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

14 February 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
  
  $$TWV(\theta) = 1 - \left[ P_{\text{Miss}}(\theta) + \beta P_{\text{FA}}(\theta) \right]$$

- 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:

<table>
<thead>
<tr>
<th>Language</th>
<th>Release</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cantonese</td>
<td>IARPA-babel101-v0.4c</td>
</tr>
<tr>
<td>Pashto</td>
<td>IARPA-babel104b-v0.4aY</td>
</tr>
<tr>
<td>Turkish</td>
<td>IARPA-babel105b-v0.4</td>
</tr>
<tr>
<td>Tagalog</td>
<td>IARPA-babel106-v0.2f</td>
</tr>
<tr>
<td>Vietnamese</td>
<td>IARPA-babel107b-v0.7</td>
</tr>
<tr>
<td>Assamese</td>
<td>IARPA-babel102b-v0.5a</td>
</tr>
<tr>
<td>Bengali</td>
<td>IARPA-babel103b-v0.4b</td>
</tr>
<tr>
<td>Haitian Creole</td>
<td>IARPA-babel201b-v0.2b</td>
</tr>
<tr>
<td>Lao</td>
<td>IARPA-babel203b-v3.1a</td>
</tr>
<tr>
<td>Zulu</td>
<td>IARPA-babel206b-v0.1d</td>
</tr>
</tbody>
</table>
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
Multilingual STT for Spoken Term Detection

Language Dependent STT - General Training Procedure

Initial Acoustic Models
"Clean" training data
CMN
ML training

Constrained Recognition

Baseline Acoustic Models
Full training data
CMN
MPE training

Alignments Parameters

Model Parameters Alignments Initial Results

Acoustic Model
Full training data
TANDEM/SAT, Hybrid
MPE training, etc etc

Acoustic Model Initialisation

• “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 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
Multilingual STT for Spoken Term Detection

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 +Δ + Δ² + Δ³ +HLDA, pitch +Δ + Δ², (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/MTTWV 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

<table>
<thead>
<tr>
<th>Language</th>
<th>System</th>
<th>TER (%)</th>
<th>MTWV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vietnamese</td>
<td>Tandem</td>
<td>55.1</td>
<td>0.423</td>
</tr>
<tr>
<td></td>
<td>Hybrid</td>
<td>54.4</td>
<td>0.418</td>
</tr>
<tr>
<td>Cantonese</td>
<td>Tandem</td>
<td>46.4</td>
<td>0.547</td>
</tr>
<tr>
<td></td>
<td>Hybrid</td>
<td>46.9</td>
<td>0.542</td>
</tr>
</tbody>
</table>

- 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
• 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
• Combine data from LLP from seven languages:
  - Cantonese, Pashto, Turkish, Tagalog, Assamese, Lao, Zulu

• Can be applied to any language (in theory ...)

CUED Single Multi-Language System

- Input Layer
- Hidden Layers
- Bottleneck Layer
- Targets
- Context Dependent

PLP

- State Position
- Vowel?
- Word Final?
- Cantonese?

Context Dependent HMMs

CMLLR/fMPE

- Bottleneck
- Pitch

CUED Lorelei Team
BABEL Program
Seminar at Edinburgh University February 2014
Multilingual STT for Spoken Term Detection

Multi-Language Features Performance

- Tandem-SAT-fMPE, Bigram LM

<table>
<thead>
<tr>
<th>Language Id</th>
<th>BN MLP</th>
<th>TER (%)</th>
<th>MTWV</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>IV</td>
<td>OOV</td>
</tr>
<tr>
<td>Assamese 102</td>
<td>UL</td>
<td>67.7</td>
<td>0.2703</td>
<td>0.0633</td>
</tr>
<tr>
<td></td>
<td>ML</td>
<td>66.2</td>
<td>0.2996</td>
<td>0.0789</td>
</tr>
<tr>
<td>Zulu 206</td>
<td>UL</td>
<td>75.1</td>
<td>0.2400</td>
<td>0.0220</td>
</tr>
<tr>
<td></td>
<td>ML</td>
<td>73.9</td>
<td>0.2521</td>
<td>0.0240</td>
</tr>
</tbody>
</table>

- 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

CUED Lorelei Team
BABEL Program

Seminar at Edinburgh University February 2014
### Multi-Language Systems Performance

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

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

- 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

<table>
<thead>
<tr>
<th>Level</th>
<th>Shape</th>
<th>Language Id</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>L101</td>
</tr>
<tr>
<td>high</td>
<td>falling</td>
<td>0</td>
</tr>
<tr>
<td>high</td>
<td>level</td>
<td>1</td>
</tr>
<tr>
<td>high</td>
<td>rising</td>
<td>2</td>
</tr>
<tr>
<td>mid</td>
<td>level</td>
<td>3</td>
</tr>
<tr>
<td>mid</td>
<td>dipping</td>
<td>—</td>
</tr>
<tr>
<td>low</td>
<td>rising</td>
<td>5</td>
</tr>
</tbody>
</table>

- ask level and shape questions in decision tree
**Language-Independent Performance**

