Hindawi
Mobile Information Systems
Volume 2022, Article ID 7427899, 11 pages
https://doi.org/10.1155/2022/7427899

Research Article

Control System and Speech Recognition of Exhibition Hall Digital Media Based on Computer Technology

Yu Zhao

School of Visual Communication Design, LuXun Academy of Fine Arts, Shenyang 110000, Liaoning, China

Correspondence should be addressed to Yu Zhao; 1631180879@xzyz.edu.cn

Received 15 June 2022; Revised 26 July 2022; Accepted 30 July 2022; Published 30 August 2022

Academic Editor: Yajuan Tang

With environmental noise in the exhibition hall, speakers tend to change their speech production to preserve intelligible communication. While great evolution has been prepared in Automatic Speech Recognition (ASR), important performance deprivation occurs in a noisy environment. The assessment of the degree of speech impairment and the efficacy of computer recognition of impaired speech are distinctly and independently executed. Convolutional Neural Networks (CNN) have been effectively employed in speech recognition and computer vision tasks. Hence, this study uses the Deep Convolutional Neural Network-based Automatic Speech Recognition Model (DCNN-ASRM) for effective speech recognition in the noisy exhibition hall. This study configures the filter sizes, poolings, and input feature map. The filter size and pooling are decreased, and the dimension of the input feature is comprehensive to permit increasing convolution layers. Furthermore, an in-depth analysis of the proposed DCNN-ASRM model discloses critical features, like fast convergence speed, compact model scales, and noise robustness in speech recognition. The simulation analysis shows that the suggested DCNN-ASRM model enhances the recognition accuracy ratio of 98.1%, performance ratio of 97.2%, and noise reduction ratio of 96.5% and reduces the word error rate by 9.2% and signal-to-noise ratio by 10.3% compared to other existing models.

1. Introduction of Speech Recognition in Exhibition Hall

Describing the acoustic environment of performance spaces like concert halls, theatres, exhibition halls, and auditoria has been one of the key topics in room acoustics and speech recognition studies during the past decades [1]. For spaces where the major function is sound-associated (e.g., spaces for performing or listening), clear criteria connected to measurable variables must be in place to evaluate acoustic performance and quality [2]. The general deliberation on inclusion ensures that the case fits the suitable building type for crowd transit (e.g., exhibition/museum spaces, shopping malls, and transportation stations/hubs) [3]. Numerous visitors, conversations, and product presentations cause noise levels in trade fairs and exhibitions to rise quickly [4, 5]. Many of the most advanced speech recognition systems use the Mel-Frequency Cepstral Coefficient (MFCC) and the Perceptual Linear Prediction (PLP) cepstral coefficient (like MFCCs) based on human auditory models, two types of cepstral coefficients that are computed utilizing discrete cosine transform on smoothed power spectra [6]. However, these systems still execute poorly in a noisy environment (e.g., in situations with additive noise), and magnified performance deprivation has been perceived under the detached (far-fields) talking state [7]. To attain robust automatic speech recognition, it is more significant to cooperatively optimize the acoustic and beam-forming model to maximize the ASR performance [8]. In noise, speakers modify their vocal efforts. For an extensive noise level range, the dependency among voice Sound Pressure Level (SPL) and noise sound pressure level is almost linear, with diverse slopes when reading text or communicating with others [9]. Vowels tend to get more attention than consonants in increasing vocal effort, which is not universal among phones [10]. The pitch rises with an increase in subglottal pressure and tension in the laryngeal musculature due to the adjustment in a vocal effort [11]. When represented in semitones and SPL, the pitch varies roughly linearly with speech intensity [12].

Speech recognition is an artificial intelligence-enhanced technology that converts a person's speech from an analog form to a digital one. An improved computer program then uses digital speech for additional processing [13]. Speech recognition is the ability to speak to a computer and have it comprehend and interpret what people speak. Spoken language is decoded and translated into machine form. A natural language such as English is used to communicate between people and computers in Natural Language Processing (NLP) [14]. Although extraordinary performance has been realized in ASR with the introduction of Convolutional Neural Networks (CNNs), the performance still reduces dramatically in far-field and noisy situations [15]. To attain robust speech recognition, multiple microphones can improve speech signals, reduce noises and reverberation, and enhance Automatic Speech Recognition performance [16]. The low signal-to-noise ratio (SNR) in these noisy situations creates CNN more susceptible to the mismatch issue [17]. CNNs have numerous benefits: firstly, speech spectrograms have local associations in both frequency and time, and CNNs are well-matched to model those connections overtly via local connectivity, whereas Deep Neural Networks (DNN) have a comparatively more complex encoding of this data [18]. Secondly, translational invariance, like frequency shift because of speaking style or speaker variation, can be more effortlessly seized by CNNs than DNNs [19, 20]. Thus, in CNNs, a more affluent set of features can be demoralized than traditional low-dimensional feature vectors, like PLP coefficients and MFCCs.

The major contribution of the research is

(i) Designing the Deep Convolutional Neural Network-based Automatic Speech Recognition Model (DCNN-ASRM) for speech recognition in the noisy exhibition hall.

(ii) Evaluating the DCNN shows noise robustness greater than other models in noisy situations.

(iii) The simulation outcomes have been executed, and the recommended DCNN-ASRM model enhances the accuracy and performance compared to other existing models.

The rest of the study is systematized: Sections 1 and 2 discuss the introduction and existing speech recognition methods. In Section 3, DCNN-ASRM has been proposed. In Section 4, experimental results have been executed. Finally, Section 5 concludes the research paper.

2. Literature Survey

Song et al. [21] proposed the Learning-to-Rescore (L2RS) mechanism for ASR. L2RS uses a wide variety of textual data from state-of-the-art NLP models and spontaneously decides their weights to restore the N-best lists for ASR systems. These characteristics included Bidirectional Encoder Representation for Transformer (BERT) sentence topic vector, embedding, perplexity score generated by n-gram Language Model (LM), topic modeling BERT LM, LM, and RNNLM to train an algorithm for scoring. According to their research, L2RS surpasses standard rescoring techniques and its Deep Neural Network equivalents by a significant enhancement of 20.67% in terms of the NDCG@10 in terms of L2RS. L2RS enables the development of a more accurate rescoring model for ASR.

Han et al. [22] suggested the ContextNet for ASR with Global Context. New CNN-RNN-transducer architecture, termed ContextNet, focuses on a new study in this paper. ContextNet has a fully convolutional encoder that adds global context data into the convolution layer by including squeeze-and-excitation modules. In addition, they presented a simple scaling mechanism for ContextNet's widths which offers a fair trade-off between computation and accuracy. CNN models previously released have had a hard time competing with this model on the LibriSpeech test. Small Automatic Speech Recognition (ASR) models may be found using the suggested design, limiting the network's breadth to a minimum. A preliminary analysis of a bigger and more difficult data set supports their findings.

