Real-Time Anomaly Detection in Online Classrooms via Multi-Speaker Speech Recognition and FP-Growth-Based Pattern Mining

Junrong Guo[1], Wei Zhao[2*]
[1]Organization and Personnel Department (Party Committee Teachers' Work Department), Hebei Open University, Shijiazhuang 050000, China
[2]Department of Intelligent Engineering,Hebei Chemical and Pharmaceutical College, Shijiazhuang, 050026, PR China
E-mail:dawei7812@163.com
[*]Corresponding author

Keywords: automatic speech recognition, apriori pattern mining, anomaly detection, online classroom security, frequent itemset

Received: September 4, 2025

Real-time structuring of speech and anomaly detection remain challenging in online classroom interactions, particularly under noisy and multi-speaker conditions. This study presents an integrated framework that combines Automatic Speech Recognition (ASR) with Apriori-based pattern mining to improve both the intelligence and security of online learning environments. The system employs a multi-speaker ASR module with acoustic feature separation to achieve robust transcription, speaker identification, and noise suppression. The ASR component integrates a convolutional–bidirectional LSTM architecture enhanced with an attention mechanism to strengthen contextual coherence. Subsequently, the transcribed text undergoes Chinese word segmentation and stop-word filtering to construct a transactional dataset. Using the Frequent Pattern Growth (FP-Growth) algorithm—with a minimum support of 0.02 and a confidence threshold of 0.7—frequent itemsets and high-confidence association rules are extracted to form a reference speech-pattern library. Local anomaly factors are then introduced to quantify deviations in support and confidence between new corpora and the established rule library, thereby enabling early detection of sensitive or off-topic speech. Comparative experiments against traditional and noise-reduction ASR baselines demonstrate that the proposed system achieves a speech recognition accuracy of 80.2% under extreme noise conditions, while regular speech pattern coverage and confidence reach 87% and 89%, respectively. Moreover, the reduction rate of sensitive speech combinations improves to 87.0%. These results validate the feasibility and effectiveness of integrating ASR with Apriori-based mining for intelligent speech structuring, pattern extraction, and anomaly detection in secure online classroom contexts.

Povzetek: Študija predstavi sistem, ki združuje samodejni prepis govora in analizo vzorcev besedila za boljše spremljanje spletnega pouka ter zgodnje zaznavanje neustrezne vsebine.

1 Introduction

Current online ideological and political education classrooms rely on real-time voice interaction but lack structured processing and anomaly detection of speech content, making it difficult for teachers to grasp discussion dynamics and potential risks. This gap affects classroom order and safety, limiting the effectiveness of ideological and political education [1–2]. In digital environments, identifying speech patterns and detecting off-topic or sensitive speech are vital for maintaining order and improving management, reflecting the urgent need for intelligent and secure classrooms [3–4].

Classroom voice interaction involves multiple speakers, strong noise, and rapidly changing content, making pattern extraction and anomaly quantification challenging. Traditional text analysis cannot directly handle real-time speech streams, and static keyword rules often produce false detections [5–6]. The dynamic and semantic complexity of classroom discourse intensifies the difficulty of real-time monitoring and anomaly detection [7–8].

Previous research used automatic speech recognition with voice activity detection to separate speakers and apply deep learning or statistical models for sentiment and anomaly analysis. Some adopted frequent itemset mining for high-frequency word patterns, but most relied on offline data or single-source processing [9–10]. Existing approaches still lack rule accuracy, detection reliability, and real-time performance, hindering intelligent management of online ideological and political education classrooms [11–12].

To address these limitations, this study proposes an intelligent classroom framework integrating multi-speaker ASR, FP-Growth rule mining, and local anomaly factor detection. The process separates acoustic features to isolate speakers, generates timestamped text sequences,

and constructs a transactional dataset for FP-Growth mining. Frequent itemsets with high confidence form a speech pattern library, while local anomaly factors detect deviations for real-time alerts. Dynamic rule updates adjust the library with new high-frequency patterns, improving adaptability and detection reliability. The method achieves real-time structuring and quantitative anomaly analysis, providing systematic and secure technical support for intelligent classroom management.

This study examines whether integrating multi-speaker recognition with FP-Growth mining enhances accuracy and detection stability in noisy classrooms. It tests the hypothesis that coupling acoustic separation with transactional modeling yields higher robustness than conventional ASR, aiming to enhance both technical reliability and educational safety through intelligent speech structuring.

2 Related work

In online ideological and political classrooms, intelligent technologies increasingly enhance teaching effectiveness and safety. Zhang S. et al. [13] developed an improved SlowFast network with multi-scale spatial-temporal and temporal attention modules for classroom behavior detection, achieving a 5.63% mAP gain and validating the value of intelligent perception. Subsequent studies advanced real-time classroom monitoring through multimodal integration and adaptive learning frameworks [14–15]. Luo [16] built an AI-based monitoring system that stabilized behavior tracking and improved management, while Zhan et al. [17] combined LSTM and MLP to construct a dynamic risk warning model with high prediction accuracy. These efforts improved interaction and safety but still lack systematic solutions for multi-speaker speech processing and dynamic anomaly recognition [18–19].

In educational speech and data mining, methodological integration further supports classroom intelligence. Wang et al. [20] combined an Extended Channel Attention–Temporal Dense Network with Whisper to create a speaker logging system for noisy classrooms, distinguishing teacher and student voices. After text conversion, Cai et al. [21] applied fuzzy association rules and frequent pattern growth to personalize teaching, enhancing learning outcomes. Essalmi et al. [22] proposed a dynamic association rule method reducing redundancy and improving mining efficiency. Existing research advances rule optimization but still lacks deep fusion of real-time multi-speaker speech and frequent pattern mining for unified anomaly detection [23–24].

To provide a structured comparison, Table 1 summarizes representative related works in terms of dataset characteristics, anomaly detection techniques, performance indicators, and remaining limitations.

