1. Title of the project (< 300 characters)

Crowdsourcing for Language Processing (CLAP): A platform for collecting labeled data for speech and language processing

2. Project Summary (< 2000 characters)

outlining briefly the scope of work, methodology to be adopted, and utility, justification and anticipated impact of the proposed technology development

Over the last decade, Artificial Intelligence (AI) has increasingly been making inroads into society and our lives. However, as with all other technologies, it is an important challenge to make such technology accessible to people from all strata of Indian society. This challenge manifests itself mostly at the interface between humans and computers. A key component of this interface that is highly sensitive to the cultural and linguistic background of the users is automatic speech recognition (ASR). To build competitive ASR systems for Indian languages, one requires large amounts of labeled speech data i.e. speech clips in different Indian languages accompanied by their corresponding text. Publicly-available repositories of labeled speech in Indian languages is currently a limited resource.

This project aims at collecting large volumes of labeled speech in a number of different Indian languages in a scalable manner using crowdsourcing. Towards this end, we undertake the following: 1) We will design a mobile application in Android and a corresponding backend server to crowdsource tasks for labeled speech in various Indian languages. Users of this app (workforce) will be given two different types of tasks to complete: a "Speak" task where users will read out prompts in their native tongues and a "Verify" task where users will be asked to confirm whether a prompt and its corresponding speech (obtained from a different user) are well-matched. 2) To collect large volumes of data, it is essential that we have effective mechanisms to recruit the workforce as well as retain them. For recruitment, we will explore Facebook/Google social media advertising, contact student bodies (as part of the National Social Service scheme), taxi driver associations etc. For incentivizing, we will employ gamification, PayTM-based money-transfers and the coupling of AI education with data collection by presenting internship opportunities for top students. We will explore various combinations of these mechanisms to determine the right scheme that gives the best return on investment. 3) Given the crowdsourced nature of collection, it is possible for poor quality data to creep into the corpora. To tackle this challenge, we will employ a host of techniques to post-process the data. Operations such as noise reduction and volume control will be applied to all the speech clips. Verify tasks coupled with majority voting could be used to catch instances of spamming in speak tasks. As an additional measure to ensure quality, we will perform automatic random checks on each user using gold standard verify tasks (where the outcome is known) to catch spammers.

We highlight that we intend to make the collected speech data available as publicly-available corpora that can be used by researchers and industry practitioners to build or bootstrap their

existing systems. We believe this would be a very valuable contribution towards furthering research on speech technologies in Indian languages.

3. Objectives (in bullet form) (< 1500 characters)

Please spell out the principal objective of your proposal (in terms of utility and novelty of the proposed end product).)

Our objective in this proposal is to collect large corpora of labelled speech in Indian languages to spur research and development of speech technologies both in India and across the world.

Towards this end,

1. We propose to build a crowdsourcing-based Android mobile application to collect labelled linguistic data from untrained contributors in a variety of noisy environments that mimic real-life settings.
2. Design incentive mechanisms to create (and retain) a network of users so as to build large corpora in many languages.
3. Devise mechanisms to clean the collected data to generate high quality corpora.

4. Current status of development in the subject domain in which the proposed new technology is being developed (within 7500 characters or 3-4 pages)

In the proposed project, we aim to 1) build high-quality, publicly-available speech datasets for Indian languages that can be used to bootstrap technologies for these languages and 2) use crowdsourcing to achieve the same in a scalable manner. We now present the current status of development with regards to both these stated objectives.

Technologies for high-resource languages like English, Mandarin are supported by large, high-quality datasets [openslr] but, unfortunately, labeled speech corpora in Indian languages are currently not available as a free resource. This has been a significant hurdle for speech and language researchers and industry practitioners alike in developing technologies for Indian languages. Our main goal in this project is to bridge this gap.

