A set of tools for analysis of speech fundamental frequency

Jolanta Bachan

Adam Mickiewicz University, Poznań

LTC 2019, Poznań, Poland

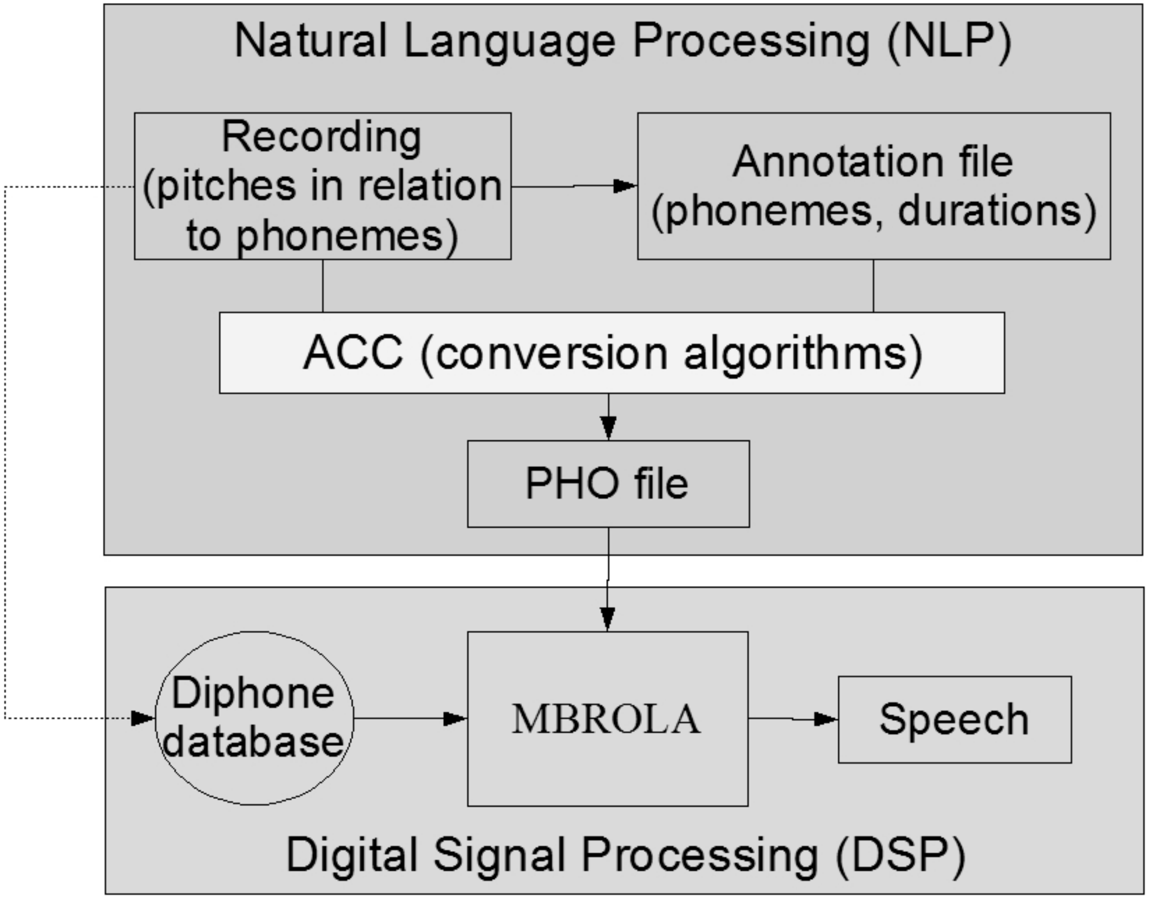
Outline

- Automatic Close Copy Speech (ACCS) synthesis
- F0 manipulation and speech resynthesis
- F0 extraction
- Extraction of prosodic information

Automatic Close Copy Speech (ACCS) synthesis

“...repeats an utterance produced by a human speaker with a synthetic voice, while keeping the original prosody” (Dutoit, 1996)

Automatic Close Copy Speech (ACCS) synthesis


Natural Language Processing (NLP)
Recording (pitches in relation to phonemes)
Annotation file (phonemes, durations)
ACC (conversion algorithms)
PHO file
Diphone database
MBROLA
Speech
Digital Signal Processing (DSP)