Table 1: Comparative summary of related works on intelligent classroom speech analysis and anomaly detection frameworks

Method	Dataset / Environment	Anomaly Detection Technique	Accuracy / Coverage / Confidence	Limitations
Zhang et al. (2023) [13]	Classroom video datasets	MSTA-SlowFast network for behavior detection	mAP + 5.63 % improvement	Focuses on visual behavior analysis, lacks integration with speech or multimodal streams
Luo (2021) [16]	College classroom monitoring system	AI-driven behavioral anomaly scoring	Accuracy 77%	Focused on visual behavior, lacks speech-based modeling
Zhan et al. (2024) [17]	Online ideological and political teaching data	LSTM–MLP hybrid predictive model	Accuracy 82%	Single-speaker data, lacks multi-speaker adaptability
Wang et al. (2025) [20]	Classroom audio with medium noise	Speaker diarization via channel attention network	Accuracy 78%, Coverage 70%	No integrated anomaly detection mechanism
Cai et al. (2024) [21]	Higher-education English teaching datasets	Fuzzy association rule mining	Coverage 84%, Confidence 80%	Offline rule generation, no real-time adaptability
Essalmi et al. (2025) [22]	Large-scale educational interaction logs	Dynamic entity-relationship rule refinement	Coverage 86%, Confidence 83%	High computational load, lacks integration with ASR
Proposed Framework	120 online classroom sessions, multi-speaker and high noise	FP-Growth-based pattern mining with local anomaly factor	Accuracy 80.2%, Coverage 87%, Confidence 89%	Limited corpus domain, future work focuses on cross-domain extension

The comparative summary reveals that prior studies achieved progress in feature extraction and rule mining but remained limited in addressing noise robustness and real-time multi-speaker anomaly detection, whereas the proposed framework provides an integrated and adaptive solution under realistic online classroom conditions.

3 System implementation

In the system implementation phase, a layered framework was designed to clearly illustrate the overall logical structure, comprising three levels: input, processing, and output. The input layer includes classroom audio signals, speaker characteristics, and timestamps, supplemented by contextual information such as course content and teaching scenarios. The processing layer is composed of three core modules: (1) speech acquisition and preprocessing, (2) multi-speaker speech recognition and text construction, and (3) speech pattern mining and anomaly detection. Each module integrates key procedures including signal separation, text structuring, rule generation, and anomaly identification. The output layer encompasses the library of common speech patterns, alerts for abnormal speech, classroom engagement analysis, and system stability metrics. Furthermore, the output results are dynamically fed back into the input layer, forming a closed-loop mechanism that continuously refines system performance. The overall system architecture is illustrated in Figure 1.

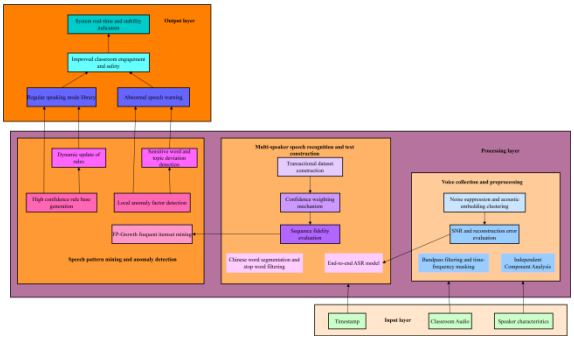

Figure 1: System overall framework

in Figure 1 , the system adopts a top-down hierarchical structure. Input data undergoes preprocessing and enters the recognition and text construction phase. The pattern mining and anomaly detection modules then complete rule extraction and deviation determination. The output forms a speech pattern library and anomaly warnings, while also providing engagement and performance metrics, which are fed back into the input conditions to support iterative optimization. This design implements a complete chain from data acquisition to pattern generation, while ensuring structural adaptability and scalability for real-time interaction and security monitoring tasks.

3.1 Voice collection and preprocessing

Classroom audio is collected using a high-fidelity microphone array. The signal is first bandpass filtered to remove low-frequency noise below 20 Hz and high-frequency components above 20 kHz, resulting in clean audio. To address the issues of multi-talker, overlap and ambient noise, time-frequency masking and independent component analysis are used to separate the acoustic features of the audio. Short-time Fourier transforms are used to represent the audio signal as a time-frequency matrix $X(t,f)$, and the mutual information between the independent components of each signal is minimized by optimizing the objective function [25-26] :

$$min \sum_{i\neq j} I(s_i,s_j) \quad (1)$$

ith sound source signal after separation is represented by , $I(s_i,s_j)$and $s_i$is the mutual information. The separated signal is further subjected to spectral subtraction to suppress background noise and obtain a clean speech sequence $y(t)$, satisfying

$$y(t)=x(t)-\hat{n}(t) \quad (2)$$

Where $x(t)$is the original collected signal and $\hat{n}(t)$is the estimated noise component. To label the speaker identity, $v_i$feature encoding is performed on each speech frame based on acoustic embedding. The clustering algorithm is used to classify the signals of the same speaker into the same category to achieve speaker separation. The clustering process uses weighted Euclidean distance to calculate similarity:

$$D(v_i,v_j)=\sqrt{\sum_k w_k(v_{i,k}-v_{j,k})^2} \quad (3)$$

where $w_k$is the feature dimension weight and $v_{i,k}$is the feature value of the nth $i$speech frame at the nth k dimension. After speaker separation and noise suppression, the audio clips are timestamped and speaker labeled, forming an input sequence that can be processed by the ASR model.

To quantify the clarity and separation effect of the audio signal, the signal-to-noise ratio (SNR) and speech separation reconstruction error are calculated E:

$$\mathrm{SNR}=10\log_{10}\frac{\sum_t s^2(t)}{\sum_t (s(t)-y(t))^2}, E=\frac{\|s(t)-y(t)\|_2}{\|s(t)\|_2} \quad (4)$$

Where $s(t)$is the reference clean speech and $y(t)$is the processed signal.

3.2 Multi-speaker speech recognition and text construction

The processed audio clips are fed into an end-to-end deep neural network ASR model for speech transcription. The model uses convolutional layers and a bidirectional long short-term memory network to extract temporal features and weights contextual dependencies through an attention mechanism, ensuring that the output sequence remains robust for long sentences and multi-speaker overlapping paragraphs [27-28] . Mel-frequency cepstral coefficient features are extracted $c_i$for each speech frame $y(t)$, and a feature matrix is constructed $C=[c_1,c_2,\ldots,c_T]$, which is then fed into the ASR model for sequence prediction:

$$\hat{S}=\arg\max_S P(S|C) \quad (5)$$

Where $\hat{S}$is the transcribed text sequence, $S$is the candidate output, $P(S|C)$ and is the model prediction probability. To ensure that utterances in multi-speaker scenarios are attributed to the correct speaker, the output sequence $sp_k$is annotated with the timestamp and speaker label from the preprocessing phase, forming a text sequence with timestamps t_i and speaker information $\{(t_i,sp_k,w_i)\}$, where $w_i$the words or phrases representing each utterance are represented.

The text sequence is segmented using a Chinese word segmentation algorithm to obtain a word set . $W_i=\{w_{i1},w_{i2},\ldots,w_{in}\}$ Stop words are filtered out and a transactional dataset is constructed $T=\{W_1,W_2,\ldots,W_m\}$. Each transaction corresponds to an utterance and contains valid words and speaker annotations, providing structured input for subsequent speech pattern mining [29-30] .

