POLITECHNIKA WROCŁAWSKA
WYDZIAŁ ELEKTRONIKI

KIERUNEK: Elektronika (EKA)
SPECJALNOŚĆ: Aparatura elektroniczna (EAE)

PRACA DYPLOMOWA MAGISTERSKA

System weryfikacji mówcy w czasie rzeczywistym

Real-time speaker verification system

AUTOR:
inż. Adam Matusiak

PROWADZĄCY PRACĘ:
dr hab. inż. Józef Borkowski Prof. PWr

OCENA PRACY:

WROCŁAW 2017

Rozdział 1

Wprowadzenie

1.1 Wprowadzenie.

Duża dynamika wzrostu produkcji danych takiego typu jak nagrania audio czy nagrania video stwarza potrzebę wdrażania nowych i niezawodnych systemów biometrycznych pozwalających na stwierdzenie tożsamości osób przy ich pomocy. Dotychczasowe techniki rozpoznawania (identyfikacji) korzystające z tego typu danych, ze względu na wysoką złożoność obliczeniową realizujących je algorytmów sprawiają, że zadanie rozpoznawania tożsamości staje się operacją długotrwałą. Powodowane jest to tym, iż wspomniane rodzaje medium charakteryzują się wysoką pojemnością informacji, a przy tym sama informacja dotycząca tożsamości osób stanowi jedynie ułamek pojemność takiego pasma. Wspomniana trudność nie stanowi problemu i jest dopuszczalna w aplikacjach typu *off-line*, tam gdzie nie są oczekiwane natychmiastowe wyniki oraz gdzie rozmiar danych jest stosunkowo mały. Sprawa jednak komplikuje się gdy w grę wchodzi przetwarzanie całych petabajtów danych lub też w aplikacjach czasu rzeczywistego, gdzie ograniczenia czasowe na otrzymanie wyniku są równie istotne jak ich poprawność.

Zdaniem autora pracy główne motywacje dla zastosowania tego typu systemów są dwie. Pierwsza z nich jest związana z administracją państwową lub też z inwigilacją społeczeństwa ze strony agencji rządowych (np. jako system przeciwdziałania terroryzmowi lub aparat policyjny) oraz dużych korporacji (np. jako system wspomagający sprzedaż produktów). Z punktu widzenia tego typu podmiotów, zainteresowanie systemami biometrycznymi wykorzystującymi tego typu dane wynika z powszechności tej informacji - dla przykładu setki milionów godzin materiału filmowego wraz z dźwiękiem umieszczane rok rocznie na serwisie internetowym youtube[**ystats**] mogą stanowić świetne wyjście dla tego typu systemu. Kolejnym argumentem przemawiającym na korzyść systemów biometrycznych wykorzystujących obraz oraz dźwięk w tym przypadku są względnie małe koszta infrastruktury monitorującej - sieci mikrofonów oraz kamer, ze względu na ich dostępność oraz małą cenę. Zwłaszcza z tego względu, że może zostać wykorzystane infrastruktura już istniejąca - sieci komórkowe czy internetowe kamery. Bardzo szybki rozwój infrastruktury na wszystkich kontynentach oraz coraz to szybciej zwiększająca się populacja ludzi potęguje ten efekt. Inne metody ustalania tożsamości takie jak systemy biometryczne wykorzystujące odciski palców, testy dna czy skany tęczówki oka wymagają zazwyczaj kooperacji ze strony osoby identyfikowanej - co w niektórych sytuacjach może to stanowić problem, szczególnie gdy system cechuje się masowością i ma w zamierzeniu pozyskać informację na temat tożsamości znaczącej części populacji danego obszaru geograficznego. Z oczywistego względu, iż człowiek nie jest w stanie podołać analizie takiej ilości danych pochodzących z rozbudowanej infrastruktury, potrzebne są systemy automatycz-

nego rozpoznawania mówcy czy też systemy automatycznego rozpoznawania twarzy. Inną możliwością dla realizacji rozbudowanych systemów inwigilacji jest wykonywanie analizy sygnały mowy już na etapie urządzeń które rejestrują sygnały mowy czy obrazy. Szeroko rozpowszechnione urządzenia wbudowane np. telefony komórkowe, telewizory czy konsole, ze względu na bardzo dynamiczny rozwój możliwości obliczeniowych są w stanie nie tylko rejestrować wspomniane dane, ale także analizować sygnał i przesyłać do systemy scentralizowanego konkretny, matematyczny model mówcy/twarzy. Ogranicza to w sposób znaczący konieczną ilość przesyłanych danych.

Drugą motywacją, nie przejawiającą już raczej żadnych wątpliwości moralnych, jest zastosowanie automatycznych systemów identyfikacji tożsamości w systemach chroniących poufne informacje lub ograniczające dostęp do posiadanego mienia np. jako lokalne punkty dostępu w budynkach lub magazynach. Tego typu systemy muszą cechować się bardzo dużą niezawodnością procesu weryfikacji oraz szybkością jej przeprowadzenia - tak aby czas potrzebny na wykonanie obliczeń niezbędnych do dokonania weryfikacji nie był obciążeniem dla użytkownika. Chociaż systemy analizujące sygnał mowy (np. poprzez połączenie telefoniczne podczas autoryzacji dla sektora bankowego) mogą być implementowane i wykonywane na dużych systemach informatycznych z ogromnymi możliwościami obliczeniowymi, ze względu na opóźnienia w komunikacji lub brak zewnętrznej sieci informatycznej istnieje potrzeba implementacja takiego systemu lokalnie, na urządzeniu wbudowanym i w czasie rzeczywistym. Przykładem takiego systemu jest system autoryzacji dostępu w biurowcach, gdzie wewnętrzna sieć urządzeń rejestrujących i analizujących sygnał mowy w czasie rzeczywistym dokonuje procesu weryfikacji mówcy i podejmuje decyzję o udzieleniu dostępu do zastrzeżonych obszarów.

Z zaprezentowanych powyżej rozważań widać, iż kluczowym zagadnieniem dotyczącym obu typów rozpatrywanych systemów rozpoznawania jest szybkość ich działania. Dla pierwszego przypadku ilość przetwarzanych danych jest ogromna i dlatego ważne jest by stosowane algorytmy były jak najmniej kosztowne obliczeniowo, oraz by sprzęt na którym są wykonywane umożliwiał ich jak najszybszą realizację po to aby była możliwa analiza całości uzyskanego materiału w rozsądnym czasie. W drugim przypadku zaś ta sama cecha umożliwia przeprowadzenie rozwiązania problemu rozpoznawania w czasie rzeczywistym bez zbędnych opóźnień. Rozwój w dziedzinie techniki cyfrowej sprawia, że problem szybkości wykonywania tego zadania jest coraz mniejszy. Dla pierwszej dziedziny umożliwione jest to dzięki temu, że duże serwery zaczynają być masowo wyposażane w jednostki graficzne (GPU) czy też układy programowalne (FPGA) co jest połączone ze wzrostem szybkości samych jednostek centralnych (CPU) oraz procesem zrównoleglenia tychże jednostek. Wykorzystanie nowych architektur sprzętowych spowodowane jest szybkich rozwojem w dziedzinie algorytmów heurystycznych takich jak sieci neuronowe oraz ich wykorzystanie dla metod sztucznej inteligencji. Podobnie jest w dziedzinie systemów wbudowanych. Systemy mikroprocesorowe osiągają wydajności obliczeniowe umożliwiające przetwarzanie w czasie rzeczywistym złożonych algorytmów. I tutaj także możliwe jest stosowanie wysoko wydajnych układów programowalnych FPGA które umożliwiają wykonywanie zadań realizowanych poprzez sieci neuronowe. Platformy sprzętowe wspomagające wykonywanie algorytmów cyfrowego przetwarzania sygnałów na czele z procesorami sygnałowymi DSP również przyczyniają się do efektywnej implementacji systemów rozpoznawania, szczególnie dla dziedziny rozpoznawania mówcy i rozpoznawania mowy.

Równie istotnym czynnikiem ograniczającym powszechne istnienie tego typu systemów jest ich skuteczność identyfikacji. Dotychczasowe rozwiązania dla systemów weryfikacji mówcy osiągały błąd sięgający 4% (ERR), (2% w przypadku fuzji systemów)[**overview**]. Szybki rozwój w dziedzinie rozpoznawania wzorca, spowodowany m. in. dynamicznym roz-

wojem w dziedzinie metod sztucznej inteligencji pozwala redukować błędy - dla problemu identyfikacji mówcy - nawet do 0.4 % (ERR) [**deepfeaturelearning2017err**]. Szeroko stosowana miara oceny systemów weryfikacji mówcy - ERR (*ang. equal error rate*) mówi o prawdopodobieństwie błędu systemu dla progu decyzji dla którego błąd odrzucenia prawdziwego mówcy jest równy błędowi akceptacji mówcy prawdziwego.

W niniejszej pracy podejmowana jest próba przedstawienia architektury oprogramowania pozwalającej efektywną obliczeniowo aplikację algorytmów realizujących zadanie weryfikacji mówcy na platformach sprzętowych systemów wbudowanych. Projekt zakłada, że przetwarzanie wejściowego sygnału mowy przeprowadzane jest w czasie rzeczywistym i umożliwia otrzymanie decyzji o autoryzacji w czasie nie dłuższym niż oczekiwany przez potencjalnych użytkowników takiego systemu. Platformy sprzętowe za pomocą których realizowane jest przetwarzanie, przewidziane są jako systemy mikroprocesorowe - ogólnego przeznaczenia (CPU), procesory sygnałowe (DSP) czy mikrokontrolery (MCU) posiadające wsparcie dla języka programowania C++ dla jego najnowszych standardów: C++11, C++14 i C++17. Propozycja architektury nie bazuje na żadnym systemie operacyjnym i może być wykorzystana również w systemach wbudowanych które nie oferują żadnego środowiska uruchomieniowego. Jednak obecność takiego systemu, zwłaszcza systemu operacyjnego czasu rzeczywistego (RTOS) w dużym stopniu może ułatwić implementację konkretnego systemu na urządzeniu. W proponowanym zastosowaniu prezentowanym przez niniejszą pracę autor korzysta ze wsparcia systemu operacyjnego linux w dystrybucji debianowej. Platformą uruchomieniową jest komputer jednopłytkowy Raspberry Pi 3.

DO ZROBIENIA: WSTĘPNY OPIS ARCHITEKTURY Projektowana architektura ma w zamierzeniu ułatwiać aplikację różnych technik realizujących weryfikację mówcy proponując narzędzia reprezentujące abstrakcję etapów ... —— Chociaż proponowane oprogramowanie przeznaczone jest jedynie dla systemów mikroprocesorowych to może być wykorzystane jako element systemu przetwarzający wstępnie dane - np. tworzący wektory akustyczne przekazywane dalej do innych systemów np. sieci neuronowej zaimplementowanej na układzie FPGA. Niewykluczone jest też użycie abstrakcji dopasowywania cech do implementacji sieci neuronowej na mikroprocesorze - co może jednak być nieefektywne. **KONIEC DO ZROBIENIA.**

1.2 Weryfikacja mówcy

{verificati

Proces weryfikacji mówcy (*ang. speaker verification*) jest związany z szerszym zagadnieniem - rozpoznawania mówcy (*ang. speaker recognition*), które charakteryzuje ogół metod wykorzystujących dane biometryczne zawarte w sygnale mowy w celu określenia tożsamości.

Sygnał mowy może być rozpatrywany jako cecha biometryczna. Sygnał mowy charakteryzowany jest przez budowę aparatu głosowego człowieka, która jest mniej lub bardziej unikatowa dla każdego człowieka, umożliwiając rozróżnienie badanej jednostki na tle populacji.

Ogólną strukturę problemu rozpoznawania mówcy można rozłożyć na na trzy elementy.[**fosr**] Po pierwsze, konieczne jest aby tworzony system dysponował modelem charakterystyk aparatu głosowego człowieka. Model taki dla przykładu może przybrać formę modelu fizyko-matematycznego aparatu głosowego człowieka. Otrzymany model musi umożliwiać parametryzację - skojarzenie z konkretną osobą. Model taki tworzony jest poprzez analizę sygnału mowy. Dopiero na tej podstawie możliwe jest porównywanie modelu utworzonego

przy użyciu testowanego sygnału z modelem odniesienia. Forma i cel tego porównania definiują podklasę problemu rozpoznawania mówcy.

Weryfikacja mówcy charakteryzuje się wykonaniem dwóch kluczowych porównań - pierwszego pomiędzy modelem utworzonym z poddanego weryfikacji sygnału mowy a pamiętanym modelem osoby której dotyczy weryfikacja. W odróżnieniu of problemu identyfikacji, podczas weryfikacji mówcy potrzebna jest więc znajomość tożsamości osoby poddanej weryfikacji. Drugie z kolei porównanie dokonywane jest pomiędzy modelem poddanym weryfikacji, a uogólnionym modelem całej populacji (*ang. background model*) lub pewnej jej podgrupy (*ang. cohort model*). Na podstawie relacji tych dwóch odległość podejmowana jest decyzja o autoryzacji.

W przypadku kiedy nie jest możliwa lub pożądana znajomość przez system tożsamości osoby weryfikowanej, przed dokonaniem autoryzacji możliwe jest zastosowanie bardziej złożonego problem identyfikacji mówcy na otwartym zbiorze(*ang. open-set speaker identification*). Proces ten można uważać jako złożenie problemu weryfikacji mówcy oraz identyfikacji mówcy na zbiorze zamkniętym (*ang. close-set speaker identification*). Polega on na przeprowadzeniu weryfikacji mówcy na modelu uzyskanym z procesu identyfikacji mówcy na zbiorze zamkniętym, która dokonuje porównania z całą dostępną bazą modeli mówców i zwraca ten najbliższy modelowi testowanemu. Problem taki jest więc obliczeniowo co najmniej tak złożony jak weryfikacja mówcy (dla bazy w której znajduje się tylko jeden mówca).

