CSE_SPEECH@LT-EDI-2023:Automatic Speech Recognition: vulnerable old-aged and transgender people in Tamil

Varsha Balaji , Archana JP & B. Bharathi
Department of CSE
Sri Sivasubramaniya Nadar College of Engineering,
Tamil Nadu, India
varsha2010399@ssn.edu.in
archana2120056@ssn.edu.in
bharathib@ssn.edu.in

Abstract

The crucial technology known as automatic speech recognition (ASR) transforms spoken language into written text and has a variety of uses, including voice commands and customer support. The lives of the elderly and the disabled are considerably improved by ASR, which is essential to the digitization of daily life. The Tamil voice recognition model presented in this paper was created by CSE_SPEECH using three pre-trained models that were improved from the XLSR Wav2Vec2 model from Facebook. The Common Voice Dataset was used to train the models, and the word error rate (WER) on the data was used to assess which model performed the best. This work explains the submission made by the team CSE_SPEECH in the shared task organized by LT-EDI at ACL 2023. The proposed system achieves a word error rate of 40%.

1 Introduction

Speech recognition, commonly referred to as speech-to-text or automatic speech recognition, is a method for turning spoken language into written text. It is a crucial tool with numerous uses, including voice search and call routing on mobile phones, customer service, emotion recognition, and, most crucially, aiding the disabled. In addition to helping deaf individuals translate words into text, it can also allow physically disabled people use voice commands to conduct tasks like typing and surfing rather than needing to manually operate a computer.

Singapore, Sri Lanka, Tamil Nadu, and Puducherry all have Tamil as their official language. Out of the more than 22 scheduled languages in India, Tamil was the first language to be categorized as a classical language. It is also of the oldest languages in the world, with an apparent origin dating back more than 2000 years.

Speech recognition is accomplished by taking Tamil's linguistic characteristics into account. The speech recognition problem is handled using the natural language processing methodology. In the Speech Recognition for Vulnerable Individuals in Tamil shared challenge, the team SANBAR_CSE_SSN came in first place with a word error rate of 37%.

Older folks and those who are physically or cognitively challenged have a tendency to speak with minor dysarthria, or slurred speech, which causes inaccurate transcription of the data. The transcription of the data varies from person to person since, in Tamil-speaking areas, people from different locations talk with non-identical dialects, accents, and speeds. The ability to effectively guess what someone from another location is saying when trained with audio from that region is not present.

The authors (Bharathi et al., 2022b) offer an overview of a collaborative project centered on Tamil automated speech recognition (ASR). Using data on spontaneous Tamil speech recorded from elderly and transgender people was the joint task. This dataset was given to the participants, who were then charged with identifying and rating the speech utterances. The information was acquired from open sources like marketplaces, hospitals, and vegetable shops. The speech corpus was split into training and testing data and included utterances from men, women, and transgender people. The Word Error Rate (WER) served as the basis for the task's evaluation. Participants used transformer-based models for ASR, and this overview paper summarises the diverse outcomes obtained utilizing several transformer models that have already been trained.

This paper serves as a submission to a conference, offering insights into the field of automatic speech recognition (B et al., 2023). It provides an overview of the conference's focus on ASR, high-

LT-EDI 2023 – Third Workshop on Language Technology for Equality, Diversity and Inclusion, pages 204–208, Varna, Bulgaria, Sep 7, 2023.
https://doi.org/10.26615/978-954-452-084-7_031

lighting relevant references such as (Bharathi et al., 2022a).

In our study, spoken audio were converted into tokens using pre-trained models created specifically for the Tamil language, which were then converted back into text. We used the Amrrs/wav2vec2-large-xlsr-53-tamil1 pre-trained model.

2 Related Works

Recently, researchers have tested a few methods to cope with speech recognition in minority languages like Tamil. The usage of a Hidden Markov Model, often known as an HMM, is suggested by the authors of Voice and speech recognition in the Tamil language (Fournier-Viger et al., 2017). This method of statistical pattern matching can produce speech using a variety of states for each model. The HMM model scales effectively and decreases the length and complexity of the recognition process because it only needs positive data.

