Multilingual Speech Processing

Chapters 2, 4, 6, 7
Outline

- Motivation and Introduction
  - Why Speech Processing in many Languages?
  - What are the Challenges and possible Solutions?
- LVCSR in many languages
- Truly Multilingual Speech Interfaces
  - Multilingual Acoustic Modeling
- Rapid Language Adaptation
  - Acoustic Models
  - Automatic Dictionary Generation
  - Language Models
  - Synthesis
- SPICE: web-based tools for rapid adaptation
Outline

- Motivation and Introduction
  - Why Speech Processing in many Languages?
  - What are the Challenges and possible Solutions?
- LVCSR in many languages
- Truly Multilingual Speech Interfaces
  - Multilingual Acoustic Modeling
- Rapid Language Adaptation
  - Acoustic Models
  - Automatic Dictionary Generation
  - Language Models
  - Synthesis
- SPICE: web-based tools for rapid adaptation
Speech Processing Systems

Phone set & Speech data

Pronunciation rules

Text data

AM

Lex

LM

NLP

MT

TTS

Input: Speech

Hello

Output: Speech & Text

สวัสดี ดรับ
Speech Recognition in 17 Languages

Word Error Rate [%]

Japanese 10
German 11.8
English 14
Thai 14
Korean 14
Ch-Mandarin 14.5
Turkish 16.9
French 18
Portuguese 19
Croatian 20
Spanish 20
Bulgarian 29
Russian 33.5
Afrikaans 20
Chinese 21.7
Arabic 23.4
Iraqi 29
Polyphone Numbers for 10 Languages

Phonotactics

The graph shows the number of sub-polyphones for 10 languages across different context widths. The x-axis represents the context width, while the y-axis shows the number of sub-polyphones. Each language is represented by a unique line and marker.

- Chinesisch
- Deutsch
- Französisch
- Kroatisch
- Portugiesisch
- Spanisch
- Türkisch
- Japanisch
- Englisch
- Koreanisch
Outline

- Motivation and Introduction
  - Why Speech Processing in many Languages?
  - What are the Challenges and possible Solutions?
- LVCSR in many languages
- Truly Multilingual Speech Interfaces
  - Multilingual Acoustic Modeling
- Rapid Language Adaptation
  - Acoustic Models
  - Automatic Dictionary Generation
  - Language Models
  - Synthesis
- SPICE: web-based tools for rapid adaptation
Truly Multilingual ASR System
Can we build a language independent ASR system?

- Language Independent Acoustic Modeling
  - International Phonetic Alphabet (IPA)
    - Simple to implement, easy to port to other languages
  - Fully data-driven procedure
    - Considers spectral properties and statistical similarities
    - Apply to context independent and dependent models

- Universal Language Model
  - Combine LMs of languages to allow code switching

- Approach:
  - Train language dependent and independent AM+LM
  - Evaluate in monolingual and multilingual mode
Goal: Speech recognition in a many different languages
Problem: Data collection is time consuming and expensive

Sound system
Speech data
(≥ 10 hours)

Pronunciation rules
Text data

AM
Lex
LM

ela /e/1/a/
eu /e/u/
sou /s/u/

eu sou
você é
ela é
Speech Recognition in Multiple Languages

Sound system
Speech data

Pronunciation rules

Text data

AM
Lex
LM

ela /e/l/a/
eu /e/u/
sou /s/u/

você é
ela é
Step 1:
- Uniform multilingual speech database
- Monolingual recognizers in many languages

Step 2:
- Combine monolingual acoustic models
- Share data across languages
Acoustic Model Combination

Sound production is human not language specific:

- International Phonetic Alphabet (IPA)

1) Build **Universal** sound inventory based on IPA
   485 sounds are reduced to 162 IPA-sound classes

2) Each sound class is represented by one “phoneme”
   which is trained through **data sharing** across languages
   - *m,n,s,l* occur in all languages
   - *p,b,t,d,k,g,f* and *i,u,e,a,o* occur in almost all languages
   - no sharing of triphthongs and palatal consonants
2) Each sound class is represented by one “phoneme” which is trained through **data sharing** across languages

