IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 8, 2021, accepted September 2, 2021, date of publication September 14, 2021, date of current version October 1, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3112535

Automatic Speech Recognition: Systematic Literature Review

SADEEN ALHARBI, MUNA ALRAZGAN, ALANOUD ALRASHED, TURKIAYH ALNOMASI, RAGHAD ALMOJEL, RIMAH ALHARBI, SAJA ALHARBI, SAHAR ALTURKI, FATIMAH ALSHEHRI, AND MAHA ALMOJIL

Department of Software Engineering, King Saud University, Riyadh 11451, Saudi Arabia

Corresponding author: Sadeen Alharbi (sadalharbi@ksu.edu.sa)

ABSTRACT A huge amount of research has been done in the field of speech signal processing in recent years. In particular, there has been increasing interest in the automatic speech recognition (ASR) technology field. ASR began with simple systems that responded to a limited number of sounds and has evolved into sophisticated systems that respond fluently to natural language. This systematic review of automatic speech recognition is provided to help other researchers with the most significant topics published in the last six years. This research will also help in identifying recent major ASR challenges in real-world environments. In addition, it discusses current research gaps in ASR. This review covers articles available in five research databases that were completed according to the preferred reporting items for systematic reviews and meta-analyses (PRISMA) protocol. The search strategy yielded 82 conferences and articles related to the study's scope for the period 2015–2020. The results presented in this review shed light on research trends in the area of ASR and also suggest new research directions.

INDEX TERMS Speech recognition, automatic speech recognition, ASR systematic review, ASR challenges.

I. INTRODUCTION

In recent decades, researchers have been increasingly interested in automatic speech recognition (ASR) since speech is a method of communication between people [1]. ASR began with simple systems that responded to a limited number of sounds and has evolved into sophisticated systems that respond fluently to natural language. Because of the desire to automate simple tasks that require human-machine interaction, there has been increasing interest in ASR technology [2]. ASR can be defined as the process of deriving the transcription of speech, known as a word sequence, in which the focus is on the shape of the speech wave [1]. In actuality, speech recognition is difficult because of the diversity in speech signals [1]. Currently, ASR is widely applied in many functions, such as weather reports, automatic call handling, stock quotes, and inquiry systems [2].

Communication can be divided into human-human communication and human-machine communication. Human-to-human communication may be limited depending on the language used, as speakers may need a third party to translate speech, such as in unified messaging systems [1]. More recently, human-machine communication has improved greatly by using speech techniques, for example, voice search, games, and interaction systems in the context of a household living room [1]. According to [1], ASR studies are affected by the following:

- **Number of Speakers.** To train a system, speech from a large number of users is needed.
- **Nature of the Speech.** The user's voice is more easily recognized in an isolated recognition system by having the speech uttered word for word with pauses in between them.
- **Vocabulary Size.** Speech recognition systems vary based on the number of words that they can recognize.
- **Spectral Bandwidth.** If bandwidth decreases, the performance of the trained ASR system will be worse, and vice versa.

In this research, we aim to help other researchers by making a systematic literature review of automatic speech recognition that will provide them with the most significant topics published in the last six years. Also, this research will help to specify the recent major challenges and the research gaps in automatic speech recognition. Moreover, it will provide them with future research directions in this area.

The associate editor coordinating the review of this manuscript and approving it for publication was Jing Liang.

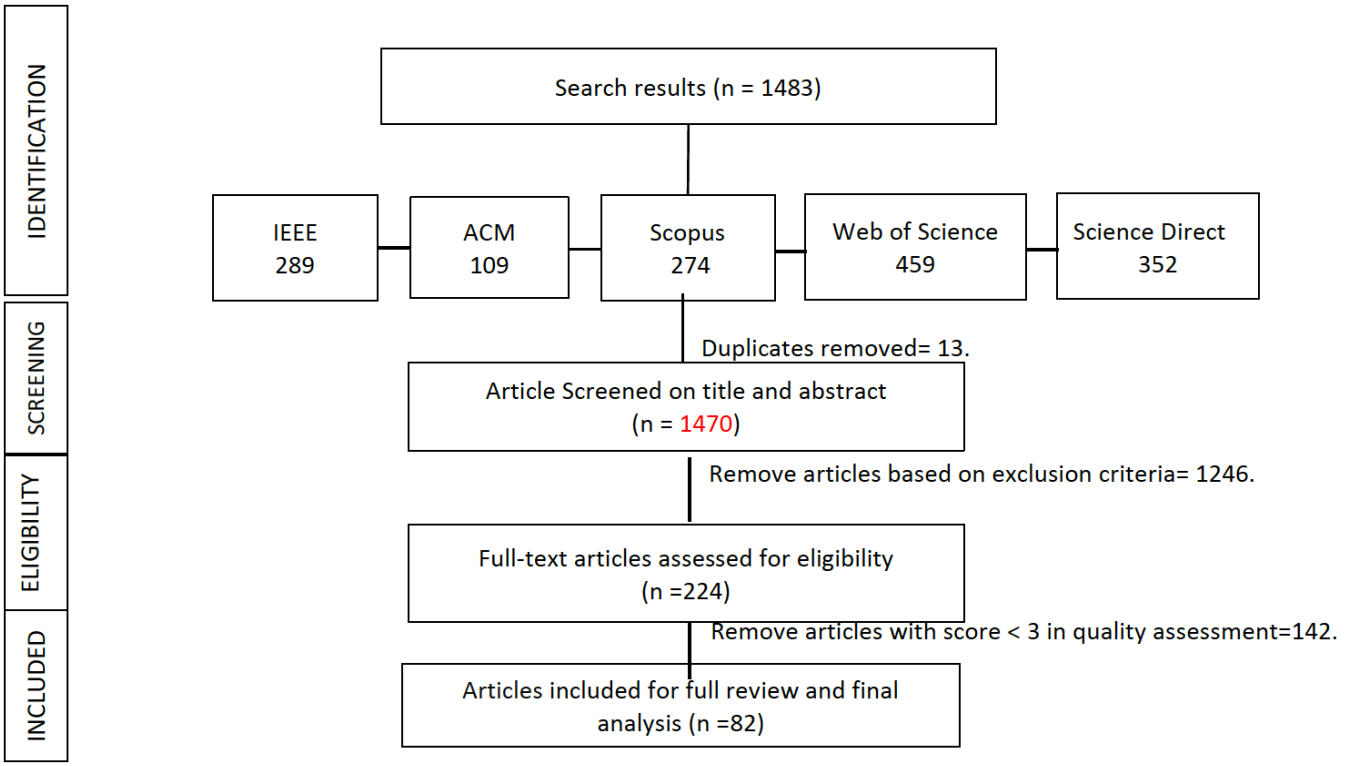

FIGURE 1. PRISMA flow chart.

The rest of the paper is organized as follows. In Section II, the research methodology is described. This is followed by a primary review of important articles in Section III. In Section IV, the selected articles are then reviewed based on their characteristics. In Section V, the research questions are answered. The limitations of this research are stated in Section VI. Finally, Section VII provides the conclusion.

II. METHODOLOGY

In this section, the method used to conduct this research is described. First, the research questions are described, followed by the search strategy. The inclusion criteria are then stated; finally, the quality assessment process is presented. The systematic review of the literature was conducted by applying the preferred reporting items for systematic reviews and meta-analyses (PRISMA) protocol [3]. The PRISMA flow chart is shown in Fig. 1. The detailed search and selection process is illustrated in Fig. 2, which was created using Lucidchart [4] (as were the rest of the figures in this section).

A. RESEARCH QUESTIONS

The first step of this systematic review was the identification of the research questions. The goal of this study was to provide a review of the recent studies in ASR with a focus on English language speech. Therefore, five research questions were defined as follows:

- RQ1: What research topics have been addressed in recent ASR research?
- RQ2: What are the major challenges in ASR?
- RQ3: What are the current research gaps in ASR?
- RQ4: What are future research directions in ASR?
- RQ5: What are the datasets used in the reviewed papers?

RQ1 aims to provide the publication trends in ASR research and the existing speech recognition issues that the authors of the articles have tried to address. RQ2 aims to review the major challenges in automatic speech recognition. RQ3 aims to investigate the current research gaps in automatic speech recognition. RQ4 aims to provide an overview of the future directions for research in automatic speech recognition. RQ5 aims to demonstrate the most applied datasets in recent articles.

B. SEARCH STRATEGY

To gain an overview of the publications in the automatic speech recognition field, the network analysis interface for literature studies (NAILS) project software [5] was used. It is free, open-source software that is used to analyze literature studies. It was used to analyze articles in ASR, from 2015 through 2020, which consists of about 4274 publications, as shown in Fig. 3. The NAILS software analyzes publication information from the Web of Science and provides information about the timeline of publication. The literature

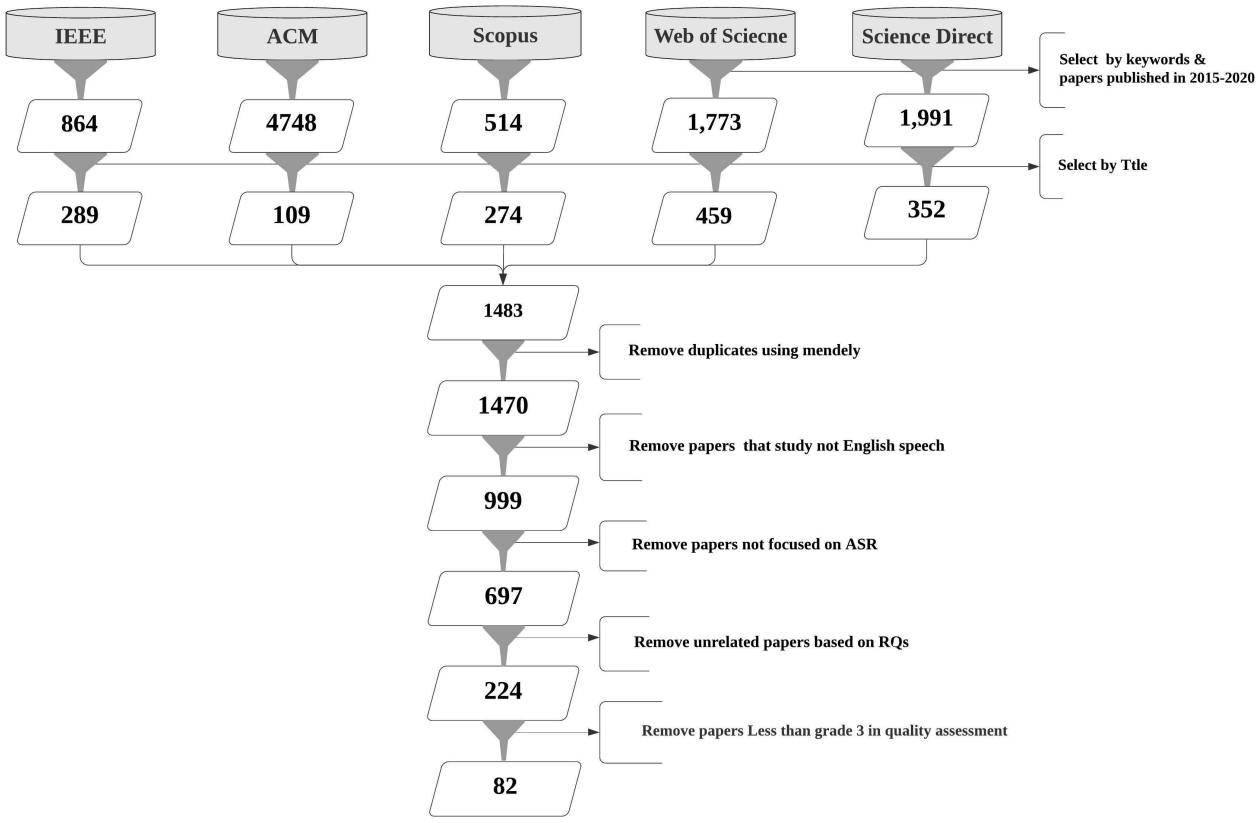

FIGURE 2. Search and selection process.

in ASR has increased in 2018. The NAILS software also provides insights about authors in the field, such as important authors (as defined by most productive authors and most cited authors) sorted by the number of articles published and by the total number of citations. In addition, it shows the important articles (as measured by the most popular articles and most cited articles) sorted by the number of published articles within a dataset and by the total number of citations. Furthermore, it provides important keywords (e.g., the most popular keywords and most cited keywords) sorted by the number of articles in which the keyword is mentioned and by the total number of citations for the keyword. Moreover, it can sort the top 25 articles using three measures of importance: (i) in-degree in the citation network, (ii) citation count provided by the Web of Science, and (iii) PageRank score in the citation network.

