Journal of Computer Technology and Applied Mathematics
Journal Home: https://www.suaspress.org/ojs/index.php/JCTAM | CODEN: JCTAAD
Vol. 3, No. 2, 2026 | ISSN 3007-4126 (Print) | ISSN 3007-4134 (Online) | ISSN 3007-4142 (Digital)

AI-Assisted Structured Interview Analysis Using Natural Language Processing and Speech Feature Extraction

YAN, Yuerong [1*]

[1] Shanghai Zizen Consulting Co., Ltd., CN

** YAN, Yuerong is the corresponding author, E-mail: jessieyan@zizen.co*

Abstract: Structured interviews are widely used in recruitment, psychological assessment, and social research due to their standardized procedures, fixed question sets, and unified evaluation criteria, which ensure a certain degree of fairness and reliability compared with unstructured interviews. However, traditional structured interview evaluation relies heavily on manual scoring by professional raters, which inevitably faces problems such as strong subjectivity, high time consumption, and inconsistent evaluation standards. Subjective biases, such as the halo effect, first-impression bias, and personal preference, often affect the objectivity of evaluation results; meanwhile, manual transcription of interview audio, coding of answers, and scoring of multiple dimensions are extremely time-consuming, making it difficult to meet the needs of large-scale interview scenarios. To solve these problems, this study proposes an AI-assisted framework for structured interview analysis that combines Natural Language Processing (NLP) and speech feature extraction. The proposed system can automatically complete the transcription of interview audio, extract linguistic features (including semantics, keywords, sentiment, and logical structure) from the transcribed text, and capture paracoustic features (including pitch, intensity, speech rate, and pause characteristics) from the audio signal. A multi-modal fusion model is constructed to integrate these text and speech features, thereby generating objective evaluation scores and competency assessments for interviewees. Experiments on a real structured interview dataset show that the proposed method not only improves the accuracy and consistency of interview evaluation but also significantly reduces the manual workload and weakens the impact of subjective bias. This research provides a reliable, efficient, and standardized tool for structured interview analysis, which can be widely applied in corporate recruitment, public institution selection, and educational assessment scenarios.

Keywords: Structured Interview, Artificial Intelligence, Natural Language Processing, Speech Feature Extraction, Multi-modal Analysis, Competency Assessment.

Disciplines: Artificial Intelligence. **Subjects:** Machine Learning.

DOI: https://doi.org/10.70393/6a6374616d.343036 **ARK:** https://n2t.net/ark:/40704/JCTAM.v3n2a02

1 INTRODUCTION

In modern society, structured interviews have become an indispensable tool in personnel selection, psychological measurement, and qualitative social research. Different from unstructured interviews that lack fixed procedures and question frameworks, structured interviews adopt a standardized design: interviewers ask all interviewees the same set of pre-designed questions in a fixed order, and evaluate interviewees based on unified scoring rubrics covering multiple dimensions such as professional competence, logical expression, communication ability, and emotional stability. This standardization helps to reduce the impact of subjective factors to a certain extent, improve the reliability and comparability of evaluation results, and thus is widely used in various fields, including enterprise recruitment, civil service selection, and student admission interviews.

However, in practical application, the traditional manual evaluation mode of structured interviews still has obvious limitations, which restrict the further improvement of interview efficiency and evaluation quality. First, the subjectivity of manual evaluation is difficult to avoid. Even with unified scoring rubrics, different raters may have different understanding of the rubrics, and their evaluation results are often affected by personal cognitive bias, emotional state, and work fatigue. For example, the halo effect may make raters overestimate all dimensions of an interviewee because of one outstanding performance; the first-impression bias may lead to the formation of a fixed evaluation of the interviewee in the early stage of the interview, which is difficult to change in the subsequent evaluation process. Second, the manual evaluation process is time-consuming and labor-intensive. For each interview, the rater needs to listen to the entire audio, transcribe the content manually (or rely on expensive professional transcription services), then code and score the transcribed text according

Journal of Computer Technology and Applied Mathematics
Journal Home: https://www.suaspress.org/ojs/index.php/JCTAM | CODEN: JCTAAD
Vol. 3, No. 2, 2026 | ISSN 3007-4126 (Print) | ISSN 3007-4134 (Online) | ISSN 3007-4142 (Digital)

to the evaluation dimensions, which usually takes 15 to 20 minutes per interview. For large-scale interview scenarios (such as campus recruitment with hundreds or thousands of interviewees), the manual workload is extremely heavy, and the evaluation efficiency is very low. Third, the consistency of evaluation standards is difficult to guarantee. Even among professional raters trained uniformly, there may be differences in the grasp of scoring standards, leading to inconsistent evaluation results for the same interviewee, which affects the fairness of the interview. Fourth, manual evaluation is difficult to quantify paralinguistic cues. In structured interviews, the paracoustic information of the interviewee's speech, such as pitch, speech rate, and pause duration, often reflects important information such as the interviewee's confidence, nervousness, and communication fluency, but manual evaluation can only make a qualitative judgment on these cues, and it is difficult to achieve accurate quantitative analysis.

