Automatic Annotation of Voice Forum Content for Rural Users and Evaluation of Relevance

Deepika Yadav
IIIT Delhi
India
deepikay@iiitd.ac.in

Mayank Gupta
IIIT Delhi
India
mayank16030@iiitd.ac.in

Malolan Chetlur
IBM Research
India
mchetlur@in.ibm.com

Pushpendra Singh*
IIIT Delhi
India
psingh@iiitd.ac.in

ABSTRACT

Voice forums are an effective intervention medium for marginalized communities to access information in a structured and localized manner. Users actively contribute by posting questions and responses in the form of audio messages, and thereby help in enriching the voice forum content. In order to build an audio library using the voice forums to disseminate information, significant manual effort is needed in analyzing and curating the data. This is one of the key impediments to the successful implementation of voice forums for knowledge dissemination and training.

In this paper, we explore the effectiveness of automated approaches to analyze and curate voice forum content in Hindi, a native language in the northern part of India. We study the use of standard techniques such as topic modeling and extractive summarization on Hindi speech transcripts (with WER of 67%) to cluster audios thematically and create summaries for individual audios respectively. These curated audios are used to build an IVR-based library for community health workers in rural India. We evaluated the relevance and preference of the automated annotation using a field trail. We find that the relevance perception varied between human and automatically generated annotations, but automatically generated summaries were still found to be useful to access the voice forum audios.

CCS CONCEPTS

• **Information systems** → **Digital libraries and archives**; **Relevance assessment**; • **Human-centered computing** → *Accessibility systems and tools*;

KEYWORDS

HCI4D; ICT4D; Interactive Voice Response; IVR; Community Health Workers; Topic Modeling; Speech Summarization

ACM Reference Format:
Deepika Yadav, Mayank Gupta, Malolan Chetlur, and Pushpendra Singh. 2018. Automatic Annotation of Voice Forum Content for Rural Users and Evaluation of Relevance. In *COMPASS '18: ACM SIGCAS Conference on Computing and Sustainable Societies (COMPASS), June 20–22, 2018, Menlo Park and San Jose, CA, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3209811.3209875

1 INTRODUCTION

Mobile phones have been a viable means of information to reach out to rural communities of developing countries. In particular, voice call communication has remained the primary form of use due to low literacy levels in these regions. This has led to the emergence of research around IVR (Interactive Voice Response) based voice forums, covering a range of application domains such as citizen journalism [17], entertainment [26], agriculture [20], job search and information portals [21, 23].

The main feature of these forums is the interactive voice application through which users can post questions and responses in form of audio messages and browse other's messages. This results in the development of the collection of messages, which further gets utilized in ways such as enriching forum content, developing digital libraries, and publishing to other media (radio, web). For these tasks, often a dedicated team of moderators are involved in analyzing and extracting the important pieces of data [8, 17]. Although human-based curation has merits on judgment, quality, and relevance aspects, it becomes difficult to scale. This necessitates the investigation of automation possibilities and their effectiveness in real-world applications.

In recent times, techniques in natural language processing and information retrieval fields have advanced for speech-based applications by the increasing use of transcripts. However, the target applications and user-base have been restricted to well-defined languages like English. Works around voice-based applications designed for marginalized communities of developing countries have been lacking. In this paper, we present an automatic approach to curate audio recordings of a Hindi language-based voice forum of community health workers in India. We used Google Speech

* corresponding author

COMPASS '18, June 20–22, 2018, Menlo Park and San Jose, CA, USA

ACM ISBN 978-1-4503-5816-3/18/06…$15.00
https://doi.org/10.1145/3209811.3209875

Recognition APIs to obtain low accuracy transcriptions of the audio recordings, and applied basic natural language processing algorithms such as topic modeling and extractive summarization to cluster and summarize audio recordings. We based our research around these questions: given the audio transcripts of low accuracy, how can the standard techniques of natural language processing and information retrieval be leveraged to automate or semi-automate the basic curation tasks of clustering, tagging and summarizing, and will that type of curation be found relevant by the end users? The specific research questions were:

- **RQ1**: Are topic models useful for clustering the voice forum dataset?
- **RQ2**: Is the relevance perception of the automatic annotation similar to manual annotation?
- **RQ3**: Is the relevance perception similar across summary types (summary as a sentence, summary as a group of keywords)?
- **RQ4**: Does the preference of the users for the type of summary change based on the annotation source (human versus machine) and annotation type (summary sentence versus group of keywords)?

We attempted to address the above research questions through a field trial with 48 community health workers in the northern part of India. Our evaluation showed that the users found topic annotation relevant with the audio clusters. In the case of annotating audios with the summary, the relevance perception varied between manual and automatic annotation, but automatically generated summaries were still found to be useful to access the voice forum audios.

Our main contributions in this paper were: (i) developing automatic annotation of voice forum data for rural users, (ii) developing a Hindi corpus of training material of community health workers and (iii) understanding user's perception, (relevance, preference) of automatic annotation in comparison with manual annotation through a field trial in northern India.

2 RELATED WORK

As a related work, we present some of the prominent works around voice forums to understand the need for automation followed by the applications of spoken content retrieval and underlying useful techniques

The proliferation of mobile phones in developing countries particularly in remote places, has led researchers to explore IVR-based voice forums as a tool for disseminating information [21, 23], training and education[30], providing social networking platform [11, 20, 26] and connecting to stakeholders such as government, NGOs etc. [8, 17]. Due to their relevance in existing scenarios of resource-constrained settings, some of the research initiatives have grown even bigger as socio-tech companies, covering a large scale of population. For instance, CGNET Swara [17] is a voice-based interface that extends the citizen journalism network through an IVR-based interface. Users can share their stories and concerns surrounding local issues which at the back-end are monitored by a dedicated team of moderators, who then perform the tasks of organizing and filtering. Likewise, Mobile Vaani [8], which provides IVR-based social media platform for discussion on a wide range of topics to users in 20 districts of India, uses the same model for moderation. Sangeet Swara [26], an entertainment forum for blind users mitigates the dependency on moderators by offloading moderation tasks to the forum users. When users listen to messages, they give their feedback in form of up votes and down votes, which are then used in computing the rank and playback order of the messages in the IVR application. So far, investigation of automatic techniques for the curation of voice forums in real-world applications has been lacking.

One of the main challenges in applying any machine-based processing is to get good quality of transcripts for the voice messages. While human-based transcription is time consuming and expensive, transcription accuracy provided by speech recognition engines is not satisfactory for local languages. Vashistha et. al [27] proposed a crowd-sourced system that enables people who speak and understand these languages particularly the low literates to transcribe in an easy manner. It works by assigning short utterances of an audio to multiple users, collecting clearer re-spoken versions and estimating the best transcript.