Implementacja systemu weryfikacji mówcy jest nazywana automatycznym systemem rozpoznawania mówcy (*ang. automatic speaker verification system*).

1.2.1 Zastosowania.

Głównymi obszarami zastosowań systemów weryfikacji mówcy są:

- Usługi bankowe: jako zdalna weryfikacja (np. telefoniczna, audio-video) dająca dostęp do danych dotyczących konta czy potwierdzenia realizacji usług. Stosowana może być także bez wiedzy zainteresowanego do zapobiegania oszustwom np. poprzez sprawdzenie czy osoba nie znajduje się w bazie osób podejrzanych.

- Zastosowania prawne: podobnie jak odciski palców czy badania DNA do weryfikacji tożsamości osób na nagraniach.

- Inwigilacja: do zapobiegania przestępstwom czy też terroryzmowi.

- Ochrona dostępu: jako lokalne punkty dostępu chroniące zasoby fizyczne (biura, magazyny, serwerownie) lub dostęp do wszelkiego rodzaju informacji (internetowe bazy danych).

- *Indeksowanie* wypowiedzi: w towarzystwie technik rozpoznawania mowy, system weryfikacji mówcy pomaga archiwizować zebrane, masowe nagrania audio i stwierdzać przynależność danych wypowiedzi do konkretnego mówcy.

1.2.2 Inne systemy biometryczne.

System weryfikacji mówcy może być skojarzony z innymi systemami rozpoznawania biometryk, szczególnie w lokalnych punktach autoryzacji gdzie mówca znajduje się fizycznie. Może okazać się to konieczne chociażby ze względu na to że część populacji jest niema. Przykładami podanymi w [**fosr**] są:

- Analiza DNA - niechybnie najpewniejsza metoda identyfikacji, jednak trwająca dużo dłużej niż weryfikacja mówcy.

- Analiza kształtu małżowiny usznej - może być użyta w połączeniu z weryfikacją mówcy w połączeniu telefonicznym na odległość np. poprzez użycie czujnika (obraz uzyskany za pomocą kamery, lub analiza akustyczna) w telefonie komórkowym. Okazuje się, że kształt małżowiny usznej różni się na tyle w obrębie populacji, iż można zastosować tą technikę jako wsparcie dla systemów weryfikacji mówcy.

- System rozpoznawania twarzy - ze względu na powszechność nagrań typu audio-wideo, system weryfikacji mówcy świetnie nadaje się do współpracy z systemami biometrycznymi wykorzystującymi rozpoznawanie mówcy. Kombinacja ta pozwala stwierdzić tożsamość osób znajdujących się na takich nagraniach.

- Skaner odcisku palca - ze względu na tanie, dedykowane czujniki, rozsądnym jest wyposażenie lokalnego punktu dostępu w system biometryczny oparty na skanie odcisku palca.

- Do innych cech biometrycznych, potencjalnie nadających się do współpracy z systemem rozpoznawania mówcy należą: wygląd i geometria dłoni, obraz tęczówki oka, obraz siatkówki oka, obraz termograficzny ciała, rozkład żył w dłoni, sposób chodu, pismo czy sposób pisania na klawiaturze. Do tego wyróżnia się systemy wielomodalne zawierające szereg podanych metod w jednym systemie.

1.3 Relacja pomiędzy rozpoznawanie mowy, a rozpoznawaniem mówcy.

Kluczowe jest odróżnienie procesu rozpoznawania mówcy od systemów rozpoznawania mowy (*ang. speech recognition*). Pomiędzy tymi dwoma rozpatrywanymi dziedzinami z zakresu analizy sygnału mowy występuje dychotomiczny podział. Wynika to z tego, że sygnał mowy jest sygnałem bogatym informacyjnie oraz że jedynie mała część tej informacji posiada znaczenie semantyczne, zaś reszta niesie wiedzę o budowie konkretnego, ludzkiego narządu mowy. W problemie rozpoznawania mowy nie jest istotna tożsamość osoby wypowiadającej się, a jedynie sens jej wypowiedzi. Zatem reszta sygnału nie zawierająca odczytywanej wiadomości jest redundantna z punktu widzenia tego zagadnienia - cała informacja biometryczna jest niewykorzystywana, co za tym idzie często filtrowana przez zaimplementowany system. Z drugiej strony, w systemach rozpoznawania mówcy, w samym sednie jego zainteresowania, abstrahuje się od treści mowy. Stanowi ona jedynie środek dla dostarczenia informacji o fizjologii aparatu mowy. Dlatego prawdopodobnie system rozpoznawania mówcy usunie treść mowy, a utworzy jedynie model aparatu głosowego. Usprawiedliwia to twierdzenie o rozłączności tych dziedzin ze względu na zainteresowanie informacją zawartą w sygnale mowy.

Okazuje się, że wspomniana wyżej zależność powoduje to, że techniki przetwarzania sygnału stosowane przy analizie obu dziedzin są w zasadzie bardzo podobne.

W przypadku rozpoznawania mówcy zależnego od wypowiadanego tekstu czy rozpoznawania mówcy z generowanym tekstem informacja semantyczna wykorzystywana jest jedynie do określenia zakresu badanych głosek czy zapobieganiu problemowi żywotności. Informacja ta zatem nie wpływa na postać stosowanych technik rozpoznawania mówcy, a jedynie na optymalny ich dobór - ujawnia kontekst użycia. Innym przykładem tego typu

jest zastosowania technik rozpoznawania treści języka naturalnego m. in. w odmianach omawianych systemów opartych na nagromadzonej wiedzy (*ang. knowledge-based systems*), których zadaniem jest jedynie wzmocnienie procesu weryfikacji oraz zapobieganie wystąpienia problemu żywotności (*ang. liveness issue*) (rozdział 1.4.1).

1.4 Klasyfikacja problemu weryfikacji mówcy.

1.4.1 Weryfikacja mówcy zależna od wypowiadanego tekstu (*ang. text-dependent speaker verification*).

{liveness}

System weryfikacji mówcy który tworzy modele mówców na podstawie jednej, ustalonej frazy zawartej w dostarczonym do niego sygnale mowy jest nazywany systemem weryfikacji mówcy zależnym od wypowiadanego tekstu. System ten oczekuje, że w fazie testowania dostarczony zostanie jako próba testowa wypowiedź o takiej samej treści. Przykładem może być system autoryzacji, oczekujący zawsze na to samo hasło. Problem tego typu jest zdecydowanie łatwiejszy do realizacji i w wyniku można od niego oczekiwać dużo lepszych wyników w porównaniu do innych rodzajów problemu weryfikacji mówcy. Inną zaletą takiego systemu jest bardzo krótki etap uczenia - wymaga się od weryfikowanego użytkownika dostarczenia jednej lub paru próbek stałego, wypowiadanego hasła. Oczywistym niebezpieczeństwem dla takiego systemu jest tzw. problemem żywotności (ang. liveness problem). Problem żywotności polega na możliwości oszukania działającego systemu weryfikacji mówcy poprzez dostarczenie na wejście spreparowany sygnał mowy - na przykład wysokiej jakości nagranie weryfikowanego mówcy, edytowane w odpowiedni sposób. Wraz z rozwojem technik audio zmylenie systemu niezabezpieczonego ze względu na ten typ ataku staje się coraz łatwiejsze. Istnieje szereg metod pozwalających na detekcję tego czy dostarczony fragment mowy pochodzi od prawdziwego mówcy - jednak takie systemy okazują się niewystarczające w przypadkach ochrony cennych danych lub krytycznych zasobów. Dobrą alternatywą dla takiego systemu może być system z wyświetlanym hasłem (1.4.3). Standardem w implementacji tego typu systemów jest użycie zastosowanie modelowania tzw. ukrytymi modelami Markowa (HMM - *ang. Hidden Markov Model*) ze względu na możliwość modelowania sekwencji zdarzeń - z czym mamy do czynienia w rozpatrywanym problemie. Z tego też powodu weryfikacja mówcy zależna od wypowiadanego tekstu jest najbardziej zbliżonym problemem do rozpoznawania mowy.

1.4.2 Weryfikacja mówcy niezależna od wypowiadanego tekstu (*ang. text-independent speaker verification*).

System weryfikacji mówcy dokonywujący weryfikacji mówcy bez względu na zawartość lingwistyczną wypowiedzi nazywa się systemem weryfikacji mówcy niezależnym od wypowiadanego tekstu. Taki system dokonuje nie dokonuje żadnych założeń co do treści wypowiedzi, dlatego też w fazie trenowania potrzebuje więcej materiału (różnorodnych wypowiedzi) od modelowanego mówcy. Problem ten jest zdecydowanie trudniejszy do realizacji w porównaniu do wcześniej omawianego. Nie jest możliwe zastosowanie pewnych upraszczających założeń jeżeli chodzi o model fizyczny aparatu głosowego człowieka - nie jest znane pobudzenie modelu - tak jak to czyni się w problemie ze znanym tekstem. Dla tego zagadnienia również towarzyszy problem żywotności - jednak nie przyjmuje takiej skali jak we wcześniej omawianym, łatwiej jest skonstruować odpowiednie systemy detekcji spreparowanej wypowiedzi. Ten typ weryfikacji cechuje także uniwersalność użycia -

nie jest wymagana kolaboracja weryfikowanego mówcy, zatem ten typ weryfikacji używany jest w systemach identyfikacji mówcy na otwartym zbiorze 1.2. Nijako standardem dla tego typu systemów stała się kombinacja technik ekstrakcji cech - melowych współczynników cepstralnych (MFCC) i mikstur Gaussowskich (GMM) będących niesekwencyjnym odpowiednikiem techniki HMM. Innym popularnym przykładem techniki modelowania jest kwantyzacja wektorów (VQ). Wskazane techinki są opisane dalej w tej pracy.

1.4.3 Weryfikacja mówcy z wyświetlanym hasłem (*ang. text-prompted speaker verification*).

{prompted}

System weryfikacji mówcy w którym w celu dokonania weryfikacji wyświetlany jest dla użytkownika tekst, który musi wypowiedzieć nazywany jest systemem weryfikacji mówcy z wyświetlanym hasłem. Jest to problem podobny do problemu weryfikacji zależnym od tekstu z rozszerzonym modelem mówcy o kolejne wypowiadane frazy. W tym wypadku czas trwania sesji treningowej znajduje się gdzieś pomiędzy długościami obu poprzednich. Tego typu system oferuje ochronę przed problemem żywotności podobną do systemu weryfikacji niezależnej od tekstu.

1.5 Sygnał mowy.

1.5.1 Produkcja sygnału mowy.

Sygnał mowy jest sygnałem akustycznym, którego medium jest powietrze a informacja w nim zawarta jest rejestrowana jako chwilowe zmiany ciśnienia akustycznego. Źródłem różnicy ciśnień są płuca człowieka, a dalej za pomocą fałd głosowych jest modulowane - powstaje tzw. ton krtaniowy (*glottal pulse*) który dalej filtrowany przez charakterystyki budowy aparatu głosowego człowieka. Ton krtaniowy i wspomniane charakterystyki różnią się ze względu na każdego mówcę i ta cech jest podstawą wykorzystania w systemie biometrycznym mówcy. Tak zaproponowany model produkcji mowy może być przedstawiony jako szeregowe połączenie tych elementów[**multidsp**]:

$$S_x(f,t) = S_r(f,t) \cdot |H(f,t)|^2 \tag{1.1}$$

gdzie $S_x(f,t)$ reprezentuje transmitancję sygnału mowy, $S_r(f,t)$ transmitancję tonu krtaniowego oraz H(f,t) funkcję transmitancji dróg głosowych. Charakterystyki dróg głosowych są zmienne w czasie i zależą od stanu 4-5 komór rezonansowych znajdujących się w aparacie głosowym człowieka.

Rozpatrywany sygnał głosowy cechuje się pasmem sygnału o szerokości nawet 8 kHz. Z tego powodu spróbkowany sygnał mowy cechuje się stosunkowo wysoką zawartością informacji. Z punktu widzenia rozpoznawania mówcy sygnał mowy zawiera w wysokim stopniu informację redundantną. Z tego powodu stosuje się parametryzację mowy w procesie ekstrakcji cech (1.7.1).

Człowiek dysponuje doskonałymi narzędziami do przeprowadzenia procesów rozpoznawania mowy oraz rozpoznawania mówcy. Analiza oraz zrozumienie mechanizmów powstawania mowy u człowieka dostarcza podstaw do tworzenia modelu produkcji mowy i umożliwia budowę takich systemów jak systemy syntezy mowy. Podobna analiza systemu percepcji mowy na który składa się aparat słuchowy oraz układ nerwowy związany z dekodowaniem sygnału mowy daje podstawy do identyfikacji cech którymi posługuje się ludzki organizm do efektywnego przeprowadzenia zadania rozpoznania mówcy.

Układ produkcji mowy oraz jej percepcji są ze sobą nierozerwalnie związane. Sposób ekstrakcji informacji przez narząd słuchu odzwierciedla fizjologię produkcji mowy - zatem może wskazać najważniejsze cechu sygnału mowy dla naturalnego procesu rozpoznania mówcy. Dlatego wydaje się właściwe prześledzenie związków pomiędzy tymi dwoma elementami.

Ludzki układ percepcji dokonuje rozróżnienia sygnałów audio poprzez rozróżnienie trzech własności: wysokości dźwięku, głośności oraz barwy dźwięku - tembru.

