ISSN: 2581-3404 (Online) **IF: 5.445 (SJIF)**

IJIRTM, Volume-7, Issue-5, October-2023.

Exploring the Challenges and Advancements in Automatic Speech Recognition: A Comprehensive Survey

Disha Joshi[1], Chetan Agrawal[2], Pawan Meena[3]

Dept. of CSE, Radharaman Institute of Technology & Science, Bhopal, India[1, 2, 3]

24jdisha@gmail.com[1], chetan.agrawal12@gmail.com[2], pawan1914@gmail.com[3]

Abstract: *Humans have evolved to communicate with one another primarily through speech since it is the most natural, effective, and favored way of communication. As a result, it is reasonable to presume that individuals are more at ease utilizing speech as a means of input for a variety of computers as opposed to other more archaic modes of communication such as keypads and keyboards. A system that uses Automatic Speech Recognition (ASR) assists us in accomplishing this objective. A method like this enables a computer to accept an audio file or direct voice from a microphone as an input and convert it into text, ideally, the text will be in the script of the language that was said. The ability to accommodate variety in speech presents the greatest difficulty for automatic speech recognition. The idea of automatic speech recognition (also known as ASR) is broken down and analyzed in this study from the perspective of pattern recognition. In this paper, several various research gaps are investigated, as well as state-of-the-art architecture models of ASR, and their classification is explained from a comprehensive perspective. The development of a speech recognizer that works in real time may encounter obstacles ranging from challenging environments to the anatomy of the human body. In addition to that, it incorporates linguistic considerations. In this study, we investigate a variety of obstacles that must be overcome while creating a reliable ASR system.*

Keywords: ASR, MFCC, CSR, Speech Analysis, Deep Learning.

Introduction

Automatic Speech Processing systems drastically improved the past few years, especially Automatic Speech Recognition (ASR) systems. It is also the case for other speech processing tasks such as speaker identification, emotion classification, etc. This success was made possible by the large amount of annotated data available combined with the extensive use of deep learning techniques and the capacity of modern Graphics Processing Units. Some models are already deployed for everyday usage such as your personal assistants on your smart phones, your connected speakers and so on.

Nevertheless, challenges remain for automatic speech processing systems. They lack robustness against large vocabulary in real-world environment: this includes noises, distance from the speaker, reverberations and other alterations. Some challenges, such as CHIME [1], provide data to let the community try to handle some of these problems. It is being investigated to improve the generalization of modern models by avoiding the inclusion of other annotated data for every possible environment.

State-Of-The-Art (SOTA) techniques for most speech tasks require large datasets. Indeed, with modern DNN speech processing systems, having more data usually imply better performances. The TED-LIUM 3 from [2] (with 452 hours) provide more than twice the data of the TED-LIUM 2 dataset. Doing so, they

obtain better results by training their model on TED-LIUM 3 than training their model over TED-LIUM 2 data. This improvement in performance for ASR systems is also observed with the LibriSpeech dataset (from [3]). V. Panayotov et al. obtain better results on the Wall Street Journal (WSJ) test set by training a model over LibriSpeech dataset (1000 hours) than training a model over the WSJ training set (82 hours) [3]. This phenomenon, of having more data imply better performances, is also observable with the VoxCeleb 2 dataset compare to the VoxCeleb dataset: [4] increase the number of sentences from 100,000 utterances to one million utterances and increase the number of identities from 1251 to 6112 compared to the previous version of VoxCeleb. Doing so, they obtain better performances compare to training their model with the previous VoxCeleb dataset.

With under-resourced languages (such as [5]) and/or tasks (pathological detection with speech signals), we lack large datasets. By under-resourced, we mean limited digital resources (limited acoustic and text corpora) and/or a lack of linguistic expertise. For a more precise definition and details of the problem you may look [6]. Non-conventional speech tasks such as disease detection (such as Parkinson, gravity of ENT cancer and others) using audio are examples of tasks under resourced. Train Deep Neural Network models in such context is a challenge for these under-resourced speech datasets. This is especially the case for large vocabulary tasks. M. Moore et al. showed that recent ASR systems are not well adapted for impaired speech [7] and M. B. Mustafa et al. showed the difficulties to adapt such models with limited amount of data [8]. Few-shot learning consists of training a model using k-shot (where shot means an example per class), where $k \geq 1$ and k is a low number. Training an ASR system on a new language, adapting an ASR system on pathological speech or doing speaker identification with few examples are still complicated tasks. We think that few-shot techniques may be useful to tackle these problems.