- $m, n, s, l$ occur in all languages
- $p, b, t, d, k, g, f$ and $i, u, e, a, o$ occur in almost all languages
- no sharing of triphthongs and palatal consonants
Acoustic Model Combination

ML-Mix

Englisch EN_M-b

Türkisch TU_M-b

Deutsch DE_M-b

Koreanisch KO_M-b

ML-Sep

Englisch M-b

Türkisch M-b/TU

Deutsch M-b/DE

Koreanisch M-b/KO

ML-Tag
Acoustic Model Combination

<table>
<thead>
<tr>
<th>Language</th>
<th>Mono</th>
<th>ML-Tag7500</th>
<th>ML-Tag3000</th>
<th>ML-Mix3000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Croatian</td>
<td>27</td>
<td>30</td>
<td>32</td>
<td>35</td>
</tr>
<tr>
<td>Japanese</td>
<td>13</td>
<td>14</td>
<td>15</td>
<td>20</td>
</tr>
<tr>
<td>Spanish</td>
<td>28</td>
<td>30</td>
<td>32</td>
<td>37</td>
</tr>
<tr>
<td>Turkish</td>
<td>20</td>
<td>21</td>
<td>21</td>
<td>29</td>
</tr>
</tbody>
</table>
Outline

- Motivation and Introduction
  - Why Speech Processing in many Languages?
  - What are the Challenges and possible Solutions?
- LVCSR in many languages
- Truly Multilingual Speech Interfaces
  - Multilingual Acoustic Modeling
- Rapid Language Adaptation
  - Acoustic Models
  - Automatic Dictionary Generation
  - Language Models
  - Synthesis
- SPICE: web-based tools for rapid adaptation
Challenges

- Algorithms are language independent but require data
  - Dozens of hours audio recordings and corresponding transcriptions
  - Pronunciation dictionaries for large vocabularies (>100,000 words)
  - Millions of words written text corpora in various domains in question
  - Bilingual aligned text corpora

- BUT: Such data are only available in very few languages
  - Audio data ≤ 40 languages, Transcriptions take up to 40x real time
  - Large vocabulary pronunciation dictionaries ≤ 20 languages
  - Small text corpora ≤ 100 languages, large corpora ≤ 30 languages
  - Bilingual corpora in very few language pairs, pivot mostly English

- Additional complications:
  - Combinatorical explosion (domain, speaking style, accent, dialect, ...)
  - Few native speakers at hand for minority (endangered) languages
  - Languages peculiarities, social factors
Solution: Learning Systems

⇒ Intelligent systems that learn a language from the user

- Efficient learning algorithms for speech processing
  - **Learning:**
    - Interactive learning with user in the loop
    - Statistical modeling approaches
  - **Efficiency:**
    - Reduce amount of data (save time and costs): by a factor of 10
    - Speed up development cycles: days rather than months

⇒ Rapid Language Adaptation from universal models

- Bridge the gap between language and technology experts
  - Technology experts do not speak all languages in question
  - Native users are not in control of the technology
Speech Processing:
Interactive Creation and Evaluation toolkit

• National Science Foundation, Grant 10/2004, 3 years
• Principle Investigator Tanja Schultz

• Bridge the gap between technology experts → language experts
  • Automatic Speech Recognition (ASR),
  • Machine Translation (MT),
  • Text-to-Speech (TTS)

• Develop web-based intelligent systems
  • Interactive Learning with user in the loop
  • Rapid Adaptation of universal models to unseen languages

• SPICE webpage http://cmuspice.org
CMU SPICE

Welcome to SPICE

Getting started
Welcome to SPICE homepage. In this page, you can develop the speech system (including Automatic Speech Recognition and Text-To-Speech) specific to your own language.

First thing to your attention:

Please ENABLE YOUR BROWSER TO ACCEPT COOKIES because in SPICE web site, we use cookies to identify different users and different projects. If your browser has not been enabled to accept cookies, then you cannot login and use SPICE at all.

Please input your name (Developer Name) and the name of your language (Language Name) and the name of the Project (Project Name) below. The basic idea is that each developer can work on several different languages, and in each language you can develop several different version of your speech system, differentiated by "Project Name".

