AI-Powered Intelligent Speech Processing: Evolution, Applications and Future Directions

Ziqing Zhang
University International College, Macau University of Science and Technology, Macau 999078, China

***Abstract*—This paper provides an overview of the historical evolution of speech recognition, synthesis, and processing technologies, highlighting the transition from statistical models to deep learning-based models. Firstly, the paper reviews the early development of speech processing, tracing it from the rule-based and statistical models of the 1960s to the deep learning models, such as deep neural networks (DNN), convolutional neural networks (CNN), and recurrent neural networks (RNN), which have dramatically reduced error rates in speech recognition and synthesis. It emphasizes how these advancements have led to more natural and accurate speech outputs. Then, the paper examines three key learning paradigms used in speech recognition: supervised, self-supervised, and semi-supervised learning. Supervised learning relies on large amounts of labeled data, while self-supervised and semi-supervised learning leverage unlabeled data to improve generalization and reduce reliance on manually labeled datasets. These paradigms have significantly advanced the field of speech recognition. Furthermore, the paper explores the wide-ranging applications of AI-driven speech processing, including smart homes, intelligent transportation, healthcare, and finance. By integrating AI with technologies like the Internet of Things (IoT) and big data, speech technology is being applied in voice assistants, autonomous vehicles, and speech-controlled devices. The paper also addresses the current challenges facing intelligent speech processing, such as performance issues in noisy environments, the scarcity of data for low-resource languages, and concerns related to data privacy, algorithmic bias, and legal responsibility. Overcoming these challenges will be crucial for the continued progress of the field. Finally, the paper looks to the future, predicting further improvements in speech processing technology through advancements in hardware and algorithms. It anticipates increased focus on personalized services, real-time speech processing, and multilingual support, along with growing integration with other technologies such as augmented reality. Despite the technical and ethical challenges, AI-driven speech processing is expected to continue its transformative impact on society and industry.**

Keywords—Intelligent speech recognition; AI speech synthesis; speech processing; AI technology

I. Introduction

With the iterative upgrading of artificial intelligence technology, the application fields of AI technology are also expanding and developing, and at the same time, intelligent voice interaction, personalized speech generation, and other technologies are getting more and more attention in this process [1]. Among them, speech processing, as an important technology and means of personalized speech generation, involves speech signal processing, artificial intelligence, pattern recognition, phonetics, and other disciplines, and is the hot spot and difficult point in the field of speech processing research, and in the rapid development of AI technology, this field has also become a key research direction. Therefore, along with the development of the Internet and big data technology, voice processing technology (TTS), speech recognition technology (ASR), and other intelligent voice technologies are gradually maturing, the production, dissemination, and storage mechanism of the sound produces changes, and the application scenarios of AI voice are increasing, and voice products such as automobile navigation, video dubbing, intelligent speakers, cell phone assistants and so on are emerging in an endless number, and have already been realized to correspond to specific application Scenarios. Nowadays, there are two main parts of intelligent voice processing according to the function of the application scenes in social life, namely, AI synthesized voice is mainly composed of "AI voice narration" scenes without human-computer interaction, and "AI voice assistant" scenes with human-computer interaction. The two parts of the scene [2].

A. Significance of the Study

Language, as a unique tool for human communication, permeates all aspects of life, contains the rich and important emotional value, and plays an important role in the construction of civilized society, while emotion, as a characteristic unique to human beings, is usually difficult to be accurately conveyed in the process of speech processing. The development of AI technology makes it possible for the part of the speech processing process concerning human-specific expressive characteristics such as emotional value to be learned and applied by the machine, which is expected to have a profound impact on speech processing. Learn and apply, which is expected to have a profound impact on speech processing [3].

Speech, as a direct carrier of language, is the most natural way of communication among human beings and a key part of information transfer in daily life. With the rapid development of artificial intelligence research, speech recognition technology, which is carried by computers, cell phones, tablets, etc., is rapidly advancing. In the field of Human-Computer Interaction (HCI), recognizing individual phonemes or utterances of specific speakers can no longer meet the needs of learning and development of AI technology, so recognizing the hidden emotional state in speech has become a new trend in speech processing research under the rapid development of current AI technology. Therefore, the study of how to use this technology to assist speech processing to achieve intelligence in the current rapid development of AI technology is an urgent problem to be solved nowadays.

B. Conceptual Identification

1) Speech processing: Broadly speaking, people change the voice in the voice of the speaker's personality characteristics of voice processing technology collectively referred to as voice processing, which can be obtained, broad voice processing can be divided into two categories: non-specific person voice processing and specific person voice processing. Non-specific speech processing refers to the technical processing that makes the converted speech no longer sound like the original speaker's voice. In practical research and application, speech processing usually refers to the technology that changes the voice personality characteristics, such as spectrum and rhythm, of one speaker, i.e., the source speaker, to make it have the personality characteristics of another specific speaker, i.e., the target speaker, while keeping the semantic information unchanged. Generally speaking, the technical difficulty of person-specific speech processing is higher than that of non-person-specific speech processing.

The current research on linguistic AI mainly focuses on language understanding and language output, which are called "Natural Language Understanding" (NLU) and "Natural Language Generation" (NLG) respectively. "Natural Language Understanding (NLU) and Natural Language Generation (NLG). Natural Language Understanding allows computers to understand human natural language (including its intrinsic meaning) through a variety of analysis and processing, while Natural Language Generation focuses on how to allow computers to automatically generate natural language forms or systems that humans can understand. The relationship between the research contents of natural language processing and the corresponding human language ability is shown in the Table I.

TABLE I. CORRESPONDENCE TABLE BETWEEN NATURAL LANGUAGE AND HUMAN LANGUAGE ABILITY

No.	Research in Natural Language Processing	Linguistic competence of counterparts
1	Information retrieval	Language Understanding
2	Information filtering	
3	Information extraction	
4	Machine translation	Linguistic Generation
5	Automatic digest	
6	Document classification	Language Understanding
7	Text/Data mining	
8	Public opinion analysis	
9	Text editing and automatic proofreading	
10	Automatic scoring of essays	
11	Question and answer system	Linguistic Generation
12	Metaphorical computing	Language Understanding
13	Optical character recognition, OCR	
14	Speech recognition	
15	Text-to-language conversion	Linguistic Generation
16	Speaker identification/authentication/verification	Language Understanding

Using the degree of development of AI technology as a classification criterion, AI can be categorized into weak AI, strong AI, and super AI. Weak AI, also known as narrow AI. Its main features are: (1) does not have self-consciousness, its action depends on the instructions given by humans; (2) and human beings are significantly different from each other, there is still a huge gap in appearance and other aspects, it is very easy to identify; (3) cannot learn and innovate on its own, it is essentially a tool that relies on human beings to be upgraded. The above characteristics determine that weak artificial intelligence should not be qualified as a legal subject. Strong artificial intelligence, also known as general artificial intelligence, is complete artificial intelligence. Strong AI can think like a human being and is not just a tool for humans. Because it possesses the consciousness of rights, claims of rights, and the ability to realize autonomy, as well as the ability to assume responsibility relatively independently, and has a certain substantive connection with human beings, it meets the conditions for having limited legal subject qualification.