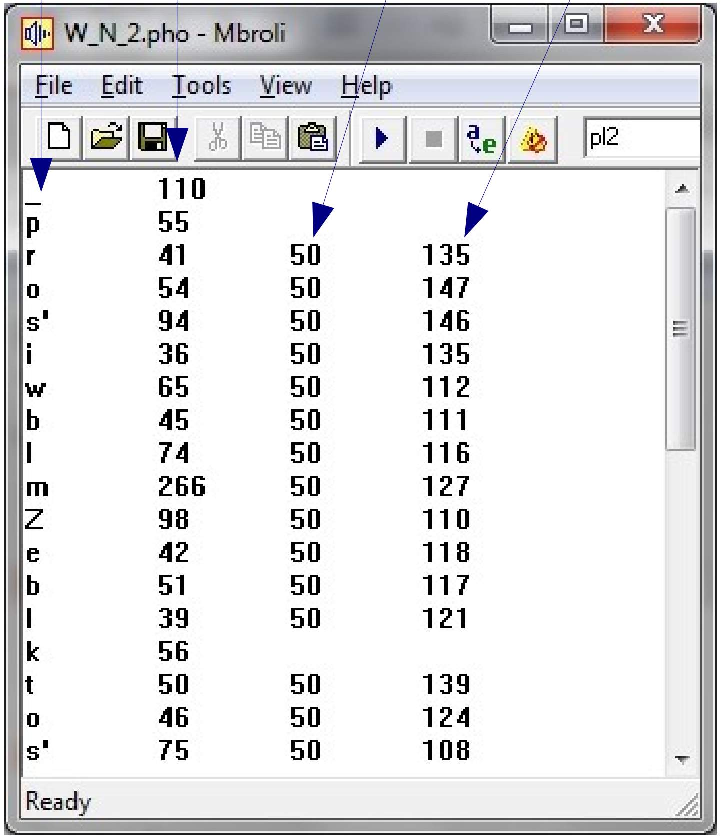
Automatic Close Copy Speech (ACCS) synthesis

- *TextGrid2pho.praat* script
- The script goes through sound and TextGrid files in a directory and creates files in the correct format for the MBROLA speech synthesiser
- PHO file format

labels

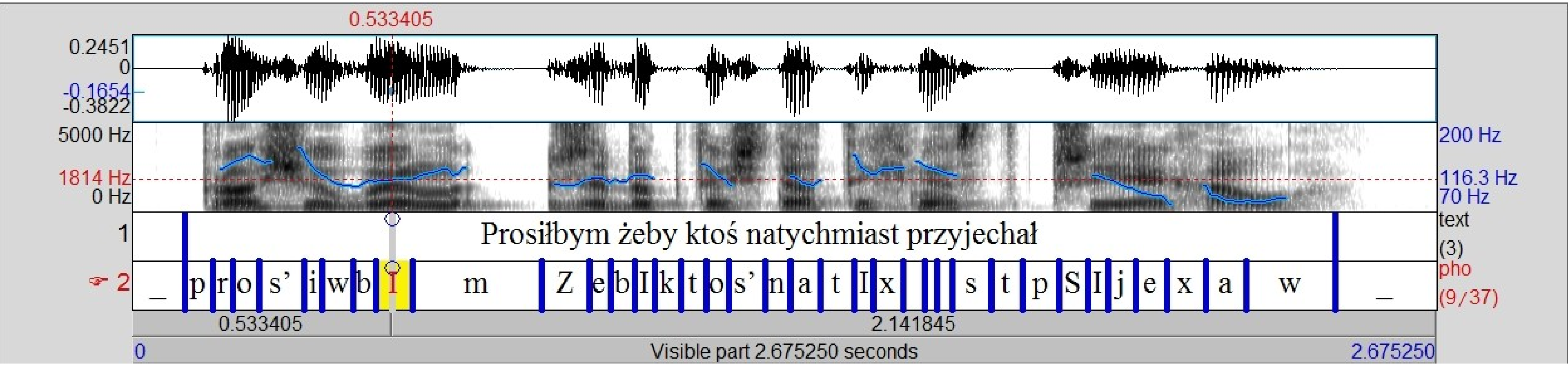
durations

F0 pairs (place + value in Hz)

