Praca magisterska

Mateusz Steliga

kierunek studiów: Informatyka

specjalność: Systemy rozproszone i sieci komputerowe

Analiza skupień segmentów sygnału akustycznego w rozpoznawaniu mowy

Opiekun: dr inż. Jakub Gałka

Kraków, 2011
Oświadczam, świadomy(-a) odpowiedzialności karnej za poświadczenie nieprawdy, że niniejszą pracę dyplomową wykonałem(-am) osobiście i samodzielnie i nie korzystałem(-am) ze źródeł innych niż wymienione w pracy.

.................................................................
(czytelny podpis)
Spis treści

1 Wstęp ............................................................... 1

2 Rozpoznawanie mowy ............................................. 3
  2.1 Główne podejścia do rozpoznawania mowy .................. 4
    2.1.1 Podejście akustyczno-fonetyczne ..................... 4
    2.1.2 Podejście oparte o rozpoznawaniu wzorców .......... 4
    2.1.3 Podejście oparte o sztuczną inteligencję .......... 5
  2.2 Mechanizm produkcji mowy ................................. 5
  2.3 Percepcja dźwięku przez człowieka ....................... 5

3 Przetwarzanie i metody analizy sygnału mowy .................... 7
  3.1 Krótkookresowa reprezentacja spektralna sygnału ........ 7
  3.2 Ciągła transformacja falkowa ............................. 8

4 Metody analizy danych wielowymiarowych ........................ 11
  4.1 Analiza skupień ............................................. 11
    4.1.1 k-means ............................................. 11
    4.1.2 Klastrowanie hierarchiczne ......................... 12
    4.1.3 OPTICS ............................................. 12
  4.2 Skalowanie wielowymiarowe ................................ 12
  4.3 Liniowa analiza dyskryminacyjna ......................... 13

5 Korpus .......................................................... 15
  5.1 Głoski ..................................................... 15

6 Analiza skupień segmentów sygnału akustycznego ................ 19
  6.1 Oznaczenia ............................................... 19
    6.1.1 Adjusted Rand Index ................................ 19
  6.2 Parametryzacja korpusu ................................... 20
6.3 Analiza metodą k-means .................................................. 24
  6.3.1 ARI dla wyników klastrowania metodą k-means .................... 25
  6.3.2 Common Cluster Factor dla k-means ................................. 26
  6.3.3 Wnioski z analizy metodą k-means ................................. 28
6.4 Klasyfikacja metodą najbliższych środków ciężkości .................. 29
  6.4.1 Współczynnik Recognized-As Factor (RAF) .......................... 30
  6.4.2 Odległość euklidesowa (EUC) ...................................... 30
  6.4.3 Znormalizowana odległość euklidesowa (NED) ...................... 33
  6.4.4 Porównanie z k-means ................................................ 35
6.5 Badanie separowalności głosek metodą LDA ............................ 35
  6.5.1 Liniowy dyskriminator Fisher’a .................................... 36
  6.5.2 Rozkład gęstości głosek ............................................. 36
  6.5.3 Rozróżnialność par głosek .......................................... 38
  6.5.4 Separowalność głosek na podstawie metod NM i LDA ............... 43
  6.5.5 Poprawa wyników klasyfikacji NM za pomocą metody LDA .......... 43
6.6 Klasyfikacja artykulacji .................................................. 44
6.7 Klasyfikacja metodą k-NN ................................................. 47
6.8 Skalowanie wielowymiarowe .............................................. 49
6.9 Analiza algorytmem OPTICS ............................................. 50

7 Podsumowanie ................................................................. 53

Literatura ................................................................. 55

Spis rysunków ................................................................. 57

Spis tablic ................................................................. 61
Wstęp

Rozpoznawanie mowy i dziedziny z tym związane są ekstensywnie badane od kilkudziesięciu lat. Mimo to nie udało się do tej pory stworzyć takiego systemu, który rozpoznawałby mowę ludzką tak efektywnie jak sam człowiek: bez względu na warunki środowiskowe oraz akcent lokalny, niedbalość czy różnice w wymowie (wynikające z niewielkich różnic w budowie narządów mowy) pomiędzy mówcami.


Celem tej pracy jest zbadanie rozłożenia skupień fonemów w 20-wymiarowej przestrzeni parametryzacji MFCC (Mel-Frequency Cepstrum Coefficients). Na tej podstawie określona zostanie między innymi separowalność różnych par zbiorów. Informacja o tym, które głoski są w danej parametryzacji słabo rozróżnialne może być pomocna w budowie systemów automatycznego rozpoznawania mowy, nie tylko na etapie analizy sygnału mowy, ale także
podczas analizy słownikowej, gdy określane jest prawdopodobieństwo z jakim dany segment reprezentuje różne fonemy.

W rozdziale 2 niniejszej pracy krótko opisano problem rozpoznawania mowy i główne podejścia do jego rozwiązania, oraz mechanizm produkcji i percepacji dźwięku przez człowieka. Rozdział 3 przedstawia główne metody analizy sygnału mowy, za pomocą których jest on konwertowany do reprezentacji parametrycznych. Rozdział 4 zawiera przegląd metod do analizy danych wielowymiarowych, a rozdział 5 informacje i podstawowe statystyki korpusu mowy, na którym przeprowadzono badania. Rozdział 6 zawiera opis i wyniki przeprowadzonych badań.
Rozpoznawanie mowy

Rozpoznawanie mowy jest procesem polegającym na konwersji mowy z formatu dźwiękowego do jej transkrypcji w formacie tekstowym. Jednym z najbardziej kłopotliwych aspektów rozpoznawania mowy przez komputer jest jej interdyscyplinarna natura. System automatycznego rozpoznawania mowy musi rozwiązywać problemy w następujących dziedzinach:

- przetwarzanie sygnałów - wydajne otrzymywanie z sygnału mowy relevantnych informacji (np.: analiza spektralna)
- fizyka (akustyka) - wykorzystanie związków pomiędzy sygnałem mowy (w sensie fizycznym), a fizjologicznymi mechanizmami (aparat głosowy człowieka), za pomocą których został on wyprodukowany i za pomocą których jest słyszany (mechanizm słuchu człowieka)
- rozpoznawanie wzorców - tworzenie wzorców elementów mowy w celu ich porównania z rozpoznawany sygnałem
- teoria komunikacji i informacji - estymacja parametrów modeli statystycznych, algorytmy kodowania i dekodowania
- lingwistyka - relacje pomiędzy dźwiękami (fonologia), słowami (syntaktyka), ich znaczeniem (semantyka) i znaczeniem kontekstowym (pragmatyka)
- fizjologia - zrozumienie mechanizmów centralnego układu nerwowego, które są wykorzystywane w produkcji i percepcji mowy.
- informatyka - efektywna implementacja algorytmów
- psychologia - zastosowanie technologii rozpoznawania mowy w praktyce

Stworzenie dedykowanych systemów rozpoznawania mowy mogłoby ułatwić interakcję człowiek-maszyna oraz człowiek-człowiek w wielu dziedzinach, np.: służba zdrowia, wojskowość (np.: nawigacja), telefonia (np.: obsługa klienta), automatyczna transkrypcja (ang. text-to-speech).
2.1 Główne podejścia do rozpoznawania mowy

Poniżej przedstawione zostały trzy główne podejścia do automatycznego rozpoznawania mowy na podstawie [13].

2.1.1 Podejście akustyczno-fonetyczne

Podejście akustyczno-fonetyczne opiera się o teorię fonetyki akustycznej, według której istnieje skończona liczba rozróżnialnych jednostek w języku mówionym (fonemów), z których każda jest opisana zbiorem właściwości objawiających się w sygnale mowy lub jego spektrum. Zakłada się, że pomimo tego, że właściwości te w wysokim stopniu różnią się pomiędzy mówcami i są zależne od sąsiedztwa każdego fonemów (efekt ten nazywany jest koartykulacją), możliwe jest zbadanie reguł tej zmienności i zastosowanie ich w praktyce. Podejście to podzielone jest na kilka etapów. Pierwszym etapem jest segmentacja fali dźwiękowej w taki sposób, aby parametry sygnału w każdym z przedziałów reprezentowały jeden (lub kilka) jednostek fonetycznych. Polega to często, na przykład, na podzieleniu sygnału na obszary o niskim poziomie zmienności pewnych jego cech, Możliwe jest otrzymanie kilku różnych segmentacji danej wypowiedzi. Następnie dla każdego z tych segmentów określany jest reprezentowany przez niego fonem (lub kilka, każdy z innym prawdopodobieństwem poprawności). Na końcu dla otrzymanej sieci fonemów ang. phoneme lattice szuka się słów (metodą słownikową), z których składa się wypowiedź.

2.1.2 Podejście oparte o rozpoznawaniu wzorców

To podejście jest oparte o rozpoznawanie wzorców mowy, bez bezpośredniego określania parametrów poszczególnych fonemów (tak jak przy podejściu akustyczno-fonetycznym) i segmentacji. Metoda ta składa się z dwóch faz: nauki wzorców (polegającej na nauce wzorców mowy na podstawie danych treningowych), oraz rozpoznawaniu wzorców (polegającym na rozpoznawaniu wzorców mowy w danych testowych za pomocą ich porównywania z tymi określonymi w fazie nauki). Wzorce mogą być rozpoznawane (i uczone) na poziomie fonemów, słów lub całych fraz. W fazie nauki system uczy się, które właściwości akustyczne danej klasy (fonemu) są miarodajne i powtarzają się we wszystkich elementach treningowych wzorca. Głównymi zaletami tego podejścia są: prostota użycia, poparcie przez teorie matematyczne i komunikacji dla procedur używanych w fazie nauki i rozpoznania, odporność na różnych mówców, zbiór parametrów, algorytmów porównania wzorców i reguł decyzyjnych; potwierdzona wysoka wydajność.
2.3 Percepcja dźwięku przez człowieka

2.1.3 Podejście oparte o sztuczną inteligencję

Podejście z wykorzystaniem sztucznej inteligencji jest hybrydą, w której wykorzystywane są, między innymi, idee stosowane w dwóch poprzednich podejściach. Metoda ta próbuje zmechanizować procedurę rozpoznawania mowy na wzór tego, jak robi to człowiek, wykorzystując wiedzę pochodzącą z różnych źródeł: akustyczną (zespół jakich cech akustycznych odpowiada konkretnym fonemom), leksykalną (jakie ciągi dźwięków tworzą poprawne słowa), syntaktyczną (jakie ciągi słów, tworzą poprawne zdania), semantyczną (które zdania mają sens w zadany kontekście) i pragmatyczną (np. w celu rozróżnienia znaczenia dosłownego od przełożonego danego zdania).

2.2 Mechanizm produkcji mowy

Narządy aparatu produkcji mowy u człowieka można podzielić na trzy grupy: oddechowe, fonacyjne i artykulacyjne. Narządy oddechowe (płuca, przepona, oskrzela i tchawica) są odpowiedzialne za przepływ powietrza, który jest niezbędny do wytworzenia dźwięku. Podczas przepływu przez krtania (aparat fonacyjny), powietrze pobudza napięte struny głosowe, co wprawia je w drgania o częstotliwości zależnej od siły ich napięcia. Drgające struny głosowe powodują powstawanie podłużnej fali dźwiękowej (naprzemiennych zagęszczeń i rozrzedzeń strumienia powietrza), która jest następnie modulowana w trzech jamach aparatu artykulacyjnego: gardłowej, ustnej i nosowej oraz przez ich narządy ruchome (wargi, język, podniebienie) i nieruchome (zęby, dziąsła, podniebienie).

Mowa jest sekwencją różnych dźwięków, według zasad określanych przez dany język. Każdemu z nich odpowiada inne położenie i kształt narządów mowy oraz inna częstotliwość drgań strun głosowych i tym podobne. Ponieważ nie jest możliwa zmiana stanu tych narządów w czasie zerowym, gdyż wymagałoby to nieskończonego przyśpieszenia, a tym samym nieskończonej energii lub zerowej masy, mowa jest procesem ciągłym. Oznacza to, że głoski się wzajemnie „przenikają” (zjawisko koartykulacji); nie można ustalić konkretnego punktu, w którym jedna głoska zamienia się w drugą. Ciągłość mowy jest główną przyczyną trudności jej rozpoznania przez maszynę.

