

Rafał Cieślak

nr albumu : 34203

kierunek studiów: Informatyka

specjalność: Systemy komputerowe i opramowanie

forma studiów: stacjonarne

Identyfikacja akustyczna rodzaju zdania w systemach dialogowych

Acoustic identification of sentence type in dialogue systems

Praca dyplomowa inżynierska wykonana pod przewodnictwem:

dr inż. Tomasz Mąka

Szczecin 2019

Spis treści

Wstęp	1
1 Wprowadzenie teoretyczne	2
1.1 Sygnał mowy	2
1.1.1 Powstawanie mowy	2
1.1.2 Reprezentacja mowy	4
1.1.3 Rozumienie mowy	4
1.1.4 Rejestrowanie sygnału mowy	5
1.2 Ton podstawowy	5
1.2.1 Definicja tonu podstawowego	5
1.2.2 Formanty	6
1.2.3 Przegląd metod estymacji	6
1.2.4 Algorytm YIN	7
1.3 Prozodia	7
1.3.1 Intonacja	7
1.3.2 Funkcje intonacji	8
1.3.3 Analiza dotychczasowych badań	9
2 Implementacja detekcji konturów częstotliwości podstawowej	12
2.1 Język programowania oraz środowisko	12
2.2 Opis możliwości aplikacji	13
2.3 Wczytanie nagrania	13
2.4 Ekstrakcja tonu podstawowego	15
2.4.1 Ramkowanie oraz ekstrakcja wartości F0	15
2.5 Wykrywanie poszczególnych segmentów	18
2.5.1 Analiza wstępna wykrytego fragmentu	21
2.5.2 Współczynniki regresji liniowej	23
2.6 Wczytywanie wartości uzyskanych za pomocą programu PRAAT	24

3	Analiza wykrytych segmentów	27
3.1	Pytania rozstrzygnięcia	29
3.2	Pytania dopełnienia	34
3.3	Zdania rozkazujące	37
3.3.1	Zdania twierdzące	44
3.4	Porównanie wyników otrzymanych z wykorzystaniem YIN i Praata . . .	46
3.4.1	Pytania rozstrzygnięcia oraz dopełnienia	46
3.4.2	Zdania rozkazujące oraz zdania twierdzące	47
3.4.3	Błędne rozpoznania	48
4	Wnioski	50

Spis rysunków

1.1	Faldy głosowe [1]	3
2.1	Interfejs graficzny aplikacji	13
2.2	Zobrazowany podział sygnału na ramki wraz zastosowaniem 30-procentowego zakładkowania. Opracowanie własne	16
2.3	Klasa stworzona w celu ekstrakcji F0 oraz przechowywania tych wartości .	16
2.4	Klasy stworzone do wykrycia poszczególnych segmentów intonacyjnych, na podstawie wszystkich wartości F0	19
2.5	Przykład usuniętego segmentu.	23
2.6	Fragment przebiegu intonacji przed i po nałożeniu linii regresji	24
3.1	Przedstawienie analizowanych cech przebiegu intonacji	28
3.2	Pytanie rozstrzygnięcia zadane przez mężczyznę	30
3.3	Schemat rozpoznawania pytań rozstrzygnięcia	31
3.4	Pytanie rozstrzygnięcia z nacechowaniem emocjonalnym	32
3.5	Schemat rozpoznawania pytań rozstrzygnięcia z tendencją rosnącą	33
3.6	Pytanie rozstrzygnięcia zadane przez kobietę	34
3.7	Pytanie dopełnienia zadane przez mężczyznę	35
3.8	Pytanie dopełnienia zadane przez mężczyznę	35
3.9	Schemat rozpoznawania pytań dopełnienia	36
3.10	Schemat wykrywania najwyżej położonego segmentu	37
3.11	Zdanie rozkazujące wypowiedziane przez kobietę	38
3.12	Schemat wykrywania zdania rozkazującego zawierającego pierwszą grupę cech	38
3.13	Zdanie rozkazujące wypowiedziane przez mężczyznę	39
3.14	Schemat wykrywania zdania rozkazującego zawierającego drugą grupę cech	39
3.15	Zdanie rozkazujące wypowiedziane przez kobietę	40
3.16	Schemat wykrywania zdania rozkazującego zawierającego trzecią grupę cech	40
3.17	Zdanie rozkazujące wypowiedziane przez mężczyznę	41