1.5.2 Aparat słuchowy.

Narząd słuchu człowieka można rozpatrywać jako transduktor, co znaczy, że mapuje zmiany ciśnienia akustycznego w powietrzu na sygnał elektryczny w układzie nerwowym. Na samym początku toru przetwarzania sygnału audio znajduje się małżowina uszna, której zadaniem jest skupienie dźwięku. Sygnał akustyczny wpadający do kanału słuchowego jest filtrowany ze względu na jego fizyczne rozmiary, usuwane są niskie częstotliwości. Zmiany ciśnienia akustycznego zamieniane są na fale mechaniczne w ciele stałym na błonie bębenkowej, a następnie wzmacniane przez układ kosteczek słuchowych - młoteczka, kowadełka i strzemiączka. Strzemiączko łączy się z uchem wewnętrznym poprzez błonę okienka owalnego (łac. *łac. fenestra vestibuli*), które jest wejściem do ślimaka, który z kolei jest częścią narządu Cortiego. Ruch membrany okienka owalnego powoduje ruch płynu nazywanego perylimfą w ślimaku. W taki sposób powstaje fala rozchodząca się w obszarze ślimaka propaguje się do szczytu tego narządu i następnie kanałem połączonym wraca w kierunku okienka okrągłego. Błona podstawna oddziela oba kanały od siebie i w zależności od długości fali w cieczy jest wyginana w innym obszarze - jest aparatem analizy częstotliwościowej sygnału. Schemat działania z rejonami które są pobudzane dla konkretnych częstotliwości przedstawia rysunek 1.1. Znajdujące się na błonie podstawnej

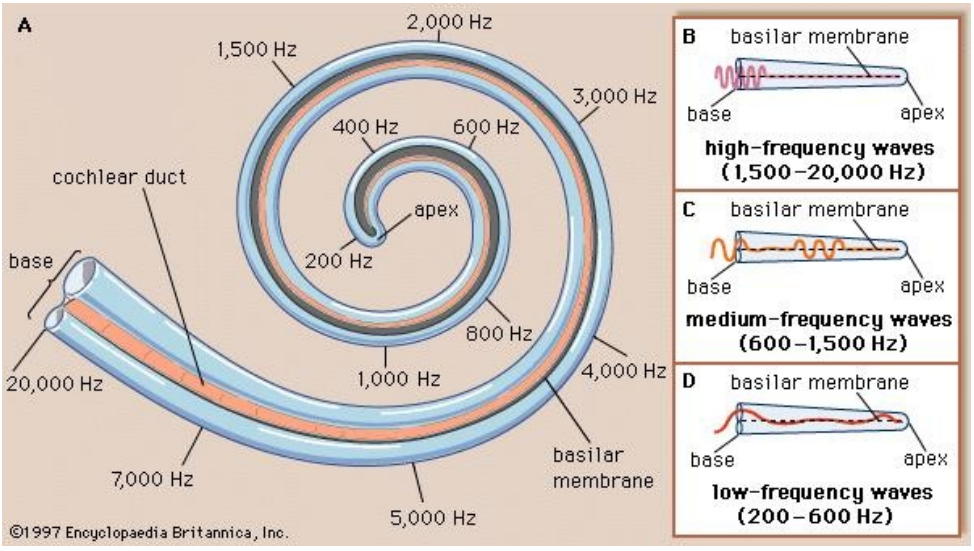


fig:ucho
Rysunek 1.1 Schemat budowy ślimaka. Źródło [**enciclopidiabritanica**].

rzęski zagłębione, ustawione w rzędzie, wzdłuż ślimaka, zamieniają wychylenie błony na impulsy elektryczne wysyłane dalej przez nerw przedsionkowo-ślimakowy do mózgu. Ze względu na różnią sztywność błony podstawnej, narząd Cortiego w różny sposób interpretuje intensywność danej częstotliwości. Inne trzy równoległe rzędy rzęsek znajdujących dostarczają sprzężenia zwrotnego ze strony mózgu pozwalając na zwiększenie rozdzielczości przeprowadzonej analizy częstotliwościowej. Rzęski zewnętrzne pobudzane są głównie

w rejonie największego ugięcia błony i poprzez to pobudzenie 'dostrajają' się do słyszanej częstotliwości.

Wykorzystanie wiedzy na temat budowy aparatu słuchowego przyniosło zdefiniowanie skal perceptualnych skal psychoakustycznych

Skala Mela. {mel}

Skala Mela jest nieliniową skalą częstotliwości powstałą poprzez subiektywne badanie psychoakustyczne. Relację pomiędzy skalą Mela a liniową skalą częstotliwości przybliża się zwykle wzorem:

$$m = 2595 \log_{10} \left(1 + \frac{f}{700}\right), \tag{1.2}$$

gdzie m oznacza wartość częstotliwości w Melach.

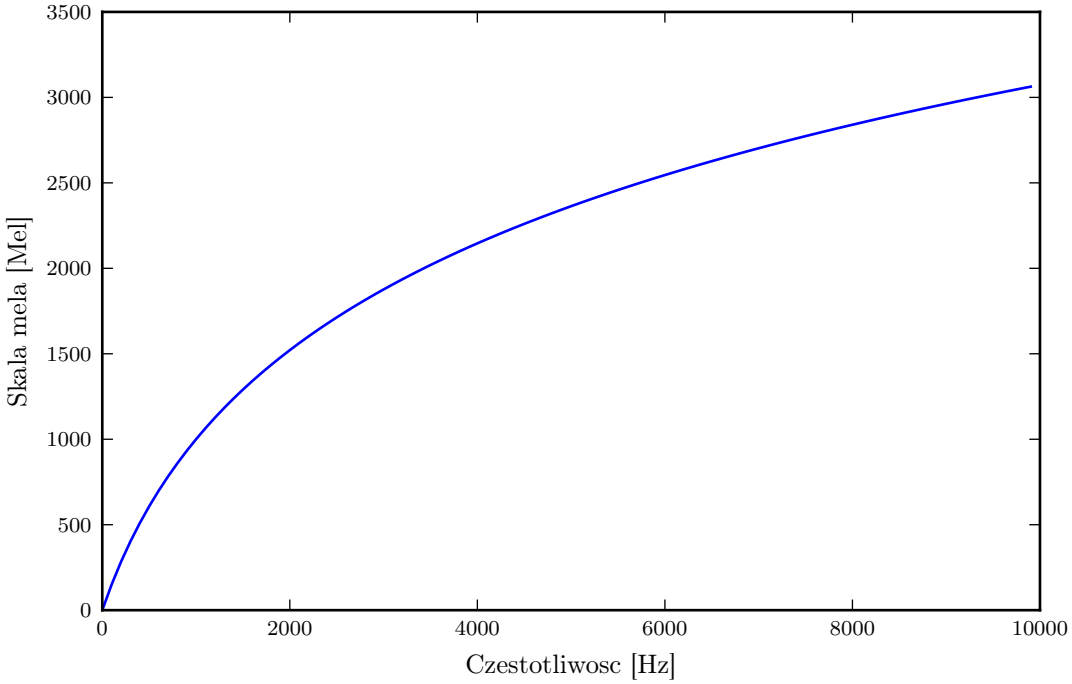


fig:mel
Rysunek 1.2 Relacja pomiędzy skalą Mela i liniową skalą częstotliwości.

Skala Barka. {bark}

Inną skalą perceptualną opartą na modelu ślimaka jako banku liniowych filtrów w skali częstotliwości jest skala Barka. Wyróżnia ona 24 pasma krytyczne. Zakłada się, że 1 bark odpowiada 100 melom, zaś funkcją przekształcenia ze skali liniowej jest:

$$B = 13 \arctan\left(0.00076 \cdot f\right) + 4.5 \arctan\left(\frac{f}{7500}^2\right), \tag{1.3}$$

gdzie B oznacza wartość w Barkach.

Krzywe jednakowej głośności

Ze względu na różną sztywność błony podstawnej (tak jak wskazano wczesniej w tym rozdziale) człowiek z inną intensywnością odbiera bodźce dźwiękowe związane z różnymi częstotliwościami. Z pomiarów subiektywnych na większej populacji wyznaczone zostały krzywe jednakowej głośności. Stosowane są one w niektórych technikach ekstrakcji cech (np. 1.6) we wstępnym wzmocnieniu sygnału mowy.

Układ nerwowy.

Dziedzina rozpoznawania głosu jest szczególnie zainteresowana naturalną aparaturą w którą wyposażony jest człowiek dla zadania rozpoznawania mowy czy identyfikowania swoich rozmówców. Dotyczy to także całego obszaru metod rozpoznawania sygnałów akustycznych. Trend jest szczególnie widoczny związku z bardzo dynamicznym rozwojem dziedziny sztucznych sieci neuronowych i sztucznej inteligencji. Tak jak wspomniano wyżej, szereg metod związanych z rozkodowywaniem sygnału mowy - ekstrakcją cech - korzysta z wiedzy anatomicznej przy okazji konstrukcji algorytmów. Niestety w dziedzinie klasyfikacji mówców to podejście nie jest stosowane - prawdopodobnie ze względu na skromną wiedzę o tym jak funkcjonuje ludzki umysł. Wydaje się jednak, że można przywołać pewną wiedzę na temat tego jak człowiek interpretuje sygnał mowy poprzez wskazanie pewnych obszarów mózgu odpowiedzialnych za zadania związane z analizą mowy - a także połączenia pomiędzy nimi. Wraz z zastosowaniem rezonansu magnetycznego (MRI) neurologia poczyniła duże postępy i można mieć nadzieję na to, że w najbliższej przyszłości uzyskana wiedza pozwoli na udoskonalenie systemów rozpoznawania głosu.

1.5.3 Klasyfikacja języka.

{formant}

Na rysunku 1.3 znajduje się przykładowe widmo pojedynczej ramki sygnału. Obszary koncentracji energii w dziedzinie częstotliwości nazywane są formantami.

Fonetyka

Fonetyka zajmuje się badaniem dźwięków produkowanych przez aparat mowy człowieka. Elementarnym dźwiękiem rozpatrywanym przez fonetykę jest głoska. Z punktu widzenia całej lingwistyki jest to najmniejszy segment mowy. Budowa i funkcjonalność narządu mowy determinują zakres produkowanych głosek. Fonetyka bada podstawowe dźwięki mowy bez rozróżniania ze względu na konkretny język czy znaczenie głoski. Fonem jest najmniejszą jednostką mowy na podstawie której możliwa jest interpretacja jej znaczenia. To znaczy, że jest semantycznie istotna. Fonem rozpatruje się ze względu na znaczenie w konkretnym języku. Głoska może być realizacją fonemu. Dwie różne głoski mogą stanowić realizację tego samego fonemu - to znaczy nieść tę samą informację semantycznie. Zbiór takich głosek realizujących dany fonem nazywany jest alofonem.

Z punktu widzenia mechanizmów powstawania dźwięków w ludzkim narządzie mowy, można wyróżnić trzy z których zbudowany jest każdy emitowany dźwięk mowy. Składa się na nie: - dźwięk rezonujący powstały w wibrującym źródle (np. drgające fałdy głosowe) i rezonujący w przestrzeni rezonansowej na którą składają się drogi oddechowe znajdującą się powyżej krtani, - dźwięk powstały przez nielaminarny przepływ powietrza, - dźwięk impulsowy powstały przez energiczne wypuszczenie powietrza z układu oddechowego.

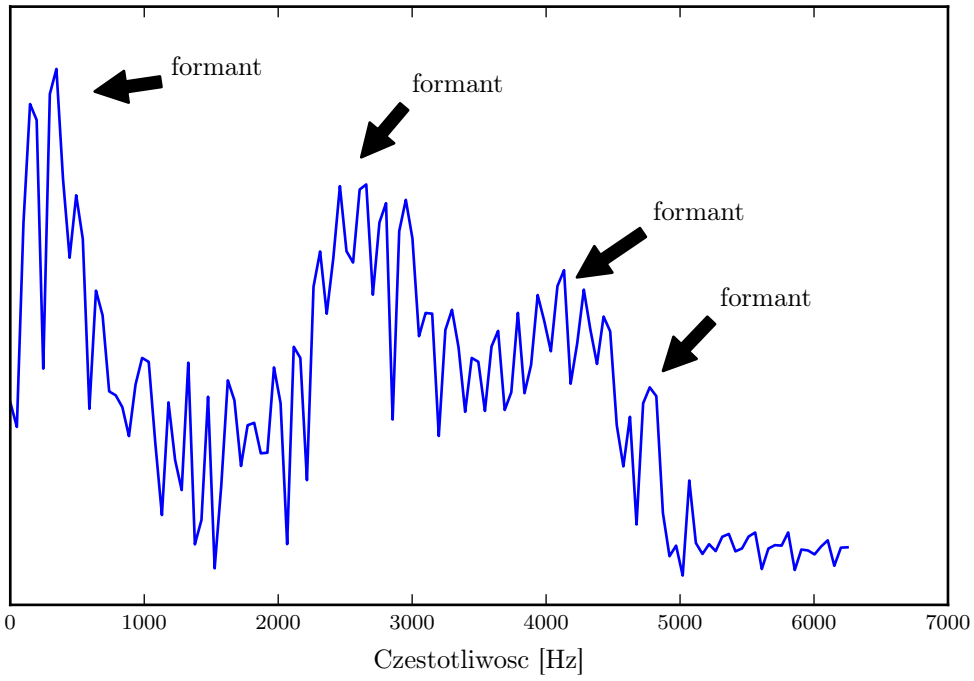


fig:spectrum
Rysunek 1.3 Przykładowe spektrum pojedynczej ramki sygnału mowy.

1.6 Wstępne przetwarzanie sygnału mowy.

1.6.1 Preemfaza.

1.6.2 Redukcja szumu.

1.6.3 System detekcji mowy.

1.7 Ekstrakcja cech sygnału mowy.

1.7.1 Przegląd.

extraction}

Celem ekstrakcji cech z sygnału mowy (*ang. feature extraction*) jest uzyskanie zbioru cech charakteryzujących sygnał ludzkiej mowy za pomocą technik cyfrowego przetwarzania sygnału. Jednocześnie jest to zamiana sygnału w którym zawarta jest redundantna informacja na sygnał o niskiej zawartości informacji znaczących dla problemu rozpoznawania mówcy/mowy.

