



International Journal of Information and Communication Technology

ISSN online: 1741-8070 - ISSN print: 1466-6642

<https://www.inderscience.com/ijict>

An automatic English pronunciation scoring model using GAN-enhanced synthetic data and active learning

Xiaoyong Long, Liying Zhang

DOI: [10.1504/IJICT.2026.10076663](https://doi.org/10.1504/IJICT.2026.10076663)

Article History:

Received:	15 September 2025
Last revised:	09 December 2025
Accepted:	13 December 2025
Published online:	13 March 2026

An automatic English pronunciation scoring model using GAN-enhanced synthetic data and active learning

Xiaoyong Long and Liying Zhang*

Chengdu College of Arts and Sciences,
Chengdu, 610000, China

Email: longxiaoyong@cdcas.edu.cn

Email: zhangliying@cdcas.edu.cn

*Corresponding author

Abstract: To address the critical challenges of limited labelled data and high annotation costs in automatic pronunciation assessment, this study proposes a novel framework that integrates generative adversarial network-based synthetic data augmentation with active learning. A conditional generative adversarial network is employed to generate synthetic speech samples with controlled phonemic and articulatory features, while a hybrid active learning strategy combining uncertainty and diversity criteria is designed to select informative samples for expert annotation. Evaluation on the L2-ARCTIC dataset demonstrates that the proposed approach achieves a Pearson correlation coefficient of 0.843, outperforming the best baseline (0.801) by 5.2%. It also reduces root mean square error and mean absolute error by 8.0% and 9.8%, respectively. The results highlight the synergistic effect of generative and selective data strategies, offering an effective solution for low-resource pronunciation scoring.

Keywords: automatic pronunciation scoring; adversarial network generation; active learning; data enhancement; English pronunciation.

Reference to this paper should be made as follows: Long, X. and Zhang, L. (2026) 'An automatic English pronunciation scoring model using GAN-enhanced synthetic data and active learning', *Int. J. Information and Communication Technology*, Vol. 27, No. 19, pp.51–69.

Biographical notes: Xiaoyong Long received her Master's degree from Sichuan International Studies University in 2010. She is currently a Lecturer at Chengdu College of Arts and Sciences. Her research interests include English language, literature education and cultural translation.

Liying Zhang received her Master's degree from Sichuan Normal University in 2015. She is currently an Associate Professor at Chengdu College of Arts and Sciences. Her research interests include AI-assisted education, active learning, data enhancement and automatic pronunciation scoring.

1 Introduction

With accelerating globalisation and the wide application of information technology in education, the status of English as a global lingua franca is becoming more and more prominent, leading to surging global demand for efficient and accurate spoken English learning tools (House, 2003). Traditional English pronunciation teaching relies heavily on

manual assessment by professional teachers, which is not only time-consuming and costly, but also difficult to ensure the consistency of evaluation standards, and fails to meet the urgent demand for large-scale, personalised language learning (Amrate and Tsai, 2024). Therefore, the development of automatic, objective, and scalable automatic pronunciation assessment (APA) systems has become a key research topic at the intersection of educational technology and computational linguistics, which is of great practical significance for promoting educational equity and realising personalised tutoring (Li et al., 2016).

In recent years, thanks to the rapid development of deep learning techniques, especially the successful application of recurrent neural networks (RNN) and transformer architecture, the performance of automatic pronunciation scoring systems has been significantly improved (Cheng and Wang, 2022). Early research focused on the hidden Markov model (HMM)-based goodness of pronunciation (GOP) algorithm, which evaluates the quality of pronunciation by comparing the likelihood of a non-native speaker's speech with that of a native speaker's acoustic model (Witt and Young, 2000). Current state-of-the-art approaches typically use end-to-end deep neural networks (DNNs) that learn complex mapping relationships directly from raw audio or acoustic features (e.g., Mel spectrograms) to predict pronunciation scores at the phoneme level, word level, or sentence level (Zahran et al., 2023). These deep models are able to capture more subtle acoustic differences and suprasegmental features (e.g., stress, rhythm, and intonation), thus providing finer-grained and more accurate assessments than traditional methods (Yan and Chen, 2024; Chen, 2015). For example, Zhang (2025) utilised DNNs to fuse multiple acoustic features to significantly improve the correlation between ratings and manual evaluations.

However, despite the great potential of deep modelling, its success relies heavily on large-scale, high-quality, and expertly fine-labelled speech data (Brahmi et al., 2024). The scarcity of data and the cost of labelling constitute the core bottlenecks facing current APA research. First, it is extremely costly to construct a database that covers multi-native language backgrounds, multi-pronunciation error patterns with consistent annotation, which makes most of the researches to be conducted only on limited datasets and limits the generalisation ability of the models (Xu et al., 2021). Second, while existing data enhancement techniques such as velocity perturbation, volume adjustment, and spectral masking (SpecAugment) can improve model robustness to some extent, these methods are essentially low-level transformations on top of existing samples, and are unable to create entirely new samples that are diverse at both the level of phoneme content and articulatory quality (Park et al., 2019). This means that the model struggles to learn unseen patterns of mispronunciation or accent characteristics, thus limiting its performance in realistically diverse environments.

In order to break through the above data dilemma, this study explores the deep fusion of two highly promising directions: generative adversarial networks (GANs) for synthetic data augmentation and active learning (AL). GANs (Goodfellow et al., 2014) have demonstrated powerful data generation capabilities in image and speech synthesis. In speech processing, GAN-based speech conversion and synthesis techniques (e.g., MelGAN, Kumar et al., 2019) are capable of generating highly realistic speech waveforms. If applied to APA, it is expected to create synthesised speech samples with specified phonemes and specified articulatory quality levels (e.g., 'good' or 'bad') by means of conditional generation, thus greatly enriching the diversity of the training data, in particular, complementing those mispronunciation patterns that are rare in small real

datasets. On the other hand, AL (Settles, 2009), as a classical machine learning paradigm, is centred on the idea of allowing the model to autonomously select the most informative unlabelled samples for querying, so as to maximise the performance of the model with minimal labelling cost. In the context of APA, this means that pronunciation samples that are the most uncertain to the model, or the most representative of unknown error patterns, can be prioritised and handed over to experts for labelling, thus allowing each labelling input to yield the highest marginal benefit (Rogti, 2025). Although the two have their own merits in data generation and sample selection, respectively, most of the existing research work treats them as independent solutions. We argue that GAN and AL are not substitutes for each other, but rather complementary to each other (Xiang and Bao, 2020): GANs can create large amounts of diverse training data ‘out of thin air’, alleviating the cold-start problem and providing an informative pool of choices for AL (Chatziagapi et al., 2019); whereas ALs can intelligently guide us to invest costly manual annotation resources on real samples that are the most challenging even for powerful generative models, thus targeting the gap between generated data and real data distributions that may exist. Models the most challenging real samples, thus targeting the gap that may exist between generated and real data distributions.

