Telephony Speech Recognition System: Challenges

Joyanta Basu
CDAC
Salt Lake, Sector–V
Kolkata- 700091
joyanta.basu@cdac.in

Rajib Roy
CDAC
Salt Lake, Sector–V
Kolkata- 700091
rajib.roy@cdac.in

Milton S. Bepari
CDAC
Salt Lake, Sector–V
Kolkata- 700091

Soma Khan
CDAC
Salt Lake, Sector–V
Kolkata- 700091

ABSTRACT

Present paper describes the challenges to design the telephony Automatic Speech Recognition (ASR) System. Telephonic speech data are collected automatically from all geographical regions of West Bengal to cover major dialectal variations of Bangla spoken language. All incoming calls are handled by Asterisk Server i.e. Computer telephony interface (CTI). The system asks some queries and users' spoken responses are stored and transcribed manually for ASR system training. In real time scenario, the telephonic speech contains channel drop, silence or no speech event, truncated speech signal, noisy signal etc along with the desired speech event. This paper describes these kinds of challenges of telephony ASR system. And also describes some brief techniques which will handle such unwanted signals in case of telephonic speech to certain extent and able to provide almost desired speech signal for the ASR system.

General Terms

Automatic Speech Recognition, Signal Processing, Telephony Application.

Keywords

Asterisk server, Interactive Voice Response, Transcription Tool, Temporal and Spectral features, Knowledge Base

1. INTRODUCTION

Modern human life is totally dependent on technology and along with these devices become more and more portable like mobile, PDAs, GPRS etc. Beside this, there is also a growing demand for some hands free Voice controlled public purpose emergency information retrieval services like Weather forecasting, Road-Traffic reporting, Travel enquiry, Health informatics etc. accessible via hand-held devices (mobiles or telephones) to fulfill urgent and on the spot requirements. But real life deployment of all these applications involves development of required modules for voice-query based easy user interface and quick information retrieval using mobiles. In fact, throughout the world the number of telephone users is much higher than that of the PCs. Again human voice or speech is the fastest communication form in our daily busy schedule that further extends the usability of such voice enabled mobile applications in emergency situations. In such a scenario, speech-centric user interface on smart hand-held devices is currently foreseen to be a desirable interaction paradigm where Automatic Speech Recognition (ASR) is the only available enabling technology.

Interactive Voice Response (IVR) systems provide a simple yet efficient way for retrieving information from computers in speech form through telephones but in most of the cases users still have to navigate into the system via Dual Tone Multiple Frequency (DTMF) input and type their query by telephone keypad. A comparative study by K.M Lee & J.Lai, 2005 [1] revealed that in spite of occasionally low accuracy rates, a majority of users preferred interacting with the system by speech modality as it is more satisfying, more entertaining, and more natural than the touch-tone modality which involves the use of hands, quite time consuming and require at least the knowledge of English alphabets.

Especially for a country like, India, with its multi-lingual requirements and not so fortunate achievements in terms of overall literacy, development of IVR application with backend ASR support for recognizing voice query and response in native languages is of major importance because these systems facilitates the common mass of the country to access the huge information available in Internet using telephones. But at the time of deployment, such an IVR based access system will definitely have to cope up with real world speech and related challenges as well. So, it is very important to detect such challenging situations and find out ways to overcome them gracefully without annoying the user.

Present paper addresses some real time challenges of telephony ASR applications. And also provides a clear picture of the above tasks in a well planned and sequential manner aiming towards the development of an IVR application in spoken Bangla language.

2. MOTIVATION OF THE WORK

A practical IVR system should be designed in such a way that it should be capable of handling real time telephony hazards like channel drop, clipping, speech truncation etc. It should also provide robust performance considering following issues:

1. **Speaker-Variability:** Handle speech from any arbitrary speaker of any age i.e., it would be a speaker-independent ASR system.
2. **Pronunciation/Accent Variability:** Different pronunciations, dialectical variations and accents within a particular Indian language
3. **Channel Variability:** Different channels such as landline versus cellular and different cellular technologies such as GSM and CDMA.
4. **Handset Variability:** Variability in mobile handsets due to differences in spectral characteristics.

