Investigation of multilingual speech-to-text systems for use in spoken term detection

Kate Knill, Mark Gales,
CUED Babel Team (Anton, Austin, Chao, Phil, Shakti, Takuya),
Lorelei Babel Team

3 April 2014

Cambridge University Engineering Department

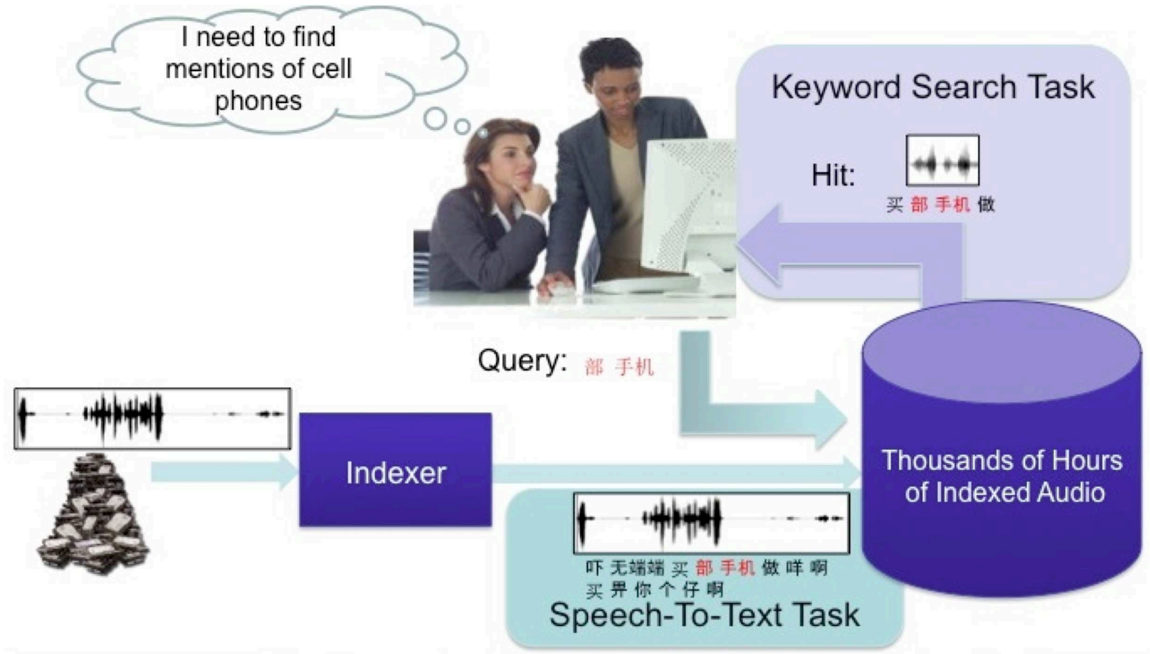
Overview

- Motivation
- IARPA Babel program
- Language Dependent speech-to-text systems
- Multi-Language systems
- Language Independent systems
- Conclusions

Motivation

- Development of speech processing systems for low/zero resource languages
  - Challenging!
  - Increase resources by using data from multiple languages
  - Enable bootstrapping when no transcribed audio data available
- Potential benefits
  - Faster and cheaper to develop
  - Better non-native performance
  - Help understanding of commonalities and differences across languages

IARPA Babel Program

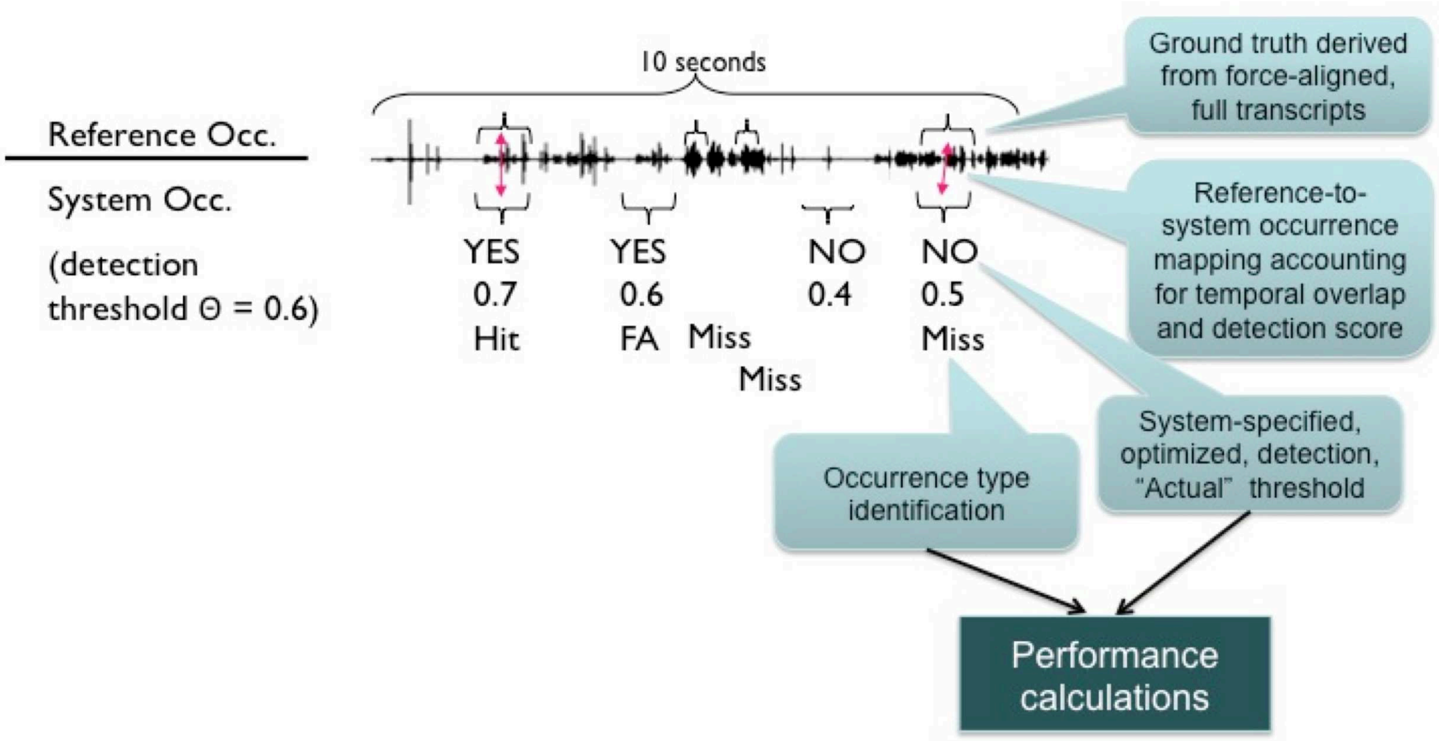


- Goal - rapidly develop spoken term detection in new languages
  - Broad set of languages with varying phonotactics, phonological, tonal, morphological and syntactic characteristics
  - Speech recorded in variety of conditions
  - Limited amounts of transcription

IARPA Babel Program Specifications

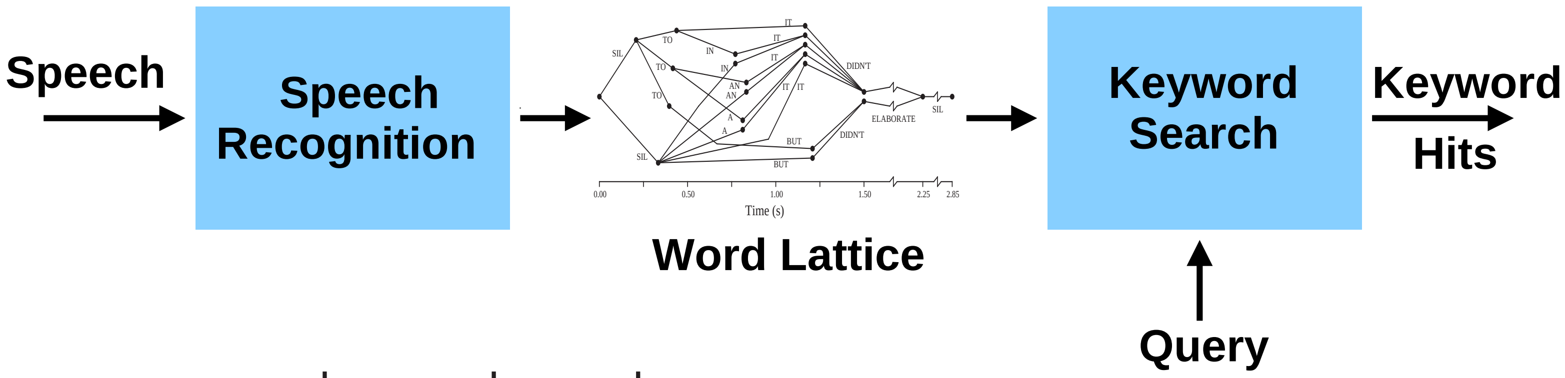
- Language Packs
  - Conversational and scripted telephone data (plus other channels)
  - Full: 60-80 hours transcribed speech (plus untranscribed speech)
  - Limited: 10 hours transcribed speech
  - 10 hour Development and Evaluation sets
  - Lexicon covering training vocabulary
  - X-SAMPA phone set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



