Language Independent and Unsupervised Acoustic Models for Speech Recognition and Keyword Spotting

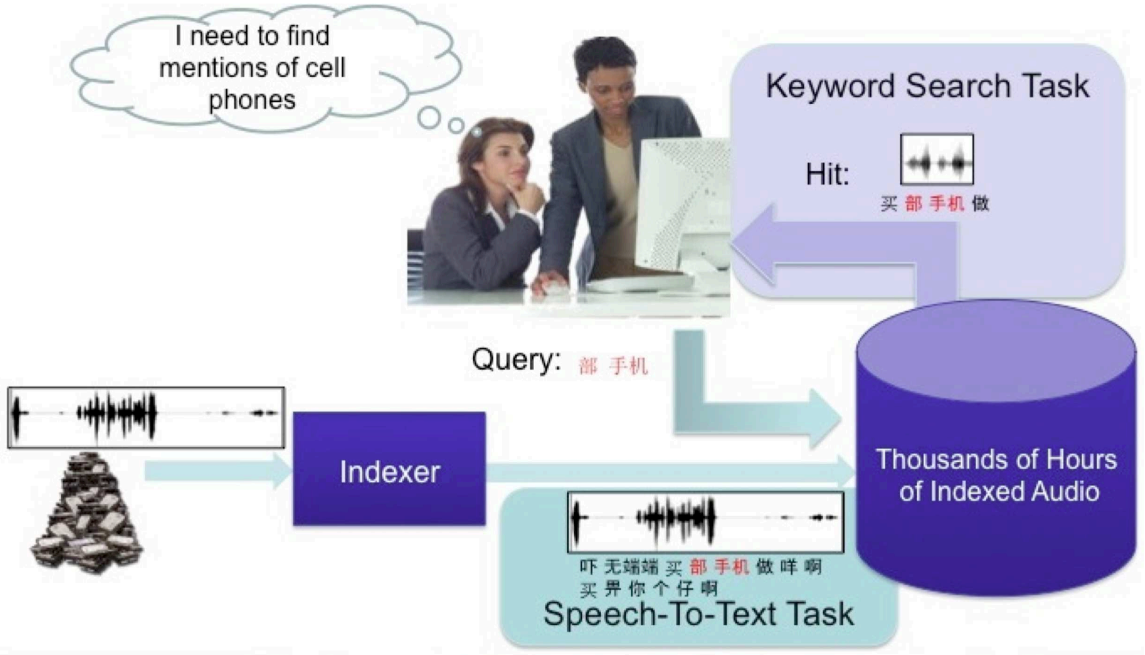
Kate Knill, Mark Gales, Anton Ragni, Shakti Rath

15 September 2014

Cambridge University Engineering Department

Interspeech 2014

IARPA Babel Program


I need to find mentions of cell phones
Keyword Search Task
Hit:
买 部 手机 做
Query: 部 手机
Indexer
Thousands of Hours of Indexed Audio
吓 无端端 买 部 手机 做 咩 啊
买 畀 你 个 仔 啊
Speech-To-Text Task

- Goal - rapidly develop spoken term detection in new languages
  - Broad set of languages with varying phonotactics, phonological, tonal, morphological and syntactic characteristics
  - Speech recorded in variety of conditions
  - Limited amounts of transcription

