Field Trial, Evaluation and Error Correction methods of an IVR based Commodity Price Retrieval System

Soma Khan
CDAC, Kolkata
Plot-E2/1, Block-GP
Salt Lake, Sector-V, Kolkata-91
soma.khan@cdac.in

Joyanta Basu
CDAC, Kolkata
Plot-E2/1, Block-GP
Salt Lake, Sector-V, Kolkata-91
joyanta.basu@cdac.in

Milton Samirakshma Bepari
CDAC, Kolkata
Plot-E2/1, Block-GP
Salt Lake, Sector-V, Kolkata-91
milton.bepari@cdac.in

Rajib Roy
CDAC, Kolkata
Plot-E2/1, Block-GP
Salt Lake, Sector-V, Kolkata-91
rajib.roy@cdac.in

Abstract

Present study illustrates the entire evaluation and improvement process of an IVR based agricultural commodity price information retrieval system developed mainly for the semi-literate or illiterate farmers. Like evaluation of any real world speech recognition application, the system also has to face challenge regarding spoken language conventions, pronunciation variations, recognition in noisy environment, limitations of human cognition, working memory and above all inexperienced users. In view of these challenges, novel evaluation strategies considering all possible perspectives and situations with new set of metrics are designed here and validated accordingly. Analysis of user feedback in successive field trials and system performance results are presented in a well organized way. Additionally to meet users' expectations, distinctive error correction methods like Signal Analysis and Decision, Confidence Measure and Polling, Complementary Information, Runtime model generation etc. are introduced and incorporated to confirm performance enhancement in the final trial. Evaluation methods and metrics used here are domain independent and applicable to similar systems.

1 Introduction

Within the past several years, extensive growth in worldwide mobile phone usage and coverage has given greater prospects to Interactive Voice Response (IVR) technologies in solving various technological challenges in the areas of health, transport, financial assistance, disaster management and education. While technological advents with newer means of communication enable multiple ways of information access for literate users, there are still fewer options for low literate users. In fact, at the end of 2011, compared to just 2.3 billion internet connections, number of mobile users have already reached to 6 billion[1] which was mainly driven by developing countries accounting more than 80% of the new connections added in that year. These figures clearly indicate the scope and potential of mobile IVR based information access applications especially for low literate users who can't use or afford computers and internet. Here Automatic Speech Recognition (ASR) can remarkably improve such IVR utilization because of its hands free usage, implicit query response, shorter call

[1] ITU statistics: http://www.itu.int/ITU-D/ict/statistics/material/pdf/2011%20Statistical%20highlights_June_2012.pdf, 2012

duration, natural conversational UI and therefore enhanced user satisfaction.

However, even in best circumstances, voice interfaces suffer a number of usability challenges regarding spoken language conventions, pronunciation variations, limitations of ASR in noisy environment, limitations of human cognition, working memory and differences between users. A number of prior studies have compared voice and touch-type input modalities with varying results (Patel et al., 2009; Sherwani et al., 2009). Practically, it is still under study that how these different modalities need to be balanced for applications involving inexperienced subjects. It is also well established, that most of the existing IVR systems designed for the developed world target neither the needs nor the usability challenges faced by people in resource and literacy-poor regions (Plauche and Nallasamy, 2007). Field trial analysis in some of the related studies (Adam et al., 2010; Plauche and Nallasamy, 2007) also focuses these issues. But besides the comparison between different modalities, often the adopted evaluation strategies are highly task or domain dependent and usually end up with performance measures like Word Error Rate (WER) and Call Success Rate (CSR) only. In such a situation, not only application development, emphasis should be given equally on designing domain independent evaluation strategies, user adaptation methods by successive field trials and usability challenges faced during entire deployment process so that novel error recovery methods can be explored and incorporated into future developments.

2 Objective of the work

Present study elaborately describes the entire evaluation process of an IVR based agricultural information retrieval system mainly targeted towards the semi literate or illiterate farmers. The core of the application is a real-time, multiple pronunciation connected word ASR system that provides market wise latest prices of commonly grown agricultural commodities on recognition of spoken district and commodity names in native (here it is Bangla) language over telephone. Figure 1 shows the functional block diagram of the initial system.