- Term Weighted Value (TWV) - official metric
  - $TWV(\theta) = 1 - [P_{Miss}(\theta) + \beta P_{FA}(\theta)]$
- 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
  - Out-of-vocabulary (OOV) terms searched at phone level
  - Phone confusability matrix used to boost OOV performance
  - Normalised posterior probabilities using “sum-to-one”

- Scored using Maximum Term Weighted Value (MTWV)

IARPA Babel releases

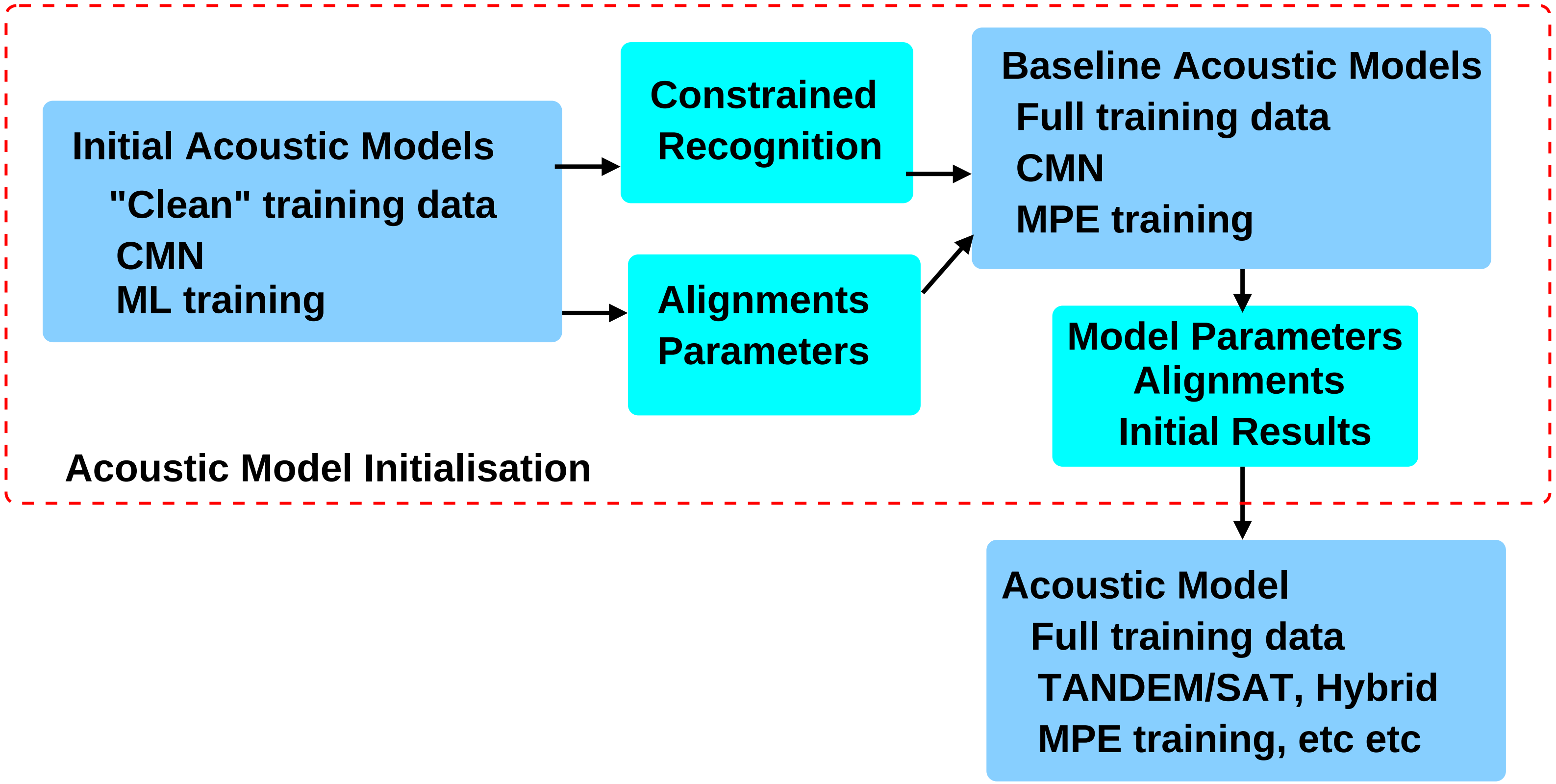
This work uses the IARPA Babel Program language collection releases:

Language	Release
Cantonese	IARPA-babel101-v0.4c
Pashto	IARPA-babel104b-v0.4aY
Turkish	IARPA-babel105b-v0.4
Tagalog	IARPA-babel106-v0.2f
Vietnamese	IARPA-babel107b-v0.7
Assamese	IARPA-babel102b-v0.5a
Bengali	IARPA-babel103b-v0.4b
Haitian Creole	IARPA-babel201b-v0.2b
Lao	IARPA-babel203b-v3.1a
Zulu	IARPA-babel206b-v0.1d

Speech-to-Text Systems

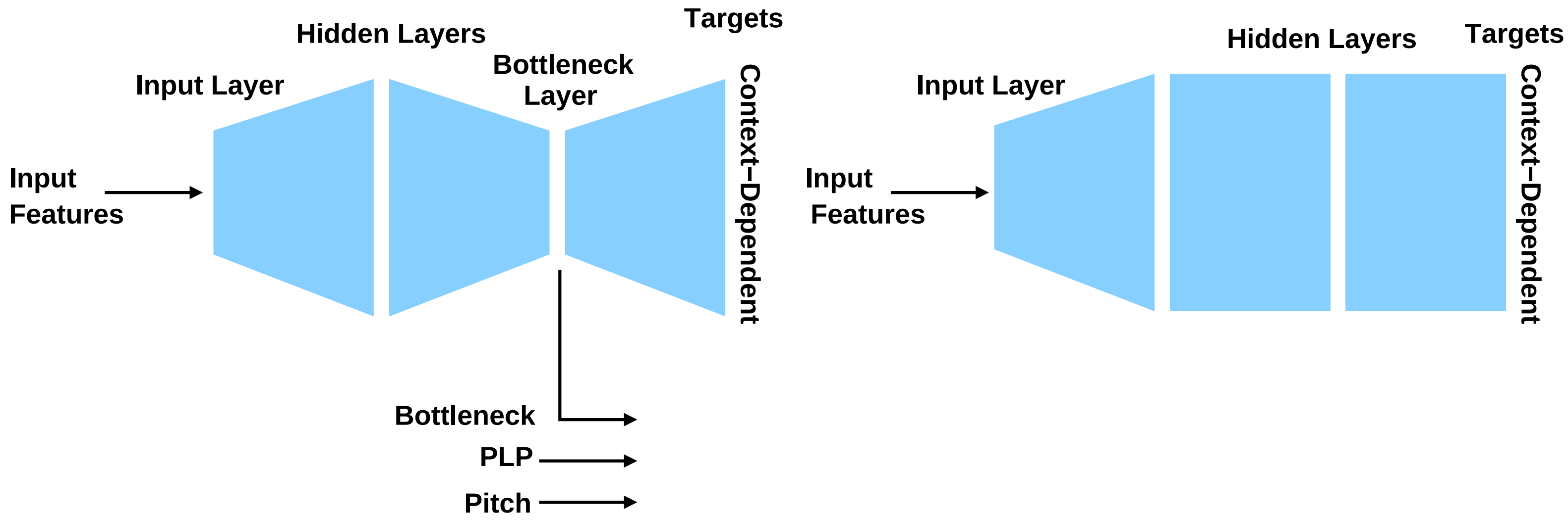
- Categorise in a similar fashion to speaker
- Language Dependent
  - Common approach taken across languages
- Multi-Language
  - Shared training data across closed set of languages
- Language Independent
  - Apply to languages outside training set