This study captures all the aspects of an ASR from the feature extraction phase to language models with the following objectives in mind:

- To understand and explain the basic structure of an ASR (shown in Fig. 1) in detail, as well as discuss how using different techniques at different stages can affect the overall performance of the system.
- Discuss in detail the various classifications of ASR (shown in Fig. 2) and classification techniques being used for the development of an ASR.
- Different Evaluate parameters and Datasets (shown in table 1) with their description and how they affect the performance of an ASR.
- Various challenges of an ASR, including different state-of-the-art classification techniques.

The rest of the paper is organized as follows: Section 2 explains about the speech recognition system and its general structure, Section 3 presents a briefly about the various classification of ASR, Section 4 explains evaluation parameter on which outcomes are computed, Section 5 discusses various challenges in developing a robust ASR system, Section 6 explains various Research Gaps in the ASR Domain, finally we conclude our survey in section 7 followed by references used in this paper.

2. Speech Recognition

Speech recognition is a special case of pattern recognition. Fig. 1 show the processing stages involved in speech recognition. There are two phases in supervised pattern recognition, viz., training and testing. The process of extraction of features relevant for classification is common to both phases. During the training phase, the parameters of the classification model are estimated using a large number of class exemplars (training data). During the testing or recognition phase, the features of a test pattern (test speech data) are matched with the trained model of each and every class. The test pattern is declared to belong to that class whose model matches the test pattern best.

An ASR can generally be divided into 4 modules: a pre-processing module, a feature extraction module, an Acoustic model, and a language model, as shown in Fig. 1. Usually the input given to an ASR is captured using a microphone. This implies that noise may also be carried alongside the audio. The goal of preprocessing the audio is to reduce the signal-to-noise ratio [9]. There are different filters and methods that can be applied to a sound signal to reduce the associated noise. Framing, normalization, end-point detection and pre-emphasis are some of the frequently used methods to reduce noise in a signal [10, 11, 12]. Preprocessing methods also vary based on the algorithm being used for feature extraction. Certain feature extraction algorithms require a specific type of pre-processing method to be applied to its input signal.

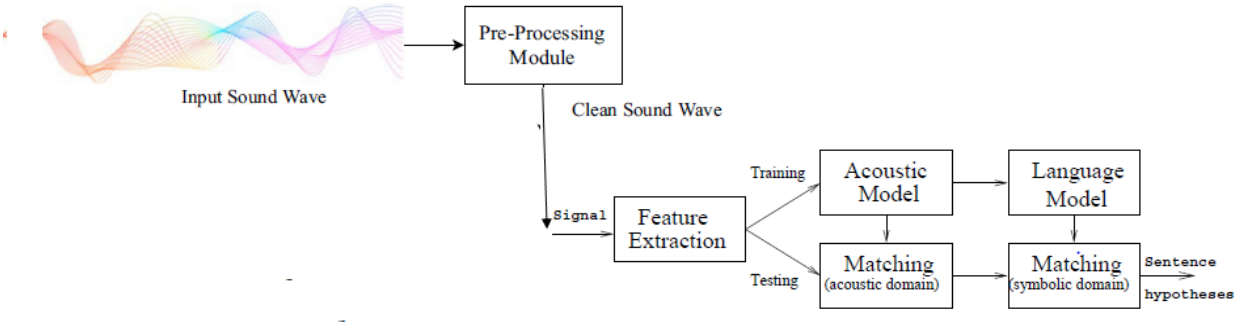

Fig.1: the general Structure of ASR.

After pre-processing, the clean speech signal is then passed through the feature extraction module. The performance and efficiency of the classification module are highly dependent upon the extracted features [13, 14]. There are different methods of extracting features from speech signals. Features are usually the predefined number of coefficients or values that are obtained by applying various methods on the input speech signal. The feature extraction module should be robust to different factors, such as noise and echo effect. Most commonly used feature extraction methods are Mel frequency cepstral coefficients (MFCCs), linear predictive coding (LPC), and discrete wavelet transform (DWT) [13, 15].