Based on this, this paper aims to propose a unified framework to systematically integrate GAN-based synthetic data enhancement with AL loops to synergistically address the data scarcity and labelling cost issues in APA. We hypothesise that such a fusion strategy can simultaneously utilise the ‘breadth’ of synthetic data and the ‘depth’ of active selection data to jointly forge a more powerful and efficient automatic pronunciation scoring model.

2 Related work

2.1 Development and evolution of automated pronunciation scoring models

APA technology centres on how to allow machines to simulate human experts to effectively assess the pronunciation accuracy and intelligibility of non-native speakers. Research in this field has evolved remarkably from traditional methods relying on hand-designed features to data-driven deep learning models. The most influential early approach was the GOP algorithm based on HMM (Witt and Young, 2000). GOP computes the likelihood of a given speech segment with respect to the HMM of a target phoneme and compares it to the likelihoods of all the other phonemes to obtain a plausibility score. Although GOP and its subsequent improved versions (e.g., extended GOP) are effective in specific scenarios, their performance relies heavily on the accuracy of the native speaker acoustic model shared with the speech recognition system and struggles to capture suprasegmental features and complex patterns of articulatory errors.

In recent years, deep learning has revolutionised the research paradigm in APA. Researchers have begun to employ various neural network architectures, such as DNNs, convolutional neural networks (CNNs), and RNNs, to learn more discriminative features directly from raw data in an end-to-end manner. For example, Yan and Chen (2024) pointed out that APA can quantify the target language pronunciation level of second language learners from multi-dimensional and multi-lingual hierarchies, while the parallel modelling frameworks in the existing studies do not take into account the hierarchical structure of linguistic units and the relevance of the pronunciation aspect,

and therefore proposed a hierarchical graph attention transformer-based (HierGAT) method, which models the input utterance. This method models the input utterance as a heterogeneous graph containing language nodes with different granularity, and optimises the evaluation through the hierarchical graph messaging mechanism and aspectual attention module. The team introduces multilingual nodes into the graph neural network for the first time to be used in APA and to do the qualitative analysis, and the experiments on Speech Ocean's 762 benchmark dataset prove the feasibility and effectiveness of this method.

Chen (2015) highlighted that multi-task learning (MTL) improves performance by sharing internal representations and is more effective when the training data is small, and the DNN hidden layer provides an ideal internal representation for sharing knowledge for acoustic modelling of automatic speech recognition (ASR). The study proposed three methods for applying MTL to acoustic modelling of DNNs without additional linguistic resources (training phoneme- and grapheme-acoustic models with the same network, training different triphoneme states with a small bound state set, and combining universal phoneme set and language-specific triphoneme modelling in a multilingual environment). The three methods for applying MTL to DNN acoustic modelling without additional linguistic resources (training phoneme and grapheme acoustic models with the same network, training different triphone states with a small set of bound states, and combining a common set of phonemes with language-specific triphones for modelling in a multilingual environment, respectively) are valid on multiple datasets, and heuristic guidelines are given in the paper for finding useful secondary tasks.

In particular, Zhang (2025) points out that spoken pronunciation is crucial for effective communication, and aims to construct a deep learning-based model for evaluating English spoken pronunciation, proposing a novel dynamic Bayesian optimisation (DBO)-conditional deep belief network (CDBN) algorithm, which is trained with an English speech dataset containing correct and incorrect pronunciation samples, and then fine-tuned with the DBO algorithm, after pre-processing by normalising the data, extracting the features by the gammatone frequency cepstral coefficients (GFCC) algorithm, and fine-tuning the CDBN architecture by the DBO algorithm. After the CDBN architecture, the model was implemented in Python and performed well in terms of precision (96.78%), recall (93.65%) and F1-score (94.53%), and comparative analyses also verified the model validity, which significantly improved the correlation between the ratings and the manual evaluations. While these deep models have dramatically improved their performance, they are inherently data hungry and their success is heavily dependent on the existence of large-scale, high-quality labelled data, which directly directs the research focus to how to solve the data scarcity problem.

2.2 *Speech data enhancement technology*

To alleviate the data scarcity problem, data augmentation has become a standard process for training robust speech processing models. Traditional data augmentation methods focus on artificially extending the dataset by performing simple transformations on the original audio signal, such as adding random noise, changing the speech rate, and adjusting the pitch or volume (Ko et al., 2015). These methods improve the robustness of the model to some extent, but they only operate on low-level acoustic signals and do not create new semantic or linguistic content.

More advanced augmentation techniques attempt to operate in feature space. The SpecAugment method proposed by Park et al. (2019) applies masking operations directly to the Mayer spectrogram, forcing the model to learn more robust representations by masking out the continuous parts in the time and frequency dimensions. This approach has been highly successful on ASR tasks and has been borrowed for APA tasks. However, the augmented samples of these approaches, both at the signal level and the feature level, are still highly correlated with the original data and do not allow for the generation of completely new samples with true diversity in both articulatory content and quality.

To this end, generative models, in particular GANs, show greater potential. GANs are able to learn high-dimensional distributions of real data through adversarial training between generators and discriminators, thus generating new samples with a high degree of fidelity (Goodfellow et al., 2014). In the field of speech synthesis, GAN-based models (e.g., MelGAN, Kumar et al., 2019) have been able to generate high-quality raw waveforms. In order to solve the problem of mismatch between training and test data in noise robust speech recognition, Qian et al. (2019) proposes to utilise GANs for data augmentation (instead of traditional noise addition to raw waveforms), firstly, we study the basic GAN (frame-by-frame generation of unlabelled speech samples based on spectral features, paired with an unsupervised learning framework for acoustic modelling), and then we introduce the conditional information to propose the conditional GAN (cGAN) (generating real labelled samples conditional on the acoustic state of speech frames or the original paired clean speech), and combining real labels and soft labels in acoustic model training; this method is validated in Aurora4 (simulated noise and channel distortion) and Audio-Visual Meeting Corpus (AMI) conference transcription (real reverberation) tasks, and augmenting the data with the basic GAN/cGAN augmentation of data on advanced acoustic models results in a relative reduction in word error rate (WER) of 6% to 14%, and improved performance in all types of noise conditions. It is still an under-explored research direction to construct a cGAN that can generate synthesised speech based on specified phonemes and preset articulatory quality levels (e.g., ‘good’ or ‘bad’) as augmentation data for the APA system. This is a direction that has not been explored enough. This is a key innovation in the field of APA that goes beyond existing enhancement techniques.

2.3 AL and its application to speech processing

AL is a paradigm in machine learning that aims to minimise the cost of annotation, with the core idea being that learning algorithms actively select the ‘most informative’ unlabelled samples to be annotated by experts rather than randomly selecting them (Settles, 2009). Common query strategies include uncertainty sampling (selecting the sample with the most uncertain model predictions), committee querying (using multiple models and selecting the most controversial sample), and diversity-based sampling (selecting the sample that best represents the entire distribution of unlabelled data).