Language Dependent STT - General Training Procedure



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

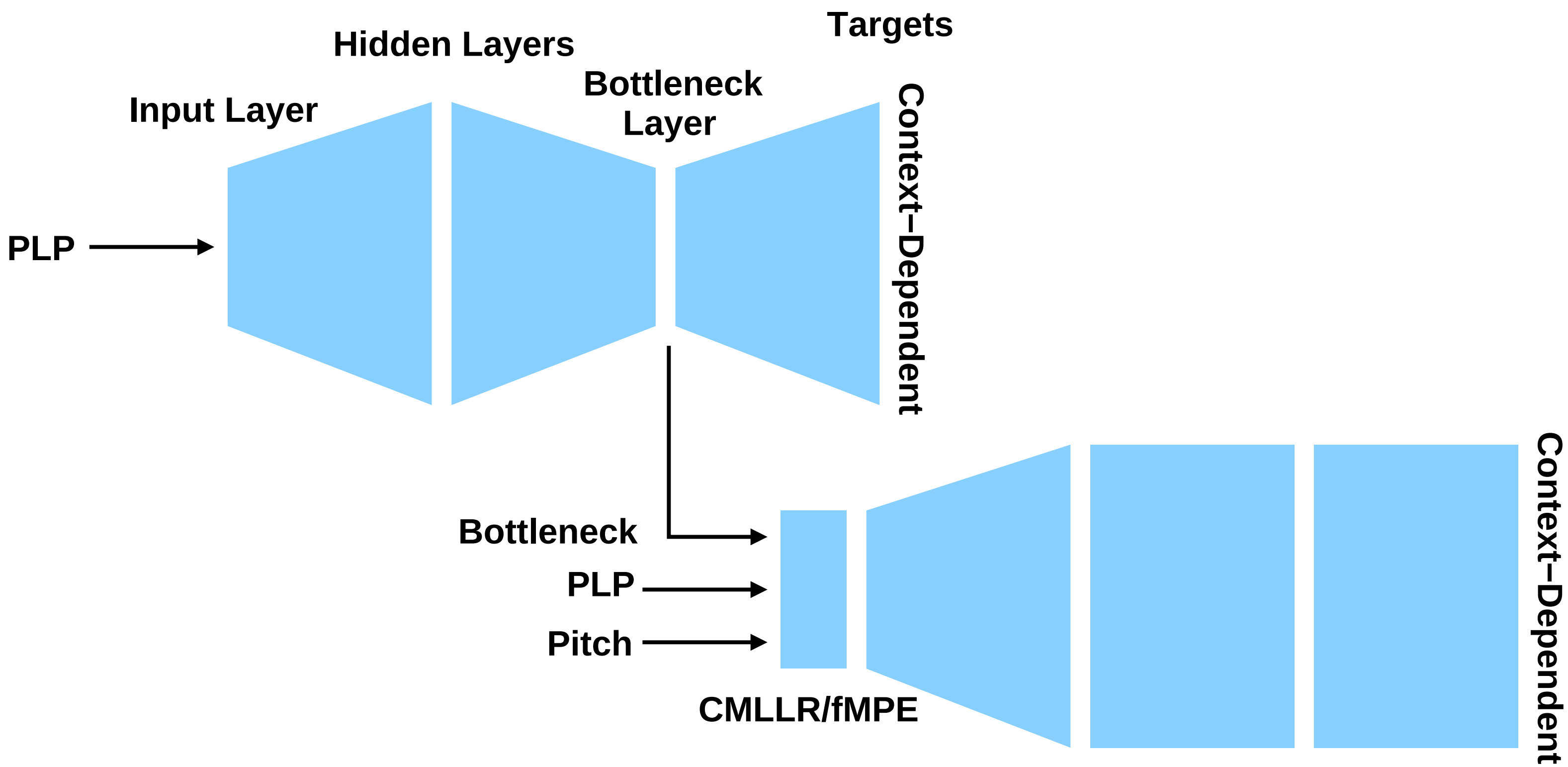
Use of (Deep) Neural Networks



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

- Possible to combine approaches uses stacking

Stacked Hybrid System

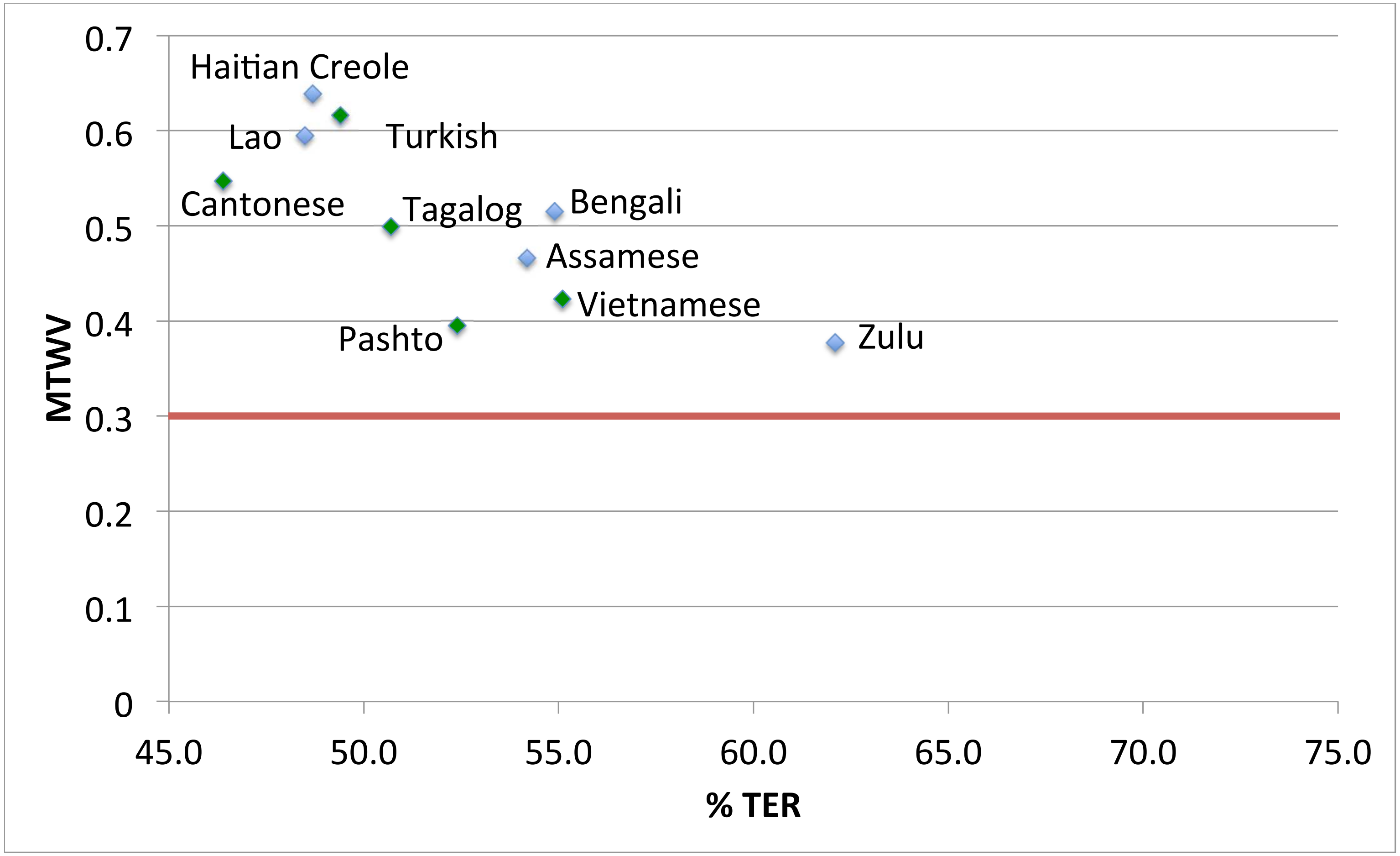


- Stacked approach used for Hybrid system development
  - configuration allows re-use of existing Tandem systems
  - use of bottleneck features improves STT (0.5% abs)
  - same context dependent labels as Tandem system