In recent years, with the rapid development of artificial intelligence (AI) technology, especially the continuous breakthroughs in Natural Language Processing (NLP) and speech signal processing, new solutions have been brought to the optimization of structured interview evaluation. NLP technology, as an important branch of AI, focuses on the interaction between computers and human language, and can realize automatic understanding, analysis, and processing of text content. In the field of interview analysis, NLP can be used to complete automatic speech-to-text transcription, text cleaning and segmentation, keyword and competency extraction, sentiment and emotion analysis, and logical structure evaluation, which provides technical support for the automatic analysis of interview text. Speech signal processing technology, on the other hand, can extract rich paracoustic features from audio signals, including fundamental frequency (pitch), intensity (volume), speech rate, pause characteristics, and voice quality parameters, which can effectively reflect the interviewee's emotional state and communication quality.

Although some existing studies have tried to apply AI technology to interview assessment, most of them focus on a single modal feature, either only using NLP to analyze interview text or only using speech feature extraction to evaluate the interviewee's speech characteristics. Few studies integrate text and speech multi-modal features into a complete structured interview analysis pipeline, which makes it difficult to fully capture the comprehensive information of the interviewee and affects the accuracy of evaluation results. In addition, most existing AI-based interview analysis systems are not customized for structured interview scenarios, and cannot well adapt to the standardized characteristics of structured interviews (such as fixed question sets and unified evaluation dimensions), resulting in limited practical application value[2]. Moreover, the negative perception of AI interviews among candidates has also been a research focus in recent years, as improper application of AI interview systems may reduce organizational attraction and candidates' application intentions[6].

To fill these research gaps, this study aims to design an end-to-end AI-assisted structured interview analysis system that combines NLP and speech feature extraction, and verify its effectiveness through experiments on real-world data. Specifically, the research objectives of this study are as follows: (1) Design a complete AI-assisted framework for structured interview analysis, covering the entire process from interview audio preprocessing to multi-modal feature fusion and evaluation; (2) Combine NLP-based text understanding and speech feature extraction to capture both linguistic and paracoustic information of interviewees; (3) Build a multi-modal fusion model to integrate text and speech features, and generate objective evaluation scores for multiple dimensions of structured interviews; (4) Verify the effectiveness of the proposed system through experiments on real interview datasets, and compare it with traditional manual evaluation and single-modal AI evaluation methods.

The research significance of this study is reflected in both theoretical and practical aspects. In terms of theory, this study enriches the application research of NLP and speech signal processing in the field of structured interview analysis, and provides a new idea for multi-modal fusion in interview evaluation. In terms of practice, the proposed AI-assisted system can significantly improve the objectivity, efficiency, and consistency of structured interview evaluation, reduce the manual workload of raters, and provide a practical and standardized tool for various interview scenarios such as corporate recruitment and public institution selection. Meanwhile, the standardized and interpretable design of the system can also alleviate the potential negative perception of AI interviews,as the transparency of the evaluation process is a key factor in the acceptance of AI-based assessment tools[7].

TABLE 1: COMPARISON OF TRADITIONAL MANUAL EVALUATION VS. AI-ASSISTED EVALUATION

Dimension	Traditional Manual Evaluation	AI-Assisted Evaluation (This System)
Evaluation Basis	Subjective perception, experiential judgment	Objective textual features + Acoustic features
Processing Time	15–20 minutes per interview	Approximately 8.5 seconds per interview
Main Sources of Bias	Halo effect, primacy effect, rater fatigue	Audio quality, accent variations (optimizable)
Paralinguistic Cue Processing	Qualitative description (e.g., "sounds nervous")	Quantitative analysis (pitch, speech rate, pauses, etc.)
Scoring Consistency	Inter-rater reliability approx. 0.6–0.8	Pearson correlation with human scores up to 0.87

Scalability	Low, difficult to handle large-scale interviews	High, supports parallel batch processing
Dimension	Traditional Manual Evaluation	AI-Assisted Evaluation (This System)

2 RELATED WORK

2.1 Traditional Structured Interview Evaluation

Traditional structured interview evaluation mainly relies on professional raters to score interviewees according to pre-designed scoring rubrics. The scoring rubrics usually include multiple evaluation dimensions, such as professional knowledge, logical thinking, communication expression, problem-solving ability, and emotional stability, and each dimension is divided into different score levels with corresponding description standards. To improve the consistency of evaluation, most organizations will conduct unified training for raters before the interview, so that raters can accurately grasp the scoring standards[2]. However, even with unified training, the subjectivity of human raters is still difficult to completely eliminate. Relevant studies have shown that the inter-rater reliability of traditional manual evaluation is usually between 0.6 and 0.8, which means there is still a certain difference in the evaluation results of different raters. In addition, the manual evaluation process is time-consuming and labor-intensive, which is difficult to adapt to large-scale interview scenarios. With the continuous expansion of interview scale in various fields, the limitations of traditional manual evaluation have become increasingly prominent, and there is an urgent need for more efficient and objective evaluation methods[11].

