SPPAS: Automatic Phonetic Annotation of SPeech

Brigitte Bigi

Keywords:

Phonetizaton Automatic Speech
Syllabification Segmentation
Alignment Prosody

What SPPAS 1.4.9 can do?

- Automatic annotations:
  - **Momel/INTSINT**: Modelisation of Mélodie
  - **IPUs segmentation**: utterance level segmentation
  - **Tokenization**: text normalization
  - **Phonetization**: grapheme to phoneme conversion
  - **Alignment**: phonetic segmentation
  - **Syllabification**: group phonemes into syllables
- Components:
  - Wav Player
  - Manual transcription based on IPUs segmentation
  - Get tiers information and Filter tiers

Key-points

- A tool dedicated to computer scientists **and** linguists

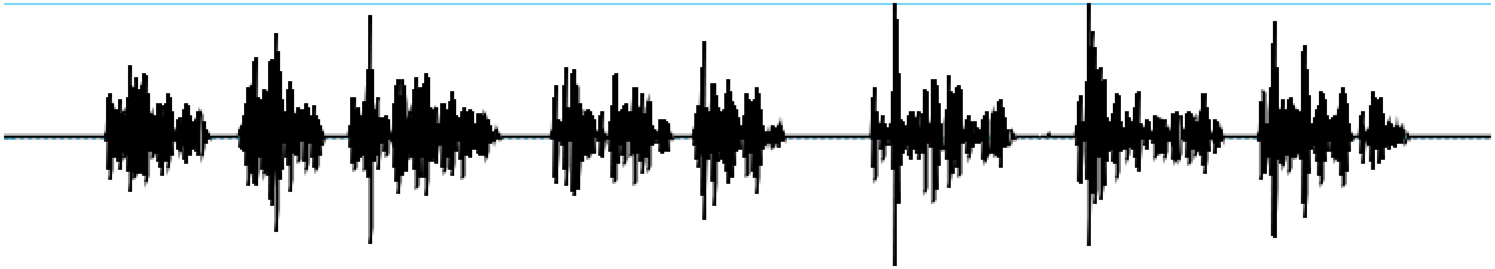
- Language-independent algorithms
  - Resources for French, English, Italian, Chinese, Taiwanese, partially Vietnamese
  - There is an easy way to add other languages

- GNU Public License

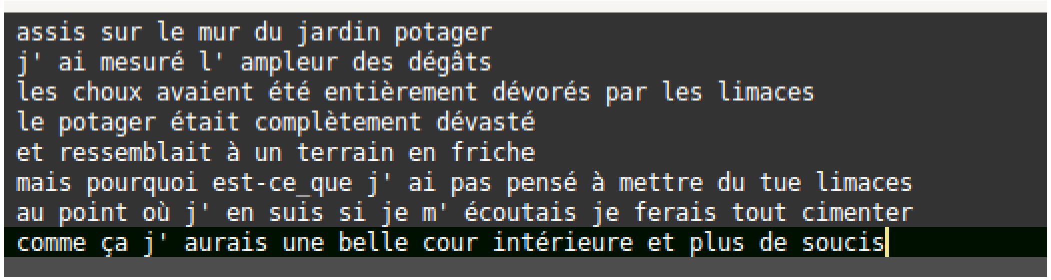
Details in [4,5]

