

**Państwowa Wyższa Szkoła Zawodowa
w Tarnowie**

Informatyka Stosowana

INSTYTUT POLITECHNICZNY



PRACA INŻYNIERSKA

PIOTR KOZŁOWSKI

**STEROWANIE GŁOSOWE ROBOTEM KAWASAKI W
WYBRANYCH GRACH PLANSZOWYCH**

PROMOTOR:

dr Robert Wielgat

Tarnów 2012

OŚWIADCZENIE AUTORA PRACY

OŚWIADCZAM, ŚWIADOMY ODPOWIEDZIALNOŚCI KARNEJ ZA POŚWIADCZENIE NIEPRAWDY, ŻE NINIEJSZĄ PRACĘ DYPLOMOWĄ WYKONAŁEM OSOBIŚCIE I SAMODZIELNIE, I NIE KORZYSTAŁEM ZE ŹRÓDEŁ INNYCH NIŻ WYMIENIONE W PRACY.

.....

PODPIS

Serdecznie dziękuję mojemu promotorowi za cenne wskazówki, które pomogły mi ukończyć tą pracę.

Spis treści

1. Wstęp	6
2. Rozpoznawanie sygnału mowy	7
2.1. Przegląd gotowych rozwiązań	7
2.2. Mowa w języku polskim	8
2.3. System rozpoznawania mowy	9
2.4. Reprezentacja sygnału mowy	10
2.5. Detekcja sygnału mowy	13
2.6. Ekstrakcja cech sygnału mowy	15
3. Ukryte Modele Markowa jako jedna z metod klasyfikacji	18
3.1. Wprowadzenie teoretyczne	18
3.2. Algorytm Bauma-Welcha	19
3.3. Algorytm Viterbiego	19
4. HTK (Hidden Markov Model Toolkit)	20
4.1. Rozpoczęcie pracy z bibliotekami	20
4.2. Trenowanie modeli	20
4.3. Rozpoznawanie	20
5. System sterujący robotem za pomocą komend głosowych	21
5.1. Schemat działania systemu	21
5.2. Użyte technologie	21
5.3. Robot przemysłowy	21
5.4. Urządzenie na którym działa cały system	21
5.5. Gry planszowe	21
5.6. Prezentacja działania systemu	21
6. Podsumowanie	22

1. Wstęp

Wiek w którym żyjemy, zwany jest często wiekiem komunikacji w różnych jej formach. Jest ona bardzo ważnym aspektem życia człowieka. Tematyka tej pracy związana jest z komunikacją werbalną na linii człowiek - maszyna.

Jeszcze do niedawna sterowanie sprzętem elektronicznym za pomocą komend głosowych, było spotykane częściej w filmach s-f niż w rzeczywistych zastosowaniach. Obecnie jest to zadanie możliwe do wykonania, niemniej jednak, ze względu na skomplikowany charakter sygnału mowy, często trudne do realizacji.

Systemy rozpoznawania mowy są rozwijane od wielu lat, zarówno jako projekty komercyjne jak i naukowe. Obecnie większość telefonów komórkowych, tzw. „smartfonów” posiada aplikacje ułatwiające korzystanie z nich, pozwalające na wybieranie głosowe połączeń lub zamieniające mowę na tekst. Liczba zastosowań systemów rozpoznawania mowy ciągle rośnie, ze względu na oczywisty fakt, że mowa jest naturalnym oraz jakże wygodnym sposobem przekazywania myśli i odczuć człowieka. Uzyskiwana jest również coraz większa skuteczność takich systemów, jednakże nie tak duża, by pozwalała na bezpieczne i efektywne wykorzystywanie ich w każdym przypadku. Co za tym idzie, poprawianie skuteczności wiąże się z dalszym rozwojem tej dziedziny a więc może być ona bardzo ciekawym obiektem zainteresowań.

Przeglądając artykuły i prace dotyczące tematyki tutaj poruszanej często spotkać się można ze stwierdzeniem że problem rozpoznawania mowy dla języka polskiego, nie został dotychczas tak dobrze opracowany jak to zostało zrobione w przypadku innych języków, tzw. języków dominujących, takich jak angielski, niemiecki, francuski, hiszpański czy chiński. [12] Ta granica powoli zaczyna się zacierać, coraz więcej rozwiązań tworzonych jest również z myślą o języku polskim.

Niniejsza praca koncentruje się na wykorzystaniu osiągnięć projektu HTK [2] i stworzeniu systemu rozpoznawania mowy dla języka polskiego na przykładzie sterowania robotem przemysłowym w grze planszowej. Rozgrywka polega na wydawaniu komend głosowych które sterują ruchami robota. Jest on zarówno przeciwnikiem użytkownika jak również wykonuje za niego wszystkie ruchy podczas rozgrywki, użytkownik skupia się zatem tylko na wydawaniu poleceń głosowych. Gra z robotem jest formą demonstracji systemu stworzonego na potrzeby tej pracy a więc jest jej poświęcone mniej uwagi niż tematyce rozpoznawania mowy.

2. Rozpoznawanie sygnału mowy

Mowa jest najbardziej naturalnym oraz skutecznym sposobem porozumiewania się ludzi. Jest znacznie szybsza niż pisanie czy gestykulacja, a co najważniejsze nie wymaga użycia rąk. Nie dziwi więc fakt, że wraz z rozwojem sprzętu komputerowego, wzrostem jego mocy obliczeniowej oraz miniaturyzacji, systemy rozpoznawania mowy mają coraz częstsze zastosowanie.

Oto lista przykładowych zastosowań takich systemów:

- sterowanie głosem urządzeń elektronicznych i elektromechanicznych,
- teleinformatyczne systemy informacyjne,
- systemy STT (ang. Speech To Text),
- urządzenia dla osób niepełnosprawnych,
- programy do nauki języków obcych,
- systemy diagnostyki medycznej i logopedycznej,
- systemy identyfikacji mówców. [18]

Warto zaznaczyć, że systemy rozpoznawania mowy, nie znajdują zastosowania tam gdzie sterowanie jest bardziej skomplikowane oraz wymaga od użytkownika szybkiej reakcji. Są sytuacje w których szybciej dotrzemy do odbiorcy (pozwolimy mu zrozumieć naszą ideę) używając rąk, np. za pomocą gestów lub rysunków niż przy pomocy komunikacji głosowej.

