Rapid Development of Speech to Speech Systems for Tourism and Emergency Services in Indian Languages

Anandaswarup V*, Karthika M*, Nagaswetha G*, PK Narne*, VV Vinay Babu*, Mrudula K*, Poornima T*, RR Patil*, CMS Raju*, Snehata T*, Azharuddin S*, Abhilash B*, P Raju*, GSC Prasad*, Sriram A*, E Veera Raghavendra*, Sachin Joshi*, Vamshi Ambati† and Kishore S Prahallad*†
* International Institute of Information Technology, Hyderabad, India
†Carnegie Mellon University, Pittsburgh, USA

***Abstract*—In this paper we deal with the development of a speech to speech system for tourism and emergency services in Indian languages.We discuss the development of the the speech synthesis (TTS), speech recognition (ASR) and machine translation (MT) subsystems of the speech to speech system.We also describe the evaluation of the system and present the results of the evaluation.**

I. Introduction

A. The diverse nature of Indian Languages

India is a diverse country with a plethora of languages.The official languages of India are Hindi and English.Apart from these two languages the following 17 languages are also recognized by the constitution of India : 1) Assamese 2) Bengali 3) Gujarati 4) Kannada 5) Kashmiri 6) Konkani 7) Malayalam 8) Manipuri 9) Marathi 10) Nepali 11) Oriya 12) Punjabi 13) Sanskrit 14) Sindhi 15) Tamil 16) Telugu and 17) Urdu. As a result of this vast diversity of languages, communication is very difficult, especially between persons from different parts of the country.Such situations demand the development of speech to speech systems (STS systems) as they ease communication between groups who do not know a common language.

B. Motivation for this work

The rich cultural heritage of India had always attracted tourists.In addition to this, India's recent emergence as an IT and ITeS hub has resulted in an influx of people from other parts of the country.As a result there has been a requirement for STS systems in Telugu and Hindi. In this paper we demonstrate the rapid development of STS systems in Telugu and Hindi, which could be extended to other Indian languages as well.

II. Context of the problem

To place the problem in context, we present a small scenario below.

- A tourist lands at the airport. The tourist has no knowledge of the local language and can communicate only in English.
- The tourist then wants to go to a hotel and needs transportation there. This entails communicating with the taxi drivers and other local transportation people at the airport.
- Once he/she is at the hotel the tourist needs to interact with the hotel and restaurant staff.
- To go site seeing at the local tourist places again entails communication with taxi drivers and other local transportation people at the hotel.
- In case of any emergency, the tourist needs to interact with emergency services.
- To get back to the airport, the tourist once again needs to communicate with the local transportation people.

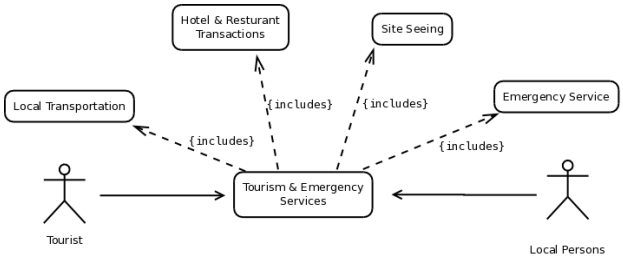

Fig. 1. Behavioral Representation of the system

A behavioral representation of the system capturing the interaction based requirements of the system is given in Figure 1. The representation comprises of two actors and four use cases describing the system's behavior.

III. Parallel Data Collection

Based on the collection of the possible usage scenarios, the broad domain of tourism and emergency services was divided into four different sub domains : 1) Local travel (D1) 2) Hotel and restaurant transactions (D2) 3) Tourism (D3) and 4) Emergency services (D4). This was done with a view to facilitate and simplify the design of the speech corpus. The quality of the translation and synthesis depends largely upon the variability and availability of the representative units. This fact has to be kept in mind while designing the

corpus. The corpus should be large enough to cover all the speech units and their variations within a reasonable size. Taking into consideration the above constraints, we designed the speech corpus the details of which are presented in Table 1.