You don't have to finish all the work at one time. For example, you can do Acoustic Model at one day, Language Model tomorrow ... etc. Every time you come back to SPICE page, Use your "Developer Name", "Language Name" and "Project Name", you can locate your previous work and continue to do the rest.

Developer Name:

Language Name:

Project Name:

Submit
Speech Processing Systems

Phone set & Speech data

Pronunciation rules

Text data

Input: Speech

Hello

Output: Speech & Text
Rapid Portability: Data

Phone set & Speech data

Input: Speech

AM

Lex

LM

NLP / MT

TTS

Output: Speech & Text

Hello

/hi /h//ai/ you /j/u/ we /w///i/ hi you you are I am

สวัสติ ดับ
Rapid Portability: Acoustic Models

Phone set & Speech data

Input: Speech

AM
Lex
LM

NLP
MT

TTS

Output: Speech & Text

Hello

สวัสดี ตรัพ
Speech Production is independent from Language  \[\Rightarrow\] IPA

1) IPA-based **Universal Sound Inventory**

2) Each sound class is trained by **data sharing**
   - Reduction from 485 to 162 sound classes
   - *m,n,s,l* appear in all 12 languages
   - *p,b,t,d,k,g,f* and *i,u,e,a,o* in almost all

**Problem:**
Context of sounds are language specific
How to train context dependent models for new languages?

**Solution:**
1) Multilingual Decision Context Trees
2) Specialize decision tree by Adaptation
Context Decision Trees

- Context dependent phones ($\pm n = \text{polyphone}$)
- Trainability vs Granularity
- Divisive clustering based on linguistic motivated questions
- One model is assigned to each context cluster
- Multi-lingual case: should we ignore language information?
- Depends on application
- Yes in case of adaptation to new target languages

Blaukraut
Brautkleid
Brotkorb
Weinkarte

-1=Plosiv?

k

lau k r a
ut k le
ot k or
in k ar

k (0)

+2=Vokal?

N
J

k (1)

ot k or

k (2)

ut k le

Context Decision Tree
Rapid Language Adaptation

Model mapping to the target language

1) Map the multilingual phonemes to Portuguese ones based on the IPA-scheme
2) Copy the corresponding acoustic models in order to initialize Portuguese models

Problem:
Contexts are highly language specific, how to apply context dependent models to a new target language

Solution:
1) Train a multilingual polyphone decision tree
2) Specialize this tree to target language using limited data
   (Polyphone decision tree specialization - PDTS)
English Polyphone Tree
Polyphone Decision Tree Specialization
Polyphone Decision Tree Specialization

Polyphones found in Portuguese
1. Tree Pruning
Select from all polyphones only the ones which are relevant for the particular language
2. Tree Regrowing
Further specialize the tree according to the adaptation data.
Rapid Portability: Acoustic Model

Word Error rate [%]

Ø Tree  ML-Tree  Po-Tree  PDTS

<table>
<thead>
<tr>
<th>Time</th>
<th>Ø Tree</th>
<th>ML-Tree</th>
<th>Po-Tree</th>
<th>PDTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>69.1</td>
<td>57.1</td>
<td>49.9</td>
<td>40.6</td>
</tr>
<tr>
<td>0:15</td>
<td>57.1</td>
<td>49.9</td>
<td>40.6</td>
<td>32.8</td>
</tr>
<tr>
<td>0:25</td>
<td>49.9</td>
<td>40.6</td>
<td>32.8</td>
<td>28.9</td>
</tr>
<tr>
<td>0:25</td>
<td>40.6</td>
<td>32.8</td>
<td>28.9</td>
<td>19.6</td>
</tr>
<tr>
<td>1:30</td>
<td>32.8</td>
<td>28.9</td>
<td>19.6</td>
<td>19</td>
</tr>
<tr>
<td>16:30</td>
<td>19</td>
<td>19</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### CMU SPICE

**Phoneme set specification**

This is a tool which will display all IPA phonemes. As a naive user, you can choose and give names to phonemes you wish your Speech Engine to use. After you have finished, you can click the "Submit" button to create the new acoustic model on the fly.