2.1. Przegląd gotowych rozwiązań

Wśród rozwiązań komercyjnych obecnie przodujących na rynku warto wymienić te przeznaczone dla urządzeń mobilnych, czyli m.in. telefonów komórkowych, które w obecnych czasach często posiadają możliwości zbliżone do komputerów osobistych. Firmy których produkty zdominowały rynek systemów przeznaczonych na urządzenia mobilne mają opracowane własne technologie rozpoznawania mowy połączone ze sztuczną inteligencją oraz syntezą mowy. I tak kolejno, Apple posiada Siri [8], Google - Google Now [10] a Microsoft - TellMe [15]. Produkty te, mają za zadanie ułatwić korzystanie z telefonu. Pełnią rolę wirtualnego asystenta, z którym można rozmawiać, zapytać o pogodę, najbliższą restaurację, poprosić o wpis do kalendarza lub przeczytanie sms'a. Produkt firmy Apple był najwcześniej

wprowadzony na rynek więc wydaje się być technologią najbardziej zaawansowaną. Natomiast Google prócz najnowszej technologii Google Now, posiada również aplikację Google Mobile App [9], która służy głównie do wyszukiwania głosowego (np. kontaktów, wiadomości lub innych rzeczy korzystając z wyszukiwarki internetowej Google). Poza produktami wymienionych wcześniej firm, powstaje coraz więcej aplikacji typu third-party (czyli tworzone przez osoby lub firmy niezwiązane z producentami systemów operacyjnych) posiadających podobne funkcjonalności (np. S-Voice firmy Samsung lub Iris, obie przeznaczone na platformę Android firmy Google).

Do najstarszych projektów komercyjnych, rozwijających technologię rozpoznawania mowy, można zaliczyć IBM ViaVoice, Microsoft Speech API, Oracle Java Speech API lub Dragon NaturallySpeaking firmy Nuance, która ostatecznie przejęła od IBM technologię ViaVoice [16]. W Polsce powstaje coraz więcej firm projektujących i wdrażających rozwiązania biznesowe oparte na rozpoznawaniu mowy, zwane systemami IVR (ang. Interactive Voice Response), umożliwiające interaktywną obsługę osoby dzwoniącej. Są to m.in. firma Primespeech (produkująca systemy przeznaczone na rynek telekomunikacyjny, np. dla Zarządu Transportu Miejskiego w Warszawie, działające na zasadzie wirtualnych konsultantów, informujących o rozkładach jazdy, cenach biletów, aktualnościach lub przyjmujący różnego rodzaju skargi klientów) [5], MagicScribe (systemy zamieniające mowę na tekst, MagicScribeMedical - system stworzony dla medycyny, wspomagający obsługę pacjenta, tworzenie dokumentacji medycznej itp. oraz MagicScribeLegal, rozwiązanie dla adwokatów, notariuszy, radców prawnych) [4], Stanusch Technologies S.A. lub VOICE LAB.

Z pośród projektów naukowych związanych z rozpoznawaniem mowy warto wymienić HTK [2], Julius [3], CMUSphinx [1] lub rodzimy projekt SARMATA (używany m.in. przez instytucje wymiaru sprawiedliwości do zarządzania dokumentacją procesową) [6].

W obecnych czasach tworzenie systemów rozpoznawania mowy jest dużo prostsze, ze względu na obecność gotowych narzędzi wspomagających, m.in. HTK czy CMUSphinx, które posiadają zaimplementowane algorytmy, pozwalające stworzyć system rozpoznawania mowy nawet osobie nie znającej szczegółów aparatu matematycznego. Użytkownik korzystający z takich bibliotek często używa ich jak „czarnej skrzynki” niemniej jednak potrafi stworzyć w pełni działający system. Na tyle skuteczny, by nie musieć zagłębiać szczegółów zaimplementowanych algorytmów. Co prawda taka niewiedza stwarza problemy, przy doborze optymalnych parametrów mających istotny wpływ na działanie systemu lecz nie przeszkadza w doborze ich w sposób eksperymentalny i osiągnięciu zadowalającej skuteczności.

2.2. Mowa w języku polskim

Alfabet języka polskiego składa się z 32 liter: a, ą, b, c, ć, d, e, ę, f, g, h, i, j, k, l, ł, m, n, ó, o, ó, p, r, s, ś, t, u, w, y, z, ź, ż. Łacińskie litery q, v i x występują jedynie w pisowni wyrazów obcych. Alfabet fonetyczny języka polskiego składa się z około 78 dźwięków. Dokładna liczba głosek zależy od sposobu traktowania wariantów. Relacja litera – głoska jest typu wiele – wiele. Tak jest również w wielu innych językach. Strukturę języka możemy przedstawić w postaci modelu warstwowego (rys. 1). Warstwa wyższa zawiera pewną liczbę elementów warstwy niższej.

myśl
wypowiedź
zdania
słowa
syłaby
głoski/fonemy

Rysunek 1: Warstwowy model języka

Badaniem struktury dźwiękowej języka zajmują się dwa pokrewne działy, fonetyka oraz fonologia. Fonetyka jest nauką o głoskach, czyli dźwiękach mowy. „Bada i opisuje dźwięki mowy ze względu na ich właściwości fizyczne, tzn. ustala artykulacyjne i akustyczne ich cechy.” [7] Fonologia natomiast bada dźwięki mowy, pod kątem pełnionych przez nich funkcji w procesie komunikacji. Podstawową jednostką fonologii jest fonem. Fonem uznaje się za najmniejszy rozróżnialny segment dźwiękowy mowy, który może odróżniać znaczenie dźwięków mowy danego języka o różnicach wynikających wyłącznie z charakteru indywidualnej wymowy lub kontekstu. Dla systemów rozpoznawania mowy istotne jest to, że każdy fonem, jest zespołem cech dystynktywnych pozwalających na odróżnienie go od pozostałych fonemów, co przekłada się bezpośrednio na parametry akustyczne sygnału. [7]

Miarę podobieństwa między sygnałem mowy możemy podzielić na dwie kategorie, podobieństwo akustyczne oraz fonetyczne.

