Mean Opinion Score Performance in Classifying Voice-enabled Emergency Communication System

International Conference on Computer and Information Sciences
2nd Wold Engineering, Science, and Technology Congress

2012 June 13

Convention Center, Kuala Lumpur, Malaysia

Nuwan Waidyanatha
Research Fellow, LIRNEasia
nuwan@lirneasia.net
Kunming, China

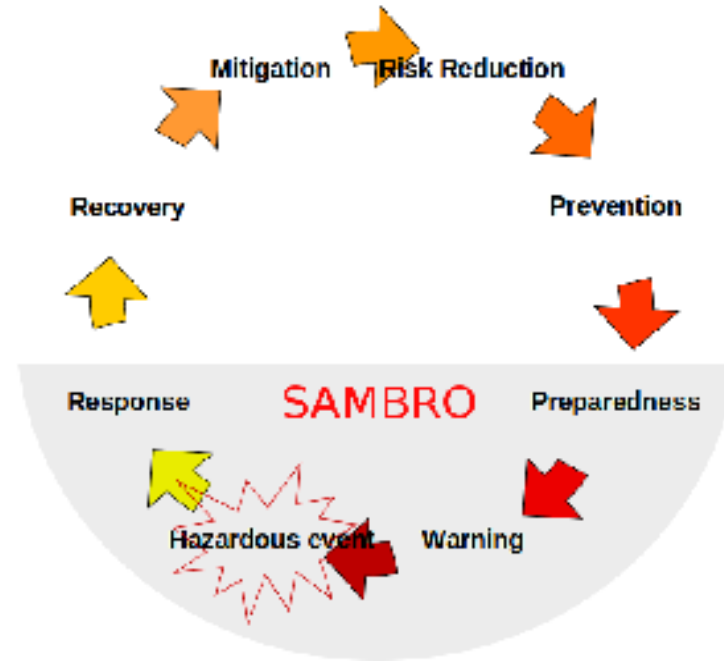
The research was carried out with a grant from The Kubatana Trust of Zimbabwe

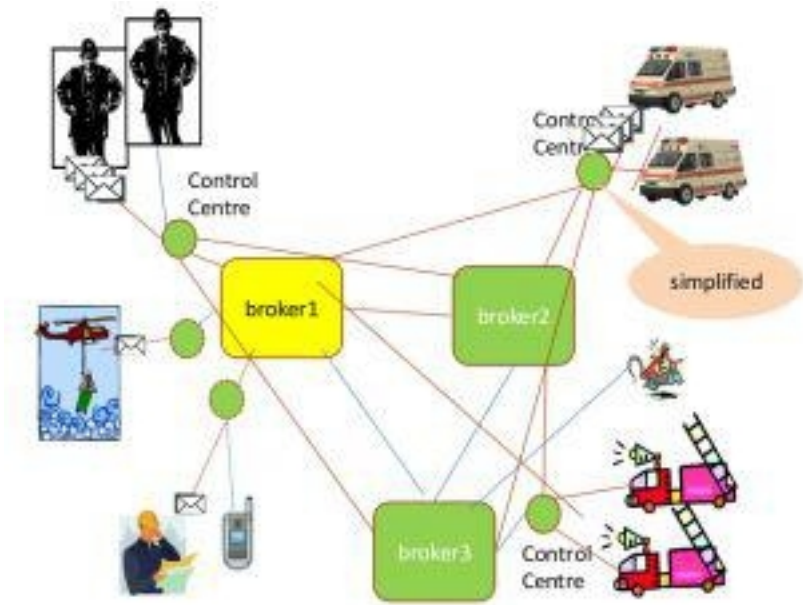
The diffusion work was carried out with a grant from the Humanitarian Innovation Fund of UK

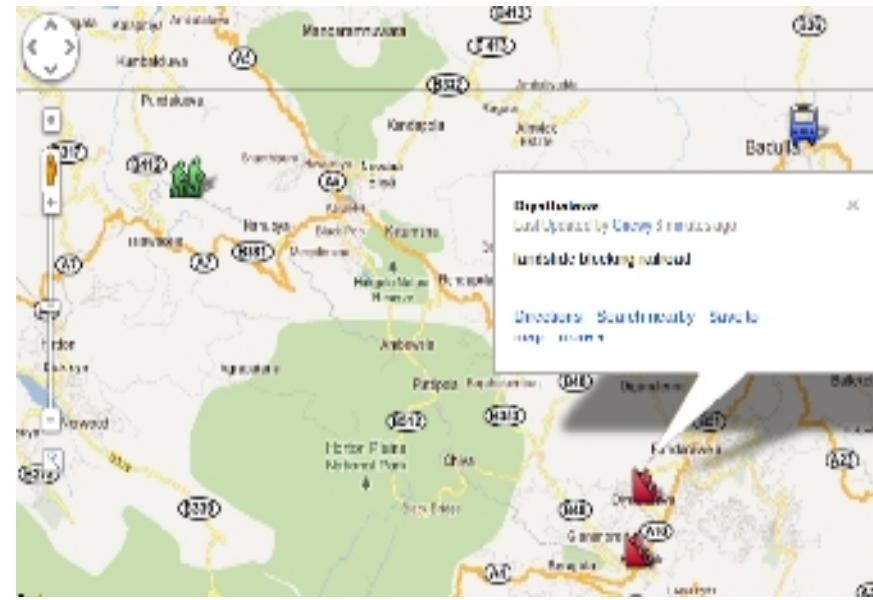
LIRNEasia
www.lirneasia.net

LIRNEasia's Emergency Communication Research Focus

- LIRNEasia is in EWS space of DM; i.e. ***HazInfo***:
  - *Projects* :: Webhamuwa, NEWS:SL, Dam-safety, LM-HWS, CB, Biosurv, FF4EXL
- Bleeds in to "preparedness" and "response"
  - "plans w/o drills and drills w/o plans are useless" → action research
- Advocate interoperability
  - common procedures (registry of alerting authorities)
  - data standards (e.g EDXL)

