Research Article

A Review on Various Approach of Speech Recognition Technique

Dr. Bhosale Rajkumar S.
Amrutvahini College of Engineering, Sangamner
Department of Information Technology
bhos_raj@rediffmail.com

Dr. Panhalkar A.R.
Amrutvahini College of Engineering, Sangamner
Department of Computer Engineering
archana10bhosale@rediffmail.com

Abstract- The Speech is most prominent & primary mode of Communication among of human being. The communication among human computer interaction is called human computer interface. Speech has potential of being important mode of interaction with computer. This paper gives an overview of major technological perspective and appreciation of the fundamental progress of speech recognition and also gives overview technique developed in each stage of speech recognition. This paper helps in choosing the technique along with their relative merits & demerits. A comparative study of different technique is done asperstages.

After years of research and development the accuracy of automatic speech recognition remains one of the promising research challenges (eg. variation of the context, speakers, and environment). The design of Speech Recognition system requires careful attentions to the following issues: Definition of various types of speech classes, speech representation, feature extraction techniques, speech classifiers, and database and performance evaluation. The problems that are existing in ASR and the various techniques to solve these problems constructed by various research workers have been presented in a chronological order.

Real-time speech recognition is a challenging task due to the variability of speech signals and the need for fast and accurate processing. Support Vector Machines (SVMs) are a popular machine learning technique that has been used for speech recognition tasks. In this paper, we present a real-time speech recognition system using SVM. The system is based on a feature extraction process that uses Mel-Frequency Cepstral Coefficients (MFCCs) to represent speech signals. The extracted features are then used as input to the SVM classifier, which is trained to recognize different speech signals. The proposed system was implemented using the Python programming language and the Scikit-learn machine learning library. The performance of the system was evaluated using a dataset of spoken digits. The results showed that the proposed system achieved high recognition accuracy and real-time performance, making it suitable for practical applications.

Speech is a unique human characteristic used as a tool to communicate and express ideas. Automatic speech recognition (ASR) finds application in electronic devices that are too small to allow data entry via the commonly used input devices such as keyboards. Personal Digital Assistants (PDA) and cellular phones are such examples in which ASR plays an important role.

Keywords: real-time speech recognition, support vector machines, Mel-Frequency Cepstral Coefficients, machine learning.

I. Introduction

The speech is primary mode of communication among human being and also the mostnatural and efficient form of exchanging information among human in speech. So, it is onlylogical that the next technological development to be natural language speech recognition forHCI. Speech Recognition can be defined as the process of converting speech signal to a sequence of words by means Algorithm implemented as a computer program. Speech processing is one of the exciting areas of signal processing. The goal of speech recognitionarea is to developed technique and system to develop for speech input to machine. Based onmajor advanced in statically modeling of speech, automatic speech recognition today findswidespread application in task that require human machine interface such as automatic call processing. Since the 1960 s computer scientists have been researching ways and means to make computers able to record interpret and under-stand human speech. Through-out the decades this has been a daunting task. Even the most rudimentary problem such as digitalizing (sampling) voice was a huge challenge in the early years. It took until the 1980sbefore the first systems arrived which could actually decipher speech. Off course these earlysystems were very limited in scope and power. Communication among the human being isdominated by spoken language, therefore it is natural for people to expect speech interfaces with computer. Automatic speech recognition systems convert a speech signal into a sequenceof words, usually based on the Hidden Markov Model (HMM) (Young 1990). Computer which can speak and recognize speech in native language. Machine reorganization of speech involves generating a sequence of words best matches the given speech signal. Someof known applications include virtual reality, Multimedia searches, auto-attendants, travel Information and reservation, translators, natural language understanding, In medical and many more Applications (Scan soft, 2004; Robertson, 1998. Support Vector Machine is analso different method which is not based on neural or HMM it is work on classification base. This is work for speaker dependent method and it is used to give promising result [11].

Now in the decade of 2000s Speech recognition is a conversion from an acoustic wave form to a written equivalent of the message information. The nature of the speech recognitionproblem is heavily depending upon the constraints placed on speaker, speaking situation andmessage context. The most of the application of speech recognition systems are many and varied; e.g. a voice operated type writer and voice communication with computers and command line interface with machine. The listening tests are conducted on a large vocabularytask, recognition accuracy by human was found to be an order of magnitude higher than machines. Though these tests are included data with varied signal qualities, human recognition performance was found to be consistent over a diverse set of conditions.

