Institute for
Infocomm Research
A*STAR

# Efficient Methods to Train Multilingual Bottleneck Feature Extractors for Low Resource Keyword Search

Chongjia Ni, Cheung-Chi Leung,
Lei Wang, Nancy Chen and Bin Ma

9 March 2017
ICASSP 2017, New Orleans

# Outline

- Introduction
- Multilingual Data Selection for Low Resource Keyword Search
- Multilingual Deep Bottleneck Feature Extractors
- Experiments on 2015 NIST Open KWS
- Conclusions

# Introduction

- Background
  - LVCSR-based keyword Search (KWS) for low resource languages
    - Multilingual DNN for rapid language adaptation
    - Bottleneck feature extraction from multilingual DNN
  - Multilingual deep bottleneck features
    - An efficient way for cross-lingual knowledge transfer
    - Not all multilingual data contribute equally to ASR/KWS performance of a target language

# Introduction

- Organization of the Paper
  - Effective multilingual data selection
    - LSTM RNN for modeling languages
    - Select utterances in multilingual training data that are acoustically close to the training data of the target language
  - Multi-lingual deep bottleneck feature (BNF) extractor
    - Comparison with previous work with submodular subset selection
    - Analysis on rapid updating existing BNF extractor vs. new BNF extractor

# Multilingual Data Selection

- Multilingual Data Selection based on Submodular function
  - GMM tokenization instead of phonetic related features

**Submodular multilingual data selection**

- Utterance representation based on GMM
- tf-idf features for each utterance based on n-gram of Gaussian index
- Based on tf-idf features, compute the probability distribution $\{p_u\}_{u \in U}$ on target language data set
- Using the following submodular function to select multilingual data

