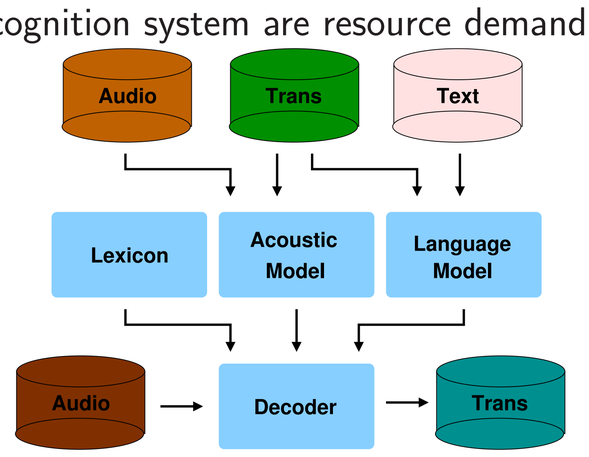
Language Independent Bootstrapping for Automatic Speech Recognition and Keyword Search in Limited Resource Conditions

A. Ragni, K.M. Knill, M.J.F. Gales, {ar527,kate.knill,mjfg}@eng.cam.ac.uk

Department of Engineering, University of Cambridge

1. Introduction

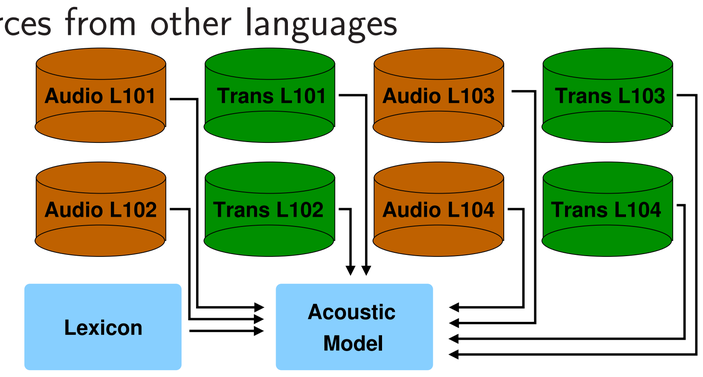
- Standard speech recognition system are resource demanding



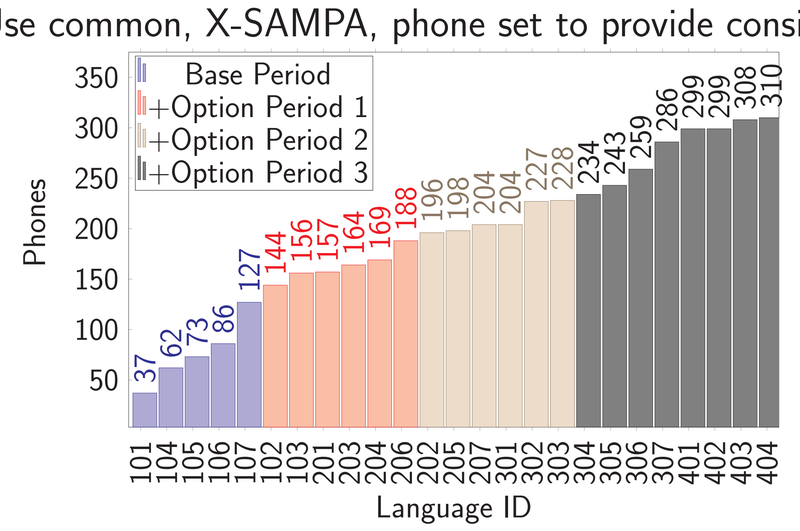
  - vast quantities of transcribed audio and text data, high-quality lexicon
- Impossible to satisfy for roughly 6,500 spoken languages in the world
- Investigate language independent speech recognition systems relying only on
  - (a) limited lexicon; (b) +limited text data; (c) +untranscribed audio

