Pronunciation Modeling

Tim Schlippe

With material from Hua Yu, Michael Finke, William Barry, Paolo Baggia, Dong Wang, Martine Adda-Decker, Lori Lamel and Juergen Schroeter

05 June 2012
Outline (I)

I. What is Pronunciation Modeling?

II. Graphemes and Phonemes
   - Graphemes
   - Phonemes
     • Pronunciation Alphabets
     • International Phonetic Alphabet
   - Relationship between Graphemes and Phonemes

III. Vocabulary
   - What is a Word?
   - Segmentation
   - Text Normalization
Outline (II)

IV. Pronunciation Dictionary Production
   - Manual
   - Statistic-based (Sequitur G2P)
   - Rule-based (Lex Learner)

VI. Pronunciation Variations

VII. Challenges
   • Proper Names
   • Automatic Pronunciation Dictionary Generation from the WWW
I. What is Pronunciation Modeling?

- Word sequence
  - Pronunciation lexicon
  - Phone sequence
  - Phonetic decision tree
  - State/model sequence

Pronunciation model
I. What is Pronunciation Modeling?

- Purpose of text collection: to learn how a language is written → language model may be constructed
- Purpose of audio collection: to learn how a language is spoken → acoustic models may be constructed

Job of pronunciation modeling is to connect these two realms

- The Pronunciation Model consists of these components:
  1. a **definition of the elemental sounds** in a language (usually phonemes).
  2. a **dictionary** that describes how words in a language are pronounced.
  3. **post-lexical rules** for altering pronunciation of words spoken in context.
I. Pronunciation Modeling – Pronunciation Dictionary

- Pronunciation dictionary: Link between the acoustic and the language models

Language-dependent resources for transcription systems
[according to Martine Adda-Decker and Lori Lamel]
I. Pronunciation Modeling – Pronunciation Dictionary

- Pronunciation Dictionary = Lexicon
- Question: how do we connect our HMMs to words?
  - with a pronunciation dictionary
    e.g. CMU DICT [www.speech.cs.cmu.edu/cgi-bin/cmudict](http://www.speech.cs.cmu.edu/cgi-bin/cmudict)

- Simplest architecture
  - each HMM state represents one phoneme in the language
  - a word is represented as a sequence of phonemes
  - Dictionaries come with different phone sets (in English usually 38-45)
  - CMU DICT defines 39 (plus SIL= silence) for American English

Example:
(excerpt from CMU DICT)

<table>
<thead>
<tr>
<th>WE</th>
<th>W IY</th>
</tr>
</thead>
<tbody>
<tr>
<td>WHE</td>
<td>W EH</td>
</tr>
<tr>
<td>WILL</td>
<td>W IH L</td>
</tr>
<tr>
<td>WIN</td>
<td>W IH N</td>
</tr>
</tbody>
</table>
I. Pronunciation Modeling – Pronunciation Dictionary

- Pronunciation Dictionary, e.g. Janus Dictionary

<table>
<thead>
<tr>
<th>Word</th>
<th>Pronunciation</th>
</tr>
</thead>
<tbody>
<tr>
<td>blue</td>
<td>{b WB} l {uw WB}</td>
</tr>
<tr>
<td>blueberries</td>
<td>{b WB} l uw b eh r iy {z WB}</td>
</tr>
<tr>
<td>blueberry</td>
<td>{b WB} l uw b eh r {iy WB}</td>
</tr>
<tr>
<td>bluebird</td>
<td>{b WB} l uw b er {d WB}</td>
</tr>
</tbody>
</table>

- If pronunciations in dictionary deviate sharply from spoken language:
  - Creation of models with great variance
  - Similar data with essentially same phonemes are distributed to different models
II. Graphemes

• Fundamental unit in written language

• Graphemes include …
  – alphabetic letters
  – Chinese characters
  – numerical digits
  – punctuation marks
  – ...