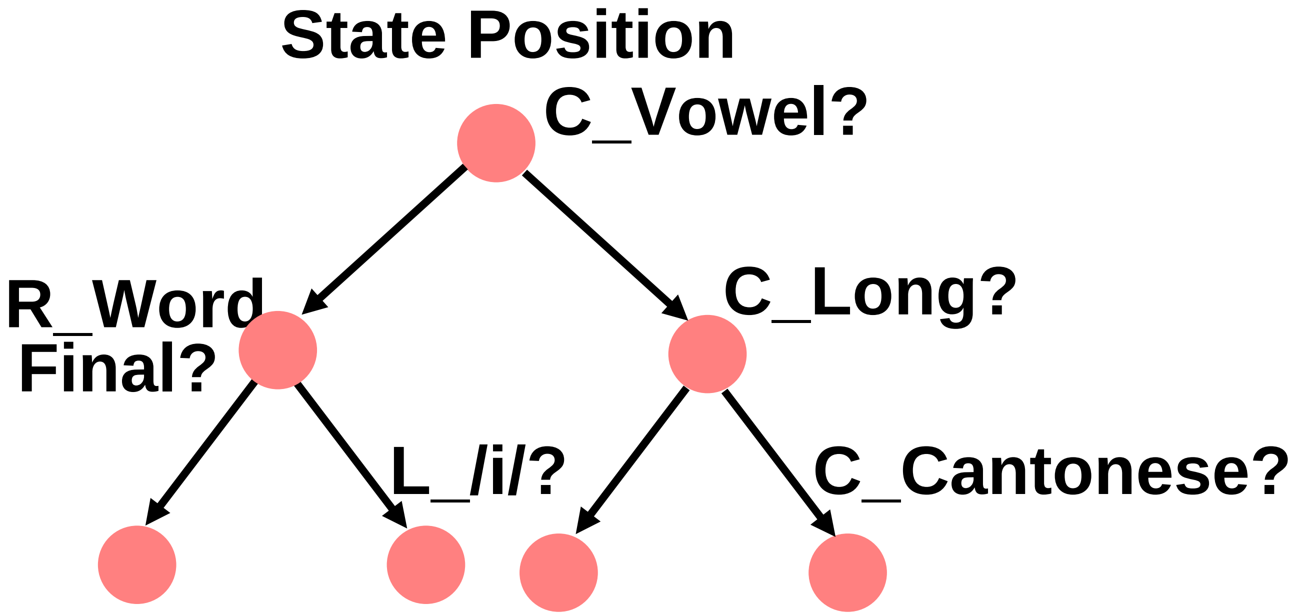
Introduction

- Assumed available data in target language
  - transcribed audio data
  - lexicon and phone set
  - language model training data
- Reduce overhead in deploying new language?
- Zero acoustic resources
  - no acoustic training data available for target language
  - limited lexicon
  - limited language model training data
- Unsupervised acoustic resources
  - target language acoustic training data without transcriptions

Zero-Resource Acoustic Models

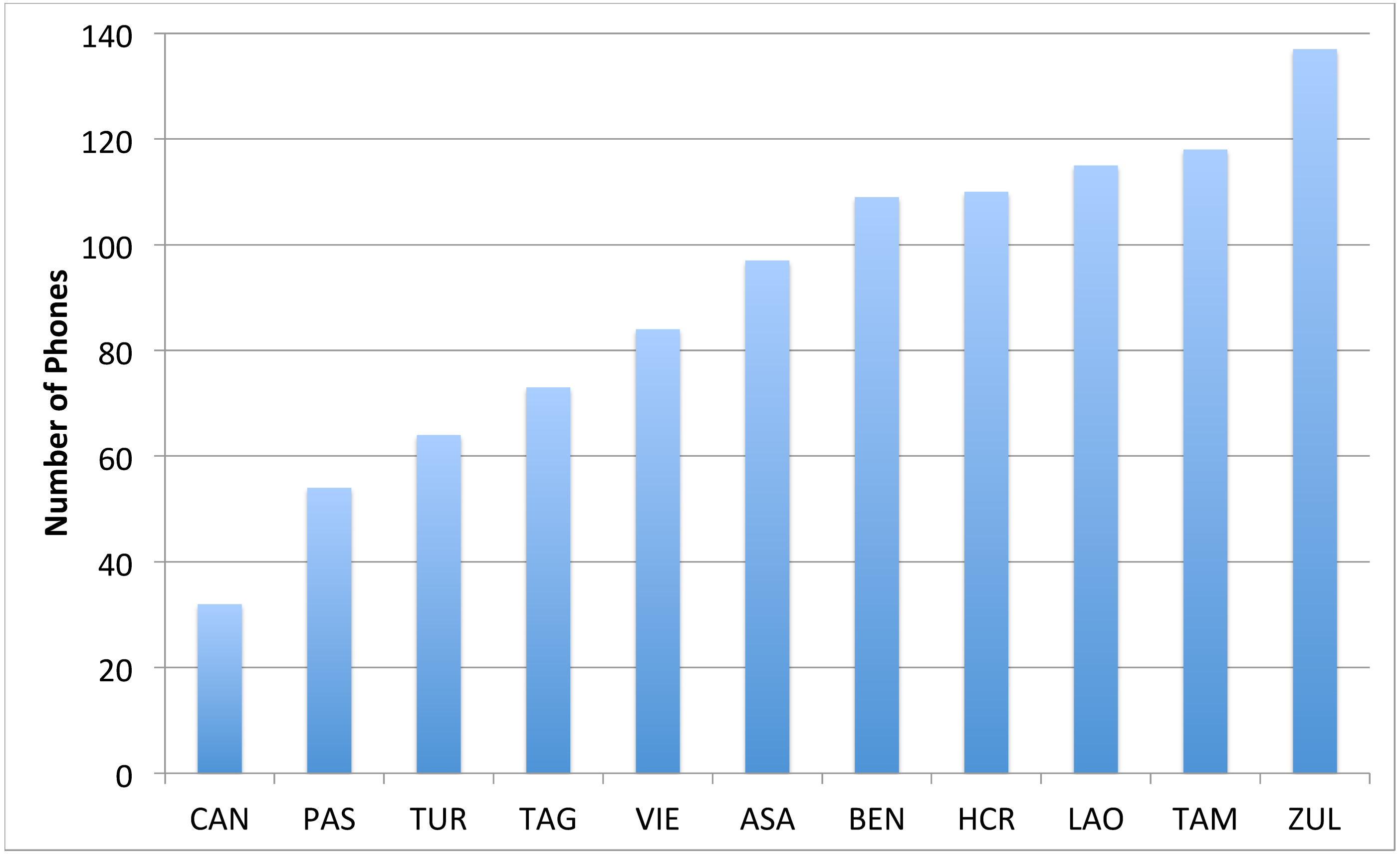
- Scenario
  - no acoustic training data available for target language
  - access to (limited) lexicon and language modelling data
- Language independent acoustic models
  - common phone-set (X-SAMPA)
  - used for both MLP (Tandem/Hybrid) and acoustic model
  - investigated ASR and KWS performance

