Automatic Speech Segmentation of French: Corpus Adaptation

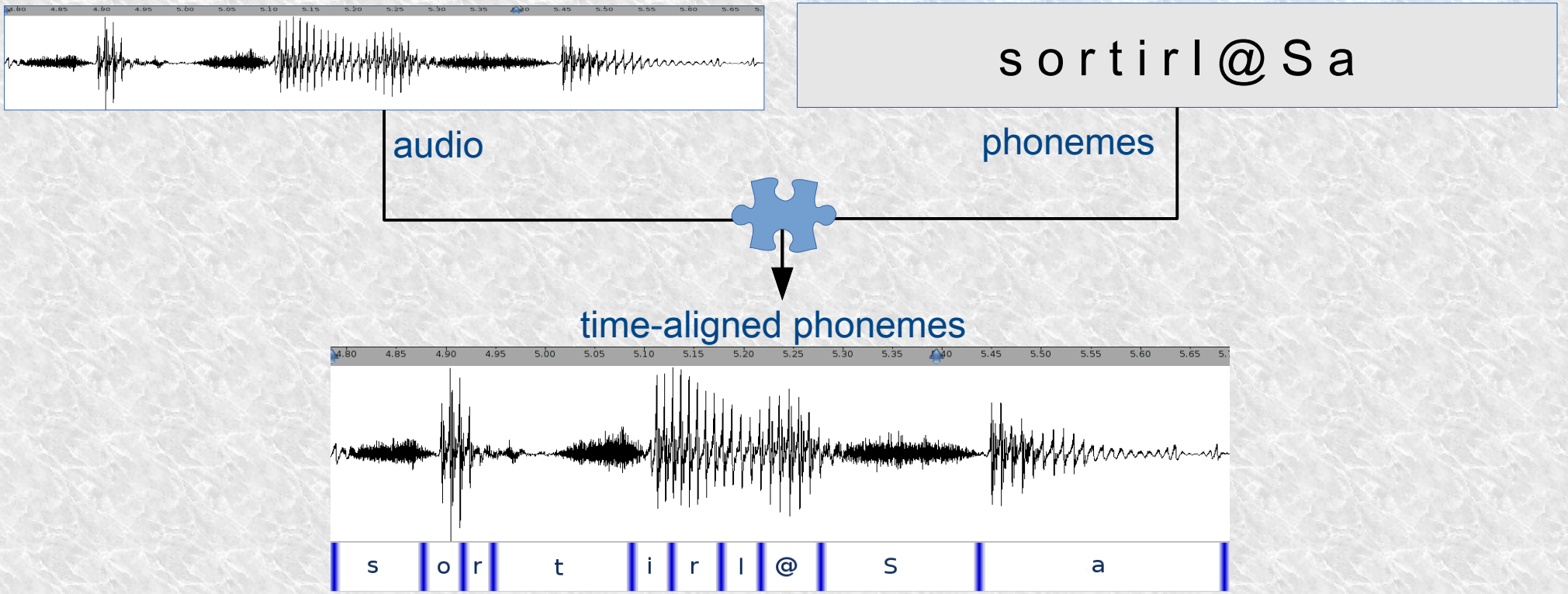
Brigitte Bigi

LPL - Aix-en-Provence - France

This work has been carried out thanks to the support of the A*MIDEX project (n° ANR-11-IDEX-0001-02)
funded by the « Investissements d'Avenir » French Government program,
managed by the French National Research Agency (ANR)

What is Speech Segmentation?

- the process of taking the phonetic transcription of an audio speech segment and determining where in time particular phonemes occur in the speech segment.



What's for?

- Determining the location of known phonemes is important to a number of speech applications:
  - When developing an ASR system, “good initial estimates … are essential” when training Gaussian Mixture Model (GMM) parameters (Rabiner and Juang, 1993, p. 370).
  - Knowledge of phoneme boundaries is also necessary in some cases of health-related research on human speech processing.
  - and other applications...

How to perform Speech Segm.?

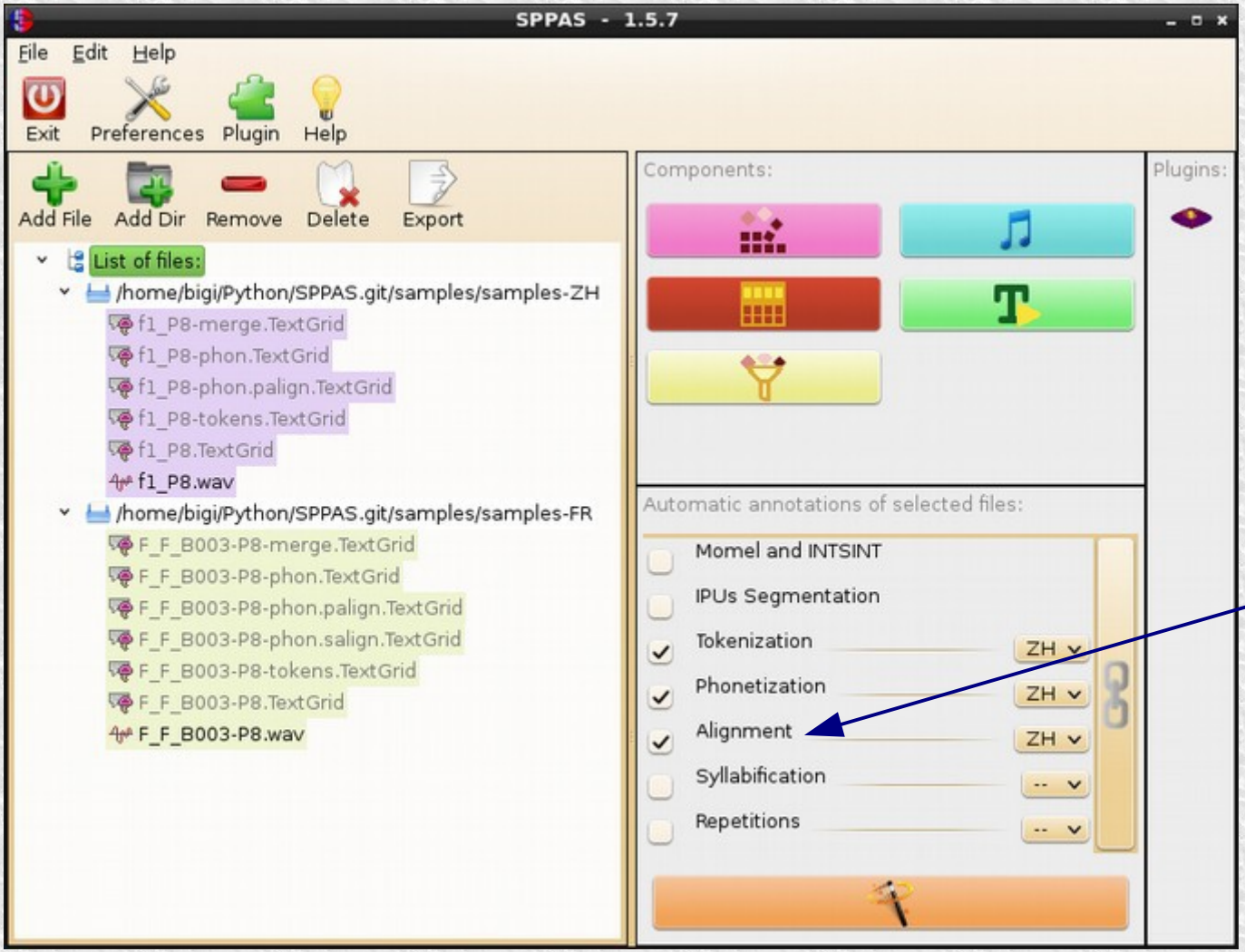
- Manually:
  - Manual alignment has been reported to take between 11 and 30 seconds per phoneme (Leung and Zue, 1984).
  - Manual alignment is too time consuming and expensive to be commonly employed for aligning *large corpora*.

