Multilingual Speech Recognition 3

10 July 2012
Outline

- Rapid Language Adaptation
- Rapid Generation of Language Models
- Text normalization with Crowdsourcing
- Code-Switching
  - SMT-based text generation for code-switching language models
- Automatic pronunciation dictionary generation from the WWW
- Multilingual Bottle Neck Features
- Multilingual Unsupervised Training
Overview – Automatic Speech Recognition

Front End (Preprocessing) → Decoder (Search) → Text

- Acoustic Model
- Lexicon / Dictionary
- Language Model

hi /h//ai/
you /j/u/
we /w//i/
hi you
you are
I am
Overview – Automatic Speech Recognition

Multilingual Bottle Neck Features

Unsupervised training

Front End (Preprocessing)

Decoder (Search)

Text

Acoustic Model

Lexicon / Dictionary

Language Model

hi /h//ai/
you /j//u/
we /w//i/

hi you
you are
I am

Web-derived prons.

Text Normalization

Crawling

language modeling in the context of code-switching
Rapid Language Adaptation

Goal:

• Build Automatic Speech Recognition (ASR) for unseen Languages/Accents/Dialects with minimal human effort

Challenges:

• No text data
• No pronunciation dictionary
• No or Few Data, i.e. no transcribed Audio Data
Rapid Generation of Language Models

(based on Vu, Schlippe, Kraus and Schultz 2010)
Overview – Automatic Speech Recognition

Front End (Preprocessing) ➔ Decoder (Search) ➔ Text

- Acoustic Model
- Lexicon / Dictionary
- Language Model

Crawling ➔ Text Normalization
Rapid Bootstrapping

- **Overview:**
  - ASR for Bulgarian, Croatian, Czech, Polish, and Russian using the Rapid Language Adaptation Toolkit (RLAT)
  - Crawling and processing large quantities of text material from the Internet
  - Strategy for language model optimization on the given development set in a short time period with minimal human effort

- **Slavic Languages and data resources**
  - Well known for their rich morphology, caused by a high reflection rate of nouns using various cases and genders (e.g. nowy student, nowego studenta, nowi studentci)
  - GlobalPhone speech data: ~20h for each language, 80% for training, 10% for dev and 10% for evaluation
Rapid Bootstrapping

• Baseline systems:
  – Rapid bootstrapping based on multilingual acoustic model inventory trained earlier from seven GlobalPhone languages
  – To bootstrap a system in a new language, an initial state alignment is produced by selecting the closest matching acoustic models from the multilingual inventory as seeds
  – Closest match is derived from an IPA-based phone mapping

  – Initial results (word error rates (WER)) with language model built with the utterances of the training transcriptions:
    – 63% for Bulgarian
    – 60% for Croatian
    – 49% for Czech
    – 72% for Polish
    – 61% for Russian
Rapid Bootstrapping for five Eastern European languages

“Quick&Dirty” Text Processing

Remove HTML tags, code fragment, empty lines
Rapid Bootstrapping for five Eastern European languages

“Quick&Dirty” Text Processing

![Graph showing perplexity over days for Bulgarian, Croatian, Czech, Polish, and Russian languages. The graph indicates improvements over time.]
Rapid Bootstrapping for five Eastern European languages

“Quick&Dirty” Text Processing

+ strong increase of ppl due to the rough text processing and strong growth of vocabulary

![Graph showing perplexity over days for different languages](image)
Rapid Bootstrapping for five Eastern European languages

“Quick&Dirty” Text Processing

Text normalization & Vocabulary Selection

process special character, digits, cardinal number, dates, punctuation
+ select most frequent words
Rapid Bootstrapping for five Eastern European languages

"Quick&Dirty" Text Processing

Development of PPL for normalized crawl data over 20 days

- Bulgarian
- Croatian
- Czech
- Polish
- Russian

Perplexity

0  5  10  15  20
day(s)
Rapid Bootstrapping for five Eastern European languages

“Quick&Dirty”

Te
Vo

\[
\text{Bulgarian} \quad \text{Croatian} \quad \text{Czech} \quad \text{Polish} \quad \text{Russian}
\]

\[\begin{align*}
\text{day[s]} & \quad 0 \quad 5 \quad 10 \quad 15 \quad 20 \\
(Y/A0) & \quad 12 \quad 10 \quad 8 \quad 6 \quad 4 \quad 2 \quad 0
\end{align*}\]
Rapid Bootstrapping for five Eastern European languages

“Quick&Dirty” Text Processing

Teş Voc

Development of WER for normalized crawl data over 20 days
Rapid Bootstrapping for five Eastern European languages

“Quick&Dirty” Text Processing

- Development of WER for normalized crawl data over 20 days

- + decrease of WER only in few days
- + enlarging the text corpus provides the generalization of LM but does not help for the specified test set
Rapid Bootstrapping for five Eastern European languages