According to year-2000 Indian census, agriculture is serving almost 600 million households in rural India (Sherwani et al., 2009). Across the developing countries, farmers and other agricultural workers constitute over 40% of the total labour force (Plauche and Nallasamy, 2007). So, targeting such a larger mass of people will definitely raise high expectations. Here field trial and evaluation of any real world application are likely to be a daunting task because most of the formal users are inexperienced who lack formal education and computer literacy. To cope up with this situation, this paper presents the outline of evaluation strategies being adopted, details of field trial processes and finally performance results in a well organized way. Additionally to meet the user expectations and real world requirements, distinctive error correction methods like Signal Analysis and Decision (SAD), Confidence Measure (CM) of recognized output, Complementary Information, Runtime model generation are introduced and evaluated respectively.

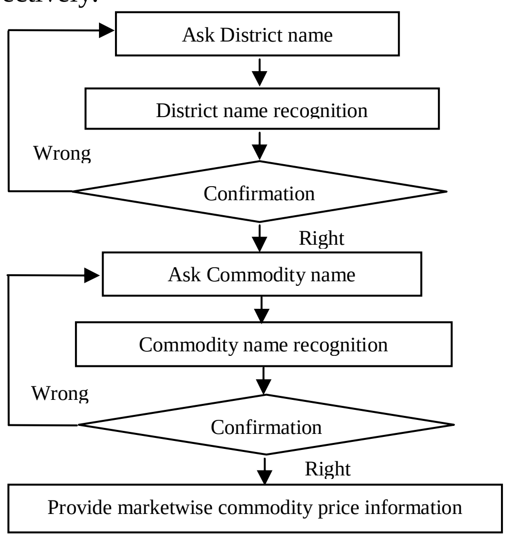

Figure1. Functional block diagram of the initial system

Rest of the paper is organized as follows. Section 3 describes the evaluation strategy being defined before starting the actual evaluation. Section 4 presents the details of experimental setup, offline evaluations, successive field trials and feedback analysis. From the user feedbacks, some novel error correction techniques have been introduced and illustrated in section 5. Section 6 depicts the final results of the evaluation processes after incorporation of the error correction techniques with discussions. Finally conclusions are summarized in section 7.

3 Evaluation Strategy

At very first, it is important to understand the various perspectives from which the system's performance needs to be evaluated so that the sequence, objective and scope of each evaluation can be defined prior to field trial. Evaluation metrics should also be decided to fulfill the expectations from different perspectives. Since user and system are the two participating entities in any automated spoken dialogue system, it is obvious to have mutual expectations from each end. Again these expectations, to what extent need to be fulfilled for successful completion of the task, should also be evaluated. So clearly, there can be three major perspectives:

3.1 User's perspective

These set of metrics are mainly designed to meet user's expectations from system, regarding quality of performance and to justify proper system functioning in field trials.

- Information adequacy: evaluates whether information provided by the system is sufficient for the purpose.
- Response time: evaluates whether the time taken by the system to respond to a particular query (including recognition time) is acceptable.
- Understandability of prompts: evaluates whether the instructions are meaningful enough to convey the information or what needs to be done at any instance.
- User friendliness: evaluates user's comfort level with the system like the sequence of queries, proper guidance throughout the interaction, ease of use etc.
- Usefulness in real-life scenario: evaluates how far the system is useful to fulfill user's real life requirement.
- Overall experience: evaluates user's overall satisfaction level with the system.

3.2 System's perspective

These evaluations are important speacially from system point of view and will serve as quantitative evaluation of system's performance in different co-operative and non co-operative situations. Here, the commonly used metrics are Percentage Correct ($\%W_C$), Percentage Error or Word Error Rate (WER) and Percentage Word Accuracy ($\% W_{Acc}$).

$\%W_C = (W_C / W_{Ref}) * 100$

$= (W_{Ref} - (W_S + W_D)) / W_{Ref} * 100$

$WER = ((W_S + W_D + W_I) / W_{Ref}) * 100$ and

$\%W_{Acc} = 100 - WER$

where,

W_C = number of correctly recognized words

W_{Ref} = total number of words in Reference

W_S = number of words Substituted

W_D = number of words Deleted

W_I = number of words Inserted

Here we will be using mostly the word accuracy.

Before starting the actual field trial, some preliminary evaluations of the initially build system is necessary to enhance the possibilities of getting acceptable results on further evaluations. These offline evaluations will include:

- Cross validation: This assessment technique can be used to check how well the model built on training dataset will generalize to new data. Generally k fold cross validation testing is used for this purpose.
- Parameter tuning: Experimental evaluation can be done by changing the parameters of Acoustic Model (AM) training and type of Language Model (LM) that best suit the amount and type of collected speech data.