How to perform Speech Segm.?

- Speech Recognition Engines that can perform Speech Segmentation:
  - HTK - Hidden Markov Model Toolkit
  - CMU Sphinx
  - Open-Source Large Vocabulary CSR Engine Julius
- Wrappers:
  - Prosodylab-Aligner: python / HTK
  - P2FA: python / HTK
  - and many others...

How to perform Speech Segm.?

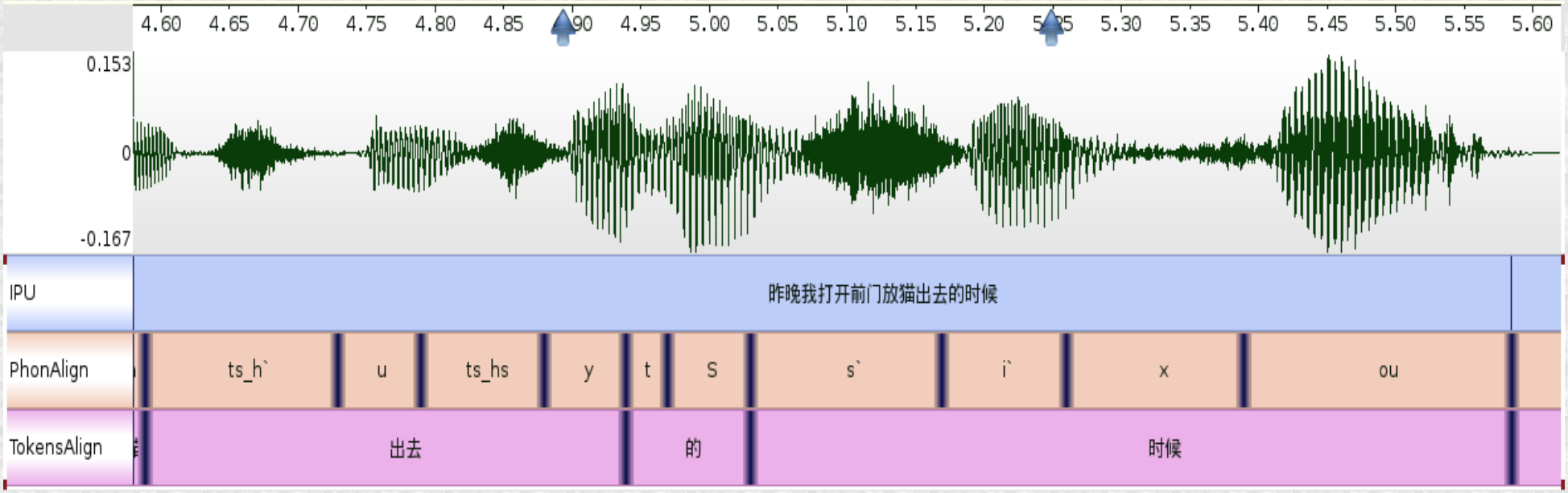
- Graphical User Interface: SPPAS (Bigi, 2012)



Speech Segm. is also called: **Alignment**

On which languages?

- SPPAS can perform speech segmentation of:
  - French, English, Italian, Spanish, Chinese, Taiwanese, Japanese.
- Requirement: **an acoustic model** for each language.

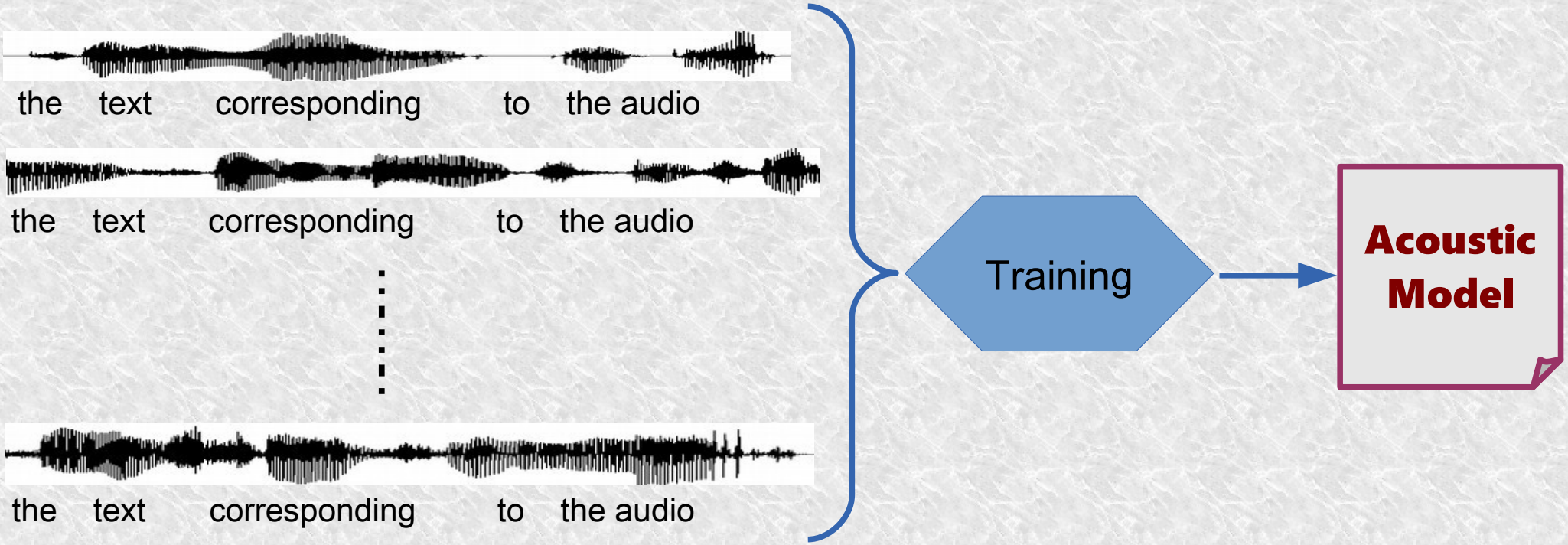


an Acoustic Model???

