MPSA-DenseNet: A novel deep learning model for English accent classification

Song, Tianyu
Mathematical Modeling Laboratory, Kyushu University

Linh Thi Hoai Nguyen
International Institute for Carbon-Neutral Energy Research (WPI-I2CNER), Kyushu University

Ton Viet Ta
Mathematical Modeling Laboratory, Kyushu University

出版情報：Computer Speech & Language. 89, pp.101676-, 2025-01. Elsevier
バージョン：

Computer Speech and Language 89 (2025) 101676

Contents lists available at ScienceDirect

Computer Speech & Language

journal homepage: www.elsevier.com/locate/csl

MPSA-DenseNet: A novel deep learning model for English accent classification

Tianyu Song [a], Linh Thi Hoai Nguyen [b], Ton Viet Ta [a,*]

[a] Mathematical Modeling Laboratory, Kyushu University, Fukuoka 819-0395, Japan
[b] International Institute for Carbon-Neutral Energy Research (WPI-I2CNER), Kyushu University, Fukuoka 819-0395, Japan

ARTICLE INFO

Keywords:
Deep learning
Accent classification
Multi-task learning
Attention mechanism

ABSTRACT

This paper presents three innovative deep learning models for English accent classification: Multi-task Pyramid Split Attention- Densely Convolutional Networks (MPSA-DenseNet), Pyramid Split Attention- Densely Convolutional Networks (PSA-DenseNet), and Multi-task- Densely Convolutional Networks (Multi-DenseNet), that combine multi-task learning and/or the PSA module attention mechanism with DenseNet. We applied these models to data collected from five dialects of English across native English-speaking regions (England, the United States) and nonnative English-speaking regions (Hong Kong, Germany, India). Our experimental results show a significant improvement in classification accuracy, particularly with MPSA-DenseNet, which outperforms all other models, including Densely Convolutional Networks (DenseNet) and Efficient Pyramid Squeeze Attention (EPSA) models previously used for accent identification. Our findings indicate that MPSA-DenseNet is a highly promising model for accurately identifying English accents.

1. Introduction

In recent years, deep learning technology has made significant advancements in the field of speech applications. Accent recognition, in particular, has received much attention, as it plays a very important role in improving current Automatic Speech Recognition (ASR) systems (Najafianand and Russell, 2020). Accents vary in the emphasis on words or syllables and are often influenced by a person's upbringing or social background.

As the most widely used language in the world, English plays a crucial role in international communication and business activities. However, different usage environments, historical and cultural backgrounds in different regions, have resulted in various accents and dialects of English, increasing the difficulty of recognizing correct content by smart voice devices. Therefore, there is a need to enhance the recognition of different accents.

To address the accent classification problem, traditional machine learning methods such as Support Vector Machines (SVM) (Pedersen and Diederich, 2007; Rizwan and Anderson, 2018; Tang and Ghorbani, 2003; Hou et al., 2010), Hidden Markov Model (HMMs) (Angkititrakul and Hansen, 2006; Kumpf and King, 1996) were used before the emergence of deep learning. Studies by Pedersen and Diederich (2007), Tang and Ghorbani (2003), and Bird et al. (2019) show successful classification results using SVM, Pairwise SVMs, and K-Nearest Neighbors (KNN) methods, respectively.

In the field of accent classification, feature extraction of speech signals plays a critical role. Mel-scale Frequency Cepstral Coefficients (MFCC) is a widely used method for speech signal feature extraction, which has been confirmed to have good

* Corresponding author.
E-mail addresses: song.tianyu.064@s.kyushu-u.ac.jp (T. Song), linh@i2cner.kyushu-u.ac.jp (L.T.H. Nguyen), tavietton@agr.kyushu-u.ac.jp (T.V. Ta).

https://doi.org/10.1016/j.csl.2024.101676
Received 23 December 2023; Received in revised form 24 May 2024; Accepted 28 May 2024
Available online 30 May 2024

performance in accent classification by many studies (Ittichaichareon et al., 2012; Hossan et al., 2010; Bhowmik et al., 2022). MFCCs output spectrograms–detailed graphical representations of audio, allowing powerful models like CNN to address speech processing challenges effectively. However, some studies have attempted to prove that MFCC may not always be the optimal representation. Biswas (2023) and Kethireddy et al. (2020) found that learning from waveforms improved the accent classification performance by 10.94% unweighted average recall (UAR), yet MFCC remained superior in specific acoustic features. Additionally, Lou and Ren (2021) incorporated characteristics such as voice onset region (VOR), vowels, and formants into the feature vector to train linear neural network models and neural networks with nonlinear classifications and two hidden layers to differentiate between British English and American English, achieving an accuracy rate of 86.67%.

With the development of deep learning, neural network models have gradually replaced traditional machine learning methods in accent recognition and classification research. Deep learning models can automatically extract features from large-scale speech data, thereby improving accent recognition ability. For instance, studies by Jiao et al. (2016) and Chionh et al. (2018) show that combining Deep Neural Network (DNN) and Recurrent Neural Network (RNN) or using Convolutional Neural Network (CNN) greatly improved the classification accuracy of accents. Additionally, a survey by Zaman et al. (2023) provides a comprehensive comparison of audio signal classification across five different deep neural networks, highlighting the efficacy of CNN in domains such as audio signal classification and speech recognition. Furthermore, research by Al-Jumaili et al. (2022) shows that using transfer learning with a lightweight neural network model achieved high-precision results on accent classification tasks. Aiming at the similarity problem between native and non-native English accents, Wubet et al. (2023) proposed a new model of native accent identification (NAI) framework based on the sharing of intrinsic native accent features, which improved the baseline method by 3.7% to 7.5% average accuracy.

Recently, Carofilis et al. (2023) proposed leveraging deep learning's feature extraction to improve classical machine learning methods' performance. Specifically, a CNN model is trained on audio file spectrograms to extract important features. The results obtained by CNNs are interpreted using gradient-weighted class activation mapping (Grad-CAM), a class-discriminative localization technique, to produce dimensionality-reduced heatmaps. Subsequently, the Grad-Transfer concatenates flattened spectrogram and the heatmaps from Grad-CAMs to form the feature matrix for classification machine learning algorithms, such as SVM, Gaussian Naive Bayes, Online Passive-Aggressive Classifier, and XGBoost Classifier. Implementation on a subset of the Voice Cloning Toolkit (VCTK) dataset reveals that this method outperforms baseline approaches that only utilize spectrograms. For instance, test Macro Average Accuracy (MAA) improves by up to 23% with Gaussian Naive Bayes, UAR improves by up to 16.24% with Passive-Aggressive Classifiers.