“Quick&Dirty” Text Processing

Text normalization & Vocabulary Selection

Day-wise Language Model Interpolation

LM was built for each day and interpolated with the LM from the previous days
Rapid Bootstrapping for five Eastern European languages

Development of WER for interpolated crawl data over 20 days

- Bulgarian_interp
- Croatian_interp
- Czech_interp
- Polish_interp
- Russian_interp

WER (%) vs. day(s)
Rapid Bootstrapping for five Eastern European languages

+ harvesting the text data from one particular website makes the crawling process fragile
Rapid Bootstrapping for five Eastern European languages

- "Quick&Dirty" Text Processing
- Text normalization & Vocabulary Selection
- Day-wise Language Model Interpolation

Text Data Diversity

Build LMs based on text data from different websites, Interpolate them with the background LM
Rapid Bootstrapping for five Eastern European languages

Final language models:

<table>
<thead>
<tr>
<th>Languages</th>
<th>OOV(%)</th>
<th>PPL</th>
<th>#Tokens</th>
<th>#Types</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bulgarian</td>
<td>1.2</td>
<td>543</td>
<td>405M</td>
<td>274K</td>
</tr>
<tr>
<td>Croatian</td>
<td>3.6</td>
<td>813</td>
<td>331M</td>
<td>362K</td>
</tr>
<tr>
<td>Czech</td>
<td>3.8</td>
<td>2,115</td>
<td>508M</td>
<td>277K</td>
</tr>
<tr>
<td>Polish</td>
<td>2.9</td>
<td>1,372</td>
<td>224M</td>
<td>243K</td>
</tr>
<tr>
<td>Russian</td>
<td>3.4</td>
<td>1,675</td>
<td>2931M</td>
<td>293K</td>
</tr>
</tbody>
</table>
Rapid Bootstrapping – Language Model optimization strategy

Figure: Speech Recognition Improvements [WER]
Rapid Bootstrapping

- **Conclusion:**
  - Crawling and processing a large amount of text material from WWW using RLAT
  - Investigation of the impact of text normalization and text diversity on the quality of the language model in terms of perplexity, out-of-vocabulary rate and its influence on WER
  - ASR systems in a very short time period and with minimum human effort

- Best systems on the evaluation set (WERs):
  - 16.9% for Bulgarian
  - 32.8% for Croatian
  - 23.5% for Czech
  - 20.4% for Polish
  - 36.2% for Russian
SMT-based Text Normalization with Crowdsourcing

(based on Schlippe, Zhu, Gebhardt and Schultz 2010)
Overview – Automatic Speech Recognition

Front End (Preprocessing) ➔ Decoder (Search) ➔ Text

Acoustic Model
Lexicon / Dictionary
Language Model

Crawling
Text Normalization
Text Normalization based on Statistical Machine Translation and Internet User Support – Web-based Interface

- Web-based user interface for language-specific text normalization
- Hybrid approach (rules + Statistical Machine Translation (SMT))

Figure: Web-based User Interface for Text Normalization
Text Normalization based on Statistical Machine Translation and Internet User Support – Experiments and Evaluation

• Experiments and Results:
  – How well does SMT perform in comparison to LI-rule (language-independent rule-based), LS-rule (language-specific rule-based) and human (normalized by native speakers)?
  – How does the performance of SMT evolve over the amount of training data?
  – How can we modify our system to get a time and effort reduction?

• Evaluation:
  – comparing the quality of 1k output sentences derived from the systems to text which was normalized by native speakers in our lab
  – creating 3-gram LMs from our hypotheses and evaluated their perplexities on 500 sentences manually normalized by native speakers
Table: Language-independent and -specific text normalization

<table>
<thead>
<tr>
<th>Language-independent Text Normalization (LI-rule)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Removal of HTML, Java script and non-text parts.</td>
</tr>
<tr>
<td>2. Removal of sentences containing more than 30% numbers.</td>
</tr>
<tr>
<td>3. Removal of empty lines.</td>
</tr>
<tr>
<td>4. Removal of sentences longer than 30 tokens.</td>
</tr>
<tr>
<td>5. Separation of punctuation marks which are not in context with numbers and short strings (might be abbreviations).</td>
</tr>
<tr>
<td>6. Case normalization based on statistics.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Language-specific Text Normalization (LS-rule)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Removal of characters not occurring in the target language.</td>
</tr>
<tr>
<td>2. Replacement of abbreviations with their long forms.</td>
</tr>
<tr>
<td>3. Number normalization (dates, times, ordinal and cardinal numbers, etc.).</td>
</tr>
<tr>
<td>5. Removal of remaining punctuation marks.</td>
</tr>
</tbody>
</table>
Text Normalization based on Statistical Machine Translation and Internet User Support – Experiments