- Podobieństwo fonetyczne dwóch słów jest ustalane na podstawie tzw. odległości Levenshteina określanej jako minimalny koszt przekształcenia jednego słowa w drugie.
- Podobieństwo akustyczne dwóch słów jest ustalane na podstawie odległości (prawdopodobieństwa identyczności) między dwoma słowami wynikającej z przyjętych cech sygnału mowy w ramce oraz przyjętej metody klasyfikacji.

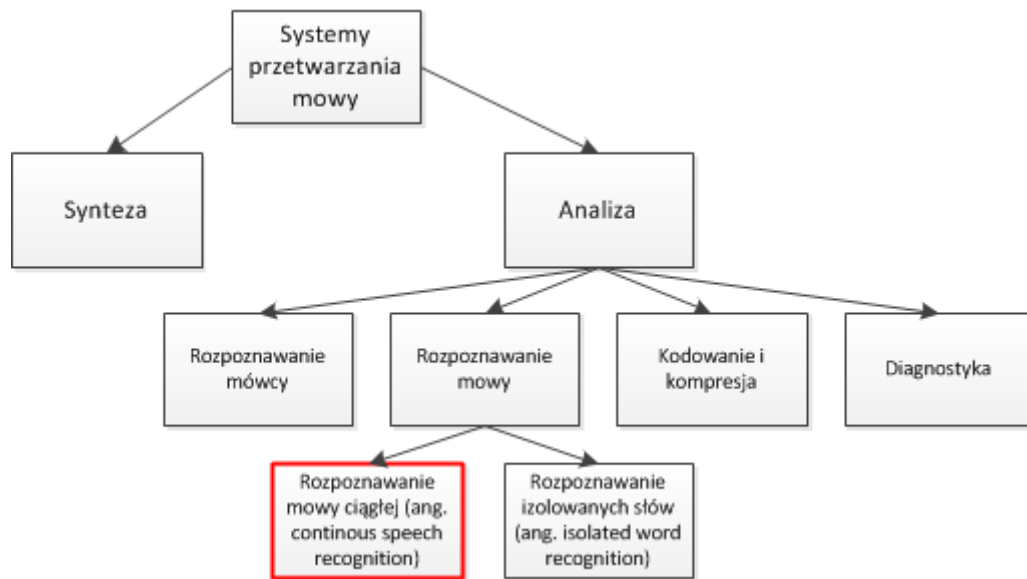
Zazwyczaj im słowa są bardziej podobne fonetycznie, tym bardziej są podobne akustycznie. Zdarzają się jednak sytuacje, w których zależność taka nie zachodzi. **[Przypis z Pana wykładów? Czy wyrzucić stąd akapit o podobieństwach?]** [18]

2.3. System rozpoznawania mowy

Na początek warto określić miejsce systemów rozpoznawania mowy na tle innych zagadnień z tej dziedziny (rys. 2). Na czerwono zaznaczono dziedzinę w której znajduje się system stworzony na potrzeby tej pracy.

Parametry wpływające na skuteczność rozpoznawania:

- stopień zależności od mówcy (systemy zależne (ang. speaker-dependent) i niezależne od mówcy (ang. speaker-independent)),
- liczba rozpoznawanych słów (mała < 20 słów, duża > 20 000 słów),
- podobieństwo akustyczne i fonetyczne,



Rysunek 2: Ogólny podział systemów przetwarzania mowy

- SNR sygnału mowy (mały < 10 dB, duży > 30 dB),
- parametry transmisji sygnału mowy (zniekształcenia analogowego sygnału mowy wnoszone przez kanał transmisyjny, liczba poziomów kwantyzacji 8-16 bitów/próbkę, częstotliwość próbkowania 8-44 kHz). [18]

Skuteczność rozpoznawania zależy również od środowiska w jakim działa dany system. Może być ona inna dla zamkniętego pomieszczenia oraz otwartej przestrzeni. Negatywny wpływ mają tutaj wszelkiego rodzaju szумы (np. ruch uliczny, odgłosy rozmów innych ludzi). Nie bez wpływu pozostają również cechy samego mówcy. Istotny jest sposób wymowy (gwara, akcent), szybkość oraz głośność. Na rysunku 3 jest widoczny ogólny schemat działania systemu rozpoznawania mowy.