Super Artificial Intelligence, defined by Nick Bostrom as an intelligence that outperforms the human brain in almost every domain, these domains include scientific innovation, general intelligence, and social skills. This definition is open-ended. The possessor of superintelligence can be a digital computer, a computer network integration, or an artificial neural organization without regard to whether it has subjective consciousness or experience. It seems that super-artificial intelligence should be given a higher level of legal status, but this idea is still debatable due to the human fear of the risks posed by the unknown and the adherence to ethics.

Therefore, another commonly accepted classification standard at present is to categorize AI based on different application modes of AI, i.e., to classify AI into generative AI and discriminative AI. According to China's Interim Measures for the Administration of Generative Artificial Intelligence Services, which came into effect on August 15, 2023, generative AI technology refers to models and related technologies that can generate content such as text, pictures, audio, video, and so on. It has become a milestone in the history of AI because it transcends the scope of traditional AI, is capable of generating natural language understandable to humans, performs tasks that should only be accomplished by human intellectual guidance, and is deeply involved in human daily life. Discriminative AI, also known as discriminative AI or decision-making AI, aims to train machine learning models to classify or predict based on input features. This approach is well suited for tasks such as regression and sequence labeling in areas such as image recognition, speech recognition, and natural language processing. One of the main advantages of discriminative AI is its ability to make efficient and accurate predictions, but it cannot typically generate new samples and an understanding of the generative mechanisms behind the data.

C. *Purpose of the Synthesis*

This paper will focus on analyzing the latest progress of intelligent speech processing technology in the wave of artificial intelligence technology and the consequent profound changes that will be brought to this research field. With the rapid development of AI technology, intelligent speech processing technology has moved from the laboratory to the road of wide application has gradually penetrated daily life, and has profoundly affected the current life and the mode of operation of various industries. The purpose of this paper is to comprehensively show the technological development of intelligent speech processing technology in terms of speech recognition accuracy, speech processing naturalness, and natural language processing intelligence by systematically combing domestic and foreign research results and application cases, and exploring how the development of AI technology will have a transformative impact on the field of intelligent speech processing, and at the same time, to look forward to the future development of intelligent speech processing. In addition, this paper will also discuss the impact of possible pitfalls in data security, privacy, and law brought about by the ongoing research in the field of intelligent speech processing, to obtain thoughts and inspiration on the impact of the development of AI technology. Overall, the research objective of this paper is to comprehensively sort out the development of intelligent speech processing technology under the development of AI technology, assess its transformative impact on the society and economy, and provide valuable reference and guidance for future research and application.

D. *Structure of the Paper*

This paper aims to study the development of an intelligent speech-processing technique based on AI technology and to discuss the technical and legal issues that are currently faced as well as the outlook for the future. The following is the Section structure of this thesis:

Section II mainly elaborates on the current status of intelligent speech processing, which firstly gives an overview of the relevant research progress at home and abroad, and then summarizes, combs, and summarizes the current research results at home and abroad, mainly including the overview of the achievements in the fields of speech recognition, speech processing, and speech learning paradigm.

Section III analyzes the key problems faced by the current intelligent speech processing technology based on AI technology, mainly from the main tasks and problems of speech processing technology, firstly, the tasks of the current speech processing technology are clarified, and then the completion of the speech processing tasks are analyzed in-depth by the current problems of speech processing technology, to the problems faced by the intelligent speech processing technology based on AI technology A detailed exposition is carried out.

Section IV describes the development issues in the field of intelligent speech processing based on AI technology, and this Section further explores the public data security and ethical and legal issues brought about by the technological development on top of the technological development issues in Section 3, and this Section analyzes the current intelligent speech processing field outside of the technology that may be encountered in the field of intelligent speech processing, mainly from the perspective of three aspects, namely, data privacy and security, data algorithmic discrimination and bias, and accountability and responsibility problems.

Section V provides an outlook on the future of intelligent speech processing based on the development of current technologies and application scenarios, mainly analyzing and looking forward to both technology and application levels.

Section VI concludes with a summary, which summarizes the significance of the development of intelligent speech processing technology, as well as further summarizes the technological development discussed in the content of the above Sections and the problems that may be faced, and discusses the future research direction, indicating the outlook for the future development of intelligent speech processing technology based on AI technology.

II. The Current Status of Research in the Field of Intelligent Speech Processing

A. *Overview of Domestic and International Research*

The work related to speech processing research can be traced back to the 1960s and 1970s, and there have been more than 50 years of research history, but it is only in the last decade or so that it has received extensive attention from both academia and industry. In recent years, the advancement of technologies such as speech signal processing and artificial intelligence deep learning, in addition, due to the explosive growth of information, the field of big data has also been developed significantly, and a considerable amount of learning and analyzing text has been obtained, therefore, the improvement of big data acquisition ability and large-scale computational performance has likewise given a strong impetus to the research and development of speech processing technology. Among them, the most important role is still the rise of speech processing methods based on artificial neural networks (Artificial neural networks, ANN), which makes the quality of speech processing further improved by the mode of deep learning. Institutions in China that have conducted speech processing research earlier include the Chinese Academy of Sciences, the University of Science and Technology of China, the National University of Defense Technology, Microsoft Research Asia, IBM China Research Institute, and so on. In recent years, many universities such as Southeast University, Nanjing University of Posts and Telecommunications, South China University of Technology, Soochow University, Harbin Institute of Technology, Northwestern Polytechnical University, Army Engineering University, and many other companies such as Tencent, KDDI, and Baidu have also begun the research on this technology, and have successively achieved several research results. In 2016, scientists in the field of speech processing from China, Japan, and the United Kingdom organized the VCC2016. VCC2016 was organized by scientists from China, Japan, Britain, and other countries, which provided a data platform and performance scale for speech processing research. 2018 VCC2018 was also held as scheduled, and based on the world's cutting-edge artificial intelligence technology, the speech processing methods were once again pushed forward, and accordingly, the quality of speech processing was once again significantly improved. Enhancement. According to the key breakthroughs in speech recognition technology, Table II summarizes the development of speech processing and recognition technology in different time periods.

