POLITECHNIKA WROCŁAWSKA
WYDZIAŁ ELEKTRONIKI

KIERUNEK: Automatyka i Robotyka (AIR)
SPECJALNOŚĆ: Technologie informacyjne w systemach automatyki (ART)

PRACA DYPLOMOWA INŻYNIERSKA

Detekcja aktywności mówcy w systemach automatycznego rozpoznawania mowy

Voice activity detection in automatic speech recognition systems

AUTOR:
Paulina Szczerbak

PROWADZĄCY PRACĘ:
Prof. dr hab. inż. Ryszard Makowski

OCENA PRACY:

WROCŁAW 2017

Spis treści

Rozdział 1

Wstęp

Celem niniejszej pracy jest zaprezentowanie wybranych metod detekcji aktywności mówcy (VAD) w systemach automatycznego rozpoznawania mowy w oparciu o napisany program w języku C++. Kolejnym etapem jest porównanie zaimplementowanych metod pod względem dokładności detekcji w separowanych wyrazach oraz w dłuższych ciągach słów.

Rozdział 2. opisuje w uproszczony sposób proces wytwarzania mowy przez człowieka. Prezentuje, w jaki sposób działa aparat mowy oraz z jakich narządów się składa. Wyjaśnione zostaje zagadnienie fonemów oraz ich wykorzystanie w polskim alfabecie. Na koniec pokazany jest matematyczny model jaki można stworzyć wzorując się na naturalnym systemie generowania mowy.

Rozdział 3. zawiera wyjaśnienie na temat detekcji aktywności mówcy - czym jest oraz gdzie jest wykorzystywana. W tym rodziale opisane są również wybrane algorytmy, które zostały zestawione w dalszej części pracy. Przedstawiona jest zasada ich działania oraz pokrótce wyjaśniona kwestia implementacyjna każdego z nich.

Rozdział 4. prezentuje wyniki działania wybranych algorytmów dla separowanych słów. Pokazane są różnice w detekcji oraz ocena każdego z algorytmów.

Rozdział 5. zawiera wyniki detekcji dla całych ciągów słów oraz porównanie w działaniu wybranych algorytmów i ich ocenę.

Rozdział 2

Generowanie sygnału mowy

2.1 Mowa w życiu człowieka

Mowa w życiu większości ludzi stanowi podstawę komunikacji interpersonalnej. Jest sygnałem akustycznym, czyli rozważany jest zakres częstotliwości słyszanych przez człowieka, to jest od 20Hz do 16kHz. Zatem mowa to nic innego jak system artykułowanych dźwięków, które układają się zgodnie z konwencją wybranego języka. Pełni ona funkcję nie tylko komunikacyjną (przekazywanie informacji drugiej osobie o tym, co doświadczyliśmy, czy czego się dowiedzieliśmy), ale również ekspresyjną (można w niej zawrzeć informacje o emocjach nadawcy) oraz regulacyjną (wydawanie i przyjmowanie dyspozycji).

2.2 Biologiczny proces generowania mowy

Wszelkie metody przetwarzania sygnału mowy muszą bazować na strukturze sygnału, a ta jest niewątpliwie uzależniona od sposobu, w jaki jest on wytwarzany. Niegdyś generowanie sygnałów mowy było domeną jedynie organizmu człowieka, czyli systemu naturalnego. W celu stworzenia systemu, który w jakiś sposób operuje na sygnałach mowy, czyli np. syntezatora mowy, systemu generującego sygnały mowopodobne, systemu automatycznego rozpoznawania mowy czy detekcji aktywności mówcy, należy mieć przynajmniej podstawową wiedzę na temat systemu naturalnego - tego, w jaki sposób działa aparat mowy człowieka.

Wytwarzanie mowy przez człowieka jest procesem niezwykle skomplikowanym, który ma swój początek w mózgu, gdzie następuje konstrukcja wypowiedzi. Później następuje sformułowanie fonetyki i artykulacja poprzez aparat mowy. Ponadto, w procesie generowania mowy można wyróżnić cztery pomniejsze etapy:

- proces psychologiczny - wymyślenie i skonstruowanie wypowiedzi,
- proces neurologiczny - pobudzenie przez układ nerwowy mięśni, które biorą udział w wytwarzaniu mowy,
- proces fizjologiczny - proces kształtowania dźwięków mowy ludzkiej,
- proces aerodynamiczny - drgania i przepływ powietrza przez aparat mowy.

Pierwszym narządem wchodzącym w skład traktu głosowego człowieka są płuca - dostarczają one powietrze do procesu artykulacji, są źródłem zmian ciśnienia akustycznego. Organ mowy człowieka jest napędzany przez wydychane powietrze. Powietrze to, jest prowadzone przez oskrzela i tchawicę do krtani, a drgające w niej struny głosowe modyfikują ciśnienie i wytwarzają dźwięczne fragmenty mowy. Następnie, dzięki wnękom rezonansowym, tworzonym przez język, podniebienie, zęby oraz wargi, dźwięk ten jest modulowany.