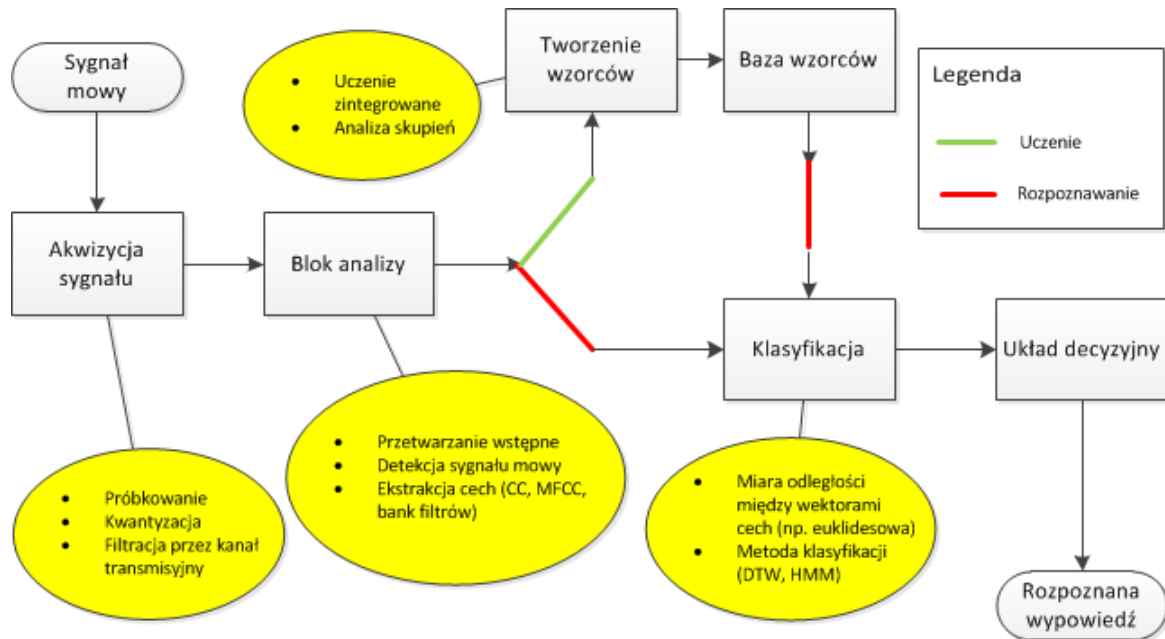
Opisać poszczególne etapy?

Systemy automatycznego rozpoznawania mowy, zwane ASR (ang. Automatic Speech Recognition) cechują się tym, że działają w czasie rzeczywistym (ang. real time). System ten ma za zadanie przeprowadzić detekcję sygnału mowy (wykryć kiedy została wypowiedziana jakaś sentencja) następnie jako wynik rozpoznawania utworzyć jej transkrypcję fonetyczną (zamienić mowę na tekst). Wszystko to powinno zostać zrobione automatycznie, tzn. bez ingerencji użytkownika. Taka sama idea przyświeca systemowi stworzonemu na potrzeby tej pracy.

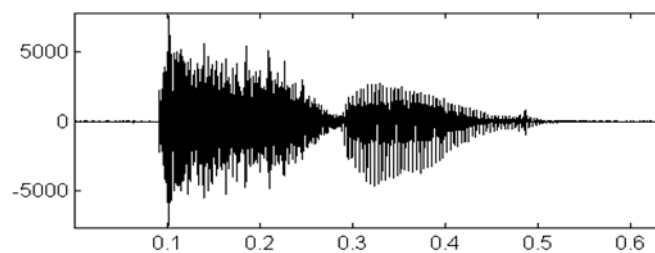
2.4. Reprezentacja sygnału mowy

Sygnał mowy może być opisywany na różne sposoby. Podstawowym i najprostszym jest opis w dziedzinie czasu (rys. 4). Ma on jednak skomplikowany przebieg, będący odzwierciedleniem złożonego charakteru artykulacji i często zawierający różne przypadkowe szумы pochodzące z otoczenia.

Matematycznie możemy go przedstawić jako spłot przebiegu czasowego sygnału źródła $u_z(t)$ i odpowiedzi impulsowej kanału głosowego $h(t)$, czyli:



Rysunek 3: Ogólny schemat blokowy rozpoznawania mowy



Rysunek 4: Reprezentacja słowa „trzy” w dziedzinie czasu

$$u(t) = \int_0^{\tau} h(t - \tau) u_z(\tau) d\tau$$

Źródłem sygnału mogą być drgania wiązań głosowych (głoski dźwięczne) jak i szum powstający w skutek przepływu powietrza przez narządy mowy (głoski bezdźwięczne). [14]

Zanim jednak otrzymamy powyższy przebieg sygnału mowy, musi być wykonana jego dyskretyzacja, czyli próbkowanie i kwantyzacja. Z punktu widzenia rozpoznawania mowy, reprezentacja czasowa sygnału mowy nie przenosi wystarczającej ilości informacji.[13] Potrzeba zatem dokonać transformaty Fouriera sygnału w celu umożliwienia jego analizy w dziedzinie częstotliwości. Wadą klasycznej transformaty Fouriera jest brak jawnych informacji o czasie w widmie sygnału, która jest bardzo istotna w analizie sygnału mowy, ponieważ jest on ciągiem pewnych zdarzeń (zmian częstotliwości, amplitudy oraz następstw fonemów i słów), których kolejność jest bardzo istotna. Dlatego stosuje się tzn. krótko-okresową transformatę Fouriera (STFT ang. Short-Time Fourier Transform) zwaną też okienkową transformatą Fouriera (ang. Windowed Fourier Transform). Polega ona na dokonywaniu transformat krótkich fragmentów sygnału wyznaczonych za pomocą okna $w(t)$. Dyskretna realizacja tego przekształcenia ma postać:

$$F(k, \tau) = \frac{1}{\sqrt{M}} \sum_{t=0}^{M-1} [f_{\tau+t} e^{-i2\pi kt/M} * w_{\tau}^t]$$

$F(k, \tau)$ jest transformatą Fouriera, dla pojedynczej ramki (o indeksie τ , częstotliwości k , oraz szerokości M) otrzymanej w chwili $t = \tau$ uzyskaną wskutek złożenia sygnału $f(t)$ z funkcją okna w .