- Non-norm. Text
- Rule-based LI norm.
- LI-rule Output Text
- SMT-based norm.
- SMT Output Text
- Human norm.
- Human Output Text
- Statistical post-editing
- Hybrid Output Text
- Language-specific rule-based with SMT post-editing (*hybrid*)
- LS-rule Output Text
- Language-specific rule-based (*LS-rule*)
- SMT approach (*SMT*)
- Manually normalized by native speakers as golden line (*human*)

Language-independent rule-based (*LI-rule*)
Text Normalization based on Statistical Machine Translation and Internet User Support – Results

Figure: Performance (edit distance) over amount of training data
Text Normalization based on Statistical Machine Translation and Internet User Support – Results

![Graph showing performance (PPL) over amount of training data.](image)

Figure: Performance (PPL) over amount of training data
Text Normalization based on Statistical Machine Translation and Internet User Support – Results

Figure: Performance (edit dist.) over amount of training data (all sentences containing numbers were removed)
Time to normalize 1k sentences (in minutes) and edit distances (%) of the SMT system
Conclusion:

- A crowdsourcing approach for SMT-based language-specific text normalization:
  Native speakers deliver resources to build normalization systems by editing text in our web interface
- Results of SMT close to LS-rule, hybrid better, close to human
- Close to optimal performance achieved after about 5 hours manual annotation (450 sentences)
- Parallelization of annotation work to many users is supported by web interface

Evaluation:

- Investigating other languages
- Enhancements to further reduce time and effort
SMT-based Text Generation for Code-Switching Language Models

(based on Blaicher 2010)
Code-Switching Speech Recognition

- Code-switching: [Pop79]
  Sometimes I’ll start a sentence in English y termino en español

- Problem:
  - Scarse code-switching data for training speech recognizers

- Solution:
  - Combine existing code-switching data, with large monolingual texts for better code-switch language models
Search & Replace (S&R)

- Build code-switch texts from SEAME train text + monolingual texts
### Search & Replace Evaluation

<table>
<thead>
<tr>
<th></th>
<th>Baseline</th>
<th>S&amp;R</th>
<th>Rel. Δ to baseline</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perplexity</td>
<td>516</td>
<td>533</td>
<td>-3.3%</td>
</tr>
<tr>
<td>Mix error rate (%)</td>
<td>50.11</td>
<td>50.29</td>
<td>-0.4%</td>
</tr>
<tr>
<td>#CS bigrams (unique)</td>
<td>34K</td>
<td>689K</td>
<td>+1926%</td>
</tr>
<tr>
<td>CSR(2)*</td>
<td>26%</td>
<td>45%</td>
<td>+73%</td>
</tr>
</tbody>
</table>

- **CS n-gram ratio (CSR):**
  - Percentage of unique CS n-grams of the dev. text, which are contained in SMT-based text

- Many new CS n-grams
- Improve probabilities
build better CS n-grams:
• Generate less CS n-grams, keep CSR high, use context info

1. **Threshold (T2)**: Replace segments, which are frequent in ST
   • Use a minimum occurrence threshold = 2
   • Higher thresholds removed nearly all segments

2. **Trigger**: Replace only segments after a CS trigger token
   [Sol08,Bur09],
   which occurred in ST before CS
   e.g. 他的 car (his car)
   a. Trigger words (trig words)
   b. Trigger part-of-speech tags (trig PoS), e.g. noun, verb, ...

3. **Frequency Alignment (FA)**: Replace found segment only until a target frequency is reached, computed from ST

   • Target frequency (hello world) = \(\frac{\#\text{segments "hello world"}}{\#\text{sentences}}\)

ST: SEAME train text
Further S&R Improvements: Results

Baseline: Train+Monol. EN/CN
S&R: Search & Replace
T2: Min. occurrence threshold=2

trig words: Trigger words
Trig PoS: Trigger part-of-speech tags
FA: Frequency alignment of Train+S&R

trig PoS and FA show improvement
Combination trig PoS + FA shows highest improvement
Automatic pronunciation dictionary generation from the World Wide Web

(based on Schlippe, Ochs, and Schultz 2010)
Overview – Automatic Speech Recognition

- Front End (Preprocessing)
- Decoder (Search)
- Text

- Acoustic Model
- Lexicon / Dictionary
- Language Model

Web-derived prons.
Web-derived Prongs. – Introduction

- World Wide Web (WWW) increasingly used as text data source for rapid adaptation of ASR systems to new languages and domains, e.g.
  - Crawl texts to build language models (LMs),
  - Extract prompts read by native speakers to receive transcribed audio data (Schultz et al. 2007)

- Creation of pronunciation dictionary
  - Usually produced manually or semi-automatically
  - Time consuming, expensive
  - Proper names difficult to generate with letter-to-sound rules