3.18 Schemat wykrywania zdania rozkazującego zawierającego czwartą grupę cech	42
3.19 Zdanie rozkazujące wypowiedziane przez kobietę	42
3.20 Zdanie rozkazujące wypowiedziane przez mężczyznę	43
3.21 Schemat wykrywania zdania rozkazującego zawierającego piątą grupę cech	43
3.22 Zdanie twierdzące wypowiedziane przez kobietę	44
3.23 Zdanie twierdzące wypowiedziane przez mężczyznę	45
3.24 Zdanie twierdzące wypowiedziane przez kobietę	45
3.25 Intonacja pytania przedstawionego na rysunku ... lecz uzyskana za pomocą PRAAT-a	47
3.26 Intonacja pytania przedstawionego na rysunku ... lecz uzyskana za pomocą PRAAT-a	47
3.27 Intonacja zdania rozkazującego przedstawionego na rysunku ... lecz uzyskana za pomocą PRAAT-a	48
3.28 Intonacja zdania twierdzącego przedstawionego na rysunku ... lecz uzyskana za pomocą PRAAT-a	48
3.29 Intonacja zdania rozkazującego przedstawionego na rysunku ... lecz uzyskana za pomocą PRAAT-a	49

Streszczenie

Niniejsza praca inżynierska dotyczy analizy przebiegu intonacji w zdaniach wypowiedzianych w języku polskim. Omówione zostały zagadnienia związane z wytwarzaniem mowy, częstotliwością podstawową oraz funkcjami intonacji. Celem pracy było zaobserwowanie charakterystycznych cech, towarzyszących każdemu z rodzajów wypowiedzi. W ramach pracy sporządzono bazę 90 nagrań, których intonacja była poddawana analizie wizualnej. Na podstawie tych obserwacji zaimplementowany został program, dokonujący klasyfikacji wypowiedzi. Do ekstrakcji intonacji użyty został algorytm YIN. W ramach pracy porównano również skuteczność zaproponowanej metody klasyfikacji dla wartości intonacji uzyskanych za pomocą programu PRAAT.

Abstract

This engineering thesis concerns analysis of intonation in utterances spoken in Polish. Among discussed topics there were ones related to speech production, fundamental frequency and functions of intonation. The purpose of this thesis was to observe the characteristic features that accompany various types of utterance. As part of a work, a database of 90 recordings was prepared. Their intonation was analysed. Basing on these observations, the classifying program has been implemented. The YIN algorithm was used as a way to isolate the intonation. As part of the work, a comparison was made between the classification results obtained using the proposed method, for intonation isolated by YIN and PRAAT.

Wstęp

Mowa jest najpowszechniejszym sposobem komunikacji międzyludzkiej. Używając jej na codzień, nie zdajemy sobie sprawy jak bardzo złożonym jest procesem. Wydaje się być czymś normalnym i oczywistym. Mimo tego, że otacza nas cały czas, nauka wciąż nie poznała dokładnie wszystkich mechanizmów stojących za jej wytwarzaniem i rozumieniem. Złożoności dodaje fakt, że w mowie nie chodzi tylko o wypowiedziane słowa. Dla naszego odbierania mowy ważna jest również cała otoczką - ton głosu, jego barwa, akcentowanie i wiele więcej cech, których postrzeganie jest subiektywne. Dawno zostało zauważone, że cechy te mogą istotnie wpływać na nasze życie. Ludzie podświadomie wybierają towarzystwo osób, których głos jest dobrze przez nich postrzegany. Bez tych otaczających wypowiedziane słowa cech, nasza mowa brzmiałaby podobnie do uzyskiwanej za pomocą syntezatorów.

Jedną z takich cech jest intonacja. Przy tym jest jednym ze słabiej poznanych zagadnień związanych z wytwarzaniem oraz percepcją mowy. Nie niesie ze sobą żadnej semantycznej treści, dotyczy tego w jaki sposób coś mówimy, a nie co mówimy. Nie zdajemy sobie na codzień z tego sprawy ale bez niej bardzo trudne byłoby zrozumienie języka mówionego i przekazywanych za jego pomocą myśli. Wpływa na nasze postrzeganie zdań wypowiedzianych przez naszego rozmówcę. Dzięki niej momentalnie wiemy czy osoba, z którą rozmawiamy zadała nam pytanie czy nakazała nam wykonanie jakiejś czynności. Nie potrzebujemy do tego znaków przestankowych, używanych w języku pisanym.