Furthermore, in recent years, deep learning has been increasingly employed to handle more complex scenarios beyond simple single label, single task learning setups. For example, Mulimani and Mesaros (2024) proposed a method for class incremental learning capable of learning new classes independently of old ones. This method addresses a series of multi-label audio classification tasks for potential overlapping sounds, achieving an average F_1 score of 40.9% across five stages. In addition, attention mechanism (Gao et al., 2021) and multi-task learning (Zeng et al., 2019) have been applied in the speech and audio fields, showing improved classification performance. For instance, Sharma (2022) used a pre-trained wav2vec 2.0 model and constructed an emotion recognition system using multi-task learning on a multilingual dataset. This method improved performance by 7.2% and 1.7%, respectively, compared to single-head PANN and wav2vec 2.0 models. Meanwhile, Naini et al. (2022) utilized CNN-based dual and single attention pooling methods to classify recording devices based on speech signals, achieving superior performance compared to the optimal baseline scheme.

While most of the current research focuses on accurately translating speech signals with different accents into correct content, our study focuses on the recognition and classification of different accents based on the country or region of the speaker's English accent. In this study, we newly introduce three novel deep learning models named Multi-DenseNet, PSA-DenseNet, and MPSA-DenseNet, which combine DenseNet (Huang et al., 2017), multi-task, and attention mechanism models, to achieve better performance compared to previous models such as DenseNet or Residual Neural Networks (ResNet) for accent identification. This study has an important impact on communication, language learning, and social interactions. It can also have practical applications in areas such as speech recognition, language translation, and forensic linguistics.

In the literature, there have been efforts to combine two of these three components for various applications. For example, a model combining multiscale learning and attention mechanisms densely connected network (CMADNet) was developed to effectively extract rain streak features and restore damaged background texture information in images (Chen and Zhu, 2023). Liu et al. developed Multi-Task Deep Neural Networks (MT-DNN) for learning representations across multiple natural language understanding tasks (Liu et al., 2019). To the best of our knowledge, our three models are the first deep learning models to combine DenseNet, multi-task learning, and attention mechanisms to address problems in audio data and speech applications.

The remainder of this paper is organized as follows. The Materials and Methods section describes the architectures of the three new deep learning models: Multi-DenseNet, PSA-DenseNet, and MPSA-DenseNet. In the Results section, we validate the effectiveness of our proposed models by comparing them with existing models. Finally, the Conclusion section provides concluding remarks and outlines future research directions.

2. Materials and methods

In this section, first, we describe the data collection and preprocessing process, which involves processing raw speech data and converting it into spectrograms. Second, we introduce our deep learning models, Multi-DenseNet, PSA-DenseNet, and MPSA-DenseNet, which combine DenseNet, multi-task, and attention mechanism models. Evaluation metrics and graphics processing unit (GPU) configuration used in the study are also given.

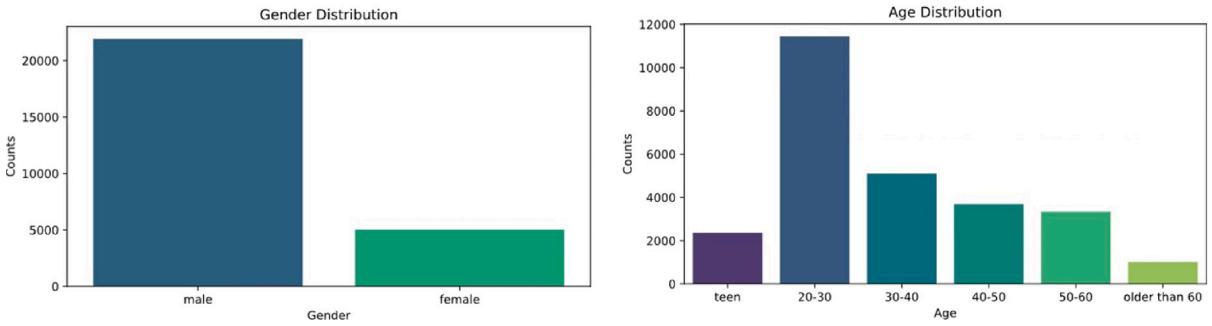

Fig. 1. Gender (left) and age (right) distributions.

Table 1
Distribution of samples across accent classes.

Accent	Total number
England	5054
U.S.A	7209
German	5345
Hong Kong	4159
India	5187

2.1. Data collection

In this research, English accent classification experiments have been conducted using various datasets collected from the Common Voice Dataset (Ardila et al., 2020). To ensure the diversity of datasets, we avoided selecting a single type of speech data as much as possible, resulting in five dialects of English in the dataset, including two native English-speaking regions (England and the United States) and three nonnative English speaking regions (Hong Kong, Germany, and India).

In addition, since we utilized a multi-task learning method, age, and gender categories as auxiliary tasks are included. This approach aimed to train a model with better robustness. The age category is divided into six groups: below 20, 20–29, 30–39, 40–49, 50–59, and 60 and above. The gender category is grouped as male and female. The distribution of data in each category is shown in Fig. 1, revealing that these distributions are uneven. Although the unevenness could generally be addressed through undersampling, we chose to retain it for two reasons. First, we did not want to affect the distribution of accent labels in the main task by undersampling the two auxiliary tasks. Second, the unevenness in these auxiliary tasks may actually contribute to making our model more robust when facing real-world challenges.

The dataset consists of 26,954 samples, with the distribution of data in each accent class shown in Table 1. Furthermore, Fig. 2 illustrates the density function of the duration distribution of samples in each class, typically ranging from 2 to 10 s, with a concentration of most samples between 5 and 6 s. (The density is estimated using the kernel density estimation (KDE) method Terrell and Scott (1992).) Subsequently, we standardize the duration of all sample data based on this distribution to ensure uniform feature shapes.

2.2. Data preprocessing

Let us describe step-by-step the speech data preprocessing flow.

2.2.1. Standardizing file formats

Audio files in various formats, including wav, mp3, and mp4, are converted into a uniform wav format compatible with Librosa (McFee et al., 2015), a Python library for audio processing. To ensure lossless conversion and minimize any potential impact on the voice files, we utilized FFmpeg (Tomar, 2006), an open-source command-line tool designed for handling diverse multimedia files and streams.

2.2.2. Standardizing voice files

We standardized the voice files by adjusting the sampling rate to 44,100 Hz, a commonly used value in the field. For audio samples recorded in mono, we converted them into stereo by duplicating the channel. Additionally, we normalized the duration of all samples to 4 s by trimming longer audio files and padding shorter ones with silence.

In Fig. 3, these steps are illustrated as Resize Channel and Resize Duration.

