IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 14 June 2024, accepted 13 July 2024, date of publication 25 July 2024, date of current version 2 August 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3430296

RESEARCH ARTICLE

Mel-Scale Frequency Extraction and Classification of Dialect-Speech Signals With 1D CNN Based Classifier for Gender and Region Recognition

HSIANG-YUEH LAI, CHIA-CHIEH HU, CHIA-HUNG WEN, JIAN-XING WU, (Member, IEEE), NENG-SHENG PAI, CHENG-YU YEH, AND CHIA-HUNG LIN

Department of Electrical Engineering, National Chin-Yi University of Technology, Taichung City 41170, Taiwan

Corresponding authors: Neng-Sheng Pai (pai@ncut.edu.tw) and Chia-Hung Lin (eechl53@gmail.com)

This work was supported by the National Science and Technology Council (NSTC) under Contract NSTC 112-2635-E-167-001 and Contract NSTC 112-2221-E-167-015 (August 2023–July 2024).

ABSTRACT Humans communicate and interact through natural languages, such as American English (AE), Taiwanese, Italian, and numerous variants of Spanish. Through automatic speech analysis and recognition technologies, human–machine interaction systems (HMISs) can be used for language learning in query systems, smart devices, and healthcare applications, emphasizing the need to enhance user interaction across different sectors. Because people differ in their basic attributes (e.g., gender, age group, and spoken dialect), an HMIS must be able to identify the speaker's gender, age group, and regional dialect on the basis of their speech signals. To achieve automatic speech recognition, we analyzed and distinguished feature patterns using a feature extraction method and identified gender and region using a convolutional neural network (CNN)-based classifier. Mel-frequency cepstral coefficients were used to extract Mel-scale frequencies (MSF) from dialect-sentence speech signals for conversion into specific feature patterns. Subsequently, a one-dimensional CNN-based classifier was used to identify these features patterns by gender and regional dialect. The proposed speech classifier was rigorously trained, tested, and validated using dialect-sentence speech corpora from AE, Italian (IT), and Spanish (SP) acoustic–phonetic continuous speech database. The experimental results indicate that the proposed model with MSF features can perform accurate gender and region recognition. The classifier was evaluated in metrics of precision (%), recall (%), F1 score, and accuracy (%).

INDEX TERMS Automatic speech recognition (ASR), Mel-scale frequency, one-dimensional convolutional neural network (CNN), dialect-sentence speech signal, acoustic-phonetic continuous speech.

I. INTRODUCTION

Human natural languages (NLs), such as Taiwanese, American English (AE), Latin, Japanese, and Chinese, are the primary means for humans to communicate during social activities. They are also communication tools that can transmit, express, describe, preserve, and exchange information and knowledge, experiences, and culture. Distinct dialectal variations have been developed for these languages across different countries, geographical environments, and regions. Even within the same country, distinct honorifics or special vocabulary may be used in communication between speakers of different genders, age groups and social classes. For example, TIMIT (Texas Instruments/Massachusetts Institute of Technology) APCSD (DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus [USA]) contains 6,300 dialect sentences (630 speakers) from eight major dialect regions (North, South, and West regions of the United States (#1-#7)

The associate editor coordinating the review of this manuscript and approving it for publication was M. Sabarimalai Manikandan.

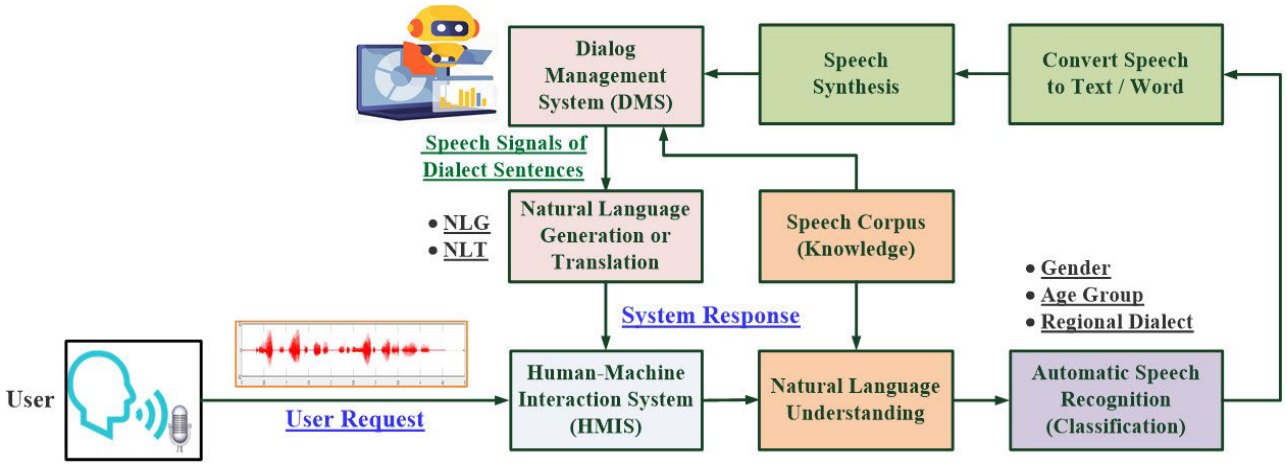

FIGURE 1. Architecture of natural language user interface system for natural language processing, understanding, and generation/translation.

and covers multiple dialects from the United States (#8)) [1], [2]. In speech recognition tasks, these dialect sentences can be used to identify the gender, age group, and regional dialect of a speaker; they can also be used to verify the performance of automatic speech recognition (ASR) systems. ASR systems and human– machine interaction systems (HMISs) [3], [4], [5], [6] have two operation modes, namely the reactive communication mode (RCM) and interactive communication mode (ICM) [4], [5], [6]. In the ICM, users can issue commands through the microphone units. The HMIS then preprocesses the input speech signals, analyzes the grammatical components (syntax analysis) and meaning of words (semantic analysis), issues the user's speech commands, and synthesizes an audio response through speech synthesis; alternatively, the machine translation function can automatically translate words or texts from one language to another [7]. Through NL generation (NLG) and NL translation (NLT), users can receive feedback messages, and the responses can be generated in subsequent interactions between any user and an HMIS; for example, the ICM-based system depicted in Figure 1 has been used in smartphone speech assistants [8] and Internet of Vehicles (IoV) language applications for speech navigation [9].

In an NL user interface (as seen in Figure 1), the HMIS interacts with a user by first identifying the user's gender, age group, and regional dialect. Subsequently, through NL understanding (NLU), ASR is applied for distinguishing the user's personal features from acoustic speech signals. An NLU module is employed to perform syntax and semantic analysis and subsequently decipher the sentence structure and meaning [7]. An ASR system is used to perform sentence classification for identifying a user's gender, age group, and regional dialect; thereafter, the input sentences are converted into texts or words, and an NLG module automatically produces the appropriate responses to the user on the basis of a speech corpus (Knowledge) [10], such responses include automated reports, product descriptions, and speech navigation information [4], [5], [6], [7]. Over several conversations with users, the reaction time of an HMIS is gradually shortened, making human–computer interactions more efficient. Consequently, the needs of different users across regions with high linguistic diversity can be met, paving the pathway to global applications. Moreover, both user satisfaction and HMIS performance can be improved.

To provide a customized and intelligent HMIS for meeting the user request, the feature extraction must be performed on incoming speech signals. Feature extraction tools and techniques include the digital filters, Mel- frequency cepstral coefficients (MFCC), YIN (name based on the Oriental concept of yin and yang), and Yet Another Algorithm for Pitch Tracking (YAAPT) [1], [11], [12], [13]. Classifier methods based on artificial intelligence (AI), deep learning (DL), and machine learning (ML) have been employed in ASR applications, such as multilayer perceptrons, support vector machine (SVM), WAV2VEC2 model (self-supervised learning [SSL]), fully convolutional network (FCN), deep neural network with hidden Markov model (DNN-HMM), recurrent neural network (RNN), and one-dimensional (1D) and two-dimensional (2D) convolutional neural networks (CNNs) [1], [4], [5], [6], [14], [15], [16], [17], [18], [19], [20]. These AI based methods are used to train a classifier on a large number of dialect–sentence speech datasets, thereby enabling the correct identification and interpretation of incoming speech signals. These methods [1], [4], [5], [6], [14], [15], [16], [17], [18], [19], [20] were trained on speech corpora from TIMIT [2], Hillenbrand [12], [13], the Ryerson Audio–Visual Database of Emotional Speech and Song (RAVDESS), and BGC (Bangladesh) [21].

In the present study, we applied ASR to dialect-speech signals to identify the gender, age group, and regional dialect of a user; this process involves digital signal preprocessing, feature extraction, and signal classification tasks. For digital

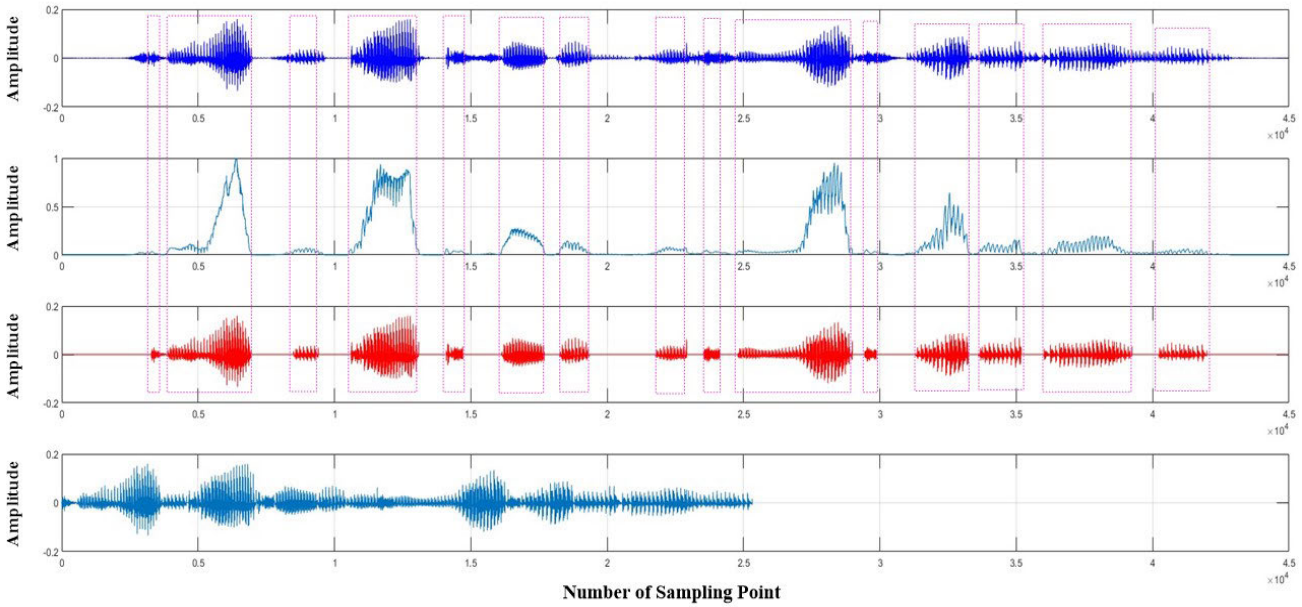

FIGURE 2. Speech signals are preprocessed during natural language processing through digital filter and endpoint detection to segment interesting content (the sentence "She had your dark suit in greasy wash water all year" was obtained from the TIMIT Speech Corpus).