Celem pracy jest zrozumienie jak bardzo wpływa na ten proces i odpowiedź na pytanie czy analizowanie samej intonacji jest wystarczające do stwierdzenia do jakiej kategorii można zakwalifikować daną wypowiedź. W tym celu podjęta została próba wykrycia charakterystycznych cech intonacji, związanych z różnymi typami wypowiedzi, których zastosowanie umożliwi automatyczne klasyfikowanie zdań przez program.

Jako, że analiza będzie się opierać w głównej mierze na próbach zauważenia charakterystycznych różnic między różnymi wypowiedziami, metodą badawczą zastosowaną w tej pracy będzie metoda obserwacyjna.

Praca została podzielona na trzy główne rozdziały. W rozdziale pierwszym przedstawione zostały istotne zagadnienia teoretyczne związane z produkcją mowy, wytwarzaniem częstotliwości podstawowej, oraz przybliżone zostały funkcje intonacji. W rozdziale drugim omówiony został sposób implementacji wykrywania poszczególnych segmentów w przebiegu całej intonacji. W rozdziale trzecim przedstawione zostały rezultaty analizy charakterystycznych cech dla poszczególnych rodzajów wypowiedzi.

Rozdział 1

Wprowadzenie teoretyczne

1.1 Sygnał mowy

Mową określamy komunikowanie się między sobą ludzi, za pomocą ukształtowanego zbioru dźwięków i reguł, zwanego językiem. Każdy język używa własnych fonetycznych kombinacji zbioru spółgłosek i samogłosek, które tworzą słowa mające semantyczne znaczenie. W czasie mówienia, osoba mówiąca poza samym wypowiedzianiem słów, nadaje wypowiedzi znaczenie również za pomocą dodatkowych aspektów, takich jak intonacja, tempo mówienia czy stopień głośności. Sama produkcja mowy jest wielokrokovym procesem zamiany myśli w ustną wypowiedź, która może być zarejestrowana jako sygnał mowy.

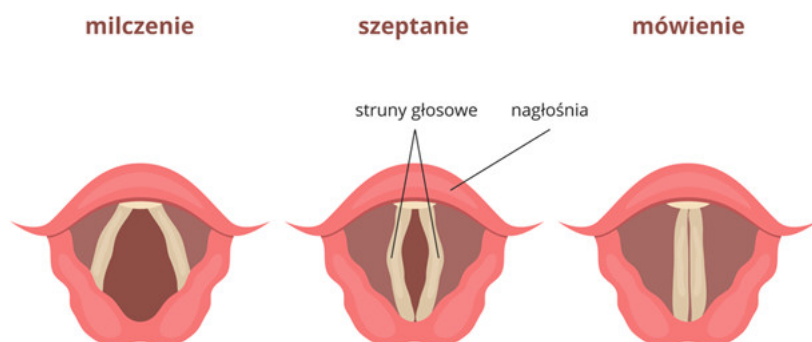
1.1.1 Powstawanie mowy

Sygnał mowy ludzkiej jest sygnałem akustycznym powstającym podczas przepływu powietrza poprzez aparat mowy, który jest definiowany jako 3 osobne grupy narządów [2].

Składowymi aparatu mowy są:

1. Aparat oddechowy. Bierze udział w początkowej fazie powstawania mowy, dostarczając kolejnym składowym strumień powietrza, który jest niezbędny do wygenerowania drgań. Dzieje się to podczas wydechu. Elementy, z których jest zbudowany to płuca, oskrzela, przepona oraz tchawica.
2. Aparat fonacyjny, którego głównym elementem jest krtąń. Jest to narząd niezbędny do wygenerowania jakiegokolwiek dźwięku, nie tylko mowy. Najważniejszym elementem krtani, w kontekście procesu powtarzania dźwięku, są fałdy głosowe. W ich skład wchodzi więzadła głosowe oraz mięśnie głosowe. Przestrzeń pomiędzy nimi nazywana jest szparą głośni. Struktury te przybliżają się i oddalają od siebie podczas powstawania dźwięku co powoduje zwanie i rozwanie szpary głośni. Podczas

oddychania oraz przy generowaniu głosek bezdźwięcznych, fałdy są rozsunięte, natomiast zwierają się i rozwierają podczas powstawania głosek dźwięcznych. Dzięki tej czynności, strumień powietrza wprowadzany jest w drgania, co postrzegamy jako dźwięczność. Cecha ta występuje wraz z każdą samogłoską oraz przy niektórych spółgłoskach. Podczas drgań generowany jest ton krtaniowy, zwany również częstotliwością podstawową, oznaczany w literaturze jako F0.



