



Performance & Analysis of Encrypted Approach on Speech to Text Converter

¹G.Sankar Babu, ²Dr.M.Anto Bennet, ³R.Kaushik Krishna, ⁴S.Jaya Prakash, ⁵B.S.Jayavignesh

^{1,2}Professor, Department of Electronics and Communication Engineering, ^{3,4,5}UG Student,
VEL TECH, Chennai-600062, Tamil Nadu, INDIA

Abstract: Speech recognition is the analysis side of the subject of machine speech processing. The synthesis side might be called speech production. These two taken together allow computers to work with spoken language. My study concentrates on isolated word speech recognition. Speech recognition, in humans, is thousands of years old. On our planet it could be traced back millions of years to the dinosaurs. Our topic might better be called automatic speech recognition (ASR). We give a brief survey of ASR, starting with modern phonetics, and continuing through the current state of Large-Vocabulary Continuous Speech Recognition (LVCSR). A simple computer experiment, using MATLAB, into isolated word speech recognition is described in some detail. We have experimented with several different recognition algorithms and I used training and testing data from two distinct vocabularies. My training and testing data was collected and recorded with both male and female voice.

Keywords: Large-Vocabulary Continuous Speech Recognition (LVCSR), Automatic Speech Recognition (ASR)

I. INTRODUCTION

Real-time speech-to-text-conversion aims at transferring spoken language into written text (almost) simultaneously. This gives people with a hearing impairment, access to the contents of spoken language in a way that they e.g. become able to take part in a conversation within the normal time frame of conversational turn taking. Another scenario for real-time speech-to-text-transfer is a live broadcast of a football match where the spoken comments of the reporter are so rapidly transferred into subtitles that they still correspond to the scene the reporter comments on. An example from the hearing world would be a parliamentary debate which ends with the electronic delivery of the exact word protocol presented to the journalists immediately after the end of the debate. (cf. Eugeni, forthcoming) This list could be easily continued. However, most people with a hearing disability do not receive real-time speech-to-text services at counselling interviews, conferences or when watching a sports event live on TV. Most parliamentary protocols are tape recorded or written steno typed and subsequently transferred into readable text[11].

Automatic speech recognition (ASR) technologies today can correctly recognize and write down more than 80% percent of a long series of spoken words for many languages. However, even this high percentage is not sufficient for speech-to-text services. So we have to provide people a more flexible way to use speech to text conversion. Language is a very fast and effective way of communicating. To use language means to express an unlimited amount of ideas, thoughts and practical information by combining a limited amount of words with the help of a limited amount of grammatical rules. The result of language production processes are series of words and structure. Series of words are produced– i.e. spoken or signed – in a very rapid and effective way. Most people use oral language for everyday communication, i.e. they speak to other people and hear what other people say. People who are deaf or hard-of-hearing do not have equal access to spoken language, for them, precondition 2 is not full filled, their ability to receive speech is impaired. If people who are severely impaired in their hearing abilities want to take part in oral Communication, they need a way to compensate their physical impairment. Hearing aids are sufficient for many hearing impairment people. However, if hearing aids are insufficient spoken language has to be transferred into a modality which is accessible without hearing, e.g. into the visual domain. Speech-to-text-translation (audiovisual translation) of spoken language into written text is an upcoming field since movies on DVDs are usually sold with subtitles in various languages. While the original language is given auditorily, subtitles provide a translated version in another language at the same time visually. The audiovisual transfer from the spoken original language into other languages which are presented in the subtitles can be called an interlingual audiovisual translation. Interlingual translation aims at transferring messages from one language into another language. This translation process combines with a transfer from spoken language patterns into written text patterns. Auditory events which are realized as noises or speech melodies would often not be transferred because normally hearing people can interpret them by themselves. Interlingual translation primarily addresses the lack of knowledge of the original language, i.e. the first precondition for understanding language. The intralingual audiovisual transfer differs in many aspects from the interlingual audiovisual translation between two languages. First of all, intralingual audiovisual transfer for

people with hearing impairments addresses primarily precondition 2, i.e. the physical ability to perceive the speech signals[12-14]