After offline testing, system's efficiency and consistency of performance needs to be evaluated on live data over a period of time. At field trials, following online evaluations can be carried out with active user participation initially with co-operative speaking:

- Guided speaking: users will be guided by field volunteers, instructed how to use or shown demonstration and provided pictures of vocabulary words prior of calling.
- Free speaking: no guidance, instruction or demonstration by field volunteers, users will handle calls by themselves and are free to say anything within the context.

Irrespective of system's expectations regarding proper usage and environment conditions, additionally some robustness evaluations also need to be done to see system's performance in extreme situations. Some of these situations can also be found in free speaking but exhibit more unusual and non co–operative speaking or some unfavorable environment condition. Examples of similar situations are:

- Total silence: no response from user end at all.
- Disfluencies: hesitation sounds like "uhm...ah.." with valid words, thinking while speaking.
- Stammering: repeating part of the valid words, due to speaking difficulty, nervousness, excitement etc.
- Repetition: speaking the same word again and again, not satisfied or conscious speaking.
- Extra words: out of vocabulary words with valid words, casual speaking.
- Distant speech: speaking from a distance, phone held on the palm.
- Extreme noise: filed noises of pump, tractors etc., background announcements in loud speaker, television or radio playing in full volume.

3.3 Overall task completion perspective

Finally, fulfilling mutual expectations of each other, overall how well the user and the system both have performed for successful completion of the task need to be evaluated. The metrics used for this purpose is percentage success out of the number of total calls and valid calls. It measures the percentage success count of the system in providing the required information to the speakers. Such information either can be the latest price that is available in the stored database or just saying that it is not available right now.

These evaluation strategies are equally applicable to any ASR supported spoken dialogue system of similar kind.

4 Evaluation Process

After deciding the outline of the evaluation strategy it is necessary to validate the metrics being thought of, through properly designed offline evaluations and successive field trials. Analysis of user feedbacks can then only be helpful to design possible error correction methods.

4.1 Experimental Setup

Prior to evaluation, system development process has been completed using the following architectural and experimental setup. Asterisk (Gomillion and Dempster, 2005) is used here as an open source IVR Server, converged telephony platform, that primarily runs on Linux. With a properly designed call flow (using PHP-AGI scripting) and local language IVR prompts, real time speech data have been collected from about 2938 speakers (mostly farmers) in all over the state. Collected data was transcribed and manually verified by using a Semi Automatic Transcription Tool designed in house. Thereafter backend ASR models were built using CMU SPHINX[2] toolkit, an open source ASR engine. Commodity prices were regularly crawled from Agmarknet website using a Web-Crawler and stored in local MySQL database using python-MySQLdb connector. With an initially designed functional call flow (figure 1) the live system is finally made ready to be used in initial field trials. But in house offline evaluations from system's perspective are carried out before going for the field trials.

4.2 Offline Evaluations

At first, a three fold cross validation testing is done by partitioning the entire training dataset into three orthogonal subsets (Part 1, 2 and 3). With default values of AM training parameters, data from any two subsets are merged and trained subsequently to have three different AM. Trigram LM is created out of 360 task vocabulary words. Table 1 shows the results of this testing. Though percent correct in each case is around 80%, percent word accuracy is heavily affected by large number of insertions. This may occur due to insufficient modeling of speech data and real world background noises. But models can be improved on successive training by adjusting the model training parameters according to the type and amount of speech data.

Next in parameter tuning, default values of acoustic model training parameters are experimentally changed. The number of tied states (senones)

[2] http://www.speech.cs.cmu.edu/

and Gaussian mixtures are varied at training time to obtain the final AM having optimal performance. The experiment is done on a randomly selected subset of the train dataset using two different LM namely Finite State Grammar (FSG) and statistical n-gram (here Trigram). Table 2 shows the experiment details and figure 2 represents the results of this evaluation.