2.3 Percepcja dźwięku przez człowieka

Na rysunku 2.1 przedstawiony został model mechanizmu słuchowego człowieka w postaci diagramu blokowego [12]. Fała dźwiękowa dociera poprzez ucho zewnętrzne do ucha wewnętrznego, gdzie bębenek i kosteczki słuchowe konwertują ją na drgania mechaniczne, przenosząc je na błonę podstawową ślimaka słuchowego (*lac. cochlea*). Każdy fragment błony jest pobudzany przez inne częstotliwości, przez co dokonywana jest wstępna analiza spektralna dźwięku.
Błona jest pokryta komórkami rzęsatymi, które zamieniają wibracje membrany na impulsy neuronowe. To jak przetwarzany jest dźwięk dalej, w mózgu, nie jest dokładnie zbadane. Przeprowadzono jednak wiele eksperymentów (z użyciem dźwięków i szumów), które pomogły poznac reakcje mechanizmu słuchowego człowieka na takie własności akustyczne jak natężenie czy częstotliwość.

Jednym z najważniejszych czynników w percepcji mowy jest "głośność". Jest ona wyrażana poprzez zestawienie rzeczywistego ciśnienia czystego tonu do jego odczuwalnej głośności. Głośność odczuwalna jest zależna od częstotliwości dźwięku, np.: aby głośność dźwięków o niskich częstotliwościach była taka sama jak tych o średnich częstotliwościach, natężenie tych pierwszych musi być większe.

Analiza częstotliwościowa dokonywana przez błonę słuchową odpowiada zbiorowi filtrów pasmowych, których szerokość pasma zwiększa się wraz ze wzrostem częstotliwości. Szerokości te są stałe dla wartości poniżej 500 Hz i coraz większe powyżej. Średnio używa się 25 filtrów pasmowych (ang. critical band filters) rozciągniętych na przedział od 0 do 20 kHz.

Nierówny podział częstotliwości na pasma w ślimaku słuchowym jest także przyczyną nie-liniowego, w stosunku do częstotliwości, rozpoznawania przez człowieka wysokości dźwięków (tonów) oraz efektu maskowania psychoakustycznego (ang. auditory masking), który polega na maskowaniu przez głośny ton cichszych tonów w jego bliskim otoczeniu.
Przetwarzanie i metody analizy sygnału mowy

Niezależnie od wybranego podejścia do rozpoznawania mowy, każdy system posiada moduł odpowiedzialny za przetworzenie sygnału mowy, czyli zamianę wejściowej fali dźwiękowej na pewnego rodzaju reprezentację parametryczną, która jest potem wykorzystywana do procesu rozpoznania mowy. Reprezentacja ta posiada zazwyczaj mniejszą ilość informacji (ang. information rate) niż oryginalny sygnał (wiele informacji nieistotnych z punktu widzenia rozpoznawania mowy jest w tym procesie odrzucane).

Istnieje wiele parametrycznych reprezentacji sygnału, należą do nich: short time energy, zero crossing rates, level crossing rates i inne. Najważniejszą jest jednak krótko-okresowa reprezentacja spektralna ang. short-time spectral envelope, która stanowi bazę wstępnej analizy sygnału dla większości systemów rozpoznawania mowy. Jest ona często stosowana w połączeniu z zestawem filtrów pasmowych w skali melowskiej [16].

3.1 Krótkookresowa reprezentacja spektralna sygnału

Reprezentacja ta opiera się o krótkookresową transformację Fouriera:

$$\hat{s}(f,b) = \int_{-\infty}^{\infty} w(t-b)s(t)e^{-2\pi ift} dt$$

gdzie $w(t-b)$ jest oknem o centrum położonym w czasie $b$, $f$ - częstotliwością dla jakiej wykonywana jest analiza, a $s(t)$ - analizowanym sygnałem. Manipulując szerokością okna można zmieniać rozdzielczość czasową i częstotliwościową dokonywanej analizy. Szerokie okno zwiększa rozdzielczość częstotliwościową kosztem czasowej, a okno o małej szerokości na odwrot.

Na rysunku 3.1 przedstawiono spektrogram dla przykładowej wypowiedzi, obliczony z wykorzystaniem dyskretniej realizacji powyższej transformacji Fouriera oraz filtrów pasmowych w skali melowskiej. Analiza została wykonana dla okna o szerokości 25 ms przesuwnanego o 10 ms. Można z niego odczytać kilka podstawowych informacji na temat wypowiedzi i rodzajów użytych w nim głosek. Przykładowo, ramki, które reprezentują głoski dźwięczne (w tym...
przypadku jest to /jέn/), mają wysokie wartości dla współczynników wymiarów (w przybliżeniu) od pierwszego do siódmego. Są to współczynniki odpowiadające za częstotliwości, z których składa się sygnał reprezentujący głoski dźwięczne (samogłoski i półsamogłoski). Rysunek ten ilustruje także zjawisko ciągłości mowy: widać, że parametry zmieniają się stopniowo: największe „nasilenie dźwięczności” przypada na początek głoski /e/ i stopniowo wygasza się, aż zanika całkiem przed /ć/ (przed którym występuje kısa pauza, w czasie której wytworzone zostaje ciśnienie potrzebne do wymówienia głoski zwartościowowej, jaką jest /ć/).

W celu otrzymania cepstrum sygnału, oblicza się transformatę Fouriera z logartymu otrzymanego wcześniej widma spektralnego.

### 3.2 Ciągła transformacja falkowa

Jedną z nieporządanych cech okienkowej transformacji Fouriera (wykorzystywanej do obliczania krótkookresowego spektrum sygnału) jest niezmienna szerokość okna. Analiza dla każdej z częstotliwości jest przeprowadzana dla takiej samej szerokości okna. Składowe sygnału o wyższych częstotliwościach można jednak analizować dla mniejszych ramek czasowych niż składowe o niższych częstotliwościach. Transformacja falkowa jest metodą analizy częstotliwościowej, która automatycznie reguluje szerokość okna w zależności od częstotliwości analizowanego sygnału.
Rysunek 3.2. Różnice w okienkowaniu dla metod analizy częstotliwościowej: Short-Time Fourier Transform (na górze) Wavelet Transform (na dole) [14]
Metody analizy danych wielowymiarowych

Isnieje wiele rodzajów metod o różnych zastosowaniach, które służą do analizy danych wielowymiarowych.

4.1 Analiza skupień

Analiza skupień (ang. clustering) jest metodą służącą do grupowania elementów zbioru danych, które charakteryzują się małymi odległościami wewnątrz grup, stanowią zagęszczenia w przestrzeni danych lub mają w niej szczególny rozkład prawdopodobieństwa. Jednym z głównych celów klastrowania jest klasyfikacja danych sugerowana przez ich naturalne zgrupowanie, co często, w przypadku danych wielowymiarowych, nie jest możliwe do wykonania za pomocą wizualizacji lub podobnych jej, prostych metod.

Analiza skupień jest przeważnie procesem interacyjnym, polegającym na szukaniu podobieństw, różnic oraz innych zależności pomiędzy danymi i wykorzystywaniu ich do grupowania danych. Algorytmy klastrowania posiadają często zestaw parametrów wejściowych, od których odpowiedniego wyboru zależy poprawność wyniku. Dlatego też analiza jest przeprowadzana kilka razy dla różnych parametrów. Możliwa jest także modyfikacja parametrów pomiędzy iteracjami (np.: zmiana parametrów algorytmu OPTICS po każdej iteracji, aż do znalezienia takich, dla których otrzymano k klastrów). Istnieje kilka różnych modeli klastrów w zależności od wybranej metody. Poniżej krótko przedstawiono najważniejsze metody analizy skupień.

4.1.1 k-means

Metoda k-means polega na podziale n punktów wejściowych na zadanych k grup (klastrów). Klastry są w niej reprezentowany jako zbiór k punktów (środków ciężkości klastrów). Każdy z n punktów jest przyporządkowywany do najbliższego środka ciężkości, co stanowi podział na k klastrów. Startowe środki ciężkości są wybierane losowo. Każda iteracja algorytmu opiera się o wybór nowych klastrów (polega to na przypisaniu punktu do klastra, którego środek ciężkości znajduje się najbliżej), a następnie ustaleniu nowych środków ciężkości na podstawie...

4.1.2 Klastrowanie hierarchiczne

Klastrowanie hierarchiczne polega na utworzeniu hierarchii klastrów. Wyróżnia się dwa rodzaje takiego klastrowania: algomeratywne, w którym każdy punkt początkowo stanowi klaster i w każdym kroku dwa klastry są grupowane w tworząc nowy poziom hierarchii; oraz dzielące, w którym początkowo wszystkie punkty należą do jednego klastra i w każdym kolejnym kroku każdy klaster jest dzielony na dwa nowe na niższym poziomie hierarchii. Istnieją różne kryteria, według których następuje podział (i algomeracja) klastrów. Na wynik klastrowania wpływ ma także wybór metryki, w której obliczane są odległości pomiędzy klastrami. W większości przypadków złożoność obliczeniowa algorytmu wynosi $O(n^3)$ dla wersji algomeratywnej i $O(2^n)$ dla wersji dzielącej.

4.1.3 OPTICS


4.2 Skalowanie wielowymiarowe

Znaczna część trudności analizy danych wielowymiarowych wynika z naszej nieumiejętności wizualizacji struktur danych wielowymiarowych. Problem staje się jeszcze trudniejszy, gdy dochodzi do tego miary podobieństwa i różnic pomiędzy obiektami lub ich grupami, które nie mają znanych odpowiedników w mierze odległości. Jednym z podejść do rozwiązania tego problemu jest próba reprezentacji zbioru danych wielowymiarowych w jakiejś przestrzeni o


**4.3 Liniowa analiza dyskryminacyjna**

Liniowa analiza dyskryminacyjna jest popularną metodą używaną do odnalezienia liniowej kombinacji cech, które najlepiej charakteryzują lub separują dwie (lub więcej klas obiektów). Wynikowa kombinacja może zostać użyta jako klasyfikator liniowy lub do redukcji ilości wymiarów. Metoda ta zakłada normalny rozkład zmiennych niezależnych (ang. independent variables) i nadaje się do stosowania dla danych o takim właśnie rozkładzie.
Korpus

Korpus mowy jest to zbiór posegmentowanych i opisanych nagrań audio. Segmentacja może być dokonana np.: na poziomie fonemów, sylab lub słów. Korpus może być użyty od celów takich jak: analiza mowy czy trening systemów rozpoznawania mowy. Do badań opisanych w tej pracy użyto korpusu "Corpora" o następujących parametrach:

- 26 mówców - tylko płci męskiej, każdy 365 wypowiedzi w języku polskim: cyfry, litery, imiona oraz zdania przygotowane przez lingwistów.
- format WAV, częstotliwość próbkowania 16 kHz
- wszystkie nagrania posegmentowane na fonemy i opisane
- segmentowanie z wykorzystaniem algorytmu programowania dynamicznego: jeden mówca posegmentowany ręcznie, reszta automatycznie na jego podstawie

Szczegółowe informacje na temat korpusu znajdują się w [6]

Rysunek 5.1 przedstawia przykładową segmentację i opis słowa “pięć”. Fala dźwiękowa jest podzielona na fragmenty, z których każdy odpowiada jednej głośce. Każde nagranie z korpusu jest posegmentowane i opisane w analogiczny sposób.

5.1 Głoski

Poniższe wykresy przedstawiają statystyki fonemów występujących w korpusie. Na wykresie 5.2 przedstawiono średni czas trwania poszczególnych fonemów, a na wykresie 5.3 ilość ich wystąpień w korpusie.