Considering the scenario of applications using well defined languages like English, there are some notable works that have developed applications for the access of spoken contents using transcripts. These are Podcastle [7] [19] for the searching of podcasts in Chinese and Japanese languages, MIT lecture browser [5] for topic-wise navigation of the course material, and NTU virtual instructor [10] for supporting on-demand learning by organizing the course lectures semantically. Typically, such works use text based techniques like topic modeling [28], clustering algorithms, content summarization, and visual encodings [1, 2] for organizing and presenting the content to the users.

Topic models have been increasingly used to characterize spoken content and various adaptations have been proposed to capture different speaking environments e.g. conference, lecture etc. Further, to provide efficient browsing over audio-only channels, audio summarization is an important task for which the standard techniques of text summarizing have been explored belonging to the categories, namely extractive and abstractive. While the former generates a summary by concatenating the important segments of the text, the latter applies linguistic methods to create condensed and syntactically correct formats. Extractive summaries are easier to create and are widely used. Both supervised and unsupervised approaches have been explored [29, 32, 33]. In comparison to well-structured forms such as broadcast news, summarization of spontaneous conversations is challenging due to the high rate of disfluencies, redundancies, and recognition errors [4]. Nevertheless, extractive summaries have been found to be effective in document retrieval [24]. Further, alternative strategies beyond transcripts such as use of information in speech signals have also been explored [9, 13].

3 DATASET

We base our implementation context around our broader goal of developing educational tools for Community Health Workers (CHWs) of rural India. In 2016-2017, we proposed and deployed a voice forum, Sangoshthi [30], for the training of ASHAs (a cadre of CHWs in India) which hosted ten talk shows on Home Based Post-Natal

Care. In this study, we used the audio dataset of Sangoshthi which was in Hindi language, and comprised two kinds of data: training material prepared by the experts (10 audios, 150 minutes) and the Q&A recordings (175 audios, 350 minutes). A Q&A audio was composed of a question asked by a health worker and its answer given by a doctor and had a duration of 2 minutes, on average. We used Google Speech Recognition engine for transcribing the audios considering its availability, ease of use, and coverage of languages [6]. The APIs offer both real time and offline transcription service along with a feature to get timestamps of the transcribed words, that we later used in constructing the audio summaries. The produced transcripts had an average word length as 129 (min 20, max 401) and average confidence score (estimation of the correctness of recognition) normalized by word count as 0.91 (min = 0.73, max = 0.93). The word error rate (WER) was computed as 67% on a subset of audios (41) that were selected randomly for manual transcription. The high noise in the transcripts was mainly due to the factors of type of communication channel (telephony), use of regional accents and the nature of speech which was spontaneous.

Further, we also created our domain specific corpus covering the training material of ASHAs of India, because corpora for Hindi language particularly of health domain are limited. The content was collected from various resources available at different websites of National Health Mission, the government organization managing the Community Health Workers program [16]. The text was cleaned and tokenized to form a corpus containing 12,513 sentences and 166201 tokens. We have made our corpus publicly available[1].

3.1 Pre-Processing

Before applying any text-processing techniques, the raw transcripts were pre-processed to filter out the noise. This involved stop-words removal, parts-of-speech (POS) tagging, stemming, and removal of words that were typical of the telephonic conversation (hello, welcome etc.). Post these steps, we selected nouns and adjectives for further processing. For tagging parts-of-speech we used the tagger for the Hindi language developed by [22].

4 METHODOLOGY

In this section, we describe the semi-automated approach of developing an IVR-based library consisting of audio recordings of the Q&As generated by the Sangoshthi voice forum. The structure of the IVR in terms of access was kept simple for this study such that the menu hierarchy consisted of only two levels. The first level was composed of the broad topics of the library and the second level that of the corresponding audio recordings. Figure 1 illustrates the steps followed, the overview of the two key steps of the curation process are:

i Theme classification - This step pertained to the task of finding themes in the dataset comprising Q&A audio recordings using topic modeling technique. This step mapped to the creation of the main menu of the IVR. Once the audio clusters were created, human intervention was used to construct labels for these themes.

[1]https://github.com/deepikay/ASHA_Corpus

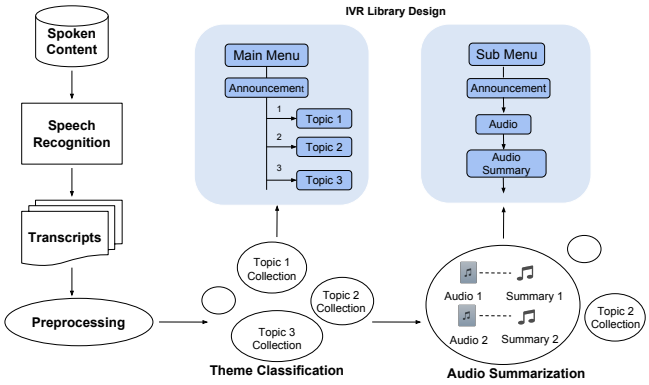

Figure 1: Curation Methodology

ii Audio Summary Creation - This step pertained to the task of constructing audio summaries of the audio recordings accessible on the second level of the IVR.

4.1 Theme Classification

Topic modeling is a widely used technique to discover themes in a collection of text documents. It is an unsupervised data-driven approach that takes bag of words as input and generates output in the form of document-topic and topic-word distributions. Topics are represented as sets of top-n words ordered by their marginal probabilities e.g. Topic 1: {newborn, breastfeed, milk, mother}, Topic 2: {hospital, medicine, doctor, vaccination}.

While deciding which topic modeling algorithm to apply on our dataset, considering its short transcripts, we had two choices: Latent Dirichlet Allocation (LDA) and Bi-Term Topic Modeling (BTM). LDA [3] is one of the most standard algorithm; however, gets affected by data sparsity found in short text documents such as messages, tweets etc. Whereas, BTM [31], which is an improvement over LDA, is able to handle the sparsity problem by modeling word co-occurrences explicitly at corpus level instead of document level as done in conventional topic models. Before adopting BTM, we verified empirically by applying both of the algorithms on our dataset. The analysis was performed using two popular metrics, namely coherence score [15] and PMI-score [18]. These metrics measure the quality of generated topics in terms of coherence between the constituent words, based on the underlying assumption that words describing a single concept tend to co-occur. Given a topic t and V^t as the list of K most probable words, the coherence score is calculated as:

$$C(t, V^t) = \sum_{k=2}^{K} \sum_{l=1}^{k-1} \log \frac{D(v_m^t, v_l^t) + 1}{D(v_l^t)} \quad (1)$$

where $D(v)$ is the number of documents in which the word v appears and $D(v, v')$ is the number of documents in which the words v and v' appear together. While the coherence score looks for word co-occurrences in the corpus on which the topic modeling is applied, which in our case were the Q&A transcripts of the Sangoshthi dataset, the PMI-score refers to external data sources, e.g., Wikipedia. For PMI-score calculation, we used two data sources,

covering the training material of the health workers(ASHAs), one was the corpus created by us and the other available in the Sangoshthi dataset. The formula of PMI-Score calculation for all unique word pairs is as follows:

$$\text{PMI-Score}(t) = \frac{1}{K(K-1)} \sum_{1 \le i < j \le K} \text{PMI}(w_i, w_j) \qquad (2)$$

where $\text{PMI}(w_i,w_j) = \log \frac{p(w_i, w_j)}{p(w_i)p(w_j)}$, $\text{p}(w_i,w_j)$ and $\text{p}(w_i)$ are the probabilities of co-occurrence of the word pair and the word, computed using a sliding window of word size 10 [18].