Crowdsourcing, as a mechanism to solve large scale problems by breaking up the problem into microtasks performed in parallel by large crowds, has received considerable attention in the past. Within this framework, one has to typically address the challenges of effective task creation, task allocation, task completion/feedback and incentive mechanisms. Many of these challenges have already been tackled by existing Internet based crowdsourcing platforms such as Mechanical Turk [mturk], Skillsforchange [schange] and Helpfromhome [hhome]. A large body of work covering various aspects of crowdsourcing has already been published. For example, the authors in [rula14] have shown that lottery based payment mechanisms have high recruitments compared to guaranteed micro-payments. However, in terms of task compliance and user effort, guaranteed micro-payments work much better. Authors in [asha17] have shown

via a pilot study that effective feedback/motivation mechanisms are crucial in improving worker performance. In this project, we will leverage existing knowledge in this space and improve upon them to achieve our desired objective. Some of the innovative mechanisms we wish to explore are outlined in the methodology/research plan section.

Crowdsourcing has also been employed to collect annotated data for speech and language technologies [eskenazi13,google,cvoice]. Most of the prior work in this domain has focused on collecting text for speech in the wild. We found this to be an ineffective strategy for Indian languages and instead focus on collecting speech corresponding to text prompts. Overall, there has been very limited work on crowdsourcing speech in Indian languages. Mozilla's Common Voice project [cvoice] is a recent solution started in 2017, that is most relevant to our proposed project. Common Voice asks for volunteers to donate their voice to build an open-source voice database. The focus of this project is on building datasets for various popular languages of the world. They are in the process of collecting labeled speech in English, German, French etc. However, only five Indian languages have been identified (Tamil, Telugu, Bengal, Odia and Assamese) and efforts are currently underway to collect text prompts; thus, the actual speech collection phase will take time to come to fruition. Collecting text prompts is a non-trivial problem, which we discuss further in our research plan. Using any freely available text from Wikipedia or Government-based websites, as Common Voice might do, will not necessarily result in good speech samples as speakers might have trouble pronouncing the words which are not typically used in daily life. Active ground work with knowledge about good sources for the Indian languages in question will make a significant difference to the quality of speech samples collected. Apart from this, there are other fundamental differences between our approach and Mozilla's initiative as outlined below.

a) Unlike Mozilla's Common Voice project, our focus is on improving speech technologies for Indian languages, which would involve accounting for sufficient representation from different dialects and different demographics of speakers of Indian languages. Keeping this in mind, we will target specific sets of users by reaching out to organizations that are diverse in terms of their user base and have wide reach including, for example, student bodies in various local colleges (as part of the National Social Service scheme) and taxi driver associations.

b) Mozilla's project employs intrinsic motivation i.e. altruistic reasons in crowdsourcing. Prior work [roth] has shown that in developing countries, extrinsic motivation i.e. monetary and social gains are a lot more effective in increasing recruitment, ensuring task compliance and sustaining user effort. We intend to leverage current knowledge in the crowdsourcing ecosystem to design innovative incentive mechanisms which will have a huge and sustainable impact from an outreach perspective. We also intend to use gamification strategies to engage more users and tap into their social circles. We elaborate some of these strategies in the methodology/research plan section.

While Mozilla Common Voice will have a strong positive impact on the speech ecosystem, we believe other parallel efforts such as ours that target Indian languages with active ground work

will add significant value. Also, our efforts are complementary to initiatives like that of Mozilla's; we can utilize their data to build initial models which will be further refined to work better for Indian languages using our curated datasets. Our project will also aim for larger breadth across Indian languages (by targeting a number of different languages including Marathi, Malayalam, etc.) which is currently not a focus for the Common Voice project.

In order to build high-quality speech datasets, our crowdsourced speech samples will need to be suitably post-processed to ensure they are relatively noise-free. We will use existing techniques to clean up the speech samples [eskenazi13]. Also, the collected speech samples should be well-matched with the underlying text prompts. This can be ensured by building preliminary ASR systems and comparing the reference text with the decoded text from the ASR system. We elaborate more about this in our research plan.

5. Background, motivation and scope of the proposed technology development including relevant milestones already achieved and technology readiness level (TRL) already reached (< 2500 characters)

A large fraction of Indians, who are unable to read or write in their native tongues, are precluded from using technology. Enabling ASR solutions in a number of different Indian languages has the potential to make technology accessible to a wide range of users with low literacy skills and make technology more inclusive. Towards this, as a first step, we would require high-quality labeled speech corpora for Indian languages. Crowdsourcing has been explored as a technique to collect labeled speech in prior work [eskenazi13], though not for Indian languages.