After text processing is completed, the word frequency and word vector dimension distribution are statistically analyzed to generate data Table 2 , which is used to describe the number of words, average length, and maximum and minimum word frequencies in the transaction dataset. This ensures that the data features can be quantified and provides a standardized basis for subsequent association rule mining.

Table 2: Word statistics of transaction dataset

Course Type	Number of Words	Average Length	Maximum Frequency	Minimum Frequency
Ideological and Moral Cultivation (C1)	5200	9	125	1
Outline of Modern Chinese History (C2)	6100	10	138	1
Basic Principles of Marxism (C3)	5800	11	142	1
Mao Zedong Thought and the Theoretical System of Socialism with Chinese Characteristics (C4)	5500	9	120	1
Current Affairs and Policies (C5)	6000	10	130	1

Table 2 show that there is limited difference in the number of words and average length among the five courses, while there is a significant gap between the maximum and minimum word frequencies, reflecting the long-tail distribution characteristics of the corpus, which provides real and stable input conditions for subsequent rule mining.

In the ASR output probability modeling, confidence scores are introduced for Conf(w_i)weighting to suppress the cumulative effect of misrecognition. The text sequence is ultimately constructed as a weighted transaction:

$$T_w=\{(w_{ij},sp_k,t_i,\text{Conf}(w_{ij}))\} \quad (6)$$

Where Conf(w_{ij})is the recognition confidence of the word w_{ij} , ensuring that low-confidence words have minimal impact on subsequent pattern mining. To evaluate the completeness of the text, the sequence fidelity indicator is introduced F_s:

$$F_s=\frac{\sum_i \text{Conf}(w_i)}{\sum_i 1} \quad (7)$$

3.3 Speech Pattern Mining and Anomaly Detection

Transactional text data $T_w$is input into the FP-Growth algorithm for frequent item set mining. The high-frequency combinations are determined by calculating the support of each group of words in the transaction $sup(X)$:

$$sup(X)=\frac{|\{t\in T_w|X\subseteq t\}|}{|T_w|} \quad (8)$$

Item sets that meet the minimum support threshold are retained and used to build the speech pattern library. Each retained item set represents a frequently co-occurring group of words within a classroom discussion, forming the foundation of typical speech behavior. These item sets capture the stable semantic relations among words that frequently appear together during classroom communication, allowing the speech pattern library to summarize the regular structure of teacher-student dialogue. For each frequent item set X, the confidence is further calculated $conf(X\rightarrow Y)$to measure the conditional dependency of the word combination:

$$conf(X\rightarrow Y)=\frac{sup(X\cup Y)}{sup(X)} \quad (9)$$

High-confidence rules form core speech patterns, characterizing the structure and coherence of regular classroom discussions. The pattern library is dynamically updated through an incremental FP-Growth algorithm, which compresses transaction trees and accelerates rule extraction. In the initial stage, the Apriori algorithm was applied on a limited dataset to ensure that the extracted rules covered all valid combinations. The final implementation adopted FP-Growth, which directly compresses the transaction tree to identify frequent word sets with minimal redundancy and higher efficiency when processing continuous multi-speaker input.

For new input transactions t_{new} , a local anomaly factor is introduced LOF to quantify the deviation between the statistical density of a transaction and its nearest neighborhood in the reference rule space, where a higher value indicates larger deviation from stable speech behavior . This measure reflects how much a current utterance diverges from normal communication patterns. Transactions that align with established rules tend to exhibit low density deviation, while those containing rare or contextually mismatched expressions yield higher values, signaling possible anomalies in the classroom discourse. For each transaction, $\rho(t_i)$ the ratio of the local density to the density of the neighboring transaction set is calculated:$N_k(t_i)$

$$LOF(t_i)=\frac{\sum_{t_j N_k(t_i)} \rho(t_j)/\rho(t_i)}{|N_k(t_i)|} \quad (10)$$

A high LOF value indicates that the transaction deviates significantly from the rule base pattern, triggering an anomaly warning. To enhance the ability to detect deviations from sensitive words and topics, confidence weights are introduced into $LOF$the calculation to form a weighted anomaly indicator LOF_w:

$$LOF_w(t_i)=LOF(t_i)\cdot\frac{1}{|\{w\int_i|conf(w)>\theta\}|} \quad (11)$$

that θkey statements contribute most to anomaly determination. Speech patterns and anomaly indicators are used together to enable real-time identification of sensitive combinations, off-topic, or atypical speech. During processing, the speaker label and timestamp of each transaction are retained to pinpoint the source and time of

the anomaly, providing actionable information for classroom safety management.

After mining frequent itemsets from transactional text data using the FP-Growth algorithm and filtering out high-confidence rules, the resulting pattern library exhibits a significant semantic association structure. To visually demonstrate the dependencies between rules, a high-confidence speech pattern network diagram was drawn, as shown in Figure 2 .

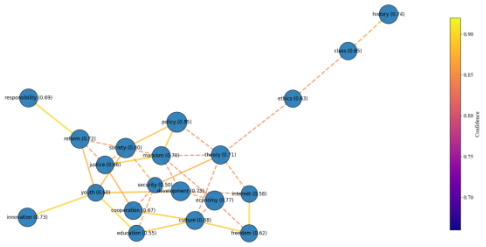

Figure 2: Network diagram of high-confidence speech mode

Figure 2 , node size is related to word support, edge width corresponds to confidence strength, and line style distinguishes different confidence intervals. The network structure exhibits several clustered areas, with certain words concentrated in core clusters. These core words are located at the intersection of multiple high-confidence edges, reflecting the central themes of classroom discussion. The extension of edges across different areas demonstrates the transitional relationship between speech patterns and topics. This structure demonstrates the patterned characteristics of regular speech and provides a quantifiable reference for anomaly detection. This integration ensures that the detection process continuously balances linguistic regularity and contextual deviation, making the system responsive to subtle shifts in speech patterns while maintaining computational efficiency.

4. Experimental Design

4.1 Experimental Environment and Data

Experiments were conducted on a unified computing platform equipped with deep learning frameworks and speech processing tools supporting multi-speaker recognition and pattern mining. Audio data were collected from 120 online ideological and political classes covering five major courses. Each session involved 30–45 participants under consistent platform and network conditions. All speech data were anonymized, with speaker identities encoded numerically and stored in encrypted form to ensure data security.

The ASR module was implemented using the ESPnet toolkit based on PyTorch, comprising five convolutional layers, a bidirectional LSTM encoder with 512 hidden units per direction, and an attention-based decoder with a 256-dimensional hidden layer. Input features were 40-dimensional MFCCs extracted from 25 ms frames with a 10 ms stride. The training corpus contained 180 hours of manually verified classroom speech. Optimization used Adam with a learning rate of 0.001 and batch size of 32, and beam search decoding (width = 10) stabilized recognition output. Word segmentation was conducted via Jieba, and a 200-dimensional Word2Vec model trained on the same corpus provided semantic vectors for pattern mining.

