

# New techniques to measure fluency in speech automatically

Nivja de Jong

Webinar EALTA July 16th, 2020



Universiteit  
Leiden



**ICLON**

# New techniques to measure fluency in speech automatically

## *Co-authors:*

- Jos Pacilly & Willemijn Heeren, Leiden University

## *Funding:*

- British Council research grant
- NWO

## *Student assistants:*

- Danique van Aalst & Katarina Stankovic

## *Programme webinar:*

- Research project on new scripts to measure fluency
- Discussion
- Brief tutorial on how to use the scripts in PRAAT

De Jong, N. H., Pacilly, J., & Heeren, W. (2020, July 3). Praat scripts to measure fluency automatically. Retrieved from [osf.io/w3r7t](https://osf.io/w3r7t)

And: <https://sites.google.com/view/uhm-o-meter/home>

# Fluency defined in applied linguistics

- *Broad notion* = speaking proficiency in a second language
  - Content
  - Wording
  - (grammatical) Accuracy
  - Pronunciation
  - Tempo / fluency
- *Narrow notion* = *Part* of speaking proficiency
  - speed of speech
  - few pauses
  - few “uhm”s

# Speaking proficiency

Views on speaking proficiency/language ability:

- Communicative competence (Hymes, Canale & Swain, Celce-Murcia)
- Language ability (Bachman & Palmer)

# Communicative speaking competence: KNOWLEDGE OF

1. Words and chunks;
2. Morphosyntax;
3. Pronunciation;
4. Nonverbal gestures;
5. Pragmatic knowledge;
6. Strategies for speaking;
7. Rules for interaction.

# Communicative speaking competence: SKILLS IN

## ***Fast access to:***

1. Words and chunks;
2. Morphosyntax;
3. Pronunciation;
4. Nonverbal gestures;
5. Pragmatic knowledge;
6. Strategies for speaking;
7. Rules for interaction.

# Fluency in language assessment

*For example:*

**ACTFL-OPI, APTIS, IELTS, TOEFL:**

**As part of their assessment of speaking proficiency**

***Judges have instructions to consider as disfluent speech:***

- Occurrence of (unnatural) filled and unfilled pauses
- Slow (or unnatural, staccato) pace

# Relating objective measures to subjective ratings

Instructed judges rate *fluency*:

- 84% of variance explained by objective measures in speech

Manipulated speed (speech rate and articulation rate):

- Same effect on ratings of native and nonnative speech

Bosker et al., 2013; Bosker et al., 2014



# Measuring fluency automatically

PRAAT-script, 2009:

- silent pauses (frequency and duration)
- speed of speech (articulation rate)

Missing:

- filled pauses (frequency and duration)
- repetitions (frequency)
- repairs (frequency)

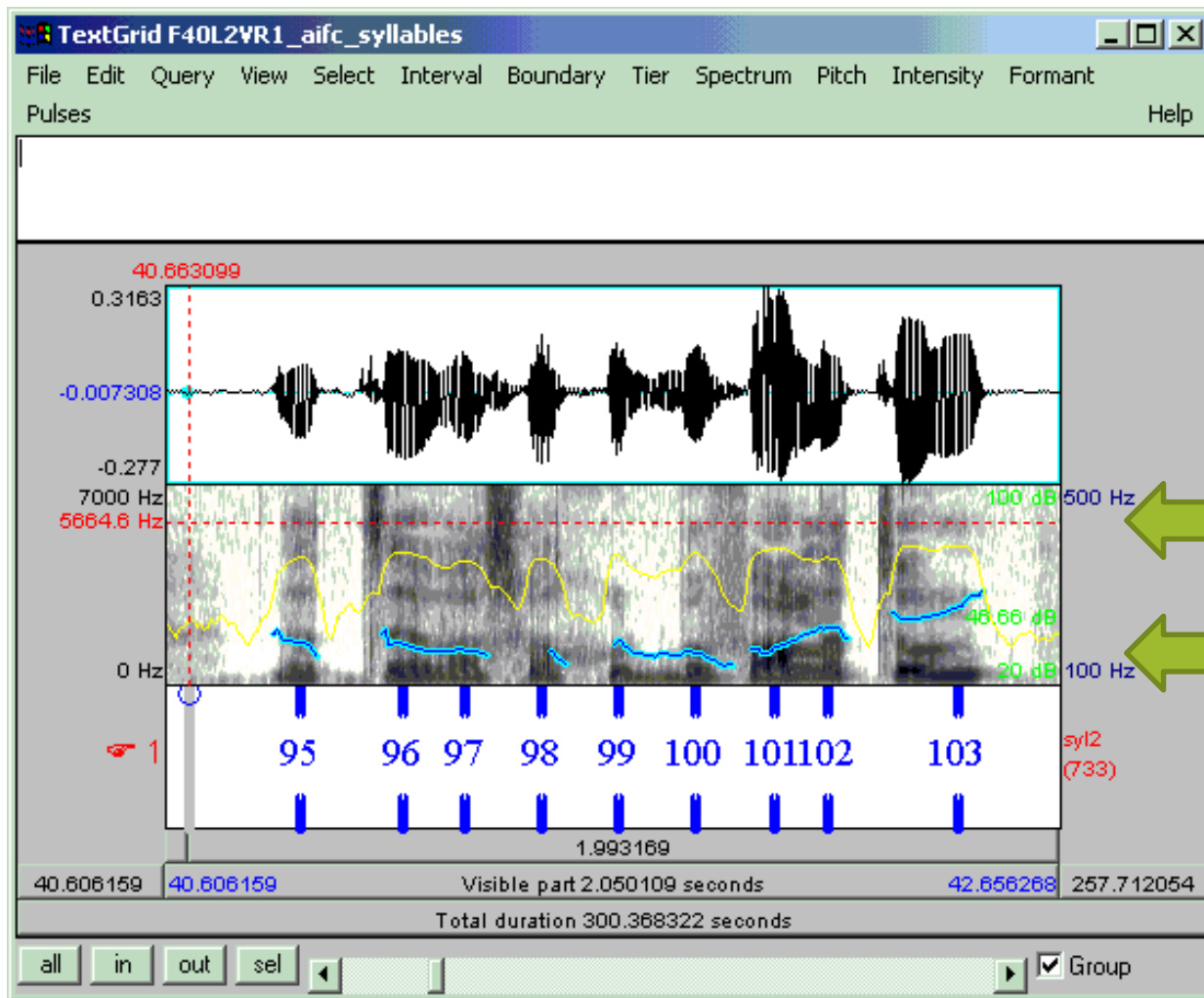
De Jong & Wempe, 2009

# Current project: add filled pauses

PRAAT-scripts, 2020:

- silent pauses (frequency and duration)
- speed of speech (articulation rate)
- *filled pauses (frequency and duration)*

