The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages
29-31 August 2018, Gurugram, India

Designing an IVR based Framework for Telephony Speech Data Collection and Transcription in Under-Resourced Languages

Joyanta Basu[1], *Soma Khan*[1], *Milton S. Bepari*[1], *Rajib Roy*[1], *Madhab Pal*[1], *Sushmita Nandi*[1], *Karunesh Kumar Arora*[2], *Sunita Arora*[2], *Shweta Bansal*[3], *Shyam Sunder Agrawal*[3]

CDAC Kolkata[1], CDAC Noida[2], KIIT College of Engineering Gurgaon[3]

{joyanta.basu, soma.khan, milton.bepari, rajib.roy, madhab.pal}@cdac.in,
nandi.sushmita@gmail.com, {karunesharora, sunitaarora}@cdac.in,
bansalshwe@gmail.com, ss_agrawal@hotmail.com

Abstract

Scarcity of digitally available language resources restricts development of large scale speech applications in Indian scenario. This paper describes a unique design framework for telephony speech data collection in under-resourced languages using interactive voice response (IVR) technology. IVR systems provide a fast, reliable, automated and relatively low cost medium for simultaneous multilingual audio resource collection from remote users and help in structured storage of resources for further usage. The framework needs IVR hardware & API, related software tools and text resources as its necessary components. Detailed functional design and development process of such a running IVR system are stepwise elaborated. Sample IVR call-flow design templates and offline audio transcription procedure is also presented for ease of understanding. Entire methodology is language independent and is adaptable to similar tasks in other languages and specially beneficial to accelerate resource creation process in under-resourced languages, minimizing manual efforts of data collection and transcription.

Index Terms: Under-resourced languages, IVR based resource creation, Audio transcription.

1. Introduction

Based on the availability of resources, languages can be categorized in well-resourced and under-resourced languages. While more than 6,900 languages exist all over the world [1], the number of well-resourced languages is quite limited. In fact, a large amount of world languages is under-resourced [2]. The term under-resource refers to languages with one or more of the following aspects: lack of a unique writing system or stable orthography, lack of linguistic expertise, lack of electronic resources for speech and language data [3].

Modern India, as per the 1991 census, has more than 1576 mother tongues, genetically belonging to five different language families. They are further rationalized into 216 mother tongues, and grouped under 114 major languages [4]. The year 2001 census identified 122 major languages in India, out of which 29 languages have more than a million native speakers. While, other 1,599 languages are spoken by smaller societies, local groups and tribes. As per Eighth Schedule, there are 22 official languages in India [5]. Though having such a rich language heritage and enormous diversity, all the Indian languages are under-resourced in terms of digitally available language resources. Designing an effective framework to create and preserve digitally available resources for under-resourced languages is very important in Indian scenario. Interactive Voice Response (IVR) technology [6] can make a difference here. Properly designed IVR systems provides a fast, reliable, automated and relatively low cost medium for real world telephony speech data collection from target users, residing anywhere in the country in their native language at convenient time [7]. Moreover, if previously programmed, IVR systems can also store the collected speech resources in a predefined format (at runtime) maintaining a structured storage hierarchy for future usage. Data collection, preservation as well as data retrieval, in both the ways, this technology is beneficial to a multilingual country like India without spending any extra man-hours for the same task. Properly maintained language resources then only can be standardized and efficiently used for development of large scale advanced speech applications in Indian languages, much alike the well-resourced languages.

2. Purpose of the Work

Speech data collection seems to be quite easy yet time taking and tedious work involving too much manual interventions in our general inception. A smartly designed IVR system can change the entire picture and the very purpose of our work in this paper is to describe how to make this happen. Uniqueness of our work lies in the following:

- We describe the inevitable components to set up an IVR based framework for real world speech data collection and transcription in under-resourced languages
- We include a detail on functional design and development process with all intermediate processes to explain related sequential and parallel tasks while working within such a running framework
- We present an offline web based audio data transcription process along with sample IVR call flow design templates to ease development processes in similar lines.

