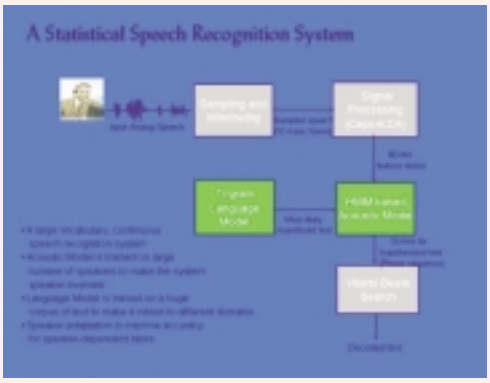
IBM India Research Lab

Speech Technologies The design and implementation of computer systems that can cover the whole range of human-like interaction by using faces and voices is one of the challenging objectives of the Speech Group at the India Research Lab (IRL). The group aims to build state-of-the-art speech recognition/synthesis systems for Indian languages and improve the existing techniques in general. IRL is currently pursuing activities in the areas of: I. *Indian Language Desktop Speech Recognition II. Indian Language Telephony Speech Recognition III. Indian-English Text-to-speech synthesis I.*

Indian Language Speech Recognition Speech recognition can be very useful as it provides an easy interface for interacting with computers for those unfamiliar with computers and/or English. Since there is no common standard for keyboard input for Indian languages, speech recognition can eliminate the need to know the keyboard mapping for the different standards and related problems. Building a speech recognition system that would help a machine understand and transcribe human speech has been a challenging field of research. IRL is addressing the problem through statistical techniques. For any given language, the computer first needs to learn the sounds of spellings in various contexts from a huge corpus of speech and the corresponding transcribed text. This is known as the training phase of the recognition system. To train a recognition system for a new language, the system needs to learn about the grammar/syntax (known as the Language Model) of the language and about its sound-spelling relation (known as the Acoustic Model). Speech is divided into phonemes, which are the different, distinguishable sounds specific to a language. Each of these phonemes is then statistically modeled (Hidden Markov Models, in our case) in the training stage. Once the training phase is complete, the system is ready for speech recognition. Given a speech utterance for recognition, several sentences are hypothesized and the probabilities for the sequence of phonemes are calculated using the Acoustic Model and the Language Model. The most likely sentence is then chosen as the text corresponding to the input speech and is the decoded output of the recognition system. Complex signal processing and statistical techniques are used to make the recognition robust to speaker speech variations and to make it work on continuous speech of a large vocabulary. The technology: At IRL, we are extending the IBM ViaVoice recognition technology to build a speech recognition system for the Hindi language. We have identified a phone set consisting of 61 phonemes for Hindi. A mapping from the English phone set to Hindi phone set was used to build the initial phone models for Hindi. Using these models, we have built an acoustic model for Hindi by training it on a large sample of Hindi acoustic data. We have the set up to collect and transcribe speech samples that are used to train the acoustic models. We have also built a statistical language model for Hindi, which captures the language statistics from a huge amount of text data. Currently we are working on increasing the accuracy of our Hindi speech recognition system by enhancing the Acoustic and Language models. The recognition system performs well over a trained context and we have a prototype demonstration of a large vocabulary continuous speech Hindi recognition system at the Lab. Our goal is to cover more Indian languages and then to build a multilingual speech recognizer for the Indian languages based on a multilingual phone set. The IBM Desktop Hindi Speech Recognition System is a continuous speech, speaker-independent recognizer. The system has a vocabulary of 65,000 words (to increase to 125,000). With speaker training, accuracy levels in excess of 90 percent can be achieved. A novel technique of building phonetic spellings of Hindi words has been used to generate the vocabulary in the system. The system displays output in Unicode Hindi characters. Algorithms have been developed to perform speaker adaptation of the system in order to improve accuracy with speaker training. The system provides capabilities to add user-specific words to the existing vocabulary. These technologies aid in adapting the IBM ViaVoice recognition system for the Hindi language.

II. Telephony Speech Recognition Speech recognition over the telephone line is a challenging task owing to the 8 kHz band limited speech signal and the noise introduced in the telephone channel.