State Position Root Phonetic Decision Trees



- Assumption: phones are consistent over languages ...
  - requires good phone-set coverage
  - requires consistent phone labelling/attributes
  - use phone attributes to handle missing phones
  - decision trees can represent target language

Phone Set Coverage



- Mapped diphthongs/triphthongs to individual phones
- CUED X-SAMPA attribute file has 215 entries (seen 64%)

Tone Modelling

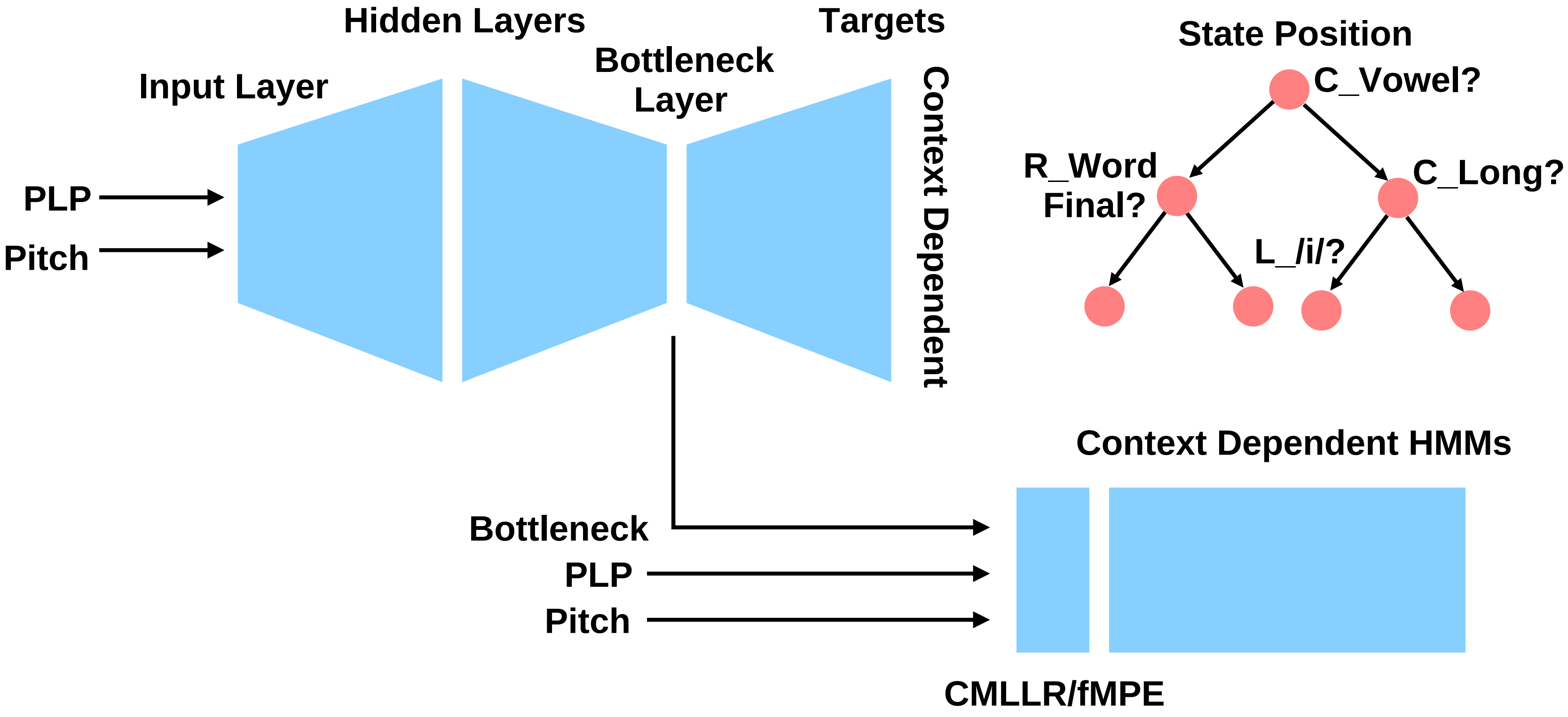
Tone			Training		Unseen
Label	Level	Shape	Can	Lao	Vie
21	high	falling	0	4	—
22	high	level	1	—	—
23	high	rising	2	2	2
32	mid	level	3	1	1
34	mid	dipping	—	—	4
41	low	falling	4	5	3
42	low	level	6	6	—
43	low	rising	5	3	—
61	creaky	falling	—	—	6
63	creaky	rising	—	—	5

