Overview of the Third Shared Task on Speech Recognition for Vulnerable Individuals in Tamil

B. Bharathi[1], **Bharathi Raja Chakravarthi**[2], **N. Sripriya**[1], **Rajeswari Natarajan**[3], **S. Suhasini**[4]
[1]Sri Sivasubramaniya Nadar College of Engineering, Tamil Nadu, India
[2]School of Computer Science, University of Galway, Ireland
[3]SASTRA University, India
[4]R.M.D. Engineering College, Tamil Nadu India
bharathib@ssn.edu.in, bharathiraja.akr@gmail.com

Abstract

The overview of the shared task on speech recognition for vulnerable individuals in Tamil (LT-EDI-2024) is described in this paper. The work comes with a Tamil dataset that was gathered from elderly individuals who identify as male, female, or transgender. The audio samples were taken in public places such as marketplaces, vegetable shops, hospitals, etc. The training phase and the testing phase are when the dataset is made available. The task required of the participants was to handle audio signals using various models and techniques, and then turn in their results as transcriptions of the provided test samples. The participant's results were assessed using WER (Word Error Rate). The transformer-based approach was employed by the participants to achieve automatic voice recognition. This overview paper discusses the findings and various pre-trained transformer-based models that the participants employed.

1 Introduction

The earliest known examples of Old Tamil writing are tiny inscriptions found in Adichanallur that date between 905 and 696 BC. Of all the Indian languages, Tamil possesses the most ancient non-Sanskritic literature. The grammar of Tamil is agglutinative, meaning that noun class, number, case, verb tense, and other grammatical categories are indicated by suffixes. Unlike other Aryan languages, which use Sanskrit as their standard language, Tamil uses Tamil for both its scholarly vocabulary and its metalinguistic terminology. Together with dialects, Tamil has multiple forms: cankattami, the classical literary style based on the ancient language; centami, the modern literary and formal style; and kotuntami, the present vernacular form. (Sakuntharaj and Mahesan, 2021, 2017). There is a stylistic continuity created by these styles merging together. For instance, one may write centami using cankattami vocabulary, or one could speak kotuntami while using forms related to one of the other types. (Srinivasan and Subalalitha, 2019; Narasimhan et al., 2018). A lexical root plus one or more affixes combine to form Tamil words. Suffixes make up the bulk of affixes in Tamil. Tamil suffixes fall into two groups: derivational suffixes, which change a word's meaning or part of speech, and inflectional suffixes, which identify certain categories like person, number, mood, tense, and so on. Agglutination can lead to huge words with multiple suffixes, needing numerous words or a phrase in English. Its length and scope are infinite. Although smart technologies have come a long way, human-machine interaction is still being developed and enhanced. (Chakravarthi et al., 2020). Automatic speech recognition (ASR) is one such recent technology that has enabled voice-based user interfaces for numerous automated systems. Many elderly and transgender people are frequently unaware of the technology (Hämäläinen et al., 2015) that is made available to help people in public places like banks, hospitals, and administrative offices. Thus, communication is the only kind of media that can assist people in getting what they want. However, these ASR systems are infrequently used by the elderly, transsexuals, and others with lower levels of education. English-language voice-based interfaces are a feature of most automated systems currently in use. Elderly people and those living in rural areas prefer to speak in their native tongue. The provision of speech interfaces in the local language for help systems designed for public usage would be advantageous to all. Information regarding spontaneous speech in Tamil is gathered from transgender and elderly people who are not able to use these programs. The aim of this challenge is to find an efficient ASR model to handle the elderly person's speech corpus.

The pertinent features will first be extracted from the speech signal using an ASR system. Acoustic models will also be produced using these features that were retrieved. Ultimately, the language model

Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 133–138
March 21, 2024

assists in converting these probabilities into grammatical words. The language model uses statistics from training data to assign probabilities to words and phrases (Das et al., 2011). It is necessary to evaluate ASR systems' performance prior to deploying them in real-time applications. On large-scale automatic speech recognition (ASR) tasks, an end-to-end speech recognition system has shown promising performance, matching or surpassing that of traditional hybrid systems. Using an acoustic model, lexicon, and language model, the end-to-end system quickly transforms audio data into tag labels (Zeng et al., 2021; Pérez-Espinosa et al., 2017). In the field of end-to-end voice recognition, there exist two extensively utilized frameworks. Frame synchronous prediction separates one input frame from the other by giving each one a target label (Miao et al., 2020; Xue et al., 2021; Miao et al., 2019; Watanabe et al., 2017). Phoneme identification can also be used to assess the efficacy using different test feature vectors and model settings. The use of acoustic models for speech recognition, which are created using the sounds of younger people, may have a substantial impact on the capacity to recognize elder speech (Fukuda et al., 2020; Zeng et al., 2020; Iribe et al., 2015). There aren't many acoustic models that can handle the voice detection task. Among the acoustic models are Japanese Newspaper Article Sentences (JNAS), Japanese Newspaper Article Sentences Read Speech Corpus of the Aged (S-JNAS), and Corpus of Spontaneous Japanes (CSJ). The CSJ model only achieves the lowest WER once the older voices are adjusted, according to a comparison of all the acoustic models in the literature (Fukuda et al., 2020). Dialect adaptation is also required in order to improve recognition accuracy (Fukuda et al., 2019). Recent advances in large vocabulary continuous speech recognition (LVCSR) technologies have led to the widespread use of speech recognition systems in several fields (Xue et al., 2021). Variations in the acoustics of individual speakers are thought to be one of the primary causes of the decline in speech recognition rates. For elder speakers to use speech recognition systems trained on typical adult speech data, the acoustic discrepancies between their speech and that of an adult should be investigated and correctly adjusted. Rather, this loss can be mitigated by an acoustic model enhanced by senior speakers' utterances, as shown by a document retrieval system. Modern voice recognition technology can reach excellent recognition accuracy while speaking while reading a written text or something comparable; nevertheless, the accuracy decreases when speaking spontaneously and freely. The main reason for this issue is that the linguistic and acoustic models used in voice recognition were mostly developed using read-aloud or written language materials. However, there are significant linguistic and auditory differences between written language and spontaneous speech(Zeng et al., 2020). Currently, it is becoming more and more popular to create ASR systems that can detect voice data from older persons. The aging population in modern society and the proliferation of smart devices, which make information freely accessible to both the young and the old, have led to a demand for improved voice recognition in smart devices (Kwon et al., 2016; Vacher et al., 2015; Hossain et al., 2017; Teixeira et al., 2014). Because of the influences of speech articulation and speaking style, speech recognition systems are often optimized for the voice of an average adult and have a lower accuracy rate when recognising the voice of an elderly person. It will surely become more expensive to adapt the current voice recognition systems to handle the speech of elderly users(Kwon et al., 2016).

2 Related Work

When a model is fine-tuned on many languages at the same time, a single multilingual speech recognition model can be built that can compete with models that are fine-tuned on individual language speech corpus. Speech2Vec expands the text-based Word2Vec model to learn word embeddings directly from speech by combining an RNN Encoder-Decoder framework with skipgrams or cbow for training. Acoustic models are designed at the phoneme/syllable level to carry out the speech recognition task. Initially, the acoustic models were created with JNAS, S-JNAS and CSJ speech corpus(Lin and Yu, 2015; Iribe et al., 2015). Later, the models were trained/fine-tuned with different speech corpus. To get a better performance and accuracy, backpropagation using transfer learning was attempted in the literature. Similar work was performed for other languages like Bengali, Japanese, etc. Also, more speech corpus is collected from young people for many languages(Zeng et al., 2020; Lee et al., 2021). However, speaker fluctuation, environmental noise, and transmission channel noise all degrade ASR performance. As the shared task is given with a separate training data set, an effective model has to be created during the training. Therefore, the hierarchical transformer-based model for large context end-to-end ASR can be used (Masumura et al., 2021). In the recent era, the environment is changing with smart systems and is identified that there is a need for ASR systems that are capable of handling the speech of elderly people spoken in their native languages. To overcome this problem, the shared task is proposed for the research community to build an efficient model for recognizing the speech of elderly people and transgenders in Tamil language. Findings of the automatic speech recognition for vulnerable individuals are given in (S and B, 2022) (B et al., 2022)("S and B, "2023") (Bharathi et al., 2023), have used transformer models used for transformer-based ASR for Vulnerable Individuals in Tamil.