A number of related studies include prior efforts, on audio resource collection and corpus creation in different world languages. [8, 9, 10, 11, 12, 13]. Similar efforts on Indian languages can be found in [14, 15, 16]. But, those mostly focus either on studio collected data or telephony data used in task specific limited vocabulary applications which are not easily adaptable to large scale application development purposes. Similar works on real world audio resource creation (speech data collection and transcription) using IVR technology is relatively scarce in Indian language context. Present paper provides a clear picture of IVR based telephony speech data collection and transcription in a well-organized

10.21437/SLTU.2018-10

way aiming towards its applicability in large scale advanced speech application development. Entire methodology is language independent and easily scalable to large scale speech resource creation task in any language with similar purpose.

3. System Overview and Architecture

Successful development of IVR systems for speech resource creation depends upon availability and functionality of three major components or resources. These are IVR server (hardware and API), related software resources and tools and text resources for software design and speech data recording.

3.1. IVR hardware & API

A typical IVR system for data collection consists of IVR hardware (generally a telephony hardware), a computer (to be used as IVR server) and application software running on that computer. IVR hardware is connected parallel to the telephone line. Functionality of IVR hardware is to lift the telephone automatically when the user calls, activate specific call contexts as written in configuration files and fire certain scripts or Call flow programs prepared using IVR specific application program Interface (API) installed in the server. The scripts then identify the input information type (either Dual Tone Multiple Frequency i.e. DTMF of touch and key pressed type input or human uttered speech), invoke specific commands to enable interaction between user and system, process DTMF and speech (after converting to digital form) inputs separately and store it in the server. Finally, it does follow up tasks or asks user for next input in speech prompts and the process continues until the call flow scripts end.

In our work, Asterisk [17] is used as an open source IVR server, converged telephony platform, designed primarily to run on Linux operating system. This is basically a software implementation of telephone private branch exchange (PBX). It allows connection to telephony services, such as the public switched telephone network (PSTN), supports voice over Internet Protocol (VoIP) services like SIP, H.323; interfaces with PSTN channels, supports linkage with various computer telephony Interface (CTI) cards. We used Sangoma made CTI card that supports parallel call handling upon thirty voice channels. Open source drivers and libraries are also available to configure the same within IVR server.

3.2. Software resources & tools

Separate software resources, like Asterisk Gateway Interface or AGI [18] library packages written in any of the high level languages (PHP, Python, Perl, JAVA etc.) are required to design and implement IVR call flow program for speech data collection. Database software like MySQL, PostgreSQL are generally used for structured storage and easy retrieval of metadata, intermediate and log files information. Other related software tools include, any popular Speech editing software (like Praat, WaveSurfer, CoolEdit etc.), Speech prompts amplitude normalizer, Speech Activity Detection (SAD) tool [19], Grapheme-to-Phoneme converter, Semi-automatic audio transcription tool, pronunciation lexicon (PL) or dictionary creation tools etc. Similarly, automatic web crawlers to fetch online text data from source websites, language specific text data checking, filtration and correction tool, spellchecker software are used to refine and prepare text resources.

3.3. Text Resources

Text resources define what data specifically we are going to collect in data collection. Conversations, free and spontaneous speech data collection do not generally require any pre-designed text data. Though, pictures, some introductory speech, audio visual content, clips of performing arts etc. are often used to trigger thought process if necessary. For read out speech data recording, well prepared text data is a must have. Now-a-days, websites with regional language content are fastly becoming the easy and straight forward source of unlimited raw text. This essentially requires removal of text noises like, typos, nonsense, syntax and semantic errors and normalization of incomplete, bracketed and code switched phrases. A meaningfully extracted language representative subset (phonetically rich or balanced) of the refined text corpus is generally used for ultimate recording purposes. Beside text corpus, prompt list, system design document, metadata sheet, standardized grapheme and phone set, grapheme to phoneme (G2P) conversion rules, lexicons, tool usage manuals are also necessary text resources.

4. Functional Design and Development

Functional design targets to include some salient features in the IVR system; like, less expectancy for manual intervention, ability of smartly handling parallel and sequential tasks in predefined systematic order, duly notifying for specific human intervention, functional flexibility to resume later, allowing module level corrections whenever needed and most importantly, still maintaining a simple user interface for ease of target users who will actually use it.

Having all the inevitable components available to build the basic architecture, functional design and development of the automated data collection system then only can be started. Figure 1 depicts the main functionalities involved in IVR based speech resource creation task in step by step processes.

