Konkani ASR

Swapnil Fadte
Discipline of Computer Science
& Technology
Goa Business School
Goa University
Taleigao 403207, India
swapnil.fadte@unigoa.ac.in

Gaurish Thakkar
Faculty of Humanities
& Social Sciences
University of Zagreb
Zagreb 10000, Croatia
gthakkar@m.ffzg.hr

Jyoti D. Pawar
Discipline of Computer Science
& Technology
Goa Business School
Goa University
Taleigao 403207, India
jdp@unigoa.ac.in

Abstract

Konkani is a resource-scarce language, mainly spoken on the west coast of India. The lack of resources directly impacts the development of language technology tools and services. Therefore, the development of digital resources is required to aid in the improvement of this situation. This paper describes the work on the Automatic Speech Recognition (ASR) System for Konkani language. We have created the ASR by fine-tuning the whisper-small ASR model with 100 hours of Konkani speech corpus data. The baseline model showed a word error rate (WER) of 17, which serves as evidence for the efficacy of the fine-tuning procedure in establishing ASR accuracy for Konkani language.

1 Introduction

Konkani belongs to the Southern Indo-Aryan language group spoken on the western coast of India. It is the official language of Goa and one of India's 22 officially recognized languages. Some people also speak Konkani in the coastal regions of Maharashtra, Karnataka, Kerala, Gujarat, and Daman & Diu. It has approximately 2.3 million speakers (Census, 2011). Devanagari is the official script for writing the language, and the Antruzi dialect spoken in Ponda Taluk is considered the standard dialect for official communication. Language is resource-scarce and lacks a lot of digital resources and tools.

In this digital era, there is a need for digital resources to be made available in all local and regional languages so that their identity is maintained. It may also help in preserving language diversities and identities. Some digital solutions require common fundamental components. The automatic speech recognition (ASR) system can be considered one of the most important digital resources for any language to survive in this digital era.

Speech is the most natural form of communication. If ASR is developed for all the native languages, it has the potential to break the barriers between technology and people. ASR has the power to make technology available to all. Some immediate beneficiaries can be a person with a vision disability, an illiterate person, or a person with reading problems. ASR solutions are well explored, and we can see many commercial solutions for resource-rich languages like English. Ironically, it is also true that ASR is not available for low-resourced languages. The main reason could be a lack of language resources like large annotated speech corpora, text corpora, etc. Here, we have attempted to develop a quick ASR for Konkani language by banking on the resources of resource-rich languages.

The paper is organized as follows: A brief introduction to Konkani language is provided in Section 1 above. A quick literature review of the ASR development is provided in Section 2. The motivation behind the current work is provided in Section 3. Properties of the dataset used in the development of ASR are provided in Section 4. The methodology used for the design of the ASR is provided in Section 5. Results and discussions based on the ASR development are provided in Section 6. Finally, the conclusion and future work are discussed in Section 7.

2 Literature Review

ASR systems have evolved from simple limited command and control word setups to today's fluent natural language processing systems (Liu et al., 2023) and Fluent End-to-end systems (Jamshid Lou and Johnson, 2020; Mhiri et al., 2020). Initial work on ASR has employed statistical modeling. The most famous was the GMM-HMM-based model (Swietojanski et al., 2013; Kumar et al., 2014; Şchiopu, 2013). This model requires vast amounts of data and takes a big team to develop the model. Later, ML models started replacing these statistical models. Still, this model takes a large amount of annotated corpus, and a good acoustic and language model. Acoustic model creation is similar to feature extraction, like GMM models. Nowadays, the large pre-trained model can be used to develop ASR solutions for low-resourced languages. This pre-trained model can be later fine-tuned to get ASR in target languages (Khare et al., 2021; Yang et al., 2023). This method, however, required annotated speech data for fine-tuning. If data for fine-tuning is not present, then it needs to be generated for a specific language.

3 Motivation of the Work

Every language holds and represents its own unique linguistic and cultural heritage. Konkani is a resource-scarce language that deserves complete preservation in this digital age. One logical step towards conservation would be to build the linguistic tools required for survival in this digital age. Here, we have attempted to develop one such tool. Through this work, we have developed an ASR system using tools and resources from resource-rich languages and available Konkani data.

4 Dataset Properties

We have used CIIL Konkani annotated speech corpus (Khandale et al., 2018). Data consists of more than 100 hours of annotated speech data. The dataset consists of ten subcategories:

- Contemporary Text (News): The text is taken from news items from 2005-2012 from websites or print newspapers. मोपाचो मोरया.कांय वर्सां पयलीं सरकारान मोपा विमानतळाची घोशणा केली...

- Creative Text: The creative text category comprises mostly six essays or short stories. एके बागेंत एक मोर रावतालो. आपुण पाखां फुलोवन नाचतकच, लोक खोशयेन नाचतात, तें पळोवन तो सामको गर्वान फुलतालो. ताणें मागीर मान 'अश्शी–अश्शी '

- Sentence: A list of isolated sentences that are not interconnected within a continuous speech. गुजरात राज्य भारताच्या पश्चिम दिशेक पडटा.

