Building ASR systems using Resource-Rich and Resource-Poor languages

Ganesh S Mirishkar

Speech Processing Laboratory, International Institute of Information Technology, Hyderabad, India

mirishkar.ganesh@research.iiit.ac.in

1. Research Question & Motivation

Nowadays information processing systems have become an integral part of human life. One such information processing system is automatic speech recognition (ASR). The role of an ASR system is to transduce/transform the given speech signal into the corresponding text. Mathematically, it can be represented as,

$$recognized\ text = \mathbf{H}(speech\ signal) \quad (1)$$

Here the role of $\mathbf{H}(\cdot)$ is to properly estimate/predict this function. These ASR systems have multiple applications in various domains like health care, automobile, IoT devices, etc. The performance of the ASR system gets affected if there are too many acoustic variations [1] in the speech signals. The main cause of these variations might be due to (1) Distortion in the microphone, electrical noise, directional characteristics, (2) Overlapping speech, background noise, reverberations, and (3) different speaking styles like stress/emotion, high or low speaking rate, stuttering, Lombard effects. In a practical scenario, building an ASR system is a challenging task as it has to address the following issues, (1) Noise [2, 3, 4], (2) Accent [5, 6, 7], (3) Single-channel, Multi-Channel (4) Low-resource [8, 9, 10] (5) Multilingual [11, 12] (6) Code-switching. I would be addressing Low-resource and multilingual problems for my doctoral research. So there arises a question what does a "low-resource" [13] mean? In general, a language is considered to be low-resource if they fall under any of the following:(1) limited web presence, (2) lack of linguistic expertise, digital resource (audio or text corpora). In order to build an ASR model to achieve human parity we need to have a tons of labeled speech corpus for training. Recently [14, 15], has shown that the English ASR system has reached human parity for a generic domain. And the amount of labeled data available for the English ASR task is 30,000 hours [16, 17]. But this is not in the case of Indic languages, and we can hardly find in the magnitude of hundreds. So in my work, I would be addressing how to procure the data from the crowd, the internet (as there is a lot of freely available) and curated it into the ASR format. In my case, I have collected around 2000 hours of Telugu Speech corpus in crowdsource manner.

So the next question is "what is multilingual"? Let's consider a country like India, where numerous languages are being spoken across different geographical locations. So building a monolingual ASR system for all the languages is complex, and handling it involves too much computation. And it also requires a good amount of training data to build. To mitigate this issue, I will be building a Joint Acoustic Model (JAM) across all the languages. For a JAM, we will be pooling the data from all the languages and build a single acoustic model so that the attributes are shared across the languages. It is observed that such kind of approach is being benefited for building low-resource multilingual ASR systems.

1.1. Aims of the dissertation

The aim of my doctoral research are as follows:

1. Building a low-resource multilingual speech recognition system in Indian context
2. Crowd-sourced strategies for the collection of a large scale speech corpus
3. How good is a high resource language good for building low resource ASR system?

2. Dissertation outcomes so far

The main goal of this thesis is to build a speech technology system that is predominately an ASR for low-resource multilingual in the Indian context. Moreover, strategies involved in collecting a large-scale speech corpus using crowd-sourcing methods. In this section, I will be presenting a brief overview of the work done so far.

2.1. Building a low-resource multilingual speech recognition system in Indian context

India is a country with enormous linguistic diversity. Approximately 2000 languages are being spoken across, out of which the constitution of India schedules 23 languages. Among those languages, automatic speech recognition (ASR) supports only a few. This is because an ASR system demands thousands of hours of annotated speech data, extensive text, and a dictionary that can stretch out all words in languages. In literature, many studies have shown that Indian languages share a common phonetic space even though they differ phonotactically[1]. In [10], phonetic sound principles and their benefits have been