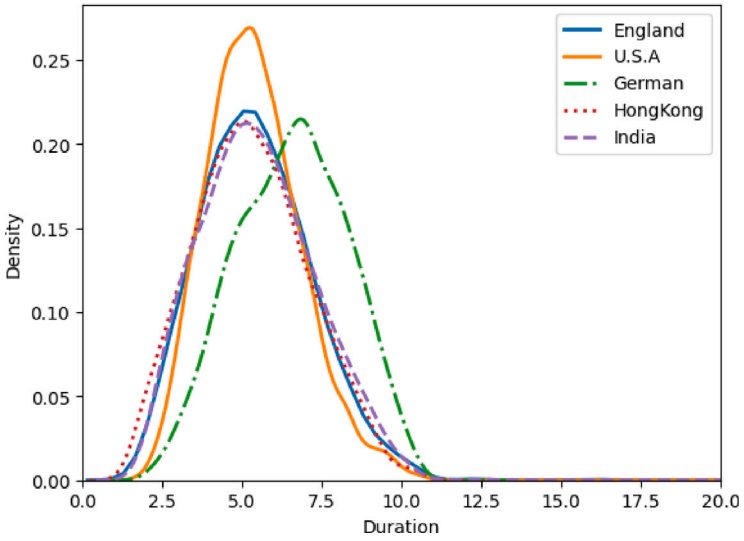

Fig. 2. Duration density by accent class.

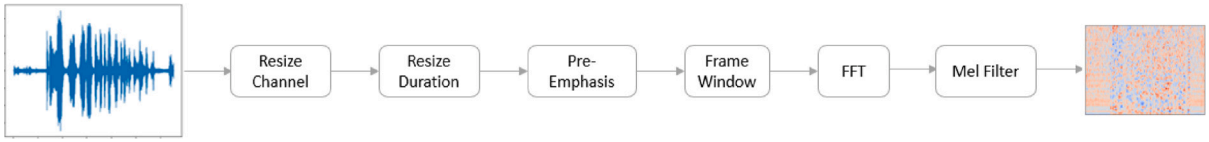

Fig. 3. Voice file standardization and spectrogram conversion with MFCCs.

2.2.3. Spectrogram conversion

To extract spectral features from the collected data of five English dialects, we utilized Mel-frequency cepstral coefficients (MFCCs). MFCCs provide a compact representation of the spectral features of an audio signal, capturing essential frequency information while discarding irrelevant details. The spectrogram conversion process using MFCCs involves four steps shown in Fig. 3.

Pre-Emphasis: Applying a first-order high-pass filter to a speech signal enhances its high-frequency components, resulting in a flatter spectrum. This process enables the extraction of the signal spectrum at the same signal-to-noise ratio.

Frame Window: The signal is divided into frames, typically lasting about 25 ms and possibly overlapping, to capture temporal information more accurately. A window function is then applied to each frame to reduce oscillations at the frame boundaries.

Fast Fourier Transform (FFT): A Fourier transform is applied to each frame to convert the signal from the time domain to the frequency domain. The resulting magnitude spectrum represents the distribution of signal energy across different frequencies.

Mel Filter: MFCCs utilize the nonlinear Mel-scale, which approximates the human auditory system's response to different frequencies. Mel filters are a set of overlapping triangular or cosine windows spaced along the Mel-scale. Each filter captures energy within a specific frequency range. The magnitude spectrum of each frame is multiplied by each filter in the Mel filters, and the resulting energies are summed to create a filter-bank energy. This process generates a set of filter-bank energies, each corresponding to one frame, representing the distribution of signal energy across different Mel-frequency banks.

The filter-bank energies undergo logarithmic scaling to compress their dynamic range and enhance robustness to signal amplitude variations. Then, a discrete cosine transform is applied to the logarithmically scaled filter-bank energies, yielding a set of cepstral coefficients.

Typically, only the lower-order coefficients are retained as MFCCs, as they effectively represent the spectral characteristics of the audio signal within each frame. These coefficients serve as features for subsequent analysis or classification tasks.

2.3. Deep learning models

In this subsection, we present our deep learning models: Multi-DenseNet, PSA-DenseNet, and MPSA-DenseNet. The backbone of these three models is DenseNet, which is elaborated on in detail in the following subsection.

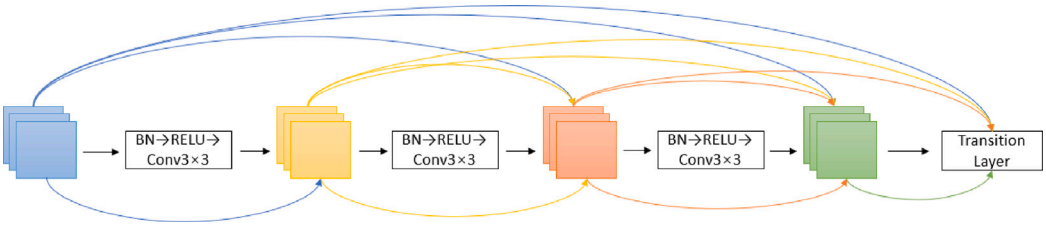

Fig. 4. Dense blocks and transition layers of DenseNet.

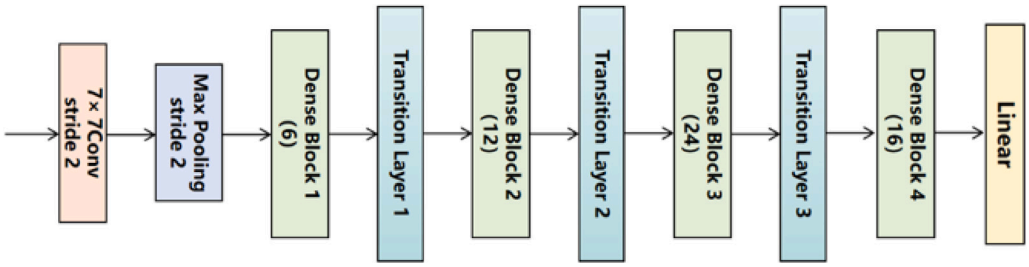

Fig. 5. DenseNet-121 architecture.

2.3.1. DenseNet model

Our research employs DenseNet as the backbone network due to its efficient feature reuse, which helps mitigate the risk of overfitting, especially with smaller datasets, and reduces the overall parameter count of the model. DenseNet (Huang et al., 2017) comprises two key components: dense blocks and transition layers. The dense blocks determine how the inputs and outputs are concatenated, while the transition layers control the number of channels to prevent it from becoming too large. The structure of the main components is illustrated in Fig. 4.

In DenseNet, each layer is connected to all the preceding layers in a feed-forward manner. For a K-layer DenseNet, there are $\frac{K(K+1)}{2}$ connections, and the input to each layer comes from the output of all previous layers. DenseNet has several main architectures, and we selected the DenseNet-121 architecture in this study, as shown in Fig. 5.

Let us delve into the details of this architecture. The input to DenseNet-121, typically an RGB image, undergoes processing through an initial convolutional layer. This layer employs a 7×7 kernel size and a stride of 2, followed by Max Pooling with a 3×3 kernel and a stride of 2, which aids in downsampling the image and extracting low-level features.

Subsequently, the output flows through four dense blocks, denoted as Dense Block 1 through Dense Block 4, each containing multiple dense layers. Within each dense block, feature maps from previous layers are concatenated, fostering feature reuse and facilitating gradient flow. These dense blocks are characterized by a high number of feature maps and smaller filter sizes.

