How to solve homophone problems in Automatic Speech Recognition?

„Based on material from Béchet (1999) et al. and Nemoto et al. (2008)“

Ivaylo Yanev

Advisor: Tim Schlippe

04-02-2011
Outline

1. Introduction
2. Solution on language model (LM) level
   2.1 Motivation
   2.2 Phase-based language model (LM)
   2.3 Cache-based language model (LM)
   2.4 Model Combination Experiments
3. Acoustic and prosodic information
   3.1 Motivation
   3.2 Perceptual Errors (Human Errors)
   3.3 Automatic Transcription Errors (Errors of ASR)
   3.4 Classification with acoustic and prosodic attributes
4. Conclusion
1. Introduction

• Challenges for ASR: Homophones

• Homophone: word having different orthography with the same phonetic transcription

• Here: Investigation for French

• Facts about homophones in French:
  • Singular/Plural inflection
  • Homophone forms in French: e.g. et ("and") and est ("is"), especially different verb forms, e.g. tuer, tué, tués... “to kill”, allez (you go), aller (to go), allé (gone (m)), allés (gone (f))
  • Each word in average belongs to homophone class of 2.2 elements (paper 1, Bechet et al.)
  • Humans distinguish homophones with context knowledge
  • But the context knowledge is very limited in ASR (Automatic Speech Recognition)

→ Homophones: One of specific problems for ASR in French
1. Introduction

- I present here two papers that describe **how to deal with homophones:**

  - **Paper 1:** *Large span statistical language models: Application to homophone disambiguation for large vocabulary speech recognition in French*  
    (Frédéric Béchet, Alexis Nasr, Thierry Spriet, Renato de Mori, EuroSpeech 1999)

  - **Paper 2:** *Speech errors on frequently observed homophones in French: Perceptual evaluation vs automatic classification*  
    (Rena Nemoto, Ioana Vasilescu, Martine Adda-Decker, LREC 2008)
2.1 Large span statistical language models: Motivation

- Confusion pairs for French homophones particularly high for some singular/plural inflections (e.g. diffusé/s)

→ **Focus of paper 1** (Béchet et al., 1999)

- Analyses of local LMs such as 3-gram or 3-class LMs with POS (Part-of-Speech), large-span LMs and combination of these LMs

- In paper 1 - **two kinds of models:**
  - Local LMs:
    - 3-gram LM (words)
    - 3-class LM with 105 POS
  - Large-span LMs:
    - Phrase-based LM: Phrase-patterns from clusters of POS sequences
    - (Homophone) Cache-based LM: Vectors with POS histories of singular and plural homophones to determine adequate form
How to solve homophone problems in ASR?

2.2 Experiments with 3-gram LM, 3-class LM and phrase-based LM

- Likelihood of sentence hypothesis = Linear combination of probabilities of 3-gram LM on words, 3-class LM on POS and 3-class LM on phrases

- 3-gram LM: on words

- 3-class LM on POS: LM with POS tags (# POS tags=105)

- 3-class LM on phrases:
  1. **tag corpus** with statistic tagger
  2. **parse tagged corpus** with finite state parser to **recognize syntactic phrases** (e.g. nominal, verbal, prepositional syntagms)
  3. **label each phrase** according to its syntactic structure → phrase patterns (larger context)
2.2 Phrase-based LM

- **Table 1**: Result of correct analysis of a sentence:
  - Words in bold: Singular/Plural homophones
  - Correctly disambiguated homophones: marked with *

3-class LM on POS realises agreement between verb „constituent“ and noun „justice“ instead of its subject „valeurs“ (false would be: „constituët“)

POS on words

`Table 1 - Parsing example`

<table>
<thead>
<tr>
<th>word</th>
<th>POS</th>
<th>phrase</th>
<th>phrase</th>
<th>phrase</th>
</tr>
</thead>
<tbody>
<tr>
<td>quand</td>
<td>COSUB</td>
<td>COSUB</td>
<td></td>
<td></td>
</tr>
<tr>
<td>d’</td>
<td>DETFP</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
| authentiques | AFP | NFP | * | * | *
| valeurs | NFP   |        | *      | *      |
| de      | PREPDE |      |        |        |
| justice | NMS   | GP     | *      | *      |
| ne      | ADVNE |        |        |        |
| constituent | V3P | VP   |        | *      |
| plus    | ADV   |        |        |        |
| le      | DETMS |        |        |        |
| fondement | NMS | NMS | * | * | *
| des     | PREPDES |    |        |        |
| lois    | NFP   | GP     | *      | *      |
| c’      | PPER3MS | PPER3MS |        |        |
| est     | VE3S  | VS     |        |        |
| souvent | ADV   |        |        |        |
| l’      | DETMS |        |        |        |
| arbitraire | NMS | NMS | * | * | *
| qui     | PRELMS | PRELMS |        |        |
| les     | PPOBJMP |      |        |        |
| remplace | V3S  | VS     | *      | *      |
Shortcomings of 3-gram LM, 3-class LM and phrase-based LM

