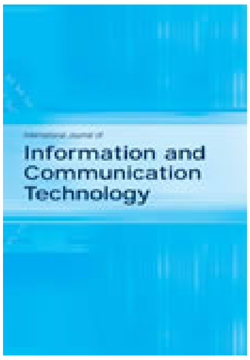


International Journal of Information and Communication Technology

ISSN online: 1741-8070 - ISSN print: 1466-6642
https://www.inderscience.com/ijict

Multimodal English corpus text recognition based on unsupervised domain adaptation

Xiaole Duan, Yan Hu

DOI: 10.1504/IJICT.2025.10070831

Article History:

Received:	12 March 2025
Last revised:	21 March 2025
Accepted:	22 March 2025
Published online:	06 May 2025

Multimodal English corpus text recognition based on unsupervised domain adaptation

Xiaole Duan*

Teaching Department of Public Courses,
Hunan Communication Polytechnic,
Changsha 410132, China
Email: hhtt2024@126.com
*Corresponding author

Yan Hu

College of Intelligent Transportation,
Hunan Communication Polytechnic,
Changsha 410132, China
Email: tytyhy120@163.com

Abstract: With the explosion of multimodal data, it is an important challenge to effectively utilise unlabeled data for cross-modal text recognition. This paper first preprocesses the text and speech data in the English corpus, and use BiLSTM and self-attention mechanism (SA) to extract important text features; and use convolutional neural network, BiLSTM and SA to extract speech features with high contribution. Subsequently, the multimodal features are modelled by graph neural networks, a two-part graph is constructed and knowledge transfer is performed, and domain-invariant features containing information about inter-domain interactions are extracted. Reducing the difficulty of domain adaptation with large inter-domain differences through unsupervised domain adaptation makes the adversarial training process smoother. Finally, the recognition results are obtained by the inference of domain invariant features by the classifier. Experimental results show that the weighted accuracy of the proposed model reaches 93.67%, which significantly improves the recognition effect.

Keywords: multimodal text recognition; self-attention mechanism; SA; unsupervised domain adaptation; UDA; graph neural network; adversarial training; AT.

Reference to this paper should be made as follows: Duan, X. and Hu, Y. (2025) 'Multimodal English corpus text recognition based on unsupervised domain adaptation', *Int. J. Information and Communication Technology*, Vol. 26, No. 11, pp.53–68.

Biographical notes: Xiaole Duan received her Master's degree at Hunan University in 2013. She is currently a Lecturer in the Hunan Communication Polytechnic. Her research interests include natural language processing, English education and applied linguistics.

Yan Hu received his Master's degree at Hunan University in 2012. He is currently an associate professor in the Hunan Communication Polytechnic. His research interests include intelligent traffic system, computer simulation and motion control algorithm.

1 Introduction

As globalisation continues to advance, the field of language learning and research has ushered in new changes, and multimodal English corpora have emerged. While traditional unimodal corpora rely only on textual data (Mirzaei et al., 2023), multimodal English corpora integrate information from multiple modalities, such as text and audio, to provide a more comprehensive resource for English language research (Tu, 2021). In the process of multimodal English corpus construction, text recognition technology is the key link. Accurately recognising the text content can effectively correlate and fuse the text with other modal information, so as to fully explore the linguistic knowledge embedded in multimodal data (Coccetta, 2018). At present, although text recognition techniques have achieved some results in general scenarios, they still face many challenges when applied to multimodal English corpora, such as the synchronisation of different modal data and the accurate recognition of text in complex contexts (Beavis, 2013). The research on text recognition of multimodal English corpus not only helps to improve the corpus construction and promote the innovation of language learning and research methods, but also provides strong support for intelligent education, translation technology and other fields, which has important theoretical and practical significance.

Early English corpus text recognition was based on a unimodal approach. Zhang et al. (2008) used a set of hand-crafted features and trained them with a support vector machine (SVM) classifier, which was unsatisfactory due to the limited expressive power of the hand-crafted features. Liu and Tsai (2021) proposed an intelligent recognition method for English fuzzy text relied on fuzzy computing and big data, and the generalisation ability of the model is relatively weak. Maroof et al. (2024) used CNN to extract character features, and then used random forest to filter the final English letters, and finally used classification to recognise the characters, but the recognition accuracy is not high. Zhong et al. (2019) proposed Gated RCNN based on recursive convolutional neural network (RCNN) and also constructed a bi-directional long-short-term memory for sequence modelling, but it is difficult to recognise irregular text. Ma et al. (2018) proposed the arbitrary orientation network (AON) model, which generates a sequence of features after a designed filter gate and finally generates a sequence of characters using an attentional decoder, but leads to a redundant representation.

In addition to recognising textual features in a single modality, the representation of acoustic features in the English corpus should not be neglected. Song (2020) used CNN and RNN to train the original input signal, extracted the spectral spatial features and temporal features of the speech signal, and used the fully connected layer (FC) for text classification with good results. Li (2020) extracted a number of rhythmic features, including fundamental frequency, energy, etc., and fed them into SVM for text categorisation, and experiments proved that text categories can be well discriminated based on rhythmic features. Leng et al. (2017) extracted relevant audio features as inputs to the hidden Markov model and achieved good experimental results on text recognition tasks.

Single-modality-based text recognition methods for English corpora may suffer from recognition effects when encountering texts in new modalities. Multimodal-based text recognition methods can simultaneously process and fuse information from different modalities, which helps to enhance the accuracy of recognition. Ivanko et al. (2018) have good recognition accuracy in English corpus by combining early fusion and spatial optimisation of text features with acoustic features. Singh et al. (2021) used bi-directional

RNN to encode text and speech, and finally based on the information from the raw data for text sentiment recognition. Liu et al. (2023) used CNN and BiLSTM to extract deep features from MFCC features of speech and text word vectors output from the glove model, and then used the attention mechanism to learn the weights of intra-modal and inter-modal interactions. To solve the issues of data scarcity and model generalisation ability in multimodal recognition, Ding et al. (2019) proposed an adversarial-based unsupervised domain adaptation (UDA) method, which extracts domain-invariant feature representations to spoof the domain discriminator in an adversarial learning manner, thus achieving feature alignment. Diao and Hu (2021) greatly improved the recognition accuracy by domain adaptation of text features and acoustic features based on feature matching and constructing intermediate domains as domain gaps.