Transition Layers are inserted between the dense blocks. These layers include batch normalization followed by a 1×1 convolutional layer and average pooling. The primary function of the transition layers is to minimize feature map dimensionality, control parameter count, and enable efficient information flow between dense blocks.

Following the final dense block, a global average pooling layer is employed, which spatially averages feature maps, resulting in a reduction of spatial dimensions to a single feature vector per channel. The last linear layer is a fully connected layer, which is typically consists of a small number of neurons, followed by a softmax activation function to convert the raw scores into classes.

2.3.2. Multi-DenseNet model

In this subsection, we present our first model, the Multi-DenseNet. This model introduces a novel approach by integrating DenseNet with multi-task learning, yielding several distinct technical advancements. This combination aims to not only reduce parameter count and memory requirements compared to the existing DenseNet model but also enhance performance across multiple tasks.

Multi-task learning involves training a neural network to simultaneously learn multiple related tasks, leveraging shared representations across these tasks (Caruana, 1997). By backpropagating gradients from all tasks together and sharing parameters at the bottom layer, the model benefits from collective learning, enhancing its generalization capability.

One significant advantage of multi-task learning is its ability to share representations across tasks, improving generalization and performance for each individual task. This approach acts as a form of regularization, encouraging the model to learn representations useful for multiple tasks simultaneously, thus preventing overfitting and enhancing generalization to unseen data. Additionally, it facilitates transfer learning, allowing knowledge gained from one task to improve performance on related tasks. Leveraging information from related tasks enriches the learning process, resulting in superior performance.

Two main categories of multi-task learning are commonly employed: hard parameter sharing and soft parameter sharing. The distinction lies in how the parameters in the bottom layer are shared among the tasks. In hard parameter sharing, only specific

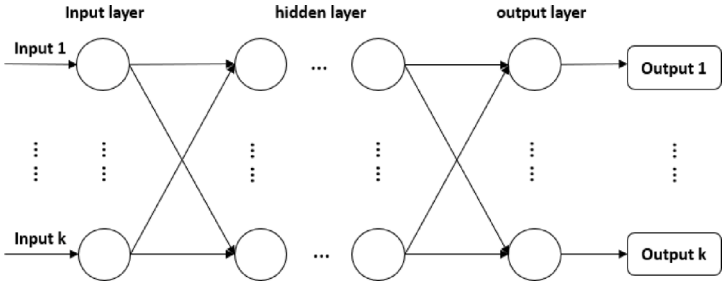

Fig. 6. Multi-task learning: hard parameter sharing.

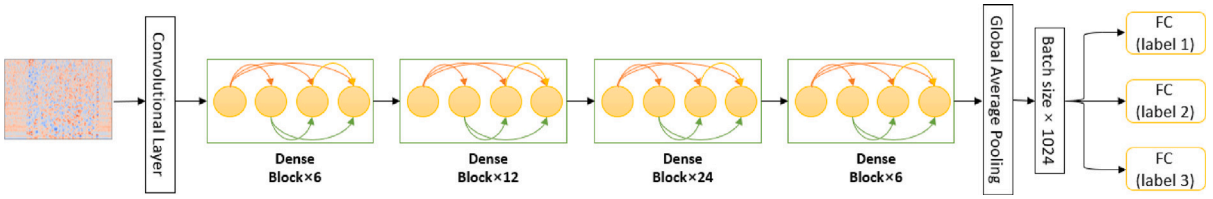

Fig. 7. Structure of Multi-DenseNet model.

output layers (such as Softmax) are added for each task, while the underlying parameters are uniformly shared across all tasks, utilizing a shared backbone. (See Fig. 6 for an illustration of hard parameter sharing.)

The hard parameter-sharing category is employed in our Multi-DenseNet model, encouraging it to learn representations useful for all tasks simultaneously, potentially leading to better generalization. We incorporate age and gender as auxiliary tasks to enhance the accent classification performance. Age and gender exhibit distinct characteristics in speech, such as pronunciation clarity and intonation, which can influence the results of the accent classification task. Additionally, by building a classifier that can determine the age and gender of the speaker's accent, we lay the groundwork for future applications in speaker recognition.

The architecture of Multi-DenseNet is depicted in Fig. 7. As seen in this figure, the feature information extracted from DenseNet is passed as input to the fully connected (FC) layers (label1, label2, label3) for multi-task learning during the training process.

Let us now define the total loss function for Multi-DenseNet. Since each task has its own loss function, and the importance of each task may vary, it is necessary to assign appropriate weights to the losses for aggregating them into a total loss function. The simplest approach for assigning multi-task loss weights is to linearly sum the individual task losses. The model's loss function, denoted as L_{total}, aggregates the individual loss function from each task:

$$L_{total} = \sum_i \omega_i L_i,$$

where L_i represents the loss function of each task, and ω_i denotes the weight assigned to each task. These weights are typically manually set as prior hyperparameters, ensuring that $\sum_i \omega_i = 1$.

Since we use the cross-entropy function (Crawshaw, 2020) for the loss function L_i of each task, the total loss function L_{total} is calculated as follows:

$$L_{total} = \omega_1 \sum_{i=1}^{C_1} y_i^{(1)} \log(\hat{y}_i^1) + \omega_2 \sum_{i=1}^{C_2} y_i^{(2)} \log(\hat{y}_i^2) + \omega_3 \sum_{i=1}^{C_3} y_i^{(3)} \log(\hat{y}_i^3).$$

Here, $y_i^{(k)}$, $\hat{y}_i^k$, and C_i represent the true label, the predicted value, and the total number of classes for each task, respectively.

2.3.3. PSA-DenseNet model

In this subsection, we introduce our second model, the PSA-DenseNet. Similar to the Multi-DenseNet model, PSA-DenseNet presents a novel approach by incorporating an attention mechanism into DenseNet. Here, we utilize a specific attention mechanism called PSA Module (Pyramid Split Attention Module), derived from EPSANet (Zhang et al., 2022). The PSA Module consists of two key components: Spatial Pyramid Convolution (SPC) module and Squeeze–Excitation Weight (SEWeight) module. Let us give more information about these two modules.

The first one, SPC Module is responsible for channel segmentation and extracting spatial information from each channel to generate multi-scale features. It achieves this by dividing the MFCC feature map into S parts based on the number of input channels (C). The resulting feature maps retain the same shape as the original, with S feature maps, each containing $\frac{C}{S}$ channels. The spatial information is then extracted through convolution operations on the feature maps of different scales. By concatenating these multi-scale feature maps, a new multi-scale fusion feature map is obtained. The structure of the SPC Module is shown in Fig. 8.