ภาษาไทย
Thai

北方話
Chinese

العربية
Arabic
II. Graphemes

- Approaches have been shown that only require orthographic representation of the vocabulary list rather than the pronunciation for each word
  [Mirjam Killer, Sebastian Stüker and Tanja Schultz: Grapheme based Speech Recognition]

- Advantages:
  - No linguistic knowledge needed
  - No further resources other than the audio training material and their transcripts are needed
  - Works good for languages for which writing systems provide some kind of a grapheme-phoneme relation

- Disadvantage:
  - Works bad for languages with few or no kind of a grapheme-phoneme relation
II. Phonemes

- Smallest linguistically distinctive unit of sound

- Selection of phonemes (e.g. from standard IPA chart)

  $\phi$  $\ddot{\sigma}$

- Names for the phonemes
  - Can match a pronunciation dictionary (if one exists)
  - Can match the user’s familiarity

- Audio feedback helpful
  - Click to hear recording of each phone
II. Graphemes and Phonemes

- In phonemic orthography: a grapheme corresponds to one phoneme

- In non-phonemic orthography:
  - Multiple graphemes may represent a single phoneme
    - digraphs (two graphemes for a single phoneme), e.g. "sh" in English
    - trigraphs (three for a single phoneme), e.g. "sch" in German
    - ...
  - A single grapheme can represent multiple phonemes, e.g. "box" has three graphemes, but four phonemes: /bɒks/, 北方話 (Běifānghuà „Northern Chinese Language“)
II. Phonemes – Phoneme Set Specification in SPICE / RLAT

http://csl.ira.uka.de/rlat-dev
II. Phonemes - Pronunciation Alphabets

- Proprietary phonetic alphabets:
  - Each speech technology provider supports several phonetic alphabets

- General/standard phonetic alphabets:
  - SAMPA
  - ASCII based (simple to write)
  - IPA – International Phonetic Alphabet
    “universally agreed system of notation for sounds of languages”
    - Covers all languages
    - Requires UNICODE to write it

- Other well-known phonetic alphabets:
  - Chinese Mandarin: Pinyin
  - Worldbet (James L. Hieronymus)
  - Arpabet (Speech Understanding Project)
II. International Phonetic Alphabet

• Introduction

– Created by International Phonetic Association (active since 1896), collaborative effort by all the major phoneticians around the world


– IPA – International Phonetic Alphabet is largely used by phoneticians, by dictionaries and phonetic resources → W3C chose to normatively reference IPA in Speech Synthesis Markup Language (SSML) and PLS (Playlists) specifications

For more details on IPA: http://www.langsci.ucl.ac.uk/ipa and to listen to sounds from languages: http://phonetics.ucla.edu/index/sounds.html
II. International Phonetic Alphabet

- **IPA Full Chart**
  - Describes the phonemes that cover all the world languages:
    - Consonants
    - Vowels
    - Other Symbols
    - Diacritics
    - Suprasegmental
    - Tones and Word Accent
  - IPA is used by phoneticians for broad and narrow transcriptions
  - IPA is used in many dictionaries
### II. International Phonetic Alphabet

- **Consonants (some)**

All these are possible Pulmonic Consonants
- The columns are “places of articulation”
- The rows are “manner of articulation”
- The gray areas are considered to be impossible to articulate
II. International Phonetic Alphabet

- Vowels:
  - A speech sound created by the relatively free passage of breath through the larynx and oral cavity, usually forming the most prominent and central sound of a syllable

- Vowel are distinguished on the basis of “Height” and “Backness”

- The IPA diagram resembles the place of articulation of the vowels

VOWELS

Front   Central   Back

Close  i  y ——>  i  u ——>  u

Close-mid  e  φ ——>  θ  θ ——>  y  o

Open-mid  e  θ ——>  θ ——>  y  o

Open  a  ae ——>  a ——>  a

Where symbols appear in pairs, the one to the right represents a rounded vowel.
II. International Phonetic Alphabet