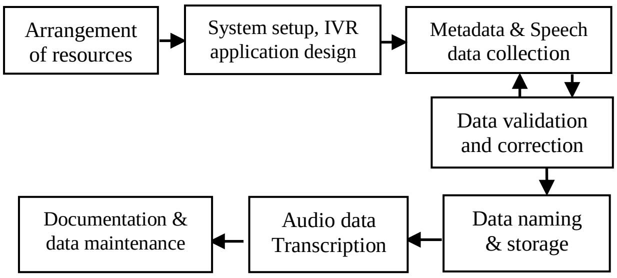

Arrangement of resources
System setup, IVR application design
Metadata & Speech data collection
Data validation and correction
Data naming & storage
Audio data Transcription
Documentation & data maintenance

Figure 1: St*eps for IVR based speech resource development*

4.1. System setup & IVR application design

Once, IVR infrastructure set up is done, application specific IVR system can then only be developed, which again needs to be completed before starting of field data collection. Any IVR application like the same for speech resource creation functionally depends on three components; Call flow scripts, IVR prompts and backend Database system. Call flow specifies step by step actions and associated effects for a target IVR application are is programmed by an experienced programmer using AGI library written in high level languages (like PHP-AGI etc.). Call flow script uses IVR speech prompts that are specially designed audio segments played over phone to enable easy interaction with end users in native language from system side. Finally call flow connects to a predesigned backend database using specific MySQL stored procedures to store, manage and retrieve collected information, metadata, logs etc. for routine maintenance and future use purposes.

4.2. Speech Data Collection

As per speech data collection call flow, following activities are done sequentially:

4.2.1. *Metadata Collection*

This is the only manual task in the entire data collection process as it involves spoken interaction between speakers and field agents or instructors to ensure source authenticity, quality and variety of the collected speech data. Metadata defines the type of collected data based on its recording condition, source type and related specific information considering various factors depending on speaker, recording hardware and background information. Necessary metadata types, recording condition factors, their representative values and possible set of options along with format of blank metadata sheets need to be already thought of and designed at the time of text resource preparation. Table 1 provides representative fields and possible options in such a metadata sheet. Field agents are instructed to carry around such pre-designed blank metadata sheets while field data collection and fill them up duly after gathering all necessary information with help from speaker.

Table 1: *Detail of required fields in metadata sheet*

Field name	Field content, values and options
Speaker ID	0001-9999 (program generated)
Speaker Info	Spk name, Address, Contact
Gender	Male / Female / Other
Age group	15-30 yrs/ 31-55 yrs /Above 55 years
Mother tongue	English / Hindi /Bengali
Education	Primary/ Secondary / Higher Secondary / Graduation / Post Graduation/ Other
Occupation	Student / Service / Business/ Self Employed / Home maker / Other
Language	Indian English / Hindi / Bengali
Dialect Region	Region A / B / C / D / E
Rec. device	Mobile Phone / Landline / VoIP / Other
Device Make	Nokia/ Samsung/ LG/ Sony/Mi etc.
Service Provider	BSNL/MTNL/Airtel/Vodafone/Jio/Aircel/MTS/IDEA/Tata/Other.
Loud Speaker	OFF / ON / Hands free
Environment	Studio/Office/Home/Roadside Other

4.2.2. *Input by Field Agents*

Field Agents or instructors are responsible for speaker selection; make them understand the objective, IVR phone call initiation, smooth continuation and successful completion of individual data collection. At the initial part (agent part) of data collection call flow program, provisions could be made for agents to enter metadata information in form of DTMF (touch or key press type) input string. Thus, system can easily store metadata information and generate related statistics automatically on interpreting field by field values in that DTMF input string. A sample DTMF input string can be like, “SSSSLRRAA”, where SSSS represents Speaker ID, L is for language code, RR is for dialect region code and AA is for agent ID. Similarly, explicit metadata information can be represented by DTMF input string like this, “GMEADMSLB”. Here each letter can have one-digit value related to appropriate options (as in Table 1) for specific metadata fields like gender, mother tongue, education, age group, device type, its make, service provider, loud speaker mode and environment. Call flow program validates the DTMF information and in case of any wrong input, it reverts back to the user for correct input. After correct processing of all metadata input, telephone is then handed over to actual speaker for data recording.