probability distribution of feature *u* estimated from target language data

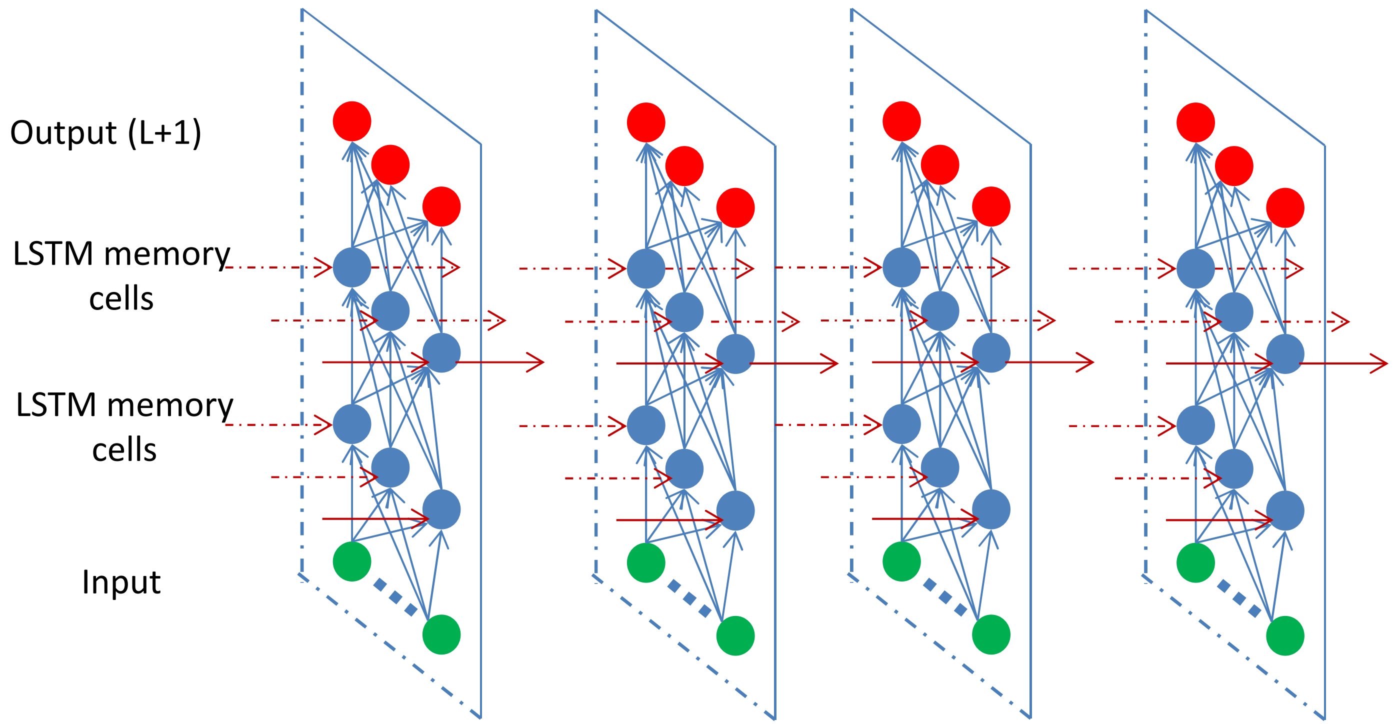
normalization of utterance length

$$f(s) = \sum_{u \in U} p_u \log\left(\sum_{s \in S} \frac{1}{l(s)} m_u(s)\right)$$

$m_u(s)$ measures the degree of feature *u* of the utterance *s*

# Multilingual Data Selection

- Multilingual Data Selection based on Language Identification
  - LSTM RNN model for language identification




*J. Gonzalez-Dominguez, I. Lopez-Moreno, H. Sak, J. Gonzalez-Rodriguez, P. J. Moreno, "Automatic Language Identification using Long Short-Term Memory Recurrent Neural Networks", Interspeech 2014*

# Multilingual Data Selection

- Multilingual Data Selection based on Language Identification
  - Utterances in multilingual training data, which have high softmax outputs for the target language, are selected.

$S$: a set of utterances

$\mathrm{p}(\mathrm{L}|\mathrm{x_t}) = \big(\mathrm{p}(\mathrm{l_0}|\mathrm{x_t}), \mathrm{p}(\mathrm{l_1}|\mathrm{x_t}), \cdots, \mathrm{p}(\mathrm{l_N}|\mathrm{x_t})\big)$ be the posterior vector for input feature $\mathrm{x_t}$ at frame t. $\mathrm{p}(\mathrm{l_i}|\mathrm{x_t})$ is the posterior of language $\mathrm{l_i}$ for input feature $\mathrm{x_t}$

$$f(S) = \sum_{s \in S} \log\left(Proj_{i_k}\left(\frac{1}{T(s)} \sum_{t=1}^{T} p(L|x_t^s)\right)\right)$$