3 Data-set Description

The dataset given to this shared task (Bharathi et al., 2022) is an Tamil conversational speech recorded from the elderly people whose average age is around 61 for male, 59 for female and 30 for transgender people which are tabulated in Table 1 . A total of 7.5 hours is collected from the elderly people. 46 audio files were recorded and each audio file is split into many subsets as transformer model does not support the large audio files. The speech is recorded with a sampling rate of 16KHZ. The audio files from Audio - 1 to Audio - 36 are used for training (duration is approximately 5.5 hours) and Audio - 37 to Audio - 48 are used for testing (duration is approximately 2 hours).

4 Methodology

The methodology used by the participants in the shared task of speech recognition for vulnerable individuals in Tamil is discussed in this section. Three teams submitted their runs for this task. All three teams have used the pre-trained models. The first team "CEN_Amrita" has used the whisper model, Whisper is a pre-trained automatic speech recognition (ASR) model trained on 680,000 hours of multilingual and multitask supervised data sourced from the web. This end-to-end transformer-based model adopts the encoder- decoder architecture.. The second team "ASR_Tamil_SSN" have used the transformer based model called 'akashsivanandan/wav2vec2-large-xls-r-300m-tamil-colab-final'. The third team " have also used the transformer based pretrained model called 'Rajaram1996/wav2vec2-large- xlsr-53-tamil'.

5 Evaluation of Results

The results submitted by the participants are evaluated based on the WER computed between the ASR hypotheses submitted by the participants and the ground truth of human speech transcription.

$$\text{WER (Word Error Rate)} = (S + D + I) / N$$

where,
S = No. of substitutions
D = No. of deletions
I = No. of insertions
N = No. of words in the reference transcription

As discussed in the methodology, different average word error rates are measured using various pre-trained transformer-based models. The participating team's WER are shown in Table. 2.

6 Conclusions

The shared challenge for vulnerable voice recognition in Tamil is covered in this overview paper. The speech corpus shared for this job was recorded from elderly persons. Getting older people's speech more accurately recognised is a difficult endeavor. In order to boost the accuracy and performance in recognising elderly people's speech, the participants have been given access to the gathered speech corpus. There were a total of seven teams participated in this joint task and turned in their transcripts of the supplied data. The team estimated the WER and then compared the outcome to the human transcripts. Three teams built their recognition systems using various Whisper models and transformer-based models. Finally, the word error rates of the three participants are 24.452, 29.297, 37.7333 respectively. Based on the observations, it is suggested that the transformer-based model and whisper model can be trained with given speech corpus which could give better accuracy than the pre-trained model, as the transformer-based model and whisper model used are trained with a common voice dataset. Also, a separate language model can also be created for this corpus.

S.No	Filename	Gender	Age	Duration(in min)
1	Audio - 1	M	72	10
2	Audio - 2	F	61	9
3	Audio - 3	F	71	11
4	Audio - 4	M	68	8
5	Audio - 5	F	59	14
6	Audio - 6	F	67	9
7	Audio - 7	M	54	8
8	Audio - 8	F	65	16
9	Audio - 9	F	55	3
10	Audio - 10	M	60	13
11	Audio - 11	F	55	17
12	Audio - 12	F	52	6
13	Audio - 13	F	53	11
14	Audio - 14	F	61	9
15	Audio - 15	F	54	1
16	Audio - 16	F	56	6
17	Audio - 17	F	52	12
18	Audio - 18	F	54	6
19	Audio - 19	F	52	8
20	Audio - 20	F	52	9
21	Audio - 21	F	62	13
22	Audio - 22	F	52	12
23	Audio - 23	F	62	13
24	Audio - 24	F	53	4
25	Audio - 25	F	65	3
26	Audio - 26	F	64	8
27	Audio - 27	F	54	6
28	Audio - 28	M	62	8
29	Audio - 29	M	54	16
30	Audio - 30	F	76	9
31	Audio - 31	F	55	9
32	Audio - 32	M	50	6
33	Audio - 33	F	63	6
34	Audio - 34	M	84	6
35	Audio - 35	F	70	6
36	Audio - 36	F	50	6
37	Audio - 37	M	53	6
38	Audio - 38	F	55	6
39	Audio - 39	M	62	6
40	Audio - 40	T	24	6
41	Audio - 41	T	22	7
42	Audio - 42	T	40	8
43	Audio - 43	T	25	11
44	Audio - 44	T	29	10
45	Audio - 45	T	35	9
46	Audio - 46	T	33	16
47	Audio - 47	F	20	5
48	Audio - 48	M	37	5

