



Online Interactive Language Tools for Bangla and English Languages



Shiblee Imtiaz Hasan

ID: 07110049 / 09241001

shiblee@ieee.org

Supervisor: Prof. Mumit Khan

Co-supervisor: Matin Saad Abdullah



Background

- Importance of Natural Language Processing and related applications
 - Automation (e.g. dialogue system, voice instructions)
 - Text summarization in large volume
 - Understanding different languages and grammars
 - Help the illiterate to learn language
 - Assist the physically impaired





Objective of the Thesis

- Build a complete rich internet application which is an aggregate software of various **open-source** language utilities available today.
- Not re-inventing but rather **optimize** for functionality and efficiency.
- A complete application that supports Bangla and English languages with full **Unicode** support.
- The application which will be **user friendly** for most physically impaired individuals.



Project Overview

- The application uses Adobe Flash as its front-end user interface.
- Data manipulation in the back-end is done by the collaboration between Adobe Flash and PHP Script in the server along with many other components.
- The application is designed to support multiple languages and utilities which can be added later on.
- Facebook application version of the utilities already released.



Quick Statistics



facebook

- **5** open-source projects are involved
 - *Bangla WordNet, BOCR, GOCR/JOOCR, Festival, Princeton WordNet*
- **2,500+** lines of code written in **7** languages
 - *ActionScript, FBML, JavaScript, PHP, Shell, SQL, (X)HTML/CSS/XML*
- **4** servers used to process, store and serve data



Utilities Provided

- **Text To Speech (TTS)**

- Text to speech or Speech synthesis is the artificial production of human speech. Currently all the utilities which consists of text generation has TTS implemented.

- **Image To Text/Speech (OCR + TTS)**

- The Image To Text/Speech conversion or Optical character recognition (OCR) is the translation of images of handwritten, typewritten or printed text (usually captured by a scanner or camera) into computer-editable text.



Utilities Provided (Contd.)

- **Speech Recognition**

- Speech recognition (also known as automatic speech recognition or computer speech recognition) converts spoken words to text. Due to availability of the Flash Media Server, this utility was not fully implemented.

- **Bangla and English WordNet**

- WordNet is a lexical database for a language. It groups words into sets of synonyms called synsets, provides short, general definitions, and records the various semantic relations between these synonym sets. Right now a dummy database has been setup for demonstration.



Facebook Application

- Allows users to use the application while using Facebook
- Users are able to invite friends to application
- Can be integrated to add Notes from scanned text or update Status Message in different languages





Technologies Used

- **Interface**

- A clean and highly interactive user interface has been designed and programmed using the Adobe Flash technology. The Flash interface can perform almost all necessary client side scripting as well as interact with PHP to get server data.

- **Backend**

- Behind the user friendly Flash, PHP provided all the necessary data processing across domains and file formats. It also accessed the data from MySQL database for the Flash.



Technologies Used (Contd.)

- **Text To Speech**

- "Festival" open-source speech synthesizer, which is a software capable of making artificial speech in place of a real human. It's engine is generating English speech developed by University of Edinburgh and Bangla speech developed at CRBLP.

- **Bangla and English WordNet**

- There has been a Bangla WordNet developed by CRBLP which contains a database of dictionary and thesaurus of Bangla. Princeton University's database has been used for the English WordNet. Data are saved in a MySQL database accessed by PHP.



Technologies Used (Contd.)

- **Image To Text/Speech**

- CRBLP's Bangla OCR or BOCR is an Optical Character Recognition tool for Bangla glyph. The open-source GOCR by Jörg Schulenburg is used to recognize English characters. The OCR can take image from local machine, upload it in server and process it to output text.

- **Speech Recognition**

- Currently the Flash engine can only detect microphone input and volume level. The presence of a Flash Media Server will enable it to record the sound and process it to detect text.



Limitations and Solutions

- **Bangla Glyph Display in Adobe Flash**

- Despite of being supported by Unicode 4.1 and above, Bangla glyphs are not displayed properly by Adobe Flash in different platforms and browsers. It was required to add additional functions to enable all users to view Bangla such as dynamic text and font embedding.

- **Cross-Domain Communication in Adobe Flash**

- Due to security issues, Adobe Flash doesn't allow data from different domains to be displayed in its Shockwave Flash files. To solve these limitations, different PHP connectors are being used.



Limitations and Solutions (Contd.)

- **Using Unicode Glyphs in Query and Other Languages**
 - Although most languages now support almost full Unicode support for string types, it is always required to test input/output to components to make sure the variables are storing the right string.
- **File Writing/Uploading in Server**
 - While multiple users are uploading, there is a possibility that the filenames are the same which might trigger an overwrite. To overcome this, the uploaded files were given random non-duplicate string filenames.



Limitations and Solutions (Contd.)

- **Multiple Input of Similar Data**

- The server executions are quite expensive in terms of time and memory. This program implements memoization to keep history of processed string and its output audio file URL.

- **Syntactic ambiguity**

- The grammar for natural languages is ambiguous, i.e. there are often multiple possible parse trees for a given sentence. Choosing the most appropriate one usually requires semantic and contextual information. Specific problem components of syntactic ambiguity include sentence boundary disambiguation.



Limitations and Solutions (Contd.)