After the overview analysis, the search process was conducted by searching for journal articles through five databases, which were:

- *IEEE Xplore Digital Library*
- *ACM Digital Library*
- *Scopus*
- *The Web of Science*
- *Science Direct*

The specific search in each database was by using titles with keywords as in the following: ''artificial intelligence'' AND (''speech recognition'' OR ''automatic speech recognition'').

C. STUDY SELECTION AND INCLUSION CRITERIA

The results of the search were retrieved and stored using the Mendeley Reference Management Software®. Inclusion and exclusion criteria were identified for articles in the study, as shown in Fig. 4. The inclusion criteria were:

- Article published during the period 2015–2020.
- The article was relevant to the topic of ASR.
- The language examined in the ASR article was English.
- The article focus was related to the RQs.
- Articles with a quality assessment grade of at least three (as defined in this study).

These criteria were applied to filter the articles. The research's focus was on recent articles in ASR; therefore, the first criterion was to include articles published during the period 2015–2020. This criterion was applied using the databases' research boundaries. The second criterion was to filter the articles by the speech language examined. The focus was on articles in which the examined speech language was English. This criterion was applied using the Mendeley software. For ASR articles that examined English, the article focus had to be related to the RQs. For this criterion, Rayyan [5] was used, which is a web application that shows an article's publication information and abstract that assisted the authors of this systematic review to collaborate

FIGURE 3. Summary of NAILS results for ASR publications for 2015–2020.

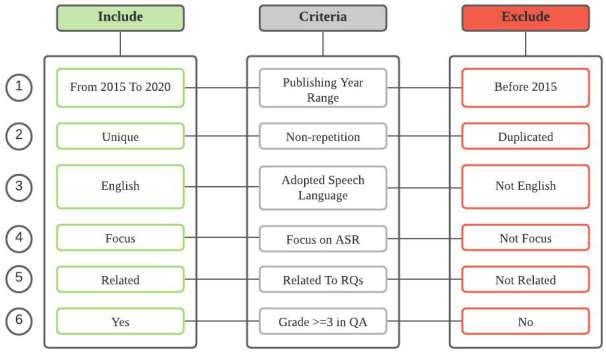

FIGURE 4. Search and selection process.

and vote on articles based on the RQ criteria. There were three voting options: include, exclude, and maybe. Moreover, Rayyan allows hiding individual voting decisions from other team members. Eight writers used the Rayyan website for this evaluation, and each article was voted on anonymously by two individuals. Each criterion was applied separately, as shown in Fig. 2. All the articles that fit the other criteria and that received two "include" votes or one "include" and one "maybe" were kept in the dataset. Articles with two "exclude" votes or one "exclude" and one "maybe" were excluded. Articles that received two "maybe" votes and one "include" and one "exclude" were examined by a third reviewer; in those cases, the third reviewer cast a deciding vote on including or excluding the article.

D. QUALITY ASSESSMENT

The quality assessment process shown in Fig. 5 was based on the following predefined quality questions [6]

- Are the aims of the research clearly stated?
- Does the article provide new techniques or contributions in ASR?
- Are there any challenges of ASR mentioned in the article?
- Does the article provide answers to the formulated RQs?

Each of these quality assessment questions answered in the affirmative was worth one point toward the quality score. All the authors were involved and assessed the articles according to the quality questions and associated research questions. The research article was selected if it had a quality score greater than or equal to three. The evaluation process was as follows: if the article answered the question fully, it received one point; it received 0.5 points if it partially answered the question; it received zero if it did not answer the question. Because the fourth question was a composite, it was divided into four sub-questions based on the individual research questions. Each sub-question had a score, and at the end, the result was divided by four to get one overall result for that item. After the scoring, the points were summed for all the quality questions. If the article received a non-integer total score, it was rounded to the nearest digit (for example, a 3.4 would be recorded as a three). Only articles with total

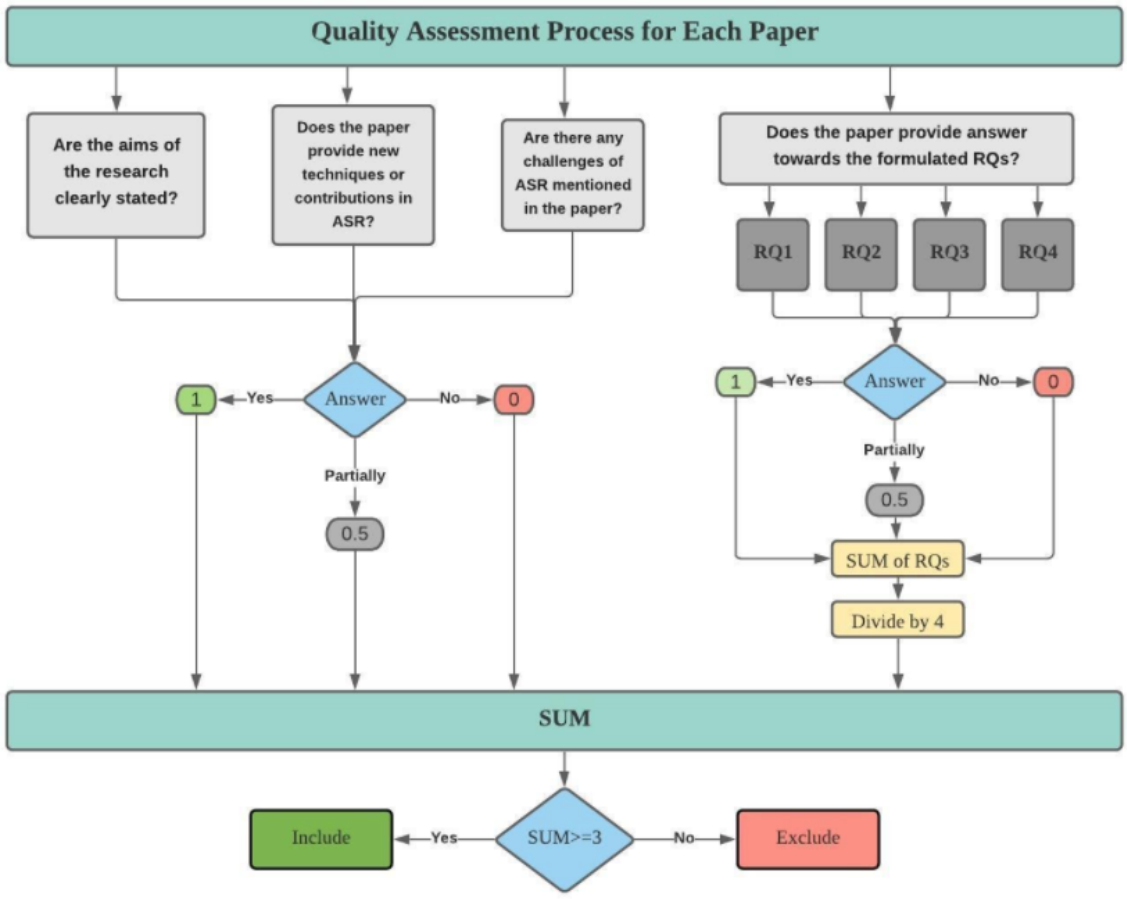

FIGURE 5. Quality assessment process.

scores of three or more after this assessment were kept as meeting this criterion.

III. CHARACTERISTICS OF THE ARTICLES

To cover the trends of the articles, an overview of those selected for the research according to their year of publication is presented in Fig. 6. The characteristics of the reviewed articles are presented in Table 1.

A. CLASSIFICATION OF ARTICLES BASED ON DOMAIN PROBLEMS

1) **Noise and Reverberation** Several techniques have been developed to detect target speech in noisy environments. A very recent study [17] proposed a hybrid-task learning system that frequently switches between multi and single-task learning (depending on whether the input is real or simulated data) to improve robust speech recognition in noisy and reverberant environments. In [22], the authors created an enhanced power-normalized cepstral coefficients algorithm to improve ASR in which there is real-world noise and other acoustic distorting conditions.

The researchers in [28] proposed a front-end speech parameterization technique that is robust with respect to both noise and pitch variations. An ASR system was trained on speech data collected from both adult and child speakers, and testing was done on both clean and noisy speech from children. The aim was to enhance the noise robustness of that ASR system. The effectiveness of that approach has been verified on an ASR system developed with DNN-HMM-based acoustic modeling. In [100], the researchers examined an ASR system with music in the background. They used two approaches. The first was based on multi-condition training of the acoustic models. The second one denoised autoencoders and then conducted acoustic model training on preprocessed data. The results showed that all the techniques they investigated could significantly improve the recognition of speech that was distorted by music.

In [62], the aim was to combine speech enhancement techniques and feature normalization methods. The researchers transformed an estimate of the noise power spectral density to the MFC domain, where they subtracted it from the noisy mel-frequency cepstral coefficients MFCCs. They showed that this process was superior to the application of CMVN alone. The improvement in performance was best in low signal-to-noise ratios.

TABLE 1. Classification-wise breakdown of the reviewed articles.

Classification of Articles Based on Domain Problems				Classification of Articles Based on Natural Language Preprocessing			Classification of Articles Based on Device Efficiency
Noise and Reverberation	*Speech Overlapping*	*Signal Processing*	*Adaptation*	*Vocabulary*	*Pronunciation*	*Dialect*	*Microphone*
[17],[22]	[18]	[20],[25]	[96]	[19],[24]	[19]	[21]	[18]
[27],[28]	[23]	[32],[37]	[98]	[30],[31]	[36]	[26]	[20]
[33],[38]	[29]	[41],[46]	[99]	[35],[42]	[49]	[72]	[27]
[39],[40]	[34]	[85],[86]		[45],[53]	[18]		[38]
[44],[47]	[71]	[87],[90]		[78],[79]	[73]		[43]
[48],[51]		[91],[92]		[81],[82]	[75]		[47]
[52],[55]		[93],[94]		[83],[84]	[75]		[50]
[56],[57]		[95]			[76]		
[58],[59]					[77]		
[60],[61]							
[62],[63]							
[64],[65]							
[66],[67]							
[68],[70]							

An approach in [63] extends a generative model-based multi-channel noise reduction approach. It was dominance-based locational and power spectral characteristics integration (DOLPHIN), and it used the generative-discriminative hybrid approach. The researchers showed that a generative-discriminative hybrid approach that incorporates a DNN-SME into DOLPHIN was beneficial for a multi-condition noise-reduction task, and it was superior to the conventional approaches.

In another technique presented by [65], Discrete Cosine Series (DCS) for noise robust ASR was proposed as a feature set. The temporal and spectral modulations were computed with only a few filters of DCS, and they were based on a short frames spacing. This reduced the effects of slowly varying noise typically accounted for by long-term frames. The results showed that individual components of the DCS algorithm were highly accurate for both reducing additive noise and reverberation.