5. **Different Background Noise:** Various kinds of environmental noise, so that it is robust to real-world application.

Considering the above said requirements, telephonic ASR is being designed in such a way that, to-some-extent it can meet the above mentioned capabilities. In the present study, speech data are mainly collected from all the geographical regions where native Bangla language speaking population is considerably high. The collected speech data is then verified and used for ASR training.

The reason behind choosing a large geographical area for data collection is to cope up with the problem of speaker variability, accentual variability. Additionally, various issues regarding the telephonic channel such as channel drop or packet lost during transmission, handset variability, service provider variability, various types of background noise such as cross-talk, vehicle noise etc. have been observed, analyzed and estimated efficiently from the collected speech data and modeling of those can improve ASR performance. These issues will not only help us to improve the system performance effectively, but also provide us very good research motivation on other telephonic applications.

3. BRIEF OVERALL SYSTEM OVERVIEW

Telephonic ASR system is designed such a way, that users get the relevant information in a convenient manner. First the system will give the user a language preference (within Hindi, Bangla and Indian English) and then onwards each time a directed question is asked, and the user would reply it with appropriate response from a small set of words. System is composed of three major parallel components. They are IVR server (hardware and API), Signal Processing Blocks with ASR engine and Information Source. Fig. 1. represents an overall block diagram of the system.

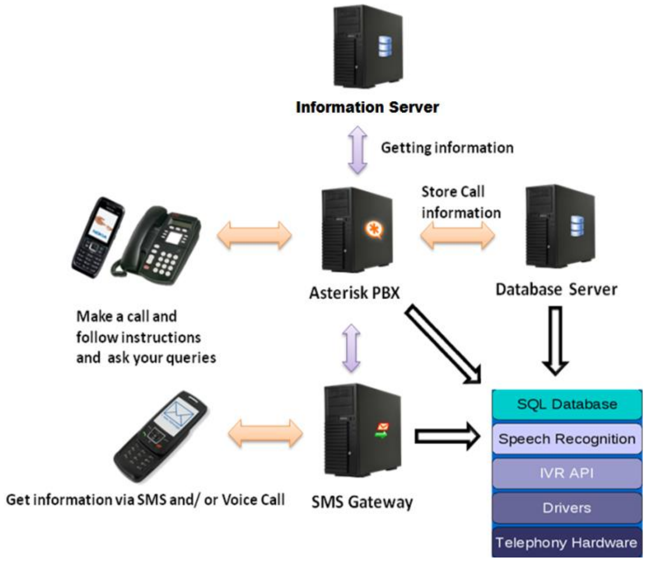


Fig. 1: Block diagram of Telephonic ASR System

3.1 IVR hardware and API

As shown in fig. 2 the Interactive Voice Response (IVR) consists of IVR hardware (generally a Telephony Hardware), a computer and application software running on the computer. The IVR hardware is connected parallel to the telephone line. The functionality of the IVR hardware is to lift the telephone automatically when the user calls, recognize the input information (like dialed digit or speech) by the user, interact with computer to obtain the necessary information, then convert the information into speech form and also convert the incoming speech into digital form and store it in the computer.

In development of telephonic ASR system, Asterisk [2][3] is used here as an open source IVR Server, converged telephony platform, which is designed primarily to run on Linux. It support VoIP protocols like SIP, H.323; interfaces with PSTN Channels, supports various PCI Cards, and also open source Drivers and Libraries are available.

3.2 Signal Processing Block and ASR engine

This block consists of three major blocks namely Speech Acquisition and Enhancement module, Signal Analysis and Decision module and ASR Engine. Block diagram of such Signal Processing Block is shown in Fig. 3.