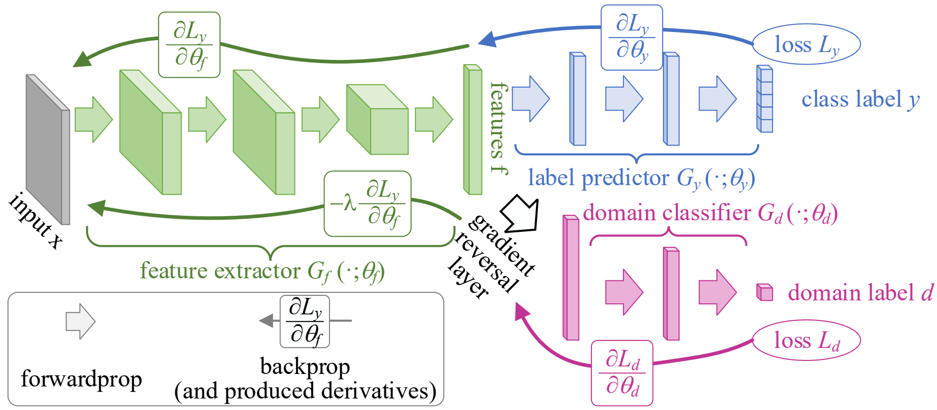
Through the specific analysis of the above research status, it can be found that the existing multimodal text recognition methods have the problems of data sparsity as well as domain shift, to cope with these issues, this paper proposes a multimodal English corpus text recognition model based on UDA. Firstly, glove algorithm and Maier cepstrum coefficient (MFCC) were used to preprocess text and speech data in English corpus, and BiLSTM and SA were used to extract text features with high contribution to text vector output from Glove model. The network framework composed of CNN, BiLSTM and self-attention mechanism (SA) is used to extract speech features with high contribution to Meir spectrum. Then the text and speech modal features are modelled by graph neural network, and the category prototypes and domain prototypes are computed as the node representations of each modality on the graph, on the basis of which a two-part graph is constructed with the training samples and knowledge transfer is carried out, and the domain-invariant features containing the interaction information between the domains are extracted. Second, domain modality-specific markers are extracted for each sample, thus bridging the text and speech domains, which have large distributional differences, and providing a buffer zone making the adversarial training (AT) process smoother. Finally, the recognition results are obtained by inference of domain invariant features by the classifier. The experimental outcome indicates that the weighted accuracy (WA) and F1 of the proposed model reach 93.67% and 91.75%, respectively, and it has a significant advantage on the multimodal English corpus text recognition task.

2 Relevant technologies

2.1 Unsupervised domain adaptation theory

Traditional supervised learning-based neural networks require a large amount of manually labelled data, which poses a serious problem of lack of human and financial resources. This paper considers migrating the trained model from one domain to another domain to achieve good results. UDA aims to use unlabeled target domain data to 'fit' the model and thus improve its performance on the target domain (Liu et al., 2022). UDA techniques are categorised into reconstruction-based approaches, distributional matching-based approaches, and generative adversarial network-based approaches (DANN) (Sicilia et al., 2023). The introduction of domain classifiers in DANN makes the model more capable of reducing the distributional differences between the source and target fields when learning feature representations (Zhang et al., 2022), as implied in Figure 1.

Figure 1 The framework of DANN (see online version for colours)



To capture domain constant characteristics, the parameters are studied by maximising the loss of the field discriminator to the feature extractor, while the parameters of the domain discriminator are learned by minimising the loss of the domain discriminator. Moreover, the loss of the label predictor is also minimised and the target function of DANN is as follows.

$$C_0\left(\theta_f,\theta_y,\theta_d\right)=\frac{1}{n_s}\sum_{x_i\in D_s}L_y\left(G_y(G_f(x_i)),y_i\right)-\frac{\lambda}{n}\sum_{x_i\in(D_s\cup D_t)}L_d\left(G_d(G_f(x_i)),d_i\right) \tag{1}$$

where $n = n_x + n_i$ and λ are trade-off parameters between the two goals of the learning process that form the characteristics. After training convergence, the optimisation function for parameter $\hat{\theta}_f$, $\hat{\theta}_y$, $\hat{\theta}_d$ is as follows.

$$\begin{cases}\left(\hat{\theta}_f,\hat{\theta}_y\right)=\arg\min_{\theta_f,\theta_y} C_0\left(\theta_f,\theta_y,\theta_d\right)\\ \left(\hat{\theta}_d\right)=\arg\max_{\theta_d} C_0\left(\theta_f,\theta_y,\theta_d\right)\end{cases} \tag{2}$$

When the distribution of source domain and target domain can be aligned successfully, domain adversarial network is the best structure for standard domain adaptation.

2.2 Graph neural network

Graph neural networks (GNN) are neural networks that operate on graph-structured data. Unlike conventional neural networks that operate on fixed-size vector inputs, GNNs can handle inputs with different sizes and structures. The core concept of GNN revolves around examining node representations by collecting and integrating information from adjacent nodes within the graph (Bessadok et al., 2021). By iteratively propagating information through the graph, GNNs can learn to capture both local and global feature information of the input graph. GNNs are classified into graph convolutional neural networks (GCN) and graph attention networks (GAT). GCN feature extraction is very strong, while GAT enables each node to perform different levels of information aggregation based on the features of its surrounding nodes by introducing SA.