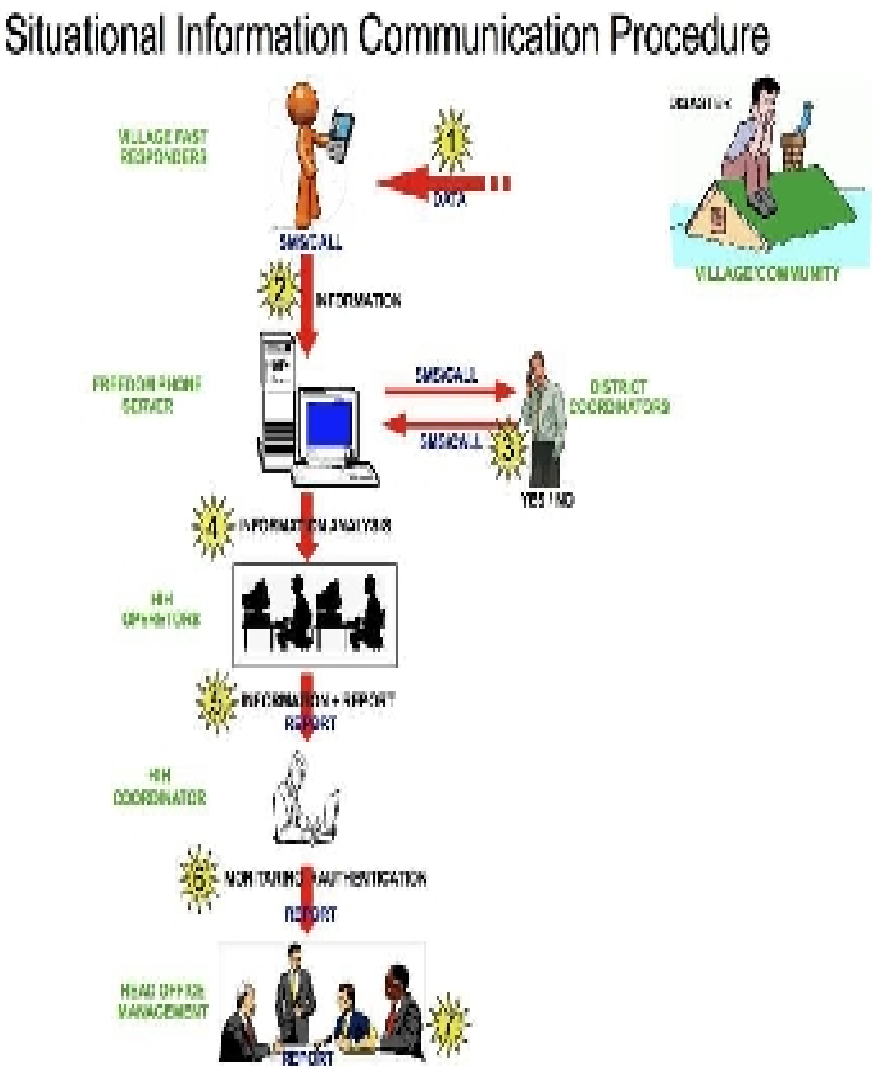






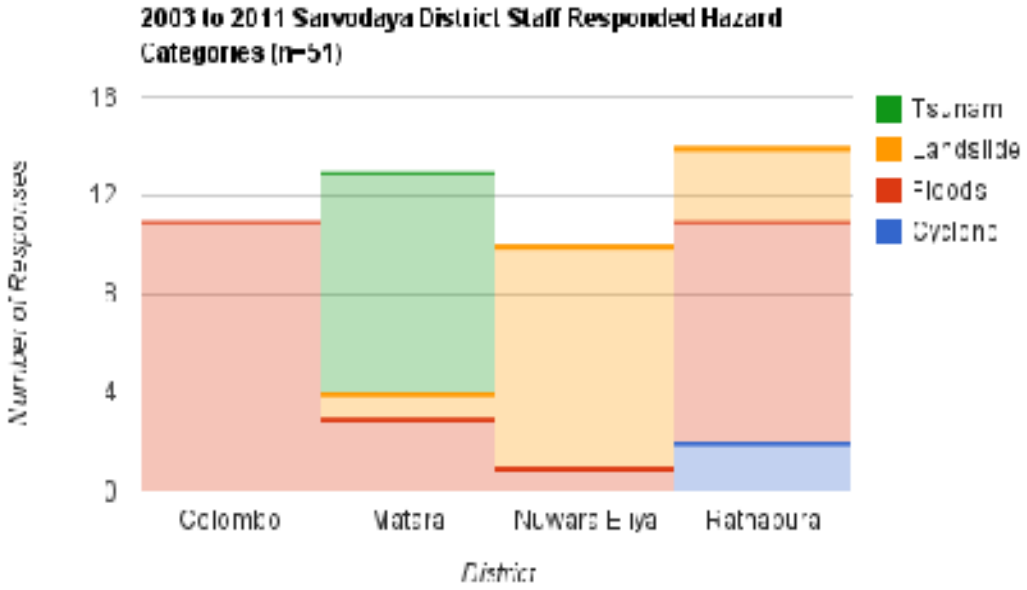
“Sarvodaya Samana Thetha” Community-based Disaster Management Center Emergency Information Needs

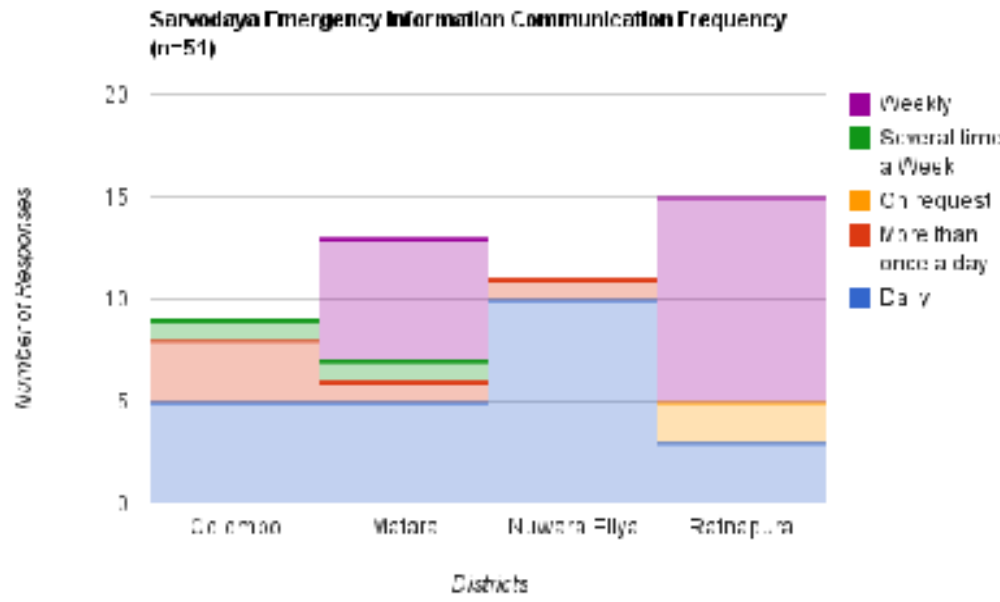
Voice for Alerting and Response >> Full Story

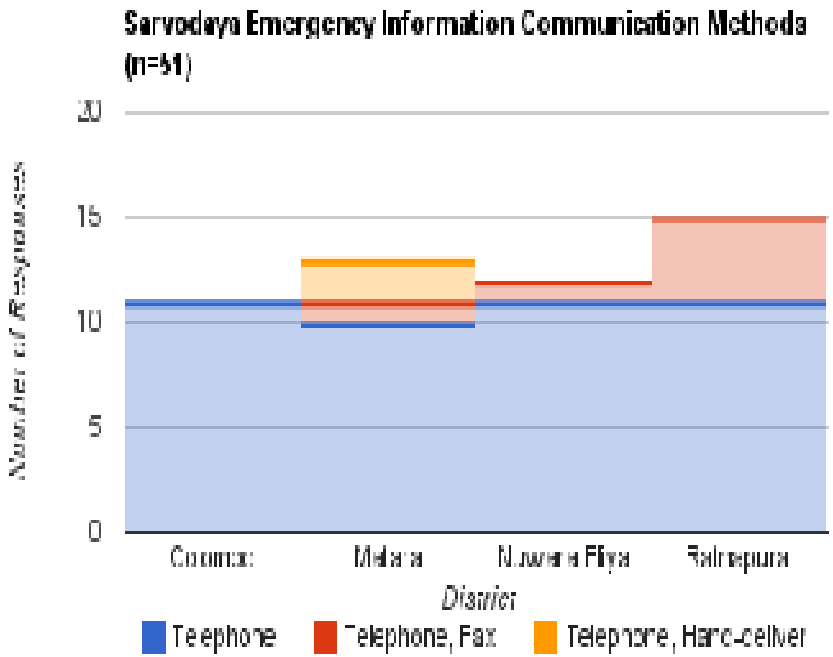


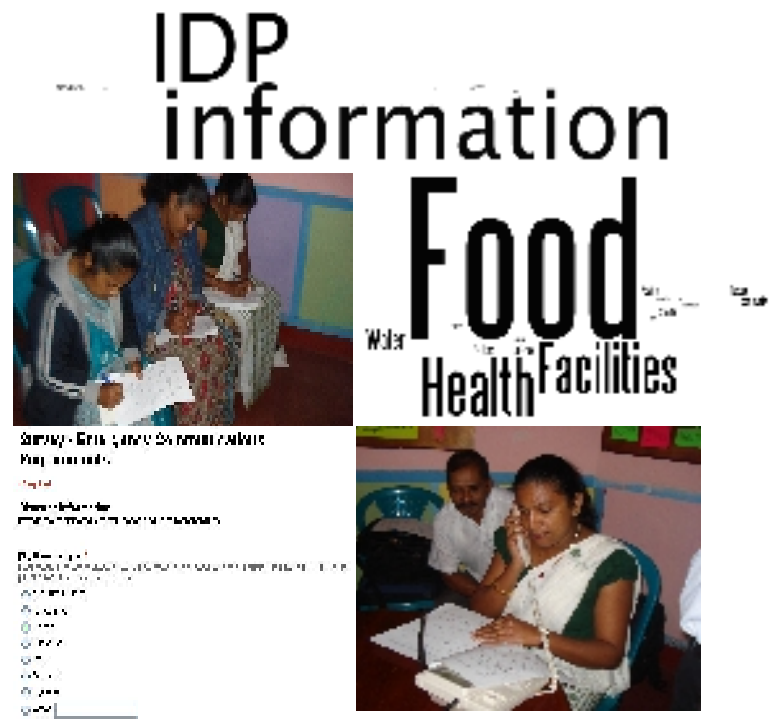
- Community members (victims or associates)
  - call the District or Head Office
  - report of incidents.
- Collect ground truth
  - 2011 Floods, SCDMC dispatched youth with cameras, laptops, and dongles
- Blog situation on www.sarvodaya.org
- Get word to media and donors
- Secure response resources