Table 1: Age, gender, and duration of the utterances of the speech corpus

S. No	Team Name	WER (in %)
1	CEN_Amrita (Jairam R, 2024)	24.452
2	ASR_TAMIL_SSN (Suhasini and Bharathi, 2024)	29.297
3	DRAVIDIAN LANGUAGE - Abirami Jayaraman (Abirami. J, 2024)	37.733

Table 2: Results of the participating system's Word Error Rate

References

Dharunika Sasikumar B. Bharathi Abirami. J, Aruna Devi. S. 2024. Dravidian language@ lt-edi 2024:pre-trained transformer based automatic speech recognition system for elderly people. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion (LT-EDI-2024)*.

Bharathi B, Dhanya Srinivasan, Josephine Varsha, Thenmozhi Durairaj, and Senthil Kumar B. 2022. SSNCSE_NLP@LT-EDI-ACL2022:hope speech detection for equality, diversity and inclusion using sentence transformers. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 218–222, Dublin, Ireland. Association for Computational Linguistics.

B Bharathi, Bharathi Raja Chakravarthi, Subalalitha Chinnaudayar Navaneethakrishnan, N Sripriya, Arunaggiri Pandian, and Swetha Valli. 2022. 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, Sripriya Natarajan, Rajeswari Natarajan, S Suhasini, 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 Language Technology for Equality, Diversity and Inclusion*, pages 31–37.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Vigneshwaran Muralidaran, Shardul Suryawanshi, Navya Jose, Elizabeth Sherly, and John P McCrae. 2020. Overview of the track on sentiment analysis for Dravidian languages in code-mixed text. In *Forum for Information Retrieval Evaluation*, pages 21–24.

Biswajit Das, Sandipan Mandal, and Pabitra Mitra. 2011. Bengali speech corpus for continuous auutomatic speech recognition system. In *2011 International conference on speech database and assessments (Oriental COCOSDA)*, pages 51–55. IEEE.

Meiko Fukuda, Ryota Nishimura, Hiromitsu Nishizaki, Yurie Iribe, and Norihide Kitaoka. 2019. A new corpus of elderly japanese speech for acoustic modeling, and a preliminary investigation of dialect-dependent speech recognition. In *2019 22nd Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, pages 1–6. IEEE.

Meiko Fukuda, Hiromitsu Nishizaki, Yurie Iribe, Ryota Nishimura, and Norihide Kitaoka. 2020. Improving speech recognition for the elderly: A new corpus of elderly japanese speech and investigation of acoustic modeling for speech recognition. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6578–6585.

Annika Hämäläinen, António Teixeira, Nuno Almeida, Hugo Meinedo, Tibor Fegyó, and Miguel Sales Dias. 2015. Multilingual speech recognition for the elderly: The aalfred personal life assistant. *Procedia Computer Science*, 67:283–292.

M Shamim Hossain, Md Abdur Rahman, and Ghulam Muhammad. 2017. Cyber–physical cloud-oriented multi-sensory smart home framework for elderly people: An energy efficiency perspective. *Journal of Parallel and Distributed Computing*, 103:11–21.

Yurie Iribe, Norihide Kitaoka, and Shuhei Segawa. 2015. Development of new speech corpus for elderly japanese speech recognition. In *2015 International Conference Oriental COCOSDA held jointly with 2015 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)*, pages 27–31. IEEE.

Premjith B Viswa M Jairam R, Jyothish Lal G. 2024. Cen_amrita@lt-edi-eacl2024 - a transformer based speech recognition system for vulnerable individuals in tamil. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion (LT-EDI-2024)*.

Soonil Kwon, Sung-Jae Kim, and Joon Yeon Choeh. 2016. Preprocessing for elderly speech recognition of smart devices. *Computer Speech & Language*, 36:110–121.

Taewoo Lee, Min-Joong Lee, Tae Gyoon Kang, Seokyeoung Jung, Minseok Kwon, Yeona Hong, Jungin Lee, Kyoung-Gu Woo, Ho-Gyeong Kim, Jiseung Jeong, et al. 2021. Adaptable multi-domain language model for transformer asr. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7358–7362. IEEE.