In the field of speech processing, AL has been applied to tasks such as speech recognition, speech emotion recognition, and speaker identification. For example, Riccardi and Hakkani-Tur (2005) earlier explored the use of AL in speech recognition to reduce transcription workload. By focusing on the application of adaptive learning in ASR), an ASR AL algorithm is proposed, which estimates the utterance confidence

scores online from the speech recogniser output grid, and selects the optimal subset of training samples after filtering by an informativeness function, aiming to reduce the number of manually-supervised and manually-labelled samples required for training ASR models, and which is applied in both batch and online learning scenarios; experiments show that this AL algorithm can reduce the amount of labelled data required to achieve word-specific accuracy by more than 60% compared to random sampling. Abdelwahab and Busso (2019) pointed out that although DNNs drive the performance improvement of speech emotion recognition, they require a large amount of labelled data, and the annotation cost for speech emotion recognition is high, the existing corpus is limited in size, and the model generalisation ability is weak; therefore, we explore a scheme to train DNNs for speech emotion recognition with AL under limited resources, assuming that new domains of unlabelled data are available and that a small amount of labelled data can be selected to achieve specific word accuracy. Assuming the availability of unlabelled data in new domains and the ability to select a small number of recordings for labelling, greedy sampling (GS) and uncertainty-based methods are used to select samples, which are evaluated in a regression problem for predicting arousal and validity scores, and the results demonstrate that AL can achieve competitive performance with limited training data.

However, relatively little research has been conducted on applying AL to automatic pronunciation scoring. The APA task is unique in that it is extremely costly to annotate (requiring linguistic expertise) and has an imbalanced data distribution (samples with severe pronunciation errors are usually in the minority). This makes AL of great value for applications in this domain: annotation efficiency and model performance can be greatly improved by intelligently querying those pronunciation samples whose models are most uncertain, or most representative of rare error patterns. Most of the existing APA systems rely on random annotation or full-volume annotation, failing to leverage AL's intelligent query mechanism to optimise the annotation process. Therefore, systematically introducing AL into the APA framework and designing a query strategy compatible with it is an important step to fill the current research gap.

In summary, existing research has made progress on the respective paths of APA modelling, data augmentation, and AL, but no work has yet been done to deeply integrate these three ideas into a synergistic framework. This study aims to bridge this gap by exploring the joint benefits of GAN synthetic data augmentation and AL in improving the performance of automatic pronunciation scoring models.

3 Methodology

In this section, the proposed APA model fusing GAN synthetic data augmentation with AL is elaborated. The framework is a well-designed iterative optimisation system, whose core idea is to synergistically utilise the breadth of data distribution brought by generative data augmentation and the depth of expert annotation guided by AL, aiming to significantly improve the generalisation ability and final performance of pronunciation scoring models at the most cost-effective annotation cost.

3.1 Problem definition and overall framework

We begin with a formal definition of the automatic pronunciation scoring task. Let the pool of unlabelled raw speech data:

$$\{(\mathbf{o}_i)\}_{i=1}^N \quad (1)$$

where $\mathbf{o}_i \in \mathbb{R}^{T_i \times D}$ represents the sequence of acoustic features for the i^{th} speech segment (e.g., T_i is the number of frames in the segment, and D is the number of dimensions in the Mel spectrum). Its corresponding, unknown true labelling is:

$$\{(y_i, \mathbf{p}_i)\}_{i=1}^N \quad (2)$$

where $y_i \in \mathbb{R}$ is the global articulatory score (or can be defined as a vector of phoneme-level scores) for the speech segment, and $\mathbf{p}_i$ is the corresponding phoneme sequence transcription. When the system is initialised, we randomly select a small subset $\mathcal{L}_0$ ($|\mathcal{L}_0| = M, M \ll N$) from $\mathcal{U}$ for costly expert labelling to obtain the initial labelled training set:

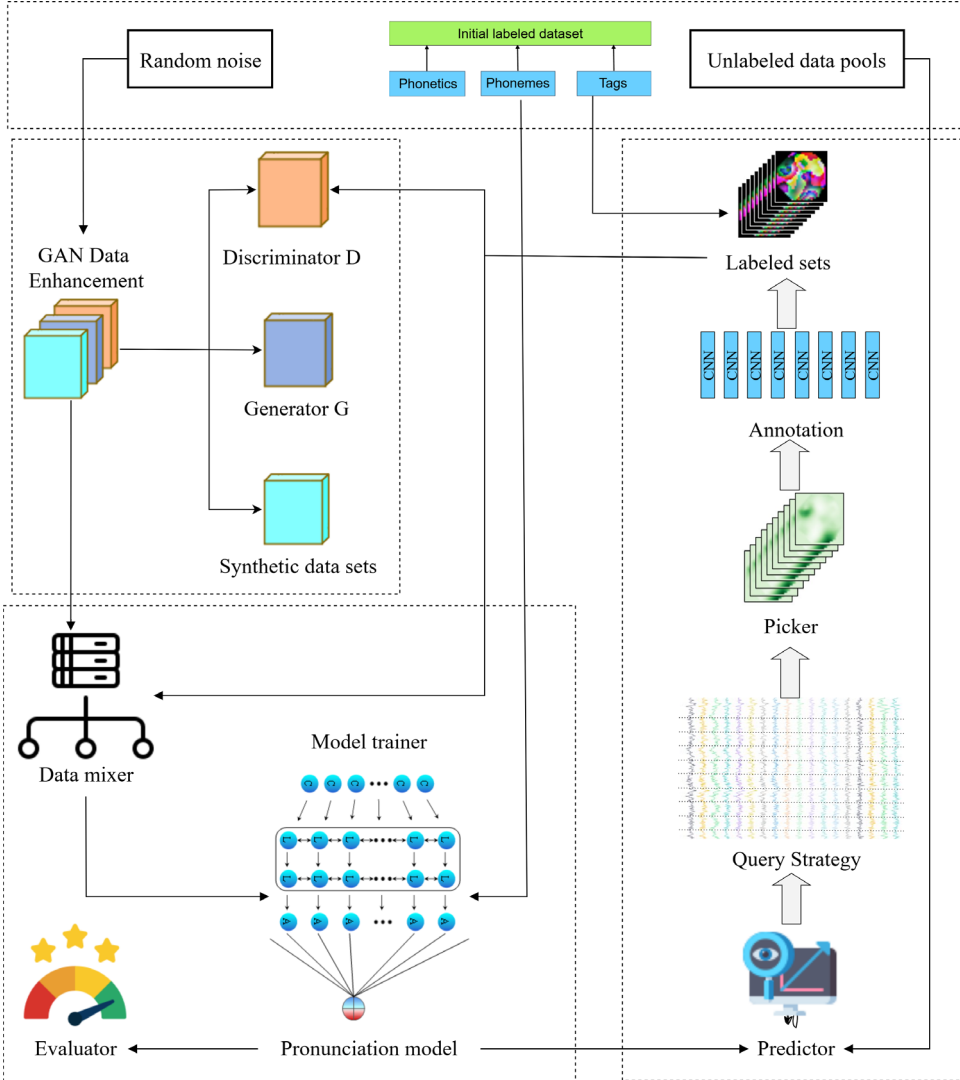
$$\mathcal{L}_0 = \{(\mathbf{o}_j, y_j, \mathbf{p}_j)\}_{j=1}^M \quad (3)$$