Rysunek 1.1: Fałdy głosowe [1]

3. Aparat artykulacyjny, w którego skład wchodzi jamy przewodu oddechowego, znajdującego się ponad krtanią. Najważniejsze z punktu widzenia artykulacji - nosowa, gardłowa oraz ustna - nazywane są nasadą. Artykulatory znajdujące się w nasadzie dzielone są na ruchome oraz nieruchome. Do ruchomych zaliczamy język, podniebienie miękkie, wargi oraz zuchwę. Nieruchomymi określamy zęby, dziąsła oraz podniebienie twarde. Ich ustawienie ostatecznie determinuje cechy wytwarzanego dźwięku.

W zależności od tego czy dana głoska jest dźwięczna czy bezdźwięczna proces powstawania dźwięku przebiega w trochę inny sposób. W obu przypadkach w początkowej fazie wzrasta ciśnienie w płucach, co prowadzi do wydechu [4]. Powietrze dostaje się do tchawicy. Na szczycie tchawicy znajduje się krtąń, należąca do aparatu fonacyjnego. W przypadku głosek dźwięcznych, w miarę przepływu powietrza przez głosnię, spada lokalne ciśnienie, co pozwala mięśniom krtani zamknąć głosnię, przerywając przepływ powietrza. To powoduje wzrost ciśnienia, prowadzący do kolejnego oddalenia się strun głosowych. Cały ten cykl zapętla się, struny wibrują tworząc dźwięk, kierowany do aparatu artykulacyjnego. Na tym etapie, poza artykulacją, zachodzi również tłumienie niektórych częstotliwości, nie będących harmonicznymi fali głosniowej. Nie wytłumione zostają tylko częstotliwości będące bliskie naturalnemu rezonansowi traktu głosowego.

Rusząc szczęką, ustami lub zmieniając położenie języka, możemy zmieniać uzyskiwany dźwięk, ponieważ zmieni się rezonans traktu głosowego, a zatem inne częstotliwości zostaną wytłumione. Gdy wypowiadane są głoski bezdźwięczne, krtani nie odgrywa istotnej roli, a modulacja dźwięku odpowiedzialna za uzyskanie brzmienia głoski odbywa się w aparacie artykulacyjnym. Jako rezultat kompletnego procesu, uzyskiwana jest fala akustyczna, wydostająca się z ust. Prawdziwość opisanych różnic między dwoma rodzajami głosek można sprawdzić w prosty sposób, przykładając palce do krtani. W czasie wypowiedzenia głosek dźwięcznych czyli wszystkich samogłosek oraz części spółgłosek, takich jak b, d, g, w, z, ż, ż, l, ł, r, m, n, j, dz, dź, dż, wyczuwalne będą wibracje, które nie wystąpią podczas wypowiedzenia głosek bezdźwięcznych, takich jak p, t, k, f, s, ś, sz, c, ć, cz, ch. Fakt, że wiele różnych narządów bierze udział w tworzeniu mowy powoduje, że zaburzenia zdrowotne każdego z nich mają istotny wpływ na cały proces. Zakres powstałych w ten sposób zaburzeń mowy jest szeroki - od drobnych wad wymowy do całkowitej utraty mowy.

1.1.2 Reprezentacja mowy

W procesie rozwoju technologii związanych z przetwarzaniem mowy, konieczne było ustalenie sposobu przedstawienia wypowiedzi za pomocą symboli reprezentujących wyprodukowany sygnał. Litery, używane w tym celu w języku pisanim, są niewystarczające, ponieważ w różnych wyrazach mogą być wymawiane na różne sposoby. Często produkowany dźwięk dla danej litery różni się w zależności od otaczających ją liter. Dla języka polskiego charakterystyczne jest występowanie tak zwanych dwuznaków, na przykład "rz,sz,ch". Dźwięk produkowany dla tych znaków jest całkowicie odmienny od dźwięków reprezentujących każdą z liter osobno. Jednym ze sposobów reprezentowania dźwięków powszechnie występujących w danym języku są fonemy. Są to najmniejsze elementy języka mówionego, pozwalające na rozróżnienie poszczególnych słów [10]. Często po zamienieniu jednego z fonemów składowych na inny, znaczenie słowa może ulec zmianie. W lingwistyce istnieją różne sposoby definiowania czym są fonemy oraz w jaki sposób dany język powinien być przez nie reprezentowany. Najczęściej jednak fonem jest rozumiany jako często powtarzający się w danym języku zbiór głosek. W języku polskim, w zależności od sposobu definiowania, liczba fonemów waha się od 31 do 42 [11].