The third and final module is the classification model; this model is used to predict the text corresponding to the input speech signal. The classification models take input of the features extracted from the previous stage to predict the text. Like the feature extraction module, there are different types of approaches that can be applied to perform the task of speech recognition. The first type of approach uses joint probability distribution formed using the training dataset, and that joint probability distribution is used to predict the future output. This approach is called a generative approach; HMM and Gaussian mixture models (GMM) are the most commonly used models based on this approach. The second approach calculates a parametric model using a training set of input vectors and their corresponding output vectors. This approach is called the discriminative approach; Support Vector Machines (SVM) and ANN are its most common examples [16]. Hybrid approaches can also be used for classification purposes; one example of such a hybrid model is that of a HMM and ANN [17].

The language model is the last module of the ASR; it consists of various types of rules and semantics of a language. Language models are necessary for recognizing the phoneme predicted by the classifier; and is also used to form trigrams, words or sentences using all of the predicted phonemes of a given input. Most modern ASRs are designed to work without Language Models as well. Such ASRs can predict words and sentences spoken in the given input, but their efficiency can be increased significantly by using a language model [18].

The goal of speech recognition is to generate the optimal word sequence subject to linguistic constraints. The sentence is composed of linguistic units such as words, syllables, phonemes. The acoustic evidence provided by the acoustic models of such units is combined with the rules of constructing valid and meaningful sentences in the language to hypothesis the sentence. Therefore, in case of speech recognition, the pattern matching stage can be viewed as taking place in two domains: acoustic and symbolic. In the acoustic domain, a feature vector corresponding to a small segment of test speech (called a frame of speech) is matched with the acoustic model of each and every class. The segment is assigned a set of well matching class labels along with their matching scores. This process of label assignment is repeated for every feature vector in the feature vector sequence computed from the test data. The resultant lattice of label hypotheses is processed in conjunction with the language model to yield the recognized sentence.

3. Classification of ASR System

Speech recognition problem can be divided into following 3 major categories as shown in the Fig. 2.

A. Classification based upon Speaker Mode

On the basis of speaker mode, speech recognition systems can be classified as speaker dependent, speaker independent and speaker adaptive systems. Speaker dependence describes the degree to which a speech recognition system requires knowledge of the speaker‘s individual voice characteristics to successfully recognize the speech.

- Speaker dependent speech recognition system: Speech recognition systems that require a user to train the system for his/her voice are known as speaker dependent systems. These systems are usually easier to develop, cheaper to buy and more accurate. But these systems are not as flexible as speaker adaptive or speaker independent systems.
- Speaker independent speech recognition system: Speech recognition systems that do not require a user to train the system are known as speaker independent systems. A speaker independent system is developed to operate for any speaker. These systems are the most difficult to develop and have less accuracy but more expense than speaker dependent systems.
- Speaker adaptive speech recognition system: A speaker adaptive system is developed to adapt its operation to the characteristics of new speakers. It lies somewhere between speaker dependent and speaker independent systems.

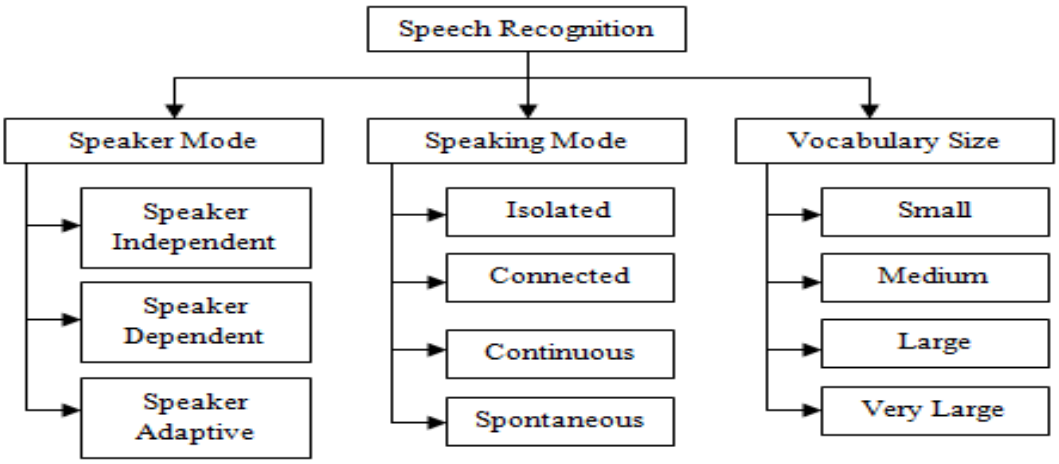

Fig.2: Classification of ASR Methods.

