ASR and KWS for Low Resource Languages: Babel Project Research at CUED

Mark Gales, Kate Knill, Anton Ragni, Shakti Rath
CUED Babel Team (X. Chen, X. Liu, P.C. Woodland,T. Yoshioka, C. Zhang),
Lorelei Babel Team

May 2014

Cambridge University Engineering Department

SLTU-14 May 2014

Overview

- IARPA Babel program
- Keyword Spotting System
- Speech-to-Text (ASR) research
  - deep neural networks
  - data augmentation
  - "zero acoustic model resource" systems
- System Performance (Option Period 1 Languages)

IARPA Babel Program

"The Babel Program will develop agile and robust speech recognition technology that can be rapidly applied to any human language in order to provide effective search capability for analysts to efficiently process massive amounts of real-world recorded speech." - Babel Program BAA

IARPA Babel Program Specifications

- Language Packs
  - Conversational and scripted telephone data (plus other channels)
  - Full: 60-80 hours transcribed speech
  - Limited: 10 hours transcribed speech (plus untranscribed speech)
  - 10 hour Development and Evaluation sets
  - Lexicon covering training vocabulary
  - X-SAMPA phone set
  - Collected by Appen (ABH)

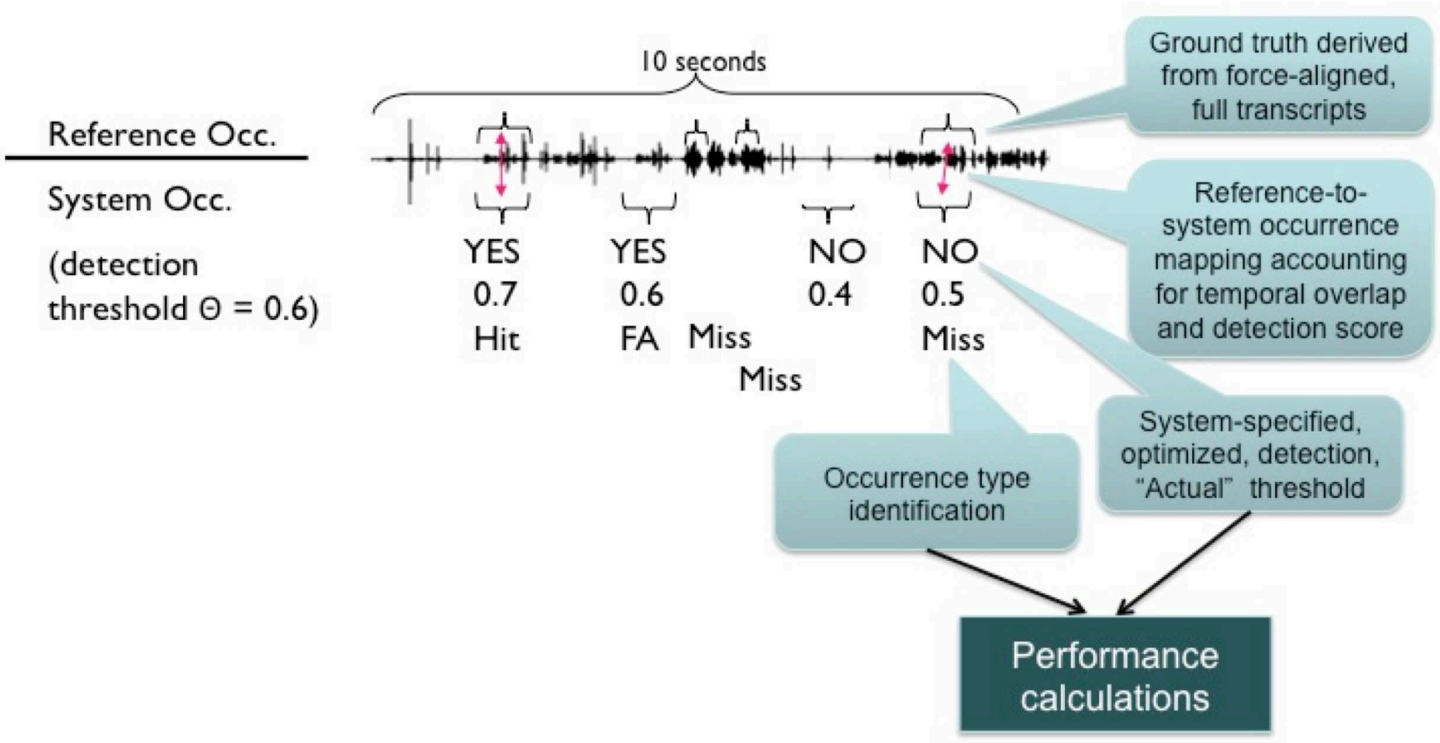
- Evaluation conditions
  - BaseLR - teams can only use data within a language pack
  - BabelLR - can use data from any language pack
  - OtherLR - can add data from other sources e.g. web

IARPA Babel releases

This work uses the IARPA Babel Program language collection releases:

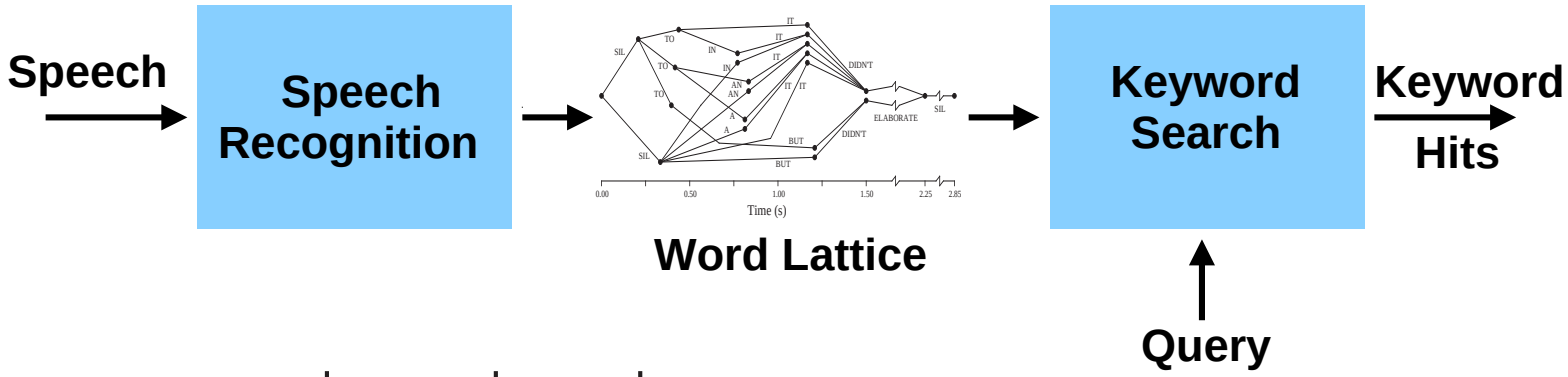
Language	Id	Release
Cantonese†	101	IARPA-babel101-v0.4c
Assamese†	102	IARPA-babel102b-v0.5a
Bengali	103	IARPA-babel103b-v0.4b
Pashto†	104	IARPA-babel104b-v0.4aY
Turkish†	105	IARPA-babel105b-v0.4
Tagalog†	106	IARPA-babel106-v0.2f
Vietnamese	107	IARPA-babel107b-v0.7
Haitian Creole	201	IARPA-babel201b-v0.2b
Lao†	203	IARPA-babel203b-v3.1a
Tamil	204	IARPA-babel204b-v1.1b
Zulu†	206	IARPA-babel206b-v0.1e

IARPA Babel Program Metric



- Term Weighted Value (TWV) - official metric ($\beta = 999.9$)
  - $TWV(\theta) = 1 - [P_{Miss}(\theta) + \beta P_{FA}(\theta)]$
- Target: achieve above 0.3000 on each language pack

Lorelei Team Spoken Term Detection



- Query terms can be words or phrases
- IBM WFST-based keyword search system
  - in-vocabulary terms searched at word level
  - normalised posterior probabilities using "sum-to-one" (STO)
- Scores quoted are Maximum Term Weighted Value (MTWV)
  - bigger is better!

