

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).

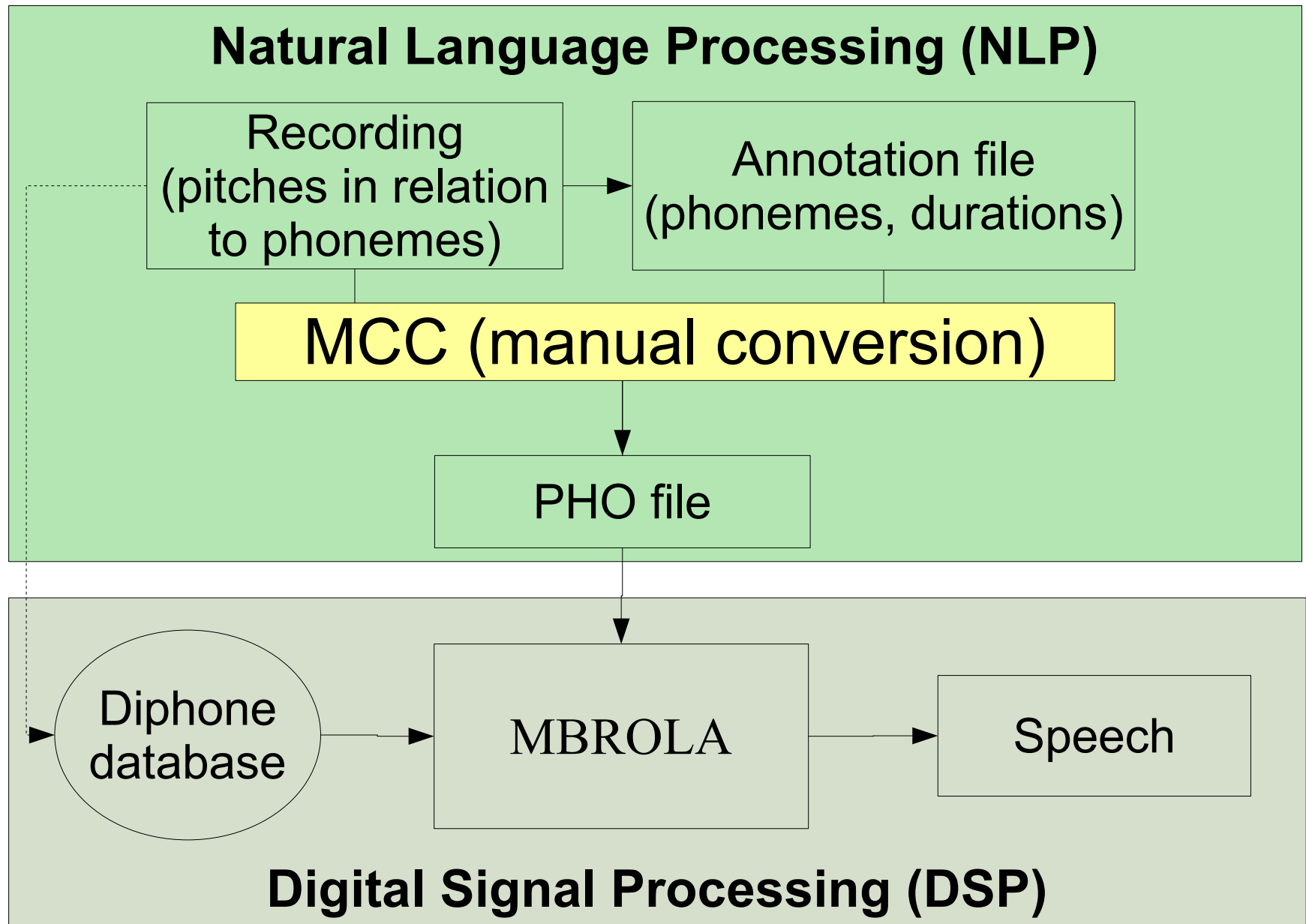
# Manual Close Copy Speech (MCCS) Resynthesis