In this experiment, we compared the performance of LDA and BTM by varying topic size from 10 to 50. Since on every run of the algorithms, the output distributions change slightly, we considered the output of 10 runs in each topic size category. Hence, the overall score for a topic size was averaged over the runs. For the number of words in a topic, we followed the convention of representing a topic via its top-10 words as usually ten words are able to convey sufficient information about a topic [15, 18]. The computed coherence scores and PMI-scores for the LDA and BTM are shown in the table 1. Higher scores represent better coherency in the topics. Cleary, BTM performed better than LDA. Further, to finalize the number of topics for the library, topic size 10 was selected due to its high score. Also, ten topics seemed appropriate according to our prior knowledge that the Q&As were around the fixed ten topics on which the training was given in the Sangoshthi deployment [30]. Typically, selecting the number of topics is a heuristic based approach where expert's knowledge is used and developing an automatic metric is still an open problem [3].

Table 1: Topic Coherence Based Comparison of LDA and BTM

Topic Size	Coherence Score		PMI-Score	
	LDA	BTM	LDA	BTM
10	-89.4 ± 1.6	-72.9 ± 1.3	0.09 ± 0.03	0.08 ± 0.03
20	-85.8 ± 2.9	-77.3 ± 1.5	0.08 ± 0.05	0.08 ± 0.04
30	-84.2 ± 2.1	-78.6 ± 1.0	0.06 ± 0.04	0.08 ± 0.03
40	-81.7 ± 2.1	-78.5 ± 0.7	0.05 ± 0.02	0.07 ± 0.04
50	-79.1 ± 2.6	-79.0 ± 1.2	0.08 ± 0.04	0.05 ± 0.04
Mean Score	-84.10	-77.30	0.07	0.08

Topic Labeling — automating the creation of intuitive labels for the topics i.e. textual phrases, to make them easier to understand is a challenging problem. For the current study, we considered to use manual approach to produce more natural constructs, where, given a set of words for the topics, our goal was to use them in constructing general sentences with caution of adding no specific meaning. This activity was performed by the one of the author who had familiarity with the domain and the data. Table 2 presents the English translation of some of the selected topics with their labels. While constructing these sentence-style representations, not all the words in a topic were used, as few of them did not contain any useful meaning. For example, in constructing the phrase for the topic 2 - "baby, matter, mother, weight, nice, day, light, card, normal, quite so" as "this topic discusses about the weight and growth related matter of the newborn" the words "quite so", "nice" were not used as they did not hold any useful meaning for the context. Also, here the domain knowledge of the coder play an important role. Since the coder knew that the word "card" represented the MCP card (Mother Child Protection Card) which is used for monitoring the growth of a baby, she then constructed the phrase accordingly.

4.2 Audio Summarization

In a document collection, after the generation of themes and their labeling, document summarization is an important task to help users access the content in an efficient way. This is particularly important for audio-only interfaces where information is rendered sequentially.

Machine-Created Summaries - In our automated method of audio summary creation, we generated two types of summary, namely keyword-based summary and sentence-based summary. While the keywords-based summary annotate an audio with a group of words, the sentence-based summary annotate with a representative extract from the audio itself. The subsections below describes the methodology of constructing the two types of summaries.

4.2.1 Sentence-Based Summary. Extractive summarization process creates a summary by extracting important sentences from the input text and then concatenating them. In the speech domain, we approached this by first identifying useful sentences in a transcript and then fetching the audio segment corresponding to the most representative sentence. Although, generally extractive summaries are composed of multiple sentences per document, we selected only one sentence per transcript because the Q&A audios in our dataset were of short duration (2 minutes). The three step process is as follows:

(1) Selection of Candidate Keywords - Since the significance of a sentence is characterized by the significance of its constituent words, the selection of candidate sentences was preceded by the identification of keywords. Typically, candidate keywords are selected following heuristic rules, which generally include stopwords removal, POS-tagging, and selection of n-grams based on some criteria. We performed these steps in the pre-processing stage that gave us a collection of words belonging to two parts of speech as nouns and adjectives. Further, to extract important words among these, we used the tf-idf (Term Frequency-Inverse Document Frequency) statistic. The top 10% of the words ranked by tf-idf weight were selected as the candidate keywords for a transcript. The tf-idf technique is an effective technique, however, in case of noisy transcripts, it can lead to weighing of mis-recognized words as high. Nevertheless, this gets compensated when the corresponding audio segments are fetched.
(2) Selection of Representative Sentence - Now that we had a set of candidate keywords for a transcript, the selection of the most representative text segment was achieved by analyzing the neighborhood of the keywords. For each keyword location in its transcript, a window of 10 words with the keyword in the middle, was checked for the presence of rest of the keywords and scored on the basis of total keywords present. The score was computed as the sum of tf-idf weights of all the keywords that appeared in the window.

Here, the sum score was normalized by the count of keywords to avoid giving preference to the text segments having multiple keywords with low tf-idf ranks as opposed to the ones having one or few high rank keywords. Finally, the window with the highest score was selected as the most representative text segment of a transcript.

(3) Selection of Audio Segment - The audio summaries were generated by extracting the relevant audio portions for the selected text segments. For this, we used the timestamp information given in the transcripts. Since the text segments were not grammatically correct sentences due to missing speech recognition, we did not directly extract the audio segments from the start and end timestamps of the first and last word of the text segments. Instead, we fetched the audio segments between two natural pauses occurring before and after the first and last timestamps respectively. While extracting the audio segments, each was checked for its duration against the limits set as minimum of 4 seconds and maximum of 12 seconds. The duration thresholds were found out by conducting a lab testing. An audio segment not fulfilling the duration criteria was discarded, and the step would get repeated for the next preferred text segment.