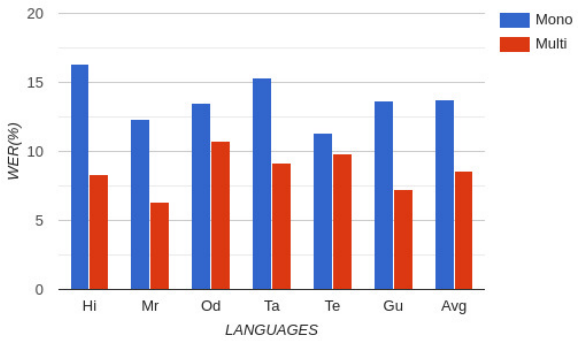

Figure 1: *Contrastive comparison of (Low-Rank TDNN with external text data) monolingual ASR system vs. multilingual ASR system reported in WER(%)*

exploited, which led to the idea of using a common label set (considering each language character as a separate entity). So in this work, we exploited a common label set (CLS) approach in building a multilingual speech recognition system for low-resource languages. We have investigated different variants of Time Delay Neural Networks (TDNN) architectures [18, 19] for building low-resource speech recognition systems using CLS approach. The efficacy of different acoustic models for low resource multilingual speech recognition tasks has been reported in six Indian languages, namely Hindi (Hi), Marathi (Mr), Odia (Od), Tamil (Ta), Telugu (Te), and Gujarati (Gu). All the experiments have been conducted on MUCS Interspeech 2021 challenge [20]. Initially, monolingual systems have been built across each language on different variants of TDNN architectures. Then external text data is included while building recurrent neural network (RNN) based language model. Later a joint acoustic model is trained on different variants of TDNN architecture, and the same process is being followed for the multi-lingual systems as well. Among all the variants of TDNN architectures, the low-rank TDNN with an external text language model has performed better across all the languages. In the Figure 1, the blue colour bars indicates the performances of monolingual ASR systems, where as the red corresponds to the multilingual ASR systems. The performance improvement in low-rank TDNN is due to the skip connections, bottleneck liner transformation layer, and batch normalization in it.

[1] In general phonotactic is defined as the study of the rules governing the possible phoneme sequences in a language.

Table 1: *Baselines for Telugu speech recognition task on the corpus collected*

Models	GMM-HMM	SGMM-HMM	DNN-HMM	TDNN-HMM	Transformer	Conformer
WER(%)	33.12	28.89	24.04	21.98	18.09	15.72

2.2. Crowd-sourced strategies for the collection of a large scale Telugu speech corpus

Building an ASR system requires a vast amount of annotated corpus, and finding such corpus for Indian languages is tricky. Furthermore, leveraging the advancement of deep neural architectures is not exploited yet. So we have attempted these lines to create a large-scale corpus in a crowd source manner. In this work, we explain the approaches incorporated while collecting the database [21]. So the primary purpose of this database collection is to mitigate the problem of resource-poor languages. At the International Institute of Information Technology Hyderabad (IIIT-H), we followed the crowdsourcing manner in collecting Telugu speech corpus using three approaches, namely

Approach-1 :: Collection through web and mobile applications from the crowd

In this approach, we have collected the database via a laptop and mobile phone. In this the crowd has been given a list from where he/she is supposed to choose a topic and asked them to speak over it. The recorded streams from mobile and laptops are being captured on to a VOIP based platform. The data captured is of 16 kHz 16 bit PCM format. Apart from it, we have conducted a Just a minute (JAM) and debate session across the universities, different work places etc., to have a diversity in the database. The data which we collected from this approach is spontaneous speech.

Approach-2 :: Data pooling through freely available resource

There is a lot of audio-video content available on the internet, but it is in an unannotated format. Due to this, there is a need to have a framework or a workbench where an annotation is performed automatically or manually. So to mitigate this issue, we have come up with a semi-automatic pipeline where the user can upload the URL or raw audio file into the pipeline as it processes the content into a specified format by the user for his/her task. We named the framework as Crowdsourced Data-collection (CSTD) pipeline. It is observed that most of the available content is in conversation mode. We have pulled the content of the Telugu language and passed it to the CSTD pipeline.

Approach-3 :: Data pooling through WhatsApp campaigns

In this approach, the crowd has to read the prompt(sentence) which is sent to their whatsapp contact number.It is made sure that each user cant exceed 4000 sentences. The precaution has been taken so that the users have unique sentences (no overlapping). The text sentence is collected from various domains(law,sports, arts etc)

2.2.1. Audio and Transcript verification

In this, data collected from all approaches is being verified by human. Four fold verification is performed through out the database. It is made sure that the verification is done by the humans who are certified in the particular language(in our case Telugu).

