Improving Cross-Modal Matching of College English Listening Materials Using Contrastive Learning

Xuan Du
School of Foreign languages, Huainan Normal University, Dongshan West Road, Huainan 232038, Anhui, China
Corresponding author: Xuan Du (e-mail: 18205546373@163.com)

ABSTRACT Low audio-text matching accuracy in college English listening materials is caused by modality gaps, pronunciation variability, and scarce annotated data. To address this problem, this paper proposes a cross-modal alignment method based on contrastive learning. Wav2Vec 2.0 and BERT are used to extract speech and text features, respectively, followed by modality-specific projection networks that map heterogeneous representations into a shared semantic space. Under extreme data scarcity with a sample size of 5, semantically constructed positive and negative pairs and an in-batch hard negative mining strategy are introduced to improve discriminability. The InfoNCE loss optimizes cross-modal similarity, while a symmetric dual-tower encoder balances semantic consistency and feature independence. Experimental results show Top-1 accuracy of 94.7% $\pm$ 1.3% and Recall@5 of 97.6% $\pm$ 0.9%, outperforming AudioCLIP, CLAP, and Wav2CLIP baselines. A similarity gap of 0.58 at 10 dB confirms a clearly separated cross-modal decision boundary. The proposed approach mitigates modality disparity and small-sample limitations, providing a robust semantic alignment framework for intelligent listening instruction and low-resource multimodal signal matching.

INDEX TERMS contrastive learning, cross-modal matching, Wav2Vec 2.0, semantic alignment, signal matching

I. INTRODUCTION

THE essence of language comprehension lies in the collaborative decoding of multimodal information; in college English listening instruction, speech and text form a natural semantic coupling system [1,2]. Learners receive speech streams through hearing and rely on text materials to reconstruct meaning. This process implies a complex cross-modal mapping mechanism [3]. Accurate speech-text alignment is not only the basis for intelligent management of teaching resources, but also a prerequisite for personalized learning path design, automatic error analysis, and semantic-level feedback [4-6].

Constructing a computational model that simulates human semantic association ability [7,8] can reveal the deep correspondence between different representational forms of language input. Such a model also supports the transformation of foreign language teaching—from experience-driven to data-semantics-driven.

In the practical application of college English listening materials, semantic matching between speech and text faces multiple structural challenges. Speech signals inherently possess temporal continuity, pronunciation variability, and prosodic diversity; the same sentence can exhibit significant acoustic differences in different contexts or by different readers [9,10]. Text, as a discrete symbol system, expresses semantics in a linear sequence and lacks paralinguistic information such as intonation, pauses, and stress. This creates an asymmetry between the two modalities in terms of expressive granularity and structural morphology [11,12].

This representation gap complicates the direct calculation of cross-modal similarity, making it difficult for models to determine whether a speech segment is semantically equivalent to a text segment, especially when synonymous substitutions, omitted expressions, or sentence reorganizations occur [13,14]. In addition, one-to-many or many-to-one mapping relationships are common in teaching corpora. For example, a recording may correspond to multiple versions of the transcribed text, or the same text may be read aloud by different voice libraries, increasing the uncertainty of the matching [15,16]. Most current systems rely on preset alignment rules or static feature splicing, lacking the ability to model semantic dynamics[17]. More importantly, in real-world teaching scenarios, well-labeled paired data is extremely limited, making it difficult for models to establish robust cross-modal associations through large-scale supervised learning[18,19]. Even if pre-trained models are introduced to extract features, the distance distribution between embedded vectors may still deviate

from the true semantic relationship if fine-grained alignment optimization is not performed in the semantic space. Under small-sample conditions, models are particularly vulnerable to noise interference. Moreover, they often lack the capacity to distinguish between semantically similar yet nonmatching samples [20]. The core goal is to enhance the model's discriminative capabilities in low-resource environments by achieving precise speech and text alignment in a shared semantic space using a contrastive learning framework. The study proposes a dual-tower symmetric architecture integrating Wav2Vec 2.0 and BERT as encoders, ensuring both feature independence and semantic comparability through a modality-specific nonlinear projection network. Training efficiency is achieved via the InfoNCE loss function, which maximizes positive similarity while minimizing negative interference. Crucially, a hard negative example enhancement strategy is introduced within the batch to dynamically select difficult non-matching samples, forcing the model to learn refined discrimination boundaries. The entire training process relies solely on original paired data, requiring no additional manual labeling. Innovations include deep integration of contrastive learning into educational cross-modal matching to reduce supervision reliance, the hard negative enhancement for improved generalization under small samples, and the symmetric dual-tower design for efficient, characteristic-preserving alignment. This method provides a scalable technical path for the semantic understanding module of intelligent listening instruction systems.

II. RELATED WORKS

In recent years, researchers have attempted to model audio-text pairing relationships using deep neural networks, promoting the development of cross-modal matching technology. Khurana S proposed a SAMU-XLSR (Semantically-Aligned Multimodal Utterance-Level Cross-Lingual Speech Representation) learning framework. By combining the XLSR and LaBSE (Language Agnostic BERT Sentence Embedding) models, sentence-level embedding of multilingual speech-to-text is achieved, supporting cross-lingual speech-to-text and speech-to-speech retrieval tasks[21]. However, such methods are prone to weakening the expressiveness of individual features due to modal interference. Other studies have turned to encoder structures, which have been applied to recognition technology[22,23]. Some studies use Deep Speech[24,25] to extract speech features or use Word2Vec[26,27] text representations for comparison. Although this improves module independence, it is difficult to ensure consistency in the cross-modal projection space. To alleviate the data bottleneck, some scholars have introduced self-supervised pre-training strategies[28], using large-scale unlabeled corpora for intra-modal modeling, such as pre-training single-modal encoders based on Masked Language Modeling[29].

Contrastive learning[30,31], as an emerging representation learning paradigm, has shown great potential in fields such as image-text matching and speech recognition. Its core idea is to construct a discriminative embedding space by narrowing the distance between positive samples and pushing the distribution of negative samples further away. In cross-modal learning, existing studies have verified the effectiveness of this framework. Ao T proposed a speech gesture generation framework based on CLIP and diffusion model, which can flexibly control gesture style through text, action clips or videos and ensure semantic accuracy. It outperforms existing methods in terms of realism, appropriateness and style matching. The system also supports fine-grained style control of body parts[32]. These results show that the contrast mechanism can effectively deal with the representation shift problem caused by modality differences. In terms of system improvement, Tan X proposed a text-to-speech system called NaturalSpeech. Through an improved variational autoencoder architecture and a number of key technologies, it achieved for the first time a synthesis quality that was statistically indistinguishable from human speech on the LJSpeech dataset[33]. In the field of educational technology, Yasuda Y verified that its pre-training features can effectively improve the pitch and accent accuracy of the Japanese end-to-end text-to-speech synthesis system by adjusting the fine-tuning strategy of PnG BERT[34]. Fan Y proposed a joint model combining audio and text embedding. By leveraging the contextual information of a pre-trained language model, it was able to generate more natural facial expressions while maintaining accurate lip synchronization, significantly improving the realistic expressiveness of speech-driven 3D facial animation[35]. However, existing applications generally have several limitations: first, hard negative example generation relies on external clustering or static sampling, failing to fully utilize the dynamic semantic distribution within the batch; second, the projection network design is too simple and fails to fully adapt to the characteristic statistical characteristics of speech and text; third, most models neglect the preservation of modality-specific structure, leading to feature homogeneity.

These issues are particularly acute in small-sample teaching scenarios. To address these bottlenecks, this paper proposes a contrastive learning framework that integrates a dual-tower encoding architecture, a shared projection network, and dynamic hard negative example enhancement. This framework aims to improve the robustness and scalability of cross-modal matching for college English listening materials.