• Some cases: difficult to process with phrase-based and with (more generally) syntactic-based LMs:
  1. Overlapping prepositional syntagms (=phrases or sentences) or relative clause, co-ordinate clauses, etc.
     → Syntactic constraints not captured by simple grammars
  2. Syntactically undecidable or really ambiguous cases

• Solution to 1: Full syntactic parsing (but such a parsing very difficult to integrate in a speech decoding process due to coverage and complexity)

• Solution to 2: Lexical or Semantic information: needed to remove ambiguities (by number agreement):
  • Example: „Le président Boris Eltsine dans un message de voeux diffusé à la télévision russe“
  • The number agreement between 'diffusé' and 'message' (singular) rather than 'voeux' (plural) can't be predicted by a syntactic model
Shortcomings of 3-gram LM, 3-class LM and Phrase-based LM

- Substitutions, insertions and deletions errors occurring during decoding process make full syntactic parsing nearly impossible

  - Example: True: Valeurs (pl.) de justice ne constituent (pl.)
  - False: Vent de justice ne constitue (sing.) (here was valeurs false confused with vent)

- But strong syntactic constraints increase WER dramatically!
  - Decision LM: Robust to speech recognition errors, which can take a decision on the number of a homophone word without strong syntactic constraints
  - Cache-based LM presented in following slides
2.3 Cache-based LM

- **Cache-based LM** stores each singular/plural homophone and its left contexts (= word histories made of last ten words stored in cache memory) as seen in training corpus
  - Each cache content vector \( C(w) \), whose components are syntactic POSs assigned to words by tagger (size of vectors = 105 (= #POS))
- **Training of LM**: Using training corpus for updating **two cache memory vectors** for each homophone \( w \):
  - \( CP(w) \): contexts of plural flexion of \( w \)
  - \( CS(w) \): contexts of singular flexion of \( w \)
- **Decoding**: Two distances computed, when two singular/plural homophones of same \( w \) in competition:
  - 1-st distance between \( CP(w) \) and the current cache and
  - 2-nd distance between \( CS(w) \) and the cache
  - The used distance is a symmetric Kullback-Leibler divergence measure
2.3 Cache-based LM

- `'diffusé': singular/plural homophone (=w) can either be singular or plural, depending on agreement with 'message' or 'voeux,'
- **Cache memory vector** \( A(w) \) based on 9 words, preceding word 'diffusé'
- \( A(w) \) is than compared with \( CP(w) \) and \( CS(w) \), associated to 'diffusé'
- \( |\text{dist}(A,CS)-\text{dist}(A,CP)| > \text{th} \)
  - When satisfied than select flexion whose vector closest to the current one (=A(w))
  - \( CS(w) \) has here minimum distance and represents the singular flexion of 'diffusé'

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Le</td>
<td>président</td>
<td>Boris</td>
<td>Eltsine</td>
<td>dans</td>
<td>un</td>
<td>message</td>
<td>de</td>
<td>voeux</td>
<td>diffusé or diffusés</td>
</tr>
<tr>
<td></td>
<td>DETMS</td>
<td>NMS</td>
<td>XPRE</td>
<td>XFAM</td>
<td>PREP</td>
<td>DETMS</td>
<td>NMS</td>
<td>PREP</td>
<td>NMP</td>
<td>VPPMS or VPPMP</td>
</tr>
</tbody>
</table>
## 2.4 Results of Model Combination Experiments

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>WA</td>
<td>90.95</td>
<td>95.36</td>
<td>89.02</td>
<td>84.59</td>
<td>96.89</td>
<td>96.14</td>
<td>96.22</td>
<td>96.98</td>
<td>97.36</td>
</tr>
</tbody>
</table>

**Disadvantages**
- M3 and M4 are less precise than n-grams
- M4 does not capture all syntactic constraints (cache small, applied only to homophones)