Isobe et al. [23] recommended the Deep Canonical Correlation Analysis (DCCA) for speech recognition in noisy environments. DCCA provides projections from two modalities into a single space to increase the correlation of projected vectors. Automatic Speech Recognition (ASR) may be more robust by employing DCCA approaches with audio and visual modalities, recovering noisy audio characteristics, and training an ASR model using the additional data. Our technique was tested using an audio-visual corpus CENSREC-1-AV and a noise database DEMAND. Our DCCA-based speech recognizers outperformed traditional ASR and featured fusion-based audio-visual speech recognition systems in terms of accuracy.

Hidayat and Winursito [24] discussed the Mel-Frequency Cepstral Coefficients for Improved Wavelet-Based Denoising (MFCC-IWBD) on Robust Speech Recognition. The Fast Fourier Transform (FFT) step of the MFCC was used. The denoising procedure utilizing Wavelet was only applied to data with noise, as determined by the FFT analysis findings. A total of eleven isolated English words were used in the research, along with various kinds of background noise. According to the research, this approach was more accurate than standard wavelet denoising methods for SNRs of 10 dB, 15 dB, and 20 dB when utilizing a Fejer Korovkin 6 wavelet type.

Based on the survey, there are several challenges to existing approaches such as Learning-to-Rescore (L2RS), ContextNet, Deep Canonical Correlation Analysis (DCCA), and Mel-Frequency Cepstral Coefficients for Improved Wavelet-Based Denoising (MFCC-IWBD) in achieving high accuracy, performance, signal-to-noise ratio, noise reduction, and word error rate (WER). The following section discusses the proposed DCNN-ASRM briefly.

3. Deep Convolutional Neural Network-Based Automatic Speech Recognition Model (DCNN-ASRM)

The most important component of the Human-Computer Interface (HCI) system is speech recognition, which translates the human auditory function. Automatic Speech Recognition (ASR) is making its way into our everyday lives at the same time as advances in computer technology are being made, and its applications are becoming more prevalent. There are two main causes for the present difficulties with speech recognition: extended range and noisy environments in the exhibition hall. This emphasizes the need for even higher accuracy systems capable of handling complex ASR tasks. The interruption of the channel and the unwanted background noise will always have a negative impact on the performance of the ASR system's recognition capabilities. Strategies for reducing noise may be applied differently to the ASR system. For instance, these techniques are speech enhancement upon the signal levels, extraction of the robust feature vector, and adjustment of the back-end acoustic model. In the real world, the problem of ambient noise cannot be considered in the preliminary phase, and it is not easy to anticipate how it will play out. The strategies for reducing noise will not depend on any suppositions about noisy settings or training variables, and they should function well across various noise circumstances. The basic objective of a feature extractor designed to be sound is to make as few assumptions as possible or none at all regarding the noise information. This is one of the difficult challenges that has recently been favoring research fields that are still being conducted. The capacity of different speech classes is appropriate and relevant with the interference of background noise and the variable nature of speaker characteristics. Hence, this paper proposes the DCNN-ASRM model for speech recognition in exhibition halls.

Figure 1 shows the proposed DCNN-ASRM model. The functionality of an ASR system can be described as the extraction of several speech variables from acoustic speech signals for every word or subword unit. The cascaded model contains two steps. The initial step is the denoising step, in which the denoised spectrogram is produced by eliminating the noise from the noisy speech signal. An Audio-Visual Speech Recognition (AVSR) model utilizes visual speech data, and the acoustic data are utilized by a normal ASR model. Audio and video data can be incorporated by feature or decision fusion. Mel-Frequency Cepstral Coefficient (MFCC) is one of the most accurate feature extraction methods used in Automatic Speech Recognition. The feature vector is extracted from the frequency spectrum of the windowed speech frame. The output is a speech signal or vocal sound accepted in all languages on a computer.

3.1. Problem Formulation. Let us preassume that a microphone input signal $y(t)$ is modeled as

$$y(t) = w(t) * g(t) + m(t). \tag{1}$$

As shown in equation (1), $w(t)$ denotes the non-reverberant speech signal, $g(t)$ indicates exhibition Room Impulse Response (RIR) among the speaker and the microphone, $*$ represents the convolution operator, and $m(t)$ signifies additive noise signal.

In the short-time Fourier transform (STFT) domain, the convolution is estimated as a multiplication operation, and equation (1) can be modified as

$$y(k, f) = J(k) \cdot w(k, f) \cdot g(k, f) + m(k, f). \tag{2}$$

As inferred from equation (2), $k \in \{0, \ldots, K-1\}$ and $f \in \{0, \ldots, F-1\}$ are frame and frequency index. The term $J(k)$ denotes the activity of $w(k, f)$ with,

$$J(k) = \begin{cases} 0, & \sum_{f=0}^{F-1} w(k, f) < Th, \\ 1, & \sum_{f=0}^{F-1} w(k, f) \geq Th. \end{cases} \tag{3}$$

As shown in equation (4), Th denotes a predefined threshold applied on the clean signals.

The automatic speech detection goal is to accurately detect frames in which the speaker is active given a noisy reverberant signal,

$$U(k) = q(J(k) = 1|y). \tag{4}$$

3.2. Convolutional Neural Network in Speech Recognition. A convolution layer does the convolution on the prior layer's feature maps utilizing filters and then augments bias scalars to the respective feature maps, trailed by nonlinear operations. Convolutions can be observed as operations employed to feature maps utilizing filters, where both feature maps and filters can be signified as a matrix. During this process, the enclosed region of feature maps is termed receptive fields, and the size is equal to the size of filters. At every stage, receptive fields execute dot products with filters and give one output, and every output collected during the progression will form fresh feature maps. The whole progression of a convolution layer is articulated as an expression given as follows:

$$g^{(k)} = \rho\left(S^{(k)} * g^{(k-1)} + a^{(k)}\right). \tag{5}$$

As discussed in equation (5), $g^{(k-1)}$ and $g^{(k)}$ denote feature map in two successive layers. $*$ represents the convolutional operator is achieved within filters $S^{(k)}$ and the feature maps $g^{(k-1)}$. The bias $a^{(k)}$ is auxiliary, and lastly, activation functions $\rho(\cdot)$, usually ReLU or sigmoid, can be employed to produce the output of the convolution layer. Expression equation (5) shows the modest setting where only one feature map occurs in prior layers. When numerous feature maps are contemporary in prior layers, the outcomes of convolution operation are initially added before accumulating biases.