- **Diacritics**
  - Small marks that can be added to a symbol to modify its value
  - Used to differentiate allophones (different sounds but same functionality) of a phoneme
  - They are very important for narrow transcriptions, which shows more phonetic details

<table>
<thead>
<tr>
<th>DIACRITICS</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \ddot{\text{o}} )</td>
<td>Voiceless</td>
</tr>
<tr>
<td>( \ddot{\text{i}} )</td>
<td>Breathy voiced</td>
</tr>
<tr>
<td>( \ddot{\text{h}} )</td>
<td>Dental</td>
</tr>
<tr>
<td>( \tilde{\text{w}} )</td>
<td>Voiced</td>
</tr>
<tr>
<td>( \tilde{\text{w}} )</td>
<td>Creaky voiced</td>
</tr>
<tr>
<td>( \tilde{\text{h}} )</td>
<td>Apical</td>
</tr>
<tr>
<td>( \tilde{\text{w}} )</td>
<td>Linguolabial</td>
</tr>
<tr>
<td>( \tilde{\text{w}} )</td>
<td>Laminal</td>
</tr>
<tr>
<td>( \ddot{\text{w}} )</td>
<td>More rounded</td>
</tr>
<tr>
<td>( \ddot{\text{w}} )</td>
<td>Labialized</td>
</tr>
<tr>
<td>( \ddot{\text{w}} )</td>
<td>Nasalized</td>
</tr>
<tr>
<td>( \ddot{\text{w}} )</td>
<td>Nasal release</td>
</tr>
<tr>
<td>( \ddot{\text{w}} )</td>
<td>Lateral release</td>
</tr>
<tr>
<td>( \ddot{\text{w}} )</td>
<td>No audible release</td>
</tr>
<tr>
<td>( \ddot{\text{w}} )</td>
<td>Centralized</td>
</tr>
<tr>
<td>( \ddot{\text{w}} )</td>
<td>Volarized or pharyngealized</td>
</tr>
<tr>
<td>( \ddot{\text{w}} )</td>
<td>Mid-centralized</td>
</tr>
<tr>
<td>( \ddot{\text{w}} )</td>
<td>Raised</td>
</tr>
<tr>
<td>( \ddot{\text{w}} )</td>
<td>Voiced alveolar fricative</td>
</tr>
<tr>
<td>( \ddot{\text{w}} )</td>
<td>Lowered</td>
</tr>
<tr>
<td>( \ddot{\text{w}} )</td>
<td>Voiced bilabial approximant</td>
</tr>
<tr>
<td>( \ddot{\text{w}} )</td>
<td>Advanced Tongue Root</td>
</tr>
<tr>
<td>( \ddot{\text{w}} )</td>
<td>Retracted Tongue Root</td>
</tr>
</tbody>
</table>
II. International Phonetic Alphabet

- Suprasegmentals and Tones

Suprasegmentals

- Aspects of speech that involve more than single phonemes
- The principal features are stress, length, tone and intonation (Sprachmelodie)

Tones and Word Accents

- Pitch variations that affect the meaning of word
  - i.e., /ma/ in Chinese Mandarin may mean “mother”, “hemp”, “horse”, or “scold”, by changing tone from “high level”, “low level”, “rising”, and “going”
III. Vocabulary

• How to select a vocabulary list?

• Recognizer vocabularies are generally defined as the $N$ most frequent words in training texts
  \[ \implies \text{guarantees optimal lexical coverage on the training data} \]

• Training text materials should be closely related (in time and topics) to the test data:
  
  – Living languages: Usage of a word can decay and increase with time and completely new items may appear (neologism and proper names) (e.g. internet, cyper café, …)
III. Vocabulary – What is a Word?