In GAT, the similarity coefficient e_{ij} between neighbouring nodes $j \in N_i$ connected to node i is computed one by one, assuming that the set of node features h = $\{h_1, h_2, \ldots, h_N\}$, for node i, N_i is the set of its neighbouring nodes.

$$e_{ij} = \vec{a}^T \left[Wh_i \,\|\, Wh_j \right] \tag{3}$$

where W is the parameter matrix, [||] is the feature concatenation, and $\vec{a}^T(\cdot)$ maps the concatenated feature to a real number, thus obtaining the relation among node i and node j. Then the correlation coefficient is normalised to get the corresponding attention coefficient.

$$\alpha_{ij} = \frac{\exp\left(Leaky\,\mathrm{Re}\,LU\left(e_{ij}\right)\right)}{\sum_{k \in N_i} \exp\left(Leaky\,\mathrm{Re}\,LU\left(e_{ik}\right)\right)} \tag{4}$$

where LeakyReLU is the activation function. Then, according to the calculated attention coefficient, the features of adjacent nodes are aggregated to obtain a new node feature, as shown below.

$$h_i' = \sigma\left(\sum_{j \in N_i} \alpha_{ij} W h_j \right) \tag{5}$$

where σ is the activation function and h_i' is the new characteristic extracted from GAT after neighbourhood information fusion.

3 Pre-processing of a multimodal English corpus

The two most common modal data in English corpus are text and audio, in multimodal English corpus text recognition, audio modality can provide additional contextual information or assist recognition, so as to improve the comprehensiveness and accuracy of the recognition results, before proceeding to the construction of the recognition model, it is necessary to pre-process the text and audio data.

3.1 Text modal pre-processing based on glove algorithm

Commonly used word embedding algorithms include Word2Vector, glove (Stein et al., 2019), but Word2Vector is unable to deal with multiple words, while glove utilises co-occurrence matrices to visually represent the relationship between words i and j, to ensure that the word vectors encapsulate as much semantic and syntactic information as feasible. Firstly, inputting the corpus to construct the co-occurrence matrix X and calculate the co-occurrence probability matrix $p_{i,j} = p(j|i) = x_{ij}/x_i$ from X, where $p_{i,j}$ represents the likelihood of words i and j occurring together in the context, $x_{i,j}$ is the amount of times word j occurs in the context of word i, and x_i is the amount of each word appearing amid the backdrop of word i. Then the approximate relationship between word vectors and X is constructed, and the correlation between word k and i and j is judged. Finally, for each word pair (i, j), the number of times they co-occur is calculated and this number is used as the value of the element in the corresponding position in the co-occurrence matrix, obtaining the pre-processed corpus text.

$$F(i,j,k) = \frac{p_{i,k}}{p_{j,k}} \quad (6)$$

3.2 Speech modality pre-processing based on MFCC

Given the speech S in the English pre-feed library, It requires undergoing processes like window insertion and frame division, In the pre-processing section of this article, a Hamming window with a duration of 25 milliseconds and a frame shift of 10 milliseconds are utilised to obtain the pre-processed audio, denoted as $S = \{s_1, s_2, \ldots, s_n\}$, where n is the entire quantity of frames into which the speech is split. Applying the Fourier transform (FFT) to each frame in S produces the representation x_t in the frequency domain.

x_t of each frame, as shown below, where M is the amount of FFT and $0 \le k \le M$, $x_t(k)$ is the k^{th} value in the x_t vector. Then the Mel transform is performed according to equation (8) to convert the frequency of x_t from a linear scale to the Mel scale, in which f represents the frequency scale, and filters are created on the Mel scale specifically to process the spectrum of each frame (Nema and Abdul-Kareem, 2018), and finally the pre-processed speech is obtained.

$$x_t(k) = \sum_{m=0}^{M} s_t(m) \exp\left(-\frac{j2\pi k}{M}\right) \quad (7)$$

$$Mel(f) = 2595 \times \lg\left(1 + \frac{f}{700}\right) \quad (8)$$

4 Multi-channel parallel-based feature extraction for multimodal English corpus

4.1 Text feature coding for English corpus based on BiLSTM and self-attention mechanism

After pre-processing the text and speech modal data in the corpus, this paper adopts CNN-BiLSTM and SA to extract the speech features with high contribution to the Mel spectrum; BiLSTM and SA are used to extract the text emotion features with high contribution to the text vectors outputted from the glove model as shown in Figure 2.

Assuming that the English text feature sequence obtained by word embedding is $x = \{x_1, x_2, \ldots, x_m\} \in R^{m \times d}$. After the word vectors output from the glove model, a BiLSTM network is used to encode the text features at the word level, and then the SA is used to extract the important text features. After the forward LSTM network channel, the text forward feature vector $\vec{T}_i$ is obtained, and after the backward LSTM network channel, the text reverse feature vector $\vec{T}_i$ is obtained.

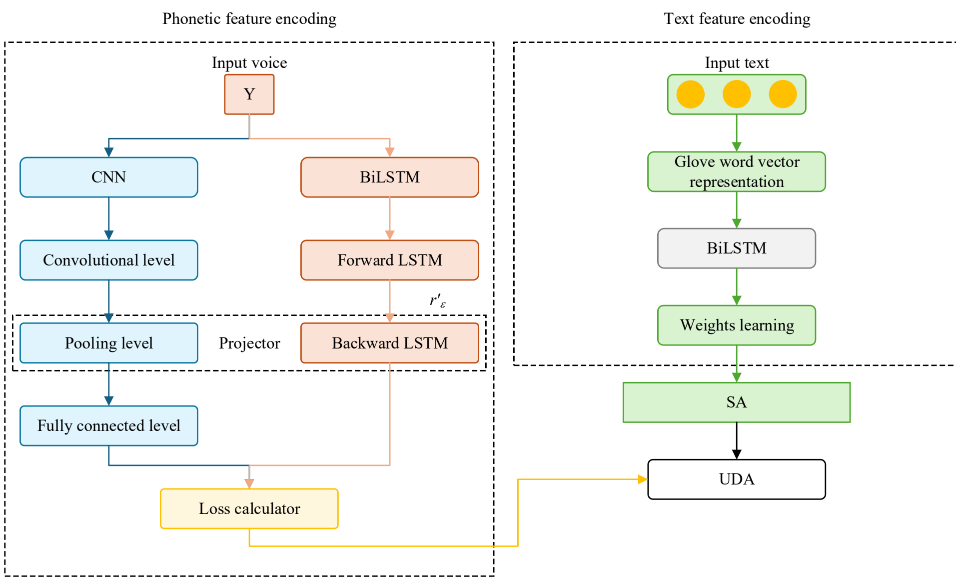
$$\vec{T}_i = \overrightarrow{LSTM}\left(x_i, \vec{T}_{i-1}\right) \quad (9)$$

$$\overleftarrow{T}_i = \overleftarrow{LSTM}\left(x_i, \overleftarrow{T}_{i-1}\right) \quad (10)$$