2.2.2. Summary & Results of the Telugu speech corpus collected

To the best of my knowledge, this is the first attempt to collect Telugu speech corpus on a large scale using the crowd sourcing approach. Despite the pandemic, we could successfully collect such a large-scale corpus without delay. The total duration of the corpus is roughly around 2000.8 hours. This corpus covers three regional dialects (Coastal Andhra, Rayalaseema, and Telangana) of the Telugu language in different speaking modes i.e., read, conversational and spontaneous. The Figure 2, depicts the statistics of the corpus which is collected. In Figure 2, the labels AP-1,AP-2, and AP-3 corresponds to the approach-1 (ie. spontaneous speech), approach-2 (ie. conversational speech) and approach-3 (ie. read speech).The database which we have collected is being evaluated on hybrid and end-to-end speech recognition systems. From Table 1, the first corresponds to the model which we have built, and the second row presents the results in terms of word error rate(WER(%)). The developed CSTD pipeline/framework tries to bridge the gap between resource-poor and resource-rich languages.

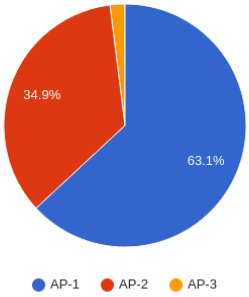

Figure 2: *Comparison of the corpora collected across different approaches*

It encourages other researchers to procure and curate the databases in other languages.

2.3. How good is a high-resource language good for building low resource ASR system?

The data which we have collected through the crowd source approach, as mentioned in section 2.2, we have built a pretrained model using the wav2vec2.0 framework [22]. To evaluate our pretrained model and understand the learned representation of it, we have fine-tuned it with different languages in a low resource setting(with labeled 10 minutes, 1 hour,10 hours, and 100 hours). Later, a set of experiments was performed on a high-resource language (in this case, Telugu). We found that this method achieves the WER of 10.2% in the same language. In the future, we would like to build a multilingual speech recognition system using this pretrained model(Telugu).

3. Contributions & Future Work

The significant contribution of my thesis revolves around building an ASR system in the Indian context for low-resource settings by investigating different architectures and lexicons (i.e., Common Phone set, Common label set, etc.). Another contribution to the speech community would be strategies incorporated while collecting the speech corpus using a crowd sourcing approach. To the extension of Section 2.3, I would like to explore the learned representations in the context of multilingual ASR task, noisy scenario (how does the noisy input affect speech recognition task?). Later, I would like to devise domain adaptation methods for low-resource settings.

4. Acknowledgements

I sincerely thank my supervisor, Dr. Anil Kumar Vuppala, and Prof. B Yegnanarayana, Speech processing laboratory, IIIT Hyderabad, for the valuable discussions and motivation for the work. Also I would like to acknowledge the funding source Technology Development for Indian Languages (TDIL), Ministry of Electronics and Information Technology (MeitY), Government of India, under the project titled “Crowd Sourced Large Speech Data Sets To Enable Indian Language Speech - Speech Solutions”.

5. References

[1] S. Furui, “Generalization problem in asr acoustic model training and adaptation,” in *2009 IEEE Workshop on Automatic Speech*

Recognition & Understanding. IEEE, 2009, pp. 1–10.

[2] M. L. Seltzer, D. Yu, and Y. Wang, "An investigation of deep neural networks for noise robust speech recognition," in *2013 IEEE international conference on acoustics, speech and signal processing*. IEEE, 2013, pp. 7398–7402.

[3] K. C. Sim, A. Narayanan, A. Misra, A. Tripathi, G. Pundak, T. N. Sainath, P. Haghani, B. Li, and M. Bacchiani, "Domain adaptation using factorized hidden layer for robust automatic speech recognition." in *Interspeech*, 2018, pp. 892–896.

[4] T. Yoshioka and M. J. Gales, "Environmentally robust asr front-end for deep neural network acoustic models," *Computer Speech & Language*, vol. 31, no. 1, pp. 65–86, 2015.