# Measuring syllable nuclei



Intensity (dB):  
peaks and dips

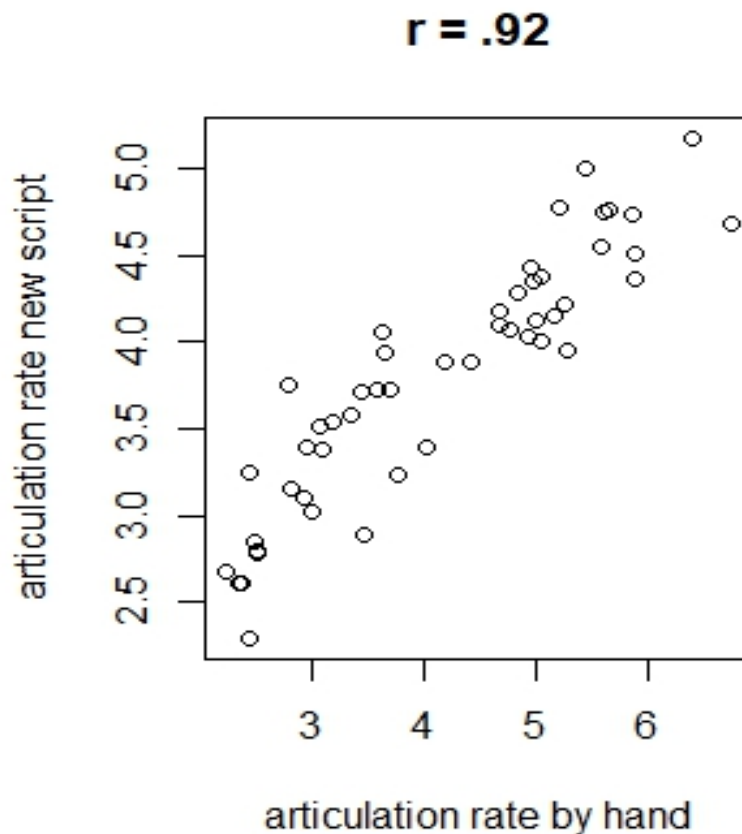
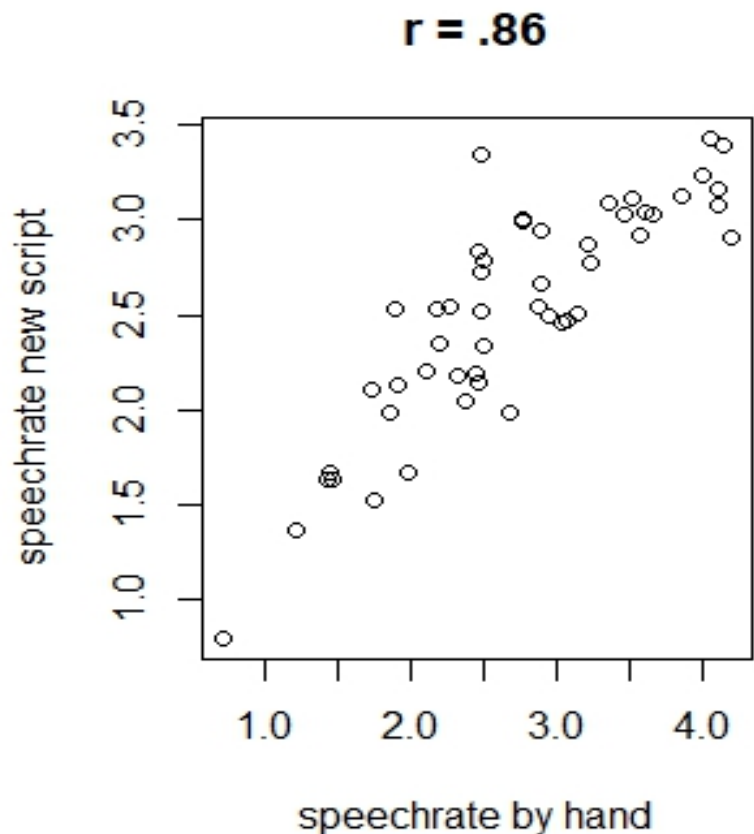
Voiced peaks

De Jong & Wempe, 2009

# Validation rewritten script for articulation rate

PRAAT-script "syllable nuclei v3":

- Preceding and next dip in intensity indication improved
- New (more efficient) PRAAT-syntax implemented



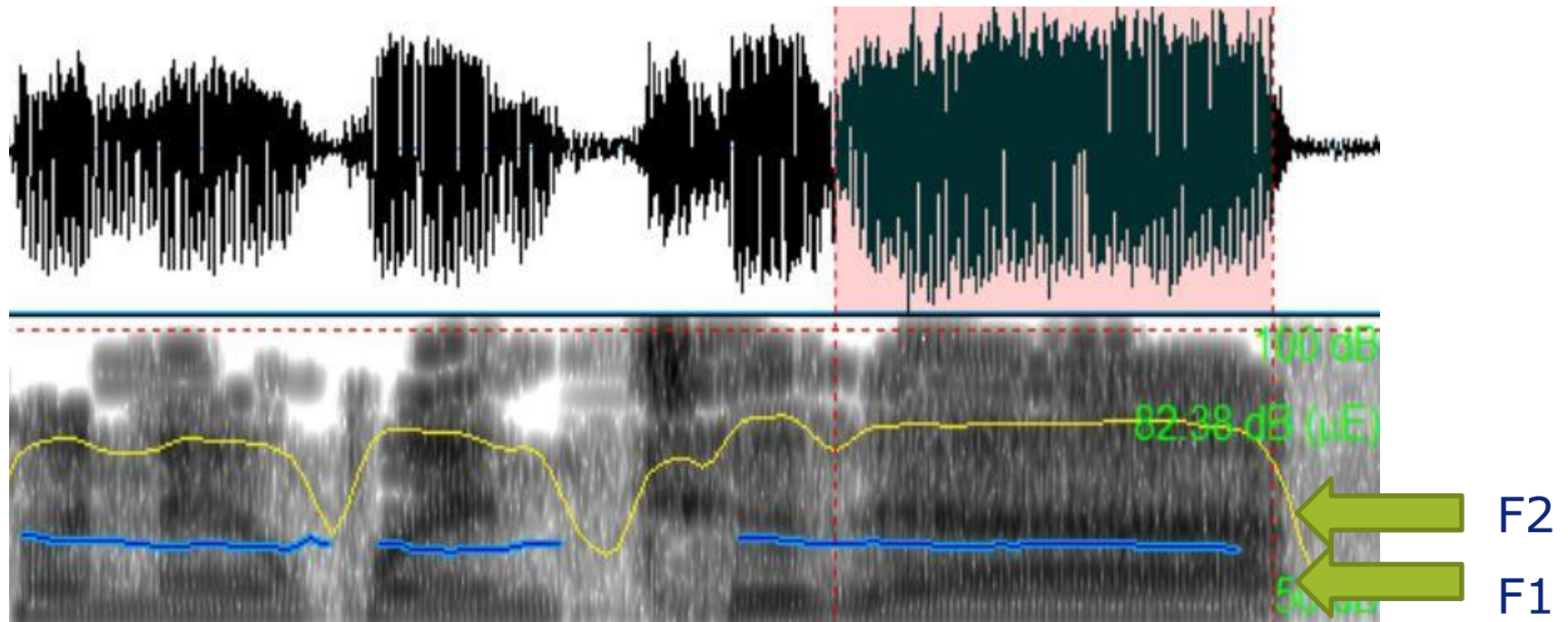
# Characteristics of filled pauses

Previous research on characteristics of filled pauses:

- Duration (long)
- Variation of F0 (little)
- Height of F0 (low)
- Variability in formants F1 through F3 (little)

e.g., Audhkhasi et al., 2009; Clark & Fox Tree, 2001;  
Hughes et al., 2016; Kaushik et al., 2010;  
Shriberg & Lickley, 1993; Stouten & Martens, 2003;  
Verkhodanova & Shapranov, 2016

# Characteristics of filled pauses



when I went to the...uuhhh

# Characteristics of filled pauses (2)

Previous research suggest filled pauses are “lazy”, or close to a schwa ([ə]) ([Wikipedia-link](#)):

- For (American) English, filled pause more like a mid-open back unrounded vowel ([ʌ]), thus distance F1 and F2 relatively small. For both [ə] and [ʌ]: F3 relatively high, with lips not rounded. ([Wikipedia-link](#))
- Little effort in articulation: the current vowel is close to the average vowel of that speaker.

