5. *Indo European IEMCT Conference*

Indo-European Conference on Multilingual Communication Technologies at C-DAC, Pune on 24-25th June, 2002

Theme

The objective of the conference was to dwell with the technologies and their applications which will help convert Digital Divide into Digital Unite. This is achieved by bringing together the researchers and their ideas from the two continents on a common platform.

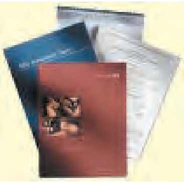
The Indo-European Conference on Multilingual Communication Technologies (IEMCT) was organised to present, discuss and explore the latest advancements in the emerging field of Multilingual Communication Technologies and its potential future directions. The key focus areas of the conference were Optical Character Recognition, Speech Technologies and Internet Tools for content creation and its retrieval in South Asian and European Languages.

The conference was organised by Centre for Development of Advanced Computing (C-DAC), India, in association with Maison des Sciences de l'Homme, France and Asociacion de Industrias de las Tecnologias Electronicas y de la Informacion del Pais Vasco, GAIA, Spain. The Conference took place on 24th and 25th of June in Pune, India. It was spread over two days, keeping in mind the varied research topics and large number of speakers. A total of 16 speakers presented their papers in three sessions; dedicated to, "Optical Character Recognition", "Text to Speech" and "Content Creation and Information Retrieval".

Inaugurating the conference, Shri Rajeeva Ratna Shah said, " Language technology has a very pivotal role to play in creating a "Digital Unite". With only 5-7% of the total population of India being computer literate owing to dominant use of the English language in computers, India can be the hub for the developments of technologies such as Text-to-Speech and Speech-to-Text. Shri Shah said that the developments in Language Technology could be in sync with the Media Lab Asia's aim of delivering a World Computer- low cost and inter-lingua system, providing Bits for all- connectivity/ bandwidth as also broadband to average Indian citizens and creating a Digital Village. Shri Shah said he believed that C-DAC has the potential to create the IT revolution in India by taking IT to the masses through its expertise in Language Technology". He further added, "It was time for C-DAC to extend its expertise to Web technologies and Speech Technology".

Delivering the Keynote Address, Dr S. Ramani, Director, HP Labs India said that at present Speech Interfaces in Indian Languages do not have widespread applications. In order to establish a stronghold in the Speech technology he felt it was necessary to create acoustic models for Indian languages, adapt Text-to-Speech systems using speech elements from Indian Languages. Over and above he felt it was essential to develop prototype applications, in parallel, anticipating improvements in Automatic Speech Recognition (ASR) and Text-to-Speech (TTS).

Dr. Augustin Benoist said, it is the first time that European Union undertakes a real initiative of co-operation between European Union and India in the field of Information system and technologies of Communication. The Programme aims to improve co-operation between Asia and Europe in the identification and implementation of information-technology solutions. We had chosen these



topics because they are of the utmost importance for the future of India. Just a word with the example of the web, but it applies to text to speech as well. Because of the web, India and Europe which are 6000 miles one from each other have become suddenly now few seconds one form each other.

Session on Text-To-Speech

Mr. Asok Bandyopadhyay's paper highlighted some aspects involved in developing Bangla speech synthesis system. The other main aspect in developing the Indian Language TTS system involving feature analysis for classification of speech under stress was presented by Dr. S. Dandapat. Mr. Amit Kumar Mishra's paper presented an approach using wavelets for speech processing during his presentation. Mr. Prachish Chugh discussed about speaker independent word recognition during his presentation.

Session on Optical Character Recognition

The nonlinear shape of the characters creates a lot of problems in recognizing the characters. Dr. Atul Negi's paper described the solution for this problem, by normalizing the characters non-linearly. The other main problem in developing the Indian Language OCR software is the presence of Roman characters in the Indian Language text. Dr. Lehal 's paper was on how to detect the Roman characters (especially in Gurumukhi script) and Mr. Aditya's paper described the method to detect and recognize the language of the character. Mr. Kunte's paper was on online character recognition of south Indian languages like Kannada, Telugu and Tamil, which is very useful in the current scenario with the increasing demand for the development of advanced man-machine interfaces. Mr. Thool's paper is was recognising the characters using statistical pattern recognition technique and Mr. Thokal's paper is on neural network based character recognition. The other problem in commercializing the OCR is, with the images containing both pictures and text, as no current Indian language OCR software supports this feature, Mr. Aditya's paper provided a solution for this.

Session on Content Creation and Information Retrieval

The session addresses the issue of creating documents in Indian languages that would go on the World Wide Web and the Search Engines, that need to be developed for these languages to retrieve documents. Mr Shashank Bhatt's paper spoke about the data representation in the Indic Web pages today, that Search Engines would have to process; the issues involved and possible solutions. Ms Rupali Sharma's paper talked about the advantages of maintaining phrase-based text representation in the Search Engines database. Mr P. C. Reghu Raj spoke about efficient document categorization in the Search Engine. Mr M. Saravanan described the technical aspects of a module in the Search Engine that summarizes a document. Dr Chantal Enguehard spoke about tools for keyword extraction from a corpus and keyword recognition. Mr Sanjiv Burman spoke about the different ways in which content can be created in Indian languages.

Conclusion

The conference underlined the need for continuous research in the three specific areas, viz. Speech Technologies, OCR & Content Creation and Information Retrieval.

The Text-To-Speech session was a very informative session. Three approaches to Speech Recognition were addressed and an approach to Speech Synthesis was also presented. The OCR session provided possible solutions for the recognition of Roman script in text containing Indic scripts, methods to improve recognition rates; methods to give a better training to the system for efficient recognition; separating images from text and the noise removal techniques. The Content Creation and Information Retrieval session dealt with different perceptions opened up by the cutting edge researches. Language independent term distribution model came out as a strong proposition for South Asian and European languages. The need was stressed to test "content identification and Semantic indexing" model on Indian and European languages. Some language tools have been developed for European languages which can very well be modified to suit Indian Languages.

The conference concluded with identifying specific areas of work of interest with relevance to Indian & European languages.