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

B. Bharathi[1], **Bharathi Raja Chakravarthi**[2],
Subalalitha C N[3], **N. Sripriya**[1], **Rajeswari Natarajan**[4], **S. Suhasini**[1], **Swetha Valli**[5]
[1]SSN College of Engineering
[2]National University of Ireland Galway
[3]SRM Institute Of Science And Technology
[4]Thiagarajar College of Engineering
[5]SASTRA University, India
bharathib@ssn.edu.in, bharathiraja.akr@gmail.com

Abstract

This paper manifests the overview of the shared task on Speech Recognition for Vulnerable individuals in Tamil(LT-EDI-ACL2023). The task is provided with a Tamil dataset, which is collected from elderly people of three different genders, male, female, and transgender. The audio samples were recorded from public locations like hospitals, markets, vegetable shops, etc. The dataset is released in two phases, the training and the testing phase. The participants were asked to use different models and methods to handle audio signals and submit the result as transcription of the test samples given. The result submitted by the participants was evaluated using WER (Word Error Rate). The participants used the transformer-based model for automatic speech recognition. The results and different pre-trained transformer-based models used by the participants are discussed in this overview paper.

1 Introduction

The earliest Old Tamil documents are small inscriptions in Adichanallur dating from 905 BC to 696 BC. Tamil has the oldest ancient non-Sanskritic Indian literature of any Indian language. Tamil uses agglutinative grammar, which uses suffixes to indicate noun class, number, case, verb tense, and other grammatical categories. Tamil's standard metalinguistic terminology and scholarly vocabulary is itself Tamil, as opposed to the Sanskrit that is standard for most Aryan languages. Tamil has many forms, in addition to dialects: a classical literary style based on the ancient language (cankattami), a modern literary and formal style (centami), and a current colloquial form (kotuntami) (Sakuntharaj and Mahesan, 2021, 2017). These styles blend into one another, creating a stylistic continuity. It is conceivable, for example, to write centami using cankattami vocabulary, or to utilize forms connected with one of the other varieties while speaking kotuntami (Srinivasan and Subalalitha, 2019; Narasimhan et al., 2018). Tamil words are made up of a lexical root and one or more affixes. The majority of Tamil affixes are suffixes. Tamil suffixes are either derivational suffixes, which modify the part of speech or meaning of the word, or inflectional suffixes, which designate categories like as person, number, mood, tense, and so on. There is no ultimate limit to the length and scope of agglutination, which might result in large words with several suffixes, requiring many words or a sentence in English. Smart technologies have advanced significantly, and they are still developing and improving human-machine connection(Chakravarthi et al., 2020). One such modern technology is automatic speech recognition (ASR), which has made it possible for many automated systems to have voice-based user interfaces. The technology that are facilitated to assist individuals (Hämäläinen et al., 2015) in public spaces like banks, hospitals, and administrative offices are often unknown to many elderly and transgender persons. Therefore, the only media that could help them meet their demands is communication. However, the elderly, transsexual, and less educated persons rarely use these ASR systems. The majority of the automated systems in use today include voice-based interfaces that are available in English. People in rural areas and the elderly prefer to communicate in their own language. All people would benefit if the assistance systems created for use in public spaces could be equipped with speech interfaces in the local tongue. The data on spontaneous speech in Tamil is collected from elderly and transgender individuals who are deprived of the opportunity to take benefit of these services.Finding an effective ASR model to handle the elderly persons speech corpus is the goal of this task. The representation of how the audio samples are collected is shown in

LT-EDI 2023 – Third Workshop on Language Technology for Equality, Diversity and Inclusion,
pages 31–37, Varna, Bulgaria, Sep 7, 2023.
https://doi.org/10.26615/978-954-452-084-7_005