Figure 1 clearly depicts the overall workflow of the fusion framework proposed in this paper, which mainly consists of two mutually coupled, co-evolving core loops:

- **GAN synthesis data enhancement loop:** A cGAN is trained using the initial or latest augmented labelling set $\mathcal{L}_t$ (t indexed by iteration rounds). The generator of this cGAN is able to synthesise realistic speech features $\tilde{\mathbf{o}}$ and their corresponding pseudo-labels $(\tilde{y}, \tilde{\mathbf{p}})$ from random noise conditional on a specified sequence of phonemes $\mathbf{p}$ and a predefined level of articulatory quality y , all of which synthetic samples constitute a large-scale synthetic dataset $\mathcal{G}$.
- **AL loop:** the current pronunciation scoring model f_θ is trained using the mixed dataset $\mathcal{L}_t \cup \mathcal{G}$. After training, the model predicts all samples in the current unlabelled pool $\mathcal{U} \setminus \mathcal{L}_t$. Subsequently, the system intelligently selects the most informative samples $\mathbf{g}$ from the large number of unlabelled samples based on a well-designed query strategy Q and submits them to a human expert for labelling. The newly acquired precious labelled data is then expanded into the training set:

$$\mathcal{L}_{t+1} = \mathcal{L}_t \cup \mathcal{Q} \quad (4)$$

The above two processes are not isolated, but form an active feedback loop. In each round of AL iterations, the GAN generator can be retrained or fine-tuned with the latest augmented and higher quality labelled set $\mathcal{L}_t$ to generate more realistic and diverse synthetic data $\mathcal{G}$. Meanwhile, the pronunciation scoring model f_θ is updated on the continuously growing hybrid dataset $\mathcal{L}_t \cup \mathcal{G}$, and its performance is thus evolved. This synergistic mechanism effectively breaks the deadlock between data scarcity and labelling cost.

Figure 1 A pronunciation scoring modelling framework incorporating GAN synthetic data enhancement and AL (see online version for colours)

3.2 Baseline articulation score model

We adopt an end-to-end DNN architecture as the baseline pronunciation scoring model f_θ . The model takes as input a variable-length sequence of speech acoustic features (e.g., an 80-dimensional Meier spectrogram, $\mathbf{o} \in \mathbb{R}^{T \times 80}$), and ultimately outputs a scalar score characterising its overall articulatory quality, $\hat{y}$.

At the core of the model is a multilayer bidirectional long short-term memory network (Bi-LSTM) encoder, which is able to efficiently capture the complex long-term contextual dependencies and temporal dynamics in speech signals:

$$\mathbf{h}_t = \text{Bi-LSTM}(\mathbf{o}_t, \mathbf{h}_{t-1}; \Theta_{\text{enc}}) \quad (5)$$

where $\mathbf{h}_t \in \mathbb{R}^H$ is the hidden state vector at time step t (H is the hidden layer dimension) and Θ_{enc} represents the parameters of the encoder.

Subsequently, we employ an improved GOP algorithm based on the posterior probability of phonemes as the core scoring mechanism of the model. For each phoneme segment $\mathbf{o}_i$ obtained by forced alignment (whose corresponding target phoneme is p_i and the number of sustained frames is T_i), we compute its GOP score:

$$\text{GOP}(p_i) = \frac{1}{T_i} \left(\log P(p_i | \mathbf{o}_i) - \frac{1}{|M|} \sum_{m \in M} \log P(m | \mathbf{o}_i) \right) \quad (6)$$

where $P(p_i | \mathbf{o}_i)$ is the model's predicted posterior probability that a given acoustic segment $\mathbf{o}_i$ belongs to the target phoneme p_i , M is the set of phonemes, and $|M|$ is the total number of phonemes. The second term $\frac{1}{|M|} \sum_{m \in M} \log P(m | \mathbf{o}_i)$ in equation, s represents the model's

average uncertainty about the phoneme to which the segment belongs, as a baseline. Thus, $\text{GOP}(p_i)$ effectively measures how far the confidence of the current pronunciation relative to the target phoneme deviates from its average uncertainty, with higher values indicating a more accurate pronunciation, closer to native speaker level.

Ultimately, the global articulatory score s for the entire utterance can be obtained by aggregating (e.g., averaging) the GOP scores for all phonemes and constraining the scores to the $[0, 1]$ interval via a sigmoid function, $\sigma(\cdot)$, to facilitate training and interpretation:

$$\hat{y} = \sigma \left(\frac{1}{N_p} \sum_{i=1}^{N_p} \text{GOP}(p_i) \right) \quad (7)$$

where N_p is the total number of phonemes contained in the utterance. The full parameters of the model θ are optimised by minimising the mean square error (MSE) between the predicted score $\hat{y}$ and the true expert rating y :

$$\mathcal{L}_{\text{score}} = \frac{1}{B} \sum_{b=1}^B (\hat{y}_b - y_b)^2 \quad (8)$$

where B represents the number of samples in the training batch.

3.3 CGAN-based synthetic data enhancement

In order to break through the limitation of the size of real labelled data, we design a cGAN to generate high-quality and controlled synthetic speech data. The main function of the generator G is to map a random noise vector $\mathbf{z} \sim p_z(\mathbf{z})$ sampled from a priori distribution (usually a standard Gaussian distribution) into a segment of realistic speech features $\tilde{\mathbf{o}}$, while the whole generation process is precisely regulated by conditional information $\mathbf{c}$.

The conditional information $\mathbf{c}$ is the joint embedding representation of the target phoneme sequence $\mathbf{p}$ and the desired global articulatory score y , i.e.,

$\mathbf{c} = \text{Embedding}(\mathbf{p}, y)$. This design allows the researcher to precisely specify the linguistic content of the generated sample and its level of articulatory quality (e.g., to generate a ‘ship’ with ‘good’ articulation or a ‘sheep’ with ‘poor’ articulation).

The generator G and the discriminator D are trained adversarial by the following min-max game:

$$\min_G \max_D V(D, G) = \mathbb{E}_{\mathbf{o}, \mathbf{c} \sim p_{\text{data}}} [\log D(\mathbf{o}|\mathbf{c})] + \mathbb{E}_{\mathbf{z} \sim p_z, \mathbf{c} \sim p_{\text{data}}} [\log(1 - D(G(\mathbf{z}|\mathbf{c})))] \quad (9)$$

where $D(\cdot|\mathbf{c})$ is a conditional discriminator whose task is to determine whether the input speech features are from the true distribution ($\mathbf{o}$) or synthesised by the generator ($G(\mathbf{z}|\mathbf{c})$) under a given condition $\mathbf{c}$.