TABLE II. SPEECH PROCESSING TECHNOLOGY DEVELOPMENT

Period	Key Technology Breakthroughs	Application Areas	Main features
1950s	Syllable Recognition System	Specific Speech Conversion	Relies on hard coding with limited recognition capabilities
1960s-1970s	Rule-based and statistical modeling	Continuous speech stream processing	Recognition improved but was still limited by noise, etc.
1980s-1990s	Deep learning techniques, neural network models	Processing complex speech streams	Automatic learning capability to handle non-standard pronunciation
Early 21st century to the present	Mobile Internet and smartphone penetration	Multi-language, multi-accent recognition	Dealing with noise, accents, and speed of speech variations, a wide range of applications

B. Synthesis of Main Research Results

1) Development of speech recognition technology: For a long time, the dominant approach for speech recognition has been the Gaussian Mixture Model-Hidden Markov Model (GMM-HMM), which is based on context-dependent generative models of GMM and HMM. Neural networks were also a popular method used for speech recognition, however not as effective as GMM-HMM. Deep learning started to make an impact in the field of speech recognition after a close collaboration between academia and industry in 2010. The collaboration began with a phoneme recognition task, and the results demonstrated the capabilities of the DNN architecture and subsequent convolutional and recurrent network architectures. Their work also demonstrated the importance of moving away from the widely used MFCC features to lower-level raw speech spectrum features. Their collaboration has also yielded good results on large-vocabulary recognition tasks. The main reason for the success of DNN on large-vocabulary speech recognition tasks is that similar to the speech units in GMM-HMM, DNN uses a large-scale output layer structure. The reason for using this structure is that speech researchers expect to take advantage of context-dependent phoneme modeling techniques, which are very effective in GMM-HMM, and that this structure allows for as little change as possible to the highly efficient decoder software architecture developed for GMM-HMM. The work on large-vocabulary speech recognition by the DNN has also demonstrated that, if a large amount of labeled data can be exploited, a pre-training process similar to that for a deep confidence network is not necessary. the training process is unnecessary. Deep learning has been successful in both industry and academia in the field of speech recognition, and this success has been due to three main factors: (1) deep learning significantly reduces the speech recognition error rate compared to the best previous GMM-HMM systems; (2) due to the use of phonemes as the output of DNNs, deploying DNN-based speech recognizers requires only a small portion of the decoder to be changed; and (3) DNNs Powerful modeling capability reduces the complexity of speech recognition systems. After the success of the DNN-HMM system for speech recognition, researchers have proposed many new architectures and nonlinear units for improving speech recognition accuracy. Yu et al. proposed a tensor version of DNN by replacing one or more layers in a traditional DNN by using dual projection layers and tensor layers. The dual projection layer projects each input vector into two nonlinear subspaces. In the tensor layer, the two subspace projections interact with each other and jointly project the next layer of the entire depth architecture. The researchers also propose a method to map the tensor layer to a traditional sigmoid layer, so the tensor layer can be trained in a similar way to the sigmoid layer. The idea of time-domain convolution originated from a time-delay neural network (TDNN), which was used as a shallow network in early speech recognition. Recently when researchers used deep convolutional neural networks for phoneme recognition tasks, they found that weight sharing in the frequency domain is more compared to weight sharing in the time domain [4]. A research report also states that convolutional neural networks help in large-vocabulary continuous speech recognition tasks and that multilayer convolutional neural networks using a large number of convolutional kernels or feature maps give greater performance gains [5]. Sainath et al. explored a large number of variations of deep convolutional neural networks and found that when combined with several new methods, deep convolutional neural networks achieved the best results on several large-vocabulary speech recognition task results [6]. The most notable deep structures for speech recognition tasks are recurrent neural networks and their deep versions [7]. Although RNNs were first successful in phoneme recognition, however, due to the complexity of training RNNs, it was difficult to scale RNNs to larger speech recognition tasks [8]. Since then the learning algorithms for RNNs have been improved and better results have been achieved using RNNs on several tasks, especially using bidirectional LSTM RNNs [9]. In addition to innovations in deep learning models for speech recognition, a large amount of work has investigated how to develop and implement better nonlinear units. The sigmoid function and the tanh function are the most commonly used nonlinear functions in DNNs, however both have limitations. For example, when the neuron nodes are close to saturation, the error function has a small value of the gradient concerning the parameters, at which point the network is slow to train. To overcome the shortcomings of the sigmoid and tanh functions, Jaitly and Hinton first used ReLU in speech recognition.

Another effective unit for speech recognition is the maxout unit, which is used to construct deep maxout networks. Deep max-out networks generate hidden layer activation values by performing a max operation or max-out operation on a fixed number of weighted inputs. This operation is the same as the maximum pooling operation in convolutional neural networks. These maximum values are the outputs of the previous layer. Thereafter, Zhang et al. generalized the max-out unit into two new types. Both of these types have been shown computationally and experimentally to work better than the ReLU units described in the previous section.

2) Research and development of speech processing technology: The goal of speech processing is to generate speech directly from text as well as additional information. To overcome the shortcomings of statistical parameter-based speech-processing algorithms, researchers have applied deep learning to the field of speech processing. Earlier speech processing techniques were mainly based on rules and rule sets, where computers converted text into speech according to predefined rules. This method requires a lot of human intervention and the results of synthesized speech are not satisfactory. Subsequently, statistical parametric speech processing appeared in the mid-1990s and became the main approach in the field of speech processing [10]. In this approach, the relationship between the text and the corresponding acoustic realizations is represented by a set of stochastic generative acoustic models, which presents three modules that are very important for speech processing: the language model, the acoustic model, and the vocoder. Among them, the task of the language model is to extract the input text into linguistic features utilizing natural language processing techniques, which have the linguistic information needed by the back-end acoustic model. The acoustic model is responsible for converting linguistic features into acoustic features, and then a separate vocoder completes the conversion of acoustic features to the original speech waveform. Hidden Markov models based on decision tree clustering, contextual correlation, and output states satisfying a Gaussian distribution are the most popular generative acoustic models. However, due to the use of shallow HMMs, this acoustic model is not adequately modeled. Some recent studies use deep learning to overcome this deficiency.

Ling et al. used Restricted Boltzmann Machine (RBM) and Deep Confidence Network as generative models to replace the traditional Gaussian model and achieved significant improvement in both subjective and objective evaluation of synthesized speech [11]. Kang et al. used the Deep Confidence Network as a generative model to represent the joint distribution of linguistic features and acoustic features and used the Deep Confidence Network to replace the decision tree and Gaussian model joint distribution of features and used deep confidence networks instead of decision trees and Gaussian models. This approach is similar to using deep confidence networks to generate digital images [12].