Ladefoged & Johnson, 2011; Reetz & Jongman, 2009;  
Shriberg, 2001; Vasilescu and Adda-Decker, 2007

# Research aims

- 1) Create a PRAAT script that measures aspects of fluency automatically, including information on filled pauses
- 2) Test the accuracy of the script with respect to filled pauses for two types of speech data (Dutch and English speaking performances in language assessment settings)
- 3) Gauge validity of the automatic measures of filled pauses for the purpose of language assessment



# Research aims

- 1) Create a PRAAT script that measures aspects of fluency automatically, including information on filled pauses**
- 2) Test the accuracy of the script with respect to filled pauses for two types of speech data (Dutch and English speaking performances in language assessment settings)
- 3) Gauge validity of the automatic measures of filled pauses for the purpose of language assessment

# Corpora

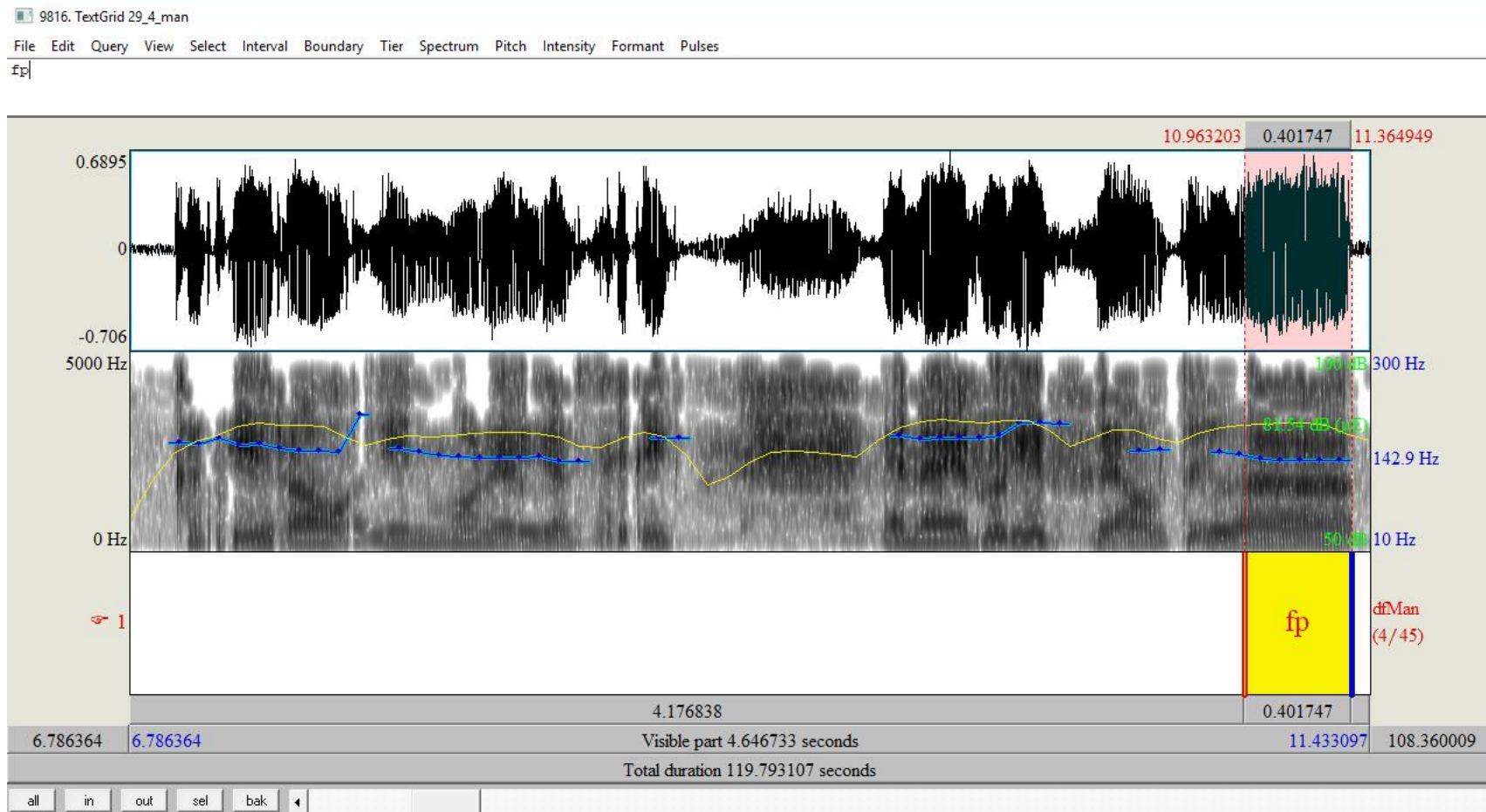
## Primary *language assessment* corpora:

- Tavakoli et al. (2017) corpus of APTIS data for English language assessment data, including manually measured filled pauses (subset; 60 files, ~120 minutes)
- Bosker et al. (2013) corpus of WISP data for Dutch language assessment-type data (114 files, ~38 minutes)

## Secondary *informal interview* corpus:

- Orr & Quené (2017): L1 Dutch and L2 English speech data from the same (F) speakers (118 files, ~240 minutes)
- Orr & Quené (2017): L1 English (mixed American and British) (12 files, ~24 minutes)

# PRAAT TextGrids: manual annotations



# Measures

For all automatically determined syllable nuclei, determine syllable boundaries and then measure for all syllable-intervals:

- Duration
- F0z: fundamental frequency, normalized per speaker
- sdF0: standard deviation of the F0;
- Distance between F1 and F2;
- F3;
- Standard deviations of F1, F2, F3;
- Mean absolute deviations of F1, F2, F3 to the globally measured F1, F2, F3 (per speaker);

# Optimal generalized linear models

- For primary corpora, on 70% of the data (training set), determine optimal models predicting which syllables are, and which syllables are not manually annotated to be filled pauses
- repeated cross-validation: for each step in the analyses, cross-validation carried out with 10 folds, repeated 10 times; outcomes evaluated in delta's

Dutch Score =  $8.62 \times \sqrt{\text{duration}} - 0.36 \times F0z - 0.11 \times (F2 - F1) + 0.21 \times F3 - 1.36 \times \sqrt{\text{standard deviation of F2}} - 1.02 \times \sqrt{\text{standard deviation of F3}} - 0.72 \times \sqrt{\text{absolute deviation of F1}} - 1.62 \times \sqrt{\text{absolute deviation of F2}}$

English Score =  $4.73 \times \sqrt{\text{duration}} - 0.29 \times F0z - 0.20 \times (F2 - F1) + 0.31 \times F3 - 0.32 \times \sqrt{\text{standard deviation of F1}} - 1.38 \times \sqrt{\text{standard deviation of F2}} - 0.10 \times \sqrt{\text{absolute deviation of F1}} - 0.80 \times \sqrt{\text{absolute deviation of F2}}$

# Determine optimal cut points

- Area under the curve (AUC; Fawcett, 2006) for the training data using cutpointr package (Thiele, 2019) in R
- Cut point Dutch: 2.7094
- Cut point English: 3.4942

