applications. These results are consistent with and expand the existing body of research on multi-task audio deep learning, and they are the first to show that joint speaker-subject classification can be effectively performed with high accuracy in the context of secondary-level education. The findings advocate the greater

use of audio-based intelligent systems in institutional settings where non-invasive, high-level surveillance of instructional activity is needed.

Conclusion

This paper was able to create and test an audio-based multi-task deep learning system that can be used to identify teachers and recognize the domains of instructional subjects in a classroom setting at the secondary level. The suggested CNN-BiLSTM model demonstrated a high accuracy of 96.8, which is significantly higher than the conventional SVM and GMM classifiers, and demonstrated high accuracy in a wide range of academic domains and low-to-moderate noise levels. The study has shown that, using deep learning, joint speaker identity and subject-domain classification using classroom audio is technically feasible and practically viable. The results render a solid empirical basis to incorporate smart audio-based monitoring systems in educational organizations, which promotes the objectives of automated attendance checks, curriculum adherence, and intelligent classroom control in a non-invasive and scalable way.

Recommendations

Future studies must consider the incorporation of more sophisticated noise suppression methods, including speech enhancement based on deep neural networks, to enhance the performance of the system in classroom environments with high levels of noise. The sample must be increased to higher education institutions and more subjects, languages, and speaking styles to increase the generalizability of the model. Researchers are advised to explore real-time deployment systems and edge computing, which allow processing on the device without the use of cloud computing. Also, the integration of multimodal inputs, e.g., audio with visual cues, can further enhance the accuracy and strength of identification in various institutional settings.

References

ad, S., Kaker, M. W. K., Rafi, S., Bibi, A., & Gul, H. (2020). The empowerment of language over the meek creature (women) through the discourse in the novel “a thousand splendid suns”. *Ilkogretim Online*, *19*(4), 5909–5915.

ad, S., Khurram, S., Smerat, A., Andleeb, N., Aysha, S., Sultan, A., & Ahmad, S. (2025). An investigation into the relationship between emotional intelligence and academic success among university students. *Pegem Journal of Education and Instruction*, *15*(1), 555–571.

ed, S., Memon, N. A., Batool, Z., & Wazir, S. (2025). Assessing the Impact of Technology Integration on Teaching and Learning in Pakistani Universities. *Journal for Current Sign*, *3*(3), 658–576.

, A., & Mohanty, A. (2023). Cultural beliefs and equity in educational institutions: exploring the social and philosophical notions of ability groupings in teaching and learning of mathematics. *International Journal of Adolescence and Youth*, *28*(1), 2270662.

orenas, A. J. (2024). Machine learning framework for audio-based content evaluation using mfcc, chroma, spectral contrast, and temporal feature engineering. *arXiv preprint arXiv:2411.00195*.

Y. (2025). Audio-to-Image Encoding for Improved Voice Characteristic Detection Using Deep Convolutional Neural Networks. *arXiv preprint arXiv:2503.05929*.

Chen, S., Zhang, Y., & Yang, Q. (2024). Multi-task learning in natural language processing: An overview. *ACM Computing Surveys*, *56*(12), 1–32.

Ehtsham, M., Ahmad, S., Anjum, H. M. S., Shaker, A. S. A., Sajid, M. K. M., Lodhi, K., & Al Anazi, N. (2023). A Study Of Teachers' And Learners' Approach Regarding Code-Mixing And Code-Switching In English Language Classrooms At College Level. *Journal of Namibian Studies*, *33*.

Fatima, N., & Ahmad, S. (2025). Formulaic language in high-stake research writing: Investigating the semantic implications of collocations and fixed expressions in postgraduate dissertation. *Research Journal in Translation, Literature, Linguistics, and Education*, *1*(4), 36–47.

Fatima, N., Memon, N. A., Muhammad, M., & Ahmad, M. S. (2025). EVALUATING THE EFFECTIVENESS OF TRANSFER LEARNING IN FEW-SHOT LEARNING SCENARIOS FOR NATURAL LANGUAGE PROCESSING TASKS. *Spectrum of Engineering Sciences*, 551–563.

Johnston, W. J., & Fusi, S. (2023). Abstract representations emerge naturally in neural networks trained to perform multiple tasks. *Nature Communications*, *14*(1), 1040.

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., & Hüllermeier, E. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and individual differences*, *103*, 102274.

Kogan, M. (2022). *Education accountability: An analytic overview*. Routledge.

Kumar, D., Ahmad, S., & Lodhi, K. (2024). Exploring the Role of Digital Technology in Enhancing Learning Experiences in Pakistani Classrooms. *International Journal of Language and Literary Studies*, *8*(2), 380–393.

., Basit, A., Daraz, A., & Jan, A. (2024). Deep causal speech enhancement and recognition using efficient long-short term memory Recurrent Neural Network. *Plos one*, *19*(1), e0291240.

on, N. A., Paracha, U., & Ahmad, M. S. (2025). THE FUTURE OF HUMAN-COMPUTER INTERACTION: A STUDY OF AI-POWERED INTERFACES AND THEIR IMPACT ON USER EXPERIENCE. *Spectrum of Engineering Sciences*, 945–958.

on, N. A., Sultana, M., Siddiqui, E. A. A., & Murtaza, M. (2025). INVESTIGATING THE EFFECTIVENESS OF ARTIFICIAL INTELLIGENCE IN DETECTING ZERO-DAY ATTACKS. *Spectrum of Engineering Sciences*, 804–817.

ı, J., Viriri, S., & Adegun, A. (2022). Long short-term memory recurrent neural network for automatic speech recognition. *IEEE access*, *10*, 30069–30079.

D. N., & Giri, K. J. (2025). AUDIRE: a comprehensive review of speech recognition technologies–methods, uses, and challenges. *International Journal of Speech Technology*, 1–27.

A., Memon, S., Nizamani, M. A., Dhomeja, L. D., Memon, N., & Charan, K. (2024). Machine Learning Techniques for Cyber Security in Internet of Robotic Things. *VFAST Transactions on Software Engineering*, *12*(3), 01–10.

eider, S., Wessels, A., & Pilz, M. (2023). Theory and practice of teaching and learning in the classroom–Lessons from Indian industrial training institutes. *Vocations and learning*, *16*(1), 99–120.

ı, M. S., Latib, N. A. A., & Sidhu, K. K. (2025). MFCC in audio signal processing for voice disorder: a review. *Multimedia Tools and Applications*, *84*(10), 8015–8035.

ı, M. K., Kumar, S., & Ranjan, R. (2026). Global Trends in Speaker Identification Under Voice

Disguise: A 25-Year Review. *Sakarya University Journal of Computer and Information Sciences*, *9*(1), 243–261.

Xin, D., Ghorbani, B., Gilmer, J., Garg, A., & Firat, O. (2022). Do current multi-task optimization methods in deep learning even help? *Advances in Neural Information Processing Systems*, *35*, 13597–13609.

Yue, S., Wei, J., Aziz, H., & Liew, K. (2023). A study on the effectiveness of self-assessment learning system of ideological and political education for college students. *Learning and Motivation*, *84*, 101929.

g, M. (2025). Modeling of Speech Recognition Based on Deep Learning. *International Journal of Advance in Applied Science Research*, *4*(2), 8–15.

g, P., Swaminathan, A., & Uddin, A. A. (2023). Pulmonary disease detection and classification in patient respiratory audio files using long short-term memory neural networks. *Frontiers in Medicine*, *10*, 1269784.