With the development of AI deep learning technology, speech processing technology has made a leaping breakthrough, and the iconic technology representative is the Tacotron model proposed by Google in 2017, which is an end-to-end speech processing model based on the self-attention mechanism, where the input side consists of text, which is generated by a text encoder to generate contextual text vectors with robustness, and the decoder side uses an attention-based mechanism based autoregressive decoder at the decoder side, which outputs N frames of Mel Spectrum speech features at a time. The so-called autoregressive decoding means that the N frames output in the first step become inputs in the second step, and this is repeated to generate the complete Mel Spectrogram. The Mel Spectrogram is then passed through the final high-speed convolution module of the Tacotron to generate a linear spectrogram, which is then passed through the Griffin-Lim algorithm to obtain the synthesized speech waveform. Recently, DeepMind announced its latest research breakthrough WaveNet in speech processing [13]. Subsequently, the Tacotron Generation 2 model proposed by Google Inc. in 2018 replaces the high-speed convolution module of the generation algorithm with a 3-layer long and short-term memory module and replaces the vocoder part from the GriffinLim algorithm with the deep-learning WaveNet algorithm, which is noteworthy for its synthesized quality, which is already able to reach the level of falsetto on the subjective evaluation. WaveNet utilizes real human voice clips and corresponding linguistic and phonetic features to train a convolutional neural network to be able to discriminate between linguistic and phonetic audio patterns. When using the WaveNet system, new textual information is input and the WaveNet system regenerates the entire original audio waveform to describe this new textual information.WaveNet is capable of simulating any human voice and generates speech that sounds more natural than the best speech-processing systems available today.WaveNet reduces the discrepancy between simulated-generated speech and the human voice by 50% or more [14].

The Tacotron model can generate high-quality speech processing, however, due to its autoregressive generation structure, the training speed and inference speed are not ideal [15]. Thus, in 2018, Transformer TTS proposed by the University of Electronic Science and Technology of China and Microsoft Research Asia, among others, utilized the self-attention mechanism Transformer to replace the original traditional content-based attention mechanism to accomplish non-autoregressive generation [16]. Subsequently, the FastSpeech1 and Fastspeech2 architectures proposed by Zhejiang University and Microsoft Research Asia in 2019 and 2020, respectively, succeeded in end-to-end non-autoregressive generation, which not only improved the inference speed, but also possessed a duration predictor, a pitch predictor, and an energy predictor that could accomplish the fine-grained control of the output speech duration, pitch, energy, etc. and at the same time improves the errors of lost words and repeated words that Tacotron2 can make [17].

The VITS model is a 2021 highly expressive speech processing model that combines variational inference, normalized streaming, and adversarial training and is currently comprised of most of the speech processors used on major self-promotion platforms. VITS is the first truly end-to-end speech processing model that does not require an additional vocoder to reconstruct the waveform and directly maps characters or phonemes to waveforms. This synthesis improves the versatility of speech processing by cascading the vocoder and acoustic model of speech processing through hidden variables instead of the spectrum of the previous model [18].

3) Learning paradigms for speech recognition: In speech recognition, model training methods mainly include supervised learning, self-supervised learning, and semi-supervised learning. Traditional supervised learning employs video speech as well as corresponding labeled data and uses the loss function of label prediction for model optimization. Self-supervised learning frameworks, on the other hand, are usually divided into two phases: pre-training and fine-tuning. In the pre-training phase, unlabeled audio and video data are first utilized to train the model based on agent tasks such as mask loss prediction or audio-video cross-modal comparison learning to extract generic deep representations from the audio and video data. Then in the fine-tuning phase, supervised learning and optimization are performed for speech recognition tasks using labeled audio and video data. Semi-supervised learning, on the other hand, utilizes a portion of the labeled data to first train the speech recognition model, followed by pseudo-labeling of unlabeled audio and video speech using the trained model, and subsequently training the model jointly with all audio and video data.

The first is supervised learning. General end-to-end speech recognition directly uses text labels to train the model, and depending on the sequence-to-sequence model framework used, CTC loss, RNNT loss, or cross-entropy loss is used as the objective function for model optimization. To further constrain the model during the model training process, the researchers proposed the use of an auxiliary objective as a regularization term for training. This approach can also be regarded as a kind of multi-task learning. Sterpu et al. proposed the use of lip shape-related articulatory action units (action units) as auxiliary training targets for visual representations based on the AVAlign model to enhance audio-video modal alignment [19]. Ma et al. proposed a method based on audio representations as auxiliary training targets in a lip recognition task [20]. In this, the audio representation target is obtained by extracting the middle layer representation of the encoder using a trained speech recognition model.

The second is self-supervised learning. With the development of the self-supervised learning paradigm in recent years, speech recognition methods based on audio and video self-supervised learning have also received more and more attention. Early audio-video self-supervised learning focuses on the learning of local audio-video features, and most of the methods utilize the natural alignment properties between audio-video modalities to supervise each other's information, representative methods include AVTS, XDC, and local-global audio-video comparative learning methods, etc. LiRA proposes a cross-modal self-supervised learning method based on the pre-trained audio self-supervised learning model PHASE extracts audio representations as targets to train a visual representation extraction module [21]. Audio self-supervised learning methods have developed rapidly in the recent past, and researchers have been inspired by them and extended them to audio-video self-supervised learning, which includes AV-HuBERT based on the HuBERT model extended to audio-video speech, and AV-data2vec based on the data2vec model extended to audio-video speech. Based on the model of AV-HuBERT, Zhu et al. further introduced additional text modalities to realize the joint learning of audio, video, and text modal representations [22]. u-HuBERT further introduced additional unimodal speech in AV-HuBERT and could theoretically introduce more modalities of data [23]. The AV2vec model and the RAVEn models, on the other hand, use a teacher model based on an exponential sliding average approach to multimodal self-distillation learning for student models in training [24]. Audio-video self-supervised speech recognition has achieved better results in lip recognition as well as speech recognition tasks, and many current studies have used pre-trained self-supervised models to initialize the parameters of speech recognition models [25].

Then comes semi-supervised learning. Semi-supervised speech recognition methods are mostly inspired by semi-supervised learning methods for audio speech recognition. The Auto-AVSR model generates pseudo-labels for unlabeled speech recognition datasets using pre-trained speech recognition models and then trains the speech recognition models using a combination of labeled and pseudo-labeled data [26]. The AV-CPL method, on the other hand, proposes a semi-supervised based on continuous pseudo-label update learning method, which continuously employs the updated model for continuous optimization of pseudo-labels on unlabeled audio and video [27]. Self-supervised learning methods are limited by less labeled data, so the results are generally worse than semi-supervised learning. However, the self-supervised learning method can be combined with semi-supervised learning, i.e., using self-supervised learning at the beginning to obtain an initial speech recognition model, then using the initial model to generate text pseudo-labels for unlabeled data, and then jointly optimizing the recognition model with labeled data, which generally achieves better speech recognition results. The summary of speech recognition learning paradigms is shown in Table III.