III. CROSS-MODAL ALIGNMENT MODELING BASED ON CONTRASTIVE LEARNING

Figure 1 illustrates a contrastive learning-driven cross-modal matching model for college English listening materials. The original audio is processed using Wav2Vec 2.0 to extract frame-level features and pool them into sentence embeddings. The text is encoded using BERT and represented as [CLS] vectors. Both are fed into a two-layer feedforward projection network with identical structure but non-shared weights, mapping the heterogeneous features

into a unified semantic space. In this space, the semantic relevance of audio-text pairs is calculated using cosine similarity. The InfoNCE loss function is introduced, and a contrastive learning objective is constructed using positive sample pairs and hard negative examples within a batch. Loss backpropagation enables joint optimization of the audio encoder, text encoder, and projection network. The overall architecture emphasizes a balance between modality-specificity preservation and semantic consistency alignment. It is suitable for teaching scenarios where annotations are scarce, effectively improving the cross-modal matching accuracy between speech and text in complex contexts, and providing a scalable semantic alignment solution for intelligent listening comprehension systems.

A. CONSTRUCTING BIMODAL INPUT REPRESENTATIONS

1) Speech modality feature extraction and sentence-level vector generation

When processing speech signals from college English listening materials, the raw audio is first preprocessed, resampled to 16kHz, and deep feature encoding is performed using a pretrained Wav2Vec 2.0 model. This model, based on the Transformer architecture, completes self-supervised training on a large-scale unlabeled speech corpus and possesses powerful joint acoustic-semantic modeling capabilities. Although Wav2Vec 2.0 and BERT are pre-trained on general corpora, their contextual modeling capabilities can effectively generalize to domain-specific terms (e.g., "woven"). This paper achieves domain adaptation by aligning phonetic and semantic representations in the projection space through contrastive learning, without requiring fine-tuning of the backbone model. The input audio is segmented into 25ms frames with a step size of 10ms. After initial features are extracted by a convolutional neural network, they are fed into the Transformer encoder, which outputs a high-dimensional sequence of frame-level hidden states. To obtain sentence-level semantic representations suitable for cross-modal matching, a temporal average pooling strategy is used to aggregate the hidden states of all valid frames. Specifically, let the frame-level feature sequence obtained after the speech input is encoded by Wav2Vec 2.0 be $H_a = [h_1^a, h_2^a, \ldots, h_T^a] \in \mathbb{R}^{T\times d}$, where T is the number of valid frames and d is the feature dimension. The final speech sentence vector v_a is defined as:

$$v_a = \frac{1}{T}\sum_{t=1}^{T} h_t^a \tag{1}$$

This operation preserves the overall semantic distribution of the speech signal in the temporal dimension while suppressing noise interference caused by local acoustic fluctuations. The parameters of the Wav2Vec 2.0 encoder are not frozen during training, but are fine-tuned end-to-end with the subsequent projection network through backpropagation to achieve adaptive optimization of speech features in the teaching context. Notably, because Wav2Vec 2.0 has already learned abstract representations at the phoneme, lexical, and even syntactic levels during pre-training, the extracted sentence vectors naturally contain rich semantic information, demonstrating strong robustness against common teaching scenarios such as synonyms, speech rate variations, and accent differences. Furthermore, to adapt to the input requirements of the subsequent projection network, all speech vectors are uniformly mapped to the target dimension via a linear transformation layer, ensuring consistency in spatial structure with the text modality.

2) Semantic Encoding and Sentence Vector Construction in the Text Modality

The pre-trained language model BERT performs context-aware semantic encoding of the text content corresponding to the listening material. After word segmentation, the input text is labeled with a special classification tag [CLS], and a standard BERT input sequence is constructed. During the model's forward propagation, each Transformer layer dynamically models the contextual dependencies between tokens, ultimately outputting a hidden state for the [CLS] position that contains global semantic information as a representative vector for the entire sentence. After BERT encoding of the text sequence, the output vector corresponding to the [CLS] tag is assumed to be $h_{[\mathrm{CLS}]}^t$, which serves as the initial text sentence vector. This vector not only captures the semantics of the words themselves but also incorporates complex linguistic phenomena such as syntactic structure, referential relations, and contextual context. It is particularly suitable for parsing ellipsis, inversion, and colloquial expressions commonly found in listening materials.

To further enhance the discriminative power of text representations in cross-modal matching tasks and avoid semantic drift caused by differences in expression between modalities, the original [CLS] vector is normalized and an independent linear projection layer is introduced to project it into a shared semantic space. Assuming the projection matrix is W_t, the final text sentence vector is represented as:

$$v_t = W_t \cdot \mathrm{LayerNorm}\left(h_{[\mathrm{CLS}]}^t\right) \tag{2}$$

LayerNorm represents a layer normalization operation, which is used to stabilize the training process and enhance the model's adaptability to texts of varying lengths and complexities. This design ensures that text features gradually approximate the distribution of speech modalities in the semantic space while preserving their linguistic structural characteristics. The entire text encoding process provides high-quality, highly comparable input representations for subsequent modality alignment without compromising BERT's original semantic understanding capabilities. Although the sentence vectors of the two modalities originate from heterogeneous data sources, unified dimensional mapping lays the foundation for accurate cross-modal similarity calculation.

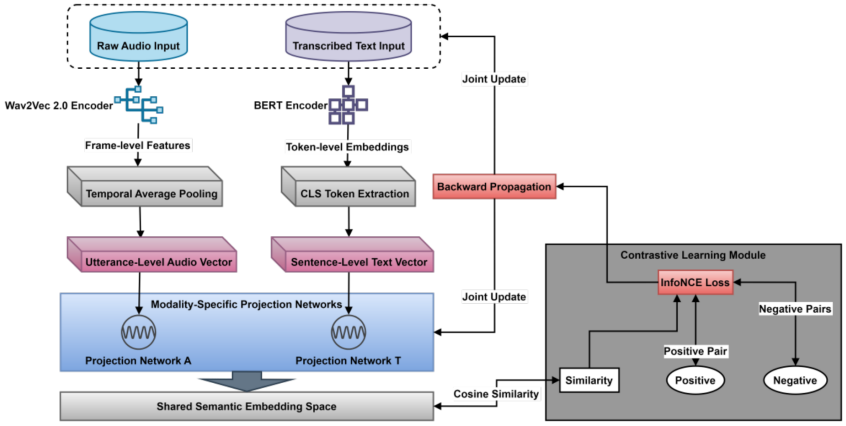

FIGURE 1. Cross-modal matching model driven by contrastive learning

B. DESIGNING SEMANTICALLY CORRELATED POSITIVE AND NEGATIVE SAMPLE PAIRS

1) Construction of Positive Sample Pairs and Semantic Coupling Mechanism

In cross-modal matching tasks, the design of positive sample pairs directly determines the model's ability to learn semantic consistency. This paper, based on real-world pairings in college English listening instruction, constructs positive sample pairs from the original audio recordings and their official transcriptions within the same corpus unit. Specifically, during the data preprocessing phase, each listening passage maintains audio-text synchronization, ensuring strict semantic alignment between the speech signal and the corresponding text. During the model input phase, the audio is encoded using Wav2Vec 2.0 to generate speech sentence vectors, while the accompanying text is extracted using BERT to generate text sentence vectors. The two are combined to form a semantically matched positive sample pair $(s(v_a, v_t))$. These pairings encompass not only standard language expressions but also acoustic variations such as natural intonation, linked speech, and weak pronunciation found in real-world teaching scenarios, allowing the model to fully engage with the diversity of language usage during training.

