

ICT for Preserving Indigenous Languages

Sarah Samson Juan

Senior Lecturer, Faculty of Computer Science and Information Technology
&
Research Fellow, Institute of Social Informatics and Technological Innovations
Universiti Malaysia Sarawak, Malaysia

Pustaka Negeri Sarawak, Kuching



Research on Speech Technology

- ▶ Speech synthesis
- ▶ **Speech recognition**
 - ▶ Speaker recognition/verification
 - ▶ Keyword spotting
- ▶ Multimodal interaction (e.g, speech + image)
- ▶ Speech to speech



Automatic Speech Recognition (ASR)



ASR applications



Languages in Malaysia



- Population: 30 million
- Official language: Malay
- Second language: English



Languages in Malaysia

Living languages

Total: 138



- Population: 30 million
- Official language: Malay
- Second language: English

Lewis, Simons, and Fennig, *Ethnologue : Languages of the world, Seventh Edition, 2014*



Languages in Malaysia



- Population: 30 million
- Official language: Malay
- Second language: English

Living languages

Total: 138

Endangered languages

In Trouble - 101

Dying - 15

Lewis, Simons, and Fennig, *Ethnologue : Languages of the world, Seventh Edition, 2014*



Languages in Malaysia



- Population: 30 million
- Official language: Malay
- Second language: English

Living languages

Total: 138

Endangered languages

In Trouble - 101

Dying - 15

Extinct languages

Total: 2

Lewis, Simons, and Fennig, *Ethnologue : Languages of the world, Seventh Edition, 2014*



How can we help to preserve or maintain languages?

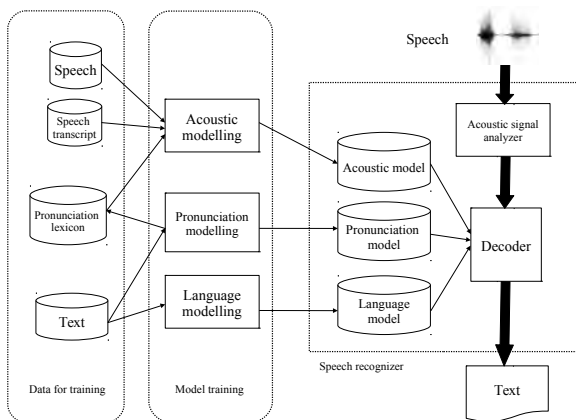
Language documentation:

- ▶ Speech in native language
- ▶ Problem: Transcribing speech manually is a tedious task
- ▶ Automatic speech recognition system can speed up the process

Similar projects: BULB, Aikuma
Local RG: Sarawak Language
Technology (SaLT), Unimas



Automatic speech recognition system (ASR)



Scientific methods in ASR for under-resourced languages

- ▶ Bootstrapping pronunciation dictionary ([Maskey, Black, and Tomokiyo, 2004], [Juan and Besacier, 2013])
- ▶ Merging acoustic models ([Tan, Besacier, and Lecouteux, 2014], [Juan et al., 2015])
- ▶ Cross-lingual and multilingual acoustic models ([Lu, Ghoshal, and Renals, 2014],[Imseng et al., 2014],[Juan et al., 2015])



Iban data - collected for PhD study



Iban Speech Transcription Workshop 2013

Sarah F. S. Juan
17 - 19 July 2013



Speech data:

- ▶ 8 hours of news data
- ▶ Collaborative workshop for collecting speech transcripts
 - ▶ Hire 8 native transcribers
 - ▶ Use Transcriber software [Barras et al., 2000]



Speech transcripts

ibf_002_003 iya madah ka pen-gawa tuk deka berengkah dik-ereja enda lama agi

ibf_002_004 sebengkah kompeni minyak ke nyulut royal dutch shell deka begempung eng-gau petroleum nasional berhad petronas leboh ti bejalai ke dua bengkah projek ngali minyak ba kandang tasik sarawak enggau sabah

ibf_002_005 tuai bagi pekara lng royal dutch shell delareventer madah ka projek tiga puluh

Audio files:

Name	Size	Type	Date Modified	Attributes
ibf_002_007.wav	481,566	Wave Sound	18/6/2015 8:22	none
ibf_002_008.wav	413,340	Wave Sound	18/6/2015 8:22	none
ibf_002_009.wav	411,634	Wave Sound	18/6/2015 8:22	none
ibf_002_011.wav	668,684	Wave Sound	18/6/2015 8:22	none
ibf_002_012.wav	251,116	Wave Sound	18/6/2015 8:22	none
ibf_002_019.wav	324,384	Wave Sound	18/6/2015 8:22	none
ibf_002_021.wav	87,088	Wave Sound	18/6/2015 8:22	none
ibf_002_022.wav	389,624	Wave Sound	18/6/2015 8:22	none
ibf_002_026.wav	45,102	Wave Sound	18/6/2015 8:22	none
ibf_002_027.wav	184,366	Wave Sound	18/6/2015 8:22	none
ibf_002_028.wav	314,414	Wave Sound	18/6/2015 8:22	none
ibf_002_029.wav	387,276	Wave Sound	18/6/2015 8:22	none
ibf_002_031.wav	325,282	Wave Sound	18/6/2015 8:22	none
ibf_002_032.wav	410,824	Wave Sound	18/6/2015 8:22	none
ibf_002_036.wav	621,260	Wave Sound	18/6/2015 8:22	none
ibf_002_039.wav	559,212	Wave Sound	18/6/2015 8:22	none
ibf_002_043.wav	538,204	Wave Sound	18/6/2015 8:22	none
ibf_002_044.wav	72,238	Wave Sound	18/6/2015 8:22	none
ibf_002_047.wav	415,756	Wave Sound	18/6/2015 8:22	none
ibf_002_051.wav	289,852	Wave Sound	18/6/2015 8:22	none
ibf_002_053.wav	335,084	Wave Sound	18/6/2015 8:22	none
ibf_002_055.wav	475,884	Wave Sound	18/6/2015 8:22	none

Iban data - collected for PhD study

Data for creating language model and pronunciation dictionary:

- ▶ Online news articles
- ▶ Obtain 7 thousand articles from 2009-2012
- ▶ 2 million words

```
projek pro dZe KK  
tanah tanah  
pelajar p@ la dZa r  
bn bien  
siti siti KK  
tuai tu waj  
anak anea KK  
kandang kande a NG  
penerang p@ n@ ra NG  
tadi ta di KK  
pesisir p@ sisi@ r  
taja ta dZ@ KK  
sepuluh s@ pulu@ h  
ketuai k@ tu waj  
iban iban
```

Figure: Iban pronunciation dictionary for ASR



Iban corpora for ASR

- ▶ Speech: 7 hours for training acoustic models, 1 hour for system evaluation
- ▶ Language model: 2 million words
- ▶ Pronunciation dictionary: 36 thousand pronunciations
- ▶ Open Source Toolkits for development: Kaldi¹, SRILM², Phonetisaurus³

¹<http://kaldi.sourceforge.net/>

²<http://www.speech.sri.com/projects/srilm/>

³<https://github.com/AdolfVonKleist/Phonetisaurus>

Iban ASR system evaluation

Tested on Iban ASR

pehin sri taib madahka perintah besai udah mega ngemen-
darka duit dua poin tiga biliun ringgit kena ngereja sekeda
projek di serata menua sarawak rambau menteri besai ti be-
jalai kin kitu di menua sarawak dalam kandang tiga taun tu

Play file: `ibf_001_014`



Iban ASR system evaluation

Tested on Iban ASR

nyadi **berikan** tadi ditusun **ramli** haji junaidi ari <**bilik**>
berita rtm kuching lalu disalin **raban jawah**

Play file: ibm_005_171



Iban ASR system evaluation

Summary of Iban ASR results

System	Accuracy (%)
Monolingual	81.25
Cross-lingual	84.85

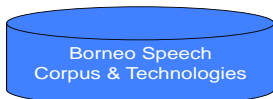
Table: Evaluation on 1 hour data (473 sentences)

- ▶ More information in conference paper [[Juan et al., 2015](#)]
- ▶ ASR accuracy is still quite low
- ▶ Domain-specific system



Future Directions - Long term

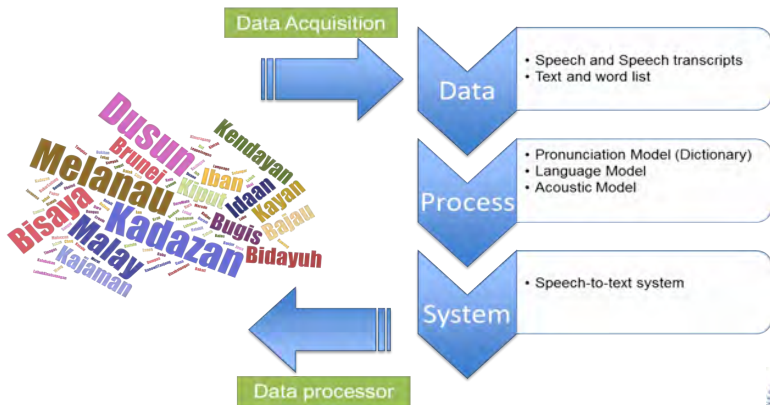
Partners:



Ongoing projects

Target language	Project
Melanau, Iban	Corpus building for Multilingual ASR
Iban, Kelabit	ASR prototypes and for mobile devices
Melanau	Pronunciation dictionary for ASR
Iban	Language modelling for low-resource language

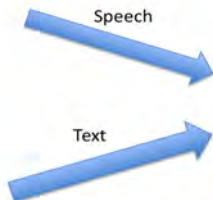
Corpus building for Multilingual ASR



Corpus building for Multilingual ASR

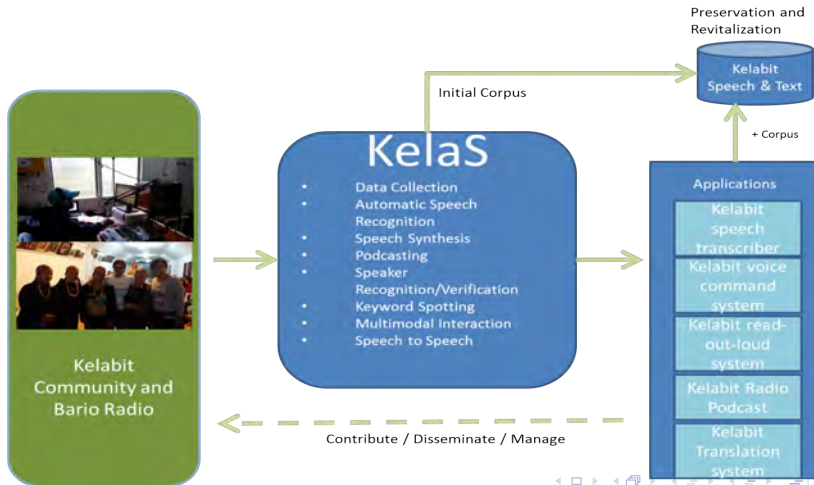


Data collection strategy for multiple languages






- How to develop a corpus building strategy for Sarawak languages?
- Can smart devices be used to collect speech and text from native speakers?
- How to systematically manage stored speech and text data in smart devices?

KelaS: Kelabit Speech Project



References I

-  Barras, C. et al. (2000). “Transcriber: development and use of a tool for assisting speech corpora production”. In: *Proceedings of Speech Communication special issue on Speech Annotation and Corpus Tools*. Vol. 33. available at : trans.sourceforge.net/en/publi.php.
-  Imseng, David et al. (2014). “Using out-of-language data to improve under-resourced speech recognizer”. In: *Speech Communication* 56.0, pp. 142–151.
-  Juan, Sarah Samson and Laurent Besacier (2013). “Fast Bootstrapping of Grapheme to Phoneme System for Under-resourced Languages - Application to the Iban Language”. In: *Proceedings of 4th Workshop on South and Southeast Asian Natural Language Processing 2013*. Nagoya, Japan.

References II

-  Juan, Sarah Samson et al. (2015a). “Merging of Native and Non-native Speech for Low-resource Accented ASR”. In: ed. by Klára Vicsi Adrian-Horia Dediu Carlos Martin-Vide. Springer International Publishing. Chap. Statistical Language and Speech Processing, pp. 255–266.
-  Juan, Sarah Samson et al. (2015b). “Using Resources from a Closely-related Language to Develop ASR for a Very Under-resourced Language: A Case Study for Iban”. In: *Proceedings of INTERSPEECH*. To appear. Dresden, Germany.
-  Lewis, M. Paul, Gary F. Simons, and Charles D. Fennig (2014). *Ethnologue : Languages of the world, Seventh Edition*. SIL International. URL: <http://www.ethnologue.com> (visited on 2013).

References III

-  Lu, Liang, Arnab Ghoshal, and Steve Renals (2014). “Cross-lingual Subspace Gaussian Mixture Models for Low-resource Speech Recognition”. In: *IEEE/ACM Transactions on Audio, Speech and Language Processing*. Vol. 22, pp. 17–27.
-  Maskey, Sameer R., Alan W Black, and Laura M. Tomokiyo (2004). “Bootstrapping Phonetic Lexicons for Language”. In: *Proceedings of INTERSPEECH*, pp. 69–72.
-  Tan, Tien-Ping, Laurent Besacier, and Benjamin Lecouteux (2014). “Acoustic model Merging using Acoustic Models from Multilingual Speakers for Automatic Speech Recognition”. In: *Proceedings of International Conference on Asian Language Processing (IALP)*.