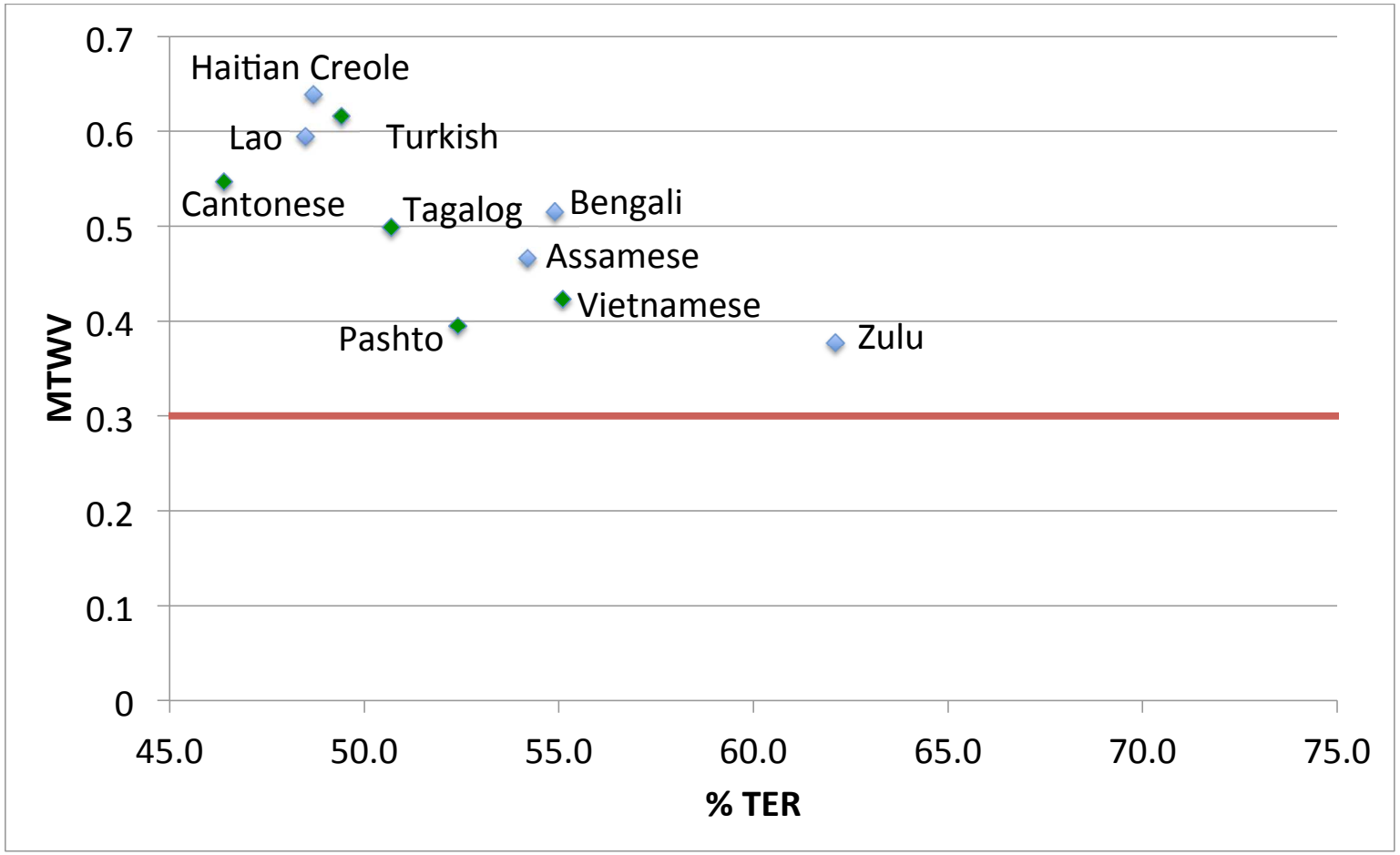
KWS Options

- Limited resources can yield high OOV rates for, e.g. agglutinative languages
  - Zulu Limited Language Pack: 61% development query terms OOV

KWS Process	MTWV		
	IV	OOV	Tot
Word	0.2655	0.0000	0.1033
+phone	0.2596	0.0970	0.1606
+cascade	0.2609	0.0970	0.1611
+lm0	0.2649	0.1338	0.1851
+morph	0.2615	0.2073	0.2287

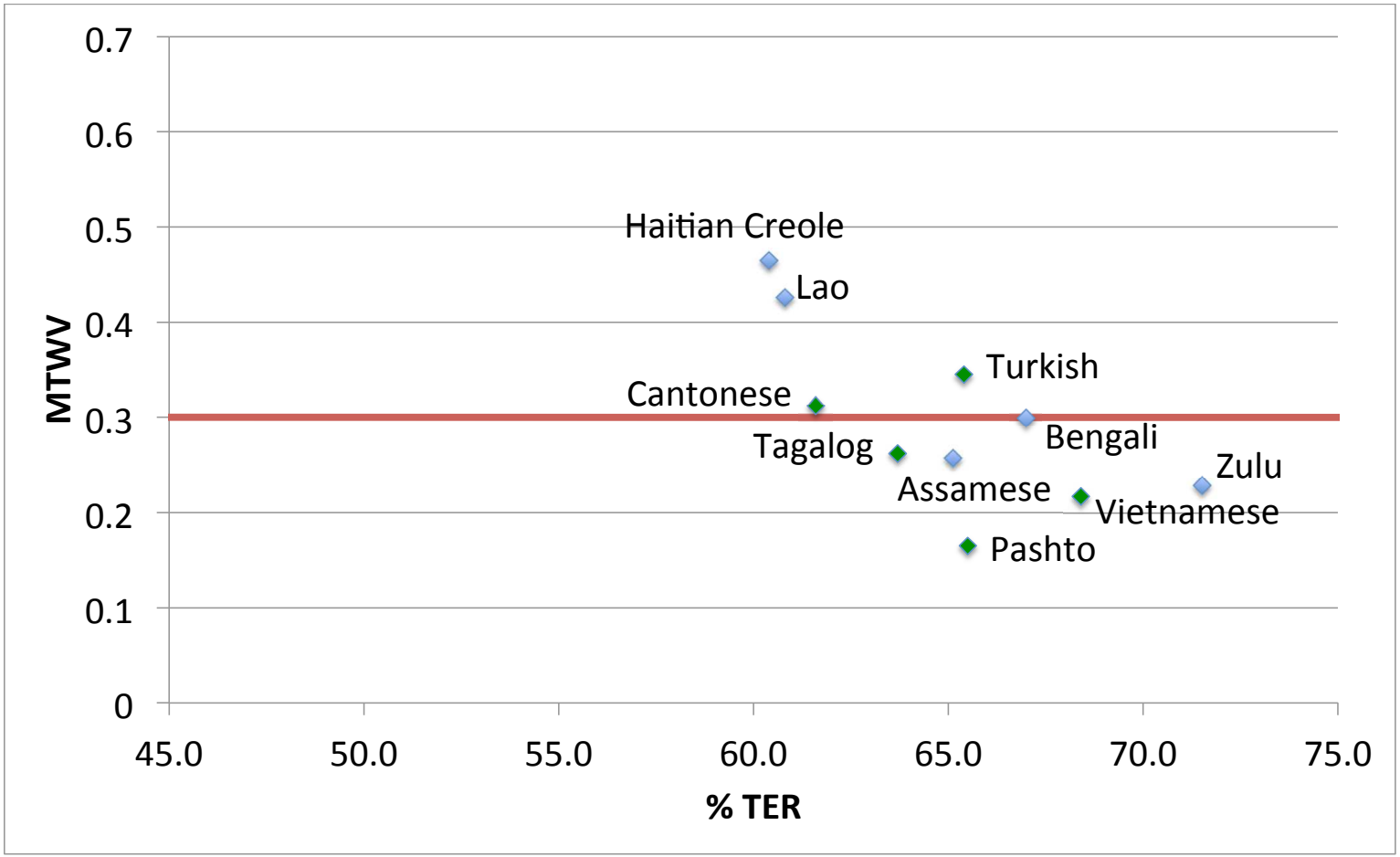
- A range of approaches developed by Lorelei team to address OOVs
  - +phone: map lattice to phones, phone KWS (with confusions)
  - +cascade: treat missed IV terms as OOV
  - +lm0: set LM scores to zero for OOV search
  - +morph: generate morph lattices, do IV search for morphs

CUED Full Language Packs (snapshot)



- green indicates Base Period languages
- blue indicates Option Period 1 languages