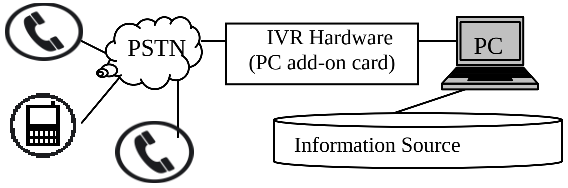


Fig. 2: Interactive voice response system

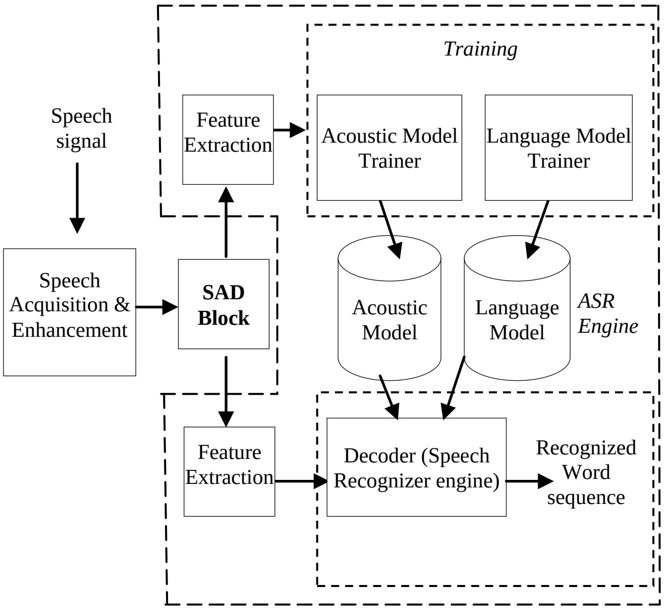


Fig. 3: Basic block diagram of Signal Processing Blocks including Automatic Speech Recognition engine

3.21 Speech Acquisition and Enhancement module

The first block, which consists of the acoustic environment plus the transduction equipment, can have a strong effect on the generated speech representations because additive noise, room reverberation, recording device type etc. are associated with the process. A speech enhancement module suppresses the above effects so that the incoming speech can easily be recognized in heavy perturbed conditions.

3.2.2 Signal Analysis and Decision module (SAD)

In this module all incoming speech waveform analyzed for the valid speech signal or not. This is the very important module of this system. This module extracts different temporal features like Zero Crossing Rate (ZCR) [4][5], Short Time Energy (STE) [6][7] and Spectral features like formant analysis etc. Then using some predefined Knowledge Base (KB) this module gives some decision valid information regarding incoming speech signal. After this module system takes some decision whether users need to re-record speech signal or not. If re-recording is not required then recorded speech signal go through ASR engine for decoding the signal.

3.2.3 ASR engine

The basic task of Automatic Speech Recognition (ASR) is to derive a sequence of words from a stream of acoustic information. Automatic recognition of telephonic voice queries requires a robust back-end ASR system. CMU SPHINX [8], an open Source Speech Recognition Engine is used here which typically consists of Speech Feature Extraction module, Acoustic Model, Language Model, Decoder [9]

3.3 Information Source

Repository of all relevant information is known as a trusted Information Source, and design architecture of the same typically depends on type of information. In current work dynamic information is mainly refereed from trusted information source or online server and all other information which does not change very much during a considerable period of time are kept in local database using web crawler. System response of any query on dynamic information must ensure delivery of latest information. To accomplish this objective a specific reference made to trusted information source. This approach ensures quick delivery of information.

4. CHALLENGES OF TELEPHONY ASR

Speech quality is affected by the transmission impairments found on telephone connections. Sometimes the intelligibility and naturalness of speech degrade to an intolerable extent [10]. These challenges include loudness loss, circuit noise, side tone loudness loss, room noise, attenuation distortion, taker echo, listener echo, quantizing distortion, phase jitter etc. In addition, such user interfaces terminating the transmission channel as mobile handsets and hands-free terminals are likely to pick up background noise, mainly including circuit noise, noise floor, impulse noise, ambient room noise and crosstalk noise.