Milestones we have already achieved include:

1. Pilot studies to assess what would be the most effective way to collect speech corpora. As described above, we narrowed down on collecting speech for text prompts extracted from story-books and novels online.

2. An initial prototype of the Android app is already in place. We used this app to conduct a set of field trials within the IITB student community. About 300 users registered with the app and used it to perform a total of 12,000 tasks (where each task, speak or verify, corresponds to a single sentence) for Hindi, Marathi, Telugu, Malayalam and Bengali. The overall feedback has been very positive.

Going forward, we would like to refine the app based on the feedback received during the trials to make it more user friendly. We will also need to handle intermittent Internet connectivity issues, which are pervasive in India. Scaling the app to a wider audience and user retention is another crucial step to increase the scope of our project. Towards this, we will design appropriate gamification/incentive/payment mechanisms and implement the same within our platform. Cleaning of collected data to generate high quality corpora is another important step,

which will also be incorporated into the platform. Some of our envisioned mechanisms along with the architecture of the app is described in more detail in our research plan.

6. Methodology, approach and research plan (15000 characters or up to 5 pages, including figure/chart/table/image; flow sheet may be included for any process development project)

Our main end-goals for the project include collecting large volumes of labeled speech in a number of different Indian languages using crowdsourcing and building speech corpora which can be used to build or bootstrap ASR systems. This can be broken down into four specific technical objectives, which we elaborate on further below:

A) Designing and implementing an Android app to crowdsource for labeled speech.

B) Devising mechanisms to evaluate the quality of the resulting speech samples and building high-quality speech corpora.

C) Devising innovative incentive mechanisms to engage larger user-bases.

D) Discovering what demographics need to be specifically targeted by building prototype ASR systems and iterating the data collection process with more targeted objectives.

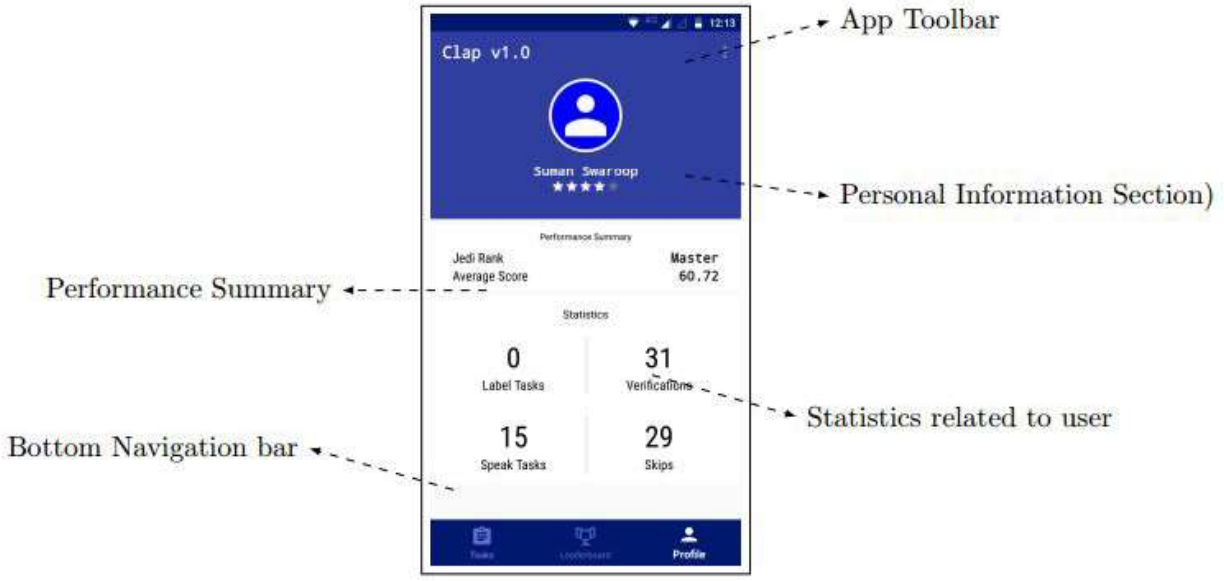
A) **Android app:** Labeled speech collection from crowds will be enabled with the help of an Android app and a corresponding backend server. The interaction between the two is as depicted in the figure (arch.pdf). For the app, the manager in the server is the main point of contact. The manager in turn will interface with relevant modules to get the information requested by the app.