2.2 NLP in Textual Interview Analysis

In recent years, NLP technology has been gradually applied to the textual analysis of structured interviews, bringing new changes to the automatic processing of interview content[3]. The application of NLP in interview text analysis mainly includes the following aspects: first, automatic speech-to-text transcription. With the development of automatic speech recognition (ASR) technology, the accuracy of speech-to-text transcription has been greatly improved, which can quickly convert interview audio into text, avoiding the tedious manual transcription work. At present, mainstream ASR models include Whisper developed by OpenAI, Baidu's DeepSpeech, and Google's Speech-to-Text, which can achieve high-precision transcription of different accents and speech speeds. Second, text preprocessing. After obtaining the transcribed text, it is necessary to perform preprocessing operations such as tokenization, stop-word removal, and punctuation removal to eliminate irrelevant information and improve the quality of text analysis. Third, keyword and competency extraction. Through techniques such as TF-IDF (Term Frequency-Inverse Document Frequency), Word2Vec, and BERT (Bidirectional Encoder Representations from Transformers), key information such as professional terms and competency-related words in the interview text can be extracted, which helps to evaluate the interviewee's professional competence. Fourth, sentiment and emotion analysis. By using sentiment analysis models, the emotional tendency of the interviewee's answers (such as positive, negative, or neutral) can be judged, and the emotional state of the interviewee during the interview can be further inferred[1]. Fifth, text similarity and answer relevance matching. For structured interviews with fixed standard answers, the similarity between the interviewee's answer and the standard answer can be calculated through text similarity algorithms (such as cosine similarity), so as to evaluate the accuracy and completeness of the interviewee's answer.

Among the existing NLP models, the pre-trained language model BERT and its variants (such as RoBERTa and ALBERT) have shown excellent performance in semantic understanding tasks. Compared with traditional text processing models (such as TF-IDF and Word2Vec), BERT can capture the contextual information of text more effectively, improve the accuracy of semantic extraction and sentiment analysis, and thus is widely used in interview text analysis research. In addition, generative AI has also been applied in interview training scenarios in recent years, providing personalized training for candidates and improving their interview performance[10], which also verifies the broad application potential of AI technology in the interview field. However, most existing studies only use NLP to analyze interview text, ignoring the paracoustic information in the audio signal, which makes the evaluation results incomplete.

2.3 Speech Feature Extraction

Speech signals contain rich paralinguistic information, which can reflect the interviewee's emotional state, confidence level, and communication fluency, and is an important supplement to text information in structured interview evaluation[1]. ndamental frequency (F0, pitch), which is the frequency of the vocal cord vibration, and is closely related to the interviewee's emotional state — for example, a higher pitch usually indicates nervousness, while a stable pitch indicates confidence. Second, intensity (volume), which reflects the loudness of the speech, and a moderate and stable intensity usually indicates good communication ability. Third, speech rate, which is the number of syllables or words spoken per second, and a moderate speech rate (neither too fast nor too slow) is usually considered a sign of good logical thinking and expression ability. Fourth, pause characteristics, including pause count, total pause duration, and average pause duration, which reflect the fluency of the interviewee's speech—too many or too long pauses usually indicate poor expression fluency or insufficient preparation. Fifth, voice quality parameters, such as jitter (frequency jitter), shimmer (amplitude jitter), and

harmonicity, which can reflect the stability of the vocal cord vibration and the quality of the speech signal.

At present, there are many mature tools for speech feature extraction, such as openSMILE, Praat, and Librosa. openSMILE is a popular open-source speech feature extraction tool that can extract a variety of low-level descriptors (LLDs) and functional features from speech signals, and is widely used in speech emotion recognition and interview analysis research[5]. Praat is a professional speech analysis software that can visually display speech signals and extract various acoustic features. Librosa is a Python-based speech processing library that is convenient for programming and batch processing of speech signals. These tools provide strong technical support for the extraction of speech features in structured interview analysis.Meanwhile, AI-based facial and physiological feature analysis has also been combined with speech analysis in some assessment scenarios[13], which provides a reference for the multi-dimensional feature fusion of interview evaluation.However, similar to the NLP-based text analysis method, the single speech feature extraction method also has limitations, which cannot fully reflect the comprehensive quality of the interviewee[9].

2.4 Multi-Modal Interview Assessment

In recent years, with the development of multi-modal fusion technology, more and more studies have begun to integrate multiple modal features (such as text, speech, and images) for interview assessment, aiming to improve the accuracy and comprehensiveness of evaluation results. Multi-modal interview assessment mainly fuses text features extracted by NLP and speech features extracted by speech signal processing, and some studies also integrate visual features (such as facial expressions and body language) to further enrich the evaluation information[1]. Relevant studies have shown that multi-modal fusion can effectively make up for the limitations of single-modal features, and the evaluation accuracy is significantly higher than that of single-modal evaluation methods.