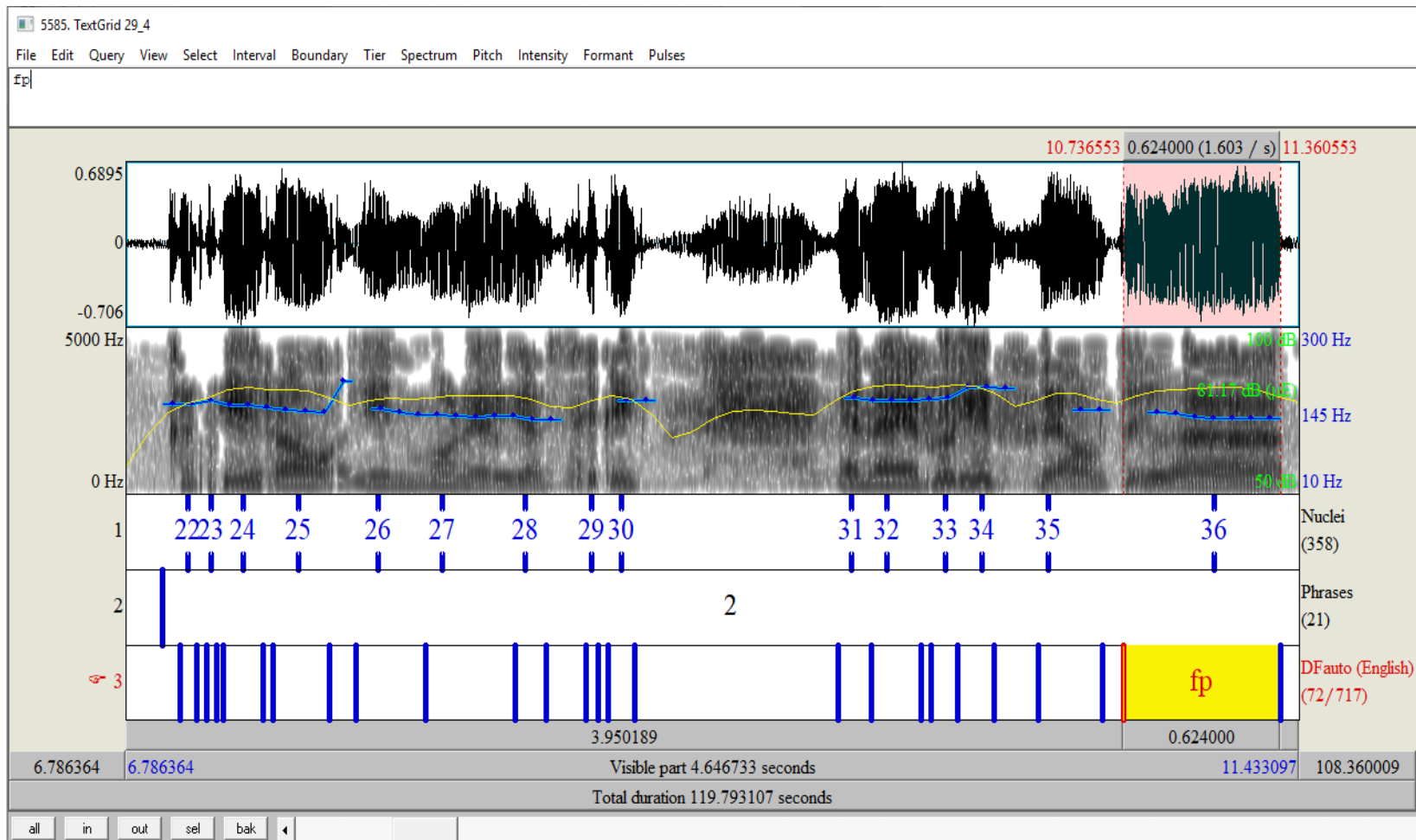
# Algorithms in lay terms

In both Dutch and English L2 , syllables that are relatively

- Long
- Low in pitch
- Stable
- Like a schwa/mid-open back vowel
- Close to average vowel

are potential filled pauses

# TextGrid of automatically determined syllables and filled pauses





# Research aims

- 1) Create a PRAAT script that measures aspects of fluency automatically, including information on filled pauses
- 2) Test the accuracy of the script with respect to filled pauses for two types of speech data (Dutch and English speaking performances in language assessment settings)**
- 3) Gauge validity of the automatic measures of filled pauses for the purpose of language assessment

# Testing *local* accuracy

On 30% test data of primary corpora, example of “confusion matrix” for Dutch test data:

	Manual categories	
Automatic categories	FP	normal
FP	107	223
Normal	58	1596

# Testing *local* accuracy

On 30% test data of primary corpora:

	Dutch test data (n = 1984)	English test data (n = 5935)
Sensitivity	0.65	0.56
Specificity	0.88	0.86
Precision	0.32	0.33
Accuracy	0.86	0.83

## Testing *local* accuracy (2)

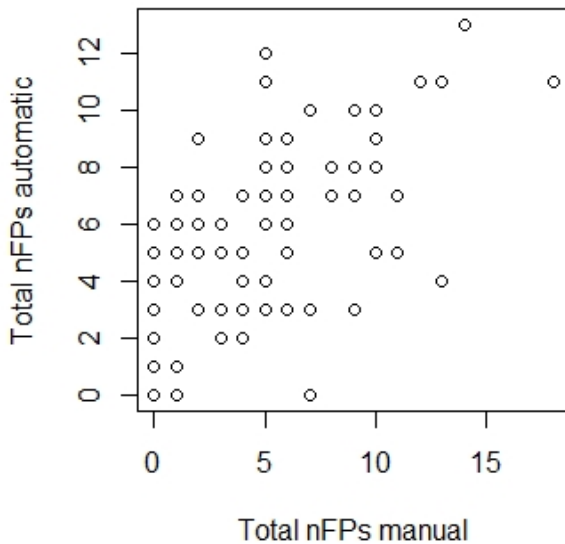
On complete secondary corpora (without training):

	Dutch L1 data (n = 24006)	English L2 data (n = 22266)	English L1 data (n = 4179)
Sensitivity	0.75	0.76	0.68
Specificity	0.91	0.85	0.87
Precision	0.28	0.21	0.20
Accuracy	0.90	0.84	0.86

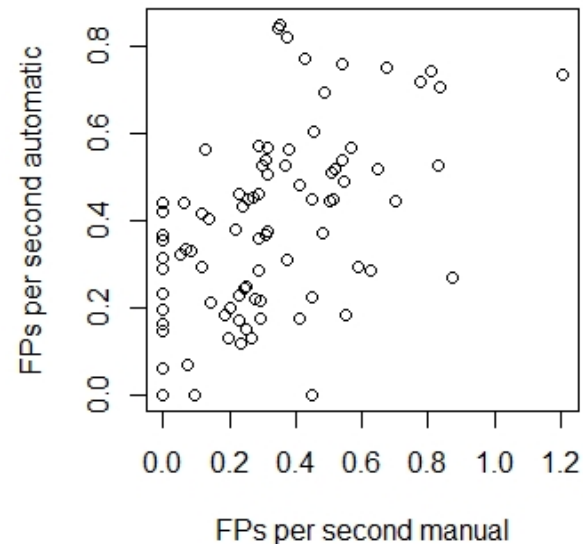
Low precision: goes up with higher cutpoint/threshold