Partitions	Hours of train data	# of test utterances	% correct	% accuracy
Train: Part 1, 2; test: part 3	65.28	20870	79.81	58.80
Train: Part 1, 3; test: part 2	65.49	20404	84.76	42.90
Train: Part 2, 3; test: part 1	64.83	21336	83.08	61.47

Table 1. Results of three fold cross validation testing

Hours of training	Number of Train utterances	Number of Test utterances	Unique words	Vocabulary size
98 hrs	138775	18504	360	4992

Table 2. Specification of offline parameter tuning experiment

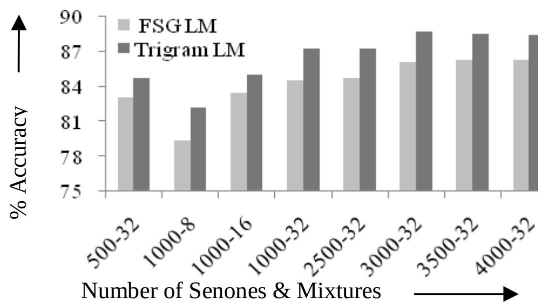

Figure 2. Results of varying LM type and AM parameters.

With increasing number of tied states, actual variations in the dataset were modeled more precisely which significantly reduced the number of insertions to boost up the word accuracy. Highest percent accuracy (88.75% using Trigram LM and 86.08% using FSG LM) is obtained using 3000 senones and 32 mixtures. Trigram LM is seen to perform better in each case and hence it has been used in the final system. Phase 1 and 2 field trials were conducted using a single AM trained according to the specification given in Table 3.

Parameter	Value	Parameter	Value
Features	39 MFCC	Pre-emphasis co-efficient	0.97
Frame Rate	100 Frames/sec	Lower cut-off frequency	130 Hz
Window Length	25 ms	Upper cut-off frequency	3500 Hz
FFT Size	256	Vocab. size	4992
Sampling Rate	8000 Hz	Number of tied states (senones)	3000
Mel Filters	31	Number of mixtures	32

Table 3. Final specification of AM training

4.3 Details of field trial

Field trial is being conducted in three phases having predefined objectives with increasing complexity level each time. Phase 1 field trial was conducted within two districts while the next two phases were carried out on fresh new speakers from all over the state. Field volunteers (agent) are sent in a team of 2-3 persons in each district parallely.

In the first phase, before introducing the system, users are first interviewed with small set of queries to answer. These questionnaires were specifically designed to understand the present scenario, analyze the need, user's response and expectations from this kind of an automated system. Some of the questions were like following:

- What are the agricultural commodities they cultivate throughout the year?
- How they currently get to know the recent market price of agricultural commodities?
- What are their ideas about any solution for this purpose?
- What would be their expectations from an automated solution for this purpose?

After this, users are introduced to the system by the field volunteers by showing a quick demonstration of the prototype system. Users are then requested to initiate a similar call by themselves with different query words (for district and commodity) and complete at least an entire cycle until getting the resultant response from system. At the end, qualitative evaluation metrics are presented and users are asked to give their judgment on each metric in a five point grading scale of remarks.

Second phase of field trial was solely meant for online quantitative evaluations. Here users are first requested for guided and then free speaking described in the previous section. This was done regularly for a two month period maintaining day-to-day system call log. Analyzing this log file, results are summarized in terms of percent accuracy and percentage success in task completion.

Finally, Phase three of the field trial was conducted after incorporation of error correction methods. This includes repeating the evaluations in both the previous phases with added non co-operative speaking scenario. This was continued for another two months on a new set of speakers.

4.4 Analysis of field trial

The preliminary question-answer session was found to be very much useful for getting bias less response from actual users. In fact, after this session, some of the metrics mentioned above in the qualitative and break testing category are refined or added for better evaluation purpose. Analysis of first two field trials revealed following facts:

- Most of the users were not aware of Agmarknet website[3]. To get the market price, presently they need to make a trip to the local market or wait for a reply from human operators on Kisan call centre toll free number. They were happy to have such an automated system.
- First time users were little bit hesitant because they are not familiar with this kind of systems. After repeated trials and getting help from field volunteers they got easy and co-operated with the system.
- Within the specified recording time, users often remained silent, spoke too fast, before the beep or spoke even after the time has passed. These often lead to truncation; recognition errors or too many insertions while silence.
- Co-operative speakers often get dissatisfied with confirmation at each recognition stage in the call flow as it needs extra effort from user and delays the process.
- While free speaking, sometimes farmers spoke the local names of commodities or queried for a different variety or grade, which did not match with any of the listed commodity in Agmarknet website and hence price can't be made available.
- In contrary, too much information often confused them like for the query commodity, when price information is provided for all the 4-5 listed (in Agmarknet) markets within a district reporting on latest date.
- In actual field trials, recorded speech data contain loud speaking or continuous background speech, heavy noises of vehicles' horn, air flow etc. Often speech information is totally lost due to clipping, truncation or channel drop.