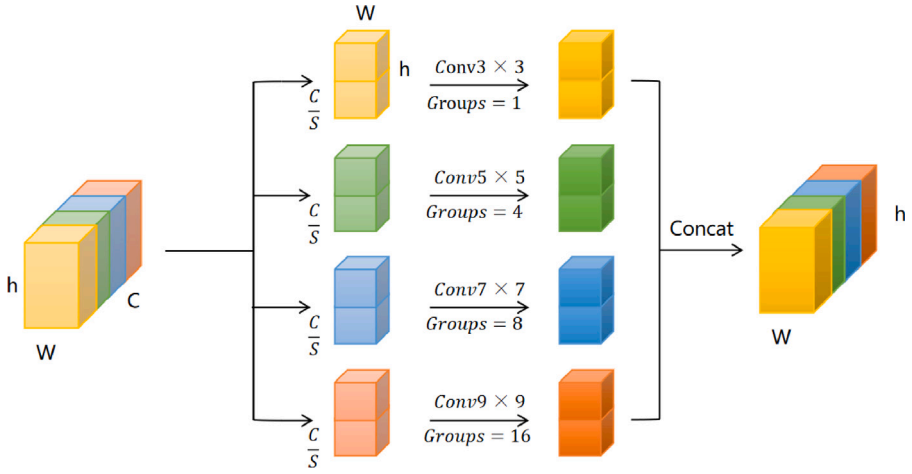

Fig. 8. Structure of SPC Module refer to the EPSANet.

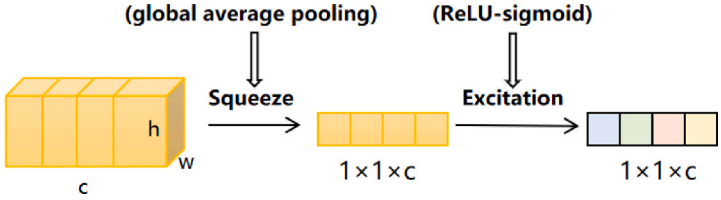

Fig. 9. Structure of SEWeight module which includes the two most important components: Squeeze and Excitation.

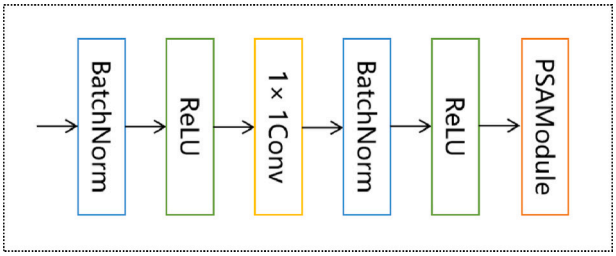

Fig. 10. PSA-DenseNet model resulting from inserting the PSA Module into DenseNet121.

The second one, SEWeight module is designed to extract attention weights. Its structure, depicted in Fig. 9, includes two essential components: Squeeze and Excitation. Global average pooling is applied to calculate the mean of each channel, compressing the feature maps into numerical information with a shape of $1 \times 1 \times C$. This information is then processed through two linear layers with Rectified Linear Unit (ReLU) activation and a sigmoid activation function to obtain the weight for each channel.

By combining the feature maps obtained after weighted channel attention, a new feature map containing multi-scale information is generated. In our implementation, we inserted the PSA Module into DenseNet121, replacing the 3×3 convolution layer in the Dense Block. It results the PSA-DenseNet model, depicted in Fig. 10.

In conclusion, the PSA-DenseNet model offers significant technical advancements, including a reduction in parameter count and computational costs compared to existing models, while enhancing the model's ability to capture critical features and establish long-term dependencies of information. The incorporation of the PSA Module enables the model to prioritize important parts of the input data and effectively utilize crucial contextual information from different parts of the input during processing, thereby improving overall performance. Additionally, by allowing the model to attend to specific regions of interest within the input, the attention mechanism enhances its ability to capture intricate patterns and relationships within the data, resulting in more accurate and robust predictions.

2.3.4. MPSA-DenseNet model

In this subsection, we present our third model, the MPSA-DenseNet. The MPSA-DenseNet model integrates both the PSA Module attention and multi-task learning into DenseNet, intending to explore their combined effect on classification performance.

Table 2
Structure of MPSA-DenseNet model.

Layers	MPSA-DenseNet	Output size
Convolution	7 × 7, stride 2, 64 channels	32 × 172
Block 1	$[\begin{smallmatrix}1\times1Conv\\PSAModule\end{smallmatrix}]\times6, 448channels$	16 × 86
Transition layer 1	$[\begin{smallmatrix}1\times1Conv\\AvgPool(size=2)\end{smallmatrix}], 208channels$	8 × 43
Block 2	$[\begin{smallmatrix}1\times1Conv, 256\\PSAModule\end{smallmatrix}]\times12, 992channels$	8 × 43
Transition layer 2	$[\begin{smallmatrix}1\times1Conv\\AvgPool(size=2)\end{smallmatrix}], 496channels$	4 × 21
Block 3	$[\begin{smallmatrix}1\times1Conv, 256\\PSAModule\end{smallmatrix}]\times24, 2032channels$	4 × 21
Transition layer 3	$[\begin{smallmatrix}1\times1Conv\\AvgPool(size=2)\end{smallmatrix}], 1016channels$	2 × 16
Block 4	$[\begin{smallmatrix}1\times1Conv, 256\\PSAModule\end{smallmatrix}]\times16, 1400channels$	2 × 10
Classification layer	GlobalAvgPool(size = 1) 5d/2d/6d fc, Softmax	1 × 1

The rationale behind this combination is that, in multi-task learning, there may be differences in the data and feature distributions across different tasks. By using PSA Module attention, the model can adaptively extract features of different scales and types based on the needs of each task, enhancing its representation ability. Additionally, the characteristics of multi-task learning and the PSA Module, as discussed in the previous subsections, suggest that the MPSA-DenseNet can significantly reduce memory requirements and the model's size while improving generalization ability without affecting performance.

The structure of MPSA-DenseNet is shown in Table 2. Its input is a sequence of MFCC feature maps, which are processed by a stack of dense blocks and transition layers. Each dense block contains several convolutional layers with batch normalization and ReLU activation functions. Meanwhile, the transition layers are used to reduce the spatial dimensions of the feature maps while increasing the number of channels.

In addition to the dense blocks and transition layers, MPSA-DenseNet also includes two auxiliary heads for the two tasks in our multi-task learning scenario. These auxiliary heads provide additional supervision for each task during training, thereby aiding in the overall optimization of the model. Each auxiliary head consists of a global average pooling layer followed by a fully connected layer with a softmax activation function. The outputs of the two heads are combined with a weight factor during training to achieve a balance between the two tasks, ensuring the effective contribution of both tasks to the final classification.

In conclusion, our MPSA-DenseNet is a comprehensive deep learning architecture that leverages attention mechanisms, multi-task learning, and DenseNet. This combination aims to achieve superior classification performance on the accent classification dataset. MPSA-DenseNet follows a bottom-up development approach, where components are added step-by-step to address identified weaknesses in previous stages while increasing model complexity. MPSA-DenseNet is particularly well-suited for accent classification due to several key features:

- Hierarchical Feature Extraction: The integrated PSA module enhances the model's ability to capture and utilize hierarchical features, which are essential for understanding the subtle variations in speech patterns across different accents.
- Rich Contextual Information: The PSA module also enables the model to learn from richer contextual information within the speech data. This allows MPSA-DenseNet to develop more discriminative features for accurate accent classification.
- Attention to Detail: The attention mechanism helps the model to focus on specific segments of the speech signal that are most informative for identifying accent-related characteristics. This targeted focus improves the overall classification accuracy.
- Enhanced Generalization: By incorporating multi-task learning, MPSA-DenseNet can learn multiple related tasks simultaneously. This process allows the model to capture a wider range of features and patterns, ultimately leading to better generalization capabilities for the task of accent classification.