Figure 2 shows the convolution and pooling operation. Pooling layers execute down-sampling on the feature map of prior layers and produce novel ones by decreased resolutions. Presently, the most famous configuration for convolutional neural networks is utilized in speech recognition, which has 2 convolution layers with 256 feature maps, 9 × 9 filters with 1 × 3 poolings in the initial convolution layer and

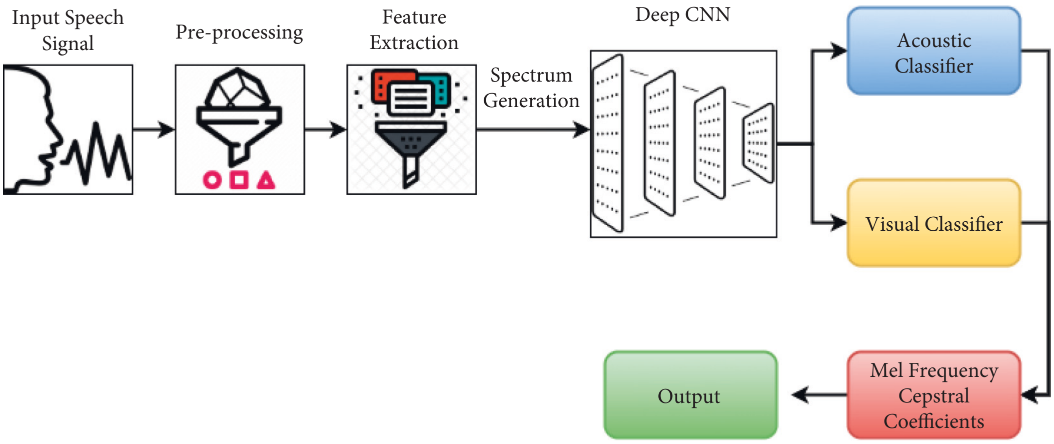
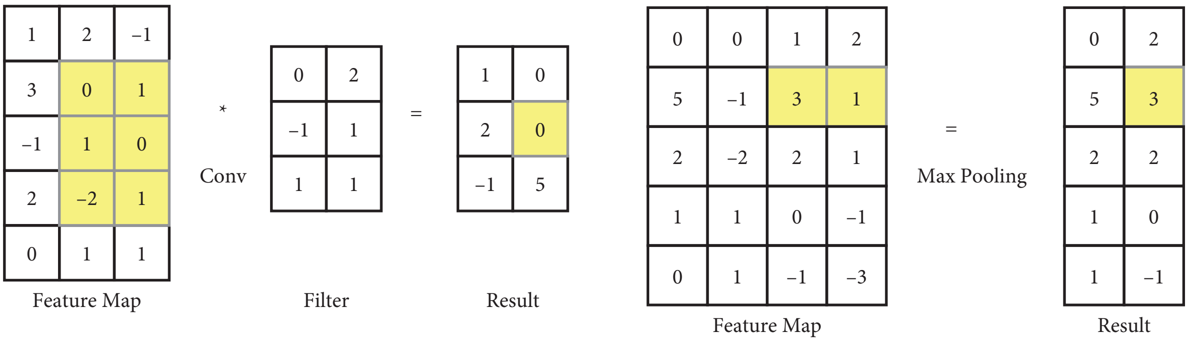


Figure 1: Proposed DCNN-ASRM model.



$$\begin{bmatrix} 1 & 2 & -1 \\ 3 & 0 & 1 \\ -1 & 1 & 0 \\ 2 & -2 & 1 \\ 0 & 1 & 1 \end{bmatrix} * \begin{bmatrix} 0 & 2 \\ -1 & 1 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 2 & 0 \\ -1 & 5 \end{bmatrix}$$

$$\begin{bmatrix} 0 & 0 & 1 & 2 \\ 5 & -1 & 3 & 1 \\ 2 & -2 & 2 & 1 \\ 1 & 1 & 0 & -1 \\ 0 & 1 & -1 & -3 \end{bmatrix} = \begin{bmatrix} 0 & 2 \\ 5 & 3 \\ 2 & 2 \\ 1 & 0 \\ 1 & -1 \end{bmatrix}$$

Figure 2: Convolution and pooling operation.

3×4 filters in the additional convolution layer and without poolings. This structure is utilized as the reference Convolutional Neural Network in this study. The number of convolution layer variables is comparatively minor since the filters are collective between every receptive field in one feature map. The variable number of one convolution layer can be computed as

$$\text{parameters}^{(k)} = \text{filtersize}^{(k)} \times n^{(k)} \times n^{(k-1)}. \tag{6}$$

As shown in equation (6), $n^{(k)}$ denotes the feature map in kth layers. Pooling layers achieve down-sampling on prior layers' feature maps and produce novel ones with a condensed resolution. In this research, max-pooling is utilized in the Convolutional Neural Network model.

Figure 3 shows the mismatches between testing and training conditions. The mismatch between testing and training conditions because of reverberation can be distinctly viewed in signals, features, and model spaces. Typical features are the Mel-Frequency Cepstral Coefficients, the formants, the pitch, the intensity of the speech signal, the vocal tract cross-sectional areas, and the speech ratio. Feature compensation and model adaptation are approaches that correspondingly work in the feature and model spaces. The notion is to decrease the mismatch between the perceived utterance and the actual speech models during utterance recognition. In feature compensation, this study maps the distorted features to evaluate the actual features so that the actual acoustic model can be utilized. This study maps the actual acoustic models to a converted model that better matches the detected utterance in model adaptation.

3.3. *Aware Factor Training (AFT).*

Aware factor training normally integrates a vector representing acoustic state data into the network training progression to standardize the nonspeech unpredictability. An LSTM mechanism has concatenated the assisting factor depictions with the acoustic feature, the energy for noise, j-vector for the speaker, and room reverberations. This study uses an auxiliary feature in a combined framework to evaluate a speaker-dependent bias rather than concatenate them with the acoustic feature. The speaker-dependent bias can be integrated into any position in a NN, e.g., convolution layers or fully connected layers. The construction of the suggested united aware factor training is

$$a^{wk} = U_2^k \rho \left(U_1^k x^w + q_1^k \right), \tag{7}$$

$$p^{wk} = \rho \left(S^k p^{k-1} + a^{wk} + a^k \right). \tag{8}$$

As inferred from equations (7) and (8), p^{wk} denotes the speaker that altered the hidden output of layers k. a^{wk} indicates speaker-dependent biases. This work uses shallow adaptations neural network with one hidden layer. U_2^k, U_1^k, and q_1^k are the weight matrix and biases for the adaptation

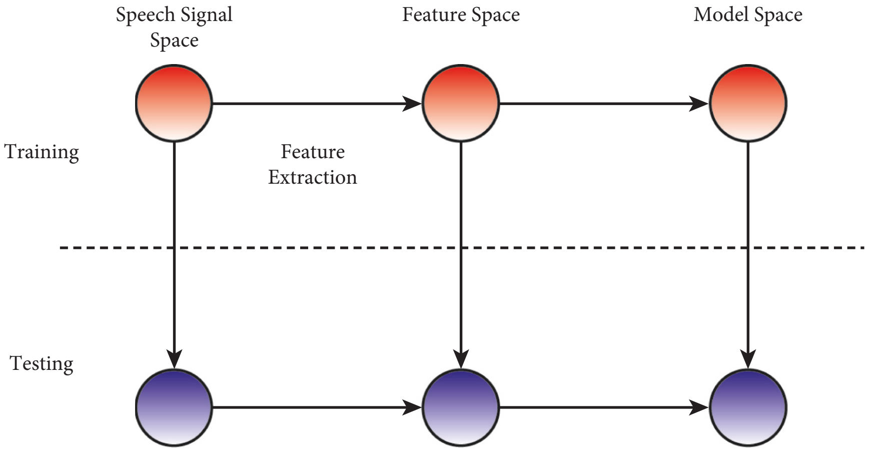


Figure 3: Mismatches among testing and training conditions.

neural network. x^w represents an auxiliary feature. S^k signifies the weight matrices employed to the output from the prior layer. a^k denotes independent speaker biases.