Sarvodaya use telephones to communicate, daily/weekly, IDP Info

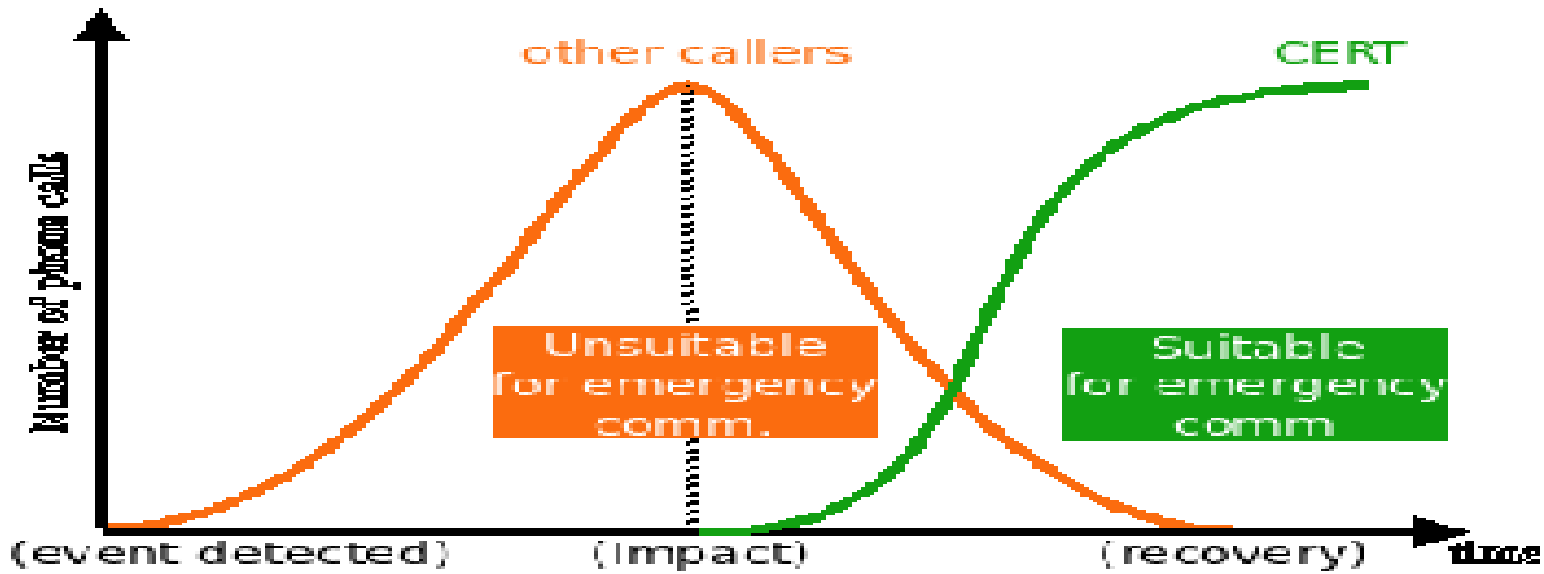






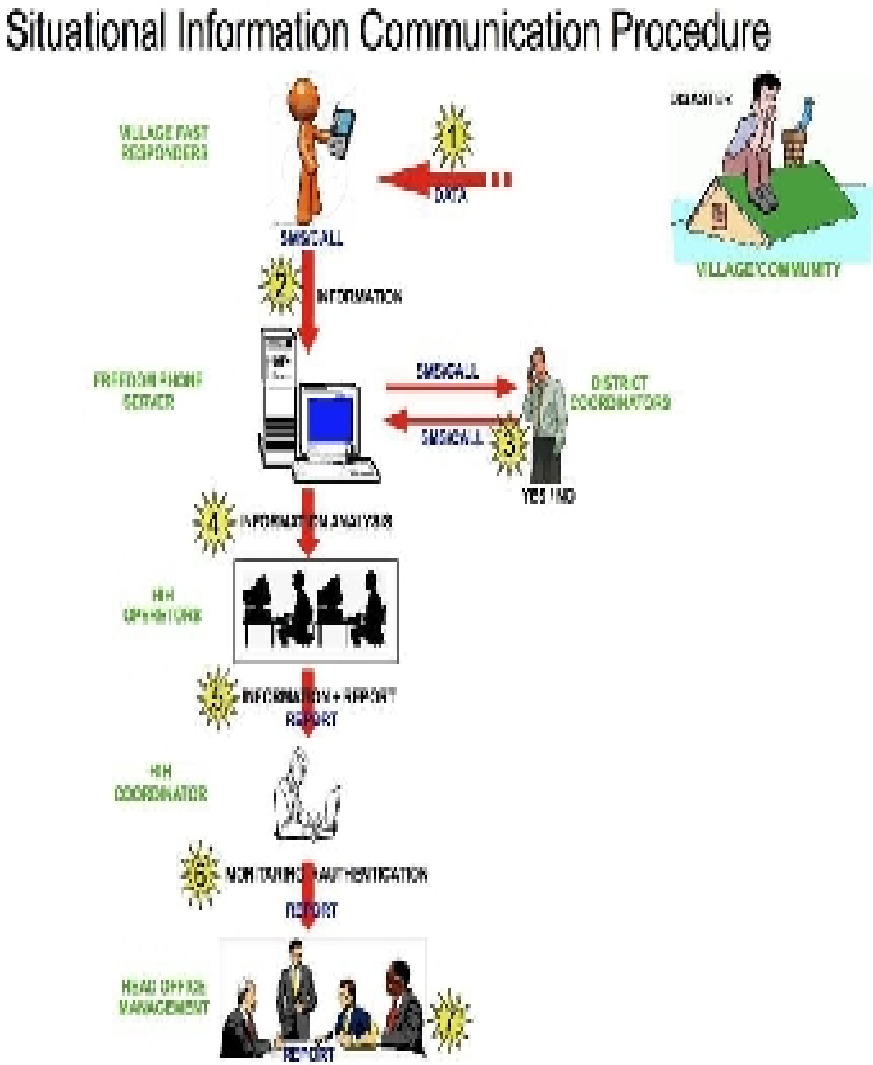


Positioning Voice for Emergency Communication



- ***Dropped calls*** are high during hazard events
  - Voice best after the *N*th hour from disaster impact
  - Recommended for ***disaster response phase*** (rescue and relief)
  - Can be used for other none critical disaster management activities
- Canada use IVR for rapid on-set tsunami warnings (< 60 min window)
  - But too expensive for developing countries
  - Definitely not for community-based organizations
- ***IVR*** removes language and computer literacy barriers
  - ***less burden on training regime***

Two key consecutive functions for community-based emergency response



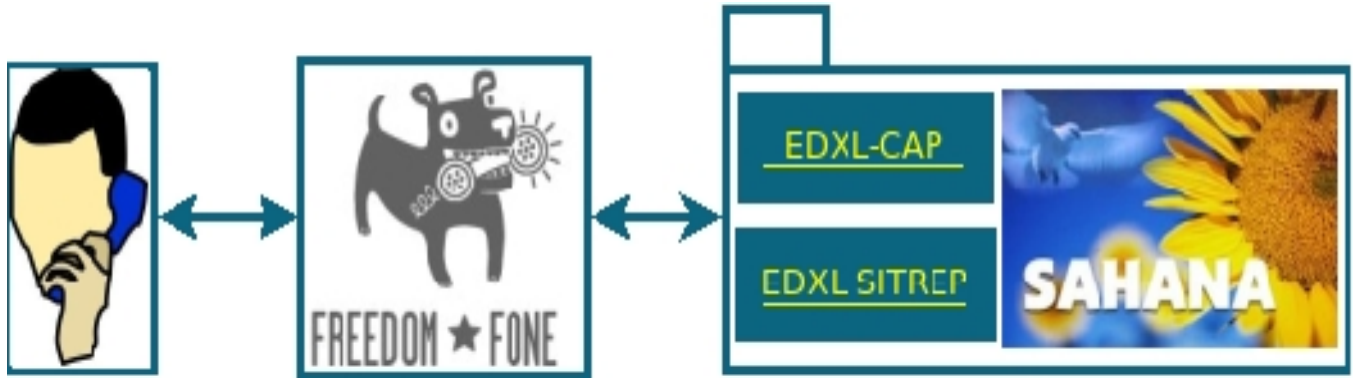
ALERTING

Following a hazard event activate CERT members and HIH operators to identify the incidents, then report the field observation