Tak jak zostało wspomniane wcześniej, problem rozpoznawania mówcy w standardowym podejściu rozpatrywany jest jako problem estymacji parametrów ustalonego modelu. W zależności od rozpatrywanej odmiany rozpoznawania mówcy dokonywane są odpowiednie założenia konkretyzujące postać problemu.

W ogólności system produkcji mowy człowieka można przedstawić jako układ regulacji automatyki z kontrolerem reprezentującym układ nerwowy wraz z kontrolowaną przezeń motoryką aparatu mowy. Sygnał sterujący kontrolera powoduje pobudzenie układu charakterystyk sygnału mowy charakteryzowany przez budowę dróg głosowych. Dopiero wyjściem tego ostatniego jest akustyczny sygnał mowy. Podczas gdy problem rozpoznawania mowy próbuje wyeliminować wpływ na formułowany model postaci układu charaktery-

styk głosowych, tak rozpoznawanie niezależne od tekstu próbuje ustalić model charakterystyk głosowych bez względu na postać układu kontrolera. Ponieważ oba rozpatrywane modele są mocno nieliniowe najtrudniejszym zadaniem w rozpoznawaniu mówcy jest ich rozdzielenie. Z tego właśnie powodu problem ze znanym wypowiadanym tekstem jest o wiele prostszym zagadnieniem od problemu rozpoznawania mówcy niezależnego od wypowiadanego tekstu. W tym pierwszym przypadku dokonujemy redukcji do tylko znanych pobudzeń oraz konkretyzować model kontrolera G_c.

Przedstawione w tym podrozdziale techniki ekstrakcji cech okazują się równie przydatne i powszechnie stosowane zarówna dla technik rozpoznawania mówcy jak i rozpoznawania mowy. Najpopularniejszą obecnie stosowaną metodą ekstrakcji cech z sygnału mowy jest współczynnikami cepstrum w dziedzinie częstotliwości Mela - MFCC (*ang. Mel-Frequency Cepstral Coeficcients*).

Ważnym elementem toru przetwarzania systemu są również elementy związane z zaszumieniem sygnału mowy, przetwornikiem elektroakustycznym oraz procesem dyskretyzowania elektrycznego sygnału mowy przez przetwornik analogowo-cyfrowy. Jednak ze względu na prostotę te zagadnienia zostaną omówione oddzielnie od problemu ekstrakcji cech.

1.7.2 MFCC.

Ekstrakcja cech za pomocą współczynników cepstrum w dziedzinie częstotliwości Mela składa się z dwóch głównych etapów: uzyskania współczynników mocy w dziedzinie częstotliwości Mela na podstawie estymacji widma mocy sygnału mowy (*freqency warping*) oraz obliczenia współczynników cepstrum na podstawie uzyskanych wcześniej współczynników mocy w skali Mela. Wynikiem przeprowadzonej operacji jest uzyskanie wektora cech x_j dla numeru ramki j.

DSTFT - dyskretna krótkookresowa transformata Fouriera

Dyskretna krótkookresowa transformata Fouriera (DSTFT - *ang. Discrete Short-time Fourier Transform*) jest dyskretną wersją ciągłej transformaty Fouriera (STFT - *ang. Short-time Fourier Transform* dla sygnału próbkowanego w czasie oraz dyskretnego w dziedzinie częstotliwości. Jest to metoda analizy czasowo-częstotliwościowej, przeprowadzając oparację dyskretnej transformaty Fouriera (DFT - *ang. Discrete Fourier Transform*) na ramkach sygnału dla których możemy założyć jego lokalną stacjonarność. Relacja pomiędzy STFT, a DSTFT jest relacją analogiczną do relacji ciągłej transformaty Fouriera, a DFT. Estymacja widma za pomocą metody DSTFT jest podstawą techniki ekstrakcji cech MFCC.

Podział sygnału na ramki.

Ze względu na własność quasi-stacjonarności sygnału mowy, aby wydobyć informację dotyczącą wypowiadanej głoski konieczne jest rozdzielenie sygnału na ramki. Liczbę próbek dla pojedynczej ramki wybiera się ze względu na konieczność uzyskania lokalnej stacjonarności - w taki sposób aby było możliwe zbadanie charakterystyki częstotliwości pojedynczej głoski. Średnia długość głoski to 80 ms. Jednak trzeba mieć na względzie, że samogłoski trwają długo w stosunku do przerw pomiędzy (trwających zwykle ok. 5 ms). Zatem aby móc uchwycić krótsze głoski oraz przerwy zwykle ustala się długość ramki na 20 do 30 ms. Jednocześnie długość ramki w ilości próbek jest funkcją częstotliwości próbkowania. Do przedstawionych założeń dochodzi zwykle warunek dotyczący użycia

algorytmu szybkiej transformaty Fouriera (FFT - *ang. Fast Fourier Transform*) wymagającej aby sygnał był długości potęgi dwójki. W przypadku pierwszej oraz ostatniej ramki stosowana jest technika uzupełniania zerami w przypadku braku próbek do zapełnienia całej ramki.

Aplikacja okna na ramki.

W wykorzystywanej technice DSTFT stosowane są zazwyczaj okna inne niż okno prostokątne o takiej samej długości jak ramki w celu zredukowania przecieku widma. Zastosowanie okna czasowego odbywa się poprzez przemnożenie wartości próbek ramki przez wartości okna w następujący sposób:

$$h_i = w_i \cdot x_i, \quad i \in 1, ..., N \tag{1.4}$$

gdzie x_i oznacza i-tą próbkę sygnału, h_i - zmodyfikowaną wartość próbki użytej dla wyznaczania współczynników DFT, w_i - i-tą wartości funkcji okna, N ilość próbek w ramce.

Najpopularniejszymi oknami stosowanymi w tym przypadku są okna czasowe Hanna, Hamminga oraz okna Gaussowskie [**fosr**].

Estymacja widma.

{stft}

Po zaaplikowaniu okien czasowych dla kolejnych ramek sygnału, kolejnym etapem procedury obliczania współczynników MFCC jest znalezienie widma ramki za pomocą dyskretnej transformaty fouriera (DFT). Kolejne współczynniki uzyskanego widma zdefiniowane są następująco:

$$H_k = \sum_{n=0}^{N-1} h_n exp(\frac{-2\pi kn}{N}) \tag{1.5}$$

W omawianej metodzie korzysta się jedynie z informacji o amplitudzie uzyskanego widma. Kolejne współczynniki widma amplitudowego zdefiniowane są w następujący sposób:

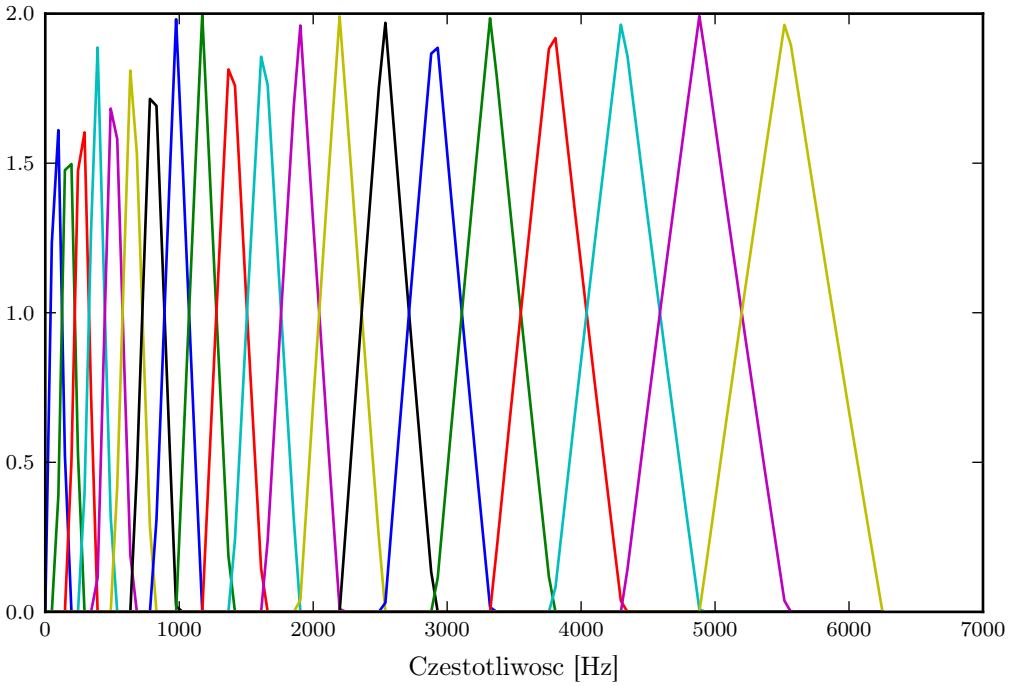
$$M_k = |H_k| = \sqrt{H_k^2} \tag{1.6}$$

Przejście do dziedziny częstotliwości Mela.

Przejście do dziedziny częstotliwości Mela (*frequency warping*) uzasadnione jest nieliniową percepcją układu słuchowego człowieka. Ponieważ procedura DFT produkuje liniowe widmo w dziedzinie widmo częstotliwościowe, dokonywana jest jego transformacja do dziedziny częstotliwości Mela. Jest to proces imitujący cechę sygnału akustycznego - wysokości tonu. Rozpatrywana operacja polega na zastosowaniu banku filtrów trójkątnych na uzyskanych amplitudach widma kHz:

$$\tilde{H}_j = \sum_{k=1}^{K} H_k \cdot F_k, \quad k \in 1...N, \quad j \in 1...M \tag{1.7}$$

gdzie $\tilde{H}_j$ oznacza j-ty współczynnik amplitudy widma w dziedzinie częstotliwości Mela, F_k, zaś M oznacza ilość filtrów. W wyniku tej operacji otrzymuje się wektor współczynników częstotliwości w skali Mela o długości równej ilości użytych filtrów. Najczęściej stosowanym filtrem jest okno trójkątne zdefiniowane w dziedzinie częstotliwości. Każdy kolejny filtr zaczyna się w środku pasma wcześniejszego. Ponieważ przetwarzany jest sygnał składający się z próbek o wartościach rzeczywistych, pokryte pasmo częstotliwości



{melfb}

Rysunek 1.4 Przykładowe 24 trójkątne filtry dla sygnału o częstotliwości próbkowania *fs=12500 Hz.*

zawiera próbki od 0-wej do próbki o oznaczonej numerem N/2. Ilość filtrów oraz ich rozmieszczenie w dziedzinie częstotliwości ustalane są na jeden z kilku sposobów. Pierwszym sposobem jest skorzystanie z 24 obszarów krytycznych zdefiniowanych w skali Barka. Każdy filtr ma niezerowe wartości w zakresie częstotliwości od k-1 częstotliwości krytycznej, aż do k+1 częstotliwości krytycznej. Maksimum zaś przypada na k-tą częstotliwość krytyczną. Drugim i najpopularniejszym podejściem jest zastosowanie ustalonej ilości filtrów i rozmieszczenie ich równomiernie w dziedzinie częstotliwości mela.

Stosowane są również filtry trapezowe oraz kosinusowe, a także zmianę wzmocnienia filtrów ze względu na psychoakustyczną krzywą jednakowej głośności.

Uzyskanie współczynników głośności.

Wartości otrzymane w procesie przejścia do dziedziny częstotliwości Mela z prążków reprezentujących widmową gęstość mocy sygnału mogą reprezentować natężenie dźwięku w danym momencie czasu. Ponieważ zadaniem ekstrakcji cech przy pomocy współczynników MFCC jest mimika ludzkiego układu percepcji konieczne jest przekształcenie wartości reprezentujących energię na głośność (*ang. magnitude warping*). Zgodnie z przedstawioną definicją **??** *głośność* jest funkcją zarówno *wysokości tonu* oraz natężenia dźwięku I. Relację z *wysokością tonu* imituje się za pomocą etapu *equalizacji* sygnału. Zatem pozostaje jedynie zdefiniować współczynniki głośności C_k [**fosr**] dla każdej ramki sygnału:

$$\tilde{C_k} = 10 \cdot \left(\frac{|\tilde{H_k}|^2}{I_0} \right) \tag{1.8}$$

Ponieważ normalizacja współczynników w postaci odniesienia wartości współczynników widma mocy do natężenia odniesienia I_0 nie wpływa na efekt identyfikacji mówcy

dlatego w praktyce stosuje się dla wygody postać (szczególnie ze względu na pokrywanie się tych wartości z elementami przetwarzanymi w procedurze otrzymywanie cepstrum sygnału 1.7.2):

$$C_k = log\left(|\tilde{H}_k|^2\right) \tag{1.9}$$

Przy użyciu tych współczynników otrzymuje się cepstrum sygnału.

Cepstrum. {cepstrum}

Funkcja cepstrum mocy dla ciągłego przekształcenia Fouriera zdefiniowana jest następująco [**fosr**]:

$$\tilde{h_{pc}} = \left[\frac{1}{2\pi}\int_{-\pi}^{\pi} log\left(|H(\omega)|^2\right)e^{i\omega t}\right]^2 \tag{1.10}$$

gdzie $|H(\omega)|^2$ oznacza funkcję widma gęstości mocy sygnału oryginalnego. Jest to odwrotne przekształcenie Fouriera zastosowane na logarytmie funkcji gęstości mocy sygnału.