Support Vector Machines (SVMs) are a powerful machine learning technique that has been used for speech recognition with promising results. SVMs are based on the idea of finding a hyperplane that best separates data points into different classes. In speech recognition, this involves training an SVM to classify speech signals based on their acoustic features [12].

One of the main advantages of SVMs is their ability to handle high-dimensional data, which is common in speech recognition tasks. This is achieved by mapping the high-dimensional data to a higher-dimensional space using a kernel function, which makes it easier to find a hyperplane that separates the data into different classes.

SVMs have been used in various aspects of speech recognition, including phoneme recognition, speaker recognition, and speech emotion recognition. In phoneme recognition, an SVM is trained to classify speech sounds into different phonemes based on their acoustic features, such as MFCCs. SVMs have been shown to achieve high accuracy in phoneme recognition, particularly when combined with other techniques, such as Hidden Markov Models (HMMs) [13].

In speaker recognition, SVMs are used to classify speech signals into different speakers based on their acoustic features. This can be useful in security applications, where it is important to verify the identity of a speaker. SVMs have been shown to be effective in speaker recognition, particularly when used with features such as MFCCs and pitch.

In speech emotion recognition, SVMs are used to classify speech signals into different emotions based on their acoustic features. This can be useful in applications such as virtual assistants or customer service, where it is important to understand the emotional state of the user. SVMs have been shown to achieve high accuracy in speech emotion recognition, particularly when used with features such as prosodic features and spectral features. However, SVMs can be computationally intensive and may require a large amount of training data to achieve good performance. Additionally, the choice of kernel function can have a significant impact on the performance of the SVM, and selecting the appropriate kernel function can be challenging [14].

II. Type Of Speech

Speech recognition system can be separated in different classes by describing what type of ullerances they can recognize. Here introduce types of speech in speech recognition. There are several types of speech that can be recognized by speech recognition software.

When it comes to speech recognition, there are a few different categories of speech that are commonly recognized.In order to accurately transcribe speech, speech recognition software is trained to recognize various types of speech, such as.Speech recognition technology has come a long way in recent years, and can now recognize a wide range of speech types, including.Whether you're dictating a document or navigating your phone with your voice, speech recognition technology relies on recognizing different types of speech, such as.

A. Isolated Word

Isolated word recognizes attain usually require each utterance to have quiet on bothside of sample windows. It accepts single words or single utterances at a time .This is having “Listen and Non Listen state”. Isolated utterance might be better name of this class.

B. Connected Word

Connected word systemare similar to isolated words but allow separateutterance tobe“run togetherminimum pausebetween them.

C. Continuous speech

Continuous speech recognizers allows user to speak almost naturally, while the computer determine the content. Recognizer with continues speech capabilities are some of the most difficult to create because they utilize special method to determine utterance boundaries.

D. Spontaneous speech

At a basic level, it can be thought of as speech that is natural sounding and notrehearsed .an ASR System with spontaneous speech ability should be able to handle a varietyofnatural speech featuresuch aswords being run together.

E. Command and Control Speech

This type of speech recognition system is designed to recognize a limited set of spoken commands, such as those used to control a device or software application.

F. Natural Language Speech

This type of speech recognition system is designed to recognize spoken language in a more natural and conversational way, allowing users to interact with computers or other devices using everyday language.

G. Speaker-Dependent Speech

This type of speech recognition system is trained to recognize the speech patterns of a specific user, and is therefore more accurate for that particular user.

H. Speaker-Independent Speech

This type of speech recognition system is not trained on a specific user's speech patterns, and is therefore more versatile and can be used by anyone. However, it may not be as accurate as a speaker-dependent system.

I. Acoustic Models

These are mathematical models that describe the relationship between speech sounds and the physical characteristics of the sound waveforms. They are used to help recognize speech in a speech recognition system.

J. Language Models:

These are statistical models that represent the probability distribution over sequences of words or phrases in a particular language. They are used to help recognize the meaning of spoken language in a speech recognition system[15].

III. Structure Of Speech Reco Gnition Techniques

The speaker recognition system may be viewed in four steps.

I. Analysis

II. Featureextraction(training)

III. Modelling

IV. Testing

The overall system shows above in four steps can be elaborated as below in detail

I. Audio input: The system receives an audio signal, which is typically captured by a microphone or other recording device.

II. Preprocessing: The audio signal is preprocessed to remove noise, enhance the signal-to-noise ratio, and convert the analog signal into a digital form that can be processed by a computer.