signal preprocessing, digital filters (100–1100 Hz bandpass Butterworth filter) or Chebyshev filters [22], [23] were used to remove redundant data, such as unwanted artifacts and environmental noise. Furthermore, endpoint detection (EPD) algorithms [24], [25] were used for signal activity detection (SAD) and weighted operations, which facilitate speech signal segmentation and the determination of the beginning and end points of speech activities in signal amplitude variations, as shown in Figure 2; for a given specific threshold value, significant signal variations can be used to directly distinguish speech signals from nonspeech parts for signal denoising and segmentation processing, allowing for substantive content to be extracted from dialect-sentence speech. For the signal segmentation processing, EPD was performed to identify the boundaries between sentences in a piece of text for real-time NL processing (as seen in Figure 2). During feature extraction, MFCC-based methods [1], [11], [12], [26] were used to extract the frequency-based parameters of speech signals, such as audio signal frequency and amplitude. MFCCs serve as an extractor that converts time-domain signals into Mel spectrograms in visualizations. Hence, in speech signal preprocessing, filtering and segmentation are performed to remove background noise and extract interesting fragments from the incoming speech-signal stream. To extract features from interesting fragments, we must consider that a native AE subject and a non-native AE subject have different speech acoustic features depending on their country, gender, and regional dialect. This also applies to other languages and is useful in applications where linguistic diversity matters, as is the case in Taiwanese, Latin American Spanish, Japanese, and Chinese.

In speech classification, a cascaded 1D CNN [1], [11], [12], [14], [26], [27] is used to train a speech classifier, such that its recognition scheme can identify users by gender and regional dialect. The proposed 1D-CNN-based classifier comprises a feature extraction layer (MFCC-based extractor), two cascaded 1D convolutional–pooling layers with multikernel windows and maximum-pooling (Max-pooling) windows, a flattening layer, and a fully connected layer (two dense layers). Two dropout layers are inserted into two dense layers, and 10% of the neurons are randomly deactivated to overcome the overfitting problem during the training stage. When a Mel spectrogram is produced as an input feature pattern, a 1D CNN is used to extract the feature patterns for training classifier. Subsequently, the Gaussian error linear unit (GeLU) activation function [28], [29] is used to activate neural nodes and to identify the possible classes from incoming speech signals. Finally, the final output is obtained using the Softmax activation function in the output layer. In the training stage, the adaptive moment estimation (ADAM) algorithm [28], [29], [30] is used to refine optimal connected weighted parameters and bias parameters through iteration computations, thereby minimizing the loss function (LF) value with the specific threshold value. In this study, a 2D-CNN model [30] with two cascaded 2D convolutional–pooling layers is also used to implement the speech classifier. Three dialect speech databases, namely the TIMIT APCSD [2], Corpora e Lessici dell' Italiano Parlato e Scritto (CLDPS) [31], and Sound-board-Learn Spanish (SLS) [32] speech corpus, are used. The content from these databases, which have AE, Italian (IT), and Spanish (SP)-language speech content, is divided into training datasets for training the classifier and testing datasets for validating the classifier's

effectiveness for performing gender and region recognition. Four indices, namely precision (%), recall (%), accuracy (%), and F1 score [12], [30], are used to evaluate the classifier's performance.

II. METHODOLOGY AND MATERIALS

A. SPEECH SIGNALS PREPROCESSING

The Butterworth filter is used for speech signal preprocessing, which is designed as a band-pass filter by using the MATLAB syntax "butter (•)" with the fifth-order filter [33]; its voltage gain is 60 dB and stable; and the cutoff frequency range is set between 100 Hz and 1100 Hz for human voice (fundamental frequency ranges of 85-255 Hz for female and male speakers [12], [13]); this range is used to exclude unwanted frequencies, such as the 60/50-Hz power-grid frequency and environment noises. For segmentation processing, the EPD is performed to extract interesting content from speech sentences, which can be expressed as follows:

$$E[n] = \sum_{k=1}^{N_k} (x_n(k))^2 \geq \theta, n = 1, 2, 3, \ldots, N - N_k, k = 1, 2, 3, \ldots, N_k \quad (1)$$

In a speech-signal stream, as seen in Figure 2, major fluctuations in amplitude changes are more pronounced in violent-activity segments than in stationary segments. Therefore, at sampling point n, where a specific threshold value is applied, as $\theta = \mathbf{max}(E[n]) \times 0.1$ for all n; which indicates amplitude changes within a short period to identify obvious signal activities and to determine the interesting content; and the length is $(N - N_k)$, where $E[n]$ denotes short-term amplitude variations at the nth sampling point, $x_n(i)$ denotes the time-domain raw speech signal, and N_k is the length of the frame processing. Thus, interesting content can be extracted from spoken dialect sentences.

B. MFCC-BASED FEATURE EXTRACTOR

An MFCC-based feature extractor is a short-term power spectrum transformation tool that can perform fast computations for directly extracting speech features from speech signals [1], [11], [20]. During this process, a discrete cosine transform (DCT) uses a finite sequence of sampling data to compute a sum of sinusoids with varying magnitudes and frequencies, and this sum is implemented on a power spectrum with a nonlinear Mel scale to enable key feature selection for pattern recognition [34]. For feature extraction processing, speech signals are analyzed frame-by-frame (framing process) by using a window function [35], [36], [37], such as Hamming, Hanning, Kaiser, Blackman, and Gaussian window functions. Each analyzed frame is segmented into overlapping segments (such as a 15–25 ms window length with a 10–15 ms overlap). The windowing process is performed to smoothen the power spectrum, for example, the function shapes of Hamming and Hanning windows and their responses are shown in Figures 3(a) and 3(b), respectively. This study uses the Hamming window, which can enhance the speech signal continuities at the beginning and ending of a frame [37] and enable the retention of characteristic frequencies. Subsequently, each analysis window is subjected to fast Fourier transform (FFT) operation to obtain a power spectrum that converts each framing signal from the time domain to the frequency domain, as expressed in Eq. (2) [11], [20], [26], [35], [36]:

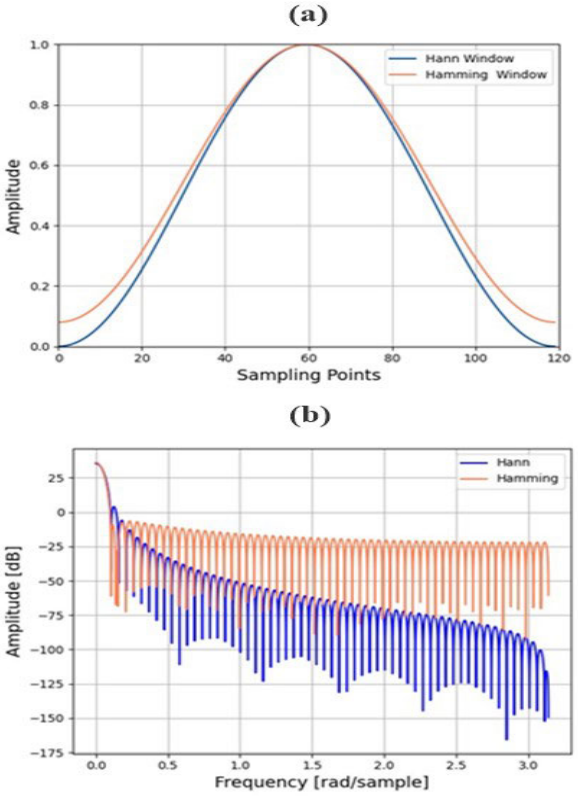

FIGURE 3. Hamming and Hanning window functions and their frequency-domain responses. (a) Function shape for Hamming and Hanning windows, (b) Windows' frequency-domain responses.

$$\mathrm{X_s\,[k]} = FFT(x_f'), \quad (2)$$

where x_f' denotes the interesting content from speech signals in the sth timing frame, $X_f[k]$ denotes the power spectrum after FFT operation, $k = 1, 2, 3, \ldots, N_k$; and N_k denotes the number of output frequency parameters. Then, multiplication operations are performed using a triangular-shaped bandpass filter (TBF) with a Mel-scale distribution, as expressed in Eqs. (3) and (4) [1], [11], [12], [26], [35], [36]:

$$\mathrm{Y\,[m]} = \log\left(\sum_{k=f_{m-1}}^{f_{m+1}} |X_s[k]|^2 B_m[k]\right), \quad (3)$$

$$\mathrm{B}_m\,[\mathrm{k}] = \begin{cases} 0, & for\ k < f_{m-1} and k > f_{m+1} \\ \dfrac{k - f_{m-1}}{f_m - f_{m-1}}, & for\ f_{m-1} \leq k \leq f_m \\ \dfrac{f_{m+1} - k}{f_{m+1} - f_m}, & for\ f_m \leq k \leq f_{m+1} \end{cases} \quad (4)$$