Convolutional neural networks (CNNs) are used by the authors of Speech Rate Control for Improving Elderly Speech Recognition of Smart Devices (Thamburaj et al., 2021) to produce feature vectors that are then fed into fully connected networks (FCs) for the classification of syllable transition boundaries frame by frame. In order to segment the syllables, the syllable transition probability is determined. They use a Synchronised Overlap-Add (SOLA) Algorithm to help them change the speech rate in accordance with the time-scale ratio that is being measured.

By utilizing transformer networks in the neural transducer, the authors of TransformerTransducer: End-to-End Speech Recognition with Self-Attention (Yeh et al., 2019) aim to create a model for end-to-end speech recognition. They suggest two approaches: shortened self-attention to enable streaming for transformer and minimize computational complexity, and VGGNet with causal convolution to add positional information and lower frame rate for efficient inference.

The authors (Madhavaraj and Ramakrishnan, 2017) construct two distinct recognition systems for phone recognition (PR) and for continuous speech recognition (CSR) using deep neural networks (DNN) in the Design and Development of a big vocabulary, continuous voice recognition system for Tamil. It has been demonstrated that the DNN-based triphone acoustic model produces noticeably improved outcomes in CSR and PR.

In (Lin et al., 2020), the research discusses the rising concern over cybersecurity and software industry security issues. It is necessary to make more improvements because the current methods for vulnerability detection are deemed insufficient. Machine learning and data mining techniques can be used to find patterns in the vast amount of open-source software code that is now available.(Madhavaraj and Ramakrishnan, 2017) Deep learning has the capacity to comprehend natural languages, as seen by the success of its applications in speech recognition and machine translation. Researchers in software engineering and cybersecurity have been encouraged by this to investigate deep learning and neural network-based methods for finding software vulnerabilities. The survey examines the use of neural approaches to comprehend code semantics and spot vulnerable patterns in the literature that is currently available in this field. The authors of (S and B, 2022) uses transformer model Rajaram1996/wav2vec-large-xlsr-53-tamil transformer model for recognizing the Tamil speech utterances of vulnerable individuals. In (Srinivasan et al., 2022) uses akashsivanandan/wav2vec-large-xlsr-53-tamil pre-trained model for recognizing the vulnerable individual's Tamil speech utterances. In this paper, however, we use a pre-trained XLSR model to transcript the audio.

3 Dataset Analysis

Tamil speech utterances are collected from old-aged people and transgender whose mother tongue is Tamil. The recorded speech utterances of old-aged people and transgender contain how those people communicate in primary locations like banks, hospitals, and administrative offices. The data set contains 51 Speakers of literates and illiterates. The duration of the corpus is 7 hours and 30 minutes. The speech files in the directories are in the WAV format. The sampling rate of the speech utterances is 44kHz. The speech corpus with 5.5 hours of transcribed speech will be released for the training, and 2 hours of speech data will be released for testing. Table 1. shows that detailed description of the collected speech utterances.

4 Methodology and data

The strategy for the discourse acknowledgment errand includes a few steps. At first, a different agent dataset of discourse recordings in Tamil

Speakers	Literate	Illiterate	Total
Male	4	9	13
Female	7	24	31
Transgender	3	4	7

Table 1: Detailed Description of speech corpus

(Chakravarthi and Muralidaran, 2021) is collected. The particular demonstration utilized for this errand is the Amrrs/wav2vec2-large-xlsr-53-tamil show, known for its adequacy in discourse acknowledgment. The collected dataset is at that point subject to information preprocessing strategies, counting sound cleaning, portioning and labeling, highlight extraction, normalization, and language-specific preprocessing, to improve the quality and appropriateness of the information.

Following this, the preprocessed dataset is separated into preparing, approval, and testing subsets. The Amrrs/wav2vec2-large-xlsr-53-Tamil demonstration is prepared utilizing the preparing dataset, with parameters optimized to play down misfortune and make strides in precision. The execution of the prepared demonstration is assessed utilizing the approval dataset, and the Word Error Rate (WER) is calculated to the degree of its precision.