To enhance the model's sensitivity to semantic equivalence, the similarity of all positive example pairs in the embedding space is explicitly optimized. Cosine similarity is used to measure the strength of cross-modal associations, defined as:

$$s(v_a, v_t) = \frac{v_a^\top v_t}{\|v_a\|\|v_t\|} \tag{3}$$

Here, v_a and v_t represent the projected speech and text vectors, respectively. This similarity value serves as the core criterion for subsequent comparative learning, driving the model to continuously narrow the distance between matching pairs during parameter updates. It is worth noting that since teaching materials often involve standardized question types (such as CET-4 and CET-6 dialogues and IELTS monologues), the sound-text correspondence has high semantic density and structural stability, providing a reliable semantic anchor for positive samples. Furthermore, all positive examples are drawn from real teaching corpora, without manual rewriting or machine translation, preserving the original language style and expression conventions, thus avoiding semantic distortion. This construction strategy based on real pairings enables the model to fully utilize limited alignment resources to establish a robust cross-modal association foundation without the additional cost of additional annotation.

2) Dynamic Screening of In-Batch Hard Negative Examples and Suppression of Semantic Interference

To improve the model's discriminative ability under small sample size conditions, this paper adopts a hard negative example construction strategy based on in-batch sampling, focusing on strengthening the ability to distinguish semantically similar but non-matching examples. Negative examples are not randomly selected.

Instead, for each speech sample in the current training batch, text vectors with high semantic relevance but mismatched content are selected from texts in different corpus units as negative examples. Specifically, for a given speech vector v_a^i, its corresponding negative text candidate set is composed of all unpaired text vectors $\{v_t^j \mid j \neq i\}$ in the same batch. By precalculating the cosine similarity between v_a^i and each non-matching text, and sorting them in descending order, the top three highly similar non-matching texts are selected as "hard negative examples." In InfoNCE loss calculation, the negative sample set for each speech query retains only the top 3 non-matching texts (i.e., hard negatives) with the highest similarity, while the rest are masked.

This strategy achieves gradient optimization by modifying the negative sample terms in the InfoNCE denominator, focusing on the most confusing semantic boundaries.

The core of this mechanism is to dynamically generate the most confusing negative samples by leveraging the semantic distribution of samples within a batch, forcing the model to learn a more refined discrimination boundary. Assuming

there are N speech-text pairs in a batch, each speech sample can obtain N-1 potential negative texts. The three most similar hard negative examples are selected from them, namely:

$$N_{\text{hard}}(v_a^i) = \text{Top-3}_{j \neq i} \left(s(v_a^i, v_t^j) \right) \tag{4}$$

s is the cosine similarity function, and Top-3 means taking the top three items with the highest similarity.

Such negative examples are usually highly similar to positive examples in terms of topics, keywords, or sentence structures. For example, different dialogue fragments under the same topic, or declarative sentences with similar grammatical structures but different semantic directions, constitute the most confusing interference items in real teaching scenarios. By continuously being exposed to such highly challenging negative samples, the model gradually learns to ignore surface vocabulary overlap and instead focuses on deep semantic consistency, significantly improving its robust matching capabilities in complex contexts. The entire process does not require additional labeling or external clustering, and relies entirely on the dynamic generation of data within the batch, which is both efficient and scalable.

Figure 2 reveals the key matching mechanisms in cross-modal contrastive learning through bimodal visualization. Figure 2(a) shows the similarity matrix distribution of speech-text sample pairs. The high-similarity regions on the diagonal represent strictly aligned positive sample pairs, and their concentrated distribution reflects the semantic consistency of the strict pairing of speech and text. The diffuse distribution in the off-diagonal regions reflects the diversity of random negative examples. The red points with significant deviations indicate hard negative examples that pass the threshold screening (similarity > 0.6)—these samples are highly similar to positive examples at the acoustic or lexical level, but are not semantically equivalent, posing a major challenge to model discrimination. Figure 2(b) deconstructs the essential characteristics of the similarity distribution from a statistical perspective: the majority of positive examples have a similarity range of 0.6-0.8, indicating semantic convergence after modal alignment; the high number of random negative examples (abbreviated as Neg) around 0.3-0.6 reveals the baseline difficulty of cross-modal matching; The significant accumulation of difficult negative examples in the 0.6-0.7 range highlights the reality of blurred semantic boundaries. This distribution pattern confirms the prevalence of synonymous paraphrasing and near-sense interference in listening materials, suggesting that the model must go beyond surface feature matching and establish a discriminative criterion based on deeper semantics.

Table 1 lists the construction and training configuration of positive and negative examples in cross-modal contrastive learning. Each speech sample is paired with a positive text example that is strictly semantically aligned with it, ensuring that the model learns the correct semantic associations. Furthermore, three hard negative examples are introduced to filter out semantically similar but content-mismatched text from the same batch, improving the model's ability to discriminate against highly confounded samples. The training batch size is set to 32, balancing computational efficiency and gradient stability. The maximum length of text input is 512 tokens, which meets BERT's processing limits. Audio is retained for a maximum of 30 seconds to prevent feature degradation caused by excessively long speech. All parameters are implemented without the use of external data, using a dynamic sampling mechanism within batches. This supports efficient training in small-sample teaching scenarios and enhances the model's ability to model complex semantic relationships in real listening materials.

TABLE 1. Positive and negative sample construction and training configuration

Parameter Name	Value	Description
Number of Positive per Sample	1	One aligned text for each audio sample
Number of Hard Negative per Samples	3	Three most similar non-matching texts per audio
Batch Size	32	Number of audio-text pairs per training batch
Maximum Text Length (tokens)	512	Maximum input length for BERT encoding
Maximum Audio Duration (seconds)	30	Maximum retained audio length for processing

C. DEFINING THE CROSS-MODAL CONTRASTIVE LOSS FUNCTION

1) Cross-Modal Similarity Optimization Mechanism Based on InfoNCE

To achieve precise alignment of speech and text in a shared semantic space, this paper constructs a contrastive learning objective function for cross-modal matching, using Information-Noise Contrastive Estimation (InfoNCE) as the core loss. This loss explicitly models the relative probabilistic advantage of positive pairs, driving the model to reconstruct semantic neighborhood structures in a high-dimensional embedding space. Specifically, in each training batch, for any positive speech-text pair (v_a^i, v_t^i), its speech vector is used as the query, and semantic search is performed against a candidate pool consisting of all text vectors in the batch. The candidate pool contains one positive example and multiple negative examples. The model learns to ensure that the similarity score of the positive example is significantly higher than that of the remaining distractors. Similarity is calculated using a normalized cosine metric to ensure that comparisons between vectors of different modalities are not affected by amplitude differences.

In this framework, the speech-to-text loss term is defined as:

$$L_{a \to t}^i = -\log \frac{\exp\left(s(v_a^i, v_t^i)/\tau\right)}{\sum_{j=1}^{N} \exp\left(s(v_a^i, v_t^j)/\tau\right)} \tag{5}$$

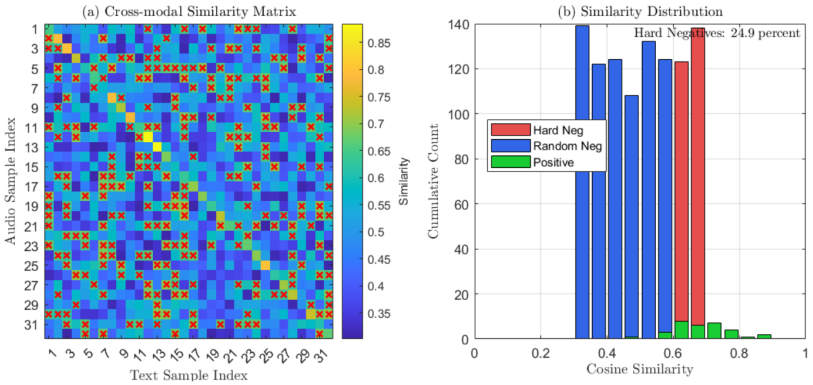

FIGURE 2. Key Matching Mechanisms in Cross-Modal Contrastive Learning

τ is the temperature coefficient, which adjusts the distribution sharpness, and N is the batch size. Formula (5) essentially takes the negative logarithm of the normalized confidence score of the positive sample in all candidate texts, forcing the model to improve the positive score while suppressing the negative response.