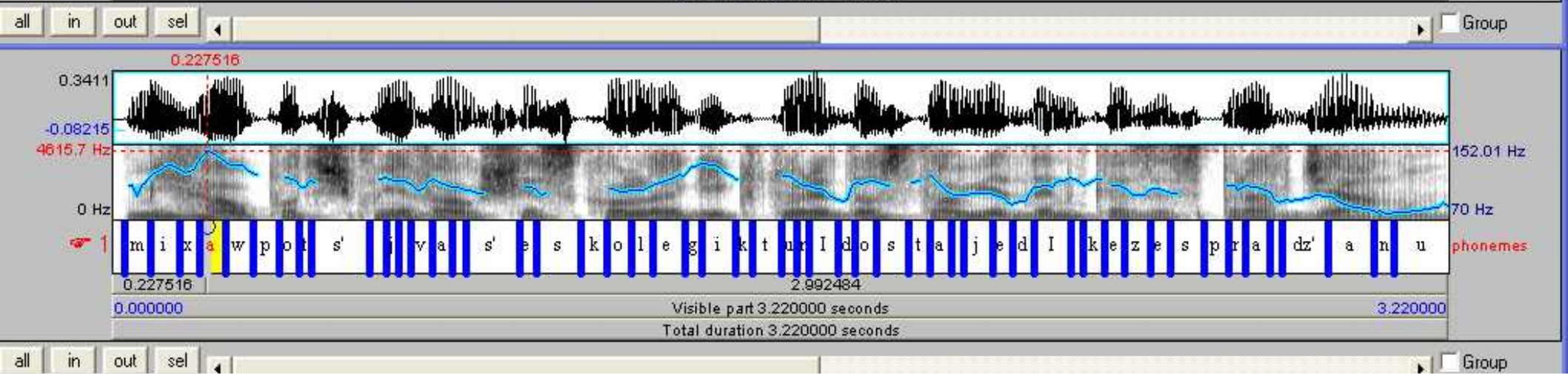
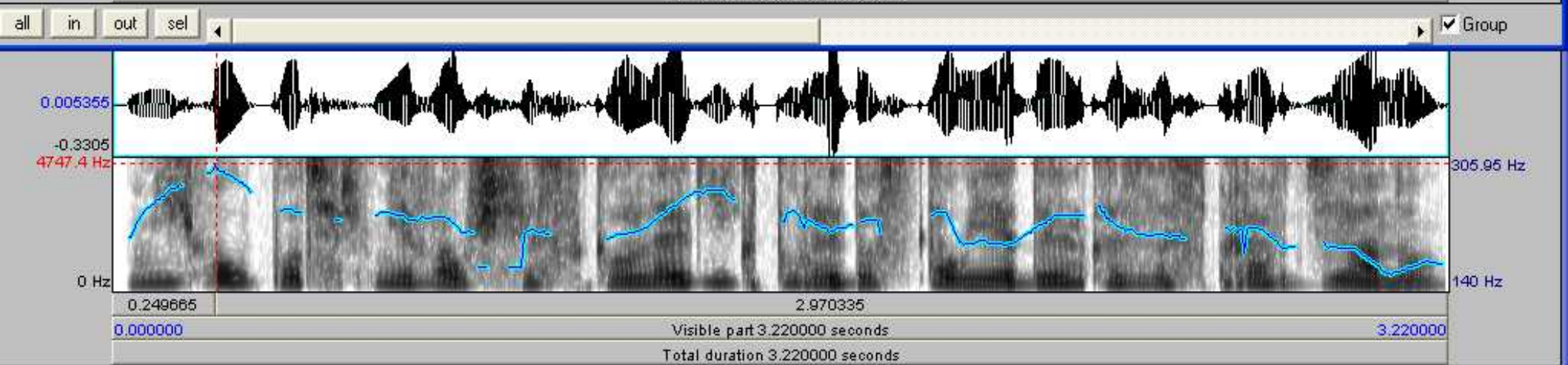
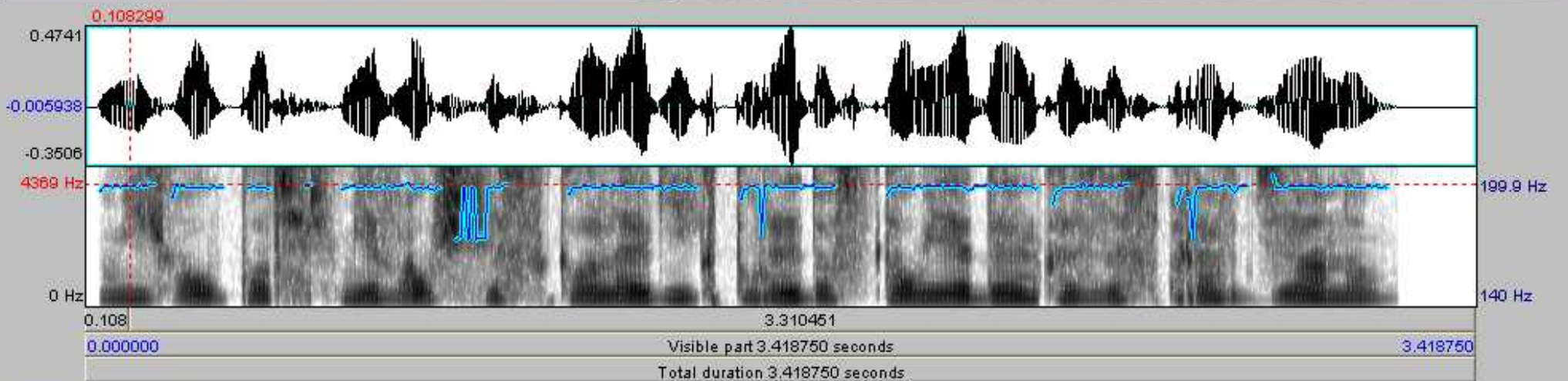
# Komponenty MCCS