REPORTING

Receive field observation reports, process them at the Hazard Information Hub to create Situational Reports

Decoupled Software Systems

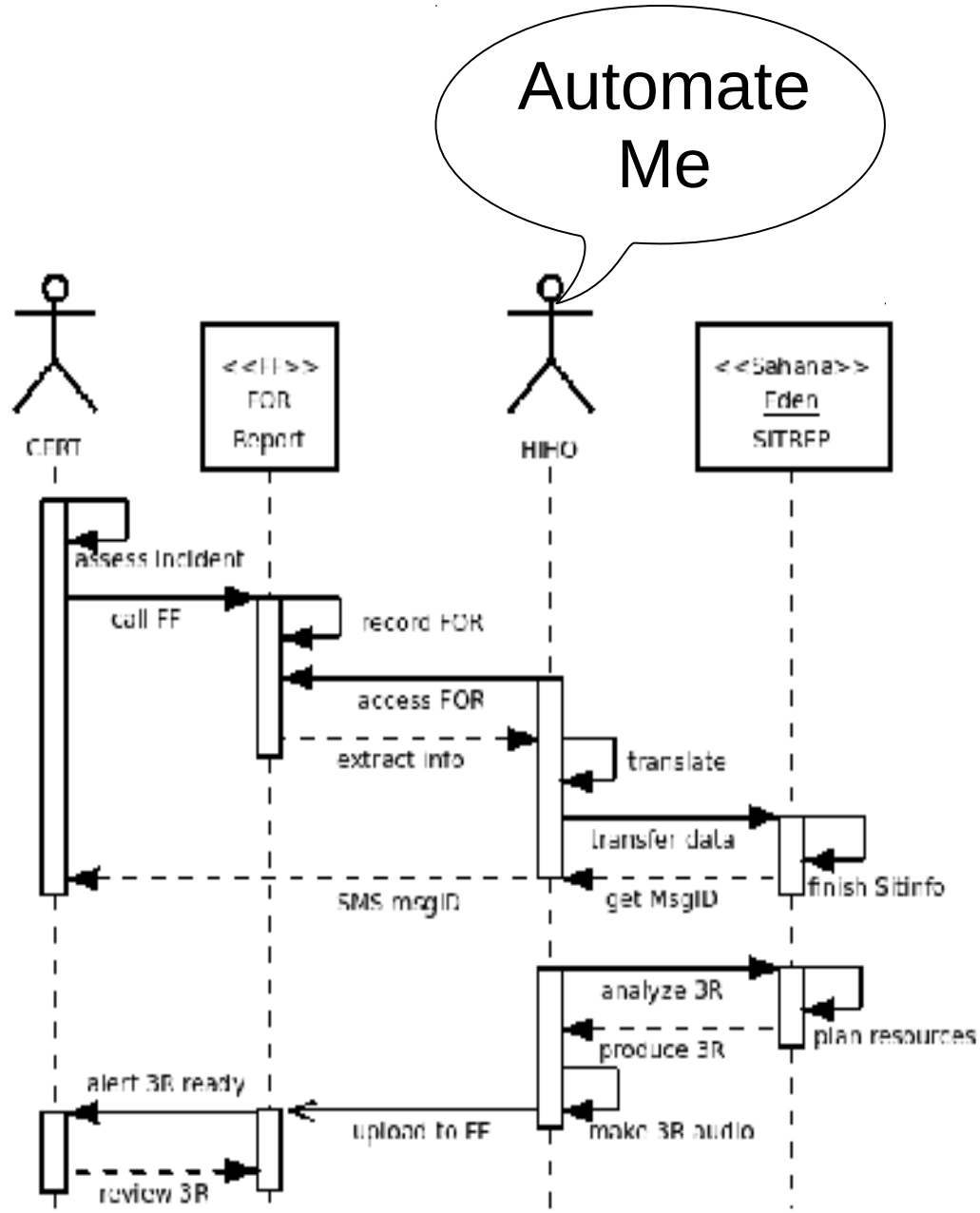


- **Sahana Disaster Management System** (www.sahanafoundation.org)
  - _FOSS humanitarian ICT tool but in several flavors: Agasti and Eden
  - _supported by a global community of software engineers and emergency managers
  - _categorical data important for decision support
  - _Believes in interoperability
- **Freedom Fone** (www.freedomfone.org)
  - _FOSS IVR tool
  - _Don't need Internet
  - _Infrastructure: GSM modem + cheap PC
  - _User Interface: standard telephones (mob + fix)
- **Audacity**
  - _FOSS tool for making audio files