The ultimate step includes testing the demonstration utilizing the isolated testing dataset to survey its generalization and real-world execution. Execution examination is conducted to distinguish qualities, shortcomings, and ranges for enhancement. The demonstration and preprocessing methods are iteratively refined based on the examination and input gotten

5 Model Description

The advanced voice recognition model ”Amrrs/wav2vec2-large-xlsr-53-tamil” was developed especially for the Tamil language. It makes use of the powerful ”wav2vec2” architecture, which is renowned for enabling the self-supervised learning of voice representation. The large scale allows the model to record complex acoustic patterns, which leads to greater performance.

The model name ”xlsr-53” implies that it was pre-trained on a varied dataset that included 53 languages. By utilizing shared representations between languages, this multilingual pretraining improves the model’s capability to comprehend and accurately transcribe Tamil speech.

This model (Yeo et al., 2022) stands out due to its focus on voice recognition for vulnerable people. The model has been improved to handle issues experienced by people with speech problems, non-native speakers, and other disadvantaged populations, despite the fact that the task’s specific requirements are unknown. Because of its flexibility, it is especially useful for supporting accurate and accessible speech recognition for users who would have trouble with more traditional systems.

The ”Amrrs/wav2vec2-large-xlsr-53-tamil” model, created by ”Amrrs,” provides a potent remedy for Tamil speech recognition. It offers a powerful tool for accurately transcribing speech in Tamil and enhancing communication by fusing cutting-edge architecture, multilingual pretraining, and customized training for vulnerable individuals. Due to its emphasis on inclusivity and accessibility, this model is a priceless tool for resolving speech recognition issues that vulnerable populations encounter.

(Yeo et al., 2022)The difficulty of low data availability for dysarthria severity categorization, which impedes research advancement in this area, is addressed in this work. The importance of language-specific traits has been disregarded, despite the fact that the cross-lingual approach has been utilized to overcome this problem. In response, the research suggests a multilingual classification system for the degree of dysarthria in Tamil, Korean, and English. The approach makes use of both language-specific and language-independent features. From different speech dimensions, such as voice quality, pronunciation, and prosody, 39 features are derived. The best feature set for each language is then determined using feature selection techniques. Shared features and distinctive features can be distinguished by comparing the results of feature selection across the three languages.The suggested method uses these two feature sets to automatically classify severity, taking into account the elimination of language-specific features to prevent adverse impacts on other languages. For classification, the eXtreme Gradient Boosting (XGBoost) technique is used since it can deal with missing data. For validation, two baseline experiments are carried out using the intersection and union sets of monolingual feature sets. With a 67.14 F1 score as opposed to 64.52 for the Intersection trial and 66.74 for the Union experiment, the data show that the proposed technique performs better. Furthermore, for all three languages, English, Korean, and Tamil, the suggested method performs better than monolingual classifications, with relative percentage improvements of 17.67, 2.28, and 7.79 for each language. These results highlight the significance of classifying the severity of cross-language dysarthria independently by taking into account common and language-specific traits.

6 Observation and Results

Upon initial analysis of the translations created by the Amrrs demonstrate, certain challenges are apparent. They demonstrate battles to precisely separate between the boundaries of words, regularly blending adjoining words or erroneously sectioning them. This issue postures troubles in accurately translating the expected meaning of talked sentences.

Moreover, the demonstration experiences challenges in precisely capturing pushed consonants, driving to mistakes in translation. Focused consonants play a noteworthy part in Tamil dialect articulation, and their error can result in mistakes and mistaken assumptions.

Moreover, the show faces confinements in recognizing and accurately translating certain vowel sounds, especially those that are overwhelmingly utilized as pushed consonants in Tamil. This error emerges due to the nearness of a going before vowel sound that's often undefined within the translations.

Whereas the Amrrs show illustrates an, by and large, translation precision of 40% Word Error Rate (WER), it is critical to encourage investigation of the particular effect of these mistakes on powerless people. Assessing client criticism and conducting focused appraisals with the aiming client bunch would give profitable bits of knowledge into the ease of use and adequacy of the demonstration for people with discourse impedances, hearing impedances, or cognitive inabilities.

It is basic to proceed with refining the Amrrs show, tending to the recognized challenges, and endeavoring for progressed accuracy in discourse acknowledgment for defenseless people within the Tamil dialect.

