Synteza mowy (TTS)
Rozpoznawanie mowy (ASR)
Optyczne rozpoznawanie znaków (OCR)

Jolanta Bachan

Synteza mowy

- System przetwarzania tekstu pisanego na mowę
- Text-to-Speech (TTS)
- TTS powinien być w stanie przeczytać każdy tekst, ale w praktyce nie jest to takie proste do zrealizowania

Synteza mowy

- Parametryczna synteza mowy
  - synteza formantowa
  - synteza artykulacyjna (*VTdemo*, vocal tract demo)
- Konkatenacyjna synteza mowy
  - synteza difonowa
  - synteza trifonowa
  - unit selection
- Synteza mowy oparta o HMM (Hidden Markov Models)
  - ang. HMM-based speech synthesis
    - Statistical Parametric Synthesis
- End-to-End (e2e) speech synthesis based on deep learning

Konkatenacyjna synteza mowy

- Konkatenacyjna synteza mowy łączy mniejsze jednostki nagranej mowy (difony, trifony, sylaby, wyrazy) w większą całość (wyrazy, zdania).
- System jest oparty na bazie nagrań mowy. Baza posegmentowana jest na mniejsze elementy (głoski, difony, wyrazy), z których później "skleja się" wypowiedzi.

Co to jest Close Copy Speech Synthesis?

Close Copy Speech Synthesis

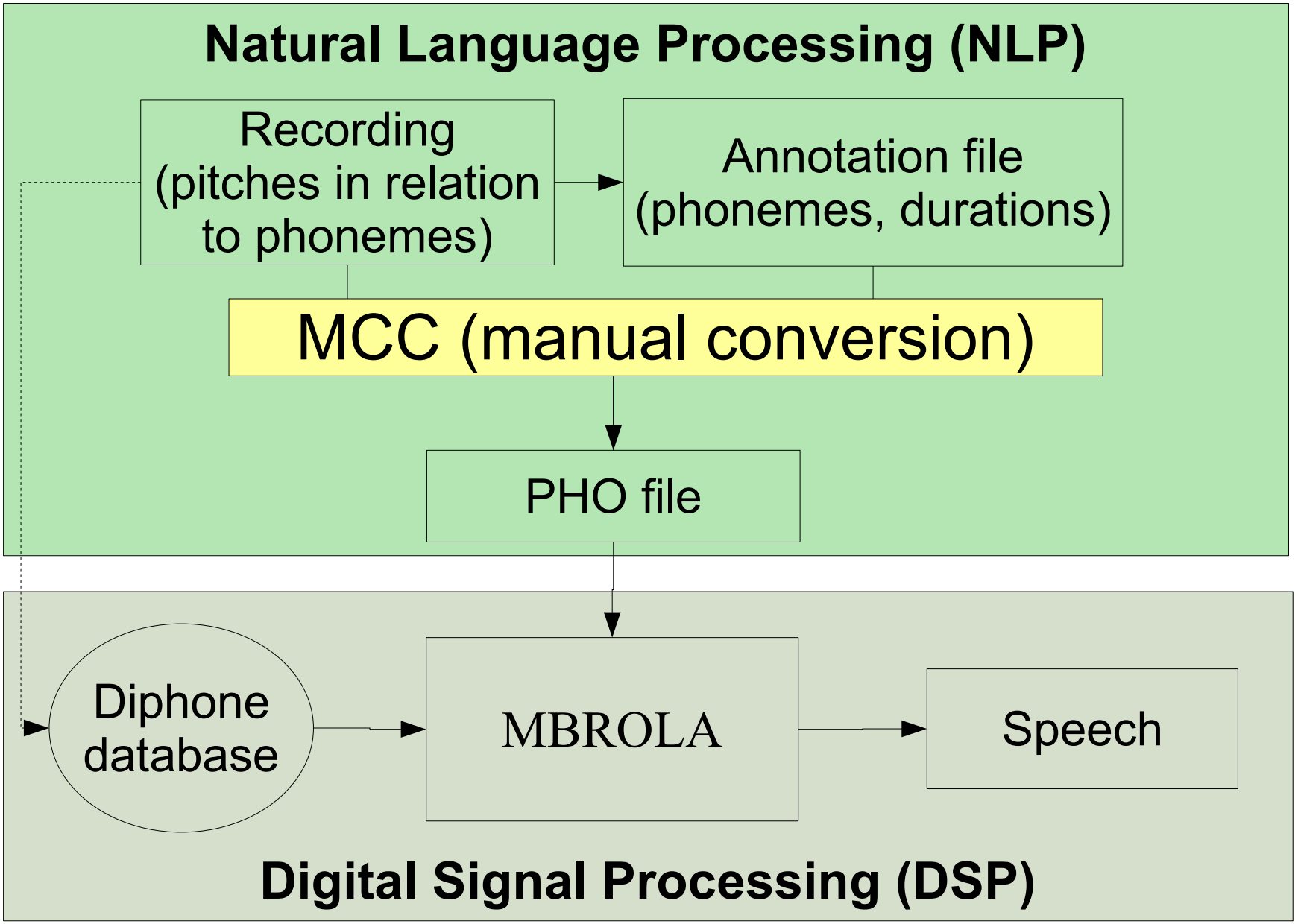
The CCS synthesis system produces a sound which “repeats an utterance produced by a human speaker with a synthetic voice, while keeping the original prosody” (Dutoit, 1996).

