Grenze International Journal of Engineering and Technology, June Issue

S2T.ai: Self Supervised Speech Recognition System

Pradeep Kumar[1], Ajay Kumar[2], Kakoli Banerjee[3], Aparna Sharma[4], Ashish Katiyar[5] and Daksh Garg[6]
[1-6]Department of Computer Science and Engineering, Jss Academy of Technical Education Noida, India
Email: Pradeep8984@gmail.com, ajaykrverma@jssaten.ac.in, Kakoli.banerjee@jssaten.ac.in, aparna.s2k1@gmail.com, ashish.kat123@gmail.com, daksh2607garg@gmail.com

***Abstract*— Speech recognition is a computational linguistics sub-field focused on enabling computers to accurately transcribe spoken words into text. It drives digital transformation, spanning education, industry, healthcare, and emerging IoT and ML applications. Research in this field is rapidly advancing as scientists endeavor to broaden computers' abilities in processing spoken language. Feature extraction which is one of the steps in the process, transforms raw audio into machine-readable data for analysis. It is vital for machine learning and pattern recognition tasks. This paper presents a groundbreaking advancement in speech recognition with the introduction of the wav2vec 2.0 model which is a self- supervised feature extractor. Departing from conventional supervised methods, this model achieves superior performance by initially learning representations from unlabeled speech audio and subsequently fine-tuning on transcribed speech. Utilizing latent space masking and a task involving contrast, the model efficiently learns contextualized representations, demonstrating remarkable adaptability on the Libri-Light dataset. Even with minimal labeled data, wav2vec2.0 outperforms previous cutting edge semi-supervised approaches, showcasing its potential for robust speech recognition in scenarios with limited labeled data—a significant breakthrough for the broader accessibility of speech recognition technology.**

***Index Terms*— self-supervised, unsupervised, supervised, neural networks, automatic speech recognition (ASR).**

I. INTRODUCTION

Speech stands out as the most vital, prevalent, and effective means of communication for individuals to interact with one another. Humans are inherently comfortable with speech, leading individuals to prefer interacting with computers through speech rather than using rudimentary interfaces like keyboard and pointing device [1]. Speech Recognition, a sub-discipline at the intersection of computational linguistics, encompasses the development of techniques and technologies aimed at facilitating computers in recognizing and translating spoken words into textual format [1].

The quest for advancing automatic speech recognition (ASR) systems has led to the emergence of innovative methodologies, and in this context, the paradigm of Self-Supervised Speech Recognition model has gained prominence. Traditional ASR systems heavily rely on extensive labeled datasets, a luxury not readily available for numerous languages and domains. In response to this challenge, the research introduces a groundbreaking approach that capitalizes on the wav2vec 2.0 model for self-supervised learning. Unlike other models, our S2T.ai model aims to learn powerful representations directly from raw speech audio, offering a conceptually simpler yet highly effective alternative. The research not only pushes the boundaries of ASR but also addresses the critical issue of limited labeled data, presenting a compelling solution for practical speech recognition applications across

Grenze ID: 01.GIJET.10.2.257

diverse linguistic landscapes [2][3].

II. Dataset

We present a novel set of audios in spoken English designed for training speech recognition systems with minimal or no supervision. This collection is sourced from open-source audiobooks from the LibriVox project, comprising over 60,000 hours of audio. To the best of our knowledge, this constitutes the most extensive freely available corpus of speech [4]. The challenges posed by the existence of non-speech data, speech data with noise or low quality, and an unequal distribution of speakers are significant and these problems can be tackled using Libri-Light dataset [6]. Utilizing the tools offered by the Libri-Light dataset, we choose 72,000 hours of read speech from English book listings and execute multiple preprocessing steps. This includes filtering samples to eliminate readings of duplicate text and corrupted audio. Additionally, we exclude all audio in which the speaker overlaps with a sample in the Libri Speech dataset [5].