TABLE III. SUMMARY OF SPEECH LEARNING PARADIGMS

Learning paradigm	Descriptive	Typical models or methods	Vintage	Drawbacks
Supervised learning	Using labeled speech data to train the model, the model learns the mapping relationship from speech signal to text	LSTM+CTC, DNN-HMM, RNN-HMM, etc.	1. can directly optimize the speech recognition performance 2. is effective in the case of sufficient annotation data	1. Labelling data is costly and time-consuming 2. Difficulty in covering all possible speech variations and noise environments
Self-supervised learning	Using unlabeled speech data to learn useful speech representations by designing assistive tasks	Siamese Networks, Generative Adversarial Networks (GANs), Self-Encoders	1. does not require large amounts of labeled data 2. can learn more generalized speech features 3. helps to improve the generalization ability of the model	1. complex model structures and training strategies may be required 2. The design of ancillary tasks has a large impact on performance
Semi-supervised learning	Combining labeled and unlabeled speech data for learning to improve the accuracy and generalization of models	Clustering-based methods, bias correction-based methods, etc.	1. exploits the richness of unlabeled data 2. enables improved performance with limited labeled data	1. need to balance the use of labeled and unlabeled data 2. may require complex model fusion strategies

III. KEY ISSUES IN INTELLIGENT SPEECH PROCESSING BASED ON AI TECHNOLOGY

A. Main Tasks of Speech Processing Technology

1) Speech matching: Speech matching refers to the automatic retrieval of all speech segments that have the same content as the query speech segment from a given speech database, so the extracted speech features are crucial for the speech-matching task. Since speech matching automatically retrieves all speech segments with the same content as the query speech segment from a given speech database, it can be regarded as a class of content-based speech retrieval applications, which have been widely used in music retrieval, song recommendation, and speech intelligence analysis. At the same time, speech matching is a class of unsupervised learning tasks, and the techniques applicable to speech matching can be applied to other unsupervised learning tasks in the field of machine learning, thus the study of speech matching algorithms has important academic value. According to the above scenario, it can be seen that the speech matching task is a typical class of speech processing tasks because the key to speech matching is the extraction of speech features.

2) Multimodal speech recognition: Human-computer interaction interfaces for intelligent machines, such as smartphones, home robots, and self-driving cars, are becoming more and more common in daily life. Speech recognition that is robust to noise is the key to achieving effective human-machine interaction [28]. Multimodal speech recognition is considered one of the effective solutions for robust speech recognition. In human-computer interaction systems, the machine can not only receive the operator's speech signal, but also observe the operator's behavioral information, such as body movement and changes in mouth shape, and this behavioral information can help the machine recognize the operator's speech signal, and the study of multimodal speech recognition has an important application value in human-computer interaction systems [29]. In addition, multimodal speech recognition is a supervised learning task that involves the fusion of information from multiple sources, which has important academic research value. Multimodal speech recognition is also a typical speech-processing task [30].

B. Technical Issues in Intelligent Speech Recognition and Processing

In recent years, with the continuous increase of current training data, deep learning models, and learning paradigms, the effect of speech recognition has been greatly improved. However, there are still many challenges in speech recognition, and future research needs to pay more attention to the following points. First, current speech recognition still performs poorly in more open and complex environments, such as home scenarios and multiple people talking freely [31]. Currently, most studies are still overly focused on the LRS2 and LRS3 datasets, but lip recognition as well as speech recognition on these two datasets are already saturated with word error rates, e.g., Auto-AVSR has a speech recognition word error rate of 0.9% on LRS3 [32]. Under other more challenging scenario test sets, such as MISP2021 and Google's YTDEV, speech recognition remains poor. In the MISP2021AVSR competition, the word error rate of the winning team was still as high as 24.6% [33]. Therefore, future research in speech recognition needs to focus on the performance of datasets with more complex conditions and avoid overfitting to single or multiple similar scene datasets [34]. Second, the performance of speech recognition in small languages with sparse data still needs to be improved. Similar to speech recognition, there are still many challenges in small-language speech recognition [35]. Compared to mono-speech, audio and video data in small languages are more difficult to capture, and many small languages even face the challenge of not being able to capture video data [36]. How to deal with this situation of missing visual data and very low resources to improve speech recognition in small languages is still a problem that needs to be further researched and solved. Third, speech recognition is relatively computationally intensive and incurs high inference computation costs in practical application deployments. Under many high signal-to-noise ratio conditions, the improvement of speech recognition relative to audio speech recognition is not significant and brings little benefit [37]. Therefore, how to further reduce the computational burden of speech recognition, achieve lightweight computation of the vision module, and at the same time realize effective dynamic computation for environmental noise conditions is one of the important issues to be solved for the practicalization of speech recognition [38]. Fourth, the robustness of speech recognition in the case of missing video modalities and faces being occluded needs to be further improved [39]. Recently, researchers have begun to pay more and more attention to such problems of speech recognition in practical applications, and the proposed methods have improved the robustness of speech recognition to some extent [40]. However, these challenges have not been fully addressed, especially when there are cases of video damage, occlusion, etc. outside the training set distribution, speech recognition may still be worse than unimodal speech recognition. Ideal speech recognition maximizes the use of incremental information in the video while avoiding the effects of distracting information that appears in the video. How best to achieve this goal remains the open question.

IV. Development Issues in the Field of Intelligent Speech Processing based on AI Technology

AI technology in the development process is accompanied by lingering controversies, such as ChatGPT since its inception has been full of controversial copyright issues, as well as the existence of a large number of the use of AI voice cloning technology on the network to clone the voices of some of the singers and secondary creation of content, which belong to the "gray area" [41 These are all "gray areas" [41]. These contents not only involve copyright issues but also bring some difficult ethical problems. The current rapid development of AI technology has led to three major problems as summarized in Table IV.

TABLE IV. AI-Based Intelligent Speech Processing Problem

Type of problem	Explicit description	Frequency/number of reports
Data Privacy and Security	Voice data is being collected on a large scale, increasing the risk of privacy breaches	In recent years, an average of dozens of related privacy breaches have occurred each year, affecting millions of users.
Algorithmic bias and discrimination	Bias in training data leads to unfair recognition results	Several studies have reported that at least half of intelligent speech-processing systems suffer from varying degrees of algorithmic bias.
Responsibility and accountability	Difficulty in defining the responsible party when there is a problem with the speech recognition system	Every year there are dozens of legal disputes caused by errors in voice recognition systems, most of which involve unclear delineation of responsibility.