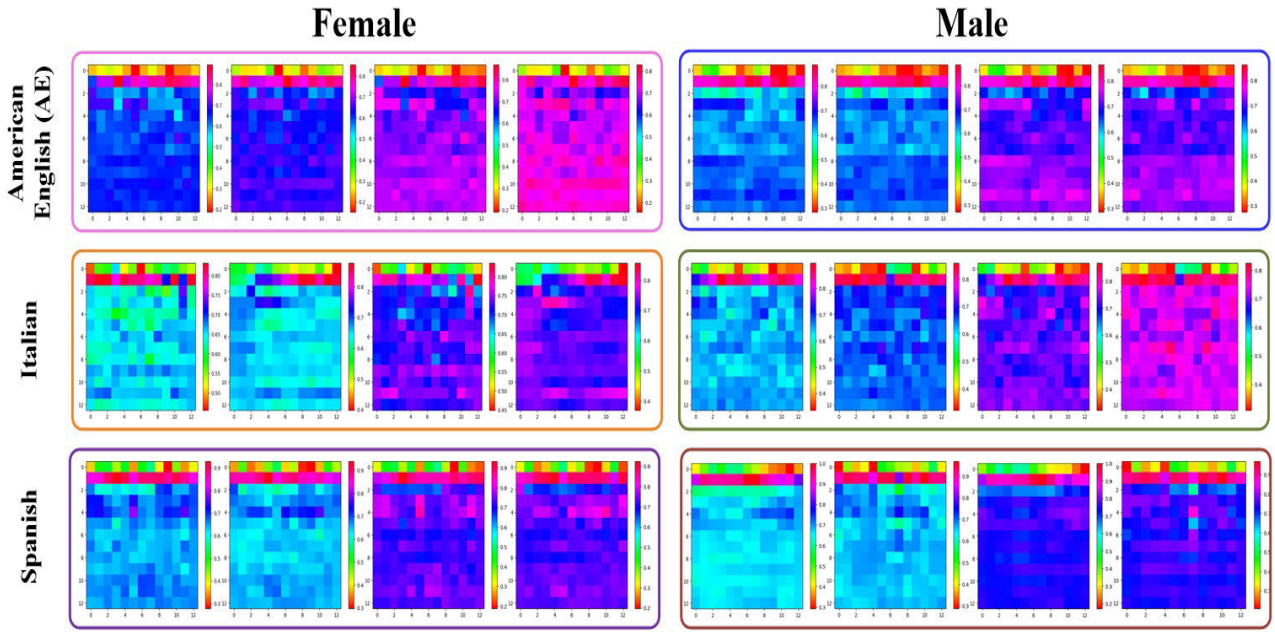

FIGURE 4. Mel spectrogram with digital filtering and MFCC processes for American English (AE), Italian (IT), and Spanish (SP) languages in female and male subjects, respectively.

where $Y[m]$ is the log-value parameter that is obtained by mapping the power spectrum to the Mel scale; $B_m[k]$ is a TBF function, whose frequency is linearly distributed below 1 kHz and increases logarithmically above 1 kHz.

The Mel frequency spectrum is obtained by applying the DCT to Y[m], as follows [11], [20], [34], [35], and [36]:

$$C_x[n] = \frac{1}{M}\sum_{m=1}^{M} Y[m]\cos\left(\frac{\pi n(m-0.5)}{M}\right), \tag{5}$$

$$\begin{cases} m = 2595\log_{10}(1+\dfrac{f}{700}) & (6) \\ f = 700(10^{m/2595}-1) & (7) \end{cases}$$

where M is the number of sampling points, which is generally half or one-third of N_n; f is the speech-signal frequency; and m is the Mel frequency. Typically, the frequency ranges of the speech signals of female and male adults are 85–180 Hz and 165–255 Hz, respectively. After DCT operation, the MFCC-based parameters are obtained when $n \geq 2$ and are used to exclude direct-current (DC) and low-frequency components. Twelve parameters (n = 2–14) related to the amplitude of the frequency provide enough obvious feature parameters for speech recognition tasks. The advantage of DCT operation can directly operate on real-part components, thereby reducing the computational complexity for frequency-domain transformation.

For example, after MFCC processes, the visualization feature patterns can be extracted from AE, IT, and SP speech signals in female and male speakers, as seen in Figure 4. To remove DC components ($n = 1$), 2D feature patterns can be produced using 12 frames with 12 feature parameters. Each visualization pattern is a 12 × 12 image in the colorful mode. In this format, a visualization pattern displays the differences in visual patterns across different genders and dialect-regional languages (as seen in Figure 4). Therefore, the MFCC-based extractor can be used to distinguish the differences in incoming speech signals for gender and region recognition.

C. CASCADED CONVOLUTIONAL NEURAL NETWORK (CNN)

A 1D CNN model can be trained as a speech classifier that combines automatic feature extraction, feature enhancement, and classification or pattern recognition [1], [12], [13], [14], [30] (as seen the multilayer structure in Figure 5). The 1D CNN has a multilayer network structure; at the convolutional operation layer, 1D convolutional operations are performed using different convolutional kernels windows with different weights for each layer. Differently weighted combinations of convolutional kernels can be used to strengthen and extract features that can increase the depth, width, dimensionality, and nonlinearity of feature patterns, thereby increasing the complexity level of feature patterns and enhancing the classifier's ability to recognize complex feature patterns. At the pooling operation layer, multi-Max-pooling processes are performed to select key feature parameters, reduce the number of feature parameters to one-fourth of the original quantity, and retain the distinctive features of incoming patterns. Thus, at the classification layer, dense networks are usually trained using the back-propagation algorithm to distinguish different feature patterns through the MFCC-based extractor (designed by Python Library).

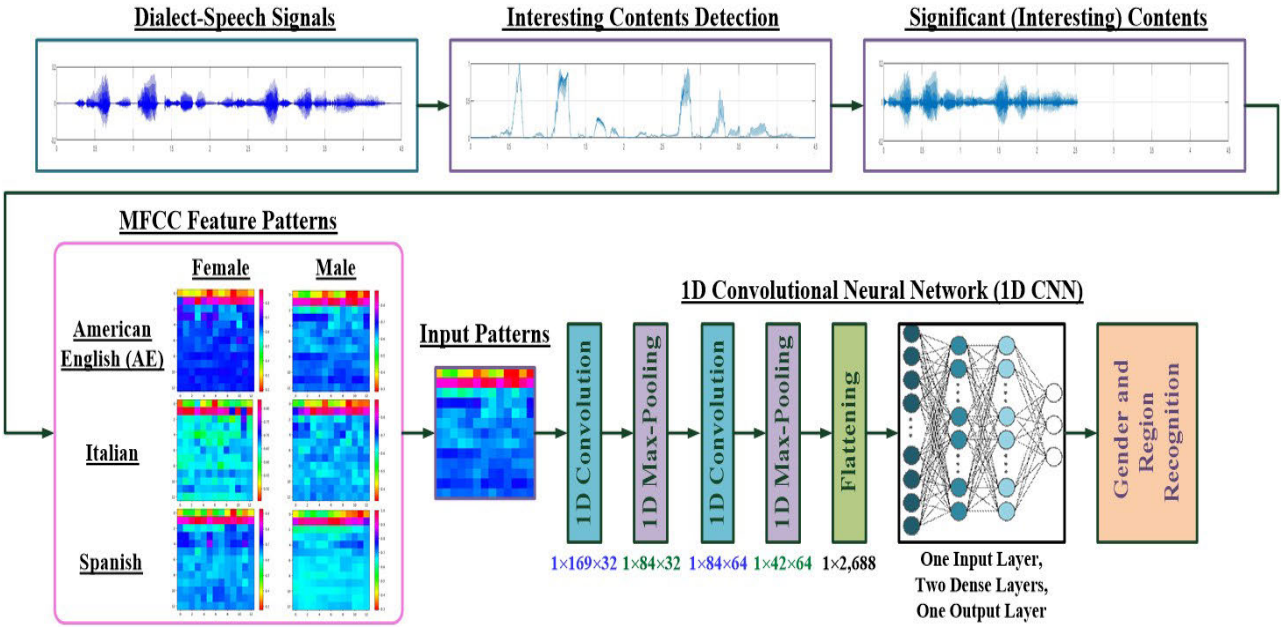

FIGURE 5. Structure of the proposed 1D-CNN classifier (two convolutional-pooling layers, one flattening layer, and two dense layers).

The present study uses the high-level API Keras development platform (an easy-to use free and open-source Python library available on Google's Cloud Services) to construct and train the DL-based classifier models (as seen the design human–machine interface in Figure 6). The 1D CNN is designed as an ASR classifier for dialect-sentence classification. The proposed classifier has multiple cascade layers: two convolutional layers (1 × 169 × 32 and1 × 84 × 64), two Max-pooling layers (1 × 84 × 32 and 1 × 42 × 64), a flattening process layer (1 × 2,688), and a fully connected layer. The fully connected layer (classification layer) comprises an input layer (2,688 nodes), two dense layers (128 and 32 nodes), and an output layer (2 nodes for gender recognition and 3 nodes for region recognition). The 2D CNN is also designed as a classifier for gender and region recognition; it consists of four convolutional layers (13 × 13 × 16, 6 × 6 × 16, and 6 × 6 × 32), a maximum pooling layer (6 × 6 × 16 and 3 × 3 × 32), a flattening process layer (1 × 288), and a fully connected layer (as seen in Figure 7). The GeLU-type and Softmax- type activation functions [30] are used in the hidden and output layers, respectively, and the implemented training scheme uses the categorical cross-entropy loss function [30] to evaluate the classifier performance for multiclass classification.

In this study, two dense layers are set (as seen in Figure 5), and the ADAM algorithm [28], [30] is used to adjust the network's bias and connecting weighted parameters. After iteration computations are completed, the values of loss function reach the predetermined convergence conditions, or the iteration computations reach the maximum iteration number, the training of the classifier is terminated. In the testing stage, the classifier outputs confusion matrices. On the basis of the actual and predicted classes, four index values are obtained: true positives (*TP*s), true negatives (*TN*s), false positives (*FP*s), and false negatives (*FN*s). On the basis of these four indices and through k-fold cross- validation, the feasibility of the classifier for gender and region recognition is verified by assessing its precision (%), recall (%), F1 score, and accuracy (%) [12], [30].

Layer (type)	Output Shape	Param #
conv1d (Conv1D)	(None, 169, 32)	128
max_pooling1d (MaxPooling1D)	(None, 84, 32)	0
conv1d_1 (Conv1D)	(None, 84, 64)	6208
max_pooling1d_1 (MaxPooling1	(None, 42, 64)	0
flatten_2 (Flatten)	(None, 2688)	0
dense_6 (Dense)	(None, 128)	344192
dropout_4 (Dropout)	(None, 128)	0
dense_7 (Dense)	(None, 32)	4128
dropout_5 (Dropout)	(None, 32)	0
dense_8 (Dense)	(None, 2)	66

Total params: 354,722
Trainable params: 354,722
Non-trainable params: 0

FIGURE 6. The 1D CNN based classifier design for gender and region recognition.