The introduction of the temperature parameter effectively alleviates the gradient sparsity problem in the initial training of the embedding space, especially enhancing the stability of the loss function under small sample conditions. Through this mechanism, the model gradually establishes a discriminative logic centered on semantic equivalence, rather than relying on surface feature matching. This loss only applies to the dynamic combination of samples within the current batch, making it naturally compatible with hard-negative example screening strategies. It enables efficient semantic contrastive learning without the need for external memory or complex sampling strategies.

2) Bidirectional Symmetric Loss Construction and Joint Parameter Update

To avoid semantic drift caused by unidirectional matching bias, this paper adopts a symmetric loss design while simultaneously optimizing the contrast objective in the text-to-speech direction. Specifically, a text vector is used as the query, and its corresponding positive examples are identified across all speech vectors.

The loss function for this direction is consistent with that for the speech-to-text direction, but the roles of query and key are swapped. Its expression is:

$$L_{t\to a}^{i} = -\log \frac{\exp\left(s(v_t^i, v_a^i)/\tau\right)}{\sum_{j=1}^{N} \exp\left(s(v_t^i, v_a^j)/\tau\right)} \tag{6}$$

This bidirectional design ensures that the model establishes equivalent semantic mapping capabilities between the two modalities, preventing unidirectional performance degradation caused by encoder capacity differences or feature distribution shifts. The final total loss is the arithmetic average of the two directions:

$$L_{\text{total}} = \frac{1}{2N} \sum_{i=1}^{N} \left(L_{a\to t}^{i} + L_{t\to a}^{i}\right) \tag{7}$$

This loss function jointly updates all trainable parameters of the speech encoder, text encoder, and modality projection layer during backpropagation, achieving end-to-end collaborative optimization. Gradients flow through the dual-tower architecture, prompting the two encoding pathways to gradually converge to a unified representation space under the constraints of the semantic discrimination task.

Figure 3 shows the training convergence curves of Bidirectional Contrastive Loss under different temperature coefficients. Figure 3(a) shows the speech-to-text direction (a→t), and Figure 3(b) shows the text-to-speech direction (t→a). As training epochs increase, the values of each temperature coefficient show a trend of rapid initial decline and gradual convergence in the later stages, which is consistent with the typical optimization dynamics of contrastive learning. The loss decreases rapidly with τ = 0.05 and reaches the lowest final stable value, indicating that the similarity distribution at this scale is the most discriminative and effectively balances gradient strength and semantic distinction. While τ = 0.01 also responds quickly, the similarity response is too sharp, leading to oscillations, suggesting that a small τ value tends to amplify noise interference. The loss decreases more slowly with a larger τ value, indicating that excessively high τ values weaken the contrast between positive and negative samples. These results demonstrate that τ = 0.05 is an optimal choice and provide guidance for model hyperparameter setting.

D. INTRODUCING THE INTERMODAL ALIGNMENT PROJECTION NETWORK

1) Design and Implementation of a Modality-Specific Projection Structure

The modality-specific projection layer refers to two MLP layers, one for speech and one for text, each with independent parameters (see Equation (8)). The shared semantic space refers to the normalized vector space that maps the output vectors of both to the same dimension, which is

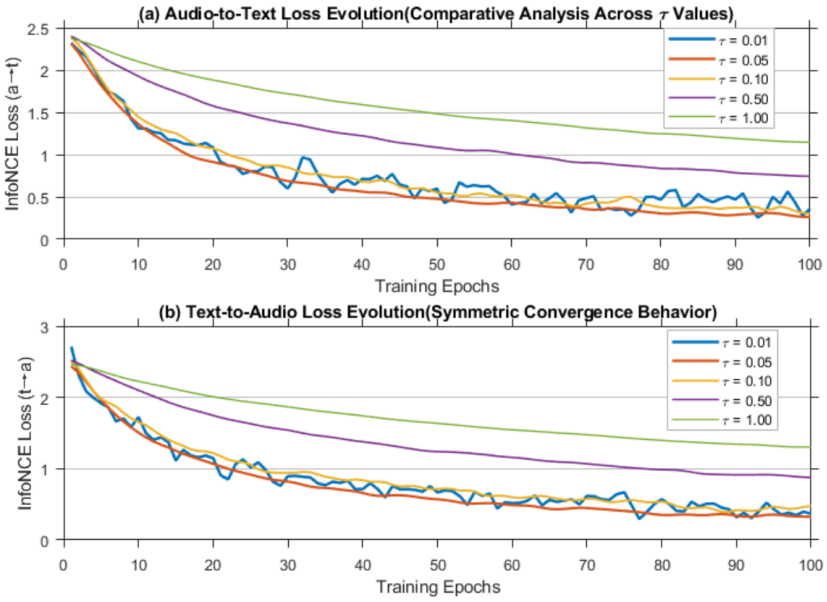

FIGURE 3. Dynamics of Bidirectional Contrastive Loss under Different Temperature Coefficients

used to calculate cross-modal similarity (Equation (3)). This design retains the modality coding specificity while forcing the two to be comparable at the semantic level, thereby achieving a balance between "feature independence" and "semantic consistency".

To address the heterogeneity of the distribution, dimensionality, and semantic granularity of encoded speech and text features, this paper constructs independent feedforward projection networks at the output of the dual-tower encoder to achieve fine alignment of the cross-modal semantic space. Specifically, a nonlinear transformation module consisting of two fully connected layers is connected to the speech sentence embeddings generated by Wav2Vec 2.0 and the text sentence embeddings extracted by BERT. This module receives the high-dimensional latent vector from the encoder as input, first maps it to the intermediate representation space through a linear transformation, then introduces nonlinear modeling capabilities through the GELU activation function, and then outputs the final projection vector through a second layer of linear transformation. Assuming that the input of the speech projection network is v_a, its projection process can be expressed as:

$$z_a = W_2^a \cdot \mathrm{GELU}\left(W_1^a v_a + b_1^a\right) + b_2^a \tag{8}$$

W_1^a and W_2^a are learnable weight matrices, b_1^a and b_2^a are bias terms, and the output z_a is the speech representation mapped to a shared semantic subspace. The text modality uses the same structure, with the projection function $z_t = f_t(v_t; W_1^t, W_2^t)$, ensuring symmetric expressiveness between the two modalities along the transformation path. Although the network structure remains consistent, the projection parameters for speech and text are completely independent, meaning that weights are not shared, thus avoiding the homogenization of modal features caused by forced parameter sharing.

2) Progressive Alignment Mechanism in a Shared Semantic Subspace

The core function of the projection network is to construct a semantically reducible intermediate representation domain, making the high-dimensional features of speech and text comparable within this space. This subspace is not predefined but rather dynamically formed through end-to-end training driven by a comparative learning objective. During the optimization process, the projection layer parameters are updated collaboratively with the encoder, ensuring that the output vectors of different modalities gradually converge when semantically equivalent, while maintaining sufficient distance when semantically divergent.

Because the original feature distributions of speech and text exhibit systematic deviations in direction, scale, and sparsity, directly calculating their similarity is susceptible to interference from non-semantic factors. The projection network reparameterizes the original vector distribution through learnable affine transformations and nonlinear activations, weakening modality-specific noise patterns and enhancing the response strength of cross-modal co-occurrence semantics.