2.3.5. Evaluation metrics

We divide the dataset into three subsets: training, validation, and test sets, in the ratio of $6:2:2$. First, we train the models on the training set and then use their classification accuracy on the validation set as the objective function to tune the hyperparameters. The parameter setting that maximizes the accuracy on the validation set is chosen. Utilizing these chosen parameter settings, we retrain the models on the union of the training set and validation set to increase the models' accuracy. The test set was used to evaluate the classification performance of the models. Despite attempts to balance the distribution of data labels during data collection, imbalances persisted, especially with age labels in multi-task learning methods where the label distribution was evident. Therefore, relying solely on the accuracy of the model training is insufficient in expressing the classification performance of our model.

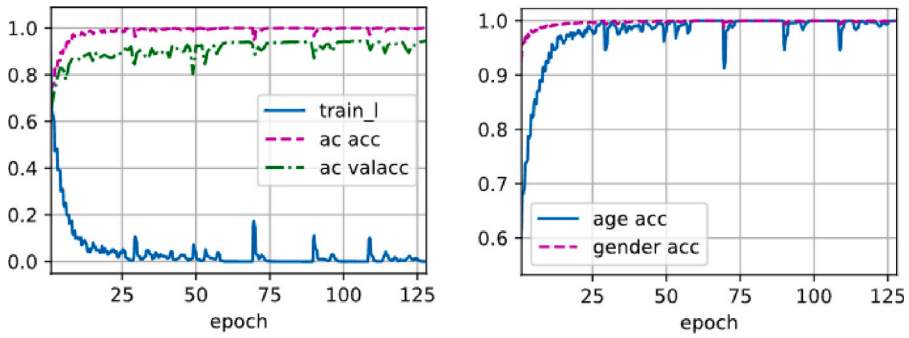

Fig. 11. MPSA-DenseNet: Training loss & training/validation accuracy (left), training gender/age accuracy (right).

To ensure a more reliable comparison of the performance of our model, we use the F_β score as an evaluation metric. F_β score is an effective measure for imbalanced data sets in classification problems, which is defined by

$$F_\beta = (1+\beta^2) \times \frac{\text{Precision} \times \text{Recall}}{(\beta^2 \times \text{Precision}) + \text{Recall}}.$$

The value of β affects the proportion of prediction and recall in the evaluation index.

2.3.6. Environment and hyperparameters

Our study uses the CUDA 11.0 environment, and the MXNET framework that provides a variety of APIs and has the advantage of saving memory. The Softmax Cross Entropy is used as the loss function of classification. To ensure a clear comparison of different models, each model is trained using the same set of hyperparameters. Specifically, the hyperparameters used in the models are 128 epochs, 16 batch sizes, and a learning rate of 0.0001. The optimization function used is Adam.

3. Results

In this section, we present the main results of our study, focusing on the performance of the three proposed deep learning models: Multi-DenseNet, PSA-DenseNet, and MPSA-DenseNet, compared to other models.

Fig. 11 illustrates the results for MPSA-DenseNet. The sub-figure on the left depicts the accuracy of accent identification on both the training and validation sets, plotted against the number of epochs. Additionally, the downward trend of losses on the training set is shown. On the right side, the sub-figure displays the training accuracy of the gender and age auxiliary tasks. It is worth noting that for the PSA-DenseNet model since we do not incorporate multiple tasks, there is no figure depicting the training accuracy of the gender and age auxiliary task.

From the figure, we observe that as the number of epochs increases, the accuracy of accent identification improves while the training loss decreases. This indicates that the models are learning and adapting to the accent classification task over time. Additionally, the training accuracy of the gender and age auxiliary tasks shows positive progress, demonstrating the effectiveness of the multi-task learning approach in these models.

To gain deeper insights into the quantitative results, we analyze the normalized confusion matrix obtained from the MPSA-DenseNet's accent classification on the test set, as depicted in Fig. 12. The diagonal elements of the matrix represent the ratio of data files correctly classified, revealing the exceptional performance of MPSA-DenseNet. The model demonstrates robust performance across multiple accents in terms of recognition accuracy, particularly excelling in the German accent with an impressive recognition percentage of 99.0%.

Fig. 12 also highlights that the highest misclassification percentage is only 6.9%, specifically misclassifying England accents as American accents, which shows that the difference between native speakers' pronunciation is small and the model needs to learn more subtle differences. Meanwhile, other misclassifications remain below 5.6%. This affirms the model's proficiency in accent recognition and its minimal misclassification ratio.

Let us now evaluate the performance of our three proposed models (MPSA-DenseNet, PSA-DenseNet, and Multi-DenseNet) against six existing architectures: EPSANet, DenseNet, Convolutional Block Attention Module with CNN (CBAM-CNN), Long Short-Term Memory (LSTM), ResNet, and MobileNetV2. Fig. 13 illustrates the results on the test set in the decreasing order of accuracy. As the figure shows, MPSA-DenseNet outperforms all six baselines and the other two proposed models (PSA-DenseNet and Multi-DenseNet) across all evaluation metrics (Micro, Macro, Accuracy). In addition, while PSA-DenseNet and Multi-DenseNet do not achieve the same level of performance as MPSA-DenseNet, they still exhibit competitive performance compared to the six baseline models.

We further investigate the peak memory usage during training for all nine models, presented in Table 3. The table reveals that models incorporating attention mechanisms, such as PSA-DenseNet and EPSANet, require more memory compared to others. This is likely due to the additional computations involved in the attention modules. Conversely, parameter sharing inherent to multi-task learning strategies reduces the total number of parameters. This is evident in the significantly lower peak memory usage of Multi-DenseNet compared to other models. MPSA-DenseNet, which also leverages multi-task learning, demonstrates a lower peak memory

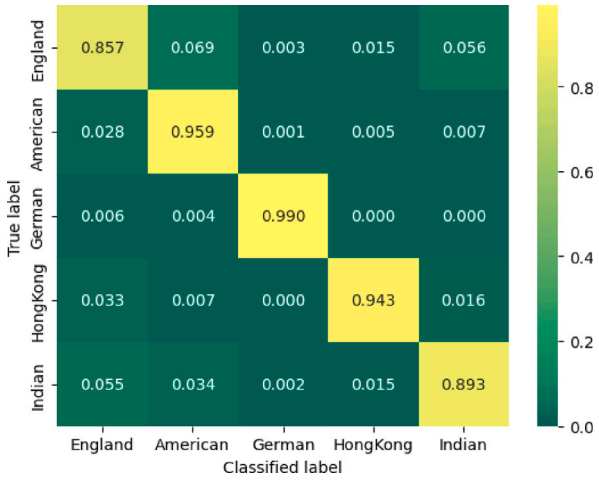

Fig. 12. Normalized confusion matrix for accent classification by MPSA-DenseNet on test set.