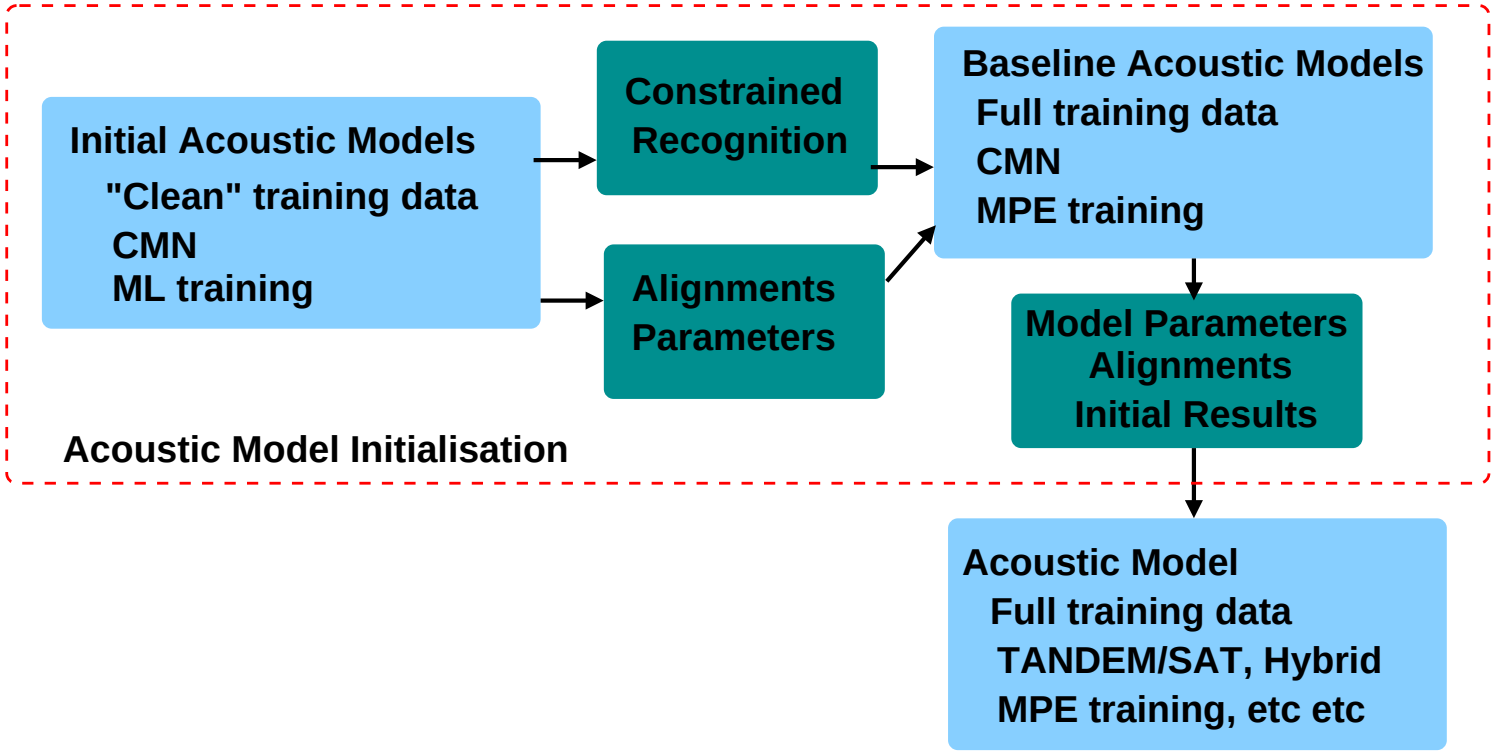
CUED Limited Language Packs (snapshot)



- green indicates Base Period languages
- blue indicates Option Period 1 languages

Speech-to-Text (ASR)

General Training Procedure

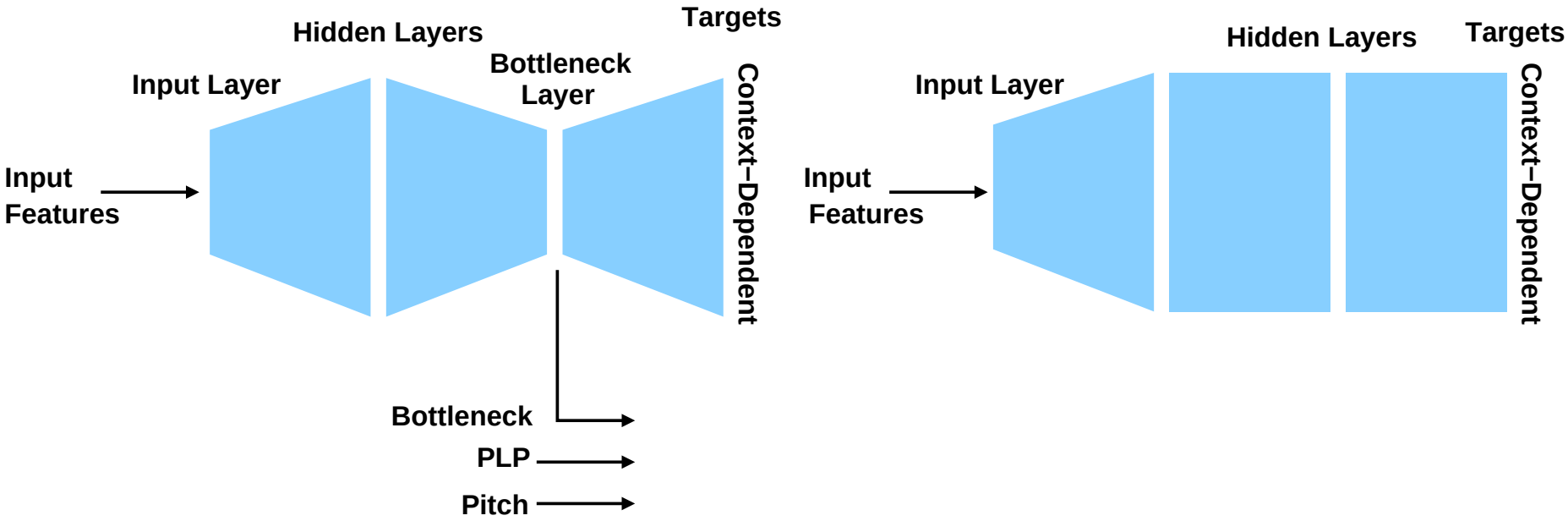


- "Clean" training data - remove segments containing:
  - unintelligible ((())), mispronounce (*WORD*), fragment (WORD-)
- Pronunciations for above symbols derived by highly constrained recognition

Speech-to-Text Systems

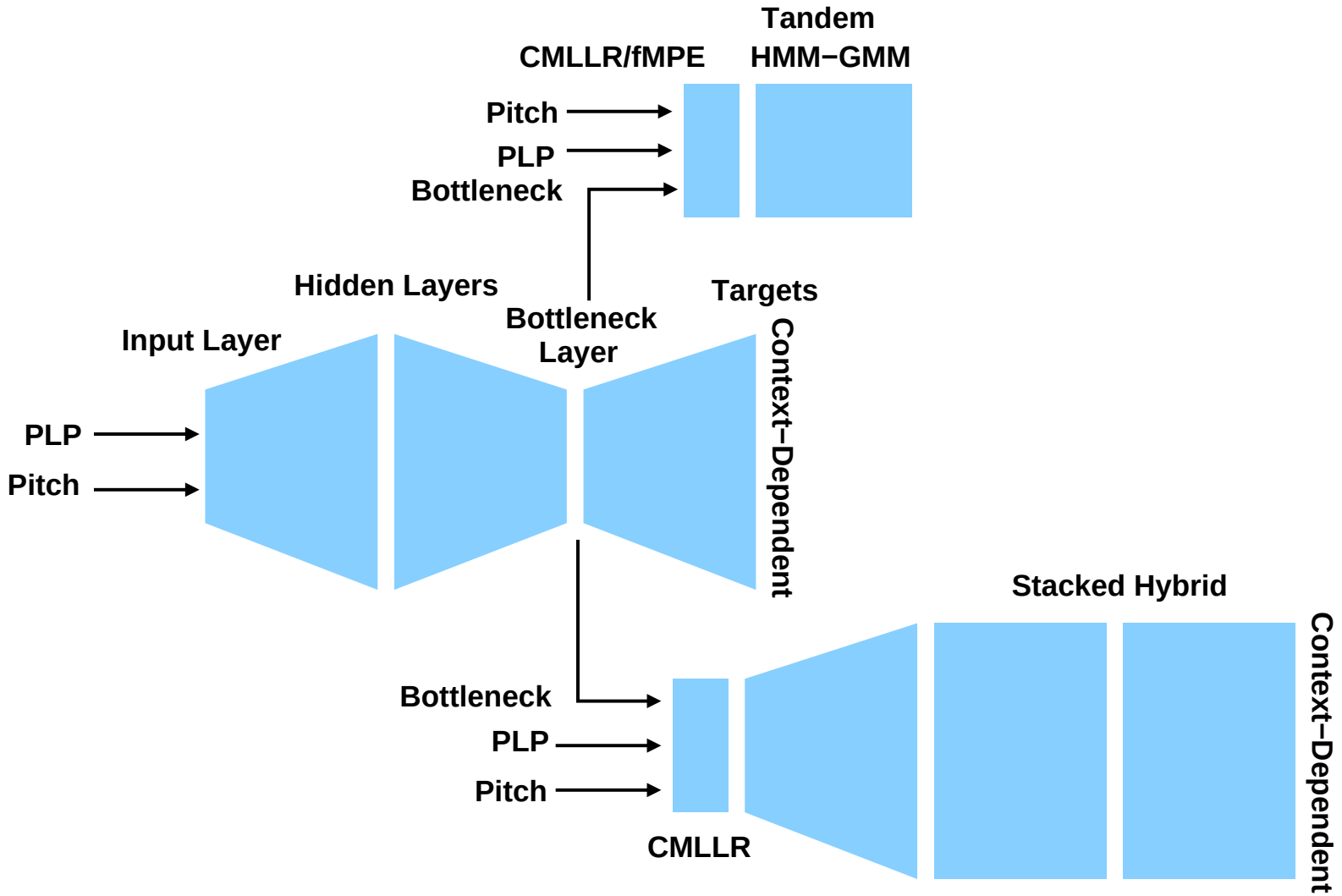
- Describe three areas investigated at CUED
- Deep Neural Network Systems
  - comparison of Hybrid and Tandem performance
- Data Augmentation
  - automatic data/transcription generation
  - multi-language resources
- "Zero Acoustic Model Resource" Systems
  - language-independent systems
  - unsupervised acoustic model training

Use of (Deep) Neural Networks



- Develop both Tandem and Hybrid system configurations
  - results are complementary (both for ASR and KWS) - see later
  - gains from techniques often apply to both set-ups
  - but systems also have different advantages