Szerokość okna determinuje rozdzielczość częstotliwościową i czasową otrzymanego widma czasowo-częstotliwościowego. Najprostszą funkcją okna jest okno prostokątne:

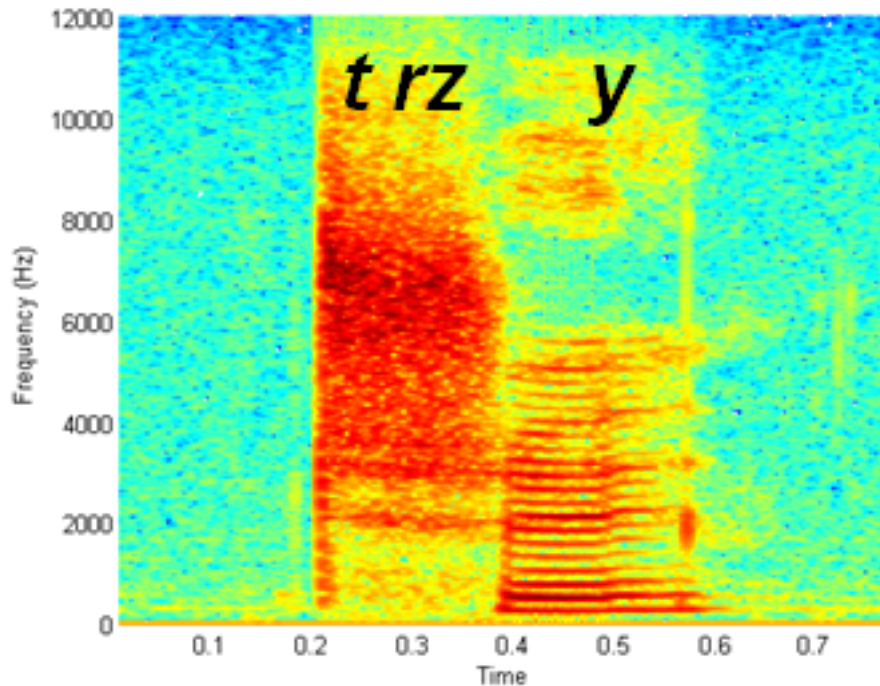
$$w_{\tau}^t = \begin{cases} 1 & \text{dla } t = \{\tau+0, \tau+1, \dots, \tau+M-1\}, \\ 0 & \text{w przeciwnym razie.} \end{cases}$$

Wprowadza ono jednak znaczne zniekształcenia widma analizowanego sygnału, które są związane z efektami brzegowymi wyciętego fragmentu. [17] To zjawisko możemy minimalizować stosując inny typ okna, np. okno Hamminga, które ma węższe widmo:

$$w_{\tau}^t = \begin{cases} 0.54 - 0.46 \cos\left(\frac{2\pi t}{M-1}\right) & \text{dla } t = \{\tau+0, \tau+1, \dots, \tau+M-1\}, \\ 0 & \text{w przeciwnym razie.} \end{cases}$$

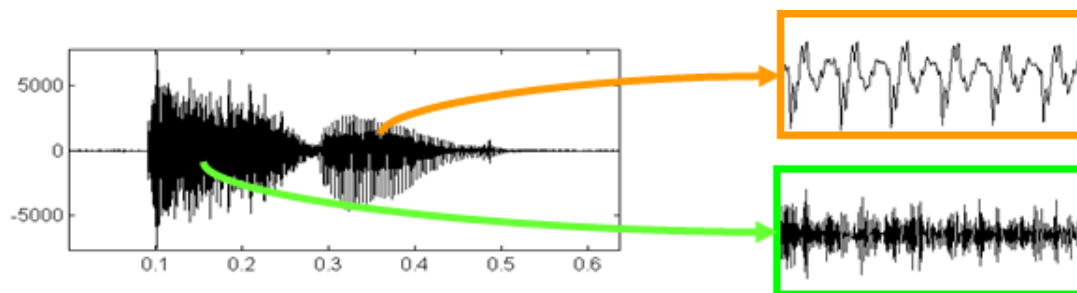
Niestety okno Hamminga posiada również pewne negatywne właściwości, mianowicie poszerza wszystkie pasma filtrów analizujących widmo co wiąże się pogorszeniem rozdzielczości częstotliwościowej prowadzonej analizy. [19]

Po dokonaniu transformaty sygnał mowy może zostać przedstawiony na spektrogramie (rys. 5).



Rysunek 5: Reprezentacja słowa „trzy” w dziedzinie częstotliwości

Patrząc na spektrogram (rys. 5) możemy zauważyć różnice pomiędzy głoskami bezdźwięcznymi a głoską dźwięczną „y”. W przypadku głosek dźwięcznych możemy wyróżnić pewną okresowość, natomiast głoski bezdźwięczne często mieszają się z szumem (rys. 6). Co za tym idzie, słowa różniące się tylko częścią bezdźwięczną trudno rozpoznać.



Rysunek 6: Segment dźwięczny i bezdźwięczny słowa „trzy”

Typowym formatem zapisu sygnału analogowego w postaci cyfrowej jest WAV (ang. wave form audio format), został on opracowany przez Microsoft oraz IBM w roku 1991. Jest to format bezstratny.

Strukturę pliku WAV (tab. 1) można opisać następująco.

- Początkowy napis „RIFF” identyfikuje rodzaj pliku jako RIFF.
- Następnie 32-bitowe pole podaje rozmiar danych = rozmiar pliku - 8 bajtów.
- Napis „WAVE” identyfikuje rodzaj pliku dźwiękowego.
- Napis „fmt” identyfikuje pole formatu danych, elementy tego formatu to:
 - rozmiar pola formatu (16 bajtów struktury WAVEFORMATEX),
 - pole „nChannels” (= 1- mono, 2- stereo),
 - pole „nSamplesPerSec” (= próbki na sekundę, np. 44 100),
 - pole „nAvgBytesPerSec” (= próbki na sekundę * liczba kanałów * rozmiar próbki / 8, zaokrąglone w górę),
 - pole „nBlockAlign” (= liczba kanałów * rozmiar próbki w bajtach / 8, zaokrąglone w górę),
 - pole „wBitsPerSample” (= liczba kanałów * rozmiar próbki w bajtach).
- Napis „data”.
- 32-bitowy format zapisu danych użytkowych.
- Wszystkie próbki danych użytkowych.