Finally, $\vec{T}_i$ and $\overleftarrow{T}_i$ are spliced together to obtain the final BiLSTM output $T_i = \left[\vec{T}_i, \overleftarrow{T}_{m-i+1}\right]$. T_i is the encoding of the i^{th} word by BiLSTM, and T_i is input to SA for weight learning as shown in equation (11), where Q is the query vector, K is the eigenvector of T_i, V is the Eigenvalue of T_i, and d_k is the dimensionality of K.

$$Attention(Q, K, V) = soft\max\left(\frac{Q \cdot K^T}{\sqrt{d_k}}\right)V \tag{11}$$

Figure 2 Dual-channel feature extraction process (see online version for colours)



4.2 *CNN-BiLSTM based speech feature coding for English corpus*

Suppose the English speech sequence is $Y = \{y_1, y_2, \ldots y_n\} \in R^{n \times a}$, where n is the amount of acoustic frames and a is the characteristic dimension. First, two 1D convolutional levels are adopted to extract local characteristics, and all convolutional levels are followed by maximum pooling levels. It is employed to decrease the dimensionality of features while preserving the key characteristics, and also prevents overfitting in order to reduce the temporal resolution and facilitate subsequent learning. Then BiLSTM is used to capture the contextual interdependence between frames of the speech signal. Finally, important speech features are extracted by SA.

$$X = ConvBlock(ConvBlock(X)) \tag{12}$$

where $ConvBlock(.) = Maxpool(Conv(\cdot))$, N are the number of acoustic frames after the second pooling level. Y gets audio forward feature vector $\vec{S}_i = \overrightarrow{LSTM}\left(y_i, \vec{S}_{i-1}\right)$ after forward LSTM and audio backward feature vector $\overleftarrow{S}_i = \overleftarrow{LSTM}\left(y_i, \overleftarrow{S}_{i-1}\right)$ after backward

LSTM. Splice $\vec{S}_i$ and $\overleftarrow{S}_i$ to get the final BLSTM output as $S_i = \left[\vec{S}_i, \overleftarrow{S}_{N-i+1}\right]$. Input S_i into SA for weight learning as shown in equation (13), where Q is the query vector, S is the eigenvector of S_i, V is the eigenvalue of S_i, d_s is the dimension of S.

$$Attention(Q, S, V) = soft\max\left(\frac{Q \cdot S^T}{\sqrt{d_S}}\right)V \tag{13}$$

5 Multimodal English corpus text recognition based on unsupervised domain adaptation

5.1 Domain modality-specific labelling based on graph convolutional neural network

After obtaining the text and speech features of the English corpus, in order to solve the data sparsity as well as domain bias problems of existing multimodal recognition methods, a graph neural network is used to model the text and speech modal domains, and a node representation is constructed on the graph for each domain by calculating the category prototypes and domain prototypes. On this basis, a two-part graph is constructed to break the barriers between multimodal domains through feature transfer to enrich the inter-domain interaction information of the samples and enhance the generalisation performance of the model in the objective field. In addition, by adding domain modality-specific markers to each sample, the AT process of the characteristic extractor and domain discriminator is smoother, and the learning difficulty of domain-invariant features is reduced to improve the text recognition efficiency. The framework of the offered recognition model is shown in Figure 3.

In this paper, domain modal specific tag f_m is used to clearly distinguish the domain category of the corpus mode to enhance the dependence of the domain discriminator on f_m. To represent the modal characteristics of each domain more generally, GCN is used to further extract the features of f_m at the domain structure level, and the features of f_m represent f_d. The mini graphs are then used as the basis for the construction of the mini graphs. The node in the mini graph of each domain is a sample feature f_d of this domain, and the adjacency matrix A_m of the mini graph is generated from the domain probabilities of the participating samples in each round of training as output by C_D. The domain probabilities denote the inter-domain similarity between the text and speech modalities. Using these inter-domain similarities, the similarity adjacency matrix $A_m = RR^T$ is constructed, where R is the inter-domain similarity matrix as follows, where softmax is the activation function.

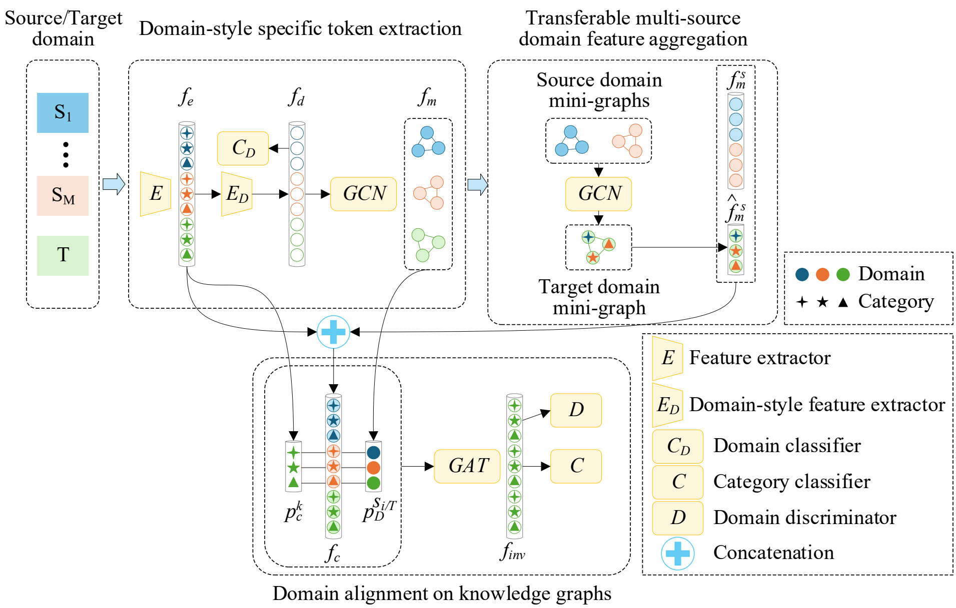
$$R = softmax\left(C_D(f_d)\right) \tag{14}$$

The mini graph of each domain is then fed into the weight-sharing GCN to learn the feature representation of f_m. f_m is computed as follows.

$$f_m = \sigma\left(A_m f_d W\right) \tag{15}$$

where σ is the activation function, W is the matrix of learnable parameters, d is the feature dimension of f_d, A_m is the adjacency matrix of mini graphs in one source or target domain.