Layer (type)	Output Shape	Param #
conv2d_12 (Conv2D)	(None, 13, 13, 16)	160
conv2d_13 (Conv2D)	(None, 13, 13, 16)	2320
max_pooling2d_6 (MaxPooling2	(None, 6, 6, 16)	0
conv2d_14 (Conv2D)	(None, 6, 6, 16)	2320
conv2d_15 (Conv2D)	(None, 6, 6, 32)	4640
max_pooling2d_7 (MaxPooling2	(None, 3, 3, 32)	0
flatten_3 (Flatten)	(None, 288)	0
dense_9 (Dense)	(None, 128)	36992
dropout_6 (Dropout)	(None, 128)	0
dense_10 (Dense)	(None, 32)	4128
dropout_7 (Dropout)	(None, 32)	0
dense_11 (Dense)	(None, 3)	99

Total params: 50,659
Trainable params: 50,659
Non-trainable params: 0

FIGURE 7. The 2D CNN based classifier design for gender and region recognition.

D. SPEECH CORPUS DESCRIPTION

Various child and adult speech databases containing AE and British English (BE) speech content (acquired from various media such as movies, news broadcasts, and audiobooks) can be used to train ASR models to identify child, female, and male speakers. These databases include the child speech datasets of Carnegie Mellon University (CMU)-KIDS (24 boys and 52 girls) and PF-STAR corpora (158 children aged 4–14 years) [38], [39] and the adult speech datasets of Librispeech (1,000 h), LibriLight (60,000 h), and LJSpeech (24 h) [16], [38], [40]. These two sets of corpora differ inherently in terms of child and adult speech signals, such as high fundamental frequencies (232.17 ± 2.03 Hz for male individuals, 234.07± 1.48 Hz for female individuals) and short vocal tract lengths [12], [16]. Through pretraining and inference processes, we can use these publicly available custom datasets to improve automatic annotations for the identification of children and adult speech in ASR and text-to-speech (TTS) research [41], [42]. In the present study, we verify the proposed speech classifier using the APCSD corpus, which comprises the TIMIT [2], Hillenbrand [12], [13], [43], CLDPS [31], and SLS [32] databases. Specifically, the TIMIT and Hillenbrand databases are used for gender recognition, and the TIMIT, CLDPS, and SLS corpora are used to identify dialect-sentence speech signals for regional dialect recognition.

For example, the TIMIT speech corpus, jointly collated by the Texas Instruments (TI), Massachusetts Institute of Technology (MIT), and Stanford Research Institute International [2], [44], contains 16-bit and 16-kHz speech signals produced by 630 speakers (438 male and 192 female individuals) and encompassing eight major dialects of AE for ASR, gender, and region classification. Each enrolled participant read ten dialect sentences, such as "She had your dark suit in greasy wash water all year" (as seen in Figure 2) or "Don't ask me to carry an oily rag like that". Thus, the TIMIT speech corpus contains 6,300 dialect sentences for training and validating ASR classifiers. The Hillenbrand corpus [12], [13] comprises AE vowel (AEV) sounds collected from 45 adult men, 48 adult women, and 46 children (27 boys and 19 girls). These AEV sounds can be categorized into 12 classes, which are used to distinguish between the speech signals of speakers from different genders and age groups. The CLIPS corpus contains 100 h of diverse spoken IT content [31], including dialogs, read speeches, television programs, telephone conversations, and special corpora, collected from 15 cities in Italy. The IT language is characterized by both linguistic and demographic diversity. The speech corpus (8 bit and 8 kHz, A-law Coding) was collected from 100 IT speakers (30 female and 70 male individuals) aged between 23 and 50 years. The SLS is a speech corpus database for autodidacts to learn basic SP phrases. It contains more than 70 short phrases [32], [34], including conversations related to the following scenarios: dining and food, shopping, getting around, making reservations at high-end hotels, and making plans for a summer beach holiday.

III. EXPERIMENTAL RESULTS

The present study used the TIMIT, Hillenbrand, CLDPS, and SLS speech corpora to train and test the ASR classifiers for gender and region recognition with respect to AE, IT, and SP. We used a multicore personal computer (Intel Q370, Intel Core i7 8700, DDR4 2400 MHz 8G*3) as the base development platform for implementing the proposed classifier (as seen in Figure 5). The TensorFlow Inception V3 platform (Keras) was used to establish various classifier models with a graphics processing unit (GPU) (NVIDIA GeForce RTX 2080 Ti, 1755 MHz, 11 GB GDDR6), thereby reducing the execution time and accelerating speech signal recognition. Hence, we applied the 12 AEV classes and used the speech corpora to test and validate gender and regional-dialect recognition; subsequently, precision (%), recall (%), F1 score, and accuracy (%) [12], [30] were used to quantify the proposed classifier performance, and the experimental results are as follows.

A. GENDER IDENTIFICATION WITH TIMIT AND HILLENBRAND SPEECH CORPUSES

We used the TIMIT and Hillenbrand speech corpora to train the proposed classifier; the training and testing datasets comprised 1,668 AEV sounds [12], [13] and 6,300 AE dialect sentences [2], respectively. In speech signal preprocessing, the Butterworth filter was designed as a band-pass filter by using MATLAB syntax "butter (•)," maintaining a reasonable balance between attenuation and phase response [45].

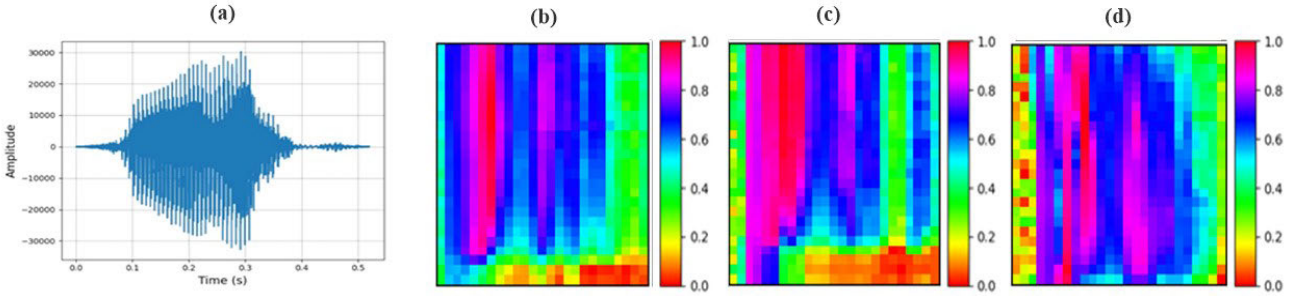

FIGURE 8. Vowel speech signal and feature patterns. (a) Speech signal for the vowel "eh" in time domain, and (b)–(d) Mel spectrogram of (b) adult male, (c) adult female, and (d) children.

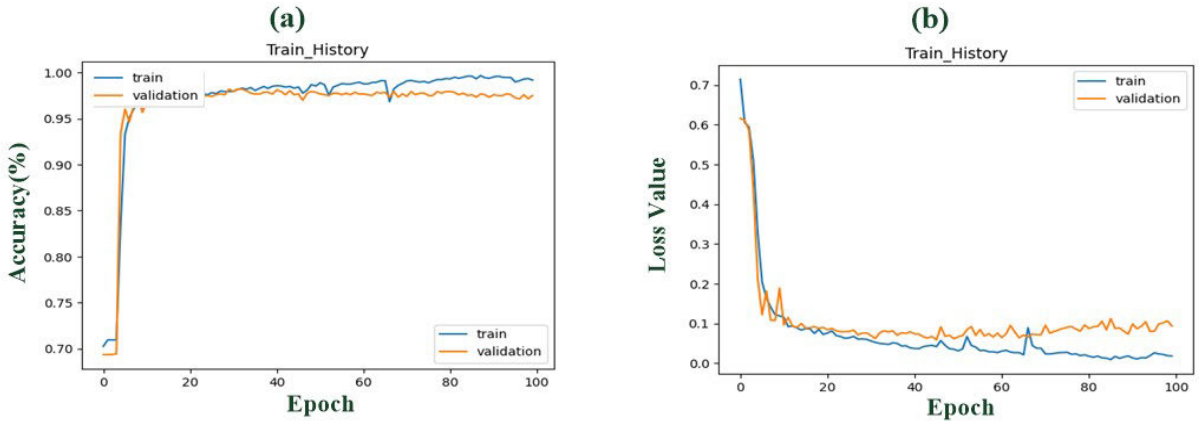

FIGURE 9. Training and validation history curves for cascaded CNN-based classifier training with TIMIT datasets. (a) Accuracy versus training epoch number for training classifier, (b) Loss function values versus training epoch number for validating classifier.

The MFCC-based extractor was used to extract the feature patterns of vowel sounds, as depicted in the Mel spectrograms for female adults, male adults, and children (as seen the vowel speech signal and feature patterns in Figures 8(a)–8(d)). For each fold cross- validation, 50% of these datasets were randomly used to train the classifier, and the remaining 50% were used for test the classifier; this is indicated by the training history curves and converging curves for training the proposed 1D classifier, which was subjected to 100 epochs of iteration computations (as seen in Figures 9(a) and 9(b), which depict "accuracy vs. number of training epochs" and "loss function values vs. number of training epochs," respectively). Average precision (%), average recall (%), and average F1 scores were estimated to assess the experimental results (as seen in Table 1) pertaining to the AEV sounds identified across different genders and age groups within vowel speeches; these values were estimated using the four indices, *TP*, *TN*, *FP*, and *FN*. Notably, the average values were greater than 90% for identified *TP*s (e.g., 98.8% for male adults, 93.8% for female adults, and 94.2% for children in terms of average F1 scores), and the overall average accuracy (%) of 95.6% was greater than 90%. That is, the experimental results indicate the feasibility of the proposed ASR model.