**Current PhoneMapFile**

Consonants (Pulmonic): Please choose the consonant sounds you’d like to have in your new acoustic models by giving it a name in the textbox next to it.

<table>
<thead>
<tr>
<th></th>
<th>Bilabial</th>
<th>Labiodental</th>
<th>Dental</th>
<th>Alveolar</th>
<th>Postalveolar</th>
<th>Retroflex</th>
<th>Palatal</th>
<th>Velar</th>
<th>Uvular</th>
<th>Pharyngeal</th>
<th>Glottal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plosive</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>p</td>
<td>p</td>
<td>p2</td>
<td>p'</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>b</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Rapid Portability: Pronunciation Dictionary

**Pronunciation rules**

```
"adios"  →  /a/ /d/ /i/ /o/ /s/
"Hallo"  →  /h/ /a/ /l/ /o/ → ???
```

**Input:** Speech

```
Hello
```

**Output:** Speech & Text

```
สวัสดี  ครับ
```

```
AM
Lex
LM
NLP / MT
TTS
```
Follow the work of Davel&Barnard

* Word list: extract from text

* G-2-P
  - explicit map rules
  - neural networks
  - decision trees
  - instance learning (grapheme context)

* Update after each $w_i$ → effective training
Currently you are logging in as tanja, you are working on language German, Project Name = 1

We use an iterative procedure to gather information to create dictionary for you.

First of all, please input an initial Grapheme to Phoneme (G2P) rule of your language.

Based on this rule, our system will "guess" the correct pronunciation of words in your language. You are able to view the predicted pronunciation, change it, delete it, or type a correct pronunciation for this word. The correct pronunciation will be saved into your dictionary and our system will make use of this information to make a better "guess" in predicting pronunciation of new words.

Now please type in Grapheme to Phoneme rule (G2P) for us. Just type one of the most common pronunciation for each grapheme. Thanks.

e:  e  ○ uppercase  ○ lowercase  ○ punctuation mark  ○ number  ○ others

n:  n  ○ uppercase  ○ lowercase  ○ punctuation mark  ○ number  ○ others

r:  r  ○ uppercase  ○ lowercase  ○ punctuation mark  ○ number  ○ others

i:  i  ○ uppercase  ○ lowercase  ○ punctuation mark  ○ number  ○ others

t:  t  ○ uppercase  ○ lowercase  ○ punctuation mark  ○ number  ○ others

s:  s  ○ uppercase  ○ lowercase  ○ punctuation mark  ○ number  ○ others

a:  a  ○ uppercase  ○ lowercase  ○ punctuation mark  ○ number  ○ others
Here's a list of all phonemes you selected in AM page:

p b t d k g m n r f v s z x i u e o a

11.35593220339% Finished
new word:

Merck

system suggested pronunciation: \texttt{m\,e\,r\,x\,k} \hspace{1em} \textbf{Listen to it} \hspace{1em} \textbf{Accept Pronunciation}

If you want to skip this word and work on it later, please click

\hspace{2em} \textbf{Skip this word}

If you don't think it's a valid word in your language, please click

\hspace{2em} \textbf{Remove this word}
Issues and Challenges

- How to make best use of the human?
  - Definition of successful completion
  - Which words to present in what order
  - How to be robust against mistakes
  - Feedback that keeps users motivated to continue

- How many words to be solicited?
  - G2P complexity depends on the language (SP easy, EN hard)
  - 80% coverage hundred (SP) to thousands (EN)
  - G2P rule system perplexity

<table>
<thead>
<tr>
<th>Language</th>
<th>Perplexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>50.11</td>
</tr>
<tr>
<td>Dutch</td>
<td>16.80</td>
</tr>
<tr>
<td>German</td>
<td>16.70</td>
</tr>
<tr>
<td>Afrikaans</td>
<td>11.48</td>
</tr>
<tr>
<td>Italian</td>
<td>3.52</td>
</tr>
<tr>
<td>Spanish</td>
<td>1.21</td>
</tr>
</tbody>
</table>
Rapid Portability: LM

Resource rich languages ↔ Resource low languages:

Internet

TV

Inquiry

Automatic Extraction

LM

Bridge Languages

Text data

Hello

Input: Speech

AM

Lex

LM

NLP / MT

TTS

Output: Speech & Text
Crosslingual LM Adaptation

Build language model

Build a language model using either a web 'crawler' or by uploading a file

We will try to grab text data from internet and build language model from those text data.