Automatic annotations: inputs

- Speech signal: wav file



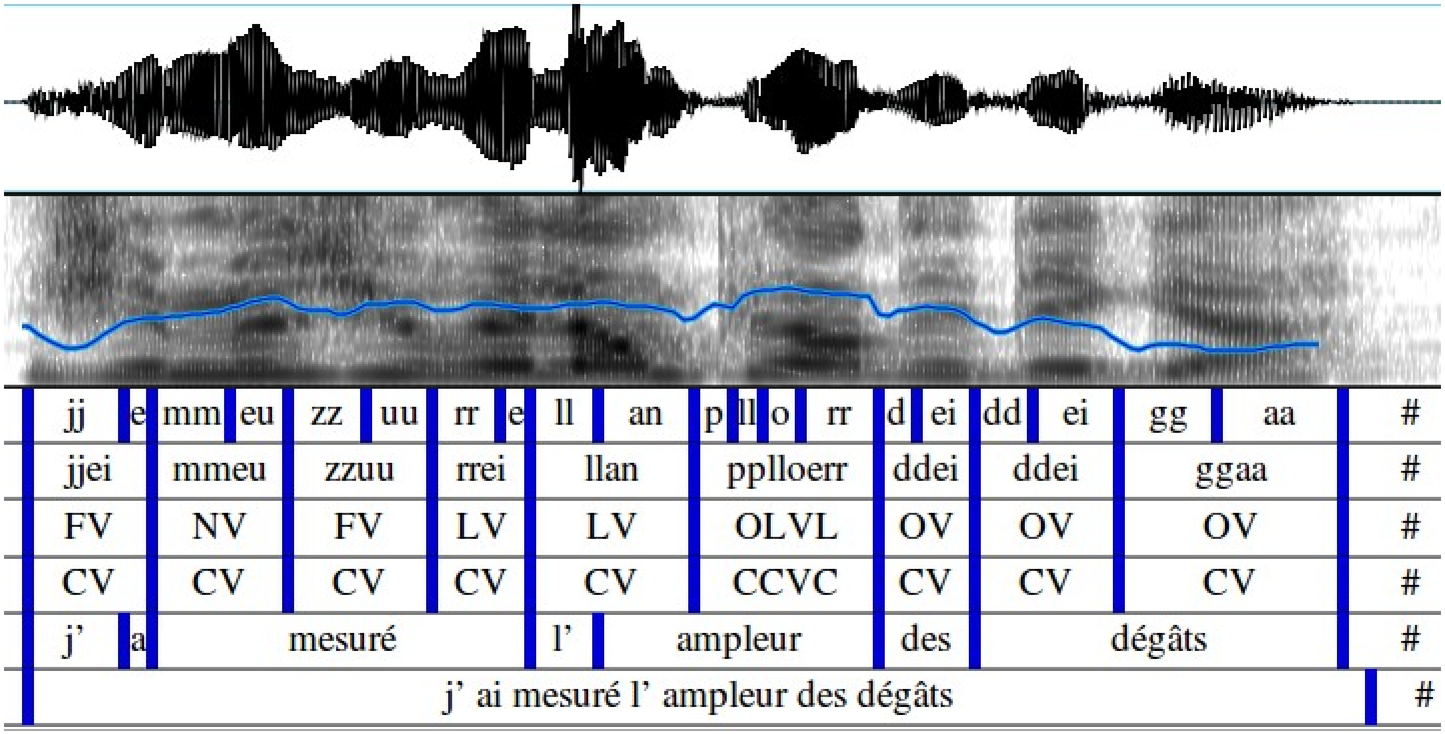
- Transcription: txt or TextGrid



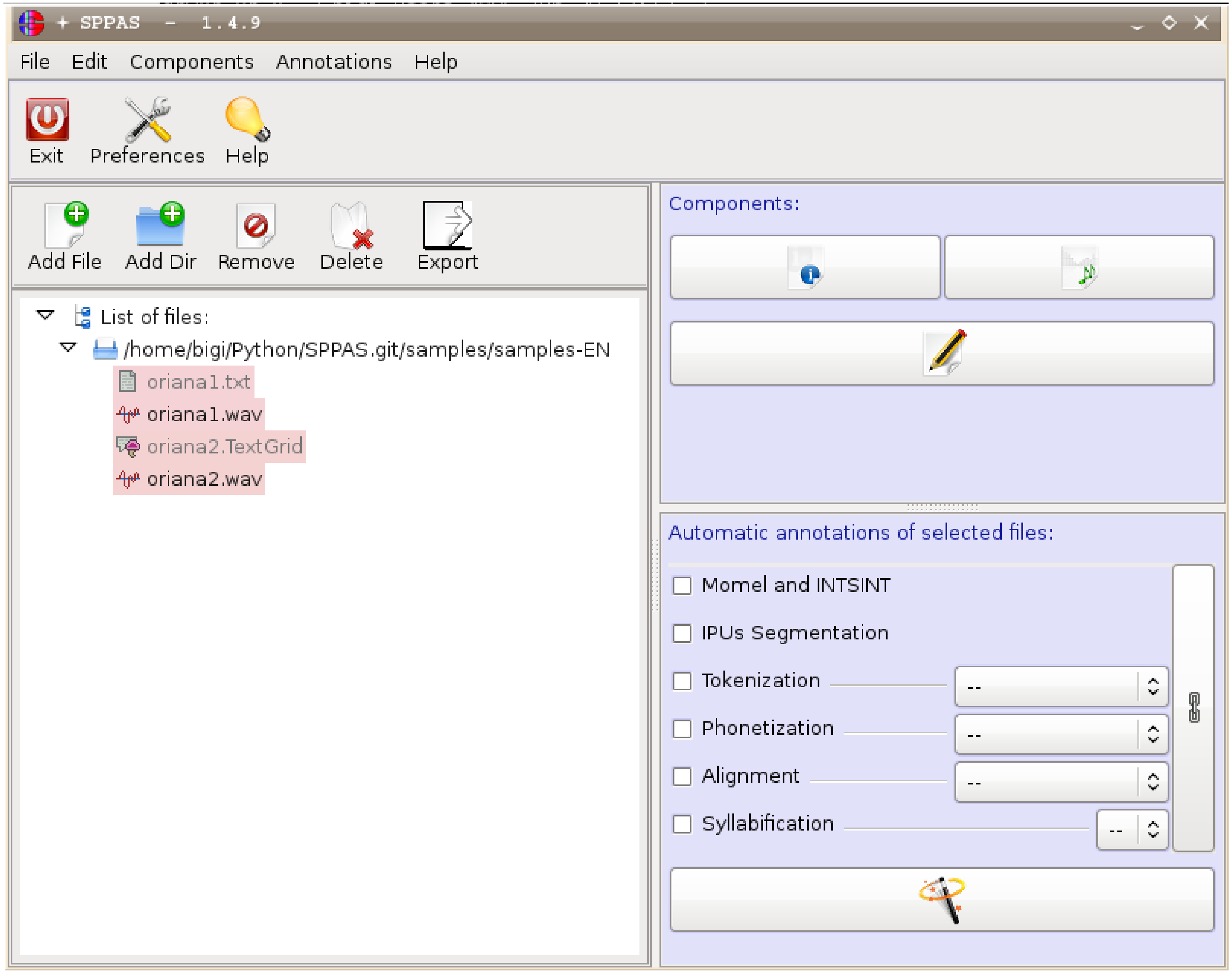
```
assis sur le mur du jardin potager
j' ai mesuré l' ampleur des dégâts
les choux avaient été entièrement dévorés par les limaces
le potager était complètement dévasté
et ressemblait à un terrain en friche
mais pourquoi est-ce_que j' ai pas pensé à mettre du tue limaces
au point où j' en suis si je m' écoutais je ferais tout cimenter
comme ça j' aurais une belle cour intérieure et plus de soucis
```

Automatic annotations: outputs

- A set of TextGrid files


jj e mm eu zz uu rr e ll an p ll o rr d ei dd ei gg aa #
jjei mmeu zzuu rrei llan pplloerr ddei ddei ggaa #
FV NV FV LV LV OLVL OV OV OV #
CV CV CV CV CV CCVC CV CV CV #
j’ a mesuré l’ ampleur des dégâts #
j’ ai mesuré l’ ampleur des dégâts #