4.2.2 Keywords-Based Summary. We explored another more abstract way of summary creation that represented the high level idea of an audio through a set of keywords. After selecting the required number of keywords for a transcript using the steps described in the previous subsection, an important question in constructing the audio form was whether to use machine synthesized speech or extract their utterances from the audios. To address this, we conducted a lab testing. Given a set of keywords for an audio, when their utterances were concatenated to form a single audio summary, it brought uneven transitions between consecutive words due to variation in pitch and background noise, leading to lack of clarity. On using Google text to speech conversion engine, we found better quality. However, the main concern in regard to text to speech conversion engine was because of the presence of the mis-recognized words which are directly used for speech production. To overcome this, we regenerated the keywords by incorporating an additional step of removing irrelevant words after the pre-processing step and before the tf-idf scoring. The preprocessed words (nouns and adjectives) per transcript were checked for their presence in the two corpora (ASHA training material) and the list of most-frequent words of the collection (taken as the top 10%). Although, this step ensured that no out-of-domain word were selected, it led to the loss of some useful words. We chose the number of keywords per audio to be seven following the guidelines by [14].

Human-Created Summaries - To evaluate the automatically generated summaries, we prepared a baseline of manually created summaries for an audio set that was later used in the field trial. This coding activity involved three participants - two authors and one master's student and was performed in two stages. In the first stage, all the coders individually generated their summary versions and in the second stage, one of the coders, who had better domain knowledge, selected the final version of the summaries. In the case of keywords-based summaries, the number of keywords per audio were consistent with the automatic method (seven). Once all the coders had tagged every audio with a set of seven keywords, the main coder selected the final set of keywords by following a procedure in which, she first hand-picked the common keywords with at least two coders followed by the addition of the remaining keywords by selecting from her set of keywords. The average number of common keywords found per audio was five. Note that, the keywords were selected from the audios and not constructed by the coders. In the case of sentence-based summary, the coders were asked to summarize every audio in one sentence which was similar to title creation. To generate the best summaries, the main coder, based on her judgment, either chose the best among the three summaries for an audio or constructed a new one by combining their idea.

5 EXPERIMENT DESIGN

In this section, we describe the details of the experiment that we conducted to evaluate our approach to curation.

5.1 User Judgments

To evaluate the usefulness of automation in real-world scenarios, we conducted a user study with the 48 community health workers, (ASHAs) in India and collected their judgments on the components of the library which were curated automatically. These judgments were of two types as follows:

- Relevance Judgments - Users' judgments of relevance were collected on two components of the library. One was on the quality of audio clustering that was associated with the audios and their allocated themes, and the other was on the quality of audio summaries. The judgments were collected on a three-point Likert scale, namely, 1—not relevant, 2—moderately relevant and 3—relevant.
- Preference judgments - Since users were made to listen two types of audio summaries (sentence and keywords), their preferences were collected on a four-point Likert scale as 1—sentence-based summary, 2—keywords-based summary, 3—both, 4—none.

5.2 Testbed

In order to do an effective evaluation, we designed our testbed in such a way that users were exposed to different testing conditions. Every user was supposed to listen to two versions of the library having audio summary types as sentence-based and keywords-based, which could be human created or machine created. Overall, we had four combinations of modalities as shown in the Table 3. Therefore, we divided 48 users into 4 groups, assigning each to one of the testing conditions randomly.

We used Freeswitch, an opensource telephony platform, to develop the IVR application of the library. Users were provided a phone number on which they could call and listen to the contents of the library. To make the calling activity free of cost for the users, the system was designed to work on the model of missed call. A user wishing to listen to the library, would have to drop a missed call to the given phone number and connect to the IVR application on immediate callback.

The users' judgments were collected in two phases corresponding to the evaluation of the two versions of the audio library. A single phase could be completed in one or more calls depending on the convenience of the users. The relevance judgments were collected while listening to the library content, and the preference judgments and overall feedback were collected at the end of the second phase.

5.3 Library Content Selection

Considering the userbase of community health workers, who have limited literacy and technological exposure, we tried to keep the design of the IVR application simple with small number of items. The first level of the IVR consisted of only three topics and the second level consisted of two audios under each topic. We selected the topics and the audios for the user study using the criteria of maximizing the coverage of audios in the dataset and keeping the minimum required number of users (four) for the evaluation of every item. Therefore, two sets of topics for the two versions of the library (sentence and keywords) were used in the study. These six topics were selected randomly from the set of ten topics generated by the topic modeling algorithm. In each of the four testing conditions (Table 3), 12 users were allocated, so, instead of using same audios for all the users, we divided the 12 users into 3 sub-groups and created three different audio sets to avoid the possible bias originating from the type of audio. The allocation of the audios to these three sets in a modality was done as follows. The first audio set consisted of the first and the second most probable audios of every topic, the second audio set consisted of the third and the fourth most probable audios and the third audio set consisted of the fifth and the sixth most probable audios. Hence, a total of 36 audios, divided into two groups, each consisting of three sub-sets of the audios were used across the testing conditions in our study. Finally, to avoid the sequence effect, we adopted the full counterbalancing strategy by altering the order of the modalities to be exposed to the users in the two phases of the assessments.

From the usability aspect, all the instructions in the IVR application were presented in a clear voice using colloquial Hindi language - the mother tongue of the participants. Instructions were repeated multiple times and had prompts to handle invalid inputs from the users while collecting judgments. The mapping of the keys to be pressed for giving the judgment ratings were made consistent everywhere in the application. Prior to the release of the application, a small pilot was conducted with two health workers and the suggested modifications were incorporated.

5.4 User Interaction Flow

We now describe the interaction of a user with the developed IVR-based library through a workflow diagram as shown in the Figure 2. On a successful call connection, the user is presented with the introduction of the library followed by an announcement of the topic list and the corresponding keys to be pressed for their selection. After the user has selected a topic, the first associated audio is played followed by its summary. At this point in time, user feedback on the summary quality is collected over three parameters as sound clarity, understandability, and relevance. Upon successful collection of ratings, the next audio and its summary are played if the content of the current topic is not finished, otherwise judgments on the association of the played audios and the assigned topic is collected. Every time on completion of a second level, the first level is brought back to present the topics list again. When all the topics are complete, the user is prompted to give an optional feedback by recording her voice. Finally, if the current call is the last call of the assessment, then preference judgment on the two modalities (keywords or sentence) of the summary is collected before exiting. To handle the cases of bad sound quality due to issues in the telephony network, user feedback on the sound quality is collected after the playback of every audio and if it is bad, the call is terminated by leaving a message to try again later.