B. Classification based upon Speaking Mode

The speech recognition systems can be categorized into several different classes such as isolated words, connected words, continuous speech and spontaneous speech. A brief of each is given as below [19]:

- Isolated word speech recognition system: Isolated word recognizers (IWR) accept single word or single utterance at a time. It usually requires each utterance to have quiet (lack of an audio signal) on both sides of the sample window. This is the simplest form of recognition to perform because end points are easier to find and the pronunciation of a word doesn‘t affect others. It is important to note that even though IWR has a very limited recognition power; it has many applications in the real world.
- Connected words speech recognition system: Connected word systems (or more correctly 'connected utterances') are similar to isolated words, but allows separate utterances to be 'run-together' with a minimal pause between them.
- Continuous speech recognition system: A continuous speech system operates on speech in which words are connected together i.e. not separated by pauses. It allows users to speak almost naturally, while the computer determines the content. Recognizers with continuous speech capabilities are somewhat difficult to create because it is difficult to find the start and end points of words.
- Spontaneous speech recognition system: A spontaneous speech recognition system has the ability to handle a variety of natural speech features such as words being run together, "ums" and "ahs", and even slight stutters.

C. Classification based upon Vocabulary

ASR can also be classified based upon the size of the vocabulary as small size vocabulary, medium size vocabulary, large vocabulary and very large vocabulary speech system. The size of the vocabulary of a speech system affects the complexity, processing requirements and the accuracy of the system. A small size vocabulary consists of up to ten words. Vocabulary of hundreds of words is used by the medium vocabulary speech system. A large vocabulary has thousands of words whereas very large vocabulary contains tens of thousands of words.

4. Evaluation Parameters of ASR

Evaluation is one of the most important aspects of a conducted research because of its importance this section explains in detail different metrics that can be used to evaluate the performance of an ASR. The performance of a speech recognition system usually depends on two factors, the accuracy of the output produced as well as the processing speed of the ASR.

4.1 Real-time factor

The real-time factor (RTF) is the most commonly used metric for calculating the speed of a proposed model. The RTF can be computed by using the following formula:

$$RTF = \frac{P}{I}$$

Where P is the time taken by the system to process the input and I is the duration of the input audio. If RTF equals 1, then the input audio was processed in “Real-Time”. RTF is a highly hardware-dependent value and it is not only limited to calculating the speed of a speech recognition model. It can be used to calculate the speed of any model that can process an audio or video input.

4.2. Accuracy

The following methods can be used to measure the accuracy of an ASR:

Word error rate: The accuracy of an ASR is hard to calculate as the output produced by the ASR may not have the same length as the ground truth. Word error rate (WER) is the commonly used metric to estimate

the performance of an ASR, as it calculates error on word level rather than phoneme level [20]. The WER can be calculated using the following formula:

$$WER = \frac{S + D + I}{N}$$

Where S is the number of substitutions performed in the output text as compared to the ground truth. D is the number of deletions performed, and I is the number of insertions performed. N is the total number of words in the ground truth.

Word recognition rate: Word Recognition Rate (WRR) is a variation of WER that can also be used to evaluate the performance of an ASR. It can be calculated using the following formula:

$$WRR = 1 - WER$$
$$= \frac{N - S - D - I}{N}$$
$$= \frac{H - I}{N}$$

Where H = N - (S + D) represents the total number of correctly guessed words.

5. Various Challenges of ASR

5.1 Psycho-Intellectual Aspect

5.1.1 Human Understanding of Speech:

Human have their own knowledge base which is resulted from the reading, experiments, experiences, examination, situation, interaction and communication. They may hear more than the speaker speaks to them. While speaking speakers have its own language model of native language. Human may understand and interpret the words or sequence of words, they never heard before.

In automatic speech recognition we need to develop the annotated corpus and language model which provide the system with the limited grammatical structures. To increase the chance of pattern matching one can use the statistical models like Hidden Markov model. At the level of human understanding the knowledge can be represented by the exhaustive content. In ASR the limited knowledge needs to build and can be trained accordingly. Vocabulary size plays a crucial role in speed of the system. Search time and resources get increase due large vocabulary. Limited vocabulary has the restricted application domain.