The aim of [66] was to establish a new method for weighting two-dimensional time-frequency representation of speech. The weighting was done using auditory saliency to create maps. Then, they modeled the mechanism for grabbing human auditory attention. Before extracting ASR, maps were used to weight the T-F representation of the speech. Experimental methods were used to determine the weight, and experiments were done on the Aurora-4 corpus. In demonstrating the effectiveness of the proposed methods. The error rate was reduced from 5.3 % to 4% compared to using a multi-stream system. Combining multi-stream systems with the proposed technique and a single stream system using conventional spectral masking techniques reduced the error rate to 0.

In [69], the researchers proposed a robust technique that parameterized ambient noise and pitch variations at the front end of speech. That approach captures a short-time magnitude spectrum by discrete Fourier transform, which uses variational mode decomposition (VMD) to break it into several components. Then, it smooths the spectrum by discarding higher-order components. After that, it reconstructs the spectrum using only the first two modes to smooth the spectrum. The smoothed spectra are used to compute MFCCs. After evaluating the new approach on the ASR, the results showed that the acoustic features were more robust with respect to ambient noise and pitch variations than conventional MFCC.

In [33] the authors employed two stages that detect and redact the environment noise to perform speech recognition in human-robot interactions. The proposed system automatically determines how to enhance speech quality based on the signal-to-noise ratio (SNR). In the second stage, independent component analysis (ICA) and subspace speech enhancement (SSE) are employed for noise reduction. Similarly, another study [39] applied the signal-to-noise ratio combined

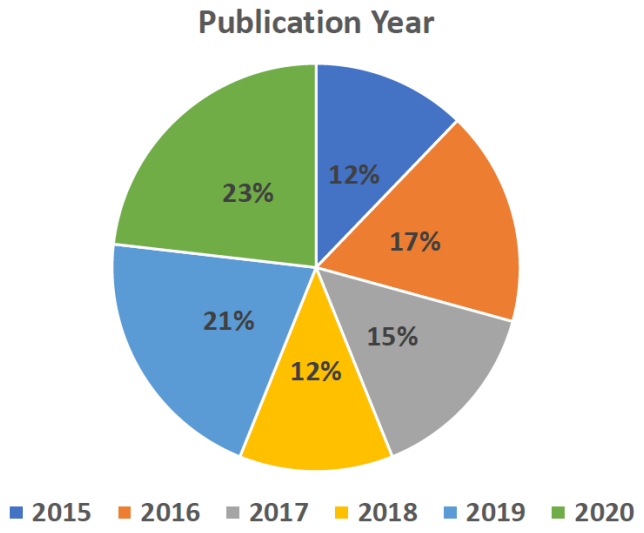

FIGURE 6. Year of publication for the selected studies.

with progressive learning (PL) in an appropriately named SNR-PL framework that provides better speech intelligibility for all SNR levels. In order to reduce the degradation of the desired signal by both additive reverberation and noise, the authors of [44] provided a state-of-the-art ASR system that enhances the signal in the feature domain and uses back-end methods for a wide range of reverberation and noise conditions. The proposed system clearly demonstrated the benefit of speech enhancement processing. Furthermore, in [48], the authors created a very deep convolutional residual network (VDCRN) that included batch normalization and residual learning for speech recognition using a combination of factor aware training (FAT) and cluster adaptive training (CAT). Another way of decreasing the environment noise is to employ spectro-temporal processing methods with the speech signal. In [51], the authors used a framework that combines spectro-temporal feature extraction and the training of neural network-based acoustic models into a single process. They proposed two improvements on recent advances in neural net technology to evaluate speech contaminated with new types of noise by using an artificial neural network (ANN) approach. Similarly, in [52], the authors studied the spectro-temporal effect on ASR. They provided a novel framework for a modulation filter to remove the spectro-temporal modulations of the speech signal using deep variational networks.

Some of the articles presented various methods to improve the performance of ASR systems in noisy environments, such as in [55], which used self-modeling multivariate autoregressive (MAR) models with Riesz envelope estimation. At the same time, MAR models are widely used in forecasting. Within the same scope in [56], a neural network can be used to calculate the audio signal's angle. A forward-bound neural network is then used to deal with the noise. The signal can then be fed into an ASR system to improve performance with robots in noisy environments. The authors of [47] discussed addressing the challenge of noisy ASR tasks by utilizing a neural beamformer in addition to proposing an architecture of multiple channels in end-to-end ASR. This allows deduction in recognizing multichannel speech to enhance it based on an ASR objective, leading to a comprehensive framework that works effectively with a noisy background. The authors of [40] sought to obtain an accurate speech estimate from noise without requiring specific knowledge about the noise. They discussed the novel realization of integrating full-sentence speech correlation with clean speech recognition to enhance conventional speech methods based on a multiframe. They used a Zero-mean Normalized Correlation Coefficient (ZNCC) as the comparison measure. The results show that the proposed approach was able to significantly outperform conventional methods that use optimized noise tracking in terms of ASR. A novel joint training framework for speech separation and recognition [57] was provided to build a larger neural network, which jointly adjusts the weights in each model by concatenating a DNN-based speech separation frontend with more noise and reverberation. In order to reconstruct noise-robust features in various noise conditions, the authors of [58] applied DNN-based speech feature enhancement (FE) using a direction-of-arrival (DOA)-constrained independent component analysis (DCICA) to obtain multichannel input signals. To solve sensitivity to the recording conditions caused by a high level of background noise, the authors of [38] provided an adapting DNN-based acoustic using an audio database recorded by wireless sensors to train an accurate model for the actual speech processing application. A trained DNN was also used in [27] to take a supervised approach that classifies each time-frequency (TF) binary into noise or speech, which, however, resulted in a degradation in ASR performance in a noisy environment. This led the authors to take an unsupervised approach that decomposes each TF binary into the sum of speech and noise by using a multichannel nonnegative matrix factorization (MNMF). Similarly, the authors of [43] used the basic version of NMF with a variational Bayesian (VB) technique to separate the target speaker's voice from background sources. In addition, they used an amplitude modulation filter bank (AMFB) that implicates prior information of speech to analyze its temporal dynamic and outperformed the commonly used frame-splicing technique of filter bank features in conjunction with a deep neural network (DNN). Another work [59] provided a new factor aware training framework called neural network-based multifactor aware joint training to reduce ASR performance degradation in noisy

environments using combined functional models in the DNN models. In addition, in [60], the authors introduced new all-convolutional networks (A-ConvNets) to reduce the number of free parameters, which consisted of little to no connection layers to solve the feature extraction problem. They found that the proposed approach is effective with 99% accuracy and works well for extended operating conditions (EOCs). Also, in [61], the authors proposed a hybrid neural model that consisted of a multi-layered neural network. It contains a two-memristor synapse to solve the diversity problem in the input data and solve the noise. They found that the average accuracy of the model was 95.4%.

In [67], the authors extended the familiar missing-data bounded marginalization technique from a static to a dynamic filter bank for robust ASR. A well-known theorem from statistics was used to show how the reliability of derivative filter bank features can be expressed in the form of a probability density function. As another contribution, the corresponding HMM state emission likelihood equation (bounded marginalization rule) for dynamic features was derived in closed form. On the CHiME corpus, the new approach showed superior accuracy compared to previous heuristics for handling missing dynamic features. To this author's best knowledge, the average accuracy of 92.58% is the best result reported so far for the 2011 CHiME Challenge.

Other articles to improve noisy environments in schools were presented in [64]. They explored the teacher-student learning approach using a parallel clean and noisy corpus to improve speech recognition in multimedia noise. They incorporated up to 8000 hours of untranscribed data for training, and they presented separate results for sequence-trained models and cross-entropy-trained ones. Compared to a sequence-trained teacher, the best sequence-trained student model reduced the word error rate (WER) by approximately 10.1%, 28.7%, and 19.6% on clean, simulated noisy, and real test sets, respectively. Another semi-supervised learning method known as ''noisy student training'' has shown improved performance for deep networks. Noisy student training is a method that depends on iterative self-training, which leverages augmentation to enhance the performance of the network. The authors in [68] worked to improve noisy student training for ASR, employing (adaptive) SpecAugment as the augmentation method. They were able to filter, balance, and augment the data generated between self-training iterations. They obtained WERs of 4.2%/8.6% on the clean/noisy LibriSpeech test sets by using only the clean 100-h subset of LibriSpeech as the supervised set and the rest (860-h) as the unlabeled set. They also achieved WERs of 1.7%/3.4% on the clean/noisy LibriSpeech test sets by using the unlab-60k subset of LibriLight as the unlabeled set for LibriSpeech 960-h. Thus, they improved upon the previous state-of-the-art clean/noisy test WERs achieved on LibriSpeech 100h (4.74%/12.20%) and LibriSpeech (1.9%/4.1%).

In distant speech recognition, [70] provided an effect of multi-channel processing with modern DNN recognizers. The researchers evaluated multi-channel methods for distant speech recognition in urban environments. The experiments were applied to the third CHiME Challenge database. They analyzed the effects of processing the stages of beamforming, dereverberation, and adaptive noise cancelation, and they discussed back-end processing components.

2) **Speech Overlapping (Simultaneous Conversation)**

Speech overlapping means that several people are talking at the same time. Researchers observed a vital degradation in the performance of ASR systems when speech contained cross-talk. A few recent articles have addressed speech overlapping in ASR. In [18], the authors suggested a target speaker extraction network (TEnet) that identifies and isolates a specific speaker's speech based on a clean voice sample of the speaker to address the problem of multiple people speaking at the same time. They concluded that the proposed method has a high performance with a word error rate (WER) of 22.5% and signal-to-distortion rate (SDR) of 15.5%. Likewise, in [23], the authors proposed a model that divides the interfering speech recognition problem in one channel into three parts: translation, speaker tracking, and speech recognition. They find that it improves by 30% the rate of speech errors.

In [71], the authors combined approaches to address the cross-talk problem called deep clustering (DPCL) by creating a hybrid acoustic model. They obtained a WER of 16.5% on the wsj0-2mix dataset, which is the best performance reported so far.

However, in [29], one of the challenges of automatic speech recognition is identifying children's speech in bilateral interactions because children have weaker communication ability. To address this problem, the authors suggested two methods: semantic response generation and lexical repetition. They concluded that it improved children's speech recognition and was applicable. In addition, in [34], to address the problems of integrating multimedia features and frame alignment between two data streams, the authors proposed WaveNet with a mutual interest for automatic voice and visual speech recognition. It improved performance as it reduced Tibetan singular syllable error by 4.5% and 39.8% on English word error in speech.

3) **Signal processing**

More recent examples of narrative studies within generating an acoustic model in ASR can be found in [32]; its authors presented an end-to-end acoustic modeling approach using convolutional neural networks (CNNs) in which the CNN takes as input a raw speech signal

and estimates the conditional probabilities of HMM states as the output. They found that the proposed approach yields a consistently better system with fewer parameters when compared with the conventional approach of cepstral feature extraction followed by ANN training.

Additionally, in [37], the authors provided a review of modeling the development process of end-to-end (E2E) ASR. They also discussed three different models: a Connectionist temporal classification (CTC)-based model, a recurrent neural network (RNN)-transducer model, and an attention-based model. They looked at the models from different aspects, which were principles, progress and research hotspots, and detailed comparisons from theoretical and experimental views. In terms of results, they found the CTC-based model had the worst effect without an external language model. The gap between it and the other models was very large because it could not learn the language model knowledge by itself; it is just made a conditional independent assumption for the output, while the RNN-transducer was greatly improved on all test sets because, compared to the CTC, it could use the prediction network to learn the language knowledge by itself. Lastly, the attention-based model had the best results of all the end-to-end models; specifically, the decoder's depth had an impact on the results, as the two-layer decoder was better than the one-layer decoder.

Connected to the related studies, in [20], the authors described the enhancement of ASR using linear, mel, and inverse-mel filter banks. They noted that the use of linear or inverse-mel filter banks improved the recognition of children's and adult females' speech.