In this work, three positions have been compared, including (i) the acoustic feature input layers, (ii) the output of the first convolution layers, and (iii) the output of the initial fully linked layers, which trails the entire Convolutional Neural Network blocks. When biases are auxiliary to a convolution layer, they must be redesigned to 3-dim tensors. The speaker-dependent dimension bias a^{wk} that adapted output feature maps are provided by

$$\begin{aligned} a^{wk} &= U_2^k \rho\left(U_1^k x^w + q_1^k\right), \\ r^{wk} &= \text{reshape}\left(a^{wk}\right), \\ p_j^{wk} &= \rho\left(\sum_{i=1}^{M} S_{j,i}^k \oplus p_i^{k-1} \oplus a_j^k + r_j^{wk}\right). \end{aligned} \tag{9}$$

As shown in equation (9), p_i^{k-1} denotes *jth* feature maps of the speaker-adapted hidden output of layers k. r^{wk} denotes reshaped tensor. r_j^{wk} is the jth element in the tensor.

Figure 4 shows the frequency masking. In a naturalistic environment, simultaneous masking is a common occurrence. Using the Minimum Masking Threshold (MMT) to obscure watermarks, audio watermarking attempts to achieve this goal. This study offers an improved MFCC by merging it with the masking effect to extract strong acoustic information from loud utterances based on this consideration. This research can use masking models in feature extraction to evaluate which frequency components are more susceptible and how much noise influence may mix into the signal without being predicted. Furthermore, it can calculate the loudness of speech.

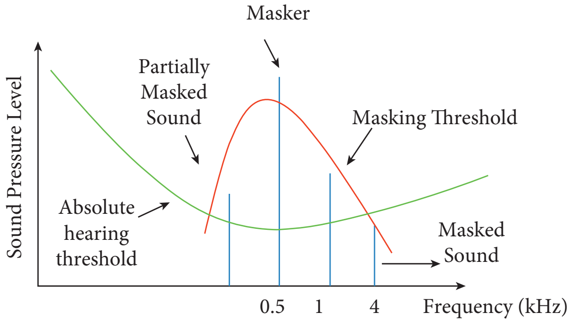


Figure 4: Frequency masking.

3.4. Adaptive Cluster Training (ACT). Figure 5 shows the adaptive cluster training. Different from making adaptations with the auxiliary feature as in the preceding subdivision, numerous factorized adaptation approaches have been suggested to straightly adapt the NN in the model domain in the past few years. Speaker-dependent matrices are utilized to form speaker-dependent hidden layers. The variance among those approaches is the model utilized to evaluate speaker-dependent matrices. Weight matrices bases have been computed as

$$N = \left\{\left\{N^1, \ldots N^k\right\}, \left\{S^1, \ldots, S^L\right\}\right\}. \tag{10}$$

As inferred from the equation (10), $N^k = [S_1^k, \ldots S_Q^k]$ denotes weight matrix basis of layer k and Q indicates the clusters. K represents the overall ACT layer. N^k signifies the weight matrices of non-ACT layer l, and L illustrates the overall non-ACT layer.

The transformation function of the speaker-dependent interpolation vectors is λ^{wk}.

$$\lambda^{wk} = \left[\lambda_1^{wk}, \ldots, \lambda_Q^{wl}\right]^T. \tag{11}$$

As discussed in equation (11), λ_c^{wk} denotes interpolation weights for *cth* clusters. The last adapted weight matrices for an assumed speaker w and the output are provided by

$$S^{wk} = \sum_{c=1}^{Q} \lambda_c^{wk} S_c^k, \tag{12}$$

$$p^{wk} = \rho\left(S^{wk} p^{k-1} + a^k\right). \tag{13}$$

The weight matrices are first decomposed by singular value decomposition (SVD), $S_{m\times n} \approx V_{m\times Q} U_{Q\times n}$, then speaker-dependent square linear layers were employed for

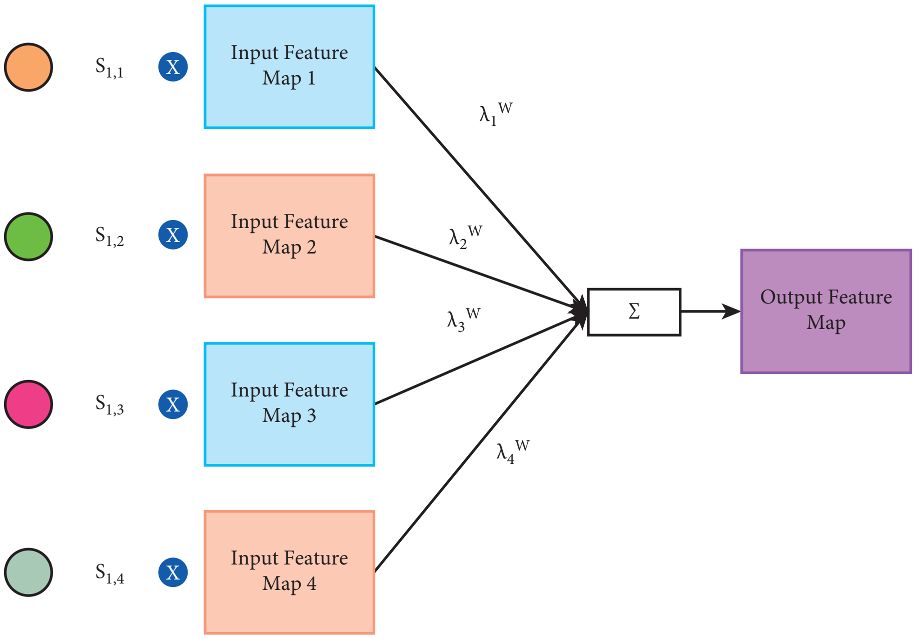


Figure 5: Adaptive cluster training.

bottlenecks. A speaker-adapted weight matrices can be found as

$$S^{w}_{m\times n} = V_{m\times Q} W^{w}_{Q\times Q} U_{Q\times n}. \quad (14)$$

As inferred from the equation (14), $W^{w}_{Q\times Q}$ indicates speaker-dependent square matrices. The speaker-dependent square matrices are decomposed into diagonal matrices plus low-rank matrices.

$$W^{w}_{Q\times Q} = Z^{w}_{Q\times Q} + Q^{w}_{Q\times c} O^{w}_{c\times Q}. \quad (15)$$

As shown in equation (15), $Z^{w}_{Q\times Q}$ represents diagonal matrix and $Q^{w}_{Q\times c}$ and $O^{w}_{c\times Q}$ are two low-rank matrices. The formulation of the diagonal part can be rewritten as follows

$$V_{m\times Q} Z^{w}_{Q\times Q} U_{Q\times n} = \sum_{c=1}^{Q} z^{w}_{c} v_c u_c^{T}. \quad (16)$$

As found in equation (16), v_c denotes *cth* column of V and u_c indicates the cth column of U^T. This structure can be understood as incorporating manifold rank-1 matrices based on speaker-dependent vectors. Motivated by these effective factorization-based adaptation models for CNNs, this study intends further to discover its potential capability for robust noise and speech recognition.