Na rysunku 5.4 przedstawiono macierz następstw fonemów. Ostatni rząd odpowiada ciąży. Może ona występować przed gloską, która rozpoczyna wypowiedź, lub przed gloską przed którą mówca robi krótką pauzę. Wartość elementu macierzy $Ms(a, b)$ wyliczana jest w następujący sposób:

$$Ms(a, b) = \frac{succ(a, b)}{A} \quad (5.1)$$
Rysunek 5.1. Segmentacja wypowiedzi słowa ‘pięć’

Rysunek 5.2. Średni czas trwania poszczególnych fonemów [sec]

Rysunek 5.3. Ilość wystąpień poszczególnych fonemów w korpusie
gdzie $\text{succ}(a, b)$ odpowiada ilości wystąpień głoski $b$ po głosce $a$ oraz $A$ to ilość wystąpień głoski $a$ w korpusie. Patrząc na rząd $a$ widzimy, przez jakie głoski następowaly po głosce $a$; natomiast w kolumnie $a$ pokazano jakie głoski występowały przed głoską $a$.

**Rysunek 5.4.** Macierz częstości występowania par głosek w korpusie.

Na rysunku 5.5 znajdują się trójkąty samogłoskowe, na których przedstawiono pewne ich cechy.

**Rysunek 5.5.** Trójkąt samogłoskowy obrazujący pionowe i poziome położenie języka przy wymowie samoglosek.

W tabeli 5.1 zebrano wszystkie fonemy występujące w korpusie i pogrupowano je na podstawie ich cech akustycznych, tak by głoski podobnie brzmiące dla ucha ludzkiego były w jednej grupie (nie dotyczy to samogłosek, które znajdują się wszystkie w jednej grupie).
Tablica 5.1. Głoski występujące w korpusie pogrupowane ze względu na wyróżniające je cechy akustyczne. W nawiasach podano oznaczenia używane na rysunkach.

<table>
<thead>
<tr>
<th>głoski</th>
<th>opis</th>
</tr>
</thead>
<tbody>
<tr>
<td>a, o, u, e, i, y</td>
<td>samogłoski</td>
</tr>
<tr>
<td>a(a), e(e)</td>
<td>samogłoski unosowione</td>
</tr>
<tr>
<td>m, n, ń(n), N</td>
<td>nosowe</td>
</tr>
<tr>
<td>b, d, g</td>
<td>zwarte dźwięczne</td>
</tr>
<tr>
<td>p, t, k</td>
<td>zwarte bezdźwięczne</td>
</tr>
<tr>
<td>dz, drz, dzi</td>
<td>zwartoszczelinowe dźwięczne</td>
</tr>
<tr>
<td>z, zi, rz</td>
<td>szczelinowe dźwięczne</td>
</tr>
<tr>
<td>s, si, sz</td>
<td>szczelinowe bezdźwięczne</td>
</tr>
<tr>
<td>c, ci, cz</td>
<td>zwartoszczelinowe bezdźwięczne</td>
</tr>
<tr>
<td>r</td>
<td>drżące</td>
</tr>
<tr>
<td>l</td>
<td>boczne</td>
</tr>
<tr>
<td>l(l), j</td>
<td>półsamogłoskowe ustne</td>
</tr>
<tr>
<td>w, f</td>
<td>szczelinowe wargowo-zębowe (dzw i bezdzw)</td>
</tr>
<tr>
<td>h</td>
<td>przydech</td>
</tr>
</tbody>
</table>
Analiza skupień segmentów sygnału akustycznego

6.1 Oznaczenia

• $F$ - zbiór wszystkich fonemów z tablicy 5.1: $F = \{a, o, u, e, i, y, a(\_), e(\_), m, n, \_u(\_), N, b, d, g, p, t, k, dz, dz, z, zi, rz, s, si, sz, c, ci, cz, r, l, \_l(\_), j, w, f, h\}$.

• $P_{cep}$ - parametryzacja korpusu: k-wymiarowa przestrzeń Mel-frequency cepstrum coefficients (MFCC).

• $P_{spec}$ - parametryzacja korpusu: k-wymiarowa przestrzeń Mel-frequency spectrum (MFS).

• $v_f^k$ - wektor parametrów w k-wymiarowej parametryzacji MFCC (cepstrum) lub MFS (spectrum) reprezentujący fonem $f \in F$.

• $V_f^k$ - zbiór wszystkich wektorów w k-wymiarowej parametryzacji MFCC (cepstrum) lub MFS (spectrum) reprezentujących fonem $f \in F$.

• $V = V_1 \cup V_2 \ldots \cup V_F$ - zbiór wszystkich analizowanych wektorów.

• $s_n$ - nieprzetworzony sygnał dźwiękowy w formacie WAV. $n \in (0, N)$, gdzie $N$ - ilość próbek w całym nagraniu (długość).

• $U$ - wynik partycjonowania zbioru $V$ taki, że $U(v_i) : v_i \in V \rightarrow p$, gdzie $1 \leq p \leq P$ to numer partycji, a $P$ ilość partycji.

• $C_A$ - wynik klasyfikacji zbioru $A$ taki, że $C_A(a) : a \in A \rightarrow f \in F$

6.1.1 Adjusted Rand Index

*Rand Index* [3] oraz jego modyfikacja *Adjusted Rand Index* [4] są wskaźnikami podobieństwa pomiędzy dwoma podziałami zbioru danych. Każdy z tych podziałów może dzielić zbiór danych na różną ilość podzbiorów (klastrów). Wskaźniki te obliczne są na podstawie informacji o parach elementów, przez co nie są ważne etykiety grup w podziałach. Mając dwa różne podziały $U_1$ i $U_2$ zbioru $V$ ustalmy:

• $a$ - ilość par które należą do tej samej grupy w podziale $U_1$ oraz w podziale $U_2$

• $b$ - ilość par, które należą do tej samej grupy w podziale $U_1$ ale różnych w podziale $U_2$

• $c$ - ilość par, które należą do tej samej grupy w podziale $U_2$ ale do różnych w podziale $U_1$
d - ilość par, które należą do różnych grup w podziale $U_2$ oraz w podziale $U_1$

Jeśli założymy, że podział $U_1$ jest właściwym podziałem zbioru $V$, a $U_2$ próbą jego klasyfikacji, to wtedy: $a$ - *true positive rate*, $b$ - *false negative rate*, $c$ - *false positive rate*, $d$ - *true negative rate*.

Tak więc $a$ i $d$ określają liczbę podobieństw pomiędzy kластrowaniami, a $b$ i $c$ liczbę różnych pomiędzy nimi. Wskaźnik *Rand Index* został zdefiniowany [3] jako

$$RI = \frac{a + d}{a + b + c + d}.$$  \hspace{1cm} (6.1)

Wartość tego wskaźnika należy do przedziału $0 \leq RI \leq 1$, osiągając wartość 1 w przypadku gdy dwa podziały są identyczne. Jednym z problemów wskaźnika $RI$ jest brak stałej wartości (np. zero) dla dwóch losowych podziałów zbioru danych. Kolejną jego wadą jest wzrost jego wartości przy wzroście ilości kластrów (grup) podziału. Z tego powodu stworzono modyfikację: *Adjusted Rand Index* [4], która dla dwóch losowych podziałów przyjmuje wartość 0. ARI zdefiniowano jako:

$$ARI = \frac{{n \choose 2}(a + d) - [(a + b)(a + c) + (c + d)(b + d)]}{{n \choose 2}^2 - [(a + b)(a + c) + (c + d)(b + d)]}.$$ \hspace{1cm} (6.2)

Jego wartość oczekiwana wynosi 0, a maksymalna 1.

6.2 Parametryzacja korpusu

Parametryzacja korpusu została zaadoptowana z HTK Speech Recognition Toolkit. Wstępnie fala dźwiękowa poddawana jest procesowi preemfazy wg wzoru $v'_n = v_n - kv_{n-1}$. Gdzie współczynnik preemfazy $k = 0.97$.

Wektory parametrów $v^k$, gdzie $k$ to wymiarowość wektora, są obliczane na oknie o szerokości $W_{\text{length}} = 25\text{ms}$ przesuwanego o $\Delta_w = 10\text{ms}$. Każdemu takiemu wektorowi przypisany jest ten fonem w którego obszarze leży centralny punkt jego okna $W_{\text{center}} = W_{\text{start}} + \frac{W_{\text{length}}}{2}$. Jeśli punkt centralny okna nie znajduje się pomiędzy początkiem i końcem jednego fonemu, wektor $v^k$ jest odrzucony. Do parametryzacji korpusu użyto okna Hamminga:

$$w(n) = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right),$$

gdzie $N$ to szerokość okna, a $n : 0 \leq n \leq N - 1$, odpowiada kolejnym próbkom sygnału.

Dla każdego okna, któremu przypisano fonem, obliczone jest spektrum sygnału.

Następnie wartości spectrum są podnoszone do kwadratu i za pomocą nachodzących na siebie $k$ trójkątnych okien jest on mapowany na skalę melowską [16], w wyniku czego otrzymujemy $k = 20$ parametrów. Z otrzymanych wartości wyciągany jest logarytm.

Tak dla każdego okna otrzymujemy wektor $V_k$ zawierający współczynniki Mel-frequency spectrum (MFS).

Kolejno, w celu otrzymania współczynników cepstrum (MFCC - Mel-frequency cepstrum coefficients), dla każdego wektora $v_k$ obliczana jest dyskretna transformata cosinusowa (DCT).

W wyniku powyższego procesu otrzymano $\bar{P} = 953267$ wektorów dla obydwu parametryzacji: MFS i MFCC. Każdemu wektorowi $v$ przypisano dodatkowo fonem który reprezentuje, mówcę oraz odległość od środka artykulacji.

Rysunek 6.2. Stosunek odległości pomiędzy parami centroidów skupień fonemów a średnią variancją ich chmur (po lewej); położenie centroidów skupień fonemów (po prawej). Parametryzacja MFCC.
Rysunek 6.3. Stosunek odległości pomiędzy parami centroidów skupień fonemów a średnią wariancją ich chmur (po lewej) oraz położenie centroidów skupień fonemów (po prawej). Parametryzacja MFS.

Na rysunkach 6.2 i 6.3 przedstawiono wzajemne położenie centroidów zbiorów poszczególnych fonemów \( V_x : x \in F \), oraz dla każdej pary zbiorów \( (V_x, V_y) \) stosunek odległości pomiędzy ich centroidami a sumą wariancji ich chmur: \( V_{\text{dist}}(x, y) \), obliczany ze wzoru:

\[
V_{\text{dist}}(x, y) = \frac{|V_x - V_y|}{\text{Var}(V_x) + \text{Var}(V_y)},
\]

gdzie, \( V_x \) jest wektorem, którego współrzędne stanowią średnią ciężkości zbioru \( V_x \), natomiast \( \text{Var}(V_x) \) jest średnią wariancją zbioru \( V_x \). Współczynnik ten daje przybliżoną informację, jak blisko położone są dwa zbiory fonemów: \( V_x \) i \( V_y \), oraz jak bardzo na siebie "zachodzą". Współczynnik jest oparty o wskaźnik serparowalności klas zaprezentowany w [10], rozdz 4.10.

Rysunek 6.4. Znormalizowana odległość euklidesowa (za normę przyjęto wariancję) pomiędzy centroidami skupień fonemów. Parametryzacja MFCC (po lewej) i MFS (po prawej)
6.2 Parametryzacja korpusu

Na rysunku 6.4 pokazano znormalizowaną odległość euklidesową (patrz wzór 6.10) pomiędzy centroidami zbiorów \( V_a \) fonemów. Czynnikiem normalizującym jest wariancja tych zbiorów (inaczej niż we wzorze 6.10, gdzie czynnikiem normalizującym jest \( \max_{0.95} \) - patrz objaśnienie wzoru).

Analizując macierze odległości na rysunkach 6.2, 6.3 i 6.4 można zauważyć, które chmury fonemów są położone blisko siebie i na siebie nachodzą, a które są bardziej od siebie odległe. Zauważalna jest też duża wartość współczynnika \( Vdist \) pomiędzy “drz,dzi,dz,cz” a pozostałymi głoskami. Może to oznaczać duże rozporoszenie zbiorów tych głosek (wskażuje na to także ich duża wariancja).