Inspired by the work of Zhu et al. (2017), in order to further improve the quality of the generated samples and their consistency with the conditional information, we introduce a cyclic consistency loss in addition to the standard adversarial loss. Specifically, we require that the generated sample $\tilde{\mathbf{o}} = G(\mathbf{z}|\mathbf{c})$ can be accurately reduced back to its conditional information $\tilde{\mathbf{c}} = E(\tilde{\mathbf{o}})$ by a pre-trained auxiliary encoder E . This reconstruction loss function is defined as:

$$\mathcal{L}_{\text{recon}} = \|\mathbf{c} - \tilde{\mathbf{c}}\|_1 \quad (10)$$

Thus, the overall objective function of the generator is the weighted sum of the adversarial loss and the reconstruction loss:

$$\mathcal{L}_G = -\mathbb{E}_{\mathbf{z}, \mathbf{c}} [\log D(G(\mathbf{z}|\mathbf{c}))] + \lambda \mathcal{L}_{\text{recon}} \quad (11)$$

where λ is a hyperparameter used to balance the two loss contributions. With this training approach, the generator is eventually able to produce high-quality synthetic data that highly matches the given conditions $\mathcal{G} = \{(\tilde{\mathbf{o}}_k, \tilde{y}_k, \tilde{\mathbf{p}}_k)\}_{k=1}^K$. These synthetic data, when mixed with the real labelled data, greatly enrich the training set of the scoring model.

3.4 AL cycle

The AL loop is the core of the entire framework for implementing intelligent annotation. In each iteration t , we train a pronunciation scoring model f_{θ_t} using the currently available mixed dataset $\mathcal{L}_t \cup \mathcal{G}$. The model is then used to predict all the samples in the currently unlabelled pool $\mathcal{U} \setminus \mathcal{L}_t$.

We design a hybrid query strategy Q that incorporates uncertainty sampling and diversity sampling to select the most valuable samples for expert labelling.

For each unlabelled sample $\mathbf{o}_i$, we first measure the cognitive uncertainty of the model’s prediction about it. We use Monte Carlo dropout (MC dropout) (Gal and Ghahramani, 2016) for Bayesian approximation. MC dropout was chosen for its combination of simplicity of implementation and theoretical rigor. From a practical point of view, MC dropout does not require any changes to the existing model architecture or training process, and only needs to turn on the dropout several times during testing and perform forward propagation, using the variance of the prediction results as a measure of uncertainty, which is an improvement with almost no additional cost. From a theoretical point of view, it provides a practical and efficient Bayesian approximation of the dropout-trained network as a Bayesian neural network, which allows for sampling the

model weights and estimating the predictive distribution. This is far more probabilistically interpretable than simple methods based on the entropy of a single model prediction, while avoiding the implementation difficulty and computational overhead associated with more complex methods such as variational inference.

Specifically, we perform S forward propagations of the same sample (with dropout randomly enabled for each forward, which is equivalent to sampling from the approximate posterior distribution) to obtain S different score predictions $\{\hat{y}_i^s\}_{s=1}^S$, and then compute the variance of these predictions as a measure of uncertainty:

$$U(\mathbf{o}_i) = \text{Variance}(\{\hat{y}_i^1, \hat{y}_i^2, \dots, \hat{y}_i^S\}) \quad (12)$$

However, selecting only the samples with the highest uncertainty may lead to clustering of the selected samples in the acoustic feature space, resulting in labelling redundancy. To address this issue, we introduce a diversity metric. We map speech samples to a low-dimensional representation space using the feature extractor ϕ of the scoring model (e.g., the last hidden state layer of the Bi-LSTM encoder). Subsequently, we employ the k-means clustering algorithm to cluster the representations across the entire unlabelled pool $\mathcal{U} \setminus \mathcal{L}_t$. The k-means algorithm has been adopted for diversity sampling, mainly weighing its computational efficiency, scalability, and compatibility with the representation space. AL requires clustering of a large unlabelled pool in each round, and the linear time complexity $O(n)$ of k-means allows it to quickly process high-dimensional feature data and meet the real-time requirements of iterative queries. Despite the existence of more advanced clustering algorithms (e.g., density-based density-based spatial clustering of applications with noise or Gaussian mixture models), k-means works well for spherical clusters in Euclidean space, and our deep feature representations usually tend to be isotropically distributed after training, which is in line with the assumed premise of k-means. Its simplicity also avoids the tuning complexity brought by introducing additional hyperparameters and ensures the stability of the whole AL loop. Our strategy is to prioritise the selection of samples from clusters that have high uncertainty and whose belonging clusters are under-represented in the current annotation set $\mathcal{L}_t$. We use the k-means clustering algorithm to select the clusters from the current annotation set.

The ultimate query strategy Q is to select a set of samples Q such that they maximise both the uncertainty of the model and the intrinsic distributional structure of the unlabelled data. This strategy deftly balances exploration (querying the most difficult samples of the model) with exploitation (ensuring the comprehensiveness of the distribution of the labelled data).

The selected query batch Q is expert labelled and its valuable labels are added to the training set: $\mathcal{L}_{t+1} = \mathcal{L}_t \cup Q$. Subsequently, the pronunciation scoring model f_θ starts a new round of training on the updated, augmented hybrid dataset $\mathcal{L}_{t+1} \cup \mathcal{G}$. This cycle is repeated until a predefined labelling budget is reached or the model performance satisfies the convergence condition.

4 Experimental validation

4.1 Experimental setup

In this study, the Second Language Automatic Speech Recognition Corpus of Texas Instruments, Massachusetts Institute of Technology, and Oregon Graduate Institute (L2-ARCTIC) English non-native speech dataset was used as the experimental platform (Turgunboyeva, 2024), which contains English speech materials recorded by 24 speakers with different native language backgrounds (Chinese, Hindi, Korean, Arabic, Spanish and Vietnamese). Each speaker read the same 1,132 sentences, totalling approximately 27,168 speech samples. We use the officially provided phoneme-level forced alignment results and manually calibrated pronunciation scores as real labels for supervised learning. In terms of data partitioning strategy, we adopt speaker-independent partitioning to ensure model generalisation ability: data from 18 speakers (three from each native language background) are randomly selected as the training/unlabelled pool, and data from the remaining six speakers (one from each native language background) are used as the test set. All speech signals were uniformly resampled to 16 kHz, and then 80-dimensional Mel spectrograms were extracted as acoustic features with a frame length of 25 ms and a frame shift of 10 ms. The Mel scale effectively simulates the human ear’s nonlinear auditory perception. The dimensionality must balance between preserving critical acoustic information and controlling the scale of model parameters. Too low a dimension (e.g., 40 dimensions) will over-smooth the spectral details and lose the fine information necessary for distinguishing between similar phonemes, while too high a dimension (e.g., 128 dimensions) introduces a large amount of redundancy, which not only increases the computational load, but also tends to overfit the model. 80 dimensions have been widely validated as the ‘sweet spot’ that can adequately capture key features such as resonance peak structure while maintaining high computational efficiency, thus making it the best choice for the human ear. The 80-dimensional model is widely recognised as the ‘sweet spot’ that can adequately capture key features such as the resonance peak structure while maintaining high computational efficiency, and is therefore the choice for this study.