Baseline CUED STT System Configuration

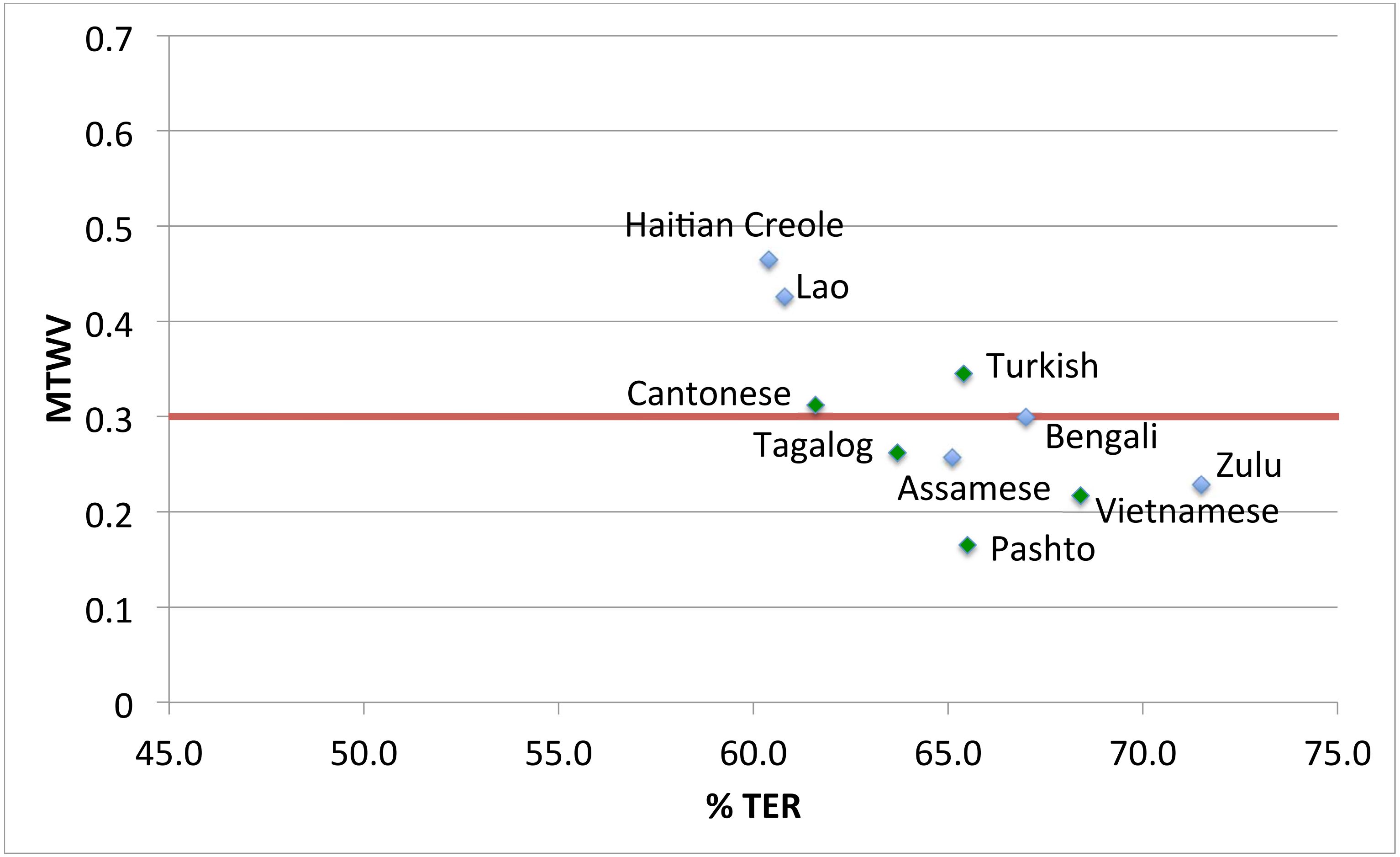
- General Configuration (both FLP and LLP)
  - ABH dictionary - word boundary/tone markers for dec. tree
  - decision-tree state-clustered cross-word triphones
  - PLP $+\Delta + \Delta^2 + \Delta^3$ +HLDA, pitch $+\Delta + \Delta^2$, (39+3)
  - Bottleneck features + SemiTied transform (26)
  - speaker adaptive training at the conversation side level
  - fMPE features and MPE acoustic model training
  - word-level bigram LM trained on acoustic data transcriptions
  - optional bigram class-based and neural network LMs
- Full Language Pack Configuration
  - 4-hidden layer plus bottleneck layer for bottleneck MLP
  - 6000 context dependent states
- Limited Language Pack Configuration
  - 3-hidden layer plus bottleneck layer for bottleneck MLP
  - 1000 context dependent states

CUED STT/MTWV Performance: Full Language Packs



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

CUED STT/MTWV Performance: Limited Language Packs



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

Tandem/Hybrid Performance

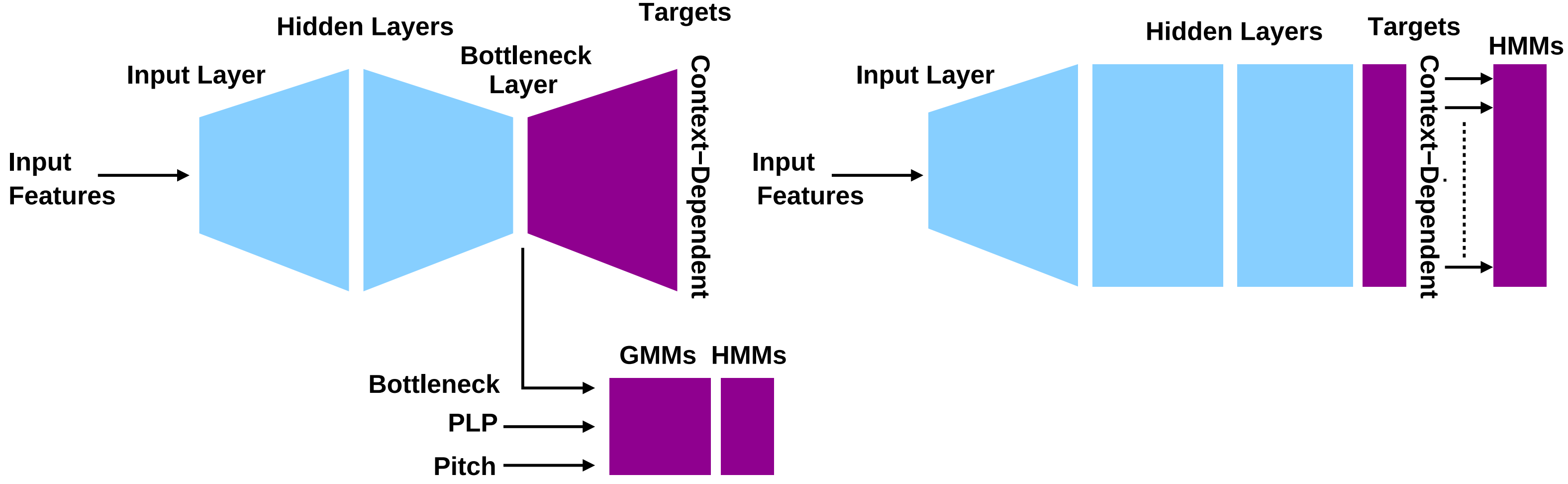
Language	System	TER (%)	MTWV
Vietnamese	Tandem	55.1	0.423
	Hybrid	54.4	0.418
Cantonese	Tandem	46.4	0.547
	Hybrid	46.9	0.542

- Hybrid currently trained using the cross-entropy criterion
- Hybrid OOV KWS sensitive to interaction acoustic/language models
  - "Zeroing" language model for OOV search yields gains
  - Also helps Tandem system
- Tandem and Hybrid systems complementary for STT and MTWV

Multi-Language Systems

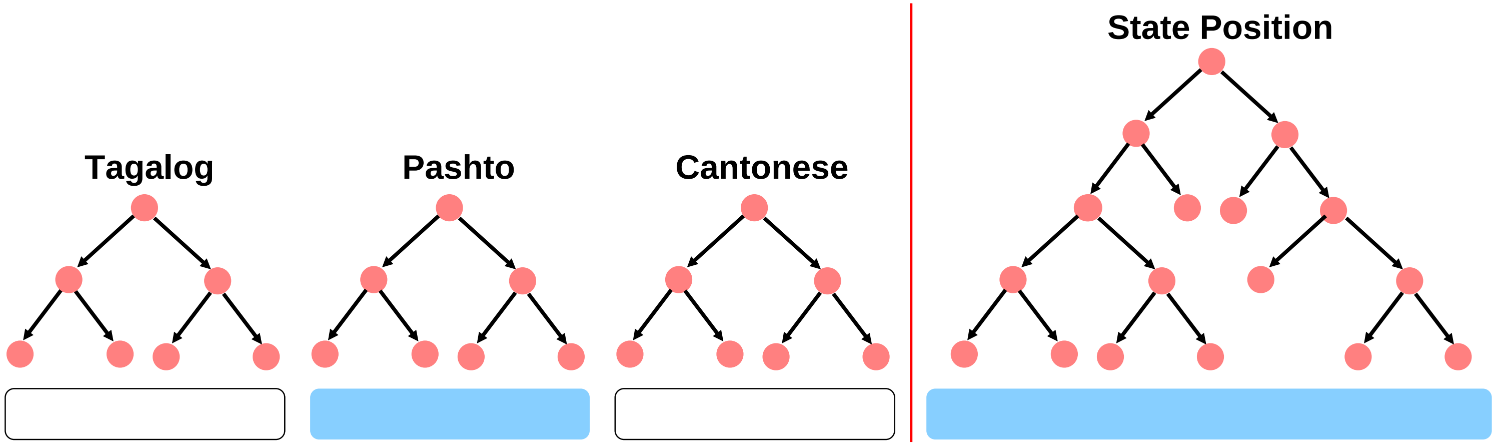
- Limited language packs - 10 hours of data
  - Limits complexity of AMs and DNN features
- To increase resources - combine training data across languages
  - CUED - LLPs, Aachen - FLPs
- Can use multi-language data in two modes:
  - Multilingual feature extraction
  - Multilingual classifiers