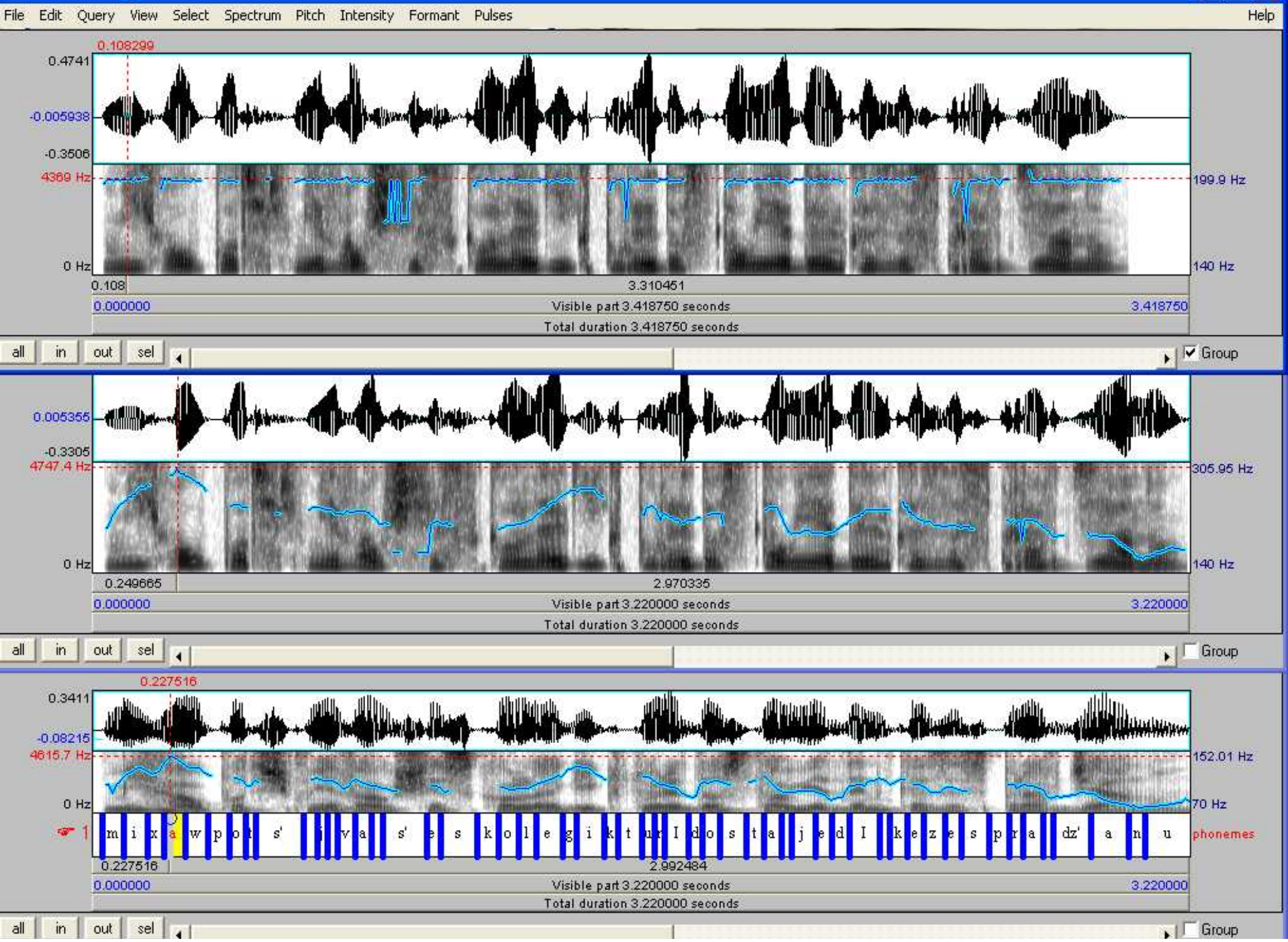
Manual Close Copy Speech (MCCS) Resynthesis