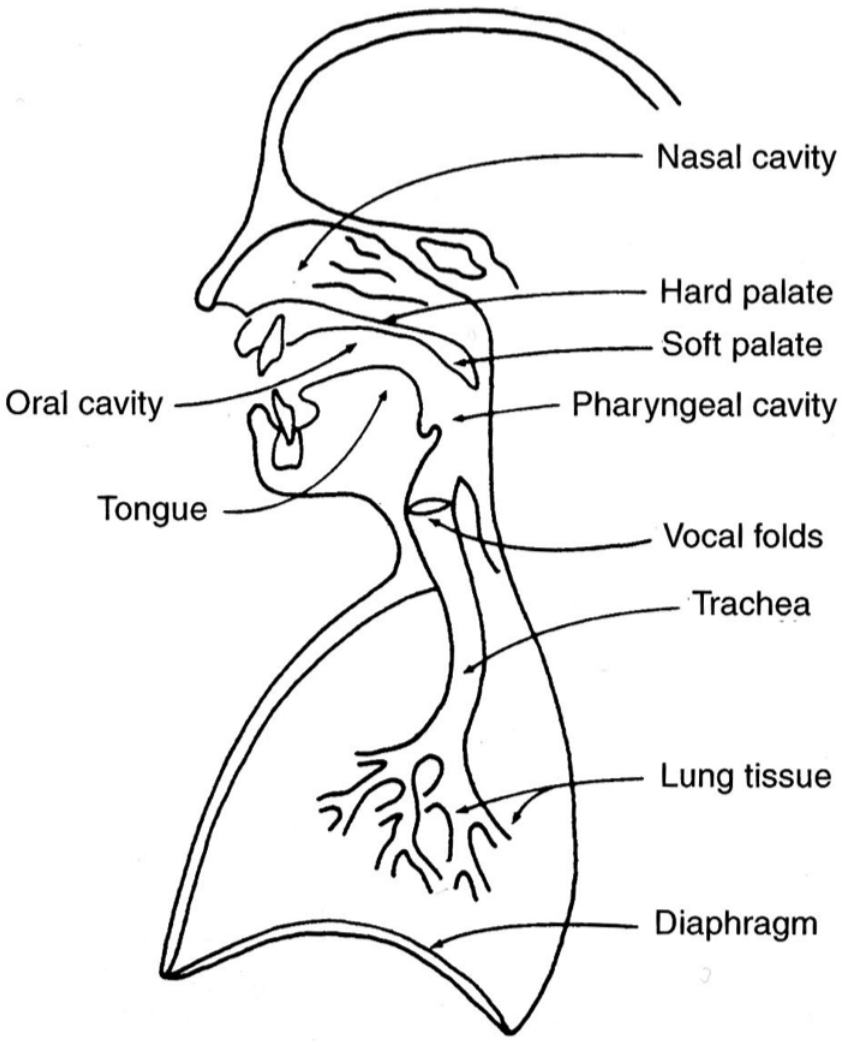

FIGURE 1-35

Schematic of the speech mechanism.

Rysunek 2.1 Aparat mowy człowieka

Niezwykle ważną rolę przy formowaniu tych wnęk, odgrywają ruchy żuchwy i policzków. Podczas generowania głosek nosowych zamknięta jama ustna spełnia rolę bocznika akustycznego, a dzięki odpowiedniemu ustawieniu języczka podniebienia miękkiego, fala dźwiękowa jest emitowana przez jamę nosową i nozdrza. Struktura traktu głosowego jest przedstawiona schematycznie na rysunku 2.1.

Ponadto, sterowanie całym systemem generowania mowy jest bardzo złożone i w dużej mierze opiera się na licznych sprzężeniach zwrotnych. Główną rolę odgrywa tutaj sprzężenie zwrotne, które poddaje jakość wydawanych dźwięków bezpośredniej ocenie poprzez analizator słuchowy. Dzięki temu proces artykulacji jest odpowiednio sterowany. Istotę tego sprzężenia zwrotnego potwierdzają trudności z mową wśród ludzi głuchych oraz ludzi słyszących, którzy tymczasowo przebywają w trudnych warunkach środowiskowych, które uniemożliwiają słyszenie własnego głosu.

SCHEMAT Z PDFA STR 23 SPRZERZENIE ZWROTNE

2.3 Jednostki fonetyczne

W celu przeprowadzania badań nad sygnałem mowy, należy wprowadzić jednostkę, która ułatwi wykonywanie operacji na całych słowach. Należy tutaj zaznaczyć, że słowo to dźwiękowy odpowiednik wyrazu, a wyraz to zapis słowa. Każde słowo zawiera w sobie przynajmniej jedną sylabę, każdą sylabę można również podzielić na mniejsze stany. W związku z tym, że każdy wyraz zawiera w sobie ciąg liter, to najczęściej wyróżnianymi elementami słowa są fonemy, zwane również głoskami. W większości przypadków na każdy fonem przypada odpowiadająca mu litera, ale są fonemy, które takiego odpowiednika nie posiadają. Istotne jest rownież, że każda litera może mieć różną reprezentację akustyczną w zależnośći od sąsiadujących z nią fonemów. Listę fonemów języka polskiego przedstawiono w tabeli 2.3.1.