Multi-Language Deep Neural Networks



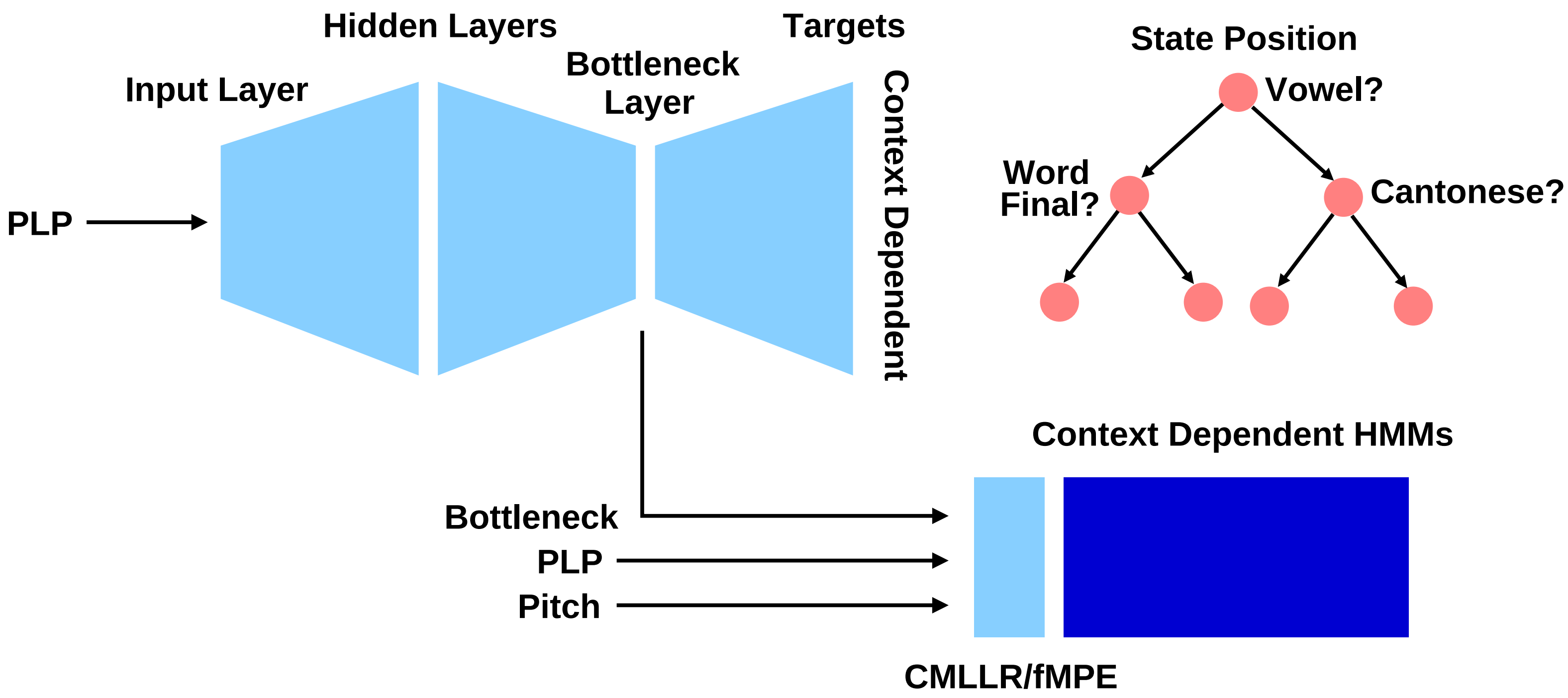
- NNs in Tandem and Hybrid act as both feature extractors and classifiers
- Can make multi-language feature extractors and/or classifiers
  - Standard option is to make multi-language feature extractor
  - Need to consider the nature of the CD targets

MLP Context Dependent Targets



- Language-specific targets (Aachen)
  - decision trees associated with targets language-specific
  - optimise MLP features to discriminate within languages
  - simple to add additional languages/tune to target language
- Global targets (Cambridge)
  - single decision tree (possible to ask language questions)
  - optimise features to discriminate all phones
  - supports unseen languages

CUED Single Multi-Language System



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

Multi-Language Features Performance

- Tandem-SAT-fMPE, Bigram LM

Language	Id	BN MLP	TER (%)	MTWV IV	MTWV OOV	MTWV Tot
Assamese	102	UL	67.7	0.2703	0.0633	0.2132
		ML	66.2	0.2996	0.0789	0.2382
Zulu	206	UL	75.1	0.2400	0.0220	0.1069
		ML	73.9	0.2521	0.0240	0.1136

- Acoustic model HMM trained on target language
  - UL configuration (only trained on target language)
- Gains from using multilingual MLP features (ML) over UL
- Further gains from using FLP training data - Aachen

Multi-Language Systems Performance

- Tandem-SAT, Bigram LM, UL trained on target language

Language	Id	AM HMM	BN MLP	TER (%)	MTWV IV	MTWV OOV	MTWV Tot
Assamese	102	UL	UL	68.8	0.2544	0.0634	0.2012
		UL	ML	66.7	0.2956	0.0681	0.2325
		ML	ML	67.9	0.2733	0.0584	0.2137
		ML-LQ	ML	66.8	0.2948	0.0732	0.2335
Zulu	206	UL	UL	76.5	0.2313	0.0205	0.1024
		UL	ML	73.8	0.2698	0.0211	0.1180
		ML	ML	74.4	0.2425	0.0186	0.1061
		ML-LQ	ML	73.8	0.2573	0.0161	0.1101

- Multilingual BN features (ML) always helped
- ML-LQ - language questions used in AM decision trees
  - Raised multilingual AM HMM to UL level