	Number of sentences		
	English	Telugu	Hindi
D1	204	204	-
D2	206	206	-
D3	316	316	-
D4	-	231	231

Table 1 : Speech corpus details

The factual data was obtained from help desks of the Andhra Pradesh Tourism Development Corporation, the Emergency Management Response Institute, local taxi agents and few restaurants and hotels of varied ratings. A few examples are :

P (English/Telugu) : $\frac{Take \quad me \quad to \quad hotel \quad Dwaraka}{dvaaraka \quad hoot:alki \quad tiisukel:lu}$

P (English/Telugu) : $\frac{What \quad rooms \quad are \quad available?}{ei \quad vidhamaina \quad ruums \quad unnaayi?}$

P (Hindi/Telugu) : $\frac{aapakii \quad samasyaa \quad baataayiye}{mii \quad samasyanu \quad cheppan:d:i}$

The speech databases used for English, Telugu and Hindi were recorded by 15 different speakers. All the recordings were done using a laptop and a standard microphone in a quiet room with minimal background noise. The sentences were read in a relaxed reading style at a moderate speaking rate.

IV. System Building

A typical speech to speech system consists of three distinct components : 1) Speech recognition system (ASR system) 2) Machine translation system (MT system) and 3) Speech synthesis system (TTS system). These three components are loosely coupled to form a speech to speech system [1].

A. ASR System

The ASR system is the speech recognition component of the speech to speech system. The function of the ASR system is to convert spoken utterances into the corresponding text. Typically ASR systems comprise of three major components : 1) Acoustic models 2) Language models and 3) Phonetic lexicon [2].

Acoustic models capture the characteristics of the basic recognition units. The recognition units can be at the word level, syllable level or at the phoneme level. For large vocabulary ASR systems phonemes are the preferred units [2]. The language model attempts to convey the behavior of the language. At the time of recognition, various words are hypothesized against the speech signal. To compute the likelihood of a word, the lexicon is referred to and the word is broken into its constituent phones. The phone likelihood is computed from the acoustic model. The combined likelihood of all the phones represents the likelihood of the word in the acoustic model. The word having the highest likelihood is selected as the result of recognition.

The open source Sphinx framework was used to build the ASR system. The Sphinx framework functions in two phases : 1) Training, which is the process by which the system learns about the sound units and 2) Decoding or recognition, which is the process of computing the most probable sequence of units based on the training.

The following components were required by the Sphinx trainer to train the system : a) Acoustic signals b) Transcript file c) Language dictionary and d) Filler dictionary. We modified the corpus file by placing delimiters and sentence labels, to ensure that the resulting transcript file conforms to the requirements of the Sphinx trainer. Initial language dictionaries were generated by the Sphinx knowledge base tool [3]. These were manually checked and modified wherever necessary to produce the language dictionaries. A standard filler dictionary was used. The Sphinx trainer was then run to train the system and generate acoustic models.

Once the training was complete the Sphinx-3 decoder was used to perform the recognition task. The inputs of the decoder were : a) the trained acoustic models b) the language model c) the language and filler dictionaries used during training and d) a set of acoustic signals which needed to be recognized. The language model was generated using the CMU SLM toolkit [4]. The Sphinx decoder was then run on the test data.

In each language, the system was tested using 200 utterances recorded by 10 different speakers, where each speaker contributed 5 utterances from each domain. The utterances were decoded using the Sphinx decoder. Evaluation of the performance was made according to the recognition accuracy and computed using the word error rate (WER) metric. The WER is calculated by aligning the decoded utterances against the original transcription and computing the number of substitutions (S), deletions (D) and insertions (I) in the decoded utterance and the number of words in the correct sentence(N).

$$WER = (S + D + I)/N \tag{1}$$

$$RecognitionAccuracy = 1 - WER \tag{2}$$

The results of this evaluation are given in Table 2.

	Recognition Accuracy		
	English	Telugu	Hindi
D1	0. 90	0. 83	-
D2	0. 97	0. 90	-
D3	0. 85	0. 84	-
D4	-	0. 90	0. 87