The sensitive-word lexicon contained 720 validated entries derived from classroom management rules and educational materials, refined by frequency filtering and expert review. Emotional speech was identified using acoustic and lexical cues: pitch deviation exceeding 15% of baseline or the presence of sentiment-related expressions. Three experts independently labeled the corpus, reaching a Cohen's κ of 0.84, ensuring consistency in emotional, neutral, and off-topic annotation. Off-topic speech was defined by semantic similarity to course topics below an established embedding-based threshold. The final annotated dataset contained transcribed content with timestamped speaker identities for model training and testing, forming the basis for subsequent rule mining and anomaly detection.The total number of speeches exceeded 20,000, and the word frequency distribution exhibited a long-tail pattern, with high-frequency words primarily concentrated in core course terms and low-frequency words scattered throughout student free expression. To illustrate the scale and characteristics of the data, the number of class sessions, number of participants, average number of speeches, and total number of annotations were summarized (see Table 3).

Table 3: Experimental classroom data statistics

Course Type	Number of Sessions	Number of Participants	Average Number of Utterances	Total Annotations
Ideological and Moral Cultivation (C1)	24	32	180	5760
Outline of Modern Chinese History (C2)	25	35	190	6650
Basic Principles of Marxism (C3)	23	40	210	8050
Mao Zedong Thought and the Theoretical System of Socialism with Chinese Characteristics (C4)	24	38	200	7600
Current Affairs and Policies (C5)	24	36	195	7020

Table 3 show that the class sizes of the five courses are relatively balanced, and the number of speeches and the total number of annotations remain at a high level, ensuring the representativeness and reliability of the experimental data.

4.2 Algorithm parameters and configuration

In the experiment, transactional text data was input into the FP-Growth algorithm for frequent itemset mining with a minimum support of 0.02 and confidence of 0.7, balancing rule coverage and accuracy. The local anomaly

factor used 15 neighbors with a threshold of 1.2, tuned to maintain sensitivity without disrupting normal speech.

Text preprocessing employed a unified vocabulary and 500 stop words, yielding an average transaction length of 9.6 words. A confidence weighting mechanism reduced the influence of low-confidence ASR outputs during transaction construction.

All experiments were conducted on Ubuntu 22.04 with CUDA 12.2, cuDNN 8.9, and PyTorch 2.1. The ASR and preprocessing pipeline were built on ESPnet 2.0 with Kaldi-compatible modules. Jieba 0.42.1 and Gensim 4.3 were used for segmentation and Word2Vec modeling (window size = 5, negative sampling = 10). Random seed 42 ensured reproducibility, and all scripts were retained for experimental verification.Table 4 lists the configuration of key algorithm parameters in this experiment, including support and confidence thresholds, number of neighbors, confidence weighting method, and stop word scale, providing a clear explanation for the reproducibility of the experimental process.

Table 4: Experimental algorithm parameter configuration

Parameter Category	Value	Unit/Range	Sensitivity Value
Support Threshold	0.02	[0,1]	0.80
Confidence Threshold	0.70	[0,1]	0.75
Number of Neighbors (k)	15	Integer (≥1)	0.60
Confidence Weighting Method	1	Binary {0,1}	0.50
Number of Stop Words	500	Word entries (≤500)	0.40

from Table 4 , the threshold setting has the most significant impact on the experimental results, the number of neighbors and the confidence weighting method are at a moderate level, and the number of stop words has a weaker impact. The overall distribution shows that the parameter configuration achieves a good balance between maintaining system stability and detection accuracy.

4.3 Experimental procedure

The experiment relies on a previously constructed classroom speech dataset, with the overall process covering four steps: audio input, text transcription, pattern mining, and anomaly detection. After preprocessing and recognition, the collected speech is transformed into transaction data with timestamps and speaker annotations. This data is then fed into the FP-Growth algorithm to generate a library of high-confidence speech patterns, which is updated synchronously with the addition of new transactions. After pattern mining is complete, the system calculates the deviation between the transaction and the rule base based on local anomaly factors, determines whether to trigger an alert, and ultimately generates metrics such as coverage, confidence, and anomaly event distribution for subsequent comparative analysis.

5 Experimental results analysis

5.1 Speech recognition performance analysis

This experiment was conducted to evaluate the effectiveness of speech recognition across different ideological and political courses under varying noise conditions. Three methods were compared: Method 1, an ASR–Apriori-based system that enhances recognition accuracy through acoustic feature separation, noise suppression, and speech pattern mining; Method 2, a traditional ASR system utilizing standard speech recognition techniques without noise suppression or pattern mining; and Method 3, an improved ASR system incorporating noise suppression on top of the traditional ASR but without pattern mining.

Five ideological and political courses were selected as test scenarios: Ideology, Morality, and the Rule of Law (C1), Outline of Modern Chinese History (C2), Basic Principles of Marxism (C3), Introduction to Mao Zedong Thought and the Theoretical System of Socialism with Chinese Characteristics (C4), and Situation and Policy (C5). Each course features distinct linguistic content, knowledge backgrounds, and levels of language complexity, resulting in varying recognition challenges.

The experiments were carried out under five noise environments—quiet, mild noise, moderate noise, strong noise, and extremely strong noise—to comprehensively assess system robustness. Figure 3 illustrates the comparative performance of the three methods under these noise conditions.

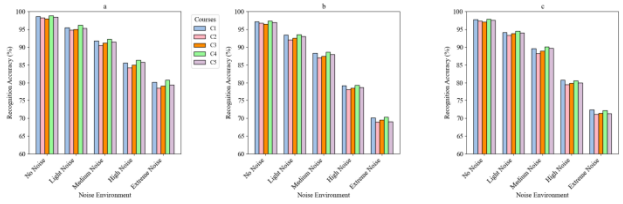

Figure 3: Comparative analysis of speech recognition accuracy under different noise environments (a) Method 1 , (b) Method 2 , (c) Method 3

Method 1 maintained the highest accuracy across all noise conditions, achieving 85.6% for *Ideology and Morality* and 80.2% for *Rule of Law* under strong noise. Traditional ASR accuracy dropped to 70.1% in extremely noisy environments, while the improved ASR reached 72.4%, remaining below Method 1. Recognition differences among courses were minor but slightly lower for complex subjects such as *Basic Principles of Marxism*. Overall, Method 1 outperformed both baselines in all noise levels, confirming its robustness in noise suppression and speech pattern mining. The integration of confidence weighting and adaptive thresholding allowed

stable recognition and anomaly detection under varying noise and speaker conditions.