To comprehensively assess the performance of pronunciation scoring models, we employ three widely used evaluation metrics for regression tasks: the Pearson correlation coefficient (PCC) is used to measure the degree of linear correlation between predicted scores and manual scores; the mean absolute error (MAE) reflects the average absolute deviation of predicted scores; and the root MSE (RMSE) is more sensitive to large errors in prediction. We prioritise the PCC as it best captures the trend agreement between predicted scores and human expert ratings. The ultimate goal of APA is to mimic an expert’s relative judgement logic (i.e., to determine whether one pronunciation is better than another), and the PCC precisely captures this ordinal association. In contrast, the MAE and the RMSE, while able to quantify absolute error, are more sensitive to the overall bias of the scoring system (bias). In contrast, MAE and RMSE quantify absolute error but are more sensitive to the overall bias of the scoring system. Therefore, a metric system based on PCC and supplemented by MAE/RMSE can comprehensively assess the accuracy of the model in terms of specific scores while maintaining the correct scoring trend, which is highly suitable for the practical application of the APA task.

We compare the proposed fusion model (GAN + AL) with four strong baseline models: baseline 1 (seed only) uses only 5% of the initial labelled data to train a standard pronunciation score model; baseline 2 (traditional aug) adds traditional data augmentation techniques to baseline 1; baseline 3 (AL only) uses only AL loops; baseline 4 (GAN only) uses only GAN synthetic data augmentation. All experiments were implemented based on the PyTorch framework, and the pronunciation scoring model was trained with 50 epochs using the Adam optimiser using a two-layer Bi-LSTM architecture. We employed the Adam optimiser due to its adaptive learning rate mechanism, which is well-suited for the multi-task hybrid training environment of this study. For pronunciation scoring networks, Adam is able to automatically adjust the learning step size for different parameters, robustly handling possible gradient scale differences between the feature extractor (Bi-LSTM) and the output layer. For GAN training, whose notoriously difficult optimisation landscape requires the optimiser to dynamically adapt to the conflicting gradient dynamics of the generator and discriminator, Adam’s built-in momentum term helps to smooth out the gradient direction, avoid oscillations, and facilitate the convergence of adversarial training. Compared with traditional stochastic gradient descent (SGD), Adam reduces the need for cumbersome learning rate scheduling, and its design with momentum shows faster convergence and better final performance than other adaptive algorithms [e.g., root mean square propagation (RMSprop)] in practice. The training batch size is set to 32, which is based on the trade-off between memory constraints and the stability of the gradient estimation: smaller batches (e.g., 16) provide more frequent model updates, but lead to too much variance in the gradient, which prolongs the convergence time; larger batch sizes (e.g., 64) improve the stability of the training, but reduce the model generalisation capability and increase the memory overhead. A widely adopted benchmark value for the training batch size is 32. It achieves the best efficiency-effectiveness balance in the single-card training setting of this task.

4.2 Comparison of main results

Table 1 demonstrates the comprehensive performance of each model on the test set. Baseline 1 (PCC = 0.712) represents the baseline performance under data-sparse conditions. Baseline 2 (PCC = 0.728) gains a limited boost through conventional data enhancement, which demonstrates the inherent limitations of simple audio transformations in creating diversity in pronunciation content. Baseline 3 (PCC = 0.785) and baseline 4 (PCC = 0.801) demonstrate the respective effectiveness of AL and generative data augmentation, with generative augmentation showing relatively greater gain potential. Our proposed fusion model (GAN + AL) achieves the best performance (PCC = 0.843), which significantly outperforms all baseline models, which fully demonstrates the synergistic effect of GAN and AL: the diversified data generated by GAN provides a richer informative environment for AL, and the high-quality and real-world data intelligently labelled by AL, in turn, enhances the training effectiveness of GAN and the accuracy of the scoring model, and the two form a positive feedback loop.

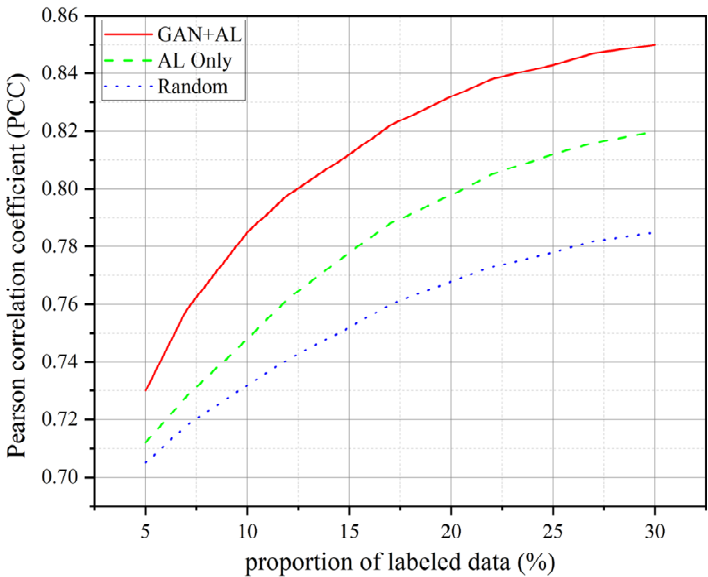
Table 1 Comparison of pronunciation scoring performance of the models on the L2-ARCTIC test set

Model	PCC	MAE	RMSE
Baseline 1 (seed only)	0.712 ± 0.015	0.321 ± 0.008	0.415 ± 0.009
Baseline 2 (traditional aug)	0.728 ± 0.012	0.308 ± 0.007	0.398 ± 0.008
Baseline 3 (AL only)	0.785 ± 0.011	0.285 ± 0.006	0.372 ± 0.007
Baseline 4 (GAN only)	0.801 ± 0.010	0.276 ± 0.005	0.361 ± 0.006
Proposed (GAN + AL)	0.843 ± 0.008	0.249 ± 0.004	0.332 ± 0.005

4.3 AL performance curve

Figure 2 illustrates the performance evolution curves of different strategies during AL. The horizontal axis represents the amount of labelled data as a percentage of the total training pool, and the vertical axis is the PCC value of the model on the test set. It can be observed that the performance curve of the GAN + AL strategy consistently outperforms the AL only and random sampling baselines throughout the AL process. In particular, the advantage of GAN + AL is especially significant in the early stage (<15%) when the annotation budget is limited, indicating that the synthetic data effectively compensates for the lack of real annotations in the cold-start phase of the model, providing the model with an initial generalisation capability and making the subsequent AL queries more accurate and efficient. Pure uncertainty sampling, although superior to random sampling, starts from a lower starting point and progresses at a relatively slower rate, highlighting the importance of providing high-quality synthetic data in the early stages of AL.