1) Data privacy and security issues: Generative AI models require a large amount of data for training at the initial stage, and data privacy issues are involved in whether the source of the data is traceable, authentic, and licensed by the information source. In the process of training generative AI models, the protection of user data is very important and must be done when conducting experiments. Several protection measures can be taken to address this issue. First, the data can be anonymized, the user data can be desensitized, and sensitive information, such as name, address, phone number, etc., can be deleted or replaced while ensuring accuracy, and encryption technology can be used to encrypt the user data to ensure that only authorized personnel can access it [42]. Second, user data can be stored in an isolated environment, data access rights can be set, and secure communication protocols can be used for transmission. Based on this, training data selection is performed by choosing appropriate training data in the dataset and avoiding the use of training data containing sensitive information, such as publicly available datasets or using processed data [43]. Again, the security of the model should be ensured, and after the model training is completed, the model needs to be evaluated for security to ensure that the model will not leak user data or have adverse effects. Finally, relevant laws and regulations, such as the Data Security Law of the People's Republic of China and the Personal Information Protection Law of the People's Republic of China, need to be strictly observed to ensure that user data are legally used and protected. In conclusion, a series of measures are needed to protect the security and privacy of user data when training and using generative AI.

2) Algorithmic bias and discrimination issues: If there is bias in the training data, then the generative AI may replicate that bias, leading to discriminatory decisions. American news commentator and social critic Lippmann famously proposed the "agenda-setting theory", which argues that "the news media influences 'the images we have in our minds about the world'", and this is also the case in generative AI [44]. For example, in 2016 Microsoft launched Tay, a conversational bot that was "portrayed" as racist only 16 hours after it went live and engaged in a conversation with a user, after which Microsoft urgently took the account offline [45]. The fact that it took only 16 hours from posting to taking the account offline suggests that a series of measures need to be taken to deal with this kind of problem. The first is the cleaning and filtering of data containing bias and discrimination during the data collection and processing phase [46]. This step involves removing or correcting inaccurate, outdated, or discriminatory data and, during the data collection process, ensuring that the data is diverse and inclusive. This means collecting data from different backgrounds, genders, ethnicities, and other characteristics to reduce the possibility of algorithmic bias and discrimination [47]. During algorithm design and training, ensure that the model is fair and unbiased. This includes the use of fair evaluation criteria, unbiased training data and algorithms, and a transparent decision-making process [48]. Algorithms also need to be audited periodically after the model has begun to perform computations to ensure that they meet the requirements of impartiality and fairness [49]. This includes examining the design of the algorithm, the training data, and actual runs, as well as assessing its impact on different populations. Bias detection and correction techniques can also be used to identify and correct bias in algorithms, where available [50]. This includes the use of techniques such as unsupervised learning, semi-supervised learning, or reinforcement learning to improve the performance of the model and reduce bias. In terms of personnel requirements, there is a need to educate and train algorithm designers and users to increase their awareness of bias and discrimination issues and equip them with the skills to recognize and address these issues.

3) Responsibility and accountability issues: When AI systems lead to undesirable consequences, how to pursue responsibility is a complex issue. Similar to how to select the author of the textual content produced by generative AI, there are many controversies around this issue, some scholars believe that the responsibility should be borne by the AI system, and some scholars believe that it should be borne by the program developer, after all, the entire algorithmic mechanism is from the developer's hand [51]. It can be seen that pursuing responsibility is a complex issue that needs to be considered from several aspects. First, it is necessary to determine who the owners and users of the AI system are, i.e., the responsible subjects. This involves multiple parties such as hardware and software vendors, data providers, application developers, and users. Second, it is necessary to understand information about the operation mechanism, algorithm design, training data, and actual application scenarios of the AI system. This helps to analyze the causes of adverse consequences and attribute responsibility. Again, it is important to assess the extent of the impact of the adverse consequences of the AI system on individuals, organizations, and society for a specific time, including economic losses, privacy leakage, security issues, etc. Finally, it is important to determine the type of responsibility, including technical, legal, ethical, and other responsibilities, based on the adverse consequences and impacts of AI systems [52]. The discussion of this issue, should not stop at determining the responsible person, but should also look for solutions, based on the type of responsibility and the actual situation, including compensation for losses, improvement of system design, correction of algorithms, and strengthening of supervision. At the same time, an accountability mechanism should be established to hold AI systems accountable for adverse consequences. There is a need to establish internal accountability systems, implement transparent and traceable management programs, and strengthen regulation and legal sanctions. In addition, there is a need to strengthen cooperation and communication, mainly among government departments, enterprises, social organizations, and other institutions, to jointly solve problems that are likely to arise from AI systems.

V. Future Outlook for Intelligent Speech Processing

A. *Forecast of Technology Development Trends*

The future development of speech-processing technology will mainly rely on the continuous development of deep learning and neural network technology. With the continuous upgrading of hardware equipment and the continuous optimization of algorithms, the quality and naturalness of speech processing technology will continue to improve. The current speech processing technology can already realize realistic speech processing, but there are still some shortcomings, such as the rhythm and rhyme of the voice is not natural. Future speech processing technology will pay more attention to the improvement of these aspects to realize more realistic speech processing. Future speech-processing technology will pay more attention to personalized services and experiences. With the continuous development of artificial intelligence technology, future speech processing technology will be able to personalize speech processing according to the user's needs and preferences, and can synthesize any person's voice based on small or zero samples, providing speech processing services that are closer to the user's needs. Future speech processing technology will pay more attention to real-time speech processing. Real-time speech processing can provide users with a more natural and smooth voice interaction experience, providing a broader application space for the development of voice interaction technology.

B. *Forecast of Application Development*

With the development of artificial intelligence technology and other related technologies, the field of intelligent speech processing will increasingly focus on the personalized needs and experiences of users. Future speech processing technologies will be able to better provide personalized services and experiences, such as personalized speech processing based on user needs and interests, thereby increasing user satisfaction and loyalty.