Screenshot



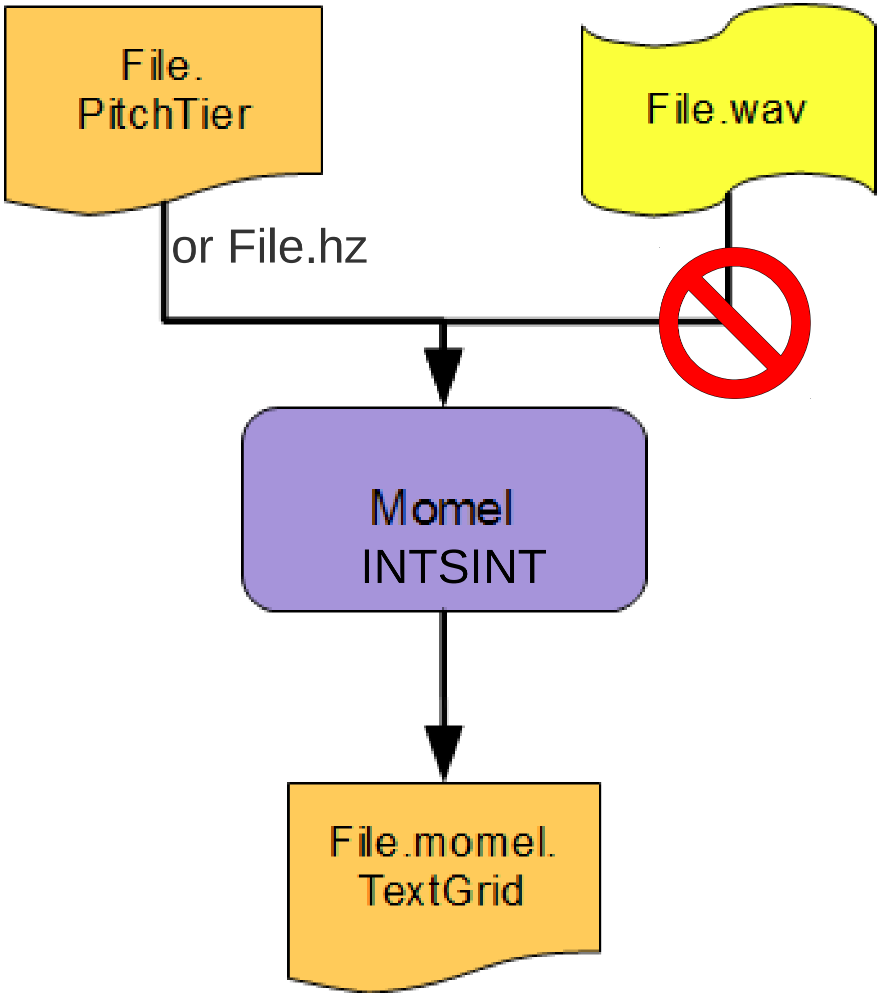
List of files:
/home/bigi/Python/SPPAS.git/samples/samples-EN
oriana1.txt
oriana1.wav
oriana2.TextGrid
oriana2.wav

Automatic annotations of selected files:

- ☐ Momel and INTSINT
- ☐ IPUs Segmentation
- ☐ Tokenization --
- ☐ Phonetization --
- ☐ Alignment --
- ☐ Syllabification --

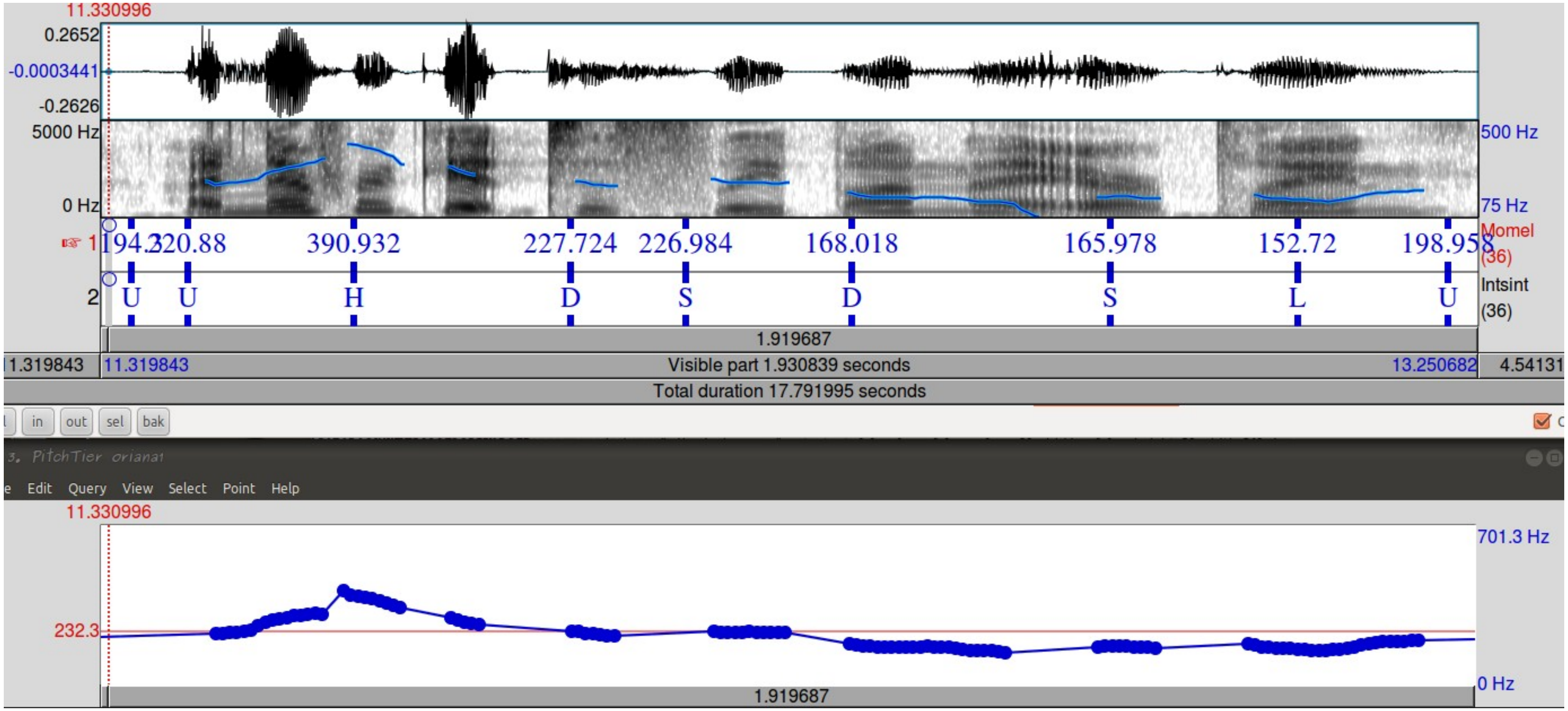
Momel and INTSINT

- SPPAS implements Momel and INTSINT: Daniel Hirst
- A file with pitch values is required (one value each 10ms).