Tandem and Stacked Hybrid System



- Common features - different classifiers

FLP Tandem and Hybrid Performance

- Hybrid currently trained using the cross-entropy criterion
  - sequence training almost done

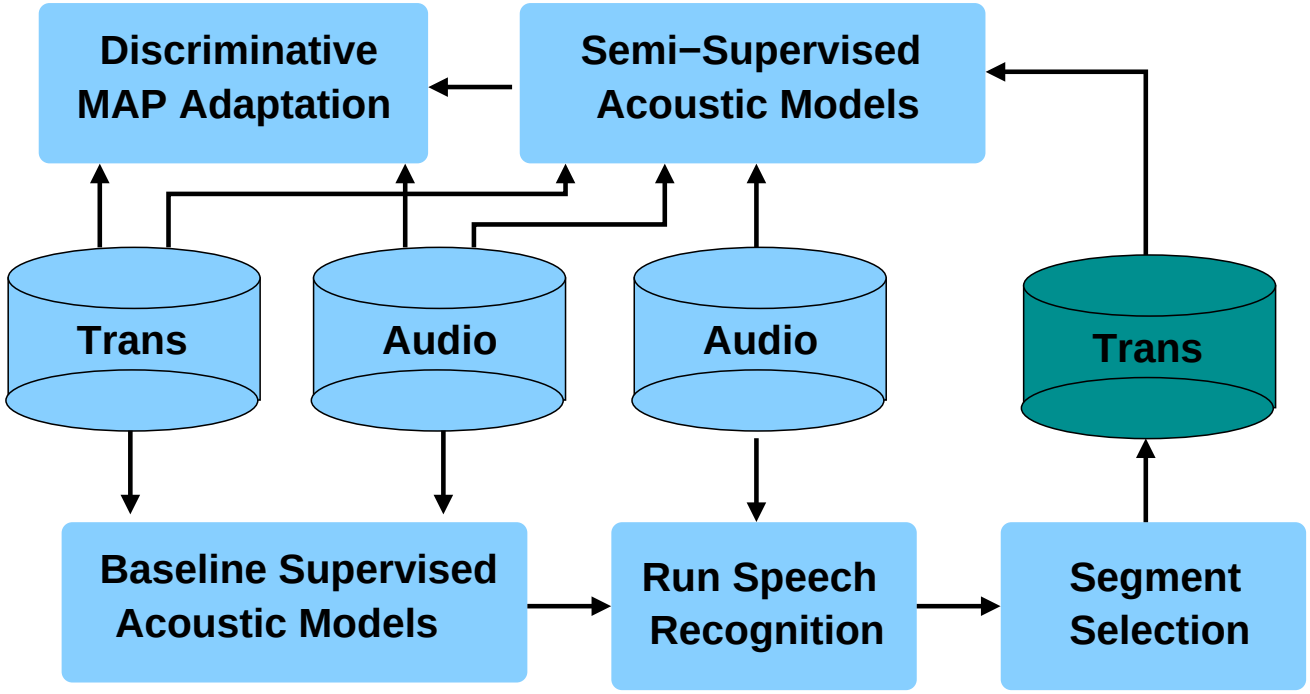
Language	System	TER (%)	MTWV
Vietnamese	Tandem	55.1	0.423
	Hybrid	54.4	0.418
Cantonese	Tandem	46.4	0.547
	Hybrid	46.9	0.542

- Similar performance for both configurations for both ASR and KWS
  - examine combination later ...

Data Augmentation

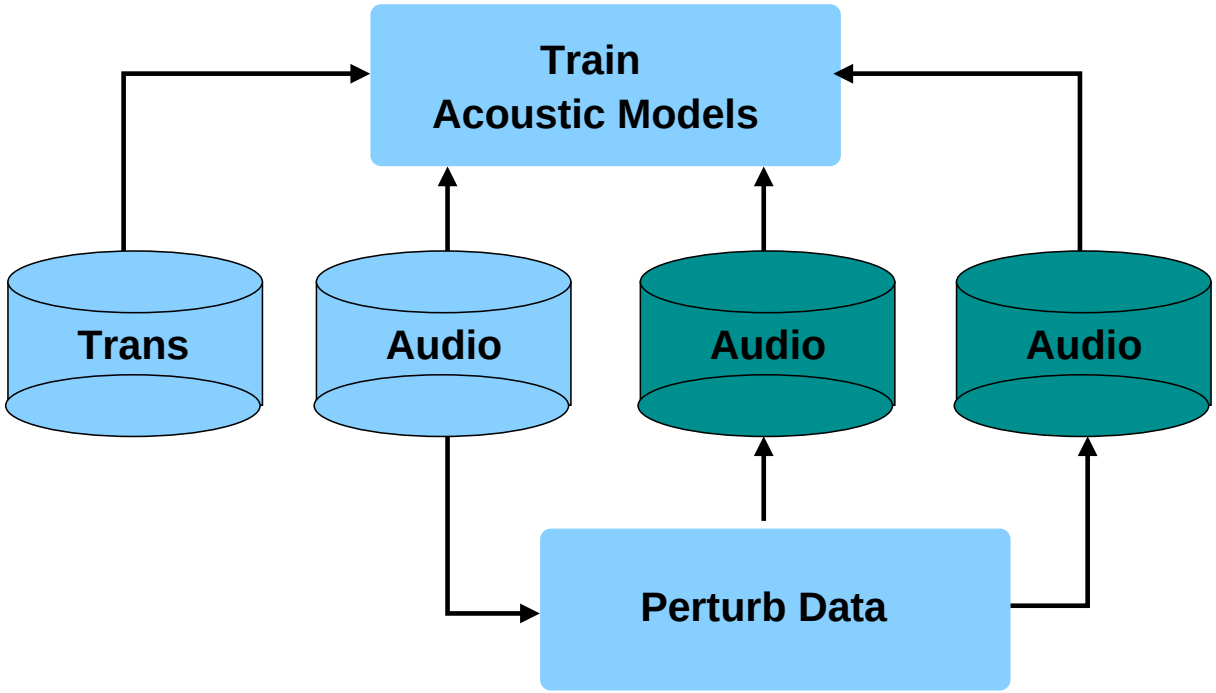
- LLP data is very limited - only 10 hours of transcribed audio
  - examined approaches to increase the quantity of transcribed data
- Scheme investigated:
  - semi-supervised - increase quantity of transcriptions
  - data augmentation - increase data given the transcriptions
  - multi-language - "borrow" data from other languages
- Also interested in parametric speech synthesis
  - generate as much data as you want!
  - not tried on the Babel data (yet)

Semi-Supervised Training



- Segment level selection of data to use
  - 50% of data selected - frame-weighted word confidences

Data Perturbation



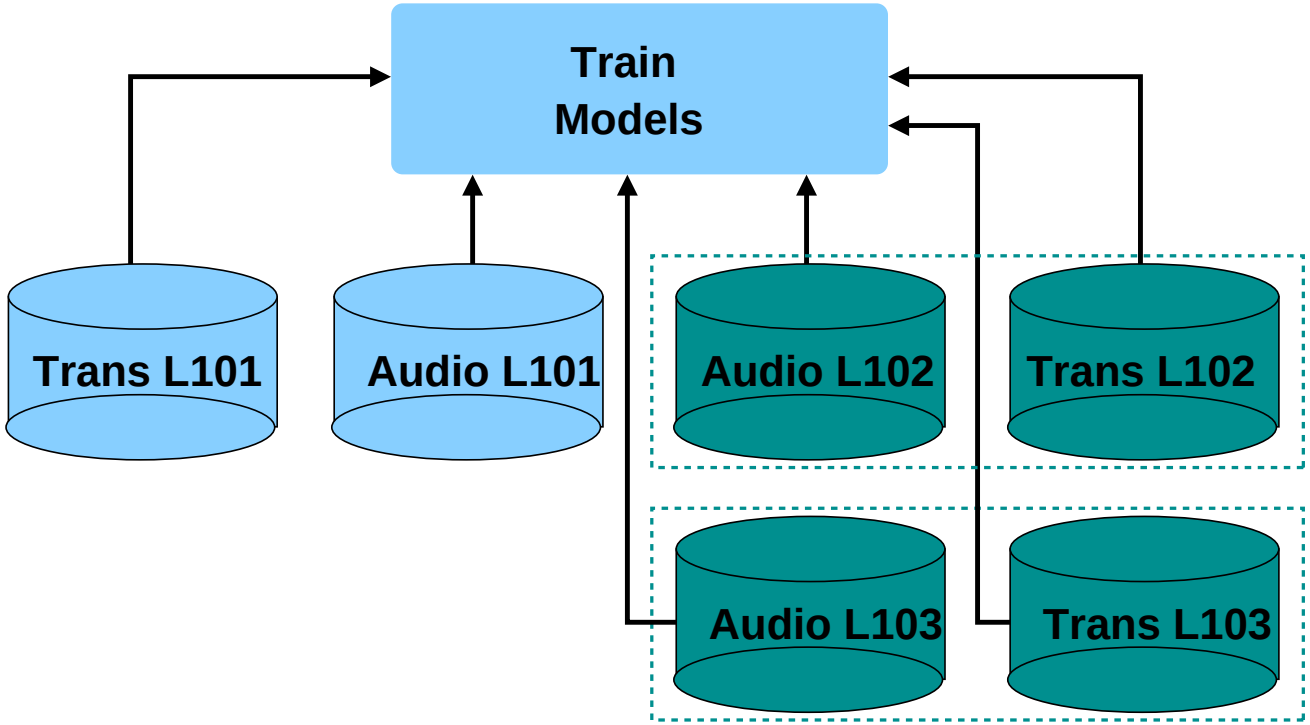
- Simplest form of perturbation - Vocal Tract Length Perturbation