• There are languages with no established writing conventions (e.g. in Amharic, any written form that reflects what is said is acceptable) → orthographic variability and its manifestation at acoustic level are challenges

• Lexical word list generally consists of a simple list of lexical items observed in running text

• Depending on the form of the training texts, different word lists can be derived
III. Vocabulary – Segmentation

เธอกำลังไปธนาคาร  (ter gum-lang pai ta-nar-karn)
She is going to a bank.

- Some written languages such as Chinese, Japanese and Thai do not have single-word boundaries
- For dictionary production definition and identification of word boundaries are required → Text segmentation
- Also compound splitting to reduce number of words in dictionary.
Mrs. Green is a member of the Greens Garden Club.
Bob Green’s car is green.
The Greens all like eating greens.
Green’s her favorite color.

<table>
<thead>
<tr>
<th>Normalization 1</th>
<th>Normalization 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>graphemic form</td>
<td>phonemic form</td>
</tr>
<tr>
<td>green</td>
<td>grin</td>
</tr>
<tr>
<td>Green</td>
<td>grin</td>
</tr>
<tr>
<td>Green’s</td>
<td>grinz</td>
</tr>
<tr>
<td>greens</td>
<td>grinz</td>
</tr>
<tr>
<td>Greens</td>
<td>grinz</td>
</tr>
</tbody>
</table>

[according to Martine Adda-Decker and Lori Lamel]

- Reduce OOV (Out-of-Vocabulary) rate
III. Vocabulary – Text Normalization

• Example for French normalization:

\[ N_0: \text{processing of ambiguous punctuation marks (hyphen -, apostrophe ’) not including compounds} \]
\[ N_1: \text{processing of capitalized sentence starts} \]
\[ N_2: \text{digit processing (110 → cent dix)} \]
\[ N_3: \text{acronym processing (ABCD → A. B. C. D.)} \]
\[ N_4: \text{decompounding (arc-en-ciel → arc en ciel)} \]
\[ N_5: \text{no case distinction (Paris → paris)} \]
\[ N_6: \text{no diacritics (énervé → enerve)} \]

„Decompounding“ rules

Keep word boundaries unchanged but reduce intraword graphemic variability

[according to Martine Adda-Decker and Lori Lamel]
III. Vocabulary – Text Normalization

• Language-independent and –specific text normalization in RLAT:

<table>
<thead>
<tr>
<th>Language-independent Text Normalization (<em>LI-rule</em>)</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 (<em>LS-rule</em>)</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>
III. Vocabulary – Text Normalization

- If high lexical variety is measured with a large proportion of long units, there are several reasons to consider reducing the variability:
  - Smaller units will provide better lexical coverage
  - Easier development of pronunciations
  - More efficient spelling normalization
  - More reliable N-grams estimated for language modeling

- Compound splitting:

  [according to Martine Adda-Decker and Lori Lamel]

<table>
<thead>
<tr>
<th>$k$</th>
<th>$W_{beg}(k)$</th>
<th>$#W_{end}(k)$</th>
<th>$#Sc(k)$</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>K</td>
<td>147731</td>
<td>62</td>
<td>Klasse, Kopf, Kritik, Kind, Köln, Kurs…</td>
</tr>
<tr>
<td>2</td>
<td>Ka</td>
<td>29068</td>
<td>41</td>
<td>Kampf, Kanzler, Kairo, Kauf, Kappe…</td>
</tr>
<tr>
<td>3</td>
<td>Kap</td>
<td>2131</td>
<td>25</td>
<td>Kapuze, Kapriolen, Kapitän, …</td>
</tr>
<tr>
<td>4</td>
<td>Kapi</td>
<td>1281</td>
<td>14</td>
<td>Kapielski, Kapillaren, Kapitel…</td>
</tr>
<tr>
<td>5</td>
<td>Kapit</td>
<td>1218</td>
<td>8</td>
<td>Kapitänt, Kapitel, Kapitulation, Kapitol…</td>
</tr>
<tr>
<td>6</td>
<td>Kapita</td>
<td>974</td>
<td>4</td>
<td>Kapital, Kapitain, Kapitän…</td>
</tr>
<tr>
<td>7</td>
<td>Kapital</td>
<td>968</td>
<td>27</td>
<td>Kapitalismus, Kapitals, K-erhöhung…</td>
</tr>
</tbody>
</table>