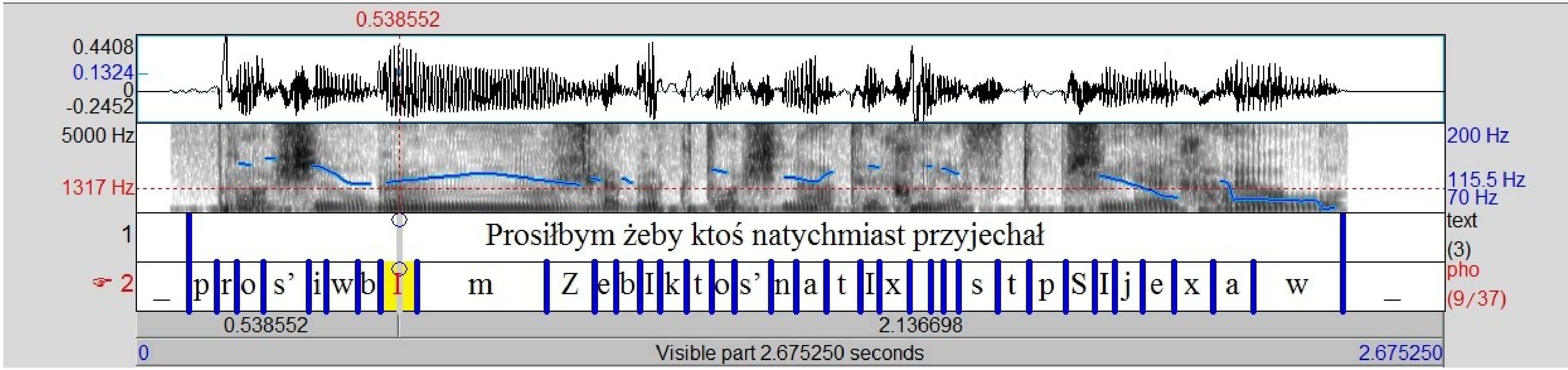


```
_       110
p       55
r       41      50      135
o       54      50      147
s'      94      50      146
i       36      50      135
w       65      50      112
b       45      50      111
l       74      50      116
m       266     50      127
Z       98      50      110
e       42      50      118
b       51      50      117
l       39      50      121
k       56
t       50      50      139
o       46      50      124
s'      75      50      108
```

Automatic Close Copy Speech (ACCS) synthesis



Prosiłbym żeby ktoś natychmiast przyjechał



Prosiłbym żeby ktoś natychmiast przyjechał

F0 manipulation and speech resynthesis

The present study was supported by the Polish National Science Centre, project no.: 2013/09/N/HS2/02358, *"Vocal schemes of verbal emotion communication in linguistic perspective"*.
Principal Investigator: Magdalena Oleśkowicz-Popiel

F0 manipulation (1st step)

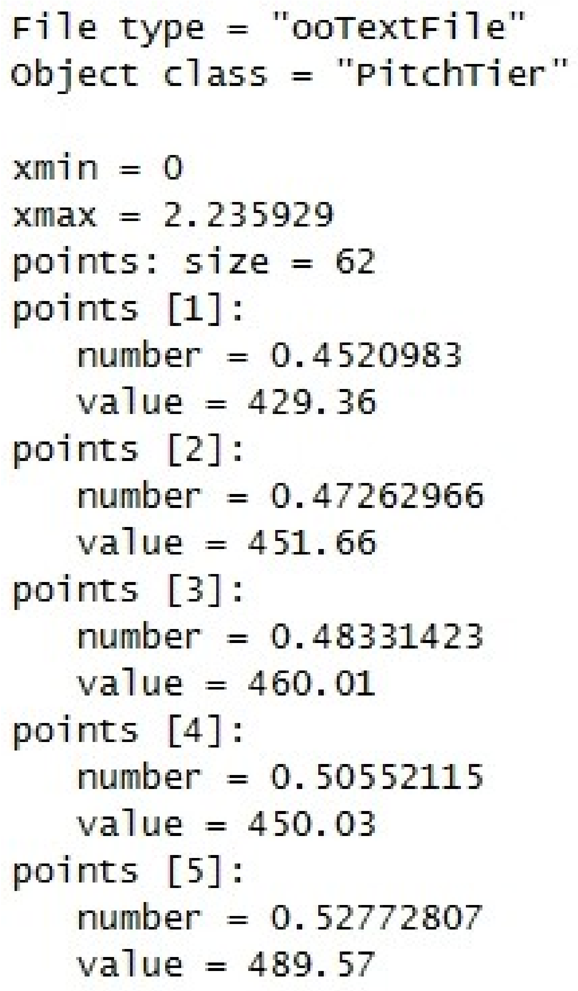
- F0 extraction from natural, emotional and smiling recordings using a Praat script
  - information extraction from TextGrid files about phone labels on “pho” tier and the phone durations
  - each phone duration is divided into 3 intervals from 0%-20%, 20%-80% and 80%-100% of phone duration and the mean pitch value is extracted for each of the intervals from a corresponding WAV file
  - data from this step are saved in text files with “.F0” extension for each of the file in a directory