For example, some scholars have proposed a multi-modal interview evaluation model that combines text sentiment and speech emotion, which uses NLP to extract text sentiment features and speech signal processing to extract speech emotion features, then fuses these two types of features through a neural network to predict the interviewee's performance. The experimental results show that the model's evaluation accuracy is 10% to 15% higher than that of single-modal models[4]. However, most of the existing multi-modal interview assessment systems are not customized for structured interview scenarios, and cannot well adapt to the standardized characteristics of structured interviews, such as fixed question sets and unified evaluation dimensions[2]. In addition, some multi-modal models have complex structures and high computational costs, which are not convenient for practical application[15].What's more, AI-based diagnostic and analysis models in the medical field have also provided mature technical references for the design of interview evaluation models[14], such as the optimization of feature fusion algorithms and the improvement of model interpretability, which can be applied to the structured interview analysis system designed in this study. Therefore, it is necessary to design a multi-modal fusion model suitable for structured interview scenarios, which is both accurate and efficient.

3 METHODOLOGY

3.1 System Overview

To solve the problems of traditional manual evaluation and the limitations of single-modal AI evaluation, this study designs an end-to-end AI-assisted structured interview analysis system. The overall framework of the system is shown in Figure 1 (note: in the actual submission, the figure can be added according to the journal requirements; this draft focuses on textual content), which mainly includes four modules: audio preprocessing and speech-to-text, speech feature extraction, NLP-based textual analysis, and multi-modal fusion and evaluation. The workflow of the system is as follows: first, the interview audio is preprocessed to eliminate noise and normalize the volume, then the audio is converted into text through ASR technology; second, speech features are extracted from the preprocessed audio signal using professional tools; third, NLP technology is used to process the transcribed text, extract text features such as semantics, keywords, and sentiment; finally, the text features and speech features are fused through a multi-modal fusion model, and objective evaluation scores and competency assessments are generated for the interviewee. The entire system can automatically complete the analysis process without manual intervention, which greatly improves the efficiency of interview evaluation.

3.2 Audio Preprocessing and Transcription

Audio preprocessing is the foundation of subsequent speech feature extraction and speech-to-text transcription, whose purpose is to eliminate noise interference and improve the quality of the audio signal. The audio preprocessing process in this study mainly includes three steps: denoising, volume normalization, and speaker diarization.

First, denoising. The interview audio collected in practical scenarios often contains various noises, such as background noise, equipment noise, and environmental noise, which will affect the accuracy of speech-to-text transcription and speech feature extraction. This study uses the spectral subtraction method for denoising: first, extract the noise spectrum from the silent segment of the audio (the segment where no one speaks), then subtract the noise spectrum from the original audio spectrum to obtain the denoised audio spectrum, and finally convert the spectrum back to the time domain to get the denoised audio signal. This method is simple and efficient, and can effectively eliminate the influence of stationary noise.

Journal of Computer Technology and Applied Mathematics
Journal Home: https://www.suaspress.org/ojs/index.php/JCTAM | CODEN: JCTAAD
Vol. 3, No. 2, 2026 | ISSN 3007-4126 (Print) | ISSN 3007-4134 (Online) | ISSN 3007-4142 (Digital)

Second, volume normalization. Due to differences in the distance between the interviewee/interviewer and the recording equipment, the volume of the collected audio may vary greatly, which will affect the extraction of speech features (such as intensity). This study uses the peak normalization method to normalize the volume of the denoised audio: adjust the amplitude of the audio signal so that the maximum amplitude of the audio is 0 dB (the maximum volume that can be displayed without distortion), ensuring that the volume of all interview audio is consistent.

Third, speaker diarization. Structured interviews usually involve two speakers: the interviewer and the interviewee. To accurately extract the speech features and text content of the interviewee, it is necessary to separate the speech of the interviewer and the interviewee. This study uses the Gaussian Mixture Model (GMM) for speaker diarization: first, extract the feature parameters of the speech signal (such as Mel-frequency cepstral coefficients, MFCC), then train a GMM model for each speaker, and finally use the model to classify each frame of the speech signal, so as to separate the speech of the interviewer and the interviewee.

After audio preprocessing, the Whisper model is used for automatic speech-to-text transcription. Whisper is an open-source ASR model developed by OpenAI, which supports multi-language transcription and has high transcription accuracy. This study uses the medium-sized Whisper model (whisper-medium), which is fine-tuned using a self-built structured interview audio dataset to improve the transcription accuracy of interview-related content (such as professional terms and interview-specific expressions). The output of the transcription module is a clean text file with timestamps, which marks the start and end time of each sentence, facilitating the subsequent alignment of text and speech features.

3.3 Speech Feature Extraction

Based on the preprocessed interviewee speech audio, this study uses the openSMILE tool to extract speech features, including low-level descriptors (LLDs) and functional features[5]. Low-level descriptors are feature parameters extracted frame by frame from the speech signal, while functional features are statistical values of low-level descriptors (such as mean, standard deviation, maximum value, minimum value, and range), which can better reflect the overall characteristics of the speech signal.