Given a link to some webpage composed of your own language, our web spider can automatically go into this web page, grab all text from this page, and also follow the hyperlink inside this page then go to another linked page to collect more text data. We need huge text data in order to build a robust language model. As we often said in speech research community, "no data is like more data"

Please provide us a URL of webpage which is composed by your language: 

Please specify how many levels would you like our web spider to get into:

- 1
- 2

Please verify that the URL you gave above is correct and valid, select the level you want the web spider to get. The deeper the value is, the longer it takes to finish the task. So our suggestion is to select a smaller value unless you care about the amount of training text more than the time to finish the task.

Another way is to submit a text file from your local machine. This file should contains large amount of text of your language. We can train Language Model from this text as well. But if you want to submit a text file, please make it such that every line contains a single sentence. This is the standard input format our LM toolkit will take.

You can also train LM via both method. Just input both the URL and the file name. Our toolkit will use both information.

Choose a file to upload: 

Browse...
Rapid Portability: TTS

Phone set & Speech data

Input: Speech

AM
Lex
LM
NLP
MT
TTS

Output: Speech & Text

Hello

สวัสดี คุณ
Parametric TTS

- Text-to-speech for G2P Learning:
  - Technique: phoneme-by-phoneme concatenation, speech not natural but understandable (Marelie Davel)
  - Units are based on IPA phoneme examples
    - PRO: covers languages through simple adaptation
    - CONS: not good enough for speech applications

- Text-to-speech for Applications:
  - Common technologies
    - Dipphone: too hard to record and label
    - Unit selection: too much to record and label
  - New technology: *clustergen* trajectory synthesis
    - Clusters representing context-dependent allophones
    - PRO: can work with little speech (10 minutes)
    - CONS: speech sounds buzzy, lacks natural prosody
Manual speaker selection

⇒ For all languages monolingual TTS performs best
⇒ Multilingual Models perform well …
    … only if knowledge about language is preserved (Multi+)
    (only small amount of sharing actually happens)
SPICE: Afrikaans – English

- Goal: Build Afrikaans – English Speech Translation System with SPICE
  - Cooperation with University Stellenbosch and ARMSCOR
  - Bilingual PhD visited CMU for 3 months
  - Afrikaans: Related to Dutch and English, g-2-p very close, regular grammar, simple morphology

- SPICE, all components apply statistical modeling paradigm
  - ASR: HMMs, N-gram LM (JRTk-ISL)
  - MT: Statistical MT (SMT-ISL)
  - TTS: Unit-Selection (Festival)
  - Dictionary: G-2-P rules using CART decision trees

- Text: 39 hansards; 680k words; 43k bilingual aligned sentence pairs;
  Audio: 6 hours read speech; 10k utterances, telephone speech (AST)
- **Good results:** ASR 20% WER; MT A-E (E-A) Bleu 34.1 (34.7), Nist 7.6 (7.9)
- **Shared pronunciation dictionaries** (for ASR+TTS) and LM (for ASR+MT)
- **Most time consuming process:** data preparation → reduce amount of data!
- **Still too much expert knowledge required** (e.g. ASR parameter tuning!)

### Time Effort

<table>
<thead>
<tr>
<th></th>
<th>Data</th>
<th>Training</th>
<th>Tuning</th>
<th>Evaluation</th>
<th>Prototype</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM (ASR)</td>
<td>5</td>
<td>8</td>
<td>7</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Lex</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LM (ASR, MT)</td>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TM (MT)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TTS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S-2-S</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Conclusion

- Intelligent systems to learn language
  - SPICE: Learning by interaction with the (naive) user
  - Rapid Portability to unseen languages
- Truly Multilingual Systems
  - Systems and data in multiple languages
  - Universal language independent models
- Other Applications and Issues in Multilinguality
  - Speech translation (see MT Vorlesung)
  - Multilingual Interfaces
    - Code switching, non-native / accented speech
  - Alternative model units (AF, Graphemes)