In addition, the authors of [86] evaluated DNN performance when trained on envelope spectrogram features that represented temporal amplitude modulations as a subband of speech signals. Their method outperformed standard DNNs that trained on different types of features, such as mel and PLP, in both TIMIT phone recognition and AURORA-4 word recognition.

The authors in [87] discussed the use of perceptually motivated subband temporal envelope (STE) features and a time-delay neural network (TDNN) denoising autoencoder (DAE) to improve DNN-based ASR. Improved ASR performance was obtained when features enhanced by TDNNDAE were used in an ASR system using DNN acoustic models. In that scenario, using STE features provided a WER reduction of 9.8% compared to using FBANK features. Another proposal by [88] investigated the possibility of optimizing acoustic models for ASR systems using a variant of evolutionary stochastic gradient descent (ESGD). In [90], the authors investigated various stream fusion methods on a multi-size window fusion. They used posterior-in-posterior-out (PIPO-BLSTMs) and employed them in the context of stream fusion for ASR. The results showed that the turbo fusion approach outperformed the single-window setup by 8.2.

In [90], the authors studied the influence that noisy spectral phase improvement had on the accuracy of ASR when corrupted speech signals were included.

In [25], the authors proposed a speech recognition evaluation method for dysarthric speech recognition systems using an adaptive neuro-fuzzy inference system (ANFIS). They found that the proposed method was effective by employing it to measure the performances of two dysarthric ARS systems based on multiple-view multiple-learner (MVML) and multiple-view single-learner (MVSL) active learning.

The authors of [41] discussed how adults' and children's speech differ significantly due to large deviations in the acoustic correlates. They used a pitch-dependent acoustic mismatch in the context of children's speech recognition on adults' speech-trained models. A low-latency adaptation approach has been explored for their GMM-HMM-based ASR system.

In [46], the authors proposed an online hybrid CTC/attention E2E ASR architecture that replaces all the offline components of a conventional CTC/attention ASR architecture with their corresponding streaming components by using LibriSpeech English and Mandarin tasks (from the Hong Kong University of Science and Technology, HKUST) to decode the speech in a low-latency and real-time manner. The researchers in [91] introduced a combined framework to integrate social signal detection (SSD) and ASR systems based on CTC, which is an end-to-end model. They studied several reference labeling methods regarding social signals, and they confirmed that the end-to-end framework by BLSTM-CTC beat the standard DNN-HMM system with a language model in both SSD and ASR performance.

Going deeper, some studies have discussed the difficulty of decoding ASR graphs. In traditional approaches, the algorithm must extend the graph to generate each newly observed n-gram when the graph is decoded with higher-order language models. This expansion process raises computation time and memory consumption. In [85], the researchers introduced a method to decode ASR graphs by applying an algorithm based on ant colonies. The algorithm used a new language model, without the need to extend it.

Recently, many companies have relied on distributed deep learning to overcome the long time needed for training modern ASR systems. The algorithm for training must be able to converge with a large mini-batch. In [92], the researchers found that Asynchronous Decentralized Parallel Stochastic Gradient Descent (ADPSGD) can run with a much larger batch size than the usually applied synchronous SGD (SSGD) algorithm. In [94], the authors proposed different types of distributed deep learning approaches for ASR,

and they evaluated them on a long short-term memory (LSTM) acoustic model on the 2000-hour switchboard (SWB2000). The experiments confirmed that the LSTM model could be trained by Asynchronous Decentralized Parallel SGD (ADPSGD) in 14 hours with 16 NVIDIA P100 GPUs to reach a 7.6% WER.

Several streaming attention-based sequence-to-sequence (S2S) models have recently been suggested to implement an online ASR system with linear-time decoding complexity. However, those models are delayed during the token generating process because of the lack of future information. To decrease latency, the researchers in [94] introduced several approaches during the training process that leveraged external hard alignments obtained from the hybrid model. In addition, the researchers in [95] introduced a fully streaming E2E ASR system based on transformer architecture, where output is produced shortly after each spoken word. This result was achieved by employing time-restricted self-attention to control the latency of the encoder. Then, the triggered attention (TA) concept is used to control the output latency of the decoder. The proposed model achieved WERs of 2.8% and 7.3% for the test-clean and test-other data sets of LibriSpeech.

4) **Adaptation**

Adaptation is often used to solve mismatch problems. Domain mismatch and robustness is one of the challenging problems for ASR. Pre-trained ASR systems can be purchased and used by companies of any size to build speech-based products. However, domain mismatch is still considered a problem in these applications for many stakeholders who need an optimal result.

In [96], the authors proposed a factorized hidden layer (FHL) adaptation method to investigate the robustness of acoustic models trained on multi-domain data on unseen domains. The authors collected speech data from various applications with different domains. Results on two unseen domains confirmed that FHL was a more useful method of adaptation than the standard fine-tuning network approach. Then, authors in [97] suggested applying domain adaptation for ASR error correction through the machine translation process. They used a machine translation model to learn how to map errors from an out-of-domain ASR to in-domain terms in the corresponding reference files.

ASR accuracy can also be affected by the mismatch in train and test datasets, so adaptation is required. In [98], the authors employed discriminative features as input derived from joint acoustic factor learning for DNN adaptation. The bottleneck (BN) layer of a DNN generates these characteristics, which are referred to as BN vectors. The authors used two types of joint acoustic component learning to generate these BN vectors, which captured speaker and auxiliary information such as noise, phone, and articulatory information of the speech. The authors showed how BN vectors could be used to adapt and enhance ASR performance. In addition, they show that adding BN vectors to standard i-vectors improved performance even further. The experiments were carried out on the Aurora-4, REVERB challenge, and AMI databases.

The authors in [99] examined the adaptation of visual signals to ASR systems. They investigated the AM and LM adaptation for ASR using a speaker face for transcribing a multimedia dataset. Results revealed a small WER enhancement in the transcription of instruction videos after applying the AM and LM adaptation with fixed-length face embedding vectors.

B. CLASSIFICATION OF ARTICLES BASED ON NATURAL LANGUAGE PREPROCESSING

1) **Vocabulary**

Several articles have dealt with improvements in WERs by adding more vocabulary to the language model (LM). This section reviews articles that dealt with vocabulary and improving word vectors to enhance the ASR system.

The authors of [78] provided a novel approach to extract valuable information from out-of-vocabulary (OOV) speech regions in ASR system output. They used a hybrid decoding network with words and subword units. The candidates for OOV regions were identified as subgraphs of subword units in the decoded lattices. They clustered the recurring OOVs to facilitate word recovery. The clustering metrics depended on a comparison of the OOV candidates' subgraphs. That approach outperformed conventional techniques that consider one best subword string to discover the repeating out-of-vocabulary words and find their graphemic representation.

A problem that was addressed in [79] is that the acoustic encoder and the language model are entangled, and this doesn't allow the language model to be trained separately from external text data. They studied two strategies to update the E2E ASR network. They found that by pre-training the subnet with the text data and then fine-tuning the entire E2E network using both labeled and text data, they introduced a new architecture that separates the decoder subnet from the encoder output. As a result, the language model could be easily updated using external text data. Experimental results showed that the new architecture benefitted more from the external text data than the conventional architecture. To improve LM, the researchers in [82] proposed a context-sensitive candidate label approach to smooth the training of recurrent neural network language models (RNNLMs), and it enhanced the ASR performance. The method helped prevent over-fitted and over-confident models, and it could distinguish plausible target words from incorrect ones.

Subwords are the most popular applied output units in an E2E ASR system [83]. The researchers in [81] tested subword regularization with both CTC-based and attention-based ASR models. They found that regularizing subwords with an attention-based model improved the performance of ASR systems. In [81], the same authors introduced a new loss function, n-gram maximum likelihood loss. It yields significant improvement over a character-based model with two different subword vocabularies and text decomposition strategies. They followed the latent sequence decomposition (LSD) framework for using subword units, but they introduced an updated loss function that allowed the ASR model to explicitly perform unit discovery. They showed that the n-gram loss function outperformed standard maximum likelihood loss in the LSD framework. They also showed that uniform greedy sampling of subword units, which is much faster than LSD, was also an effective decomposition strategy when combined with n-gram loss. In [83], the researchers investigated the regularizing influence of the subword segmentation sampling approach on a streaming task of E2E ASR. They evaluated the contribution of subword regularization that relied on the training dataset size, and the results suggested that subword regularization provided a consistent reduction of WER.

A Google group in [84] produced a large vocabulary ASR system for both adults and children by experimenting and comparing the results of applied long short-term memory (LSTM) recurrent networks to convolutional LSTM deep neural networks (CLDNN).

Other recent studies have sought to improve E2E ASR with word embedding learned from text-only data. In [42], the researchers chose to adopt word embedding because off-the-shelf word embedding carrying semantic information learned from a vast amount of text can be easily obtained. An autoregressive decoder was generally used to predict the transcription corresponding to the input speech. The results showed the benefits that word embedding can bring to this type of sequence-to-sequence ASR mode.

In [19], the authors compared several graph-based algorithms and proposed the prior-regularized measure propagation (pMP) algorithm. They evaluated two different frameworks for integrating graph-based learning into state-of-the-art DDN-based speech recognition systems. The first framework utilizes graph-based learning in parallel with a DNN classifier within a lattice-rescoring framework, whereas the second framework relies on an embedding of graph neighborhood information into continuous space using an autoencoder. They showed experimental results on several large vocabulary continuous speech recognition (LVCSR) tasks and showed consistent improvements in WERs under a variety of conditions.

The authors of [24] proposed a fast-learning method for multilayer perceptrons (MLPs) on large vocabulary continuous speech recognition (LVCSR) tasks. The method is suitable for humanoid robots whose CPU/GPUs and memories are limited. The basic concept of this method is to pre-adjust the initial MLP and then train it using an unconventional back propagation (BP) algorithm after restructuring weight matrices via singular value decomposition (SVD). The researchers found that the method accelerates the training processes to about 2.0 times faster with improvements in the cross-entropy loss and frame accuracy. The method can accelerate the training processes to around 3.5 times faster with just a negligible increase of the cross-entropy loss and with a tiny loss of the frame accuracy. In addition, in [30], the authors discussed brain-inspired spiking neural networks (SNN) for speech recognition and evaluated their performance on several large vocabulary recognition scenarios. They mentioned that SNN-based ASR systems achieved competitive accuracy on par with their ANN counterparts across phone recognition, low-resourced ASR, and large-vocabulary ASR tasks. The results of their work are that an SNN-based acoustic model has been revealed as a compelling prospect for rapid inference and unprecedented energy efficiency in a neuromorphic approach. Furthermore, to detect errors and make the estimation accurate for the vocabulary, the authors of [35] discussed deep bidirectional recurrent neural networks (DBRNNs) as classifiers for error detection and accuracy estimation. In addition, in [53], the authors suggested a topic similarity score to specify the variation among word topic distributions and identical sentences, in addition to another word-discourse score, to measure the word appearance probability in a sentence using the word vector and discourse vector that was produced internally to grade the n-best hypotheses of an ASR system. In this work, they tested two types of semantic features: linear discriminant analysis (LDA) topic features and global vector (GloVe) continuous word representations. They achieved 0.29% and 0.51% reductions in WERs.

A Viterbi approximation of latent word language models (LWLMs) for ASR was proposed by the authors of [31]; they concluded that the combination of an n-gram approximation method and the Viterbi approximation method improved ASR performance.

Confusing words is another factor that affects the understanding of speech. In [45], the authors suggested an N-best rescoring system that integrates attentional information for locally confusing words extracted from alternative hypotheses in a conventional speech recognition system. A top-down selective attention (TDSA) mechanism was used to adapt the input feature by maximizing the log-likelihood of the feature given confusing words. They used a designed neural network

to output data-dependent rescoring weights in the proposed CM (class model), and it was optimized by minimizing the WER. They found that emphasis in the proposed system can be applied so that ASR systems can generate competing hypotheses and provide the gradient of the input feature for confusing words.