- Idea: Leverage off the internet technology and crowdsourcing

→ Is it possible to generate pronunciations based on phonetic notations found in the WWW?
Web-derived Prons. – Wiktionary

- At hand in multiple languages
- In addition to definitions of words, many phonetic notations written in the International Phonetic Alphabet (IPA) are available
- Quality and quantity of entries dependent community and the underlying resources

The ten largest Wiktionary language editions (July 2010)
(http://meta.wikimedia.org/wiki/List_of_Wiktionaries)

<table>
<thead>
<tr>
<th>No.</th>
<th>Language</th>
<th>“Good” Entries</th>
<th>Admins</th>
<th>Active Users</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>French</td>
<td>1,786k</td>
<td>21</td>
<td>286</td>
</tr>
<tr>
<td>2</td>
<td>English</td>
<td>1,770k</td>
<td>100</td>
<td>1047</td>
</tr>
<tr>
<td>3</td>
<td>Lithuanian</td>
<td>542k</td>
<td>4</td>
<td>14</td>
</tr>
<tr>
<td>4</td>
<td>Turkish</td>
<td>268k</td>
<td>6</td>
<td>50</td>
</tr>
<tr>
<td>5</td>
<td>Chinese</td>
<td>257k</td>
<td>9</td>
<td>31</td>
</tr>
<tr>
<td>6</td>
<td>Russian</td>
<td>246k</td>
<td>6</td>
<td>139</td>
</tr>
<tr>
<td>7</td>
<td>Vietnamese</td>
<td>229k</td>
<td>5</td>
<td>31</td>
</tr>
<tr>
<td>8</td>
<td>Ido</td>
<td>171k</td>
<td>2</td>
<td>13</td>
</tr>
<tr>
<td>9</td>
<td>Polish</td>
<td>165k</td>
<td>25</td>
<td>79</td>
</tr>
<tr>
<td>10</td>
<td>Portuguese</td>
<td>156k</td>
<td>6</td>
<td>112</td>
</tr>
</tbody>
</table>
2.1 Data – Wiktionary

**Wiktionary**

**sein**

*Inhaltsverzeichnis [Verbergen]*

1 sein (Deutsch)
   1.1 Hiflswerb
       1.1.1 Übersetzungen
       1.2 Possessivpronomen, 3. Person Singular m, n
       1.2.1 Übersetzungen
   2 sein (Französisch)
      2.1 Substantiv, m
      2.1.1 Übersetzungen

**Hilfsverb**

Anmerkung:
Alle Verbindungen mit sein schreibt man nach neuer Rechtschreibung getrennt (da sein, weg sein, zusammen sein).

Silbentrennung:
sein, Präteritum: war, Partizip II: ge-sein

Aussprache:
IPA: [zən], [tzn], bist: [bist], ist: [ist], sind: [zont], seid: [zont]

**sein (Französisch)**

*Substantiv, m*

Silbentrennung:
sein, Plural: seins

Aussprache:
IPA: [sɛ̃]
Web-derived Prons. – GlobalPhone

- For our experiments, we build ASR systems with GlobalPhone data for English, French, German, and Spanish.
- In GlobalPhone, widely read national newspapers available on the WWW with texts from national and international political and economic topics were selected as resources.
- Vocabulary size and length of audio data for our ASR systems:

<table>
<thead>
<tr>
<th></th>
<th>English</th>
<th>French</th>
<th>German</th>
<th>Spanish</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vocab size</td>
<td>58k</td>
<td>122k</td>
<td>38k</td>
<td>30k</td>
</tr>
<tr>
<td>Audio train</td>
<td>15.4 h</td>
<td>24.9 h</td>
<td>14.9 h</td>
<td>17.5 h</td>
</tr>
<tr>
<td>Audio test</td>
<td>0.5 h</td>
<td>2.0 h</td>
<td>1.5 h</td>
<td>1.6 h</td>
</tr>
</tbody>
</table>

- GlobalPhone dictionaries
  … had been created in rule-based fashion, manually cross-checked
  … contain phonetic notations based on IPA scheme
  mapping between IPA units obtained from Wiktionary and GlobalPhone units is trivial (Schultz, 2002)
Web-derived Prons. – Experiments and Results

• Quantity Check:
  – Given a word list, what is the percentage of words for which phonetic notations are found in *Wiktionary*?
    • Quantity of pronunciations for *GlobalPhone* words
    • Quantity of pronunciations for proper names (e.g. New York)

• Quality Check:
  – How many pronunciations derived from *Wiktionary* are identical to existing *GlobalPhone* pronunciations?
  – How does adding *Wiktionary* pronunciations impact the performance of ASR systems?
Web-derived Prons. – Experiments and Results

– Extraction

• Manually select in which *Wiktionary* edition to search for pronunciations

• Our Automatic Dictionary Extraction Tool takes a vocab list with one word per line