In addition, the TIMIT speech corpora [2] were used to train the proposed 1D classifier, including its signal preprocessing, feature extraction, interesting content detection, and gender identification, and the same classifier structure was implemented (as seen in Figure 5). After Mel spectrogram extraction was performed, the 50 × 50 feature patterns of eight dialect regions for the same sentence ("She had your dark suit in greasy wash water all year."), as spoken by female and male adults, were obtained (as seen Mel spectrograms in Figure 10). For pattern recognition, each feature pattern was resized from 50 × 50 to 13 × 13 (1 × 169) and then fed into the 1D CNN. In the present study, the data contributed by the enrolled participants to the TIMIT speech corpora were subdivided into training and testing datasets, with approximately 70%–80% being used for training and the remaining 20%–30% being used for testing. Specifically, we randomly selected 4,620 sentences (3,260 and 1,360 sentences spoken by male and female individuals, respectively) for the training datasets and 1,680 sentences (1,120 and 560 sentences spoken by male and female individuals, respectively) for the testing datasets, which were used to train and validate the proposed classifier. The criteria for gender identification during AE dialect-sentence classification were the estimated results for average

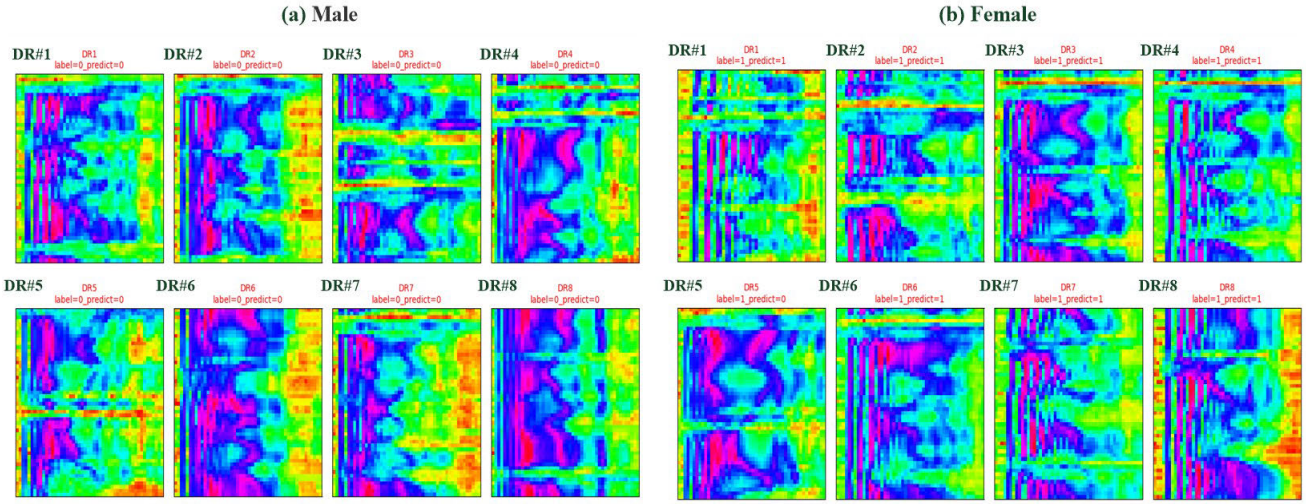

FIGURE 10. The 50 × 50 Mel spectrograms of 8 major regions for Female and Male subjects (obtained from TIMIT Speech Corpuses) with the sentence "She had your dark suit in greasy wash water all year."

precision (%), average recall (%), and average F1 score (as seen in Table 1) For example, the average F1 score was 98.0% for male adults and 96.0% for female adults, with an average accuracy of 95.6% (>95.0%) for distinguishing genders in AE dialect sentences.

B. DIALECT-REGION RECOGNITION WITH TIMIT, CLDPS, AND SLS SPEECH CORPUSES

The CLIPS speech corpus contains approximately 100 h of diverse spoken IT audio files [31], [46], including dialogues, read speeches, radio and television programs, telephone conversations, and special corpora, all of which were collected from 15 cities in Italy. The content includes map tasks, read sentences, broadcasts, talk shows, and commercials. This speech corpus (8 bit and 8 kHz, A-law Coding) was collected from 100 IT speakers (30 female and 70 male individuals) aged between 23 and 50 years. The SLS speech corpus [32] is a database for autodidacts for learning basic SP phrases; it contains more than 70 short phrases, with conversations related to situations such as dining and food, shopping, getting around, making reservations at luxurious hotels, and making plans for a summer beach holiday. We employed three databases (TIMIT, CLIPS, and SLS speech corpora) for dialect-speech classification involving gender and region recognition. We randomly selected 144 AE (72 female individuals [code: 0]; 72 male individuals [code: 1]), 144 IT (72 female and 72 male individuals), and 36 SP (24 female and 12 male individuals) sentences for training the gender classifier; 96 AE (48 female and 48 male individuals), 96 IT (48 female and 48 male individuals), and 24 SP (12 female and 12 male individuals) sentences were used to validate the classifier. With 200 iteration computations, the training history curves and converging curves for training and validating the 1D CNN classifier are presented in Figures 11(a) and 11(b), respectively (Figures 11(a), "accuracy vs. number of training epochs"; Figures 11(b), "loss values vs. number of training epochs"). The experimental results for gender recognition based on the 3 speech corpora are presented in Table 1. For example, the average F1 scores were 90.4% for male adults and 91.1% for female adults, and the average accuracy was 90.7%. For region recognition, we randomly used 144 AE (code: 0), 144 IT (code: 1), and 36 SP (code: 2) sentences to train the region classifier with the 1D CNN model and 300 iteration computations (as seen the training history curves and converging curves in Figures 11(c) and 11(d), respectively). We used 96 AE, 96 IT, and 24 SP sentences to validate the classifier; the experimental results are presented in Table 1. Therefore, for the TIMIT, CLDPS, and SLS speech corpora, the average F1 score and accuracy were more than 90% (98.5% for AE, 98.4% for IT, and 100.0% for SP), and the average accuracy was 99.1%; these results were used to quantify the classifier's performances.

The 2D CNN model (as seen in Figure 7) was also used to train an ASR classifier for identifying AE, IT, and SP. We randomly selected 120 AE (56 female and 64 male individuals), 120 IT (64 females and 56 male individuals), and 120 SP (60 female and 60 male individuals) sentences for training the gender and region classifier; 80 AE (38 female and 42 male individuals), 80 IT (42 female and 38 male individuals), and 80 SP (40 female and 40 male individuals) sentences were used to validate the classifier. The experimental results for region recognition are presented in Table 1. The average F1 score was 87.8% for male individuals and 85.1% for female individuals, and the overall average accuracy was 86.6%. Among the languages, the average F1 score was 98.7% for AE, 97.9% for IT, and 100.0% for SP, and the overall average accuracy was 98.1%. These experimental results highlight the promising performance of the 1D CNN and its superiority over the 2D CNN model for gender and region recognition.

TABLE 1. The proposed classifier for gender and region recognition in different speech corpuses (databases) and purposes.

Classifier	Speech Corpuses	Purpose
1D CNN	Hillenbrand Speech Corpuses [12-13]	**Gender Identification for Vowel Speech Classification** Average Recall(%): 98.9% for Male, 93.9% for Female, 93.9% for Children Average Precision(%): 98.7% for Male, 93.7% for Female, 94.4% for Children Average F1 Score: 98.8% for Male, 93.8% for Female, 94.2% for Children Average Accuracy: 95.6%
1D CNN	TIMIT Speech Corpuses [02]	**Gender Identification for AE Dialect Sentences Classification** Average Recall(%): 98.2% for Male, 95.5% for Female Average Precision(%): 97.8% for Male, 96.4% for Female Average F1 Score: 98.0% for Male, 96.0% for Female Average Accuracy: 95.6%
1D CNN	TIMIT, CLDPS, and SLS Speech Corpuses [02, 31-32]	**Gender Identification for Dialect Sentences Classification** Average Recall(%): 94.0% for Male, 87.9% for Female Average Precision(%): 87.0% for Male, 94.4% for Female Average F1 Score: 90.4% for Male, 91.1% for Female Average Accuracy: 90.7%
2D CNN		**Gender Identification for Dialect Sentences Classification** Average Recall(%): 80.6% for Male, 95.4% for Female Average Precision(%): 96.3% for Male, 76.9% for Female Average F1 Score: 87.8% for Male, 85.1% for Female Average Accuracy: 86.6%
1D CNN	TIMIT, CLDPS, and SLS Speech Corpuses [02, 31-32]	**Region Identification for Dialect Sentences Classification** Average Recall(%): 96.9% for AE, 100.0% for IT, 100.0% for SP Average Precision(%): 96.9% for AE, 96.9% for IT, 100.0% for SP Average F1 Score: 98.5% for AE, 98.4% for IT, 100.0% for SP Average Accuracy: 99.1%
2D CNN		**Region Identification for Dialect Sentences Classification** Average Recall(%): 96.0% for AE, 100.0% for IT, 100.0% for SP Average Precision(%): 100.0% for AE, 95.8% for IT, 100.0% for SP Average F1 Score: 98.0% for AE, 97.9% for IT, 100.0% for SP Average Accuracy: 98.1%