Komponenty MCCS

- Wejście: Mowa
  - nagrania mowy
  - anotacja nagrań mowy
- Syntezator mowy (tu: MBROLA)
  - baza difonów (tzw. głos MBROLI)
  - silnik syntezatora

MCCS synthesis




File Edit Query View Select Spectrum Pitch Intensity Formant Pulses Help
0.108299
0.4741
-0.005938
-0.3506
4369 Hz
0 Hz
199.9 Hz
140 Hz
0.108
3.310451
0.000000
Visible part 3.418750 seconds
3.418750
Total duration 3.418750 seconds
all in out sel
Group
0.005355
-0.3305
4747.4 Hz
305.95 Hz
0 Hz
140 Hz
0.249665
2.970335
0.000000
Visible part 3.220000 seconds
3.220000
Total duration 3.220000 seconds
all in out sel
Group
0.227516
0.3411
-0.08215
4615.7 Hz
152.01 Hz
0 Hz
70 Hz
1
m i x a w p o t s' j v a s' e s k o l e g i k t u r I d o s t a j e d I k e z e s p r a dz' a n u
phonemes
0.227516
2.992484
0.000000
Visible part 3.220000 seconds
3.220000
Total duration 3.220000 seconds
all in out sel
Group

Synteza MCCS

- Monotonny

Monotone

- Synteza MCCS

MCCS

- Oryginał

Original recording

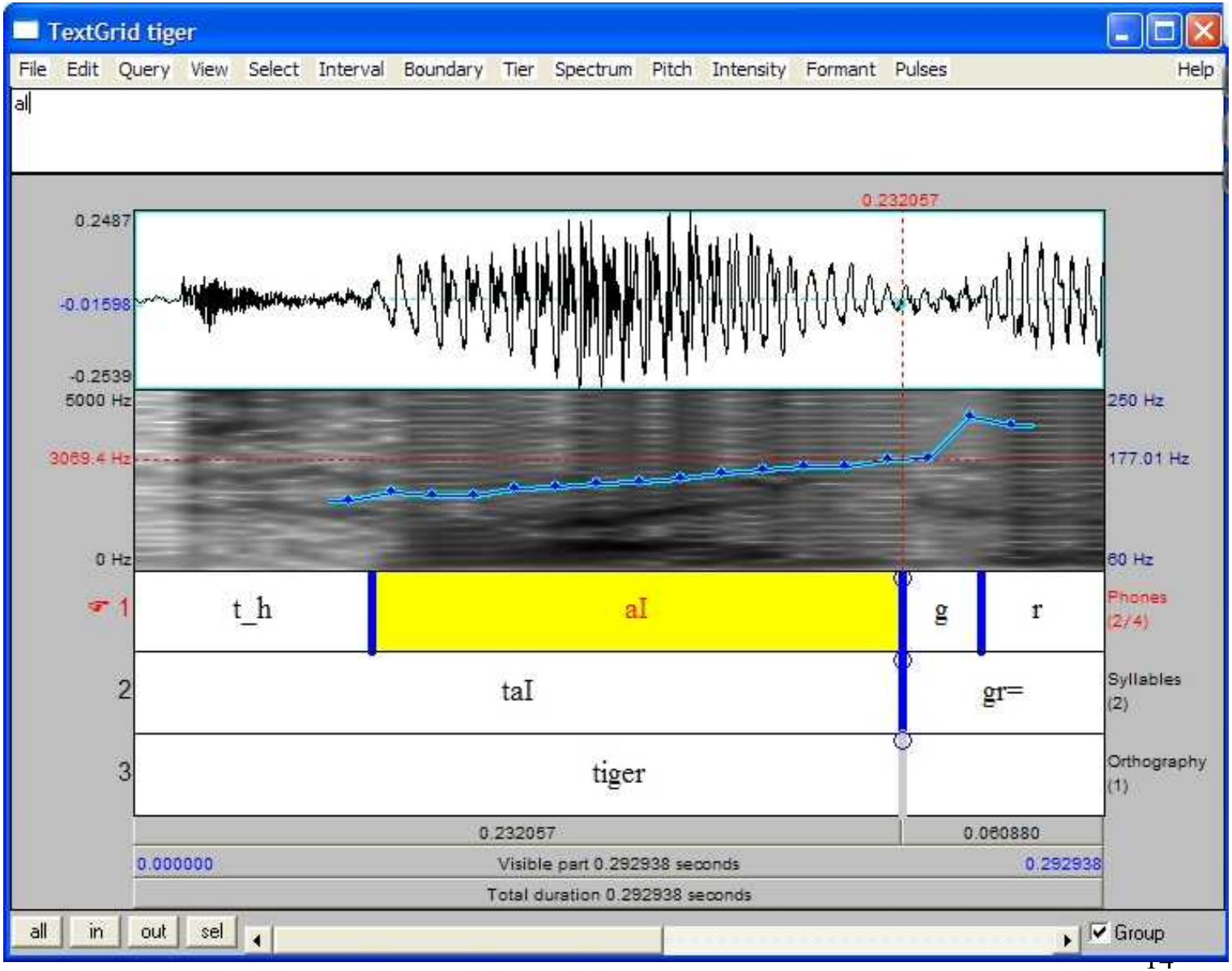
Zastowowania MCCS

- Tworzenie bodźców na potrzeby testów percepcyjnych mowy
- Narzędzie w nauczaniu fonetyki
- Eksperymentowanie z prozodią
- Sprawdzanie anotacji