III. Feature extraction: The preprocessed signal is then analyzed to extract features that are relevant to the speech recognition task. Common features include Mel-frequency cepstral coefficients (MFCCs), which represent the spectral envelope of the speech signal.

IV. Acoustic modeling: The extracted features are used to train a statistical model that can map acoustic features to phonemes or other linguistic units. This model is typically trained using a large corpus of labeled speech data.

V. Language modeling: In addition to the acoustic model, a speech recognition system also requires a language model that can predict the probability of word sequences given the context of the speech. This model is typically trained using large text corpora.

VI. Decoding: The final stage of the speech recognition system involves decoding the input speech signal into a sequence of words or other linguistic units. This is done by combining the output of the acoustic and language models, and selecting the most likely word sequence given the input signal and the language model.

Overall, the structure of a speech recognition system involves a combination of signal processing, statistical modeling, and machine learning techniques, which work together to convert spoken language into a machine-readable form [16].

A. Speech analysis technique

Speech data contain different type of information that shows a speaker identity. This includes speaker specific information due to vocal tract, excitations our and behavior feature. The information about the behavior feature also embedded in signal and that can beused for speaker recognition. The speech analysis stage is deals with stage with suitable frame size for segmenting speech signal for further analysis and extracting.

There are several techniques used in speech analysis, including:

1. Acoustic analysis: Acoustic analysis involves measuring the physical properties of speech sound, including their frequency, duration, and amplitude.
2. Phonetic transcription: Phonetic transcription involves representing speech sounds using a set of standardized symbols, known as the International Phonetic Alphabet (IPA).
3. Pitch analysis: Pitch analysis involves studying the pitch (highness or lowness) of speech sounds and how it changes over time. This can provide insights into the emotional state of the speaker, as well as the emphasis and intonation patterns used in speech.
4. Formant analysis: Formant analysis involves studying the frequencies of the resonance peaks in speech sounds. This can help identify different vowel sounds and provide information about the speaker's accent or dialect.
5. Prosodic analysis: Prosodic analysis involves studying the rhythm, intonation, and stress patterns of speech. This can provide information about the speaker's emotional state, as well as the communicative goals of the speech.
6. Discourse analysis: Discourse analysis involves studying the way language is used in context, including social and cultural factors that influence communication. This can include the analysis of patterns of language use, such as the use of specific vocabulary or nonverbal cues.
7. Content analysis: Content analysis involves studying the content of speech, including the themes, topics, and ideas that are communicated. This can be done by analyzing the words used, as well as the structure and organization of the speech.

These techniques are often used in combination to gain a comprehensive understanding of speech communication.

B. Feature Extraction Technique

Speech feature extraction techniques involve the identification and extraction of specific features or characteristics from speech signals. Some commonly used techniques for speech feature extraction include:

1. Mel-Frequency Cepstral Coefficients (MFCC): MFCC is a widely used technique for speech feature extraction. It involves breaking down speech signals into small frames and analyzing the frequency spectrum of each frame. The resulting features are then transformed using the Discrete Cosine Transform (DCT) to create a set of coefficients that are used to represent the speech signal.
2. Linear Predictive Coding (LPC): LPC is a technique that involves modelling the spectral envelope of speech signals using a series of linear filters. The resulting coefficients are used to represent the speech signal.
3. Perceptual Linear Prediction (PLP): PLP is a technique that involves modelling the human auditory system to create a set of features that are more closely related to the way that humans perceive speech. It involves applying a series of non-linear transformations to the speech signal to create a set of features.
4. Wavelet Transform: Wavelet Transform is a technique that involves analyzing the frequency content of a speech signal at different scales using wavelets. The resulting features are used to represent the speech signal.
5. Short-Time Fourier Transform (STFT): STFT is a technique that involves analyzing the frequency content of speech signals over small time intervals. The resulting features are used to represent the speech signal.

These techniques are often used in combination to create a comprehensive set of features that can be used for speech analysis, speech recognition, and other applications [17].

The speech feature extraction in a categorization problem is about reducing the dimensionality of the input vector while maintaining the discriminating power of the signal. As we know from fundamental formation of speaker identification and verification system, that the number of training and test vector needed for the classification problem grows with the dimension of the given input so we need feature extraction of speech signal.