II. LITERATURE SURVEY

One of the main problems in dynamic time-warping (DTW) based speech recognition systems are the preparation of reliable reference templates for the set of words to be recognized. This paper presents a simple novel technique for preparing reliable reference templates to improve the recognition rate score [14-20]. The developed technique produces templates called crosswords reference templates (CWRTs). It extracts the reference template from a set of examples rather than one example. This technique can be adapted to any DTW-based speech recognition systems to improve its performance. The speaker-dependent recognition rate, as tested on the English digits, is improved from 85.3%, using the traditional technique to 99%, using the developed technique [1]. Smart phones with touch screen have become very popular and it has changed the behaviour of using handsets as many applications and features are inbuilt into it. But this smart phones are not safe to handle at the time of driving as the user have to touch the small icons of the screen and while doing that he may lose the concentration which increase the accident risk. So to avoid this problem this paper proposes the no-touch mechanism with the help of proximity sensor and Text to Speech and speech to text systems. And just in case any accident happens it will drop the message to the emergency number. As there is no touch mechanism we can initiate the applications without touching the screen by just waving the hand on proximity sensor twice. Text to Speech (TTS) and Speech to Text (STT) Systems are used for notification of the incoming messages through the voice and if user wishes to call to the particular contact in his contact list he can just call the name of that person and the call will get connected [2].

Automatic recognition of spontaneous speech to text transformation is one of the most challenging tasks today. Whilst the recognition of read or formal speech (e.g. dictation) is possible for several languages by high accuracy rates allowing also commercial exploitation, the automatic recognition of spontaneous speech is a harder task, yielding at most about 50% accuracy. This is due to the characteristics of spontaneous speech, which shows an 'irregular' behaviour: insertions, abbreviations, truncations, uncompleted sentences, higher variability of both the acoustic and the linguistic features of speech. In this paper, automatic speech recognition of spontaneous and semi-spontaneous speech is evaluated and compared, using two speech databases to train the acoustic models and using statistical language models adapted for spontaneous speech and covering spontaneous and semi spontaneous job interview tasks. The involved speech databases are the HuComTech multi-modal database (training of acoustic and language models for spontaneous speech) and the Hungarian Reference Speech Database (training of non spontaneous acoustic models). Results show as expected that the recognition of spontaneous speech is less effective even in case of adapted language models and that considerable difference of speech recognition accuracy can be traced back both to the acoustic and language models, depending on the type and characteristics of the database they come from [3]. This paper presents techniques for speech-to-text and speech-to-speech automatic summarization based on speech unit extraction and concatenation. For the former case, a two-stage summarization method consisting of important sentence extraction and word-based sentence compaction is investigated. Sentence and word units which maximize the weighted sum of linguistic likelihood, amount of information, confidence measure, and grammatical likelihood of concatenated units are extracted from the speech recognition results and concatenated for producing summaries. For the latter case, sentences, words, and between-filler units are investigated as units to be extracted from original speech. These methods are applied to the summarization of unrestricted-domain spontaneous presentations and evaluated by objective and subjective measures. It was confirmed that proposed methods are effective in spontaneous speech summarization [4].

Speech understanding applications where a word based output of the uttered sentence is not needed can benefit from the use of alternative lexical units. Experimental results from these systems show that the use of non-word lexical units bring us a new degree of freedom in order to improve the system performance (better recognition rate and lower size can be obtained in comparison to word based models). However, if the aim of the system is a speech-to-text translation, a post-processing stage must be included in order to convert the non-word sequences into word sentences. In this paper a technique to perform this conversion as well as an experimental test carried out over a task oriented Spanish corpus are reported. As a conclusion, we see that the whole speech-to-text system neatly outperforms the word-constrained baseline system[5].

The access to communication technologies has become essential for the handicapped people. This study introduces the initial step of an automatic translation system able to translate visual speech used by deaf individuals to text, or auditory speech. A such a system would enable deaf users to communicate with each other and with normal-hearing people through telephone networks or through Internet by only using telephone devices equipped with simple cameras. In particular, this paper introduces automatic recognition and conversion to text of Cued Speech for French. Cued speech is a visual mode used for communication in the deaf society. Using hand shapes placed in different positions near the face as a complement to lip reading, all the sounds of a spoken language can be visually distinguished and perceived. Experimental results show high recognition rates for both isolated word and continuous phoneme recognition experiments in Cued Speech for French [6]. We summarize