• For each word, the matching *Wiktionary* page is looked up (e.g. http://fr.wiktionary.org/wiki/abandonner)

• If the page cannot be found, we iterate through all possible combinations of upper and lower case

• Each web page is saved and parsed for IPA notations:
  – Certain keywords in context of IPA notations help us to find the phonetic notation (e.g. `<span class="API" title="prononciation API">/a.bɑ̃.dɔ.nɛ/</span>`)
  – For simplicity, we only use the first phonetic notation, if multiple candidates exist
  – Our tool outputs the detected IPA notations for the input vocab list and reports back those words for which no pronunciation could be found
Web-derived Prons. – Experiments and Results – Quantity Check

- Quantity of pronunciations for *GlobalPhone* words

![Bar chart showing quantity of pronunciations for words in different languages.](image)

- Searched and found pronunciations for words in the *GlobalPhone* corpora

- Morphological variants in the word lists could also be find in Wiktionary

- French *Wiktionary* has highest match, possible explanations:
  - Strong French internet community (e.g. “Loi relative à l’emploi de la langue française”)
Web-derived Prons. – Experiments and Results
– Quantity Check

- Quantity of pronunciations for **proper names**
  - Proper names can be of diverse etymological origin and can surface in another language without undergoing the process of assimilation to the phonetic system of the new language (Llitjós and Black, 2002)
  
  → important as difficult to generate with letter-to-sound rules

- Search pronunciations of 189 international city names and 201 country names to investigate the coverage of proper names:

![Graph showing pronunciation coverage for city and country names in English, French, German, and Spanish]
Web-derived Prons. – Experiments and Results – Quantity Check

• Quantity of pronunciations for **proper names**
  - Results of only those words that keep their original name in the target language:

![Bar chart showing the percentage of city and country names that keep their original name in different languages.](chart.png)

- **City Names**
  - English (25/149)
  - French (40/145)
  - German (35/144)
  - Spanish (17/136)

- **Country Names**
  - English (53/104)
  - French (52/58)
  - German (57/80)
  - Spanish (17/80)

# found prons. for country names that keep their original name
# names which keep the original name in target language
Web-derived Prons. – Experiments and Results
– Quality Check

- **Impact of new pronunciation variants on ASR Performance**

Approach I: Add all new *Wiktionary* pronunciations to *GlobalPhone* dictionaries and use them for training and decoding

(*System1*)

<table>
<thead>
<tr>
<th>No.</th>
<th>Language</th>
<th># pron.</th>
<th>% equal</th>
<th># new</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>French</td>
<td>114k</td>
<td>74%</td>
<td>30k</td>
</tr>
<tr>
<td>2</td>
<td>Spanish</td>
<td>2k</td>
<td>50%</td>
<td>1k</td>
</tr>
<tr>
<td>3</td>
<td>German</td>
<td>7k</td>
<td>28%</td>
<td>5k</td>
</tr>
<tr>
<td>4</td>
<td>English</td>
<td>12k</td>
<td>26%</td>
<td>9k</td>
</tr>
</tbody>
</table>

Amount of *GlobalPhone* pronunciations, percentage of identical *Wiktionary* pronunciations, and amount of new *Wiktionary* pronunciation variants

<table>
<thead>
<tr>
<th></th>
<th>WER baseline (%)</th>
<th>WER System1 (%)</th>
<th>rel. improv.*</th>
</tr>
</thead>
<tbody>
<tr>
<td>French</td>
<td>23.43%</td>
<td>23.25%</td>
<td>0.79%</td>
</tr>
<tr>
<td>English</td>
<td>21.51%</td>
<td>22.46%</td>
<td>-4.44%</td>
</tr>
<tr>
<td>German</td>
<td>21.60%</td>
<td>21.67%</td>
<td>-0.31%</td>
</tr>
<tr>
<td>Spanish</td>
<td>14.68%</td>
<td>14.42%</td>
<td>1.76%</td>
</tr>
</tbody>
</table>

Impact of using all *Wiktionary* pronunciations for training and decoding

*Improvements are significant at a significant level of 5%*

→ How to ensure that new pronunciations fit to training and test data?
Impact of new pronunciation variants on ASR Performance

Approach II: Use only those *Wiktionary* pronunciations in decoding that were chosen in training (*System2*)

- *Wiktionary* pronunciations chosen in training during forced alignment are of good quality for training data
- Assumption: Similarity of training and test data in speaking style and vocabulary

<table>
<thead>
<tr>
<th>Language</th>
<th># wikt prons</th>
<th>% wikt prons</th>
</tr>
</thead>
<tbody>
<tr>
<td>French</td>
<td>3,000</td>
<td>10.11%</td>
</tr>
<tr>
<td>English</td>
<td>845</td>
<td>9.86%</td>
</tr>
<tr>
<td>German</td>
<td>1,439</td>
<td>27.02%</td>
</tr>
<tr>
<td>Spanish</td>
<td>259</td>
<td>22.90%</td>
</tr>
</tbody>
</table>