To find out the problems we have designed one Semi Automatic Transcription Tool [11]. This tool has been designed for offline transcription of recorded speech data, such that all transcriptions during data collection can be checked, corrected and verified manually by human experts. Automatic conversion of text to phoneme (phonetic transcription) is necessary to create pronunciation lexicon which will help the ASR System training.

The methodology for Grapheme to Phoneme (G2P) conversion in Bangla is based on orthographic rules. In Bangla G2P conversion sometimes depends not only on orthographic information but also on Parts of Speech (POS) information and semantics [12]. G2P conversion is an important task for data transcription.

From where, many information were gathered regarding telephonic speech data. At the time of transcription we have to give some transcription remark tags and also noise tags. It's totally human driven task. Descriptions and measurement of the remarks are given in Table 1 and Table 2 shows the different types of noise tags.

From Table 1 it has been seen that S_UTTR, CPR_UTTR, TR_UTTR and R_UTTR are the rejection tag sets. Speech files marker by other tag set may be accepted and considered at the time of ASR training.

Table 1: Description and measurement of Remarks

Wave Remarks (WR)	Description
A_UTTR (Amplitude)	Amplitude (Partly or fully) will be modified
C_UTTR (Clean)	Speech Utterance may contain some non overlapping non-speech event
Transcription Remarks (TR)	**Description**
CLPD_UTTR (Clipped)	Clipping of speech Utterance
CPC_UTTR (Channel Problem Consider)	Channel drop occurs randomly in silence region which have not affect speech region
CPR_UTTR (Channel Problem Reject)	Some word or phonemes dropped randomly
I_UTTR (Improper)	In this case the utterance is slightly different than prompt in phoneme level
MN_UTTR	Noise within speech
MP_UTTR	Reasonable silence (pause) within speech
R_UTTR (Reject)	In this case speech signal is wrongly spelt or may be too many noise or may be non-sense words, can't be able to understand
S_UTTR (Silence)	No speech Utterance present
TA_UTTR (Truncate Accept)	Truncation of not so significant amount (may be one or two phoneme) speech Utterance
TR_UTTR (Truncate Reject)	Truncation of significant amount speech Utterance
W_UTTR (Wrong)	In this case the utterance is totally different from the corresponding prompt

Table 2: Types of noise tags

Tag	Explanation / examples
<air>	Air flow
<animal>	Animal Sound
<bang>	Sudden (impulsive) noise due to banging of door
<beep>	Telephonic Beep sound
<bird>	Sound of Bird
<bn>	General Background Noise
<br>	breath noise
<bs>	Background speech (babble)
<bsong>	Background Song
<bins>	Background Instrument

<burp>	burp
<cough>	cough
<cry>	Children Cry
<ct>	Clearing of throat
<horn>	Horn noise of vehicles
<ht>	Hesitation
<laugh>	laugher
<ln>	Line noise
<ls>	Lip smack
<ns>	hiccups, yawns, grunts
<pau>	Pause or silence
<ring>	Phone ringing
<sneeze>	sneeze
<sniff>	sniff
<tc>	tongue click
<vn>	Vehicle Noise

5. DESCRIPTION OF REJECTION TAG SETS

In this study S_UTTR, CPR_UTTR, TR_UTTR and R_UTTR are the main rejection tag sets. Below some brief descriptions of those tag sets are given.

5.1 TR_UTTR (Truncate Reject)

In this kind of rejection, actual speech segment is truncated from the beginning of the signal or may be sometimes end of the speech signal. The speech signal is truncated in such a way that it loses its information and it is not suitable for the ASR system testing. This situation may arise when the user does not co-operate well with the system and often speaks either before or after the specified time for recording. Fig. 4 shows time domain and also spectra domain view of TR_UTTR problem.