REPORTING

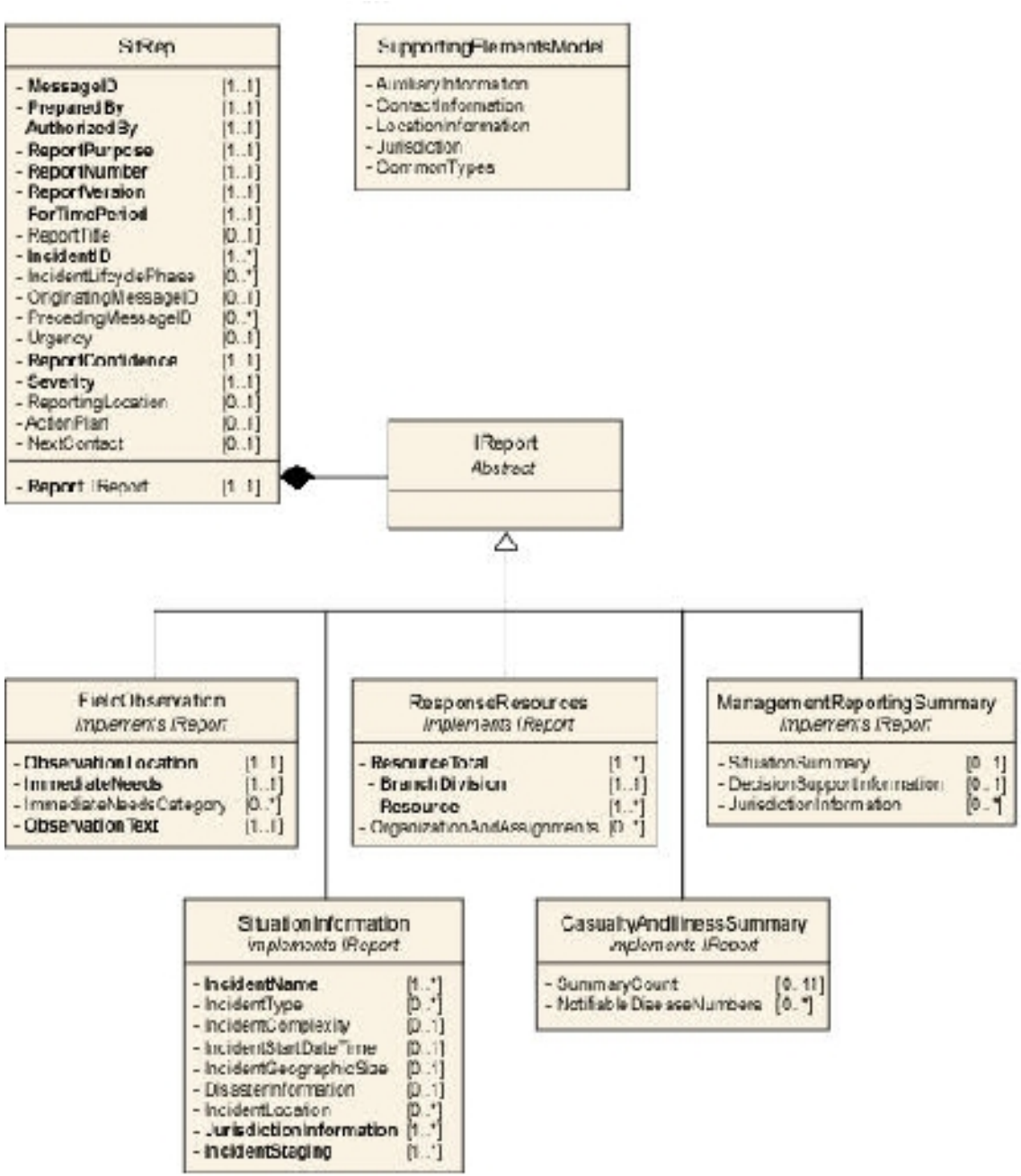
Upstream communication sequence

1) CERT members record, local lingo Field-Observation voice reports in Freedom Fone IVR
2) Incident coordinators (HazInfoHub Operators) translate/transform those voice messages to EDXL-SITREP categorical info
3) Then enter that plus other info in Sahana SitRep module
4) Incident managers derive Response Resource reports for actions



REPORTING

Emergency Data Exchange Language (EDXL) Situational-Reporting (SITREP)



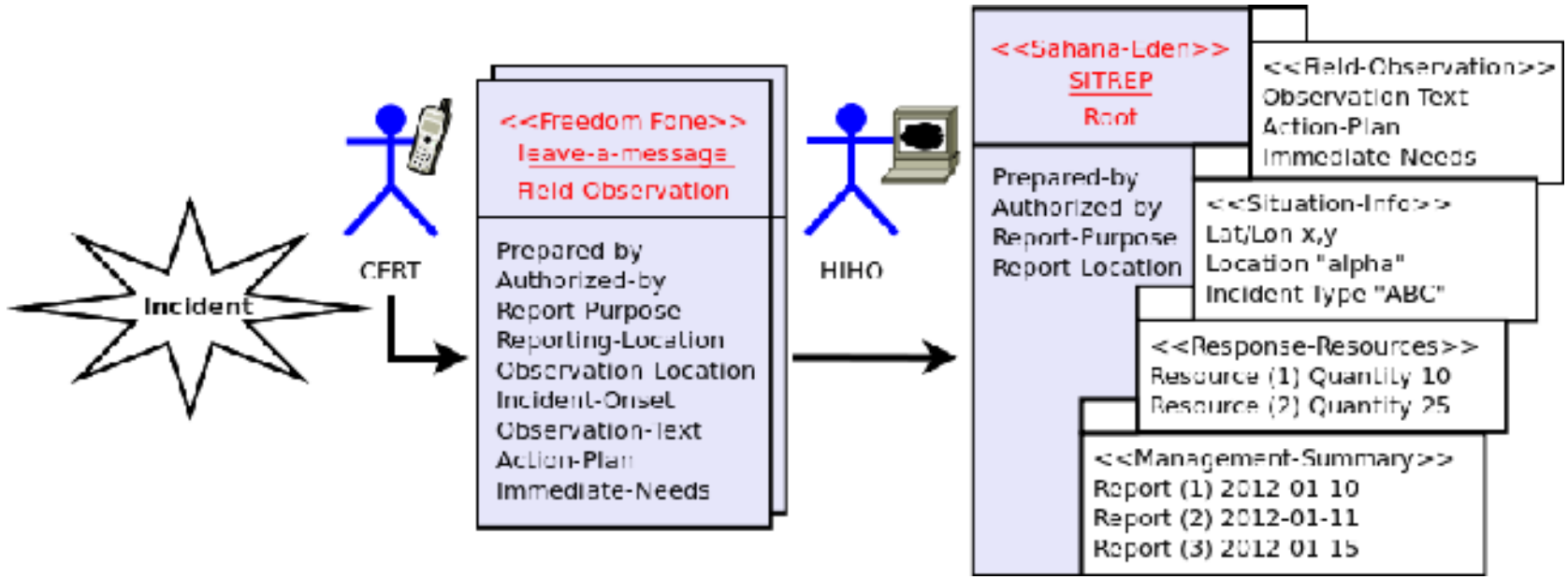
1) ***SitRep***: root element with qualifying elements

2) ***iReport***: the type of report

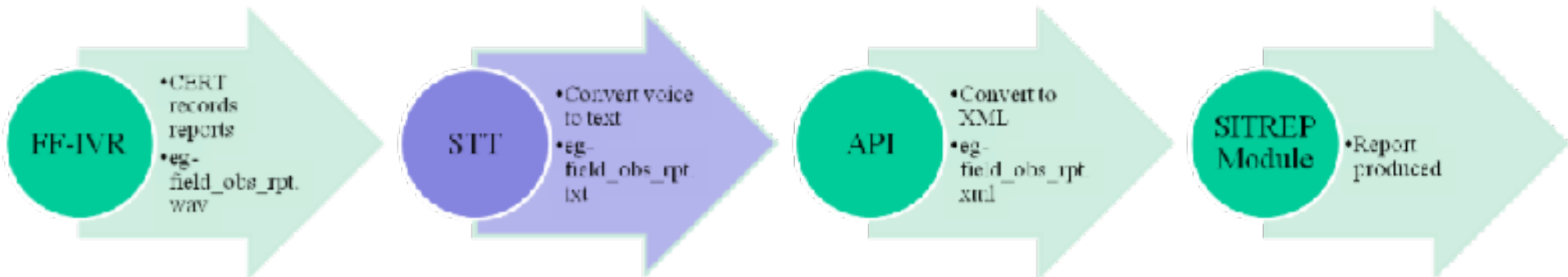
- ***Field Observation***: report sent by CERT members identifying incidents
- ***Situation Information:*** additional information for comprehensive information
- ***Response Resources:*** derived resources to deploy
- ***Casualty/Illness Summary:*** injury and health related information
- ***Management Summary:*** periodic summary of overall picture

REPORTING

Software components

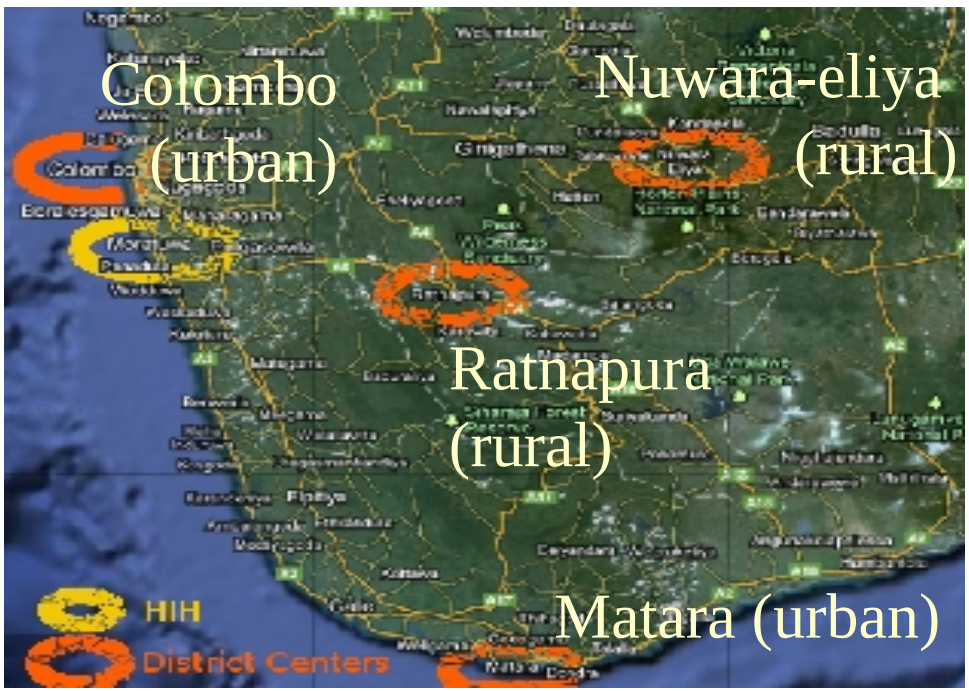


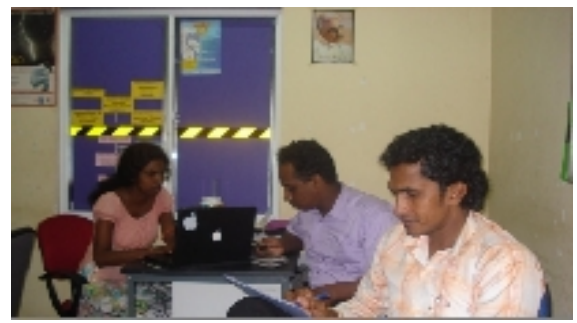
Automation required to reduce the human work load