~h "S"
<BEGINHMM>
<NUMSTATES> 5
<STATE> 2
<MEAN> 25
3.865123e+00 -2.796230e+00 -2.741646e+00 -2.575907e+00 -2.209618e+00 -5.850142e+00 -3.059854e+00 2.294439e+00 6.802940e-01 -2.800637e+00 -1.763918e+00 3.845190e-01
1.286
847e+00 -1.407083e+00 -1.252665e+00 -1.862736e+00 -3.524270e-01 4.247507e-01 -1.773855e-02 7.232670e-01 -3.501371e-01 -8.653453e-01 -1.168209e+00 -5.176944e-01 1.447603e+
00
<VARIANCE> 25
1.297570e+01 2.348404e+01 3.699827e+01 3.013035e+01 4.785572e+01 4.348248e+01 4.807753e+01 4.529767e+01 4.452133e+01 4.717181e+01 5.047903e+01 4.394471e+01
5.295042e+00
3.326635e+00 3.577229e+00 3.221893e+00 6.327312e+00 4.562069e+00 5.920639e+00 7.081470e+00 5.766568e+00 5.546420e+00 5.610922e+00 4.105053e+00 1.246813e+00
<GCONST> 1.085982e+02
<STATE> 3
<MEAN> 25
4.182722e+00 -5.747316e+00 -5.573908e+00 -3.280269e+00 7.250799e-01 -1.220587e+00 7.397585e-02 4.036344e+00 5.651740e-01 -3.612718e+00 -3.532877e+00 -1.029424e+00
7.7643
20e-02 -1.490477e-01 -1.060979e-01 8.130542e-02 2.693116e-01 4.773618e-01 2.419368e-01 -1.171875e-01 -1.453947e-01 3.595677e-03 -1.755375e-01 -1.827260e-01 -9.910033e-02
<VARIANCE> 25
1.229548e+01 1.833777e+01 3.330074e+01 3.391322e+01 4.468183e+01 4.548661e+01 5.034616e+01 4.177621e+01 4.829255e+01 4.718935e+01 4.383722e+01 3.838983e+01
5.534610e-01
9.874231e-01 1.471683e+00 1.390052e+00 2.534417e+00 2.351494e+00 2.433162e+00 2.457205e+00 2.317599e+00 2.229505e+00 2.289994e+00 2.051025e+00 4.103379e-01
<GCONST> 9.480565e+01
<STATE> 4
<MEAN> 25
4.170075e+00 -3.602696e+00 -3.229792e+00 -2.666616e+00 -5.769264e-01 -2.755867e+00 -6.961405e-01 2.032978e+00 1.096958e-01 -2.195134e+00 -2.524131e+00 -9.696913e-01
7.72
3407e-02 1.414706e+00 1.097951e+00 8.257185e-01 -3.040556e-01 -2.347561e-02 -2.900199e-01 -1.342138e+00 -5.801741e-01 3.527923e-01 4.388814e-01 3.887816e-02 -1.326638e+00
<VARIANCE> 25
1.412758e+01 2.168075e+01 4.145230e+01 3.500136e+01 6.340505e+01 5.574141e+01 5.442813e+01 4.434394e+01 4.613047e+01 4.639702e+01 4.196549e+01 4.127845e+01
1.312419e+00
1.832024e+00 2.573012e+00 2.434281e+00 3.214828e+00 3.160381e+00 3.389642e+00 3.730893e+00 3.638973e+00 3.536761e+00 3.276227e+00 2.968326e+00 1.121088e+00
<GCONST> 1.025482e+02
<TRANSP> 5
0.000000e+00 1.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00
0.000000e+00 4.490560e-01 5.509440e-01 0.000000e+00 0.000000e+00
0.000000e+00 0.000000e+00 6.871416e-01 3.128584e-01 0.000000e+00
0.000000e+00 0.000000e+00 0.000000e+00 4.482542e-01 5.517458e-01
0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00
<ENDHMM>

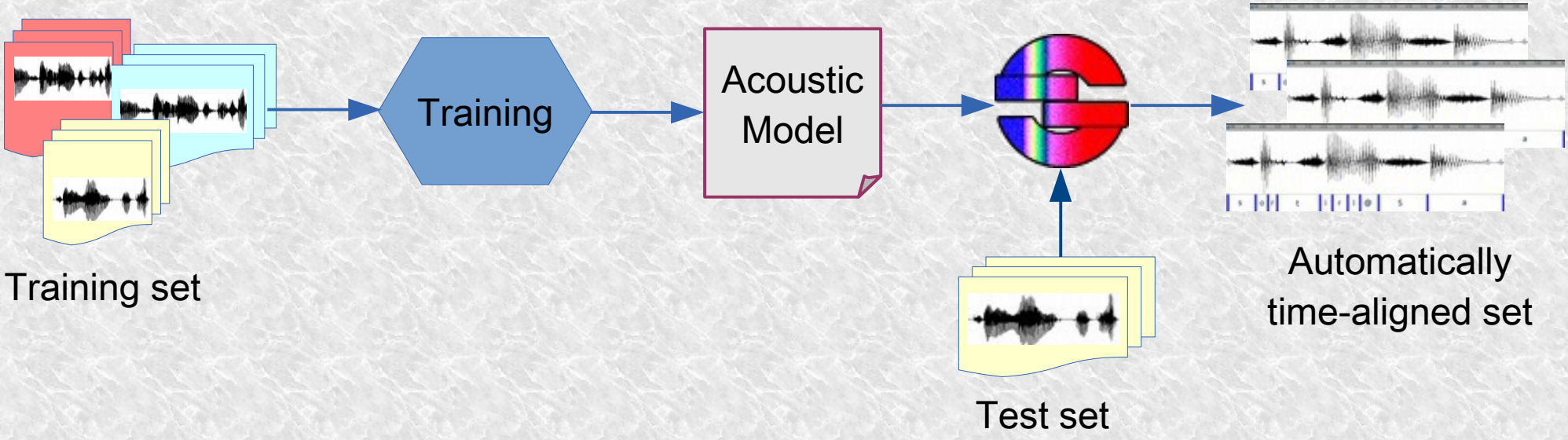
Yes, an Acoustic Model!

- It's a probability distribution (a 5-states HMM, blah blah blah). But, don't matter! It's not necessary to understand.
- The model is trained from data



Impact of the training data on the Speech Segmentation

- Measure:
  - **the impact of the quality vs quantity**
  - **the impact of the speech style**
- How to measure the impact of the training set on speech segmentation?