5.1.2 Spoken Language and Written Language:

Spoken language that has been spoken by the people is not similar to that of written language. Spoken languages are less complicated than written languages. In spoken language people use shortened unit, repetition of words to emphasize important ideas and information, use of more familiar words to increase audience understanding, In ASR this problem needs to identify and address. Speech is a two way communication and is dialogue oriented. The dialogue is delivered and is interpreted in the form of feature vector and to understand the meaning of words, we adopt and train the receiver speech data. In written communication the dialogue is one way and is bounded by the complex structure of language. The grammatically spoken language is quite different from written language at many different levels.

Some difference we pointed are as follows:

- In spoken language there is often a radical reduction of morphemes and words in pronunciation.
- The frequencies of words collection and grammatical construction are highly different between spoken and written language.

- The grammar and semantic of spoken language is also significantly different from that of written language: 30-40% of all utterance consists of short utterances of 1-2-3 words with no predicative verb.
- Use of more colloquial words and shortened form make the conversation lively
- Much less use of terms and phrases that work in writing but can lose their meaning or become confusing in speaking. Examples: "as mentioned above," "the former… the latter," and "respectively."

The speech recognition system should address these issues at psycho-intellectual level, as spoken language and written language is different from each other.

5.1.3 Multimodality in human-human communication:

The indispensable need for speech system design is to understand the multimodal communication between human. Human contributes more with the body language while he communicates. It includes eye movement, postures, symbolic gestures, hand movement, human-human communication which is mediated by computer [25]. When knowledge from one modality is weak, it may be complimented by another modality. Auditory and -visual information can be used to its optimum for speech recognition. Speech recognizer can use this cue to ensure improvement in speech recognition accuracy.

5.1.4 Background Noise:

In a natural environment, human speaks the words and sentences which incorporate the background noise. The noise is unwanted information in the clearest speech signal. It may be an echo effect, another speaker, speaking in the background, playing instruments in the background and so on. The speech signals mixed with noise is difficult to recognize accurately. Noise contaminated speech signal gives rise to speech variability and recognition become difficult in real time events. Reverberation and noise elimination are most challenging tasks. A robust noise reduction algorithm can be deployed dynamically for this type of challenge.

5.2 Continuous Speech

The real time ASR system works with the continuous speech. Phonetic, syllabic and word level recognition is not complex as recognition performed in isolation, but when the sequence of words/sentence is spoken out recognition become more critical. We need to deal with the word boundary ambiguity which is a difficult problem for ASR. One way to address this difficulty is to give the pauses while speaking so that adjacent words can be identified clearly, but it tends to loss naturalness in the speaking style and also it needs the training of the speaker while speaking for increased length of the sentences. Current Continuous Speech Recognition (CSR) systems with large vocabulary are strictly based on the principles of statistical pattern recognition [20].

5.3 Channel Variability and Noise

The acoustic of the speech signal is affected by the noise as well as a microphone that is used for acquiring the speech signal. This is called as channel variability. This phenomenon can be addressed by applying the filters as well as using the high end microphones.

5.4 Speaker Variability

All speakers have their unique voice as fingerprint due to the unique anatomy of the vocal tract and personality the speaker variability can be,

5.4.1 Phone-Realization:

The way of speaking the same word may result in different pronunciation every time. Hence the acoustic wave of the speech may vary over time for the same utterance. This is the phone realization of speech.

5.4.2 Accent:

While investigating the variability between speakers through statistical analysis methods the first two principal components of variation correspond to the gender (and related to physiological properties) and accent respectively [21]. Indeed, compared to native speech recognition, performance degrades when

recognizing accented speech and non-native speech [22] In fact accented speech is associated with a shift within the feature space [23]. The accents of speaker differentiate with respect to their personality. Each one is having the unique way of pronunciation and emphasize. The speaking accent may vary according to the social and personal situations. Speaker uniqueness results from the complex combination from pschysiological and cultural aspects. [21] We may vary in accent while speaking to parents, friends. It also affects with emotional level, which may affect the loudness, pressure while speaking.