recent progress in automatic speech-to-text transcription at SRI, ICSI, and the University of Washington. The work encompasses all components of speech modelling found in a state-of-the-art recognition system, from acoustic features, to acoustic modelling and adaptation, to language modelling. In the front end, we experimented with nonstandard features, including various measures of voicing, discriminative phone posterior features estimated by multilayer perceptrons, and a novel phone-level macro-averaging for cepstral normalization. Acoustic modelling was improved with combinations of front ends operating at multiple frame rates, as well as by modifications to the standard methods for discriminative Gaussian estimation. We show that acoustic adaptation can be improved by predicting the optimal regression class complexity for a given speaker. Language modelling innovations include the use of a syntax-motivated almost-parsing language model, as well as principled vocabulary-selection techniques. Finally, we address portability issues, such as the use of imperfect training transcripts, and language-specific adjustments required for recognition of Arabic and Mandarin [7]. The current work proposes a prototype model for speech recognition in Assamese language using Linear Predictive Coding (LPC) and Mel frequency cepstral coefficient (MFCC). The speech recognition is a part of a speech to text conversion system. The LPC and MFCC features are extracted by two different Recurrent Neural Networks (RNN), which are used to recognize the vocal extract of Assamese language- a major language in the North Eastern part of India. In this work, decision block is designed by a combined framework of RNN block to extract the features. Using this combined architecture our system is able to generate 10% gain in the recognition rate than the case when individual architectures are used [8]. We have developed a speech-to-text input method for web systems. The system is provided as a JavaScript library including an Ajax-like mechanism based on a Java applet, CGI programs, and dynamic HTML documents. It allows users to access voice-enabled web pages without requiring special browsers. Web developers can embed it on their web page by inserting only one line in the header field of an HTML document. This study also aims at observing natural spoken interactions in personal environments.

We have succeeded in collecting 4,003 inputs during a period of seven months via our public Japanese ASR server. In order to cover out-of-vocabulary words to cope with some proper nouns, a web page to register new words into the language model are developed. As a result, we could obtain an improvement of 0.8% in the recognition accuracy. With regard to the acoustical conditions, an SNR of 25.3 dB was observed [9]. The paper describes recent advances at LIMS in Mandarin Chinese speech-to-text transcription. A number of novel approaches were introduced in the different system components. The acoustic models are trained on over 1600 hours of audio data from a range of sources, and include pitch and MLP features. N-gram and neural network language models are trained on very large corpora, over 3 billion words of texts; and LM adaptation was explored at different adaptation levels: per show, per snippet, or per speaker cluster. Character-based consensus decoding was found to outperform word-based consensus decoding for Mandarin. The improved system reduces the relative character error rate (CER) by about 10% on previous GALE development and evaluation data sets, obtaining a CER of 9.2% on the P4 broadcast news and broadcast conversation evaluation data [10].

III. PROPOSED SYSTEM

Speech to text conversion is the process of converting spoken words into written texts. This process is also often called speech recognition. Although these terms are almost synonymous, Speech recognition is sometimes used to describe the wider process of extracting meaning from speech, i.e. speech understanding. The term voice recognition should be avoided as it is often associated to the process of identifying a person from their voice shown in fig 1.

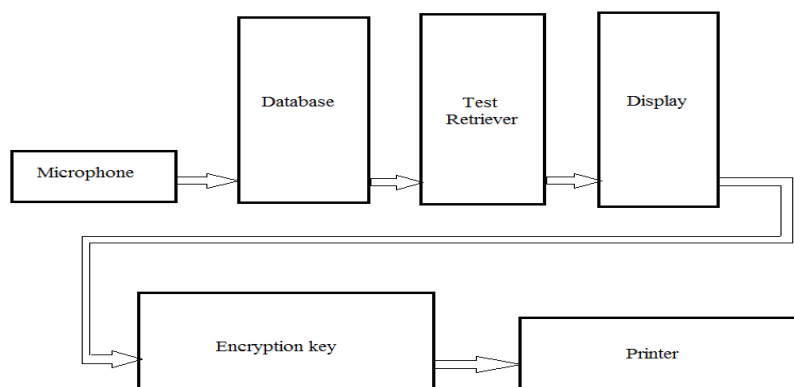


Figure 1 Speech to Text Conversion

IV. THE COMPLETE SPEECH-TO-TEXT CONVERSION PROCESS

1. You speak to the software via an audio feed Then software identifies the audio segments containing speech
2. The device you're speaking to creates a wave file of your words