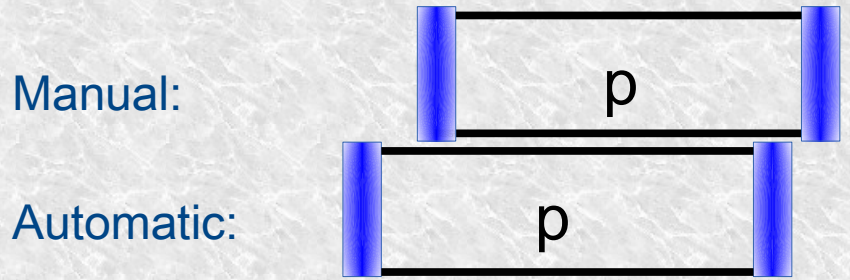


Evaluating Automatic Speech Segm.?

- Compare automatic segm.

  with a human segm.

- What to compare:
  - Duration
  - Position of phoneme boundaries
  - Middle of the phoneme

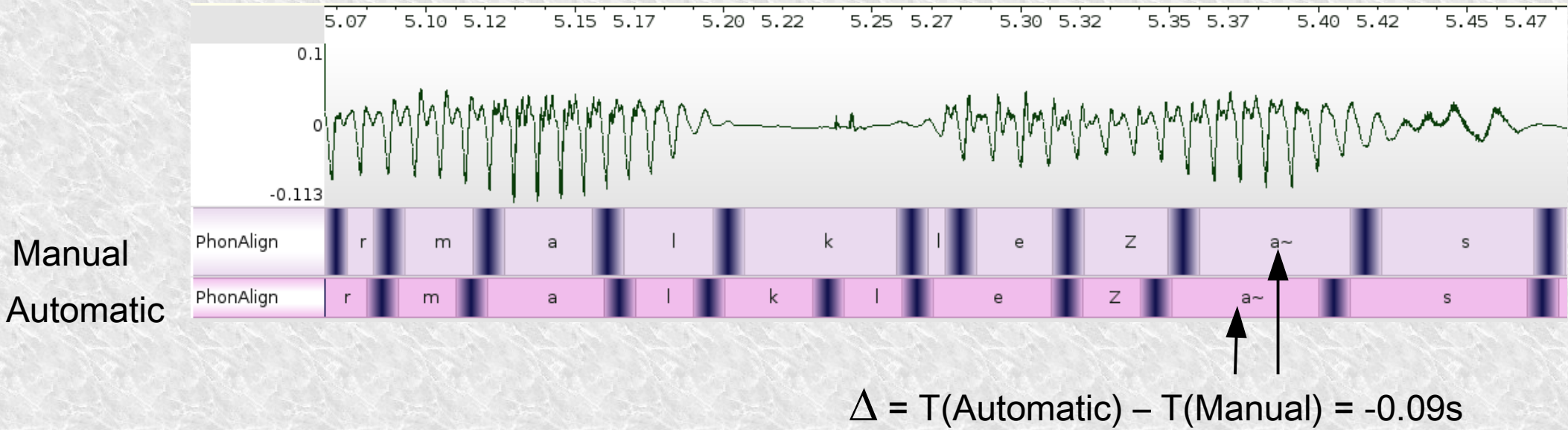


Evaluating Automatic Speech Segm.?

- Measure what percentage of the automatic-alignment boundaries are within a given time threshold of the manually-aligned boundaries.

Agreement of humans on the location of phoneme boundaries is, on average, 93.78% within 20 msec on a variety of English corpora (J-P. Hosom, 2008).

Manual vs Automatic



Δ = T(Automatic) – T(Manual) = -0.09s

- I preferred to evaluate **the center of the phonemes**

French Phoneset

Vowels	Consonants		Others
a	S	p	H
a~	Z	t	j
E	f	k	w
e	s	b	
i	v	d	sil is silence
o clusters /o/ and /O/	z	g	sp is short pause
o~			fp is filled pause
EU clusters /2/ and /@/	m		gb is garbage
EU9 is /9/	n		@@ is laugher
u			dummy
y	l		
U~ clusters /e~/ and /9~/	r clusters /r/ and /R/		

Training corpus

- The difficulties are that corpora are:
  1. from various file formats
  2. speech is segmented at various levels (phones, tokens, utterances)
  3. ortho. transcriptions are of various qualities
  4. corpora are of various speech styles

- Points 1 and 2 are solved by “scripting the data”
- Point 3 and 4 are the purpose of this study.