In addition, multilingual support and cross-cultural communication will be facilitated by deep learning from a large amount of text and data. Speech-processing technology will be expected to support more languages and dialects so that it can better meet the needs of users from different countries and regions and realize cross-cultural communication [53]. Further, speech processing technology will also realize automatic translation and conversion between multiple languages, providing users with more convenient and diversified services. At the same time, the future speech processing technology will be able to better combine with other integrated media technologies, such as images, videos, text, etc., to realize richer and more vivid forms of media expression [54]. For example, in TV news, speech processing technology can be combined with video and text to realize more vivid and intuitive news broadcasting. Based on the above applications, it is foreseeable that future speech processing technology is expected to be combined with augmented reality technology to realize a more intelligent and convenient user experience. Based on the trend of deep integration of speech processing technology and artificial intelligence technology, the key applications of technology in the future may be concentrated in the following areas, as shown in Table V:

TABLE V. KEY APPLICATION AREAS FOR INTELLIGENT SPEECH PROCESSING

Application Areas	Specific application	Key technologies
smart home	Smart speaker, smart appliance control	Speech Recognition, Speech Synthesis, Internet of Things
smart city	Intelligent Transportation, Intelligent Security	Deep learning, computer vision, big data analytics
health care	Disease diagnosis, drug development	Natural language processing, machine learning, big data analytics
financial	Intelligent Risk Control, Intelligent Investment	Machine learning, natural language processing, big data analytics
teach	Intelligent Tutoring, Language Learning	Speech Recognition, Natural Language Processing, Intelligent Recommender Systems

VI. CONCLUSION

Intelligent speech processing technology has made significant progress under the impetus of AI technology, which not only realizes a qualitative leap in speech recognition and speech processing accuracy but also promotes the innovation of human-computer interaction. Through advanced technologies such as deep learning, intelligent speech processing systems can more accurately understand the emotional and semantic information in human speech and generate natural and smooth speech. With the continuous maturation of technology and the expansion of application scenarios, intelligent speech processing technology will play an important role in more fields, and at the same time, its wide application has also brought about profound impacts on social structure, occupational structure, and privacy laws. Looking ahead, intelligent speech-processing technology will continue to innovate and develop, bringing more convenience and possibilities to human society.

REFERENCES

[1] Bostrom, N. (1998). How long before superintelligence?. *International Journal of Futures Studies*.

[2] Hinton, G., & Shallice, T. (1991). Lesioning an attractor network: investigations of acquired dyslexia. *Psychological Review*, 98(1), 74-95.

[3] Yu, D., Deng, L., & Seide, F. (2013). The deep tensor neural network with applications to large vocabulary speech recognition. *IEEE Transactions on Audio Speech and Language Processing*, 21(2), 388-396.

[4] Abdel-Hamid, O., Mohamed, A. R., Jiang, H., & Penn, G. (2012). Applying Convolutional Neural Networks concepts to hybrid NN-HMM model for speech recognition. *IEEE International Conference on Acoustics*. IEEE.

[5] Sainath, T. N., Kingsbury, B., Mohamed, A. R., Dahl, G. E., Saon, G., & Soltau, H. (2013). Improvements to deep convolutional neural networks for LVCSR. *Automatic Speech Recognition & Understanding*. IEEE.

[6] Pardede, H. F., Adhi, P., & Zilvan, V. R. A. K. D. (2023). Deep convolutional neural networks-based features for indonesian large vocabulary speech recognition. *IAES International Journal of Artificial Intelligence*, 12(2), 610-617.

[7] Graves, A., Mohamed, A. R., & Hinton, G. (2013). Speech recognition with deep recurrent neural networks. Acoustics, Speech, and Signal Processing, 1988. ICASSP-88. *1988 International Conference on* (Vol.38). IEEE.

[8] Wei, A., Zhang, H., & Zhao, E. (2025). DereflectFormer: Vision Transformers for Single Image Reflection Removal. *International Conference on Pattern Recognition*. Springer, Cham.

[9] Graves, A., Jaitly, N., & Mohamed, A. R. (2013). Hybrid speech recognition with deep bidirectional lstm. *IEEE*.

[10] Tokuda, K., Nankaku, Y., Toda, T., Zen, H., Yamagishi, J., & Oura, K.(2013). Speech synthesis based on hidden markov models. *Proceedings of the IEEE*, 101(5), 1234-1252.

[11] Ling, Z. H., Deng, L., Yu, D. (2013). Modeling spectral envelopes using restricted Boltzmann machines and deep belief networks for statistical parametric speech synthesis. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(10): 2129-2139.

[12] Oord, A. V. D., Dieleman, S., Zen, H., Simonyan, K., & Kavukcuoglu, K. (2016). Wavenet: a generative model for raw audio. DOI:10.48550/arXiv.1609.03499.

[13] Miao, Y., Metze, F., Rawat, S. (2013). Deep max out networks for low-resource speech recognition. *IEEE Workshop on Automatic Speech Recognition and Understanding*: 398-403.

[14] Zhang, X., Trmal, J., Povey, D., & Khudanpur, S. (2014). Improving deep neural network acoustic models using generalized maxout networks. *2014 IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP), 215–219. doi.org/10.1109/ICASSP.2014.6853589

[15] Humphrey, E. J., & Bello, J. P. (2012). Rethinking automatic chord recognition with convolutional neural networks. *2012 11th International Conference on Machine Learning and Applications* (ICMLA), 2, 357–362. doi.org/10.1109/ICMLA.2012.232

[16] Humphrey, E. J., Bello, J. P., & LeCun, Y. (2012). Moving beyond feature design: Deep architectures and automatic feature learning in music informatics. Proceedings of the 13th International Society for Music Information Retrieval Conference, 403–408.

[17] Lee, H., Grosse, R., Ranganath, R., & Ng, A. Y. (2011). Unsupervised learning of hierarchical representations with convolutional deep belief networks. *Communications of the ACM*, 54(10), 95–103. doi.org/10.1145/2001269.2001295

[18] Sterpu, G., Saam, C., & Harte, N. (2020). How to teach DNNs to pay attention to the visual modality in speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28, 1052–1064. doi.org/10.1109/TASLP.2020.2983042

[19] Ma, P. C., Petridis, S., & Pantic, M. (2022). Visual speech recognition for multiple languages in the wild. *Nature Machine Intelligence*, 4, 930–939. doi.org/10.1038/s42256-022-00523-1

[20] Ma, P. C., Wang, Y. J., Petridis, S., & Pantic, M. (2022). Training strategies for improved lip-reading. In 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 8472–8476). *IEEE*. doi.org/10.1109/ICASSP43922.2022.9747620

[21] Korbar, B., Tran, D., & Torresani, L. (2018). Cooperative learning of audio and video models from self-supervised synchronization. arXiv preprint arXiv:1807.00230.

[22] Alwassel, H., Mahajan, D., Torresani, L., & Ghanem, B. (2019). Self-supervised learning by cross-modal audio-video clustering. arXiv preprint arXiv:1911.12667.

[23] Ma, S., Zeng, Z. Y., McDuff, D. J., & Song, Y. (2021). Contrastive learning of global and local video representations. *In Advances in Neural Information Processing Systems* (Vol. 34, pp. 7025–7040).

[24] Ma, P. C., Mira, R., Petridis, S., & Pantic, M. (2021). LiRA: Learning visual speech representations from audio through self-supervision. *In Interspeech 2021* (pp. 3011–3015).