Research Design

- Principal: Lanka Jathika **Sarvodaya** Shramadana Sangamaya
  - Sri Lanka's largest community development NGO
  - Also provide humanitarian services
- Hazard Information Hub @ Community Disaster Management Center, Moratuwa, HIH Manager, 3 HIH Operators
- Four Districts: Colombo, Matara, Nuwara-eliya, Ratnapura, ~ 10 - 15 CERT members from each district: Divisional/District Coordinators, Staff

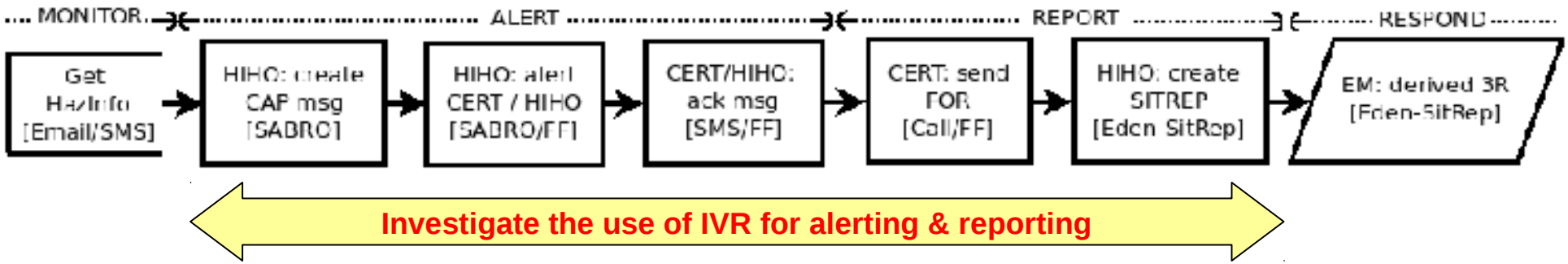




Formative Evaluation Method

Controlled Exercises

- Discussed operating procedures (goal, intention, action)
- Executed those procedures (execution, perceiving, interpreting SoW)
- Evaluated the outcomes (Performance, Usability)



Complexity:

- Interaction techniques (HCI)
- Reliability - mean time to completion & **voice quality (ITU-T)**

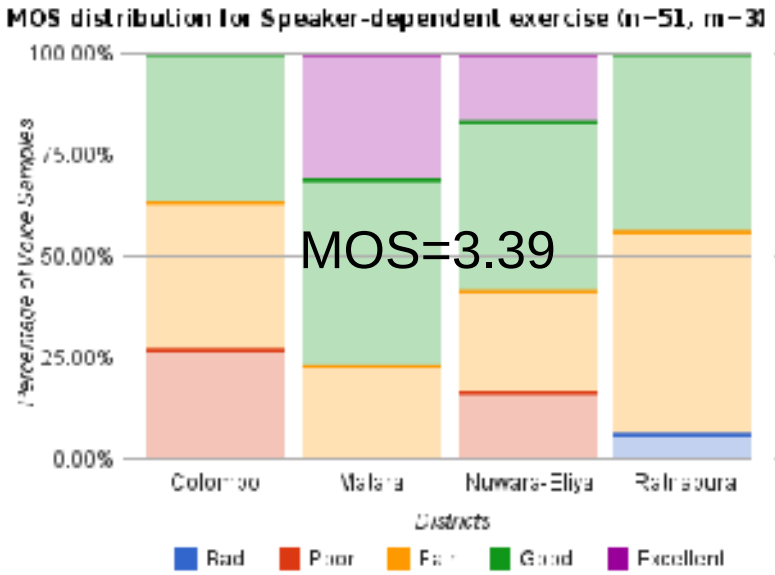
Usability:

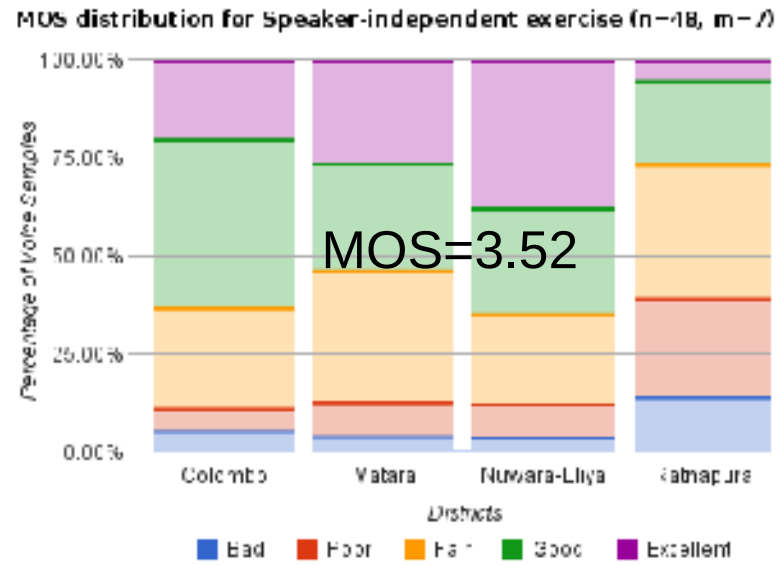
- Human action cycle (HCI)
- Gulf of execution/evaluation (HCI - what system allows and understanding of SoW)

Utility:

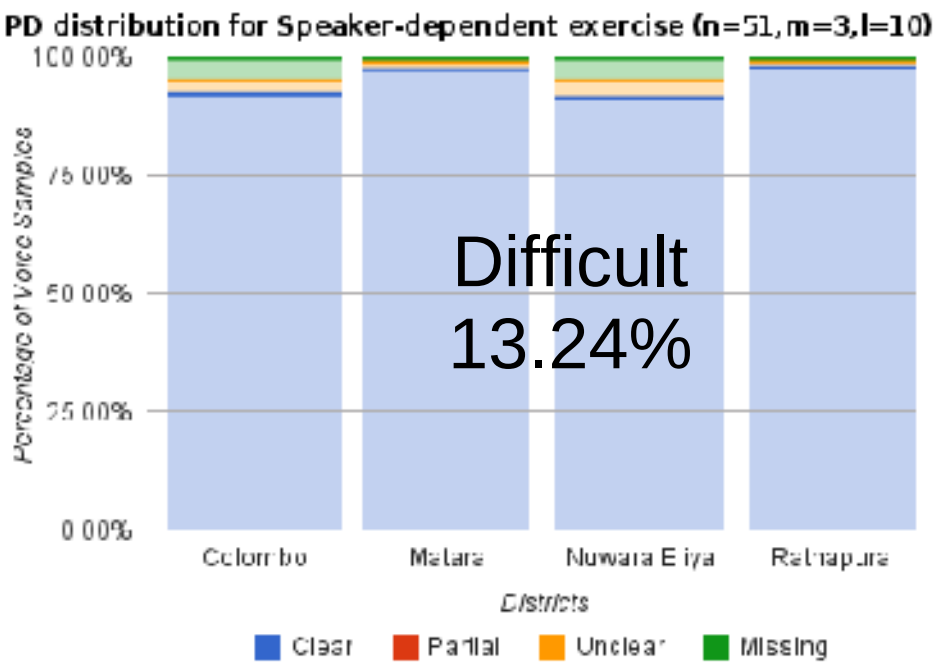
- Ease-of-Use, Usefulness, and Attitude (TAM)