Pojęcie cepstrum pierwotnie zostało wprowadzone do badań nad echem sygnałów akustycznych [**hdsp**]. Taki sygnał opisywany jest przez wyrażenie

$$x(t) = s(t) + \alpha s(t-\tau) \tag{1.11}$$

gdzie α jest współczynnikiem osłabienia sygnału echa przesuniętego w czasie o τ względem sygnału oryginalnego. Reprezentacja częstotliwościowa zaś przyjmuje w takim wypadku postać

$$|S(f)|^2\left[1+\alpha^2+2\alpha cos(2\pi f\tau)\right]. \tag{1.12}$$

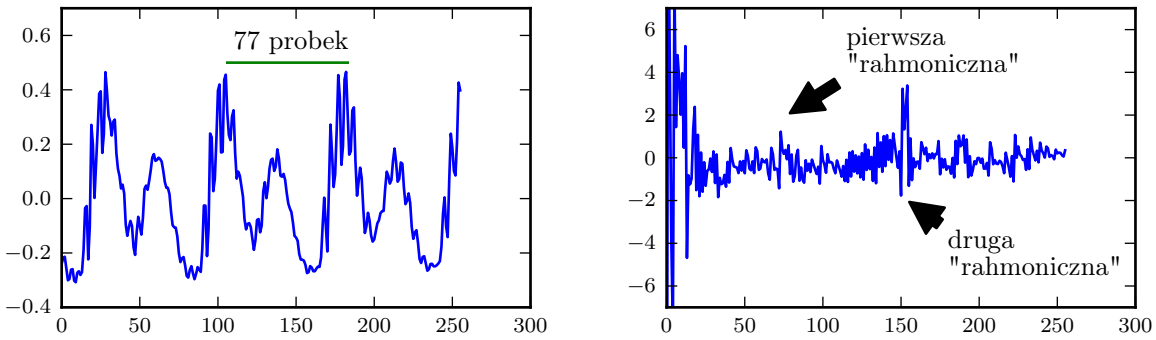
Jak widać z tej postaci, widmo sygnału oryginalnego $S(f)$ modulowane jest przez zmienny w częstotliwości komponent $[1+\alpha^2+2\alpha cos(2\pi f\tau)]$. Znajomość tego w jaki sposób modulowana jest obwiednia widma pozwala na określenie współczynników α oraz τ definiujących sygnał echa. Ponieważ problem taki jest doskonale opisany w dziedzinie czasu przy analizie sygnałów zmodulowanych, narzucającym się rozwiązniem jest skorzystanie z tych znanych już technik. Korzystając z własności logarytmu możliwe jest rozdzielenie dwóch komponentów - zamianę mnożenia na dodawanie, a poprzez zastosowanie odwrotnego przekształcenia Fouriera otrzymuje się sumę dwóch funckji w dziedzinie tzw. "quefrency":

$$\begin{aligned}\mathcal{F}^{-1}\{log(X(f))\} = \mathcal{F}^{-1}\{log(S(f))\} + \mathcal{F}^{-1}\{log(1+\alpha^2)\}+ \\ +\mathcal{F}^{-1}\{log(1+\frac{2a}{1+a^2}cos(\omega\tau))\}.\end{aligned} \tag{1.13}$$

Ostatni składnik powyższej sumy objawia się w postaci widocznego w cepstrum skupionego impulsu którego położenie w skali cepstrum określa przesunięcie τ zaś jego amplituda jest związana z czynnikiem osłabienia α.

Okazuje się, że problem wykrycia wysokości tonu (*ang. pitch detection*) sygnału mowy jest podobny do problemu analizy echa sygnału akustycznego. Zaproponowana została więc wersja wykorzystująca cepstrum dla analizy STDFT[**noll**]. Tak jak wspomniano wcześniej w tym rozdziale %TODO ODNOŚNIK prosty model produkcji sygnału mowy zakłada, że ten jest wynikiem splotu odpowiedzi impulsowej dróg głosowych h_1 oraz quasi-okresowym sygnałem impulsowym h_2 (*ang. quasi-periodic pulse train*) - wymuszeniem produkowanym przez głośnię:

$$h(t) = h_1(t) * h_2(t) \tag{1.14}$$



e_cepstrum}

Rysunek 1.5 Ramka sygnału wypowiadanej samogłoski 'i' oraz jej cepstrum.

Z własności ciągłego przekształcenia Fouriera wynika, że powyższe równanie w dziedzinie częstotliwości przybiera postać iloczynu transformat $H_1(\omega) \cdot H_2(\omega)$. W dziedzinie cepstrum rozdziela się natomiast na sumę składników:

$$\tilde{h}(t) = log(H_1(\omega)) + log(H_2(\omega)) \tag{1.15}$$

Poprzez obliczenie cepstrum sygnału mowy otrzymujemy zatem rozdzielenie komponentów pochodzących od funkcji $h_1(t)$ i $h_2(t)$. Proces ten nazywany jest również dekonwolucją (*homomorphic deconvolution*)[**hdsp**] i zobrazowany jest na rysunku 1.5. Niskie współczynniki cepstrum opisują charakterystykę aparatu głosowego, zaś widoczne w okolicach próbki 77 i 150 piki, stanowią harmoniczne w dziedzinie cepstrum (*ang. rahmonics*) i świadczą o okresie sygnału wymuszenia głośni (h_2), którą można odczytać z pierwszego wykresu jako odległość w próbkach pomiędzy najwyższymi szczytami - 77 próbek.

Należy zwrócić uwagę, że wykorzystywany w tej metodzie jest logarytm z liczb rzeczywistych oraz współczynniki mocy widma. Spowodowane jest to, że podczas procesu rozpoznawania mówcy nie interesuje nas rekonstrukcja sygnału, anie nie korzysta się z informacji zachowanej w fazie sygnału. Z tego też powodu klasycznie w ostatnim etapie obliczania współczynników MFCC zamiast odwrotnej dyskretnej transformaty Fouriera (IDFT) stosuje się dyskretną transformację kosinusową (DCT). Kolejną korzyścią jest otrzymanie w wyniku przeprowadzenia tej transformacji współczynników będącymi liczbami rzeczywistymi. W tej pracy korzysta się z następującej definicji dyskretnej transformaty kosinusowej:

$$H_k = \sum_{n=0}^{N-1} h_n cos\left(\frac{\pi(2n+1)k}{2N}\right) \tag{1.16}$$

gdzie

$$a_k = \begin{cases} 1/N & \text{dla} \quad k = 0 \\ 2/N & \forall \quad k > 0. \end{cases} \tag{1.17}$$

Uzyskane współczynniki MFCC noszą nazwę wektora akustycznego (*ang. acoustic vector*) reprezentującą cechy pojedynczej ramki i są w procesie weryfikacji mówcy dalej wykorzystywane w tzw. procesie dopasowywania cech opisanym w sekcji 1.8.

W wektorze cech zwykle stosuje się od 12 do 15 najniższych współczynników uzyskanych z DCT. Zwykle uzupełnia się go dodatkowymi współczynnikami Δ oraz Δ^2 zawierające dodatkową informację o dynamice sygnału.

Współczynniki Delta cepstrum oraz Delta-Delta cepstrum

Tak jak opisano dalej w tej pracy w 1.7.4, duże znaczenie w procesie rozpoznawania mówcy ma informacja temporalna zawarta w sygnale mowy - związana z dynamiką zmian współczynników kolejnych ramek wektorów cech. Najpopularniejszą metodą na dodanie informacji tego rodzaju jest dołączenie do wektora cech MFCC współczynników *Delta* - Δ oraz *Delta-Delta* - Δ^2 które kolejno reprezentują pierwszą i drugą pochodną dyskretną dla kolejnych ramek sygnału. Okazuje się, że dodanie wspomnianych współczynników daje znaczącą poprawę procesu rozpoznawania, szczególnie w systemach rozpoznawania mówcy z hasłem lub zależnego od wypowiadanego tekstu [**delta**]. Najczęściej, stosuje się następującą formułę do obliczenia cech delta cepstrum:

$$\Delta_{t,k} = \frac{\sum_{n=1}^{N} n(c_{t+n,k} - c_{t-n,k})}{2\sum_{n=1}^{N} n^2} \tag{1.18}$$

gdzie N przyjmuje typową wartość 2, $C_{t,k}$ oznacza współczynnik cepstrum dla ramki t i numerze współczynnika w ramce k. Współczynniki Delta-Delta obliczane są w ten sam sposób za wyjątkiem użycia współczynników Delta w miejsce współczynników cepstrum:

$$\Delta^2_{t,k} = \frac{\sum_{n=1}^{N} n(\Delta_{t+n,k} - \Delta_{t-n,k})}{2\sum_{n=1}^{N} n^2} \tag{1.19}$$

Do wektora cech MFCC dodaje się zwykle taką samą ilość współczynników Δ oraz Δ^2 co współczynników cepstrum c_k. W systemach rozpoznawania mówcy bez znajomości wypowiadanego tekstu proponuje się ograniczyć ilość współczynników delta w stosunku do współczynników cepstrum i współczynników delta-delta w stosunku do współczynników delta [**fosr**].

1.7.3 LPCC

{LPCC}

Współczynniki liniowego kodowania predykcyjnego cepstrum - LPCC (*ang. linear predicting ceptstral coding*) są alternatywną metodą ekstrakcji cech z sygnału mowy. Jest to synteza metody kodowania LPC wraz z zastosowaniem współczynników cepstrum. Liniowe kodowanie predykcyjne (LPC) jest techniką opartą na zastosowaniu modeli autoregresyjnych (AR). Główną różnicą w stosunku do techniki MFCC jest więc sposób aproksymacji widmowej gęstości mocy przy pomocy współczynników $|H_k(\omega)|^2$. Kroki wstępnego przetwarzania poprzedzające estymację, łącznie z aplikacją na ramkę okna czasowego, są identyczne jak w przypadku procedury obliczania współczynników MFCC. Zamiast posługiwać się dyskretnym przekształceniem Fouriera (DFT), korzysta się z modelu autoregresyjnego, kształtując widmo przy pomocy transmitancji $H(z)$ posiadającej same bieguny (*ang. all-pole*) i jednym zerem w punkcie $z = 0$:

$$\tilde{H}(z) = \frac{\tilde{G}}{1 - \sum_{k=1}^{p} a_k z^{-k}} \tag{1.20}$$

{transar}

gdzie $a_k = \frac{\alpha_k}{\alpha_0}$, $\tilde{G} = \frac{1}{\alpha_0}$, a współczynniki $\alpha_0, ..., \alpha_p$ oznaczają współczynniki modelu AR, nazywane również w rozpatrywanym przypadku jako współczynniki predykcji - PC (*ang. predictor coefficients*).

W dziedzinie czasu modele liniowej predykcji (LP) określają zależność próbki sygnału s_n poprzez liniową kombinację wcześniejszych p próbek przeskalowanych przez współczynniki predykcji - PC:

$$s_n = -\sum_{k=1}^{p} a_k \cdot s_{n-k} + \tilde{G} \cdot u_n \tag{1.21}$$

W metodach liniowej predykcji zazwyczaj ignoruje się wartość u_n oznaczającą obecną próbkę wejściową - w przypadku mowy oznaczającą sygnał tonu krtaniowego. W ten sposób otrzymuje się wektor współczynników $a_1, ..., a_n$ modelujących danego mówcę. Popularną metodą estymacji tych wartości jest minimalizowanie wartości błędu średniokwadratowego wartości resztowych (rezydua):

$$V = \sum_n e_n^2 = \sum_n \left\{ s_n + \sum_{k=1}^{p} \alpha_k \cdot s_{n-k} \right\}. \tag{1.22}$$

Kryterium minimalizacji

$$\frac{\delta E}{\delta \alpha_i} = 0, \quad i = 1, .., p \tag{1.23}$$

daje w wyniku zależność:

{yweq}
$$\sum_{k=1}^{p} a + k \cdot \sum_n s_{n-k} s_{n-i} = -\sum_n s_n s_{n-i}, \quad i = 1, ..., p. \tag{1.24}$$

Lewa suma $\sum_n s_{n-k} s_{n-i}$ jest funkcja autokorelacji o opóźnieniu $\tau = n-k$, czyli $r(|n-k|)$, zaś analogicznie prawa jest funkcją autokorelacji o opóźnieniu $\tau = n$ tzn. $r(n)$, gdzie w obu przypadkach $i = 1, ..., p$. Układ tych równań nazywany jest równaniami Yule'a-Walkera (*Yule-Walker Equations*). Otrzymana z tych równań macierz autokorelacji $\boldsymbol{R}$ przyjmuje postać macierzy Toeplitza - posiada te same wartości na poszczególnych przekątnych. Rozwiązanie problemu minimalizacji przedstawia się jako:

$$\boldsymbol{\alpha} = \boldsymbol{R}^{-1}\boldsymbol{r}, \tag{1.25}$$

gdzie wektor $\boldsymbol{r}$ jest wektorem kolumnowym prawych sum równań 1.24 tzn. autokorelacji $r(n)$, zaś wektor $\boldsymbol{\alpha}$ wektorem kolumnowym współczynników $\alpha_1, ..., \alpha_n$. Wspomniana własność macierzy $\boldsymbol{R}$ jako macierzy Toeplitza w dużym stopniu upraszcza procedurę odwracania macierzy poprzez zastosowanie rekursywnego algorytmu *Levinsona-Durbina*[**durbin**]. Możliwy do zastosowania algorytm Shura[**shur**] pozwala przyspieszyć obliczenia poprzez paralelizację. Obliczone wartości α_i reprezentują współczynniki LPC. Wraz z wartością:

$$\tilde{G} = r(0) - \sum_{i=1}^{p} a_i r(i) \tag{1.26}$$

umożliwiając aproksymację funkcji gęstości widma przy pomocy równania 1.20.