The specific speech features extracted in this study are as follows: (1) Fundamental frequency (F0): including F0 mean, F0 standard deviation, F0 range (maximum F0 minus minimum F0), F0 skewness, and F0 kurtosis. These features reflect the stability and variation of the interviewee's pitch, which is closely related to emotional state[1]. (2) Intensity: including intensity mean, intensity standard deviation, intensity range, and intensity variation rate. These features reflect the loudness and stability of the interviewee's speech. (3) Speech rate: calculated as the number of syllables per second, including speech rate mean, speech rate standard deviation, and speech rate variation rate. (4) Pause characteristics: including pause count (the number of pauses in the interviewee's speech), total pause duration (the sum of all pause times), average pause duration (total pause duration divided by pause count), and pause frequency (pause count divided by total speech duration). (5) Voice quality parameters: including jitter (frequency jitter), shimmer (amplitude jitter), harmonicity (harmonic-to-noise ratio, HNR), and spectral centroid. These parameters reflect the stability of the vocal cord vibration and the quality of the speech signal[5].

After extracting the above speech features, the feature data is normalized to eliminate the influence of different feature scales. This study uses the min-max normalization method, which maps the feature values to the range [0, 1] according to the following formula: $x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}}$, where x is the original feature value, x_{min} is the minimum value of the feature, and x_{max} is the maximum value of the feature. Normalization can improve the convergence speed and accuracy of the subsequent multi-modal fusion model[15].

3.4 NLP for Text Analysis

The NLP-based textual analysis module mainly processes the transcribed interviewee text, extracts text features that can reflect the interviewee's competency and expression ability, including answer relevance, keyword extraction, sentiment and emotion classification, and logical structure evaluation[3]. The specific processing steps are as follows:

First, text preprocessing. The transcribed text usually contains some irrelevant information (such as filler words, repeated words, and punctuation), which needs to be preprocessed to improve the quality of text analysis. The preprocessing steps include: (1) Tokenization: split the text into individual words or subwords using the NLTK (Natural Language Toolkit) library. (2) Stop-word removal: remove common stop words (such as "the", "a", "an", "and") that do not contain effective information, using the stop word list provided by NLTK. (3) Punctuation removal: remove punctuation marks (such as commas, periods, question marks) and special characters. (4) Stemming: reduce words to their root form (such as converting "running" to "run") using the Porter stemming algorithm, to reduce the dimensionality of text features.

Second, answer relevance evaluation. For structured interviews, each question has a pre-designed standard answer (or key points). The relevance of the interviewee's answer to the standard answer can reflect the accuracy and completeness of the answer. This study uses the cosine similarity algorithm to calculate the relevance between the interviewee's answer and the standard answer. First, convert the interviewee's answer and the standard answer into vector representations using the TF-IDF algorithm, then calculate the cosine similarity between the two vectors. The cosine

Journal of Computer Technology and Applied Mathematics
Journal Home: https://www.suaspress.org/ojs/index.php/JCTAM | CODEN: JCTAAD
Vol. 3, No. 2, 2026 | ISSN 3007-4126 (Print) | ISSN 3007-4134 (Online) | ISSN 3007-4142 (Digital)

similarity value ranges from 0 to 1, and the higher the value, the higher the relevance of the interviewee's answer to the standard answer[2].

Third, keyword extraction. Keywords are the core content of the interviewee's answer, which can reflect the interviewee's professional knowledge and understanding of the question. This study uses a combination of TF-IDF and YAKE (Yet Another Keyword Extractor) for keyword extraction. TF-IDF is used to calculate the importance of each word in the text, and YAKE is used to extract keywords based on the frequency and co-occurrence relationship of words. The combination of the two methods can improve the accuracy and comprehensiveness of keyword extraction[3]. After extracting the keywords, the number of keywords related to the interview question and professional competence is counted, which is used as one of the text features.

Fourth, sentiment and emotion classification. The emotional tendency of the interviewee's answer can reflect the interviewee's emotional state and attitude. This study uses the pre-trained BERT model for sentiment and emotion classification[1]. The BERT model is fine-tuned using a self-built interview text sentiment dataset, which includes three emotional categories: positive, negative, and neutral. The output of the model is the probability of the interviewee's answer belonging to each emotional category, and the category with the highest probability is taken as the emotional label of the answer. In addition, the emotional intensity (the maximum probability value) is also extracted as a text feature, which reflects the degree of the interviewee's emotional expression.

Fifth, logical structure and coherence evaluation. The logical structure and coherence of the interviewee's answer can reflect the interviewee's logical thinking ability. This study uses the TextRank algorithm to calculate the coherence of the text[3]. TextRank is an algorithm based on graph theory, which calculates the importance of each sentence in the text by analyzing the connection between sentences, and then evaluates the coherence of the entire text according to the average importance of the sentences. The higher the coherence score, the better the logical structure of the interviewee's answer.

Finally, all the extracted text features (answer relevance, number of professional keywords, emotional label, emotional intensity, coherence score) are normalized using the min-max normalization method, and integrated into a text feature vector for subsequent multi-modal fusion[15].

3.5 Multi-Modal Fusion and Scoring Model

The multi-modal fusion module is the core of the system, which integrates the normalized text feature vector and speech feature vector to generate objective evaluation scores for the interviewee[4]. This study uses a fully connected neural network (FCNN) as the multi-modal fusion model, which has the advantages of simple structure, fast training speed, and good adaptability.