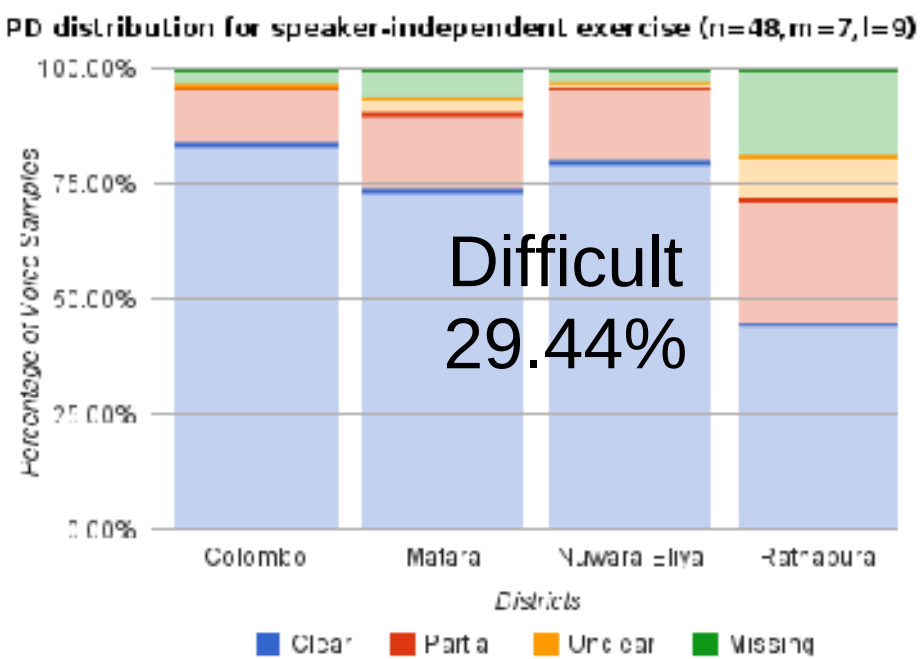
Overall MOS and DS





Speaker-dependent ⟵ Emulates ⟶ Speaker-Independent





Methodology for validating MOS as a diagnostic test

From the 9 data elements 5 were selected as necessary and sufficient; i.e. {Report-purpose, Observation-location, Observation-text, Action-plan, and Immediate-needs}

MOS and DS values of the 5 data elements were used to evaluate MOS as a diagnostic test

Mean Opinion Score (MOS) Classifier = {1, 2, 3, 4, 5}
*Enum :: 1=Poor, 2=Fair, 3=Moderate, 4=Good, 5=Excellen*t

Difficulty Score (DS) predictor = {*clear, unclear, partial, missing*}
Enum :: 1=missing, 2=unclear, 3=partial, 4=clear

{Clear} ---> normal, {Unclear, Partial, Missing} ---> distorted

The two-by-two contingency table

	Actual Value	
Prediction Value	***TP***: MOS classifier $\geq x$ and DS predictor = {normal}	***FP***: MOS classifier $\geq x$ and DS predictor = {distorted}
	FN: MOS classifier $< x$ and DS predictor = {normal}	***TN***: MOS classifier $< x$ and DS predictor = {distorted}

TP=True Positive, FP=False Positive, FN=False negative, TN=True Negative

How to calculate the 2 x 2 contingency table

Filter the evaluators', five key, difficulty score responses to the data elements

District	3 Report Purpose	5 Observation Location	7 Observation text	8 Action Plan	9 Immediate Needs	What is the overall MOS?
Colombo	Partial	Clear	Clear	Clear	Clear	4
Colombo	Clear	Clear	Clear	Clear	Clear	4
Matara	Partial	Unclear	Unclear	Partial	Partial	3
Matara	Clear	Partial	Clear	Partial	Clear	3

Cut-Point=1

District	3 Report Purpose			
	TP	FP	FN	TN
Colombo	0	1	0	0
Colombo	1	0	0	0

Calculate the TP, FP, FN, & TN for each Cut-point 1, 2, 3, 4, 5

Aggregate the TP, FP, FN, and TN for each cut-point and district; calculate Sensitivity and Specificity

	TP	FP	FN	TN	Sensitivity = TP/(TP+FN)	Specificity = TN/(TN+FP)
Cutto01						
CO	140	25	7	3	.9524	.1071
MH	252	78	0	15	1.0000	.1613
NE	192	38	0	10	1.0000	.2083
RN	179	146	4	51	.9781	.2589

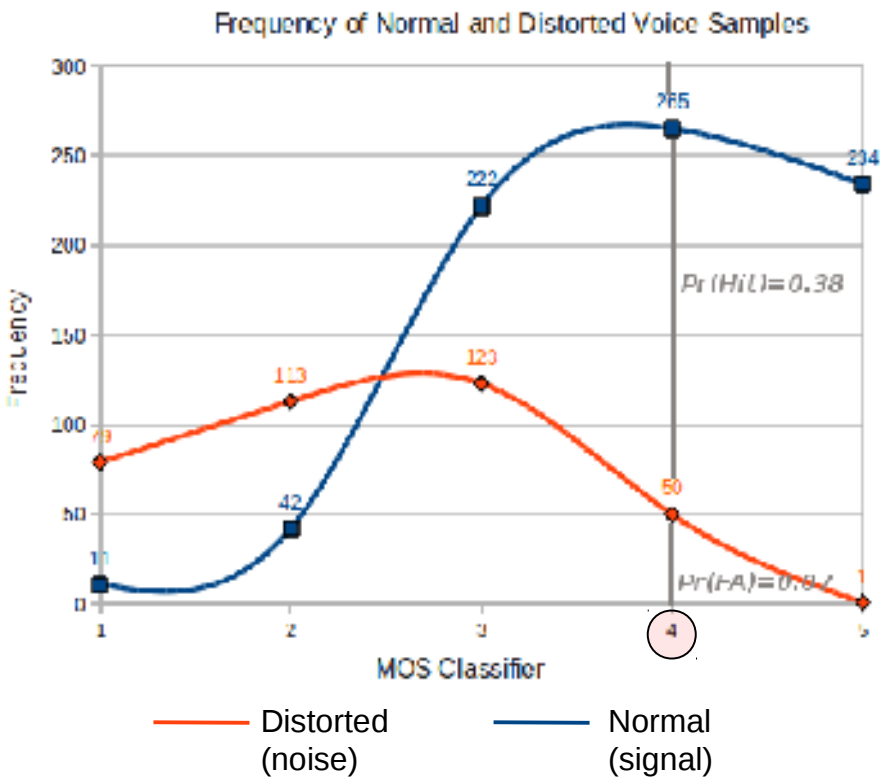
Justifying the Cut-Point

CERT members (subjects) generated voice samples records = 48

Those evaluated by people: m = 7

Clea records available for analysis = 228 of 336

After spiting the record in to five data element, samples: n = 1140



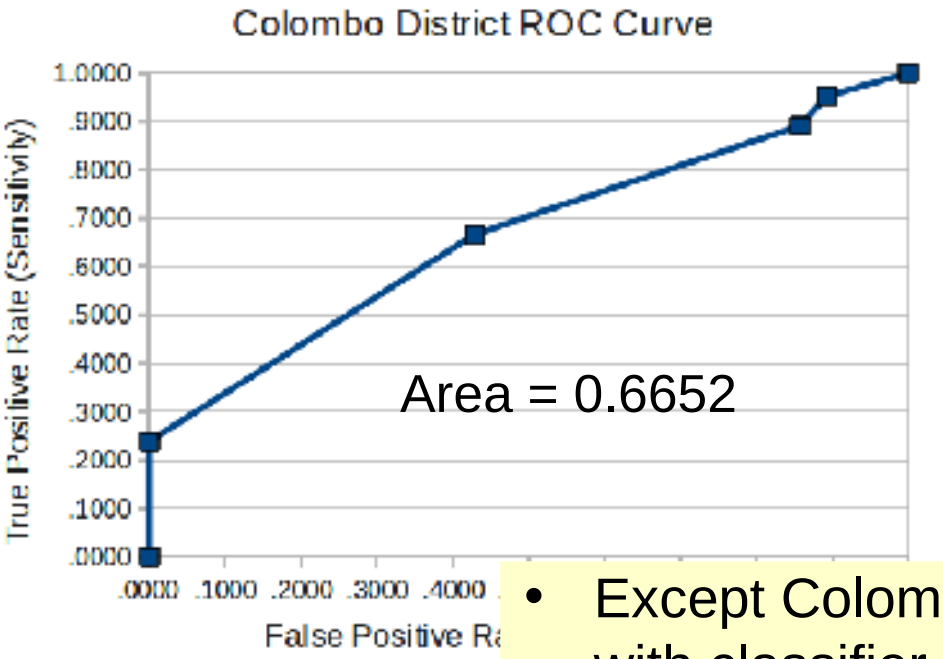
Summary of normal and distorted curve overlapping areas