# Testing *global* accuracy: Dutch correlations – higher threshold

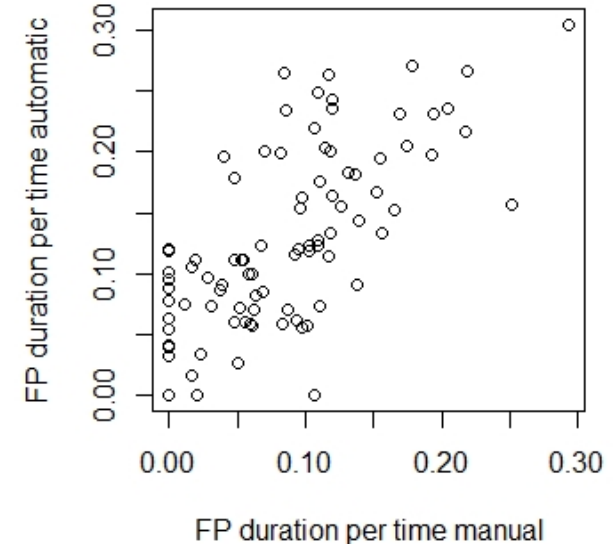
Pearson  $r = 0.57$



Pearson  $r = 0.53$

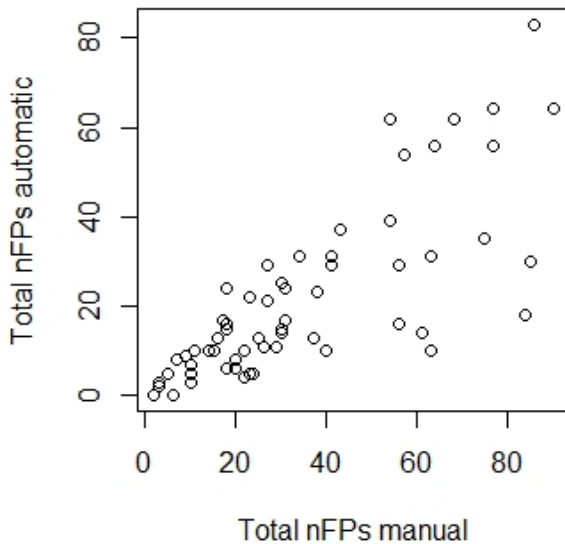


Pearson  $r = 0.69$

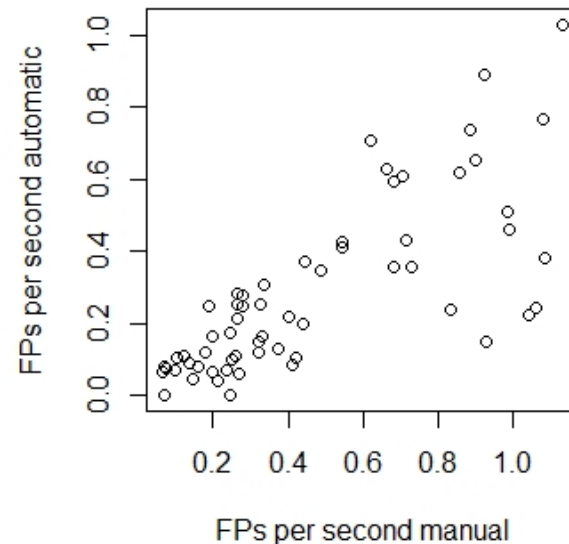


# Testing *global* accuracy: English correlations – higher threshold

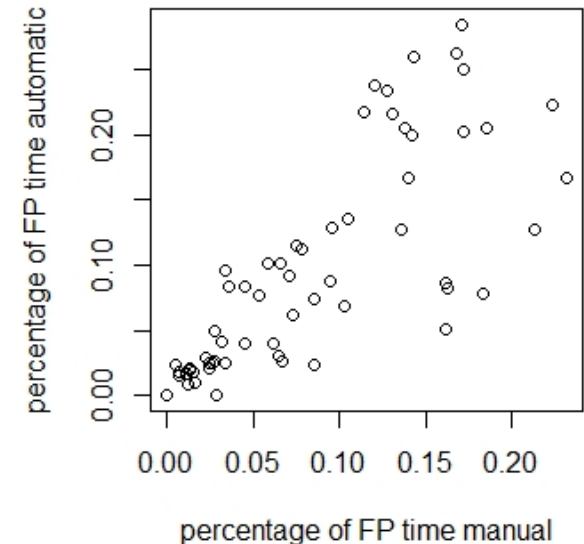
Pearson  $r = 0.78$



Pearson  $r = 0.75$



Pearson  $r = 0.77$



# Research aims

- 1) Create a PRAAT script that measures aspects of fluency automatically, including information on filled pauses
- 2) Test the accuracy of the script with respect to filled pauses for two types of speech data (Dutch and English speaking performances in language assessment settings)
- 3) Gauge validity of the automatic measures of filled pauses for the purpose of language assessment**

# Predicting human fluency ratings

Primary corpora have been judged on fluency:

- WISP corpus (Bosker et al., 2013) specific instructions for raters about fluency
- APTIS corpus (Tavakoli et al., 2017) only those speaking performances chosen where fluency judgement is the same as holistic judgement

<b>Total R<sup>2</sup></b>	<b>Dutch</b>	<b>English</b>
<b>Predictors</b>	<b>(n = 90)</b>	<b>(n = 60)</b>
<b>Manual, only FP-measures</b>	0.15	0.01 (ns)
<b>Manual, including speech rate</b>	0.75	0.43
<b>Automatic, only FP-measures</b>	0.16	0.02 (ns)
<b>Automatic, including speech rate</b>	0.53	0.32



# Limitations

## *Limitations*

Dutch primary L2 corpus: small, with short excerpts

English primary L2 corpus: quality recordings below standards for precise phonetic analyses

## *Nevertheless...*

Performance on secondary corpora similar/promising, even for L1 data.

## *However*

Too many false positives (leading to low 'precision'). Perhaps they are hesitated/lengthened syllables (like example on slide 14)

# Discussion

## *Accuracy of algorithms:*

Difficult to compare numbers of sensitivity, specificity, and precision to those of other automatic systems

## *Validity of algorithms:*

For the English data, neither the automatic nor the manual filled pauses could predict the human fluency ratings...

For the Dutch data, automatic/manual filled pause measurements equally predicted human fluency ratings

-> Dutch raters were specifically instructed to take into account filled pauses, but English raters were not

# BRIEF TUTORIAL

How to run the two scripts to get a fluency report?

# Step 0: download PRAAT

from <http://www.fon.hum.uva.nl/praat/>

NB: you need a new (2020) version of PRAAT (6.1.1.X) if you already have PRAAT on your computer.

# Step 1a: save PRAAT-scripts

Open PRAAT. Then for each script separately, copy-paste the script from this page:

<https://sites.google.com/view/uhm-o-meter/scripts>

For each script, select everything from the first line starting with "###" to the bottom of the page (but scripts end with a line stating "endproc").

## Step 1b: save PRAAT-scripts

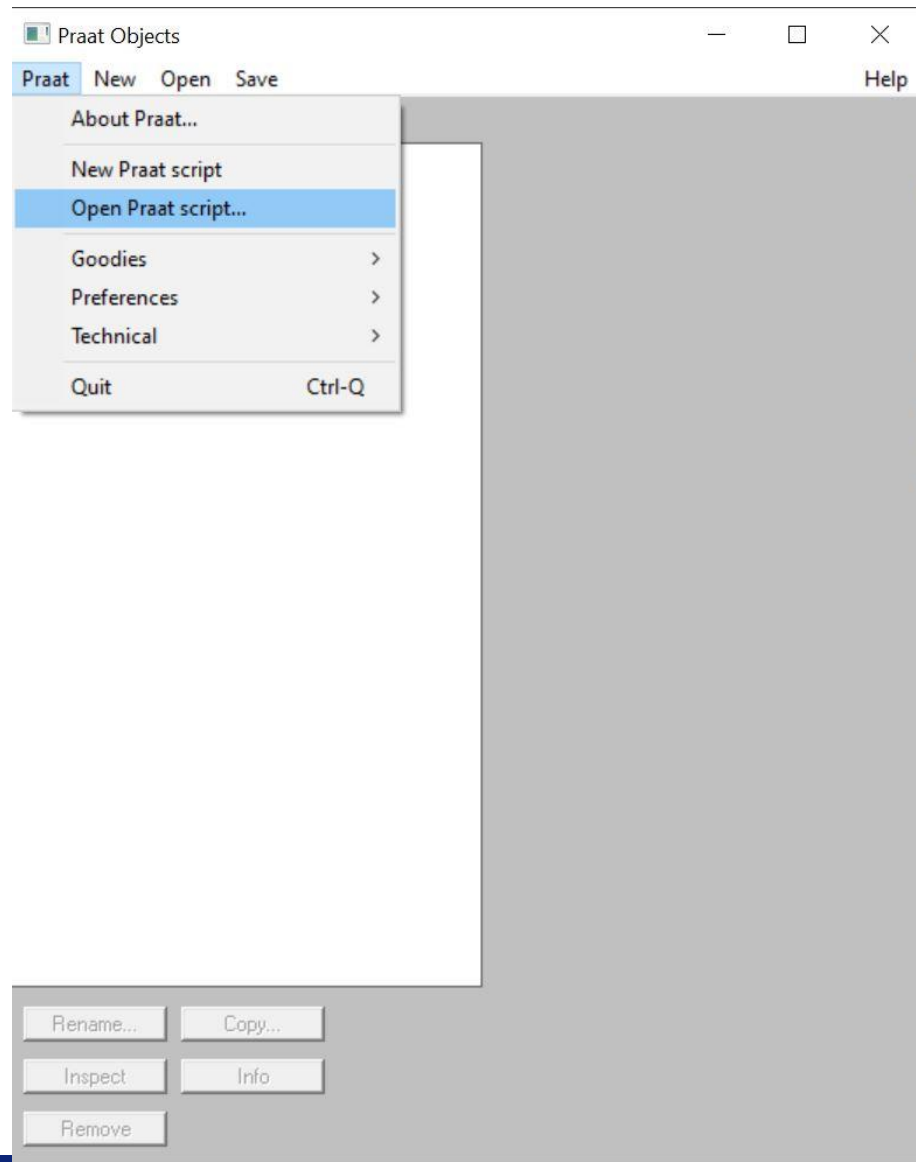
Again for each script separately, choose "Praat - New praat script": in PRAAT (Click on "Praat" top-left, then choose "New Praat script").