Anotacja mowy w Praacie

Anotacja dla syntezy CCS

REQUIRED
FOR
SPEECH
RE-SYNTHESIS



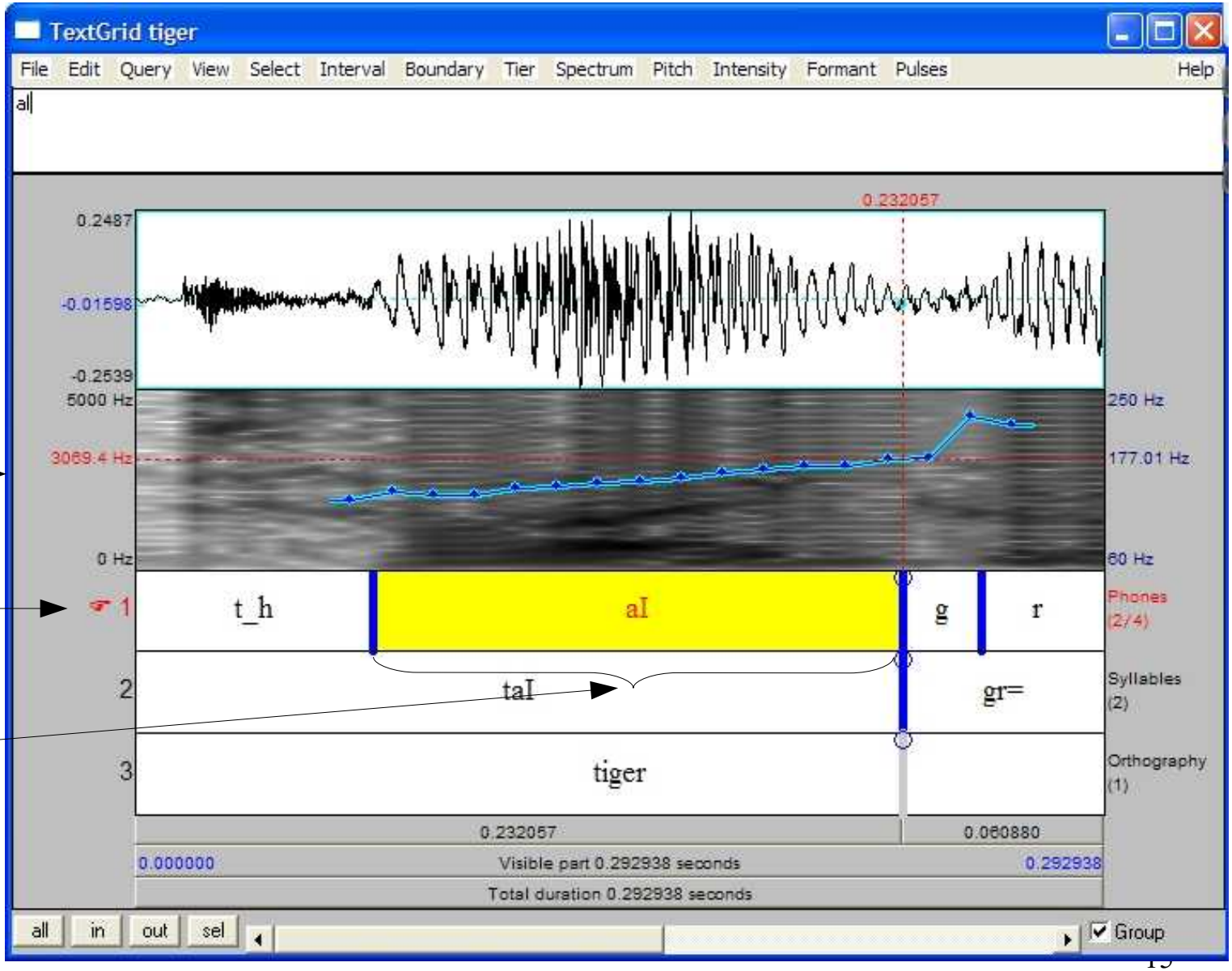
Anotacja dla syntezy CCS

REQUIRED FOR SPEECH RE-SYNTHESIS

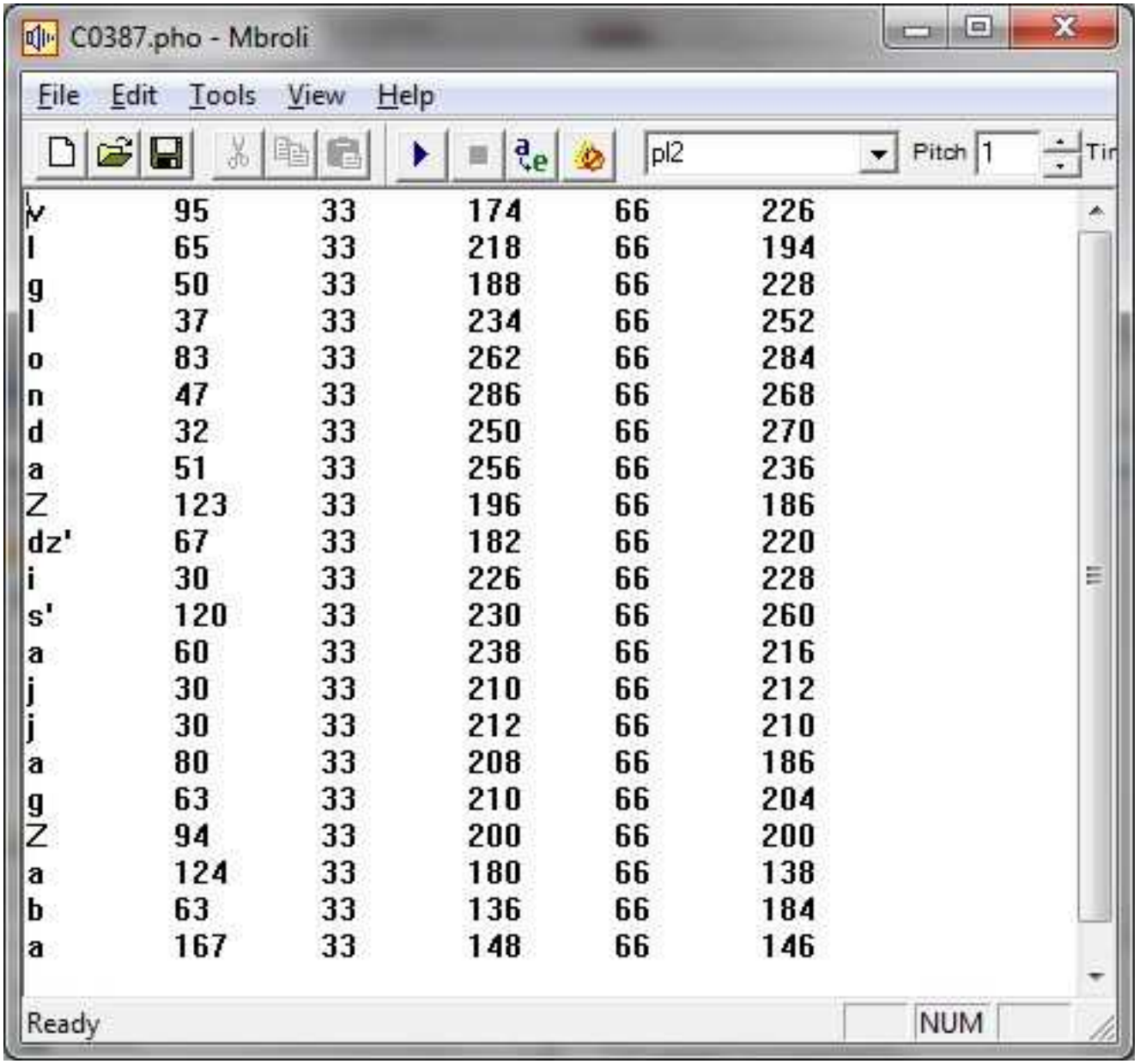
Pitch

Phones

Durations



Format pliku PHO



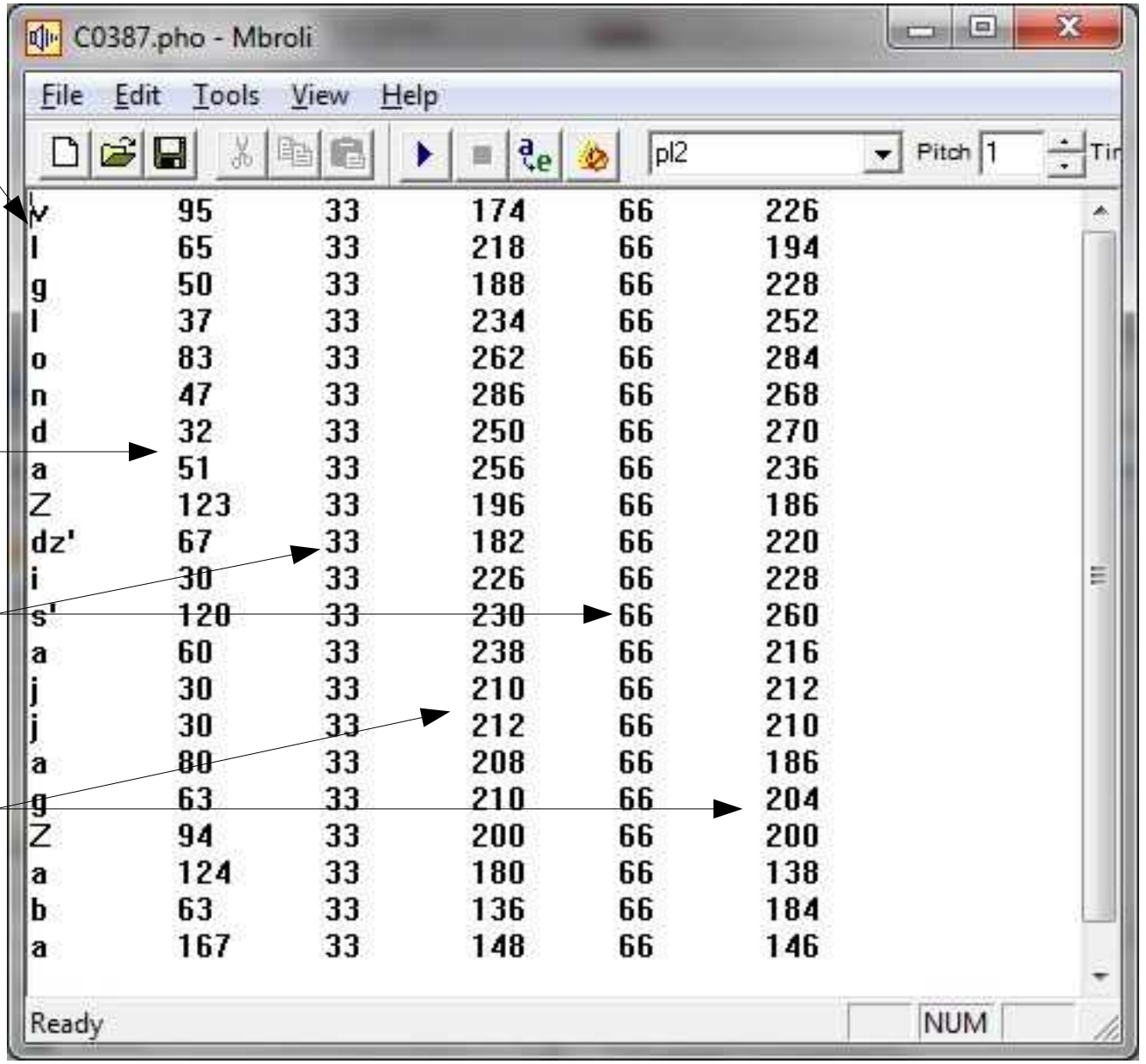