Zulu Limited Language Pack Experiments

- Used Tandem DNN configuration
  - choice to augment BN features and or HMM parameters

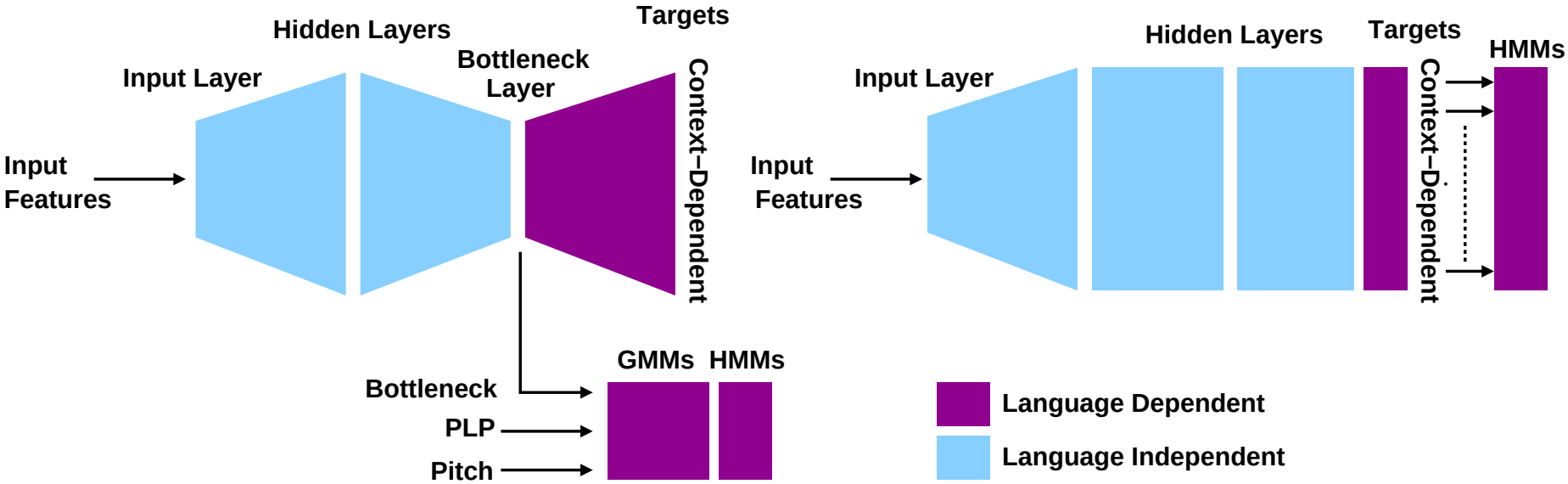
Data Augmentation HMM	Data Augmentation BN-MLP	TER (%)	MTWV Tot
—	—	78.4	0.1362
—	vtlp	77.1	0.1496
—	semi	77.7	0.1468
—	semi+vtlp	76.7	0.1446
semi	semi	76.9	0.1490
semi	semi+vtlp	76.1	0.1441
semi+vtlp	semi+vtlp	76.1	0.1454

Multi-Language Data



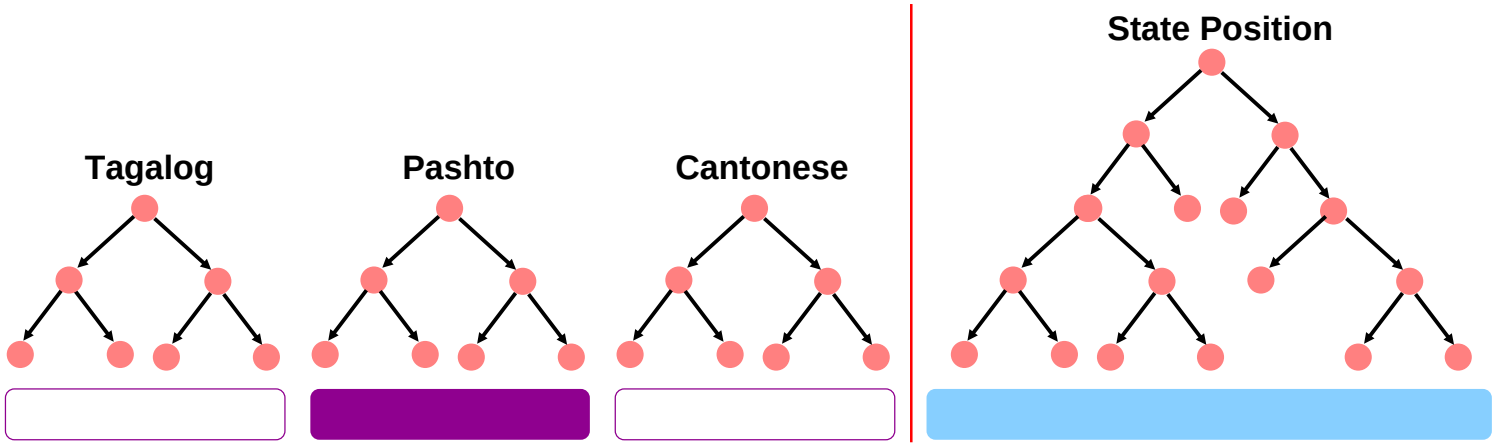
- Data from non-target language used to train model:
  - train complete acoustic model (see later)
  - train DNN to extract multi-language features

Multi-Language Deep Neural Networks



- Both Tandem and Hybrid systems can be used as feature extractors+classifier
  - aim to make feature extractor language independent
- Tandem: language-dependent GMM-based HMM and DNN targets
- Hybrid: language dependent soft-max classifier
  - classifier integrated with language-dependent HMM

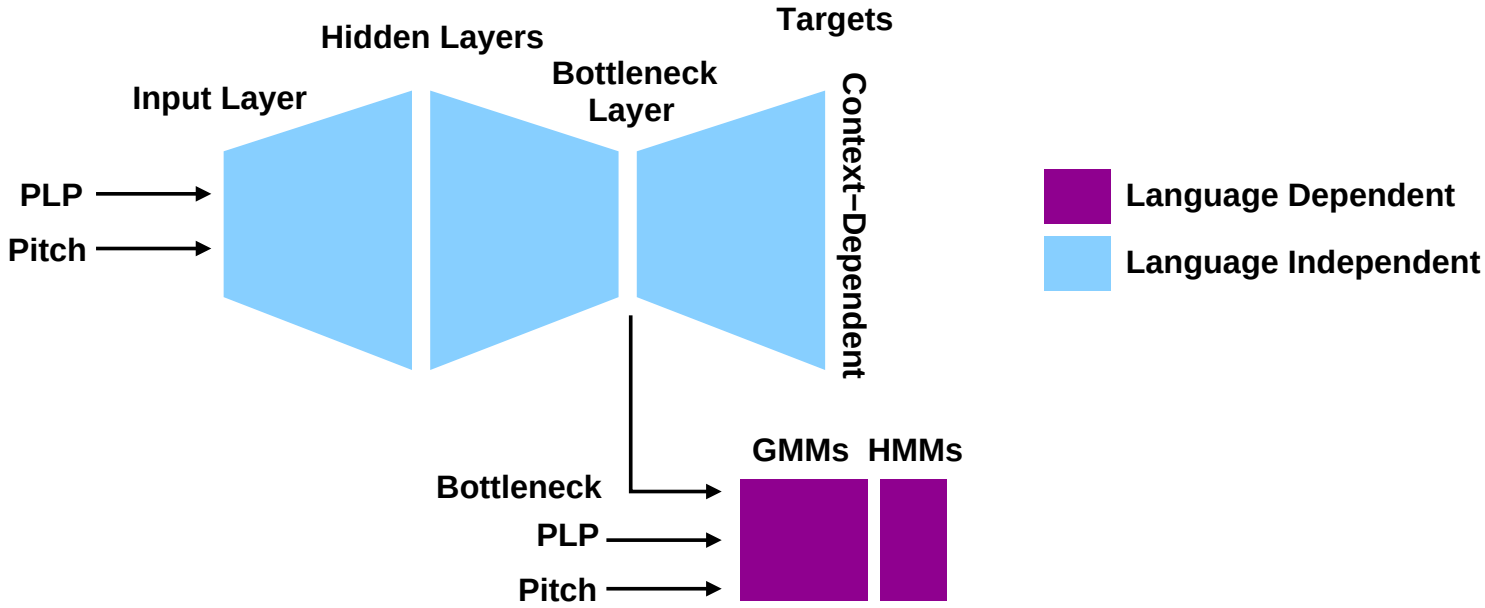
Language-Independent DNN Features



- Language-dependent context-dependent DNN targets
  - optimise MLP features to discriminate within languages
  - simple to add additional languages/tune to target language