**Advantages**
- M3 and M4 cover some useful cases not covered by the other models
3.1 Acoustic and prosodic information - Motivation (I)

- Perceptual vs automatic transcription errors
- Focus of paper 2 (Nemoto et al., 2008) in terms of *et* / *est* homophones
- „et“ (conjunction) and „est“ (verb) different part of speech
  - Occupy distinct positions in sentences
  - Different prosodic realization of words (e.g. the duration of the words and the fundamental frequency)
- Make acoustic analysis of appropriate *acoustic and prosodic* attributes
- **Prosody**: the rhythm, stress and intonation of speech ([www.wikipedia.org](http://www.wikipedia.org))
- **Reflect various features** of speaker (emotional state) or utterance (statement, question, command)
3.1 Acoustic and prosodic information - Motivation (II)

• Use of the French Technolangue-ESTER corpus: broadcast news shows from different francophone (French and Moroccan) radio stations

• Extraction of automatic transcription errors by the LIMSI speech recognition system
3.1 Acoustic and prosodic information - Motivation (III)

- **Fundamental frequency** ($f_0$ (or $F_0$)): the lowest frequency of a periodic waveform (www.wikipedia.org)
- **Formants** (defined by Fant): the resonance frequencies of an acoustic tube (vocal tract)
- **In practice, only the first few formants are of interest**
  - $F_1$: major resonance of the **pharyngeal cavity**
  - $F_2$: major resonance of the **oral cavity**
3.2 Perceptual Errors (Human Errors)

- Select stimuli comprising the target et/est homophones in limited n-gram contexts
- The test material consisted in 83 chunks extracted from the ESTER development corpus
- **Chunk**: a 7-word string with the target word as center
- Stimuli illustrate different types of errors: et/est confusion, insertions, deletions, substitutions of the target word(s) only or together with surrounding words (target word within a syntagm)

### 7-gram language model (4-gram left and 4-gram right)

| Ex. 1          | REF: rhume de cerveau est la maladie virale |
|               | HYP: rhume de cerveau et la maladie virale |
| Ex. 2          | REF: sur les salaries est si formidable que |
|               | HYP: sur les salaries ici formidable que |
| Ex. 3          | REF: politique aujourd’hui il est essential d’approfondir |
|               | HYP: politique aujourd’hui il d’approfondir essential |
### 3.2 Perceptual Errors (Human Errors)

#### Types of chunks giving rise or not to automatic transcriptions errors

<table>
<thead>
<tr>
<th>Types of chunks</th>
<th>Types of errors</th>
</tr>
</thead>
<tbody>
<tr>
<td>5 distractors</td>
<td>Stimuli without et/est in the middle</td>
</tr>
<tr>
<td>10 corrects</td>
<td>Stimuli with et/est correctly transcribed by the system</td>
</tr>
<tr>
<td>20 et/est symmetric confusions</td>
<td>Stimuli with symmetric ASR confusions of et/est</td>
</tr>
<tr>
<td>48 other errors (errors of the target homophone word + surrounding context)</td>
<td>Stimuli with other errors: insertions, deletions, erroneous transcription of target word alone or within a syntagm</td>
</tr>
</tbody>
</table>

Results of ASR system
3.2 Perceptual Errors (Human Errors)

- Test protocol:
  - 60 native French subject divided into two sub-groups (40 and 20) with different test conditions:

  1. **Acoustic+language model (AM+LM) = audio test**, condition test with 40 subjects:
     - Provided to listeners 7-gram chunks and had to transcribe entire chunk
     - 83 stimuli submitted to 2 groups of 20 subjects via a web available interface
     - Each group of 20 subjects transcribed half of the stimuli
     - duration of the test is less than one hour
     - The two groups were comparable in terms of age and background
3.2 Perceptual Errors (Human Errors)

2. A local language model (LM) condition test (=the written version of the stimuli focusing on the symmetric et/est confusion):
   - Subjects had to fill „et“ or „est“ using 3-word left and 3-word right contexts

This test assumes perfectly homophony for the target

```
          et
Rhume de cerveau       la maladie virale
          est
```

1) Syntactic/semantic information of the written sequence contributes to solve ambiguity

2) Humans explicitly focus on local ambiguity
3.2 Perceptual Errors (Human Errors)

- AM+LM (audio) condition test results:
  - Humans produce more errors on stimuli misrecognized by the ASR system
  - Reversely, humans are almost error free on correctly decoded stimuli
  - Weighting the achieved perceptual results: humans 4-5 times more accurate than ASR
  - Humans produce more errors for stimuli for which ASR missed the target word