Table 2 : Recognition accuracy of ASR

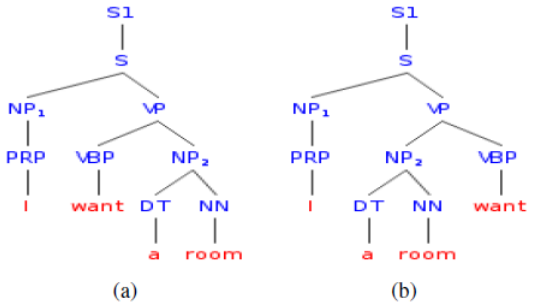

Fig. 2. (a) Original Parse Tree (b) Parse Tree after rearrangement

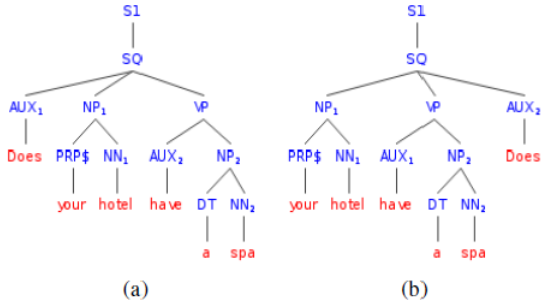

Fig. 3. (a) Original Parse Tree (b) Parse Tree after rearrangement

B. MT system

The MT system translates text in one language into another language. It can be defined as the task of automatically converting one natural language into another, preserving the meaning of the input text, and producing fluent text in the output language. Commonly used approaches to building MT systems include : 1) Rule based machine translation 2) Statistical machine translation and 3) Example based machine translation. Rule based machine translation paradigm includes transfer based MT, interlingual MT and dictionary based MT paradigms [5].

The main issues in the development of the MT system to translate English into Indian languages were the differences in the structure and grammar of the languages. English is a SVO (Subject-Verb-Object) language, whereas Indian languages (especially Telugu and Hindi) are SOV (Subject-Object-Verb) languages [6]. Therefore, some amount of syntactic analysis and reordering of the corpus sentences was necessary before proceeding to dictionary based translation.

Syntactic analysis of the corpus sentences was done using the Charniak parser. This analysis was used to develop a transfer grammar to reorder the English sentences in order to perform English to Indian language translation. This grammar follows two rules. The first is reversal of child nodes of the verb phrase (VP). We demonstrate this by means of a small example. Consider the sentence : *I want a room*. Syntactic analysis of the sentence using the Charniak parser yields the parse tree shown in 2(a). After applying this rule the reordered parse tree of the sentence is shown in 2(b). The second is transfer of the auxiliary verbs(AUX) to the end of the sentence. We demonstrate this rule by means of another example. Consider the sentence : *Does your hotel have a spa*. Syntactic analysis yields the parse tree shown in 3(a). After applying this rule the reordered parse tree of this sentence is shown in 3(b).

In parallel with the development of the transfer grammar, a lexicon was also built from the corpus. The lexicon served as the basis of the dictionary based MT system. During translation, the sentences were first reordered using the transfer grammar. The lexicon was then used to substitute each word with its corresponding Indian language counterpart. English to Indian language translation was performed in this manner.

Indian language to English translation was achieved by means of an example based MT system. Every sentence in the Indian language was uniquely identified using a keyword. A hash table was used to map each keyword to the corresponding English sentence.

Indian language to Indian language translation was effected by using a lexicon based technique. This was possible because both languages under consideration are SOV languages. An example would be : *mii samasyanu cheppan:d:i*, which is in Telugu is translated into Hindi as *aapakii samasyaa baataayiye*.

The system was tested by taking 60 sentences in English, 80 sentences in Telugu and 20 sentences in Hindi. These test sentences were translated by running the system on them. An evaluation of the of the MT system was done by calculating the translation accuracy using WER (1).

$$TranslationAccuracy = 1 - WER \tag{3}$$

The results of this evaluation are presented in Table 3.

	Translation Accuracy			
	E-T	T-E	H-T	T-H
D1	0. 74	0. 75	-	-
D2	0. 80	0. 78	-	-
D3	0. 74	0. 74	-	-
D4	-	-	0. 75	0. 75

Table 3 : Translation accuracy of MT System.
(E : English, T : Telugu, H : Hindi)

C. TTS system

The function of a TTS system is to convert the given text into a spoken waveform. This conversion involves text processing and speech generation processes. Data driven synthesis is the approach most commonly used to build TTS systems. This approach seeks to develop strategies for concatenating stored speech segments as a means of synthesizing speech. Sub-word units, such as syllables or diphones, in which co-articulation between adjacent phonemes are preserved, are considered as satisfactory units, under this approach to synthesizing speech [7].

The Festvox framework was used to develop an Indian voice, which was then used in the Festival speech synthesis system to synthesize speech in Telugu and Hindi. The first step in building an Indian language voice was to define a phoneset along with grapheme to phoneme conversion rules for Telugu and Hindi. These were used to label the recorded speech database at the phone level, which was achieved using the labeler provided by Festvox. Since accurate duration knowledge was not available for Telugu and Hindi phones, the label boundaries generated by the labeler were not accurate and had to be corrected manually. Festvox was then used to extract the pitch markers and the Mel-Cepstral coefficients which were utilized by Festvox to build a decision tree for each unit based on the the phonemic and prosodic content of that unit.

