Proof of Concept

**US English**

Audio In

**Automatic Speech Recognition**

**Text to Speech Synthesis**

Audio Out

**Cebuano**

Audio In

**Machine Translation**

Audio Out

**Text to Speech Synthesis**

**Automatic Speech Recognition**

*Fully on board speech-to-speech translation system to facilitate cross-language communication in disaster relief scenarios*
Sponsors and Stakeholders

PTC Honolulu
PTC Australia
Vonwiller Foundation
Open Systems Education Trust
Kirby Foundation

Pro bono contributors
Dr Julie Vonwiller, Project Manager
Dr James Nealand, Project Advisor

RedR Australia
Habitat for Humanity Australia
Project Steering Committee
Project Tasks

Task 1  Identify specific domains, and develop scenarios and vocabulary

Task 2  Collect and annotate databases of Cebuano in-domain audio and Cebuano & English text; build matching Cebuano & English lexicons – APPEN

Tasks 3 & 4

Develop CEB speech recognition module – ASR
Develop US ENG speech recognition module

Task 5 & 6

Develop CEB <> ENG translation modules

Task 7  Develop CEB synthesis module

Task 8  Integrate US ENG synthesis module

Task 9  Develop basic user interface and integrate with modules

Task 10  Conduct trials and final testing
Domains

• Different operational stages can be classified in language processing as *domains*
  – Examples include Medical & Health, Water & Sanitation, Security

• For this project, we selected the **Needs Assessment** domain.
  – The first activity undertaken by relief workers to gain a rapid overview of the situation, in order to efficiently manage the response to the disaster

• Speech recognition and Machine Translation can be optimised by training the computer algorithms on the scenarios, dialogue and vocabulary of actual disaster relief operations
  – Tune language models the specific domain
  – Ensure that all relevant vocabulary is covered by the system
  – Ensure that the expected topics and vocabulary are well covered by parallel corpora used to build machine translation

• Actual scenarios and training data for the Needs Assessment domain was captured by participation in RedR’s training programs for relief workers
Data Collection

- A Cebuano language training corpus was collected by Appen over four months in the Philippines in the Visayan region
- 100 native speakers, balanced by age and gender
- Speakers were recorded in role playing scenarios as disaster victims and first responders
- The recording environment and background noise matched likely field conditions
- Processing the data post-recording was carried out by Appen in Davao and Sydney
Automatic Speech Recognition (ASR)

• Development of ASR components was performed by Assistant Professor Khe Chai Sim of the National University of Singapore

• The open source package Pocketsphinx from Carnegie Mellon University was selected as the ASR engine:
  – Full tool-chain available for developing new language components
  – Supports fast model architectures suitable for embedded systems – runs completely on board
  – Existing Android build available

• Speech recognition relies on two main components for a language: the acoustic model and language model
  – For US English a standard acoustic model released by CMU was used, and a custom, domain specific language model built.
  – For Cebuano, both a custom acoustic model and language model were built.

• All processing is performed on board with no reliance on network connectivity. Instances of the decoder in both languages are kept loaded so that there is no latency in switching between languages.
Machine Translation (MT)

- Text-to-Text translation has been developed for both directions:
  - Cebuano → English
  - English → Cebuano
- The Open Source MOSES software package was selected:
  - Typically server based software, to the best of our knowledge we are the first to have ported Moses to Android
- Moses model development was performed at Carnegie Mellon University by Andrew Wilkinson
- The training data consisted of parallel text corpora (English and Cebuano), derived from the domain-specific recordings and general Cebuano text
- The MT package is by far the largest component in the system. As such, we operated the MT software as a separate process and communicate with the application using XML-RPC.
Speech Synthesis (TTS)

- The Text-To-Speech system uses the CMU Festival Lite (Flite) system
  - The English TTS is an existing US English voice
  - The Cebuano TTS was created by Prof Alan Black and Andrew Wilkinson of CMU, using training data provided by Appen
  - New voices are added by copying the voice file to micro-SD
- The training data included high-quality recordings of 2000 phonetically rich sentences by a female Cebuano voice talent, supplemented by phonemic lexicons of general and domain-specific Cebuano words
- An existing build of Festival Lite for Android existed, however we modified the Java Native Interface (JNI) to add a number of features:
  - Faster switching between voices – essentially two voices are kept loaded simultaneously
  - Voices can be added by copying files to micro-SD
Integration

• The application is a regular Android application (shipped as a .apk file).
• ASR and TTS packages are built as shared libraries with a Java Native Interface (JNI) layer for communication with the application.
  – Existing builds of Pocketsphinx and Festival Lite were available for Android so these were modified as needed
• Moses runs as a separate process and communicates with the application using XML Remote Procedure Calls (XML-RPC).
  – The components of Moses needed for our application were ported to Android.
  – To our knowledge we are the first to build Moses for Android
• ASR, TTS, and MT models are shipped as a set of files on a micro-SD card. Logs are written to micro-SD
  – Convenient during development and trials
  – Ultimately we will add the ability to fetch components from a remote server and also to store components on internal flash
• The User Interface presents unique challenges
  – Two users interacting with the system in different languages
  – While the aid worker can be trained to use the application, the victim will be seeing an unfamiliar system for the first time.
Modular Performance

- **ASR**
  - On a portion of the Appen corpus held out from training, ASR word accuracy of 80.31% was achieved for US English and 85.39% for Cebuano
  - English speakers were not native US English speakers
  - Preliminary experiments show that we can achieve a 2% absolute improvement in word accuracy using Maximum Likelihood Linear Regression adaptation. This would require users to engage in a short enrolment session

- **MT**
  - English -> Cebuano BLEU 38.7%
  - Cebuano -> English BLEU 47.9%

- **TTS**
  - Mean Opinion Score (MOS) of 4.48 (out of 5), which is well within the target performance.
  - The Word Error Rate was assessed using semantically unpredictable sentences (SUS) played to a native Cebuano speaker, who was required to write down what he heard.
  - The SUS word error score was 4.5%, also within the target error rate of <30%
The Future

• End-End performance testing
  – Using pre-recorded audio as well as live field tests

• Software engineering
  – Improvement of user interface
  – Complete port of Moses to enable use of models in binary form
  – Add support for speaker adaptation

• Languages and Domains
  – Text and Audio data collection in an additional domains and languages
  – Prove that the recipe developed can be used to rapidly build out new languages and domains

• Partnerships and Deployment