Analiza wymagań problemu wyrównania tekstu i mowy

streszczenie

Rozdział 1: Wstęp – wprowadzenie do zagadnienia

Rozdział 2: Sygnał mowy

Wprowadzenie do teorii sygnału dźwiękowego pod kątem przetwarzania mowy. Próbuję przedstawić tutaj co jest dźwiękiem wytwarzanym przez ludzi, który jest nazywany mową i w jaki sposób ludzie go odbierają. Opisuję standardowe praktyki przetwarzania sygnału w problemach rozpoznawania mowy, takie jak konwersja to skali melodycznej, transformacja Fouriera do spektrum częstotliwości, i w końcu transformacja do postaci cepstrum. Podsumowuję ten rozdział przedstawiając tworzenie cech mowy ze strumienia audio w bibliotece Sphinx.

Rozdział 3: Modelowanie języka.

Opisuję tutaj podejścia do podziału mowy na części składowe, czyli głoski i fonemy, jako wiedzę potrzebną do modelowania mowy. Następnie próbuję przedstawić dwa sposoby reprezentacji fonemów służące automatycznemu ich rozróżnianiu, liczenie odległości pomiędzy cechami dźwięku od próbki zakwalifikowanej do fonemu, oraz modelowanie fonemów za pomocą rozkładu normalnego i ukrytych modelów Markowa. Przedstawiam również sposoby trenowania rozkładu takiego modelu za pomocą metody analitycznej (rozkład Gaussa) oraz metody EM.

Rozdział 4: Proste wyrównanie oparte o pauzy i czas.

Opisuję tutaj proste podejście do problemu z zadaną grubą granulacją, ale za to z minimalną wymaganą wiedzą dodatkową. Rozpoczynam od sposobu detekcji mowy przy użyciu metod statystycznych, co jest potrzebne do rozpoznania pauz pomiędzy wypowiedziami, które powinny być skorelowane z podziałem tekstu na znaki przystankowe. Następnie opisuje algorytm niezbyt precyzyjnego wyrównania za pomocą algorytmu dynamicznego, który przypisuje wykryte kawałki mowy do tekstu na podstawie jego długości i estymacji czasu poszczególnych zdań. Kończę ten rozdział prezentując wyniki testu przeprowadzonego na tekście „Doktor Piotr”.

Rozdział 5: Wyrównania na podstawie modelu audio.

W tym rozdziale próbuję wykorzystać modele audio dla języków rosyjskiego i angielskiego aby wyrównać tekst z dokładnością co do słowa. Na wstępie prezentuję w jaki sposób takie modele są wykorzystywane w problemie rozpoznawania mowy oraz badam różnice między językami na podstawie różnic w zbiorze fonemów wykorzystanych w danych modelach audio. Pokazuję również prosty sposób generowania słowników zapisu fonetycznego słów w danym języku za pomocą gramatyk konwersji. Na koniec prezentuję statystyki porównań ręcznego wyrównania do uzyskanego za pomocą powyższych modeli.

Rozdział 6: Wyrównanie fonetyczne.

Opisuję tutaj dwa sposoby wyrównania z dokładnością co do fonemu, używając zbioru i modelu fonemów dla języka rosyjskiego oraz uproszczony dla języka polskiego wytrenowany przy użyciu wyników z rozdziału 5-tego. Algorytm trenowania modelów jest oparty na prostym użyciu rozkładów normalnego oraz algorytmu wyrównania opartego o technikę dynamicznego programowania. Testuję powyższe podejścia przy użyciu korpusu Corpora zawierającego zestaw krótkich nagrań otagowanych fonemami oraz prezentuję wyniki testów.

Rozdział 7: Trenowanie modelu fonemów z dużych plików audio

Badam w tym rozdziale sposób wytrenowania modelu audio oraz wykorzystania jego do uzyskania wyrównania z dokładnością co do słowa, bez użycia jakiegokolwiek wcześniej wytrenowanego modelu, a jedynie z minimalną wiedzą o fonetyce i interpunkcji danego języka. Wykorzystuję tutaj nieprecyzyjne wyrównanie z rozdziału 4-tego jako danych wejściowych wspomagających trenowanie modelu fonemów jako rozkładów normalnych niewiele różniącego się od tego z rozdziału 6-tego. Następnie opisuję metodą w jaki sposób wykorzystać takie modele do uzyskania otagowania słowami, która ze względów czasowych jest zmodyfikowaną metodą również z rozdziału 6-tego i bardzo przypomina sposób rozpoznawania mowy z biblioteki sphinxa, a mianowicie algorytm dynamicznego programowania z kolejką priorytetową o ograniczonej długości. Na koniec porównuje uzyskane wyniki z tymi uzyskanymi w rozdziale 5-tym.

Rozdział 8: Testowania wyrównania przy pomocy syntezatora mowy.

Prezentuję tutaj przykładowe wykorzystanie wyrównania z dokładnością co do fonemu, pod postacią prostego syntezatora mowy, z uwzględnieniem zastosowanie wyników syntezatora, jako sposobu testowania wyników algorytmu z rozdziału 7-mego. Oczywiście prezentuję tutaj algorytmy syntezujące mowę, które zlepiają części strumienia audio otagowane sekwencją fonemów odpowiadających tym uzyskanym z tekstu wejściowego. Wyniki syntezatora oceniam ręcznie, ponieważ testowanie automatyczne jest tu dość problematyczne. Skupiam się przy tym na wnioskach, które mogę wysnuć na temat jakości wyrównania.

Rozdział 9: Podsumowanie.

Streszczenie osiągniętych wyników oraz wnioski jakie można wysnuć na temat wymaganej wiedzy w problemie otagowania tekstu, a także kilka propozycji na kontynuację pracy.