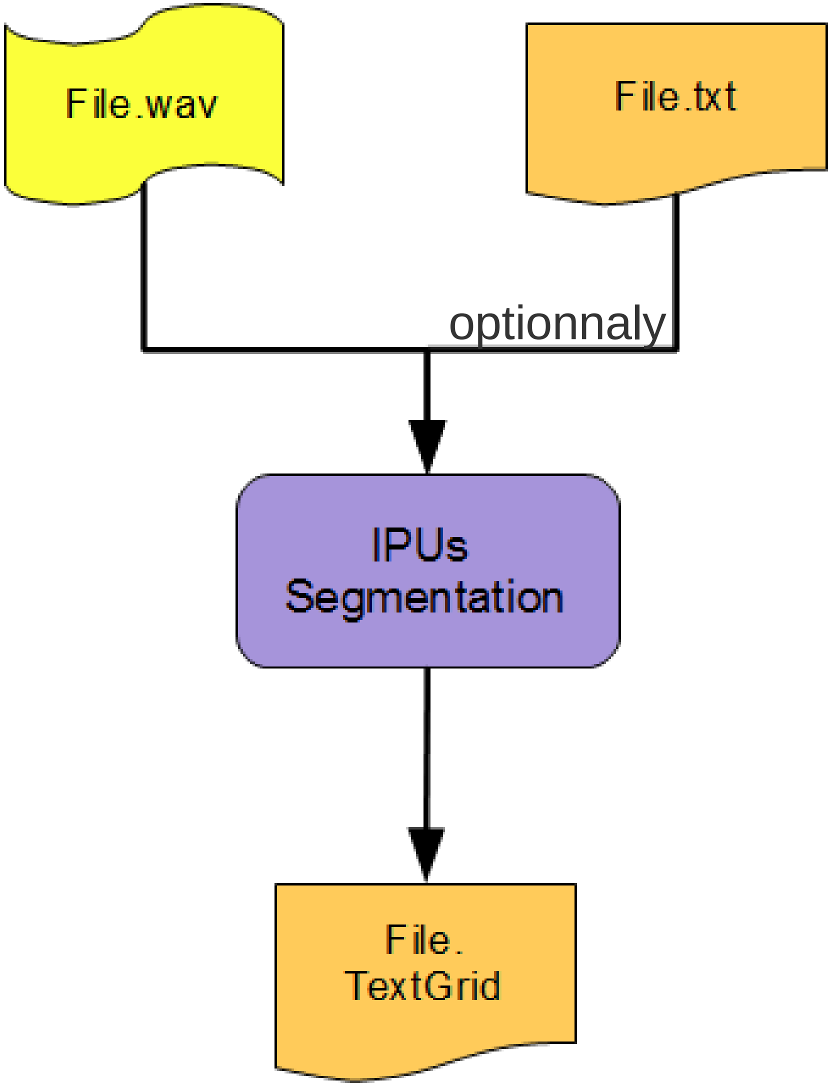
Momel/INTSINT: example

- Output: a TextGrid file with 2 tiers
  - Momel targets (pitch values)
  - INTSINT annotation of these targets



IPUs segmentation

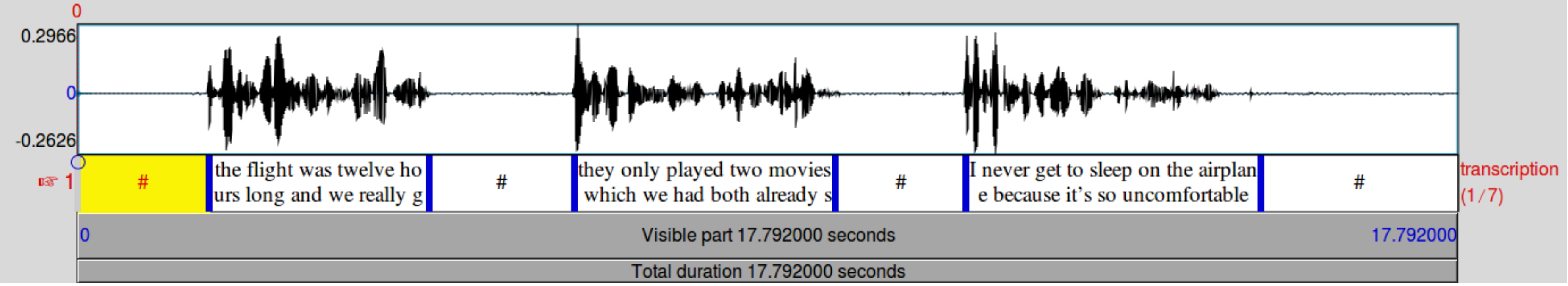
- Inter-Pausal Units segmentation
- The algorithm computes a heuristics based on the detection of silences, by using:
    - volume
    - min silence duration
    - min speech duration



IPUs segmentation: example

Transcription: silences are indicated by newlines or '#'

```
the flight was twelve hours long and we really got bored
they only played two movies which we had both already seen
I never get to sleep on the airplane because it's so uncomfortable
```



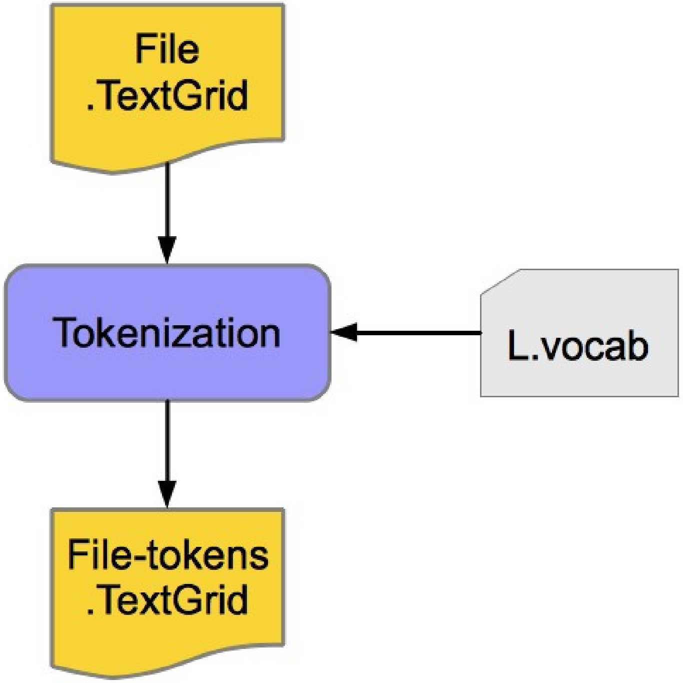
0
0.2966
0
-0.2626
1
transcription
(1/7)

#	the flight was twelve ho urs long and we really g	#	they only played two movies which we had both already s	#	I never get to sleep on the airplan e because it's so uncomfortable	#