2) **Pronunciation**

Many articles in the field of ASR have addressed pronunciation problems that could affect the work. Regular ASR usually employs phoneme-based pronunciation lexicons provided by linguistic experts. If hand-crafted pronunciations cannot include the vocabulary of a new domain, then the best solution is to use a grapheme-to-phoneme (G2P) converter to obtain pronunciations for new words. In [73], the researchers suggested a probabilistic framework approach that is grapheme-based ASR. As a stage in ASR system training, lexical modeling combines pronunciation learning and employs both lexical resources and acoustics. That approach was tested on four lexical resource-constrained ASR tasks, and it was compared with the conventional two-stage approach, where G2P training is followed by ASR system development.

Recent studies have proved that grapheme-based acoustic modeling outperformed phoneme-based methods in E2E and hybrid ASR, even in English, which is considered a non-phonemic language. Nevertheless, graphemic ASR systems still have difficulty with words that have low frequency, such as entity names. In [74], the authors introduce a novel approach to train a statistical grapheme-to-grapheme (G2G) model on text-to-speech data that could rewrite sequences with an arbitrary character to be more consistent phonetically.

Standard Named Entity Correction (NEC) algorithms apply single-stage grapheme- or phoneme-level editing to search and replace named entities misrecognized by the ASR system. In [76], the researchers suggested a hierarchical multi-stage NEC algorithm. Since longer-named entities are not easily processed by a single-stage correction, they proposed a three-stage NEC. The first stage is word-level matching, followed by phonetic double-metaphone-based matching. Finally, a grapheme-level candidate is selected. The authors also suggested a new NE rejection technique that maintains the NEs recognized correctly to ensure that they are not changed. That suggestion was evaluated on call and music domains, and it minimized the WER by 14% for music and 63% for calls.

In [77], the researchers examined whether an ASR system could predict phoneme confusion in human listeners. DNN-ASRs and listening tests with six normal-hearing subjects provided phoneme-specific response rates. The accuracy of the correlation of phoneme recognition from ASR and human speech recognition (HSR) is the measure for model quality.

In [19], the authors investigated graph-based semi-supervised learning (SSL) in DNN-based acoustic models for speech recognition. They compared several graph-based learning (GBL) algorithms and proposed the pMP algorithm. Their proposed method is likely to be useful for adapting ASR systems to data-sparse test conditions, such as noisy environments or accented speech, and for developing ASR technology for low-resource languages.

To address mispronunciation detection and diagnosis (MDD) issues, such as the unrecognition of phone errors that are missing from extended recognition networks (ERNs), the authors of [49] proposed an acoustic-graphemic phonemic model (AGPM) that uses DNNs. They found that when compared with an approach using ERNs, the results showed that the proposed approach is simpler and more effective. It achieved an 11.1% phone error rate (PER), while the ERN approach achieved 16.8%, and the free-phone recognition for L2 English speech obtained a PER of 25.6%.

Similarly, in [36], the authors presented a detailed account of the anatomy of modern ASR, with examples of how it has been used in speech-language pathology research. They presented the architecture of a modern speech recognizer and the probabilistic framework underlying this technology. They took into consideration a pronunciation model since words may have multiple correct pronunciations, as they are influenced by factors such as the speaker's accent, speaking style, and neighboring words.

The authors of [54] presented vocal characteristics of whispered speech and discussed the problems for the recognition of whispered speech in different conditions. They provided a new pre-process method with cepstral features based on a deep denoising autoencoder (DDAE) to improve whisper recognition.

There were other experiments conducted by [75] to study why the recognition process was more difficult on children's speech than on adult speech. The answer suggested by the authors was that the errors in ASR came from predictable phonological effects correlated with language acquisition. They experimented with phone recognition on hand-labeled data for children. A comparison of the results for children and adults on TIMIT data showed higher phone substitution rates for children.

3) **Dialect**

ASRs work well with native English but poorly on non-native English data. Different articles have discussed the problem of dialect recognition in ASR to improve performance and accuracy. To improve the performance of a native English ASR on non-native English data, the authors in [26] proposed a DNN-based pseudo-likelihood correction (PLC) technique. They experimented with DNN-based PLC

mapping to improve the ASR performance for Indian English speakers with varied mother tongues. They proposed a novel objective function to learn the parameters. The experiments revealed that optimizing PLC mapping using standard MSE objective function was detrimental to non-native ASR performance. On the contrary, the proposed objective function showed significant improvement in WER compared to native model performance.

In [21], the authors provided an ASR system to address the problem of mixed dialects in input utterances in ASR using two main methods: the maximization of recognition likelihoods and the integration of recognition results. The proposed system statistically trains transformation rules between a common language and dialects and simulates a dialect corpus for ASR based on a machine translation technique. As a result, they demonstrated that the maximization of recognition likelihoods showed the best performance, while the integration of the recognition method showed slightly smaller accuracy. In [72], the researchers highlighted the demand for ASR systems that can account for dialectal variations behind acoustic modeling. Researchers evaluated two ASR systems—DeepSpeech and Google Cloud Speech—to examine how well ''are'' is recognized in African American English (AAE), namely habitual ''be''. They found that the habitual ''be'' was more likely to be an error than the unhabitual ''be'' and the words surrounding it.

C. CLASSIFICATION OF ARTICLES BASED ON DEVICE EFFICIENCY

1) **Microphone**

Many articles described the enhancement of ASR from the perspective of the speaker's problems and the microphone that captures the speaker's voice. Usually, the results of the ASR are good when the training and testing data are matched. However, the results are much worse when they differ in the number and arrangement of microphones. In [20], the authors suggested an unsupervised spatial clustering approach to microphone array processing. This approach, known as Model-based EM Source Separation and Localization (MESSL). While using MESSL's outputs for spatial covariance estimates of the noise improved ASR performance compared to a standard baseline.

The authors of [27] proposed a method that used multichannel nonnegative matrix factorization (MNMF) to estimate the spatial covariance matrix (SCM) of speech and noise in an unsupervised manner and generated an enhanced speech signal with beamforming. They found that the proposed methods were more robust in an unknown environment than the state-of-the-art beamforming method with DNN-based mask estimation. Moreover, in [18], the researchers proposed a target speaker extraction network (TEnet) to isolate the speech of a specific speaker. They relied on the auxiliary speaker characteristics provided by an anchor (a clean audio sample of the target speaker). They demonstrated that the proposed TEnet can outperform the single short anchor baseline by about 22.5% on WER and 15.5% on the SDR.

Sensitivity to recording conditions can be caused by a high level of background noise and a mediocre or poor-quality microphone installed on the sensors. The authors of [38] provided an adapting DNN-based acoustic model. They used an audio database recorded by wireless sensors to train an accurate model for the actual speech processing application. They found that joint training was not significantly better than training on the sensor-recorded noisy database subset, while the DNN adaptation turned out to perform significantly better. In [43], the authors provided an ASR system that employs various methods to address noisy acoustic scenes in public environments using an NMF with VB technique to separate the target speaker's voice from background sources and a time-varying minimum variance distortionless response (MVDR) to detect failure in the microphone channel. They use the AMFB that implicates prior information of speech to analyze its temporal dynamics. The proposed system achieved an absolute WER of 5.67% on the real evaluation test data. Also, in [47], the authors suggested extending an existing attention-based encoder-decoder framework to address the challenging noisy ASR tasks using a neural beamformer. In addition, they proposed an architecture of multiple channels in end-to-end ASR that allows the deduction of recognizing multichannel speech to enhance it based on an ASR objective. Their comprehensive framework works effectively with a noisy background. They found that the suggested framework results exceeded the end-to-end baseline with noisy input. Furthermore, successful learning was achieved by the beamformer of the noisy suppression. To improve the prediction accuracy of speakers, the authors of [50] proposed a hybrid method for automatic speaker identification using an ANN. The recognition is performed using Bayesian regularization and MLPs. The features are extracted using the mel frequency cepstral coefficient (MFCC). They found that the proposed method provides the best discrimination and has a high accuracy of 93.33%.

IV. RESULTS

A. WHAT RESEARCH TOPICS HAVE BEEN ADDRESSED IN RECENT ASR RESEARCH? (RQ1)

According to the aforementioned studies, most of the research published during 2015–2020 focused on addressing the major problems that degraded ASR system performance, such as various dialects, background noise, and speech interference. The research orientations applied DNN techniques to address

these problems by training the model with various background noises, large vocabularies, speech pronunciations, and dialects. Moreover, many research works have applied audio-visual techniques to make ASR systems more robust. Furthermore, some research was devoted to employing ASR in order to support the medical and education sectors.

B. WHAT ARE THE MAJOR CHALLENGES IN ASR? (RQ2)

Based on the literature, problems related to speech capture in general were detailed, such as domain and device effects. In addition, problems related to speech pre-processing, which were vocabulary, pronunciation problems, and English dialects, were described.

One of the biggest challenges for ASR is getting a suitable performance even in the presence of background noise. ASR system performance is highly degraded in the presence of noisy environments [52]. Although many algorithms have been proposed to perform ASR, most of them frequently fail in real-world environments with noise [58]. For instance, if the environment is very noisy, acoustical features will be severely degraded, especially in short-frame detection [40]. Moreover, noisy environments contribute to spectro-temporal absence in input signals, which leads to missed time-frequency correlations of the underlying speech signal [55]. Although many noise reduction methods have been developed, these methods cannot work unless the noises are known. However, noise signals can have many properties in real-world situations [33]. Due to this variety, creating a database that is responsive to external noise is a big challenge for DNN-based systems [38].

Another challenge caused by the environment is speech overlapping or simultaneous conversation. This happens when more than one person talks at the same time and is known as the ''cocktail party'' problem. Consequently, ASR systems face difficulties in detecting the target speech. This problem is still one of the most difficult in ASR [23].

The first step of the ASR process is to capture the speech by microphone. Consequently, ASR system performance is directly affected by the device hardware. Poor and mediocre microphones are one of the factors that reduce the quality of ASR system performance [38]. Furthermore, background noise can cause telephone channel distortions; suitable system performance in the presence of background noise requires high-quality microphone manufacturing [52]. In addition, to apply the approaches that use beamforming for speech segregation, the number of microphones has to be larger than the number of sound sources [56].

In addition, ASR systems have to contend with speech pre-processing challenges. One of the dilemmas in natural language processing (NLP) is the diversity of dialects. While the diversity of commonly spoken language is caused by the speakers themselves and their parents' business and residence histories, this variety makes detecting common language tremendously difficult for ASR systems [21]. Furthermore, a large vocabulary causes an increase in the computational cost of an ASR system [24], [37], which decreases the system response. Moreover, audio-visual fusion models are always facing difficulty in detecting continuous speech with a large vocabulary.

Another challenge occurs when dealing with spontaneous speech. Pronunciation problems directly affect ASR system performance, and this is particularly noticeable when the content of the speech is less predictable. Differences in pronunciation may be due to a health symptom, such as stuttering, or may be from children with limited speech ability [29].

C. WHAT ARE THE CURRENT RESEARCH GAPS IN ASR? (RQ3)

Researchers have faced several limitations in ASR. Limited datasets for ASR research was one of the major limitations noted during the application of the proposed methods [17], [34]. Moreover, single-channel speech recognition is considered another major limitation in ASR, especially with NN-based methods. In [27], the authors discussed only single-channel magnitude spectrograms in the evolution of the field as neglecting important data, such as interaural level differences (ILDs) and interaural phase differences (IPDs), which lead to unbalanced results. Similarly, in [23], the authors mentioned the problem of single-channel for overlapped speech recognition, which is derived from crossed speech when multiple people speak at the same time.