Feature extraction is a process of extracting important and relevant features from raw data to represent it in a meaningful way that can be used for analysis or classification. There are several techniques for feature extraction, some of which are [22]:

1. Principal Component Analysis (PCA): PCA is a widely used technique for feature extraction that reduces the dimensionality of data by finding the principal components that explain the most variance in the data. PCA is particularly useful for dealing with high-dimensional data.
2. Linear Discriminant Analysis (LDA): LDA is a supervised learning technique that finds the linear combination of features that maximizes the separation between different classes. It is commonly used for dimensionality reduction and feature extraction in classification problems.
3. Independent Component Analysis (ICA): ICA is a technique that separates a multivariate signal into independent, non-Gaussian components. It is often used for signal processing, image processing, and feature extraction.
4. Wavelet Transform: Wavelet transform is a mathematical technique that decomposes a signal into a set of wavelets with different frequency ranges. It is often used for feature extraction in time-series analysis, image processing, and pattern recognition.
5. Histogram of Oriented Gradients (HOG): HOG is a feature extraction technique that extracts local image gradients and computes histograms of the orientations of these gradients. It is often used for object detection in computer vision [23].
6. Scale-Invariant Feature Transform (SIFT): SIFT is a feature extraction technique that detects local features in an image that are invariant to scaling, rotation, and translation. It is often used for object recognition and image matching.
7. Speeded Up Robust Features (SURF): SURF is a feature extraction technique that is similar to SIFT but is faster and more robust to noise and image deformations. It is often used for object recognition and image matching.
8. Convolutional Neural Networks (CNNs): CNNs are a type of deep neural network that automatically learn hierarchical representations of data. They are often used for image and speech recognition, and can be used for feature extraction in other domains as well.

These techniques are some are more computationally expensive than others and require more data, while others are more robust to noise and variations in the data. The choice of technique depends on the specific problem and data at hand [24].

Following are some feature extractions.

Table 1: List of technique with their properties For Feature extraction

Sr. no.	Method	Property	Procedure for Implementation
1.	Principal Component analysis (PCA)	Nonlinear feature extraction method Linear map, fast, eigen vector-based.	Traditional, eigen vector base method, also known as karhuneu-Loeve expansion; good for Gaussian data.
2.	Linear Discriminate Analysis (LDA)	Non-linear feature extraction method, Supervised linear map; fast, eigenvector-based	Better than PCA for classification [9]
3.	Independent Component Analysis (ICA)	Non-linear feature extraction method, Linear map, iterative non-Gaussian	Blind course separation, used forde-mixingnon-Gaussian distributed sources(features)
4.	Linear Predictive coding	Static feature extraction method,10 to16 lower order coefficient,	It is used for feature Extraction at lower Order.
5.	Cepstral Analysis	Static feature extraction method, Power spectrum	Used to represent spectral envelope [9]
6.	Mel-frequency scale analysis	Static feature extraction method, Spectral analysis	Spectral analysis is done with a fixed resolution along a Subjective frequency Scalei.e.Mel-frequencyScale.
7.	Filter bank analysis	Filters tuned required frequencies	
8.	Mel-frequency cepstrum (MFFCs)	Power spectrum is computed by Performing Fourier Analysis	This method issued for find our features.

9.	Kernel based feature extraction method	Nonlineartransformations	Dimensionality reduction leads to better classification and it is used to redundant features, and improvement in classification error [11].
10.	Wavelet	Better time resolution than Fourier Transform	It replaces the fixed bandwidth of Fourier transform with one proportional to frequency which allow better time resolution at high frequencies than Fourier Transform
11.	Dynamic feature extractions i) LPC ii)MFCCs	Acceleration and delta coefficients i.e. II and III order derivatives of normal LPC and MFCCs coefficients	It is used by dynamic or run time Feature
12.	Spectral subtraction	Robust Feature extraction method	It is used basis on Spectrogram [4]
13.	Cepstral mean subtraction	Robust Feature extraction	It is same as MFCC but working on Mean statically parameter
14.	RASTAfiltering	For Noisy speech	It is find out Feature in Noisy data
15.	Integrated Phoneme subspace method (Compound Method)	A transformation based onPCA+LDA+IC	Higher Accuracy than the existing Methods [14].
16.	Support Vector Machine	Nonlinear Feature extraction Method and use One to one classifier for testing	Gives Higher Accuracy by high pressure Microphone

C. *ModelingTechnique*

The objective of modeling technique is to generate speaker models using speaker specific feature vector. The speaker modeling technique divided into two classification speaker recognition and speaker identification. The speaker identification technique automatically identifies who is speaking on basis of individual information integrated in speech signal the speaker reorganization is also divided into two parts that means speaker dependent and speaker independent [21]. In the speaker independent mode of the speech reorganization the computer should ignore the speaker specific characteristics of the speech signal and extractthe intended message .on the other hand in case of speaker reorganization machine should extract speaker characteristics in the acoustic signal. The main aim of speaker identification iscomparing a speech signal from an unknown speaker to a database of known speaker .Thesystem can recognize the speaker, which has been trained with a number of speakers. Speaker recognition can also be divide into two methods, text-dependent and text independent methods. In text dependent method the speaker say key words or sentences having the sametext for both training and recognition trials. Whereastext independent does not rely on a specific text being spoken [18].