5.4.3 Gender of the Speaker and Anatomy of Vocal Tract:

Men and women have different fundamental frequencies while speaking. It is because women have shorter vocal tract than men have. The fundamental frequency of a woman's voice is roughly two times higher than a man's voice. The shape and length of the vocal cords, formation of cavities and size of the lungs may change over time and it may cause the speaker variability. Children have shorter vocal tract and vocal folds compared to adults. This results in higher positions of formants and fundamental frequency. The high fundamental frequency is reflected in a large distance between the harmonics, resulting in poor spectral resolution of voiced sounds. The difference in vocal tract size results in a non-linear increase of the formant frequencies. In order to reduce these effects, previous studies have focused on the acoustic analysis of children's speech [24] this variability issue has been addressed by vocal tract length normalization [25] as well as spectral normalization [26]

5.4.4 Speed of the speech:

The speed while speaking may vary and depend upon the situations, physical stress. Rate-of-speech (ROS) is considered as an important factor which makes the mapping process between the acoustic signal and the phonetic categories more complex. Timing and acoustic realization of syllables are affected due in part to the limitations of the articulator machinery, which may affect pronunciation through phoneme reductions, time compression/expansion, changes in the temporal patterns, as well as smaller-scale acoustic–phonetic phenomena. Due to rate of speech significant degradation in performance has been reported.

5.5 Regional and Social Dialects

Dialects are group related variant within language. Regional dialects are the features of pronunciation, vocabulary and grammar which differ according to the geographical area of the speaker. [27]

In many cases we may be forced to consider dialects as another language in ASR due to large differences between two dialects. Dialects of the specific language differ from each other, but they are still understandable by the speaker of another dialect of the same language. Differences among dialects are mainly due to the social and regional factor. During the past few years there have been significant attempt to automatically recognize the dialect or accent of the a speaker given his her speech utterances [28, 29], Recognition of dialects or accents of speakers prior to automatic speech recognition (ASR) helps in improving performance The ASR systems by adapting the ASR acoustic and/or language models appropriately [30].

5.6 Amount of Data and Search Space

Day to day communication produces the largest amount of speech data. This has to match with the group of phones, sounds, the words and the sentences.

The quality of input can be regulated by the number of samples of the input signal, but the quality of the speech signal will decrease with the lower sampling rate.

5.7 Ambiguity

Spoken language have the ambiguity i.e. the set of words have different meaning. There is ambiguity that needs to deal with speech recognition are homophones and word boundary ambiguity.

5.7.1 Homophones:

Homophones are the words that sound the same but have different orthography. The words or sounds same, but the meaning is different. How ASR will distinguish between homophones? It is impossible at the world level is ASR, we need a larger context to decide which is intended.

5.7.2 Word boundary ambiguity:

When the sequences of groups of phones are put into a sequence of words, we sometimes encounter word boundary ambiguity. Word boundary ambiguity occurs when there are multiple ways of grouping phones into words. There are some examples which naturally occur during the spoken communication which ends with the rhyming sounds. This can be viewed as a specific case of handling the continuous speech where even a human can have problems with finding the word boundaries.

6. Research Gap in ASR

Research Gap in Automatic Speech Recognition (ASR):

- Data Efficiency: While deep learning has shown remarkable progress in ASR, these models often require large amounts of labeled data for training. Research is needed to develop more data-efficient ASR models that can achieve high accuracy with smaller training datasets. Techniques such as transfer learning and few-shot learning in the context of ASR remain relatively unexplored.
- Self-Supervised Learning: Self-supervised learning has been successful in other domains, but its application to ASR is still in its early stages. There is a gap in exploring self-supervised approaches for training ASR models, which could reduce the reliance on manual transcription efforts.
- Unsupervised ASR: Current ASR systems mostly rely on supervised learning with transcribed data. Research is needed to advance unsupervised ASR, where models can learn from transcribed audio data, making ASR more accessible for low-resource languages and dialects.
- Multilingual and Cross-Lingual ASR: Developing machine learning-based ASR systems that can handle multiple languages or perform cross-lingual recognition remains a challenge. Research should focus on improving the accuracy and efficiency of multilingual and cross-lingual ASR models.
- Robustness to Adverse Conditions: ASR systems often perform poorly under adverse conditions, such as noisy environments or heavy accents. More research is needed to develop machine learning techniques that enhance ASR robustness in real-world scenarios.
- Neural Architecture Search (NAS): NAS has shown promise in optimizing neural network architectures. Applying NAS techniques to ASR could lead to the discovery of novel architectures that outperform handcrafted designs, but this area requires more exploration.
- Continual Learning and Adaptation: ASR models should be able to continually adapt to new speakers, accents, and environments. Research on continual learning techniques for ASR, where models can learn incrementally without catastrophic forgetting, is a valuable research direction.
- Multimodal Integration: Integrating ASR with other machine learning modalities, such as natural language processing and computer vision, is an area ripe for exploration. Research is needed to develop effective techniques for combining multiple modalities to enhance ASR accuracy and understanding.
- Interpretable ASR Models: ASR models are often considered black boxes. Research is needed to develop interpretable ASR models that can provide insights into their decision-making processes, making them more transparent and trustworthy.
- Low-Resource ASR: Many languages and dialects lack sufficient data for training high-performance ASR models. Developing machine learning techniques that can leverage limited resources effectively is a critical research gap in low-resource ASR.