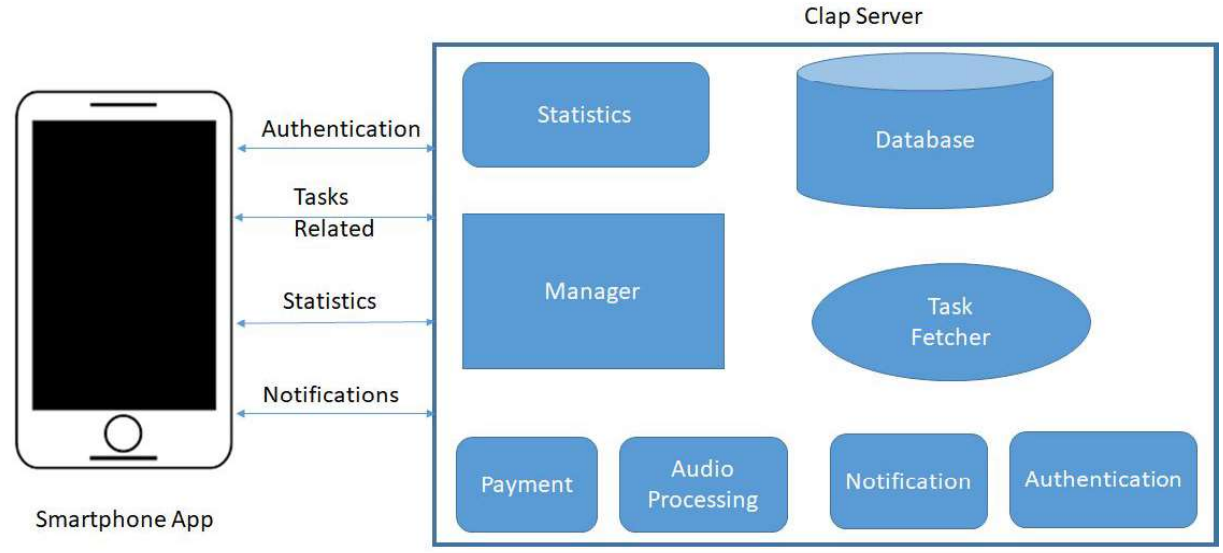
We now explain the typical workflow involved when using our system.

1. On installing the app, a user can sign up using his or her Facebook/Google ID or create an independent account. The user is accordingly verified by the authentication module in the server. First-time users are prompted to enter relevant information needed for assigning tasks such as language preference, task type preference, limit of outstanding tasks (the server never allocates total tasks exceeding this number), etc. First-time users are also made to walk through a flashcard-based tutorial so as to get familiarized with the system. They are then directed to the task page (see tasks.pdf). Repeat users directly land on the task page.
2. The task page on the app shows the various tasks assigned to the user. The app will contact the server to get this information. The manager, with the help of the task fetcher, will determine the right kind of tasks and allocate the same to the user. Currently our system supports three types of tasks: Speak, Label and Verify. Speak tasks prompt the user to speak the text shown in the box (see speak.pdf). Label tasks prompt the user to