Table 1.3. Given a word start $W_{beg}(k)$ of length $k$, the number of character successors $\#Sc(k)$ generally tends towards zero with $k$. A sudden increase of $\#Sc(k)$ indicates a boundary due to compounding. $\#W_{end}(k)$ indicates the number of words in the vocabulary sharing the same word start.
IV. Pronunciation Dictionary Production

Language-independent processing steps for pronunciation dictionary generation [according to Martine Adda-Decker and Lori Lamel]

- Designing pronunciation dictionaries:
  1. Definition of words in the particular target language
  2. Selection of a finite set of words
  3. Determining how each of these words is pronounced
IV. Pronunciation Dictionary Production

Pronunciation dictionary development for ASR systems [according to Martine Adda-Decker and Lori Lamel]
IV. Pronunciation Dictionary Production

- Statistical
  - Supervised (Sequitur G2P)
  - Unsupervised
- Rule-based
  - Manual (linguists)
  - Automatical / Semi-automatical (LexLearner)
IV. Pronunciation Dictionary Production

• Completely manual
  
  – Developer (often expert in linguistics or phonetics) types in the phone sequence for each lexical entry
  
  → Only viable for relatively small vocabulary tasks and poses problem of pronunciation consistency

• Manually supervised
  
  – Given an existing dictionary, rules are used to infer pronunciations of new entries
  
  → Requires reasonably sized starting dictionary and is mainly useful to provide pronunciations for inflected forms and compound words
IV. Pronunciation Dictionary Production

- Grapheme-to-phoneme rules:
  - Usually developed for speech synthesis
  - Ensure that text normalization is consistent with the pronunciation rules

- Manually supervised grapheme-to-phoneme rules:
  - Manual supervision is particularly important for languages with ambiguous written symbol sequences
  - For any language, proper names – particularly those of foreign origin – may not be properly spelled or may require multiple pronunciations
Pronunciation Dictionary Production – Pronunciation Generation Tool

Pronunciation Generation Tool
[according to Martine Adda-Decker and Lori Lamel]
IV. Pronunciation Dictionary Production – Lex Learner

- Rules-based
- Semi-automatical

http://csl.ira.uka.de/rlat-dev
IV. Pronunciation Dictionary Production – Sequitur G2P

- A trainable data-driven Grapheme-to-Phoneme converter developed at RWTH Aachen

- Statistical:
  - Apply graphone (or grapheme-phoneme joint-multigram) approach to the alignment problem,
    e.g. pronunciation of “speaking” may be regarded as a sequence of five graphones:

    \[
    \text{“speaking”} = s \ p \ \text{ea} \ k \ \text{ing} \\
    \text{[spi:kɪŋ]} = [s] \ [p] \ [i:] \ [k] \ [ɪŋ]
    \]

- Use standard language modeling techniques to model transcription probabilities
### V. Pronunciation Modeling – Pronunciation Variations

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>counting</td>
<td>ka⁵ntiŋka⁵nŋ</td>
</tr>
<tr>
<td>interest</td>
<td>IntrIst IntəIst InəIst</td>
</tr>
<tr>
<td>amortization</td>
<td>æmərtəζəSxn æmərtəζəSxn æmərtəζəSxn æmərtəζəSxn</td>
</tr>
<tr>
<td>company</td>
<td>kampəni kampni</td>
</tr>
<tr>
<td>coupon</td>
<td>kjupan kupan</td>
</tr>
<tr>
<td>excuse</td>
<td>Ekskjuz Ekskjus</td>
</tr>
</tbody>
</table>

Examples of alternate valid pronunciations for American English

[according to Martine Adda-Decker and Lori Lamel]
V. Pronunciation Modeling – Pronunciation Variations

- Problem:
  The same word can be pronounced differently in different situations due to:
  - Context (coarticulation effects)
  - Dialects (something vs. somephin, thang vs. thing, ...)
  - Various correct pronunciations (the + w/vowel, a = AE or A = EY, ...)
  - Correct pronunciation differs from most commonly used ones (February vs. Febyury, Monday vs. Mondy, ...)