IV. QUANTITATIVE ANALYSIS OF CROSS-MODAL MATCHING PERFORMANCE

A. EXPERIMENTAL DATA

The data used in the experiment are all drawn from a real corpus of university English teaching materials.

They cover the listening modules of the "New Horizon College English" and "New Era Interactive English" textbook series, and include audio and corresponding text for

typical test questions such as CET-4 and CET-6, and IELTS. The audio sampling rate is 16kHz, and each track lasts between 30 seconds and 3 minutes. The audio covers a variety of language styles, including dialogues, monologues, and news broadcasts. The speakers come from different native language backgrounds, preserving the pronunciation differences and speech rate variations seen in real-world teaching. The text is standardized and cleaned to remove timestamps and non-semantic symbols to ensure accurate correspondence with the audio content.

B. IMPACT OF MODALITY-SPECIFIC PROJECTION NETWORKS

Real speech-text paired data was collected and raw feature vectors (eigenvalues) were extracted using pre-trained Wav2Vec 2.0 and BERT models, respectively. Speech features were then subjected to temporal average pooling to obtain sentence-level representations, while text features were encoded using [CLS] tags.

Subsequently, the raw features were fed into a independently designed modality-specific projection network, and projected features were obtained through forward propagation. Finally, the statistical distributions of the input and output features were calculated and the probability densities were plotted. Kernel density estimation was used to smooth the distributional changes.

Figure 4 reveals the critical role of modality-specific projection networks in heterogeneous features through a comparison of probability density distributions. Figures 4(a) and 4(c) show the distribution of original speech and text features, respectively. The speech features exhibit a left-skewed distribution, reflecting the Wav2Vec 2.0 encoding process's sensitivity to asymmetric features such as acoustic energy. The text features exhibit a pronounced bimodal structure, corresponding to the BERT model's discretized modeling of linguistic phenomena (such as affirmative/negative semantics and active and passive voice). This bimodality is essentially a mathematical representation of the complexity of natural language grammatical structure. After processing by the projection network, Figures 4(b) and 4(d) show a systematic shift in the distributions of the two modalities: both are compressed to the interval [-1.5, 1.5].

C. TRAINING DYNAMICS OF THE HARD NEGATIVE EXAMPLE AUGMENTATION STRATEGY

A benchmark dataset containing matched speech-text pairs was constructed. After extracting modal features using Wav2Vec 2.0 and BERT, the cosine similarity matrix of all samples in the batch was calculated in each round of training. For each speech sample, the three most similar non-matching texts were selected as hard negative examples. Contrastive loss was calculated using both basic negative example sampling and hard negative example enhancement strategies. The InfoNCE loss and Recall@1 metric were simultaneously recorded for each round.

Figure 5 compares the training process of the baseline method (random negative sampling) with the hard negative enhancement strategy, revealing the optimization mechanism of dynamic negative example screening for cross-modal contrastive learning. The InfoNCE loss curve on the left vertical axis shows that the hard negative enhancement strategy exhibits rapid convergence in the early stages of training (the first 20 epochs), and later falls below 0.5, indicating that the model rapidly learns more discriminative feature representations by being exposed to semantically similar difficult examples. In contrast, the baseline method's loss is higher, indicating that random negative examples provide a weak gradient signal, making it difficult to effectively distinguish highly confusing samples. The Recall@1 curve on the right vertical axis further demonstrates that the hard negative strategy significantly improves the model's retrieval performance, ultimately stabilizing at a higher level. This phenomenon reflects the persistent effect of hard negative examples: by forcing the model to repeatedly distinguish samples that are superficially similar but semantically mismatched, it gradually strengthens its ability to capture deep semantic relationships, rather than relying on superficial features such as lexical overlap.

D. MATCHING ACCURACY AND RECALL IN SMALL SAMPLE SETTINGS

Top-1 Accuracy and Recall@5 are used as evaluation metrics to measure the model's ability to identify the correct text on the first try and the probability of including the correct match in the top five results, respectively. Testing was conducted with only 5, 10, or 20 labeled examples per category. Sample size refers to the number of labeled samples per semantic category in a small sample evaluation setting. The model is trained using self-supervised contrastive learning on a complete unannotated corpus, without manual annotation. The proposed contrastive learning (abbreviated as Proposed CL) was compared with the latest AudioCLIP (Audio Contrastive Language-Image Pretraining), CLAP (Contrastive Language-Audio Pretraining), and Wav2CLIP (Waveform to CLIP). The model's ability to correctly match audio from candidate text was evaluated.

Figure 6 compares the audio-text matching performance of the proposed method and a mainstream cross-modal pretraining model on College English listening materials under small sample conditions. The horizontal axis represents different annotated sample sizes (5, 10, and 20), while the vertical axis represents the Top-1 Accuracy and Recall@5 metrics, respectively. Results show that the proposed CL significantly outperforms AudioCLIP, CLAP, and Wav2CLIP in all settings. With a sample size of 5, its Top-1 Accuracy and Recall@5 achieve 94.7%$\pm$1.3% and 97.6%$\pm$0.9%, respectively. All models show improved performance with increasing sample size, with proposed model demonstrating enhanced data efficiency and generalization.

This advantage stems primarily from its meticulously decoupled modeling of speech and text modalities, as well as

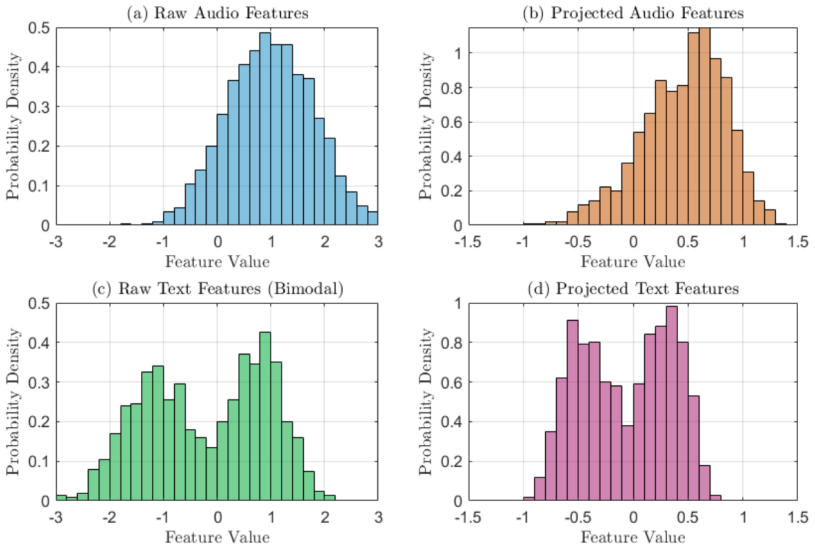

FIGURE 4. The Impact of Modality-Specific Projection Networks

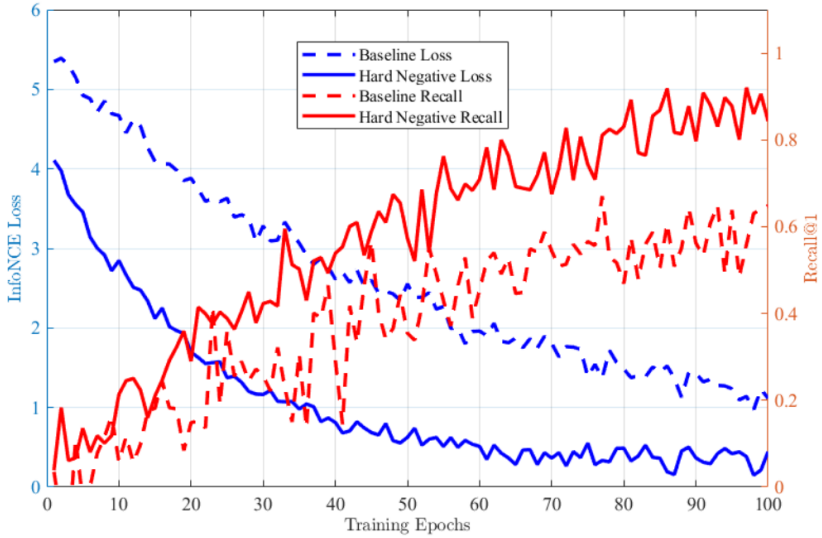

FIGURE 5. Dynamic Comparison of Training with the Hard Negative Enhancement Strategy: Loss Convergence and Recall Improvement