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

<table>
<thead>
<tr>
<th>Language</th>
<th>Id</th>
<th>AM</th>
<th>BN</th>
<th>TER (%)</th>
<th>MTWV Tot</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>HMM</td>
<td>MLP</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bengali</td>
<td>103</td>
<td>UL</td>
<td>UL</td>
<td>69.1</td>
<td>0.2106</td>
</tr>
<tr>
<td></td>
<td></td>
<td>UL</td>
<td>ML</td>
<td>67.8</td>
<td>0.2290</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ML</td>
<td>ML</td>
<td>83.2</td>
<td>0.1172</td>
</tr>
<tr>
<td>Haitian-Creole</td>
<td>201</td>
<td>UL</td>
<td>UL</td>
<td>63.1</td>
<td>0.4035</td>
</tr>
<tr>
<td></td>
<td></td>
<td>UL</td>
<td>ML</td>
<td>62.2</td>
<td>0.4205</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ML</td>
<td>ML</td>
<td>78.6</td>
<td>0.1943</td>
</tr>
</tbody>
</table>

- 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 (1)

- **red** indicates held-out languages (L107, L103, L201)
- **green** indicates tonal training languages
**Analysis on Use of Multilingual Tree (2)**

- 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

<table>
<thead>
<tr>
<th>AM</th>
<th>Tree</th>
<th>107</th>
<th>103</th>
<th>201</th>
</tr>
</thead>
<tbody>
<tr>
<td>UL</td>
<td>UL</td>
<td>77.8</td>
<td>76.0</td>
<td>71.6</td>
</tr>
<tr>
<td>ML→UL</td>
<td>ML</td>
<td>82.0</td>
<td>78.0</td>
<td>73.8</td>
</tr>
<tr>
<td>ML</td>
<td>ML</td>
<td>91.4</td>
<td>89.4</td>
<td>85.8</td>
</tr>
</tbody>
</table>

- Adaptation improved all systems
  - Vietnamese is more sensitive to tree
Multilingual STT for Spoken Term Detection

Bootstrapping with Multi-Language Systems

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

- Approx 25hrs (/66hrs) unsupervised training data selected based on confidence scoring of trigram CN output

<table>
<thead>
<tr>
<th>System</th>
<th>Stage</th>
<th>WER (%)</th>
<th>MTWV</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Language Dependent</td>
<td>fMPE</td>
<td>62.3</td>
<td>0.4485</td>
</tr>
<tr>
<td>Language Independent</td>
<td>fMPE</td>
<td>77.5</td>
<td>0.2227</td>
</tr>
<tr>
<td>Unsupervised</td>
<td>ML</td>
<td>70.9</td>
<td>0.3062</td>
</tr>
<tr>
<td></td>
<td>MPE</td>
<td>73.0</td>
<td>0.2895</td>
</tr>
<tr>
<td></td>
<td>fMPE</td>
<td>73.5</td>
<td>0.2722</td>
</tr>
</tbody>
</table>

- Maximum likelihood (ML) Unsupervised system achieves target MTWV for in-vocabulary queries
- Discriminative training degrades performance
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

<table>
<thead>
<tr>
<th>BN</th>
<th>TER (%)</th>
<th>MTWV</th>
</tr>
</thead>
<tbody>
<tr>
<td>LLP UL</td>
<td>64.0</td>
<td>0.1834</td>
</tr>
<tr>
<td>LLP ML</td>
<td>62.6</td>
<td>0.2498</td>
</tr>
<tr>
<td>LLP ML + LLP UL</td>
<td>60.9</td>
<td>0.2541</td>
</tr>
<tr>
<td>FLP ML</td>
<td>57.6</td>
<td>0.2902</td>
</tr>
<tr>
<td>FLP ML + LLP UL</td>
<td>57.1</td>
<td>0.3170</td>
</tr>
</tbody>
</table>

- 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

<table>
<thead>
<tr>
<th>Language</th>
<th>Id</th>
<th>AM</th>
<th>BN</th>
<th>TER (%)</th>
<th>MTWV Tot</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vietnamese</td>
<td>107</td>
<td>UL</td>
<td>UL</td>
<td>69.1</td>
<td>0.1882</td>
</tr>
<tr>
<td></td>
<td></td>
<td>UL</td>
<td>ML</td>
<td>68.5</td>
<td>0.2121</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ML</td>
<td>ML</td>
<td>88.3</td>
<td>0.0171</td>
</tr>
<tr>
<td>Bengali</td>
<td>103</td>
<td>UL</td>
<td>UL</td>
<td>69.1</td>
<td>0.2106</td>
</tr>
<tr>
<td></td>
<td></td>
<td>UL</td>
<td>ML</td>
<td>67.8</td>
<td>0.2290</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ML</td>
<td>ML</td>
<td>83.2</td>
<td>0.1172</td>
</tr>
<tr>
<td>Haitian-Creole</td>
<td>201</td>
<td>UL</td>
<td>UL</td>
<td>63.1</td>
<td>0.4035</td>
</tr>
<tr>
<td></td>
<td></td>
<td>UL</td>
<td>ML</td>
<td>62.2</td>
<td>0.4205</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ML</td>
<td>ML</td>
<td>78.6</td>
<td>0.1943</td>
</tr>
</tbody>
</table>