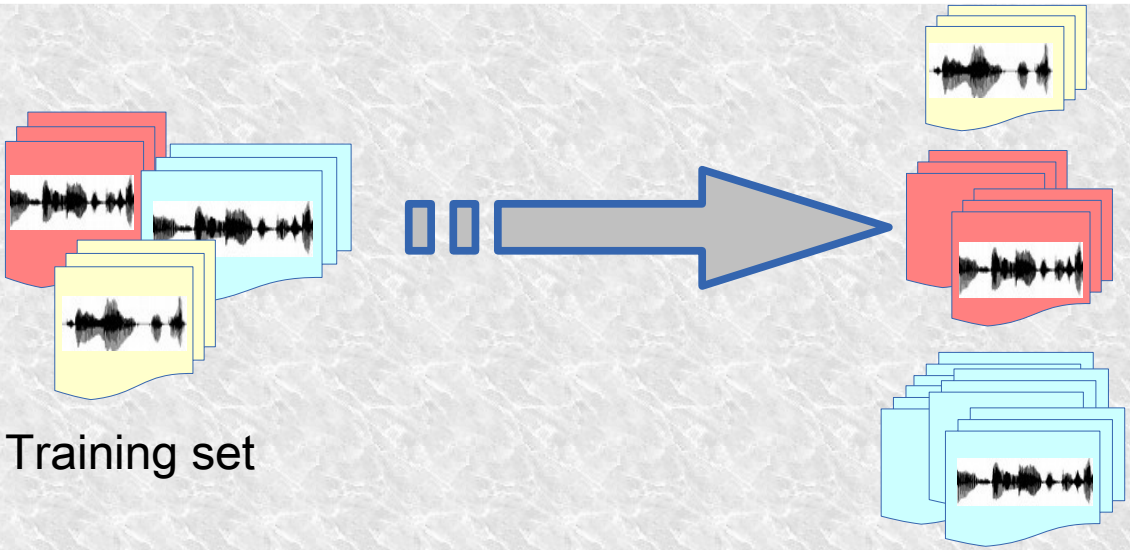
Training corpus

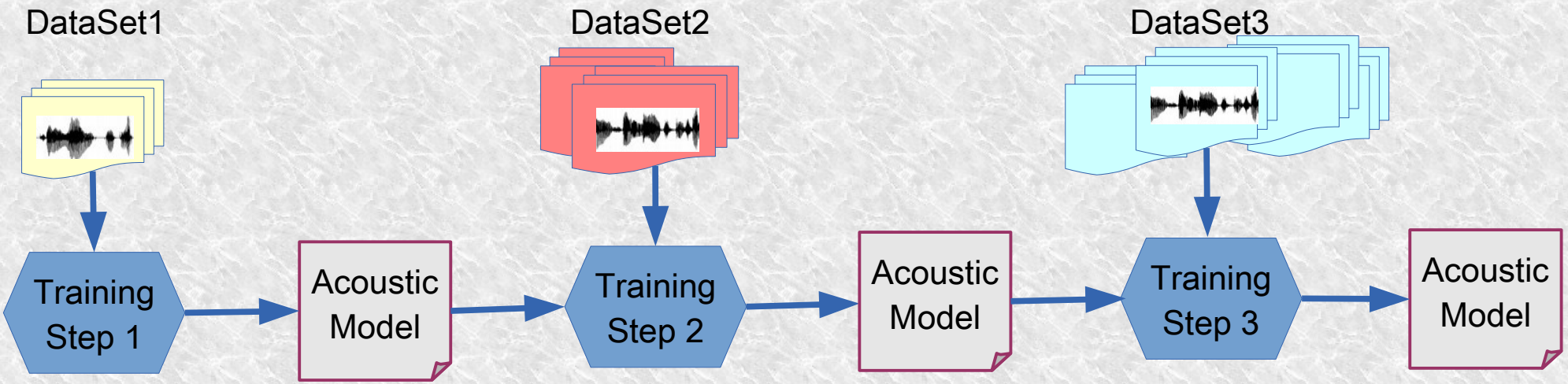
Corpus name	Transcription	Speech Duration	Style
Europe	Manually phonetized	40 min	Political debate
Eurom1	Ortho. standard manually tokenized	26 min	Read paragraphs
Read-Speech	Ortho. standard	98 min	Read sentences
AixOx	Ortho. standard	122 min	Read paragraphs
CID	Enriched ortho.	7h30min	Conversation
MapTaskAix	Standard ortho.	2h48min	Conversation Task-oriented

Test corpus

- Read Speech:
  - about 2 minutes of AixOx (1748 phonemes)
- Spontaneous Speech:
  - about 2 minutes of CID (1854 phonemes)
- Manually phonetized and segmented:
  - By one expert, then revised by another one.
- the test consists in:
  - Automatic segm. of the phonemes of each sentence;
  - Compare with the manual segmentation:
    - The time threshold is fixed to 40 ms.

Training procedure

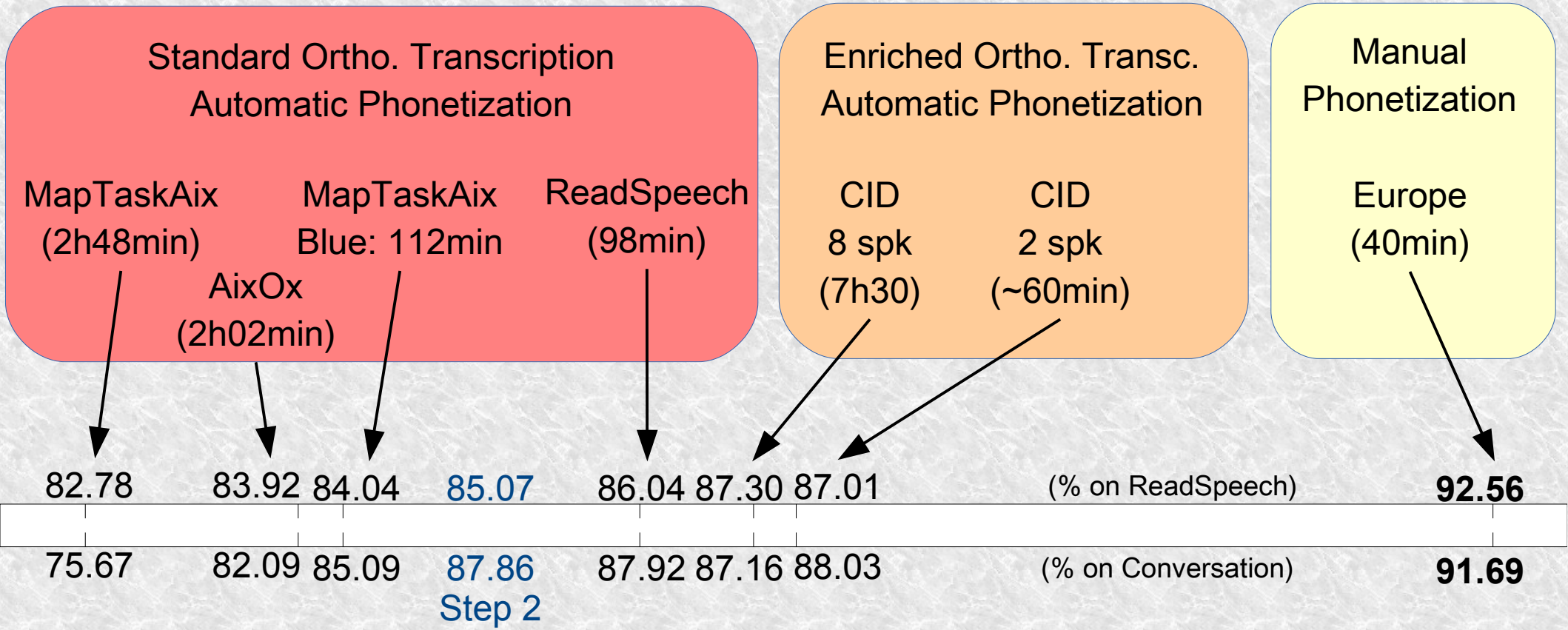




Question 1: quality vs quantity

- Perform step 1 from DataSet1 (3 min)
  - Δ < 40 ms:
    - Read speech 82.61%
    - Conversation 81.44%
- Perform step 2 from DataSet2 (42 min)
  - Δ < 40 ms:
    - Read speech 85.07%
    - Conversation 87.86%
- Split DataSet3:
  - perform as many step 3 as sub-sets.