its contrastive learning mechanism driven by hard negative examples, which enables it to construct a highly discriminative shared semantic space even with limited supervision. In contrast, baseline comparison methods, due to their reliance on large-scale pre-training alignment or lack of adaptive design for educational corpora, have blurred semantic discrimination boundaries in low-resource scenarios and are susceptible to interference from pronunciation variation and expression diversity. Furthermore, the high Recall@5 value of the Proposed CL demonstrates its excellent retrieval and ranking capabilities, with the top five candidates almost always containing the correct text, making it suitable for the fault-tolerant matching requirements of actual teaching systems.

E. SEMANTIC ALIGNMENT IN THE EMBEDDING SPACE

The paper calculated the Mean Positive Similarity (MPS) of the positive sample pairs and the mean negative similarity (MNS) of the negative sample pairs, and calculated the similarity gap (SG) between the two to measure the marginal clarity of the model in semantic discrimination. In audio environments with different signal-to-noise ratios (SNR=10dB, 20dB, 30dB, 40dB, 50dB), the alignment effects of contrastive learning in this paper are compared with the latest AudioCLIP, CLAP and Wav2CLIP.

Figure 7 shows the evolution of semantic alignment quality in the embedding space for each model under different signal-to-noise ratio conditions. It presents the similarity between positive and negative pairs, as well as the

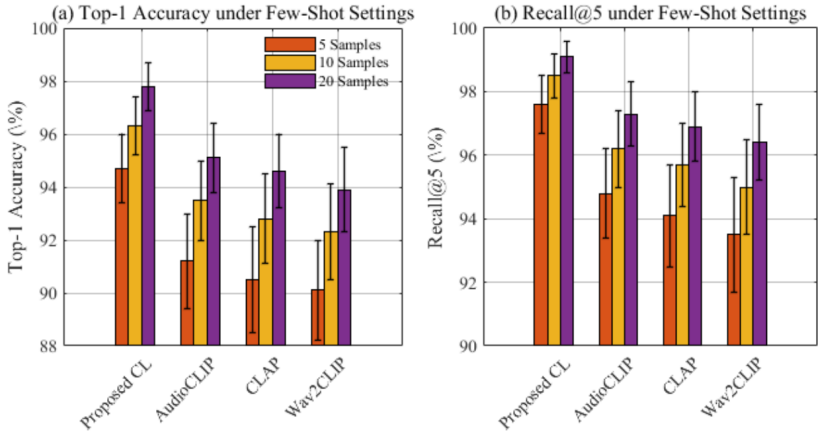

FIGURE 6. Cross-modal Matching Accuracy and Recall

gap between the two, systematically revealing the model's semantic focus capability under noise interference. The proposed contrastive learning method maintains a high MPS in low signal-to-noise ratio environments, reaching 0.89 at 10 dB, demonstrating the robustness of its speech encoder to acoustic degradation, attributed to the explicit modeling of semantic invariance during contrastive learning.

Furthermore, the method significantly outperforms the baseline model in its ability to suppress negative samples, achieving an MPS of 0.31 at 10 dB, demonstrating that the hard-negative training strategy effectively strengthens the cross-modal discrimination boundary. As the signal-to-noise ratio increases, the similarity gap continues to widen, reflecting the model's ability to further enhance semantic separation in clear speech. At 10dB, the SG is 0.58. In contrast, general pre-trained models such as AudioCLIP and CLAP exhibit a smaller positive and negative similarity gap at low SNRs, with a significant decrease in semantic resolution, exposing their limited adaptability to the specific noise structures of educational scenarios.

F. DISCRIMINATION ABILITY OF HARD NEGATIVE SAMPLES

To further evaluate the model's discriminative boundary clarity in semantically confusing environments, this paper constructs two sets of highly distracting negative samples to systematically examine the model's recognition ability and decision robustness in the presence of semantically similar distractors. The first set of negative samples is selected from non-matching texts with a lexical overlap rate exceeding 60%. These samples simulate adversarial interference with highly similar expressions but divergent semantics, reflecting common synonym substitutions and sentence variations in spoken language. The second category is based on unpaired examples with a topic similarity exceeding 50% (calculated using cosine similarity from a pre-trained semantic encoder and manually verified), constructing highly confusing scenarios with related topics but irrelevant content. Building on this, the paper introduces Hard Negative Recall@1 to measure how often the model mistakenly ranks hard negative examples first in cross-modal retrieval. It also defines the False Match Rate (FMR), which measures the proportion of negative examples with a significantly higher similarity than positive examples, to quantify the model's tendency to mismatch. All compared methods were tested five times on the same pool of hard negative examples, and the mean and standard deviation were calculated to ensure the statistical reliability of the evaluation results. This reveals the underlying differences between different architectures in recognizing subtle semantic differences and robust modeling.

Table 2 compares the proposed method's ability to discriminate against hard negative examples in a highly distracting semantic environment with a mainstream cross-modal model. The evaluation covers two types of highly confounding negative examples: high lexical overlap (>60%) and topic similarity (>50%), respectively simulating distractors with superficial formal similarity and deep semantic associations but mismatched content. Hard Negative Recall@1 and FMR are used as core metrics. Results show that the proposed CL significantly outperforms AudioCLIP, CLAP, and Wav2CLIP in both metrics, with average Hard Negative Recall@1 values no higher than 11.5% and average FMR no higher than 12.3%, demonstrating superior semantic denoising capabilities and sharpness of discrimination boundaries. In contrast, baseline methods, due to their over-reliance on word-unit matching or global semantic alignment, are prone to high-confidence false positives when faced with semantically similar interference. The superior performance of the proposed CL is attributed to its in-batch hard negative example enhancement mechanism and modality-specific projection structure introduced during training. These effectively enhance the model's sensitivity to subtle semantic differences and suppress misleading information caused by superficial similarities. These results demonstrate the method's robustness to diverse expressions and semantic ambiguity in real-world teaching scenarios, providing a reliable guarantee for high-precision speech-to-

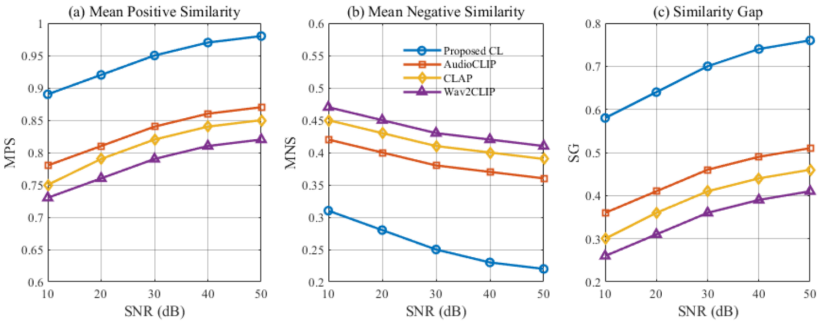

FIGURE 7. Embedding Space Alignment under Different Signal-to-Noise Ratio Conditions

text alignment.