[5] X. Yang, K. Audhkhasi, A. Rosenberg, S. Thomas, B. Ramabhadran, and M. Hasegawa-Johnson, "Joint modeling of accents and acoustics for multi-accent speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 1–5.

[6] B. Li, T. N. Sainath, K. C. Sim, M. Bacchiani, E. Weinstein, P. Nguyen, Z. Chen, Y. Wu, and K. Rao, "Multi-dialect speech recognition with a single sequence-to-sequence model," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 4749–4753.

[7] U. Nallasamy, "Adaptation techniques to improve asr performance on accented speakers," Ph.D. dissertation, Carnegie Mellon University, 2016.

[8] S. Dalmia, R. Sanabria, F. Metze, and A. W. Black, "Sequence-based multi-lingual low resource speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4909–4913.

[9] J.-Y. Hsu, Y.-J. Chen, and H.-y. Lee, "Meta learning for end-to-end low-resource speech recognition," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7844–7848.

[10] V. M. Shetty and S. Umesh, "Exploring the use of common label set to improve speech recognition of low resource indian languages," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 7228–7232.

[11] B. Sen, A. Agarwal, M. S. Ganesh, and A. K. Vuppala, "Reed: An approach towards quickly bootstrapping multilingual acoustic models," in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 272–279.

[12] H. K. Vydana, K. Gurugubelli, V. V. R. Vegesna, and A. K. Vuppala, "An exploration towards joint acoustic modeling for indian languages: Iiit-h submission for low resource speech recognition challenge for indian languages, interspeech 2018." in *INTERSPEECH*, 2018, pp. 3192–3196.

[13] G. Mirishkar, A. Yadavalli, and A. K. Vuppala, "An investigation of hybrid architectures for low resource multilingual speech recognition system in indian context," in *Proceedings of the 18th International Conference on Natural Language Processing (ICON)*, 2021, pp. 205–210.

[14] G. Saon, G. Kurata, T. Sercu, K. Audhkhasi, S. Thomas, D. Dimitriadis, X. Cui, B. Ramabhadran, M. Picheny, L.-L. Lim *et al.*, "English conversational telephone speech recognition by humans and machines," *arXiv preprint arXiv:1703.02136*, 2017.

[15] H. Yadav and S. Sitaram, "A survey of multilingual models for automatic speech recognition," *arXiv preprint arXiv:2202.12576*, 2022.

[16] D. Galvez, G. Diamos, J. Ciro, J. F. Cerón, K. Achorn, A. Gopi, D. Kanter, M. Lam, M. Mazumder, and V. J. Reddi, "The people's speech: A large-scale diverse english speech recognition dataset for commercial usage," *arXiv preprint arXiv:2111.09344*, 2021.

[17] G. Chen, S. Chai, G. Wang, J. Du, W.-Q. Zhang, C. Weng, D. Su, D. Povey, J. Trmal, J. Zhang *et al.*, "Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio," *arXiv preprint arXiv:2106.06909*, 2021.

[18] M. Sugiyama, H. Sawai, and A. H. Waibel, "Review of tdnn (time delay neural network) architectures for speech recognition," in *1991., IEEE International Sympoisum on Circuits and Systems*. IEEE, 1991, pp. 582–585.

[19] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[20] A. Diwan, R. Vaideeswaran, S. Shah, A. Singh, S. Raghavan, S. Khare, V. Unni, S. Vyas, A. Rajpuria, C. Yarra, A. Mittal, P. K. Ghosh, P. Jyothi, K. Bali, V. Seshadri, S. Sitaram, S. Bharadwaj, J. Nanavati, R. Nanavati, K. Sankaranarayanan, T. Seeram, and B. Abraham, "Multilingual and code-switching asr challenges for low resource indian languages," *Proceedings of Interspeech*, 2021.

[21] G. S. Mirishkar, M. D. Naroju, S. Maity, P. Yalla, A. K. Vuppala *et al.*, "Cstd-telugu corpus: Crowd-sourced approach for large-scale speech data collection," in *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2021, pp. 511–517.

[22] C. Yi, J. Wang, N. Cheng, S. Zhou, and B. Xu, "Applying wav2vec2. 0 to speech recognition in various low-resource languages," *arXiv preprint arXiv:2012.12121*, 2020.