- Active Learning for ASR: Active learning strategies can help select the most informative data for training ASR models, reducing annotation costs. Research on effective active learning methods specific to ASR is needed.
- Fairness and Bias Mitigation: Addressing bias and fairness issues in ASR systems is crucial. Research should focus on developing machine learning techniques that can detect and mitigate biases in ASR models, ensuring equitable performance across diverse user groups.
- Energy-Efficient ASR: As ASR systems find applications in resource-constrained devices, energy-efficient ASR models are needed. Research should explore techniques to reduce the computational and energy demands of ASR models while maintaining accuracy.
- Deployment and Scalability: The transition from research prototypes to real-world deployment of ASR systems can be challenging. Research should address the scalability, robustness, and reliability aspects of ASR models in production environments.

7. Conclusion

This survey paper evaluated and examined several methods and strategies that are applied to the voice recognition task. An ASR is dependent on three modules: the language model, the classification module, and the feature extraction module, according to the explanation of an ASR's fundamental architecture. As technology develops, new applications that are essential to human life are created. The difficulties of creating a speech recognition system that mimics a human are still being worked on. In spite of the difficulties relating to speaker variability, channel variability, and environment variability, enhancing the accuracy in speech recognition systems is a matter of mitigating the impacts. To achieve the resilience in ASR, characterizing the effects might be examined first. This paper discussed the numerous issues that, if correctly resolved, can lead to real-time, reliable ASR. Modern speech recognition technology is focused on modeling speech variance parameters, including psycho-intellectual characteristics, dialects, accents, rate of speech, speaker gender, age, health, and emotional state, as well as some physiological variables, like variations in vocal tract length. Due to these variances, it is challenging to simulate large, reliable speaker-independent systems. If humans desire to communicate naturally with computers, all of the obstacles and hurdles must be overcome.

References

1. J. Barker, S. Watanabe, E. Vincent, and J. Trmal, “The Fifth ’CHiME’ Speech Separation and Recognition Challenge: Dataset, Task and Baselines,” in Interspeech 2018. ISCA, Sep. 2018, pp. 1561–1565.
2. F. Hernandez, V. Nguyen, S. Ghannay, N. Tomashenko, and Y. Est`eve, “TED-LIUM 3: Twice as much data and corpus repartition for experiments on speaker adaptation,” in Speech and Computer - 20th International Conference, vol. 11096, Sep. 2018, pp. 198–208.
3. V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: An ASR corpus based on public domain audio books,” in 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). South Brisbane, Queensland, Australia: IEEE, Apr. 2015, pp. 5206–5210.
4. J. S. Chung, A. Nagrani, and A. Zisserman, “VoxCeleb2: Deep Speaker Recognition,” in Interspeech 2018. ISCA, Sep. 2018, pp. 1086–1090.
5. B. Deka, J. Chakraborty, A. Dey, S. Nath, P. Sarmah, S. R. Nirmala, and S. Vijaya, “Speech corpora of under resourced languages of northeast india,” in 2018 Oriental COCOSDA - International Conference on Speech Database and Assessments, Miyazaki, Japan, May 7-8, 2018, 2018, pp. 72–77.
6. L. Besacier, E. Barnard, A. Karpov, and T. Schultz, “Automatic speech recognition for under-resourced languages: A survey,” Speech Communication, vol. 56, pp. 85–100, Jan. 2014.