Rysunek 6.5. Wartość bezwzględna współczynnika skośności (skewness) dla każdego wymiaru poszczególnych fonemów

Na Rysunku 6.5 przedstawiono wartość bezwzględu współczynnika skośności dla każdej głoski i każdego wymiaru. Skośność jest wskaźnikiem asymetrii elementów zbioru wokół jego środka ciężkości (skośność rozkładu normalnego wynosi zero). Obliczana jest ze wzoru:

\[
s = \frac{E(x - \mu)^3}{\sigma^3}
\]  

(6.5)

gdzie \( E(t) \) to wartość oczekiwana zbioru danych \( t \), \( \sigma \) standardowa odchylenie zbioru \( x \), a \( \mu \) średnia wartość zbioru \( x \).

Na Rysunku 6.6 po lewej stronie przedstawiono średnią wariancję zbiorów \( V_f : f \in F \) (każdego fonemu). Po prawej znajduje się macierz wariancji fonemów w każdym z wymiarów znormalizowana do 1 (według największej wariancji w danym wymiarze).
6.3 Analiza metodą k-means

Parametry wejściowe dla metody k-means:

- \( k \) - ilość klaстрów
- \( S_i : 1 \leq i \leq k \) - punkt startowy dla każdego klastra
- \( I \) - liczba iteracji

Metodę k-means wykonano w następujących konfiguracjach:

1. \( P_{cep}^{20} \) i \( P_{spec}^{20} \), \( k = \mathcal{F} = 37 \), \( S_i = \mathbf{V}_i \): (środek ciężkości zbioru wektorów \( \mathbf{V}_i \) jest początkowym punktem klastra \( S_i \))
2. \( P_{cep}^{20} \) i \( P_{spec}^{20} \), \( k = 37 \), dwa etapy; pierwszy: \( I_1 = 100 \), \( S_i \) - inicjalizacja losowa, \( \mathbf{V}' \) - losowo wybrana próbka o rozmiarze \( \mathbf{V}' = 0.1\mathbf{V} \); drugi etap: \( I_2 = 40 \), \( S_2 \) - centroidy klastrów otrzymane po pierwszym etapie
3. \( P_{cep}^{20} \), \( k \in 10 \ldots 40 \), punkty startowe \( S \) - wybierane losowo
Analiza skupień metodą k-means dała bardzo zbliżone wyniki dla parametryzacji MFS(спектrum) i MFCC(cepstrum). Większość wyników przedstawiono tylko dla parametryzacji MFCC.

6.3.1 ARI dla wyników kластrowania metodą k-means

Rysunek 6.7. Współczynniki poprawności kластrowania k-means w zależności od $k$. $P^{20}_{cep}, I_1 = 100, I_2 = 40, 10 \leq k \leq 40$

Na rysunku 6.7 przedstawiono cztery różne indeksy poprawności kластrowania w zależności od liczby klastrów $k$ ustawionej jako parametr wejściowy dla metody k-means. Poza zdefiniowanymi powyżej RI oraz ARI, wykreślone zostały: Mirkin Index zdefiniowany jako:

$$MI = \frac{b + c}{a + b + c + d},$$  \hspace{1cm} (6.6)

c o odpowiednia ilości różnic pomiędzy dwoma podziałami; Hubert Index zdefiniowany jako:

$$HI = RI - MI$$  \hspace{1cm} (6.7)

Analizując wartości wyznaczonych wskaźników dla wyników metody k-means przy różnej ilości klastrów można stwierdzić, że wynik kластrowania nie oddaje rzeczywistego podziału zbioru na różne fonemy. Średnia wartość $\overline{AIR} = 0.14$ (oraz niewielka jego wariancja dla różnej ilości klastrów) wskazuje, że ilość klastrów nie wpływa na wynik kластrowania skupień fonemów w przestrzeniach MFCC i MFS. Współczynniki $RI$ i $MI$ nie są w tym przypadku miarodajne. Powodem tego jest stosunkowo duża ilość podobieństw pomiędzy klastrami. Wartość współczynnika $d$ (true negative rate, patrz definicja ARI), jest znacznie większa od pozostałych $a, b, c$. Kontrastując niską wartość $AIR$ z wysoką wartością wskaźnika $RI$ można także wyciągnąć wniosek, że wiele par elementów zbioru, które reprezentują ten sam fonem zostało przyporządkowanych do tej samej grupy, jednak elementy tych zbiorów zostały podzielona pomiędzy kilka partycji (efekt ten zobrazowany jest na Rysunku 6.8).
Rysunek 6.8. Rzut wyników klastrowania fonemu 'a' na dwa pierwsze wymiary reprezentacji MFCC; każda klasa została oznaczona innym kolorem. (po lewej). Histogram klas do jakich został przyporządkowany ten fonem (po prawej). Parametry: \( P_{cp} = 20 \), \( I = 40 \), \( k = 37 \)

Na rysunku 6.8 przedstawiono rezultat wyniku klastrowania na fonemie “a” oraz histogram przestawiający ilość wektoró w reprezentujących ten fonem przyporządkowanych do każdej z klas. Początkowe punkty centralne klastrów (przed pierwszą iteracją metody k-means) zostały ustawione jako środki ciężkości zbiór fonemów \( \{V_f : f \in F\} \). W tym przypadku klastrów o numerze 1 ustawiono środek będący środkiem ciężkości chmury fonemu “a”. Jak widać po 40 iteracjach tylko 27.1% punktów tego zbioru zostało przypisanych do odpowiedniego zbioru.

6.3.2 Common Cluster Factor dla k-means

Rysunek 6.9 przedstawia macierz współczynnika Common Cluster Factor (CCF). Współczynnik CCF dla pary fonemów \( V_x \) i \( V_y \) oraz wyniku klastrowania \( R \) obliczany jest następująco:

\[
CCF(V_x, V_y) = \sum_{1 \leq i < N} \min\left( \frac{V_x \cap R_i}{V_x}, \frac{V_y \cap R_i}{V_y} \right).
\]  \hspace{1cm} (6.8)

Współczynnik CCF przyjmuje wartości z przedziału 0 \( \leq CCF \leq 1 \): \( CCF(V_x, V_y) = 1 \) gdy elementy zbiorów \( V_x \) i \( V_y \) są identycznie rozłożone pomiędzy klastry \( R_i \), i \( CCF(V_x, V_y) = 0 \) gdy żadna para ich elementów \( (v_x, v_y) \) nie należy do tego samego klastra.

Macierz CCF pokazuje zbiory których par fonemów nachodzą na siebie co powoduje, że ich element są trudne do rozróżnienia. Analizując macierz współczynników CCF widać, że największe wartości współczynnika (kolor czerwony, żółty i pomarańczowy) skupione są przy przekątnej macierzy. Jest to wynik takiego uporządkowania fonemów, w którym im bliżej siebie położone są dwa fonemy tym są bardziej podobne: mniej cech akustycznych je rozróżnia, lub cechy te są słabo różnice. Dla przykładu: dźwięczność/bezdźwięczność jest cechami wykluczającymi się, dobrze rozróżniającymi głoski (np.: “p” i “b”), natomiast dwuwargowość i zębów cechami słabiej rozróżniającymi głoski (np.: “b” i “d”). Przykład podany jest na...
6.3 Analiza metodą k-means

Rysunek 6.9. Współczynnik Common-Cluster Factor dla każdej pary fonemów. $P_{cep}^{20}$, $I = 40$, $k = 37$

Rysunek 6.10. Ułożenie języka mowy przy produkcji głoski przedniojęzykowo-zębowej (po lewej) oraz wargowej (w środku) oraz nosowej (po prawej).

Rysunek 6.11 przedstawiono jak zmieniają się macierze współczynników CCF w zależności od $k$. Jak widać przy zwiększaniu liczby klastrów zwiększa się rozróżnialność par...
Rysunek 6.11. Macierze współczynnika Common Cluster Factor $k$ o wartościach kolejno rzędami: $10, 13, 17, 21, 25, 29, 33, 37, 40$. Parametry: $P_{cep}^{20}, I_1 = 100, I_2 = 40$

fonemów położonych dalej od diagonali macierzy (tj. fonemów mniej podobny do siebie pod względem cech akustycznych).

6.3.3 Wnioski z analizy metodą k-means

Wyniki badań pokazały, że metoda klastryzacji k-means nie daje ciekawych rezultatów dla zbioru danych $P_{cep}^{20}$. Dokładniejsze wyniki można otrzymać klasyfikując fonemy metodą najbliższych środków ciężkości (patrz następny rozdział). Przeprowadzona również próbę klastrowania metodą fuzzy c-means, jednak bez zadowalającego rezultatu. Wnioski płynące z analizy macierzy współczynników CCF (równanie 6.8) oraz Vdist (równanie 6.4) przedstawiono poniżej.

Rysunek 6.12. Średnia wartość współczynnika CCF dla poszczególnych głosek

- Spółgłoski tylne (a,o,u) są całkiem dobrze odseparowane od spółgłosek przednich (e, i, y).
6.4 Klasyfikacja metodą najbliższych środków ciężkości

- “i” jest wyjątkowo dobrze separowalna wśród spółgłosek, spółgłosek unosowionych, oraz samogłosek nosowych (m,n,ń,N), a także pozostałych głosek.
- Następujące grupy spółgłosek są słabo separowalne wewnątrz grup: (p,t,k), (b,d,g), (m,n,ń,N), (dzi,drz), (dz,z), (s,c), (ci,cz), (sz,cz), (f,h), (r,c), (l,i), (j,i)
- spółgłoski (r,l,ł) są słabo separowalne z samogłoskami (poza i), ich nasalizacjami (ą,ę) oraz spółgłoskami nosowymi
- głoska “w” jest słabo separowalna od większości głosek
- głoski (p,t,k) są słabo separowalne z głoskami (f,h). Możliwe, że jest to spowodowane aspiracją występującą przy wypowiadaniu wszystkich tych głosek.

6.4 Klasyfikacja metodą najbliższych środków ciężkości

W następującym rozdziale opisano wyniki analizy skupień metodą Nearest Mean (NM) zdefiniowanej następująco:

1. wyznaczyć środki ciężkości dla każdego zbioru \( V_x \).
   \[ C_x = \overline{V_x} \]
2. każdy fonem \( a \in V \) dodać do takiego zbioru \( R_x \), którego środek \( C_x \) znajduje się najbliżej niego w zadanej metryce:
   \[ a \in R_x : \min_{x \in F}(\text{dist}(a, C_x)) \]

   gdzie \( \text{dist} \) to zadana miara odległości.

Wszystkie badania zostały przeprowadzone dla niezależnych mówców. Przeprowadzona \( m \) testów, gdzie \( m \) równie jest liczbie różnych mówców. W każdym teście środki ciężkości oraz inne parametry (jak np.: wariancja dla każdego fonemu w każdym z wymiarów) wyznaczane były za pomocą danych pochodzących od wszystkich pozostałych mówców. Wynik testów jest uśrednionym wynikiem z wszystkich \( m \) testów.

Pośrednim wynikiem metody NM są wektory \( PC_a : a \in V \) odległości pomiędzy każdym fonemem a i środkiem ciężkości każdej z klas fonemów \( C_x : x \in F \). Wektor \( SPC_a \) zawiera indeksy posortowanego rosnąco wektora \( PC_a \), tzn: \( \forall_{1<i<P\bar{C}_a} PC_a(SPC_a(i)) <= PC_a(SPC_a(i+1)) \). W poniższej analizie przedstawiono następujące wyniki:

- Macierz współczynników RAF (opis w następnym podrozdziale) dla wyników klasyfikacji
- Wartość oczekiwana pozycji poprawnej klasy fonemu \( a \) w wektorze \( SPC_a : a \in V_f \) dla każdej klasy fonemów.
- Wartość oczekiwana najmniejszej wartości z wektora \( \min(PC_a) : a \in V_f \) (odległość do środka ciężkości najbliższej klasy) dla każdej klasy fonemów. (zielony na wykresie)
- Wartość oczekiwana odległości do środka ciężkości poprawnej klasy: \( PC_a(f) : a \in V_f \) dla każdej klasy fonemów. (czerwony na wykresie)
Analiza skupień segmentów sygnału akustycznego
• Różnica dwóch poprzednich wartości: \( \min(\text{PC}_a) - \text{PC}_a(f) : a \in V_f \) (żółty na wykresie)
• Histogram pozycji poprawnej klasy w wektorze \( \text{PC}_a \).