Paste the script into the window that opens and save the scripts as "syllablenucleiv3.praat" and "filledpauses.praat", respectively. You now have two PRAAT-scripts in one folder on your computer.

# Step 2a: run PRAAT-scripts

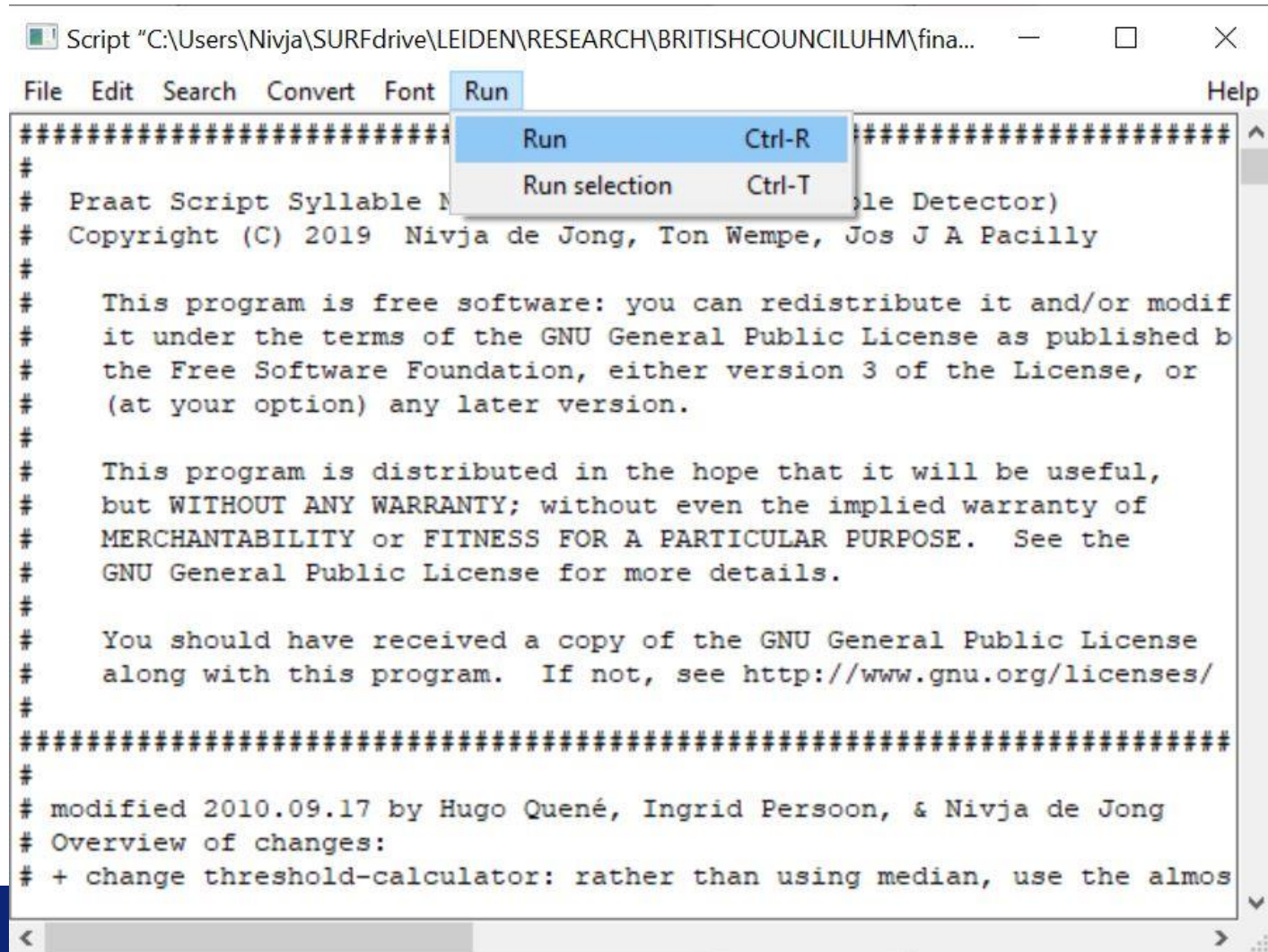
Run PRAAT script Syllable Nuclei v3: in the PRAAT Objects window, click "Praat – Open PRAAT script" and then browse to the directory where you downloaded the file.

[Or you have the scripts still open on your desktop, go to the script syllable nuclei v3]



# Step 2b: run PRAAT-scripts

Run the script by clicking on "Run" or CTRL-R



The screenshot shows a Praat script editor window titled "Script 'C:\Users\Nivja\SURFdrive\LEIDEN\RESEARCH\BRITISHCOUNCILUHM\fin...". The menu bar includes "File", "Edit", "Search", "Convert", "Font", "Run", and "Help". The "Run" menu is open, showing two options: "Run" (with keyboard shortcut "Ctrl-R") and "Run selection" (with keyboard shortcut "Ctrl-T"). The script text in the editor includes a header with the title "Praat Script Syllable N...", copyright information for Nivja de Jong, Ton Wempe, and Jos J A Pacilly, and a GNU General Public License notice. The script also contains a modification note from 2010.09.17 by Hugo Quené, Ingrid Persoon, and Nivja de Jong, mentioning a change to the threshold-calculator.

```
#####  
#  
# Praat Script Syllable N... (Syllable Detector)  
# Copyright (C) 2019 Nivja de Jong, Ton Wempe, Jos J A Pacilly  
#  
# This program is free software: you can redistribute it and/or modify  
# it under the terms of the GNU General Public License as published by  
# the Free Software Foundation, either version 3 of the License, or  
# (at your option) any later version.  
#  
# This program is distributed in the hope that it will be useful,  
# but WITHOUT ANY WARRANTY; without even the implied warranty of  
# MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the  
# GNU General Public License for more details.  
#  
# You should have received a copy of the GNU General Public License  
# along with this program. If not, see http://www.gnu.org/licenses/  
#  
#####  
#  
# modified 2010.09.17 by Hugo Quené, Ingrid Persoon, & Nivja de Jong  
# Overview of changes:  
# + change threshold-calculator: rather than using median, use the almos
```



# Step 3a: enter values

First select the input:  
either a selected sound-  
file in the PRAAT-objects  
window, a list of sound  
files in the PRAAT-objects  
window, or a "filespec",  
e.g.

"C:/Allwavfiles/\*.wav".  
This third option would  
read and analyze all .wav  
files in the specified  
directory.

