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 rozpoznawania mówcy osiągały skuteczność sięgającą zaledwie 90 - 95%. Jednak szybki rozwój w dziedzinie dopasowywania wzorca (*and. pattern matching*) spowodowany m. in. dynamicznym

rozwojem w dziedzinie metod sztucznej inteligencji pozwala dzisiaj osiągać skuteczności - dla problemu identyfikacji mówcy - sięgające nawet 99%.

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 najnowsych 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.

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.

———————— biometryk nie da się zgubić choć część populacji jest niema i to również trzeba mieć na uwadze.

1.2 Weryfikacja mówcy

Weryfikacja mówcy (*speaker verification*).

Proces weryfikacji mówcy jest związany z szerszym zagadnieniem - rozpoznawania mówcy (*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 (*background model*) lub pewnej jej podgrupy (*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.(*open-set speaker identification*). Proces ten można uważać jako złożenie problemu weryfikacji mówcy oraz identyfikacji mówcy na zbiorze zamkniętym (*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).

1.3 relacja pomiędzy rozpoznawanie mowy, a rozpoznawaniem mówcy

Kluczowe jest odróżnienie procesu rozpoznawania mówcy od systemów rozpoznawania mowy (*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 (*knowledge-based systems*), których zadaniem jest jedynie wzmocnienie procesu weryfikacji oraz zapobieganie wystąpienia problemu żywotności (*liveness issue*).

1.4 Klasyfikacja weryfikacji mówcy ...?

text-dependent speaker recognition [**fosr**]

Problem żywotności polega na możliwości oszukania działającego systemu weryfikacji mówcy poprzez dostarczenie na wejście takiego systemu 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.

text-independent speaker recognition [**fosr**]

text-prompted speaker recognition [**fosr**]

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

%TODO zastosowania

%TODO system weryfikacji mówcy może być skojarzony z innymi systemami rozpoznawania biometryk

%TODO opis zastosowania weryfikacji mówcy w systemie wbudowanym

1.5 SYGNAŁ MOWY

Człowiek dysponuje doskonałymi narzędziami do przeprowadzenia procesów rozpoznawania mowy oraz rozpoznawania mówcy. Analiza oraz zrozumienie mechanizmów powstawania mowy u człowieka dostarcza podstaw do sformułowania metod syntezy języka naturalnego. 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 rozpoznawania 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.

1.5.1 Produkcja sygnału mowy.

Aparat mowy człowieka.

Tembr głosu mówcy ustalony jest przez budowę jego dróg głosowych.

1.5.2 Percepcja sygnału mowy przez człowieka

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. *fenestra vestibuli*), którego zadaniem jest wzbudzenie drgań

Układ nerwowy

1.5.3 Lingwistyka a rozpoznawania mówcy.

Występuje silny związek pomiędzy językiem a procesem rozpoznawania mówcy. Obszarami lingwistyki, które szczególnie dotyczą badanej kwestii są: fonetyka, fonologia oraz prozodia.

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 rozróżnianie 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. Dla przykładu Zbiór takich głosek 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.

Fonologia

Fonologia bada

Prozodia

Prozodia zajmuje się brzmieniowymi właściwościami mowy na które składają się trzy elementy - intonacja, akcent oraz iloczas.

1.6 Ekstrakcja cech sygnału mowy.

1.6.1 Przegląd.

Celem ekstrakcji cech z sygnału mowy (*feature extraction*) jest uzyskanie zbioru cech charakteryzujących sygnał ludzkiej mowy za pomocą technik cyfrowego przetwarzania sygnału. Jednocześnie jest to zamiana sygnału z 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 charakterystyk 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 Gc.

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 (*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.

1.6.2 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 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 jx dla każdej ramki j.

Podział sygnału na ramki.

Ze względu na własność quasi-stacjonarności sygnału mowy, aby wydobyć informację dotyczącą wypowiadanej głoski konieczne jest rozdzielenie sygnału na ramki (*framing the signal*). 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 (*Fast Fourier Transform, FFT*) wymagającej aby sygnał był długości potęgi dwójki. W przypadku pierwszej oraz ostatniej ramki stosowana jest techinka uzupełniania zerami w przypadku braku próbek do

zapełnienia całej ramki.

Aplikacja okna na ramki.