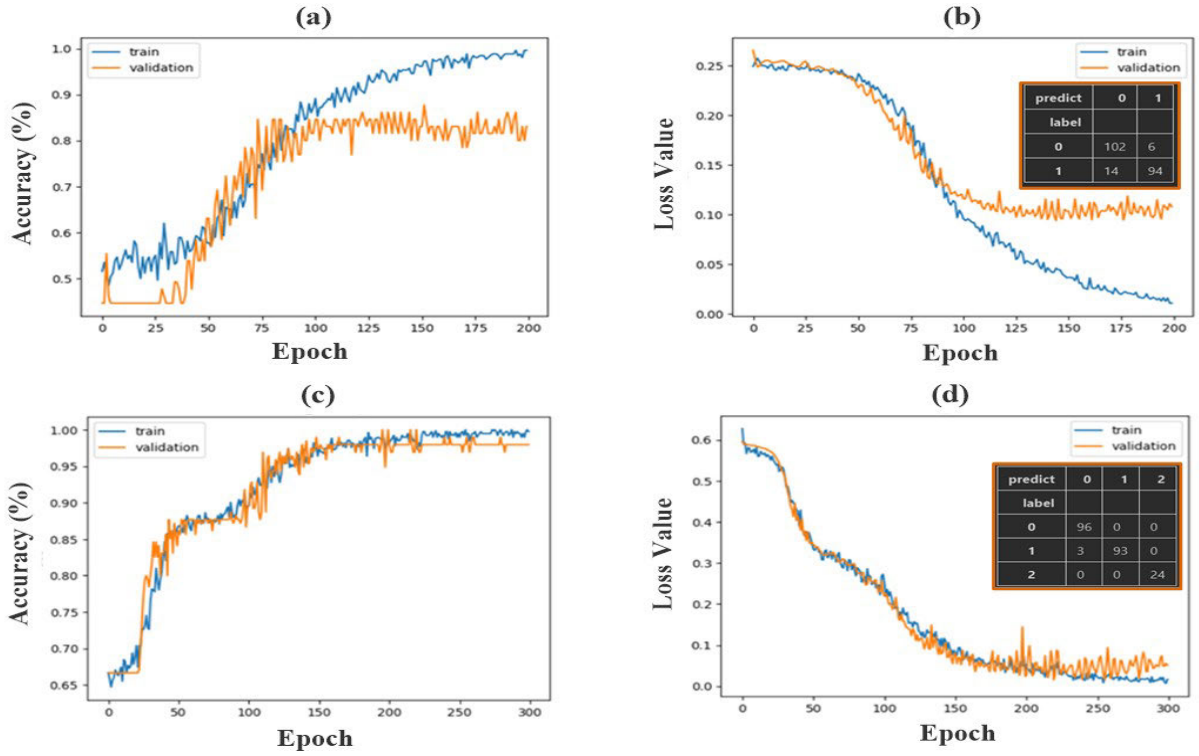

FIGURE 11. The training history curves and converging curves for training and validating 1D CNN classifier with TIMIT, CLDPS, and SLS speech corpus. (a) and (b) The training history curves and converging curves for gender identification; (c) and (d) The training history curves and converging curves for region identification.

TABLE 2. Comparisons of DL-based classifiers for ASR in gender and region recognition with different speech corpuses, methods, and purposes.

Literature	Speech Corpuses	Method	Purpose
[16]	Child Speech Dataset: MyST (My Science Tutor), PF-STAR, CMU_KIDS [16, 39-40, 47] Adult Speech Dataset: LibriSpeech, Librilight, LibriTTS [16, 38, 40]	Self-Supervised Learning (SSL): WAV2VEC2 Model [16-17]	**Child Speech Recognition** Word Error Rate (WER): 7.42% for MyST, WER: 2.91% for PF-STAR, WER: 12.77% for CMU_KIDS
[20]	TIMIT Speech Corpuses (280 Files) [02, 44]	MFCC+SVM (Support Vector Machine) based Classifier	**Gender Identification for Voice Signals Classification** Accuracy: 96.25%
[13]	Hillenbrand Speech Corpuses [12-13]	YAAPT [12-13], YIN, FCN [18], Novel PD-based HDM	**Fundamental Frequency Detection for Gender Identification** GPE (Gross Pitch Error) = 17.18 for YAPPT, GPE = 16.71 for YIN, GPE = 9.10 for FCN, GPE = 7.65 for HDM
	TIMIT Speech Corpuses [02, 44]		GPE = 16.00 for YAPPT, GPE = 43.88 for YIN, GPE = 5.82 for FCN, GPE = 6.39 for HDM
[19]	Uzbek Database (1,281 Speakers) [19]	DNN-HMM	**Automatic Speech Recognition** Model Training and Testing: Training Accuracy: 96% Testing Accuracy: 93% WER: 14.30% Character Error Rate (CER): 5.41%
[21]	TIMIT [02], RAVDESS (NA), and BGC (Bangladesh) [1, 21] Speech Corpuses	• Feature Extraction: Multiple Filtering Methods (High Pass Filter, Savltzky-Golay Filter) + MFCC + Linear Predictive Coding (LPC) • Training and Classification: Multilayer 1D CNN (ADAM Optimizer and Mean-Squared-Error (MSE) Loss Function)	**Gender and Region Detection:** Training Accuracy: 99.45% **for Gender Detection** and 99.65% **for Region Detection** Validation Accuracy: 93.01% **for Gender Detection** and 97.07% **for Region Detection** **Gender Detection:** M: Recall: 0.93; Precision: 0.94; F1: 0.93 W: Recall: 0.93; Precision: 0.92; F1: 0.93 **Three Regions Detection:** AE: Recall: 0.95; Precision: 0.93; F1: 0.95 NA: Recall: 0.99; Precision: 0.99; F1: 0.99 Banglades: Recall: 0.90; Precision: 0.94; F1: 0.92
Proposed Method	TIMIT, CLDPS, and SLS Speech Corpuses [02, 31-32]	(1) Band-Pass Butterworth Filter + MFCC + 1D CNN (2) Band-Pass Butterworth Filter + MFCC + 2D CNN	**Gender Recognition** (1) MFCC + 1D CNN: Average Recall(%): 94.0% for Male, 87.9% for Female Average Precision(%): 87.0% for Male, 94.4% for Female Average F1 Score: 90.4% for Male, 91.1% for Female Average Accuracy: 90.7% (2) MFCC + 2D CNN: Average Recall(%): 80.6% for Male, 95.4% for Female Average Precision(%): 96.3% for Male, 76.9% for Female Average F1 Score: 87.8% for Male, 85.1% for Female Average Accuracy: 86.6%
			Region Recognition (1) MFCC + 1D CNN: Average Recall(%): 96.9% for AE, 100.0% for IT, 100.0% for SP; Average Precision(%): 96.9% for AE, 96.9% for IT, 100.0% for SP; Average F1 Score: 98.5% for AE, 98.4% for IT, 100.0% for SP; Average Accuracy: 99.1% (2) MFCC + 2D CNN: Average Recall(%): 96.0% for AE, 100.0% for IT, 100.0% for SP; Average Precision(%): 100.0% for AE, 95.8% for IT, 100.0% for SP; Average F1 Score: 98.0% for AE, 97.9% for IT, 100.0% for SP; Average Accuracy: 98.1%

C. COMPARISON AND DISCUSSION

Comparisons presented in Table 2 indicate that different feature extraction methods and AI-based classifiers have been used in ASR applications [1], [12], [13], [18], [20], [21]. For gender and regional dialect identification, previous studies [1], [13], [18], [20] have proposed DL- and ML-based methods, including the "MFCC+SVM [20]", "YAAPT/YIN [12], [13]+FCN / Novel PD-based HDM [18]", and "MFCC+linear predictive coding (LPC)+1D-CNN [1]". These methods have been used to train various models of gender classifiers using the TIMIT [2], Hillenbrand [12], [13], RAVDESS, and BGC [1], [21] speech corpora. For example, multiple filtering methods and multilayer 1D CNN were combined [1] to perform gender

and region detection, and the experimental results were as follows: recall = 0.93, precision = 0.94, and F1 = 0.93 for male individuals; recall = 0.93, precision = 0.92, and F1 = 0.93 for female individuals; recall = 0.95, precision = 0.93, and F1 = 0.95 for AE; recall = 0.99, precision = 0.99, and F1 = 0.99 for North America (NA); and recall = 0.90, precision = 0.94, and F1 = 0.92 for BGC. Unwanted sound noise and distortion were removed using a high-pass filter with a cut-off frequency threshold near the audible range of speech signals [1]. The MFCC and LPC methods were used to extract feature datasets for training a 1D CNN classifier (1,433 audio files used for training), which was combined with the 3-layer feature extraction method for detecting gender-based and regional differences in dialect-speech signals (615 audio files) from the TIMIT (16 male and 16 female individuals) [2], RAVDESS (NA; 12 male and 12 female individuals), and BGC (3 male and 3 female individuals) speech corpora [21].

In the present study, data on fundamental frequency, spectral entropy, spectral flatness, and mode frequency were used as feature datasets. These frequency features were preprocessed before training of the multilayer 1D-CNN classifier. Because of the different characteristics of human speech from different geographical regions, a study [1] extracted a sufficient number of features by using three layers (frequency features+MFCC+LPC) during feature extraction, and a 1D-CNN with multiple convolution layers was used (filter sizes: 64, 128, 512, and 128; kernel sizes: 2 and 3). In addition, multiconvolution layers, batch normalization, Max-pooling, and dropout layers (10% – 30% rate) were used in the study to mitigate overfitting problems. This model employed the ADAM optimizer and the mean-squared-error (MSE) loss function to establish a classifier for both gender- and region-based classification of human speech. However, this multi- output based 1D CNN had a complex structure and an excessive number of feature datasets, which necessitated more CPU time for training its ASR classifier.

A study [16] employed the WAV2VEC2 model, comprising a feature encoder, context network, and quantization module, to extract representations from raw speech signals in an SSL scheme for child and adult ASR. These representations were used to train numerous unlabeled speech datasets in a pretraining step, and notable amounts of unlabeled child speech datasets were obtained by finetuning labeled datasets with connectionist temporal classification (CTC) during the second training step of the study. For child [16], [39], [40], [47] and adult [16], [38], [40] speech datasets (custom datasets), the experimental results for child speech recognition (CSR) were as follows: word error rate (WER) = 7.42% for MyST (My Science Tutor), 2.91% for PF-STAR, and 12.77% for CMU_KIDS. However, the WAV2VEC2 model can only separate child speech signals from adult speech signals. Therefore, in the present study, the proposed multilayer classifier was combined with the signal preprocessing (digital filtering process), an MFCC extractor, and 1D-CNN/2D-CNN models; the classifier could distinguish speech signals by gender, age group, and regional dialect for the AE, IT, and SP languages. Compared with conventional DL-based classifiers, the established small-scale training models could reduce the number of network layers (two cascaded convolutional–pooling layers) and the computational load in pattern recognition tasks. In addition, the dimensionality of the feature patterns could also be effectively reduced, thus lowering the large number of training datasets used to mitigate overfitting during training. For native AE (AE and BE) and non-native AE (IT and SP) dialect-speech signals, the EPD method could easily detect the obvious signal activities with filtering and segmentation; subsequently, the MFCC method extracted the distinguishable feature patterns to separate the native and non-native AE individuals by their gender and country. In feature extraction, the DCT operation enabled the extraction of the real part of speech signals while discarding the imaginary part, which reduced the computational load in feature extraction tasks. The proposed classifier performed well and promisingly for its intended purpose.