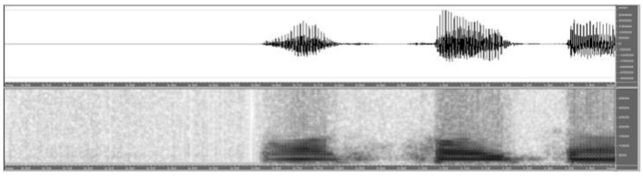


Fig. 4: Sample speech file of TR_UTTR

5.2 S_UTTR (Silence)

Sometimes due to network connection problem or may be users' input related problem this type of situations may occur. It's not ideal SILENCE, but no active speech zone is found in the entire wave file. Fig. 5 shows time domain and also spectra domain view of S_UTTR problem.

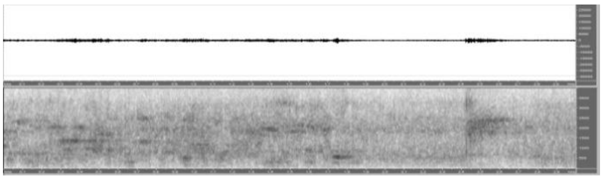


Fig. 5: Sample speech file of S_UTTR

5.3 CPR_UTTR (Channel Problem Reject)

At the time of transmission through telephone channel, due to channel problem we have faced this kind of problem. Due to channel problem sometimes we got pure silence. But often it occurs within the active speech zone and drop out the actual speech information partially or totally. Fig. 6 shows time domain and also spectra domain view of CPR_UTTR problem.

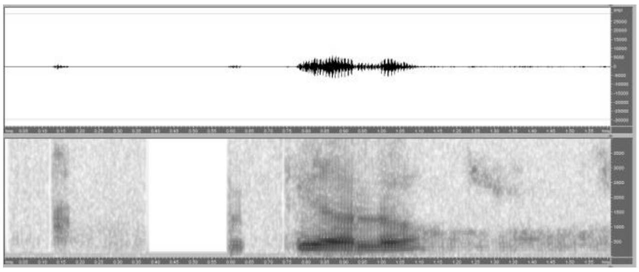


Fig. 6: Sample speech file of CPR_UTTR

5.4 R_UTTR (Reject)

Among the all rejection tag set this kind of tag set is very difficult for marking and automatically extraction. It happens when users' spoke speech segment but those are nonsense words or may be too many crosstalk etc. Fig. 7 shows time domain and also spectra domain view of R_UTTR problem. Where we have seen that few portions are speech segment and some portion are noisy. To find out R_UTTR automatically is really big problem.

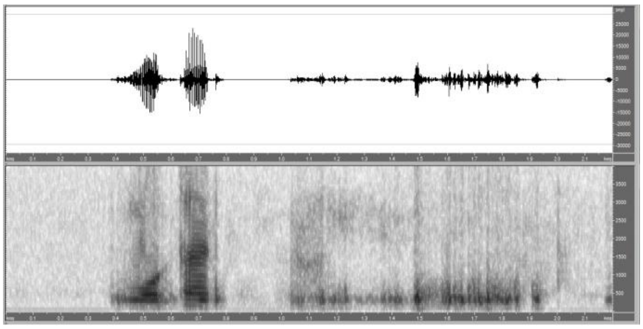


Fig. 7: Sample speech file of R_UTTR

6. OBSERVATIONS AND RESULTS

For this work we have collected almost 60 hours telephonic speech data from nineteen districts of West Bengal and transcribed manually. This transcribed data helps to build up the KB. Table 3 represents the distribution of totally collected speech data according to the variations of Speakers' Gender, Age, Education Qualification, Recording Handset model, Service provider and Environment in terms of percentage of occurrences in each criteria. Table 4 described the percentage of occurrence of major noise tags.