- Date: A list of sentences to capture how a date is spoken. आयज तारिख कितें?

- Command and control words: These include imperative sentences, optative sentences as well as other controlling phrases which may come as a reply to an interrogative sentence. सांगचें न्हय

- Place names: This set includes Indian place names. These include main cities, district names and popular tourist destinations from all over India. आगशी

- Person names: The names include individuals from diverse spheres such as politicians, film actors and directors, writers, monarchs, astrologers, historical figures, scientists, and athletes. आगुस्तिन

- Most frequent words: List of the regularly and repeatedly used list of words. जावप

- Phonetically balanced words: It is a list of words in which the occurrence of a phoneme in initial medial and final positions of that language can be represented. अमर

- Form and function words: The Form and Function dataset includes Grammatical function words, numerals, kinship terms, measurement terms, list of colors, days, months, seasons, directions, zodiac sings, body parts, planets etc. ना

Data is recorded from 504 Speakers from 4 geographical regions (districts): North Goa and South Goa districts in Goa, Sindhudurgh district in Maharashtra, and Uttara Kannada district in Karnataka. Table 1 shows speaker distribution and characteristics.

5 Methodology

Data Pre-processing: The CIIL speech data is already transcribed and annotated with corresponding text labels for a total of 72937 records. Resampling was conducted on all audio files, reducing the sampling rate from 48KHz to 16 kHz. All phrases that exceeded 448-byte pairs were removed through the process of filtering, as the model supports a maximum of 448-byte pairs for the text input. The data was split into a train-evaluation-test split in 80/10/10 ratio.

Table 1: Speaker distribution based on gender, location, and speech. M-Male, F-Female, UK-Uttara Kannada, SD-Sindhudurgh.

Age Group	Total Speakers	Gender		District							
				North-Goa		South-Goa		UK		SD	
		F	M	F	M	F	M	F	M	F	M
16-20	71	42	29	14	9	16	16	10	4	2	0
21-50	304	160	144	66	38	47	58	46	47	1	7
50+	129	65	64	11	9	31	27	23	28	0	0
Total	504	267	237	91	56	94	101	79	73	3	7

Whisper-Small (Radford et al., 2023), a pre-trained ASR model, is selected as the base model for fine-tuning. The model has 12 layers with 12 heads, resulting in 244 million parameters. The model is pre-trained on a vast quantity of labeled audio-transcription multilingual datasets, which include data from various languages, enabling it to learn general speech representations. Konkani language was not included in the set of languages where the whisper-small model was initially pre-trained.

The Whisper tokenizer received pre-training using transcriptions from 96 languages that were used for pre-training. As a result, it possesses a comprehensive byte pair suitable for nearly all multilingual ASR applications.

The fine-tuning process of the Whisper model involves loading the tokenizer specific to the language and utilizing it for fine-tuning without making any additional adjustments. Due to the absence of a Konkani tokenizer, the Marathi tokenizer is used. The system design can be referred to in Figure 1.

We fine-tuned the Whisper-Small model using the CIIL speech dataset. During fine-tuning, the model's weights are adjusted by utilizing Konkanispeech and text pairs while preserving the previously acquired representations from the multilingual pre-training stage. Transfer learning (Pan and Yang, 2009; Weiss et al., 2016) enables the model to effectively adjust to the distinctive phonetic and linguistic attributes inherent in Konkani language. The system was trained for three epochs, with an early stopping criterion for evaluation loss, using a V100 16 GB GPU. Hyperparameter tuning was performed using the validation set. The average training time for the system was around 5–6 days.

Evaluation Metric: The evaluation metric employed to assess the performance of the ASR system on the Konkani test set is the word error rate (WER). The WER is a metric used to determine the percentage of errors at the word level in the ASR output when compared to the ground-truth transcription. This value indicates the average number of errors per reference word. The lower the value, the better the performance of the ASR system, with a WER of 0 being a perfect score.

Table 2: Hyperparameters for Fine-Tuning Whisper-Small

Hyperparameter	**Value**
Learning rate	0.8e-5
Batch size	16
Number of epochs	3
Model dropout Rate	0.3
Warm-up steps	500

6 Results and Discussion

The fine-tuned whisper-small ASR model is evaluated on the test set. The obtained WER of 17 serves as evidence for the efficacy of the fine-tuning procedure in enhancing ASR performance for Konkani language. Additionally, we conducted experiments using the Hindi tokenizer. However, our findings indicated that utilizing the Marathi tokenizer yielded better scores than Hindi. To check the model's performance in a real-world case, we conducted manual testing using 70 newly curated sentences. Two native speakers were tasked with recording and noting down the output of the model. The number of space-separated tokens in sentence collection ranged from 3 to 14. Following are some examples from the dataset used to test the final model.

- शी! हें कितें करता?