Hui Lin and Yibiao Yu. 2015. Acoustic feature analysis and conversion of age speech. In *IET Conference Proceedings*. The Institution of Engineering & Technology.

Ryo Masumura, Naoki Makishima, Mana Ihori, Akihiko Takashima, Tomohiro Tanaka, and Shota Orihashi.

2021. Hierarchical transformer-based large-context end-to-end asr with large-context knowledge distillation. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5879–5883. IEEE.

Haoran Miao, Gaofeng Cheng, Changfeng Gao, Pengyuan Zhang, and Yonghong Yan. 2020. Transformer-based online ctc/attention end-to-end speech recognition architecture. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6084–6088. IEEE.

Haoran Miao, Gaofeng Cheng, Pengyuan Zhang, Ta Li, and Yonghong Yan. 2019. Online hybrid ctc/attention architecture for end-to-end speech recognition. In *Interspeech*, pages 2623–2627.

Anitha Narasimhan, Aarthy Anandan, Madhan Karky, and CN Subalalitha. 2018. Porul: Option generation and selection and scoring algorithms for a tamil flash card game. *International Journal of Cognitive and Language Sciences*, 12(2):225–228.

Humberto Pérez-Espinosa, Juan Martínez-Miranda, Ismael Espinosa-Curiel, Josefina Rodríguez-Jacobo, and Himer Avila-George. 2017. Using acoustic paralinguistic information to assess the interaction quality in speech-based systems for elderly users. *International Journal of Human-Computer Studies*, 98:1–13.

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.

Suhasini "S and Bharathi" B. "2023". "asr_ssn_cse 2023@lt-edi-2023: Pretrained transformer based automatic speech recognition system for elderly people". In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, "Varna, Bulgaria ". "Recent Advances in Natural Language Processing".

Ratnasingam Sakuntharaj and Sinnathamby Mahesan. 2017. Use of a novel hash-table for speeding-up suggestions for misspelt Tamil words. In *2017 IEEE International Conference on Industrial and Information Systems (ICIIS)*, pages 1–5.

Ratnasingam Sakuntharaj and Sinnathamby Mahesan. 2021. Missing word detection and correction based on context of Tamil sentences using n-grams. In *2021 10th International Conference on Information and Automation for Sustainability (ICIAfS)*, pages 42–47.

R Srinivasan and CN Subalalitha. 2019. Automated named entity recognition from tamil documents. In *2019 IEEE 1st International Conference on Energy, Systems and Information Processing (ICESIP)*, pages 1–5. IEEE.

S Suhasini and B Bharathi. 2024. Asr_tamil_ssn@ lt-edi-2024: Automatic speech recognition system for elderly people. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion (LT-EDI-2024)*.

António Teixeira, Annika Hämäläinen, Jairo Avelar, Nuno Almeida, Géza Németh, Tibor Fegyó, Csaba Zainkó, Tamás Csapó, Bálint Tóth, André Oliveira, et al. 2014. Speech-centric multimodal interaction for easy-to-access online services–a personal life assistant for the elderly. *Procedia computer science*, 27:389–397.

Michel Vacher, Frédéric Aman, Solange Rossato, and François Portet. 2015. Development of automatic speech recognition techniques for elderly home support: Applications and challenges. In *International Conference on Human Aspects of IT for the Aged Population*, pages 341–353. Springer.

Shinji Watanabe, Takaaki Hori, Suyoun Kim, John R Hershey, and Tomoki Hayashi. 2017. Hybrid ctc/attention architecture for end-to-end speech recognition. *IEEE Journal of Selected Topics in Signal Processing*, 11(8):1240–1253.

Jiabin Xue, Tieran Zheng, and Jiqing Han. 2021. Exploring attention mechanisms based on summary information for end-to-end automatic speech recognition. *Neurocomputing*, 465:514–524.

Jiazhong Zeng, Jianxin Peng, and Yuezhe Zhao. 2020. Comparison of speech intelligibility of elderly aged 60–69 years and young adults in the noisy and reverberant environment. *Applied Acoustics*, 159:107096.

Zhiping Zeng, Haihua Xu, Yerbolat Khassanov, Eng Siong Chng, Chongjia Ni, Bin Ma, et al. 2021. Leveraging text data using hybrid transformer-lstm based end-to-end asr in transfer learning. In *2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pages 1–5. IEEE.