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



# MCCS synthesis





# Synteza MCCS

- Monotonny

Monotone

- Synteza MCCS

MCCS

- Oryginał

Original recording

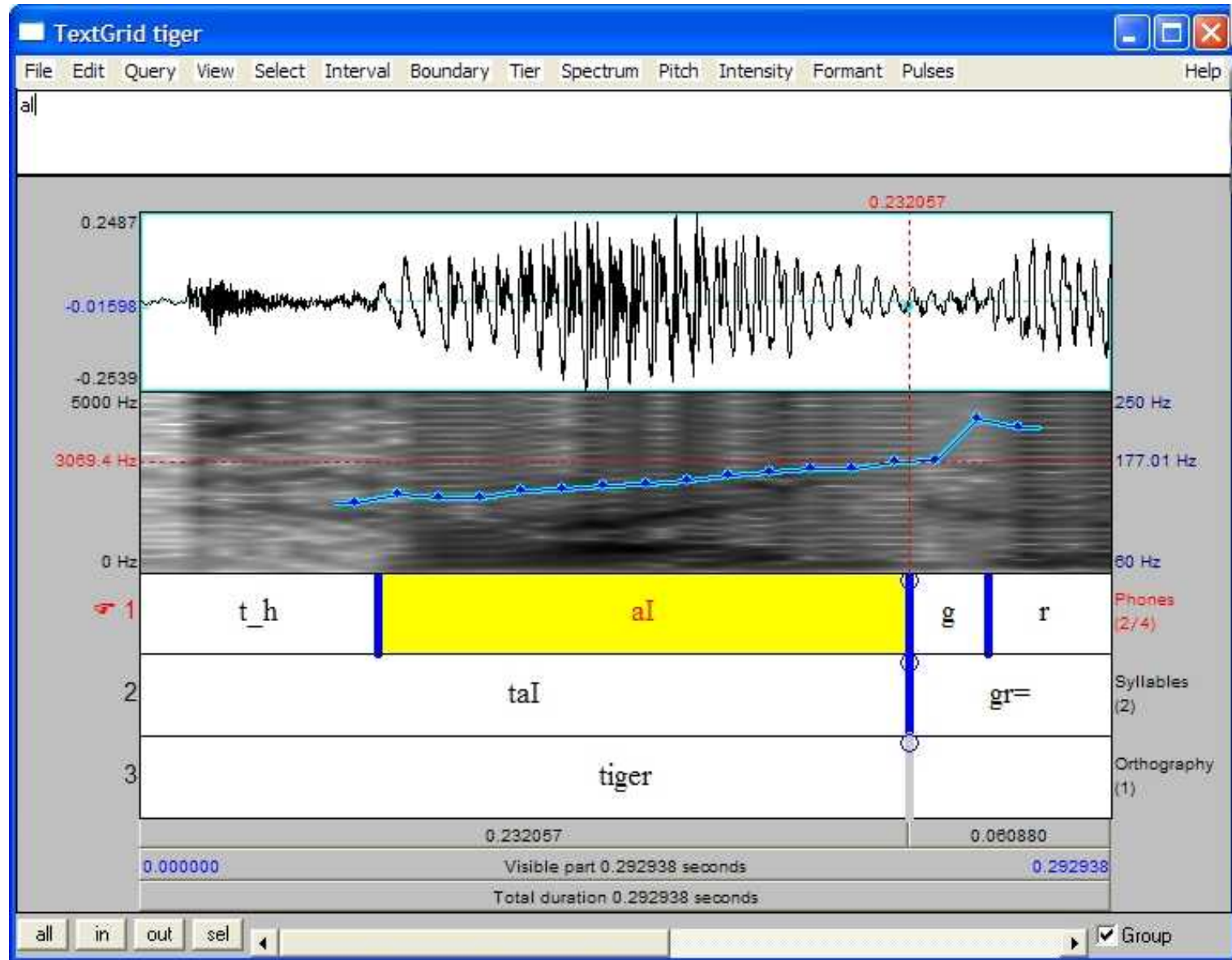
# Zastowowania MCCS

- 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



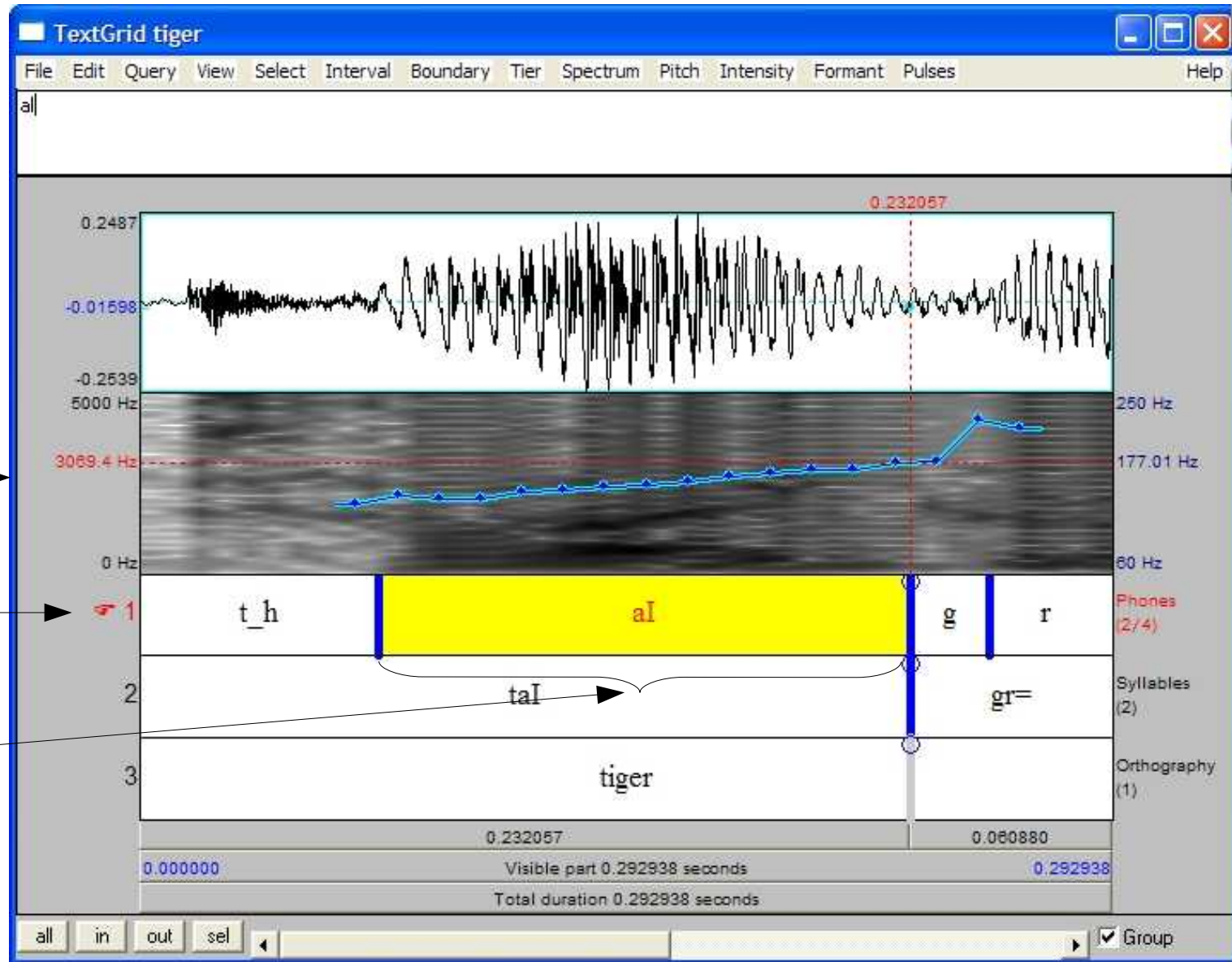
# Anotacja dla syntezy CCS

REQUIRED  
FOR  
SPEECH  
RE-SYNTHESIS