5 Error Correction Methods

From the field trial analysis in first two phases, following error correction methods are designed and incorporated into the system and then presented for evaluation in phase 3:

5.1 Signal Analysis and Decision (SAD):

This module will act like a filter to sort out the undesirable speech input before sending it to recognition module so that mis-recognitions due to silence, truncation, clipping, channel drop, extreme noise can be avoided to a certain extent. Here, the incoming speech signal is first analysed by extracting some temporal (zero crossing rate, short time energy etc.) and spectral (formants) features and then the accept-reject decision is taken by comparing those features with a previously stored knowledge base created from transcription of the training data. This knowledge base actually stores the information extracted (in terms of the above mentioned features) from the train utterances with rejection remarks at the time of transcription and set the allowable range accordingly. For rejection decision, user is prompted to speak again otherwise the recoded signal is directly send for recognition. Details of the process can be found in (Basu et al., 2012).

5.2 Confidence Measure (CM) and Polling:

To avoid confirmation in each recognition stage, system must be confident enough so that the ulti-

[3] Agricultural Marketing Information Network - AGMARKNET, http://agmarknet.nic.in.

mate ASR output is reliable to proceed further. To do this, multiple ASR decoders can be built using different training dataset, features or models. At runtime, these decoders will generate parallel hypothesis, among which a single hypothesis having high confidence will be polled out, based on a predefined logic or majority voting (Gada et al., 2013; Mantena et al., 2011; Jiang, 2003).

Here in our approach, the relative occurrence frequencies of hypothesis words are calculated from the decoded output parallely generated by three acoustic models trained from orthogonal partitions of the training dataset. Comparing the highest relative frequency with an empirically set threshold of confidence level, a decision of confidence and next plan of action is decided whether to skip explicit confirmation, ask for confirmation or to re-prompt the user.

5.3 Complementary Information:

The key idea here is to provide the related but necessary information to complement the required exact information which is then not available in the database. Our present system regularly crawls the market price from the website http://www.agmarknet.nic.in/ (managed by Ministry of Agriculture, Government of India) [4] and stores into local database. Sometimes this list may not include the latest price information for an exact district-commodity combination queried by user. Whereas the latest price is then available say for a different variety or grade of the same commodity or for the query commodity in a neighboring district. In this situation, a nearby district or nearby commodity searching is very much useful so that from this information, user can at least make a rough estimate of the required price rather than having a negative response like “price not available”.

5.4 Runtime model generation:

This is being thought of for organizing the searched information in an intelligent way so that, user will never be presented too much information to get confused. To do this, price information disbursement for at the most two markets is allowed in our final system. When price of the desired commodity at more than two markets are known, users are prompted to select and say any one from the available mandi names that the system will recognize and provide price information. Instead of having all market names in a static language model, runtime generation of LM with search vocabulary having only the names of the available markets at latest date will make this process faster, easier and more accurate.

Beside these, some minor changes in prompt level were also included like a “how to use” prompt is added; “after the beep sound…” phrase is emphasized by putting it right at the beginning of the query request prompts etc.

6 Final Results and Discussions

Before introducing the confidence measure & polling logic into the final system to be evaluated in phase-3 field trial, an initial offline testing is conducted with the entire training dataset. Three decoders were built around the partitions used in three fold cross validation testing with trigram LM in each. Individual decoder outputs are programmatically merged and provided for polling to generate the final output and finally aligned with the reference. As per table 4, applying the confidence measure and polling logic a slight increase in percentage word accuracy (85.8) is noticed which is quite encouraging and hence acceptable to the final system in phase 3 field trial. From the % confidence of ASR output obtained in this offline evaluation, the threshold for deciding the confidence level is empirically set to 50% and above (high confidence, no confirmation), 20% to below 50% (medium confidence, ask for confirmation) and below 20% (low confidence, prompt user to repeat the query).