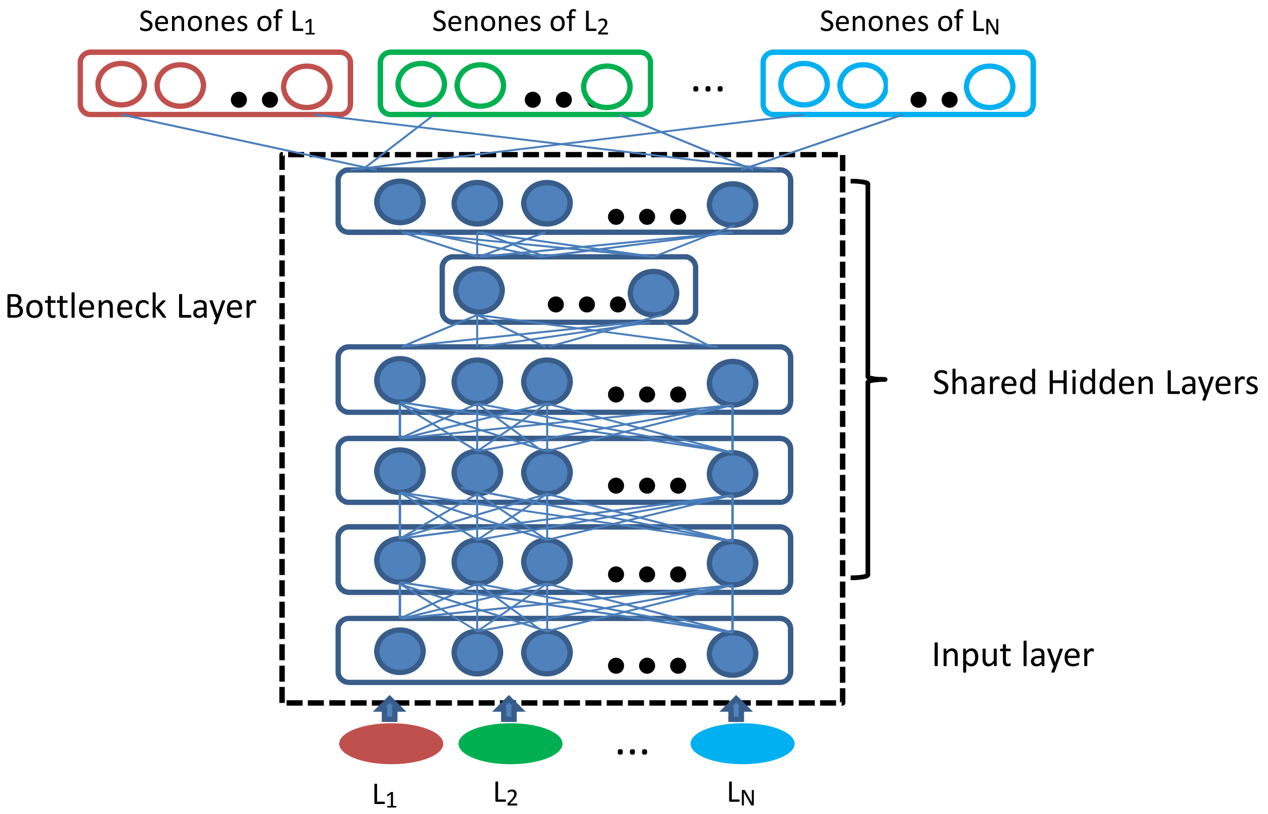
$Proj_{i_k}(\cdot)$ is the projection function in order to get the $i_k$ component for a vector, and $i_k$ is the target language index.

$T(s)$ is the number of frames of utterance $s$

  - Select those utterances which are classified into the target language with high probability (acoustically similar to the training data of the target language)

# Multilingual Deep Bottleneck Feature Extractors

- Shared-hidden-layer Multilingual DNN for Bottleneck Features




*J.-T. Huang, J. Li, D. Yu, L. Deng, and Y. Gong, "Cross-language Knowledge Transfer using Multilingual Deep Neural Network with Shared Hidden Layers", ICASSP 2013*

# Experimental Setup

- Keyword Search Task for Low Resource Languages
  - NIST Open Keyword Search 2015 Evaluation
    - Swahili as target language
    - Language packs of 23 other languages released by IAPRA Babel Program
    - VLLP (3H training set) + 10H development set *Dev10h* + 15H evaluation set *Evalpart1*
  - Feature extraction
    - 117 features including 22 fbank + 3 pitch + $\Delta$ + $\Delta\Delta$ + 42 BNF
    - Multilingual deep BNF extractor (6 hidden layers, 42 hidden units for bottleneck layer, 1500 hidden units for other hidden layers)
  - Acoustic modeling
    - Hybrid DNN (6 hidden layers, 1,024 hidden units for each hidden layer, 2,207 senones)
    - Discriminative trained GMM-HMM for alignment
    - Cross-entropy training + sMBR criterion for sequence training
  - Language modeling
    - 3-gram Web-data LM which was interpolated with 3-gram LM trained using VLLP transcription
    - Interpolation optimized by minimizing perplexity on the transcription *Dev10h*
  - Keyword search
    - 4,454 keywords (260 OOV to LM with Web-data and 2,667 OOV to LM with training transcription
    - ATWV (actual term weighted value) and WER for measuring the performance