Table 3: Variations in collected Speech data

Gender (in %)	Male	66
	Female	34
Age (in %)	Child : 0-15	11
	Adult : 15-30	56
	Medium : 30-50	28
	Senior : 50-99	5
Qualification (in %)	Primary	12
	Secondary	46
	Post-Secondary	38
	Others	4
Handset Model (in %)	Nokia	46
	Samsung	20
	LG	5
	Reliance	2
	Sony	2
	Others	25
Service provider (in %)	BSNL	11
	Airtel	10
	Vodafone	29
	Reliance	3
	Aircel	11
	MTS	15
	IDEA	17
	Others	4
Environment (in %)	Noise	2
	Clean	85
	Babble	9
	Music	4

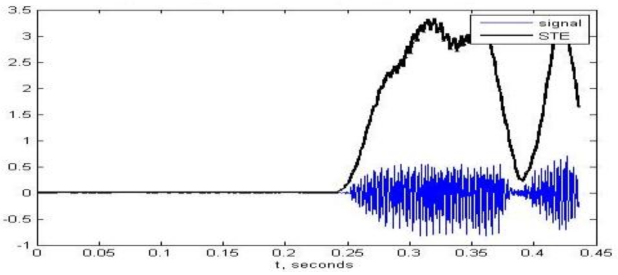
Table 4: % of occurrence of the major noise tag set

Tag	% of Occurrence
<air>	7.92
<bang>	1.09
<beep>	5.59
<bird>	5.22
<bn>	17.91
<br>	0.4
<bs>	15.9
<cough>	6.1
<ct>	0.03
<horn>	1.02
<laugh>	8.08
<ln>	6.06
<ring>	6.08

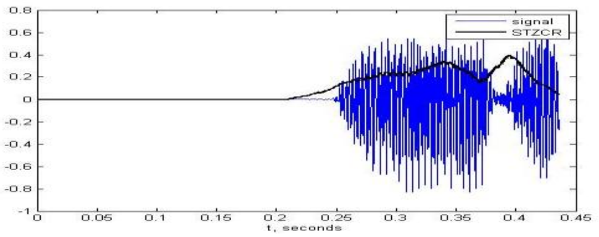
Fig. 8 shows the observation result of TR_UTTR, S_UTTR and CPR_UTTR speech utterances, with related STE and ZCR plots respectively. It has been observed from the figures that (i) In case of Truncate Problem (i.e. TR): signals ends with high STE and high SCR value may be at the end of the signal or sometimes occurs at the beginning of the signal. It happens because users some time delayed speaking within specific time span or sometimes starts so early. (ii) In case of S_UTTR: signals are generally channel noise and mostly unvoiced sounds, that's why high ZCR and low STE observed almost entire time span of the recording. (iii) In case of CPR_UTTR: sometimes channel packets have been lost due to may be bad network strength of the users or may be type of handset. It's very common problem for the telephonic applications. It has been observed that ZCR value change suddenly from low to high or high to low and same thing for STE also. Formant analysis also has been done for all these rejection cases. To find out S_UTTR automatically from the signal formant may be the one good spectral feature. But for other cases of rejection, formant analysis is not so well.

Fig. 9 shows the one of the observation of R_UTTR, where it has been seen that many voiced and unvoiced regions are there, but those are basically background speech, not the expected reply from the uses. And also we have seen that natural STE and ZCR plots are there. So, it's really difficult to automatically extract R_UTTR always. But sometimes, when overall STE is below the expectation level of the incoming speech signal then Speech Analysis module can mark it as a R_UTTR. Basically a performance of automatic extraction of R_UTTR depends on the background noise or unwanted speech. More work is going on this rejection type.