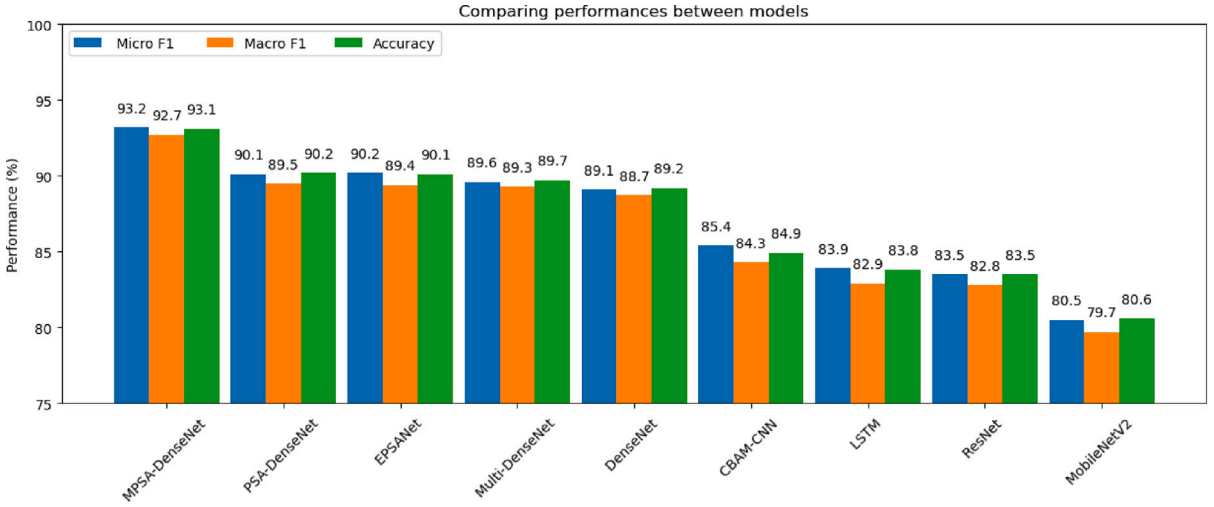

Fig. 13. Micro F_1, Macro F_1, and accuracy of nine models on test set.

Table 3
Peak memory required during training.

Model	Peak memory usage (MB)
MPSA-DenseNet	5432.759
PSA-DenseNet	5743.988
Multi-DenseNet	3141.635
EPSANet	5631.294
DenseNet	3955.400
CBAM-CNN	4497.276
LSTM	2511.745
ResNet	3711.592
MobileNetV2	2687.320

footprint than PSA-DenseNet and EPSANet, benefiting from both the reduced parameter count and potentially more efficient memory management within its architecture.

Furthermore, to assess the reliability and stability of our models, we calculate Micro and Macro F_1 scores for MPSA-DenseNet on test sets with varying dataset sizes. (All other parameters such as the ratio between training, validation, test sets are the same as stated in the Materials and Methods section.) Fig. 14 illustrates that as the training set size increases, these scores also increase. This suggests that MPSA-DenseNet is stable and reliable.

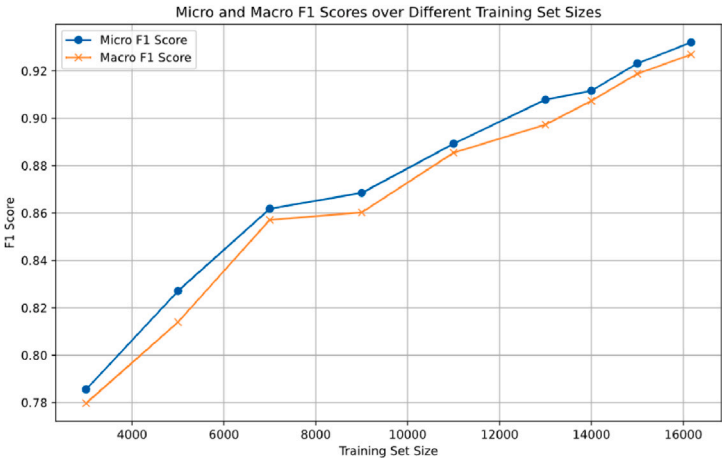

Fig. 14. Impact of training set size on Micro and Macro F_1 scores.

In conclusion, MPSA-DenseNet demonstrates outstanding performance, while Multi-DenseNet and PSA-DenseNet show competitive results, outperforming established models such as EPSANet, DenseNet, CBAM-CNN, LSTM, Resnet, and MobileNetV2 in accent classification tasks.

The superior performance of MPSA-DenseNet is achieved through the synergistic roles of its three key components: DenseNet, multi-task learning, and attention mechanisms. Dense layers provide robust feature extraction and representation learning capabilities. Multi-task learning enables the model to learn shared representations and enhances generalization by training on multiple related tasks simultaneously. The attention mechanism improves the model's ability to focus on the most relevant features, dynamically adjusting its focus to enhance performance on individual tasks.

4. Conclusion

In this study, we have introduced three novel models, MPSA-DenseNet, PSA-DenseNet, and Multi-DenseNet, designed to accurately classify English accents from both native and non-native speakers. By synergizing multi-task learning and the PSA module attention mechanism with DenseNet, we have achieved a remarkable improvement in classification accuracy, especially with the MPSA-DenseNet model, which outperforms all other models.

Our findings underscore the power of combining different components, as each model performs better than using the components individually. The MPSA-DenseNet model shows exceptional generalization capabilities by achieving high accuracy across all five English accent types present in the dataset.

While acknowledging the significant memory resources required for training the DenseNet architecture combined with multi-task learning and attention mechanism, we propose exploring lighter-weight network architectures, such as CondenseNetV2 (Yang et al., 2021), in future research endeavors.

In summary, our study presents an efficient approach to English accent classification, paving the way for further research in this area. The results highlight the potential of the three models, especially the MPSA-DenseNet model, as a promising solution for accent recognition tasks. Furthermore, they are likely applicable to various other domains, such as call-based bird classification in ecology or audio-based transport vehicle classification in engineering.

CRediT authorship contribution statement

Tianyu Song: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Linh Thi Hoai Nguyen:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Ton Viet Ta:** Writing – review & editing, Supervision, Project administration, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare no conflicting interest in this study.

Data availability

Data will be made available on request.

Acknowledgments

The authors heartily express their gratitude to the anonymous reviewers for suggestions, which greatly improved the paper. The second author (LTH Nguyen) was supported by I^{2}CNER's Start-up research funds WPI Academy UFJG050103, UFJG060103. The last author (TV Ta) was supported by the Tokuo Fujii Research Fund – Support for Article Processing Charge of International Academic Papers.