Obliczanie współczynników LPCC

W praktyce [**fosr**] jako wektory cech najczęściej używa się współczynników cepstrum obliczonych przy pomocy uzyskanych współczynników LPC. Rekursywny algorytm zaproponowany przez Rabinera i Juanga [**rabinerjuangfosr60**] oblicza współczynniki cepstrum w następujący sposób:

1. $c_0 = log(\tilde{G}^2)$,
2. $c_d = \alpha_d + \sum_{i=1}^{d-1} c_i a_{d-i}$ **dla** $1 \leqslant d \leqslant p$
3. $c_d = \sum_{i=1} d + \sum_{i=1}^{d-1} c_i a_{d-i}$ **dla** $d > p$

Rekomendowana ilość p dla obliczania współczynników LPCC powyższym sposobem to zakres od 8 do 16 [**fosr**]. Najczęściej stosuje się liczbę 8 do 20 elementów w wektorze cech.

1.7.4 Inne metody ekstrakcji cech.

Najczęściej stosowanymi metodami ekstrakcji cech są następujące: MFCC, LPCC, LSF, PLP [**overview**]. Dwie pierwsze zostały omówione wcześniej bardziej szczegółowo. W tym zaś podrozdziale pokrótce zostały omówione dwie kolejne oraz inne wybrane metody, które pojawiają się w literaturze fachowej oraz artykułach użytych jako źródła w tej pracy jako te metody które mogą mieć znaczenie w dalszym rozwoju omawianej dziedziny.

Metody bazujące na LP

Istnieje szereg innych metod opartych na współczynnikach LP. Zwykle cechuje je nieliniowa transformacja dziedziny czasu/częstotliwości oraz amplitud współczynników mocy widma, która na celu ma przybliżyć sposób funkcjonowania percepcji ludzkiego aparatu słuchowego w konstruowanym systemie.

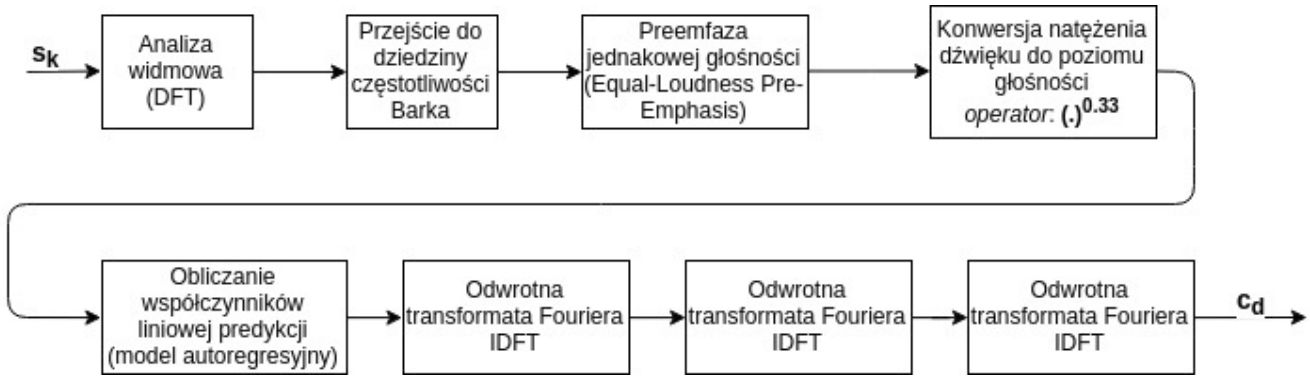
- PARCORs (*ang. partial correlation coefficients*) - *współczynniki częściowej korelacji* są współczynnikami cech κ_k interpretowanymi jako wartość reprezentującą nieredundantną informację o autokorelacji[**fosr**]. Wielkości te są obliczane jako wartości pośrednie w algorytmie obliczania współczynników LPC we wcześniej wspomnianym (1.7.3) algorytmie Levinsona-Durbina. Współczynniki te również mają interpretację fizyczną jako modelujące wartości współczynników odbicia w modelu traktu głosowego jako cylindrów o różnych średnicach przez które przepływa strumień powietrza z płuc[**flanganfosr21**]. Z tego też powodu nazywane są także *współczynnikami odbicia* (*ang. reflection coefficients*).

- LAR *ang. Log Area Ratios* - wielkości wynikające z modelu cylindrów tak jak w metodzie PARCORs. Współczynniki g_i otrzymuje się jako stosunki logarytmów średnic kolejnych cylindrów modelu. Tak obliczone współczynniki cechują się bardziej równomierną wrażliwością w dziedzinie częstotliwości niż współczynniki odbicia[**campbell**].

- PLP (*ang. Perceptual Linear Predictive analysis* - metoda ekstrakcji cech oparta na współczynnikach predykcji liniowej (LP) inspirująca się metodą MFCC. Zamiast zastosowania skali mela, w tej metodzie używa się skali Barka 1.5.2, stosuje się zmodyfikowaną preemfazę ze względu na krzywą jednakowej głośności (*ang. equal-loudness pre-emphasis*) 1.5.2 oraz dokonuje się transformacji widma amplitudowego poprzez podniesienie wartości do potęgi $\frac{1}{3}$ co powoduje wygładzenie widma - zmniejszeniem dynamiki - i jest nazywana konwersją natężenia dźwięku do poziomu głośności (*ang. Intensity to Louadness Conversion*) [**hoening**]. Po tych etapach, z tak przekształconego widma uzyskiwane są współczynniki predykcji liniowej (model AR). Tak jak w metodzie MFCC także i tutaj używa się procedury DFT i IDFT. Współczynniki cepstrum uzyskiwane są tak jak w metodzie LPCC, np. przy wykorzystaniu metody Levinsona-Durbina. Na rysunku 1.6 przedstawiono etapy analizy PLP tak jak w pracy źródłowej [**fosr30hermansky**]. {plp}

- LSF (*ang. line spectral frequencies* lub LSP (*ang. line spectral pairs*)) - metody reprezentacji sygnału mowy w których zaproponowana jest alternatywna postać mianownika transmitancji z 1.20. Oznaczając mianownik jako $A(z)$ przyjmowana jest reprezentacja:

$$\begin{aligned} A(z) &= \frac{1}{2}[P(z) + Q(z)] \\ P(z) &= A(z) + z^{p+1}A(Z^{-1}) \\ Q(z) &= A(z) - z^{p+1}A(Z^{-1}). \end{aligned} \tag{1.27}$$



Rysunek 1.6 Etapy analizy PLP.

gdzie wielomian $P(z)$ jest interpretowany fizycznie jako odpowiadający odpowiedzi impulsowej dróg głosowych przy głośni zamkniętej zaś wielomian $Q(z)$ przy otwartej. Pierwiastki wielomianów P i Q leżą na okręgu jednostkowym na płaszczyźnie zespolonej z. Techinka ta jest nazywana *ang. reverse-filtering*.

Metody wykorzystujące transformację falkową.

Tak jak wcześniej opisano w rozważaniach o krótko-czasowej transformacie Fouriera (STFT) 1.7.2, niestacjonarny sygnał mowy możemy analizować metodami czasowo-częstotliwościowymi z odpowiednio małym oknem dobranym na podstawie czasu trwania pojedynczych głosek. Segmentacja sygnału na takie ramki pozwalała zastosować metody estymacji widma dla sygnały stacjonarnego. Metody estymacji widma z zastosowaniem modeli autoregresyjnych (AR) również przyjmują, że działają na sygnale stacjonarnym. Transformacja falkowa dostarcza szeregu ortogonalnych funkcji jądra dla uogólnionego szeregu Fouriera. Wspomniane funkcje powstają poprzez rozciąganie i przemieszczanie funkcji matki dla danej transformacji falkowej. Warto zwrócić uwagę na to, że transformacja Gabora, będącą transformacją falkową z funkcją matką równą iloczynowi jądra transformaty Fouriera $exp(-2\pi jf\tau)$ oraz funkcją Gaussowską $exp(-\pi(\tau - t)^2)$ stanowi przypadek krótko-czasowej transformaty Fouriera (STFT) dla której użyto okna Gaussowkiego. Zatem widać, że istnieje pole do zastosowania analizy czasowo-częstotliwościowej w postaci transformacji falkowej w miejsce STFT czy estymacji za pomocą modeli autoregresyjnych. Odpowiednikiem dyskretnej transformaty Fouriera (DFT) jest dyskretna transformata falkowa (DWT -*ang. discrete wavelet transform*).

1. MFCWC - *Mel-Frequency Discrete Wavelet Coefficients* - technika oparta na generacji współczynników MFCC - jest praktycznie identyczna w kolejnych krokach za wyjątkiem wykorzystania transformaty cosinusowej (DCT) w zamian za użycie dyskretnej transformaty falkowej (DWT). Okazuje się, że zastosowanie DWT przyczynia się do polepszenia rezultatów w 'agresywnie' hałaśliwym środowisku[**fosr**].

2. WOCOR - *Wavelet Octave Coefficients Of Residues* - jest wektorem cech wyekstrahowanych za pomocą transformacji falkowej z funkcją matką związaną z tonem wypowiadanej głoski, co wiąże się z rozmiarem analizowanej ramki. Jest to metoda oparta na współczynnikach liniowej predykcji (LP) rezyduł sygnału w ramce. Rozważana technika jest efektywna w użytku z sygnałem mowy zawierającym języki oparte na intonacji (*ang. pitch-based*) takie jak dialekty języka Chińskiego - Mandaryński czy Kantoński [**fosr**].

erfeatures}

Inne metody ekstrakcji cech.

Istnieje szereg metod bazujących na przekształceniu dziedziny czy wartości współczynników funkcji gęstości widmowej sygnału za pomocą banku filtrów o różnych właściwościach. Zamiast banku filtrów mela (MFCC) lub Barka (PLP) konstruowane są filtry bazujące na innych, bardziej specyficznych modelach ludzkiego narządu słuchu [**fosr**].

Większość opisanych wcześniej techniki ekstrakcji cech z sygnału mowy są metodami określanymi jako widmowe i krótko-czasowe (*ang. Short-Term Spectral Features*)**??** z ramkami sygnału o długości od 20 do 30 ms.

Cechy temporalno-widmowych(*ang. Spectro-Temporal Features*)[**overview**] są zbiorem wielkości dostarczających informacji na temat tożsamości mówcy zawartej w przejściach pomiędzy formantami (1.5.3) czy modulacją amplitudy współczynników widma w czasie. Jednym ze sposobów na uzyskanie tego typu informacji jest zastosowanie opisanych wcześniej współczynników Delta (Δ) i Delta-Delta (Δ^2) dla cech MFCC **??**. Innymi skutecznymi technikami jest użycie współczynników regresji liniowej czy wartościami aproksymacji wielomianami punktów dla kolejnych kilku wektorów cech (zwykle 2 lub 3) [**multidsp**]. Nowszymi rozwiązaniami jest badanie wolnych modulacji amplitudy i częstotliwości w ramkach zawierających po ok 10wektorów cech, jak np. w metodzie TDCT (*Temporal Discrete Cosine Transform*) [modulacja amplitud].

Istnieją również cechy wysokiego poziomu, takie jak cechy prozodyczne (*ang. prosodic features*) które mogą identyfikować mówcę za pomocą rytmu, akcentu czy intonacji wypowiedzi. Opisywane w literaturze są również takie zestawy cech które operą nie na charakterystykach takich związanych ze stylem wypowiedzi czy zasobem słownictwa[**overview**].

1.8 Metody klasyfikacji sygnału mowy. Modelowanie mówcy.

{featuremat

Procesy rozpoznawania mówcy, a w szczególności weryfikacji mówcy są podklasą szerszego zagadnienia znanego z techniki jako dopasowania wzorca (*ang. pattern matching*). W ogólności celem przeprowadzonego zadania dopasowania wzorca jest określenie czy w dostarczonej na wejściu systemu sekwencji danych można znaleźć pewien znany wzorzec oraz określenie ilościowej relacji stopnia podobieństwa. W rozpatrywanym w tej pracy zadaniu weryfikacji mówcy oczekuje się, że element systemu odpowiedzialny za proces dopasowania wzorca zwróci wynik reprezentujący miarę podobieństwa wejściowego wektora cech do weryfikowanego modelu mówcy. Takie modele mówcy powstają z wektorów akustycznych dostarczonych jako zestaw danych uczących. W praktycznym systemie weryfikacji mówcy tak skonstruowane modele przechowywane są w postaci zaszyfrowanych danych. Następnie używane są do podjęcia decyzji o tym czy podający się za weryfikowaną osobę otrzyma dostęp do chronionych zasobów.

Obecnie dla etapu dopasowania wzorca - w klasycznym podejściu - stosowane są metody[**campbell**] [**overview**]: ukryte modele Markowa (*ang. HMM - hidden Markov model*, dynamiczne dopasowanie czasowe (*ang. DTW - Dynamic Time Warping*), kwantyzacja wektorów (*VQ - vector quantization*, mikstury Gaussowskie (*GMM - Gaussian mixture model*), (*SVM - support vector machine*) oraz sztuczne sieci neuronowe (*ANN - artificial neural network*).

Wyróżnia się dwie grupy technik dopasowania wzorca: techniki deterministyczne (*ang. template models*) oraz techniki probabilistyczne (*probabilistic models*)[**campbell**].

W grupie technik deterministycznych, w trakcie trwania fazy modelowania (treno-

wania), z wprowadzonych danych uczących generowane są wektory wzorcowe w sposób zależny od wybranego algorytmu - mogą to być z góry narzucone wartości, uśrednione wartości wektorów czyli centroidy (*ang. centroids*), np:

{centroid}
$$\tilde{x} = \mu = \frac{1}{N}\sum_{i=1}^{N} x_i. \tag{1.28}$$

lub takie wzorcowe wektory mogą być ustalone w inny sposób. W tej klasie metod dokonuje się porównania w ten sposób uzyskanych wektorów wzorcowych z wektorami cech dostarczonych jako dane wejściowe w celu przeprowadzenia testowania, przy czym zakłada się, że [**overview**] [**campbell**] każdy z nich jest nieidealną repliką drugiego, których stopień podobieństwa określany jako funkcja miary dopasowania (*ang. match function*), która również różni się w zależności od zastosowanej metody. Do tego typu technik należy niezmodyfikowana metoda kwantyzacji wektorów - VQ, a także dynamiczne dopasowanie czasowe - DTW.