Format pliku PHO

Głoski: SAMPA

Iloczas

Pozycja F0

Wartość F0



v	95	33	174	66	226
l	65	33	218	66	194
g	50	33	188	66	228
l	37	33	234	66	252
o	83	33	262	66	284
n	47	33	286	66	268
d	32	33	250	66	270
a	51	33	256	66	236
Z	123	33	196	66	186
dz'	67	33	182	66	220
i	30	33	226	66	228
s'	120	33	230	66	260
a	60	33	238	66	216
j	30	33	210	66	212
j	30	33	212	66	210
a	80	33	208	66	186
g	63	33	210	66	204
Z	94	33	200	66	200
a	124	33	180	66	138
b	63	33	136	66	184
a	167	33	148	66	146

Start z MBROLĄ

- Zainstaluj syntezator MBROLA
- Przez „Control Panel” w Mbrola Tools dodaj bazy difonowe pl1 i en1.
- Wykonaj syntezę MCCS w Mbroli.exe
  - Przygotuj plik PHO dla poanotowanego pliku dla języka polskiego lub angielskiego
- Stwórz własny plik PHO

Ewaluacja syntezy mowy

- zrozumiałość
- naturalność
- skala MOS – *Mean Opinion Score*, od 1 do 5

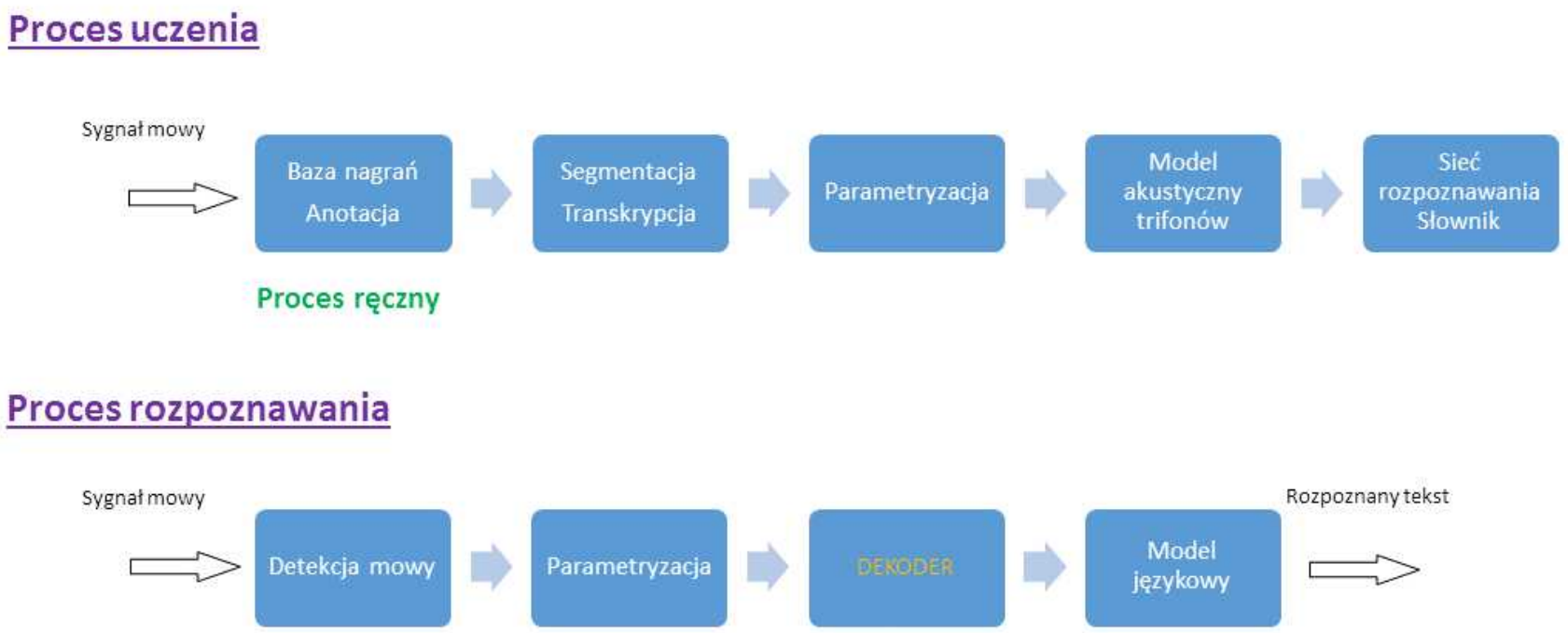
Automatyczne Rozpoznawanie Mowy (ARM)