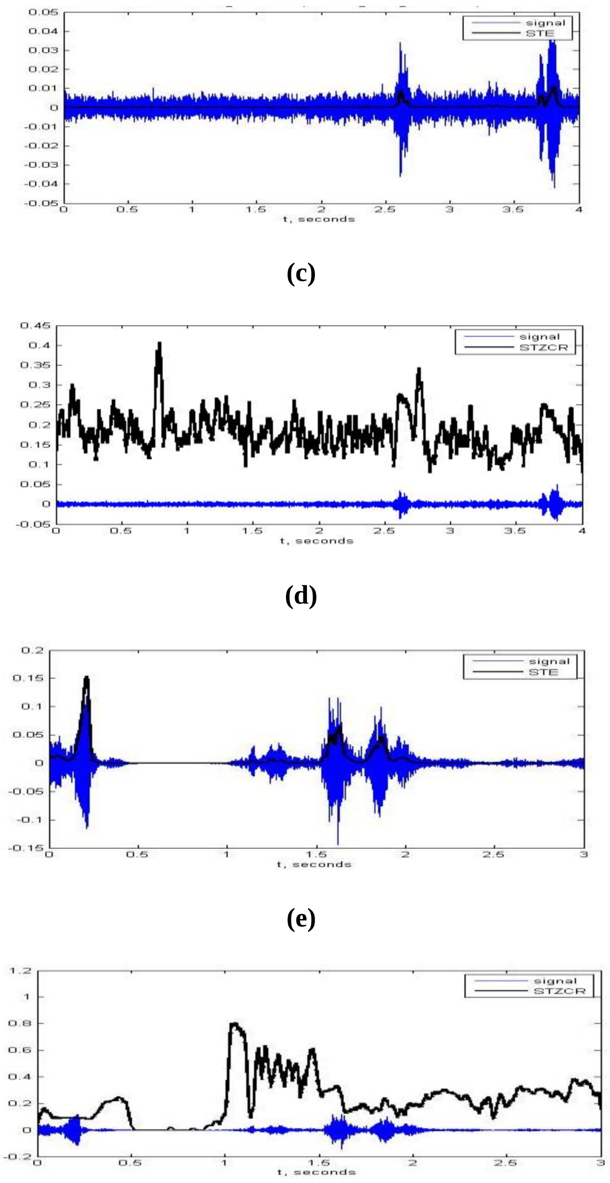
From the above observations, signal analysis module are designed for telephonic ASR system and tested in real life scenarios. On-the-fly pattern of rejection remarks extraction is the main objective of signal analysis module. We have analyzed 887 numbers of incoming calls from users and total utterances are 10327 numbers. Table 5 shows result of automatic extraction of the rejection remarks except R_UTTR. To find out R_UTTR remarks system requires some manual intervention.



(a)



(b)



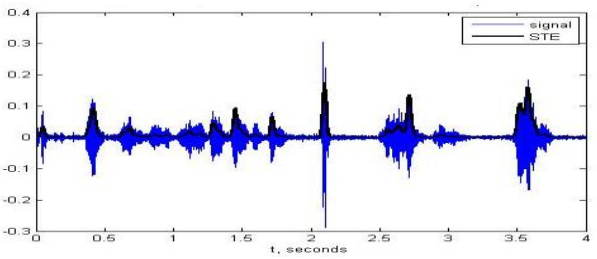
(c)

(d)

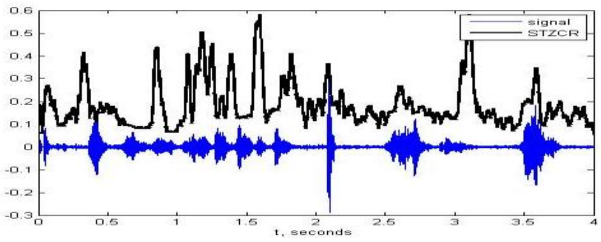
(e)

(f)

Fig. 8: TR_UTTR: (a) Signal vs. STE (b) Signal vs. ZCR
S_UTTR: (c) Signal vs. STE(d) Signal vs. ZCR
CPR_UTTR: (e) Signal vs. STE (f) Signal vs. ZCR



(a)



(b)