1.1.3 Rozumienie mowy

Rozumieniem mowy nazywany jest proces, w trakcie którego wypowiedziana mowa jest słyszana, interpretowana oraz rozumiana przez człowieka. Badania nad postrzeganiem mowy są ściśle związane z lingwistyką oraz psychologią poznawczą i próbują odpowiedzieć

na pytanie w jaki sposób ludzie rozpoznają dźwięki mowy i na ich podstawie rozumieją mówiony język. Rezultaty tych poszukiwań mają swoje zastosowania w tworzeniu systemów komputerowych służących rozpoznawaniu mowy. Rozumienie mowy w danym języku jest ściśle związane z rozpoznawaniem przez mózg fonemów charakterystycznych dla tego języka. Z tego powodu często ludzie uczący się obcego języka znacznie łatwiej przyswajają język w formie pisanej niż mówionej.

1.1.4 Rejestrowanie sygnału mowy

Dźwięk opuszczający aparat mowy może zostać zarejestrowany przez mikrofon w celu poddania szczegółowej analizie. Aby możliwe było przetwarzanie sygnału przez program komputerowy, konieczne jest przetworzenie sygnału z postaci analogowej do cyfrowej. W tym celu pobiera się próbki sygnału. Wartość określającą ilość próbek w jednostce czasu nazywamy częstotliwością próbkowania. Najczęściej spotykana wartość to 44,1 kHz. Oznacza to, że podczas sekundy pobierane jest 44100 wartości sygnału ciągłego. Liczba ta została przyjęta jako standard przy nagrywaniu audio na płytach CD. Tak pobrane próbki, po poddaniu procesowi kwantyzacji, tworzą sygnał cyfrowy. Sygnał dźwiękowy może być nagrywany w wersji monofonicznej lub stereofonicznej. Oznacza to użycie jednego lub dwóch (lewy,prawy) kanałów. Nagrania rejestrowane tymi sposobami różnią się od siebie diametralnie, zarówno w kontekście subiektywnych odczuć słuchacza, jak i podczas przetwarzania sygnału. Kanały w wersji stereofonicznej mogą różnić się od siebie wartościami próbek, zwłaszcza w widmie sygnału.

1.2 Ton podstawowy

1.2.1 Definicja tonu podstawowego

W literaturze własność bywa również nazywana częstotliwością podstawową lub po prostu oznaczana jest jako F_0 . Pod tym pojęciem rozumiane są vibracje strun głosowych, tworzące powstawaniu głosek dźwięcznych. Powstałe w ten sposób częstotliwości mieszczą się w zakresie 85-180Hz dla mężczyzn oraz w zakresie 165-255Hz dla kobiet. Wartości te mogą być wyższe gdy osoba mówiąca znajduje się pod wpływem silnych emocji. Poza płcią oraz stanem emocjonalnym, zależne są również od wieku, budowy i kształtu strun głosowych ogólnego stanu zdrowia oraz rodzaju wypowiedzi. Badania nad częstotliwością podstawową produkowaną przez mężczyzn pokazały, że jej średnie wartości spadają po osiągnięciu 35 roku życia, by ponownie ulec wzrostowi po przekroczeniu 55 roku życia [12]. W przypadku kobiet, wartości F_0 zaczynają spadać w okresie menopauzy, osiągając finalne wartości około 70 roku życia [13]. Badania nad wpływem palenia papier-