Direct training of the integrated neural network quickly falls into a local optimum as the gradients for the DCNN and acoustic model have different dynamic ranges. The training should be performed in sequence for a robust estimation of the model variables as shown in Algorithm 1. The calculation of improved speech presence probability is a recursive process, which concurrently possesses a strong track/adapt capability and a precise estimation ability for connected statistics by integrating the strong prior data of the speech and noise signal from the exhibition hall. The proposed DCNN-ASRM model enhances the recognition accuracy, performance, signal-to-noise ratio, and noise reduction ratio and decreases the word error rate compared to other existing models.

4. Simulation Results and Discussion

This paper presents the DCNN-ASRM model for accurate speech recognition in the exhibition hall. This study utilizes the speech signal from the CHiME-3 challenge data set [25]. The CHiME-3 environment is Automatic Speech Recognition for multimicrophone tablet devices utilized in an everyday, noisy environment. It signifies an important step forward in terms of realism concerning the previous CHiME challenges. Speech is seized by six microphones embedded in the frame and recorded 24 bits at a multitrack field recorder. The audio was consequently down-sampled to 16 bit 16 kHz for dissemination. This study discusses the recognition accuracy, performance, signal-to-noise ratio, noise reduction ratio, and word error rate compared to other popular models.

The challenge features are as follows:

(i) 6-channel microphone array information,

(ii) Actual acoustic mixing, i.e., talkers speaking in a challenging noisy environment,

(iii) There are five varied noise settings: cafe, exhibition, street junction, public transports, and pedestrian zone.

4.1. Recognition Accuracy Ratio. An audio-visual feature extraction technique based on the DCNN model presented in this research takes feature concatenation a step further by learning to automatically align the two media, resulting in more accurate visual and auditory representations. After the

```
Input: One noisy speech utterance
Output: improved speech presence probability estimation
Initialize the statistics at the initial frame for every frequency
for every time frame $k$do
    for every frequency $l$do
        Compute the mask estimation using equations (12) and (13)
        Compute the posterior SNR using the noise estimation in equation (15)
        Compute the priori SNR
        Compute the gain function or improved speech presence probability based on CNN input, hidden and output layer
    end for
end for
```

Algorithm 1: Deep Convolutional Neural Network Algorithm.

convolutional layer, an additional layer called a pooling layer is added, to be more specific, after a nonlinearity (such as ReLU) is added to the feature maps generated by a convolutional layer. The adoption of the wavelet denoising strategy in the MFCC significantly surpassed the identification accuracy, notably at low SNRs, although at high SNR values, the recognition accuracy remained almost unaltered. SNR levels were averaged to obtain the average identification accuracy in various noise environments. All models received validation using noise-free utterances between every training epoch, and training was terminated when recognition accuracy on the data set declined from the prior epoch or achieved 100%. Two to six epochs of training were sufficient for all models. The recognition accuracy has been computed from equation (10). The suggested DCNN-ASRM model achieves high recognition accuracy ratio of 98.1% compared to other existing models. Figure 6 signifies the recognition accuracy ratio.

4.2. Performance Ratio. Despite considerable success in speech recognition, MFCCs' performance is still unsatisfactory for many real-world applications, despite their widespread use in speech recognition systems. Ambient noise is one of the most common problems with popular spectral characteristics. There is a lot of background noise in many places where automated speech recognition applications would be perfect, making voice-activated devices unfeasible in many cases. Mean values of the proposed method's performance are presented here for noise sources to analyze and verify its performance at all input SNRs. To improve the performance of DCNN, the dropout method is used. There are 100 audio signals chosen for analyzing the performance of the proposed DCNN-ASRM model in speech recognition at the exhibition hall. A Deep Convolutional Neural Network (DCNN) has been employed for training and testing the database because of its superior performance. Training the system thoroughly through weight connectivity, local connectivity, and polling achieves excellent testing performance. This study introduces acoustic features based on higher-order indicators of the speech signal. When paired with MFCCs, these features have been proven to generate greater recognition accuracies in a noisy environment. The performance ratio has been computed from equations (12) and (13). The suggested DCNN-ASRM model achieves a high performance ratio of 97.2% compared to other systems. Figure 7 illustrates the performance ratio.

4.3. Signal-to-Noise Ratio. The first metric for evaluating the effectiveness of speech enhancement and speech recognition algorithms is one based on the signal-to-noise ratio (SNR). However, the typical SNR measure does not correspond with the quality of the speech since the average across the entire signal duration may remove key information. As a result, the SNR is estimated in short segments and then averaged to address this issue. The term "segmental SNR" refers to the method used to calculate SNR. SegSNR is a step up from traditional SNR. When SNR is computed, an upper and lower threshold is specified to replace any frames with an abnormally high or low signal-to-noise ratio. Quality estimates are based on the average SNR of all frames in a sequence. The SNR ratio has been computed from equation (14). The suggested DCNN-ASRM model achieves less SNR ratio of 10.3% than other models. Figure 8 indicates the signal-to-noise ratio.

4.4. Word Error Rate. Word error rate (WER) is a typical statistic used to measure the accuracy of speech conversion performed by Automatic Speech Recognition (ASR) systems. A 5–10% WER is considered good quality and is ready to use. The WER is the number of errors divided by the total amount of words. To obtain the WER, add the substitutions, insertions, and deletions in a series of recognized words. Gaussian white noise may cause word error rates to rise from less than 5% to more than 20% when applied to Large Standard Vocabulary Continuous Speech Recognition (LVCSR) tasks such as recognizing the 5,000 words in the Wall Street Journal corpus, even using compensatory strategies. Even for continuous digit identification, the word error rate often exceeds 10% in the presence of high noise levels. The word error rate has been computed from equation (15). The suggested DCNN-ASRM model achieves a lower WER rate of 9.2% than other models. Figure 9 demonstrates the word error rate.