In other studies, different limitations affected their results. In [44], the authors adopted offline implementations to test the proposed enhancement methods. Therefore, any locative property changes of the sound during the observation of received signals were caused by the change in the acoustic channel, which resulted in decreasing the performance and affecting the outcomes. In [56], the authors proposed an embedded cognition method to improve ASR for robots, using microphone arrays to locate the speech sources. They then separated the speech signals from background noise. Accordingly, the proposed method requires prior knowledge of the number of sound sources, which is a limitation.

D. WHAT ARE FUTURE RESEARCH DIRECTIONS IN ASR? (RQ4)

Automatic speech recognition is one of the areas that has received a great deal of interest and attention from researchers. It is considered one of the longstanding problems in the field of artificial intelligence. ASR systems can be found in mobile devices, desktop computers, and also as virtual assistants in call centers. Still, this technology has many challenges and problems that researchers seek to investigate. Studies recently published in 2020 have presented some future research directions in ASR. Most of those studies have focused on the enhancement of the proposed methods and increasing the accuracy of speech recognition in different environments. For example, in [50], the authors proposed future work using other deep recurrent neural networks, such as a restricted Boltzmann machine, deep Boltzmann machine, and hybrid neuro-fuzzy genetic algorithm (GA). Similarly, as future work in [17], the authors seek to investigate other

TABLE 2. Future directions in ASR.

Year	Paper Ref. No	Proposals for future Articles
2020	[50]	The proposed future work is to extend the work by using other deep recurrent neural networks like Restricted Boltzmann Machine, Deep Boltzmann Machine, and Hybrid Neuro- Fuzzy-GA.
2020	[17]	In future work, the authors seek to investigate other auxiliary tasks for the proposed HTL setup, for instance generating only the noise as the auxiliary task (as opposed to the DAE) also evaluating HTL performance on other databases and feature combinations other than real and simulated data. Also, the authors want to investigate more deeply the impact that the convergence of the auxiliary task has on the main task, and then select and train auxiliary tasks accordingly.
2020	[29]	In the future, the authors are interested in automatically learning adaptation weights (possibly unique to each n-gram) to minimize WER without using a held-out set, considering limited availability of in-domain data. In the case of semantic response generation, it would be useful to learn to select context adult utterances relevant for each target child utterance. This would require both incorporating long-term context (using longer encoder timesteps or hierarchical networks) and including an attention mechanism in the seq2seq network. Considering the high baseline WER, further work will also continue to focus on developing a robust generic child ASR.
2020	[26]	The authors suggest a deeper investigation for robustness for the proposed method to the noise to increase detection accuracy further.
2020	[46]	The authors suggested in future work, they are exploring more parallelizable neural network architectures because it is difficult to parallelize the training and decoding processes of the current neural network-based architectures and adopt teacher-student learning approaches to reduce the latency and maintain the recognition accuracy at the same time.
2019	[30]	The authors' future work is to improve the energy efficiency of SNN-based ASR systems.
2019	[53]	The proposed future work is to explore higher-order embeddings that can represent more words. Also, exploring the effect of the smoothing parameter. The authors believes that the fallibility weight can be used in other rescoring experiments.
2019	[20]	The authors suggested the use of pitch-adaptive spectral estimation to improve recognition performance as a future work
2019	[22]	The authors suggested many different adaptations, tests, and experiments. They seek to try different methods to enhance the proposed system's performance and to evaluate the proposed system's behaviour under different conditions. Also, the authors suggest evaluating noisier conditions such as reverberation noise effects, colored noises, background music, and mixtures of environmental noises.
2019	[25]	The authors suggested as a future work an extenstion of evaluation metrics as well as the improvement of current MVML for dysarthric ASR. Weighting of evaluation metrics and fuzzy rules can be considered in ANFIS ASR evaluator.
2019	[37]	The authors suggested for future work to take advantage of the end-to-end model. This model can be improved in the following aspects: first, model delay: Attention-based models can effectively improve the recognition performance, but it is not monotonous and has long delay. Second, language knowledge learning: the end-to-end model needs to improve its learning of language knowledge while maintaining the end-to-end structure.
2019	[27]	The authors propose a to use the recently proposed semi-supervised speech enhancement methods based on NMF or MNMF with a DNN based before speech spectra.
2019	[18]	They proposed for future work to feed the extracted target speech as the anchor and train a feedback TEnet.
2019	[56]	The authors suggest as future work to use vision as bootstrapping mechanism for training the neural layers in an online fashion. Since this way will make the entire architecture can be trained with an unsupervised earning approach.
2018	[36]	The authors suggested for the future work of research on ASR systems for pathological speech should be focused in two complementing directions. The first should be focused on generating larger corpora of pathological speech, with all the major pathologies and different manifestation of each pathology. This will not only be useful in advancing the data-driven study of the field but will also enable the retraining and adaption of the current state-of-the-art speech recognisers. The second direction should be aimed at proposing different and alternative models to the current acoustic modelling, pronunciation modelling and language modelling, which will take into consideration the different variabilities of pathological speech.
2018	[33]	The authors proposed in future work to combined with several research fields such as acoustic processing, technique of sound source localization, design of home-care service robot, and multimedia analysis, to provide more user-friendly services in application of human-robot interactions.
2018	[41]	The authors suggested a similar fast adaptation approach for the DNN-based systems will be explored in the future.
2018	[23]	The authors suggest integrating modern technologies in the initialization stage into each unit. Besides, improving the sequencing level by applying different standards and end-to-end modeling (co-modeling) in the explicit integration of the language model.
2017	[35]	The article suggested future work to develop techniques for dealing with unbalanced label frequencies to achieve further improvement in detecting such infrequent error labels.
2017	[49]	The authors suggest a for future work, Generate a phonetic inventory of L2 like units from observed data in the future. Also, investigating the relationships between an L2 phone-like unit and the native phone unit(s).
2017	[40]	Future research can study integrating more advanced clean speech recognition into the system for further improved performance and the possible extension of the new method for robust speech recognition.
2017	[47]	The article focuses on improving adaptation techniques of an end-to-end framework to solve data sparseness problem by integrating linguistic resources.
2016	[58]	The authors suggest a for future work, apply this evaluation on real data.
2015	[21]	The article focuses on future work to introduce differences in acoustic features in dialects and apply them to speech data mining.
2015	[51]	The author suggests in the future work improve them neural net framework still further. First, they will improve the system's ASR performance by simulating the features with neurons, as they did when they simulated the filtering with neurons. Second, they like to determine whether it is possible to combine this framework with multi-band speech recognition efficiently. The features obtained from different bands could then be fed to separate classifiers, and after, the resulting posterior values could be combined and become the input for another neural net.

auxiliary tasks for the proposed hybrid-task learning (HTL) setup, for instance, generating only the noise as the auxiliary task. Also, the authors want to investigate more deeply the impact that the convergence of the auxiliary task has on the

TABLE 3. Applied datasets in the reviewed papers. Some papers used more than one dataset. Therefore, the same paper has been repeated in different datasets which increase the total number of papers.

Dataset	Number of papers	Percentage %
Real speech datasets	15	13%
NOT MENTIONED	12	11%
none	13	12%
CHiME & REVERB Challenge database	12	11%
TIMIT database	9	8%
videos dataset	8	7%
Aurora	7	6%
LibriSpeech	7	6%
SWB2000	6	5%
AMI dataset	5	4%
Language dataset	4	4%
Google dataset	2	2%
various corpus	3	3%
others	9	8%
total	112	

main task and then select and train auxiliary tasks accordingly. In the future work of [29], the authors are interested in automatically learning adaptation weights (possibly unique to each n-gram) to minimize the WER without using a held-out set (considering the limited availability of in-domain data). They will also continue to focus on developing robust generic child ASR. In [26], the authors suggest a deeper investigation for robustness with the proposed method for noise to increase detection accuracy further. In [21], the authors suggested adapting acoustic models with more speech data for each dialect. This was applied later in [72], where the researchers highlighted the demand for ASR systems that could consider dialectal variations behind acoustic modeling.

The future direction of some of the related articles in ASR is presented in Table 2, which describes the plans and directions that researchers seek to investigate in the field of ASR.

E. WHAT DATASETS ARE USED IN THE REVIEWED PAPERS? (RQ5)

As shown in Table 3, in 13% of the papers, the authors tried to create their own recorded dataset to use in their experiments and test their methodology. After that, the most popular datasets were the CHiME and REVERB Challenge database, the TIMIT database, Aurora, LibriSpeech, SWB2000, and AMI. These constituted 41% of the papers. Another 12% did not use any dataset; they explained their approach without validations, and 11% did not mention which dataset they used. Several datasets were used only once. They were grouped as ''other'', and they comprised 8% of the papers. Given the explanations in this paper, it is recommended that researchers in the future either create their own ASR data or use established data sets (CHiME & REVERB Challenge database, the TIMIT database, Aurora, LibriSpeech, SWB2000, and AMI).

V. LIMITATIONS

This research was conducted with a focus on selected ASR studies that examined English speech only. The search process was performed using a limited set of keywords that targeted an overview of ASR. This research focused on publications for a finite period from 2015 to 2020. However, it does provide an overview of the challenges and recent research trends in ASR.

VI. CONCLUSION

The most important way for humans to communicate with each other and acquire information is through speech. This paper provides a systematic literature review of automatic speech recognition with the most significant topics published in the last six years. A total of 82 conferences and articles studies were reviewed from five research databases: IEEE Xplore Digital Library, ACM Digital Library, Scopus, the Web of Science, and Science. First, a brief introduction to ASR was provided. The methodology of this research, including research questions, search strategy, and quality assessment process, was then described. After that, a review of the selected studies published from 2015 to 2020 in ASR based on their characteristics was organized and presented. The publication trends in speech recognition were then detailed. In addition, the major challenges and current research gaps in ASR were reviewed. Future directions for research in ASR were investigated. It is expected that this examination will help other researchers, as it provides a review of ASR studies published in recent years. Finally, statistics of mostly utilized datasets in reviewed papers were provided.

ACKNOWLEDGMENT

The authors thank the Deanship of Scientific Research and RSSU at King Saud University for their technical support.

REFERENCES

[1] D. Yu and L. Deng, *Automatic Speech Recognition*. London, U.K.: Springer, 2016.

[2] B.-H. Juang and L. R. Rabiner, ''Automatic speech recognition—A brief history of the technology development,'' Georgia Inst. Technol., Atlanta Rutgers Univ. Calif. St. Barbara, Tech. Rep., 2005, p. 67, vol. 1.

[3] D. Moher, A. Liberati, J. Tetzlaff, and D. G. Altman, ''Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement,'' *PLoS Med.*, vol. 6, pp. 1–6, Jul. 2009, doi: 10.1371/journal.pmed.1000097.