rosów na wartości F_0 pokazały, że wieloletnie palenie również doprowadza do obniżenia tych wartości, jako że nawyk ten wpływa negatywnie na krtąń [14]. Przebieg częstotliwości podstawowej w dużym stopniu odzwierciedla intonację wypowiedzi. Gdyby ten przebieg był stały, mowa byłaby odbierana jako monotonna lub brzmiąca maszynowo. Pełni istotną funkcję w językach tonalnych, w których wielu słów jest zapisywanych tak samo, a jedynie nadawany im ton pozwala rozróżnić ich znaczenie. Z tego powodu też poprawna estymacja F_0 jest konieczna w systemach rozpoznawania mowy dla języków tonalnych. Dla idealnie okresowego sygnału, częstotliwość podstawowa byłaby po prostu odwrotnością okresu. Okresem nazywamy czas pomiędzy kompletnym cyklem otwarcia i zamknięcia głosu. Jednak sygnał mowy jest sygnałem bardzo dynamicznym, co sprawia, że estymacja F_0 przestaje być zadaniem trywialnym. Dodatkowo transformacja sygnału analogowego do postaci dyskretnej, wiążąca się zawsze z utratą danych oraz towarzyszący nagraniemu głosowi szum wpływają negatywnie na dokładność estymacji.

1.2.2 Formanty

Częstotliwość podstawowa powiązana jest w największej mierze z intonacją. Jednak w badaniach związanych z technologią przetwarzania mowy, wyznaczane z sygnału mowy są również inne częstotliwości, związane z rezonansem innych części traktu głosowego. Nie są one bezpośrednio związane z intonacją, lecz wiedza na ich temat jest istotna dla każdego badania związanego z sygnałami mowy. Są to formanty. Pod tym pojęciem rozumiane są skupiska energii akustycznej, zgromadzone wokół konkretnej częstotliwości w sygnale mowy. [15] Istnieje kilka formantów, lecz zazwyczaj wyznaczane są cztery - F_1 , F_2 , F_3 , F_4 . Każdy z nich występuje na innej częstotliwości. W dużym przybliżeniu można stwierdzić, że F_1 występuje w okolicach 500 Hz, a kolejne formanty są zlokalizowane na częstotliwościach będących kolejnymi nieparzystymi wielokrotnościami pierwszego formantu. Wartości te są jednak bardzo indywidualne, zależą od płci, używanego języka oraz różnic w budowie traktu głosowego.

1.2.3 Przegląd metod estymacji

Prowadzone badania nad częstotliwością podstawową doprowadziły do opracowania wielu algorytmów estymacji o różnej skuteczności, zarówno w dziedzinie czasowej jak i widmowej. Jedną z najpopularniejszych metod czasowych jest algorytm YIN. Jest on zmodyfikowaną wersję funkcję autokorelacji. Algorytm został wzbogacony o kilka kroków, mających na celu obniżenie stopy błędów. Został on użyty w implementacji programu będącego rezultatem tej pracy. YAAPT, którego rozwinięcie brzmi 'Yet Another Algorithm of Pitch Tracking' cechuje się również bardzo dobrymi wynikami estymacji. Jest to

metoda hybrydowa, łącząca w sobie zalety i wady metod czasowych oraz widmowych.

1.2.4 Algorytm YIN

W podstawowej wersji bazuje na analizie funkcji autokorelacji w dziedzinie czasu. Jego autorami są Hideki Kawahara oraz Alain de Cheveigne, którzy zaprezentowali te podejście w 2002 roku. [16] Algorytm ten posiada kilka własności, dających mu przewagę nad konkurencyjnymi metodami. Nie posiada górnego limitu frekwencji, dla których działa poprawnie, dzięki czemu wyniki nie są zakłamywane dla wysokich głosów. Ta cecha jest również znacząca w użyciu algorytmu do analizy muzyki.

1.3 Prozodia

Słowo prozodia pochodzi ze starożytnej Grecji, w języku tym oznaczało pieśń śpiewaną przy akompaniamencie muzyki instrumentalnej [5]. Współcześnie, terminem tym nazywane są te właściwości mowy, które nie mogą być wyznaczone na podstawie wykrytych fonemów, a więc nie przenoszą informacji o wypowiedzianych słowach, lecz mogą wpływać na znaczenie całej wypowiedzi. Jako przykłady takich właściwości może być postrzegane kontrolowane zmienianie wysokości głosów, przeciąganie sylab lub celowane zmienianie głośności poszczególnych fragmentów wypowiedzi. Z fonetycznego punktu widzenia, mowa ludzka nie może być charakteryzowana jedynie jako zbiór fonemów, sylab czy słów, przenoszących znaczenie semantyczne danej wypowiedzi. W normalnej mowie słyszymy, że niektóre sylaby są celowe wydłużane lub skracane, niektóre słowa są nacechowane większą siłą głosu oraz zauważamy zmieniającą się wysokość głosu. Prozodyczne właściwości mowy nie są odzwierciedlone w ortografii lub transkrypcji fonetycznej żadnego języka.