7. M. Moore, H. Venkateswara, and S. Panchanathan, “Whistle-blowing ASRs: Evaluating the Need for More Inclusive Speech Recognition Systems,” in Interspeech 2018. ISCA, Sep. 2018, pp. 466–470.
8. M. B. Mustafa, S. S. Salim, N. Mohamed, B. Al-Qatab, and C. E. Siong, “Severity-Based Adaptation with Limited Data for ASR to Aid Dysarthric Speakers,” PLoS ONE, vol. 9, no. 1, p. e86285, Jan. 2014.
9. Yegnanarayana, B., and Raymond NJ Veldhuis. "Extraction of vocal-tract system characteristics from speech signals." IEEE transactions on Speech and Audio Processing 6, no. 4 (1998): 313-327.
10. Mporas, Iosif, Todor Ganchev, Mihalis Siafarikas, and Nikos Fakotakis. "Comparison of speech features on the speech recognition task." Journal of Computer Science 3, no. 8 (2007): 608-616.
11. O'Shaughnessy, Douglas. "Interacting with computers by voice: automatic speech recognition and synthesis." Proceedings of the IEEE 91, no. 9 (2003): 1272-1305.
12. Saha, G., Sandipan Chakroborty, and Suman Senapati. "A new silence removal and endpoint detection algorithm for speech and speaker recognition applications." In Proceedings of the NCC, pp. 56-61. 2005.
13. Krishnan, VR Vimal, and P. Babu Anto. "Features of wavelet packet decomposition and discrete wavelet transform for malayalam speech recognition." International Journal of Recent Trends in Engineering 1, no. 2 (2009): 93.
14. Zamani, Behzad, Ahmad Akbari, Babak Nasersharif, and Azarakhsh Jalalvand. "Optimized discriminative transformations for speech features based on minimum classification error." Pattern Recognition Letters 32, no. 7 (2011): 948-955.
15. Ranjan, Shivesh. "A discrete wavelet transform based approach to Hindi speech recognition." In 2010 International Conference on Signal Acquisition and Processing, pp. 345-348. IEEE, 2010.
16. Bernardo, J. M., M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West. "Generative or discriminative? getting the best of both worlds." Bayesian statistics 8, no. 3 (2007): 3-24.
17. Tang, Xian. "Hybrid Hidden Markov Model and artificial neural network for automatic speech recognition." In 2009 Pacific-Asia Conference on Circuits, Communications and Systems, pp. 682-685. IEEE, 2009.
18. Chan, William, Navdeep Jaitly, Quoc Le, and Oriol Vinyals. "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition." In 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4960-4964. IEEE, 2016.
19. M. A. Anusuya, and S. K. Katti, ―Speech Recognition by machine: A review,‖ Int. J. Computer Science and Information Security, vol. 6, no. 3, pp. 181-205, 2009.
20. Rabiner, Lawrence R., and Biing-Hwang Juang. "Speech recognition: Statistical methods." Encyclopedia of language & linguistics (2006): 1-18.
21. Nolan, Francis, and Harry Hollien. "The Phonetic Bases of Speaker Recognition by Francis Nolan." (1985): 817-817.
22. Lawson, Aaron D., David M. Harris, and John J. Grieco. "Effect of foreign accent on speech recognition in the NATO N-4 corpus." In Eighth European Conference on Speech Communication and Technology. 2003.
23. Van Compernolle, Dirk. "Recognizing speech of goats, wolves, sheep and… non-natives." Speech Communication 35, no. 1-2 (2001): 71-79.
24. Lee, Sungbok, Alexandros Potamianos, and Shrikanth Narayanan. "Acoustics of children’s speech: Developmental changes of temporal and spectral parameters." The Journal of the Acoustical Society of America 105, no. 3 (1999): 1455-1468.

25. Das, Subrata, Don Nix, and Michael Picheny. "Improvements in children's speech recognition performance." In Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181), vol. 1, pp. 433-436. IEEE, 1998.
26. Lee, Li, and Richard C. Rose. "Speaker normalization using efficient frequency warping procedures." In 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings, vol. 1, pp. 353-356. IEEE, 1996.
27. Akmajian, Adrian, Ann K. Farmer, Lee Bickmore, Richard A. Demers, and Robert M. Harnish. Linguistics: An introduction to language and communication. MIT press, 2017.
28. Liu, Gang A., and John HL Hansen. "A systematic strategy for robust automatic dialect identification." In 2011 19th European Signal Processing Conference, pp. 2138-2141. IEEE, 2011.
29. Nerbonne, John. "Linguistic variation and computation (invited talk)." In 10th conference of the european chapter of the association for computational linguistics. 2003.
30. Liu, Mingkuan, Bo Xu, Taiyi Hunng, Yonggang Deng, and Chengrong Li. "Mandarin accent adaptation based on context-independent/context-dependent pronunciation modeling." In 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 00CH37100), vol. 2, pp. II1025-II1028. IEEE, 2000.