Analizując wyniki dla różnych miar odległości można stwierdzić, która miara najlepiej nadaje się do analizy skupień segmentów sygnału akustycznego w parametryzacji MFCC. Im większe wartości na elementach diagонаli macierzy współczynnika RAF tym więcej fonemów znajduje się w danej metryce bliżej środka ciężkości skupienia, do którego w rzeczywistości należy, na podstawie etykietowania posegmentowanych nagrań wypowiedzi. Ponieważ kolejność głosek w macierzy jest wybrana tak, by głoski do siebie podobne były stosunkowo blisko siebie, większe wartości współczynnika RAF będą pojawiać w większości się bliżej diagonalii.

6.4.1 Współczynnik Recognized-As Factor (RAF)

W celu zobrazowania i ułatwienia analizy skupień i wyników klasyfikacji fonemów wprowadzono współczynnik RAF zdefiniowany jako:

\[
\text{RAF}(a, b) = \frac{|V_a \cap R_b|}{V_a}
\]

gdzie \( a, b \in F \) to oznaczenia klas fonemów, \( V_a \) to zbiór wszystkich fonemów klasy \( a \); \( R_b \) to zbiór fonemów sklasyfikowanych jako fonem \( b \). Ty samym wartość \( \text{RAF}(a, b) \) mówi jaka część wszystkich fonemów \( a \) została sklasyfikowana jako fonem \( b \). Macierz RAF można interpretować również w taki sposób, że wiersz \( a \) pokazuje jako jakie fonemy został sklasyfikowany fonem \( a \), natomiast kolumna \( a \) pokazuje jakie fonemy zostały sklasyfikowane jako fonem \( a \).

6.4.2 Odległość euklidesowa (EUC)

Na rysunkach 6.13 i 6.14 przedstawiono wyniki klasyfikacji metodą NM z użyciem miary euklidesowej: \( \text{dist}(a, b) \) - jest równy odległości euklidesowej, obliczonej ze wzoru:

\[
\text{dist}(a, C_x) = \sqrt{\sum_{0 \leq i \leq N} (a(i) - C_x(i))^2}
\]

gdzie \( a(i) \) - i-ta współrzędna punktu \( a \), \( C_x(i) \) - i-ta współrzędna środka ciężkości fonemów klasy \( x \).

Obserwacje

Średnia wartość na diagonalii macierzy współczynnika RAF wynosi \( \overline{\text{RAF}(a, a)} = 0.242 \) co oznacza, że średnio 24% fonemów jest przyporządkowywanych do prawidłowej klasy. Średnia pozycja środka ciężkości prawidłowej klasy w wektorze \( SPC \) dla każdej głoski jest przedstawiona na Rysunku 6.14. Średnia wartość liczona dla wszystkich głosek wynosi: 7.26. Po prawej stronie znajduje się histogram pozycji rzeczywistej głoski w wektorze \( SPC \).
6.4 Klasyfikacja metodą najbliższych środków ciężkości

Rysunek 6.13. RAF dla NM z użyciem odległości euklidesowej

Rysunek 6.14. Po lewej: niebieski - średnia pozycja prawidłowej klasy w wektorze SPC; czerwony - średnia odległość od centrum prawidłowej klasy; zielony - odległość do najbliższej klasy; żółty - różnica poprzednich dwóch odległości. Po prawej: histogram pozycji prawidłowej klasy w wektorze SPC.
**Samogłoski**

Z grupy samogłosek najlepiej rozpoznawalne są “e” (63%) oraz “a” (53%). Trzeba jednak zauważyć, że 27% głosek “o” jest rozpoznawanych jako “a” oraz 39% głosek “y” i 32% głosek “ę” jest rozpoznawanych jako “e”. Dobrze rozróżnialną samogłoską (rzadko myloną z pozostałymi) jest “i”, jest ona jednak często mylona ze spółgłoską “j”, co jest prawdopodobnie spowodowane podobieństwem w brzmieniu tych głosek. Około 5% głosek “i” jest klasyfikowanych jako spółgłoski szczelinowe-dźwięczne zmiękczone (dzi,zi) oraz spółgłoska nosowa zmękczona “ń” (biorąc pod uwagę dobrą rozróżnialność pozostałych samogłosek od tych spółgłosek jest to dużo).

Samogłoski unoszone są słabo rozpoznawalne: “ą” jest najczęściej rozpoznawana jako “o”, natomiast “ę” jako “e”.

**Spółgłoski nosowe**

Spółgłoski nosowe są rozpoznawalne w większości jako “m” (pionowa linia na rysunku 6.1 od pola (m,m) w dół).

**Spółgłoski zwarte**

Spółgłoski zwarte-dźwięczne (b,d,g) są najczęściej rozpoznawane jako “b”, natomiast bezdźwięczne (p,t,k) jako “p”.

**Spółgłoski zwarto-szczelinowe**

Spółgłoski zwarto-szczelinowe dźwięczne są bardzo słabo rozpoznawalne: maksymalne wartości w rzędach (dz,drz,dzi) leżą poza diagonalą i są to kolejno kolumny (z,rz,z). Spółgłoski zwarto-szczelinowe bezdźwięczne (c,ci,cz) są również słabo rozpoznawalne. Około 24% z nich jest klasyfikowane kolejno jako ich ’niezwarte odpowiedniki’: (s,si,sz). Ponadto 20% głosek “cz” jest klasyfikowanych jako “p” i “sz”.

**Spółgłoski szczelinowe**

Spółgłoski szczelinowe dźwięczne (z,zi,rz) są dobrze rozpoznawalne. 20% głosek “zi” jest klasyfikowane w 20% jako “rz”. Spółgłoski szczelinowe bezdźwięczne (s,si,sz) są również dobrze rozpoznawalne: współczynnik RAF na diagonalach waha się od 37% do 47%. Jednak 23% głosek “si” jest klasyfikowanych jako “sz”.

**Pozostałe spółgłoski**

Spółgłoski (r,l,j,w,f) są słabo rozpoznawalne (od 12% do 15%). Ilość poprawnie rozpoznanych spółgłosek “l” i “h” to odpowiednio: 26% i 30%. 19% głosek “j” jest klasyfikowanych jako...
“e” powodem tego może być stosunkowo duża liczba wystąpień dźwięku “je” w korpusie (patrz rysunek 5.4).

### 6.4 Klasyfikacja metodą najbliższych środków ciężkości

#### 6.4.3 Znormalizowana odległość euklidesowa (NED)

Na rysunkach 6.15 i 6.16 przedstawiono wyniki klasyfikacji metodą NM z użyciem miary euklidesowej: $\text{dist}(a, b)$ to znormalizowana odległość euklidesowa obliczana ze wzoru:

$$
\text{dist}(a, C_x) = \sqrt{\sum_{0 \leq i \leq N} \left( \frac{a(i) - C_x(i)}{\max_{0.95}(C_x, i)} \right)^2}
$$

(6.10)

gdzie $a(i)$ - i-ta współrzędna punktu $a$, $C_x(i)$ - i-ta współrzędna środka ciężkości fonemów klasy $x$, norma: $\max_{0.95}(C_x, i)$ - maksymalna odległość od środka ciężkości po i-tej współrzędnej z 95% punktów położonych najbliżej środka ciężkości $x$.

Podobne wyniki otrzymano normalizując odległość euklidesową w wymiarze $k$ wariancją danej klasy w tym wymiarze: $\text{Var}(C_x, k)$.

### Obserwacje

Porównując macierze RAF dla EUC i NED zauważamy wiele zmian: wartości na diagonali są wyższe dla odległości NED: średnia wartość wynosi $\overline{\text{RAF}}(a, a) = 0.35$, co oznacza wzrost o 10 punktów procentowych. Występuje znacznie mniej artefaktów jak dla trójkąt fonemów (p,t,k) przy odległości EUD. Samogłoski są o lepiej rozpoznawalne. Średnia pozycja środka ciężkości prawidłowej klasy w wektorze $\text{SPC}$ dla każdej głoski jest przedstawiona na Rysunku 6.16. Średnia wartość liczona dla wszystkich głosek wynosi: 4.64.
6 Analiza skupień segmentów sygnału akustycznego

Rysunek 6.15. RAF dla NM z użyciem znormalizowanej odległości euklidesowej

Rysunek 6.16. Po lewej: niebieski - średnia pozycja prawidłowej klasy w wektorze SPC; czerwony - średnia odległość od centrum prawidłowej klasy; zielony - odległość do najbliższej klasy; żółty - różnica poprzednich dwóch odległości. Po prawej: histogram pozycji prawidłowej klasy w wektorze SPC.

Rysunek 6.17. Wartość RAF dla elementów na diagonali macierzy.
Rozpoznawalność samogłosek jest bardzo wysoka i osiąga 66% dla głosek “a” oraz “i”, przy czym 16% głosek “i” jest klasyfikowane jako “j”. Najniższą wartość przyjmuje “y”: 44%. Samogłoska nosowa “ą” jest mylona z “o” (19%), a “ę” z “e” (21%). Pary głosek o podobnym brzmieniu i wyższym współczynniku RAF to: (dz,z), (l,u), (j,i),(t,k),(k,h), (r,l), (c,s), (cz,sz). Stosunkowo często też głoska “n” jest rozpoznawana jako głoska “l”. Może to wynikać z podobieństwa w ułożeniu niektórych narządów mowy przy produkcji tych głosek. Głoski o najniższej rozpoznawalności to: (t,drz,w,f).

6.4.4 Porównanie z k-means

Wynik klasterowania metodą k-means dla $P_{cep}^{20}, I = 40, k = 37$ (punkty startowe klasterów ustawiono na środki cieczą zbiór fonemów) porównano z wynikami klasyfikacji NM.

Patrząc na macierze widzimy, że analiza metodą k-means jest wyraźnie mniej efektywna niż analiza metodą Nearest Means. Świadczą o tym większe wartości współczynnika CCF dla elementów dalej położonych od diagonali. Dodatkowo metoda NM działa o wiele szybciej niż k-means.

6.5 Badanie separowalności głosek metodą LDA

W tym rozdziale przedstawiono wyniki analizy skupień metodą LDA.

Liniowa analiza dyskryminacyjna (Linear Discriminant Analysis) jest używana do znalezienia kombinacji liniowej cech, która charakteryzuje lub separuje dwie lub więcej klas obiektów.
6 Analiza skupień segmentów sygnału akustycznego

W poniższych badaniach użyto liniowego dyskryminatora Fisher’a do separacji elementów dwóch różnych klas głosek \( V_a \) i \( V_b \): \( a, b \in F \).

6.5.1 Liniowy dyskriminator Fisher’a

Obliczanie liniowego klasyfikatora według [11]. Dla dwóch klas głosek \( V_a \) i \( V_b \) w \( R^n \), gdzie \( a, b \in F \) ich rozdzielność (ang. class separability) w kierunku \( w \) zdefiniowana jest jako

\[
F(w) = \frac{w S_B w}{w S_W w} 
\]

(6.11)

gdzie \( S_B \) to macierz rozporoszenia pomiędzy klasami (ang. between-class scatter matrix):

\[
S_B = (\mu_a - \mu_b)(\mu_a - \mu_b)^T, \quad \mu_f = \frac{1}{|V_f|} \sum_{i} x_i 
\]

a \( S_W \) to macierz rozproszenia wewnątrz klasy (ang. within-class scatter matrix)

\[
S_W = S_1 + S_2, \quad S_f = \sum_{1 \leq i \in |V_f|} (x_i - \mu_f)(x_i - \mu_f)^T 
\]

dla \( \mu_f \) będącego środkiem cieżkości zbioru \( V_f \), a \( |V_f| \) moczą zbioru \( V_f \).