W stosowanej, zredukowanej techince STFT stosowane są okna inne niż okno prostokątne o długości ramki lh w celu polepszenia właściwości widmowych dla dalszej analizy. Zastosowanie okna czasowego odbywa się poprzez przemnożenie wartości próbek ramki przez wartości okna w następujący sposób:

Najpopularniejszymi oknami stosowanymi w tym przypadku są okna czasowe Hanna, Hamming oraz okna Gaussowskie.

Estymacja widma.

Po zaaplikowaniu okien czasowych dla kolejnych ramek, kolejnym etapem do otrzymania współczynników MFCC jest znalezienie widma ramki za pomocą dyskretnej transformaty fouriera. Kolejne współczynniki uzyskanego widma zdefiniowane są następująco: W omawianej metodzie korzysta się jedynie z informacji o amplitudzie uzyskanego widma. Wartość wielkości amlitud otrzymywane są w następujący sposób:

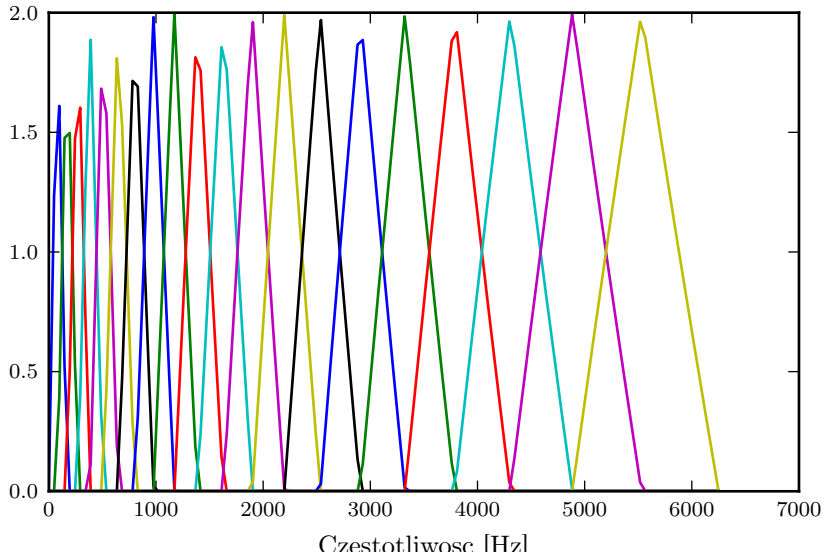
Przejście do dziedziny częstotliwości Mela.

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: 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 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. Innym proponowanym podejściem jest znalezienie filtrów poprzez proces uczenia maszynowego, uczoną ze względu na

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



{melfb}

Rysunek 1.1 Przykładowe 24 trójkątne filtry dla sygnału o częstotliwości próbkowania *fs=12500 Hz.*

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:

$$\tilde{C_k} = 10 \cdot \left(\frac{|\tilde{H_k}|^2}{I_0} \right) \tag{1.1}$$

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 1.6.2):

$$C_k = log\left(|\tilde{H_k}|^2\right) \tag{1.2}$$

Przy użyciu tych współczynników otrzymuje się cepstrum sygnału.

Cepstrum.

{cepstrum}

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

$$\tilde{h_{pc}} = \left[\frac{1}{2\pi} \int_{-\pi}^{\pi} log\left(|H(\omega)|^2\right) e^{i\omega t} \right]^2 \tag{1.3}$$

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

$$x(t) = s(t) + \alpha s(t - \tau) \tag{1.4}$$

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

postać

$$|S(f)|^2 \left[1 + \alpha^2 + 2\alpha cos(2\pi f\tau)\right]. \tag{1.5}$$

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":

$$\begin{aligned}\mathcal{F}^{-1}\{log(X(f))\} = \mathcal{F}^{-1}\{log(S(f))\} + \mathcal{F}^{-1}\{log(1+\alpha^2)\} + \\ + \mathcal{F}^{-1}\{log(1 + \frac{2a}{1+a^2} cos(\omega\tau))\}.\end{aligned} \tag{1.6}$$