- Ask *label*, *level* and *shape* questions in decision tree

Training and Test Languages

Language	Release	# Missing	
		Phones	Tones
Cantonese	IARPA-babel101-v0.4c	—	—
Assamese	IARPA-babel102b-v0.5a	—	—
Bengali	IARPA-babel103b-v0.4b	12	—
Pashto	IARPA-babel104b-v0.4aY	—	—
Turkish	IARPA-babel105b-v0.4	—	—
Tagalog	IARPA-babel106-v0.2f	—	—
Vietnamese	IARPA-babel107b-v0.7	7	3
Haitian Creole	IARPA-babel201b-v0.2b	2	—
Lao	IARPA-babel203b-v3.1a	—	—
Tamil	IARPA-babel204b-v1.1b	4	—
Zulu	IARPA-babel206b-v0.1e	—	—

CUED Language Independent System



- Combine data from LLP from seven languages:
  - Cantonese, Pashto, Turkish, Tagalog, Assamese, Lao, Zulu
- ASR and KWS gains observed using LI bottleneck features

CUED Zero Acoustic Resources System

System		TER (%)	MTWV		
			IV	OOV	Tot
Haitian Creole					
LD	fMPE	61.7	0.4673	0.2347	0.4317
LI	fMPE	77.2	0.2250	0.0966	0.2058
Bengali					
LD	fMPE	68.5	0.3173	0.0987	0.2504
LI	fMPE	81.1	0.1929	0.0775	0.1573
Vietnamese					
LD	fMPE	69.3	0.1962	0.1081	0.1851
LI	fMPE	87.6	0.0255	0.0268	0.0257
Tamil					
LD	fMPE	79.9	0.1540	0.0422	0.1149
LI	fMPE	93.5	—	—	—