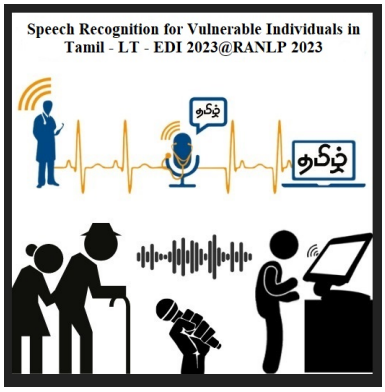

Figure 1: Speech corpus collected from vulnerable individuals in Tamil language

Fig:1

An ASR system will initially extract the relevant features from speech signal. These extracted features will also be used to generate acoustic models. Finally, the language model helps to turn these probabilities into words of coherent language. The language model, assigns probabilities to words and phrases based on statistics from training data(Das et al., 2011). Before using ASR systems in real-time applications, their performance must be assessed. An end-to-end speech recognition system has demonstrated promising performance on large-scale automatic speech recognition (ASR) tasks, placing it on par with conventional hybrid systems. The end-to-end system converts acoustic data into tag labels instantly using an acoustic model, lexicon, and language model(Zeng et al., 2021; Pérez-Espinosa et al., 2017). There are two widely used frameworks in the area of end-to-end speech recognition. Frame synchronous prediction, which assigns one target label to each input frame, distinguishes one from the other(Miao et al., 2020; Xue et al., 2021; Miao et al., 2019; Watanabe et al., 2017). With alternative test feature vectors and model settings, the effectiveness can also be evaluated in terms of phoneme recognition. The ability to recognise senior speech may be significantly influenced by the use of acoustic models for speech recognition, which are produced using the voices of younger persons(Fukuda et al., 2020; Zeng et al., 2020; Iribe et al., 2015). Few acoustic models have been developed to perform the voice recognition problem. Japanese Newspaper Article Sentences (JNAS), Japanese Newspaper Article Sentences Read Speech Corpus of the Aged (S-JNAS), and Corpus of Spontaneous Japanes (CSJ) are a few examples of the acoustic models. All of the acoustic models are compared in the literature, and it is discovered that the CSJ model only obtains the lowest WER after the adaption of the elderly voices(Fukuda et al., 2020). The same goes for dialect adaptation, which is necessary to increase recognition accuracy(Fukuda et al., 2019). Speech recognition systems are now widely used in a range of fields as a result of recent advancements in large vocabulary continuous speech recognition (LVCSR) technologies(Xue et al., 2021). One of the main reasons for the fall in speech recognition rates is assumed to be variances in the acoustics of different speakers. The acoustic differences between the speech of senior speakers and those of a typical adult should be examined and appropriately adjusted in order for older speakers to use speech recognition systems trained on normal adult speech data. Instead, as demonstrated by a document retrieval system, an acoustic model improved using utterances of senior speakers can lessen this degradation. Using cutting-edge voice recognition technology, high recognition accuracy can be achieved for speech reading a written text or anything similar; nevertheless, the accuracy declines for freely uttered spontaneous speech. The primary cause of this problem is that read speech or written language texts were predominantly used in the development of the acoustic and linguistic models used in speech recognition. However, both linguistically and acoustically, spontaneous speech and written language differ greatly(Zeng et al., 2020).

Creating ASR systems to recognise elderly people's voice data is becoming increasingly commonplace today. The need to improve voice recognition in smart devices has arisen as a result of the ageing population in contemporary society and the expansion of smart gadgets, allowing both the elderly and the younger generations to easily access information(Kwon et al., 2016; Vacher et al., 2015; Hossain et al., 2017; Teixeira et al., 2014). Speech recognition systems are frequently optimised for an average adult's voice and have a reduced accuracy rate when recognising an elderly person's voice due to the effects of speech articulation and speaking style. The cost of modifying the already existing voice recognition systems to handle the speech of older users will undoubtedly increases(Kwon et al., 2016).