2.5. Detekcja sygnału mowy

W każdym systemie analizy mowy możemy wyróżnić początkowe etapy analizy tzn. przetwarzanie wstępne do którego mogą należeć następujące kroki:

- preemfaza, stosuje się ją w celu wzmocnienia wyższych częstotliwości w sygnale mowy osłabionych na skutek filtracji przez kanał transmisyjny,

Tablica 1: Struktura pliku WAV [19]

Adres bazowy	Liczba bajtów	Dane
0000	4	Napis „RIFF”
0004	4	Rozmiar danych = rozmiar pliku -8
0008	4	Napis „WAVE”
000C	4	Napis „fmt”
0010	4	Rozmiar formatu (16 bajtów)
0014	2	wf.wFormatTag = WAVEFORMAT_PCM = 1
0016	2	wf.nChannels
0018	4	wf.nSamplesPerSec
001C	4	wf.nAvgBytesPerSec
0020	2	wf.nBlockAlign
0022	2	wf.wBitsPerSample
0024	4	Napis „data”
0028	4	Format danych użytkowych
002C		DANE użytkowe

- usuwanie zakłóceń impulsowych z sygnału za pomocą fitru medianowego,
- normowanie parametrów zależnych od mówcy,
- odszumianie sygnału mowy (metoda odejmowania widmowego (ang. spectral subtraction), filtr Kalmana, filtr Wienera, metody perceptualne),
- inne rodzaje filtracji. [18]

Opisać wszystkie podpunkty?

W widmie mowy więcej energii znajduje się w niskich częstotliwościach, to niekorzystne z punktu widzenia przetwarzania sygnałów zjawisko nazywane jest przekrzywieniem widma (ang. spectral tilt). Aby usunąć skutki przekrzywienia stosuje się w/w preemfazę według następującego wzoru:

$$y(n) = x(n) - a * x(n - 1),$$

gdzie a to współczynnik preemfazy $0.9 \leq a \leq 1$. [19]

Detekcję sygnału mowy wykonuje się w celu podniesienia skuteczności rozpoznawania oraz skrócenia czasu obliczeń. Dokonuje się jej po uprzednim podziale sygnału mowy na ramki. Można ją przeprowadzić w oparciu o kilka parametrów (tab. 2).

Tablica 2: Parametry detekcji sygnału mowy

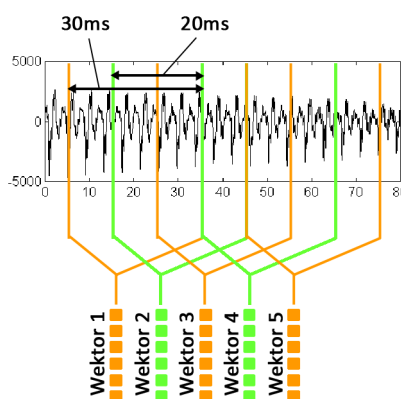
Parametr	Kryterium detekcji
Moc sygnału w ramce $P = 10 \log(\varepsilon + \frac{1}{N} \sum_{i=1}^N x_i^2)$	$P \geq P_{tr}$
Liczba przejść przez zero w ramce L	$L < L_{tr1}$ lub $L > L_{tr2}$
Entropia widma E $H = - \sum_{k=1}^N w_k p_k \log p_k$ gdzie $p_k = s(f_k) / \sum_{i=1}^N s(f_i), k = 1, \dots, N$	$H \geq H_{tr}$

W każdej ramce obliczany jest wybrany parametr a następnie na podstawie jego wartości podejmowana jest decyzja o zaakceptowaniu ramki jako sygnału mowy lub jej odrzuceniu. W przypadku stosowania entropii widmowej najpierw dokonuje się obliczenia modułu FFT. Wagi w_k dobierane są doświadczalnie. Odrzucane są najmniejsze i największe wartości p_k .

Możemy też przyjąć zasadę, że po wstępnej detekcji sygnału mowy odrzuca się fragmenty sygnału krótsze niż 100 ms, chyba że leżą w odległości czasowej mniejszej niż 100 ms od innych fragmentów sygnału. [Przypis z Pana wykładów?] [18]

2.6. Ekstrakcja cech sygnału mowy

W wyniku ekstrakcji cech otrzymujemy wartości parametrów zawierających informację o treści wypowiedzi i będących niezależnymi od indywidualnych cech głosu mówcy. Parametry te tworzą wektory cech, na podstawie których dokonuje się klasyfikacji sygnału. Przykładowy podział na ramki pokazano na rysunku 7, wektory 6 parametrów ekstrahowanych w ramkach o czasie trwania 30 ms, zachodzących na siebie na odcinkach 20 ms.



Rysunek 7: Ekstrakcja cech

Najczęściej stosowane cechy sygnału mowy:

- bank filtrów,
- współczynniki LPC (ang. Linear Prediction Coefficients),
- parametry cepstralne,
- parametry mel-cepstralne (MFCC ang. Mel-frequency cepstral coefficients),
- parametry mel-spektralne (MFC ang. Mel-frequency cepstrum),
- parametry dyskretnej transformacji falkowej (ang. DWT - Discrete Wavelet Transform). [18]

Każdy zestaw cech może być uzupełniony o tzn. cechy dynamiczne, czyli współczynniki różnicowe. Polega to na obliczeniu pierwszej i drugiej pochodnej powyższych współczynników (tzw. współczynników delta oraz delta-delta) względem kilku ramek.

Pakiet HTK może korzystać zarówno z cech mel-cepstralnych jak i LPC. Opisane zostaną jedynie cechy mel-cepstralne oraz mel-spektralne ponieważ są ze sobą powiązane. W systemie stworzonym na potrzeby tej pracy zostały użyte cechy mel-cepstralne.