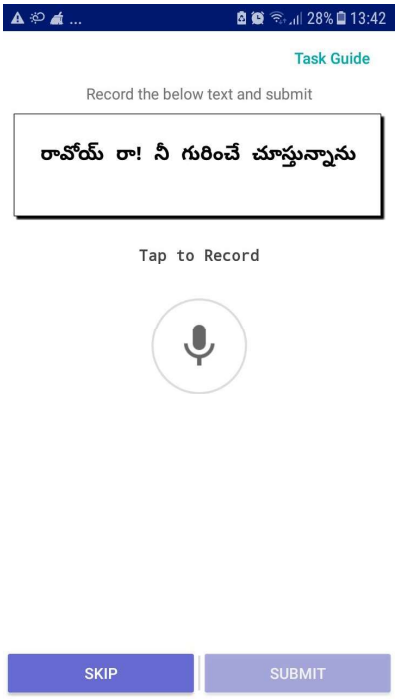
transcribe the given speech shown in the box. Verify tasks prompt the user to verify an already completed speak or label task i.e. verify if the text matches the speech and vice-versa. In the future, we may also support a Construct task, which will prompt users to construct sentences in the given language. Once a task is submitted by the user, the server updates the database and assigns new tasks to the user so as to keep the queue at the user always full.

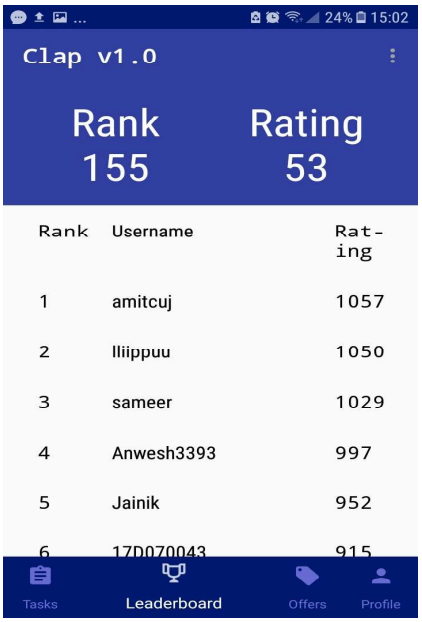
3. A user is often interested in the progress made so far. This information can be viewed by the user in the profile page (see profile.pdf). Our system also supports a leaderboard (see leader.pdf) where a user can see their standing relative to others. This is done as a means to motivate users to contribute more. The server uses the statistics module to calculate all this information.
4. An administrator who is managing the dataset collection needs appropriate tools to a) upload text or speech data for labelling b) track the progress of volunteers c) monitor the quality of the data collected and d) chart the progress made so far for various languages in the collection. All of this is provided by the server via a frontend dashboard and with the help of the statistics module.
5. A few other modules in the server ensure proper operation of the entire workflow. The server consists of a notification unit that sends periodic notifications to the users and prompts them to complete assigned tasks if they have not logged into the app for a long time. The audio processing unit filters the speech samples collected from users (i.e. removes loud background noises, increases the volume of clips that are inaudible, etc.) before passing them on for verification via the verify tasks.


App Toolbar
Clap v1.0
Suman Swaroop
Personal Information Section)
Performance Summary
Jedi Rank
Average Score
Master
60.72
Performance Summary
Statistics
0
Label Tasks
31
Verifications
Statistics related to user
15
Speak Tasks
29
Skips
Bottom Navigation bar
Profile


Clap Server
Authentication
Tasks
Related
Statistics
Notifications
Statistics
Database
Manager
Task
Fetcher
Payment
Audio
Processing
Notification
Authentication
Smartphone App


App Toolbar
Clap v1.0
Speak text
Language: Hindi
Speak text
Language: Hindi
Tasks(Vertical scrollable)
Speak text
Language: Hindi
Space for more tasks
Verify
Language: Hindi
Verify
Language: Hindi
Filter tasks
FILTERS
Tasks


28% 13:42
Task Guide
Record the below text and submit
రావోయ్ రా! నీ గురించే చూస్తున్నాను
Tap to Record
SKIP
SUBMIT



Our architecture needs input data in different modalities depending on the type of crowdsourcing task (i.e. text data for speak tasks and speech data for label tasks). Although our framework supports label tasks, our initial experiments showed this approach to be less effective for Indian languages. Firstly, access to Indian speech corpora was difficult and we could only obtain broadcast news data which does not mimic real-life speech. We also found large variance in quality across transcripts obtained from the volunteers (in terms of spelling errors). All volunteers found the transcribing task to be very difficult and tedious to carry out. Given these constraints, we predominantly focus on speak tasks and experiment with collecting speech corresponding to text prompts. Volunteers found the speaking tasks to be significantly easier. Further, the resulting speech was found to be more realistic with characteristics of spontaneous speech rather than read speech.