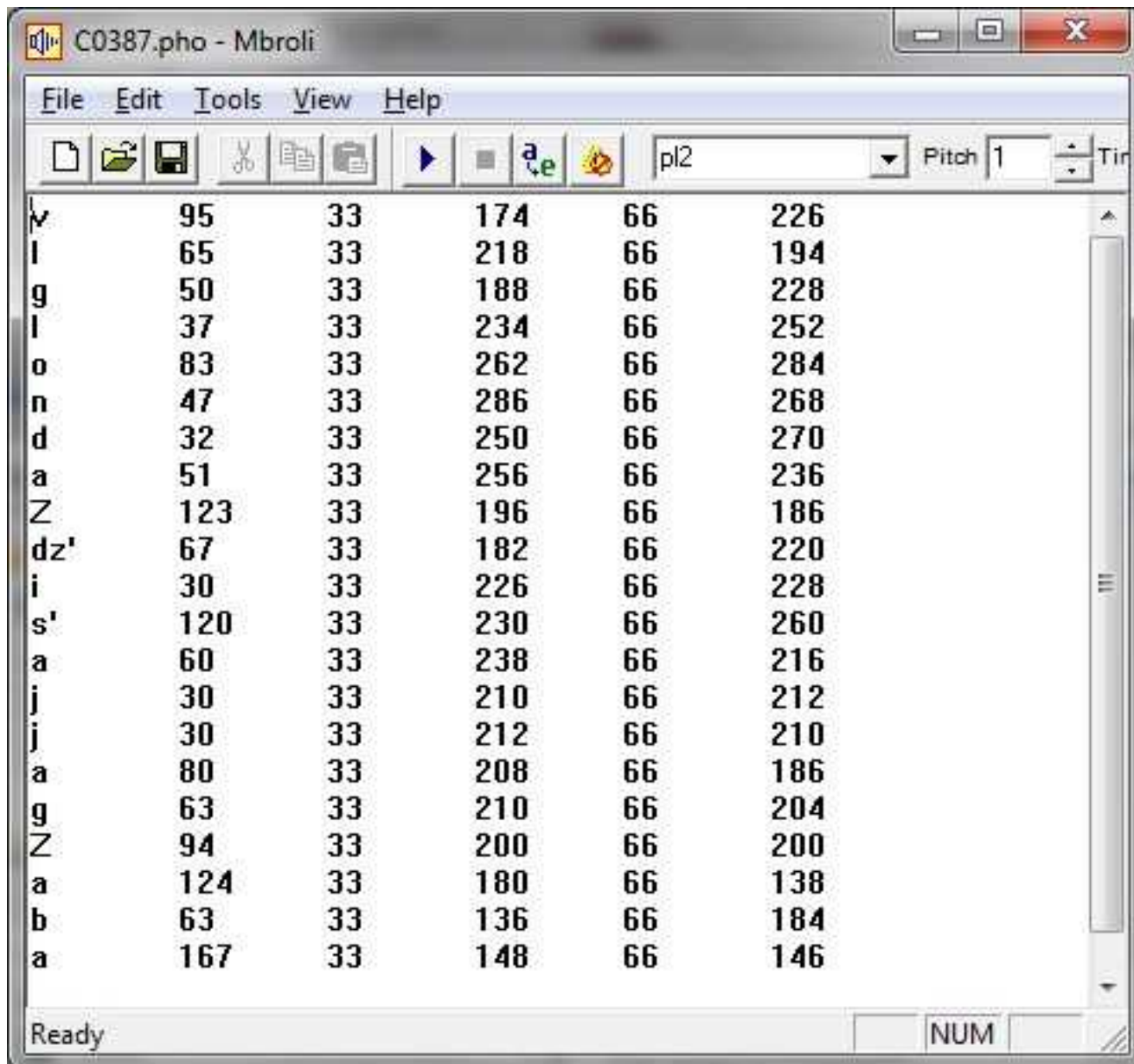
Pitch

Phones

Durations



# Format pliku PHO



The screenshot shows a window titled "C0387.pho - Mbroli" with a menu bar (File, Edit, Tools, View, Help) and a toolbar. The main area contains a table of phonetic data. The status bar at the bottom shows "Ready" and "NUM".

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



# Format pliku PHO

Głoski:  
SAMPA