Amount and percentage of *Wiktionary* pronunciations selected in training

<table>
<thead>
<tr>
<th>Language</th>
<th>WER baseline</th>
<th>WER System1</th>
<th>rel. improv.*</th>
<th>WER System2</th>
<th>relative improvement*</th>
</tr>
</thead>
<tbody>
<tr>
<td>French</td>
<td>23.43%</td>
<td>23.25%</td>
<td>0.79%</td>
<td>23.16%</td>
<td>1.17%</td>
</tr>
<tr>
<td>English</td>
<td>21.51%</td>
<td>22.46%</td>
<td>-4.44%</td>
<td>23.39%</td>
<td>-8.76%</td>
</tr>
<tr>
<td>German</td>
<td>21.60%</td>
<td>21.67%</td>
<td>-0.31%</td>
<td>21.07%</td>
<td>2.44%</td>
</tr>
<tr>
<td>Spanish</td>
<td>14.68%</td>
<td>14.42%</td>
<td>1.76%</td>
<td>13.62%</td>
<td>7.22%</td>
</tr>
</tbody>
</table>

*Improvements are significant at a significant level of 5%
Web-derived Prons. – Conclusion

• We proposed an efficient data source from the WWW that supports the rapid pronunciation dictionary creation

• We developed an Automatic Dictionary Extraction Tool that automatically extracts phonetic notations in IPA from *Wiktionary*

• Best quantity check results: French *Wiktionary* (92.58% for *GlobalPhone* word list, 76.12% for country names, 30.16% for city names)

• Best quality check results: Spanish *Wiktionary* (7.22% relative word error rate reduction)

• Particular helpful for pronunciations of proper names

• Results depend on community and language support

• *Wiktionary* pronunciations improved all system but the English one
Overview – Automatic Speech Recognition

- **Front End (Preprocessing)**
- **Decoder (Search)**
- **Text**

- **Acoustic Model**
- **Lexicon / Dictionary**
- **Language Model**

- **Multilingual Bottle Neck Features**
Multilingual Bottle Neck Features

(based on Vu, Metze and Schultz, 2012)
Introduction

• Integration of Neural Network in ASR in different levels
• Multilayer Perceptron features e.g. Bottle-Neck features
• Many studies in multilingual and cross-lingual aspects e.g. K.Livescu (2007), C.Plahl (2011)
  → Some language-independent info can be learned
• How to initialize MLP training?
• How to train an MLP with very little training data?
• Idea:
  Apply multilingual MLP to MLP training for new languages
Bottle-Neck Features (BNF)

MFCC
13*11
= 143

LDA
42 dim

AM
Dictionary
LM
Bottle-Neck Features (BNF)

MFCC
13 * 11
= 143

Multilayer Perceptron (MLP)

1 4 3
4 1 5 0 0
2 4 1 5 0 0

LDA
42 dim

Dictionary

Bottle-Neck
42 * 5 = 210
Train a MLP with multilingual data
→ more robust due to amount of data
→ combine knowledge between languages
Initialize MLP training for a new language

MFCC

13 * 11

= 143

→ Select phones of target language from multilingual phone set based on IPA

→ All the weights and bias are used to initialize MLP training

→ What happens with **uncovered phones**?
Our idea: Extend the output layer to cover all phones in IPA

MFCC
13* 11
= 143

#phones in IPA

→ How to train weights and bias for the phones which do not appear in the training data?
“Open target language” MLP

→ **Our solution:** randomly select the data of the phones which have at least one articulatory feature of the new phone

MFCC

\[
\begin{align*}
13 \times 11 &= 143 \\
\end{align*}
\]

# phones in IPA
Experimental Setup

- Data corpus: GlobalPhone database
- Train a multilingual MLP with English (EN), French (FR), German (GE), and Spanish (SP)
- Integration BNF into EN, FR, GE and SP ASR
- Adapt rapidly to Vietnamese (VN):
  → Using all 22h of training data
  → Using only ~2h of training data
## Experimental Setup

### Frame Accuracy on Cross-validation data for MLP Training

<table>
<thead>
<tr>
<th></th>
<th>EN</th>
<th>FR</th>
<th>GE</th>
<th>SP</th>
</tr>
</thead>
<tbody>
<tr>
<td>RandomInit</td>
<td>70.98</td>
<td>76.73</td>
<td>63.93</td>
<td>71.75</td>
</tr>
<tr>
<td>MultiLingInit</td>
<td>73.46</td>
<td>78.57</td>
<td>68.87</td>
<td>74.02</td>
</tr>
</tbody>
</table>