<table>
<thead>
<tr>
<th>Stimuli</th>
<th>WER (word error rates)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ASR AM+LM</td>
</tr>
<tr>
<td>5 distractors</td>
<td>0%</td>
</tr>
<tr>
<td>10 corrects (perfectly decoded)</td>
<td>0%</td>
</tr>
<tr>
<td>20 et/est symmetric confusions</td>
<td>100%</td>
</tr>
<tr>
<td>48 other errors</td>
<td>100%</td>
</tr>
</tbody>
</table>

Local linguistic ambiguity is problematic for both (ASR system and humans)
3.2 Perceptual Errors (Human Errors)

LM (only text) test conclusion:

→ Humans and ASR system are equally competitive for distractors (stimuli without et/est in the middle)

→ But both leave also unresolved ambiguities

LM condition test results

<table>
<thead>
<tr>
<th>Stimuli</th>
<th>WER (word error rates)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ASR AM+LM</td>
</tr>
<tr>
<td>5 distractors</td>
<td>0%</td>
</tr>
<tr>
<td>10 corrects (perfectly decoded)</td>
<td>0%</td>
</tr>
<tr>
<td>20 et/est symmetric confusions</td>
<td>100%</td>
</tr>
<tr>
<td>48 other errors</td>
<td>100%</td>
</tr>
</tbody>
</table>
3.3 Automatic Transcription Errors (Errors of ASR)

- „et“ and „est“ have extracted from different French broadcast news (BN) channels from Technolangue-ESTER corpus
- Several acoustic and prosodic parameters automatically extracted
- Concern duration, fundamental frequency, formants and surrounding contexts (pauses preceding/following the target word)
- Measures (for pitch and formant values): 5 ms frame by frame
- Computed voicing ratio (for each segment) and mean values for the parameters and formants over all voiced frames (of the segment)

<table>
<thead>
<tr>
<th>Words</th>
<th>Occurences</th>
<th>Phonemes</th>
</tr>
</thead>
<tbody>
<tr>
<td>et</td>
<td>19.1k /e/</td>
<td>[e]</td>
</tr>
<tr>
<td>est</td>
<td>14.5k /ε/</td>
<td>[ε] 5.0k, [e] 9.5k</td>
</tr>
</tbody>
</table>

Occurences of „et“ and „est“ in the BN corpus

\[
\text{Voicing ratio} = \frac{\# \text{ voiced frames}}{\# \text{ all frames}}
\]
3.3 Automatic Transcription Errors (Errors of ASR)

- **Acoustic analysis:**
  - In ASR acoustic and prosodic parameters can differentiate the homophones „et“ and „est“
  - Consider duration, voicing characteristics and pauses before/after these one

- **Duration:**
  - Duration range (30-200 ms) for both words
  - Comparison of distribution shows differences between two target words
  - „et“ has relatively flat distribution (durations above 80 ms) whereas „est“ an almost bellshaped distribution (centered on 60 ms)
  - On average „et“ lasts longer than „est“

**Duration distribution of the homophones et/est:** et (in red) and est (in blue) ( /e/ in clear green and /ɛ/ in dark green). Different lines correspond to number (in %) of occurrences per duration threshold
3.3 Automatic Transcription Errors (Errors of ASR)

- For analyzing voicing ratio are defined 3 classes:
  1. Devoiced (% of voicing): 0-20%
  2. Partial voicing: 20%-80%
  3. Voicing: 80%-100%

- Proportion of segments are shown for each class
- Two bars added for „est“ to separate [ɛ] from [e] pronunciation
- Devoiced class contains small amount of data for both homophones
- In the „partial voicing“ class: „et“ better represented than „est“
- In the „voicing“ class: „est“ more frequent than „et“

→ Result: „et“ less voiced than „est“
3.3 Automatic Transcription Errors (Errors of ASR)

• **Left/Right pause co-occurrences:**
  - The pauses play an important role in the process of automatic prosodic information extraction (*Lacheret-Dujour, Beaugendre, 1999*)
  - **Relation between the class „pause“ (i.e. silences, breaths and hesitations) and the two homophones:** left/right pause co-occurrences with the target words „et“ and „est“
  - The main difference between the two homophones concerns the amount of pause occurrences (in particular left pauses)