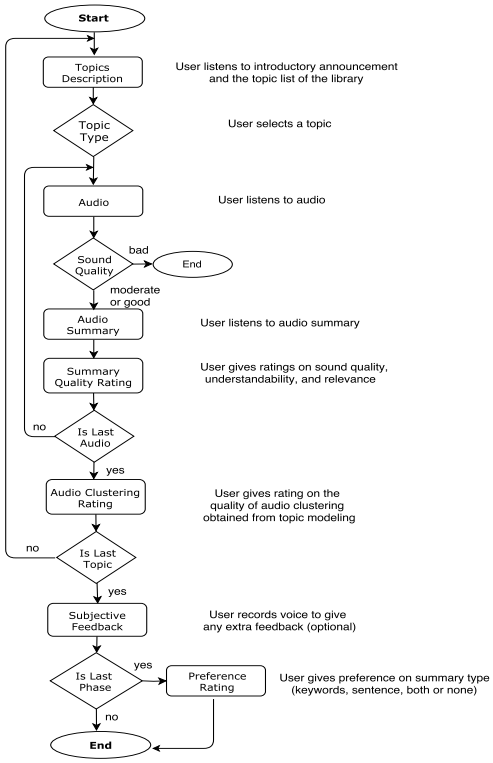

Figure 2: User Interaction Flow

6 PARTICIPANTS

We reached out to community health workers in the northern region of India with the help of an Non-Government Organization, SWACH [25]. Out of the 60 ASHAs (lowest cadre of CHWs) enrolled initially, only 32 could successfully complete the assigned tasks. The rest had to be dropped from the study due to reasons as follows: 8 ASHAs lacked the understanding of IVR systems, 10 ASHAs had call connectivity issues in their areas, and the other 10 had availability issues. Later, we enrolled 24 ANMs (Auxiliary Nurse Midwife), a higher cadre of CHWs, out of which 16 could complete both of their assessment phases. In total, we had 48 participants in our study. Among the ASHAs group, the average age

Table 2: Topics and Labels

Topic	Words Distribution	Label
1	baby, matter, milk, mother, night, nice, month, know, day, hospital	this topic discusses on the cases related to newborn going to hospital, care of newborn in the first month and feeding problems
	बच्चा, बात, दूध, माँ, रात, अच्छी, महीना, पता, दिन, अस्पताल	इस विषय में चर्चा हुई है, बच्चे के अस्पताल जाने के केसो के बारे में, बच्चे की पहले महीने में देख रेख के बारे में और दूध पिलाने के बारे में ।
2	baby, matter, mother, weight, nice, day, light, card, normal, quite so	this topic discusses about the weight and growth related topics of the newborn
	बच्चा, बात, माँ, वजन, अच्छी, दिन, हल्का, कार्ड, नॉर्मल, ठीक	इस विषय में चर्चा हुई है बच्चे के वज़न और ग्रोथ से जुडी बातो के बारे में ।
4	baby, matter, delivery, milk, mother, day, stomach, quite so, problem, water	this topic discusses about issues related to delivery, problems related to mother or baby and newborn feeding
	बच्चा, बात, डिलीवरी, दूध, माँ, दिन, पेट, ठीक, प्रॉब्लम, पानी	इस विषय में चर्चा हुई है माँ की डिलीवरी से जुडी बातो के बारे में, माँ या बच्चे की परेशानियों के बारे में और बच्चे को दूध पिलाने के बारे में ।
8	baby, matter, milk, mother, family, nice, important, quite so, water, problem	the topic discusses about important aspects related to the families of newborn, newborn feeding and problems related to mother and newborn
	बच्चा, बात, दूध, माँ, परिवार, अच्छी, जरूरी, ठीक, पानी, प्रॉब्लम	इस विषय में चर्चा हुई बच्चे के परिवार से जुडी बातें को बारे में, शिशु को दूध पिलाने के बारे में और माँ या बच्चे की परेशानियों के बारे में ।
9	baby, milk, matter, mother, day, month, harm, reason, important, problem	this topic discusses about newborn care in the first month, newborn feeding, and problems related to mother and baby
	बच्चा, दूध, बात, माँ, दिन, महीना, नुकसान, कारण, जरूरी, परेशानी	इस विषय में चर्चा हुई है बच्चे की पहले महीने में देखरेख के बारे में, बच्चे को दूध पिलाने और माँ या बच्चे की परेशानियों के बारे में ।
10	baby, matter, milk, day, harm, beautiful, nice, reason, mohter, eyes	this topic discusses about the five senses of newborn, newborn feeding, and problems related to mother and baby
	बच्चा, बात, दूध, दिन, महीना, नुकसान, खूबसूरत, अच्छी, कारण, माँ, आंखों	इस विषय में चर्चा हुई है बच्चे की इन्द्रियों के बारे में, बच्चे को को दूध पिलाने के बारे में और माँ या बच्चे की परेशानियों के बारे में ।

Table 3: Testing Conditions

Condition	Type 1	Type 2
1	human keywords	human sentence
2	human keywords	machine sentence
3	machine keywords	human sentence
4	machine keywords	machine sentence

was 38 (min = 25, max = 47), with distribution of educational backgrounds as 10%—8th standard, 47%—10th standard, 23%—12th standard and 20%— Bachelor's degree. Their experience ranged from 3 to 15 years (average = 9 years) with an exception of one having only 3 months. In the ANM group, the average age was 36 (min = 26, max = 56), with educational background as 25%—10 standard, 50%—12th standard, 12.5%—Bachelor's degree and 12.5%—Master's degree. The experience ranged from 1 to 31 years with an exception of one having only 6 months.

7 RESULTS

7.1 Users Perceptions of Summary Relevance

To effectively automate the process of creating audio summaries, it is necessary that users find them relevant enough so that the dependency on human moderators can be removed. Given the users' relevance judgments, we statistically analyzed the relationship between the relevance perceptions of the users for different types of summaries used in the study. Since the data were the subjective judgments collected on a 3-point Likert scale, we applied Fisher Exact test of independence between two categorical variables: relevance and summary−type, for different combinations of the testing conditions. Here, the contingency table for the two variables contains the distribution of frequencies for the three categories of the relevance measure, as "not relevant", "moderately relevant" and "relevant". The null hypothesis states that *the relevance perceptions of the users for a summary do not depend upon its type i.e. there exists no relationship.*

In the first test, we considered the testing combinations as {human keywords, machine keywords} (Table 4) and {human sentence, machine sentence} (Table 5) to evaluate the effectiveness of the automated approach with respect to the baseline of human-based approach. Here it can been noticed that these testing combinations were not directly assigned to the users; in all of the four testing conditions used in our study (Table 3), it was necessary that every user listened to both sentence and keywords type of summary. Therefore, we explicitly took out the frequency distributions from different groups for this test. For both the combinations, statistically significant results were found (p = 0.000, two-sided fisher-exact test and p = 0.000, two-sided fisher-exact test respectively). Thus, we reject the null hypothesis and accept the alternative hypothesis that relevance judgments gets affected by the type of summary where in this case, one was generated by human and other by machine. Changing the source of summary creator, change users judgments of relevance. This suggests the scope of improvement in the automated approach. In the current study, we applied basic techniques for the summary creation (extractive summarization), in future work, we would like to investigate the effectiveness of other advanced techniques.