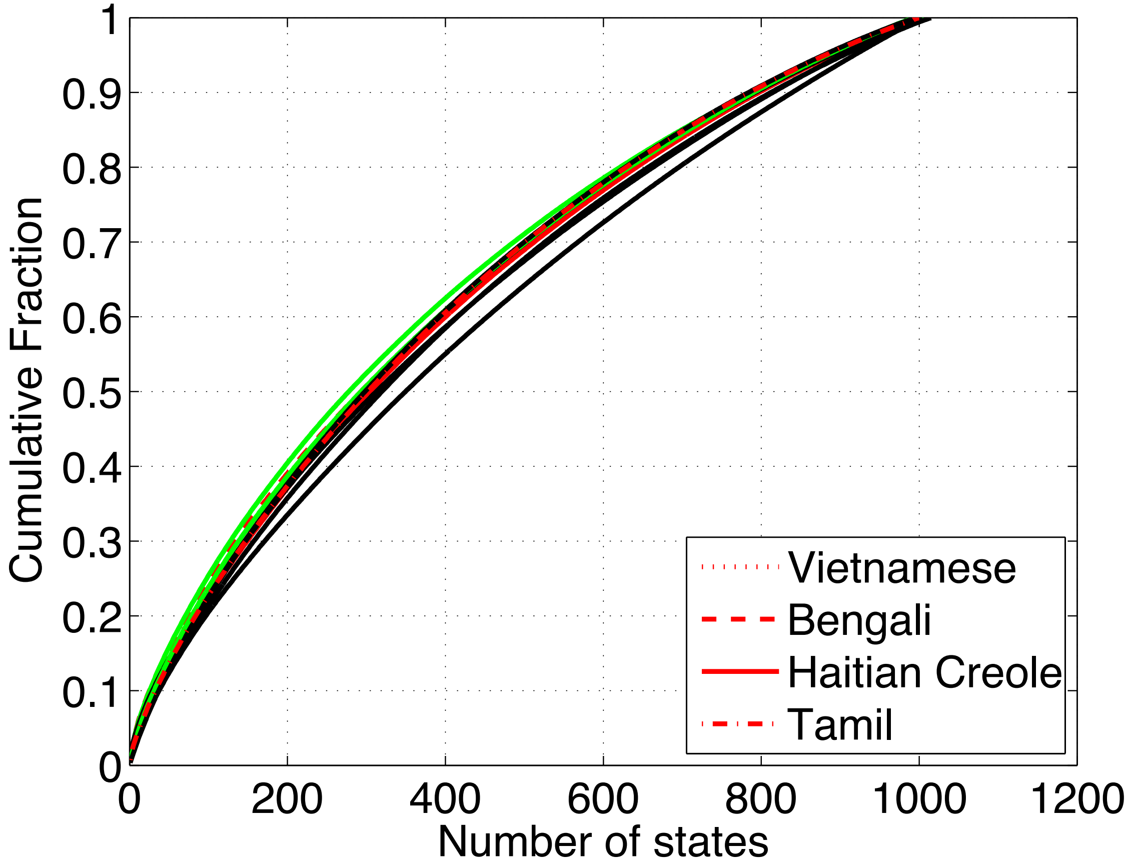
Analysis on Use of Decision Trees

- Possible causes of performance degradation include
  - acoustic realisation mismatch between languages
  - decision trees unrepresentative of target language

- Investigation of decision tree mismatch
  - mismatched - highly uneven distribution of data to leaves
  - large number of contexts mapped to a single leaf

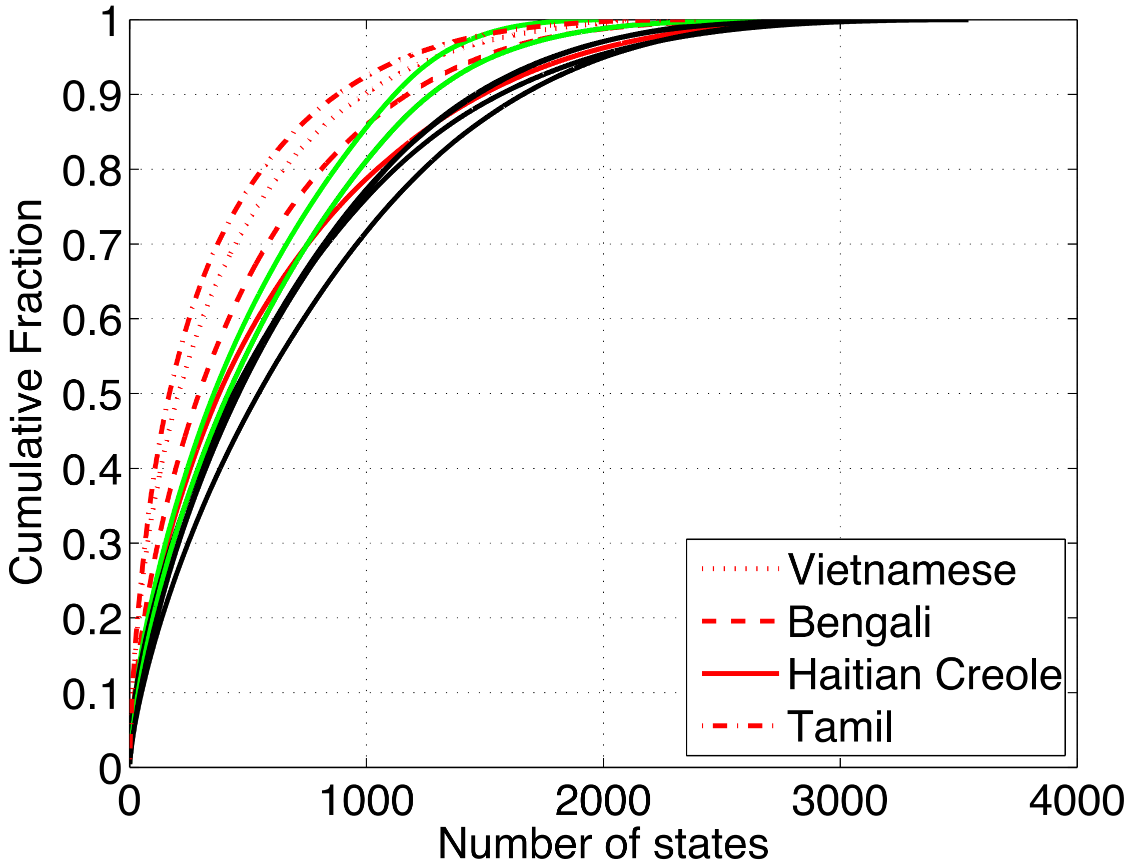
- Approach
  1. Rank order leaf observation counts of individual languages
  2. Plot cumulative distribution against number of states