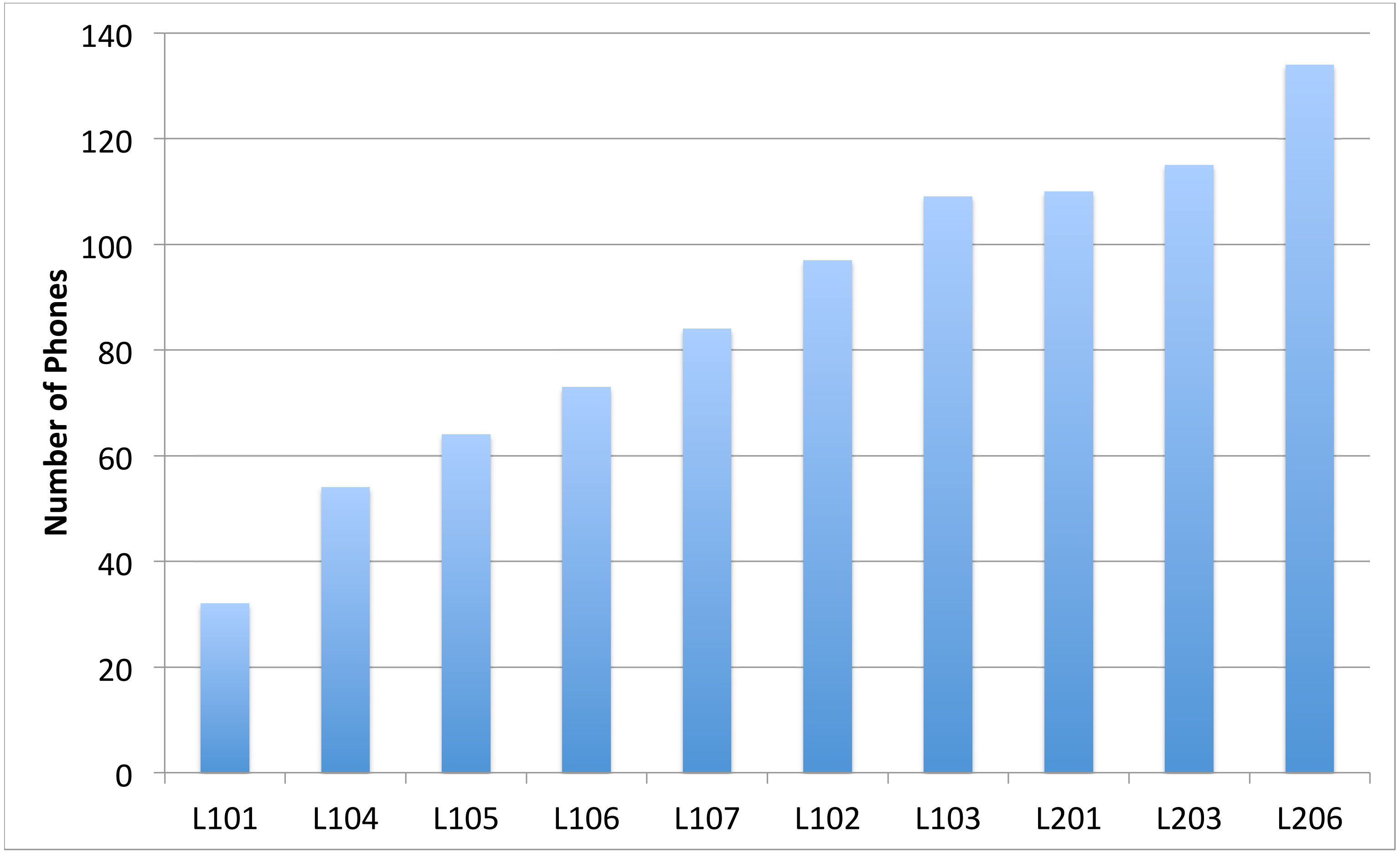
Language Independent Systems

- 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
- Bootstrap using Multi-Language system
  - Target language acoustic training data without transcriptions

Language Independent System Requirements

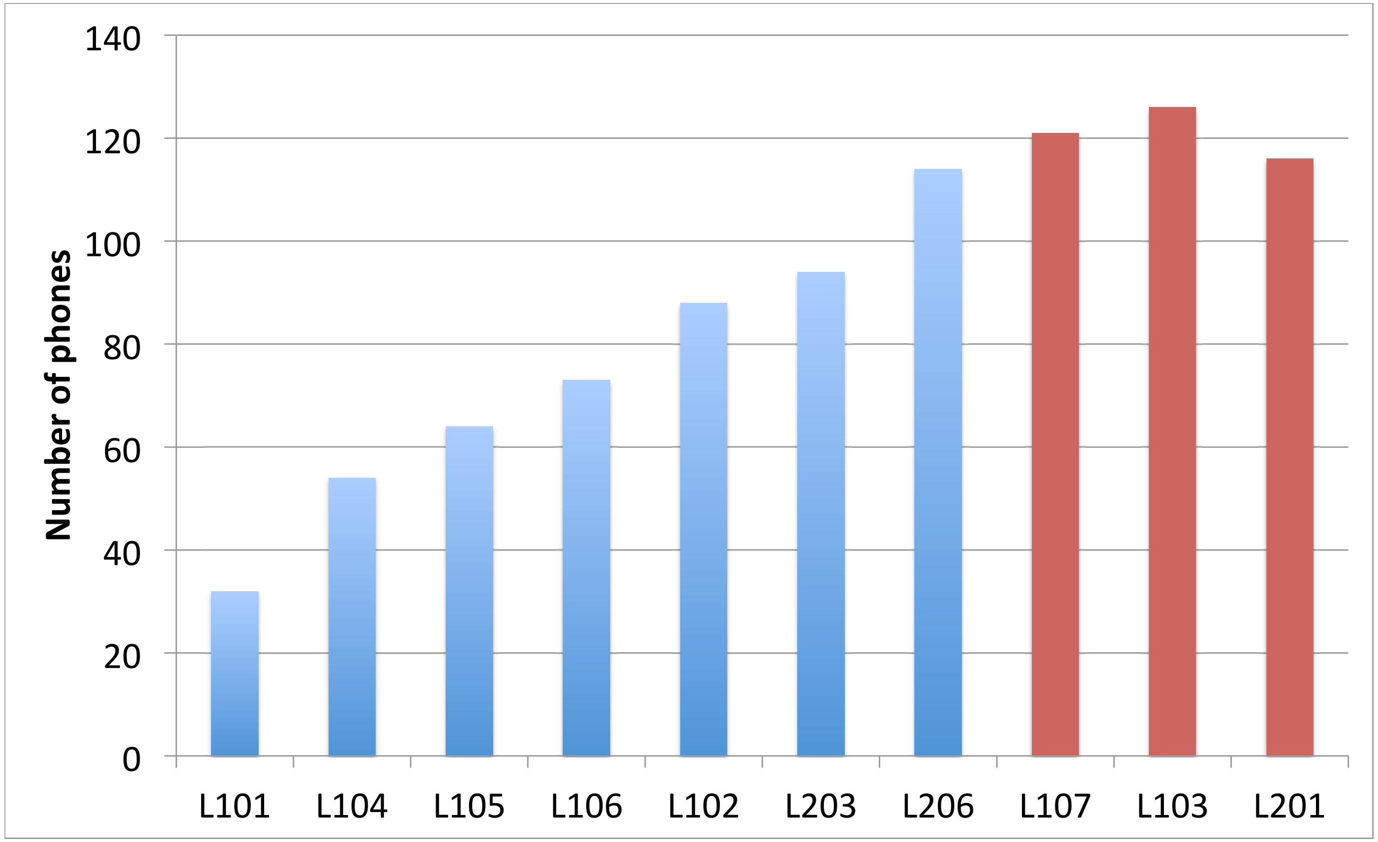
- Access to (limited) lexicon and language modelling data
- Phones are consistent across languages ...
  - requires good phone-set coverage
  - requires consistent phone labelling/attributes
  - use phone attributes to handle missing phones

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

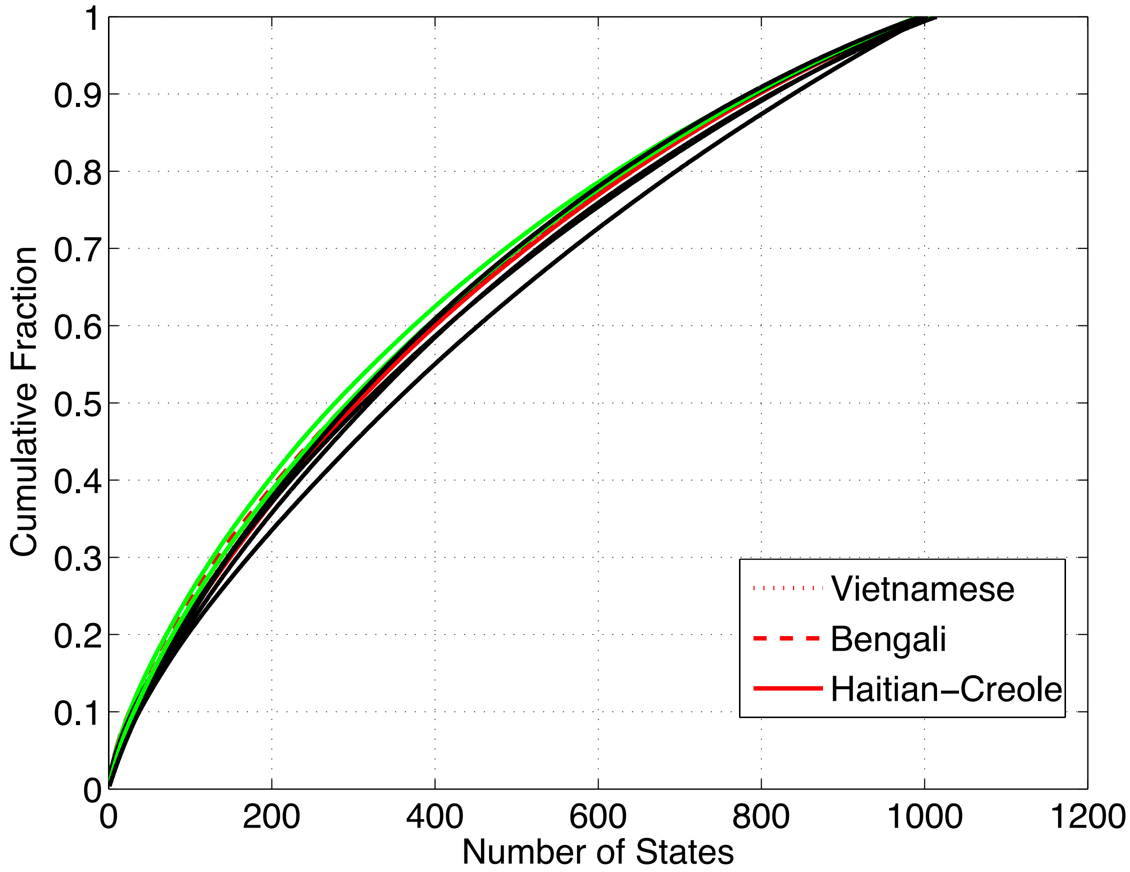
Language-Independent Performance

- Tandem-SAT, Bigram LM, UL trained on target language

Language	Id	AM HMM	BN MLP	TER (%)	MTWV Tot
Bengali	103	UL	UL	69.1	0.2106
		UL	ML	67.8	0.2290
		ML	ML	83.2	0.1172
Haitian-Creole	201	UL	UL	63.1	0.4035
		UL	ML	62.2	0.4205
		ML	ML	78.6	0.1943