Performance	Decoder 1	Decoder 2	Decoder 3	Polling
% Accuracy	81.3	74.8	66.5	85.8
% confidence of ASR output	All matched		Any two matched	No match
	62.0		27.0	11

Table 4. Results of offline polling experiment

6.1 Qualitative Evaluation

In qualitative evaluation, user’s judgment based on the predefined metrics is taken on a 5 point grading scale of remarks described as 1: very bad, 2:

[4] Agricultural Marketing Information Network - AGMARKNET, http://agmarknet.nic.in.

bad, 3: satisfactory, 4: good, 5: Very good. This evaluation is first introduced in Phase 1 of field trial on 238 subjects which was again conducted in Phase 3 field trial on new 106 subjects after incorporating the error correction methods. For both the trials, percentage of occurrences of all the remarks against each metric are presented in figure 3 and figure 4 respectively.

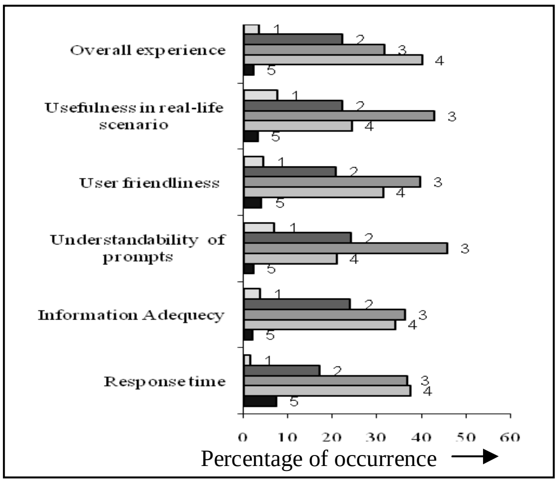

Figure 3. Results of Qualitative evaluation in phase 1.

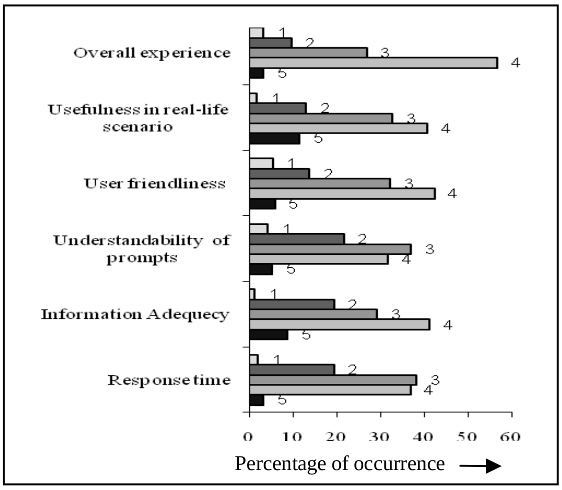

Figure 4. Results of Qualitative evaluation in phase 3

Effects of incorporating error correction methods can easily be understood by comparing figure 3 and 4. In initial Phase 1 trial (figure 3), most of the subjects have opted for 'satisfactory' (grade 3) remarks on four out of total six metrics. But, the number of 'good' grade increased from 2 (out of 6 metrics) in the first phase to 4 in the phase 3 (figure 4). In the final field trial, significant performance enhancement has been noticed regarding information adequacy, usefulness in real-life scenario, user friendliness and overall experience. Whereas increased latency due to integration of recovery modules in final system resulted into increased response time. This might be the reason for decreased 'very good' remarks for response time. But majority still have voted for 'satisfactory' remark on this metric and hence the delay can be negotiable.

6.2 Online Quantitative Evaluation

These evaluations were carried out in phase 2 and phase 3 field trials. Table 5 and 6 shows the results of both the trials in terms of percentage word accuracy and percentage success in overall task completion respectively. Figures in Table 5 consider only the evaluations done in co-operative speaking scenario while Table 6 includes overall statistics of both co-operative and system break testing (in Phase 3).

Comparing the evaluation results of phase 2 and phase 3 field trials in table 5 and table 6 it is noticeable that inclusion of error correction methods had left positive effects on system's performance enhancement. Percentage word accuracy (89.60 and 75.50) and percentage success (86.6 and 67.04), both have increased in phase 3 field trial. With proper guidance and co-operative speaking, system can perform quite well with around 90% accuracy (table 5, phase 3 guided results).