- Common Solution:
  - Add multiple entries into the pronunciation lexicon
  - When building HMMs for a word, build multiple, alternative state sequences

Example:
(excerpt from real dictionary)

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>THE(1)</td>
<td>TH AH</td>
<td>WHEN(1)</td>
<td>W EH N</td>
</tr>
<tr>
<td>THE(2)</td>
<td>TH AX</td>
<td>WHEN(2)</td>
<td>HH W EH N</td>
</tr>
<tr>
<td>THE(3)</td>
<td>TH IY</td>
<td>WHEN(3)</td>
<td>W IH N</td>
</tr>
<tr>
<td></td>
<td></td>
<td>WHEN(4)</td>
<td>HH W IH N</td>
</tr>
</tbody>
</table>
V. Pronunciation Modeling – Pronunciation Variations

How to find Pronunciation Variants?

- Linguist or phonetician writes them down
  Problem: acoustic sounds do not necessarily correspond to linguistic units
- ASR-expert adds variants to the dictionary
  Problem: Often most successful approach but also very expensive one
- In both cases we might face a lot of inconsistencies

⇒ Write down rules and apply the rules to existing pronunciations
Rules could be e.g. \( \text{... ing} \rightarrow \text{... IX NG, ... AX NG, ... AE NG, ... IH N} \)

**Automatic Dictionary Learning** (e.g. Tilo Sloboda, 1996):
1) Train a phone based recognizer (vocabulary consists of phonemes)
2) Run recognition on word-based segmented training data
3) Single out the top-N most frequent phoneme sequences per word
   4) Add these sequences as pronunciation to the dictionary

Problems with this approach:

- accuracy of phone recognizer,
- requires word-based segmentation (how to get it without dictionary?)
- requires enough examples per word to find reliable candidates
V. Flexible Tree Clustering for Pronunciation Modeling

• Case Study: The Flapping of T

  – Pronunciation variants may be used to model the flapping of the T

    BETTER       B EH T   AXR
    BETTER(2)    B EH DX AXR

  – Another approach:
    Use only the main entry in the dictionary (single pronunciation) and let phonetic decision tree choose whether or not to flap T
V. Mode Dependent Pronunciation Modeling

- Pronunciation Model $p(r|w) = p(r|q)$

<table>
<thead>
<tr>
<th>Transcription</th>
<th>I</th>
<th>AM</th>
<th>GOING</th>
<th>TO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Canonical Transcription</td>
<td>AY</td>
<td>AEM</td>
<td>GOW</td>
<td>IX</td>
</tr>
<tr>
<td>Surface Form</td>
<td>AY</td>
<td>-</td>
<td>M</td>
<td>G</td>
</tr>
<tr>
<td></td>
<td>I’M</td>
<td>GONNA</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Mode Dependent Pronunciation Model $p(r|m,w)$

<table>
<thead>
<tr>
<th>Transcription</th>
<th>I</th>
<th>AM</th>
<th>GOING</th>
<th>TO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Canonical Transcription</td>
<td>AY</td>
<td>AEM</td>
<td>GOW</td>
<td>IX</td>
</tr>
</tbody>
</table>

| Surface Form | AY | - | M | G | AO | - | N | - | AX |
|---------------| I’M | GONNA |

| Word Lattice Acoustic Model | duration, speaking rate, F0 |

[http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.45.9489]
V. Mode Dependent Pronunciation Modeling