Language Dependent Decision Trees



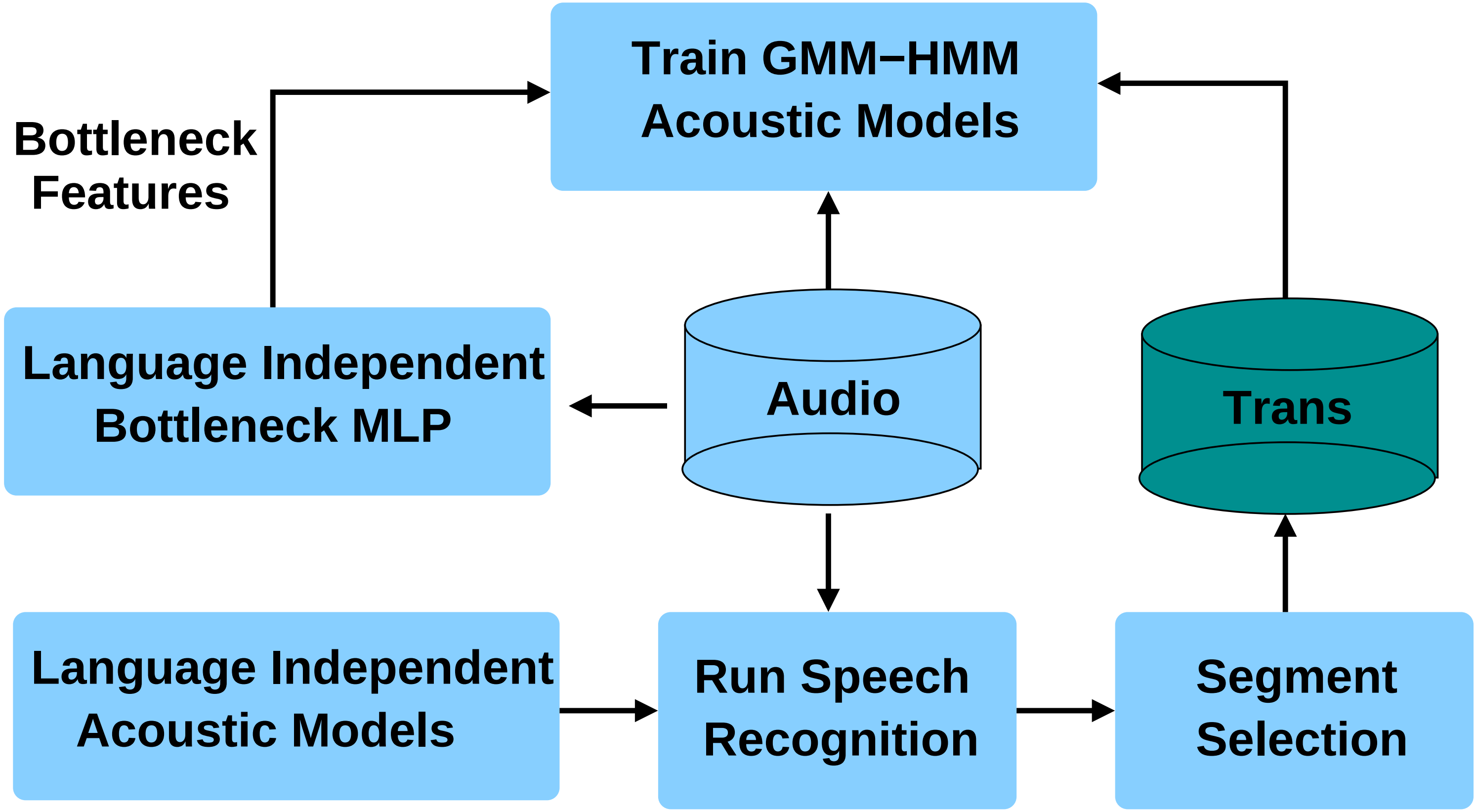
- Distribution of data to leaves relatively even
  - tonal languages (green) slightly less even

Language Independent Decision Trees



- CDF plots follow the WER/KWS performance
  - good indicator of discriminative ability