0
Visible part 17.792000 seconds
17.792000
Total duration 17.792000 seconds

Tokenization

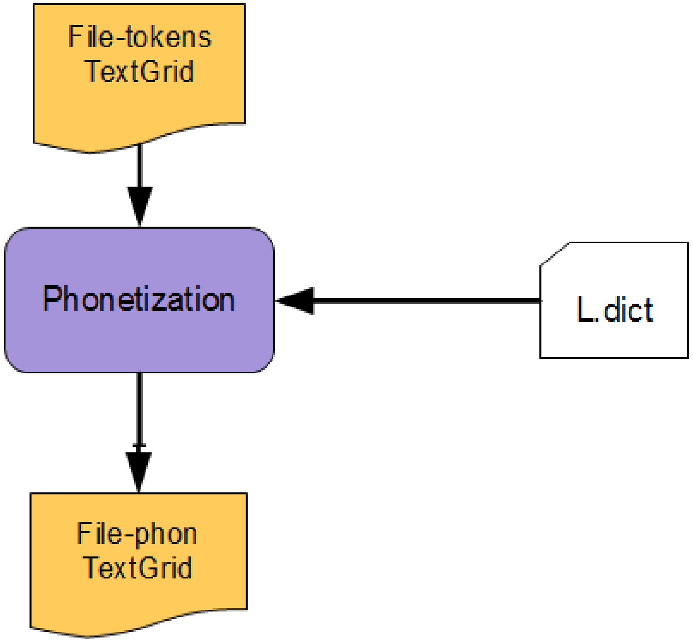
- Process of normalizing text:
  - Remove punctuation, comment etc.
  - Convert numbers to letters:
    - 2 → deux
  - Convert characters to the lower form
  - Segment into words/tokens:
    - parce que → parce_que
- An algorithm as language independent as possible (dict-based)
  - But... *bien que* tu sois là, je pars / c'est *bien que* tu sois là.



Details in [2]

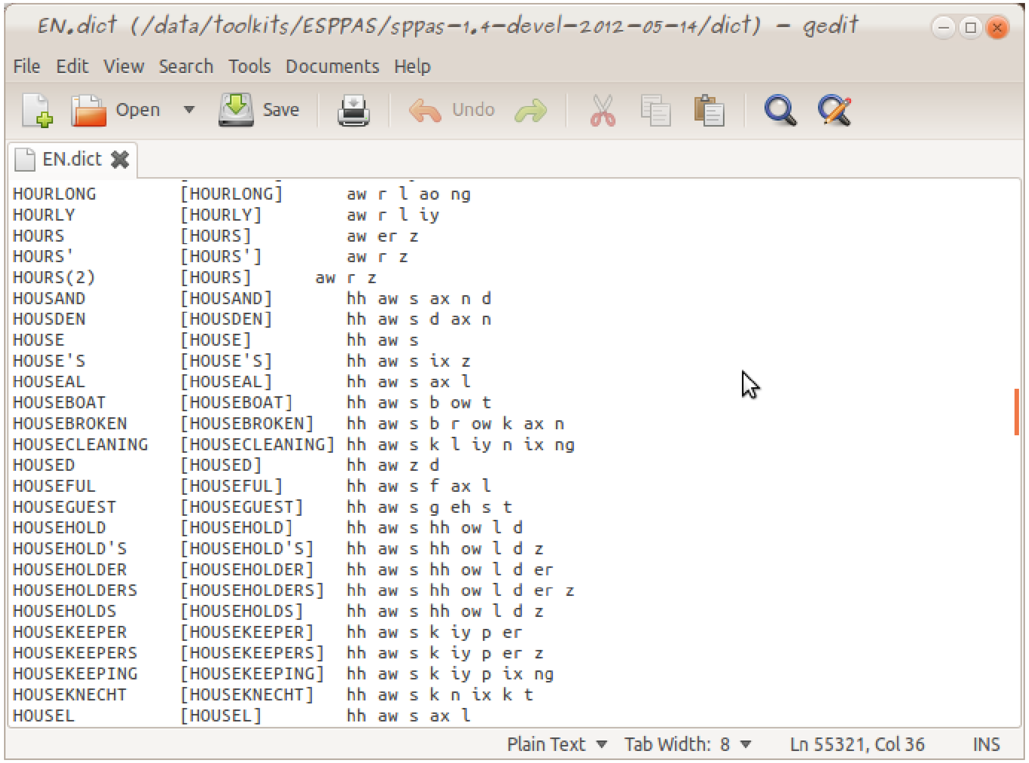
Phonetization

- Process of representing sour with phonetic signs
- The phonetization is the equivalent of a sequence of dictionary look-ups.
- Phonetic variants:
  - no rules are applied, all possibilities are storec



Phonetization: example

- Resources:
  - a dictionary (HTK-ASCII format)


EN.dict (/data/toolkits/ESPPAS/sppas-1.4-devel-2012-05-14/dict) - gedit
File Edit View Search Tools Documents Help
Open
Save
Undo
EN.dict
Plain Text
Tab Width: 8
Ln 55321, Col 36
INS

```
HOURLONG        [HOURLONG]      aw r l ao ng
HOURLY          [HOURLY]        aw r l iy
HOURS           [HOURS]         aw er z
HOURS'          [HOURS']        aw r z
HOURS(2)        [HOURS]     aw r z
HOUSAND         [HOUSAND]       hh aw s ax n d
HOUSDEN         [HOUSDEN]       hh aw s d ax n
HOUSE           [HOUSE]         hh aw s
HOUSE'S         [HOUSE'S]       hh aw s ix z
HOUSEAL         [HOUSEAL]       hh aw s ax l
HOUSEBOAT       [HOUSEBOAT]     hh aw s b ow t
HOUSEBROKEN     [HOUSEBROKEN]   hh aw s b r ow k ax n
HOUSECLEANING   [HOUSECLEANING] hh aw s k l iy n ix ng
HOUSED          [HOUSED]        hh aw z d
HOUSEFUL        [HOUSEFUL]      hh aw s f ax l
HOUSEGUEST      [HOUSEGUEST]    hh aw s g eh s t
HOUSEHOLD       [HOUSEHOLD]     hh aw s hh ow l d
HOUSEHOLD'S     [HOUSEHOLD'S]   hh aw s hh ow l d z
HOUSEHOLDER     [HOUSEHOLDER]   hh aw s hh ow l d er
HOUSEHOLDERS    [HOUSEHOLDERS]  hh aw s hh ow l d er z
HOUSEHOLDS      [HOUSEHOLDS]    hh aw s hh ow l d z
HOUSEKEEPER     [HOUSEKEEPER]   hh aw s k iy p er
HOUSEKEEPERS    [HOUSEKEEPERS]  hh aw s k iy p er z
HOUSEKEEPING    [HOUSEKEEPING]  hh aw s k iy p ix ng
HOUSEKNECHT     [HOUSEKNECHT]   hh aw s k n ix k t
HOUSEL          [HOUSEL]        hh aw s ax l
```