Ucho człowieka reaguje nieliniowo na zmieniającą się częstotliwość dźwięku. Częstotliwości powyżej 1 kHz są słabiej odczuwane niż różnice w zakresie niskich częstotliwości. Dlatego im wyższa częstotliwość tym są potrzebne coraz większe odstęp między kolejnymi pasmami dla zrekompensowania nieliniowości. Po to właśnie wprowadzono skalę melową (Mel) zamiast hercowej (Hz). [19]

$$\omega_{Mel} = 2595 \log(1 + \frac{\omega}{700Hz})$$

Widma ramek sygnału po uprzednim przeprowadzeniu DFT poddawane są filtracji za pomocą melowego banku filtrów pasmowo przepustowych. W systemach rozpoznawania mowy zazwyczaj stosuje się banki 26 filtrów melowych. Filtry tworzone są dla kolejnych pasm częstotliwości, rozmieszczonych w nieliniowy sposób wyznaczony przez skalę Mel. Zdefiniowane są w dziedzinie częstotliwości co umożliwia łatwe wymnożenie ich przez przekształcony sygnał. Do wyznaczenia cech mel-spektralnych dla każdej ramki sygnału wykorzystuje się zbiór l trójkątnych filtrów $D(l, k)$.

$$MFC(l, \tau) = \sum_{k=0}^{M-1} [D(l, k) * FC(k, \tau)], l=1, \dots, L$$

Wartość pojedynczego współczynnika MFC odpowiada ważonej sumie wartości FC należących do zakresu trójkątnego filtra pasmowego odpowiadającego danemu MFC.

Chcąc wyliczyć spektrum energii dla każdej ramki, stosujemy FFT dla każdego splotu sygnału z kolejnym oknem i obliczamy kwadrat amplitudy każdego współczynnika zespolonego Fouriera.

$$FC(k, \tau) = | \sum_{t=0}^{M-1} [x(\tau + t) e^{-i2\pi kt/M} * w_{\tau}(t)] |^2 \text{ dla } k = 0, \dots, M-1$$

Współczynniki mel-cepstralne możemy wyznaczyć ze wzoru:

$$MFCC(k, \tau) = \sum_{l=0}^{L-1} [\log MFC(l, \tau) * \cos(\frac{k * (2l+1)\pi}{2L})], k = 1, \dots, K$$

Natężenie dźwięku jest odczuwane przez ludzi w skali logarytmicznej dlatego przy obliczaniu cepstrum sygnał otrzymany po przejściu przez bank filtrów melowych jest logarytmowany.

„Ponieważ układ głosu ma charakter ciągły, zatem poziomy energii w sąsiednich pasmach są skorelowane. Dlatego też niezbędna do tego transformata odwrotna Fouriera (tu wystarczy przekształcenie cosinusowe) zamienia zbiór logarytmów energii na nieskorelowane ze sobą współczynniki cepstralne.” [19]

Aby usunąć szkodliwy wpływ podstawowych drgań krtaniowych na zestaw cech możemy zastosować przekształcenie zwane liftowaniem.

$$c_n^{lift} = (1 + \frac{\lambda}{2} \sin(\frac{\pi n}{\lambda})) c_n \text{ dla } n = 1, \dots, \lambda,$$

gdzie c_n to n -ty współczynnik MFCC a stała λ odpowiada indeksowi cechy związanej z częstotliwością podstawową.

Podstawowy wektor cech możemy uzupełnić o tzn. cechę „energetyczną”, która pomaga odróżnić sygnał cisy od sygnału mowy a także słabe bezdźwięczne spółgłoski od silnych dźwięcznych samogłosek. Sumaryczną energię w ramce sygnału τ obliczamy sumując kwadraty próbek w dziedzinie czasu według następującego wzoru:

$$E(\tau) = \sum_{i=1}^M f_i^2 \quad [19][13]$$

Podsumowując, wyznaczanie współczynników MFCC możemy przestawić w następujących krokach:

- blokowanie sygnału w ramki, okienkowanie oknem Hamminga,
- przeprowadzenie FFT na zokienkowanych ramkach sygnału,
- filtracja za pomocą melowego banku filtrów,
- obliczenie mocy FFT w określonych pasmach częstotliwościowych,
- obliczenie logarytmu zakumulowanych współczynników widmowych,
- przeprowadzenie DCT na zlogarytmowanych współczynnikach widmowych,
- opcjonalnie, wyznaczenie cech dynamicznych.

DCT (ang. discrete cosine transform) przeprowadzamy według następującego wzoru:

$$x(n) = c(n) \sum_{k=0}^{K-1} \ln(S_k) \cos\left(\frac{\pi(2k+1)n}{2K}\right),$$

$$\text{gdzie } c(0) = \sqrt{\frac{1}{K}}, \text{ dla } n > 0 \text{ } c(n) = \sqrt{\frac{2}{K}} \quad [17]$$

Otrzymany wektor cech MFCC zawiera liczbę elementów równą liczbie pasm melowych. Do dalszego przetwarzania bierzemy zazwyczaj pierwsze 12 współczynników, do których dokładamy energię sygnału w ramce oraz jeśli chcemy uwzględnić dynamikę zmian współczynników w czasie (wielkość zmian oraz ich tempo), poszerzamy nasz wektor o ich przyrosty kolejno pierwszego i drugiego rzędu (czyli współczynniki delta oraz delta-delta). W wyniku tych operacji otrzymujemy 39-elementowy wektor cech mel-cepstralnych.

Czy pod rysunkami z Pana wykładów dawać przypisy?

3. Ukryte Modele Markowa jako jedna z metod klasyfikacji

Współczesne systemy rozpoznawania mowy zakładają, że sygnał mowy jest sekwencją pewnych elementarnych jednostek fonetycznych (np. głosek, fonemów, słów). Charakteryzuje się on również pewną przypadkowością. Wypowiedź tego samego mówcy za każdym razem posiada niepowtarzalny przebieg. A co za tym idzie nie możemy przyporządkować jej jednoznacznej sekwencji jednostek mowy. Wiąże się to z tym, że nagrany sygnał mowy często bywa zaszumiony i zniekształcony.

Metody klasyfikacji wykorzystywane w systemach rozpoznawania mowy możemy podzielić na dwie kategorie:

- metody deterministyczne, gdzie obliczane są błędy porównując dźwięk ze wzorcem,
- metody niedeterministyczne, gdzie obliczamy wartości prawdopodobieństw, które reprezentują dopasowanie dźwięku do stosowanych modeli probabilistycznych. [11]

Do metod deterministycznych należy m.in. nieliniowa transformata czasowa w skrócie DTW (ang. dynamic time wrapping), metoda skuteczna dla izolowanych słów.