Unsupervised Acoustic Model Training



- Segments - frame-weighted mapped confidence scores

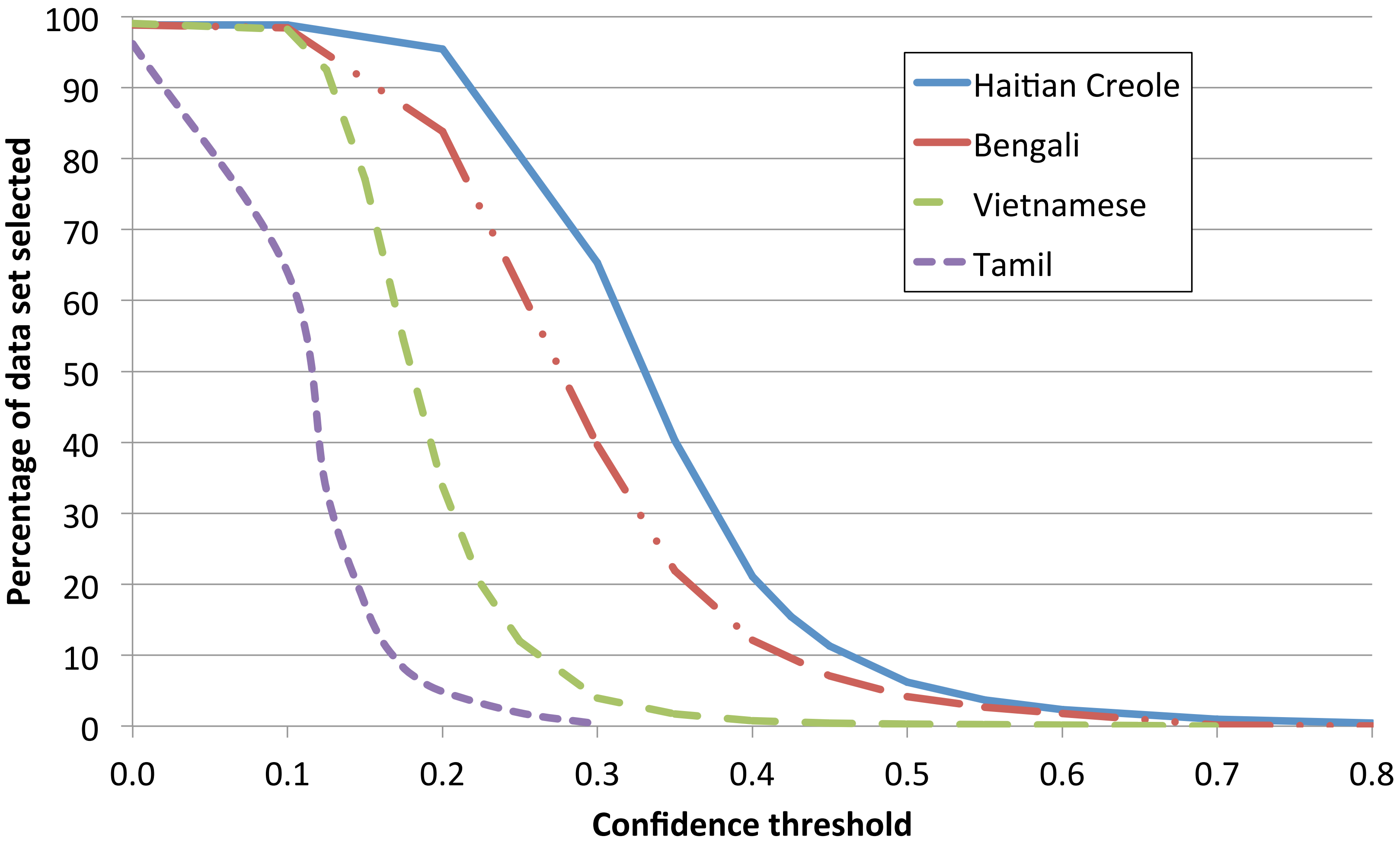
Unsupervised Training Language Resources

- Recognise 60hours full language pack conversational data
  - 10 hours limited language pack (LLP) data excluded
- Language model trained on LLP transcripts
- X-SAMPA lexicon covering LLP training vocabulary