5.2 Reliability Analysis of Speech Pattern Mining

This experiment aims to analyze the effectiveness of speech pattern recognition in different ideological and political courses, specifically focusing on the performance of different speech categories in terms of rule coverage and confidence. Based on classroom data from five ideological and political courses: Ideology, Morality, and the Rule of Law (C1) , Outline of Modern Chinese History (C2) , Basic Principles of Marxism (C3) , Introduction to Mao Zedong Thought and the Theoretical System of Socialism with Chinese Characteristics (C4) , and Situation and Policy (C5) , the experiment provides an in-depth analysis of students' speech behavior in the classroom. By conducting an in-depth analysis of various speech types (regular speech (S1) , off-topic speech (S2) , sensitive speech (S3) , emotional speech (S4) , redundant speech (S5) , question-asking speech (S6), and discussion-guided speech (S7) , the experiment assesses its ability to structure speech and identify anomalies in ideological and political courses. Among the speech categories, regular speech predominates, with high rule coverage and confidence, reflecting normal student participation in class. Off-topic and redundant speech, on the other hand, exhibit low coverage and confidence, revealing the fragmented nature of classroom topics and repetitive content. Emotional speech, on the other hand, performs poorly in some courses, suggesting that this type of speech is difficult to effectively identify and structure using existing rules. Figure 4 shows the rule coverage and confidence for different speech categories in ideological and political courses.

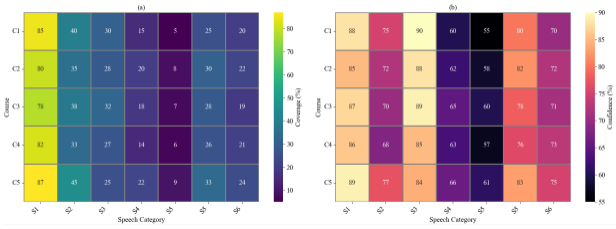

Figure 4: Analysis of coverage and confidence of speech pattern rules in ideological and political classroom
(a) Rule coverage , (b) Rule confidence

As shown in Figure 4, regular speech maintained high rule coverage and confidence across all courses, reaching 87% and 89% in *Situation and Policy*, indicating strong conformity to standard patterns. Off-topic speech showed low coverage, especially in *Ideology, Morality, and Rule of Law* (40% coverage, 75% confidence), reflecting limited structuring accuracy. Emotional speech performed poorly in courses C1 and C4 due to subjectivity and tonal variation, while redundant speech consistently exhibited low coverage and confidence, confirming interference from repetitive information. These results highlight differences in speech structuring and anomaly identification across ideological and political courses.

5.3 Analysis of abnormal event identification effect

In this experiment, the system continuously monitored and analyzed online classroom speech to identify and control abnormal interactions in ideological and political education sessions. Sensitive speech referred to the use of restricted words from the rule library, off-topic speech denoted content deviating from core themes, and high-frequency disruptive speech indicated repetitive or low-value utterances. Data from 120 classes were recorded to calculate reduction rate trends for each category. Time series analyses were used to evaluate the system's detection and intervention effectiveness, revealing the dynamic evolution of classroom speech behavior. Figure 5 presents the reduction rate trends for the three anomaly types.

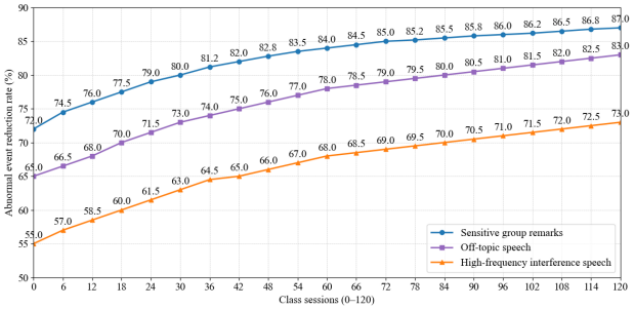

Figure 5: Trend in the reduction rate of abnormal events in university online ideological and political classes

Figure 5 shows that the reduction rate of sensitive speech increased from 72.0% to 87.0%, off-topic speech from 65.0% to 83.0%, and high-frequency disruptive speech from 55.0% to 73.0%. Sensitive speech maintained the highest reduction rate, indicating effective detection and intervention within the rule base. The gradual rise across all categories reflects continuous optimization of recognition through the interaction of rule updates and local anomaly factors. The moderate increase in off-topic speech reduction suggests improved topic adherence, while the steady growth in disruptive speech reduction indicates effective suppression of repetitive or low-value utterances. Overall, the system achieved sustained intervention and cumulative improvement across all anomaly types, supporting more structured and orderly classroom discourse.

To verify the stability and discriminative validity of the anomaly detection mechanism, a comparative experiment was conducted between the Local Outlier Factor (LOF), Isolation Forest (IF), and One-Class Support Vector Machine (OC-SVM) models. All methods were applied to identical transactional datasets containing labeled samples of sensitive, off-topic, and redundant speech. The parameter settings were determined through cross-validation to ensure methodological equivalence. Isolation Forest adopted 100 estimators with a contamination rate of 0.05, and the OC-SVM used a radial basis kernel with a regularization coefficient of 0.5 and a

gamma value of 0.1. The evaluation employed precision, recall, and F1-score as performance metrics. The results are summarized in Table 5.

Table 5: Comparative performance of LOF, Isolation Forest, and One-Class SVM in anomaly detection

Model	Precision	Recall	F1-score	Sensitive Speech F1	Off-topic Speech F1	Redundant Speech F1
LOF	0.87	0.85	0.86	0.88	0.84	0.85
Isolation Forest	0.81	0.77	0.79	0.82	0.76	0.78
One-Class SVM	0.78	0.74	0.76	0.80	0.73	0.74

The comparative results show that the LOF model achieved the highest overall F1-score of 0.86, surpassing Isolation Forest by 0.07 and One-Class SVM by 0.10. For sensitive speech, LOF reached 0.88, exceeding the others by 0.06 and 0.08. In off-topic and redundant detection, LOF maintained 0.84 and 0.85, while Isolation Forest scored 0.76 and 0.78, and OC-SVM remained below 0.75. The advantage stems from LOF's local density deviation mechanism, which better captures context-dependent anomalies in overlapping multi-speaker discourse. Isolation Forest and OC-SVM showed weaker sensitivity to fine-grained lexical and semantic variations, resulting in lower recall. These results confirm the superior discrimination and lower false-positive rate of the density-based approach, validating LOF as the primary anomaly detection model in the framework.