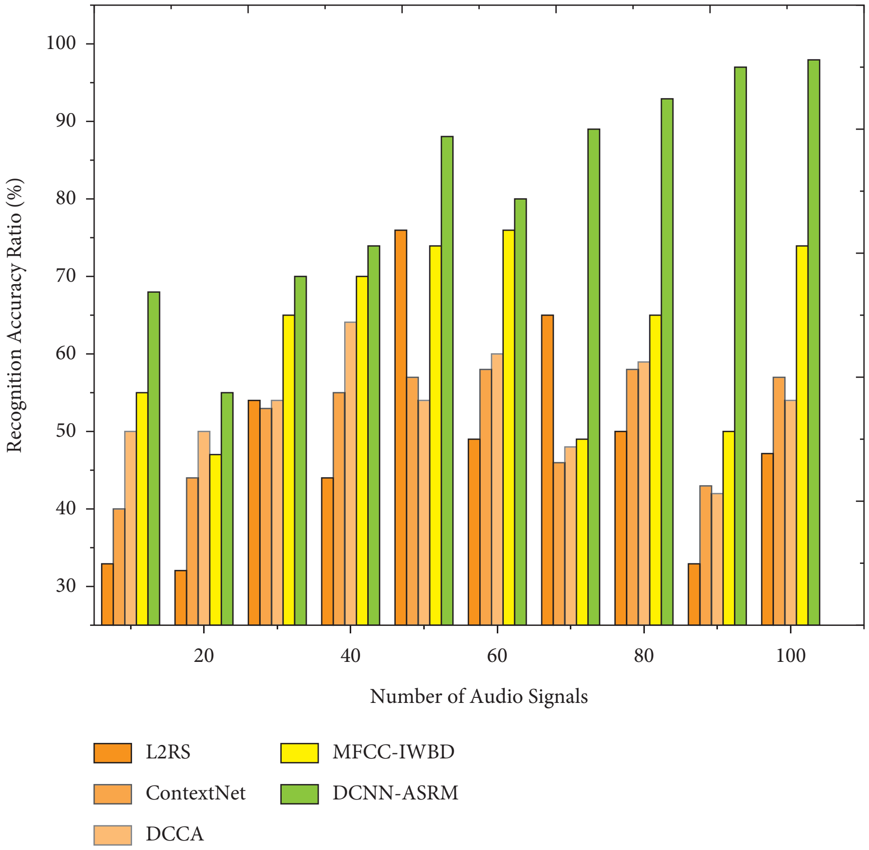


Figure 6: Recognition accuracy ratio.

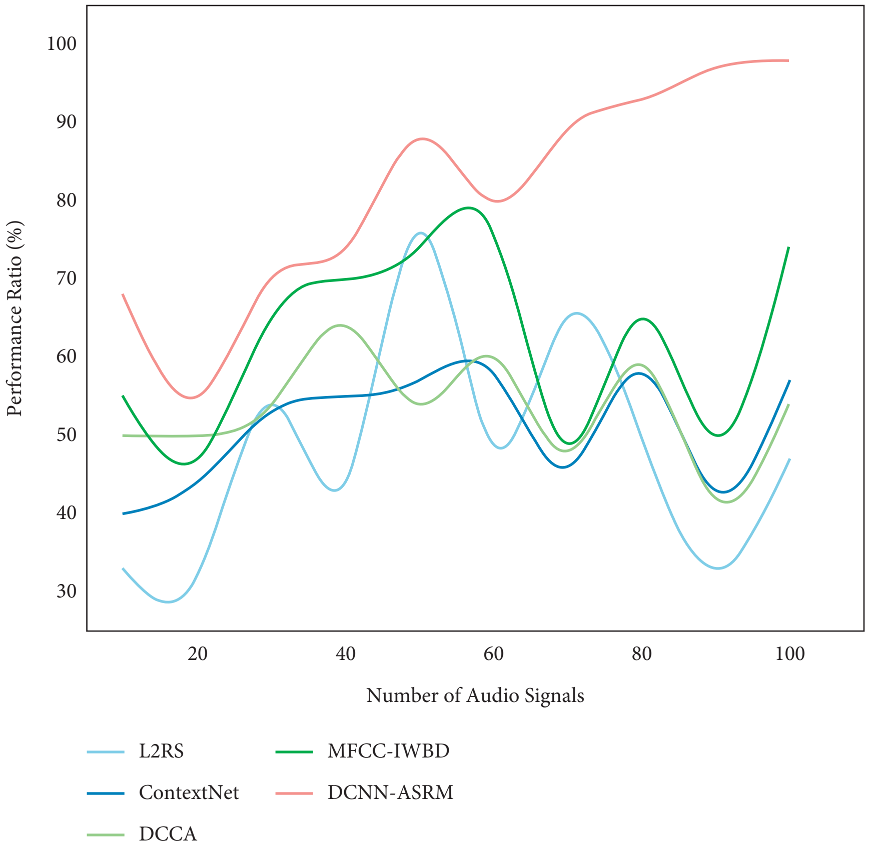


Figure 7: Performance ratio.

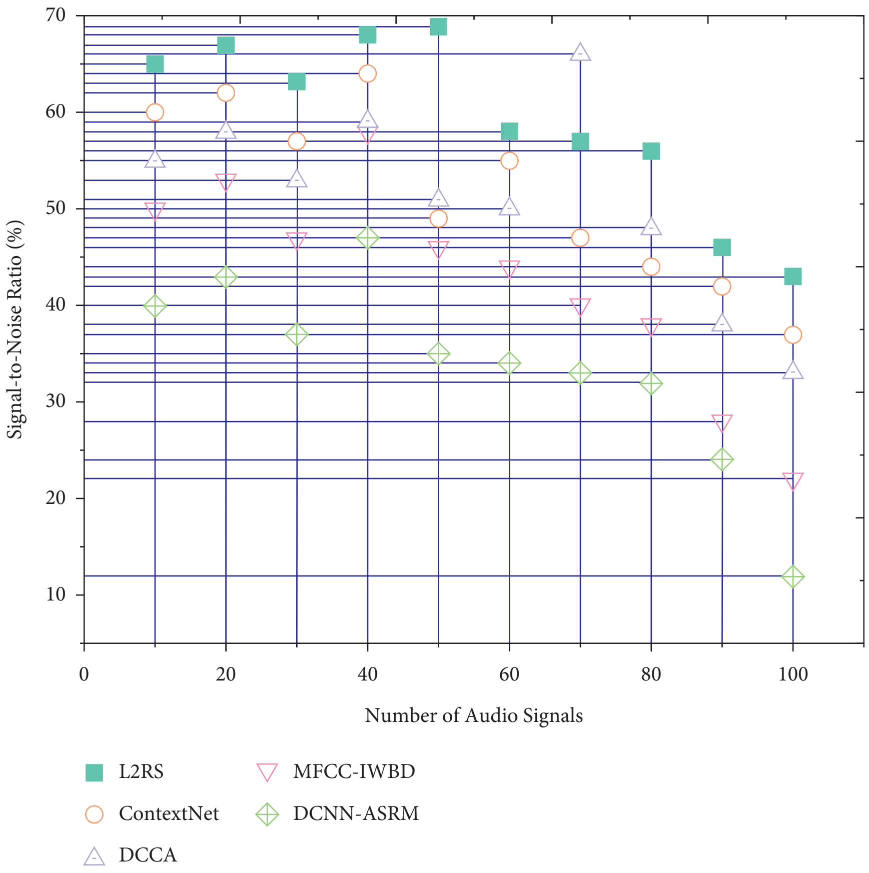


Figure 8: Signal-to-noise ratio.