4.2.3. *Data collection from speakers*

In this part of data collection, speech data is recorded by IVR system in response of some general, guided and free quarries as uttered by target speakers in field conditions. It is always good to start the automated data collection with some general queries to make speakers familiar with the system and act as expected while intended recording of guided or read out speech data in the next part. Data can be collected from new speakers as well as already registered speakers, if speakers can’t complete the entire recording in one sitting. For these, IVR only needs to save the last recording status into the linked backend database server.

4.3. Online recording verification and correction

This facility has been introduced to validate and correct the recordings at runtime by agents while recording by speaker is on the go. On sequential recording by IVR system over a long period, if speaker feels to have some rest, then the system will hold for some specific time duration. Besides, by any means, if the speaker overlooks or misses out something or speaks out any wrong input at some certain point, the same can also be corrected by pressing interrupt key, then re-record at that point and resume till end.

4.4. Speech data naming and storage

In parallel to speech data collection, proper naming and storage of the same is required so that in future it can be used properly for related applications. The storage hierarchy starts with a particular parent folder and then makes branches language wise, session wise and finally ends with speaker specific folders having the collected raw utterances within it. Besides, the entire IVR call including talks of both system and user sides are recorded using IVR call monitoring facility and stored automatically within a separate system specific folder. Individual call logs are also maintained in backend databases after automatic entry of the same via IVR call flow program. These logs include metadata and other inputs as given by field agents along with recording status information (like completed or intermediate stage reached etc.), call initiation and call end timings etc. These logs are helpful to generate data collection statistics, check recording status and verify metadata.

4.5. Audio Data Transcription

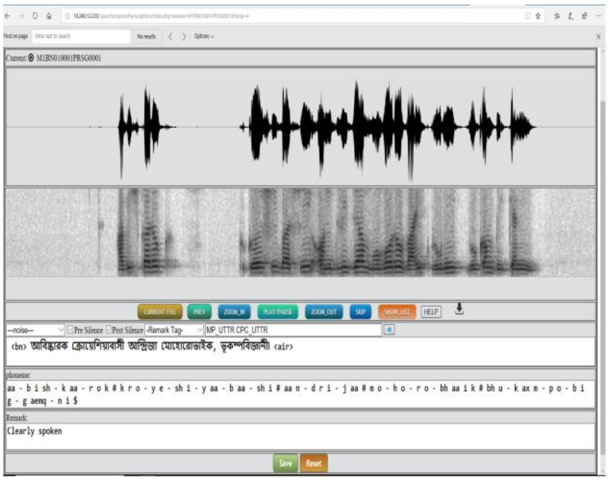
Goal of audio data transcription is to label the recorded auditory scene thereby preparing the audio data as per specific speech related application. Depending upon the target speech based application, these label can be sentence, phrase, prosody markers, word, syllable, phoneme or can be physical pauses, specific type of noise, speaker, language, emotion, speaking mode etc. Most of these applications (except automatic speech recognition or ASR), also require marking timing boundaries specific to audio events in a pre-defined format and that is primarily an offline manual process. Different strategies are discussed below to reduce manual efforts in transcription.

4.5.1. *Automatic Transcription (Online)*

Online transcription helps in phone level speech data transcriptions as required for PL creation in ASR and TTS applications. This automatic process nullifies the hassle of text data entering, reduces the chances of typo errors thereby making the entire task easy and faster than completely manual transcriptions. For read out speech data recording, default phonetic transcriptions can be provided automatically by IVR call flow program. But it’s not sure that speakers will always exactly pronounce the default way that is being labeled. Again correct transcription in word and phonetic level is a major resource for training correct and robust ASR models. Hence, an offline manual verification of the online transcriptions is necessary. Though, free speech and conversations are better to be transcribed offline with complete manual efforts.

4.5.2. Semi-automatic Transcription (Offline using tool)

Offline transcription is the next important activity, where output of online default transcriptions is checked, verified and corrected aligning to that of actually spoken in recorded audio. A semi-automatic transcription validation tool is used here to facilitate manual verification and correction process. Figure 2 shows snapshot of such a running transcription verification tool. The language specific resources that make this process fully functional are standardized phone set, implemented G2P conversion rules (like [20] for Bengali), application specific default PL, transcription remarks and possible noise tag set [21] (for speaker and environment noises).