- Language-independent context-dependent DNN targets
  - single decision tree (possible to ask language questions)
  - optimise features to discriminate all phones (unseen languages)

Language Independent Bottleneck System



- Language independent bottleneck features trained on seven languages
  - Cantonese, Assamese, Pashto, Turkish, Tagalog, Lao, Zulu
  - language-specific HMMs trained on target language
- BN features also evaluated on "held-out" languages
  - Bengali, Haitian Creole, Vietnamese

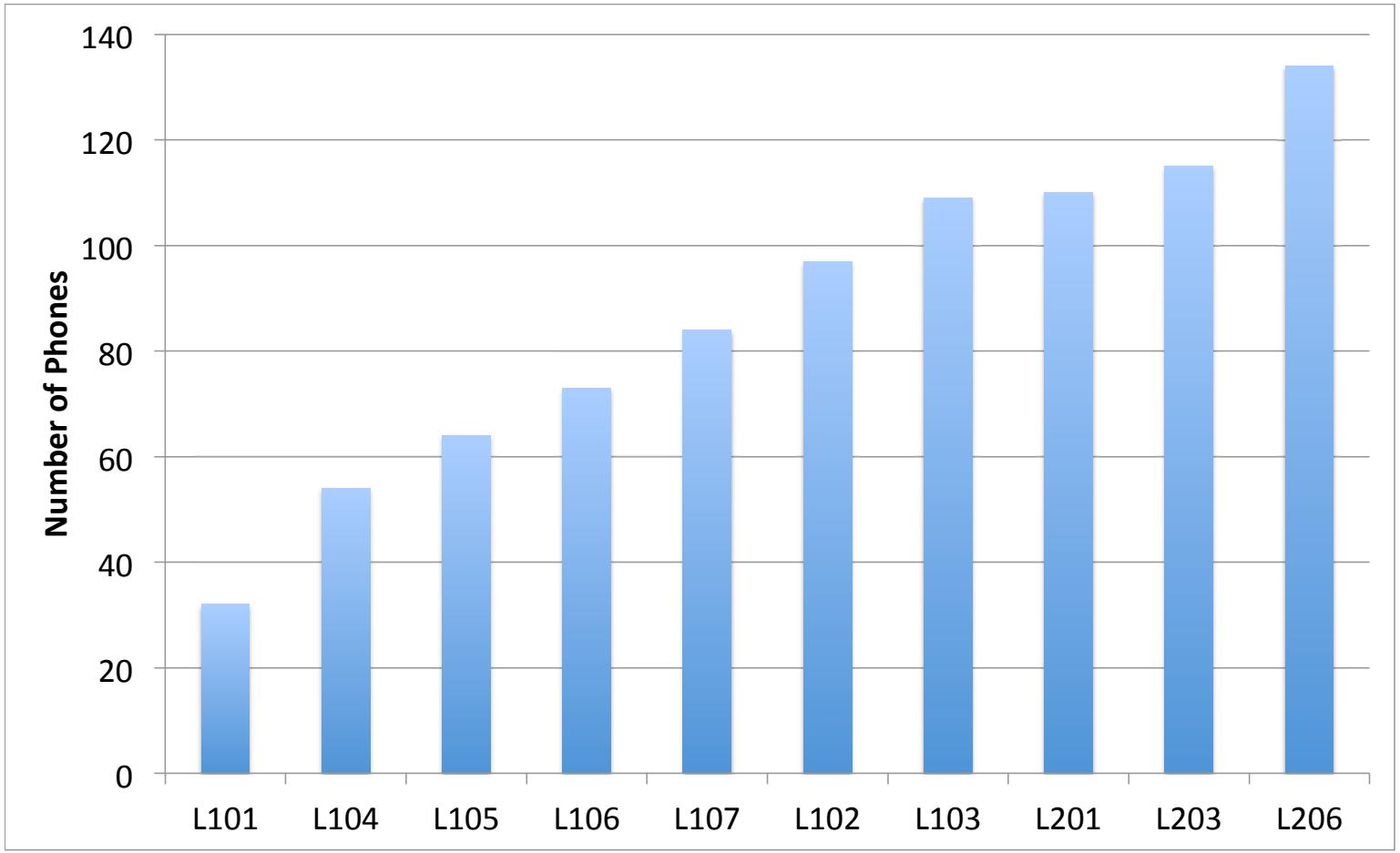
Language-Independent Features performance

Language	Id	BN MLP	TER (%)	MTWV Tot
Assamese†	102	UL	68.0	0.2132
		ML	66.4	0.2382
Zulu†	206	UL	75.8	0.1274
		ML	74.4	0.1396
Bengali	103	UL	68.6	0.2392
		ML	67.0	0.2551
Haitian Creole	201	UL	62.2	0.4054
		ML	61.1	0.4266
Vietnamese	107	UL	69.3	0.1851
		ML	68.2	0.1908

Language Independent Systems

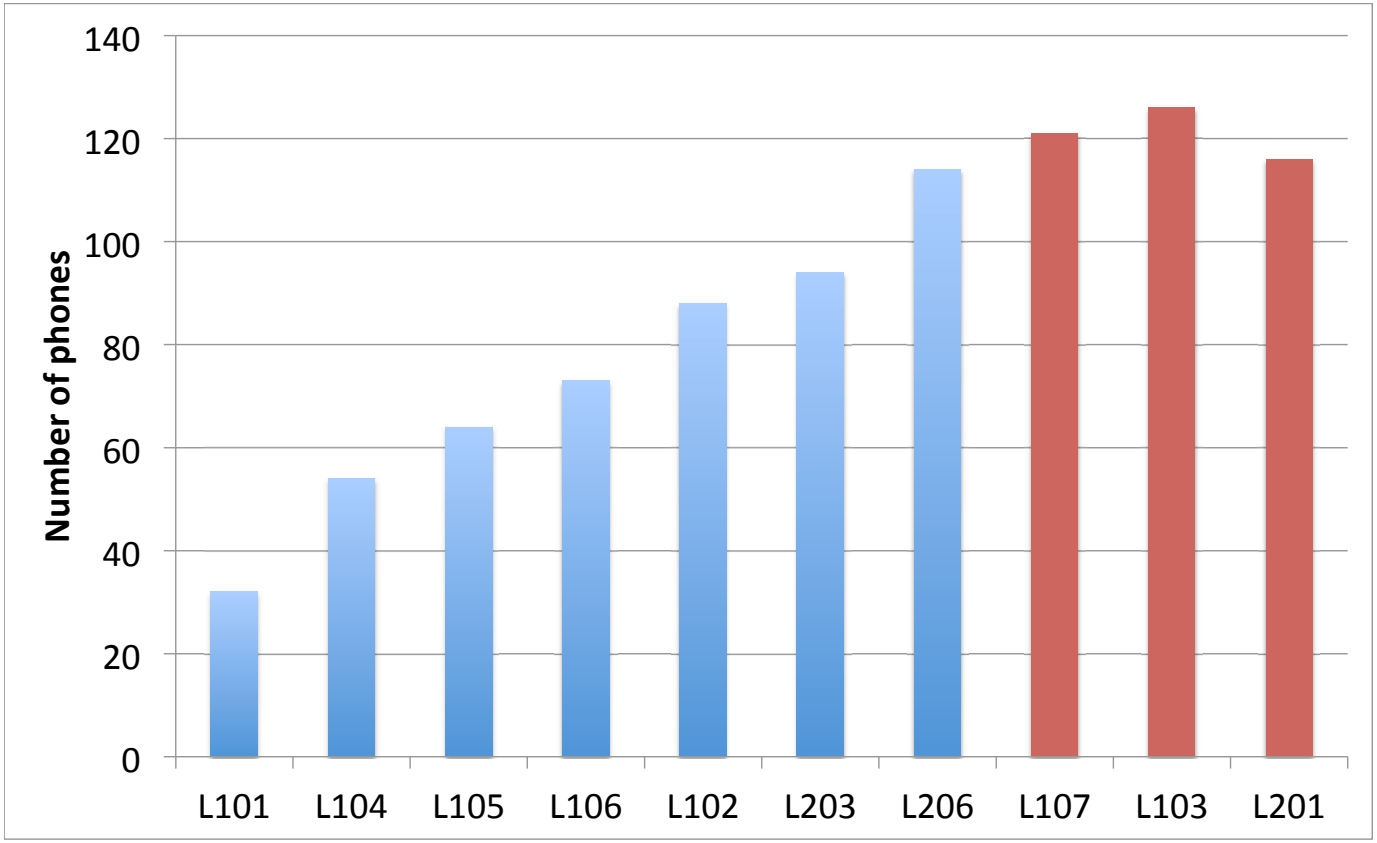
- So far assumed available data in target language
  - transcribed audio data
  - lexicon and phone set
  - language model training data
- Reduce overhead in deploying new language?
- Language Independent Acoustic Models
  - no acoustic training data available for target language
  - limited lexicon (limited language pack)
  - limited language model training data
- Bootstrap using Multi-Language system
  - target language acoustic training data without transcriptions

Phone Set Coverage



- CUED X-SAMPA attribute file has 215 entries (seen 62%)