Table 4: Relevance Judgments on Human Keywords vs Machine Keywords

	human keywords	machine keywords
not relevant	2	6
moderately relevant	15	39
relevant	127	99

* p = 0.000, relevance judgments are dependent on the type of summary

Table 5: Relevance Judgments on Human Sentence vs Machine Sentence

	human sentence	machine sentence
not relevant	2	3
moderately relevant	10	42
relevant	132	99

* p = 0.000, relevance judgments are dependent on the type of summary

The second test had the testing combinations as {human keywords, human sentence} (Table 6), and {machine keywords, machine sentence} (Table 7). Here, the summary type changes in its modality (keywords-based or sentence-based) but the source of creation either machine or human, remains the same within each combination. No significant results were found (p = 1, two-sided fisher-exact test and p = 0.179, two-sided fisher-exact test respectively). Thus, we fail to reject the null hypothesis and conclude that when the source of summary creation (human or machine) is same for the two modalities (keywords-based or sentence-based), then

Table 6: Relevance Judgments on Human Keywords vs Human Sentence

	human keywords	human sentence
not relevant	1	0
moderately relevant	3	4
relevant	68	68

*p=1, relevance judgments are independent of the type of summary

Table 7: Relevance Judgments on Machine Keywords vs Machine Sentence

	machine keywords	machine sentence
not relevant	5	1
moderately relevant	23	20
relevant	44	51

*p=0.179, relevance judgments are independent of the type of summary

users relevance perceptions for both the keywords and sentence type of summary remain similar.

Finally, the third test had the testing combinations as {human keywords, machine sentence} (Table 8) and {machine keywords, human sentence} (Table 9), where the summary type changes both in its modality and source of creation. No significant results were found on the relationship of summary type and the relevance judgments (p-value = 0.080, two-sided fisher-exact test and p = 0.052, two-sided fisher-exact test respectively). For the combination {human keywords, machine sentence} (p = 0.08), it can be inferred that representative extract of an audio as a summary is at least as good as the human created summary which is composed of keywords. Whereas, for the combination {machine keywords, human sentence} (p=0.052), though the result was non-significant, it was greater that the significance level by a small margin, indicating the effect of difference in quality of the type of summary. The machine-created keywords-base summary has downsides due to use of synthetic voice presentation and lack of structure (non-sentence) as compared to the best case of human created sentence-base summaries.

Table 8: Relevance Judgments on Human Keywords vs Machine Sentence

	human keywords	machine sentence
not relevant	1	2
moderately relevant	12	22
relevant	59	48

*p=0.080, relevance judgments are independent of the type of summary

7.2 Users Preferences for Summary Type

At the end of the second phase of assessment, when the users had been exposed to both of the assigned modalities (keywords-based and sentence-based), preference judgments were collected on a 4-point scale as 1-"sentence", 2-"keywords", 3-"both" and 4-"none". The overall distribution for all four testing conditions is shown in the Table 10. To test whether the observed distributions were different from the expected distributions of having an equal chance of choosing each judgment category, a multinomial goodness-of-fit test was applied. For the testing conditions as {human keywords, human sentence}, {human keywords, machine sentence}, {machine keywords, machine sentence}, non-significant results (p = 0.080, p = 0.169, and p = 0.407) were found which indicated the occurrence of the preference distributions by chance. For the testing condition of {machine keywords, human sentence}, though the result was significant (p = 0.03) supporting the occurrence of the sum distribution not by chance, the post-hoc tests (Exact Binomial Test with Holm method of adjustment) for the categories (sentence, keywords, both, none) individually gave non-significant results with p-values as 0.050, 1.000, 1.000, 0.130. Thus, we cannot infer the preferences of the users statistically.

Table 9: Relevance Judgments on Machine Keywords vs Human Sentence

	machine keywords	human sentence
not relevant	1	2
moderately relevant	16	6
relevant	55	64

*p=0.052, relevance judgments are independent of the type of summary

Therefore, to get some insight, we qualitatively analyzed the preference views of the 16 users who were interviewed (selected randomly). Out of these 16 users (ASHAs), 11 users mentioned their preference for sentence type of summaries, 3 users for keywords type and 2 for both. The users who preferred sentence-based summary commonly said that sentence modality is clearer in understanding, because it gives complete information. One of the user of this group highlighted a negative aspect for the keywords-based summary by saying that *"keywords-based summary is not good because the constituent words do not match with the audio"*. Whereas, the other user group who favored keywords-based summary considered the spacing between the consecutive words as a positive point towards better understanding. A supporting quote is as follows *"I found the keywords-based summary to be better as they are presented cleanly in form of separate words as compared to the sentence-based summary which can start and end abruptly"*. One of the users who favored both said *"for me both types of summary are good enough; however for other ASHAs, sentence summary is better as many of them have low literacy and comprehension skills"*. Overall, the tendency of the users was towards the sentence-based summary.

7.3 Effectiveness of Theme Classification

In our design of the library, the first level in the IVR structure presented three broad topics using abstract descriptions around the words generated by the topic modeling algorithm. User feedback on the association of the topics and the audios was collected at the

Table 10: Preference Judgments

	human keywords, human sentence	human keywords, machine sentence	machine keywords, human sentence	machine keywords, machine sentence
sentence	5	5	7	4
keywords	2	4	3	5
both	5	3	2	2
none	0	0	0	1

end of the second level when all the audios and their corresponding summaries had been played. The measuring Likert scale was 3-point as 1—"not relevant", 2—"moderately relevant", 3—"relevant".

The observed distribution of the relevance judgments is given in the Table 11 below. A multinomial goodness-of-fit test gave statistically significant result, thus, rejecting the possibility of occurrence by chance (p = 0.000, significance level = 0.05), followed by similar results of the post hoc tests for all the three categories individually (not relevant, p = 0.0; moderately relevant, p = 0.0; relevant, p = 0.0). Overall, the clustering of the audios to the topics was found to be relevant by a majority of the users.

Table 11: Relevance Judgments for Theme Classification

not relevant	moderately relevant	relevant
7	50	221

7.4 Relationship of Sound Quality Judgments with Summary Modality and Relevance Judgments

We further analyzed the sound quality ratings given by the users. First, we tested the relationship between the sound quality judgments and summary type. On applying the test of independence between the two variables, we found significant results for the combinations - {machine keywords, machine sentence} (p = 0.031, two-sided fisher-exact test) and {machine keywords, human sentence} (p=0.0, two-sided fisher-exact test). The corresponding contingency tables are given in the Table 12 and Table 13 respectively. Thus, we can reject the null hypothesis and conclude the dependency between the sound quality judgments and the type of summary. As we can observe, both the conditions have machine keywords modality, this indicates the effect of synthesized speech on the perception of sound quality.