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:

$$h(t) = h_1(t) * h_2(t) \tag{1.7}$$

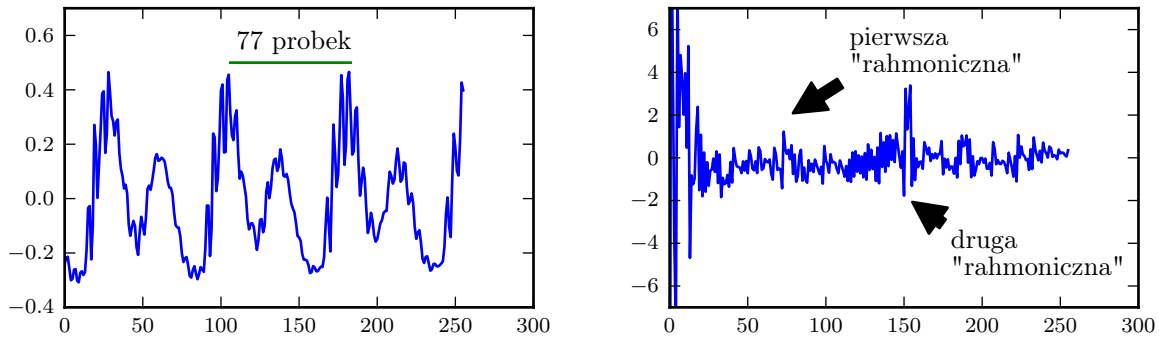
Z własności ciągłego przekształcenia Fouriera wynika, że powyższe równanie w dziedzinie częstotliwości przybiera postać iloczynu transformat $H_1(\omega)\cdot H_2(\omega)$. W dziedzinie cepstrum rozdziela się natomiast na sumę składników:

$$\tilde{h}(t) = log(H_1(\omega)) + log(H_2(\omega)) \tag{1.8}$$

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 1.2. 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:

$$H_k = \sum_{n=0}^{N-1} h_n cos\left(\frac{\pi(2n+1)k}{2N}\right) \tag{1.9}$$



e_cepstrum}

Rysunek 1.2 Ramka sygnału wypowiadanej samogłoski 'i' oraz jej cepstrum.

gdzie

$$a_k = \begin{cases} 1/N & \text{dla} \quad k = 0 \\ 2/N & \forall \quad k > 0. \end{cases} \tag{1.10}$$

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 1.7.