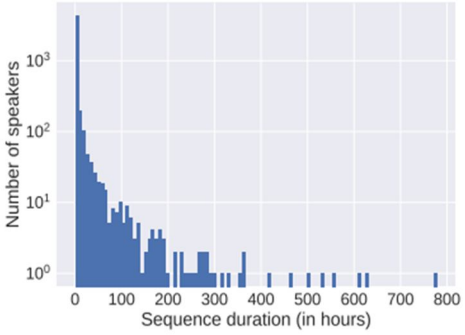

Fig 1: Duration in hours per speakers [4]

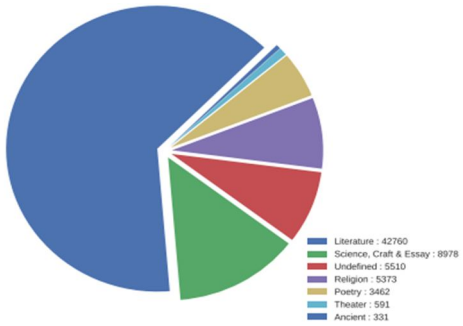

Fig 2: Durations for the 25 most frequent genres [4]

III. Related Work

Speech recognition has seen advancements in pre-training techniques for neural networks, particularly beneficial in scenarios with limited labeled data. These techniques leverage general representations to enhance performance on tasks with scarce data, focusing on tasks such as emotion recognition, speaker identification, and phoneme discrimination. However, despite efforts in unsupervised learning for speech, these representations have yet to be fully exploited for enhancing supervised speech recognition [3].

Innovations beyond frame-wise phoneme classification have emerged, with researchers leveraging learned representations to bolster robust supervised ASR systems. Notably, Wav2vec introduces a fully convolutional architecture, enabling efficient parallelization on modern hardware and deviating from previous recurrent models [3].

Additionally, the Omni-Temporal Classification (OTC) method offers a solution for ASR training systems with data that is weakly supervised, like transcripts that are not verbatim. By enabling the model to learn the mapping between speech and text alignment within the WFST framework, OTC effectively handles errors in transcripts. Results through experiment on datasets like LibriSpeech and LibriVox demonstrate that OTC-trained models maintain reasonable ASR performance even with transcripts containing maximum of 70.0% errors, significantly reducing human effort in data preparation for ASR system training [16].

Furthermore, the introduction of w2v-BERT for self-supervised speech representation learning presents a promising approach. Consisting of a contrastive module aimed at discretizing continuous speech and a masked prediction module dedicated to masked language modeling, w2v-BERT optimizes both modules jointly. Pre-trained on 60k hours of unlabeled speech data sourced from the Libri-Light corpus, w2v-BERT demonstrates superiority or parity with cutting-edge systems like HuBERT, w2v-Conformer, and wav2vec 2.0. This performance improvement extends to a more demanding internal dataset, highlighting the potential of self-supervised speech representation learning in advancing ASR systems [17].

IV. Applications Of Speech Recognition

Speech recognition applications offer developers extensive possibilities, providing users with more options without the constraints of limited keypads. Users no longer need to memorize complex numerical figures, enabling faster and more efficient interactions [8]:

A. People With Disabilities

Speech recognition's most significant impact lies in aiding individuals with disabilities. Recent advancements have tailored technology to meet their specific needs. Blind individuals were early beneficiaries, gaining valuable support from speech recognition. It also benefits those with hand-use challenges, repetitive stress injuries, and other disabilities, offering alternative input methods for computer access [8].

B. Healthcare

Speech recognition technology revolutionizes modern healthcare, streamlining information capture and enhancing doctor-patient interactions. By enabling spoken-word interaction with electronic health records and simplifying dictation tasks, it promises significant benefits. Although still evolving, it shows great potential in fields like radiology and pathology, exemplified by solutions like Dragon Naturally Speaking Medical. Future advancements may integrate medical path technology and coding databases for further healthcare optimization [8].

C. Military

Speech recognition technology is applied in military operations to improve communication, control systems, and operational efficiency. It enables personnel to verbally command and control systems in command centers, assists pilots in managing aircraft systems and communication, and facilitates real-time language translation for improved communication with individuals speaking different languages.[7]

D. Smart Home Devices

Voice-activated smart home devices use speech recognition for controlling lights, thermostats, and other connected appliances.

E. Finance

Speech recognition enhances accessibility for banking customers by enabling voice-activated banking services and transactions. Additionally, analysts utilize speech recognition for hands-free data analysis and reporting.
These applications demonstrate the diverse and evolving role of speech recognition technology in improving efficiency, accessibility, and user experiences across various sectors.

V. Challenges Faced By Speech Recognition Models

A. Noisy Environment

Research has established that a primary challenge faced by the majority of Speech Recognition Systems is the impact of environmental noise, negatively affecting system performance. Extracting features during the conversion of speech to on-screen text becomes particularly challenging due to this drawback.