[4] Lucidchart. *Online Diagram Software & Visual Solution*. Accessed: Dec. 3, 2020. [Online]. Available: https://www.lucidchart.com
[5] *Nails Project*. Accessed: Dec. 3, 2020. [Online]. Available: http://nailsproject.net/
[6] B. Liao, Y. Ali, S. Nazir, L. He, and H. U. Khan, "Security analysis of IoT devices by using mobile computing: A systematic literature review," *IEEE Access*, vol. 8, pp. 120331–120350, 2020, doi: 10.1109/ACCESS.2020.3006358.
[7] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, Nov. 2012, doi: 10.1109/MSP.2012.2205597.
[8] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-28, no. 4, pp. 357–366, Aug. 1980, doi: 10.1109/TASSP.1980.1163420.
[9] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989, doi: 10.1109/5.18626.
[10] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Comput. Speech Lang.*, vol. 9, no. 2, pp. 171–185, Apr. 1995, doi: 10.1006/csla.1995.0010.
[11] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 1, pp. 30–42, Jan. 2012, doi: 10.1109/TASL.2011.2134090.
[12] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Commun.*, vol. 12, no. 3, pp. 247–251, Jul. 1993, doi: 10.1016/0167-6393(93)90095-3.
[13] J. P. Campbell, Jr., "Speaker recognition: A tutorial," *Proc. IEEE*, vol. 85, no. 9, pp. 1437–1462, Sep. 1997, doi: 10.1109/5.628714.
[14] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Commun.*, vol. 52, no. 1, pp. 12–40, 2010, doi: 10.1016/j.specom.2009.08.009.
[15] M. S. Gazzaniga, "Cerebral specialization and interhemispheric communication: Does the corpus callosum enable the human condition?" *Brain*, vol. 123, no. 7, pp. 1293–1326, Jul. 2000, doi: 10.1093/brain/123.7.1293.
[16] L. Parra and C. Spence, "Convolutive blind separation of non-stationary sources," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 3, pp. 320–327, May 2000, doi: 10.1109/89.841214.
[17] G. Pironkov, S. U. Wood, and S. Dupont, "Hybrid-task learning for robust automatic speech recognition," *Comput. Speech Lang.*, vol. 64, Nov. 2020, Art. no. 101103, doi: 10.1016/j.csl.2020.101103.
[18] W. Li, P. Zhang, and Y. Yan, "TEnet: Target speaker extraction network with accumulated speaker embedding for automatic speech recognition," *Electron. Lett.*, vol. 55, no. 14, pp. 816–819, Jul. 2019, doi: 10.1049/el.2019.1228.
[19] Y. Liu and K. Kirchhoff, "Graph-based semisupervised learning for acoustic modeling in automatic speech recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 11, pp. 1946–1956, Nov. 2016, doi: 10.1109/TASLP.2016.2593800.
[20] M. I. Mandel and J. Barker, "Multichannel spatial clustering for robust far-field automatic speech recognition in mismatched conditions," in *Proc. INTERSPEECH*, Sep. 2016, pp. 1991–1995.
[21] N. Hirayama, K. Yoshino, K. Itoyama, S. Mori, and H. G. Okuno, "Automatic speech recognition for mixed dialect utterances by mixing dialect language models," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 23, no. 2, pp. 373–382, Feb. 2015, doi: 10.1109/TASLP.2014.2387414.
[22] M. Tamazin, A. Gouda, and M. Khedr, "Enhanced automatic speech recognition system based on enhancing power-normalized cepstral coefficients," *Appl. Sci.*, vol. 9, no. 10, p. 2166, May 2019, doi: 10.3390/app9102166.
[23] Z. Chen, J. Droppo, J. Li, and W. Xiong, "Progressive joint modeling in unsupervised single-channel overlapped speech recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 1, pp. 184–196, Jan. 2018, doi: 10.1109/TASLP.2017.2765834.
[24] C. Cai, Y. Xu, D. Ke, and K. Su, "A fast learning method for multilayer perceptrons in automatic speech recognition systems," *J. Robot.*, vol. 2015, pp. 1–7, Feb. 2015, doi: 10.1155/2015/797083.
[25] A. Asemi, S. S. B. Salim, S. R. Shahamiri, A. Asemi, and N. Houshangi, "Adaptive neuro-fuzzy inference system for evaluating dysarthric automatic speech recognition (ASR) systems: A case study on MVML-based ASR," *Soft Comput.*, vol. 23, no. 10, pp. 3529–3544, May 2019, doi: 10.1007/s00500-018-3013-4.
[26] A. Rajpal, A. R. Mv, C. Yarra, R. Aggarwal, and P. K. Ghosh, "Pseudo likelihood correction technique for low resource accented ASR," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 7434–7438, doi: 10.1109/ICASSP40776.2020.9053647.
[27] K. Shimada, Y. Bando, M. Mimura, K. Itoyama, K. Yoshii, and T. Kawahara, "Unsupervised speech enhancement based on multichannel NMF-informed beamforming for noise-robust automatic speech recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 5, pp. 960–971, May 2019, doi: 10.1109/TASLP.2019.2907015.
[28] S. Shahnawazuddin, K. T. Deepak, G. Pradhan, and R. Sinha, "Enhancing noise and pitch robustness of children's ASR," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 5225–5229.
[29] M. Kumar, S. H. Kim, C. Lord, T. D. Lyon, and S. Narayanan, "Leveraging linguistic context in dyadic interactions to improve automatic speech recognition for children," *Comput. Speech Lang.*, vol. 63, Sep. 2020, Art. no. 101101, doi: 10.1016/j.csl.2020.101101.
[30] J. Wu, E. Yılmaz, M. Zhang, H. Li, and K. C. Tan, "Deep spiking neural networks for large vocabulary automatic speech recognition," *Frontiers Neurosci.*, vol. 14, p. 199, Mar. 2020, doi: 10.3389/fnins.2020.00199.
[31] R. Masumura, T. Asami, T. Oba, H. Masataki, and S. Sakauchi, "Viterbi approximation of latent words language models for automatic speech recognition," *J. Inf. Process.*, vol. 27, pp. 168–176, 2019, doi: 10.2197/ipsjjip.27.168.
[32] D. Palaz, M. Magimai-Doss, and R. Collobert, "End-to-end acoustic modeling using convolutional neural networks for HMM-based automatic speech recognition," *Speech Commun.*, vol. 108, pp. 15–32, Apr. 2019, doi: 10.1016/j.specom.2019.01.004.
[33] S.-C. Lee, J.-F. Wang, and M.-H. Chen, "Threshold-based noise detection and reduction for automatic speech recognition system in human-robot interactions," *Sensors*, vol. 18, no. 7, p. 2068, Jun. 2018, doi: 10.3390/s18072068.
[34] H. Wang, F. Gao, Y. Zhao, and L. Wu, "WaveNet with cross-attention for audiovisual speech recognition," *IEEE Access*, vol. 8, pp. 169160–169168, 2020, doi: 10.1109/ACCESS.2020.3024218.
[35] A. Ogawa and T. Hori, "Error detection and accuracy estimation in automatic speech recognition using deep bidirectional recurrent neural networks," *Speech Commun.*, vol. 89, pp. 70–83, May 2017, doi: 10.1016/j.specom.2017.02.009.
[36] J. Keshet, "Automatic speech recognition: A primer for speech-language pathology researchers," *Int. J. Speech-Lang. Pathol.*, vol. 20, no. 6, pp. 599–609, Oct. 2018, doi: 10.1080/17549507.2018.1510033.
[37] D. Wang, X. Wang, and S. Lv, "An overview of end-to-end automatic speech recognition," *Symmetry*, vol. 11, no. 8, p. 1018, Aug. 2019, doi: 10.3390/sym11081018.
[38] G. Gosztolya and T. Grósz, "Domain adaptation of deep neural networks for automatic speech recognition via wireless sensors," *J. Electr. Eng.*, vol. 67, no. 2, pp. 124–130, Apr. 2016, doi: 10.1515/jee-2016-0017.
[39] Y.-H. Tu, J. Du, T. Gao, and C.-H. Lee, "A multi-target SNR-progressive learning approach to regression based speech enhancement," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 1608–1619, 2020, doi: 10.1109/TASLP.2020.2996503.
[40] J. Ming and D. Crookes, "Speech enhancement based on full-sentence correlation and clean speech recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 3, pp. 531–543, Mar. 2017, doi: 10.1109/TASLP.2017.2651406.
[41] S. Shahnawazuddin and R. Sinha, "A fast adaptation approach for enhanced automatic recognition of children's speech with mismatched acoustic models," *Circuits, Syst., Signal Process.*, vol. 37, no. 3, pp. 1098–1115, Mar. 2018, doi: 10.1007/s00034-017-0586-6.
[42] A. H. Liu, T.-W. Sung, S.-P. Chuang, H.-Y. Lee, and L.-S. Lee, "Sequence-to-sequence automatic speech recognition with word embedding regularization and fused decoding," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 7879–7883, doi: 10.1109/ICASSP40776.2020.9053324.
[43] N. Moritz, K. Adiloğlu, J. Anemüller, S. Goetze, and B. Kollmeier, "Multi-channel speech enhancement and amplitude modulation analysis for noise robust automatic speech recognition," *Comput. Speech Lang.*, vol. 46, pp. 558–573, Nov. 2017, doi: 10.1016/j.csl.2016.11.004.

[44] A. H. Moore, P. P. Parada, and P. A. Naylor, "Speech enhancement for robust automatic speech recognition: Evaluation using a baseline system and instrumental measures," *Comput. Speech Lang.*, vol. 46, pp. 574–584, Nov. 2017, doi: 10.1016/j.csl.2016.11.003.

[45] H.-G. Kim, H. Lee, G. Kim, S.-H. Oh, and S.-Y. Lee, "Rescoring of N-best hypotheses using top-down selective attention for automatic speech recognition," *IEEE Signal Process. Lett.*, vol. 25, no. 2, pp. 199–203, Feb. 2018, doi: 10.1109/LSP.2017.2772828.

[46] H. Miao, G. Cheng, P. Zhang, and Y. Yan, "Online hybrid CTC/attention end-to-end automatic speech recognition architecture," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 1452–1465, 2020, doi: 10.1109/TASLP.2020.2987752.

[47] T. Ochiai, S. Watanabe, T. Hori, J. R. Hershey, and X. Xiao, "Unified architecture for multichannel end-to-end speech recognition with neural beamforming," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 8, pp. 1274–1288, Dec. 2017, doi: 10.1109/JSTSP.2017.2764276.

[48] T. Tan, Y. Qian, H. Hu, Y. Zhou, W. Ding, and K. Yu, "Adaptive very deep convolutional residual network for noise robust speech recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 8, pp. 1393–1405, Aug. 2018, doi: 10.1109/TASLP.2018.2825432.

[49] K. Li, X. Qian, and H. Meng, "Mispronunciation detection and diagnosis in L2 English speech using multidistribution deep neural networks," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 1, pp. 193–207, Jan. 2017, doi: 10.1109/TASLP.2016.2621675.

[50] K. J. Devi, N. H. Singh, and K. Thongam, "Automatic speaker recognition from speech signals using self organizing feature map and hybrid neural network," *Microprocessors Microsyst.*, vol. 79, Nov. 2020, Art. no. 103264, doi: 10.1016/j.micpro.2020.103264.

[51] G. Kovacs and L. Toth, "Joint optimization of spectro-temporal features and deep neural nets for robust automatic speech recognition," Tech. Rep., p. 19.

[52] P. Agrawal and S. Ganapathy, "Modulation filter learning using deep variational networks for robust speech recognition," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 2, pp. 244–253, May 2019, doi: 10.1109/JSTSP.2019.2913965.

[53] C. Liu, P. Zhang, T. Li, and Y. Yan, "Semantic features based N-best rescoring methods for automatic speech recognition," *Appl. Sci.*, vol. 9, no. 23, p. 5053, Nov. 2019, doi: 10.3390/app9235053.

[54] T. Grozdić and S. T. Jovičić, "Whispered speech recognition using deep denoising autoencoder and inverse filtering," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 12, pp. 2313–2322, Dec. 2017, doi: 10.1109/TASLP.2017.2738559.

[55] S. Ganapathy, "Multivariate autoregressive spectrogram modeling for noisy speech recognition," *IEEE Signal Process. Lett.*, vol. 24, no. 9, pp. 1373–1377, Sep. 2017, doi: 10.1109/LSP.2017.2724561.

[56] J. Davila-Chacon, J. Liu, and S. Wermter, "Enhanced robot speech recognition using biomimetic binaural sound source localization," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 1, pp. 138–150, Jan. 2019, doi: 10.1109/TNNLS.2018.2830119.

[57] Z. Q. Wang and D. Wang, "A joint training framework for robust automatic speech recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 4, pp. 796–806, Apr. 2016, doi: 10.1109/TASLP.2016.2528171.