Phone-Set Coverage - Experimental Configuration



- Vietnamese (L107) missing phones: 7
- Bengali (L103) missing phones: 12
- Haitian Creole (L201) missing phones: 2

Multi-language Lexical Entries

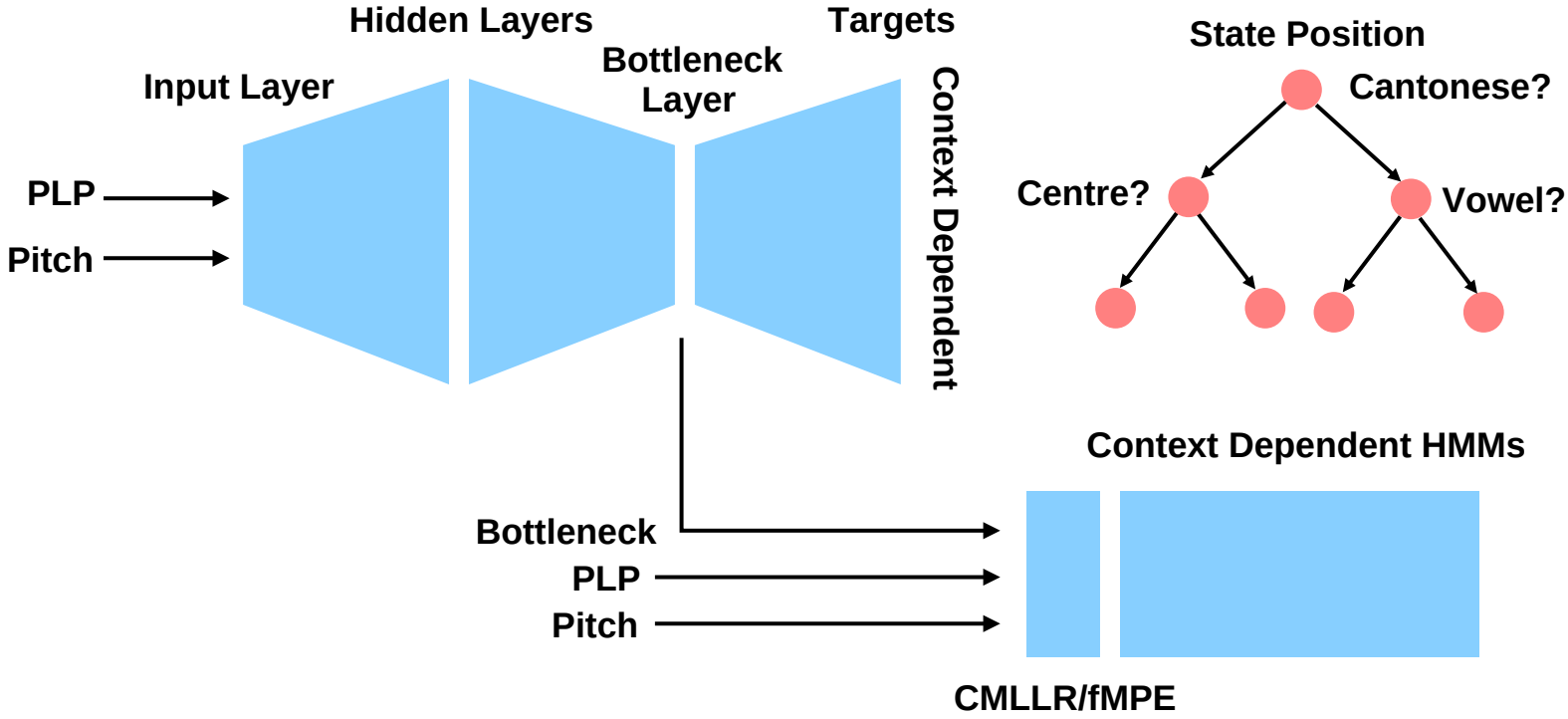
- Modifications to supplied ABH lexicon phone entries:
  - mapped diphthongs/triphthongs to individual phones
  - minor changes to map ABH to X-SAMPA labels

- ABH language-specific tone lexical labels - ignores attributes

Label	Level	Shape	Language Id		
			L101	L107	L203
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	—
43	low	rising	5	—	3

  - ask *level* and *shape* questions in decision tree

CUED Language Independent System

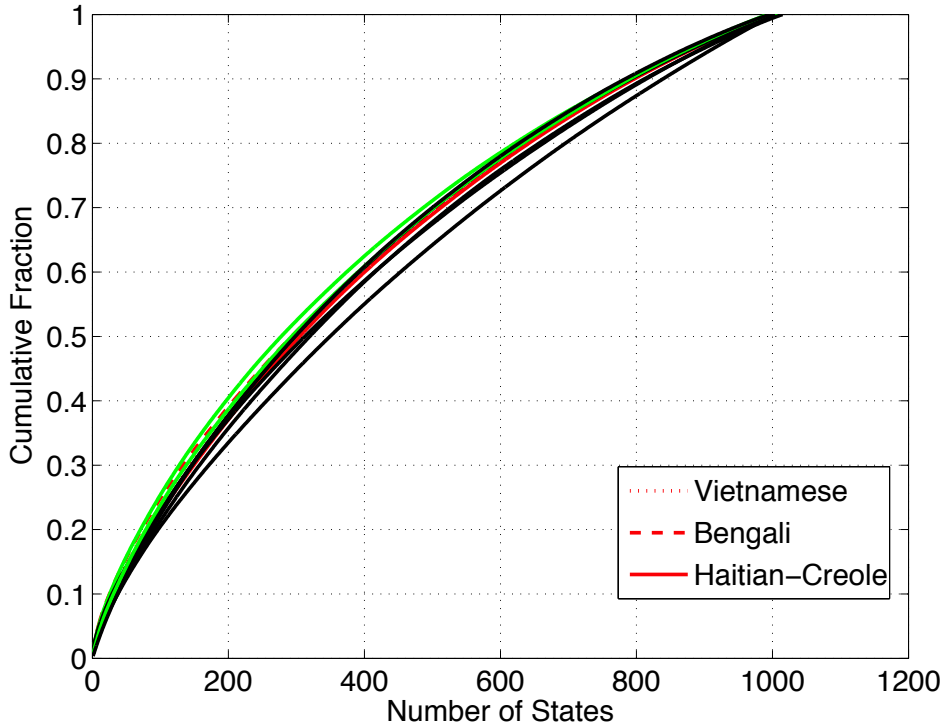


- Combine data from LLP from seven languages:
  - Cantonese, Pashto, Turkish, Tagalog, Assamese, Lao, Zulu
- Can be directly applied to any language (in theory ...)

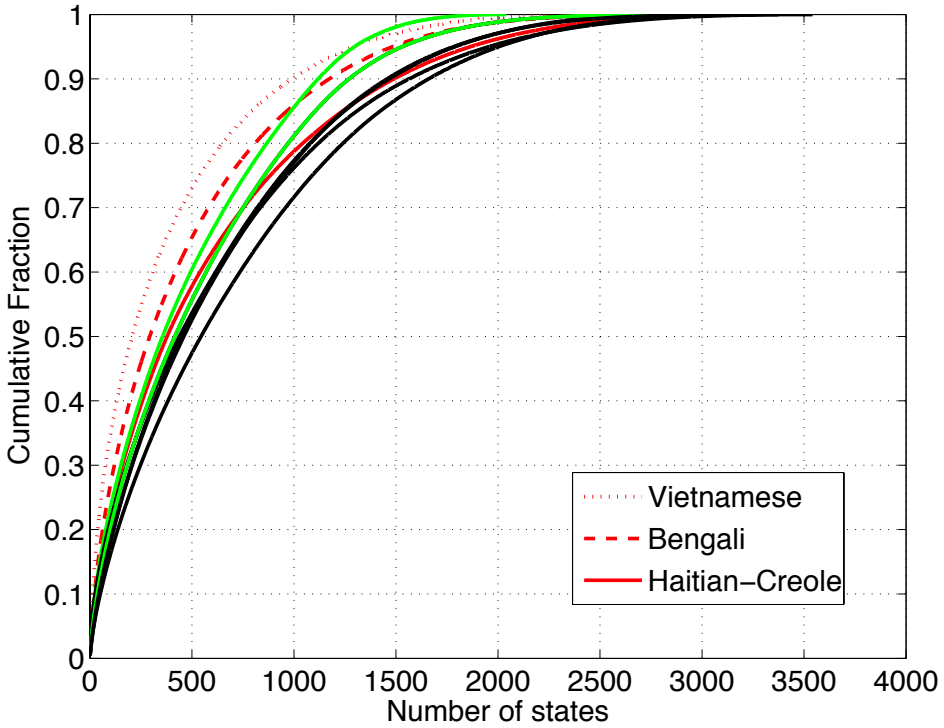
CUED Language Independent System

System		TER (%)	MTWV		
			IV	OOV	Tot
Haitian Creole (201)					
LD	fMPE	61.7	0.4673	0.2347	0.4317
LI	fMPE	77.2	0.2250	0.0966	0.2058
Bengali (103)					
LD	fMPE	68.5	0.3173	0.0987	0.2504
LI	fMPE	81.1	0.1929	0.0775	0.1573
Vietnamese (107)					
LD	fMPE	69.3	0.1962	0.1081	0.1851
LI	fMPE	87.6	0.0255	0.0268	0.0257