Table 12: Sound Quality Judgments on Machine Keywords vs Machine Sentence

	machine keywords	machine sentence
low	9	10
moderate	28	14
high	35	48

*p=0.031, sound quality judgments and summary type are dependent

Next, we tested the relationship between the sound quality and relevance judgments of the summaries for the four modalities of a summary (human keywords, human sentence, machine keywords and machine sentence). Statistically significant results were found for the modalities - machine sentence (p=0.0, two-sided fisher-exact test) (Table 14) and machine keywords (p=0.0, two-sided fisher-exact test) (Table 15). Whereas, for the modalities - human keywords and human sentence, the results were non-significant (p=0.144, two-sided fisher-exact test and p=0.0632, two-sided fisher-exact test respectively). For conciseness, we present the contingency table only for the significant results. We can observe that the sound quality and relevance judgments are dependent on each other for machine-based summaries as opposed to human-based summaries. This again highlights the impact of the machine produced voice and summaries which needs to be further investigated through a detailed experiment.

Table 13: Sound Quality Judgments on Machine Keywords vs Human Sentence

	machine keywords	human sentence
low	12	1
moderate	22	12
high	38	66

*p=0.0, sound quality judgments and summary type are dependent

Table 14: Judgments on Sound Quality vs Relevance for Machine Sentence

	sound quality	relevance quality
low	16	3
moderate	24	42
high	104	99

*p=0.0, judgments on sound quality and relevance are dependent

Table 15: Judgments on Sound Quality vs Relevance for Machine Keywords

	sound quality	relevance quality
low	21	6
moderate	50	39
high	73	99

*p=0.0, judgments on sound quality and relevance are dependent

8 LIMITATIONS

There were certain limitations in the study, that were related to the training need of health workers, cellular infrastructure, sample size and data analysis. We deployed our application with no prior training on the use of IVR systems to the community health workers due to which we had to discard some of the collected users data. The ASHAs, who serve the lowest cadre of community health workers, are only village women with low education backgrounds, some of them found difficulty in using the IVR application. For instance, one health worker after listening to the instruction of choosing a topic from the list used to press all the keys assigned,

few others, instead of pressing keys gave their inputs vocally and misunderstood the instructions during the activity of judging the quality of audio summaries. In order to avoid the inclusion of such users in the analysis, we cross-verified by interviewing all the users and then discarded the data of those who were not able to explain the instructions correctly or seemed confused (a total of eight users).

During this study, few users experienced challenges with connectivity and poor call quality. This has been a known challenge in India [12]. Because these users had to reconnect multiple times while using the application, they ultimately stopped using it. Data from 10 users had to be excluded from the analysis due to call quality issues. The other limitation of our study was the sample size of the users. Though we reached out to 84 health workers, only 48 could successfully complete their assessments.

Finally, during the field trial, though we collected understandability feedback from the users for the audio summaries, we did not analyze it completely and did not report in the paper. In the future, we would like to investigate the correlation between the users' perception about the three parameters - sound quality, understandability and relevance in greater detail.

9 DISCUSSION

Our experiment towards automating the automatic annotation of the voice forum data to reduce the manual effort raised some interesting future directions that we discuss as follows:

9.1 Use of Topic Models for Voice Forums

We used topic modeling in finding out themes in the audio collection and presenting the catalog of the library by labeling them. There are multiple facets of topic modeling that pose both the challenges and opportunities for research as the scale of voice forum increases. The first, is the task of deciding the number of clusters/topics. When the scope of a voice forum is narrow then heuristic based estimation of the number of topics seem feasible. For instance, in our study, given the dataset of Q&A of the Sangoshthi forum, we had the prior knowledge of the number of topics on which these Q&As were based that we used in finalizing the topic size. However, in forums like CGNetSwara [17] and MobileVaani [8], which pertain to a spectrum of topics, dynamic updation of the clustering process with right parameters is a challenging task. The second, is the task of creating interpretable representations of topic words. In the current implementation, we manually created topic representations by constructing abstract sentences for the topic words. However, we would like to further explore other better ways. While for the natural language processing and machine learning research, there are opportunities to come up with techniques to automatically generate better representations, the design research can explore the usability aspects for the user groups of context similar to rural communities of developing countries.

9.2 Human-in-the-loop

In this study, as we see that human intervention is required to fill the gaps that machines cannot do effectively. We need to think of better frameworks for involving human effort. One potential way in our application context could be crowd-sourcing of the tasks that can improve the performance of the automated techniques. For example, providing representative phrases for the topics, tagging audio contents with keywords, etc. While designing such crowd-sourcing applications, it is necessary to address the challenges of it as well as design mechanisms that incur minimum overhead to the system. For instance, instead of developing a crowdsourcing application for curation tasks in silo, it is more appropriate to engage the users during their application use. SangeetSwara is a nice example that collected up votes and down votes from the forum users while presenting the contents to them.

Further, going forward, with increasing penetration of smartphones in developing countries, we will also have to think of better modality to present the automatically curated information on visual interfaces.

10 CONCLUSION

In this paper our goal was to explore the automatic annotation of the voice forum data towards building an IVR library. Voice forums and its applications are actively studied for developing countries. However, automatic annotation of voice forum content is still an unsolved task. This paper evaluates the effectiveness of techniques of natural language processing and information retrieval fields to address this challenging problem.

We applied standard techniques to voice forum data, and specifically topic modeling for generating themes, and extractive summarization for constructing summaries. The transcript extraction had a high noise (WER = 67%), and we employed additional preprocessing steps to extract relevant keywords. Our results are promising in terms of users relevance for audio clustering and summary annotation. The evaluation results are encouraging for using automatic annotation in voice-based applications for rural users.

11 ACKNOWLEDGMENTS

We would like to thank the NGO-SWACH, all the health workers and our funding partners as ITRA project, DEITy, Government of India, under grant with Ref. No. ITRA/15(57)/Mobile/HumanSense/01 and Visvesvarya Young Faculty Fellowship to support this research.

REFERENCES

[1] Fahmi Abdulhamid and Stuart Marshall. 2013. Treemaps to visualise and navigate speech audio. In *Proceedings of the 25th Australian Computer-Human Interaction Conference: Augmentation, Application, Innovation, Collaboration.* ACM, 555–564.

[2] Eric Alexander and Michael Gleicher. 2016. Assessing topic representations for gist-forming. In *Proceedings of the International Working Conference on Advanced Visual Interfaces.* ACM, 100–107.