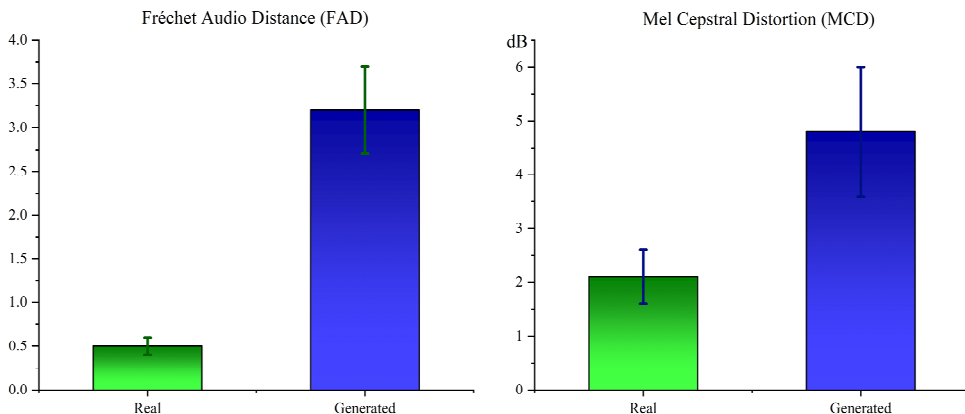
Figure 2 Convergence curves of model performance under different AL strategies (see online version for colours)



4.4 Generate data quality analysis

In order to quantitatively evaluate the quality of the generated speech, two objective metrics, Fréchet audio distance (FAD) and Mel cepstral distortion (MCD), were computed, and the results are shown in Figure 3. The FAD value measures the similarity between the generated audio and the real audio in terms of the distribution of the deeper features, and the lower the value, the better. The MCD value measures the difference in spectral details between the generated audio and the target audio, and the lower the value, the better. The MCD value measures the difference in spectral details between the generated audio and the target audio in dB, and its value is also as low as possible. The results show that the FAD (3.2) and MCD (4.8 dB) values of the generated speech remain low. Although the values are slightly higher than the baseline of real speech (FAD: 0.5, MCD: 2.1 dB) which represents the natural fluctuation of human articulation, the difference is within acceptable limits. In particular, the MCD values were well below what is commonly considered to be the threshold of auditory acceptability (5–6 dB), indicating that the generated speech was highly close to the real target in terms of phoneme clarity and spectral structure. These objective data provide strong evidence that cGANs are capable of generating acoustically sound, high-quality data, thus providing valuable training samples for articulatory scoring models.

Figure 3 Comparison of synthesised speech and real speech quality based on FAD and MCD (see online version for colours)



4.5 Ablation experiment

To explore the contribution of different components in query strategies, we designed ablation experiments comparing three strategies: pure uncertainty sampling, pure diversity sampling, and a hybrid strategy combining the two. The experimental results show that the hybrid strategy performs consistently better than any single strategy in the middle and late stages of AL. This confirms that diversity sampling effectively avoids clustering of query samples in the feature space, prompting the model to explore regions of the data distribution that are not yet adequately represented, leading to a more comprehensive performance improvement. In addition, we analyse the performance differences on speakers with different native language backgrounds and find that the

model performs better on native language groups that are adequately represented in the training data, pointing to the need for future research to further improve the model's cross-linguistic generalisation capabilities.

4.6 *Experimental results and analysis*

This study validates the superior performance of fusing GAN synthetic data augmentation with an AL framework in an APA task through systematic experiments. The experimental results clearly show that the fusion approach achieves significant improvements in three key metrics, namely, PCC, MAE, and RMSE, as compared to the traditional data enhancement methods and strategies using GAN or AL alone. This finding echoes the findings of Zhang (2025), who found that deep fusion of multiple features and strategies can effectively improve the robustness of pronunciation scoring systems. More importantly, our study further reveals the synergistic effect between GAN and AL: the large amount of diverse synthetic data generated by GAN not only directly expands the training set, but more importantly provides a more informative selection environment for the AL process, enabling the query strategy to more accurately identify the most informative samples; in turn, the high-quality real labelled data obtained through AL can in turn enhance the conditional generation quality of the GAN, resulting in a self-enhancing virtuous cycle. This finding provides a new idea for solving the long-standing data scarcity problem in speech processing.

However, there are several limitations to this study. First, the quality of GAN-generated data relies heavily on the quality and quantity of the initial seed data. Although the cyclic consistency loss (Zhang et al., 2021) guarantees the consistency of the generated content with the conditional information to a certain extent, the diversity of the generated data may be limited in very low resource scenarios (e.g., initial labelled data <2%). In this regard, in the future, we can explore the sample less generation method based on pre-trained speech representation model (e.g., wav2vec2.0), or introduce the diffusion model in speech synthesis (diffusion model) to alleviate the problem of pattern collapse under low resources through a better prior distribution, and at the same time, adopt the course learning strategy to gradually increase the complexity of the generated data. Second, the current framework has a relatively high computational cost involving GAN training, multiple rounds of model updating, and prediction of large-scale unlabelled pools, which may become a constraint in real-world application scenarios with limited computational resources. Improvement directions include designing lighter generator architectures (e.g., lightweight GANs after knowledge distillation), adopting early-stopping strategies to reduce the number of AL rounds, and developing batch AL methods based on the core set to reduce the computational overhead and improve training efficiency. Finally, while the model shows good performance on different native language groups in the L2-ARCTIC dataset, its generalisation ability on completely unfamiliar native language contexts or dialect variants still needs to be further verified. Future work will perform cross-domain evaluation on more diverse multi-dialect datasets (e.g., common voice) and introduce domain adaptive techniques (e.g., adversarial domain adaptation) and meta-learning frameworks to enable the model to quickly adapt to new accent variants, thus enhancing its robustness and generalisability in practical applications.

In addition, we propose the following future research directions: exploring the application of more advanced generative models (e.g., diffusion models) in speech data

synthesis to further improve the quality and diversity of the generated data. Investigate more efficient knowledge distillation techniques to compress knowledge from large models into lightweight models to reduce deployment costs and enable the system to run in real-time on mobile devices. Explore the fusion of multimodal data (e.g., lip shape video, articulatory organ movement data) to improve the accuracy and interpretability of pronunciation assessment through multimodal information. The framework will be extended to a wider range of languages and learning scenarios to validate its cross-linguistic and cross-domain generalisation capabilities, and ultimately promote the universal development of intelligent language education technology.

The theoretical contributions of this study are mainly in three aspects: first, we for the first time place generative data augmentation and AL in a unified framework, and empirically verify the synergistic effect that exists between the two in the pronunciation scoring task, which provides a new theoretical perspective for understanding the interactions between different data strategies. Second, we design a hybrid query strategy that incorporates uncertainty and diversity, which effectively solves the common sample selection bias problem in AL and provides new ideas for improving the query mechanism (Settles, 2009). Finally, we realise cGAN-based speech generation with controlled articulatory quality, which extends the boundaries of generative modelling applications in the field of speech education.

In a practical sense, this study provides a feasible technical path for developing a low-cost intelligent pronunciation teaching system. Based on the framework of this study, educational institutions can use limited expert annotation data to train the initial generation model at the initial stage of system deployment, and then rapidly expand the system's coverage of common pronunciation errors through synthetic data; subsequently, the system performance can be continuously optimised by collecting learners' difficult pronunciation samples through the AL mechanism and accurately annotating them. This 'generate-annotate-optimize' cycle significantly reduces the reliance on expert annotation, and enables high-quality pronunciation assessment services to reach a wider range of learners (Xu et al., 2021). In addition, the method proposed in this study is also applicable to other low-resource speech processing tasks, such as accent recognition, speech disorder detection, etc., and has good generalisation value.