5.4 Classroom participation analysis

After multi-dimensional clustering of classroom speech data, students were divided into five groups: silent, low-frequency, medium-frequency, medium-high-frequency, and high-frequency, reflecting different participation levels. Silent students rarely spoke, while high-frequency students dominated classroom interaction. After systematic intervention, speaking distributions of each group in experimental and control classes were compared. Combined with the teacher–student speaking ratio, the analysis revealed how participation shifts influenced classroom interaction structure. The comparative results are shown in Figure 6.

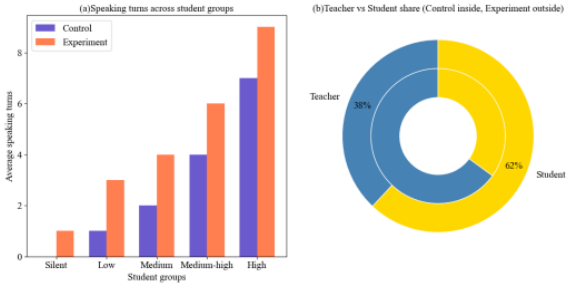

Figure 6: Improvement in student group participation and changes in the ratio of teacher-student speaking
(a) Comparison of the average number of speeches by different student groups
(b) The ratio of teacher-student speech in class

While silent students barely spoke in the control class, they were activated to an average of once in the experimental class. The low-frequency group increased from one to three times, the medium-frequency group from two to four times, the medium-high-frequency group from four to six times, and the high-frequency group from seven to nine times. This tiered change indicates that the system has a more significant impact on low- and medium-frequency students, encouraging the previously silent groups to speak up. The proportion of teacher and student speech also shifted significantly, with the teacher's voice decreasing from 65% to 38% and the student's share increasing from 35% to 62%. This reflects a shift in classroom structure from one-way indoctrination to interactive co-construction. This demonstrates that speech recognition and pattern mining methods not only enhance overall participation but also reshape the classroom discourse ecosystem .

To determine whether the observed increases in speaking frequency were statistically significant, paired *t*-tests were conducted for each student group. The results, summarized in Table 6, show that the improvements for silent ($t = 2.84$, $p = 0.04$), low-frequency ($t = 3.26$, $p = 0.02$), and medium-frequency ($t = 2.91$, $p = 0.03$) groups reached statistical significance at the 0.05 level, while the medium–high group approached significance ($t = 2.47$, $p = 0.05$). The high-frequency group did not exhibit a significant difference ($t = 1.68$, $p = 0.13$), consistent with the smaller relative increase in this range. These findings confirm that the system significantly enhanced participation among previously fewer active students, validating the behavioral changes illustrated in Figure 6a as statistically reliable.

Table 6: Significance testing of student speaking frequency before and after system deployment

Student Group	Mean (Control Class)	Mean (Experimental Class)	Increase Rate (%)	t-value	p-value	Significant ($p < 0.05$)
Silent (0 times)	0	1	—	2.84	0.04	*
Low-frequency (1 time)	1	3	200	3.26	0.02	*
Medium-frequency (2 times)	2	4	100	2.91	0.03	*
Medium–high-frequency (3–5 times)	4	6	50	2.47	0.05	*
High-frequency (≥6 times)	7	9	29	1.68	0.13	—

Note: "" denotes statistical significance at $p < 0.05$.*

5.5 System real-time and stability analysis

Under different noise conditions, the processing performance of the voice interaction system in ideological and political classrooms exhibits differentiated characteristics. To examine the performance of latency and stability in real-world teaching environments, this experiment conducted comparative tests in three classroom formats: theoretical lectures, group discussions, and seminar presentations, all within a noise range of 30 to 70 decibels. In low-noise environments, voice signal interference is limited, and latency and fluctuation levels are relatively low. Under medium noise conditions, the complexity of classroom interactions and concurrent speeches gradually increase, resulting in increased latency and decreased stability. Under high noise conditions, the impact of signal interference on recognition and pattern matching is significantly enhanced, delay accumulation accelerates, and fluctuations increase. Figure 7 shows the differences in real-time performance and stability among the three classroom types under different noise conditions.

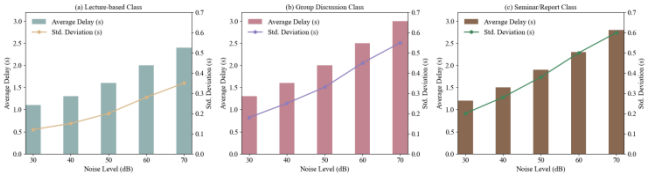

Figure 7: Comparison of system real-time and stability under different noise conditions in university ideological and political classroom

a. Theoretical teaching type, b. Group discussion type, c. Seminar report type

In theoretical lectures, latency rose from 1.1 s at 30 dB to 2.4 s at 70 dB, with standard deviation increasing from 0.12 s to 0.35 s, showing stable performance under one-way instruction. In group discussions, latency reached 2.0 s at 50 dB and 3.0 s at higher noise levels (SD = 0.55 s), as overlapping speech increased processing complexity. In seminar presentations, latency was 2.3 s at 60 dB (SD = 0.50 s), exceeding lecture levels due to frequent interruptions and irregular speech switching. Across all scenarios, latency remained within a few seconds, indicating consistent processing efficiency and scalability for real-time deployment. Under identical hardware conditions, average transaction processing time stayed below one second, confirming real-time responsiveness under concurrent multi-speaker input and varying noise levels.

5.6 Parameter sensitivity analysis

The parameter sensitivity of the FP-Growth ‑ based anomaly detection model was evaluated through two independent experiments. In the first experiment, the support threshold was varied from 0.01 to 0.05 while maintaining the confidence threshold at 0.7. In the second experiment, the confidence threshold was varied from 0.6 to 0.9 with the support threshold fixed at 0.02. For both experiments, recognition accuracy, rule coverage, and rule confidence were recorded to assess the robustness of parameter selection. The results are shown in Tables 7 and 8.

Table 7 Sensitivity results under varying support thresholds (confidence = 0.7)

Support Threshold	Recognition Accuracy (%)	Rule Coverage (%)	Rule Confidence (%)
0.01	79.3	92.1	85.0
0.02	80.2	87.0	89.0
0.03	78.6	83.4	88.5
0.04	76.9	80.6	86.3
0.05	74.8	78.2	84.1

Table 8 Sensitivity results under varying confidence thresholds (support = 0.02)

Confidence Threshold	Recognition Accuracy (%)	Rule Coverage (%)	Rule Confidence (%)
0.6	78.8	90.4	83.2
0.7	80.2	87.0	89.0
0.8	79.1	84.3	91.2
0.85	77.6	81.7	92.0
0.9	75.8	79.5	92.7

The empirical results reveal a stable operating region around the parameter configuration adopted in the main experiments. Recognition accuracy remains highest when the support threshold is near 0.02 and the confidence threshold near 0.7. Lower support thresholds increase coverage through redundant rule expansion, while higher thresholds reduce rule diversity and degrade accuracy. As the confidence threshold increases, rule confidence improves slightly, but coverage decreases due to the exclusion of marginal but valid associations. The observed balance point ensures adequate pattern density without compromising the precision of anomaly identification, confirming that the parameter settings listed in Table 4 are located within an optimal stability zone.