Step 3. Compare sub-sets



The quality plays a decisive role

The sooner the better

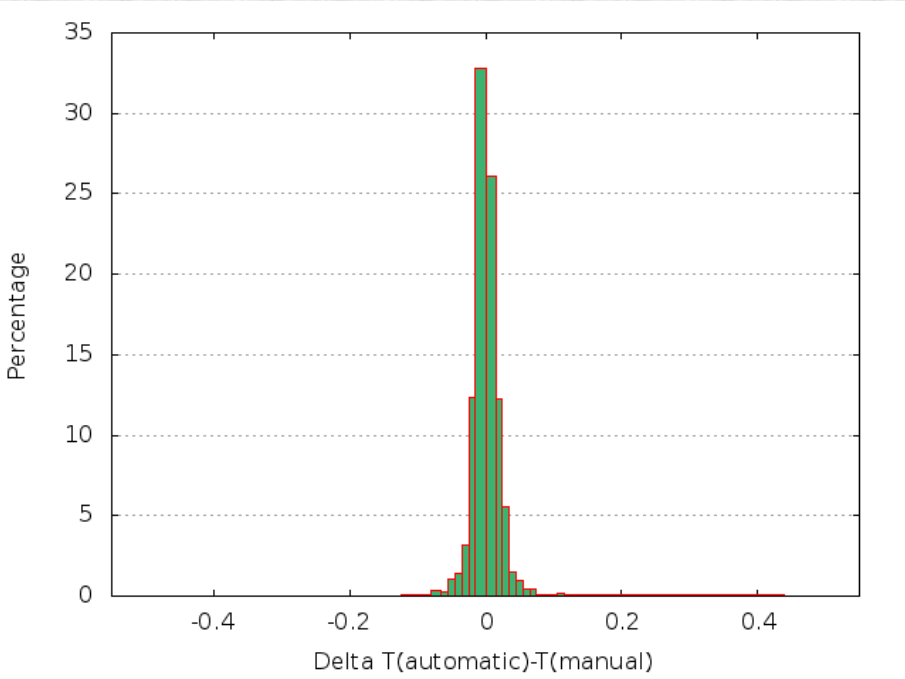
- Introduce all manually annotated data as soon as possible in the training procedure.
- Re-Perform steps 1 and 2:
  - Δ < 40 ms:
    - Read Speech: 94.16%
    - Conversational Speech: 92.77%
  - This model is (now) pretty stable.
- DataSet3:
  - perform as many step 3 as sub-sets.

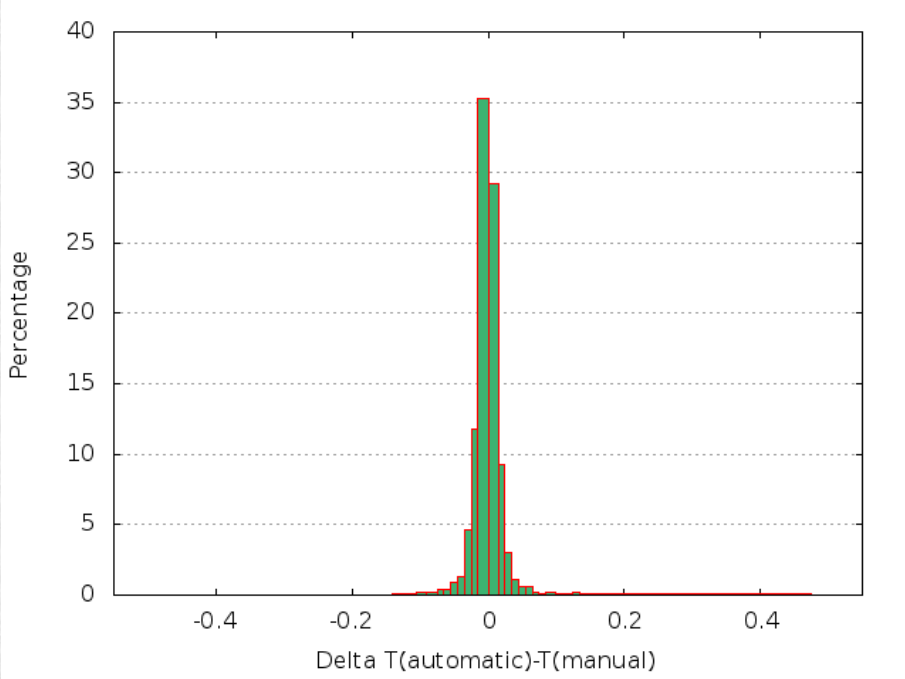
Question 2: speech style

	Δ < 40 ms Read Speech (%)	**Δ < 40 ms Conversational Speech (%)**
Step 2	**94.16**	92.77
Step 3. Read Speech	93.02	92.99
Step 3. Read Speech + AixOx	91.59	90.40
Step 3. MapTaskAix	89.93	89.21
Step 3. CID	93.25	92.23
Step 3. Read Speech + CID	93.36	**93.42**

The Acoustic Model

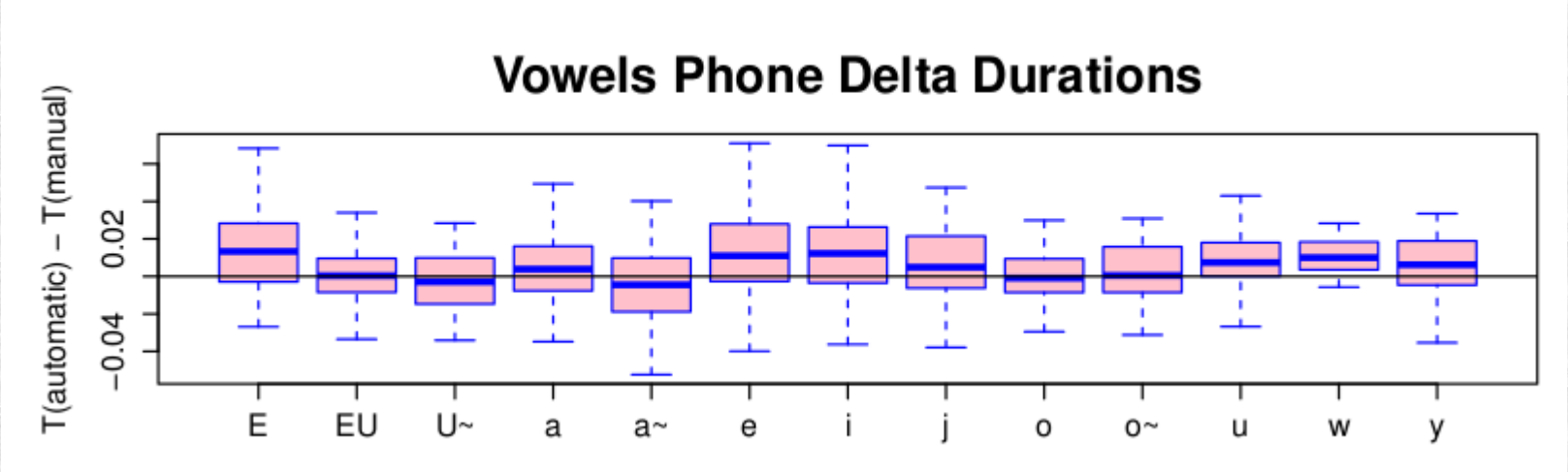
- The selected sub-sets of DataSet3 are useful to perform a 4th step to train a Triphone model:
  - Δ < 40 ms:
    - Read Speech: 95.08%
    - Conversational Speech: 95.42%