Finding text prompts for these speak tasks is however a challenging problem since we want the text to come from story-books/short stories/novels which are of a free license or come with permission from the author to be used within our app. We have already leveraged one such online resource from Pratham Books [sweaver] in our pilot studies which contains children's stories in various Indian languages. However, this will not suffice for large-scale data collection. We will explore more ways of generating text including 1) crowdsourcing for text prompts themselves from users, and 2) using machine translation tools to automatically convert prompts in English to text in different Indian languages. (Since the resulting prompts will not be error-free, they will undergo a round of error correction with the help of native speakers in the corresponding Indian languages.)

B) **Evaluating the quality of crowdsourced speech samples:** Crowdsourcing is an inherently noisy technique due to the diversity and variability of the crowds involved. In the case of crowdsourced speech, the noise mainly stems from environmental interference in the speech

samples (e.g. speech samples from users recording in extremely noisy settings) and mismatch between the text prompt and the resulting speech sample. Background noise can be reduced using a post-processing step of applying standard denoising techniques to the speech samples. Mismatch between the text prompt and the speech sample can be determined using an ASR system, which is outlined further below.

Reliability of the users is another important variable to be taken into consideration when using crowdsourcing techniques. We could identify and eliminate spammers early in the process, so that we can effectively focus on the more motivated and dedicated users. There has been a lot of prior work on how to achieve a certain amount of reliability in the crowdsourced labels from a diverse crowd and how to identify adversarial workers in crowdsourced labeling tasks [karger14,jaga17]. Majority voting is a popular strategy that is frequently employed to converge on a single label for an item from many crowdsourced labels. We will also make use of majority voting across verify labels that we receive for a single sentence to determine whether or not a speech sample is of high quality. We will also assess the reliability of workers by inserting speech samples for which we know the ground-truth labels and asking the users to label them.

Prior work on crowdsourcing has focused on labeling for multi-class classification tasks where users are presented with an item that has an underlying ground-truth discrete label. Our setting is different in that we want to collect a speech sample (which is a sequence) that exactly matches a given text prompt. This introduces the notion of getting the sequence partially correct, unlike for discrete labels where the user is either correct or wrong. To compute this measure of partial correctness, we will estimate an edit-distance based score between the reference text sequence and a decoded text sequence from an initial ASR system built for this language using the existing speech data.

C) **Innovative incentive mechanisms:** In order to build large corpora spanning a number of different languages, it is essential to reach out to a large audience and also retain the recruited users. Incentive mechanisms thereby play a very important role. In the crowdsourcing community, two mechanisms are often employed: intrinsic (based on satisfaction or altruistic purposes) and extrinsic (based on monetary or social gains). In this proposal, we plan to explore a variety of mechanisms to determine the right scheme that gives the best return on investment in the longer run. Future field trials will be based on the best identified scheme. We intend to explore the following schemes during the course of the project:

1. In this scheme, we will harness the intrinsic motivation among volunteers to contribute to this noble cause of building speech corpora for Indian languages. We will explore large advertising platforms such as Google and Facebook to reach out to audiences based on their mother tongue, locality, etc. We will also incorporate mechanisms within our app so that users can recruit other users among their social circles. Gamification will also be explored to ensure better retention rates among those targeted.

One of the gamification strategies we intend to incorporate within our app is a storytelling feature. Instead of the text prompts being randomly presented to users as part of their speak tasks, we will present a sequence of sentences from a story to the user. This qualifies as an intrinsic motivation tactic as users could be motivated to read more of the story. We will give users the option to opt out of the story-telling phase at any point and revert back to providing speech samples for randomly chosen text prompts. Users who read out an entire story will receive high scores which will increase their standing on the leaderboard.

2. In this scheme, we will explore extrinsic motivation via PayTM-based guaranteed returns, wherein users are paid a nominal amount for completing tasks successfully. One demographic that we intend to recruit from are taxi driver associations since this workforce has access to smartphones and also some leisure during wait times.