- Probability of encountering pronunciation variants is defined to be a function of the speaking style.
- This probability function is learned through decision trees from rule-based generated pronunciation variants.
V. Pronunciation weights

- Pronunciations variants may lead to confusability between words with similar pronunciations

- Idea: include information about probability of occurrence for pronunciations

- Applying pronunciation probabilities estimated via forced alignment of the training corpora

- Lead to improvement in Arabic ASR
  [Al-Haj et al. 2009: Pronunciation Modeling for Dialectal Arabic Speech Recognition]
VI. Challenges - Proper Names

• Pronunciations of 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)
  
  – Examples in German: Los Angeles, New York
  
  → difficult to generate with letter-to-sound rules
VI. Challenges – Web-derived Pronunciations

• Previous approaches
  – Manually
  – Letter-to-phoneme (L2P) rules (hand-crafted, machine-learned)

=> Expensive / can be of variable quality

• Novel strategy:
  Mining huge quantities of pronunciation info on the Web

• Ongoing work at Google

• Ongoing work at CSL
  [Schlippe et al., 2010: Wiktionary as a Source for Automatic Pronunciation Extraction]
VI. Automatic Pronunciation Dictionary Generation from the WWW

• Ongoing work at Google:
  – Derive IPA and ad-hoc transcriptions from the Google’s web and News page repositories (in English and from non-EU countries) [Ghoshal et al. 2009: Web-Derived Pronunciations]

  – Using IPA
    + unambiguously denoting English phoneme
    - IPA requires skills

  – Ad-hoc transcription
    + Based on simpler spelling
      (e.g. bruschetta pronounced broo-SKET-uh)
    + no special skills
    - no phonemic transcription
VI. Automatic Pronunciation Dictionary Generation from the WWW

- Ongoing work at CSL
  - Derive pronunciations using IPA field included in Wiktionary

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>
VI. Automatic Pronunciation Dictionary Generation from the WWW

Wiktionary Article Growth from Jan 2003 till Jan 2008
(meta.wikimedia.org/wiki/List of Wiktionaries)
sein

Inhaltsvorzeichen

| Verben | 1. sein (Deutsch)
|--------|----------------
| 1.1 Hilfsverb | sein (Deutsch)
| 1.11 Übersetzungen | sein (Deutsch)
| 1.2 Possessivpronomen, 3. Person Singular m, n | sein (Deutsch)
| 1.21 Übersetzungen | sein (Deutsch)

2. sein (Französisch)
2.1 Substantiv, m
2.11 Übersetzungen

sein (Deutsch) [Bearbeiten]

Hilfsverb [Bearbeiten]

Anmerkung:

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

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

Aussprache:
IPA: [zæn], [zain], [ters], [bär], [is], [zint], [seid: [zart]], Präteritum: [war], Partizip II: [ge- ze- sen]

sein (Französisch) [Bearbeiten]

Substantiv, m [Bearbeiten]

Silbentrennung:
sein, Plural: ser

Aussprache:
IPA: [sɛ], [sɛr], [seid: [sért]}

Horbeispiel: sein, Plural:
VI. Automatic Pronunciation Dictionary Generation from the WWW

Results of the Quantity Check for GlobalPhone words

<table>
<thead>
<tr>
<th>Language</th>
<th>Number</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>189</td>
<td>19.79%</td>
</tr>
<tr>
<td>French</td>
<td>201</td>
<td>93.27%</td>
</tr>
<tr>
<td>German</td>
<td>35</td>
<td>19.67%</td>
</tr>
<tr>
<td>Spanish</td>
<td>25</td>
<td>7.71%</td>
</tr>
</tbody>
</table>

Results of the Quantity Check for proper names

<table>
<thead>
<tr>
<th>Language</th>
<th>Number</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>189</td>
<td>19.79%</td>
</tr>
<tr>
<td>French</td>
<td>201</td>
<td>93.27%</td>
</tr>
<tr>
<td>German</td>
<td>35</td>
<td>19.67%</td>
</tr>
<tr>
<td>Spanish</td>
<td>25</td>
<td>7.71%</td>
</tr>
</tbody>
</table>
VI. Quality Check (ASR Performance)