As INPUT, use either:

- 1: the selected Sound object(s).
- 2: the selected Strings object(s) enumerating the Sound files, or
- 3: the FileSpec below

FileSpec:

---

Parameters Syllable Nuclei:

Pre processing:

Silence threshold (dB):

Minimum dip near peak (dB):

Minimum pause duration (s):

---

Parameters Filled Pauses:

Detect Filled Pauses

Language:

Filled Pause threshold:

---

Destination of OUTPUT:

Data:

DataCollectionType:  OverWriteData  
 AppendData

Keep Objects (when processing files)

# Step 3b: enter values

Optionally change the default settings for syllable nuclei:

pre-process the data, change the silence threshold, change the dip in dB between syllable peaks, and can change the minimum duration of a silent

Run script: Detect Syllables and Filled Pauses in Speech Utterances

As INPUT, use either:

- 1: the selected Sound object(s).
- 2: the selected Strings object(s) enumerating the Sound files, or
- 3: the FileSpec below

FileSpec:

---

Parameters Syllable Nuclei:

Pre processing:

Silence threshold (dB):

Minimum dip near peak (dB):

Minimum pause duration (s):

---

Parameters Filled Pauses:

Detect Filled Pauses

Language:

Filled Pause threshold:

---

Destination of OUTPUT:

Data:

DataCollectionType:  OverWriteData  
 AppendData

Keep Objects (when processing files)

# Step 3c: enter values

Optionally:

check the box "Detect Filled Pauses", set the language (English or Dutch), change the default threshold to detect filled pauses.

Run script: Detect Syllables and Filled Pauses in Speech Utterances

As INPUT, use either:

- 1: the selected Sound object(s).
- 2: the selected Strings object(s) enumerating the Sound files, or
- 3: the FileSpec below

FileSpec:

---

Parameters Syllabe Nuclei:

Pre processing:

Silence threshold (dB):

Minimum dip near peak (dB):

Minimum pause duration (s):

---

Parameters Filled Pauses:

Detect Filled Pauses

Language:

Filled Pause threshold:

---

Destination of OUTPUT:

Data:

DataCollectionType:  OverWriteData  
 AppendData

Keep Objects (when processing files)

# Step 3d: enter values

Choose where and how to save the output:

if you choose to save a .txt-file or a .Table, this will be saved in the folder where you have located the PRAAT-scripts.

Run script: Detect Syllables and Filled Pauses in Speech Utterances

As INPUT, use either:

- 1: the selected Sound object(s).
- 2: the selected Strings object(s) enumerating the Sound files, or
- 3: the FileSpec below

FileSpec:

---

Parameters Syllable Nuclei:

Pre processing:

Silence threshold (dB):

Minimum dip near peak (dB):

Minimum pause duration (s):

---

Parameters Filled Pauses:

Detect Filled Pauses

Language:

Filled Pause threshold:

---

Destination of OUTPUT:

Data:

DataCollectionType:  OverWriteData  
 AppendData

Keep Objects (when processing files)

Standards Cancel Apply OK

## Step 4: WAIT

Wait until the scripts are both done. This may take a while, especially if you have a folder with quite a few .wav or .flac or .MP3-files.

# Step 5a: inspect results

Check the following questions

- a) Are the settings to detect sound and silence ok? (Too much speech identified as silence? -> change "Silence threshold (dB)" -25 for instance to -20)
- b) Are the settings for detecting syllables ok? (Too many syllables that are actually quite long syllables identified as multiple syllables? -> change "Minimum dip near peak (dB)" from 2 to for instance 4)
- c) Is the threshold to detect filled pauses ok? (Too many regular syllables detected as filled pauses? -> change "Filled pause threshold" from 1 to for instance 1.2)

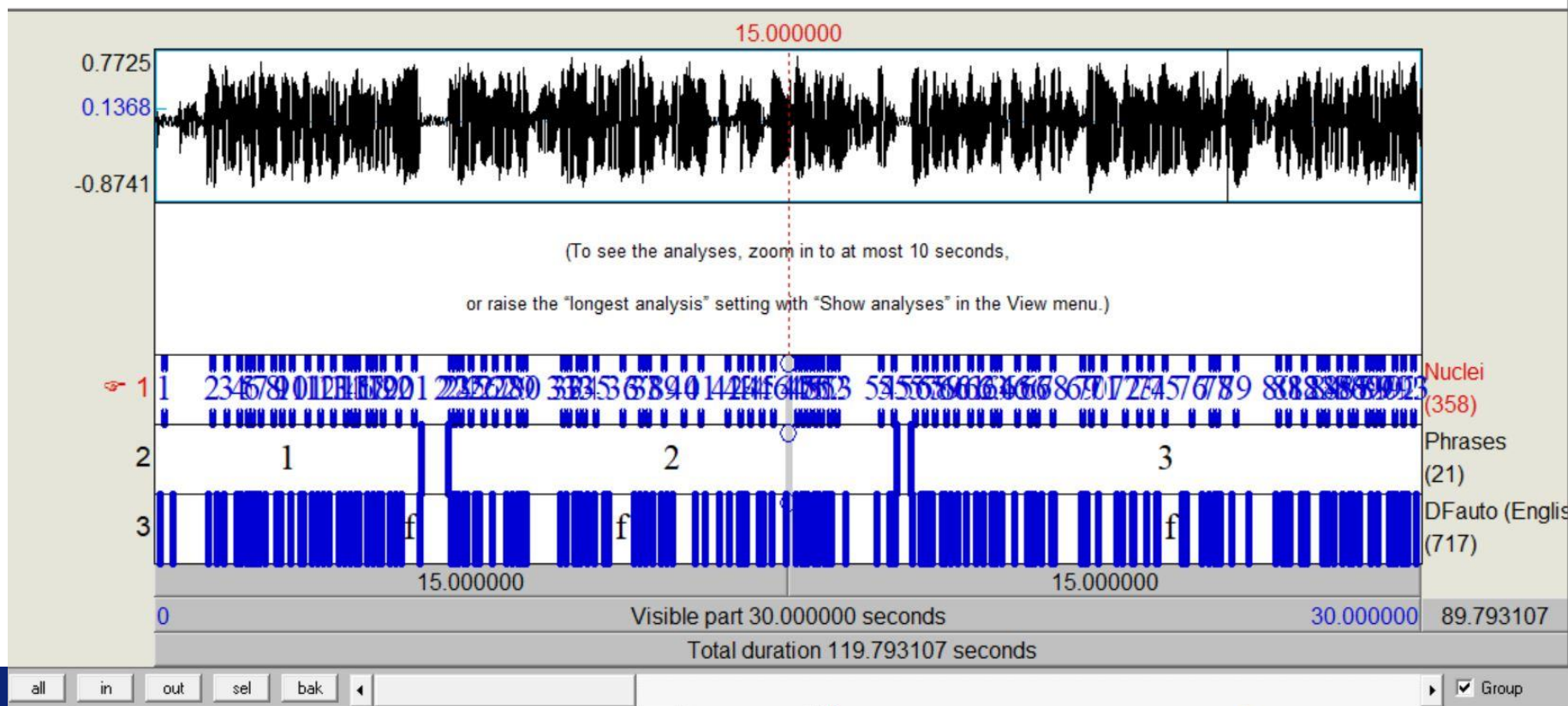
# Step 5b: inspect results

You inspect the results by selecting the Soundfile with its corresponding created TextGrid and then zoom in.