### WER on GlobalPhone database

<table>
<thead>
<tr>
<th></th>
<th>EN</th>
<th>FR</th>
<th>GE</th>
<th>SP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>11.5</td>
<td>20.4</td>
<td>10.6</td>
<td>11.9</td>
</tr>
<tr>
<td>BNF.RandomInit</td>
<td>11.1</td>
<td>20.3</td>
<td>10.5</td>
<td>11.6</td>
</tr>
<tr>
<td>BNF.MultiLingInit</td>
<td>10.2</td>
<td>20.0</td>
<td>9.7</td>
<td>11.2</td>
</tr>
</tbody>
</table>
Language Adaptation for Vietnamese (I)

Frame Accuracy on Cross-validation data for MLP Training and Syllable Error Rate (SyllER) for 22h Vietnamese ASR

<table>
<thead>
<tr>
<th></th>
<th>FrameAcc</th>
<th>SyllER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>-</td>
<td>12.0</td>
</tr>
<tr>
<td>BN.RandomInit</td>
<td>65.13</td>
<td>11.4</td>
</tr>
<tr>
<td>“Open target language” MLP</td>
<td>67.09</td>
<td>10.1</td>
</tr>
</tbody>
</table>
### Language Adaptation for Vietnamese (II)

Frame Accuracy on Cross-validation data for MLP Training and Syllable Error Rate (SyllER) for 2h Vietnamese ASR

<table>
<thead>
<tr>
<th></th>
<th>FrameAcc</th>
<th>SyllER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>-</td>
<td>26.0</td>
</tr>
<tr>
<td>BN.Multi.NoAdapt</td>
<td>37.23</td>
<td>25.3</td>
</tr>
<tr>
<td>BN.Multi.Adapt</td>
<td>57.54</td>
<td>22.8</td>
</tr>
<tr>
<td>“Open target language” MLP</td>
<td>58.32</td>
<td>21.6</td>
</tr>
</tbody>
</table>
Summary

- Multilingual MLP is a good initialization for MLP training
- We could save about 40% of the training time
- Using BNF from MLP initialized with multilingual MLP we could improve consistently ASR performance
- Up to 16.9% relative improvement by using multilingual BNF for adaptation to Vietnamese
Overview – Automatic Speech Recognition

Front End (Preprocessing) → Decoder (Search) → Text

- Acoustic Model
- Lexicon / Dictionary
- Language Model

Unsupervised training
Multilingual Unsupervised Training

(based on Vu, Kraus and Schultz 2010, 2011)
Problem Description

• Fast and efficient portability of existing speech technology to new languages is a practical concern

• Standard approach:
  – Collect large amount of speech data
  – Generate manual transcriptions
  – Train ASR system

→ Problem of time consumption and cost
  (especially generation of transcriptions)

• Idea:
  – Use existing recognizers to avoid effort of transcription generation
Motivation

- If we have a number of recognizers, why not use them to build additional recognizers for new languages with little effort?
  - 3 main components: acoustic model, language model, and dictionary
  - Language model ([VuSchlippe2010]) and dictionary ([SchlippeOchs2010]) can be built
  - In this work: concentration on acoustic model

- Acoustic Model: requires audio data with transcriptions
  - Audio data is easily available
  - Transcriptions are expensive, error-prone, time consuming...

→ Use unsupervised training approach
Unsupervised Training

- Standard approach for unsupervised training:
  - Decode untranscribed audio data
  - Select data with high confidence
  - Select appropriate confidence measure
  - Use selected data to train or adapt recognizer

- Requirements:
  - Need existing recognizer
    → multilingual unsupervised training
  - Reliable confidence scores
Multilingual Unsupervised Training

- Develop multilingual framework to generate transcriptions for the available audio data
Cross-Lingual Transfer

- Basic principle:
  Use acoustic models of language A (source) as acoustic models for language B (target)
Confidence Measure – Overview

- Indicate „sureness“ of a speech recognizer
- Word-based confidence measures calculated from a word lattice
- In this work:
  - **Gamma** = $\gamma$-probability of forward-backward algorithm
  - **A-stabil** = *acoustic stability* determines frequency of a word over several hypotheses

\[
\text{astabil score} = \frac{\#\text{occurrences (reference word)}}{h}
\]
Problem

- A-Stabil, gamma work well for well trained Acoustic Models (AM)
- But not for poorly estimated Ams
  → NO option for Confidence Threshold
Multilingual A-Stabil

\[ \text{multilingual astabil score} = \frac{\#\text{occurencies (reference word)}}{h} \]
Multilingual A-Stabil Performance