# Experimental Setup

- Selected Multilingual Training Data for BNF Extractors

| | |
|---|---|
| **Baseline-Multilingual-509h** | Cantonese (175.2 hours), Pashto (111.1 hours), Turkish (107.4 hours), Tagalog (115.7 hours) while 4 languages were randomly selected from 23 FLPs |
| **Baseline-Multilingual-14h-Submodular** | 3.5 hours from each language selected from **Baseline-Multilingual-509h** based on submodular subset selection |
| **Baseline-Multilingual-14h-LID** | 3.5 hours from each language selected from **Baseline-Multilingual-509h** based on proposed multilingual data selection |
| **Submodular-Multilingual-96h** | Zulu (20.1 hours), Pashto (35.0 hours), Vietnamese (27.6 hours), Cantonese (13.3 hours) selected from 23 FLPs based on submodular subset selection |
| **Proposed-Multilingual-96h** | Haitian Creole (29.7 hours), Zulu (21.6 hours), Dholuo (23.9 hours), Vietnamese (20.7 hours) selected from 23 FLPs based on proposed multilingual data selection |
| **Proposed-Multilingual-14h** | 3.5 hours from each language selected from **Proposed-Multilingual-96h** based on proposed multilingual data selection |
| **Creole-14h** | Haitian Creole (14 hours) selected based on proposed multilingual data selection |

# Experiments

**Table 1. Performance of baseline KWS systems on *Evalpart1*.**

| BNF extractor | Data set for training BNF extractor | Web-data LM | | Training transcription LM | |
|---|---|---|---|---|---|
| | | WER | ATWV | WER | ATWV |
| **Baseline Monolingual** | VLLP-TL | 67.4 | 0.308 | 69.3 | 0.194 |
| **Baseline Multilingual** | Baseline-Multilingual-509h | 64.5 | 0.361 | 69.0 | 0.216 |

- Better performance by using a large amount of multilingual data even they are not carefully against the target language.

# Experiments

**Table 2. The performance of different KWS systems on *Evalpart1* by rapidly updating the baseline multilingual BNF extractor using 14 hours of multilingual data.**

| BNF extractor | Data set for updating BNF extractor | Web-data LM | | Training transcription LM | |
|---|---|---|---|---|---|
| | | WER | ATWV | WER | ATWV |
| **R1** | Baseline-Multilingual-14h-LID + VLLP-TL | 62.1 | 0.396 | 66.7 | 0.239 |
| **R2** | Baseline-Multilingual-14h-Sub + VLLP-TL | 62.3 | 0.390 | 67.1 | 0.238 |
| **R3** | Proposed-Multilingual-14h + VLLP-TL | **61.4** | **0.397** | **66.0** | **0.242** |
| **R4** | Creole-14h + VLLP-TL | 61.6 | 0.389 | 66.3 | 0.231 |

# Experiments

**Table 3. The performance of different KWS systems on Evalpart1 by training multilingual BNF extractors from scratch.**

| BNF extractor | Data set for training BNF extractor | Web-data LM | | Training transcription LM | |
|---|---|---|---|---|---|
| | | WER | ATWV | WER | ATWV |
| S1 | Baseline-Multilingual-509h + VLLP-TL | 61.2 | 0.413 | 65.7 | 0.243 |
| S2 | Proposed-Multilingual-96h | 60.9 | 0.407 | 65.6 | 0.239 |
| S3 | Proposed-Multilingual-96h + VLLP-TL | 60.7 | 0.416 | 65.6 | 0.244 |
| S4 | Submodular-Multilingual-96h | 61.3 | 0.399 | 65.8 | 0.237 |
| S5 | Submodular-Multilingual-96h + VLLP-TL | 61.1 | 0.402 | 65.7 | 0.237 |
| S6 | Creole-14h + VLLP-TL | 65.1 | 0.372 | 69.5 | 0.221 |

- Combining speech data of target language with multilingual data for building the BNF extractor gives a significant improvement.
- Training a new BNF extractor using the proposed data selection provided good performance.
- The amount of selected data also affects the performance of the BNF extractor.