IV. CONCLUSION

This study developed a combined MFCC-based feature extractor and CNN-based classifier for classifying dialect-speech signals by gender and region. The MFCC extractor extracted key feature parameters from dialect-speech signals and transformed them into 1D or 2D visual patterns in the frequency domain; subsequently, the 1D- or 2D-CNN-based classifier could identify dialect-sentence features. Time-domain speech signals were preprocessed using the Butterworth filter and EPD algorithm to detect signal activities for obtaining interesting content from raw speech sentences. The Mel frequency reflected the ability of the human ear to perceive various frequency changes. Thus, MFCC features with 1D or 2D feature patterns in visualization representations were used to enhance the speech recognition accuracy of the proposed classifiers. Several speech corpora (TIMIT, Hillenbrand, CLDPS, and SLS databases) were used to train, test, and validate the proposed 1D-CNN and 2D-CNN classifiers for effectively distinguishing speech signals by gender, age group, and regional dialect. The developed classifiers achieved favorable precision (%), recall (%), and F1 score values of >90% for gender and regional dialect recognition. The ASR classifiers developed in the present study can be used in an HMIS to accurately identify speech-associated characteristics, such as the speaker's gender, age group, and regional dialect. Subsequently, the HMIS can output correct information to users to continue human–computer interactions and conversations through NLG and NLT functions. In the future, the developed classifiers can be used in daily applications, such as language learning, smart-home activities, healthcare services, hearing impairment-related assistance, smart healthcare services, smart assistants, and smart transportation. In addition, for applications involving high linguistic diversity, corpuses on other languages, such as Chinese, Japanese [48], Arabic [49],

BGC [1], [21], and Uzbek [19], can also be used to establish a multilingual speech classifier in gender and region (dialect and accents) recognition, which could mitigate identification errors and improve the linguistic diversity of HMIS in global applications.

DATA AVAILABILITY STATEMENT

Not applicable.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

ABBREVIATIONS

AE	American English.
IT	Italian.
SP	Spanish.
BE	British English.
AEV	American English Vowel.
HMIS	Human-Machine Interaction System.
ASR	Automatic Speech Recognition.
CNN	Convolutional Neural Network.
1D CNN	One-Dimensional CNN.
2D CNN	Two-Dimensional CNN.
MFCC	Mel-Frequency Cepstral Coefficients.
APCSD	Acoustic-Phonetic Continuous Speech Database.
NL	Natural Language.
TIMIT	Texas Instruments / Massachusetts Institute of Technology.
RCM	Reactive Communication Mode.
ICM	Interactive Communication Mode.
NLG	Natural Language Generation.
IoV	Internet of Vehicles.
NLU	NL Understanding.
NLT	NL Translation.
DL	Deep Learning (.
ML	Machine Learning.
MT	Machine Translation.
AI	Artificial Intelligence.
RNN	Recurrent Neural Network.
EPD	Endpoint Detection.
SAD	Signal Activity Detection.
Max-Pooling	Maximum-Pooling.
GeLU	Gaussian Error Linear Unit.
ADAM	Adaptive Moment Estimation.
CLDPS	Corpora e Lessici dell' Italiano Parlato e Scritto.
SLS	Sound-board-Learn Spanish.
DCT	Discrete Cosine Transform.
HAS	Human Auditory System.
FFT	Fast Fourier Transform.
CCE	Categorical Cross-Entropy.
CMU-KIDS	Carnegie Mellon University-KIDS.
TTS	Text-to-Speech.
GPU	Graphics Processing Unit.
YAAPT	Yet Another Algorithm for Pitch Tracking.
SSL	Self-Supervised Learning.
SVM	Support Vector Machine.
FCN	Fully-Convolutional Network.
DNN-HMM	Deep Neural Network with the Hidden Markov Model.
LPC	Linear Predictive Coding.
RAVDESS	Ryerson Audio- Visual Database of Emotional Speech and Song.
BGC	Bangladesh.
NA	North America.
CTC	Connectionist Temporal Classification.
CSR	Child Speech Recognition.

REFERENCES

[1] M. A. Uddin, R. K. Pathan, M. S. Hossain, and M. Biswas, "Gender and region detection from human voice using the three-layer feature extraction method with 1D CNN," *J. Inf. Telecommun.*, vol. 6, no. 1, pp. 27–42, Jan. 2022.

[2] (2023). *The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus (TIMIT) Training and Test Data*. [Online]. Available: https://goo.gl/l0sPwz

[3] J. Vavrek, P. Viszlay, M. Lojka, J. Juhár, and M. Pleva, "Weighted fast sequential DTW for multilingual audio query-by-Example retrieval," *J. Intell. Inf. Syst.*, vol. 51, no. 2, pp. 439–455, Oct. 2018.

[4] J. Novoa, J. Wuth, J. P. Escudero, J. Fredes, R. Mahu, and N. B. Yoma, "DNN-HMM based automatic speech recognition for HRI scenarios," in *Proc. 13th ACM/IEEE Int. Conf. Human-Robot Interact. (HRI)*, Chicago, IL, USA, Mar. 2018, pp. 150–159.

[5] Y. Zhang, D. S. Park, W. Han, J. Qin, A. Gulati, J. Shor, A. Jansen, Y. Xu, Y. Huang, S. Wang, and Z. Zhou, "BigSSL: Exploring the frontier of large-scale semi-supervised learning for automatic speech recognition," *IEEE J. Sel. Topics Signal Process.*, vol. 16, no. 6, pp. 1519–1532, Oct. 2022.

[6] S.-A. Li, Y.-Y. Liu, Y.-C. Chen, H.-M. Feng, P.-K. Shen, and Y.-C. Wu, "Voice interaction recognition design in real-life scenario mobile robot applications," *Appl. Sci.*, vol. 13, no. 5, p. 3359, Mar. 2023.

[7] (2023). *Natural Language Processing*. [Online]. Available: https://www.hyperscience.com/knowledge-base/natural-language-processing/

[8] J. Guo, M. Liu, Y. Guo, and T. Zhou, "An AR/VR-hybrid interaction system for historical town tour scenes incorporating mobile internet," in *Proc. Int. Conf. Digit. Soc. Intell. Syst. (DSInS)*, Chengdu, China, Dec. 2021, pp. 21–24.

[9] J. Gao, C. Peng, T. Yoshinaga, G. Han, S. Guleng, and C. Wu, "Digital twin-enabled Internet of Vehicles applications," *Electronics*, vol. 13, no. 7, p. 1263, Mar. 2024.

[10] Z. Ke, J. Sheng, Z. Li, W. Silamu, and Q. Guo, "Knowledge-guided sentiment analysis via learning from natural language explanations," *IEEE Access*, vol. 9, pp. 3570–3578, 2021.

[11] H. A. Sánchez-Hevia, R. Gil-Pita, M. Utrilla-Manso, and M. Rosa-Zurera, "Age group classification and gender recognition from speech with temporal convolutional neural networks," *Multimedia Tools Appl.*, vol. 81, no. 3, pp. 3535–3552, Jan. 2022.

[12] C. Lin, H. Lai, P. Huang, P. Chen, and C. Li, "Vowel classification with combining pitch detection and one-dimensional convolutional neural network based classifier for gender identification," *IET Signal Process.*, vol. 17, no. 5, pp. 1–14, May 2023.

[13] C. Parlak and Y. Altun, "Harmonic differences method for robust fundamental frequency detection in wideband and narrowband speech signals," *Math. Problems Eng.*, vol. 2021, pp. 1–17, Oct. 2021.

[14] F. Li, M. Liu, Y. Zhao, L. Kong, L. Dong, X. Liu, and M. Hui, "Feature extraction and classification of heart sound using 1D convolutional neural networks," *EURASIP J. Adv. Signal Process.*, vol. 2019, no. 1, pp. 1–11, Dec. 2019.

[15] S. Dua, S. S. Kumar, Y. Albagory, R. Ramalingam, A. Dumka, R. Singh, M. Rashid, A. Gehlot, S. S. Alshamrani, and A. S. AlGhamdi, "Developing a speech recognition system for recognizing tonal speech signals using a convolutional neural network," *Appl. Sci.*, vol. 12, no. 12, p. 6223, Jun. 2022.

[16] R. Jain, A. Barcovschi, M. Y. Yiwere, D. Bigioi, P. Corcoran, and H. Cucu, "A WAV2VEC2-based experimental study on self-supervised learning methods to improve child speech recognition," *IEEE Access*, vol. 11, pp. 46938–46948, 2023.

[17] A. Baevski, S. Schneider, and M. Auli, "Vq-wav2vec: Self-supervised learning of discrete speech representations," 2019, *arXiv:1910.05453*.

[18] L. Ardaillon and A. Roebel, "Fully-convolutional network for pitch estimation of speech signals," in *Proc. Interspeech*, Graz, Austria, Sep. 2019, pp. 2005–2009.

[19] A. Mukhamadiyev, I. Khujayarov, O. Djuraev, and J. Cho, "Automatic speech recognition method based on deep learning approaches for uzbek language," *Sensors*, vol. 22, no. 10, p. 3683, May 2022.

[20] S. Chaudhary and D. K. Sharma, "Gender identification based on voice signal characteristics," in *Proc. Int. Conf. Adv. Comput., Commun. Control Netw. (ICACCCN)*, Greater Noida, India, Oct. 2018, pp. 869–874.

[21] M. A. Uddin, M. S. Hossain, R. K. Pathan, and M. Biswas, "Gender recognition from human voice using multi-layer architecture," in *Proc. Int. Conf. Innov. Intell. Syst. Appl. (INISTA)*, Novi Sad, Serbia, Aug. 2020, pp. 1–7.

[22] K. H. Abdullah and M. Bilal Er, "Lung sound signal classification by using cosine similarity-based multilevel discrete wavelet transform decomposition with CNN-LSTM hybrid model," in *Proc. 4th Int. Conf. Artif. Intell. Speech Technol. (AIST)*, Delhi, India, Dec. 2022, pp. 1–4.

[23] D. H. Mundri, S. S. Solanki, and K. Mahto, "An approach to finding the most suitable algorithm for noise reduction in a neonatal cry signal," in *Proc. 14th Int. Conf. Comput. Commun. Netw. Technol. (ICCCNT)*, Delhi, India, Jul. 2023, pp. 1–4.