2 Task Description

This shared task tackles a difficult problem in Automatic Speech Recognition: vulnerable elderly and transgender individuals in Tamil. People in their senior years go to primary places such as banks, hospitals, and administrative offices to meet their daily needs. Many elderly persons are unsure of how to use the devices provided to assist them. Similarly, because transgender persons are denied access to primary education as a result of societal discrimination, speech is the only channel via which they may meet their needs. The data on spontaneous speech is collected from elderly and transgender people who are unable to take advantage of these services. For the training set, a speech corpus containing 5.5 hours of transcribed speech will be released, as well as 2 hours of speech data for testing test. The participants have to submit the text transcriptions for the test utterances in a separate text file.

3 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 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 the 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 the 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, 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 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), have used transformer models used for transformer based ASR for Vulnerable Individuals in Tamil.

4 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 6 hours and 42 minutes 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 - 47 are used for testing (duration is approximately 2 hours).

5 Methodology

The methodology used by the participants in shared task of speech recognition for vulnerable individuals in Tamil is discussed in this section. Three teams submitted their runs for this task. Different types of pre-trained transformer models used by the participants in this shared task are as follows:

- IIT Madras transformer ASR model - It is work based on espnet.nets.pytorch backend.e2e asr transformer:E2Eself-attention mechanism[1]
- anuragshas/wav2vec2-xlsr-53-tamil [2]
- Amrrs/wav2vec2-large-xlsr-53-tamil [3]

The above mentioned second and third models are fine tuned on facebook/wav2vec-large-xlsr-53 [4] pre-trained model using multilingual common

[1] https://asr.iitm.ac.in/demo/

[2] https://huggingface.co/anuragshas/wav2vec2-xlsr-53-tamil

[3] https://huggingface.co/Amrrs/wav2vec2-large-xlsr-53-tamil

[4] https://huggingface.co/facebook/wav2vec2-large-xlsr-53

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

S.No	Filename	Gender	Age	Duration(in secs)
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

S. No	Team Name	WER (in %)
1	SANBAR_CSE_SSN (”S and B, ”2023”a)	37.7144
2	ASR_SSN_CSE_2023 (”S and B, ”2023”b)	39.8091
3	CSE_Speech (”Balaji et al., ”2023”)	40.7562

Table 2: Results of the participating systems in Word Error Rate

voice dataset. To fine-tune the model, they had a classifier representing the downstreams task's output vocabulary on top of it and train it with a Connectionist Temporal Classification (CTC) loss on the labelled data. The models used are based on XLSR wav2vec model, this XLSR model is capable of learning cross-lingual speech data, where the raw speech waveform is converted to multiple languages by pre-training a single model.

6 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.

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 rate are measured using various pre-trained transformer based models.

Performance of the ASR submitted by the participants are tabulated in Table 2. From Table 2, IIT Madras transformer ASR model is work based on espnet.nets.pytorch backend.e2e asr transformer:E2Eself- attention mechanism model produces less WER compared to other models.

7 Conclusion

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 endeavour. In order to boost the accuracy and performance in recognising the elderly people's speech, the participants have been given access to the gathered speech corpus. There were two people that 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. Both participants built their recognition systems using various transformer-based models. Finally, the word error rates of the three participants are 37.7144, 39.8091 & 40.7462 respectively. Based on the observations, it is suggested that the transformer based model can be trained with given speech corpus which could give a better accuracy than the pre-trained model, as the transformer based model used are trained with common voice dataset. Also, a separate language model can also be created for this corpus.

References

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.

Varsha ”Balaji, Archana J P, and Bharathi” B. ”2023”. ”cse_speech@lt-edi-2023:automatic speech recognition: Vulnerable old-aged and transgender people 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. 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.

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.

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.

Saranya "S and Bharathi" B. "2023"a. "sanbar@lt-edi-2023:automatic speech recognition: vulnerable old-aged and transgender people in tamil". In *"Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages"*, "Varna, Bulgaria ". "Recent Advances in Natural Language Processing".

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"b. "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.

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.