# Experimental analysis

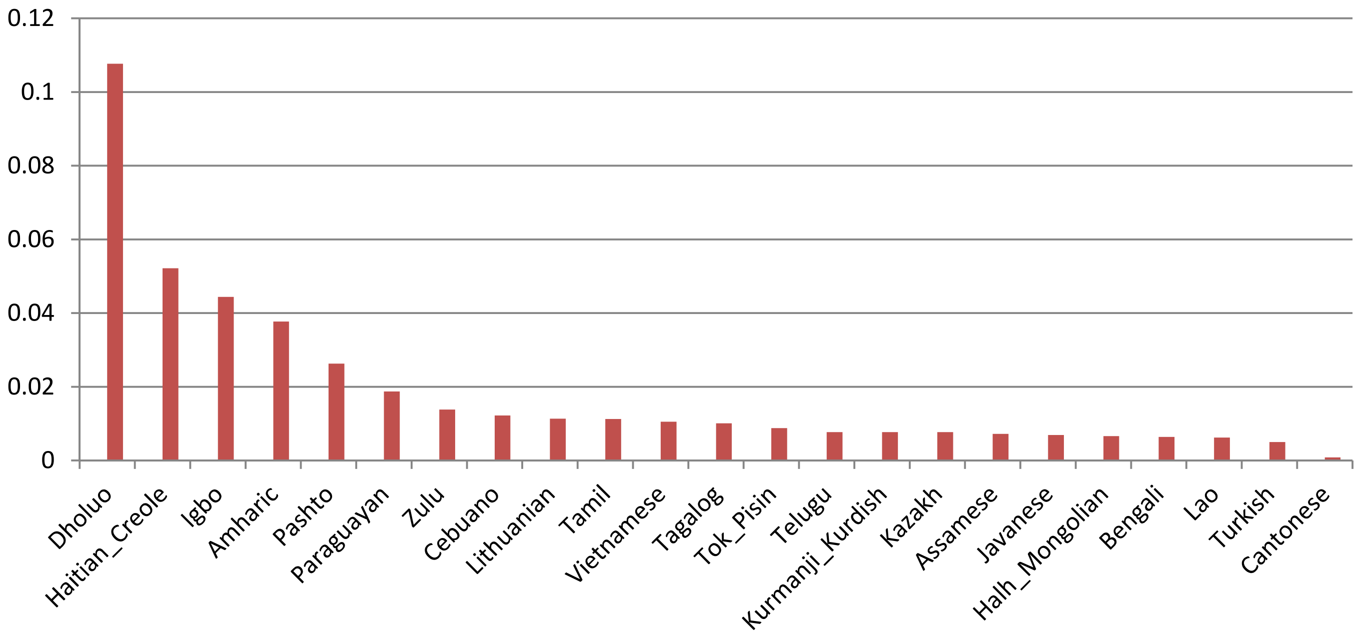



**Fig. 1. Similarity measure between different source languages and target language (Swahili). The vertical axis denotes the average misclassification posterior probability of all utterance of each language.**

- Top two languages are overlapped with the four languages in "Proposed-Multilingual-96h" (Haitian Creole, Zulu, Dholuo, Vietnamese).
- Not all the utterances in a language have equal similarity to the target language.

# Conclusions

- Studied effective methods to train multilingual bottleneck features extractors for keyword search task for low resource languages.
- Not all multilingual data can contribute equally to the KWS performance. The utterances that are acoustically similar to the target language data set are more useful.
- LSTM RNN based language identification is effective and efficient for multilingual data selection.
- Combining speech data of target language with multilingual data for building the BNF extractor gives an improvement for KWS of the target language.

# Thank you