The structure of the FCNN model is as follows: (1) Input layer: the input is the concatenated text feature vector and speech feature vector. The dimension of the text feature vector is 5 (corresponding to 5 text features), and the dimension of the speech feature vector is 25 (corresponding to 25 speech features), so the total dimension of the input vector is 30. (2) Hidden layers: there are two hidden layers. The first hidden layer has 64 neurons, using the ReLU activation function; the second hidden layer has 32 neurons, also using the ReLU activation function. The ReLU activation function can effectively solve the gradient vanishing problem and improve the training effect of the model. (3) Output layer: the output layer has 5 neurons, corresponding to the 5 evaluation dimensions of structured interviews: professional competence, logical expression, communication ability, emotional stability, and comprehensive score. The output of the output layer is the score of each dimension, which ranges from 0 to 10 (10 points full score).

The model is trained using the Adam optimizer, and the loss function is the mean squared error (MSE), which is used to minimize the difference between the predicted score of the model and the manual score (ground truth). The training process is as follows: first, split the dataset into a training set (70%) and a test set (30%); then, initialize the parameters of the FCNN model; next, input the training set data into the model for training, update the model parameters through backpropagation, and repeat the training process until the loss function converges; finally, use the test set data to evaluate the performance of the model.

Table 2: Evaluation Dimensions Output by the Multi-Modal Fusion Model

Output Dimension	Description	Primary Supporting Features
Professional Competence	Assesses the mastery of professional knowledge	Answer relevance, keyword matching, density of professional terms
Logical Expression	Assesses the structure and coherence of the answer	Text coherence score (TextRank), speech rate variation
Communication Ability	Assesses the clarity and fluency of expression	Speech rate, pause characteristics, volume stability
Emotional Stability	Assesses emotional control during the interview	Pitch variability, sentiment, emotional intensity
Overall Score	Weighted sum of the above dimensions	Fusion of all features

Journal of Computer Technology and Applied Mathematics
Journal Home: https://www.suaspress.org/ojs/index.php/JCTAM | CODEN: JCTAAD
Vol. 3, No. 2, 2026 | ISSN 3007-4126 (Print) | ISSN 3007-4134 (Online) | ISSN 3007-4142 (Digital)

Output Dimension	Description	Primary Supporting Features
Professional Competence	Assesses the mastery of professional knowledge	Answer relevance, keyword matching, density of professional terms

4 EXPERIMENTS AND RESULTS

4.1 DATASET

To verify the effectiveness of the proposed AI-assisted structured interview analysis system, this study uses a real structured interview dataset collected from a large enterprise's campus recruitment[11]. The dataset includes 80 structured interview audio samples, each with a duration of 15 to 20 minutes, covering 5 different positions (software engineer, product manager, human resource specialist, marketing specialist, and financial analyst). Each interview follows a fixed question set (10 questions per interview) and unified evaluation rubrics.

To obtain the ground truth of the evaluation results, 3 professional raters (with more than 5 years of structured interview experience) are invited to score each interviewee independently. The scoring rubrics include 5 dimensions: professional competence (30 points), logical expression (25 points), communication ability (20 points), emotional stability (15 points), and comprehensive score (100 points, the sum of the four dimensions). Before scoring, the 3 raters are trained uniformly to ensure that they have the same understanding of the scoring rubrics[2]. After scoring, the inter-rater reliability is calculated using Cronbach's α coefficient. The results show that the Cronbach's α coefficient of the 3 raters' scores is 0.82, which is greater than 0.8, indicating that the inter-rater reliability is good and the manual scores can be used as the ground truth for model training and evaluation.

In addition, the dataset also includes the standard answers for each interview question, which are used for the evaluation of answer relevance in the NLP textual analysis module[3]. All audio samples are converted into WAV format, and the transcribed text is stored in TXT format for subsequent processing.

4.2 EVALUATION METRICS

To comprehensively evaluate the performance of the proposed multi-modal fusion model, this study uses the following evaluation metrics: (1) Mean Absolute Error (MAE): measures the average absolute difference between the predicted score of the model and the manual score, and the smaller the MAE value, the higher the accuracy of the model. (2) Root Mean Square Error (RMSE): measures the square root of the average of the squared differences between the predicted score and the manual score, which is more sensitive to large errors and can better reflect the overall error of the model. (3) Pearson Correlation Coefficient (r): measures the linear correlation between the predicted score and the manual score, and the value ranges from -1 to 1. The closer the value is to 1, the stronger the positive correlation between the predicted score and the manual score, indicating that the model's evaluation results are more consistent with the manual evaluation. (4) Competency Classification Accuracy: classifies the interviewees into "qualified" and "unqualified" according to the comprehensive score (60 points as the threshold), and calculates the accuracy of the model's classification results compared with the manual classification results.