[25] Hsu, W. N., Bolte, B., Tsai, Y. H. H. (2021). HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29, 3451–3460. doi.org/10.1109/TASLP.2021.3122291

[26] Petridis, S., & Pantic, M. (2016). Deep complementary bottleneck features for visual speech recognition. In 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 2304–2308). *IEEE*. doi.org/10.1109/ICASSP.2016.7472081

[27] Chung, J. S., Senior, A., Vinyals, O., & Zisserman, A. (2017). Lip reading sentences in the wild. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 3444–3453). *IEEE*. doi.org/10.1109/CVPR.2017.367

[28] Afouras, T., Chung, J. S., Senior, A., Vinyals, O., & Zisserman, A. (2022). Deep audiovisual speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12), 8717–8727. doi.org/10.1109/TPAMI.2021.3058582

[29] Serdyuk, D., Braga, O., & Siohan, O. (2022). Transformer-based video front-ends for audio-visual speech recognition for single and multi-person video. *In Interspeech* 2022 (pp. 2833–2837).

[30] Makino, T., Liao, H., Assael, Y., et al. (2019). Recurrent neural network transducer for audio-visual speech recognition. In 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU) (pp. 905–912). *IEEE*. doi.org/10.1109/ASRU46091.2019.9003751

[31] Ma, P. C., Petridis, S., & Pantic, M. (2021). End-to-end audiovisual speech recognition with conformers. In 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 7613–7617). *IEEE*. doi.org/10.1109/ICASSP39728.2021.9413580

[32] Petridis, S., Stafylakis, T., Ma, P. C., & Pantic, M. (2018). Audio-visual speech recognition with a hybrid CTC/attention architecture. In 2018 IEEE Spoken Language Technology Workshop (SLT) (pp. 513–520). *IEEE*. doi.org/10.1109/SLT.2018.8639589

[33] Assael, Y. M., Shillingford, B., Whiteson, S., & de Freitas, N. (2016). LipNet: Sentence-level lipreading. arXiv preprint arXiv:1611.01599.

[34] Wand, M., Koutník, J., & Schmidhuber, J. (2016). Lipreading with long short-term memory. 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 6115–6119. https://doi.org/10.1109/ICASSP.2016.7472852

[35] Stafylakis, T., & Tzimiropoulos, G. (2017). Combining residual networks with LSTMs for lipreading. *Interspeech 2017*, 3652–3656. doi.org/10.21437/Interspeech.2017-85

[36] Gao, R. H., & Grauman, K. (2021). VisualVoice: Audio-visual speech separation with cross-modal consistency. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (CVPR), 15490–15500. doi.org/10.1109/CVPR46437.2021.01525

[37] Ma, P. C., Martinez, B., Petridis, S., & Pantic, M. (2021). Towards practical lipreading with distilled and efficient models. 2021 *IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP), 7608–7612. doi.org/10.1109/ICASSP39728.2021.9413581

[38] Prajwal, K. R., Afouras, T., & Zisserman, A. (2022). Sub-word level lip reading with visual attention. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (CVPR), 5152–5162. doi.org/10.1109/CVPR52688.2022.00508

[39] Boulanger-Lewandowski, N., Bengio, Y., & Vincent, P. (2013). Audio chord recognition with recurrent neural networks. *Proceedings of the 14th International Society for Music Information Retrieval Conference* (ISMIR 2013), 335–340.

[40] Van den Oord, A., Dieleman, S., & Schrauwen, B. (2013). Deep content-based music recommendation. *Advances in Neural Information Processing Systems* (NIPS 2013), 2643–2651.

[41] Lin, M., Chen, Q., & Yan, S. (2013). Network in network. arXiv preprint arXiv:1312.4400.

[42] Zeiler, M. D., & Fergus, R. (2013). Stochastic pooling for regularization of deep convolutional neural networks. arXiv preprint arXiv:1301.3557.

[43] He, K., Zhang, X., Ren, S., & Sun, J. (2015). Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(9), 1904–1916. doi.org/10.1109/TPAMI.2015.2389824

[44] Ouyang, W., Luo, P., Zeng, X., Qiu, S., & Wang, X. (2014). DeepID-Net: Multi-stage and deformable deep convolutional neural networks for object detection. arXiv preprint arXiv:1409.3505.

[45] Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. *In European Conference on Computer Vision* (ECCV 2014) (pp. 818–833). Springer. doi.org/10.1007/978-3-319-10590-1_53

[46] Chen, H., Du, J., Hu, Y., & Dai, L. (2021). Correlating subword articulation with lip shapes for embedding aware audiovisual speech enhancement. *Neural Networks*, 143, 171–182. doi.org/10.1016/j.neunet.2021.07.021

[47] Hu, D., Li, X. L., & Lu, X. Q. (2016). Temporal multimodal learning in audiovisual speech recognition. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3574–3582). Las Vegas, NV, USA: IEEE. doi.org/10.1109/CVPR.2016.389

[48] Ninomiya, H., Kitaoka, N., Tamura, S., Iribe, Y., & Takeda, K. (2015). Integration of deep bottleneck features for audio-visual speech recognition. *In Proceedings of Interspeech 2015* (pp. 563–567). Dresden, Germany: ISCA.

[49] Mroueh, Y., Marcheret, E., & Goel, V. (2015). Deep multimodal learning for audio-visual speech recognition. *In 2015 IEEE International Conference on Acoustics, Speech and Signal Processin*g (pp. 2130–2134). South Brisbane, Queensland, Australia: IEEE. doi.org/10.1109/ICASSP.2015.7178370

[50] Takashima, Y., Aihara, R., Takiguchi, T., & Ariki, Y. (2016). Audio-visual speech recognition using bimodal-trained bottleneck features for a person with severe hearing loss. *In Proceedings of Interspeech 2016* (pp. 277–281). San Francisco, CA, USA: ISCA.

[51] Yu, J. W., Zhang, S. X., Wu, J., & Xie, L. (2020). Audio-visual recognition of overlapped speech for the LRS2 dataset. *In ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 6984–6988). Barcelona, Spain: IEEE. doi.org/10.1109/ICASSP40776.2020.9054551

[52] Li, J. H., Li, C. D., Wu, Y. F., & Xie, L. (2023). Robust audio-visual ASR with unified cross-modal attention. *In ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech, and Signal Processing* (pp. 1–5). Rhodes Island, Greece: IEEE.

[53] Wang, J. D., Qian, X. Y., & Li, H. Z. (2022). Predict-and-update network: Audio-visual speech recognition inspired by human speech perception. arXiv preprint arXiv:2209.01768.