Specifications	**Phase 2**		**Phase 3**	
Number of subjects	35 speakers, 50 words/speaker		50 speakers, 20 words/speaker	
Type of speaking	Guided	Free	Guided	Free
% Accuracy	78.95	68.08	89.60	75.50

Table 5. System's performance in field trials

Head	**Phase 2**	**Phase 3**
Total calls	302	270
Valid calls	251	209
Invalid calls (nothing spoken in two turns, only background speech or call cut)	51	61
Successful calls (correct information)	167	181
Failure (recognition error, incorrect information)	84	28
Success % of valid calls	66.5	86.6
Success % of total calls	55.29	67.04

Table 6. Task completion statistics in system field trials

Task completion statistics in table 6 are also important to understand the impact of the error correction methods. Reduced number of speech recognition errors due to implementation of CM with polling and nearby district-commodity searching for complementary information concept has resulted into increased number of successful calls (from 167 to 181) in phase 3. Whereas, incorporation of SAD module has restricted most of the invalid calls to get into the system which in turn has decreased the number of failures (from 84 to 28). Hence, percentage success out of the valid calls has enhanced from 66.5% in phase 2 to 86.6% in phase 3 which is quite acceptable for live system.

Finally, to explicitly understand the performance of the live system in different adverse situations, system break testing is conducted in phase 3 only on 10 subjects. As these exceptional cases occurred rarely or by chance in field trials, it was separately done on 140 live calls (double trial per metric, so 7*2*10 = 140) with respect to the metrics defined in break testing strategy (in section 3). In fact in some cases, situations like distant speech and extreme noise are intentionally created if not available naturally. Recognition errors due to undesirable speaking like disfluencies, stammering, repetition, extra word were tackled by parallel decoder and polling while pure silence, loud speaking, extreme noises were mostly rejected by SAD module. Overall 63.2% word accuracy is achieved with these calls.

7 Conclusion

Evaluation of an IVR based ASR system for commodity price retrieval is elaborately discussed in this paper. Starting from the need analysis, scope of the present work includes defining evaluation strategies, conducting sequential field trials, feedback analysis, designing and implementing novel error correction methods and finally system's performance analysis as per the evaluation results. Most of the evaluation methods and metrics are domain independent and applicable to similar systems. Encouraging results in final trial indicate deep introspection into the recovery methods which need further explorations to enhance future developments.

Acknowledgement

This work is a part of ASR Consortium (Phase 1) project funded by DeitY, Ministry of Communication and Information Technology, Government of India. The authors would like to thank Prof. S. Umesh (Consortium Leader, IIT Madras) and all other consortium members for their constant support and cooperation during the project work.

References

D. Gomillion and B. Dempster. 2005. Building Telephony System with Asterisk, ISBN: 1-904811-15-9, Packet Publishing Ltd.

Gautam V. Mantena, S. Rajendran, Suryakanth V. Gangashetty, B. Yegnanarayana and Kishore Pra-hallad. 2011. Development of a Spoken Dialogue System for accessing Agricultural Information in Telugu. In Proceedings of 9th International Confe-rence on Natural Language Processing (ICON), Anna University, Chennai, India.

H. Jiang. 2003. Confidence measures for speech recognition: A survey, Speech communication, Volume 45, pages 455-470.

J. Gada, P. Rao and K. Samudravijaya. 2013. Confidence Measure for Detecting Speech Recognition Errors. In Proc. of NCC, pages 1-5.

Jahanzeb Sherwani, S. Palijo, S. Mirza, T. Ahmed, N. Ali and R. Rosenfeld. 2009. Speech vs. touch-tone: Telephony interfaces for information access by low literate users, In Proc. Int. Conference on Information and Communications Technologies and Development, pages 447-457.

Joyanta Basu, Milton S. Bepari, Rajib Roy and Soma Khan. 2012. Real Time Challenges to Handle the Telephonic Speech Recognition System. In Proc. of forth ICSIP, Coimbatore, India, pages 395-408.

L. Adam, M. Ward, and S. Amarasinghe. 2010. Evaluation of IVR Data Collection UIs for Untrained Rural Users. Pro.of ACM DEV'10, London, UK.

M. Plauche and U. Nallasamy. 2007. Speech interfac-es for equitable access to information technology, Volume 4, Number 1. MIT Press Cambridge, MA, USA, pages 69–86.

N. Patel, S. Agarwal, N. Rajput, A. Nanavati, P. Dave and T. S. Parikh. 2009. A comparative study of speech and dialed input voice interfaces in rural india, In Proc. of ACM CHI, pages 51-54.