TABELA Z FONEMAMI MOWY POLSKIEJ

2.4 Matematyczny model procesu generowania mowy

Rozdział 3

Wybrane metody detekcji aktywności mówcy

3.1 Detekcja aktywności mówcy

+ schemat jak w ksiazce RM, w ktorym miejscu jest vad w systemie ARM

Detekcja aktywności mówcy (Voice Activity Detection - VAD) jest powszechnie stosowana w systemach automatycznego rozpoznawania mowy. Podczas rejestrowania wypowiedzi do późniejszego przetwarzania jej przez system ARM, zostaje zarejestrowana cała wypowiedź mówcy, włącznie z częścią, która nie zawiera mowy. Jeżeli we fragmencie jest zawarty sygnał mowy, mówimy, że mówca jest aktywny. Aktywnością mówcy nazywa się emitowny przez niego dźwięk. Zawartość semantyczna wypowiedzi jest zawarta w głównej mierze we fragmentach, kiedy mówca jest aktywny. Analizowanie całego zarejestrowanego sygnału mowy, bez wykorzystania systemu VAD, jest oczywiście możliwe, aczkolwiek niepotrzebnie zwiększa czas obliczeń oraz istnieje prawdopodobieństwo, że fragment, gdy mówca nie jest aktywny, zostanie błędnie zaklasyfikowany jako jakiś konkretny fonem - zatem w dużej mierze może popsuć jakość rozpoznania. Detekcja aktywności mówcy w ogólnym przypadku zakłada, że sygnał może występować w dwóch stanach: tylko szum (brak sygnału mowy), szum + sygnał mowy. Korzystając z zagadnienia hipotez ze statystyki, możemy pierwszy stan oznaczyć jako hipotezę H_0, a drugi jako H_1, dzięki czemu możemy przedstawić to w następujący sposób:

$$\begin{aligned} H_0 &: f(n) = x(n) \\ H_1 &: f(n) = v(n) + x(n) \end{aligned} \tag{3.1}$$

Przy takim rozumowaniu konieczne jest określenie statystyki $S(n)$ sygnału, dzięki czemu możliwe będzie dokonywanie detekcji, a w dalszej kolejności zastosowanie kryterium decyzyjnego. Kryterium decyzyjne zwykle polega na porównaniu wartości $S(n)$ z progiem detekcji, który w mniej skomplikowanych algorytmach przyjmuje stałą wartość. Natomiast w tych bardziej złożonych, może występować np. jako funkcja czasu. Wartość stałej wartości progu jest ustalana w wyniku teoretycznych rozważań lub empirycznie. Zatem detekcja $\gamma(n)$, w ogólnej postaci, będzie prezentować się następująco:

$$\begin{aligned} S(n) &\geqslant \gamma(n) \rightarrow H_1 \\ S(n) &< \gamma(n) \rightarrow H_0 \end{aligned} \tag{3.2}$$

3.2 Algorytm bazujący na energii pojedynczej ramki

3.3 Algorytm bazujący na obwiedni sygnału

3.4 Algorytm Sohn

Donec cursus nulla vitae pede. Etiam quam pede, aliquet ut, pellentesque sed, sagittis non, est. Quisque egestas malesuada risus. Maecenas ultricies libero a quam. Nullam feugiat arcu. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. In interdum, risus ut gravida sollicitudin, leo sapien commodo dui, non consectetuer nisl nunc ac massa. Mauris a orci in eros venenatis euismod. Curabitur orci. Quisque pharetra, dui sed dignissim hendrerit, nibh ante malesuada eros, sed tincidunt magna lorem a tellus. Aliquam erat volutpat. Aenean pulvinar, metus et mattis dictum, massa lacus semper purus, quis vehicula augue mi et leo. Ut eu ipsum. Sed dictum dapibus nisi. Cras mattis. Nulla sed augue ac sem tempus condimentum.

Rozdział 4

Wyniki dla pojedynczych słów

4.1 Sposób oceny

Donec cursus nulla vitae pede. Etiam quam pede, aliquet ut, pellentesque sed, sagittis non, est. Quisque egestas malesuada risus. Maecenas ultricies libero a quam. Nullam feugiat arcu. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. In interdum, risus ut gravida sollicitudin, leo sapien commodo dui, non consectetuer nisl nunc ac massa. Mauris a orci in eros venenatis euismod. Curabitur orci. Quisque pharetra, dui sed dignissim hendrerit, nibh ante malesuada eros, sed tincidunt magna lorem a tellus. Aliquam erat volutpat. Aenean pulvinar, metus et mattis dictum, massa lacus semper purus, quis vehicula augue mi et leo. Ut eu ipsum. Sed dictum dapibus nisi. Cras mattis. Nulla sed augue ac sem tempus condimentum.

Rozdział 5

Wyniki dla ciągów słów

Rozdział 6

Podsumowanie

Spis rysunków

Spis tabel