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

Iloczas

Pozycja F0

Wartość F0

# Start z MBROLA

- Zainstaluj syntezytor **MBROLA**
- Przez „Control Panel” w Mbrola Tools dodaj bazy difonowe **pl1** i **en1**.
- Wykonaj syntezę MCCS w Mbrola.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)

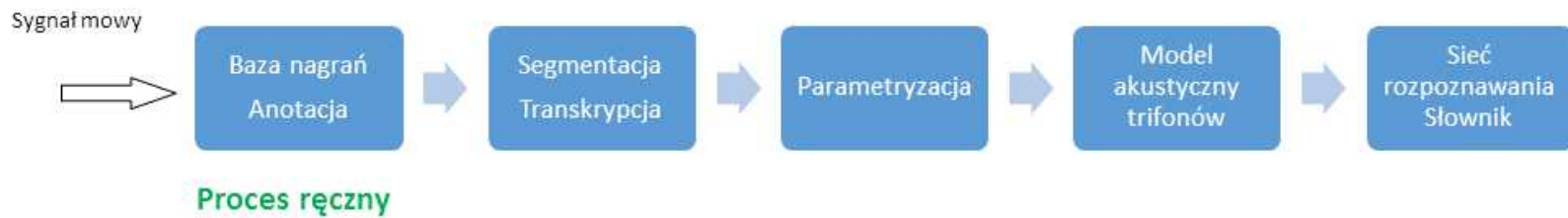
- Automatic Speech Recognition (ASR)
- Konwersja sygnału mowy na tekst

# 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

## Proces uczenia



## Proces rozpoznawania



Demenko, G., Cecko, R. Szymański, M., Owsiany, 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
- Dzielnice 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<sup>23</sup>)

# 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>
- Przygotuj się na test zaliczeniowy



Do zobaczenia za tydzień  
na teście zaliczeniowym!