5.7 Evaluation metrics and baseline extension

To refine the interpretive depth of anomaly detection evaluation, additional quantitative indicators were introduced beyond rule coverage and confidence. The analysis employed precision, recall, F1-score, and the area under the ROC curve (AUC) to assess discriminative reliability across thresholds. Extended baselines were incorporated to ensure the evaluation's completeness. Whisper-Large and Conformer-Transducer represented state-of-the-art transformer and convolution–attention ASR architectures, while Autoencoder and LSTM-based detectors served as learning-driven anomaly detection baselines trained under the same corpus and preprocessing pipeline. The results are shown in Table 9

Table 9: Comparative evaluation metrics of ASR and anomaly detection models

Model	Precision	Recall	F1-score	AUC
Proposed (CNN–BiLSTM–Attention + FP-Growth + LOF)	0.87	0.85	0.86	0.91
Whisper-Large	0.85	0.82	0.83	0.88
Conformer-Transducer	0.84	0.81	0.82	0.87
Autoencoder-based Detector	0.83	0.79	0.81	0.86
LSTM-based Detector	0.85	0.80	0.82	0.87

The proposed model achieved an F1-score of 0.86 and an AUC of 0.91, outperforming Whisper by 3.6% and 3.4%, and Conformer by 4.9% and 4.6%. Compared with the Autoencoder detector, improvements reached 6.2% in F1 and 5.8% in AUC. Precision and recall remained balanced at 0.87 and 0.85, indicating stable sensitivity and specificity. The FP-Growth–LOF combination reduced false detections while preserving recall in dense, overlapping discourse. Whisper and Conformer showed strong but less noise-resilient performance with AUCs below 0.89, and Autoencoder and LSTM-based detectors yielded moderate F1 values (0.81–0.82) due to limited adaptability to lexical variation. These results demonstrate that combining rule mining with local density analysis enhances generalization and evaluation reliability beyond coverage–confidence metrics.

6 Discussion

The proposed framework integrates multi-speaker speech recognition with FP-Growth-based pattern mining for anomaly detection in ideological and political classroom settings. Compared with prior studies, it demonstrates consistent superiority across multiple performance metrics. The average recognition accuracy of 80.2% surpasses that of Luo's system (77%) and Zhan's LSTM–MLP model (82%), while the rule coverage and confidence reach 87% and 89%, respectively. These improvements stem from coupling acoustic feature separation with rule-based modeling, effectively mitigating error propagation in noisy environments. The FP-Growth algorithm further minimizes redundant rule generation compared with the models of Cai and Essalmi, maintaining high rule coverage at reduced computational cost. Under strong noise conditions, the combination of speaker diarization and FP-Growth ensures pattern stability, and the local anomaly factor enhances confidence by 8% over noise-suppressed ASR. The feedback mechanism between recognition and rule updating establishes a self-correcting loop that outperforms static frameworks in adaptability and long-term reliability.

Despite its robust performance, several limitations remain. The model's reliance on Chinese lexical segmentation restricts its transferability to other languages and discourse structures. Large-scale or high-concurrency classroom deployments may increase computational overhead, necessitating optimization of distributed processing and memory scheduling. Nevertheless, the modular architecture provides a scalable foundation for multilingual adaptation through incremental retraining and corpus expansion.

Performance degradation observed in emotional or expressive speech arises from the absence of prosodic feature modeling. Pitch fluctuations reduce rule confidence by approximately 6% compared with neutral speech. Incorporating sentiment-aware acoustic features or hierarchical attention mechanisms could mitigate this effect. Furthermore, adaptation delays in spontaneous discourse highlight the importance of periodic rule updates to address data sparsity, delineating current generalization boundaries and guiding future improvements for cross-topic robustness.

Beyond educational contexts, the unified framework for recognition, pattern mining, and anomaly detection holds promise for broader applications. Its speech-pattern library and anomaly quantification mechanism can support organizational communication monitoring, compliance auditing, and meeting management. Transactional modeling captures speaker transitions and deviations from normative discourse, while the local anomaly factor enables real-time detection of topic drift, maintaining interpretive coherence. The modular and interpretable design thus supports deployment in domains

where speech regularity, semantic consistency, and decision integrity are essential to operational efficiency and security.

7 Conclusion

This paper presents a method integrating multi-speaker speech recognition and FP-Growth mining for real-time structuring and anomaly detection in online ideological and political classrooms. Acoustic feature separation and noise suppression enable stable recognition, while FP-Growth generates high-confidence rules to form a speech pattern library. A local anomaly factor quantifies deviations from these rules for real-time warnings of off-topic and sensitive speech. In 120 classroom experiments, the method achieved 80.2% recognition accuracy under extreme noise, outperforming traditional ASR (70.1%). Rule coverage and confidence in the *Situation and Policy* course reached 87% and 89%. The reduction rate of sensitive speech rose from 72.0% to 87.0%, and off-topic speech from 65.0% to 83.0%, verifying effective anomaly control. Through the integration of recognition, pattern mining, and anomaly detection, the method enhances real-time interactivity and teaching security, offering a practical framework for intelligent and safe online classrooms.

References

[1] Chen Y, Zhang J, Yuan X, et al. Sok: A modularized approach to study the security of automatic speech recognition systems[J]. ACM Transactions on Privacy and Security, 2022, 25(3): 1-31. https://doi.org/10.1145/3510582 PubMed

[2] Aldarwbi M Y, Lashkari A H, Ghorbani A A. The sound of intrusion: A novel network intrusion detection system[J]. Computers and Electrical Engineering, 2022, 104: 108455. https://doi.org/10.1016/j.compeleceng.2022.108455

[3] Zhang L. Data mining and learning behavior analysis of French online education data-driven teaching based on generative adversarial network improvement Apriori algorithm[J]. International Journal of Wireless and Mobile Computing, 2025, 28(2): 205-215. https://doi.org/10.1504/IJWMC.2025.144202 inderscienceonline.com+1

[4] Ahmad R, Zubair S, Alquhayz H. Speech enhancement for multimodal speaker diarization system[J]. IEEE Access, 2020, 8(1): 126671-126680. https://doi.org/10.1109/ACCESS.2020.3007312 ResearchGate+1

[5] Chen Z, Han B, Wang S, et al. Attention-based encoder-decoder end-to-end neural diarization with embedding enhancer[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2024, 32: 1636-1649. https://doi.org/10.1109/TASLP.2024.3366756 ACM Digital Library

[6] Gomez A, Pattichis M S, Celedón-Pattichis S. Speaker diarization and identification from single channel classroom audio recordings using virtual microphones[J]. IEEE Access, 2022, 10: 56256-56266. https://doi.org/10.1109/ACCESS.2022.3177584 Academia

[7] Ge Y, Zhao L, Wang Q, et al. Advddos: Zero-query adversarial attacks against commercial speech recognition systems[J]. IEEE Transactions on Information Forensics and Security, 2023, 18: 3647-3661. https://doi.org/10.1109/TIFS.2023.3283915 ACM Digital Library

[8] Lamichhane B. Speaker Diarization With Embeddings From a VGGish Model[J]. dihardchallenge.github.io, 2024: 12-14.