Analysis on Use of Decision Tree



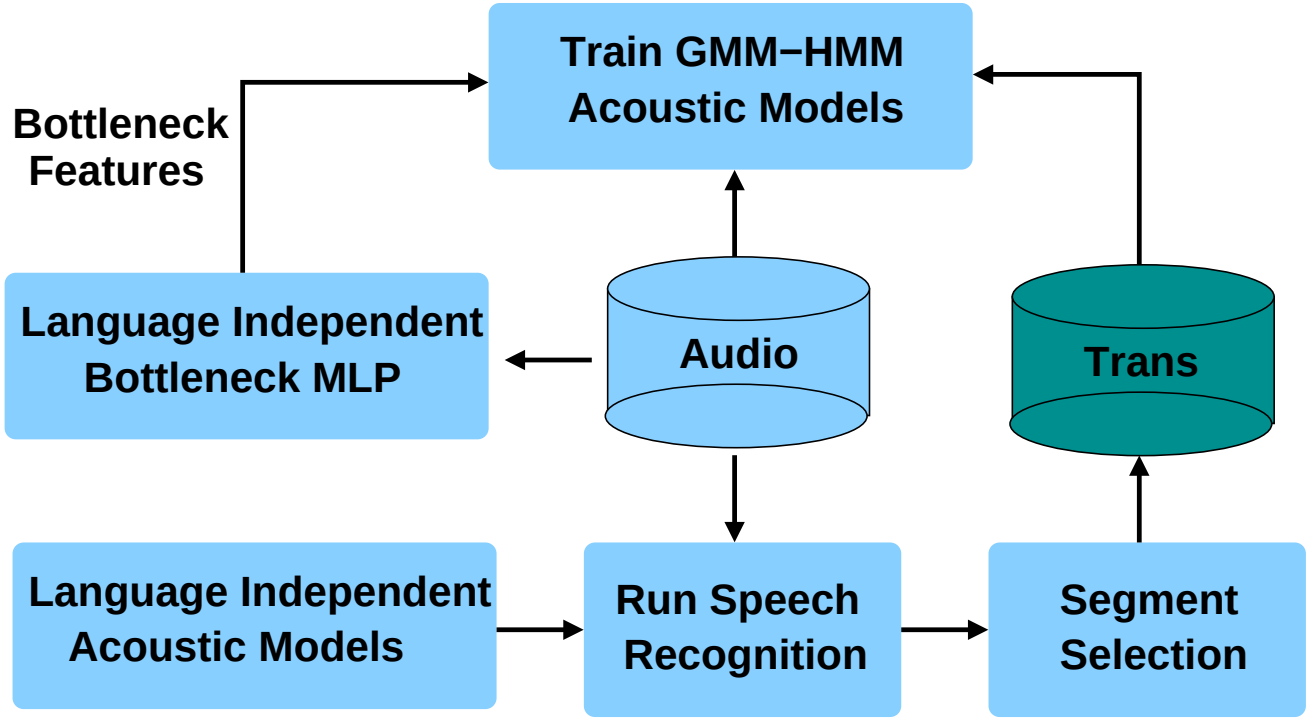
Language Dependent Tree



Language Independent Tree

- Sort state by occupancy and then accumulate
  - red indicates held-out languages (L107,L103,L201)
  - green indicates tonal training languages

Unsupervised Acoustic Model Training



- Segment level selection of data to use
  - approximately 20hours of data used

CUED Language Independent System

System		TER (%)	MTWV		
			IV	OOV	Tot
Haitian Creole (201)					
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 (103)					
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 (107)					
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

System Performance (Option Period 1 Languages)

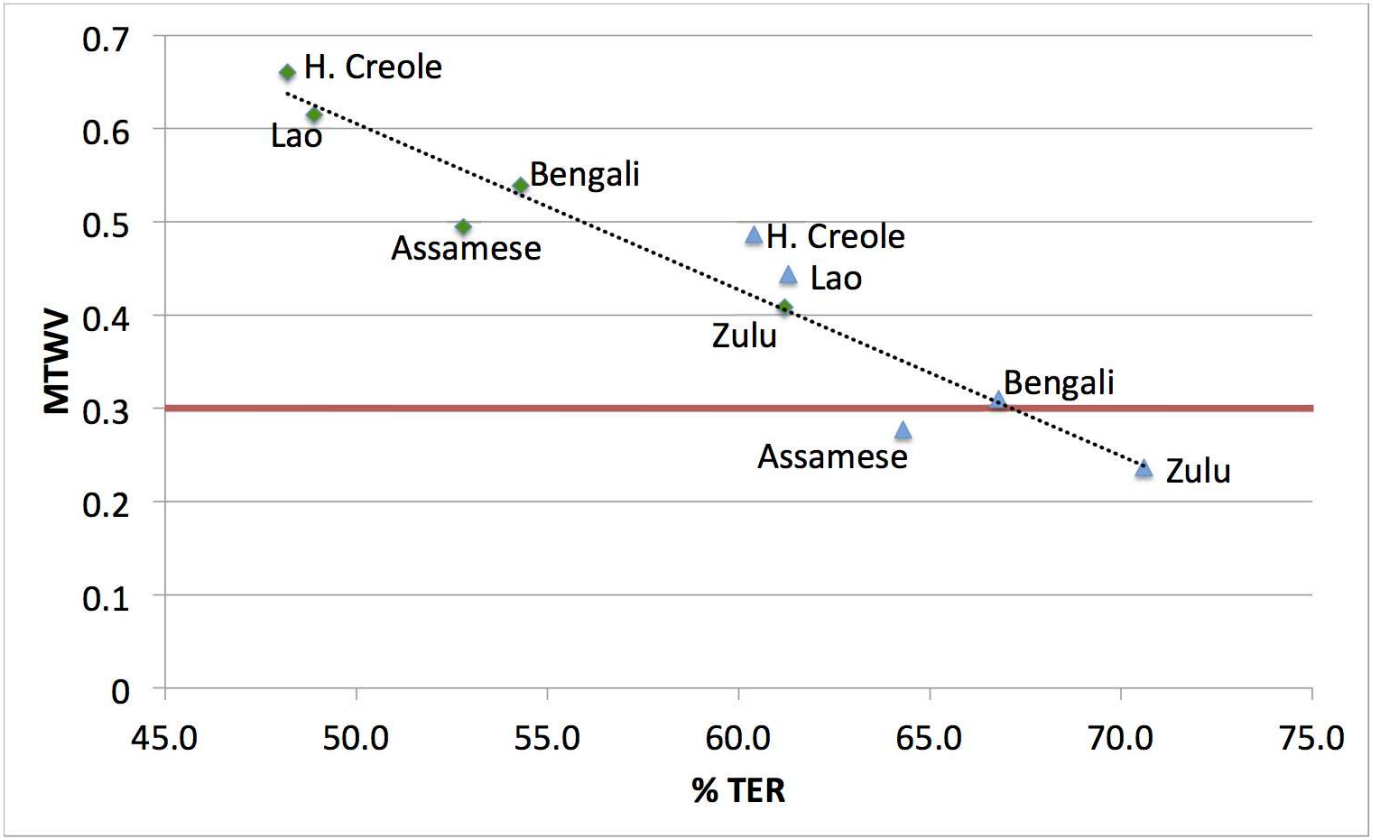
Tandem and Hybrid ASR Combination

Language	Id	LP	TER (%)		
			Tandem	Hybrid	CNC
Assamese	102	FLP	54.2	55.1	52.8
		LLP	65.1	67.8	64.3
Bengali	103	FLP	54.9	56.6	54.3
		LLP	67.0	69.5	66.8
Haitian Creole	201	FLP	48.7	50.3	48.2
		LLP	60.5	63.4	60.4
Lao	203	FLP	48.5	51.9	48.9
		LLP	61.2	65.8	61.3
Zulu	206	FLP	62.1	64.4	61.2
		LLP	71.5	74.1	70.6


Tandem and Hybrid KWS Combination

Language	Id	LP	MTWV		
			Tandem	Hybrid	Merge
Assamese	102	FLP	0.4660	0.4730	0.4946
		LLP	0.2569	0.2360	0.2771
Bengali	103	FLP	0.5151	0.5121	0.5388
		LLP	0.2992	0.2615	0.3100
Haitian Creole	201	FLP	0.6387	0.6329	0.6602
		LLP	0.4648	0.4336	0.4867
Lao	203	FLP	0.5951	0.5881	0.6149
		LLP	0.4262	0.3790	0.4439
Zulu	206	FLP	0.3770	0.3654	0.4084
		LLP	0.2287	0.1924	0.2366

Combined ASR/KWS Performance



Conclusions

- Constructing ASR/KWS system using limited data highly challenging
  - high word error rates greater than 50%
- Data augmentation yields significant gains
  - data perturbation (vocal tract length)
  - semi-supervised training
  - multi-lingual features
- Language Independent ("zero acoustic model resources")
  - current systems insufficiently language independent!
  - able to perform unsupervised acoustic model training

Acknowledgements

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.

Questions?

mjfg@eng.cam.ac.uk