2. Language Independent Models

- Leverage resources from other languages

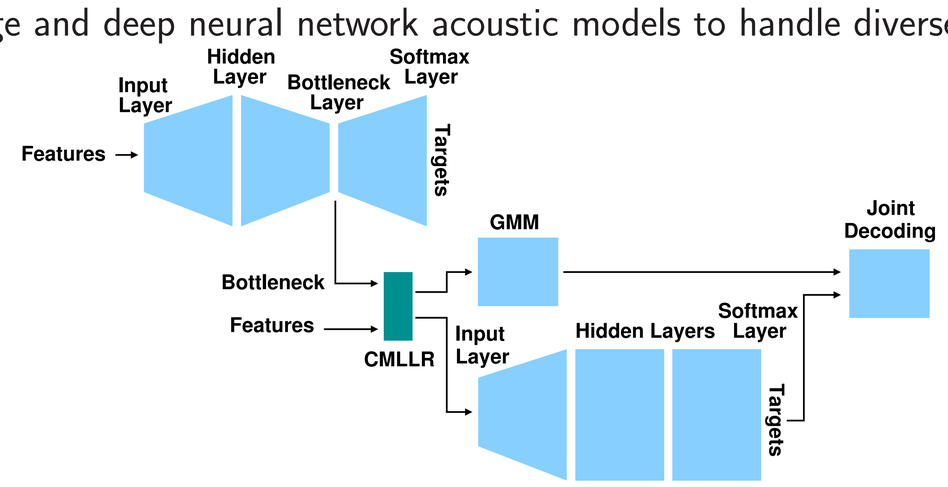


  - need to ensure can generalise to unseen languages
- Use common, X-SAMPA, phone set to provide consistent mapping



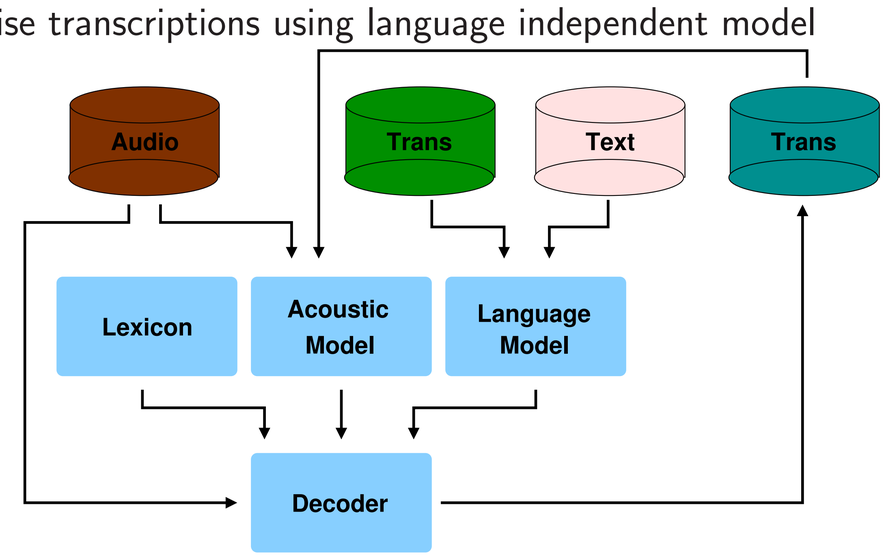
101	Cantonese	207	Tok Pisin
104	Pashto	301	Cebuano
105	Turkish	302	Kazakh
106	Tagalog	303	Telugu
107	Vietnamese	304	Lithuanian
102	Assamese	305	Guarani
103	Bengali	306	Igbo
201	Creole	307	Amharic
203	Lao	401	Mongolian
204	Tamil	402	Javanese
206	Zulu	403	Dholuo
202	Swahili	404	Georgian
205	Kurdish		

- Use large and deep neural network acoustic models to handle diverse data



4. Language Independent Bootstrapping

- Hypothesise transcriptions using language independent model

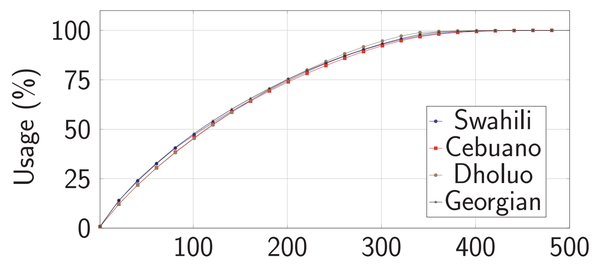


  - use confidence scores to filter out errorful transcripts
- No human annotated transcripts used unlike semisupervised training