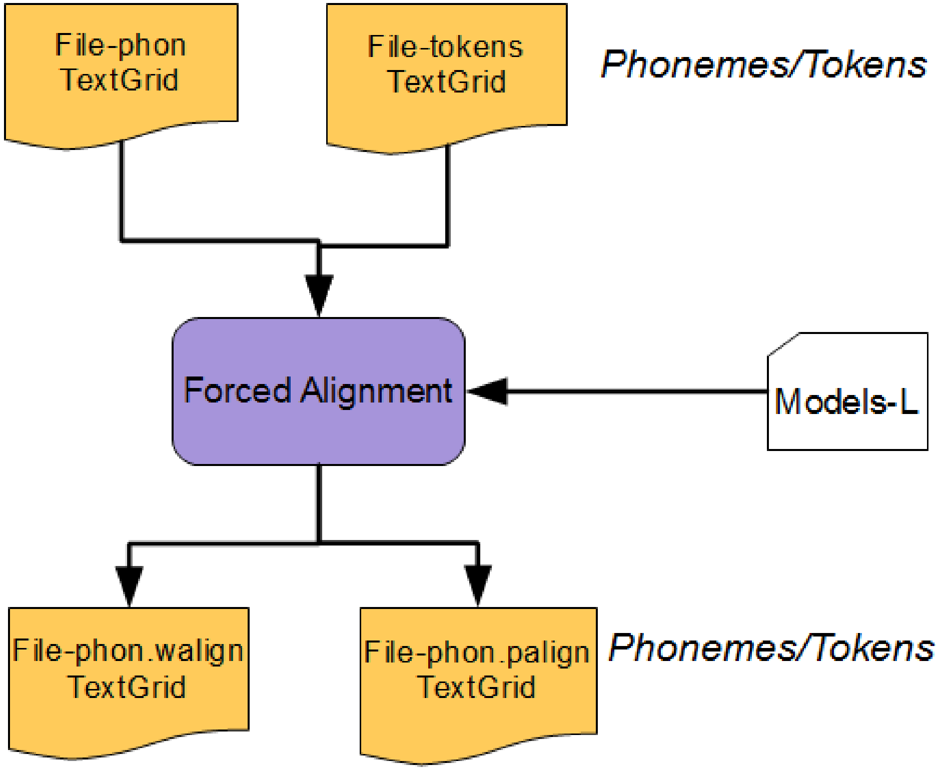
```
the flight was twelve hours long and we really got bored
```

is phonetized as follow:

```
dh.ax|dh.ah|dh.iy   f.l.ay.t   w.aa.z|w.ah.z|w.ax.z|w.ao.z   t.w.eh.l.v
aw.er.z|aw.r.z  l.ao.ng  ae.n.d|ax.n.d  w.iy  r.ih.l.iy|r.iy.l.iy  g.aa.t
b.ao.r.d
```

Alignment

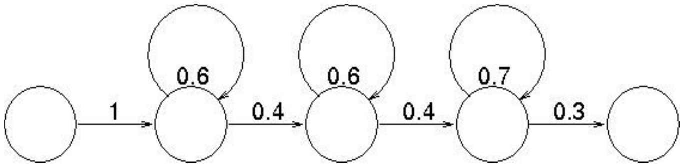
- A time-matching between a given speech utterance along with a phonetic representation of the utterance
- Forced-alignment in SPPAS is based on the **Julius** Speech Recognition Engine

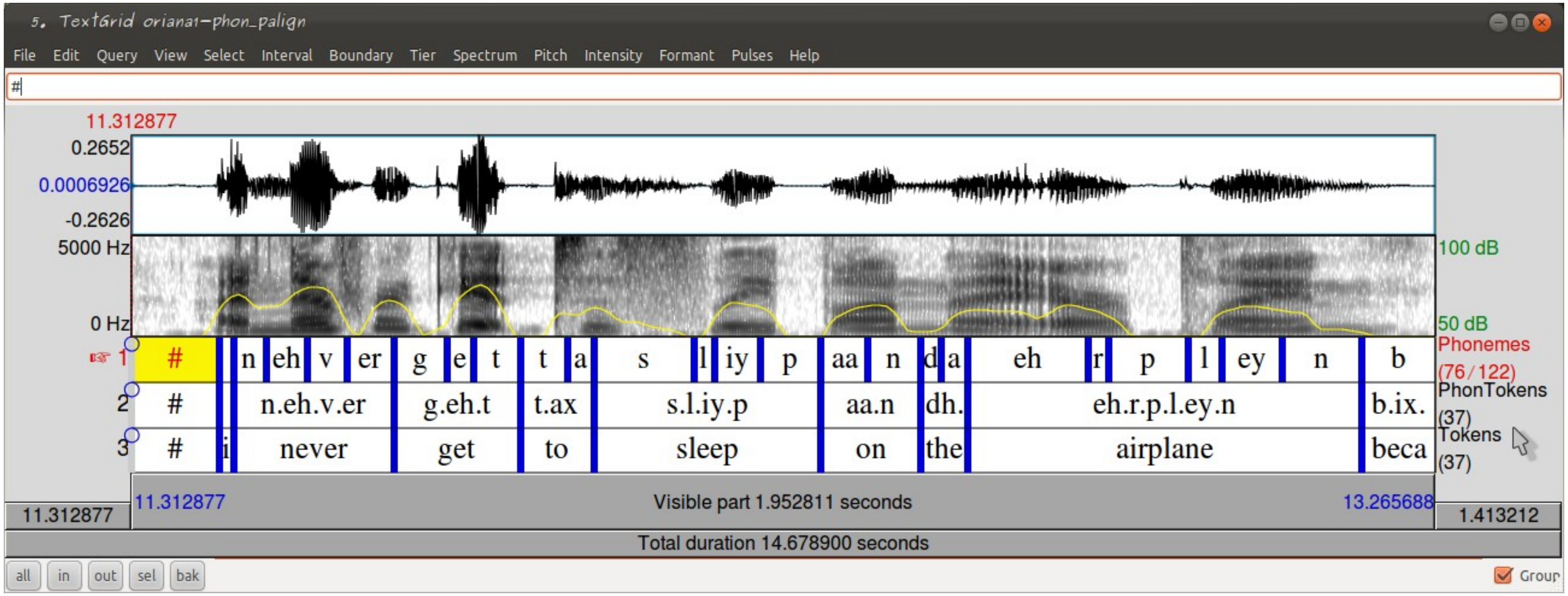


- The alignment task is a 2-step process:
  - the first one: choose the phonetization;
  - the second one: perform the segmentation.