3. The wave file is cleaned by removing background noise and normalizing volume. Then it recognizes the language being spoken
4. The resulting filtered wave form is then broken down into what are called phonemes. (Phonemes are the basic building block sounds of language and words. English has 44 of them, consisting of sound blocks such as “wh”, “th”, “ka” and “t”. Each phoneme is like a chain link and by analyzing them in sequence, starting from the first phoneme, the ASR software uses statistical probability analysis to deduce whole words and then from there, complete sentences)
5. Your ASR, now having “understood” your words, can respond to you in a meaningful way.
6. Finally it converts the speech segments to text

IV.COMPONENTS REQUIRED

- * Microphone
- * Laptop
- * Printer

Speech Recognition (SR) is the translation of spoken words into text. It is also known as "automatic speech recognition" (ASR), "computer speech recognition", or just "speech to text" (STT). The term voice recognition or speaker identification refers to identifying the speaker, rather than what they are saying. Recognizing the speaker can simplify the task of translating speech in systems that have been trained on a specific person's voice or it can be used to authenticate or verify the identity of a speaker as part of a security process. You can use your voice to control your computer. You can say commands that the computer will respond to, and you can dictate text to the computer shown in fig 2.

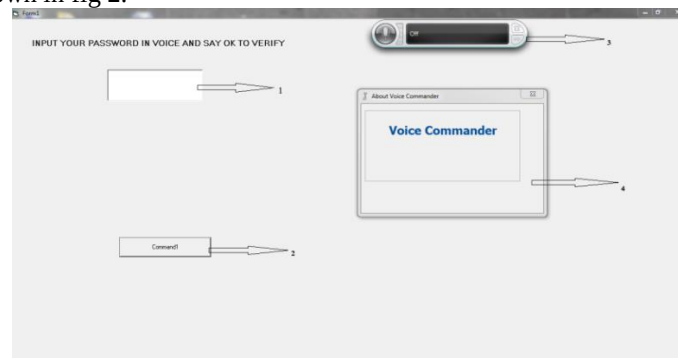


Figure 2 voice commander

- 1.Password
- 2.Command Button
- 3.Voice Recognition
- 4.Voice Training Window



Figure 3 Log in Speech To Text Conversion

- 1.Password strength(min 2 to max 8)

Before you get started using Windows Speech Recognition, you'll need to connect a microphone to your computer. Once you've got the microphone set up, you can train your computer to better understand you by creating a voice profile that your computer uses to recognize your voice and spoken commands for information about setting up your microphone shown in fig 3.

V. MICROPHONE SET UP FOR SPEECH RECOGNITION

Before you can start using Windows Speech Recognition, you need to set up a microphone. For specific information about installing your particular microphone, check the information that came with it or go to the manufacturer's website shown in fig 4 & 5.

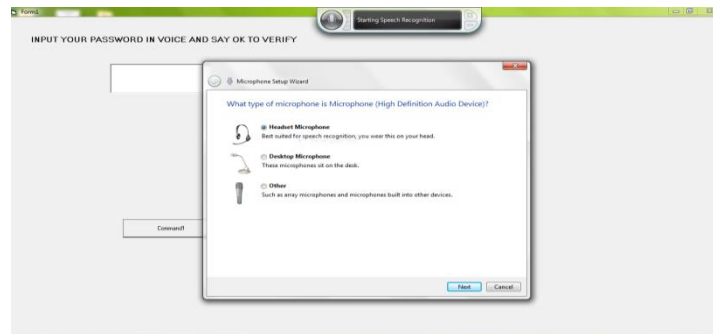


Figure 4 Microphone Setup Wizards



Figure 5 Microphone Setup

1. Make sure the microphone and speakers are properly connected to your computer.
2. Open Speech Recognition by clicking the Start button, clicking Control Panel, clicking Ease of Access, and then clicking Speech Recognition.
3. Click "Setup microphone".
4. Position the microphone about 3cm from your mouth, off to one side
5. Do not breathe directly into the microphone.
6. Make sure the mute button is not set to mute.