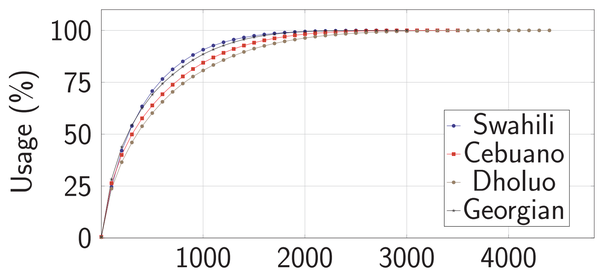
4. Experimental Results: Language Independent Models

- IARPA Babel program – conversational telephone speech
  - training: Cantonese, Pashto, Tagalog, Turkish, Vietnamese, Assamese, Bengali, Haitian Creole, Lao, Tamil, Zulu
  - testing: Swahili, Cebuano, Dholuo, Georgian
- Cumulative state data usage statistics for 3-hour configuration (VLLP)

(a) Language Dependent



(b) Language Independent



  - half of data modelled by 100 states (LD) versus 250 states (LI)
- Lexicon and language model statistics

Language	Size VLLP	Size Web	Vocabulary VLLP	Vocabulary +Web	VLLP Int	OOV VLLP	OOV +Web
Swahili	24703	10M	5423	170529	0.90	13.8	6.6
Cebuano	31959	58M	3642	101188	0.98	9.0	5.8
Dholuo	33762	0.6M	4469	13366	0.97	8.5	6.4
Georgian	23078	116M	5859	270976	0.84	17.3	3.2

  - expect Georgian and Swahili web data to be useful
- Web-only, limited and web language models

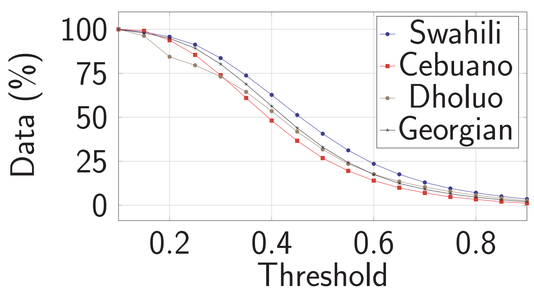
Sys	Audio (hrs)	LM	WER (%) Swahili	Cebuano	Dholuo	Georgian
LD	3	VLLP	65.6	69.8	63.8	71.5
		+Web	63.5	69.8	64.1	67.6
LI	0	Web	78.2	82.2	83.0	81.6
		VLLP	68.2	69.4	70.4	77.0
		+Web	66.3	69.1	71.6	74.0

  - poor performance of web data only LMs, high impact of VLLP LMs

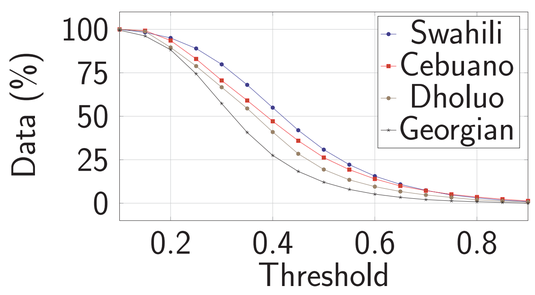
5. Experimental Results: Language Independent Bootstrapping

- Confidence-based untranscribed data selection

(a) Language Dependent



(b) Language Independent



  - 50-60% data above 0.4 threshold for LD versus 25-50% for LI
- Automatic speech recognition performance

Sys	Audio (hrs)	WER (%) Swahili	Cebuano	Dholuo	Georgian
LD	3	63.5	69.8	64.1	67.6
+SS	+60	57.3	66.0	58.7	60.3
LI	0	66.3	69.1	71.6	74.0
+US	+30-60	56.7	66.8	64.3	63.2
FLP	60	44.7	52.3	46.0	49.4

  - significant error reduction from language independent bootstrapping
- Keyword search performance

Sys	Audio (hrs)	MTWV (%) Swahili	Cebuano	Dholuo	Georgian
LD	3	**0.3611**	0.2495	0.2933	**0.4011**
+SS	+60	**0.4394**	0.2809	**0.3591**	**0.5007**
LI	0	**0.3111**	0.2843	0.1823	**0.3301**
+US	+30-60	**0.4466**	**0.3133**	**0.3119**	**0.4916**
FLP	60	0.5442	0.4196	0.5487	0.6030

  - hit 0.3 MTWV IARPA Babel program target on all test languages

6. Conclusions

- Raw performance of language independent models unsatisfactory
  - more normalisation and adaptation, more languages
- Might be suitable for certain downstream tasks