The unit selection algorithm of Festival was then used to select an appropriate decision tree and search for a suitable manifestation of the unit such that the cost of joining two adjacent units was minimized. The joining of all the units resulted in synthesized speech. Telugu and Hindi speech were synthesized by the Festival system in this manner. The speech synthesized by the TTS system was perceived to be fairly intelligible and natural by native speakers of the language.

A perceptual evaluation conducted on the TTS system for the 4 sub-domains by 20 random listeners yielded the following results on a five point scale (1:Bad, 2:Poor, 3:Fair, 4:Good, 5:Excellent).

	Mean Perceptual Scores		
	English	Telugu	Hindi
D1	3. 24	3. 24	-
D2	3. 50	3. 40	-
D3	3. 60	3. 20	-
D4	-	3. 75	3. 75

Table 4 : Perceptual scores of TTS

V. Evaluation of the speech to speech system

A simple and aesthetically appealing user interface (Figure 3) was created for evaluation of the system. The speech to speech system was subjected to a perceptual evaluation by 20 random listeners. Listeners were asked to grade the output of the system on a five point scale (1:Bad, 2:Poor, 3:Fair, 4:Good, 5:Excellent). The results of the perceptual evaluation are presented in Table 5.

Fig. 4. User Interface

Language	Mean Perceptual Scores
E-T	3. 19
T-E	3. 20
H-T	3. 43
T-H	3. 24

Table 5 : Perceptual scores of the speech to speech system. (E : English, T : Telugu, H : Hindi)

In each language, the system was tested using 200 utterances recorded by 10 different speakers, where each speaker recorded 5 sentences from each domain. An evaluation of the experiment was done by calculating the system translation accuracy using WER (1).

$$SystemTranslationAccuracy = 1 - WER \quad (4)$$

The results of this evaluation are presented in Table 6. From the results we see that the speech to speech system has a fairly reasonable accuracy. The system has also scored reasonably well on the perceptual evaluations.

Language	Translation Accuracy
E-T	0. 78
T-E	0. 74
H-T	0. 75
T-H	0. 73

Table 6 : Translation accuracy of the speech to speech system. (E : English, T : Telugu, H : Hindi)

VI. Challenges

Difficulties in developing SST systems include variations in pronunciation of words (especially in English) and grammatically ill formed test sentences. Specific challenges include generation of language dictionaries, phonesets and grapheme to phoneme conversion rules for Indian languages.

VII. Conclusions and scope for future work

In this paper, we described the development of Indian language speech to speech systems for tourism and emergency services. We discussed the parallel data collection and design of the speech corpus. We developed an Indian language speech recognition system using the Sphinx framework, which has good accuracy. A translation system was developed to perform English to Indian language translations with good accuracy. We developed an Indian language voice using Festvox and built an Indian language TTS system. The output of the TTS system was found to be reasonably fair by native speakers of the language. Finally, we integrated the three components to form the speech to speech system. The system was subjected to perceptual evaluations and the output of the system was found to be fair by the listeners.

Future work will directed towards the development of a large scale robust system for the current prototype. Specific areas of focus include improving the quality of the TTS synthesizer. Experiments using syllable like units as the basic units of concatenation can be performed in order to incorporate more naturalness in the synthesized speech signal. Future research can also focus on incorporating noise recognition in the ASR system. This would reduce the impact of external noise on the accuracy of the speech recognition.

Acknowledgment

We thank all the people who had contributed to the development of the system and participated in the perceptual evaluations.

References

[1] Kishore Prahallad, "A Direct Approach for Speech to Speech Translation, A Project Report in Machine Translation Course", LTI-CMU, Spring 2005.

[2] Anumanchipalli Gopalakrishna et al. , "Development of Indian Language Speech Databases for Large Vocabulary Speech Recognition Systems", in Proc. Int. Conf. Speech and Computer (SPECOM), Patras, Greece, October 2005.

[3] http://www. speech. cs. cmu. edu/tools/lmtool. html.

[4] http://www. speech. cs. cmu. edu/SLM_info. html.

[5] Hutchins, W. John; and Harold L. Somers. "An introduction to machine translation", London:Academic Press, 1992.

[6] Shachi Dave et al. , "Interlingua-based English-Hindi machine translation and language divergence", Kluwer Academic Publishers, 2003.

[7] S. P. Kishore, Rohit Kumar and Rajeev Sangal. , "A Data-Driven Synthesis approach for Indian Languages using Syllable as Basic Unit", in Proceedings of International Conference on Natural Language Processing (ICON), 2002.