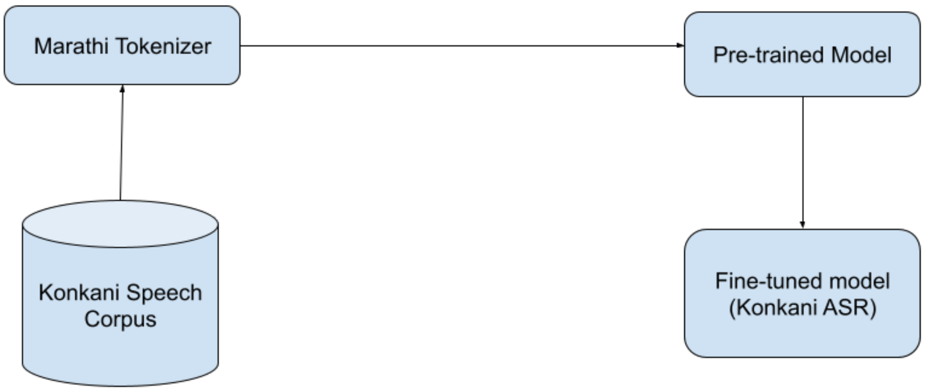

Figure 1: System development stages

Table 3: Results. The table displays Word Error Rate (WER) values for models utilizing various language tokenizers, both before and after the fine-tuning process. FT: fine-tuned model

Model	WER
whisper-small + Hindi tokenizer	73.57
whisper-small + Hindi tokenizer + FT	23.97
whisper-small + Marathi tokenizer	54.86
whisper-small + Marathi tokenizer +FT	17.16

- कितलेशेच दीस जाले तरी ताच्या तोंडांत पेजेचें एक कूंय लेगीत वचूंक नासलें.

Comparing the original and predicted speech instances shows that merging two words into one is a common error. The following is an example of such error.

- Original sentence: ताचो मांव बरो ना खंय. (His father-in-law is not well.)
- Prediction:
  - Speaker-1: ताचो माव बरो नाकांय.
  - Speaker-2: थाचूं भाव बरो नाका.
- Original sentence: हांगा तुजें वय बरय.. (Write your age here.)
- Prediction:
  - Speaker-1: हांव तुचें 'वैवपूय'.
  - Speaker-2: हांगा तुजें 'वैबरय'.

The words 'ना खंय' are merged into a single word, 'नाकांय' and 'नाका'. During manual testing of the ASR, we observed that errors are more likely to occur in the output if there is background noise. This can be improved by pre-processing voice signals for noise removal, and it will be considered in our future work. We also observed that ASR does not recognize some words. This could be because of the Marathi tokenizer used in training the model. This was done because the unavailability of the Konkani tokenizer and other tokenizers, like the Hindi tokenizer, was not helping in the improvement of ASR performance. The creation of the Konkani tokenizer will be attempted in our future work.

7 Conclusion and Future work

Through this work, we have created an ASR system for Konkani language spoken on the western coast of India. It has been demonstrated that a pre-trained model can be successfully used to build ASR by fine-tuning the model with annotated speech data in a less-resourced target language. ASR showed a baseline WER of 17 on the test data. In the future, we want to use better modeling architectures to improve the ASR model.

References

Census. 2011. Census of india 2011. PAPER 1 OF 2018 LANGUAGE INDIA, STATES AND UNION TERRITORIES (Table C-16).

Paria Jamshid Lou and Mark Johnson. 2020. End-to-end speech recognition and disfluency removal. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2051–2061, Online. Association for Computational Linguistics.

Bhageshree K. Khandale, Saurabh Varik, Rajesha N., Manasa G., Narayan Choudhary, and L. Ramamoorthy. 2018. Konkani raw speech corpus.

Shreya Khare, Ashish R Mittal, Anuj Diwan, Sunita Sarawagi, Preethi Jyothi, and Samarth Bharadwaj. 2021. Low resource asr: The surprising effectiveness of high resource transliteration. In *Interspeech*, pages 1529–1533.

Ankit Kumar, Mohit Dua, and Tripti Choudhary. 2014. Continuous hindi speech recognition using gaussian mixture hmm. pages 1–5.

Wei Liu, Kaiqi Fu, Xiaohai Tian, Shuju Shi, Wei Li, Zejun Ma, and Tan Lee. 2023. An asr-free fluency scoring approach with self-supervised learning. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Mohamed Mhiri, Samuel Myer, and Vikrant Singh Tomar. 2020. A low latency asr-free end to end spoken language understanding system.

Sinno Jialin Pan and Qiang Yang. 2009. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR.

Pawel Swietojanski, Arnab Ghoshal, and Steve Renals. 2013. Revisiting hybrid and gmm-hmm system combination techniques. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6744–6748.

Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. 2016. A survey of transfer learning. *Journal of Big data*, 3(1):1–40.

Hao Yang, Min Zhang, Shimin Tao, Miaomiao Ma, and Ying Qin. 2023. Chinese asr and ner improvement based on whisper fine-tuning. In *2023 25th International Conference on Advanced Communication Technology (ICACT)*, pages 213–217.

Daniela Şchiopu. 2013. Using statistical methods in a speech recognition system for romanian language. *IFAC Proceedings Volumes*, 46(28):99–103. 12th IFAC Conference on Programmable Devices and Embedded Systems.