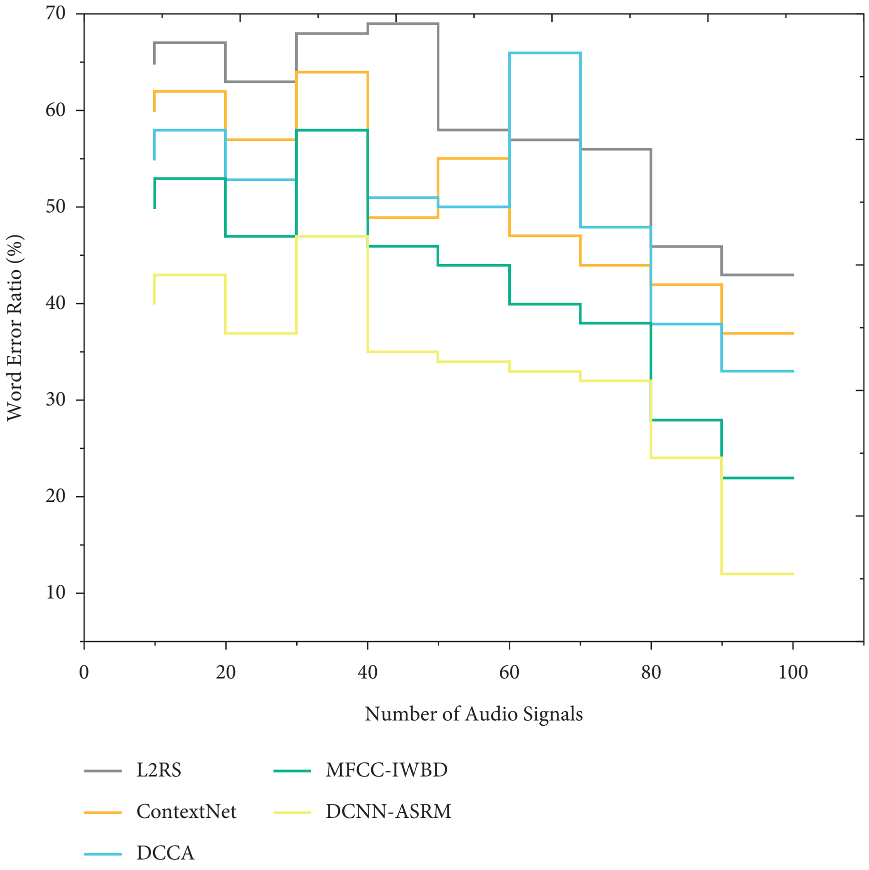


Figure 9: Word error rate.

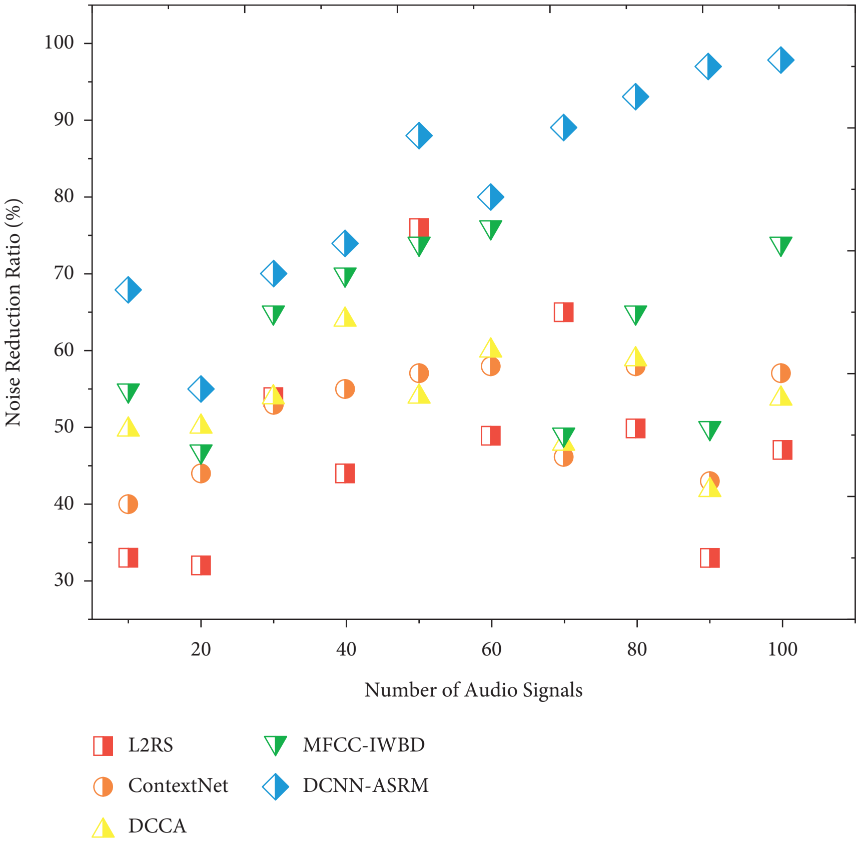


Figure 10: Noise reduction ratio.

4.5. Noise Reduction Ratio. For low SNRs and noise types that are not steady, many of the mentioned deep learning models aim for high noise attenuation and, as a result, may still impair voice quality. To solve this issue, it is proposed that noise reduction be used initially, followed by the restoration of natural-sounding speech. Even when the model was evaluated with data and noise not included in the training set, the findings show that the recommended CNN structure gives higher denoising capabilities than filtering in noise reduction. The suggested system with a single skip connection, on the other hand, has far greater noise-reducing capabilities. The noise reduction ratio has been computed from equation (16). The suggested DCNN-ASRM model attains a high noise reduction ratio of 96.5% compared to other existing models. Figure 10 shows the noise reduction ratio.

The proposed DCNN-ASRM model enhances the recognition accuracy, performance, signal-to-noise ratio, and noise reduction ratio and decreases the word error rate compared to other existing Learning-to-Rescore (L2RS), ContextNet, Deep Canonical Correlation Analysis (DCCA), and Mel-Frequency Cepstral Coefficients for Improved Wavelet-Based Denoising (MFCC-IWBD) methods.