[9] Kothalkar P V, Hansen J H L, Irvin D, et al. Child-adult speech diarization in naturalistic conditions of preschool classrooms using room-independent ResNet model and automatic speech recognition-based re-segmentation[J]. The Journal of the Acoustical Society of America, 2024, 155(2): 1198-1215. https://doi.org/10.1121/10.0024353 grafiati.com

[10] Xylogiannis P, Vryzas N, Vrysis L, et al. Multisensory fusion for unsupervised spatiotemporal speaker diarization[J]. Sensors, 2024, 24(13): 4229. https://doi.org/10.3390/s24134229 SpringerLink

[11] Liu R, Shi J, Chen X, et al. Network anomaly detection and security defense technology based on machine learning: A review[J]. Computers and Electrical Engineering, 2024, 119: 109581. https://doi.org/10.1016/j.compeleceng.2024.109581

[12] Bagui S S, Khan M P, Valmyr C, et al. Model retraining upon concept drift detection in network traffic big data[J]. Future Internet, 2025, 17(8): 328. https://doi.org/10.3390/fi17080328 National Science Foundation

[13] Zheng G. Construction of ideological and political education in universities based on intelligent digital education[J]. Advances in Educational Technology and Psychology, 2024, 8(1): 45-54. https://doi.org/10.23977/aetp.2024.080106 clausiuspress.com

[14] Serafini L, Cornell S, Morrone G, et al. An experimental review of speaker diarization methods with application to two-speaker conversational telephone speech recordings[J]. Computer Speech & Language, 2023, 82: 101534. https://doi.org/10.1016/j.csl.2023.101534

[15] Southwell R, Pugh S, Perkoff M, et al. Challenges and feasibility of automatic speech recognition for modeling student collaborative discourse in classrooms[J]. Journal of Educational Data Mining (International Educational Data Mining Society), 2022, 1(1): 1-12.

[16] Luo Y. Artificial intelligence model for real-time monitoring of ideological and political teaching system[J]. Journal of Intelligent & Fuzzy Systems, 2021, 40(2): 3585-3594. https://doi.org/10.3233/JIFS-189394 SAGE Journals

[17] Zhan H, Meng X, Asif M. Risk early warning of a dynamic ideological and political education system based on LSTM-MLP: Online education data processing and optimization[J]. Mobile Networks and Applications, 2024, 29(2): 1-13. https://doi.org/10.1007/s11036-024-02439-0 SpringerLink

[18] Cheng M, Lin Y, Li M. Sequence-to-sequence neural diarization with automatic speaker detection and representation[J]. IEEE Transactions on Audio, Speech and Language Processing, 2025.

[19] Bois A, Tervil B, Oudre L. A persistent homology-based algorithm for unsupervised anomaly detection in time series[J]. Transactions on Machine Learning Research, 2024.

[20] Wang J, Dudy S, He X, et al. Optimizing speaker diarization for the classroom: Applications in timing student speech and distinguishing teachers from children[J]. Journal of Educational Data Mining, 2025, 17(1): 98-125. https://doi.org/10.5281/zenodo.14871875 jedm.educationaldatamining.org+1

[21] Cai J, Li Y. Fuzzy association rule mining for Personalized English Language Teaching from higher education[J]. Journal of Computational Methods in Sciences and Engineering, 2024, 24(6): 3617-3631. https://doi.org/10.1177/14727978241296748 SAGE Journals+2cahaya-ic.com+2

[22] Essalmi H, El Affar A. Dynamic algorithm for mining relevant association rules via meta-patterns and refinement-based measures[J]. Information, 2025, 16(6): 438-467. https://doi.org/10.3390/info16060438 ResearchGate+1

[23] Sun Z, Peng Q, Mou X, et al. An artificial intelligence-based real-time monitoring framework for time series[J]. Journal of Intelligent & Fuzzy Systems, 2021, 40(6): 10401-10415. https://doi.org/10.3233/JIFS-200366 PMC+3SAGE Journals+3ACM Digital Library+3

[24] Fan L, Zhang J, Mao W, et al. Unsupervised anomaly detection for intermittent sequences based on multi-granularity abnormal pattern mining[J]. Entropy, 2023, 25(1): 123. https://doi.org/10.3390/e25010123 MDPI+2PubMed+2

[25] Janský J, Koldovský Z, Málek J, et al. Auxiliary function-based algorithm for blind extraction of a moving speaker[J]. EURASIP Journal on Audio, Speech, and Music Processing, 2022, 2022(1): 1. https://doi.org/10.1186/s13636-021-00231-6 SpringerLink+1

[26] Schwartz A, Schwartz O, Chazan S E, et al. multi-microphone simultaneous speakers' detection and localization of multi-sources for separation and noise reduction[J]. EURASIP Journal on Audio, Speech, and Music Processing, 2024, 2024(1): 50. https://doi.org/10.1186/s13636-024-00365-3 SpringerLink+1

[27] Tian J, Yu J, Weng C, et al. Improving Mandarin end-to-end speech recognition with word n-gram language model[J]. IEEE Signal Processing Letters, 2022, 29: 812-816.

[28] Kim M, Jang G J. Speaker-attributed training for multi-speaker speech recognition using multi-stage encoders and attention-weighted speaker embedding[J]. Applied Sciences, 2024, 14(18): 8138. https://doi.org/10.3390/app14188138 MDPI+1

[29] Huang Z, Delcroix M, Garcia L P, et al. Joint speaker diarization and speech recognition based on region proposal networks[J]. Computer Speech & Language, 2022, 72: 101316. https://doi.org/10.1016/j.csl.2021.101316

[30] He M K, Du J, Liu Q F, et al. ANSD-MA-MSE: Adaptive neural speaker diarization using memory-aware multi-speaker embedding[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2023, 31: 1561-1573. https://doi.org/10.1109/TASLP.2023.3265199