MP_UTTR CPC_UTTR

<bn> আবিষ্কারক ক্রোয়েশিয়াবাসী আন্দ্রিজা মোহোরোভাইক, ভূকম্পবিজ্ঞানী <air>

aa - b i sh - k aa - r o k # k r o - y e - sh i - y aa - b aa - sh i # aa n - d r i - j aa # m o - h o - r o - bh aa i k # bh u - k ax m - p o - b i g - g aemq - n i $

Clearly spoken

Figure 2: *Web Based Speech Data Transcription Tool*

5. Sample IVR call flow

A call flow diagram is a graph that specifies series of tasks in an IVR application. It consists of elements (nodes representing specific stages in call flow) and transitions among elements. Call flow diagrams are organized as a tree structure where a parent element can contain several other child elements. Figure 3 shows such a call flow diagram specially designed for telephony speech data collection. It shows language selection as first element, then activity selection and after completing activity specific actions, returning to parent element by pressing specific DTMF inputs. Table 2 presents much detail actions related to every selected activity. The initial information code is the DTMF input to be given by instructors and this enables speaker specific speech data recording initiation, resume, correction, checking activities.

6. Conclusion

A detailed design framework for IVR based large scale speech data collection and transcription process is described in this paper. We are working on development of similar system to create speech resources in Bengali, Hindi and Indian English languages. IVR applications in Indian context are presently the need of time as more and more people are getting familiar with IVR based automated services. This is an added advantage for speech data collection from non-tech savvy speakers, along with maintenance and access of multilingual speech resources by IVR system at the same time. Easy availability of language specific resources also enables development of audio and text processing tools in more numbers which is really required for digitally under-resourced languages. Using this framework, we hope to work in similar line including some other Indian languages.

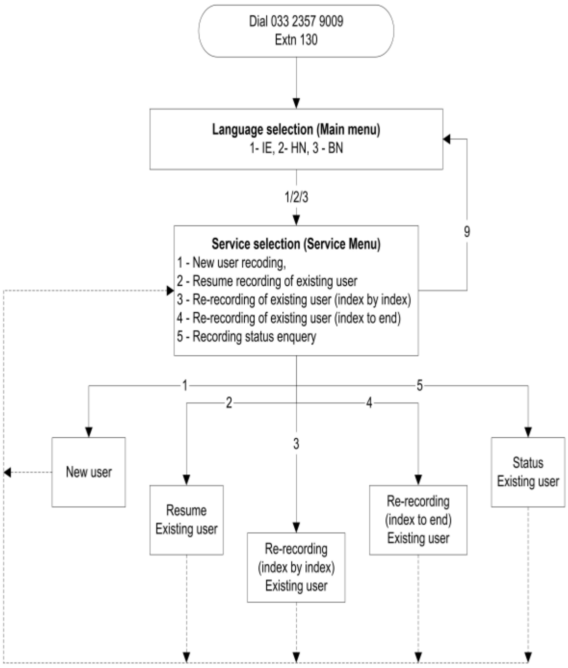

Dial 033 2357 9009
Extn 130
Language selection (Main menu)
1- IE, 2- HN, 3 - BN
1/2/3
Service selection (Service Menu)
1 - New user recoding,
2 - Resume recording of existing user
3 - Re-recording of existing user (index by index)
4 - Re-recording of existing user (index to end)
5 - Recording status enquery
9
1
2
3
4
5
New user
Resume
Existing user
Re-recording
(index by index)
Existing user
Re-recording
(index to end)
Existing user
Status
Existing user

Figure 3: *Primary Call Flow for speech data collection*

Table 2: *Detail Call Flow for Speech Data Collection*

A. Dial XXXXX with Extension 123
B. Language Selection: 1 for English / 2 for Hindi / 3 for Bengali (Language menu)
C. Activity Selection: (Service menu)
    a. New User: Press 1
        i. Press Information code
        ii. Recording
    b. Existing User: Press 2
        i. Press Information code
        ii. Recording
    c. Redo for specific node/serial number: Press 3
        i. Press Information code
        ii. Recording
    d. Redo from specific node/serial number till end: Press 4
        i. Press Information code
        ii. Press Serial number from where rerecording will start and continue till end
        iii. Recording
    e. Cross Validation of Recording: Press 5
    f. Return to language menu: Press 9

7. Acknowledgement

This work is a part of ongoing project “Indian Languages Speech Resources Development for Speech Applications” funded by MeitY, Govt. of India. The authors are thankful to all other members of speech corpora consortium for their constant support and co-operation.