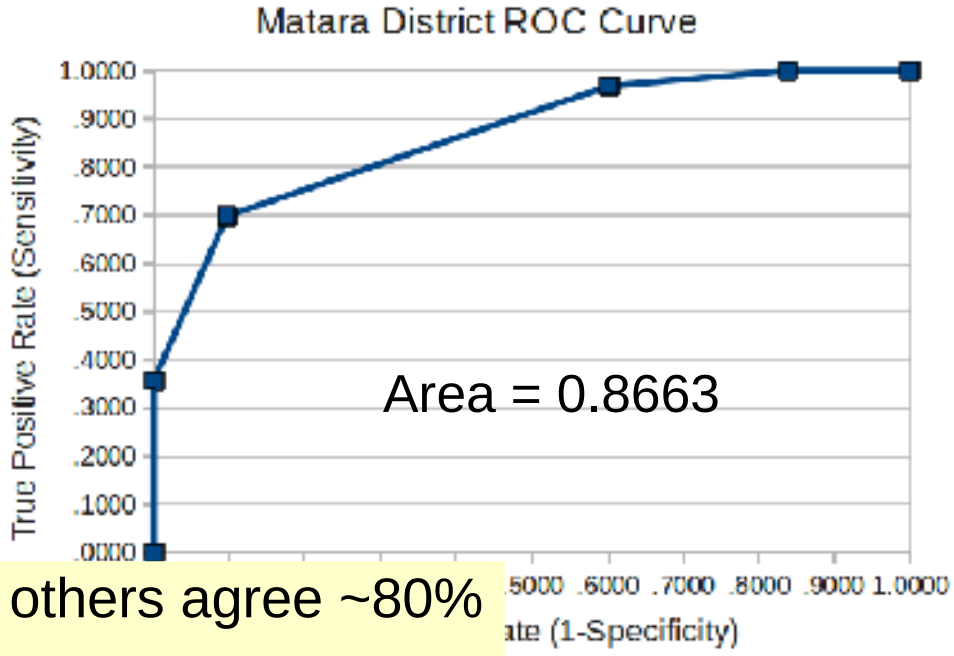
Cut-Point	Area under normal curve	Area under distorted curve	Overlapping area %
1	26.50	96.00	100.00%
2	1132.00	118.00	70.55%
3	243.50	86.50	34.36%
4	249.50	25.50	7.82%
5	0.00	0.00	0.00

FP rate = 7.82% is tolerable, any uncertainties can be reconfirmed

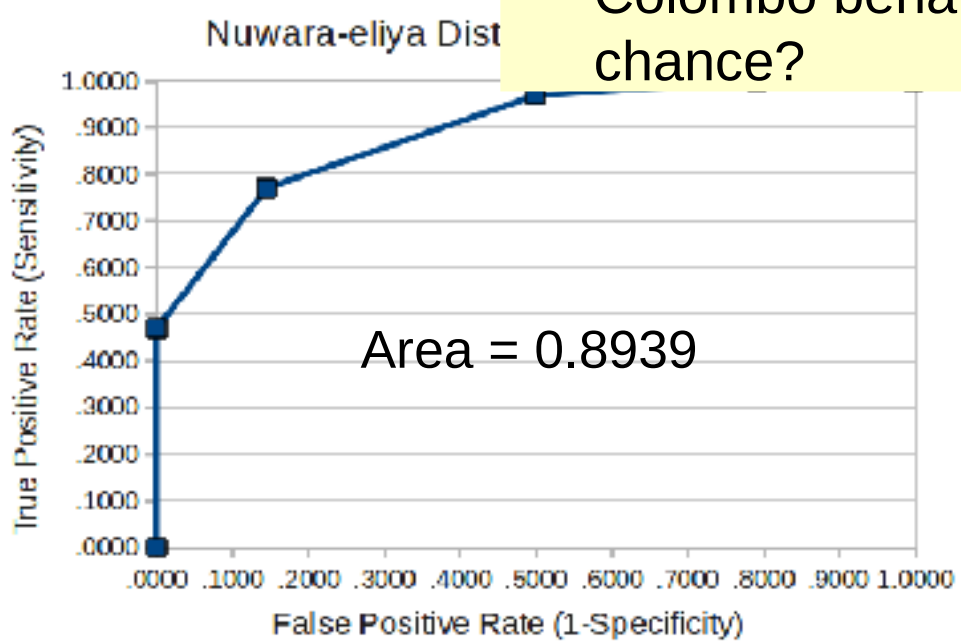
TP rate = 38% not good enough

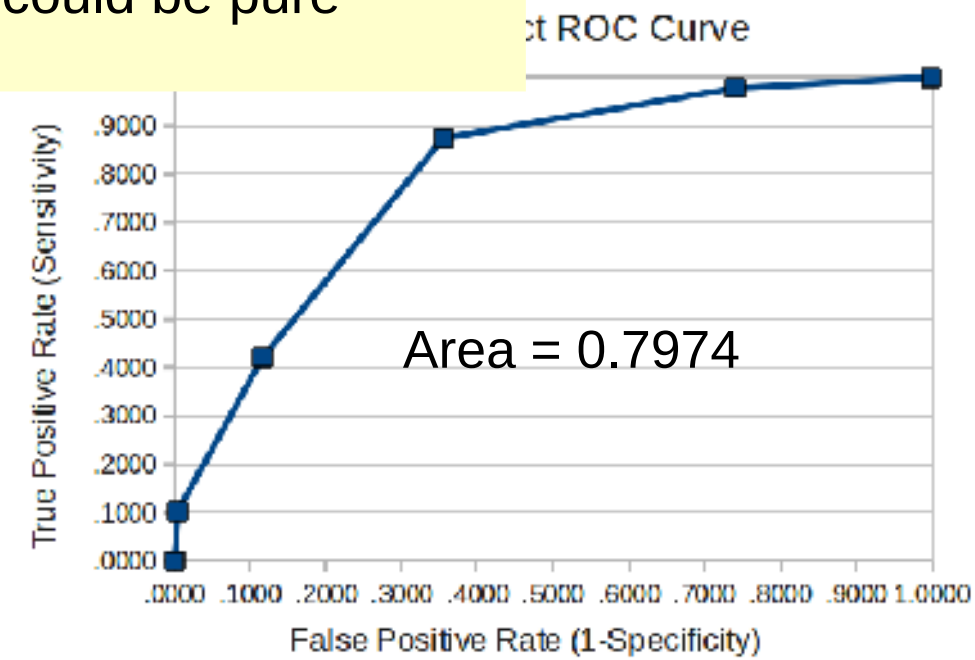
District-wise ROC curves





- Except Colombo all others agree ~80% with classifier
- Colombo behavior could be pure chance?





Conclusions

- MOS is a quick and easy diagnostic test to verify the reliability of voice-enabled technologies for emergency communication; however, it subjective.
- A cut-point of MOS=4.0
    - liberal but would improve the sensitivity making the positive test of voice recording to be strict
    - Need to be strict to remove ambiguity and inefficiencies in emergency communications.
- Forget about Speech-To-Text or Text-To-Speech
    - STT/ASR still in development for Sri Lankan local languages
    - GSM Voice quality too bad for automating transformation
    - FCC has suspended TTS for EDXL-CAP voice messaging
- VoIP quality is better (i.e. MP3)
    - Lookout for 4G voice broadcasts
    - may takes us closer towards automation with TTS/ASR
    - Don't need to invest in GSM modems (unreliable and expensive)

Proposal, URS, SRS, evaluation methodology, technical report
See Project Page: http://lirneasia.net/projects/2010-12-research-program/ff4edxl/

Thank You