B. Intensive Use of Computer Power

The execution of statistical models essential for speech recognition necessitates substantial computational effort from the computer's processor. This demand arises partly from the requirement to retain information at each stage of the word recognition search. This retention is essential in case the system needs to backtrack to accurately identify the right word.

C. Accent

Speaking accents vary based on social and personal contexts, encompassing physiological and cultural aspects. In comparison to recognizing native speech, the performance of speech recognition systems tends to decline when dealing with accented or non-native speech. Research has indicated that individuals exhibit variations in accent when speaking to their parents as opposed to when interacting with friends.

D. Speed of Speech

Speech recognition systems encounter challenges in distinguishing segments of rapidly spoken continuous speech signals. The rate of speech can vary, influenced by different situations and physical stress. This variation in speaking pace may impact pronunciation, leading to phoneme reduction, as well as time expansions and compressions.

E. Recognition of Punctuation Marks

Observations reveal that during the conversion of speech to on-screen text, punctuation marks are often not recognized in their proper form; instead, they are identified as regular words. Numerous strategies have been

devised to address this challenge, including adjusting the dictation speed for these punctuation marks. However, a more effective solution is still in the process of development.

F. Homophones

Homophones, words that sound the same but have different meanings (e.g., "There" and "Their," "Be" and "Bee"), pose a significant challenge in Speech Recognition Systems. Distinguishing the correct intended word at the word level becomes particularly difficult. Current observations indicate that Speech Recognition Systems typically achieve an accuracy ranging from 94 to 99 percent, with variations attributed to different accents.
These are some challenges faced by speech recognition models [9].

VI. Basic Techniques Used In Speech Recognition Models

Speech recognition involves converting spoken language into text, and it relies on various techniques to achieve accurate and efficient results. The basic techniques of speech recognition can be categorized into several key steps, including analysis, feature extraction, preprocessing, and classification:

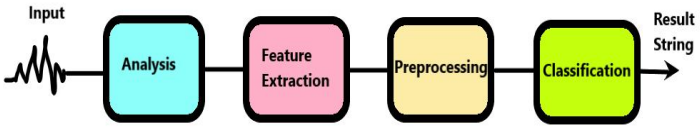

Fig 3: Process of Speech Recognition

A. Speech Signal Analysis:

The first step in speech recognition is the analysis of the input speech signal. The speech signal is sampled and divided into small frames, typically ranging from 20 to 30 milliseconds. The purpose is of breaking the signal into frames helps in capturing the dynamic nature of speech, as speech signals change over time.

B. Feature Extraction:

In speech recognition, feature extraction is a crucial step that involves transforming the raw audio signal into a set of features that can be used for further analysis and processing. The goal is to capture the relevant information from the speech signal in a form that is suitable for a machine learning model or other pattern recognition algorithms.
Various domains possess associated feature extractors widely utilized, including ResNet, BERT, and GPT-x. Typically, above models undergo pre-training on extensive sets of not labeled data through self-supervision, proving effective for downstream tasks. In the domain of speech, wav2vec 2.0 is demonstrating its robust representation capabilities and the viability of ultra-low resource speech recognition, particularly evident on the Librispeech corpus, which falls within the audiobook domain [2].
Here is a comparison with previously employed feature extractors and an explanation for the adoption of self-supervised feature extractors:

a. Supervised Pre Training

Supervised pre training used labeled data and to achieve satisfactory performance, supervised feature extraction demands extensive training with thousands of hours of speech with transcriptions. However, obtaining such a vast amount of labeled data is impractical for the vast majority of the approximately 7,000 languages spoken worldwide [10].

b. Unsupervised Pre Training

Recently, the practice of pre-training neural networks has proven to be a valuable technique, particularly in scenarios where labeled data is limited. The fundamental concept involves acquiring general representations in an environment where there is an abundance of labeled or unlabeled data. These learned representations are then harnessed to enhance performance in a subsequent task, especially when the available data for that specific task is constrained.
Utilizing unsupervised pre-training to enhance supervised speech recognition involves leveraging unlabeled audio data, which is more easily obtainable compared to labeled data. Wav2vec, is a convolutional neural network designed to process raw audio inputs. It computes a comprehensive general representation that can be served to a speech recognition system as input [3].