Inne ograniczenia metod deterministycznych:

- duża zajętość pamięci potrzebnej na przechowywanie wzorców (dla dużych słowników),
- czas rozpoznawania proporcjonalny do rozmiaru słownika. [18]

Biorąc pod uwagę to, że bardziej naturalnym dla człowieka jest wypowiadanie słów ciągiem (bez znaczących przerw między nimi) oraz, że przy analizie sygnału mowy warto uwzględnić jego przypadkowość, metody niedeterministyczne osiągają lepszą skuteczność rozpoznawania. Do takich właśnie metod należą Ukryte Modele Markowa, w skrócie HMM (ang. hidden markov models).

3.1. Wprowadzenie teoretyczne

„Ukryte Modele Markowa stanowią serce większości współczesnych systemów rozpoznawania mowy.” [12] Sprawdzają się również w wielu innych zastosowaniach, tam gdzie mamy doczynienia z sygnałem niedeterministycznym, np.:

- rozpoznawanie obrazów,
- modelowaniu DNA,

- oraz modelowaniu danych ekonomicznych.

Pierwsze prace na temat Ukrytych (niejawnych) Modeli Markowa prowadził Baum wraz ze współpracownikami w latach 60-tych oraz 70-tych. Dotyczyły one metody estymacji parametrów modeli HMM nazwanej algorytmem Bauma-Welcha. Algorytm ten jest oparty na kryterium (ML) (Maximum Likelihood) maksymalizacji prawdopodobieństwa wygenerowania sekwencji obserwacji przez model HMM. [11]

W bloku klasyfikacji następuje porównanie nadchodzących ciągów obrazów wypowiedzi ze znajdującymi się w pamięci wzorcami, które stanowią uogólniony (usredniony) opis klas dźwięków. Klasa mogą być fonemy, wyrazy lub nawet całe zdania. Obrazy wzorcowe są tworzone w procesie uczenia.

W przypadku rozpoznawania mowy rozważa się jednak nieco odmienną sytuację, gdy łańcuch q uznawany jest za niejawny (ukryty). Znana jest natomiast sekwencja obserwacji O . W przypadku mowy, obserwacja jest to wektor cech wyekstrahowanych z pojedynczej ramki nagrania.

3.2. Algorytm Bauma-Welcha

3.3. Algorytm Viterbiego

4. HTK (Hidden Markov Model Toolkit)

4.1. Rozpoczęcie pracy z bibliotekami

4.2. Trenowanie modeli

4.3. Rozpoznawanie

5. System sterujący robotem za pomocą komend głosowych

5.1. Schemat działania systemu

5.2. Użyte technologie

5.3. Robot przemysłowy

5.4. Urządzenie na którym działa cały system

5.5. Gry planszowe

5.6. Prezentacja działania systemu

6. Podsumowanie

Bibliografia

- [1] *CMUSphinx*. <http://cmusphinx.sourceforge.net>.
- [2] *HTK (Hidden Model Markov Toolkit)*. <http://htk.eng.cam.ac.uk/>.
- [3] *Julius*. http://julius.sourceforge.jp/en_index.php.
- [4] *MagicScribe*. <http://www.magicscribe.pl>.
- [5] *Primespeech*. <http://www.primespeech.pl>.
- [6] *SARMATA*. <http://www.dsp.agh.edu.pl/doku.php?id=pl:resources:asr>.
- [7] Ostaszewska D. *Fonetyka i Fonologia współczesnego języka polskiego*. Wydawnictwo Naukowe PWN, Warszawa, 2000.
- [8] Apple Inc. *Siri*. <http://www.apple.com/ios/siri>.
- [9] Google Inc. *Google Mobile App*. <http://www.google.com/mobile/google-mobile-app>.
- [10] Google Inc. *Google Now*. <http://www.google.com/landing/now>.
- [11] Szostek K. *Rozpoznawanie mowy metodami niejawnych modeli Markowa*. PhD thesis, AGH Kraków, 2006.
<http://winntbg.bg.agh.edu.pl/rozprawy/9792/full9792.pdf>.
- [12] Gałka J. *Optymalizacja parametryzacji sygnału w aspekcie rozpoznawania mowy polskiej*. PhD thesis, AGH Kraków, 2008.
<http://winntbg.bg.agh.edu.pl/rozprawy2/10009/full10009.pdf>.
- [13] Kowalski A. B. Kasprzak W. Analiza sygnału mowy sterowana danymi dla rozpoznawania komend głosowych. *Postępy Robotyki, WKiŁ*, 2006. http://www.ia.pw.edu.pl/~wkasprza/PAP/kkr06_rkg.pdf.
- [14] Grad L. Obrazowa reprezentacja sygnału mowy. *Biuletyn Instytutu Automatyki i Robotyki WAT nr: 8*, 1997. http://www.ita.wat.edu.pl/~l.grad/sieci%20neuronowe/ekstrakcja_cech_mowy.pdf.
- [15] Microsoft. *TellMe*. <http://www.microsoft.com/en-us/Tellme>.

- [16] Nuance. *Dragon NaturallySpeaking*. <http://www.nuance.com/dragon/index.htm>.
- [17] Zieliński T. P. *Cyfrowe przetwarzanie sygnałów od teorii do zastosowań*. WKiŁ, Warszawa, 2005.
- [18] Wielgat R. Wykłady z przedmiotu Techniki Multimedialne. PWSZ Tarnów., 2012.
- [19] Kasprzak W. *Rozpoznawanie obrazów i sygnałów mowy*. Oficyna Wydawnicza Politechniki Warszawskiej, Warszawa, 2009.

Dodatek A

Tu, trzeba zamieszczać treść dodatkową np. fragmenty kodu aplikacji itp.