Different types of Modeling Techniques are used and it is given below.

I. The acoustic-phonetic approach
II. Pattern Recognition approach
III. Template based approaches
IV. Dynamic time warping
V. Statistical based approaches
VI. Learning based approaches
VII. The artificial intelligence approach
VIII. Stochastic Approach

IV. Testing Of Pattern

The engine compares the incoming digital-audio signal against a pre-recorded template i.e. training sample of the word. There are two techniques used for matching incoming audio signal. Matching whole word and matching part of the word. First technique takes much less processing than sub-word matching, but it requires that the user (or someone) prerecord every word that will be recognized - sometimes several hundred thousand words. Whole-word templates also require large amounts of storage Testing part are used for to match the input test data with feature extracted data.

Testing of pattern refers to the evaluation of the performance of a speech recognition system. It involves measuring the accuracy and efficiency of the system in recognizing spoken words and phrases.

There are several techniques used for testing the pattern in speech recognition systems, including:

Word Error Rate (WER): This is a commonly used metric to measure the accuracy of a speech recognition system. It is the percentage of words that are incorrectly recognized by the system. The lower the WER, the better the performance of the system [19].

Recognition Confidence Score: This is a measure of the system's confidence in its recognition of a particular word or phrase. The score is based on the likelihood that the recognized word or phrase is correct, and the system assigns a higher score to more confident recognitions.

Precision and Recall: These are metrics used to evaluate the performance of the system in detecting specific words or phrases. Precision measures the percentage of correctly detected instances, while recall measures the percentage of instances that were correctly detected out of all the possible instances [20].

F1 Score: This is a measure of the overall performance of the system, taking into account both precision and recall. It is calculated as the harmonic mean of precision and recall, and provides a single score that summarizes the system's performance.

Cross-validation: This is a technique used to evaluate the performance of the system on different datasets. It involves splitting the dataset into multiple subsets, and testing the system on each subset. This allows for a more robust evaluation of the system's performance, as it is tested on different data.

V. Conclusion

The technique developed in each stage of speech recognition system. Here also presented the list of technique with their properties for Feature extraction. Through this review it is found that MFCC is used widely for feature extraction of speech and GHM and HMM also SVM is found best among all modeling technique and in future it is possible to extendthis work for machine interface to operate on command or operation of system will control through speech and in like such other application.

We presented a real-time speech recognition system using SVM. The system uses MFCCs to represent speech signals and an SVM classifier to recognize different speech signals. The proposed system achieved high recognition accuracy and real-time performance, making it suitable for practical applications. The system can be extended to recognize other types of speech signals, such as words or phrases, by training the SVM classifier on a larger and more diverse dataset.

References

[1]. Chulhee Lee, Donghoon Hyun, Euisun Choi, Jinwook Go, and Chungyong Lee” Optimizing Feature Extraction for Speech Recognition” *IEEE Transactions On Speech And Audio Processing, Vol.11, No.1, January-2003.*

[2]. Belgacem Ben Mosbah Tsi, “Speech Recognition for Disabilities People”, *Thlkcoms Paris46 ruede Barrault 75013Paris.*

[3]. Boonserm Kijsirikul and Nitiwut Ussivakul, “Multiclass Support Vector Machine Using Adaptive Directed Acyclic Graph”, *IEEE2002, pp 980-985.*

[4]. Xin Dong, WuZhaohuiand Pan Yunhe, “A New Multi-Class Support Vector Machines”, *IEEE-2001, pp 1673-1676.*

[5]. Chen Junil, Jiao Licheng, “Classification Mechanism of Support Vector Machines”, *IEEE Transaction 2000.*

[6]. Ahmad A. M. Abushariah, Teddy S. Gunawan, Othman O. Khalifa “English Digits Speech Recognition System Based on Hidden Markov Models”, *International Conference on Computer and Communication Engineering (ICCCE 2010), 11-13 May 2010, Kuala Lumpur, Malaysia.*