Figure 3 The architecture of the proposed recognition model (see online version for colours)



5.2 *Migratable multimodal domain feature aggregation*

To adapt to the multi-modal domain adaptive task, this chapter designs a graph convolution operator for cross-modal domain. The operator can transfer the class semantic information and the aligned domain mode-specific tags from multiple modal domains to the target domain to gain a new objective domain mode-specific tag $\hat{f}_m^T$.

$$\hat{f}_m^T = f_m^T + \gamma\sigma\left(A_t f_m^S W_t\right) \tag{16}$$

where W_t is the weight of the convolution layer of the graph, f_m^S is the feature of the source domain, and f_m is the super parameter. γ is the transfer matrix, representing class-level correspondence from multiple source domains to target domains. A_t is the transfer matrix, which represents the class-level correspondence from multiple source domains to the target domain. In order to obtain a more robust multimodal domain similarity, this paper calculates the similarity between samples by using the domain prototypes between different domains, and obtains the multimodal domain similarity matrix $A_t(i,j)$.

$$A_t(i,j)=\begin{cases}\dfrac{\exp\left(\cos\left(p_D^{S_j},p_D^T\right)\right)}{\sum_{i=1}^{M}\exp\left(\cos\left(p_D^{S_j},p_D^T\right)\right)}, & \hat{y}_i=y_j\\ 0, & otherwis\end{cases} \tag{17}$$

where y_j and $\hat{y}_i$ are the labels of the source domain samples and the pseudo-labels of the target domain samples respectively, p^s and p^T are the domain prototypes of the source and target domains, and cos is the cosine similarity between the two prototypes. This method computes the domain prototypes of each domain by domain modality-specific markers, which in unsupervised learning refer to the centroid of the dataset and represent the overall characteristics of the dataset. For the domain prototype $p_D^{S_i}$ of the i^{th} source domain, define it as the average of the domain modality-specific markers of all samples in the i^{th} source domain.

$$p_D^{S_i}=\frac{1}{N_{S_i}}\sum\nolimits_{x_j\in S_i} GCN\left(E_D\left(E\left(x_j\right)\right)\right) \tag{18}$$

Similarly, the domain prototype p_D^T for the target domain is defined as follows.

$$p_D^T=\frac{1}{N_T}\sum\nolimits_{x_j\in T} x_j\in T \tag{19}$$

5.3 Multimodal unsupervised domain adaptation and text recognition

After the aggregation of multimodal features, it makes the features of text and speech modal domains more compact. In order to better utilise the features of the multimodal domains, this method constructs a bipartite graph on the category prototype and domain prototype, which realises the dissemination of semantic similarity information. The AT approach is also used to train a domain discriminator to extract features with domain invariance. The UDA process for the proposed recognition model is shown in Figure 4.

The new feature D is first obtained by text features T_i, speech features S_i and f_m, where [||] is feature splicing.

$$f_c=\left[T_i \parallel S_i \parallel f_m\right] \tag{20}$$

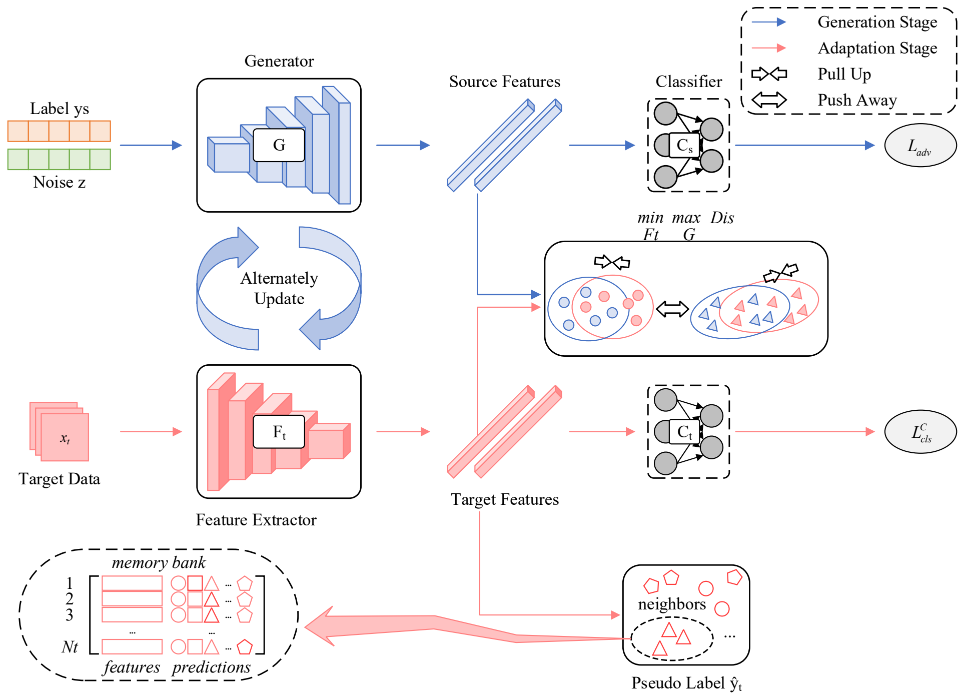
For the k^{th} semantic category prototype p_C^k, it is defined as the average of the features f_e of the k^{th} class of samples over the cross-modal domain as follows, where N_k is the number of samples in the k^{th} class.