Performance of Multilingual A-Stabil
English Reference applied to Czech Data

- 1 Language – EN
- 2 Languages – EN, SP
- 4 Languages – EN, FR, GE, SP

WER in % vs. Confidence Score Threshold
Multilingual Framework – Overview
• Stopping criterion: less than 5% (relative) additional data is selected in an iteration
Cross Language Transfer

**Original CLT**
- Phoneme mapping $EN \rightarrow CZ$ (phone set of language $CZ$)
- Select acoustic model of $EN$ for each phoneme of $CZ$
- Context-independent acoustic model

**Modified CLT**
- Phoneme mapping $CZ \rightarrow EN$ (phone set of language $EN$)
- Map phonemes in dictionary
- Context-dependent acoustic model (with context of $EN$)
Cross Language Transfer – Comparison

- Comparison of original and modified cross language transfer (WER on Czech devSet)
  - Slavic languages

<table>
<thead>
<tr>
<th>Language</th>
<th>Original CLT</th>
<th>Modified CLT</th>
<th>Delta</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bulgarian</td>
<td>66.98%</td>
<td>61.04%</td>
<td>5.9%</td>
</tr>
<tr>
<td>Croatian</td>
<td>68.03%</td>
<td>57.19%</td>
<td>10.8%</td>
</tr>
<tr>
<td>Polish</td>
<td>67.68%</td>
<td>55.83%</td>
<td>11.8%</td>
</tr>
<tr>
<td>Russian</td>
<td>72.45%</td>
<td>64.26%</td>
<td>8.2%</td>
</tr>
</tbody>
</table>

- Resource rich languages

<table>
<thead>
<tr>
<th>Language</th>
<th>Original CLT</th>
<th>Modified CLT</th>
<th>Delta</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>87.35%</td>
<td>99.82%</td>
<td>-12.5%</td>
</tr>
<tr>
<td>French</td>
<td>84.52%</td>
<td>95.15%</td>
<td>-10.6%</td>
</tr>
<tr>
<td>German</td>
<td>75.30%</td>
<td>75.21%</td>
<td>0.1%</td>
</tr>
<tr>
<td>Spanish</td>
<td>85.42%</td>
<td>87.15%</td>
<td>-1.7%</td>
</tr>
</tbody>
</table>
Experiments – Slavic Languages

- WER development of Slavic languages over iterations (on Czech dev set)

<table>
<thead>
<tr>
<th>Iteration</th>
<th>base</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>boot1</th>
<th>boot2</th>
</tr>
</thead>
<tbody>
<tr>
<td>WER_BL</td>
<td>61.04</td>
<td>40.95</td>
<td>35.34</td>
<td>35.66</td>
<td>23.52</td>
<td>22.77</td>
</tr>
<tr>
<td>WER_HR</td>
<td>57.19</td>
<td>38.12</td>
<td>33.91</td>
<td>34.24</td>
<td>23.19</td>
<td>22.71</td>
</tr>
<tr>
<td>WER_PL</td>
<td>55.83</td>
<td>36.10</td>
<td>32.16</td>
<td>32.36</td>
<td>23.12</td>
<td>22.41</td>
</tr>
<tr>
<td>WER_RU</td>
<td>64.26</td>
<td>40.79</td>
<td>34.18</td>
<td>33.39</td>
<td>23.35</td>
<td>22.93</td>
</tr>
</tbody>
</table>

Czech baseline (supervised): 21.8% WER
Experiments – Resource Rich Languages

- WER development of resource rich languages over iterations (on Czech dev set)

<table>
<thead>
<tr>
<th>Iteration</th>
<th>base</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>boot1</th>
<th>boot2</th>
</tr>
</thead>
<tbody>
<tr>
<td>WER_EN</td>
<td>87.35</td>
<td>59.47</td>
<td>52.67</td>
<td>52.75</td>
<td>26.87</td>
<td>23.90</td>
</tr>
<tr>
<td>WER.FR</td>
<td>84.52</td>
<td>64.21</td>
<td>59.07</td>
<td>57.86</td>
<td>27.27</td>
<td><strong>23.30</strong></td>
</tr>
<tr>
<td>WER_GE</td>
<td>75.30</td>
<td>53.04</td>
<td><strong>50.77</strong></td>
<td>50.77</td>
<td>26.40</td>
<td>23.50</td>
</tr>
<tr>
<td>WER_SP</td>
<td>85.42</td>
<td>58.86</td>
<td>54.31</td>
<td>53.53</td>
<td>25.74</td>
<td>23.70</td>
</tr>
</tbody>
</table>

Czech baseline (supervised): 21.8% WER
Conclusion

- Multilingual a-stabil is robust toward poorly trained acoustic models
  → It is able to select reasonable adaptation data despite high WER

- Multilingual framework allows successful construction of a recognizer without using any transcribed training data

- Approach works for similar source languages as well as for different source languages
  → in both experiments the best recognizer came close to the baseline system
Thanks for your interest!