<table>
<thead>
<tr>
<th>No.</th>
<th>Language</th>
<th># prons.</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>

Impact on ASR performance:
- Approach I: Using all Wiktionary pronunciations for training and decoding (System1)
- Approach II: Using only those Wiktionary pronunciations in decoding that were chosen in training (System2)

GlobalPhone dictionaries
... had been created in rule-based fashion, manually cross-checked (validated quality)
... contain phonetic notations based on IPA scheme
→ mapping between IPA units obtained from Wiktionary and GlobalPhone units is trivial (Schultz, 2002)

### Impact on ASR performance:

<table>
<thead>
<tr>
<th></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 significance level of 5%
VI. Quality Check (g2p Models)

**g2p model consistency check** (Generalization ability of the g2p models) within dictionary (in n-fold cross validation):

1. Extracting word-pronunciation pairs for testing
2. Extracting sets of word-pronunciation pairs (disjunctive to test sets)
3. Generating g2p models with training sets
4. Applying g2p models to words in test set
5. Computing phoneme error rates (PERs) between the new and the original pronunciations.
VI. Quality Check (g2p Models)

Consistency of GP

Better consistency with more training data

Model size increase comes with marginal consistency improvement

GP g2p model complexity

Consistency of GP saturation with model size increase

Model size

# phonemes
VI. Quality Check (g2p Models)

Consistency of Wikt

Better consistency with more training data

Wikt g2p model complexity

Model size increase comes with marginal consistency improvement

# phonemes
VI. Quality Check (g2p Models)

Consistency of GP and wiktOnGP

GP closer to validated GlobalPhone pron. than wiktOnGP for all languages
SA/BA/DA/MA: Web-derived Pronunciations

Aufgaben:
• Finden und Extrahieren von Aussprachen im WWW
• Sicherstellung der Qualität
• Auswertung des Einflusses auf Spracherkennungssysteme

Benötigte Kenntnisse:
• Grundlagenwissen Spracherkennung
• Programmierkenntnisse, z. B. in Perl oder PHP
• Spaß an Informatik und Linguistik

Ab sofort bei:
Tim Schlippe (tim.schlippe@kit.edu)
“Crowdsourcing represents the act of a company or institution taking a function once performed by employees and out-sourcing it to an undefined (and generally large) network of people in the form of an open call.“


Wie können uns die Internet-Technologie und Crowdsourcing bei der Entwicklung von Spracherkennungssystemen unterstützen?

Aufgaben:
• Konzepte von Crowdsourcing-Ansätzen für die Sammlung von Sprachressourcen (z.B. Texte vorzulesen, hochzuladen, Aussprachen für Wörter aufzuschreiben usw.) recherchieren und analysieren
• Implementieren einer Lösung
• Auswertung der Lösung

Benötigte Kenntnisse:
• Grundlagenwissen Spracherkennung
• Programmierkenntnisse, z.B. in Perl oder PHP
• Spaß an Informatik und Linguistik

Ab sofort bei:
Tim Schlippe (tim.schlippe@kit.edu)


Pronunciation Modeling

The chance that e happens. For example, if e is the English string "I like snakes," then P(e) is the chance that a certain person at a certain time will say "I like snakes" as opposed to saying something else.

P(f | e) -- conditional probability

The chance of f given e. For example, if e is the English string "I like snakes," and if f is the French string "maison bleue," then P(f | e) is the chance that upon seeing e, a translator will produce f. Not bloody likely, in this case.

P(e,f) -- joint probability

The chance of e and f both happening. If e and f don't influence each other, then we can write P(e,f) = P(e) * P(f). If e and f do influence each other, then we had better write P(e,f) = P(e) * P(f | e). That means: the chance...
Thanks for your interest!