$$p_C^k=\frac{1}{N_k}\sum\nolimits_{i=1}^{m}\sum\nolimits_{(x_j,y_k)\in S_i} f_e \tag{21}$$

After obtaining p_C^k and $p_D^{S_i/T}$, the bipartite graph is constructed together with f_e. The vertex set V_{KG} of the bipartite graph can be decomposed into two disjoint subsets, i.e., V_{KG} = $V_P \cup V_f$. Each vertex in V_p is connected to each vertex in V_f. $V_p=\{p_D^{S_1}, p_D^{S_2}, \ldots,$

$p_D^{S_M}, p_D^T, p_C^1, p_C^2, ..., p_C^k\}$ is the set of all domains and class prototypes, and $V_f = \{f_{c,1}^{S_1}, f_{c,2}^{S_1}, ..., f_{c,B}^{S_1}, ..., f_{c,B}^{S_2}, ..., f_{c,B}^{T}\}$ is the set of corpus features for all domains, where $f_{c,i}^{S_i}$ is the f_c feature in the i^{th} source domain, B is the last sample of the current batch size, and $f_{c,i}^{T}$ is the f_c feature of the target domain.

Figure 4 The UDA process for the proposed recognition model (see online version for colours)



The interaction between different modal domains is subsequently realised using GAT messaging on the graph as follows.

$$f_{inv} = GAT(V_K, A_K) \tag{22}$$

where V_K is the prototype of the domain and A_K is the adjacency matrix to model the relationship between the prototype vertex set V_P and the vertices in the sample feature vertex set V_f, defined as follows.

$$A_K(i,j) = \begin{cases} \alpha_{ij} = \dfrac{\exp\left(Leaky\operatorname{Re}LU\left(\vec{a}^T\left[W\vec{v}_j PW\vec{v}_i\right]\right)\right)}{\sum_{V_k \hat{I} V_p} \exp\left(Leaky\operatorname{Re}LU\left(\left(\vec{a}^T\left[W\vec{v}_j PW\vec{v}_k\right]\right)\right)\right)}, & v_i \in V_p \text{ and } v_j \in V_f \\ 0, & otherwise \end{cases} \tag{23}$$

where GAT is computed to obtain the weights α_{ij} of the edges connecting the two vertices V_i and V_j, $\vec{v}_j$ and $\vec{v}_i$ are the feature embeddings of V_j and V_i, respectively, and W

is the weight matrix. $\vec{a}^T$ is the parameterised weight vector. LeakyReLU is the activation function. According to α_{ij}, the node features are aggregated to get the new node features $\vec{v}_i$, and the aggregated features are output through softmax to recognise the text.

$$\vec{v}_i' = \sigma\left(\sum\nolimits_{j \in V_p} \alpha_{ij} W \vec{v}_j\right) \tag{24}$$

Finally, this paper utilises the domain discriminator D for AT. AT is a commonly used domain adaptation method (Zhao et al., 2022), in which a feature extractor and a domain discriminator are trained to make progress together by confronting each other, and the two learn iteratively until the Nash equilibrium is reached, when the features extracted by the feature extractor are considered to be domain-invariant features. The strategy for the loss function of AT is as follows.

$$\begin{aligned} L_{adv} = & \sum_{i=1}^{M} E_{x^{S_i} \sim S_i} \omega\left(x^{S_i}\right) \text{ cross entropy}\left(D\left(f_{inv}^{S_i}\right), i-1\right) \\ & + E_{x^T \sim T} \omega\left(x^T\right) \text{ cross entropy}\left(D\left(f_{inv}^{T}\right), M\right) \end{aligned} \tag{25}$$

where $\omega(x) = 1 + e^{-H(x)}$, $H(x)$ are the predicted entropy of classifier C. To make the domain-invariant features more discriminative, it is also necessary to train a linear classifier C based on domain-invariant features for all source domains, with the classification loss function defined as follows, where y^{S_i} is the semantic label of the English corpus sample.

$$L_{cls}^{C} = \sum_{i=1}^{M} E_{x_i^S \sim S_i} \text{ cross entropy}\left(C\left(f_{inv}^{S_i}\right), y^{S_i}\right) \tag{26}$$

The ultimate goal of training the model is to find the optimal parameters for the proposed method, and the entire target function of the model is obtained by combining L_{adv} and L_{cls}^{C} as follows.

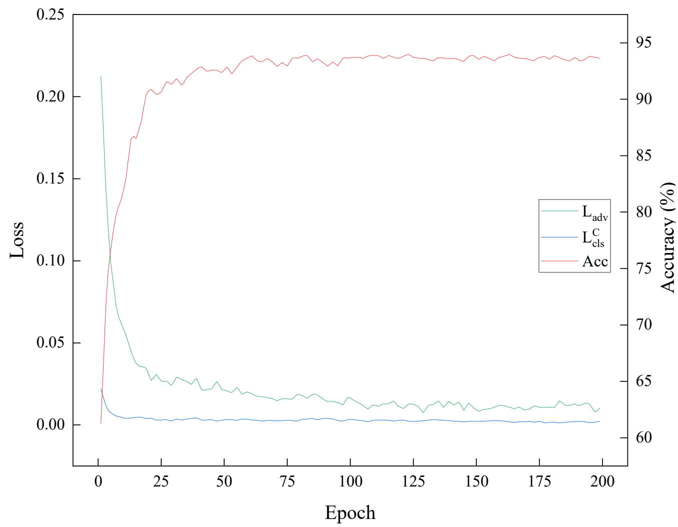
$$L_{total} = L_{cls}^{C} + \alpha L_{adv} \tag{27}$$

where α is the loss coefficient, this method constrains the training process through this objective function to find the optimal parameters of the model.