- **Imperfect or irregular input**

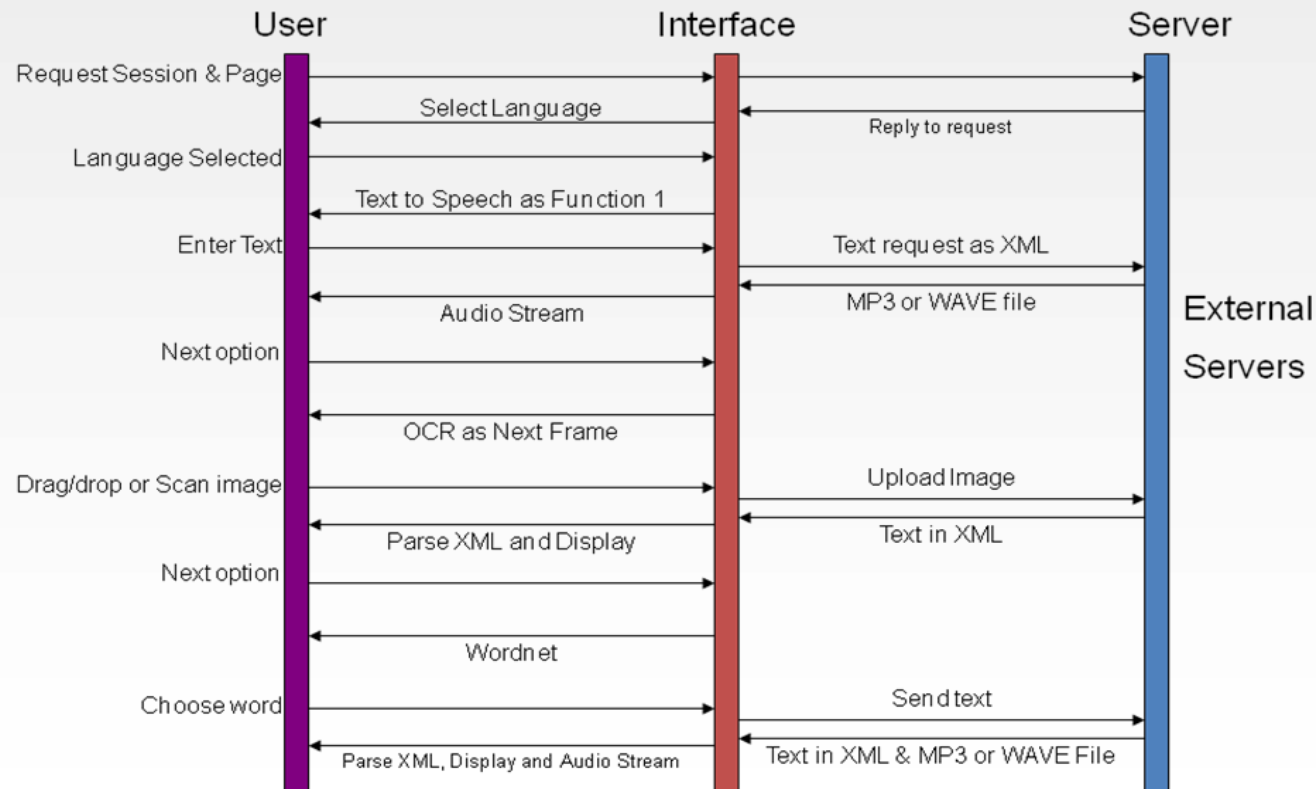
- Includes Foreign glyphs in text, accents and vocal impediments in speech; typing or grammatical errors, OCR errors in texts. Mixed input of different languages will be cut off and put into different voices for TTS.

- **Speech acts and plans**

- A sentence can often be considered an action by the speaker. The sentence structure alone may not contain enough information to define this action.



Server and Interface Communications





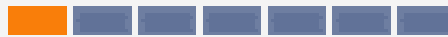
Future Developments

- The following features are planned to be implemented for the next Facebook platform application update:
 - **Streaming Sound** – to allow faster audio playback
 - **Adobe Media Server** – to implement the Online Speech Recognition
 - **Full WordNet Database** – to move from demo data to beta release
 - **Avoid Multiple Servers** – to reduce data processing time
 - **Add New Features** – dictionaries, translations and transliterations
 - **Add New Languages**



Live Demonstration

loading...



labs.com.bd/voice
&
apps.facebook.com/language_tools



References

- **CRBLP.** (2009). Centre for Research on Bangla Language Processing. [Accessed 16 December 2009]. Available from World Wide Web: <<http://crblp.bracu.ac.bd>>.
- **Wikipedia.** (2009). WordNet. [Accessed 16 December 2009]. Available from World Wide Web: <<http://en.wikipedia.org/wiki/Wordnet>>.
- **Wikipedia.** (2009). Natural language processing. [Accessed 16 December 2009]. Available from World Wide Web: <http://en.wikipedia.org/wiki/Natural_language_processing>.
- **Wikipedia.** (2009). Adobe Flash. [Accessed 16 December 2009]. Available from World Wide Web: <http://en.wikipedia.org/wiki/Adobe_Flash>.
- **Wikipedia.** (2009). PHP. [Accessed 16 December 2009]. Available from World Wide Web: <<http://en.wikipedia.org/wiki/Php>>.