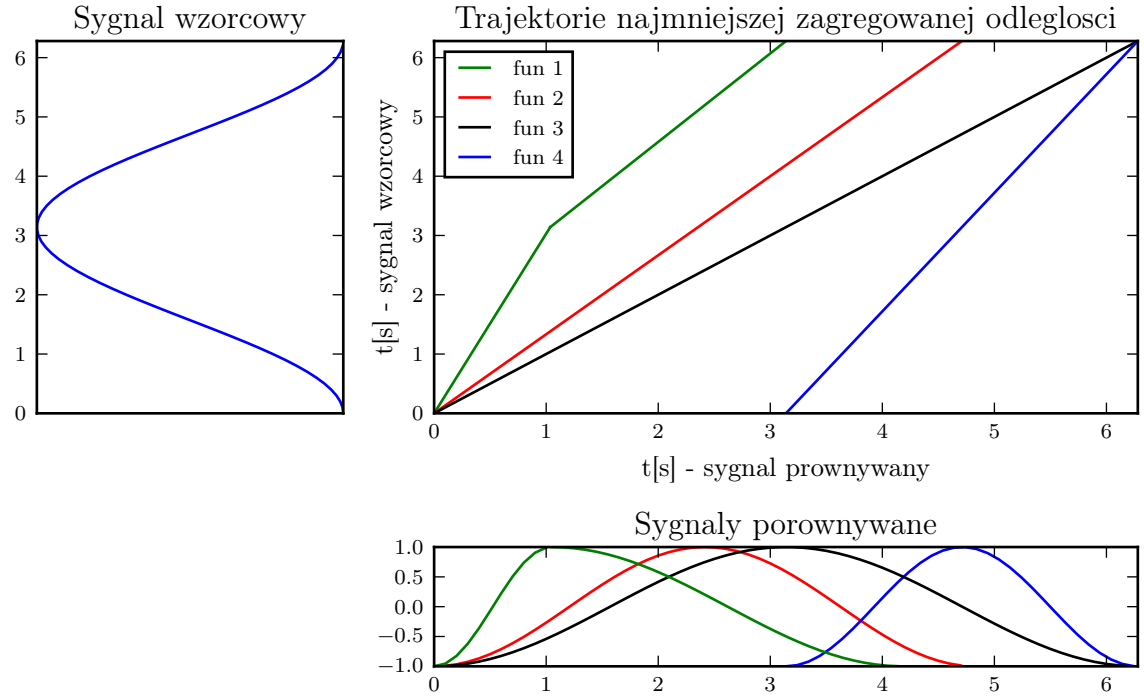
W drugiej grupie - technik probabilistycznych każdego mówce modeluje się jako funkcję gęstości prawdopodobieństwa $p(X|\lambda)$, gdzie λ oznacza model mówcy. W fazie uczenia systemu dokonuje się estymacji parametrów takiej funkcji gęstości prawdopodobieństwa na podstawie danych wejściowych - wektorów cech reprezentujących próbkę wypowiedzi mówcy. W fazie testowania najczęściej oszacowuje się prawdopodobieństwo tego czy zbiór lub sekwencja wektorów pochodzących z ekstrakcji testowanej wypowiedzi należy do zarejestrowanego modelu reprezentowanego przez ustaloną wcześniej funkcję gęstości prawdopodobieństwa. Może to odbywać się poprzez obliczenie iloczynu prawdopodobieństw wartości uzyskanych z funkcji gęstości w realizacjach zmiennej losowej reprezentowanych przez wektory wejściowe przy założeniu, że są zdarzeniami niezależnymi. Do zbioru technik probabilistycznych należą m.in. techniki ukrytych model Markova - HMM oraz mikstur Gaussowskich - GMM.

1.8.1 Dynamiczne dopasowanie czasowe - DTW.

Dynamiczne dopasowanie czasowe (DTW) jest jedną z najstarszych metod dopasowania wzorca [**multidsp**] wykorzystywana zarówno w dziedzinie rozpoznawania mowy jak i rozpoznawania mówcy. Metoda ta służy szczególnie w wypadkach kiedy potrzebna jest kompensacja różnic w długości porównywanych sygnałach. Dane wejściowe są traktowane jako sekwencja, a więc ważna jest kolejność ich dostarczania - jest to metoda zależna od czasu (*time dependent*). Z tego względu możliwe jest jej zastosowanie tylko w rozpoznawaniu mówcy zależnym od tekstu (*text-dependent*) lub z wyświetlanym hasłem (*time-prompted*). Regiony Podczas fazy testowania sekwencja wektorów akustycznych weryfikowanego mówcy $(\tilde{x_1}, .., \tilde{x_N})$ jest porównywana z z sekwencją wzorcową $(\tilde{x_1}, .., \tilde{x_M})$. Wymiary M i N najczęściej różnią się od siebie dla problemu rozpoznawania mowy oraz mówcy, co czyni tę metodę użyteczną dla tego zastosowania. W przypadku gdy $M = N$, problem redukuje się do obliczenia odległości dla pewnej zdefiniowanej metryki $d(x_i, \tilde{x})$ na przestrzeni rozpatrywanych sygnałów. W metodzie dynamicznego dopasowania czasowego wartość dopasowania sygnału wyraża się jako asynchroniczna suma:

{z}
$$z = \sum_{i=1}^{M} d(x_i, x_{\tilde{j(i)}}). \tag{1.29}$$

gdzie: $i \in M, j \in N$, oraz $j(i)$ jest funkcją indeksu i.

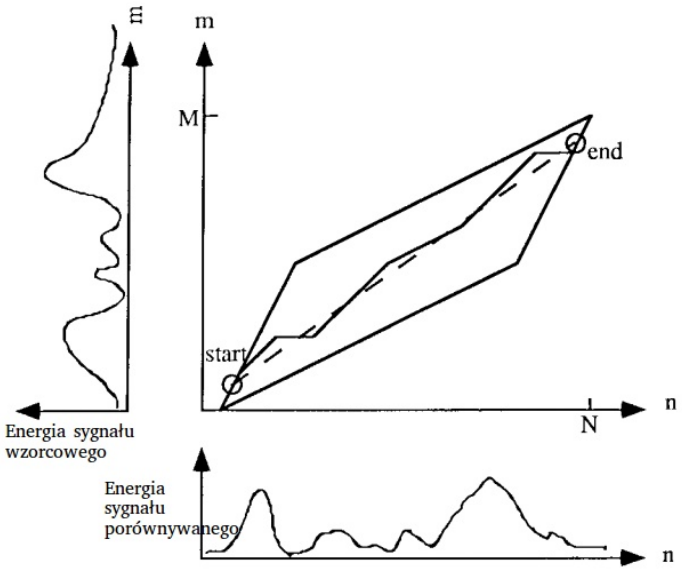


Rysunek 1.7 Przykładowe trajektorie dla przebiegów podobnych przebiegów czasowych w technice DTW. Każdy punkt sygnałów może reprezentować pojedynczy wektor cech.

{dtw}

Opisane przekształcenie dokonuje mapowania: $j(i)$ - które tworzy pary iloczynów próbek sygnału wzorcowego i sygnału porównywanego poprzez zmianę indeksowania kolejnych próbek tego drugiego. Sposób tego przekształcenia jest zależny od wybranego algorytmu realizującego technikę DTW. Najczęściej jest to krzywa szukająca najmniejszej zagregowanej odległości. Założeniem jest wybór takiego sposobu przekształcenia (trajektorii na wykresie) aby minimalizować funkcję miary dopasowania **(1.29)**. Zatem uzyskana droga powinna prezentować sekwencje iloczynów wspomnianych dwóch sygnałów, które w sumie dają najmniejszy wynik. Wyidealizowane (idealnie dobrana droga przy założeniu bardzo dużej ilości próbek) trajektorie reprezentujące takie przekształcenia zaprezentowane są na rysunku 1.7. Wykres reprezentuje iloczyn kartezjański dziedzin obu sygnałów. Na tej przestrzeni rozpięta jest funkcja metryki $d(\tilde{x}(t), x(\tau))$. Porównywane sygnały są przesuniętym i przeskalowanym sygnałem wzorcowym w dziedzinie czasu. Sygnał *fun 3.* jest sygnałem wzorcowym zatem przekształcenie j(i) jest przekształceniem identycznościowym. Sygnał *fun 2.* jest przeskalowany w czasie o wsp. 1.5 i staje się funkcją liniową. *Fun 3.* jest dodatkowo przesunięty względem sygnału wzorcowego w czasie, zaś sygnał *fun 1.* jest w obu swoich połowach przeskalowany przez inny czynnik, z tego powodu jego trajektoria jest linią łamaną na wykresie. W ogólności znalezione przekształcenie nie jest funkcją - algorytm zazwyczaj pozwala na krok w kierunku równoległym do osi czasu sygnału wzorcowego. Dla sygnału uzyskana trajektoria nie jest linią prostą co widać na rysunku **??** gdzie sygnałami wzorcowym i porównywanym są bardziej rzeczywiste przebiegi. Na tym rysunku widać również często stosowane ograniczenia dla algorytmu wykrywania drogi - otóż grubszą linią zaznaczony jest obszar dozwolony przez algorytm do prowadzenia trajektorii. Dodatkowo punkty początkowe i końcowe są ustalone "na sztywno". Klasycznym i prostym podejściem algorytmicznym jest wybór punku startowego z ograniczonego obszaru w pobliżu na osi współrzędnych w pobliżu punktu (0,0). Następnie

szukanie najmniejszej wartości spośród sąsiadujących punktów o indeksach określonych jako: $(i, j) + \Delta$. gdzie $\Delta \in \{[1, 0], [0, 1], [1, 1]\}$. Ze względu na wspomniane ograniczenia w



Rysunek 1.8 Rezultat działania metody DTW na bardziej skomplikowanej sekwencji. Adaptacja z artykułu [**campbell**]. Na głównym wykresie widać trajektorię z zaznaczonym ograniczeniem dla ścieżki algorytmu.

zastosowaniu podana metoda nie jest już chętnie stosowana w systemach rozpoznawania mówcy. Dodatkową wadą jest duża złożoność obliczeniowa, szczególnie w przypadku dużej ilości sekwencji do sprawdzenia. Jego skuteczność maleje wraz ze wzrostem długości sekwencji. Zaletą tego rozwiązania jest jednak prostota implementacji i w zastosowaniach z detekcją krótkich haseł metoda może okazać się być skuteczna.

1.8.2 Kwantyzacja wektorów - VQ.

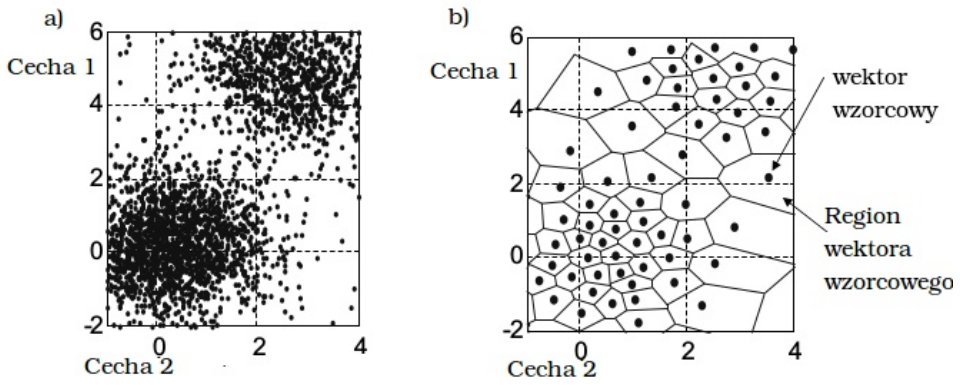
Inną deterministyczną metodą (*ang. template model*) stosowaną w etapie dopasowania wzorca w problemach rozpoznawania mówcy jest kwantyzacja wektorów (*ang. vector quantization*). W odróżnieniu do wcześniej omawianej metody DTW, ta technika jest jest niezależna od kolejności analizowanych wektorów cech - tzn. jest niezmienny w czasie (*ang. time-invariant).* Z tego powodu możliwe jest zastosowanie tej techniki dla systemów rozpoznawania mówcy niezależnego od wypowiadanego tekstu (*ang. text-independent*). Technika kwantyzacji wektorów należy do klasy problemów z zakresu technik dopasowania wzorca jaką jest uczenie bez nadzoru (*ang. unsupervised learning*). Tego typu problem nazywany jest także [**fosr**] klasteryzacją (*ang. clustering*). Jest to zagadnienie analizy skupień to znaczy dokonywane jest grupowanie przestrzeni elementów na którą składa się większa ilość - w przypadku rozpoznawania mówcy - wektorów cech na mniejszą ilość regionów. Regiony te reprezentowane są przez pewien ustalony wektor wzorcowy $\tilde{x}$. Zatem każdy punkt rozpatrywanej przestrzeni wektorów cech - x, należy do regionu reprezentowanego przez wzorcowy wektor, względem którego ów punkt leży najbliżej w rozumieniu wybranej, zdefiniowanej na tej przestrzeni metryki $d(x, \tilde{x})$. W fazie testowania, przeprowadzana jest analiza przynależności wektorów wejściowych do regionów reprezentujących weryfikowanego mówcę, to znaczy obliczana jest suma ich odległości do najbliższego wek-

tora wzorcowego:

$$z = \sum_{j=1}^{L} \min_{\tilde{x} \in C} \{d(x_j, \tilde{x})\} . \tag{1.30}$$

Funkcja ta wyraża miarę dopasowania wektorów wejściowych do weryfikowanego modelu mówcy. W tej metodzie zbiór wektorów reprezentujących poszczególnych mówców nazywany jest książką kodową (*ang. codebook*), zaś każdy wektor wchodzący w jej skład nazywa się kodem lub hasłem (*ang. codeword*).