TABLE 2. Hard Negative Sample Discrimination Ability

Method	Evaluation Condition	Hard Negative Recall@1 (%)	False Match Rate (%)
Proposed CL	Lexical Overlap >60%	8.2 ± 0.9	9.7 ± 1.1
Proposed CL	Topic Similarity >50%	11.5 ± 1.3	12.3 ± 1.4
AudioCLIP	Lexical Overlap >60%	23.6 ± 1.8	26.8 ± 2.1
AudioCLIP	Topic Similarity >50%	28.4 ± 2.3	31.1 ± 2.5
CLAP	Lexical Overlap >60%	29.7 ± 2.5	33.5 ± 2.7
CLAP	Topic Similarity >50%	34.2 ± 2.8	37.9 ± 3.0
Wav2CLIP	Lexical Overlap >60%	32.1 ± 2.9	36.4 ± 3.2
Wav2CLIP	Topic Similarity >50%	36.8 ± 3.1	40.2 ± 3.4

G. CROSS-TEXTBOOK TRANSFER MATCHING PERFORMANCE

To test the model's generalization ability across different teaching systems, a cross-textbook transfer experiment was designed, using non-overlapping domain data for external validation. The training set was derived from the listening corpus of New Horizon College English, including both audio and text. The test set was drawn entirely from independent modules of New Era Interactive English. The two models exhibit systematic differences in topic distribution, speaking speed and style, reading characteristics, and textual expression conventions. Evaluation metrics include In-Domain Accuracy and Cross-Dataset Accuracy. Cross-Dataset Accuracy measures the model's absolute matching performance on unseen textbooks. The Domain Gap Ratio (DGR) is also introduced. This is defined as the ratio of the model's performance degradation on the same textbook validation set to the cross-textbook test set, quantifying the generalization error during the transfer process. All models were optimized under the same training conditions, with only the target domain data switched during the test phase to avoid any data leakage.

Table 3 compares the generalization performance of the proposed method with mainstream cross-modal models in a cross-textbook transfer scenario. The proposed CL achieves a cross-dataset accuracy of 95.7% ± 1.1%, significantly higher than the comparison model, and a DGR of only 2.5 ± 0.3%, indicating that its semantic alignment ability is minimally affected by the textbook structure. In contrast, AudioCLIP, CLAP, and Wav2CLIP, due to their reliance on large-scale general corpus pre-training, have limited representation transferability in specific educational contexts and exhibit significant performance degradation when faced with textbook-level style changes. The superior performance of the proposed CL is due to its detailed modeling of hard negative examples and modality-specific projection structure, which allows it to learn more abstract and stable cross-modal mapping relationships even with limited annotations. The low standard deviation further verifies the robustness of its training process, indicating that the method has consistent generalization behavior under different data partitions. This result confirms that the contrastive learning framework designed for educational tasks has greater adaptability and practical value in cross-textbook deployment.

TABLE 3. Cross-Textbook Transfer Matching Performance

Method	In-Domain Accuracy	Cross-Dataset Accuracy	Domain Gap Ratio
Proposed CL	98.2 ± 0.8	95.7 ± 1.1	2.5 ± 0.3
AudioCLIP	96.8 ± 1.3	89.3 ± 1.9	7.7 ± 0.6
CLAP	95.9 ± 1.5	86.1 ± 2.2	10.2 ± 0.8
Wav2CLIP	94.6 ± 1.7	82.4 ± 2.5	12.8 ± 1.0

V. CONCLUSION

This paper addresses the insufficient cross-modal matching accuracy in college English listening instruction.

This limitation stems from two key factors: the significant disparity between speech and text modalities, and the scarcity of annotated data. By constructing independent encoding pathways for speech and text and incorporating a modality-specific nonlinear projection network, the model achieves effective alignment in the semantic space while preserving the respective representational properties. The introduction of a within-batch hard negative example enhancement strategy strengthens the model's ability to discriminate semantically similar distractors, improving generalization performance under small sample conditions. The symmetric

design of the InfoNCE loss function further optimizes the cross-modal similarity distribution and enhances semantic sensitivity during retrieval. Experiments validate the superiority of this method across multiple evaluation dimensions, including matching accuracy in low-resource scenarios, semantic separation in the embedding space, robustness in noisy environments, and stability in cross-textbook transfer. Results demonstrate that the proposed method not only effectively addresses the expression diversity and acoustic variation commonly found in teaching corpora, but also demonstrates excellent domain adaptability. It should be noted that this model only utilizes the frozen encoders of Wav2Vec 2.0 and BERT during training (without updating their parameters), with minimal computational overhead for projection networks and contrastive loss optimization. During deployment, it requires only feedforward inference, enabling real-time audio-to-text matching on standard GPUs or edge devices to meet the latency requirements of the teaching system. This research offers a scalable technical path for semantic-level content understanding in intelligent listening instruction systems. It advances the intelligent management of foreign language listening resources and enables personalized learning support.

AUTHOR CONTRIBUTIONS

Xuan Du designed, collected and analyzed the data, and drafted the manuscript. Xuan Du conducted the study, critically revised the manuscript for important intellectual content, and gave final approval of the version to be published. Xuan Du participated fully in the work, take public responsibility for appropriate portions of the content, and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

CONFLICTS OF INTEREST

The author declares no conflict of interest.

FUNDING

This work was sponsored in part by The Quality Engineering Project of Anhui Province: Blended online and offline courses: College English for Arts and Physical Education(2022xsxx207).

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] M. Gao, W. Song, C. Zhang, Q. Zhou, and P. Su, "Situational teaching based evaluation of college students' English reading, listening, and speaking ability," International Journal of Emerging Technologies in Learning (iJET), vol. 17, no. 8, pp. 140-154, 2022, doi: 10.3991/ijet.v17i08.30561.

[2] H. Liu, R. Chen, S. Cao, and H. Lv, "Evaluation of college English teaching quality based on grey clustering analysis," International Journal of Emerging Technologies in Learning (iJET), vol. 16, no. 2, pp. 173-187, 2021, doi: 10.3991/ijet.v16i02.19727.

[3] S. C. Tsai, "Learning with mobile augmented reality-and automatic speech recognition-based materials for English listening and speaking skills: Effectiveness and perceptions of non-English major English as a foreign language students," Journal of Educational Computing Research, vol. 61, no. 2, pp. 444-465, 2023, doi: 10.1177/07356331221111203.

[4] N. Li, "A fuzzy evaluation model of college English teaching quality based on analytic hierarchy process," International Journal of Emerging Technologies in Learning (iJET), vol. 16, no. 2, pp. 17-30, 2021, doi: 10.3991/ijet.v16i02.19731.

[5] T. Hao, H. Sheng, Y. Ardasheva, and Z. Wang, "Effects of dual subtitles on Chinese students' English listening comprehension and vocabulary learning," The Asia-Pacific Education Researcher, vol. 31, no. 5, pp. 529-540, 2022, doi: 10.1007/s40299-021-00601-w.

[6] H. H. H. Ali, "The importance of the four English language skills: Reading, writing, speaking, and listening in teaching Iraqi learners," Humanities & Natural Sciences Journal, vol. 3, no. 2, pp. 154-165, 2022, doi: 10.53796/hnsj3210.

[7] R. Al-Jarf, "Mobile Audiobooks, Listening Comprehension and EFL College Students," Online Submission, vol. 9, no. 4, pp. 410-423, 2021, doi: 10.29121/granthaalayah.v9.i4.2021.3868.

[8] J. Yuan, Y. Liu, X. Han, A. Li, and L. Zhao, "Educational metaverse: an exploration and practice of VR wisdom teaching model in Chinese Open University English course," Interactive Technology and Smart Education, vol. 20, no. 3, pp. 403-421, 2023, doi: 10.1108/ITSE-10-2022-0140.