F0 manipulation (1st step)

Phone	Duration	Start time	Time 20%	Time 80%	End Time	Mean F0 0%-20%	Mean F0 20%-80%	Mean F0 80%-100%
s	0.0892	0.3431	0.3609	0.4145	0.4323	undefined	undefined	undefined
k	0.0554	0.4323	0.4434	0.4767	0.4878	undefined	429.36	451.66
o	0.0645	0.4878	0.5007	0.5394	0.5523	460.01	450.03	489.57
Z	0.0732	0.5523	0.5670	0.6109	0.6256	436.69	373.13	429.97
y	0.0367	0.6256	0.6329	0.6550	0.6623	466.49	493.85	480.03
s	0.0700	0.6623	0.6763	0.7183	0.7323	459.49	480.92	undefined
t	0.0519	0.7323	0.7427	0.7739	0.7842	undefined	418.60	416.97
a	0.0580	0.7842	0.7959	0.8307	0.8423	381.68	368.52	407.55
w	0.0413	0.8423	0.8506	0.8754	0.8836	473.97	441.40	425.48

F0 manipulation (2nd step)

- The duration is taken from a “neutral” file and the F0 values are extracted from the “emotional” recording



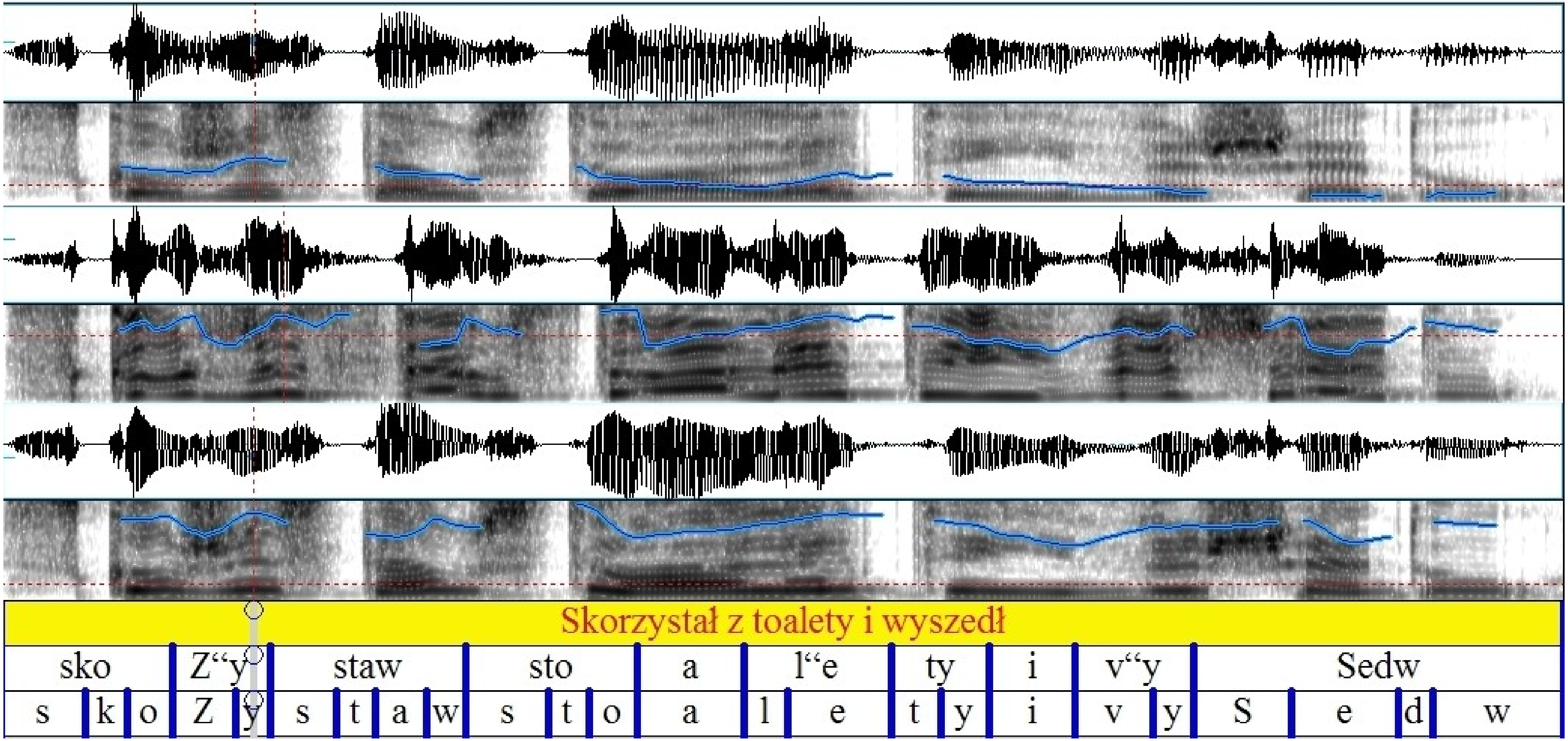
```
File type = "ooTextFile"
Object class = "PitchTier"

xmin = 0
xmax = 2.235929
points: size = 62
points [1]:
    number = 0.4520983
    value = 429.36
points [2]:
    number = 0.47262966
    value = 451.66
points [3]:
    number = 0.48331423
    value = 460.01
points [4]:
    number = 0.50552115
    value = 450.03
points [5]:
    number = 0.52772807
    value = 489.57
```