6 Experimental results and analyses

The system software platform used for the experiments in this paper is CentOS 7.6, Python version 3.8, cudatoolkit version 11.6, and the deep learning framework is Pytorch 1.12. The system hardware platform is NVIDIA RTX 3090, and the CPU is Xeon(R) Gold 6226R. The popular multimodal English corpus Spoken English Corpus (SEC) was used as the experimental dataset, which contains 31 text categories in 11 domains, totaling 14,792 audio and text data. A 10-fold cross-validation with randomised scores is used on the SEC dataset with a 9:1 ratio of training set to test set. The model uses a maximum frame length of 500 for the speech Mel spectrum, the maximum word length of glove is set to 100, and each word is represented by a 300-dimensional vector. When training the model, a mini-batch stochastic gradient descent optimiser algorithm is used, with the learning rate set to 0.0001 and the batch size to 100.

Figure 5 The loss function and target domain recognition accuracy of GUDA (see online version for colours)

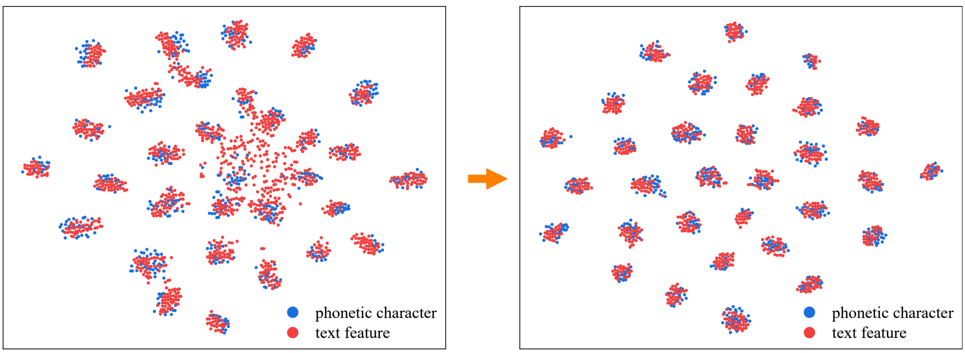


The proposed recognition model is denoted as GUDA, and the variation of the loss function and target domain recognition accuracy of GUDA with the training process is shown in Figure 5, where the green line denotes the adversarial loss L_{adv}, the blue line denotes the classification loss L_{cls}^{C}, and the red line denotes the recognition accuracy Acc. L_{cls}^{C}, declines faster than L_{adv}. This is due to the fact that it is easier to bring two different domains of data closer together, and more difficult to obtain clear classification boundaries. On the other hand, it can be seen that L_{cls}^{C}, stabilises and the classification accuracy further increases as L_{adv} decreases, indicating that AT plays an important role in the model convergence process.

To more intuitively see the changes in the feature vectors of the proposed method after UDA, this chapter takes the digital dataset as an example for visualisation experiments, as shown in Figure 6. Each point in the figure represents the output feature vector of a sample data after the feature extractor, and each colour represents a category, the left side is the visualisation of different categories of samples in the target domain before UDA, and the right side is the visualisation of different samples after UDA. The distance between the different categories of the pre-UDA target domain data is small and difficult to recognise. After UDA, the data of the same kind in the target domain are more concentrated, and the distance between different kinds of data increases and the boundaries are clearer, which makes it easier for the classifier to realise the classification of the data in the target domain, and thus obtains a higher recognition accuracy.

To further measure the recognition performance of GUDA, WA, unweighted accuracy (UA), F1, and mean absolute error (MAE) are used to compare the recognition performance of GUDA, RCNN (Zhong et al., 2019), CNN-RNN (Song, 2020), CBiLSTM (Liu et al., 2023), FMUDA (Diao and Hu, 2021 for comparison experiments and the results are shown in Table 1. The WA and UA of GUDA are 93.67% and 90.55%, respectively, which are at least 2.75% and 2.42% higher compared to the other four models, respectively. Comparing the reconciled mean F1 of recall and precision again, GUDA reaches 91.75%, and both CBiLSTM and FMUDA are above 85%, with all three models showing better recognition performance. The F1 of CNN-RNN is 83.98 and the recognition performance is average. The F1 value of RCNN is only 78.54% and the recognition performance is the worst.

Figure 6 Characteristic distribution results before and after UDA (see online version for colours)



Then comparing the recognition accuracy index MAE, the MAE of GUDA is 0.1714, which is at least 24.09% lower compared to the other four models. The RCNN model only considers the text features of a single modality and does not enhance the important features, resulting in the lowest recognition accuracy. Although CNN-RNN mines spatio-temporal features of speech, it does not consider multimodal features, which leads to incomplete mining of features. CBiLSTM considers multimodal features, but does not investigate the modal variability of text and speech. FMUDA uses the UDA method of feature matching to align the features of text and speech, and achieves better recognition results, but does not consider the domain deviation, and the recognition performance is not as good as that of GUDA. GUDA not only comprehensively considers multimodal features, but also improves the recognition effect by adding domain modality-specific markers to each corpus sample, which makes the AT process of the feature extractor and domain discriminator smoother.

Table 1 Comparison of recognition performance metrics

Model	*WA/%*	*UA/%*	*F1/%*	*MAE*
RCNN	80.39	76.94	78.54	0.3815
CNN-RNN	84.86	81.21	83.98	0.3129
CBiLSTM	88.15	87.24	85.33	0.2516
FMUDA	90.92	88.13	88.69	0.2258
GUDA	93.67	90.55	91.75	0.1714

7 Conclusions

With the increasing reliance on corpora for English learning, the accuracy of multimodal English corpus text recognition becomes more and more important. To solve the issues of data sparsity and domain shift in existing studies, this paper proposes a multimodal English corpus text recognition model based on UDA, and the main work is summarised as bellow.