5 Conclusions

In this study, we propose and validate an automatic English pronunciation scoring model that fuses GAN synthetic data augmentation with AL. Through comprehensive experiments on the L2-ARCTIC dataset, we confirm that the fusion framework significantly improves the accuracy and robustness of the pronunciation scoring model, while significantly reducing the reliance on expert labelled data. Specifically, the method achieves a performance of 0.843 on the PCC, which significantly outperforms all comparison baselines. This result suggests that the synergistic integration of generative data augmentation and AL is an effective way to address the data scarcity problem in speech processing.

This study's significance lies not only in the proposed high-performance model but also in demonstrating a practical machine learning framework design: performance optimisation under limited resources is achieved by cleverly combining generative

methods and selective annotation strategies. This methodological innovation provides a valuable reference for other low-resource machine learning tasks. With the in-depth application of AI technology in the field of education, we believe that this technical route of integrating data generation and intelligent annotation will provide important support for the development of a more inclusive and efficient language learning system, which will ultimately help to eliminate barriers in language learning and promote cross-cultural communication and understanding.

Declarations

All authors declare that they have no conflicts of interest.

References

- Abdelwahab, M. and Busso, C. (2019) 'Active learning for speech emotion recognition using deep neural network', *Affective Computing and Intelligent Interaction*, Vol. 1, pp.1–7.
- Amrate, M. and Tsai, P-h. (2024) 'Computer-assisted pronunciation training: a systematic review', *ReCALL*, Vol. 37, pp.1–21.
- Brahmi, Z., Mahyoob, M., Al-Sarem, M., Algaraady, J., Bousselmi, K. and Alblwi, A. (2024) 'Exploring the role of machine learning in diagnosing and treating speech disorders: a systematic literature review', *Psychology Research and Behavior Management*, Vol. 17, pp.2205–2232.
- Chatziagapi, A., Paraskevopoulos, G., Sgouropoulos, D., Pantazopoulos, G., Nikandrou, M., Giannakopoulos, T., Katsamanis, A., Potamianos, A. and Narayanan, S. (2019) 'Data augmentation using GANs for speech emotion recognition', *Interspeech*, Vol. 5, pp.171–175.
- Chen, D. (2015) 'Multi-task learning deep neural networks for automatic speech recognition', *Hong Kong University of Science and Technology*, Vol. 12, p.541.
- Cheng, Z. and Wang, Z. (2022) 'Automatic scoring of spoken language based on basic deep learning', *Scientific Programming*, Vol. 2022, No. 1, p.6884637.
- Gal, Y. and Ghahramani, Z. (2016) 'Dropout as a bayesian approximation: representing model uncertainty in deep learning', *Machine Learning Research*, Vol. 12, pp.1050–1059.
- Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y. (2014) 'Generative adversarial nets', *Advances in Neural Information Processing Systems*, Vol. 27, p.4534.
- House, J. (2003) 'English as a lingua franca: a threat to multilingualism?', *Journal of Sociolinguistics*, Vol. 7, No. 4, pp.556–578.
- Ko, T., Peddinti, V., Povey, D. and Khudanpur, S. (2015) 'Audio augmentation for speech recognition', *Interspeech*, Vol. 2015, p.3586.
- Kumar, K., Kumar, R., De Boissiere, T., Gestin, L., Teoh, W.Z., Sotelo, J., De Brebisson, A., Bengio, Y. and Courville, A.C. (2019) 'MelGAN: generative adversarial networks for conditional waveform synthesis', *Advances in Neural Information Processing Systems*, Vol. 32, p.1454.
- Li, K., Qian, X. and Meng, H. (2016) 'Mispronunciation detection and diagnosis in L2 English speech using multidistribution deep neural networks', *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 25, No. 1, pp.193–207.
- Park, D.S., Chan, W., Zhang, Y., Chiu, C-C., Zoph, B., Cubuk, E.D. and Le, Q.V. (2019) 'SpecAugment: a simple data augmentation method for automatic speech recognition', *Electrical Engineering and Systems Science*, Vol. 3, p.1541.

- Qian, Y., Hu, H. and Tan, T. (2019) 'Data augmentation using generative adversarial networks for robust speech recognition', *Speech Communication*, Vol. 114, pp.1–9.
- Riccardi, G. and Hakkani-Tur, D. (2005) 'Active learning: theory and applications to automatic speech recognition', *IEEE Transactions on Speech and Audio Processing*, Vol. 13, No. 4, pp.504–511.
- Rogti, M. (2025) 'Automatic pronunciation evaluation feedback and peer correction for shaping English pronunciation accuracy and interpersonal communication', *Innovation in Language Learning and Teaching*, Vol. 1, pp.1–17.
- Settles, B. (2009) *Active Learning Literature Survey*, January, Computer Sciences Technical Report 1648, University of Wisconsin-Madison, Madison, WI, USA.
- Turgunboyeva, M. (2024) 'Using speech corpora to identify common pronunciation challenges for ESL learners', *Mental Enlightenment Scientific-Methodological Journal*, Vol. 5, No. 9, pp.262–269.
- Witt, S.M. and Young, S.J. (2000) 'Phone-level pronunciation scoring and assessment for interactive language learning', *Speech Communication*, Vol. 30, Nos. 2–3, pp.95–108.
- Xiang, Y. and Bao, C. (2020) 'A parallel-data-free speech enhancement method using multi-objective learning cycle-consistent generative adversarial network', *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 28, pp.1826–1838.
- Xu, X., Kang, Y., Cao, S., Lin, B. and Ma, L. (2021) 'Explore wav2vec 2.0 for mispronunciation detection', *Interspeech*, Vol. 2021, pp.4428–4432.
- Yan, B.-C. and Chen, B. (2024) 'An effective hierarchical graph attention network modeling approach for pronunciation assessment', *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 32, pp.3974–3985.
- Zahran, A.I., Fahmy, A.A., Wassif, K.T. and Bayomi, H. (2023) 'Fine-tuning self-supervised learning models for end-to-end pronunciation scoring', *IEEE Access*, Vol. 11, pp.112650–112663.
- Zhang, X., Wang, J., Cheng, N., Xiao, E. and Xiao, J. (2021) 'CycleGEAN: cycle generative enhanced adversarial network for voice conversion', *Automatic Speech Recognition and Understanding Workshop*, Vol. 1, pp.930–937.
- Zhang, Z. (2025) 'Construction of an English oral pronunciation evaluation model based on deep learning algorithms', *International Journal of High Speed Electronics and Systems*, Vol. 6, p.2540307.
- Zhu, J.-Y., Park, T., Isola, P. and Efros, A.A. (2017) 'Unpaired image-to-image translation using cycle-consistent adversarial networks', *Computer Vision*, Vol. 7, pp.2223–2232.