F0 manipulation (3rd step)

- The final step is to replace the neutral pitch tier with the newly created pitch tier with the emotional F0 values and re-synthesise the neutral recording using the overlap-add synthesis in another Praat script.
  - Figure: On top “neutral” recoding, in the middle “emotional” recording and at bottom the resynthesised “neutral” recoding using overlap-add synthesis with the imposed “emotional” contour and anotation for the “neutral” recording.

F0 manipulation (3rd step)



neutral + emotional→ neutral durations + emotional F0

F0 extraction

The present study was supported by the Polish National Science Centre, international project Harmonia: no.: 2014/14/M/HS2/00631, *“Automatic analysis of phonetic convergence in speech technology systems.”*
Principal Investigator: Grażyna Demenko
Partners of the project:
Saarland University, Department of Computational Linguistics and Phonetics

F0 extraction

The script was dedicated for dialogue recordings (around 5 minutes) to study the phonetic convergence between speakers in different stages of conversation:

- initial (I, 0-25% of time duration)
- initial-medial (IM, 25-50%)
- medial-final (MF, 50-75%)
- final (F, 75-100%)

and extracts the F0 for these intervals

F0 extraction

- Males: 60-300 Hz
- Females: 110-500 Hz

```
To Pitch... 0.001 60 300
tmin = startTime + (step - 1) * 0.01
tmax = tmin + 0.01
mean = Get mean: tmin, tmax, "Hertz"
minimum = Get minimum: tmin, tmax, "Hertz", "Parabolic"
maximum = Get maximum: tmin, tmax, "Hertz", "Parabolic"
stdev = Get standard deviation: tmin, tmax, "Hertz"
```