1. TextGrid 29\_4

File Edit Query View Select Interval Boundary Tier Spectrum Pitch Intensity Formant Pulses

Help

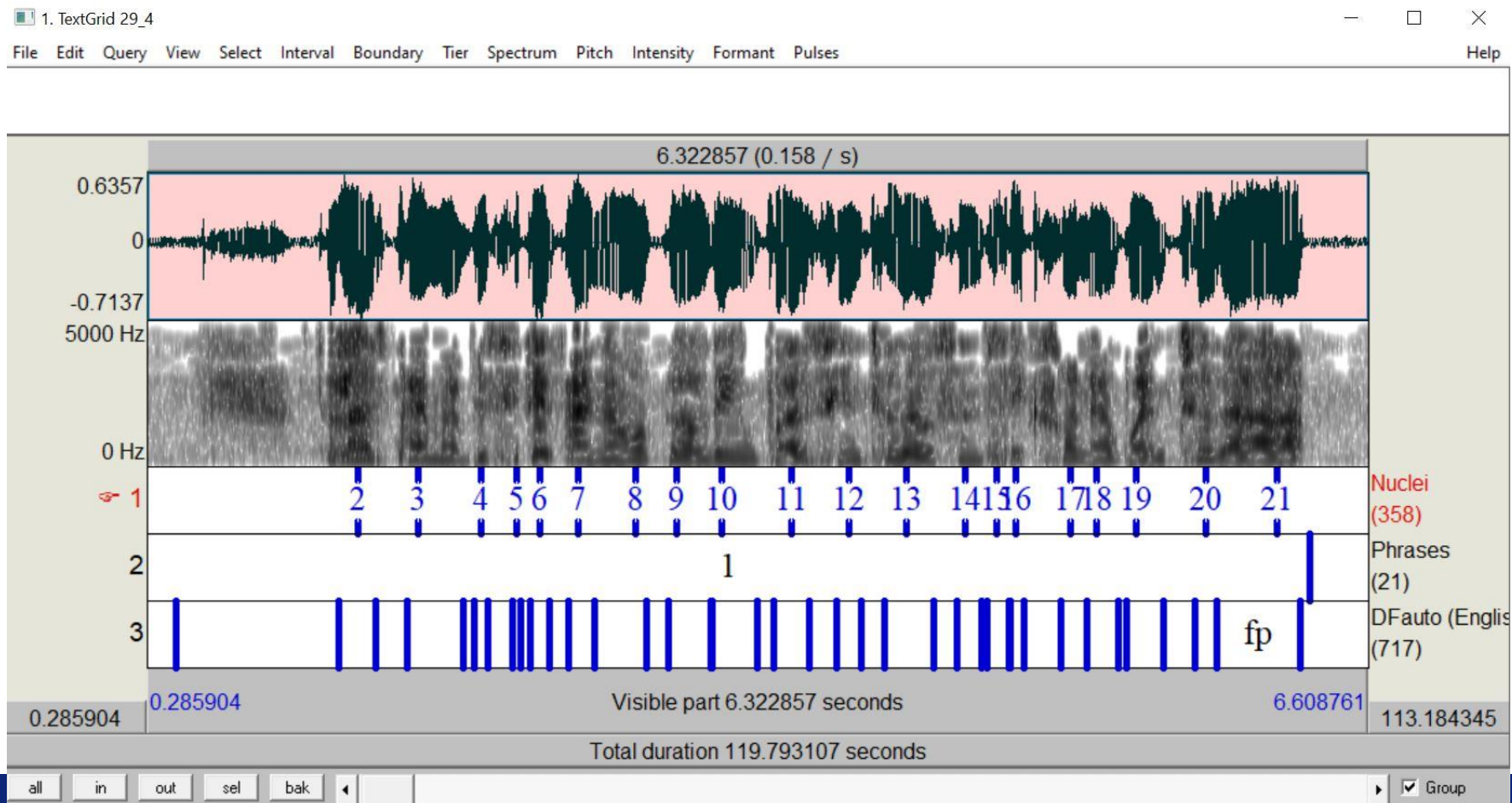






# Step 5b: inspect results

You inspect the results by selecting the Soundfile with its corresponding created TextGrid and then zoom in.



# Tip

Try out the settings for a number of soundfiles and check the results before you run the scripts on an entire folder with soundfiles.

For valid comparison across soundfiles, use the same settings on similar (e.g. with respect to quality, task type, ...) types of soundfiles.

# THANK YOU!

## Questions?

### **Research in collaboration with:**

Jos Pacilly  
Willemijn Heeren  
Danique van Aalst  
Katarina Stankovic

### **Research funded by:**

British Council Assessment Research  
Awards and Grants programme 2018

Netherlands Organization for Scientific  
Research (NWO VIDI grant 276-75-010)  
for secondary corpora

### **Questions:**

[n.h.de.jong@hum.leidenuniv.nl](mailto:n.h.de.jong@hum.leidenuniv.nl)



Universiteit  
Leiden



# ICLON

## References

- Audhkhasi, Kandhway, K., Deshmukh, O. D., & Verma, A. (2009). Formant-based technique for automatic filled pause detection in spontaneous spoken English, ICASSP, pp.4857-4860, *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009.
- Boersma, P., & Weenink, D. (2016). PRAAT: Doing phonetics by computer [Computer program]. Version 6.0.23. Retrieved from <http://www.PRAAT.org/>.
- Bosker, H. R., Pinget, A. F., Quené, H., Sanders, T., & De Jong, N. H. (2013). What makes speech sound fluent? The contributions of pauses, speed and repairs. *Language Testing*, 30(2), 159-175.
- Bosker, H.R., Quené, H., Sanders, T. and de Jong, N.H. (2014), The Perception of Fluency in Native and Nonnative Speech. *Language Learning*, 64, 579-614. doi:10.1111/lang.12067
- Clark, H. H., & Fox Tree, J. E. (2002). Using uh and um in spontaneous speaking. *Cognition*, 84(1), 73-111.
- De Jong, N. H., & Wempe, T. (2009). PRAAT script to detect syllable nuclei and measure speech rate automatically. *Behavior research methods*, 41(2), 385-390.
- Fawcett, T. (2006). An introduction to ROC analysis. *Pattern recognition letters*, 27(8), 861-874.
- Hughes, V., Foulkes, P., & Wood, S. (2016). Formant dynamics and durations of um improve the performance of automatic speaker recognition systems. In *Proceedings of the 16th Australasian Conference on Speech Science and Technology (ASSTA)*, 25 – 28.
- Kaushik, M., Trinkle, M., Hashemi-Sakhtsari, A. (2010). Automatic detection and removal of disfluencies from spontaneous speech. *Proc. 13th Australasian Int. Conf. on Speech Science and Technology Melbourne*, 98-101.
- Krikke, T. F., & Truong, K. P. (2013). Detection of nonverbal vocalizations using gaussian mixture models: Looking for fillers and laughter in conversational speech. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 163–167.
- Ladefoged, P., & Johnson, K. (2011). *A course in phonetics*. Boston, MA: Wadsworth.
- Orr, R., & Quené, H. (2017). D-LUCEA: Curation of the UCU Accent Project data. In *CLARIN in the Low Countries*, edited by J. Odiijk and A. van Hessen (Ubiquity Press, Berkeley), pp. 177–190.
- Reetz, H., & Jongman, A. (2009). *Phonetics: Transcription, production, acoustics, and perception*. Chichester: Wiley-Blackwell.
- Shriberg, E. (2001). To ‘errrr’ is human: Ecology and acoustics of speech disfluencies. *Journal of the International Phonetic Association*, 31(1), 153–169.
- Shriberg, E. E., & Lickley, R. J. (1993). Intonation of clause-internal filled pauses. *Phonetica*, 50(3), 172–179.
- Stouten, F., & Martens, J. P. (2003). A feature-based filled pause detection system for Dutch. In *2003 IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 309-314. IEEE.
- Tavakoli, P., Nakatsuhara, F., & Hunter, A. M. (2017). Scoring validity of the Aptis Speaking Test: Investigating fluency across tasks and levels of proficiency. *ARAGs Research Reports Online*.
- Vasilescu I., & Adda-Decker M. (2007). A cross-language study of acoustic and prosodic characteristics of vocalic hesitations. In Esposito A., Bratanić M., Keller E., Marinaro M. (Eds.), *Fundamentals of Verbal and Nonverbal Communication and the Biometric Issue* (pp. 140–148). IOS Press.
- Verkhodanova, V. & Shapranov, V. (2016). Experiments on Detection of Voiced Hesitations in Russian Spontaneous Speech. *Journal of Electrical and Computer Engineering*, vol. 2016, Article ID 2013658, 8 pages.