→ „est“ is less frequently preceded by a pause than „et“

<table>
<thead>
<tr>
<th>Words</th>
<th>et</th>
<th>est</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Left pause</strong></td>
<td>49%</td>
<td>9%</td>
</tr>
<tr>
<td><strong>Right pause</strong></td>
<td>7%</td>
<td>5%</td>
</tr>
</tbody>
</table>
3.4 Classification with acoustic and prosodic attributes

Attribute definition:

- **Intra-phonemic attributes** (33) (model the target word):
  - Duration attributes, f₀, voicing ratio, first three formants (global mean values by segments and begin, center, end values):
    -> Calculated also the differences (Δ) between begin-center, center-end, and begin-end for the f₀ and the formants

- **Inter-phonemic attributes** (8) (model its relation to the context):
  - Duration attributes (measured as the difference between center segment duration of target word and center segment duration of a previous/following vowel), f₀, pauses
    -> Δ values calculated as the difference between the mean values of the target word vowel and the previous/following vowel
    -> left-right pause attributes were added too

Example:

![Diagram showing comparison between "est ami" and "et ami" with arrows and nodes labeled with "est ami" and "et ami".](image-url)
The experiments used a cross-validation method: a technique for assessing how the results of a statistical analysis will generalize to an independent data set.

The 10 best attributes are almost as discriminatory as the 41 attributes.

The word accuracy with the LMT algorithm is 77%.

LMT is the best of all here represented algorithms.

<table>
<thead>
<tr>
<th>Treated attribute numbers</th>
<th>Word accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>77.0%</td>
</tr>
<tr>
<td>41</td>
<td>78.0%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Best algorithm (LMT = Logistic Model Trees)</th>
<th>Word accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>77.0%</td>
</tr>
</tbody>
</table>

| Mean of best 10 tested algorithms          | 75.7%         |
| Mean of all 25 tested algorithms          | 70.2%         |
3.4 Classification with acoustic and prosodic attributes

- **10 attributes** (from 41) are more discriminatory than the others and have been selected thanks to the **LMT** (Logistic Model Trees) algorithm (provided the highest result)

<table>
<thead>
<tr>
<th>intra-phonemic attributes and inter-phonemic attributes (in bold)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>words</strong></td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>7</td>
</tr>
<tr>
<td>8</td>
</tr>
<tr>
<td>9</td>
</tr>
<tr>
<td>10</td>
</tr>
</tbody>
</table>
4. Conclusion (I)

- **Focus on paper 1** (Béchet et al., 1999):

- **Solution on language model level**
  - Benefits of language model combinations for homophone disambiguation (different models capture complementary properties)
  - **Local LMs:**
    - 3-gram LM (on words) (M1)
    - 3-class LM with 105 POS (M2)
  - **Large-span LMs:**
    - **Phrase-based LM (M3):** Phrase-patterns from clusters of POS sequences
    - **(Homophone) Cache-based LM (M4):** Vectors with POS histories of singular and plural homophones to determine adequate form
  - Language models **M3 and M4 are less precise than n-grams:**
    - M3 uses only 70 classes
    - M4 does not capture all syntactic constraints (the cache is small and is applied only to homophones)
    - M3 and M4 cover some useful cases not covered by the other models
    - The four LMs can be used for refining word hypothesis
4. Conclusion (II)

- Focus on paper 2 (Nemoto et al., 2008):
- Classification with acoustic and prosodic attributes
  - Discussion on the perceptual evaluation:
    - Error rates varies strongly with the type of local context
    - Contexts with symmetric et/est errors and contexts with target word+surrounding context: highly ambiguous for ASR system (in these cases humans are 4-5 times more accurate)
    - Humans achieved better results for stimuli with et/est correctly transcribed by the ASR system
    - „est“ is more frequently misrecognized by the human listeners than „et“ (25% vs 10%, see duration distribution of „est“ and „et“)
    - Humans listeners deal with local ambiguity more efficiently than ASR system
4. Conclusion (III)

- Focus on paper 2 (Nemoto et al., 2008):
- **Classification with acoustic and prosodic attributes**
  - **Discussion on the automatic classification:**
    - Different acoustic realizations of „et“ and „est“
    - homophones may differ in their prosodic realization
    - defined 41 intra- and inter- phonemic acoustic and prosodic attributes of the two homophone words „et“ and „est“ and tested different algorithms
    - the best algorithm is LMT (Logistic Model Trees) with 77% of correct word identification
    - particularly robust attributes concerning intra- and inter-segmental duration, voicing and differences in f0 between the target segment and the close context
Thanks for your interest!