- Automatic Speech Recognition (ASR)
- Konwersja sygnału mowy na tekst

Poznański System Rozpoznawania Mowy Polskiej ARM

- https://speechlabs.pl/oferta/arm/
- Projekt *„Zaawansowany system automatycznego rozpoznawania i przetwarzania mowy polskiej na tekst, dedykowany dla służb odpowiedzialnych za bezpieczeństwo państwa”.*
- Dwa główne procesy:
  - proces uczenia się – proces ten rozpoczyna się od zebrania odpowiedniej bazy nagrań, która następnie zostaje poddana procesowi anotacji (segmentacja i transkrypcja mowy). Parametryzacja tak przygotowanej bazy pozwala na zbudowanie modelu akustycznego trifonów, który staje się podstawą Dekodera w procesie rozpoznawania mowy. Obok modelu akustycznego budowany jest Słownik (lista słów, obecnie ok. 450.000) z teksów pisanych języka polskiego (transkrypcja nagrań, artykuły z gazet, teksty prawnicze takie jak wyroki, akty prawne, umowy)
  - proces rozpoznawania: jądrem procesu rozpoznawania mowy jest Dekoder, który zamienia parametry akustyczne wyekstrahowane z sygnału mowy na tekst (wyrazy znajdujące się w Słowniku systemu ARM), a następnie model językowy koryguje błędy Dekodera (na podstawie n-gramów) i tak wygenerowany i skorygowany tekst jest prezentowany użytkownikowi

Schemat budowy systemu rozpoznawania mowy polskiej ARM



Demenko, G., Cecko, R. Szymański, M., Owsianny, M., Francuzik, P., Lange, M. (2012). Polish speech dictation system as an application of voice interfaces. W: Dziech, A., Czyżewski, A. (Red.) Proceedings of 5th International Conference on Multimedia Communications, Services and Security, Kraków 2012 (pp. 68–76). Springer for Research and Development.

OCR
Optical Character Recognition

- Pre-processing
  - usuwanie szumu i niwelacja zniekształceń obrazu
  - redukcja kolorów do czerni i bieli
- Dzielnie znaków
  - wykorzystanie algorytmów heurystycznych
- Rozpoznawanie znaków:
  - porównywanie wektorowe (pattern matching)
  - porównywanie rastrowe (pixel by pixel)
  - porównywanie słownikowe (near-neighbor analysis)
- Post-processing
  - formatowanie i układ tekstu (źródło: slajdy M. Koziarskiego)

Zadanie domowe

- Przeczytaj o reCAPTCHA
  - https://pl.wikipedia.org/wiki/ReCAPTCHA
- Przetestuj dowolny system OCR
  - np. FreeOCR
  - https://www.dobreprogramy.pl/FreeOCR.net,Program,Windows,12517.html

Do zobaczenia za tydzień
na teście zaliczeniowym!