c. Self-Supervised pre training

Self-supervised learning has become a prominent approach for deriving generalized data representations from unlabeled samples, followed by fine-tuning the model with labeled data [10]. Given its strong association with phoneme units, self-supervised learning in automatic speech recognition shows promising potential to significantly improve speech recognition performance [12]. Wav2vec 2.0 introduces a self-supervised learning framework to obtain speech representations by masking latent features in the raw waveform and addressing a contrastive task involving quantized speech representations. Our experiments highlight the considerable promise of pre-training on unlabeled data in speech processing. Remarkably, with only 10 minutes of labeled training data or 48 recordings averaging 12.5 seconds each, our model achieves state-of-the-art results on the complete Librispeech benchmark for noisy speech. In the clean 100-hour Librispeech setup, wav2vec 2.0 not only surpasses the previous best result but also accomplishes this with just 100 times less labeled data.

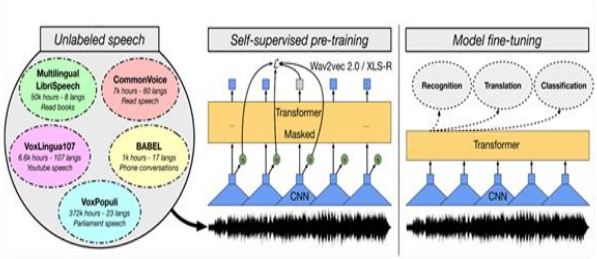

Fig 4: Working of self-supervised pre training [10]

Notably, this approach remains effective even when ample labeled data is available [10].

C. Preprocessing:

Preprocessing in speech recognition involves the application of various techniques to the raw audio data before it is fed into a speech recognition system. These preprocessing steps are essential to enhance the quality of the input and enhance the overall system performance. Some common preprocessing steps in speech recognition include:

- Sampling
- Normalization
- Noise Reduction
- Filtering
- Segmentation
- Windowing

D. Classification:

The classification step in the speech recognition process is a pivotal stage where the features are extracted from the preprocessed audio data are assigned to specific linguistic units, such as phonemes, words, or phrases. This step involves training a machine learning model, often a classifier, to distinguish between different classes of speech patterns based on the extracted features.

Several studies have been conducted to identify the optimal classifier capable of accurately recognizing speech segments across diverse conditions [11].

a. Acoustic Phonetic Approach

In the Acoustic Phonetic approach to speech recognition, the foundation lies in the identification of speech sounds and assigning them appropriate labels. This approach hypothesizes the existence of distinct, fixed phonetic units known as phonemes. These phonemes are thought to be characterized by specific acoustic properties present in speech. Despite the considerable variability in the acoustic characteristics of phonetic units, both across speakers and in proximity to other sounds, the acoustic-phonetic approach assumes that the rules governing this variability are straightforward and can be easily comprehended by a machine [1].

b. Hidden Markov Models

The Hidden Markov Model (HMM) stands out as the most successful and widely employed approach for the classification stage in Automatic Speech Recognition (ASR) systems. The prevalence of HMMs can be primarily attributed to their capability to model the temporal variation in speech signals effectively. Furthermore, HMMs are built on a versatile model that is easy to adjust based on the desired structure. Both the training methods and

the recognition process with HMMs are straightforward to perform. As a result, HMMs offer an efficient approach that is highly practical to implement in ASR systems [11].

c. Artificial Neural Networks

Artificial Neural Networks (ANNs) represent the second most commonly employed method at the classification stage of an Automatic Speech Recognition (ASR) system. They are utilized either independently or in combination with Hidden Markov Models (HMMs), with the latter approach being extensively adopted. Combining ANNs with HMMs allows for leveraging the strengths of both technologies, resulting in a more comprehensive and effective approach to ASR classification [11].
Five ANN architectures are widely used which include

- Multilayer Perceptron
- Self-organizing maps
- Radial basis function
- Recurrent neural network
- Fuzzy neural network

VII. Results

The evaluation of our Self-Supervised Speech Recognition (S2T.AI) model, particularly leveraging the wav2vec 2.0 architecture, demonstrates promising outcomes across various scenarios:

- Achieved a Word Error Rate (WER) of 1.8/3.3 on the clean/other test sets using all labeled data, showcasing high accuracy in standard speech recognition benchmarks.
- Outperformed previous state-of-the-art models on a 100-hour subset with only one hour of labeled data, showcasing the model's efficiency in scenarios with limited annotated speech data.
- Impressively achieved a WER of 4.8/8.2 using just ten minutes of labeled data and pre-training on an extensive 53,000 hours of unlabeled data. This underscores the model's adaptability to ultra-low resource conditions.
- Surpassed the performance of the best semi-supervised methods, demonstrating that our self-supervised approach, starting with unsupervised pre-training and fine-tuning on transcribed data, outperforms existing paradigms.