[24] J. Wu, G. Chong, W. Pang, and L. Wang, "Speech endpoint detection based on EMD and improved spectral subtraction," in *Proc. 5th Int. Conf. Natural Lang. Process. (ICNLP)*, Guangzhou, China, Mar. 2023, pp. 126–130.

[25] Q. Wu and Y. Liu, "A speech endpoint detection method based on cascaded speech enhancement," in *Proc. Int. Conf. Electron. Inf. Technol. Smart Agricult. (ICEITSA)*, Huaihua, China, Dec. 2021, pp. 1–6.

[26] G. Li, Y. Liu, and X. Wang, "Speech emotion recognition based on 1D CNN and MFCC," in *Proc. IEEE 5th Int. Conf. Civil Aviation Saf. Inf. Technol. (ICCASIT)*, Dali, China, Oct. 2023, pp. 956–960.

[27] A. Nfissi, W. Bouachir, N. Bouguila, and B. L. Mishara, "CNN-n-GRU: End-to-end speech emotion recognition from raw waveform signal using CNNs and gated recurrent unit networks," in *Proc. 21st IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Nassau, Bahamas, Dec. 2022, pp. 699–702.

[28] J. Ma and D. Yarats, "Quasi-hyperbolic momentum and Adam for deep learning," in *Proc. ICLR*, New Orleans, LA, USA, 2019, pp. 1–38.

[29] J. Bergstra, O. Breuleux, F. Bastien, P. Lamblin, R. Pascanu, G. Desjardins, J. Turian, D. Warde-Farley, and Y. Bengio, "Theano: A CPU and GPU math compiler in Python," in *Proc. Python Sci. Conf.*, vol. 4, 2010, pp. 1–7.

[30] F. Z. Zhang, C. H. Lin, P. Y. Chen, N. S. Pai, C. M. Su, C. C. Pai, and H. W. Ho, "Number of convolution layers and convolution kernel determination and validation for multilayer convolutional neural network: Case study in breast lesion screening of mammographi images," *Processes*, vol. 10, p. 1867, Sep. 2022.

[31] *CLDPS (Corpora E Lessici Dell' Italiano Parlato E Scritto)*. Accessed: 2024. [Online]. Available: http://www.clips.unina.it/en/index.js

[32] *Soundboard.com—Create & Download, Search Results for 'Learn Spanish'*. Accessed: 2024. [Online]. Available: https://www.soundboard.com/search/Learn%20Spanish%20-

[33] *Syntax: Butte' Cbutterworth Filter Design—MATLAB Butter*. Accessed: 2024. [Online]. Available: https://www.mathworks.com/help/signal/ref/butter.html

[34] N. Ahmed, T. Natarajan, and K. R. Rao, "Discrete cosine transform," *IEEE Trans. Comput.*, vol. COM-100, no. 1, pp. 90–93, Jan. 1974.

[35] D. Eringis and G. Tamulevicius, "Improving speech recognition rate through analysis parameters," *Electr., Control Commun. Eng.*, vol. 5, no. 1, pp. 61–66, May 2014.

[36] M. R. Firmansyah, R. Hidayat, and A. Bejo, "Comparison of windowing function on feature extraction using MFCC for speaker identification," in *Proc. Int. Conf. Intell. Cybern. Technol. Appl. (ICICyTA)*, Bandung, Indonesia, Dec. 2021, pp. 1–5.

[37] R. Hidayat, A. Bejo, S. Sumaryono, and A. Winursito, "Denoising speech for MFCC feature extraction using wavelet transformation in speech recognition system," in *Proc. 10th Int. Conf. Inf. Technol. Electr. Eng. (ICITEE)*, Jul. 2018, pp. 280–284.

[38] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 5206–5210.

[39] A. Batliner, M. Blomberg, S. D'Arcy, D. Elenius, D. Giuliani, M. Gerosa, C. Hacker, M. Russell, S. Steidl, and M. Wong, "The PF STAR children's speech corpus," in *Proc. Interspeech*, 2005, pp. 2761–2764.

[40] J. Kahn, M. Riviere, W. Zheng, E. Kharitonov, Q. Xu, P. E. Mazaré, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen, T. Likhomanenko, G. Synnaeve, A. Joulin, A. Mohamed, and E. Dupoux, "Libri-light: A benchmark for ASR with limited or no supervision," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Switzerland, May 2020, pp. 7669–7673.

[41] R. Peinl and J. Wirth, "Quality assurance for speech synthesis with ASR," in *Proc. SAI Intell. Syst. Conf.* Cham, Switzerland: Springer, 2022, pp. 739–751.

[42] V. Kadyan, H. Kathania, P. Govil, and M. Kurimo, "Synthesis speech based data augmentation for low resource children ASR," in *Speech and Computer* (Lecture Notes in Computer Science), vol. 12997. Cham, Switzerland: Springer, 2021, pp. 317–326.

[43] J. Hillenbrand, L. A. Getty, K. Wheeler, and M. J. Clark, "Acoustic characteristics of American English vowels," *J. Acoust. Soc. Amer.*, vol. 95, p. 2875, May 1994.

[44] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1," U.S. Dept. Commerce Technol. Admin. Nat. Inst. Standards Technol. Comput. Syst. Lab. Adv. Syst. Division, Gaithersburg, MD, USA, Tech. Rep. NISTIR 4930, Feb. 1993.

[45] P. Podder, M. M. Hasan, M. R. Islam, and M. Sayeed, "Design and implementation of Butterworth, Chebyshev-I and elliptic filter for speech signal analysis," *Int. J. Comput. Appl.*, vol. 98, no. 7, pp. 12–18, Jul. 2014.

[46] F. Schiel, M. Stevens, U. Reichel, and F. Cutugno, "Machine learning of probabilistic phonological pronunciation rules from the Italian CLIPS corpus," in *Proc. Interspeech*, Lyon, France, Aug. 2013, pp. 1414–1418.

[47] W. Ward, R. Cole, and S. Pradhan, "My science tutor and the MyST corpus," Boulder Learn. Inc., Carbondale, CO, USA, 2019.

[48] (2024). *Linguistic Data Consortium, Asian Spoken Language Sampler*. [Online]. Available: https://catalog.ldc.upenn.edu/LDC2010S07

[49] (2024). *Arabic Speech Corpus*. [Online]. Available: https://en.arabicspeechcorpus.com/

HSIANG-YUEH LAI received the Ph.D. degree from the Department of Engineering and System Science, National Tsing Hua University, Hsinchu, Taiwan, in 2010. She is currently an Associate Professor with the Department of Electrical Engineering, National Chin-Yi University of Technology, Taichung City, Taiwan.

Her current research interests include intelligent systems, embedded systems, and solid-state electronic devices.

CHIA-CHIEH HU received the B.S. degree from the Department of Electrical Engineering, Lunghwa University of Science and Technology, Taoyuan, Taiwan, in 2023.

He is currently enrolled with the Department of Electrical Engineering, National Chin-Yi University of Technology, Taichung City, Taiwan, where has been, since 2023. His research interests include signal processing, speech recognition, and artificial intelligence applications.

CHIA-HUNG WEN was born in Taichung City, Taiwan, in 1993. He received the B.S. degree from the Department of Information Management, National Taichung University of Science and Technology, Taichung City, in 2015. He is currently enrolled with the Department of Electrical Engineering, National Chin-Yi University of Technology, Taichung City, where has been with Tzu Chi Foundation, since 2022, with research interests include neural network computations and the application of information technology in healthcare.

JIAN-XING WU (Member, IEEE) was born in 1985. He received the B.S. and M.S. degrees in electrical engineering from the Southern Taiwan University of Science and Technology, Tainan City, Taiwan, in 2007 and 2009, respectively, and the Ph.D. degree in biomedical engineering from National Cheng Kung University, Tainan City, in 2014.

He was a Postdoctoral Research Fellow with the X-Ray and IR Imaging Group, National Synchrotron Radiation Research Center, Hsinchu, Taiwan, from 2014 to 2017. He was a Postdoctoral Research Fellow with the Department of Niche Biomedical LLC, California NanoSystems Institute, UCLA, Los Angeles, CA, USA, from 2017 to 2018. Currently, he is an Assistant Professor with the Department of Electrical Engineering, National Chin-Yi University of Technology, Taichung City, Taiwan, where has been, since 2019. His research interests include artificial intelligence applications in electrical engineering and biomedical engineering, biomedical signal processing, medical ultrasound, medical device design, and X-ray microscopy.

NENG-SHENG PAI received the B.S. and M.S. degrees from the Department of Automatic Control Engineering, Feng Chia University, Taichung City, Taiwan, in 1983 and 1986, respectively, and the Ph.D. degree from the Department of Electrical Engineering, National Cheng Kung University, Tainan City, Taiwan, in December 2002.

He was the Chairperson of the Department, from 2004 to 2007, and was also the Chairperson of the Computer Center, National Chin-Yi University of Technology, Taichung City, from 2013 to 2017. He is currently a Professor with the Department of Electrical Engineering, National Chin-Yi University of Technology. His current research interests include fuzzy systems, artificial intelligence, image processing, advanced control systems, and microprocessor systems.

CHENG-YU YEH received the B.S. and M.S. degrees in electrical engineering and the Ph.D. degree from the Graduate Institute of Mechanical and Electrical Engineering, National Taipei University of Technology, Taipei City, Taiwan, in 2000, 2002, and 2006, respectively. Currently, he is a Professor with the Department of Electrical Engineering, National Chin-Yi University of Technology, Taichung City, Taiwan. His research interests include speech and image signal processing, machine learning, deep learning, and its applications.

CHIA-HUNG LIN was born in Kaohsiung City, Taiwan, in 1974. He received the B.S. degree in electrical engineering from the Tatung Institute of Technology, Taipei City, Taiwan, in 1998, and the M.S. and Ph.D. degrees in electrical engineering from the National Sun Yat-sen University, Kaohsiung City, in 2000 and 2004, respectively.

He was a Professor with the Department of Electrical Engineering, Kao Yuan University, Kaohsiung City, from 2004 to 2017. He has been a Professor with the Department of Electrical Engineering, National Chin-Yi University of Technology, Taichung City, Taiwan, where has been, since 2018. His research interests include neural network computing and its applications in biomedical engineering, smart grid, and infosecurity applications, biomedical signal and image processing, healthcare, hemodynamic analysis, and pattern recognition.