Fig 9: R_UTTR: (a) Signal vs. STE (b) Signal vs. ZCR

Table 5: Output of Signal Analysis Module and its decision

No. of Incoming Calls	No. of Utterances	CPR_UTTR	S_UTTR	TR_UTTR	R_UTTR
887	10327	659	795	248	458

7. CONCLUSION

In the proposed work, challenges of telephony ASR have been described. This work try to address various issues regarding the telephonic channel such as channel drop or packets lost during transmission and try to handle real world speech related problematic issues. We have observed mainly four types of rejection tag sets namely CPR_UTTR, S_UTTR, TR_UTTR and R_UTTR. We have seen that automatic extraction of CPR_UTTR, S_UTTR and TR_UTTR by SAD module is encouraging. But automatic extraction of R_UTTR is not so easy. Currently more research is going on for this particular utterance type. More importantly, this kind of real voice based information retrieval application useful especially to the people having no access to computers and Internet, people who may not have the required computer skills or even reading/writing abilities and also the visually challenged personal. After successful completion of the present work, it will enable development of similar speech-based access systems for other (like Medical, tourism, transport, Emergency services) public domain applications.

8. REFERENCES

[1] Kwan Min Lee, Jennifer Lai, “Speech vs. Touch: A Comparative Study of the Use of Speech and DTMF Keypad for Navigation”, International Journal of Human Computer Interaction IJHCI, Vol. 19, No. 3, 2005.

[2] Gomillion D, Dempster B, “Building Telephony System with Asterisk”, ISBN: 1-904811-15-9, Packet Publishing Ltd.

[3] Meggelen J V, Madsen L, Smith J, “Asterisk: The Future of Telephony”, ISBN-10: 0-596-51048-9, ISBN-13: 987-0-596-51048-0, O’REILL

[4] Yiu-Kei Lau; Chok-Ki Chan; , "Speech recognition based on zero crossing rate and energy," Acoustics, Speech and Signal Processing, IEEE Transactions on, vol.33, no.1, pp. 320- 323, Feb 1985.

[5] Aye, Y.Y.; , "Speech Recognition Using Zero-Crossing Features," Electronic Computer Technology, 2009

International Conference on , vol., no., pp.689-692, 20-22 Feb. 2009.

[6] Swee, T.T.; Salleh, S.H.S.; Jamaludin, M.R.;, "Speech pitch detection using short-time energy," Computer and Communication Engineering (ICCCE), 2010 International Conference on, vol., no., pp.1-6, 11-12 May 2010.

[7] Erdol, N.; Castelluccia, C.; Zilouchian, A.; "Recovery of missing speech packets using the short-time energy and zero-crossing measurements," Speech and Audio Processing, IEEE Transactions on, vol.1, no.3, pp.295-303, Jul 1993. http://www.speech.cs.cmu.edu/.

[8] Joyanta Basu, Soma Khan, Rajib Roy and Milton Samirakshma Bepari, "Designing Voice Enabled Railway Travel Enquiry System: An IVR Based Approach on Bangla ASR", ICON 2011, Anna University, Chennai, India, pp – 138-145, December, 2011.

[9] Guoyu Zuo; Wenju Liu; Xiaogang Ruan; "Telephone speech recognition using simulated data from clean database," Robotics, Intelligent Systems and Signal Processing, 2003. Proceedings. 2003 IEEE International Conference on, vol.1, no., pp. 49- 53 vol.1, 8-13 Oct. 2003.

[10] Joyanta Basu, Milton Samirakshma Bepari, Rajib Roy and Soma Khan, "Design of Telephonic Speech Data Collection and Transcription Methodology for Speech Recognition Systems", FRSM 2012, pp- 147-153, KIIT, Gurgaon.

[11] Basu, J, Basu T, Mitra M, Das Mandal S, "Grapheme to Phoneme (G2P) conversion for Bangla," Oriental COCOSDA International Conference, pp.66-71, 10-12 Aug. 2009.