[7]. Yu Shao, Member, IEEE, and Chip-Hong Chang, Senior Member, IEEE“Bayesian Separation with Sparsity Promotion in Perceptual Wavelet Domain for Speech Enhancement and Hybrid Speech Recognition”, *IEEE Transactions On Systems, Man, And Cybernetics—Part A: Systems And Humans, Vol. 41, No. 2,March-2011.*

[8]. Nelson Morgan,“Deep and Wide: Multiple Layers in Automatic Speech Recognition”, *IEEE Transactions On Audio, Speech, And Language Processing, VOL.20,NO. 1, JANUARY2012.*

[9]. Musfir Mohammed, Edet Bijoy K, Carrol xavier C, Yasif K A, Rahamathulla and SupriyaV, “Robust Automatic Speech Recognition System: Hmm Versus Sparse”*2012 Third International Conference on Intelligent Systems Modelling and Simulation.*

[10]. Gil HoLee, Shin Jae Kang, Chang WooHan and Nam SooKim, “Feature Enhancement Error Compensation For Noise Robust Speech Recognition”,*2012, 9TH International Conference on Systems, Signals and Devices.*

[11]. L. Mošner, M. Wu, A. Raju, S.H.K. Parthasarathi, K. Kumatani, S. Sundaram, R. Maas, and B. Hoffmeister, “Improving noise robustness of automatic speech recognition via parallel data and teacher-student learning,” *In 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6475-6479, 2019, May.

[12]. X. Liu, R. Sadeghian, and S.A. Zahorian, “A modulation feature set for robust automatic speech recognition in additive noise and reverberation,” *In 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5230-5234, 2017, March.

[13]. C.T. Do, and Y. Stylianou, “Weighting Time-Frequency Representation of Speech Using Auditory Saliency for Automatic Speech Recognition,” *In INTERSPEECH*, pp. 1591-1595, 2018.

[14]. M. K. uhne, “Handling Derivative Filterbank Features in Bounded-Marginalization-Based Missing Data Automatic Speech Recognition,” *In Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[15]. D.S. Park, Y. Zhang, Y. Jia, W. Han, C.C. Chiu, B. Li, Y. Wu, and Q.V. Le, “Improved noisy student training for automatic speech recognition,” *arXiv preprint:2005*.09629, 2020.

[16]. I.C. Yadav, S. Shahnawaz uddin, D. Govind, and G. Pradhan, “Spectral Smoothing by Variational mode Decomposition and its Effect on Noise and Pitch Robustness of ASR System,” *In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5629-5633, 2018, April.

[17]. T.H. Dat, J. Dennis, L.Y. Ren, and N.W.Z. Terence, “A comparative study of multi-channel processing methods for noisy automatic speech recogni-tion in urban environments,” In 2016 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6465-6469, 2016, March.

[18]. T. Menne, I. Sklyar, R. Schlüter, and H. Ney, “Analysis of deep clustering as pre-processing for automatic speech recognition of sparsely overlapping speech,” arXiv preprint:1905.03500, 2019.

[19]. J.L. Martin, and K. Tang, “Understanding Racial Disparities in Automatic Speech Recognition: the case of habitual “be”,” *Proc. Inter speech*, pp.626- 630, 2020.

[20]. A. Mani, S. Palaskar, N.V. Meripo, S. Konam, and F. Metze, “ASR error correction and domain adaptation using machine translation,” In ICASSP 2020-2020 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6344-6348, 2020, March.

[21]. K.C. Sim, A. Narayanan, A. Misra, A. Tripathi, G. Pundak, T.N. Sainath, P. Haghani, B. Li, and M. Bacchiani, “Domain Adaptation Using Factorized Hidden Layer for Robust Automatic Speech Recognition,” *In Interspeech*, pp. 892-896, 2018.

[22]. N. Moritz, T. Hori, and J. Le, “Streaming automatic speech recognition with the transformer model,” *In 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6074-6078, 2020, May.

[23]. H. Inaguma, Y. Gaur, L. Lu, J. Li, and Y. Gong, “Minimum latency train-ing strategies for streaming sequence-to-sequence ASR,” *In 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6064-6068, 2020, May.

[24]. W. Zhang, X. Cui, U. Finkler, B. Kingsbury, G. Saon, D. Kung, and M. Picheny, “Distributed deep learning strategies for automatic speech recognition,” *In 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5706-5710, 2019, May.