3. Every year thousands of students across the country look for internship opportunities in top-tier colleges such as IITB. In this scheme, we will leverage this desire of the students and couple it with educational initiatives to recruit students. One of our ideas includes offering an AI-based MOOC course for interested students. To complete the course, the students will have to deliver a certain number of completed speech tasks (with help from family/friends); watch and learn from videos that teach advanced AI/ML techniques and finish an in-depth auto-graded lab project based on applying AI techniques on publicly available speech data. Students with top scores will be shortlisted. After an interview, 5-10 students will be awarded an internship opportunity at IITB. We believe that given the huge demand for building AI/ML skills in the country, even if a few hundred students apply and if each can recruit 20 more among their family and friends to complete say 100 tasks, we can build substantial corpora. We note here that those missing the internship opportunity will still benefit from the lab-based MOOC.

4. Like in the previous scheme, in this scheme we will once again attempt to tap into the substantial engineering student body in the country to build a large user base for our application. NSS (National Social Service) is a nation-wide program run in many colleges and universities with the goal to inculcate social service among students. The students enrolled in the scheme are required to donate a certain number of hours in a semester towards a social cause. We will approach the coordinators of this scheme in various colleges and request them to incorporate our crowdsourcing task as one of the tasks the students could attempt.

These are the main incentive mechanisms we would like to explore. Based on initial trials and feedback from the users, these schemes will be accordingly modified.

D) **Streamlining the data collection phase with the help of initial ASR prototypes:** Indian languages are rich in dialects and are also prosodically diverse in terms of different accents in which speech is rendered depending on where the native speakers reside. Speech accents are

one of the key deterrents in the widespread deployment of speech recognition systems in multilingual societies such as India [huang04]. Even if speech data from specific accents are not present in a corpus, in recent work by the PI to appear in the proceedings of Interspeech 2018, we show that properties of related and similar accents can be leveraged to improve speech recognition performance [jain18]. Such questions have not been systematically examined for Indian languages. We will study how speech accents across a specific Indian language (say Hindi) can be used to improve speech recognition performance on an unseen accent of Hindi. In the first phase of speech data collection, we intend to reach out to native speakers far and wide to collect as many diverse samples as possible. We will build ASR systems using this initial dataset, and evaluate performance on specific accents to find out which ones suffer the most in terms of error rates. In the second phase of data collection, we will specifically target users from specific parts of the country who speak in these underrepresented accents and further investigate how much recognition performance improves and which related accents help the most.

7. Justification and novelty (< 2500 characters), including the scope of patent application or utilizing an existing patent

(Explain why your should approach or method succeed and what are the novelty components.)

Speech in Indian languages can be very diverse across its native speakers hailing from different parts of the country (e.g. Rajasthan vs Hyderabad). A speech corpus in an Indian language should attempt to sufficiently capture these variations across speakers. This is why we believe that crowdsourcing for speech is a good idea for Indian languages, as opposed to collecting speech recordings from a limited set of volunteers in a studio environment. Creating a mobile app for this purpose will potentially allow us wide access across speakers in the country. We will also take care to incorporate design mechanisms in our app that will help maximize user engagement. For example, text prompts will be extracted from stories and fed to users in sequence so that the users can narrate an entire story, which is likely to keep them more engaged than using isolated text prompts. This gamification strategy was elaborated on in our research plan.

Developing speech recognition techniques for languages with limited labeled speech (also referred to as low-resource or under-resourced languages) is a fairly well-studied area [besacier14]. However, building ASR models in low-resource settings that generalize well to speech with a lot of regional and dialectal variation has limited prior work. We believe our corpora will help address these type of problems by leveraging similarities between speech accents across regions and combining deep neural-network based models with this a priori linguistic knowledge [jyothi15, jain18].

8. Benchmark, milestones and time-frame (5000 characters or within 2 pages) - including table, pie or flowchart, bar diagram, etc.