Rozwiązaniem problemu, czyli znalezieniem wektora \( w \), w kierunku którego dwa zbiory są najlepiej separowalne (\( w = \arg \max_w F(w') \)) jest

\[
w = S_W^{-1}(\mu_a - \mu_b) 
\]

Poza transformacją liniową \( w \) dla każdej pary obliczane jest także odchylenie \( d \) (ang. bias) spełniające równanie:

\[
w \mu_a + d_{ab} = -(w \mu_b + d_{ab}) 
\]

Klasyfikacja z użyciem \( w_{ab} \) i \( d_{ab} \) przebiega następująco:

1. dla wejściowego wektora \( x \in V_a \cup V_b \) obliczana jest wartość funkcji:

\[
f(x) = w_{ab} x + d_{ab} 
\]

(6.12)

2. jeśli \( f(x) \geq 0 \), to wtedy \( x \) klasyfikowane jest jako element zbioru \( V_a \)

3. w przypadku gdy \( f(x) < 0 \), \( x \) klasyfikowane jest jako element zbioru \( V_b \)

6.5.2 Rozkład gęstości głosek

Na Rysunku 6.19 pokazano rozkład gęstości fonemów “d”, “drz”, oraz “rz” w 20 wymiarach parametryzacji MFCC. Jak widać dla niektórych wymiarów pary tych głosek są bardzo słabo separowalne (np.: dla “d” i “rz” wymiar 6.), a dla niektórych lepiej (np.: dla “d” i “rz” wymiar
6.5 Badanie separowalności głosek metodą LDA

Poza tym w wymiarze 1. i 2. głoska “drz” (kolor zielony) posiada dwa zagęszczenia skupień, które pokrywają się z zagęszczeniami pozostałych dwóch głosek (d i rz). To rozciągnięcie elementów zbioru jest przyczyną dużej jego wariancji, która objawia się na rysunkach: 6.2, 6.3 i 6.6 oraz powodem słabej separowalności od głosek “d” i “rz”.


6.5.3 Rozróżnialność par głosek

Metodę LDA opisaną powyżej można wykorzystać do klasyfikacji i określenia rozdzielności par zbiorów głosek \( V_f : f \in F \). Przeprowadzono następującą analizę:

Dane wejściowe: \( V_f : f \in F \), Dane wyjściowe: macierz \( RAF(a, b) \)

**Algorytm** Dla każdej pary głosek \( a, b \in F \):

1. obliczyć \( w_{ab} \) oraz \( d_{ab} \).
2. utworzyć zbiór \( V_{ab} = V_a \cup V_b \)
3. przeprowadzić klasyfikację każdego elementu zbioru \( V_{ab} \) z użyciem \( w_{ab} \) oraz \( d_{ab} \). Wynikiem są dwie klasyfikacje: \( R_a \) i \( R_b \) (zawierające kolejno elementy sklasyfikowane jako \( a \) i \( b \)).
4. obliczyć współczynniki \( RAF \) dla par \( (a, b) \) oraz \( (b, a) \) (wzór 6.9): \( RAF(a, b) \) jest ilorazem ilości elementów zbioru \( V_a \) sklasyfikowanych jako \( b \) do mocy zbioru \( V_a \).

**Rysunek 6.21.** Macierz RAF dla par fonemów rozróżnianych metodą LDA.

Na Rysunku 6.21 przedstawiony wynik analizy w postaci macierzy RAF.

**Analiza macierzy RAF dla separacji par metodą LDA**

**Średnia wartość kolumn i wierszy**

Na wykresach 6.22 przedstawiono średnie wartości kolumn i wierszy na poszczególnych głosek. Większa średnia wartość w kolumnie macierzy odpowiadającej danej głosce onazcza, że
6.5 Badanie separowalności głosek metodą LDA

Rysunek 6.22. Średnia wartość w kolumnach (wykres górny) i rzędach (wykres dolny) macierzy RAF dla par fonemów (LDA)

więcej innych głosek jest klasyfikowanych jako dana głoska. Jednym z powodów może być to, że dana głoska ma relatywnie niską wariancję (jest bardziej od innych zagęszczona dla wartości oczekiwanej swojego rozkładu) co powoduje, że w "wygrywa" w obszarach w których "zachodzi się" z innymi głosekami. Kolejnym powodem może być to głoska ma relatywnie wysoką wariancję (rozciąga się na duży obszar), przez co "zachodzi się" z zbiorami wielu innych głosek "przejmując" niewielką część ich elementów. Przykładem pierwszego przypadku mogą być głoski nosowe (niska wariancja), a drugiego "drz".

Wysoka średnia wartość w wierszu macierzy danej głoski mówi, jak często głoska ta "przegrywa" z innymi głosekami (jest klasyfikowana jako inne głoski). Można zauważyć zbieżność faktów, że głoski o dużej wariancji mają tę wysoką (c, ci, cz, dz, drz, dz, zwarto-wybuchowe).

Niesymetryczność macierzy

Macierz RAF jest niesymetryczna co oznacza, że dla dwóch głosek słabo od siebie rozróżnialnych (dla przykładu d oraz drz) jedna z nich (w tym przypadku drz) jest częściej klasyfikowana jako ta druga (d) niż na odwrót. tego jest rozkład chmur każdej z głosek w analizowanej przestrzeni (MFCC). Jak widać na rysunku 6.19 rozkład elementów $V_j$ jest w większości wymiarów (w szczególności w 2., 8., 10.) o wiele bardziej gęsty (świadczy o tym mała szerokość
oraz duża wysokość "szczytów" obrazujących rozkład gęstości). Klasyfikacja metodą LDA punktem \( v \in V \) "wpadającym w obszar" dużego zagęszczenia glosek \( d \) będzie przypisywała etykietę \( d \) (\( v \in R_d \)). Dlatego też gloski posiadające dużą wariancję: \( \text{dz, drz, dzi, c, ci, cz} \) są często mylone z pozostałącznymi.

### gloski \( U \) i \( L \)
Głoski \( U \) i \( L \) są trudno separowalne (około 32% elementów każdej z nich jest klasyfikowane niepoprawnie jako ta druga z pary). Ich kolumny i rzędy macierzy RAF są do siebie bardzo podobne. Poza tym ze wszystkich samogłosek (nieunosowionych) "\( U \)" jest najczęściej rozpoznawana jako samogłoska nosowa: średnie RAF(u,samogłoska nosowa) wynosi 10% (kolejna taką samogłoską jest "Y")

### gloski \( I \) i \( J \)
Średnia wartość współczynnika RAF dla tej pary wynosi: \( \text{RAF}(i,j) = 29\% \), co oznacza, że gloski są bardzo słabo separowalne (z błędem na poziomie 29%). Zauważalne jest też podobieństwo ich kolumn i rzędów macierzy RAF, poza faktem, że "\( j \)" jest (w przeciwieństwie do "\( i \)") słabo rozpoznawalne z samogłosek "e" i "y" (około 14%) - może to być spowodowane częstym występowaniem w korpusie samogłoski "e" po glosce "j" oraz podobieństwem samogłosek "e" i "y" (najszlachta separowalność dla pary samogłosek).

### gloska \( A \) i \( E \)
Głoska \( A \) jest często klasyfikowana jako następujące gloski: o, u, a, e, n, N, l. Przyczyną tego może być fakt, że wymowa gloski "a" jest często upraszczana do wymowy: "o-n(N,m)" oraz "o-u(L)". Głoska \( E \) jest często klasyfikowana jako samogłoska (z pominięciem "i") lub samogłoska nosowa, a także "r".

### spółgłoski nosowe \( \{m, n, Ń, N\} \)
Samogłoski nosowe są słabo separowalne pomiędzy sobą (średnia wartość współczynnika RAF pomiędzy nimi przekracza 20%). Ponadto, 11% glosek "ń" jest klasyfikowane jako "i", co biorąc pod uwagę, że samogłoski nosowe, rzadko są klasyfikowane jako samogłoski jest wysokim wynikiem. Również dyftong "N" jest dosyć często klasyfikowany jako samogłoska, kolejno (u - 14%, o 10%, a 7%). Ciekawym może wydawać się fakt, że samogłoski nosowe są często mylone z samogłoskami zwartymi dźwięcznymi: \( \text{b, d, g} \) i na odwrot.

### spółgłoski zwarte dźwięczne \( \{b, d, g\} \)
Średnia wartość RAF pomiędzy nimi wynosi około 25% co oznacza, że są słabo separowalne. Ponad 15% tych spółgłosek jest mylone z samogłoskami nosowymi, a poniżej 10% z samogło-
6.5 Badanie separowalności głosek metodą LDA

skami - może to być przyczyną faktu, że po samogłoskach wybuchowych najczęściej występują samogłoski (patrz 5.4).

spółgłoski zwarte bezdźwięczne \(p, t, k\)

Średnia wartość RAF osiąga 30%, co czyni tę grupę najslabiej rozróżnialną (wewnątrz grupy) ze wszystkich głosek występujących w korpusie. Są często mylone z głoskami \(w, f, h\); szczególnie para \((k, h)\) około 25%. Poza tym są mylone na poziomie \(\sim 14\%\) z samogłoskami zwartymi dźwięcznymi i na poziomie \(\sim 10\%\) z nosowymi.

spółgłoski zwarto-szczelinowe dźwięczne \(dz, drz, dzi\)

Głoski te są słabo separowalne z następującymi grupami: głoski nosowe \((m, n, ni)\) - \(\sim 20\%\), głoski zwarte dźwięczne \((b, d, g)\) - \(\sim 28\%\). Separowalność między nimi samymi wynosi około 20%. Można zauważyć także bardzo słabą rozpoznawalność tych głosek z głoską “d” (30%), oraz odpowiadającymi im głoskami szczelinowymi dźwięcznymi: \((dz, z)\) 39\%, \((drz, rz)\) 34\%, \((dzi, zi)\) 26%; podczas gdy rozpoznawalność w drugą stronę jest dużo lepsza i wynosi średnio 15\% (patrz wyżej: paragraf niesymetryczność macierzy).

spółgłoski szczelinowe \((z, ź, rz, s, si, sz)\)

Spógłoski te są dobrze rozróżnialne od pozostały głosek, a także między sobą. Dwoma wyjątkami są: para \(RAF(s, c) = 26\%\) oraz pary \((si, ci)\) oraz \((si, ci)\) - \(\sim 18\%\)

spółgłoski zwarto-szczelinowe bezdźwięczne \(c, ci, cz\)

Spógłoski te są słabo separowalne pomiędzy sobą oraz od pozostałych grup głosek. Poniżej podano średnie wartości współczynnika RAF:

\[
\begin{align*}
\overline{RAF}(c – ci – cz, p – t – k) &= 25\% \\
\overline{RAF}(c – ci – cz, b – d – g) &= 11\% \\
\overline{RAF}(c – ci – cz, m – n – ni) &= 9\% \\
\overline{RAF}(ci, cz) &= 27\% \\
\overline{RAF}(c, s) &= 37\% \\
\overline{RAF}(ci, si) &= 31\% \\
\overline{RAF}(cz, sz) &= 29\% \\
\overline{RAF}(c, ci – cz) &= 16\% \\
\end{align*}
\]

spółgłoski \(r, l\)

Spógłoski te słabo separowalne z samogłoskami (poza “i”), spółgłoskami nosowymi (szczególnie “l”) i niewiele lepiej z głoskami \((d, g)\).
Spółgłoska “w” należy do głosek z którymi są często mylone pozostałe głoski (duża średnia wartość w kolumnie macierzy RAF). Jest ona słabo rozpoznawalna z głoskami zwartymi dźwięcznymi oraz trochę lepiej z nosowymi i samogłoskami. Głoski “f” i “h” są słabo rozróżnialne z głoskami zwartymi bezdźwięcznymi (co może być spowodowane aspiracją następującą po tych głoskach). “h” jest głoską często myloną z każdą inną.

Pary najsłabiej rozróżnialne

Zbiory głosek słabo ze sobą rozróżnialnych metodą LDA. Po strzałce podano głoskę, która jest częściej rozpoznawana oraz wynik ilorazu \( \frac{RAF_{LDA}(b,a)}{RAF_{LDA}(a,b)} \), gdzie \( a \) to głoska częściej rozpoznawana z pary głosek \( (a,b) \).