References

Al-Jumaili, Z., Bassiouny, T., Alanezi, A., Khan, W., Al-Jumeily, D., Hussain, A.J., 2022. Classification of spoken english accents using deep learning and speech analysis. In: Intelligent Computing Methodologies: 18th International Conference, ICIC 2022, Xi'an, China, August 7–11, 2022, Proceedings, Part III. Springer International Publishing, Cham, pp. 277–287.

Angkititrakul, P., Hansen, J.H., 2006. Advances in phone-based modeling for automatic accent classification. IEEE Trans. Audio Speech Lang. Process. 14 (2), 634–646.

Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F.M., Weber, G., 2020. Common voice: A massively-multilingual speech corpus. In: Proceedings of the 12th Conference on Language Resources and Evaluation. LREC 2020, pp. 4211–4215.

Bhowmik, T., Choudhury, A., Sharma, A., Verma, A., Kanthalia, B., Roy, B., 2022. A comparative study on native and non-native english accent classifications. In: 2022 International Conference on Futuristic Technologies. INCOFT, IEEE, pp. 1–6.

Bird, J.J., Wanner, E., Ekárt, A., Faria, D.R., 2019. Accent classification in human speech biometrics for native and non-native english speakers. In: Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments. pp. 554–560.

Biswas, A., 2023. The role of audio features in accent recognition: A comparative analysis. In: 2023 International Workshop on Intelligent Systems. IWIS, IEEE, pp. 1–5.

Carofilis, A., Alegre, E., Fidalgo, E., Fernández-Robles, L., 2023. Improvement of accent classification models through grad-transfer from spectrograms and gradient-weighted class activation mapping. IEEE/ACM Trans. Audio Speech Lang. Process. 31, 2859–2871.

Caruana, R., 1997. Multitask learning. Mach. Learn. 28, 41–75.

Chen, H., Zhu, S., 2023. Combining multiscale learning and attention mechanism densely connected network for single image deraining. Signal Image Video Process. 17, 2645–2652.

Chionh, K., Song, M., Yin, Y., 2018. Application of Convolutional Neural Networks in Accent Identification. Project Report, Carnegie Mellon University, Pittsburgh, Pennsylvania.

Crawshaw, M., 2020. Multi-task learning with deep neural networks: A survey. arXiv preprint arXiv:2009.09796.

Gao, Q., Wu, H., Sun, Y., Duan, Y., 2021. An end-to-end speech accent recognition method based on hybrid ctc/attention transformer asr. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, IEEE, pp. 7253–7257.

Hossan, M.A., Memon, S., Gregory, M.A., 2010. A novel approach for MFCC feature extraction. In: 2010 4th International Conference on Signal Processing and Communication Systems. pp. 1–5.

Hou, J., Liu, Y., Zheng, T.F., Olsen, J., Tian, J., 2010. Multi-layered features with SVM for Chinese accent identification. In: 2010 International Conference on Audio, Language and Image Processing. pp. 25–30.

Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4700–4708.

Ittichaichareon, C., Suksri, S., Yingthawornsuk, T., 2012. Speech recognition using MFCC. In: International Conference on Computer Graphics, Simulation and Modeling, Vol. 9.

Jiao, Y., Tu, M., Berisha, V., Liss, J.M., 2016. Accent identification by combining deep neural networks and recurrent neural networks trained on long and short term features. Interspeech 2388–2392.

Kethireddy, R., Kadiri, S.R., Gangashetty, S.V., 2020. Learning filterbanks from raw waveform for accent classification. In: 2020 International Joint Conference on Neural Networks. IJCNN, IEEE, pp. 1–6.

Kumpf, K., King, R.W., 1996. Automatic accent classification of foreign accented Australian english speech. In: Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96, Vol. 3, IEEE, pp. 1740–1743.

Liu, X., He, P., Chen, W., Gao, J., 2019. Multi-task deep neural networks for natural language understanding. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Florence, Italy, pp. 4487–4496.

Lou, Z., Ren, Y., 2021. Investigating issues with machine learning for accent classification. In: Journal of Physics: Conference Series ECNCT 2020. vol. 1738, IOP Publishing.

McFee, B., Raffel, C., Liang, D., Ellis, D.P., McVicar, M., Battenberg, E., Nieto, O., 2015. Librosa: Audio and music signal analysis in python. In: Proceedings of the 14th Python in Science Conference. vol. 8, pp. 18–25.

Mulimani, M., Mesaros, A., 2024. Class-incremental learning for multi-label audio classification. arXiv preprint arXiv:2401.04447.

Naini, A.R., Singhal, B., Ghosh, P.K., 2022. Attention pooling network for recording device classification using neutral and whispered speech. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, IEEE, pp. 8487–8491.

Najafianand, M., Russell, M., 2020. Automatic accent identification as ananalytical tool for accent robust automatic speech recognition. Speech Commun. 122, 44–55.

Pedersen, C., Diederich, J., 2007. Accent classification using support vector machines. In: 6th IEEE/ACIS International Conference on Computer and Information Science. ICIS 2007, pp. 444–449.

Rizwan, M., Anderson, D.V., 2018. A weighted accent classification using multiple words. Neurocomputing 277, 120–128.

Sharma, M., 2022. Multi-lingual multi-task speech emotion recognition using wav2vec 2.0. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, IEEE, pp. 6907–6911.

Tang, H., Ghorbani, A.A., 2003. Accent classification using support vector machine and hidden Markov model. In: Advances in Artificial Intelligence: 16th Conference of the Canadian Society for Computational Studies of Intelligence, AI 2003, Halifax, Canada, June 11–13, 2003, Proceedings, Vol. 16. Springer, Berlin Heidelberg, pp. 629–631.

Terrell, G.R., Scott, D.W., 1992. Variable kernel density estimation. Ann. Statist. 1236–1265.

Tomar, Suramya, 2006. Converting video formats with FFmpeg. Linux J. 2006 (146), 10.

Wubet, Y.A., Balram, D., Lian, K.Y., 2023. Intra-native accent shared features for improving neural network-based accent classification and accent similarity evaluation. IEEE Access 11, 32176–32186.

Yang, L., Jiang, H., Cai, R., Wang, Y., Song, S., Huang, G., Tian, Q., 2021. Condensenet v2: Sparse feature reactivation for deep networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3569–3578.

Zaman, K., Sah, M., Direkoglu, C., Unoki, M., 2023. A survey of audio classification using deep learning. IEEE Access.

Zeng, Y., Mao, H., Peng, D., Yi, Z., 2019. Spectrogram based multi-task audio classification. Multimedia Tools Appl. 78, 3705–3722.

Zhang, H., Zu, K., Lu, J., Zou, Y., Meng, D., 2022. E(psanet): An efficient pyramid squeeze attention block on convolutional neural network. In: Proceedings of the Asian Conference on Computer Vision. pp. 1161–1177.