Alignment: example

- Resources:
  - A finite state grammar that describes sentence patterns to be recognized (created by SPPAS);
  - An acoustic model.




5. TextGrid oriana1-phon_palign
File Edit Query View Select Interval Boundary Tier Spectrum Pitch Intensity Formant Pulses Help
#
11.312877
0.2652
0.0006926
-0.2626
5000 Hz
0 Hz
100 dB
50 dB
Phonemes (76/122)
PhonTokens (37)
Tokens (37)
11.312877
11.312877
Visible part 1.952811 seconds
13.265688
1.413212
Total duration 14.678900 seconds
all in out sel bak
Group

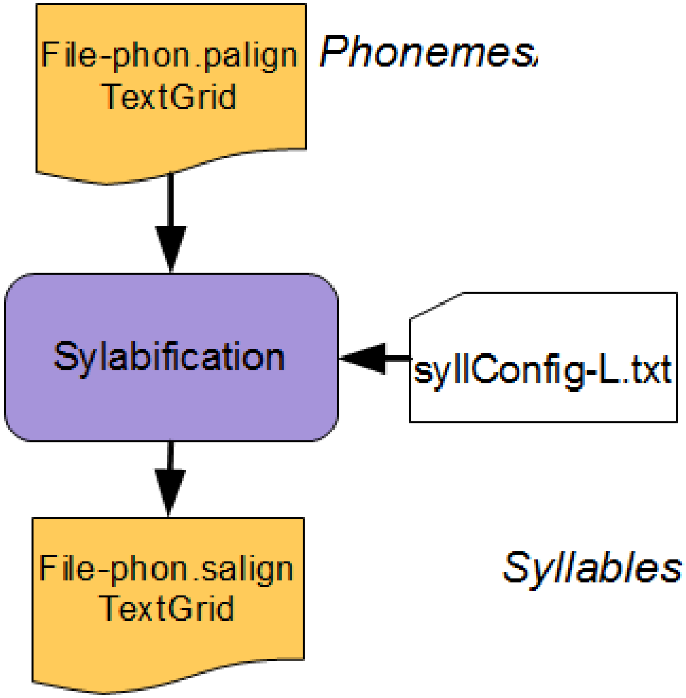
1 # n eh v er g e t t a s l iy p aa n d a eh r p l ey n b

2 # n.eh.v.er g.eh.t t.ax s.l.iy.p aa.n dh. eh.r.p.l.ey.n b.ix.

3 # i never get to sleep on the airplane beca

Syllabification

- Development of a Rule-Based System for automatic syllabification of phonemes' strings
- The syllabification is based on 2 principles:
    - a syllable contains a vowel, and only one;
    - a pause is a syllable boundary.



