Synteza mowy (TTS)
Rozpoznawanie mowy (ARM)
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.

Automatyczne Rozpoznawanie Mowy (ARM)

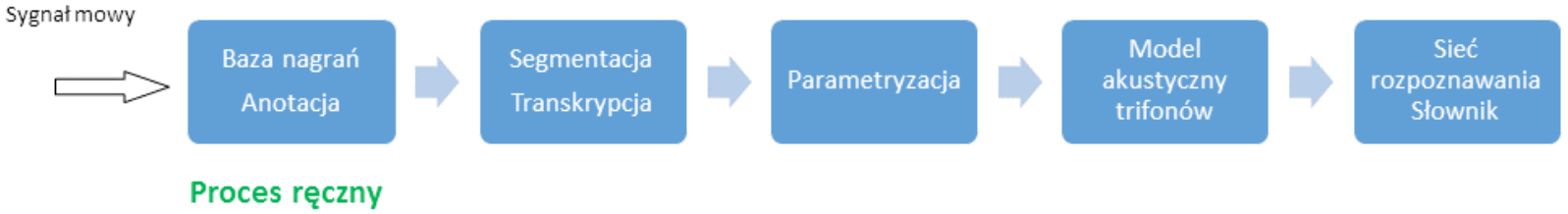
- Automatic Speech Recognition (ASR)
- Konwersja sygnału mowy na tekst

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., 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
- Przygotuj się na test zaliczeniowy

Do zobaczenia za dwa tygodnie na teście zaliczeniowym!