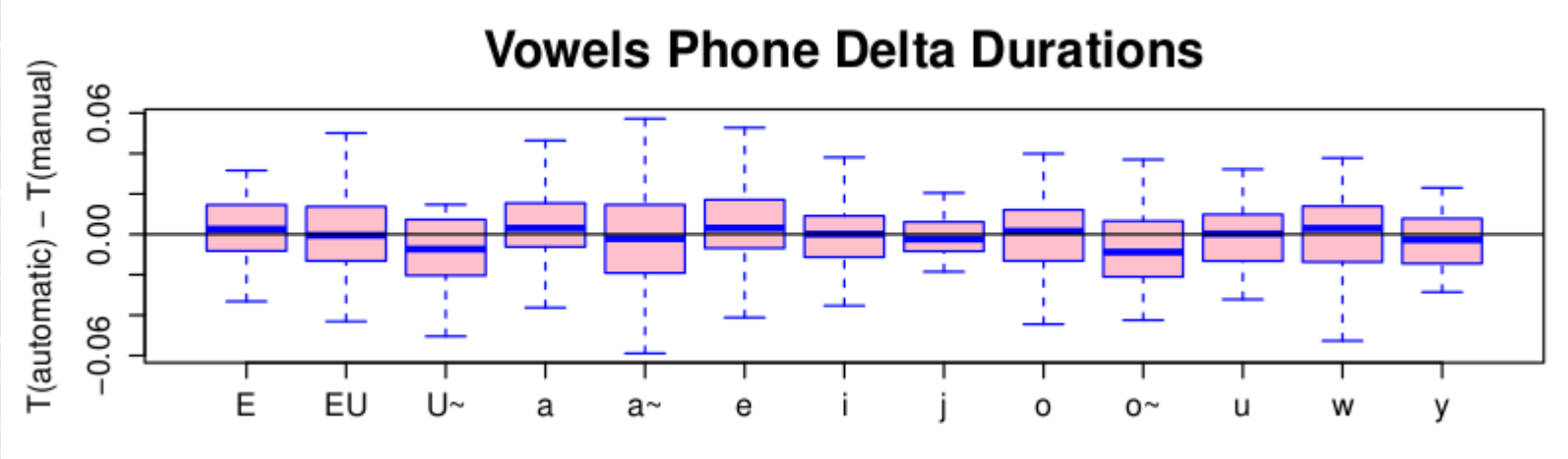


Other measures: Duration

read speech

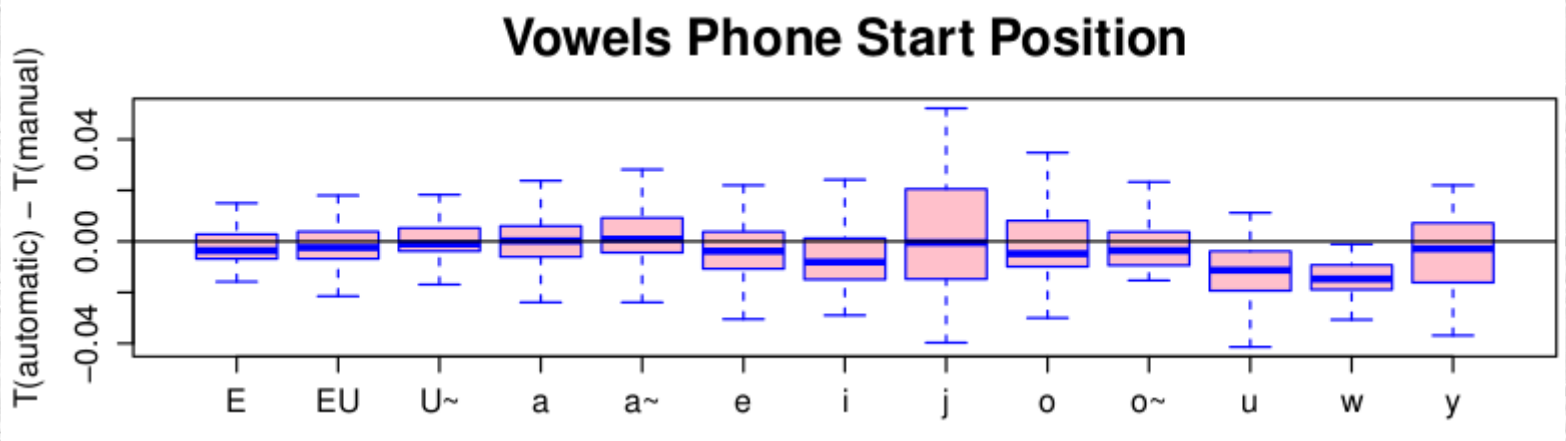


spontaneous speech

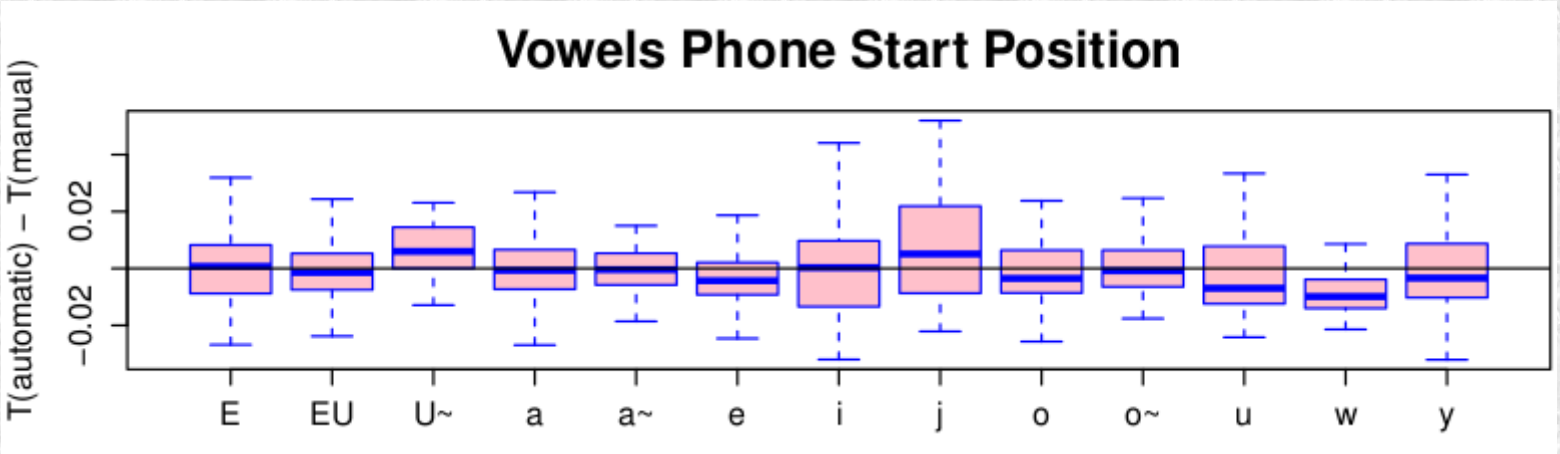


Other measures: start boundary

read speech

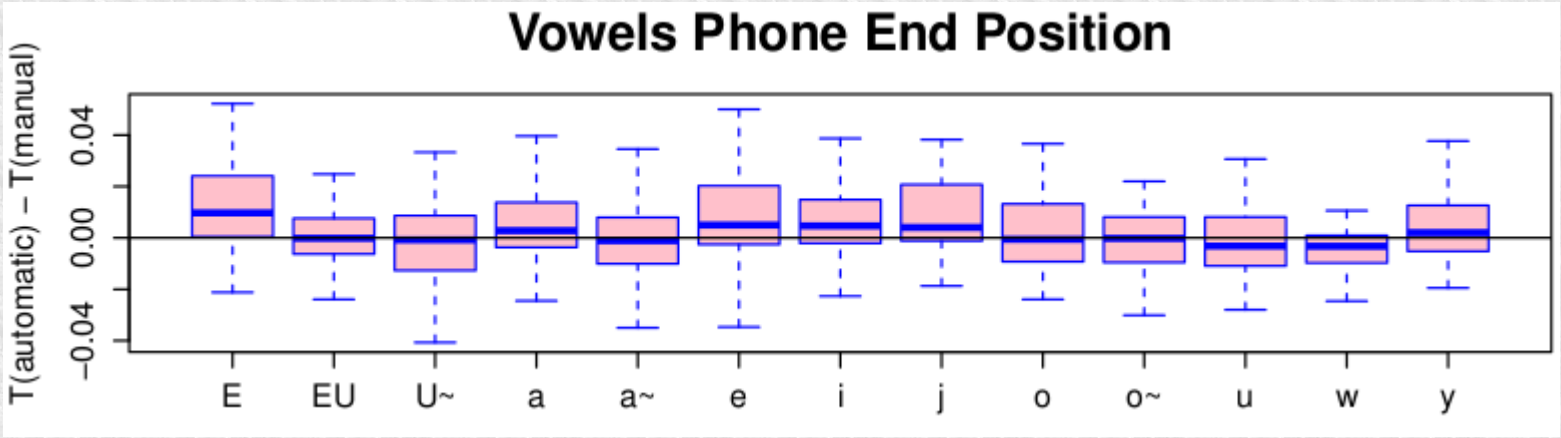


spontaneous speech

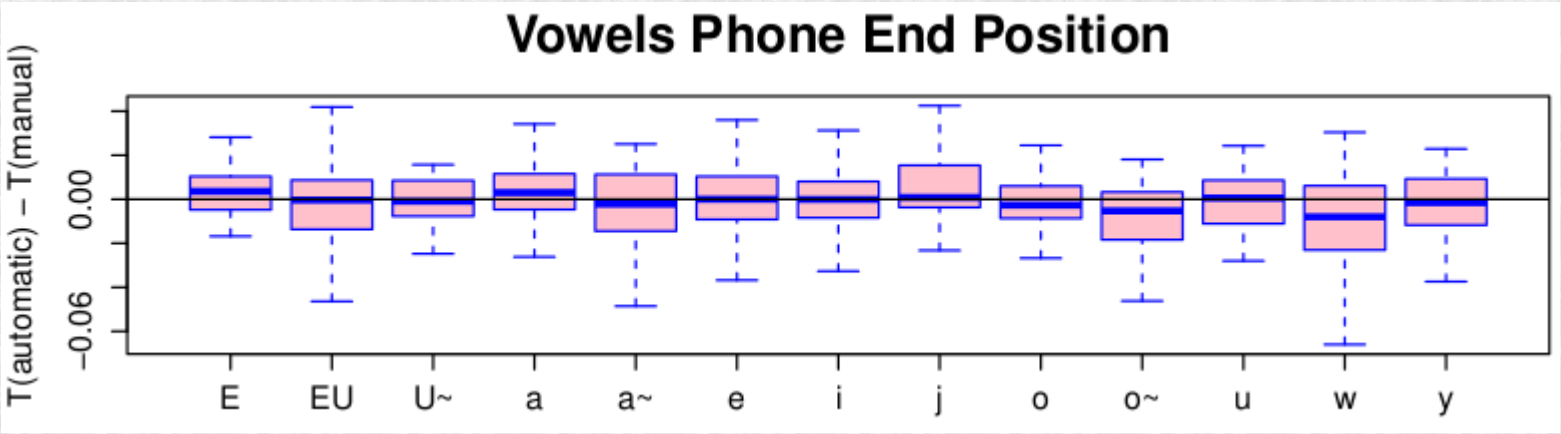


Other measures: end boundary

read speech



spontaneous speech



Conclusion

- This work enables advices to be given to data producers:
  - Requirements for a Monophone Acoustic Model:
    - at least 3 minutes of time-aligned data
    - 30-60 minutes manually phonetized data
  - Requirements for a Triphone Acoustic Model:
    - a pronunciation dictionary
    - at least 8 hours of “well”-transcribed speech
- From these data, I can train an acoustic model and add the new language in SPPAS!

Perspectives: Variamu Project

- Forced Alignment on Children Speech (FACS)
  - FA = Phonetization + Speech Segmentation (Bigi, 2011)
  - EVALITA 2014.

- Multilingual model:
  - speech segmentation of an un-trained language

References

- Hosom, J. P. (2009). Speaker-independent phoneme alignment using transition-dependent states. Speech Communication, 51(4), 352-368.
- Rabiner, L. R., & Juang, B. H. (1993). Fundamentals of speech recognition (Vol. 14). Englewood Cliffs: PTR Prentice Hall.
- Zue, V., Seneff, S., & Glass, J. (1990). Speech database development at MIT: TIMIT and beyond. Speech Communication, 9(4), 351-356.
- Bigi, B. (2012). SPPAS: a tool for the phonetic segmentation of speech. In LREC (Vol. 8, pp. 1748-1754).
- Bigi, B., Péri, P., & Bertrand, R. (2012). Orthographic Transcription: which Enrichment is required for phonetization?. In LREC (Vol. 8, pp. 1756-1763).
- Bigi, B. (2012). The SPPAS participation to Evalita 2011. In EVALITA 2011: Workshop on Evaluation of NLP and Speech Tools for Italian.