Crowdsourced speech with accompanying transcriptions from our app will be used to train ASR systems for each Indian language. Standard evaluation metrics for ASR, such as word error rates, can be used to assess the quality of our crowdsourced dataset and create a benchmark. We intend to release code recipes for our ASR systems, so that the baseline error rates on our crowdsourced data can serve as a benchmark for other research groups and organizations.

Important milestones for this project, along with a rough timeline, have been outlined below:

Year 1: Incorporate feedback received from the initial trials conducted on the IITB campus. Design and implement the various incentive schemes we proposed and evaluate the settings under which different schemes work best. In the later half of the year, start trials in different parts of India.

Year 2: Based on the extended trails, modify the architecture/mechanisms to achieve better recruitment, retention and task compliance. Discover what populations and dialects need to be specifically targeted. Implement mechanisms to clean the collected data and create corpora that are ready to be publicly disseminated.

9. Nature and evidence of industry participation indicating the type and quantum of support (within 2000 characters) and upload letter, if any

Microsoft India R&D Ltd. has committed 5 lakhs towards this project. The distribution of their funding across different budget heads is flexible. We have uploaded a letter of support from them. For the letter to contain the exact funding amount, Microsoft India required a memorandum of understanding (MoU) to be signed with IITB. The MoU is currently under preparation and will be available to us soon.

10. Equipment, infrastructure and facilities available at the host organization(s) (<2500 characters)

We will be using the services of our graduate students (and a few undergraduate students) and project staff to carry out all the relevant research work and deployment tasks. We plan to host the backend (that interfaces with the app) on servers in the IITB CSE department and the mobile app will be hosted on Google PlayStore.

Being on a large and diverse college campus gives us easy access to students/staff volunteers who can help us with testing trial versions of our app. The large scale trials will be conducted with the help of volunteers across universities and organizations across the country. All the data collected from these trials will be saved on a local server at IITB with large storage space. All our compute needs will also be handled by a compute server in the IITB CSE department.

11. Evidence of earlier records of technology development (300 words or 1 page) –

already published paper, patent, letter, certificate, photograph) may be uploaded

1. Kameswari Chebrolu, Bhaskaran Raman, Vinay Chandra Dommeti, Akshay Veer Boddu, Kurien Zacharia, Arun Babu, Prateek Chandan, "SAFE: Smart Authenticated Fast Exams for Student Evaluation in Classrooms", SIGCSE 2017 (see also http://safe.cse.iitb.ac.in/)

2. BodhiTree: A Platform for Improving Learning Outcomes using Online Interactive Courses and Assessed Labs (http://bodhitree.cse.iitb.ac.in/; 40+ courses offered so far in 10+ colleges touching 15000+ people)

12. Potential user or agency who would be keen to utilize your development (< 1000 characters)

One of the main deliverables of our project would be well-curated, labeled speech corpora in different Indian languages. There are many target users who would benefit from this resource. They would include beneficiaries from academia i.e. research labs and groups across universities and institutes in India, who could use this resource as a learning tool to learn more about how to build speech technologies for Indian languages. Such a resource would also be useful to industry partners (like Microsoft, Microsoft Research, etc.) and start-up companies (like Slang Labs, etc.) who can use this data to either augment their existing datasets to train models for Indian languages or use it as an evaluation testbed. Hosting such datasets on well-established language consortiums such as Linguistic Data Consortium (LDC) would also have the benefit of exposing speech researchers worldwide to Indian languages.

13. Final outcome and deliverables (<2500 characters)

Our final deliverables will include:

A) An Android app, hosted on the Google Playstore, that can be used to crowdsource for labeled speech in Indian languages. A backend server that interfaces with the app to crowdsource tasks and process the collected data. This app will support a variety of incentive and gamification mechanisms described earlier in our research plan.

B) Labeled speech, packaged as a corpus, which will be made publicly available. We plan to host the corpus on consortiums such as Linguistic Data Consortium (LDC) so that it meets the standardization criteria adopted worldwide for corpora and reaches a wide audience.

C) A report based on our field trials that will provide an in-depth analysis of our data collection process, along with detailing tradeoffs between the various recruitment/incentive/gamification mechanisms explored. In addition, we will also include surveys that assess the usability of our system by the end users.