- \((o,ą) \rightarrow o \ (1.41)\)
- \((u,ł) \rightarrow ł \ (1.06)\)
- \((e,e') \rightarrow e \ (2.23)\)
- \((i,j) \rightarrow i \ (1.25)\)
- \((y,l) \rightarrow y \ (1.15)\)
- \((m,ni) \rightarrow m \ (1.26)\)
- \((n,l) \rightarrow n \ (1.09)\)
- \((b,d) \rightarrow b \ (1.02)\)
- \((b,g) \rightarrow b \ (1.23)\)
- \((d,g) \rightarrow d \ (1.27)\)
- \((d,drz) \rightarrow d \ (2.38)\)
- \((g,w) \rightarrow w \ (1.05)\)
- \((p,t) \rightarrow p \ (1.15)\)
- \((p,k) \rightarrow p \ (1.31)\)
- \((t,k) \rightarrow t \ (1.28)\)
- \((t,h) \rightarrow h \ (1.12)\)
- \((k,h) \rightarrow h \ (1.23)\)
- \((dz,z) \rightarrow z \ (1.66)\)
- \((drz,rz) \rightarrow rz \ (3.00)\)
- \((dzi,zi) \rightarrow zi \ (1.94)\)
- \((s,c) \rightarrow s \ (1.46)\)
- \((si,ci) \rightarrow si \ (1.73)\)
- \((sz,cz) \rightarrow sz \ (1.66)\)
- \((r,l) \rightarrow r \ (1.01)\)
- \((f,h) \rightarrow f \ (1.34)\)
### 6.5.4 Separowalność głosek na podstawie metod NM i LDA

![Rysunek 6.23. Separowalność fonemów na podstawie metod NM i LDA](image)

Na podstawie wyników metod NM i LDA obliczono separowalność każdej z głosek od pozostałych. Im wyższa separowalność fonemu $f \in F$, tym zbiór $V_f$ jest lepiej oddzielony od pozostałych, co oznacza, że głoska ta jest rzadko rozpoznawana jako inna głoska, oraz inne głoski są rzadko rozpoznawane jako głoska $f \in F$. Rozpoznawalność fonemów obliczona ze wzoru:

$$\text{Sep}(f) = \frac{\text{RAF}_{\text{NM}}(f,f)}{\sum_i \text{RAF}_{\text{LDA}}(f,i) + \sum_i \text{RAF}_{\text{LDA}}(i,f)}$$

gdzie licznik to część (wyrażona w procentach) zbioru $V_f$, która została poprawnie sklasyfikowana metodą NM; pierwszy składnik sumy z mianownika szacuje jak często głoska ta jest klasyfikowana jako jakaś inna, a drugi składniej tej sumy szacuje jak inne głoski są klasyfikowane jako głoska $f$.

### 6.5.5 Poprawa wyników klasyfikacji NM za pomocą metody LDA

Poniżej zaprezentowano wyniki hybrydowej klasyfikacji fonemów: $NM_{\text{NED}} + LDA_k$. Klasyfikacja taka przebiega następująco dla każdego $v \in V$:

1. oblicz wektor $SPC$ zawierający indeksy kolejnych najbliższych środków ciężkości klas fonemów (w znormalizowanej odległości euklidesowej (NED))
2. dokonaj klasyfikacji metodą LDA każdej możliwej pary z $k$-pierwszych klas fonemów z wektora $SPC$ i wybierz tę klasę, która najwięcej razy będzie wynikiem metody

Powyżej opisaną klasyfikację przeprowadzono dla różnych $k$ i najlepsze wyniki osiągnięto dla $k \approx 8$. Metoda ta daje wynik lepszy o 6 punktów procentowych niż sama metoda NM:
Analiza skupień segmentów sygnału akustycznego

\[ \text{RAF}_{\text{MFCC}}(a,a) = 0.4115 \]

Wartość współczynnika *Adjusted Rand Index* dla wyniku klasyfikacji wynosi: \( \text{AR} = 0.3956 \).

\[ \text{Rysunek 6.24. Macierz RAF dla } N_{\text{NMED}}; \text{LDA}_8 \]

Na Rysunku 6.24 zaprezentowano wynik działania powyższej metody w formie macierzy współczynników RAF. Można zauważyć znaczącą poprawę klasyfikacji. Do fonemów najsłabsze rozpoznawalnych (poniżej 20%) należą: “ę” (16% prawidłowo rozpoznanych) oraz “drz” (15%). Występują także pary głosek, które są mylone częściej niż pozostałe: (ą,o), (ę,e), (dz,z), (drz,rz), (dzi,zi), (c,s), (ci,si), (cz,sz), (l,u), (j,i). Można również zauważyć “wyspy”, obszary na których wartość współczynnika RAF jest zwiększona np.: (p,t,k), (b,d,g), (m,n,ń), (dz,drz,dz; b,d,g), (c,ci,cz; p,t,k).

Rysunek 6.25 przedstawia wyniki metody w postaci macierzy współczynników CCF. Umożliwia to porównanie z wynikami metody k-means: Rysunek 6.9.

### 6.6 Klasyfikacja artykulacji

W poniższej sekcji zamieszczono wynik klasyfikacji artykulacji dla niezależnych mówców. Artykulacja jest to zbiór elementów z \( V_f : f \in F \), taki że elementy te są ciągłym fragmentem wypowiedzi i każdy z nich odpowiada tej samej głosce. Klasyfikacja artykulacji polega na stwierdzeniu jaki jeden fonem reprezentuje zbiór wektorów parametryzacji.

Zaproponowano dwie następujące metody klasyfikacji artykulacji:
6.6 Klasyfikacja artykulacji

Wejście algorytmu: \( U = (u_1, u_2, \ldots, u_n) : u_i \in V; n = |U| \) - ciąg stanowiący artykulację.

Algorytm 1:

1. Obliczany jest środek ciężkości artykulacji, \( U_c \) będący środkiem ciężkości zbioru \( U \).
2. \( U_c \) jest klasyfikowany metodą \( N_{NED} + LDA_8 \).

Algorytm 2:

1. Każdy \( u \in U \) jest klasyfikowany metodą \( N_{NED} + LDA_8 \) wynikiem czego jest klasyfikacja \( C : u \in U \rightarrow f \in F \)
2. Następnie obliczany jest środek ciężkości artykulacji, \( U_c \) będący środkiem ciężkości zbioru środków ciężkości fonemów, z których złożona jest artykulacja, każdy o wadze równej ilości wystąpień.

3. \( U_c \) jest klasyfikowany metodą \( NM_{NED} + LDA_8 \).

Skuteczność alorytmu 2. wynosi: \( \overline{RAF_{MFCC}(a, a)} = 0.5789 \).

Skuteczność algorytmu pierwszego jest o 8 punktów procentowych gorsza od drugiego: \( \overline{RAF_{MFCC}(a, a)} = 0.4996 \).

Rysunek 6.27 przedstawia wynik klasyfikacji artykulacji za pomocą algorytmu drugiego w postaci macierzy RAF.

W trzecim kroku algorytmu metoda NM generuje dla każdego \( U_c \) ciąg SPC, który zawiera indeksy fonemów posortowane według rosnącej odległości od punktu \( U_c \). Na Rysunku 6.28 (po prawej) pokazano średnią pozycję prawidłowej głoski w tym wektorze: jest ona dla większości głosek mniejsza od 3. Po lewej stronie zamieszczony został histogram poprawnych pozycji głosek w ciągach SPC. Pokazuje on, że w znacznej większości przypadków prawidłowy wektor znajduje się na pierwszej pozycji ciągu SPC. Wykres ten sugeruje, że wektor SPC mógłby być być podstawą do określenia prawdopodobieństwa klasyfikacji artykulacji jako dana głoska z dobrym rezultatem.
6.7 Klasyfikacja metodą k-NN

Metoda k-najbliższych sąsiadów służy do klasyfikacji obiektów na podstawie najbliższych mu obiektów w przestrzeni cech. Jej algorytm wygląda następująco:

Dane wejściowe: k - ilość najbliższych sąsiadów na podstawie, których klasyfikowany jest obiekt; \( V_l \) - zbiór danych treningowych; \( V_t \) - zbiór danych testowych;

Algorytm: Dla każdego elementu zbioru \( v_i \in V_l \):

1. stworzyć ciąg k-najbliższych sąsiadów obiektu \( v_i \) w zbiorze \( V_l \): \( kNN(v_i) \)
2. określić klasę, której obiektów znajduje się najczęściej w ciągu \( kNN(v_i) \): \( c \) (może być więcej niż jedna taka klasa)
3. jeśli takich klas jest kilka wybrać tę, której obiekt znajduje się najbliżej obiektu \( v_i \): \( c_{\text{nearest}} \)
4. obiektowy \( v_i \) przypisać klasę wybraną w punkcie drugim, lub, jeśli jest więcej niż jedna, to tę wybraną w punkcie trzecim algorytmu.

Rysunek 6.28. Średnia pozycja prawidłowej głoski w ciągu SPC dla artykulacji (po prawej). Histogram pozycji (po lewej).

Rysunek 6.29. Średnia wartość na diagonali macierzy RAF w zależności o parametr k w metodzie kNN.
Odległości pomiędzy obiektami jest odległością euklidesową. Testy z użyciem znormalizowanej odległości euklidesowej (różnica współrzędnych jest dzielona przez odchylenie standardowe w danym wymiarze) dały o wiele gorsze wyniki.

Na Rysunku 6.29 przedstawiono skuteczność klasyfikacji (średnią na diagonali macierzy RAF) w zależności do parametru \(k\) (ilości najbliższych sąsiadów branych pod uwagę przy klasyfikacji). Skuteczność rośnie wraz ze wzrostem \(k\). Z kolei na rysunku 6.30 pokazano wynik klasyfikacji w postaci macierzy RAF.

Pary fonemów słabo separowalnych przez metodę kNN są podobne to tych słabo separowalnych metodami NM, należą do nich: (ł,u), (i,j), (ą,o), (ę,e), (y,e), (N,n), (p,t), (dz,z), (c,s).

Do najsłabiej rozpoznawanych głosek należą: ę, Ń, N, dz, drz, dzi, c, ci.
6.8 Skalowanie wielowymiarowe

Rysunek 6.32. Położenie w 2. i 3. wymiarze parametryzacji MFCC środków ciężkości zbiorów $V_f$

Poniżej, na rysunku 6.32, przedstawiono rzutowanie na 2. i 3. wymiar parametryzacji MFCC środków ciężkości zbiorów $V_f : f \in F$. Można zauważyć, że zbiory podobnych głosek układają się w podobne kształty "półksiężyca". Dla przykładu: 4 zbiory głosek: (c,ci,cz), (z,zi,rz), (s,si,sz), (dz,dzi,drz). Wewnątrz nich każda kolejna głoskę można opisać cechami: (przedniojęzykowa, zębowa, twarda; środkowojęzykowa, paralatalna; przedniojęzykowa, dziąsłowa, twarda). Pierwsza z nich występuje zawsze w górnym roku "półksiężyca", druga w jego środku, a trzecia w dolnym rogu. Samogłoski oraz półsamogłoski tworzą kształt liter V, co odpowiada kształtowi na wykresie dwóch pierwszych formantów dla samogłosek oraz kolejnym poziomu jakie zajmuje język w jamie gębowej wypowiadając je w takiej kolejności.

Na Rysunku 6.33 przedstawiono wynik skalowania wymiarowego, z dwudziestu do dwuch wymiarów, środków ciężkości zbiorów $V_f : f \in F$. Jako wejściowa macierz różnica podana została macierz znormalizowanych według wariancji odległości euklidesowych pomiędzy środkami ciężkości.

Po zastosowaniu skalowania wielowymiarowego (do dwóch wymiarów) utracone zostały pewne właściwości wzajemnego położenia środków ciężkości tych zbiorów (np. zbiory zostały obrócone), zachowano jednak ich kształty i kolejność w jakim występują w pogrupowane głoski (wymienione w poprzednim paragrafie). Tym samym skalowanie wielowymiarowe może być
stosowane w celu zmniejszenia wymiarowości parametryzacji MFCC. Uwzględnienie przy tworzeniu macierzy różnic innych właściwości i podobieństwa głosek możliwe byłoby ciekawszego wyniku.