[9] Y. Wang, "Artificial intelligence technologies in college English translation teaching," Journal of psycholinguistic research, vol. 52, no. 5, pp. 1525-1544, 2023, doi: 10.1007/s10936-023-09960-5.

[10] N. Noviana and L. Oktaviani, "The correlation between college student personality types and English proficiency ability at Universitas Teknokrat Indonesia," Journal of English Language teaching and learning, vol. 3, no. 1, pp. 54-60, 2022, doi: 10.33365/jeltl.v3i1.1709.

[11] Y. Zhang, "The research on critical thinking teaching strategies in college English classroom," Creative Education, vol. 13, no. 4, pp. 1469-1485, 2022, doi: 10.4236/ce.2022.134090.

[12] C. Y. Jao, H. C. Yeh, W. R. Huang, and N. S. Chen, "Using video dubbing to foster college students' English-speaking ability," Computer Assisted Language Learning, vol. 37, no. 4, pp. 585-607, 2024, doi: 10.1080/09588221.2022.2049824.

[13] R. L. Fitriani, "The development of English speaking proficiency to increase students' communication skill in a business and technology college," KOMVERSAL, vol. 4, no. 2, pp. 90-112, 2022, doi: 10.38204/komversal.v4i2.1041.

[14] P. Gou, "Teaching English using mobile applications to improve academic performance and language proficiency of college students," Education and Information Technologies, vol. 28, no. 12, pp. 16935-16949, 2023, doi: 10.1007/s10639-023-11864-9.

[15] R. Gao, "The vocabulary teaching mode based on the theory of constructivism," Theory and Practice in Language Studies, vol. 11, no. 4, pp. 442-446, 2021, doi: 10.17507/tpls.1104.14.

[16] C. T. Y. Yang, S. L. Lai, and H. H. J. Chen, "The impact of intelligent personal assistants on learners' autonomous learning of second language listening and speaking," Interactive Learning Environments, vol. 32, no. 5, pp. 2175-2195, 2024, doi: 10.1080/10494820.2022.2141266.

[17] R. Al-Jarf, "Learning vocabulary in the app store by EFL college students," Online Submission, vol. 5, no. 1, pp. 216-225, 2022, doi: 10.47191/ijsshr/v5-i1-30.

[18] L. Diao and P. Hu, "Deep learning and multimodal target recognition of complex and ambiguous words in automated English learning system," Journal of Intelligent & Fuzzy Systems, vol. 40, no. 4, pp. 7147-7158, 2021, doi: 10.3233/JIFS-189543.

[19] D. Thi Ngu, D. T. Huong, D. T. N. Huy, P. T. Than, and E. S. Dongul, "Language teaching application to English students at master's grade levels on history and macroeconomic-banking management courses in universities and colleges," Journal of Language and Linguistic Studies, vol. 17, no. 3, pp. 1457-1468, 2021, doi: 10.52462/jlls.105.

[20] H. C. Yeh, W. Y. Chang, H. Y. Chen, and L. Heng, "Effects of podcast-making on college students' English speaking skills in higher education," Educational technology research and development, vol. 69, no. 5, pp. 2845-2867, 2021, doi: 10.1007/s11423-021-10026-3.

[21] S. Khurana, A. Laurent, and J. Glass, "Samu-xlsr: Semantically-aligned multimodal utterance-level cross-lingual speech representation," IEEE Journal of Selected Topics in Signal Processing, vol. 16, no. 6, pp. 1493-1504, 2022, doi: 10.1109/JSTSP.2022.3192714.

[22] S. Ge, J. Ren, Y. Shi, Y. Zhang, S. Yang, and J. Yang, "Audio-Text Multimodal Speech Recognition via Dual-Tower Architecture for Mandarin Air Traffic Control Communications," CMC-COMPUTERS MATERIALS & CONTINUA, vol. 78, no. 3, pp. 3215-3245, 2024, doi: 10.32604/cmc.2023.046746.

[23] A. Chakhtouna, S. Sekkate, and A. Adib, "Efficient bimodal emotion recognition system based on speech/text embeddings and ensemble learning fusion," Annals of Telecommunications, vol. 80, no. 5, pp. 379-399, 2025, doi: 10.1007/s12243-025-01088-y.

[24] S. Sekkate, M. Khalil, and A. Adib, "A statistical feature extraction for deep speech emotion recognition in a bilingual scenario," Multimedia Tools and Applications, vol. 82, no. 8, pp. 11443-11460, 2023, doi: 10.1007/s11042-022-14051-z.

[25] V. Poluboina, A. Pulikala, and A. N. Pitchaimuthu, "Deep Speech Denoising with Minimal Dependence on Clean Speech Data," Circuits, Systems, and Signal Processing, vol. 43, no. 6, pp. 3909-3926, 2024, doi: 10.1007/s00034-024-02644-y.

[26] S. J. Johnson, M. R. Murty, and I. Navakanth, "A detailed review on word embedding techniques with emphasis on word2vec," Multimedia Tools and Applications, vol. 83, no. 13, pp. 37979-38007, 2024, doi: 10.1007/s11042-023-17007-z.

[27] G. Di Gennaro, A. Buonanno, and F. A. N. Palmieri, "Considerations about learning Word2Vec," The Journal of Supercomputing, vol. 77, no. 11, pp. 12320-12335, 2021, doi: 10.1007/s11227-021-03743-2.

[28] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, et al., "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," IEEE Journal of Selected Topics in Signal Processing, vol. 16, no. 6, pp. 1505-1518, 2022, doi: 10.1109/JSTSP.2022.3188113.

[29] R. Mao, Q. Liu, K. He, W. Li, and E. Cambria, "The biases of pre-trained language models: An empirical study on prompt-based sentiment analysis and emotion detection," IEEE transactions on affective computing, vol. 14, no. 3, pp. 1743-1753, 2022, doi: 10.1109/TAFFC.2022.3204972.

[30] L. Xu, H. Xie, Z. Li, F. L. Wang, W. Wang, and Q. Li, "Contrastive learning models for sentence representations," ACM Transactions on Intelligent Systems and Technology, vol. 14, no. 4, pp. 1-34, 2023, doi: 10.1145/3593590.

[31] J. Yu, X. Xia, T. Chen, L. Cui, N. Q. V. Hung, and H. Yin, "XSimGCL: Towards extremely simple graph contrastive learning for recommendation," IEEE Transactions on Knowledge and Data Engineering, vol. 36, no. 2, pp. 913-926, 2023, doi: 10.1109/TKDE.2023.3288135.

[32] T. Ao, Z. Zhang, and L. Liu, "Gesturediffuclip: Gesture diffusion model with clip latents," ACM Transactions on Graphics (TOG), vol. 42, no. 4, pp. 1-18, 2023, doi: 10.1145/3592097.

[33] X. Tan, J. Chen, H. Liu, J. Cong, C. Zhang, Y. Liu, et al., "Naturalspeech: End-to-end text-to-speech synthesis with human-level quality," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 46, no. 6, pp. 4234-4245, 2024, doi: 10.1109/TPAMI.2024.3356232.

[34] Y. Yasuda and T. Toda, "Investigation of Japanese PnG BERT language model in text-to-speech synthesis for pitch accent language," IEEE Journal of Selected Topics in Signal Processing, vol. 16, no. 6, pp. 1319-1328, 2022, doi: 10.1109/JSTSP.2022.3190672.

[35] Y. Fan, Z. Lin, J. Saito, W. Wang, and T. Komura, "Joint audio-text model for expressive speech-driven 3d facial animation," Proceedings of the ACM on Computer Graphics and Interactive Techniques, vol. 5, no. 1, pp. 1-15, 2022, doi: 10.1145/3522615.