1.7 Dopasowanie wzorca.

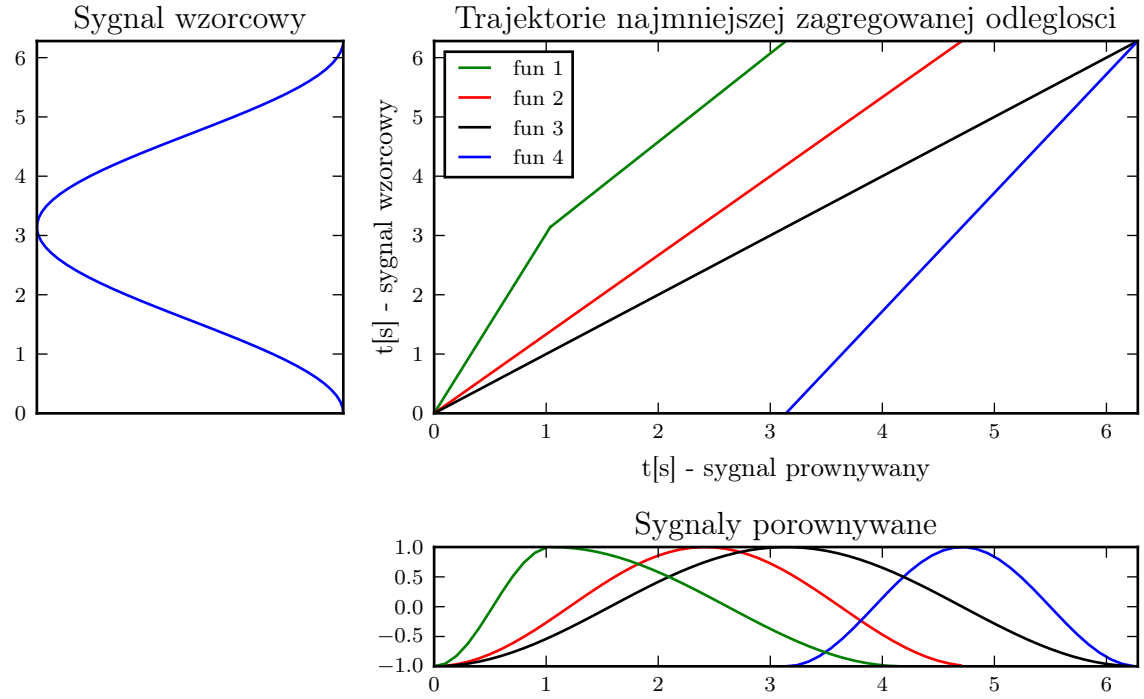
{featuremat

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 fazie modelowania z wprowadzonych danych uczących generuje się zestaw wzorcowych regionów na podstawie wybranego algorytmu. Przestrzeń wektorów wyekstrahowanych cech jest podzielona przez te regiony, które reprezentowane są przez wzorcowe wektory - $\tilde{x}$. Zatem każdy punkt rozpatrywanej przestrzeni cech - x, należy do regionu reprezentowanego przez wzorcowy wektor do którego ów punkt leży najbliżej w rozumieniu wybranej, zdefiniowanej na tej przestrzeni metryki $d(x, \tilde{x})$. W



{dtw} Rysunek 1.3 Wykresy funkcji przekształcenia dziedziny sygnałów porównywanych.

fazie testowania, przeprowadzana jest analiza przynależności wektorów wejściowych do regionów reprezentujących weryfikowanego mówcę, to znaczy obliczana jest suma ich odległości:

$$\tilde{x} = \mu = \frac{1}{N} \sum_{i=1}^{N} x_i. \tag{1.11}$$

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 pewną przyjętą funkcję gęstości prawdopodobieństwa. 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 sumy prawdopodobieństw wartości uzyskanych z funkcji gęstości prawdopodobieństwa w realizacjach zmiennej losowej reprezentowanych przez wektory wejściowe. Do zbioru technik probabilistycznych należą m.in. techniki ukrytych model Markova - HMM oraz mikstur Gaussowskich - GMM.

1.7.1 Dynamiczne dopasowanie czasowe - DTW.

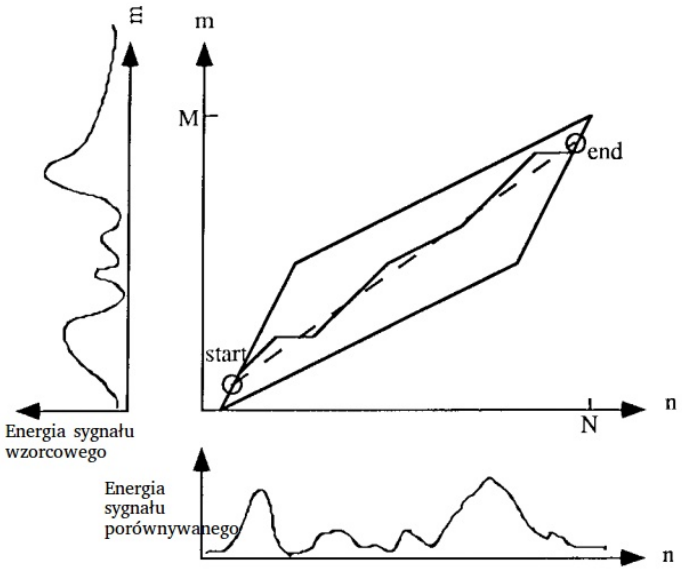
Dynamiczne dopasowanie czasowe (DTW) jest jedną z najstarszych metod dopasowania wzorca [**multidsp**] wykorzystywana zarówno w dziedzinie rozpoznawania mowy jak i rozpoznawania mówcy. Metoda ta służy szczególnie w wypadkach kiedy potrzebna jest kompensacja różnic w długości porównywanych sygnałach. 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*). 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:

$$z = \sum_{i=1}^{M} d(x_i, x_{\tilde{j(i)}}). \tag{1.12}$$ {z}

gdzie: $i \in M, j \in N$, oraz $j(i)$ jest funkcją indeksu i. 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 1.12. 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 1.3. 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 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 Klasyfikacja i teoria decyzji



Rysunek 1.4 Rezultat działania metody DTW na bardziej skomplikowanej sekwencji. Adaptacja z artykułu [**campbell**]

Spis treści