Przykładowy proces klasteryzacji w procesie uczenia się został przedstawiony na rysunku 1.8.2. Z puli 5000 wektorów akustycznych cech został stworzony model mówcy reprezentowany przez 64 wektory wzorcowe - tzw. *codebook*.



{VQjpg}

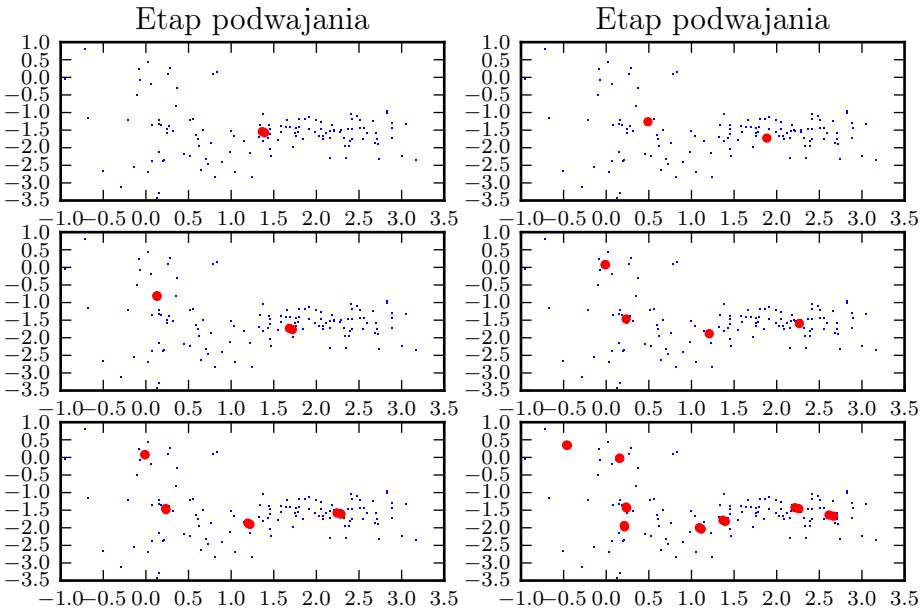
Rysunek 1.9 Klasteryzacja przestrzeni. Adaptacja z artykułu [**overview**]. Na rysunku zostały przedstawione tylko 2 arbitralnie wybrane wymiary z wektora cech. Dane uczące składające się z 5000 wektorów zostały przekształcone na model 64 wektorów wzorcowych w procesie kwantyzacji wektorów.

Metoda ta cechuje się tym, że nie przechowuje informacji związanej z sekwencją wypowiedzi, więc nie jest w stanie wychwycić pewnych cech związanych z konkretnym przejściem pomiędzy kolejnymi segmentami mowy. Z drugiej strony upraszcza w sposób znaczący implementację systemu.

W tej pracy metoda kwantyzacji wektorów została użyta w przykładowej implementacji systemu weryfikacji mówcy w postaci realizującego nią algorytmu LGB.

Algorytm LGB.

Algorytm Linde, Buzo i Greya - LGB [**linde**] przedstawia sposób uzyskania wektorów wzorcowych (książki kodów - "codebook") reprezentujących mówcę, uzyskanego z danych uczących zawierających dużą ilość wektorów cech wyekstrahowanych z wypowiedzi modelowanego mówcy. Jest to algorytm szeroko znany i chętnie wykorzystywany [**minidsp**] w zadanie rozpoznawania mówcy. Zaczynając od jednego punktu, np. uśrednionego wektora cech (jak we wzorze 1.28) produkowane są kolejne centroidy poprzez podział. Ten algorytm więc produkuje książkę kodów o rozmiarze K gdzie K jest potęgą dwójki. W kolejnych etapach algorytmu dokonuje się detekcji najbliższej centroidy dla każdego wektora cech danych trenujących. Ze względu na nowo powstałe regiony aktualizuje się położenie centroid. Po każdej aktualizacji jest prawdopodobne, że kolejne wektory cech zmienią swój region klasteryzacji, a zatem proces powtarza się aż do ustabilizowania się tychże regionów. Następnie powtarza się podwajanie wektorów wzorcowych aż do uzyskania



Rysunek 1.10 Wizualizacja procesu podwajania i aktualizacji pozycji w algortymie LBZ. `lgbvis`

pożądanej ich ilości K. Poniżej podsumowano w sposób szczegółowo kolejne etapy algorytmu, na rysunku 1.8.2 przedstawiono schemat blokowy algorytmu, a na rysunkach 1.10 przedstawiono wizualnie kolejne etapy przetwarzania algorytmu dla dwóch arbitralnie wybranych wymiarów z wektorów cech.

Kolejne kroki algorytmu Linde, Buzo i Greya.[minidsp]

1. Utwórz jedno-elementową książkę kodów - zawierającą uśrednioną centroidę jak we wzorze **1.28** dla wszystkich treningowych wektorów cech.

2. Podwój ilość elementów w książce kodów. Zmodyfikuj ich poszczególne wartości ze względu na wybrany, mały czynnik ϵ w następujący sposób, dla każdego **k** gdzie $k \in \{1, ..., K\}$:

   - $\tilde{\boldsymbol{x}}_{\boldsymbol{k}} = (1 - \epsilon) \cdot \tilde{\boldsymbol{x}}_{\boldsymbol{k}}$,
   - $\tilde{\boldsymbol{x}}_{\boldsymbol{k+1}} = (1 + \epsilon) \cdot \tilde{\boldsymbol{x}}_{\boldsymbol{k}}$.

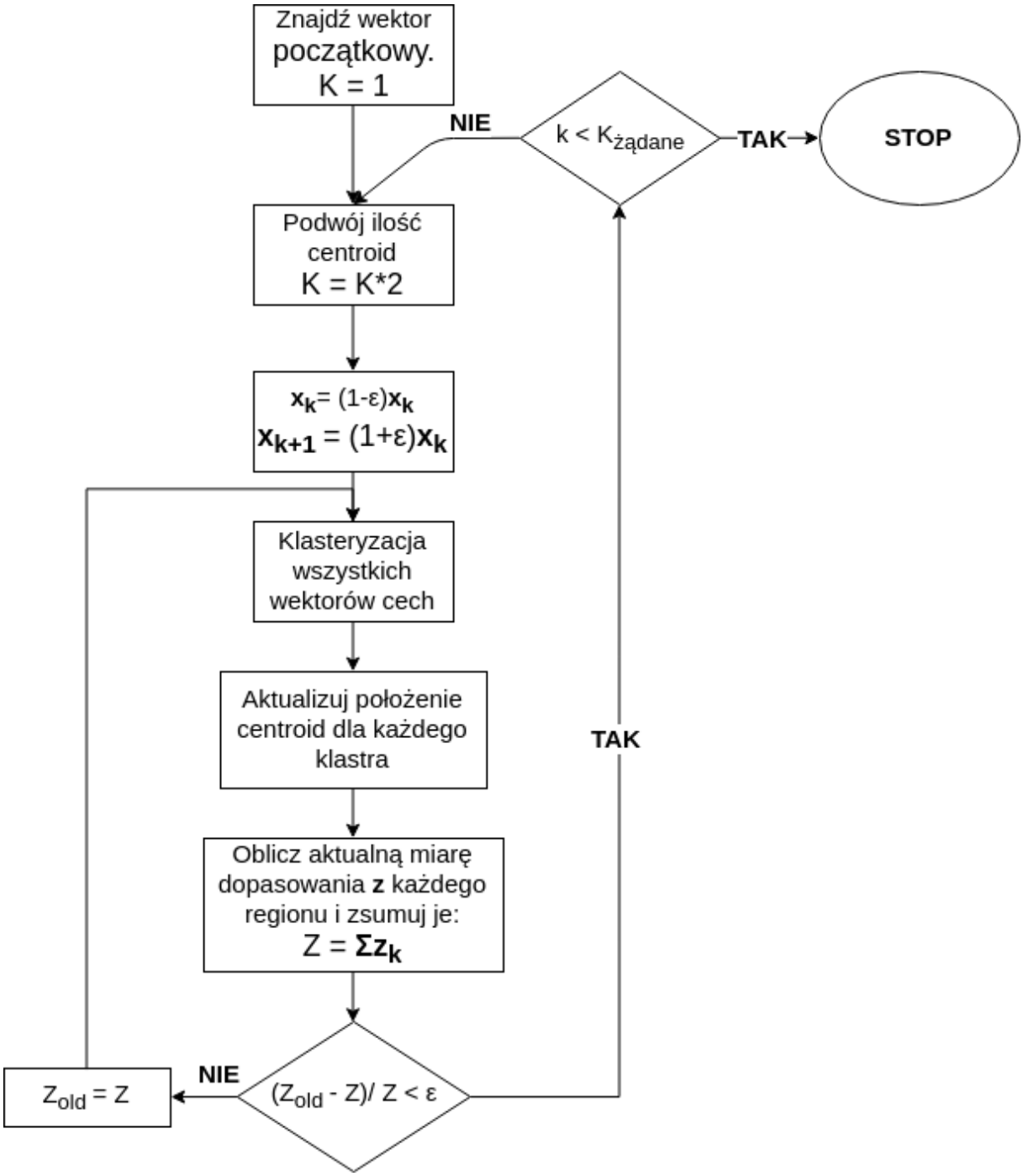
   Co oznacza przesunięcie je w przeciwnych kierunkach (w N-wymiarowej przestrzeni).

3. Dokonaj klasyfikacji wszystkich wektorów trenujących - znajdź dla każdego najbliższą centroidę w rozumieniu przyjętej metryki $d(\tilde{\boldsymbol{x}}_{\boldsymbol{k}}, \boldsymbol{x})$.

4. Zaktualizuj położenie każdej centroidy jako średniej **1.28** ze wszystkich wektorów cech sklasyfikowane jako leżące najbliżej.

5. Powtarzaj kroki 3 i 4 dopóki położenie nowych centroid nie ustabilizuje się (różnica przesunięcia dwóch ostatnich iteracji jest poniżej ustalonego progu).

6. Powtarzaj kroki 2, 3, 4 i 5 do momentu otrzymania słownika o żądanej długości **K**.



{lgbblock}

Rysunek 1.11 Schemat blokowy algorytmu Linde, Buzo i Greya. **[minidsp]**

1.8.3 Mikstury Gaussowskie - GMM.

Techinka modelowanie miksturami Gaussowskimi *ang. Gaussian mixture modeling* jest metodą stochastyczną rozpoznawania wzorca. GMM jest sposobem modelowania uważanym za swego rodzaju standard w metodach rozpoznawania mówcy niezależnym od wypowiadanego tekstu. Tak jak wspomniano w rozważaniach ogólnych (1.8), w metodach probabilistycznych należy ustalić funkcję gęstości prawdopodobieństwa $p(X|\lambda)$ charakteryzującą danego mówcę. Dla problemu rozpoznawania mówcy niezależnego od tekstu, nie ma żadnej wiedzy apriori, którą można by było wykorzystać do założenia jakiegoś

rozkładu. Dlatego też metoda GMM proponuje dokonania aproksymacji tzw. miksturami Gaussa (*ang. Gaussian mixtures*). Okazuje się, że taka reprezentacja funkcji gęstości prawdopodobieństwa jest równie skuteczna co dużo bardziej skomplikowana reprezentacja ukrytymi modelami Markova (HMM) dla problemu rozpoznawania bez znajomości tekstu [**reynolds**], ze względu na brak przydatności wiedzy o sekwencji wektorów cech. Dla K-wymiarowych wektorów cech funkcja gęstości prawdopodobieństwa dla modeli GMM wyraża się jako:

$$p(X|\lambda) = \sum_{i=1}^{M} w_i p_i(X). \tag{1.31}$$

gdzie: w_i są współczynnikami wagowymi ustalanymi w fazie uczenia oraz $\sum_{i=1}^{M} w_i = 1$. M oznacza zaś liczbę wykorzystanych misktur Gaussowskich do aproksymacji funkcji gęstości prawdopodobieństwa dla modelu λ. Kolejna mikstura Gaussowska przybiera postać wielowymiarowej funkcji rozkładu Gaussa:

$$p_i(X) = \frac{1}{(2\pi)^{K/2} \cdot |\boldsymbol{\Sigma_i}|^{\frac{1}{2}}} \cdot \exp\left\{-\frac{1}{2}(X - \boldsymbol{\mu_i})'(\boldsymbol{\Sigma_i})^{-1}(X - \boldsymbol{\mu_i})\right\} \tag{1.32}$$

gdzie: $\boldsymbol{\Sigma_i}$ oznacza macierz kowariancji wektora cech X, μ_i jest wektorem średnim (K-wymiarową wartością oczekiwaną). Pełny opis modelu λ zapewnia zbiór parametrów $\lambda = \{w_i, \boldsymbol{\mu_i}, \boldsymbol{\Sigma_i}\}$. Przy pomocy wektorów cech danych trenujących dokonywana jest estymacja parametrów dla poszczególnych mikstur Gaussowskich. W tym celu najczęściej używa się się metod maksymalnego prawdopodobieństwa a posteriori MAP (*maximum a posteriori probability*) oraz algorytmu EM (*ang. expectation-maximization*).

1.9 Klasyfikacja i teoria decyzji

Spis treści