[58] H.-Y. Lee, J.-W. Cho, M. Kim, and H.-M. Park, "DNN-based feature enhancement using DOA-constrained ICA for robust speech recognition," *IEEE Signal Process. Lett.*, vol. 23, no. 8, pp. 1091–1095, Aug. 2016.

[59] Y. Qian, T. Tan, and D. Yu, "Neural network based multi-factor aware joint training for robust speech recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 12, pp. 2231–2240, Dec. 2016, doi: 10.1109/TASLP.2016.2598308.

[60] S. Chen, H. Wang, F. Xu, and Y. Q. Jin, "Target classification using the deep convolutional networks for SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 8, pp. 4806–4817, Aug. 2016, doi: 10.1109/TGRS.2016.2551720.

[61] M. A. Rafique, B. G. Lee, and M. Jeon, "Hybrid neuromorphic system for automatic speech recognition," *Electron. Lett.*, vol. 52, no. 17, pp. 1428–1430, Aug. 2016, doi: 10.1049/el.2016.0975.

[62] R. Rehr and T. Gerkmann, "Cepstral noise subtraction for robust automatic speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 375–378.

[63] H. Meutzner, S. Araki, M. Fujimoto, and T. Nakatani, "A generative-discriminative hybrid approach to multi-channel noise reduction for robust automatic speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 5740–5744.

[64] L. Mosner, M. Wu, A. Raju, S. H. K. Parthasarathi, K. Kumatani, S. Sundaram, R. Maas, and B. Hoffmeister, "Improving noise robustness of automatic speech recognition via parallel data and teacher-student learning," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 6475–6479.

[65] X. Liu, R. Sadeghian, and S. A. Zahorian, "A modulation feature set for robust automatic speech recognition in additive noise and reverberation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 5230–5234.

[66] C.-T. Do and Y. Stylianou, "Weighting time-frequency representation of speech using auditory saliency for automatic speech recognition," in *Proc. INTERSPEECH*, Sep. 2018, pp. 1591–1595.

[67] M. Kŭhne, "Handling derivative filterbank features in bounded-marginalization-based missing data automatic speech recognition," in *Proc. 16th Annu. Conf. Int. Speech Commun. Assoc.*, Sep. 2015.

[68] D. S. Park, Y. Zhang, Y. Jia, W. Han, C.-C. Chiu, B. Li, Y. Wu, and Q. V. Le, "Improved noisy student training for automatic speech recognition," 2020, *arXiv:2005.09629*. [Online]. Available: http://arxiv.org/abs/2005.09629

[69] I. C. Yadav, S. Shahnawazuddin, D. Govind, and G. Pradhan, "Spectral smoothing by variationalmode decomposition and its effect on noise and pitch robustness of ASR system," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 5629–5633.

[70] T. H. Dat, J. Dennis, L. Y. Ren, and N. W. Z. Terence, "A comparative study of multi-channel processing methods for noisy automatic speech recognition in urban environments," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 6465–6469.

[71] T. Menne, I. Sklyar, R. Schlüter, and H. Ney, "Analysis of deep clustering as preprocessing for automatic speech recognition of sparsely overlapping speech," 2019, *arXiv:1905.03500*. [Online]. Available: http://arxiv.org/abs/1905.03500

[72] J. L. Martin and K. Tang, "Understanding racial disparities in automatic speech recognition: The case of habitual," in *Proc. INTERSPEECH*, 2020, pp. 626–630.

[73] R. Rasipuram, M. Razavi, and M. Magimai-Doss, "Integrated pronunciation learning for automatic speech recognition using probabilistic lexical modeling," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 5176–5180, doi: 10.1109/ICASSP.2015.7178958.

[74] D. Le, T. Koehler, C. Fuegen, and M. L. Seltzer, "G2G: TTS-driven pronunciation learning for graphemic hybrid ASR," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 6869–6873, doi: 10.1109/ICASSP40776.2020.9054257.

[75] E. Fringi, J. F. Lehman, and M. Russell, "Evidence of phonological processes in automatic recognition of children's speech," in *Proc. 16th Annu. Conf. Int. Speech Commun. Assoc.*, 2015, pp. 1–4.

[76] A. Garg, A. Gupta, D. Gowda, S. Singh, and C. Kim, "Hierarchical multi-stage word-to-grapheme named entity corrector for automatic speech recognition," in *Proc. INTERSPEECH*, Oct. 2020, pp. 1793–1797.

[77] M. Exter and B. T. Meyer, "DNN-based automatic speech recognition as a model for human phoneme perception," in *Proc. INTERSPEECH*, Sep. 2016, pp. 615–619.

[78] E. Egorova and L. Burget, "Out-of-vocabulary word recovery using FST-based subword unit clustering in a hybrid ASR system," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 5919–5923.

[79] V. T. Pham, H. Xu, Y. Khassanov, Z. Zeng, E. S. Chng, C. Ni, B. Ma, and H. Li, "Independent language modeling architecture for end-to-end ASR," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 7059–7063.

[80] J. Drexler and J. Glass, "Subword regularization and beam search decoding for end-to-end automatic speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 6266–6270.

[81] J. Drexler and J. Glass, "Learning a subword inventory jointly with end-to-end automatic speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 6439–6443, doi: 10.1109/ICASSP40776.2020.9053736.

[82] M. Song, Y. Zhao, S. Wang, and M. Han, "Learning recurrent neural network language models with context-sensitive label smoothing for automatic speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 6159–6163, doi: 10.1109/ICASSP40776.2020.9053589.

[83] E. Lakomkin, J. Heymann, I. Sklyar, and S. Wiesler, "Subword regularization: An analysis of scalability and generalization for end-to-end automatic speech recognition," 2020, *arXiv:2008.04034*. [Online]. Available: http://arxiv.org/abs/2008.04034
[84] H. Liao, G. Pundak, O. Siohan, M. K. Carroll, N. Coccaro, Q.-M. Jiang, T. N. Sainath, A. Senior, F. Beaufays, and M. Bacchiani, "Large vocabulary automatic speech recognition for children," in *Proc. INTERSPEECH*, Sep. 2015, pp. 1–5.
[85] B. Lecouteux and D. Schwab, "Ant colony algorithm applied to automatic speech recognition graph decoding," in *Proc. INTERSPEECH*, Sep. 2015, pp. 1–6.
[86] D. Baby and H. Van Hamme, "Investigating modulation spectrogram features for deep neural network-based automatic speech recognition," in *Proc. 16th Annu. Conf. Int. Speech Commun. Assoc.*, Sep. 2015, pp. 2479–2483.
[87] C.-T. Do and Y. Stylianou, "Improved automatic speech recognition using subband temporal envelope features and time-delay neural network denoising autoencoder," in *Proc. INTERSPEECH*, Aug. 2017, pp. 3832–3836.
[88] X. Cui and M. Picheny, "Acoustic model optimization based on evolutionary stochastic gradient descent with anchors for automatic speech recognition," 2019, *arXiv:1907.04882*. [Online]. Available: http://arxiv.org/abs/1907.04882
[89] T. Lohrenz and T. Fingscheidt, "BLSTM-driven stream fusion for automatic speech recognition: Novel methods and a multi-size window fusion example," in *Proc. INTERSPEECH*, Oct. 2020, pp. 26–30.
[90] J. Fahringer, T. Schrank, J. Stahl, P. Mowlaee, and F. Pernkopf, "Phase-aware signal processing for automatic speech recognition," in *Proc. INTERSPEECH*, Sep. 2016, pp. 3374–3378.
[91] H. Lnaguma, M. Mimura, K. Inoue, K. Yoshii, and T. Kawahara, "An end-to-end approach to joint social signal detection and automatic speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 6214–6218.
[92] W. Zhang, X. Cui, U. Finkler, G. Saon, A. Kayi, A. Buyuktosunoglu, B. Kingsbury, D. Kung, and M. Picheny, "A highly efficient distributed deep learning system for automatic speech recognition," 2019, *arXiv:1907.05701*. [Online]. Available: http://arxiv.org/abs/1907.05701
[93] W. Zhang, X. Cui, U. Finkler, B. Kingsbury, G. Saon, D. Kung, and M. Picheny, "Distributed deep learning strategies for automatic speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 5706–5710.
[94] H. Inaguma, Y. Gaur, L. Lu, J. Li, and Y. Gong, "Minimum latency training strategies for streaming sequence-to-sequence ASR," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 6064–6068.
[95] N. Moritz, T. Hori, and J. Le, "Streaming automatic speech recognition with the transformer model," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 6074–6078.
[96] K. C. Sim, A. Narayanan, A. Misra, A. Tripathi, G. Pundak, T. Sainath, P. Haghani, B. Li, and M. Bacchiani, "Domain adaptation using factorized hidden layer for robust automatic speech recognition," in *Proc. INTERSPEECH*, Sep. 2018, pp. 892–896.
[97] A. Mani, S. Palaskar, N. V. Meripo, S. Konam, and F. Metze, "ASR error correction and domain adaptation using machine translation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 6344–6348.
[98] S. Kundu, G. Mantena, Y. Qian, T. Tan, M. Delcroix, and K. C. Sim, "Joint acoustic factor learning for robust deep neural network based automatic speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 5025–5029.
[99] Y. Moriya and G. J. F. Jones, "Multimodal speaker adaptation of acoustic model and language model for ASR using speaker face embedding," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 8643–8647, doi: 10.1109/ICASSP.2019.8683724.
[100] J. Malek, J. Zdansky, and P. Cerva, "Robust automatic recognition of speech with background music," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 5210–5214.

SADEEN ALHARBI received the master's degree in software engineering from Queensland University of Technology (QUT), Australia, in 2014, and the Ph.D. degree in computer science from The University of Sheffield, U.K., in 2019. She is currently an Assistant Professor with the Department of Software Engineering, College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia. Her research interests include automatic speech recognition and synthesis, natural language processing, machine learning, and deep learning to the healthcare domain.

MUNA ALRAZGAN received the Ph.D. degree in information technology from George Mason University, VA, USA. She is currently an Associate Professor in software engineering with the College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia. Her research interests include data mining, machine learning, artificial intelligence, educational data mining, and assistive technologies.

ALANOUD ALRASHED received the bachelor's degree in computer science from King Saud University, Riyadh, Saudi Arabia, where she is currently pursuing the master's degree in software engineering.

TURKIAYH ALNOMASI received the bachelor's degree in software engineering from the University of Hail, Saudi Arabia, in 2015. She is currently pursuing the master's degree with the Department of Software Engineering, King Saud University, Riyadh, Saudi Arabia.

RAGHAD ALMOJEL received the bachelor's degree in software engineering from King Saud University, Riyadh, Saudi Arabia, in 2018, where she is currently pursuing the master's degree with the Department of Software Engineering. Her research interests include software engineering and machine learning.

RIMAH ALHARBI received the bachelor's degree in software engineering from the University of Hail. She is currently pursuing the master's degree with the Department of Software Engineering, King Saud University, Riyadh, Saudi Arabia.

SAJA ALHARBI received the bachelor's degree in electrical and computer engineering from King Abdulaziz University. She is currently pursuing the master's degree in software engineering with King Saud University, Riyadh, Saudi Arabia. She is interested in the AI research field.

SAHAR ALTURKI received the bachelor's degree in software engineering from King Saud University, Riyadh, Saudi Arabia, in 2019, where she is currently pursuing the master's degree.

FATIMAH ALSHEHRI received the bachelor's degree in information technology from King Saud University, Riyadh, Saudi Arabia, where she is currently pursuing the master's degree in software engineering. She has a four years work experience in software analysis.

MAHA ALMOJIL received the bachelor's degree (Hons.) in software engineering from King Saud University, Riyadh, Saudi Arabia, in 2020, where she is currently pursuing the M.Sc. degree in software engineering.