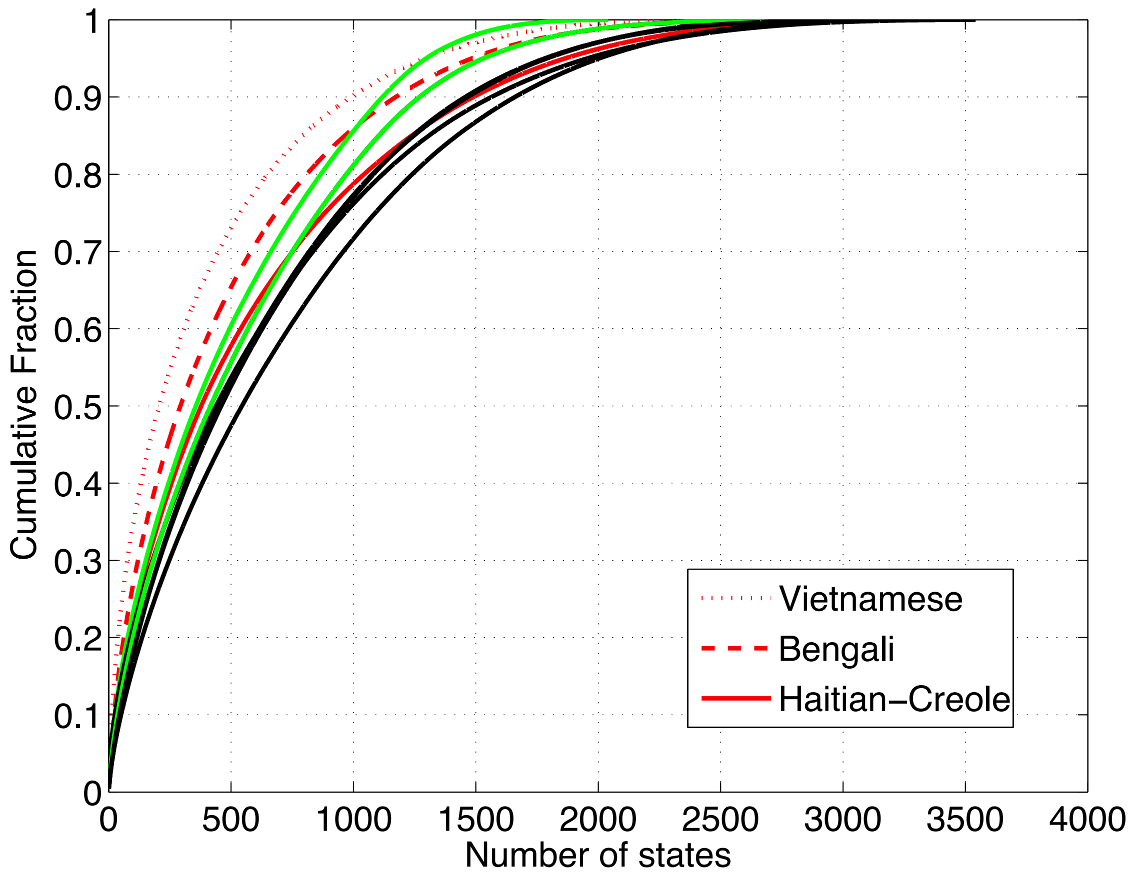
- ML bottleneck features yielded performance gain (UL/ML)
  - similar observation for Vietnamese
  - need to contrast with language-specific targets
- Baseline language-independent system performed poorly
  - Vietnamese even worse (!): TER 88.3%, MTWV 0.0171

Analysis on Use of Unilingual Trees



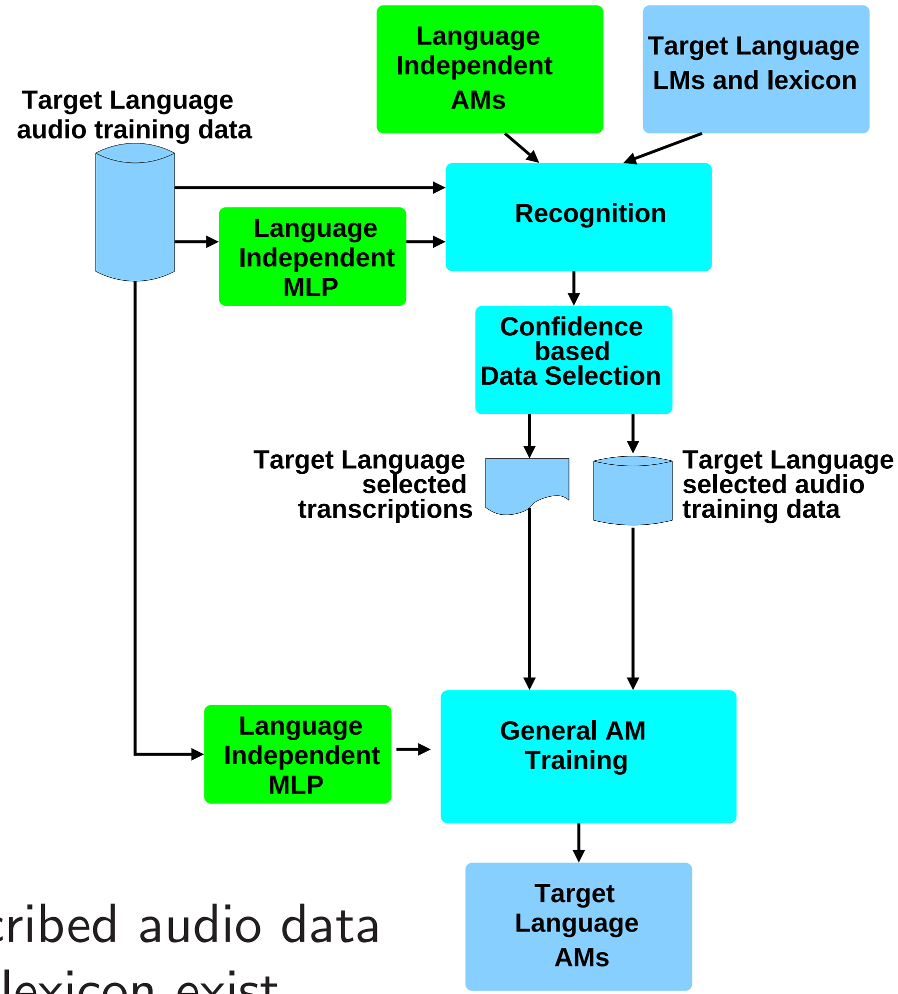
- red indicates held-out languages (L107,L103,L201)
- green indicates tonal training languages