5. Conclusion

This paper presents the DCNN-ASRM model for speech recognition in an exhibition hall. The objective of ASR research is to address the different issues relating to speech recognition. This paper offered a new auditory filter modeling-based feature extraction technique for noisy speech recognition. Based on higher-order sub-band filter speech signal indicators, a new acoustic feature extraction technique was proposed. Preliminary tests demonstrate that combining these features with MFCCs can enhance the robustness of the speech recognition model in exhibition hall noise conditions. The baseline CNN gained a high overall accuracy. Paralleled to the conventional Convolutional Neural Network structure in Automatic Speech Recognition, filter and pooling are small, and the more extensive input feature map. This modification permits us to improve the number of convolution layers to ten. A comprehensive analysis of padding, pooling, and input feature map assortment is executed. The simulation analysis validates that the suggested DCNN-ASRM model enhances the recognition accuracy ratio of 98.1%, performance ratio of 97.2%, and noise reduction ratio of 96.5% and reduces the word error rate by 9.2% and signal-to-noise ratio by 10.3% compared to other existing models.

Data Availability

No data were used to support this study.

Conflicts of Interest

The author declares that there are no conflicts of interest regarding the publication of this article.

References

[1] F. Abreu Araujo, M. Riou, J. Torrejon et al., "Role of non-linear data processing on speech recognition task in the framework of reservoir computing," *Scientific Reports*, vol. 10, no. 1, pp. 1–11, 2020.

[2] H. Bourouba and R. Djemili, "Feature extraction algorithm using new cepstral techniques for robust speech recognition," *Malaysian Journal of Computer Science*, vol. 33, no. 2, pp. 90–101, 2020.

[3] N. Saleem and M. I. Khattak, "Regularized sparse decomposition model for speech enhancement via convex distortion measure," *Modern Physics Letters B*, vol. 32, no. 22, Article ID 1850262, 2018.

[4] Y. M. Kang and Y. Zhou, "Fast and Robust Unsupervised Contextual Biasing for Speech Recognition," 2020, https://arxiv.org/abs/2005.01677.

[5] T. Iio, Y. Yoshikawa, M. Chiba, T. Asami, Y. Isoda, and H. Ishiguro, "Twin-robot dialogue system with robustness against speech recognition failure in human-robot dialogue with elderly people," *Applied Sciences*, vol. 10, no. 4, p. 1522, 2020.

[6] R. H. Gifford, J. H. Noble, S. M. Camarata et al., "The relationship between spectral modulation detection and speech recognition: adult versus pediatric cochlear implant recipients," *Trends in Hearing*, vol. 22, Article ID 2331216518771176, 2018.

[7] C. Spille, B. Kollmeier, and B. T. Meyer, "Comparing human and automatic speech recognition in simple and complex acoustic scenes," *Computer Speech & Language*, vol. 52, pp. 123–140, 2018.

[8] A. Baruwa, M. Abisiga, I. Gbadegesin, and A. Fakunle, "Leveraging End-To-End Speech Recognition with Neural Architecture Search," 2019, https://arxiv.org/abs/1912.05946.

[9] C. X. Qin, W. L. Zhang, and D. Qu, "A new joint CTC-attention-based speech recognition model with multi-level multi-head attention," *EURASIP Journal on Audio Speech and Music Processing*, vol. 2019, no. 1, pp. 1–12, 2019.

[10] U. Sharma, S. Maheshkar, A. N. Mishra, and R. Kaushik, "Visual speech recognition using optical flow and hidden Markov model," *Wireless Personal Communications*, vol. 106, no. 4, pp. 2129–2147, 2019.

[11] A. B. Nassif, I. Shahin, I. Attili, M. Azzeh, and K. Shaalan, "Speech recognition using deep neural networks: a systematic review," *IEEE Access*, vol. 7, pp. 19143–19165, 2019.

[12] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, and T. Alhussain, "Speech emotion recognition using deep learning techniques: a review," *IEEE Access*, vol. 7, pp. 117327–117345, 2019.

[13] Y. Dokuz and Z. Tufekci, "Mini-batch sample selection strategies for deep learning based speech recognition," *Applied Acoustics*, vol. 171, Article ID 107573, 2021.

[14] R. Haeb-Umbach, S. Watanabe, T. Nakatani et al., "Speech processing for digital home assistants: combining signal processing with deep-learning techniques," *IEEE Signal Processing Magazine*, vol. 36, no. 6, pp. 111–124, 2019.

[15] H. Meng, T. Yan, F. Yuan, and H. Wei, "Speech emotion recognition from 3D log-mel spectrograms with deep learning network," *IEEE Access*, vol. 7, pp. 125868–125881, 2019.

[16] A. Hannun, A. Lee, Q. Xu, and R. Collobert, "Sequence-to-sequence Speech Recognition with Time-Depth Separable Convolutions," 2019, https://arxiv.org/abs/1904.02619.

[17] M. Alhussein and G. Muhammad, "Voice pathology detection using deep learning on mobile healthcare framework," *IEEE Access*, vol. 6, pp. 41034–41041, 2018.

[18] C. X. Qin, D. Qu, and L. H. Zhang, "Towards end-to-end speech recognition with transfer learning," *EURASIP Journal on Audio Speech and Music Processing*, vol. 2018, no. 1, pp. 1–9, 2018.

[19] B. Kanisha, S. Lokesh, P. M. Kumar, P. Parthasarathy, and G. Chandra Babu, "Speech recognition with improved support vector machine using dual classifiers and cross fitness validation," *Personal and Ubiquitous Computing*, vol. 22, no. 5, pp. 1083–1091, 2018.

[20] A. Shewalkar, D. Nyavanandi, and S. A. Ludwig, "Performance evaluation of deep neural networks applied to speech recognition: RNN, LSTM and GRU," *Journal of Artificial Intelligence and Soft Computing Research*, vol. 9, no. 4, pp. 235–245, 2019.

[21] Y. Song, D. Jiang, X. Zhao et al., "L2RS: A Learning-To-Rescore Mechanism for Automatic Speech Recognition," 2019, https://arxiv.org/abs/1910.11496.

[22] W. Han, Z. Zhang, Y. Zhang et al., "Contextnet: Improving Convolutional Neural Networks for Automatic Speech Recognition with Global Context," 2020, https://arxiv.org/abs/2005.03191.

[23] S. Isobe, S. Tamura, and S. Hayamizu, "Speech recognition using deep canonical correlation analysis in noisy environments," in *Proceedings of the 10th International Conference on Pattern Recognition Applications and Methods ICPRAM*, pp. 63–70, Vienna, Austria, February, 2021.

[24] R. Hidayat and A. Winursito, "A modified MFCC for improved wavelet-based denoising on robust speech recognition," *International Journal of Intelligent Engineering and Systems*, vol. 14, no. 1, pp. 12–21, 2021.

[25] Spandh, "Chime_challenge," 2015, http://spandh.dcs.shef.ac.uk/chime_challenge/chime2015/.