Also, since the speaker is more spontaneous in the conversations, the disfluencies and spontaneity of the input speech also need to be handled. However, by training a recognition system on the context in which it is to be used, we can increase the recognition accuracy of the system for that particular domain. For example, telephony speech recognition can be used in automating/improving some of the telephone banking, directory assistance, inquiry systems. Since the context in each of the tasks is fairly restricted (e.g. banking terms are very different from directory assistance) as the caller is looking for some specific information, recognition accuracy can be improved by making the system specific to the task. Such an application is based on understanding the recognized speech, taking a corresponding action (by a back end server) and then generating and playing back an appropriate response to the caller through speech synthesis. Following are some of the advantages of using such speech-based solutions: I. More access devices: There is more telephone usage than Internet, and it's rapidly increasing. Different kinds of access devices are possible, such as wired, cordless or mobile phones, WAP, handhelds, etc. II. Ease of use: Access is possible through easier to use devices. More people are telephone literate than Internet literate. Information can be provided in a natural manner, such as speech or on WAP phone display. III. New revenue source: These new devices and modes of communication bring with them additional sources of revenue. The technology: We are working on a system that would help develop telephony applications using Hindi and Indian-Hindi speech recognition to automate the process of these applications. Speech can serve as a means of input as compared to punching digits on a telephone keypad, which is used in most Interactive Voice Response (IVR) systems today. The main research challenges in the area of telephony Hindi speech recognition are to handle the spontaneity of the speaker, the telephone noise conditions and the various accents in order to make the system completely independent of speaker voice variations. The group also aims to build a Hindi speech synthesizer to make the telephony application system completely automatic. We have built an early prototype of an acoustic model for Hindi. For the telephony applications, one approach is to eliminate the need for a language model by using prompt-specific grammar. The grammar, of course, needs to capture all the variations that the user could say for a particular prompt. When the system asks for a request from the user, it expects the answer to be in the domain of all the possible answers to that question. A grammar for such probable answers can be created and used along with the Acoustic Model to decode the speech in response to a given prompt. This way we have a grammar for each question asked by the dialogue manager. In order for the system to handle queries that have a broader domain, Natural Language Understanding (NLU) for the Hindi language is required and NLU is one of the research goals of the group.

III. Indian-English Text-to-Speech Synthesis IRL is also working on building text-to-speech synthesis systems for Indian languages. We have developed a prototype of a concatenative Indian English text-to-speech synthesis system which is capable of synthesizing speech in Indian accent and can pronounce Indian names with very high accuracy. The system uses a multilingual phone set consisting of phones from both English and Hindi. A large number of sentences are recorded in the voice of a processional speaker and this database is processed through Indian English speech recognition system.

Related Publications:

[1]. Mohit Kumar, Nitendra Rajput, Ashish Verma, "A large vocabulary continuous speech recognition system for Hindi," *IBM Journal of Research and Development (Special Issue on IBM Research in Asia)*, to appear.

[2]. Ashish Verma, Nitendra Rajput , L. V. Subramaniam, "Using Viseme based Audio HMM for Audio Driven Facial Animation," to appear in *International Conference on Acoustics, Speech and Signal Processing* (ICASSP) 2003, HongKong, April, 06-10, 2003.

[3]. Ashish Verma, L. V. Subramaniam, Nitendra Rajput, Tanveer Faruquie, Chalapathy Neti, "Animating expressive faces across languages," *IEEE Transactions on Multimedia*, to appear.

[4]. Nitendra Rajput, L. Venkata Subramaniam, Ashish Verma, "Adapting Phonetic Decision Trees Between Languages for Continuous Speech Recognition," Proceedings: *IEEE International Conference on Spoken Language Processing* (ICSLP 2000), Beijing, China, Oct 16-20, 2000.

[5]. Niloy Mukherjee, Nitendra Rajput, L. Venkata Subramaniam, Ashish Verma, "On Deriving a Phoneme Model for a New Language," *Proceedings: IEEE International Conference on Spoken Language Processing* (ICSLP 2000), Beijing, China, Oct 16-20, 2000.