1. Glove algorithm and MFCC are used to pre-process the text and speech data respectively, and BiLSTM and SA are used to extract text features with high contribution to the text vectors output from the Glove model; CNN, BiLSTM and SA are used to extract speech features with high contribution to the Mel spectrum.
2. The text and speech modal features are aggregated and represented by modelling through GNN, and the category prototype and domain prototype are obtained through calculation as the node representation of each mode on the graph. On this basis, a bipart graph is constructed with training samples and knowledge transfer is carried out to extract domain invariant features containing inter-domain interaction information.
3. UDA is used to reduce the difficulty of domain adaptation with great differences between domains, and the domain information in semantic features is extracted into domain modal specific tags. Modal specific tags are more likely to fool the domain discriminator because of the alignment of the domain discriminator information. At the later stage of training, the purified semantic features are obtained for recognition.
4. The experimental outcome implies that the WA of the proposed model is 93.67%, which is better than the benchmark model, and significant performance improvement is achieved in the multimodal English corpus text recognition task.

Declarations

All authors declare that they have no conflicts of interest.

References

Beavis, C. (2013) 'Literary English and the challenge of multimodality', *Changing English*, Vol. 20, No. 3, pp.241–252.

Bessadok, A., Mahjoub, M.A. and Rekik, I. (2022) 'Graph neural networks in network neuroscience', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 45, No. 5, pp.5833–5848.

Coccetta, F. (2018) 'Developing university students' multimodal communicative competence: field research into multimodal text studies in English', *System*, Vol. 77, pp.19–27.

Diao, L. and Hu, P. (2021) 'Deep learning and multimodal target recognition of complex and ambiguous words in automated English learning system', *Journal of Intelligent and Fuzzy Systems*, Vol. 40, No. 4, pp.7147–7158.

Ding, X., Shi, Q., Cai, B., Liu, T., Zhao, Y. and Ye, Q. (2019) 'Learning multi-domain adversarial neural networks for text classification', *IEEE Access*, Vol. 7, pp.40323–40332.

Ivanko, D., Karpov, A., Fedotov, D., Kipyatkova, I., Ryumin, D., Ivanko, D., Minker, W. and Zelezny, M. (2018) 'Multimodal speech recognition: increasing accuracy using high speed video data', *Journal on Multimodal User Interfaces*, Vol. 12, pp.319–328.

Leng, Y., Zhou, N., Sun, C., Xu, X., Yuan, Q., Cheng, C., Liu, Y. and Li, D. (2017) 'Audio scene recognition based on audio events and topic model', *Knowledge-Based Systems*, Vol. 125, pp.1–12.

Li, H. (2020) 'Text recognition and classification of English teaching content based on SVM', *Journal of Intelligent and Fuzzy Systems*, Vol. 39, No. 2, pp.1757–1767.

Liu, L. and Tsai, S-B. (2021) 'Intelligent recognition and teaching of English fuzzy texts based on fuzzy computing and big data', *Wireless Communications and Mobile Computing*, Vol. 14, No. 2, pp. 1–10.

Liu, X., Wei, F., Jiang, W., Zheng, Q., Qiao, Y., Liu, J., Niu, L., Chen, Z. and Dong, H. (2023) 'MTR-SAM: visual multimodal text recognition and sentiment analysis in public opinion analysis on the internet', *Applied Sciences*, Vol. 13, No. 12, p.7307.

Liu, X., Yoo, C., Xing, F., Oh, H., El Fakhri, G., Kang, J-W. and Woo, J. (2022) 'Deep unsupervised domain adaptation: a review of recent advances and perspectives', *APSIPA Transactions on Signal and Information Processing*, Vol. 11, No. 1, pp.5–13.

Ma, J., Shao, W., Ye, H., Wang, L., Wang, H., Zheng, Y. and Xue, X. (2018) 'Arbitrary-oriented scene text detection via rotation proposals', *IEEE Transactions on Multimedia*, Vol. 20, No. 11, pp.3111–3122.

Maroof, R., Usmani, I.A. and Feroze, A. (2024) 'English text recognition based on convolutional neural network (CNN)', *Sir Syed University Research Journal of Engineering and Technology*, Vol. 14, No. 2, pp.73–78.

Mirzaei, A., Azizi Farsani, M. and Chang, H. (2023) 'Statistical learning of L2 lexical bundles through unimodal, bimodal, and multimodal stimuli', *Language Teaching Research*, Vol. 9, pp.11–17.

Nema, B.M. and Abdul-Kareem, A.A. (2018) 'Preprocessing signal for speech emotion recognition', *Al-Mustansiriyah Journal of Science*, Vol. 28, No. 3, pp.157–165.

Sicilia, A., Zhao, X. and Hwang, S.J. (2023) 'Domain adversarial neural networks for domain generalization: when it works and how to improve', *Machine Learning*, Vol. 112, No. 7, pp.2685–2721.

Singh, P., Srivastava, R., Rana, K. and Kumar, V. (2021) 'A multimodal hierarchical approach to speech emotion recognition from audio and text', *Knowledge-Based Systems*, Vol. 229, p.107316.

Song, Z. (2020) 'English speech recognition based on deep learning with multiple features', *Computing*, Vol. 102, No. 3, pp.663–682.

Stein, R.A., Jaques, P.A. and Valiati, J.F. (2019) 'An analysis of hierarchical text classification using word embeddings', *Information Sciences*, Vol. 471, pp.216–232.

Tu, H. (2021) 'A study on the construction of emotion recognition based on multimodal information fusion in English learning cooperative and competitive mode', *Frontiers in Sychology*, Vol. 12, p.767844.

Zhang, W., Yoshida, T. and Tang, X. (2008) 'Text classification based on multi-word with support vector machine', *Knowledge-Based Systems*, Vol. 21, No. 8, pp.879–886.

Zhang, Z., Shao, M., Ma, C., Lv, Z. and Zhou, J. (2022) 'An enhanced domain-adversarial neural networks for intelligent cross-domain fault diagnosis of rotating machinery', *Nonlinear Dynamics*, Vol. 108, No. 3, pp.2385–2404.

Zhao, W., Alwidian, S. and Mahmoud, Q.H. (2022) 'Adversarial training methods for deep learning: a systematic review', *Algorithms*, Vol. 15, No. 8, p.283.

Zhong, Z., Sun, L. and Huo, Q. (2019) 'An anchor-free region proposal network for Faster R-CNN-based text detection approaches', *International Journal on Document Analysis and Recognition* (*IJDAR*), Vol. 22, pp.315–327.