VI.VOICE TRAINING PROCESS

1. Green mark will mention to 50 % of voice training completed shown in fig 6 & 7.

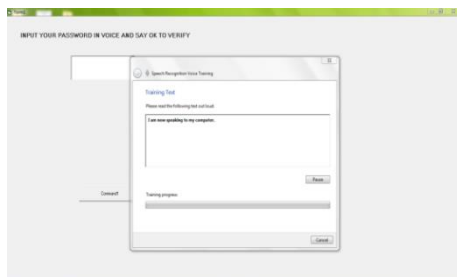


Figure 6 Voice Training Process

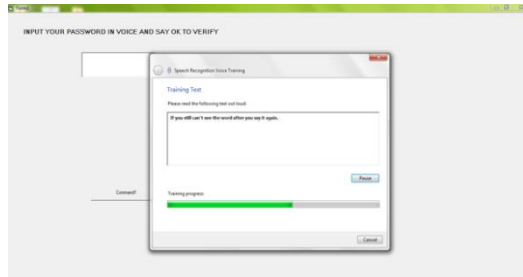


Figure 7 Voice Training Process 50% Completed

VI. SPEECH TO TEXT VOICE PRINT OPTION



Figure 8 Voice Print Window voice print option & speech analyzer

For language learning, speech recognition can be useful for learning a second language. It can teach proper pronunciation, in addition to helping a person develop fluency with their speaking skills. Students who are blind (Blindness and education) or have very low vision can benefit from using the technology to convey words and then hear the computer recite them, as well as use a computer by commanding with their voice, instead of

having to look at the screen and keyboard. Students who are physically disabled or suffer from Repetitive strain injury/other injuries to the upper extremities can be relieved from having to worry about handwriting, typing, or working with scribe on school assignments by using speech-to-text programs. They can also utilize speech recognition technology to freely enjoy searching the Internet or using a computer at home without having to physically operate a mouse and keyboard. Speech recognition can allow students with learning disabilities to become better writers. By saying the words aloud, they can increase the fluidity of their writing, and be alleviated of concerns regarding spelling, punctuation, and other mechanics of writing. Also, see Learning disability. Voice Recognition Software's use, in conjunction with a digital audio recorder, a personal computer and Microsoft Word has proven to be positive for restoring damaged short-term-memory capacity, in stroke and craniotomy individuals shown in fig 8.

VII. CONCLUSION

Speech recognition technology has made a remarkable progress in the past 5 - 10 years. Based on the progress, various application systems have been developed using dictation and spoken dialogue technology. One of the most important applications is information extraction and retrieval. Using the speech recognition technology, broadcast news can be automatically indexed, producing a wide range of capabilities for browsing news archives interactively. Since speech is the most natural and efficient communication method between humans, automatic speech recognition will continue to find applications, such as meeting/conference summarization, automatic closed captioning, and interpreting telephony. It is expected that speech recognizer will become the main input device of the "wearable" computers that are now actively investigated. In order to materialize these applications, we have to solve many problems. The most important issue is how to make the speech recognition systems robust against acoustic and linguistic variation in speech. In this context, a paradigm shift from speech recognition to understanding where underlying messages of the speaker, that is, meaning/context that the speaker intended to convey are extracted, instead of transcribing all the spoken words, will be indispensable.