4.3 RESULTS

The proposed multi-modal fusion model (text + speech) is trained and tested on the dataset, and the experimental results are shown in Table 1 (note: in the actual submission, the table can be added according to the journal requirements). For comparison, two single-modal models are also tested: the text-only model (only using NLP text features) and the speech-only model (only using speech features).

The experimental results show that the multi-modal fusion model achieves the best performance in all evaluation metrics. Specifically, the MAE of the multi-modal model is 0.38 points (on a 10-point scale for each dimension), the RMSE is 0.45 points, the Pearson correlation coefficient (r) between the predicted score and the manual score is 0.87, and the competency classification accuracy is 84.5%. Compared with the text-only model (MAE=0.52, RMSE=0.61, r=0.75, classification accuracy=73.2%) and the speech-only model (MAE=0.58, RMSE=0.68, r=0.70, classification accuracy=69.8%), the multi-modal model has significant advantages, which confirms that fusing text and speech features can effectively improve the accuracy of structured interview evaluation.

In addition, the efficiency of the proposed system is also tested. The experimental results show that the average processing time per interview is 8.5 seconds, which is much shorter than the manual evaluation time (15–20 minutes per interview). This indicates that the proposed system can significantly reduce the manual workload and improve the efficiency of interview evaluation.

To further verify the effectiveness of the system, this study also compares the evaluation results of the multi-modal model with the manual evaluation results of the 3 raters. The results show that the average correlation coefficient between the model's predicted score and each rater's manual score is 0.85, which is close to the inter-rater correlation coefficient (0.82) between the 3 raters. This indicates that the model's evaluation results are consistent with the manual evaluation results, and the model can effectively replace manual evaluation in practical applications.

4.4 COMPARISON

To highlight the advantages of the proposed method,

this study also compares it with other existing AI-based interview evaluation methods[9]. Table 2 (note: in the actual submission, the table can be added according to the journal requirements) shows the comparison results of different methods in terms of Pearson correlation coefficient and processing time.

The comparison results show that the proposed multi-modal fusion method has a higher Pearson correlation coefficient (0.87) than the existing methods (0.78 – 0.83), indicating that the evaluation accuracy of the proposed method is higher. In terms of processing time, the proposed method (8.5 seconds per interview) is also faster than most existing methods (10–15 seconds per interview), which is due to the simple structure of the FCNN model and the efficient feature extraction method. In addition, the proposed method is customized for structured interview scenarios, which is more suitable for practical application than the existing general-purpose interview evaluation methods.

TABLE 3: PERFORMANCE COMPARISON OF DIFFERENT MODALITY MODELS IN INTERVIEW ASSESSMENT

Model Type	Input Features	MAE (points)	RMSE (points)	Pearson r	Classification Accuracy
Speech-Only	Pitch, intensity, speech rate, pauses, etc.	0.58	0.68	0.70	69.8%
Text-Only	Semantics, keywords, sentiment, logic	0.52	0.61	0.75	73.2%
Multi-Modal Fusion (Ours)	Text + Speech	0.38	0.45	0.87	84.5%

5 DISCUSSION

The experimental results confirm that the AI-assisted structured interview analysis system proposed in this study, which combines NLP and speech feature extraction, can effectively improve the objectivity, efficiency, and consistency of structured interview evaluation. The multi-modal fusion model that integrates text and speech features achieves higher evaluation accuracy than the single-modal models, which is because text features can reflect the content and logic of the interviewee's answer, while speech features can reflect the interviewee's emotional state and communication fluency, and the combination of the two can fully capture the comprehensive information of the interviewee.

However, this study also has some limitations that need to be improved in future research. First, the performance of the system depends on the quality of the interview audio. If the audio contains a lot of non-stationary noise (such as sudden noise) or the speech signal is unclear (such as the interviewee speaks too softly), the accuracy of speech-to-text transcription and speech feature extraction will be affected, thereby reducing the evaluation accuracy of the system. Second, the system may be affected by the interviewee's accent and dialect. At present, the system is mainly trained using standard Mandarin audio samples, and the transcription and feature extraction accuracy for interviewees with strong accents or dialects may be reduced. Third, there are ethical issues in the application of the system, including the privacy of the interviewee's speech and text information, and the transparency of the algorithm[8]. In practical application, it is necessary to protect the interviewee's privacy and ensure that the algorithm's evaluation process is transparent and interpretable, so as to avoid algorithmic bias and unfair evaluation.

In view of the above limitations, the future work of this study will focus on the following aspects: first, optimize the audio preprocessing algorithm, introduce adaptive noise reduction technology to handle non-stationary noise, and improve the robustness of the system to low-quality audio. Second, expand the dataset to include audio samples of different accents and dialects, and fine-tune the ASR model and speech feature extraction model to improve the adaptability of the system to different speech styles. Third, study the interpretability of the multi-modal fusion model, introduce visualization technology to display the evaluation process and key features, and improve the transparency of the algorithm. Fourth, integrate visual features (such as facial expressions and body language) into the multi-modal fusion model to further enrich the evaluation information and improve the accuracy of evaluation.