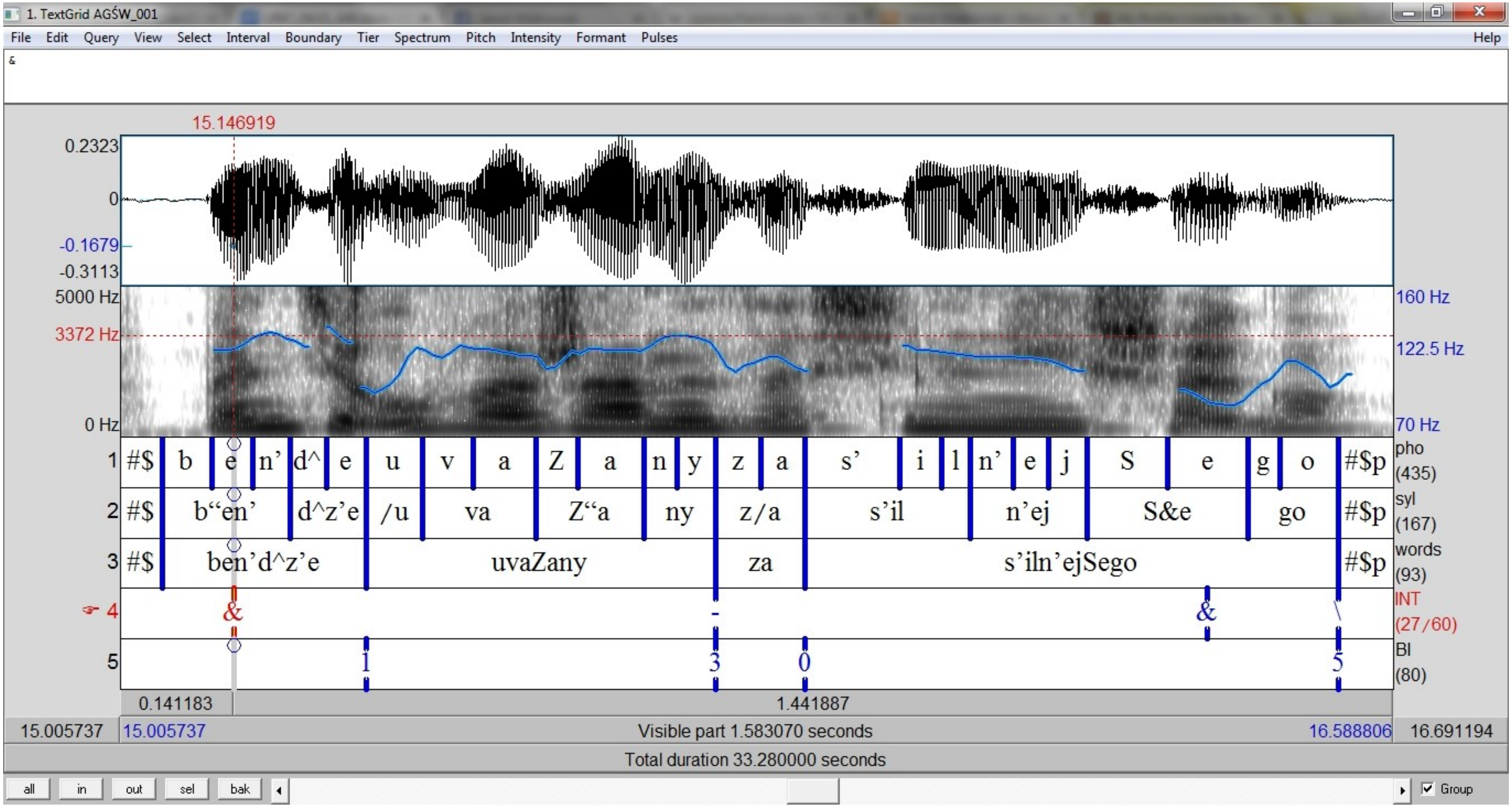
F0 extraction

filename	F0 mean	F0 median	F0 most common	F0 most common count	len F0 mean list	F0 max	F0 min	F0 std	len F0 values
N1_0_25	296	288	294	22	10	495	111	65	1330
N1_25_50	244	237	232	52	28	489	111	61	3080
N1_50_75	262	252	237	46	22	492	110	68	2350
N1_75_100	254	244	223	36	22	499	110	76	2457

Extraction of prosodic information

This research has been carried out in the scope of the project *"Rhythmic structure of utterances in the Polish language: A corpus analysis"* supported by National Science Centre (NCN) grant no. 2013/11/D/HS2/04486. Principal Investigator: Agnieszka Wagner

Extraction of prosodic information



Extraction of prosodic information: input data

- pho - phonemes (IntervalTier) - labels extracted from annotations in Annotation Pro
- syl - syllables (IntervalTier) - labels extracted from annotations in Annotation Pro
- words (IntervalTier) - labels extracted from annotations in Annotation Pro
- INT - intonational features and prominence (PointTier) - only start end end times of this tier could be ***generated automatically***
- BI (from Break Index) – prosodic structure (PointTier) - the time of the break points is equal to the end of the word tier, so for each word on the word tier, a break point was ***created automatically*** on the BI tier.

Extraction of prosodic information: output data

- start & end times and durations of the events (syllables)
- pauses in the preceding and following context (and their durations)
- prosodic information from the INT and BI tiers concerning the degree of stress, break index indicating the prosodic constituency, pitch accent or prominence
- positional/structural features
  - syllable position in clitic group (initial, medial or final)
  - phonological phrase position in the intonational phrase
  - prosodic constituents length, e.g., intonational phrase length as number of phonological phrases and clitic groups.

Extraction of prosodic information: output data

- The output can be:
  - directly used as an input to
    - rhythm analysis
    - general prosody analysis
  - can be further processed to obtain additional information that is required for other studies

Summary

- The tools take empirical models directly from authentic utterances, rather than filtering them through abstract models or human manipulation.
- The tools are easy to use and have saved a lot of time and effort in procedures which in many previous studies have been performed manually.

Thank you!