[3] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3, Jan (2003), 993–1022.

[4] Sadaoki Furui, Tomonori Kikuchi, Yosuke Shinnaka, and Chiori Hori. 2004. Speech-to-text and speech-to-speech summarization of spontaneous speech. *IEEE Transactions on Speech and Audio Processing* 12, 4 (2004), 401–408.

[5] James Glass, Timothy J Hazen, Scott Cyphers, Igor Malioutov, David Huynh, and Regina Barzilay. 2007. Recent progress in the MIT spoken lecture processing project. In *Eighth Annual Conference of the International Speech Communication Association.*

[6] Google. 2018. Google Cloud Speech API. https://cloud.google.com/speech/reference/rest/ [Online; accessed 02-March-2018].

[7] Masataka Goto, Jun Ogata, and Kouichirou Eto. 2007. Podcastle: A web 2.0 approach to speech recognition research. In *Eighth Annual Conference of the International Speech Communication Association.*

[8] Gramvaani. 2018. How Mobile Vaani Works. http://gramvaani.org/?page_id=15 [Online; accessed 19-April-2018].

[9] Guillaume Gravier, Nathan Souviraa-Labastie, Sébastien Campion, and Frédéric Bimbot. 2014. Audio thumbnails for spoken content without transcription based on a maximum motif coverage criterion. In *Annual Conference of the International Speech Communication Association*.

[10] Sheng-yi Kong, Miao-ru Wu, Che-kuang Lin, Yi-sheng Fu, and Lin-shan Lee. 2009. Learning on demand-course lecture distillation by information extraction and semantic structuring for spoken documents. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*. IEEE, 4709–4712.

[11] Zahir Koradia, Piyush Aggarwal, Aaditeshwar Seth, and Gaurav Luthra. 2013. Gurgaon idol: a singing competition over community radio and IVRS. In *Proceedings of the 3rd ACM Symposium on Computing for Development*. ACM, 6.

[12] Zahir Koradia, Goutham Mannava, Aravindh Raman, Gaurav Aggarwal, Vinay Ribeiro, Aaditeshwar Seth, Sebastian Ardon, Anirban Mahanti, and Sipat Triukose. 2013. First impressions on the state of cellular data connectivity in India. In *Proceedings of the 4th Annual Symposium on Computing for Development*. ACM, 3.

[13] Sameer Maskey and Julia Hirschberg. 2005. Comparing lexical, acoustic/prosodic, structural and discourse features for speech summarization. In *Ninth European Conference on Speech Communication and Technology*.

[14] George A Miller. 1956. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review* 63, 2 (1956), 81.

[15] David Mimno, Hanna M Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. In *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, 262–272.

[16] Government of India Ministry of Health & Family Welfare. 2018. About ASHA. http://nhm.gov.in/communitisation/asha/about-asha.html [Online; accessed 2-March-2018].

[17] Preeti Mudliar, Jonathan Donner, and William Thies. 2012. Emergent practices around CGNet Swara, voice forum for citizen journalism in rural India. In *Proceedings of the Fifth International Conference on Information and Communication Technologies and Development*. ACM, 159–168.

[18] David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 100–108.

[19] Jun Ogata and Masataka Goto. 2009. PodCastle: Collaborative training of acoustic models on the basis of wisdom of crowds for podcast transcription. In *Tenth Annual Conference of the International Speech Communication Association*.

[20] Neil Patel, Deepti Chittamuru, Anupam Jain, Paresh Dave, and Tapan S Parikh. 2010. Avaaj otalo: a field study of an interactive voice forum for small farmers in rural india. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 733–742.

[21] Agha Ali Raza, Farhan Ul Haq, Zain Tariq, Mansoor Pervaiz, Samia Razaq, Umar Saif, and Roni Rosenfeld. 2013. Job opportunities through entertainment: Virally spread speech-based services for low-literate users. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2803–2812.

[22] Siva Reddy and Serge Sharoff. 2011. Cross Language POS Taggers (and other Tools) for Indian Languages: An Experiment with Kannada using Telugu Resources. In *Proceedings of the Fifth International Workshop On Cross Lingual Information Access*. Asian Federation of Natural Language Processing, Chiang Mai, Thailand, 11–19. http://www.aclweb.org/anthology/W11-3603

[23] Jahanzeb Sherwani, Nosheen Ali, Sarwat Mirza, Anjum Fatma, Yousuf Memon, Mehtab Karim, Rahul Tongia, and Roni Rosenfeld. 2007. Healthline: Speech-based access to health information by low-literate users. In *Information and Communication Technologies and Development, 2007. ICTD 2007. International Conference on*. IEEE, 1–9.

[24] Damiano Spina, Johanne R Trippas, Lawrence Cavedon, and Mark Sanderson. 2017. Extracting audio summaries to support effective spoken document search. *Journal of the Association for Information Science and Technology* 68, 9 (2017), 2101–2115.

[25] India SWACH, Panchkula. 2018. Survival for Women & Children Foundation. http://www.swach.org/ [Online; accessed 2-March-2018].

[26] Aditya Vashistha, Edward Cutrell, Gaetano Borriello, and William Thies. 2015. Sangeet swara: A community-moderated voice forum in rural india. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 417–426.

[27] Aditya Vashistha, Pooja Sethi, and Richard Anderson. 2017. Respeak: A Voice-based, Crowd-powered Speech Transcription System. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 1855–1866.

[28] Furu Wei, Shixia Liu, Yangqiu Song, Shimei Pan, Michelle X Zhou, Weihong Qian, Lei Shi, Li Tan, and Qiang Zhang. 2010. Tiara: a visual exploratory text analytic system. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 153–162.

[29] Shasha Xie and Yang Liu. 2010. Improving supervised learning for meeting summarization using sampling and regression. *Computer Speech & Language* 24, 3 (2010), 495–514.

[30] Deepika Yadav, Pushpendra Singh, Kyle Montague, Vijay Kumar, Deepak Sood, Madeline Balaam, Drishti Sharma, Mona Duggal, Tom Bartindale, Delvin Varghese, et al. 2017. Sangoshthi: Empowering Community Health Workers through Peer Learning in Rural India. In *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 499–508.

[31] Xiaohui Yan, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng. 2013. A biterm topic model for short texts. In *Proceedings of the 22nd international conference on World Wide Web*. ACM, 1445–1456.

[32] Jian Zhang and Pascale Fung. 2007. Speech summarization without lexical features for Mandarin broadcast news. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*. Association for Computational Linguistics, 213–216.

[33] Justin Jian Zhang and Pascale Fung. 2012. Active learning with semi-automatic annotation for extractive speech summarization. *ACM Transactions on Speech and Language Processing (TSLP)* 8, 4 (2012), 6.