Rysunek 6.33. Wynik skalowania wymiarowego dla standaryzowanej odległości euklidesowej pomiędzy środkami ciężkości zbiorów $V_f$

6.9 Analiza algorytmem OPTICS


Algorytm ten działa dużo gorzej dla głosek, których zbiory są “zazębione”. Wtedy ważny jest dobór parametrów, na które algorytm jest czuły i które ciężko oszacować. Dla zbiorów
słabo separowalnych przy dobrym doborze parametrów (oraz obiektów na których wykonywany jest test) algorytm potrafi wyróżnić centra skupień, w których gęstość jest największa.

Rysunek 6.34. Wynik działania algorytmu OPTICS głosek dla trzech 1000-elementowych zbiorów głosek: a,p,s.
Podsumowanie

W niniejszej pracy przedstawiono wyniki analizy skupień segmentów sygnału akustycznego. Dane do analizy zostały wyekstrahowane z korpusu mowy polskiej Corpora według algorytmu stosowanego przez uznaną zestaw narzędzi do rozpoznawania mowy HTK. Dokonano analizy następującymi metodami: k-means, OPTICS, Nearest Mean, LDA, k-Nearest Neighbours, MDS oraz hybrydowej metody NM+LDA. Podjęto także próby, z zadowalającym rezultatem, klasyfikacji artykułow z wykorzystaniem metody NM+LDA. W celu ułatwienia wizualizacji i analizy wyników tych metod opracowano parę współczynników: RAF oraz CCF. Wykorzystano także standardowe wskaźniki podobieństwa wyników partycjonowania danych: Adjusted Rand Index. Większość badań została przeprowadzona dla obu parametryzacji MFCC oraz MFS. Przedstawiono jednak głównie wyniki dla MFCC.

W wyniku wyżej wymienionych badań otrzymano wiele informacji na temat wzajemnego położenia skupień fonemów w przestrzeni MFCC oraz dodatkowo ich struktury przestrzennej.

Badania metodą k-means dostarczyły informacji na temat grup fonemów, których skupienia leżą blisko siebie i mocno się przenikają. Stanowią je głoski, różniące się cechami akustycznymi, które mają mały wpływ na odmienność w brzmieniu głoski. Taką grupę stanowią np.: (/p/,/t/,/k/) i (/b/,/d/,/g/). Wyniki tej metody były bardzo ogólne i nie dawały możliwości głębszej analizy podobieństwa między głoskami.

Badania metodą Nearest Mean pokazały, że odległość euklidesowa znormalizowana za pomocą wariancji w każdym z wymiarów dużo lepiej nadaje się do mierzenia odległości w przestrzeni MFCC (tj. licząc odległość do centroidu skupienia głosek x normalizuje się wariancję tego skupienia). W razie mierzenia odległości pomiędzy punktami, o których przynależności do zbiorów nie mamy informacji, warto jest zastosować odległość Mahalanobisa. Analiza wyników tej metody w połączeniu znormalizacją odległości pozwoliła bardziej szczegółowo zbadać rozłożenie skupień w przestrzeni MFCC. Przykładowo stwierdzono duże podobieństwo (w stosunku do innych par) takich par jak np.: (/l/,/u/), (/i/,/j/), (/n/,/l/), (/dz/, /z/). Klasyfikacja segmentów tą metodą jest szybka w działaniu i łatwa w zrównolegleniu.
Osiągnięto prawie dwukrotne przyśpieszenie \((S_p = 1.75)\) przez rozproszenie obliczeń na trzy rdzenie procesora.

Poprawność klasyfikacji metodą k-NN była porównywalna z metodą Nearest Means. Do jej głównej wady należy jednak duża złożoność obliczeniowa oraz bardzo duża ilość informacji jaką posiadać musi system klasyfikujący: są to wszystkie elementy zbioru treningowego, podczas gdy metoda NM potrzebuje tylko środków ciężkości oraz wariancji w każdym z wymiarów dla wszystkich fonemów.

Metodą LDA zbadano wzajemne przenikanie się zbiorów par głosek. Otrzymane kombinacje liniowe dwudziestu wymiarów MFCC, separujące wszystkie możliwe pary fonemów, w połączeniu z metodą NM, posłużyły za dodatkowe narzędzie do ich klasyfikacji: metodę NM+LDA. Dodatkowa analiza rozłożenia gęstości chmur fonemów w każdym z wymiarów dowiodła, że niektóre głoski nie są rozłożone według rozkładu normalnego w przestrzeni MFCC; np. /drz/ posiada dwa zagęszczenia: jedno w pobliżu centrum zagęszczenia głoski /d/, drugie głoski /rz/. Było to powodem, słabej separowalności tych głosek.

Podstawą algorytmu klasyfikacji artykulacji było wyznaczenie środka ciężkości artykulacji na podstawie klasyfikacji składających się na niego segmentów. Klasyfikacji artykulacji potwierdziła fakt ciągłość mowy, w tym efekt koartykulacji, w przestrzeni MFCC. Poprawa o prawie 20 punktów procentowych wyniku klasyfikacji dla całych artykulacji (względem klasyfikacji pojedynczych segmentów) oznacza, że głoski nie składają się z segmentów o identycznych wartościach parametrów. Parametry te oscylują wokół środka ciężkości chmury danego fonemu. Potwierdza to także analiza algorytmem OPTICS, z której wynika że w centrum skupień znajdują się segmenty położone bliżej środka artykulacji.
Literatura

5. Jorge M. Santos, Mark Embrechts *On the Use of the Adjusted Rand Index as a Metric for Evaluating Supervised Classification*, ISEP - Instituto Superior de Engenharia do Porto, Portugal; Rensselaer Polytechnic Institute, Troy, New York, US
14. J. Gałka *Optymalizacja parametryzacji sygnału w aspekcie rozpoznawania mowy polskiej*, AGH University of Science and Technology, Kraków 2008
15. J. Gałka, *Distance Measures for wavelet representation of speech segments*, AGH University of Science and Technology, Koninki 2006
Literatura
Spis rysunków

2.1 Model schematyczny mechanizmu słuchowego u człowieka (za [12]) 6

3.1 Spektrogram Mel-frequency dla wypowiedzi słowa “pięć”. 8

3.2 Różnice w okienkowaniu dla metod analizy częstotliwościowej: Short-Time Fourier Transform (na górze) Wavelet Transform (na dole) [14] 9

5.1 Segmentacja wypowiedzi słowa ’pięć’ 16

5.2 Średni czas trwania poszczególnych fonemów [sec] 16

5.3 Ilość wystąpień poszczególnych fonemów w korpusie 16

5.4 Macierz częstości występowania par głosek w korpusie 17

5.5 Trójkąt samogłoskowy obrazujący pionowe i poziome położenie języka przy wymowie samogłosek. 17

6.1 Diagram procesu parametryzacji korpusu. Kolorem czerwonym zaznaczono okno oraz jego przesunięcia. 20

6.2 Stosunek odległości pomiędzy parami centroidów skupień fonemów a średnią wariancją ich chmur (po lewej); położenie centroidów skupień fonemów (po prawej). Parametryzacja MFCC 21

6.3 Stosunek odległości pomiędzy parami centroidów skupień fonemów a średnią wariancją ich chmur (po lewej) oraz położenie centroidów skupień fonemów (po prawej). Parametryzacja MFS 22

6.4 Znormalizowana odległość euklidesowa (za normę przyjęto wariancję) pomiędzy centroidami skupień fonemów. Parametryzacja MFCC (po lewej) i MFS (po prawej) 22

6.5 Wartość bezwzględna współczynnika skośności (skewness) dla każdego wymiaru poszczególnych fonemów 23

6.6 Średnia wariancja (po prawej) oraz wariancja dla każdego wymiaru (znormalizowana do 1) dla każdej głoski. (MFCC) 24
6.7 Współczynniki poprawności klastrowania k-means w zależności od k. $P_{20}^{ cep}$
$I_1 = 100, I_2 = 40, 10 \leq k \leq 40$ ................................................................. 25
6.8 Rzut wyników klastrowania fonemu 'a' na dwa pierwsze wymiary reprezentacji
MFCC; każda klasa została oznaczona innym kolorem. (po lewej). Histogram
klas do jakich został przyporządkowany ten fonem (po prawej). Parametry:
$P_{cep}^{ 20}, I = 40, k = 37$ ................................................................. 26
6.9 Współczynnik Common-Cluster Factor dla każdej pary fonemów. $P_{cep}^{ 20}, I = 40,$
$k = 37$ ................................................................. 27
6.10 Ułożenie języka mowy przy produkcji głoski przedniojęzykowo-zębowej (po
lewej) oraz wargowej (w środku) oraz nosowej (po prawej). .................. 27
6.11 Macierze współczynnika Common Cluster Factor $k$ o wartościach kolejno
rzędami:10, 13, 17, 21, 25, 29, 33, 37, 40. Parametry: $P_{cep}^{ 20}, I_1 = 100, I_2 = 40$ . 28
6.12 Średnia wartość współczynnika CCF dla poszczególnych głosek .................. 28
6.13 RAF dla NM z użyciem odległości euklidesowej ................................. 31
6.14 Po lewej: niebieski - średnia pozycja prawidłowej klasy w wektorze SPC;
czerwony - średnia odległość od centrum prawidłowej klasy; zielony - odległość
do najbliższej klasy; żółty - różnica poprzednich dwóch odległości. Po prawej:
histogram pozycji prawidłowej klasy w wektorze SPC. ................................. 31
6.15 RAF dla NM z użyciem zmormalizowanej odległości euklidesowej ........... 34
6.16 Po lewej: niebieski - średnia pozycja prawidłowej klasy w wektorze SPC;
czerwony - średnia odległość od centrum prawidłowej klasy; zielony - odległość
do najbliższej klasy; żółty - różnica poprzednich dwóch odległości. Po prawej:
histogram pozycji prawidłowej klasy w wektorze SPC. ................................. 34
6.17 Wartość RAF dla elementów na diagonali macierzy.............................. 34
6.18 Macierz RAF dla klasyfikacji metodami (kolejno rzędami): k-means, NM
(odległość ekulidesowa), NM (znormalizowana odległość euklidesowa), NM +
LDA. ................................................................. 35
6.19 Rozkład gęstości dla głosek 'd'(czerwony),'drz'(zielony),'rz'(niebieski) dla
kolejnych 20 wymarów (rzędami) ............................................................... 37
6.20 Rozkład gęstości dla głosek 'd'(czerwony),'drz'(niebieski) rzutowanych na
jeden wymiar metodą LDA. Przerywanymi liniami wykreślono dystrybuantę... 37
6.21 Macierz RAF dla par fonemów rozróżnianych metodą LDA................. 38
6.22 Średnia wartość w kolumnach (wykres górny) i rzędach (wykres dolny)
macierzy RAF dla par fonemów (LDA) ..................................................... 39
6.23 Separowalność fonemów na podstawie metod NM i LDA ................. 43
6.24 Macierz RAF dla $NM_{NED}; LDA_8$ ............................................. 44
6.25 Macierz CCF dla $NM_{NED}; LDA_8$ ............................................. 45
Spis rysunków

6.26 Diagonala macierzy RAF dla $NM_{NED}; LDA_{8}$ .......................................... 45
6.27 Wynik klasyfikacji artykulacji (macierz RAF) ............................................. 46
6.28 Średnia pozycja prawidłowej głoski w ciągu SPC dla artykulacji (po prawej).
    Histogram pozycji (po lewej). ................................................................. 47
6.29 Średnia wartość na diagonali macierzy RAF w zależności o parametr $k$ w
    metodzie kNN ......................................................................................... 47
6.30 Wynik klasyfikacji metodą kNN dla $k=20$ w postaci macierzy RAF .......... 48
6.31 Wynik klasyfikacji metodą kNN dla $k=20$ w postaci diagonali macierzy RAF .. 48
6.32 Położenie w 2. i 3. wymiarze parametryzacji MFCC środków ciężkości zbiorów
    $V_f$ ........................................................................................................... 49
6.33 Wynik skalowania wymiarowego dla standaryzowanej odległości euklidesowej
    pomiędzy środkami ciężkości zbiorów $V_f$ ............................................... 50
6.34 Wynik działania algorytmu OPTICS głosek dla trzech 1000-elementowych
    zbiorów głosek: a,p,s. ............................................................................... 51
Spis rysunków
Spis tablic

5.1 Głoski ................................................................. 18
6.1 Podobne głoski ......................................................... 33