File-phon.palign TextGrid

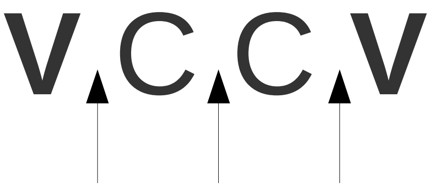
Phonemes

Sylabification

syllConfig-L.txt

File-phon.salign TextGrid

Syllables

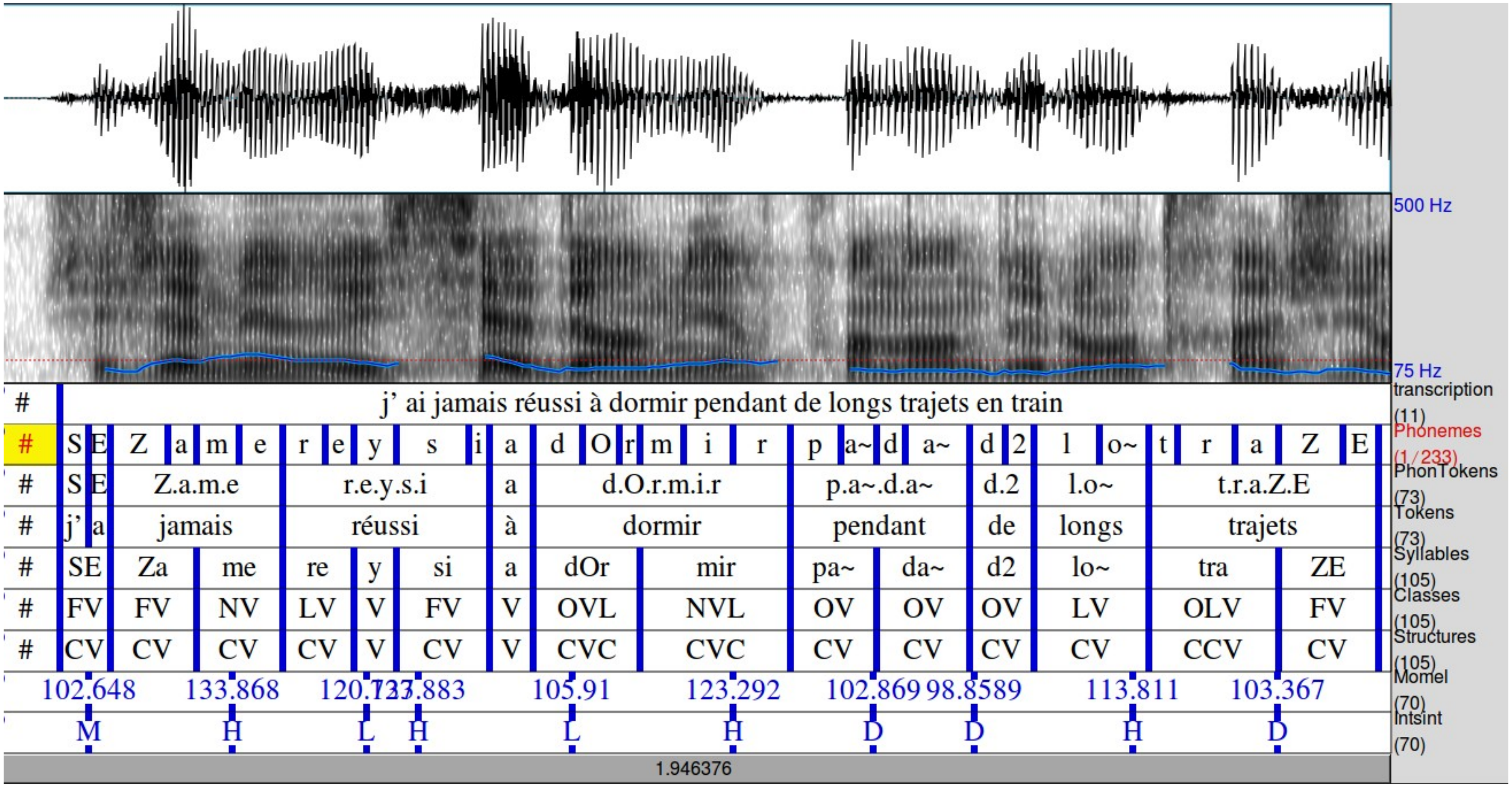


V C C V

Details in [1]

Syllabification: example

- Resources (FR and IT):
  - a configuration file with the phoneme set, the classes and all rules



j' ai jamais réussi à dormir pendant de longs trajets en train

S E Z a m e r e y s i a d O r m i r p a~ d a~ d 2 l o~ t r a Z E

S E Z.a.m.e r.e.y.s.i a d.O.r.m.i.r p.a~.d.a~ d.2 l.o~ t.r.a.Z.E

j' a jamais réussi à dormir pendant de longs trajets

SE Za me re y si a dOr mir pa~ da~ d2 lo~ tra ZE

FV FV NV LV V FV V OVL NVL OV OV OV LV OLV FV

CV CV CV CV V CV V CVC CVC CV CV CV CV CCV CV

102.648 133.868 120.7133.883 105.91 123.292 102.869 98.8589 113.811 103.367

M H L H L H D D H D

Resources summary

	FR	IT	ZH	EN
Dictionary : Number of entries	350k words and 300k variants	390k words and 5k variants	88k words (350 syllables)	121k words and 10k variants
Acoustic model: Data to train	Triphones - 7h30 CID +30min read	Triphones - 3h30 map-task	Monophones - 90min read	Triphones See voxforge.org

SLDR forge

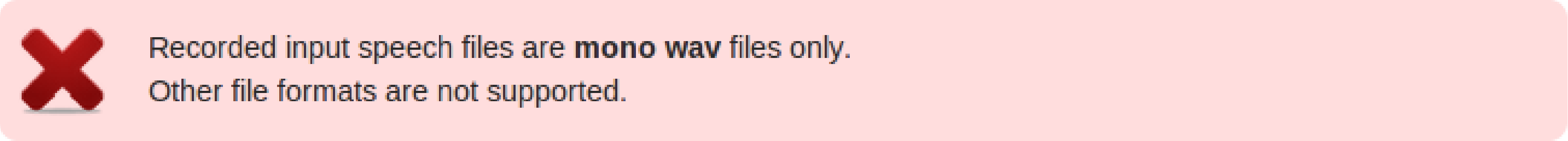
Evalita 2011

Eurom1

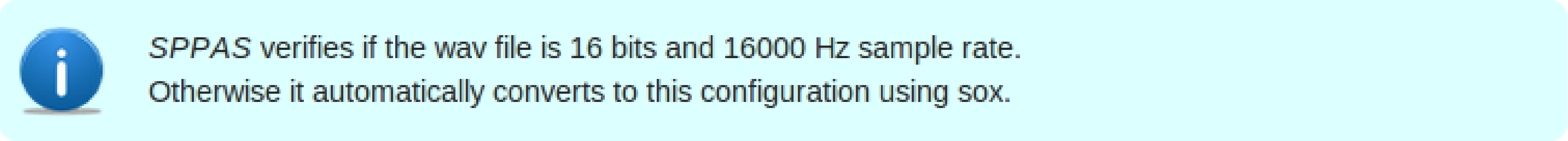
CMU dictionary

A few words about technical stuff...

- The transcription encoding must correspond to that of SPPAS dictionary:
  - UTF-8 for French, Chinese or Italian,
  - us-ascii for English.
- The transcription and the audio files must have the same name (except for the extension)



Recorded input speech files are **mono wav** files only.
Other file formats are not supported.



SPPAS verifies if the wav file is 16 bits and 16000 Hz sample rate.
Otherwise it automatically converts to this configuration using sox.

About

- URL: http://www.lpl-aix.fr/~bigi/sppas/
- Supported by the Equipex ORTOLANG
- SPPAS can achieve a set of automatic phonetic annotations of speech; results are depending on...
    - The resources quality;
    - The input wav quality;
    - The transcription quality...

References

[1] B. Bigi, C. Meunier, I. Nesterenko, R. Bertrand. *Automatic detection of syllable boundaries in spontaneous speech.* Language Resource and Evaluation Conference (LREC), pp 3285-3292, La Valetta, Malte, 2010.

[2] Brigitte Bigi (2011). *A Multilingual Text Normalization Approach.*2nd Less-Resourced Languages workshop, 5th Language Technology Conference, Poznàn (Poland). 2011.

[3] B. Bigi. *The SPPAS participation to Evalita 2011.* Working Notes of EVALITA 2011, Rome, Italy, ISSN: 2240-5186, January 2012.

[4] B. Bigi, D. Hirst. *SPeech Phonetization Alignment and Syllabification: a tool for the automatic analysis of speech prosody.* Speech Prosody, Shanghai, China, May 2012, .

[5] B. Bigi. *SPPAS: a tool for SPeech Phonetization Alignment and Syllabification.* Language Resource and Evaluation Conference (LREC), Istanbul, Turkey, May 2012. .

[6]. Bigi. *Forced Alignment on Spontaneous Speech for Italian: the SPPAS tool.* B. Magnini et al. (Eds.): EVALITA 2012, LNCS 7689, pp. 312--321. Springer, Heidelberg.

http://www.lpl-aix.fr/~bigi/sppas/