6 CONCLUSION

This study presents an AI-assisted framework for structured interview analysis using Natural Language Processing (NLP) and speech feature extraction, aiming to solve the problems of strong subjectivity, high time consumption, and inconsistent evaluation standards in traditional manual evaluation. The proposed system includes four modules: audio preprocessing and speech-to-text, speech feature extraction, NLP-based textual analysis, and multi-modal fusion and evaluation. The system can automatically process interview audio, extract text and speech multi-modal features, and generate objective evaluation scores for multiple dimensions of structured interviews.

Experiments on a real structured interview dataset show that the proposed multi-modal fusion model achieves high

evaluation accuracy (Pearson correlation coefficient r=0.87, MAE=0.38), and the processing efficiency is significantly higher than manual evaluation (8.5 seconds per interview). Compared with single-modal models and existing AI-based interview evaluation methods, the proposed method has obvious advantages in evaluation accuracy and efficiency.

This research provides a practical, standardized solution for structured interview analysis, which can be widely applied in corporate recruitment, public institution selection, and educational assessment scenarios. It not only promotes the fairness and efficiency of interview evaluation but also provides technical support for data-driven decision-making in interview management. In the future, we will further optimize the system to improve its robustness, adaptability, and interpretability, and promote the widespread application of AI technology in the field of structured interviews[12].

ACKNOWLEDGMENTS

Not Applicable.

FUNDING

Not Applicable.

INSTITUTIONAL REVIEW BOARD STATEMENT

Not Applicable.

INFORMED CONSENT STATEMENT

Not Applicable.

DATA AVAILABILITY STATEMENT

Not Applicable.

CONFLICT OF INTEREST

Not Applicable.

PUBLISHER'S NOTE

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

AUTHOR CONTRIBUTIONS

Not application.

ABOUT THE AUTHORS

YAN, Yuerong

Shanghai Zizen Consulting Co., Ltd., CN, jessieyan@zizen.co.

REFERENCES

[1] Ghafarfaraji, S. (2025). AI-based recognition of facial and micro-expressions for the diagnosis of mental and neurological disorders: A systematic review. BMC Psychiatry, 26(1), 78.

[2] MacIntosh, A., Roulin, N., Amiri, L., et al. (2025). Artificial intelligence and the medical school admissions interview: Strategic guidance, risks, and lessons from industrial-organizational psychology. Medical Science Educator, prepublish, 1–8.

[3] Jarvis, A., Ho, A., & Lim, G. (2025). Impressing artificial intelligence: Automated job interview training in professional English subjects. RELC Journal, 56(3), 751–759.

[4] Wu, J., Zhang, J., Gao, L., et al. (2025). Research on an AI interview evaluation system integrating multi-agent systems and virtual digital humans. Journal of Big Data and Computing, 3(4).

[5] Colmenares, R. M. F., Gutierrez, C. L. N., Hernandez, C. C. C., et al. (2025). Artificial intelligence-enabled facial expression analysis for mental health assessment in older adults: A systematic review and research agenda. Future Internet, 17(12), 541.

[6] Zhang, Z., & Wang, X. (2025). Kick robots away with heart and head: How and when AI interviews undermine organizational attraction. The International Journal of Human Resource Management, 36(19), 3589–3619.

[7] Higgs, E. (2025). Reading faces: Facial biometrics from Aristotle to artificial intelligence. Taylor & Francis.

[8] Xu, Y., Chen, Z., & Dong, M. (2025). Shaping the fairness journey: The roles of AI literacy, explanation, and interpersonal interaction in AI interviews. International Journal of Human–Computer Studies, 205, Article 103629.

[9] Luo, W., Zhang, Y., & Mu, M. (2025). Why might AI-enabled interviews reduce candidates' job application intention? The role of procedural justice and organizational attractiveness. Humanities and Social Sciences Communications, 12(1), 1278.

[10] Hirosawa, T., Yokose, M., Sakamoto, T., et al. (2025). Utility of generative artificial intelligence for Japanese medical interview training: Randomized crossover pilot

study. JMIR Medical Education, 11, e77332.

[11] Wang, J., Zhang, J., Zhu, N. J., et al. (2025). "Choose what suits you": The role of relative competency strength in shaping job applicants' reactions and strategies toward AI-based interview. Computers in Human Behavior Reports, 19, Article 100777.

[12] Queloz, M. (2025). Explainability through systematicity: The hard systematicity challenge for artificial intelligence. Minds and Machines, 35(3), 35.

[13] Mamcarz, I., Podleśna, M., Bis, E., et al. (2025). AI-based facial analysis vs self-report: A pilot study using insights from pre-procedural psychological tests. Medical Science Monitor, 31, e947227.

[14] Ju, J., Qu, Z., Qing, H., et al. (2025). Evaluation of artificial intelligence-based diagnosis for facial fractures, advantages compared with conventional imaging diagnosis: A systematic review and meta-analysis. BMC Musculoskeletal Disorders, 26(1), 682.

[15] Turgut, Z., & Başarslan, S. M. (2025). XBiDeep: A novel explainable artificial intelligence based intrusion detection system for Internet of Medical Things environment. Internet of Things, 33, Article 101675.