EVALITA 2011 is the third evaluation campaign of Natural Language Processing and Speech tools for Italian, supported by the NLP working group of AI*IA (Italian Association for Artificial Intelligence) and AISV (Italian Association of Speech Science)

http://www.evalita.it/2011

The SPPAS participation to Evalita 2011

Brigitte Bigi
ma che cos'è SPPAS?

SPeech Phonetization Alignment and Syllabification
Main description

- A new tool to produce automatically annotations which includes utterance, word, syllabic and phonemic segmentations from a recorded speech sound and its transcription.

- Language-independent.

- Currently designed for French, English, Italian and Chinese and there is an easy way to add other languages.

- Distributed under GPL license.
Introduction

SPPAS Screenshots

SPPAS - Version 1.3
SPEech Phonetization Alignment and Syllabification

File selection
Directory selection

IT
EN
ZH
FR

Exit
About...
Preferences
Start processing...

About SPEech Phonetization Alignment and Syllabification
SPEech Phonetization Alignment and Syllabification 1.3
SPPAS is a tool to produce automatically annotations which includes utterance, word, syllabic and phonemic segmentations from a recorded speech sound and its transcription. SPPAS is distributed under GNU Public License.
Copyright (C) 2011 LPL Laboratory
http://www.lpl.cri.fr/~bigi/sapas/

Activated
Macro-Segmentation
Phonetization
Alignment
Syllabification
Help
Cancel
Validate

Brigitte Bigi - LPL
Evalita task

- "Forced Alignment on Spontaneous Speech":
  - Phone segmentation;
  - Word segmentation.
- Data: ✓ Closed; · Opened.
- Dialogues, map-tasks:
  - 3h30 speech;
  - 15% phones: pauses, filled-pauses, garbage.
- SPPAS Forced-Alignment is 2 sub-tasks:
  - phonetization + alignment
Phonetization

- The process of representing sounds by phonetic signs.
- There are two general ways to construct a phonetization process:
  - rule based systems (with rules based on inference approaches or proposed by expert linguists);
  - dictionary based solutions which consist of storing a maximum of phonological knowledge in a lexicon.
SPPAS uses the dictionary-based approach.

The phonetization is the equivalent of a sequence of dictionary-look-ups:

- Input transcription needs to be word-segmented
- It is supposed that all words of the transcription are mentioned in the pronunciation dictionary.

A specific phone to represent filled pauses.
Dictionary

- The dictionary contains a set of possible pronunciations of words, including accents as "perché" pronounced as /b e r k e/, and reduction phenomena as /p e k/.

- Made of:
  - 390k words;
  - 5k variants.

- From:
  - Festival + Evalita training corpus.

- Manual corrections
Phonetization: variants

- No rules are applied:
  - all possibilities are proposed to the aligner.
- For example, the sentence "del mio frigorifero" will produce the following list of words with associated pronunciations:
  - d.e.l | d.E.l
  - m.jo | m.i
  - f.r.i.g.o.r.i.f.e.r.o | f.r.i.g.o.r.i.f.E.r.o | f.r.i.g.o.r.i.f.e.r | f.r.i.g.o.r.i.f.e.r.o
Alignment

- A time-matching between a given speech utterance along with a phonetic representation of the utterance.
SPPAS Alignment

- Alignment in SPPAS is based on the Julius Speech Recognition Engine (SRE):
  - A finite state grammar that describes sentence patterns to be recognized;
  - An acoustic model.

- The alignment task is a 2-steps process:
  - the first one choose the phonetization;
  - the second one perform the segmentation.
Grammar

- A grammar: constraints on what the SRE can expect as input. It is a list of words that the SRE listens for. Each word has a set of associated list of phonemes.

<table>
<thead>
<tr>
<th>Dictionary</th>
<th>Grammar</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 w_0 d e l</td>
<td>0 2 1 0 1</td>
</tr>
<tr>
<td>0 w_0 d E l</td>
<td>1 1 2 0 0</td>
</tr>
<tr>
<td>1 w_1 m jo</td>
<td>2 0 3 0 0</td>
</tr>
<tr>
<td>1 w_1 m i</td>
<td>3 -1 -1 1 0</td>
</tr>
<tr>
<td>2 w_2 f r i g o r i f e r o</td>
<td>0 2 1 0 1</td>
</tr>
<tr>
<td>2 w_2 f r i g o r i f E r o</td>
<td>1 1 2 0 0</td>
</tr>
<tr>
<td>2 w_2 f r i g o r i f e r</td>
<td>2 0 3 0 0</td>
</tr>
<tr>
<td>2 w_2 f r i g o r i f e r 0</td>
<td>3 -1 -1 1 0</td>
</tr>
</tbody>
</table>
Acoustic Model

- HMM, 5-states
- Triphones
- Trained with HTK:
  - from the proposed phonetized transcription, without using the phonetic time-alignment;
  - using 16 bits, 16000 hz wav files;
  - Mel-frequency cepstrum coefficients (MFCC) along with their first and second derivatives were extracted from the speech in the standard way: MFCC_D_N_Z_0.
Results

- Development corpus:
  - 200 utterances; 12 min 04 sec.
  - 2373 words, 6282 phonemes, including:
    - 689 “_” (pauses);
    - 246 “#” (garbage);
    - both represent 14.88%.

- Evaluate separately:
  - Alignment;
  - Phonetization obtained after the alignment.
Results alignment

- The availability of our system to align the good phoneme sequence.
- On the basis of the manual phonetization.
- Sclite using the time-alignment option:
  - a correct rate of 89.8%, with 7.6% substitutions, 2.6% deletions and 2.6% insertions.
Results alignment

- Pauses, filled pauses and garbages: greatest ranges
- Vowels: automatic shorter than manual
- Consonants: automatic higher than manual
Results phonetization

- The availability of our system to propose the expected phoneme sequence.
- Sclite *without* using the time-alignment option:
  - a correct rate of 89.5%, with 8.1% substitutions, 2.3% deletions and 6.9% insertions.
- Most frequent errors are due to the garbage manual annotation. Example: *bravissimo a questo*
  - Automatic: `bravisimo a k westo`
  - Manual: `br # s # k westo`
  - 5 insertions, 2 substitutions!
SPPAS final results

- Official results estimated using sclite:
  - 88.4% good phoneme alignments:
    - This score contains both phonetizations and alignments errors.
  - 96.7% good word alignments.
Example
Example
Conclusion

- **SPPAS**: a tool to perform the forced-alignment task during the Evalita 2011 campaign, on Italian map-task dialogues.

- **SPPAS** was not specifically devoted to Italian: it can deal with various languages: French, English, Chinese.