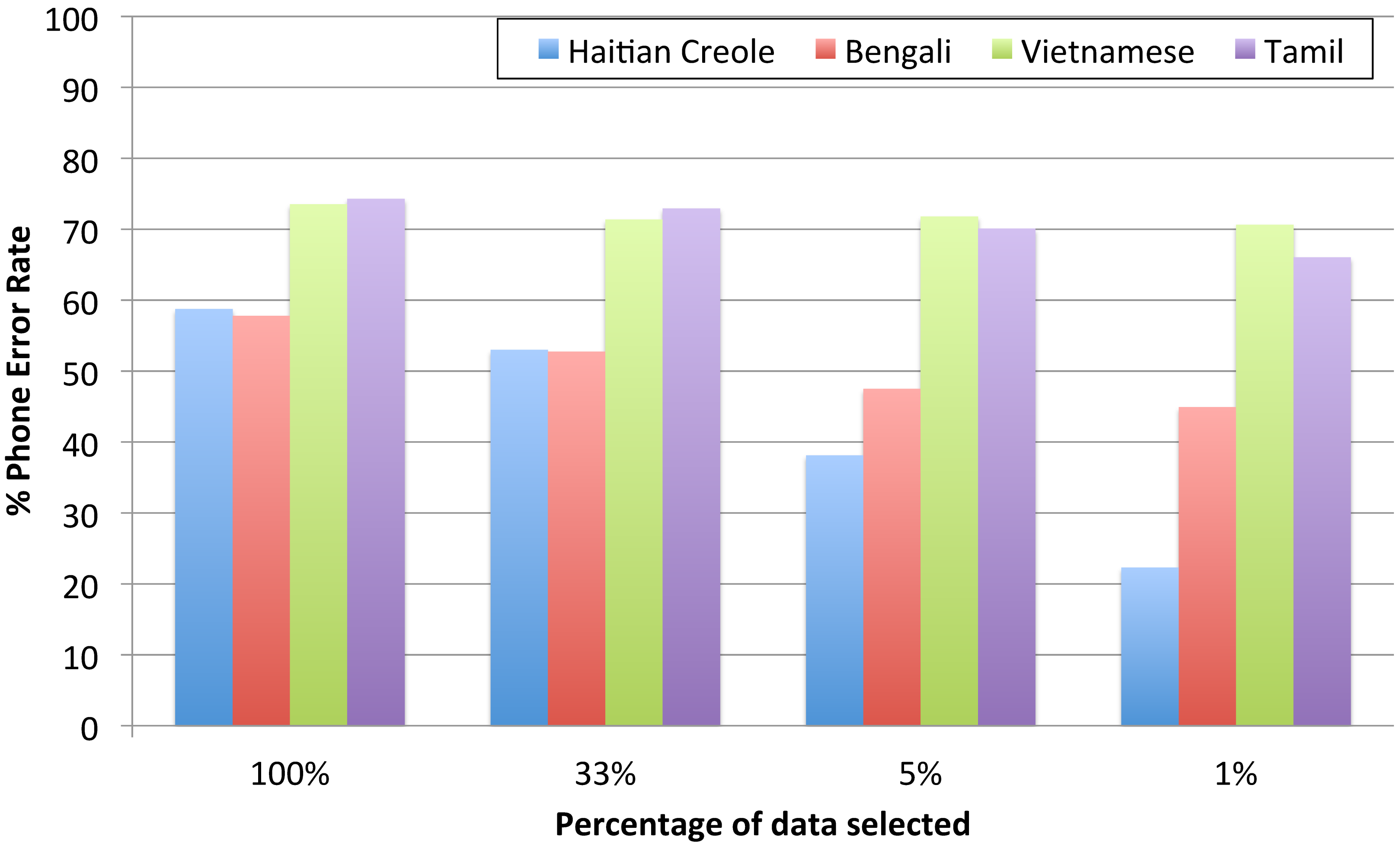
Language	# Words († syllables)	Vocab Size	Bigram LM	
			PPL	%OOV
Haitian Creole	104193	5711	172.5	4.93
Bengali	82406	9511	306.0	8.85
Vietnamese†	122010	3565	173.1	1.56
Tamil	77556	16288	443.3	14.13

- LM in-domain but weakly constrained
  - at least 5-10x fewer words/increase in %OOV compared to literature

Unsupervised Confidence-based Data Selection



Phone Recognition Accuracy



Unsupervised Acoustic Model Training

System		TER	MTWV		
		(%)	IV	OOV	Tot
Haitian Creole					
LD	fMPE	61.7	0.4673	0.2347	0.4317
LI	fMPE	77.2	0.2250	0.0966	0.2058
UN	ML	71.4	0.2907	0.1462	0.2691
Bengali					
LD	fMPE	68.5	0.3173	0.0987	0.2504
LI	fMPE	81.1	0.1929	0.0775	0.1573
UN	ML	75.9	0.2068	0.0913	0.1723
Vietnamese					
LD	fMPE	69.3	0.1962	0.1081	0.1851
LI	fMPE	87.6	0.0255	0.0268	0.0257
UN	ML	84.9	0.0086	0.0357	0.0174

Conclusions

- Zero resource acoustic models
  - consistency of mappings (phone sets, decision trees) required
  - observed uneven distribution of leaf node occupancy
  - results highly variable depending on target language

- Unsupervised acoustic model training
  - transcription quality constrained by LM and decision trees
  - need to make better use of confidence scores

Questions?

This work was supported in part by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Defense U.S. Army Research Laboratory (DoD / ARL) contract number W911NF-12-C-0012. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoD/ARL, or the U.S. Government.