7 Conclusion

In conclusion, the assessment of the Amrrs/wav2vec2-large-xlsr-53-tamil show for discourse acknowledgment among helpless people within the Tamil dialect gives profitable bits of knowledge. The ponder uncovers challenges in word division, focused consonant acknowledgment, and separation of certain vowel sounds, coming about in translation mistakes. In spite of these confinements, the demonstrate accomplishes a Word Error Rate (WER) of 40%, demonstrating its potential in discourse acknowledgment. In any case, encourage investigate is essential to investigate the particular affect on defenseless people, counting those with discourse impedances or cognitive inabilities. This consider emphasizes the significance of tending to the one of a kind needs of powerless populaces to create comprehensive and available discourse acknowledgment frameworks. Client input and engagement with the target bunch will be instrumental in moving forward convenience and viability. By progressing discourse acknowledgment innovation in Tamil for powerless people, this investigate contributes to making more comprehensive arrangements and cultivating availability in dialect handling applications.

References

Bharathi B, Bharathi Raja Chakravarthi, Subalalitha Chinnaudayar Navaneethakrishnan, Sripirya N, Rajeswari Natarajan, Suhasini S, and Swetha Valli. 2023. Overview of the second shared task on speech recognition for vulnerable individuals in tamil. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, Varna, Bulgaria. Recent Advances in Natural Language Processing.

B Bharathi, Bharathi Raja Chakravarthi, Subalalitha Chinnaudayar Navaneethakrishnan, N Sripriya, Arunaggiri Pandian, and Swetha Valli. 2022a. Findings of the shared task on Speech Recognition for Vulnerable Individuals in Tamil. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*. Association for Computational Linguistics.

B Bharathi, Bharathi Raja Chakravarthi, Subalalitha Cn, N Sripriya, Arunaggiri Pandian, and Swetha Valli. 2022b. Findings of the shared task on speech recognition for vulnerable individuals in tamil. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 339–345.

Bharathi Raja Chakravarthi and Vigneshwaran Muralidaran. 2021. Findings of the shared task on hope speech detection for equality, diversity, and inclusion. In *Proceedings of the first workshop on language technology for equality, diversity and inclusion*, pages 61–72.

Philippe Fournier-Viger, Jerry Chun-Wei Lin, Rage Uday Kiran, Yun Sing Koh, and Rincy Thomas. 2017. A survey of sequential pattern mining. *Data Science and Pattern Recognition*, 1(1):54–77.

Guanjun Lin, Sheng Wen, Qing-Long Han, Jun Zhang, and Yang Xiang. 2020. Software vulnerability detection using deep neural networks: a survey. *Proceedings of the IEEE*, 108(10):1825–1848.

A Madhavaraj and AG Ramakrishnan. 2017. Design and development of a large vocabulary, continuous speech recognition system for tamil. In *2017 14th IEEE India Council International Conference (INDICON)*, pages 1–5. IEEE.

Suhasini S and Bharathi B. 2022. SUH_ASR@LT-EDI-ACL2022: Transformer based approach for speech recognition for vulnerable individuals in Tamil. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 177–182, Dublin, Ireland. Association for Computational Linguistics.

Dhanya Srinivasan, Bharathi B, Thenmozhi Durairaj, and Senthil Kumar B. 2022. SSNCSE_NLP@LT-EDI-ACL2022: Speech recognition for vulnerable individuals in Tamil using pre-trained XLSR models. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 317–320, Dublin, Ireland. Association for Computational Linguistics.

Kingston Pal Thamburaj, Kartheges Ponniah, Ilangkumaran Sivanathan, and Muniisvaran Kumar. 2021. An critical analysis of speech recognition of tamil and malay language through artificial neural network.

Charles D Yeh, Christopher D Richardson, and Jacob E Corn. 2019. Advances in genome editing through control of dna repair pathways. *Nature cell biology*, 21(12):1468–1478.

Eun Jung Yeo, Kwanghee Choi, Sunhee Kim, and Minhwa Chung. 2022. Cross-lingual dysarthria severity classification for english, korean, and tamil. In *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 566–574. IEEE.