These results affirm the efficacy of the proposed S2T.AI model, especially in scenarios where labeled data is scarce. The wav2vec 2.0 architecture, with its masked latent space and contrastive learning, proves instrumental in achieving state-of-the-art performance in various speech recognition benchmarks. The adaptability to limited labeled data positions our approach as a promising avenue for advancing speech recognition in resource-constrained settings.

VIII Conclusion

In conclusion, our research introduces a groundbreaking approach to Speech Recognition through the lens of Self-Supervised Learning, specifically employing the wav2vec 2.0 model. The results obtained underscore the potential of self-supervised methods in mitigating the challenges associated with the scarcity of labeled speech data. The comparison with traditional semi-supervised methods emphasizes the conceptual simplicity and efficacy of our approach. Leveraging unsupervised pre-training and subsequent fine-tuning on transcribed data, our model learns powerful representations, demonstrating the feasibility of speech recognition with limited labeled data. However, it is essential to acknowledge limitations, and further research is needed to explore the generalization of the model across languages and real-world spoken scenarios. Despite these challenges, our S2T.AI model, driven by the wav2vec 2.0 architecture, opens new avenues for advancing speech recognition technology, especially in resource-constrained environments. As speech recognition continues to be a critical technology in diverse applications, our research lays a foundation for future advancements in building robust, efficient, and scalable systems that can cater to a wide range of linguistic and environmental constraints

Acknowledgement

Producing a research paper is a collaborative effort, and we owe a debt of gratitude to those who have supported and guided us along the way. We are especially thankful to our supervisor, Dr. Pradeep Kumar, for his invaluable assistance in shaping the methodology of this project. Without their guidance, cooperation, and encouragement, our progress would not have been possible.

We express our appreciation to J.S.S ACADEMY OF TECHNICAL EDUCATION, NOIDA for providing us with this opportunity. Finally, We want to express our gratitude to our project team members, whose valuable suggestions and guidance have been instrumental at various stages of completing the research paper. Their assistance has been invaluable.

REFERENCES

[1] Survey Paper On Different Speech Recognition Algorithm: Challenges And Techniques
[2] APPLYING WAV2VEC2.0 TO SPEECH RECOGNITION IN VARIOUS LOW-RESOURCE LANGUAGES
[3] WAV2VEC: UNSUPERVISED PRE-TRAINING FOR SPEECH RECOGNITION
[4] LIBRI-LIGHT: A Benchmark For Asr With Limited Or No Supervision
[5] END-TO-END ASR: FROM SUPERVISED TO SEMI-SUPERVISED LEARNING WITH MODERN ARCHITECTURES
[6] TOWARDS UNSUPERVISED LEARNING OF SPEECH FEATURES IN THE WILD
[7] Artificial Intelligence Applications For Speech Recognition
[8] Speech Recognition Technology: Applications & Future
[9] SPEECH RECOGNITION SYSTEMS - A COMPREHENSIVE STUDY OF CONCEPTS AND MECHANISM
[10] Wav2vec 2.0: A Framework For Self-Supervised Learning Of Speech Representations
[11] Comparative Study Of Automatic Speech Recognition Techniques
[12] Why Does Self-Supervised Learning For Speech: Recognition Benefit Speaker Recognition?
[13] AUTOMATIC SPEECH RECOGNITION: A COMPREHENSIVE SURVEY
[14] Pushing The Limits Of Semi-Supervised Learning For Automatic Speech Recognition
[15] On The Limit Of English Conversational Speech Recognition.
[16] LEARNING FROM FLAWED DATA: WEAKLY SUPERVISEDAUTOMATIC SPEECH RECOGNITION
[17] W2V-BERT: COMBINING CONTRASTIVE LEARNING AND MASKED LANGUAGE MODELING FOR SELF-SUPERVISED SPEECH PRE-TRAINING