Analysis on Use of Multilingual Tree



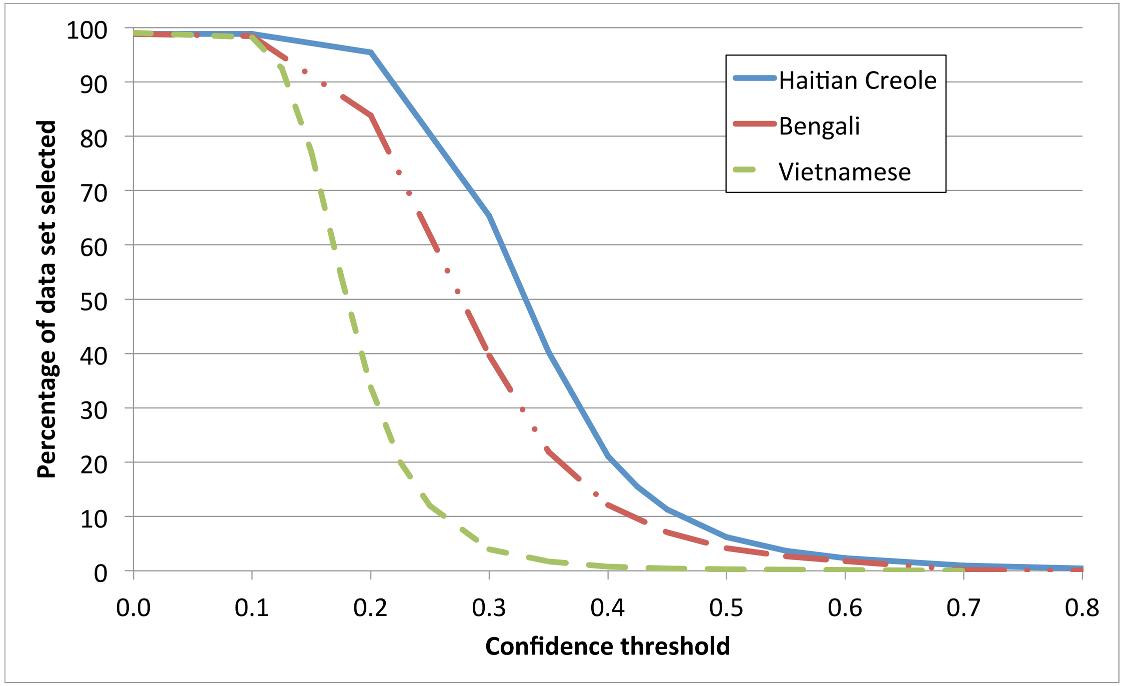
- red indicates held-out languages (L107,L103,L201)
- green indicates tonal training languages

Bootstrapping with Multi-Language Systems



- Assumptions
  - Set of untranscribed audio data
  - Phone set and lexicon exist
  - Text data exists to generate language model

Unsupervised Data Selection



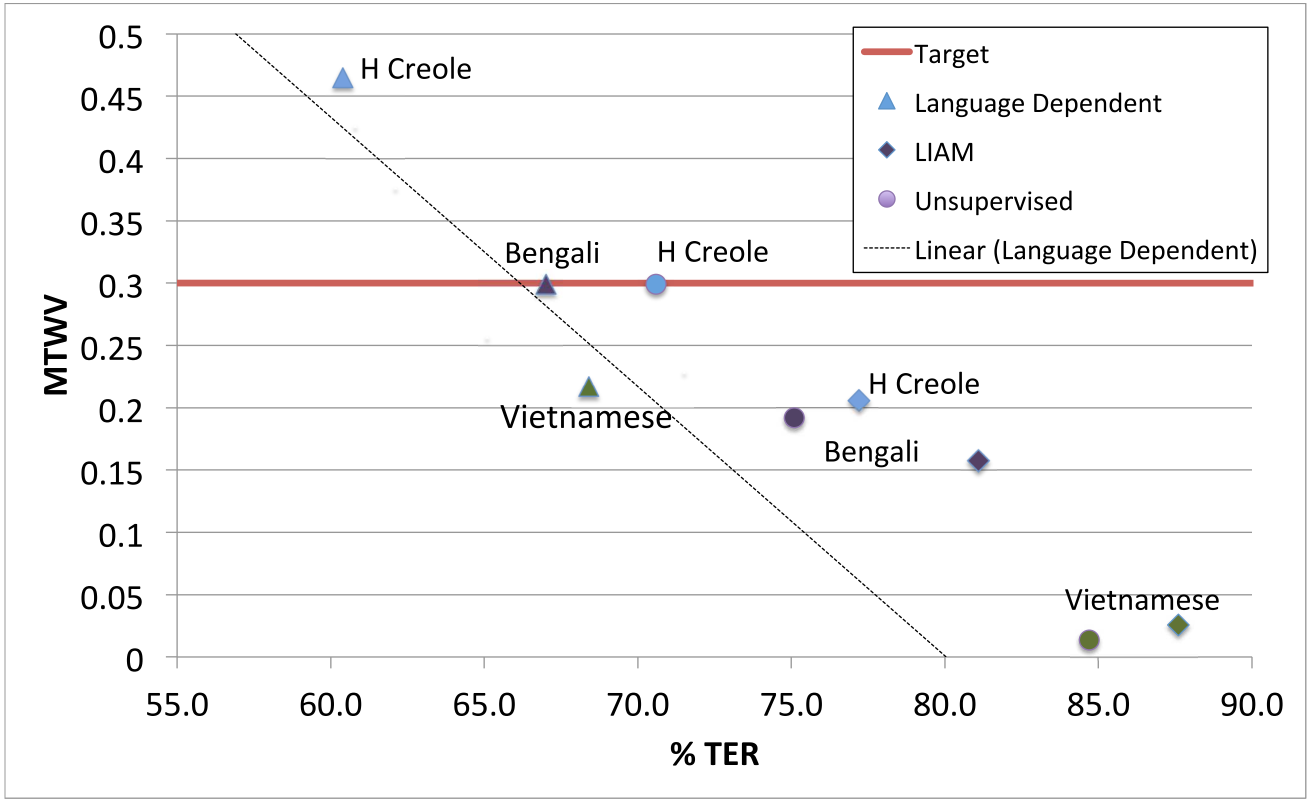
- Confidence scoring of trigram CN output
- ∼30%/25 hours selected for unsupervised training
- ∼3%/2.5 hours selected for MAP adaptation

Haitian Creole Bootstrapping

System	Stage	WER (%)	MTWV IV	MTWV OOV	MTWV Tot
Language Dependent	fMPE	61.7	0.4673	0.2347	0.4317
Language Independent	fMPE	77.2	0.2250	0.0756	0.2058
Unsupervised	ML	70.4	0.3118	0.1560	0.2880
	MPE	71.7	0.3021	0.1682	0.2815
	fMPE	71.3	0.2956	0.1524	0.2736
	ML-MAP	70.6	0.3123	0.1723	0.2911

- All systems use SAT
- Maximum likelihood (ML) Unsupervised system achieves target MTWV for in-vocabulary queries
- Discriminative training degrades performance
- MAP adaptation of ML system with 2.5hrs gives small gain

LD vs LIAM vs Unsupervised



- Bengali slightly smaller gains than Haitian Creole
- Vietnamese TER improves but MTWV very poor - but on trend!

Conclusions

- Multi-Language DNN features yield significant gains over Language Dependent
  - Improve languages within training set and outside
  - Useful to fine tune features to a language
  - Open question as to the optimum nature of the targets
- Multi-Language classifiers can help - results inconclusive to date
- Language Independent
  - Current systems insufficiently language independent!
  - Possible($*$) to achieve program goals bootstrapping from ML system

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?

kate.knill@eng.cam.ac.uk

Aachen Multi-Language Features Performance

- Language-specific targets, Tandem-SAT-MPE, Vietnamese

BN	TER (%)	MTWV
LLP UL	64.0	0.1834
LLP ML	62.6	0.2498
LLP ML + LLP UL	60.9	0.2541
FLP ML	57.6	0.2902
FLP ML + LLP UL	57.1	0.3170

- Fine tuning used above - generally gave gains
- Including FLPs instead of LLP: 9% rel. TER improvement over the unilingual features, >40% improvement in MTWV
- Similar but slightly less gain if fast developed BNs are used

Language-Independent Performance

- Tandem-SAT, Bigram LM, UL trained on target language

Language	Id	AM HMM	BN MLP	TER (%)	MTWV Tot
Vietnamese	107	UL	UL	69.1	0.1882
		UL	ML	68.5	0.2121
		ML	ML	88.3	0.0171
Bengali	103	UL	UL	69.1	0.2106
		UL	ML	67.8	0.2290
		ML	ML	83.2	0.1172
Haitian-Creole	201	UL	UL	63.1	0.4035
		UL	ML	62.2	0.4205
		ML	ML	78.6	0.1943

Analysis on Use of Multilingual Tree

- PLP, ML-trained, Bigram LM

- Three systems compared for impact of ML tree:
  - UL: uni-language (target) performance
  - ML→UL: mllr+map of ML system to target language
  - ML: multi-language performance

AM	Tree	107	103	201
UL	UL	77.8	76.0	71.6
ML→UL	ML	82.0	78.0	73.8
ML	ML	91.4	89.4	85.8

- Adaptation improved all systems
  - Vietnamese is more sensitive to tree