8. References

[1] Stephen R. Anderson, "How many languages are there in the world?", *Linguistic Society of America*

[2] L. Besacier, V. B. Le, C. Boitet and V. Berment, "ASR and Translation for Under-Resourced Languages," *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, Toulouse, 2006. doi: 10.1109/ICASSP.2006.1661502

[3] Laurent Besacier, Etienne Barnard, Alexey Karpov, Tanja Schultz, "Automatic speech recognition for under-resourced languages: A survey", *Speech Communication*, Volume 56, January 2014, pp 85–100

[4] http://mhrd.gov.in/language-education

[5] Languages of India: https://en.wikipedia.org/wiki/Languages_of_India

[6] Joyanta Basu, Milton Samirakshma Bepari, Rajib Roy and Soma Khan, “Resource Building Methodology for Designing IVR based Bangla Speech Recognition System”, in Proc. of Oriental COCOSDA 2012, 9th-12th December, 2012, University of Macau, China, pp 101-106.

[7] Joyanta Basu, Soma Khan, Rajib Roy and Milton Samirakshma Bepari, “Commodity Price Retrieval System in Bangla: An IVR Based Application”, *APCHI '13 Proceedings of the 11th Asia Pacific Conference on Computer Human Interaction*, Pages 406-415, Bangalore, India — September 24 - 27, 2013, ISBN: 978-1-4503-2253-9

[8] Linguistic Data Consortium (LDC), various corpus resources on http://www.ldc.upenn.edu

[9] European Language Resources Association (ELRA). http://www.elra.info/

[10] R. Duncan and J. Picone, "A Unix-based speech data collection platform," Southeastcon '99. Proceedings. IEEE, Lexington, KY, 1999, pp. 29-31.

[11] D. Langmann, R. Haeb-Umbach, L. Boves and E. den Os, "FRESCO: the French telephone speech data collection-part of the European Speechdat(M) project," Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on, Philadelphia, PA, 1996, pp. 1918-1921 vol.3.

[12] C. Veaux, J. Yamagishi and S. King, "The voice bank corpus: Design, collection and data analysis of a large regional accent speech database," 2013 International Conference Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE), Gurgaon, 2013, pp. 1-4.

[13] S. Nakamura et al., "Data collection and evaluation of AURORA-2 Japanese corpus [speech recognition applications]," 2003 IEEE Workshop on Automatic Speech Recognition and Understanding (IEEE Cat. No.03EX721), 2003, pp. 619-623.

[14] Pukhraj P. Shrishrimal, Ratnadeep R. Deshmukh, Vishal B. Waghmare, "Indian Language Speech Database: A Review", *International Journal of Computer Applications (0975 – 888)* Volume 47– No.5, June 2012, pp - 17 – 21

[15] http://tdil-dc.in/index.php?option=com_vertical&parentid=58&lang=en

[16] S. S. Agrawal, "Developments in Text, Speech Corpora for Different Indian Languages and their Applications", *Country Report -India*, O-COCOSDA 2016

[17] Gomillion D, Dempster B., “Building Telephony System with Asterisk”, ISBN: 1-904811-15-9, Packet Publishing Ltd., 2005

[18] Asterisk Gateway Interface (AGI): https://www.voip-info.org/asterisk-agi/

[19] G. Gelly and J. L. Gauvain, "Optimization of RNN-Based Speech Activity Detection," *in IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 3, pp. 646-656, March 2018

[20] Joyanta Basu, Tulika Basu, Mridusmita Mitra, Shyamal Kr Das Mandal, “Grapheme to Phoneme (G2P) Conversion for Bangla”, *O-COCOSDA 2009*, pp. 66 – 71.

[21] J. Basu, M. S. Bepari, S. Nandi, S. Khan and R. Roy, "SATT: Semi-automatic transcription tool," *2013 International Conference Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)*, Gurgaon, 2013, pp. 1-6.