REFERENCES

1. Dr. AntoBennet, M, Sankar Babu G, Natarajan S, "Reverse Room Techniques for Irreversible Data Hiding ", Journal of Chemical and Pharmaceutical Sciences 08(03): 469-475, September 2015.
2. Dr. AntoBennet, M , Sankaranarayanan S, Sankar Babu G, " Performance & Analysis of Effective Iris Recognition System Using Independent Component Analysis ", Journal of Chemical and Pharmaceutical Sciences 08(03): 571-576, August 2015.
3. Dr. AntoBennet, M, Suresh R, Mohamed Sulaiman S, "Performance & analysis of automated removal of head movement artifacts in EEG using brain computer interface", Journal of Chemical and Pharmaceutical Research 07(08): 291-299, August 2015.
4. AntoBennet, M & JacobRaglend, "Performance Analysis Of Filtering Schedule Using Deblocking Filter For The Reduction Of Block Artifacts From MPEQ Compressed Document Images", Journal of Computer Science, vol. 8, no. 9, pp. 1447-1454, 2012.
5. AntoBennet, M & JacobRaglend, "Performance Analysis of Block Artifact Reduction Scheme Using Pseudo Random Noise Mask Filtering", European Journal of Scientific Research, vol. 66 no.1, pp.120-129, 2011.
6. Anto Bennet, M, Mohan babu, G, Rajasekar, C & Prakash, P, "Performance and Analysis of Hybrid Algorithm for Blocking and Ringing Artifact Reduction", Journal of Computational and Theoretical nanoscience vol.12,no.1,pp.141-149,2015
7. AntoBennet, M & JacobRaglend, "Performance and Analysis of Compression Artifacts Reduction for MPEQ-4 Moving Pictures Using TV Regularization Method", Life Science Journal vol. 10, no. 2, pp. 102-110, 2013
8. AntoBennet, M & JacobRaglend, 'A Novel Method Of Reduction Of Blocking Artifact Using Machine Learning Metric approach', Journal of Applied Sciences Research, vol.8,no.5, pp. 2429-2438, 2012
9. Dr. Anto Bennet , M, Manimaraboopathy M, Srinath R, Maragathavalli P, Mekala S, "Frame Rate Up-Conversion Using Trilateral Filtering For Video Processing", International Journal of Engineering Research & Technology (IJERT), Vol. 3 Issue 9, pp-448-457, September- 2014
10. Dr. Anto Bennet, M, Srinath R, Maragathavalli, P Subhashini B, "Performance & Analysis of Blocking Artifacts Reduction Using Guassian Adaptive Spatial Low Pass Filter", Journal of Theoretical and Applied Information Technology, Volume68 No 01 pp-149-155, October 2014
11. Dr. AntoBennet, M, Vijayalakshmi G, Maragathavalli, P, Lokesh S, Dinesh Kumar T R, "Performance & Analysis of Blocking & Ringing Artifacts Reduction Using Generalized Least Square (GLS) Algorithm, Middle-East Journal of Scientific Research 21 (11): 2194-2202, October 2014. Abdulla, W.H., Dept. of Electr. & Electron. Eng "Cross-words reference template for DTW-based speech recognition systems" IEEE (Volume:4) Page(s): 1576 - 1579 Date Oct. 2003.
12. Admane, M.P., Jasutkar, R., Comput. Sci. Dept., " Speech to text and accelerometer based smart phone interaction system" IEEE Page(s):1 - 4, Date:27-28 Feb. 2014.
13. Szaszak, G. , Tundik, A.M., Vicsi, K. Dept. of Telecommun. & Media Inf., "Automatic speech to text transformation of spontaneous job interviews on the HuComTech database" IEEE Page(s):1 - 4, Date 7-9 July 2011.
14. Furui, S., Kikuchi, T., Shinnaka, Y., Hori, C., Dept. of Comput. Sci., "Speech-to-text and speech-to-speech summarization of spontaneous speech" (Volume: 12, Issue: 4) IEEE Page(s): 401 - 408, Date : July 2004
15. Penagarikano. M., Bordel, G., Dept. Electr. & Electron., "Speech-to-text translation by a non-word lexical unit based system" IEEE Page(s):111 - 114 (vol.1) Date: 22 Aug 1999.
16. Heracleous, P. ; Ishiguro, H. ; Hagita, N., Dept, Telecommunications., "Visual-speech to text conversion applicable to telephone communication for deaf individuals" IEEE, Page(s):130 - 133 Date: 8-11 May 2011.
17. Stolcke, A.; Chen, B.; Franco, H. ; Gadde, V.R.R., " Recent innovations in speech-to-text transcription at SRI-ICSI-UW" (Volume:14, Issue: 5) IEEE, Page(s):1729 - 1744, Date : Sept. 2006
18. Dutta, K ; Sarma, K.K., Dept. of ECE., " Multiple feature extraction for RNN-based Assamese speech recognition for speech to text conversion application" IEEE, Page(s):600 - 603, Date:28-29 Dec. 2012.
19. Nisimra, R. ; Miyake, J.; Kawahara, H.; Irino, T., Dept. of Syst. Eng., " Speech-to-text input method for web system using JavaScript" IEEE Page(s):209 - 212, Date :15-19 Dec. 2008
20. Lamel, L. Gauvain, J. Le, V.; Oparin, I.; Dept, Spoken Language Process Group, LIMSI-CNRS, "Improved models for Mandarin speech-to-text transcription" IEEE Page(s): 4660 - 4663 Date:22-27 May 2011