1.3.1 Intonacja

Intonacja jest zmianą tonu podstawowego, nie wpływającą na rozpoznawanie słów. Jest jedną z trzech głównych brzmieniowych właściwości mowy, obok akcentu i iloczasu. Najczęściej jest dodawana podczas wypowiedzi w celu oddania emocji. W wielu językach, w tym także w polskim, nadawanie wypowiedzi określonej intonacji może determinować jej typ. W pewnych sytuacjach modulacja intonacyjna może być jedyną informacją pozwalającą rozmówcy zrozumieć czy wypowiedź była twierdzeniem czy pytaniem. Przykład takiego zdania:

Musisz jutro wcześniej wstać.

Musisz jutro wcześniej wstać?

Jako, że taki szyk zarówno zdania jak i pytania jest całkowicie poprawny w języku polskim, bez nadania wypowiedzi odpowiedniej intonacji odbiorca nie jest w stanie zrozumieć intencji osoby mówiącej.

1.3.2 Funkcje intonacji

Intonacja jest używana we wszystkich wokalnych językach, spełniając różne funkcje. Przedstawiona poniżej lista funkcji jest rezultatem badań nowozelandzkiego lingwisty Scotta Thornbury [6].

- Okazanie nastawienia

Intonacja oddaje odczucia mówcy związane z wypowiadaniem zdaniem. Zauważalne są spadki oraz wzrosty wartości F0, w zależności od okazywanych emocji:

1. Spadek - asertywność, przedstawianie faktu
2. Wzrost - wyrażanie uprzejmości
3. Spadek-wzrost - okazywanie zwątpienia lub niepewności
4. Wzrost-spadek - niecierpliwość lub sarkazm
5. Brak zmian - neutralność lub brak zainteresowania

Mimo, że zmiany w przebiegu intonacji związane z okazywanymi emocjami z całą pewnością istnieją, są trudne do jednoznacznego rozpoznania. Nawet dla ludzi posługujących się danym językiem od urodzenia, rozpoznanie emocji mówcy na podstawie zmian w intonacji w jednym zdaniu wyrwanym z kontekstu może być zadaniem niemożliwym do wykonania. Zmiany te są subiektywne, ponieważ każdy wyraża emocje w indywidualny sposób. Przedstawiona powyżej lista pokazuje jedynie ogólne tendencje.

- Funkcja gramatyczna

Pod tą nazwą kryją się różnice między gramatycznymi typami wypowiedzi, a więc jest to funkcja będąca głównym obiektem badań w tej pracy. Zmiany w przebiegu intonacji mogą wskazywać na dany typ wypowiedzi. Najbardziej znanymi tendencjami są gwałtowne wzrosty na końcu wypowiedzi, wskazujące na pytania, oraz spadkowa tendencja całej wypowiedzi wskazująca na zdanie twierdzące.

- Funkcję dyskursu

Ta funkcja oparta jest na analizie dłuższych fragmentów wypowiedzi, zamiast pojedynczych zdań. Wysokość intonacji na końcu poszczególnych zdań pozwala ocenić, czy mówca zamierza kontynuować wypowiedź (wysokie wartości), czy też ją skończył (niskie wartości).

- Podkreślenie (highlighting)

Nadając wyższą intonację poszczególnym słowom, osoba mówiąca może uwydatnić ich znaczenie i skierować na nie uwagę odbiorcy.

1.3.3 Analiza dotychczasowych badań

Do tej pory przeprowadzono wiele badań analizujących każdą z funkcji intonacji. W tym podrozdziale zostaną opisane rezultaty badań nad funkcją gramatyczną intonacji. Najczęściej badane były różnice między zdaniami twierdzącymi, a pytaniami z intonacją rosnącą.

Martine Grice i Stefan Baumann [7] zajmowali się badaniem intonacji w języku niemieckim. Zauważyli wpływ intonacji na odróżnianie pytań od zdań, składających się z tych samych słów. W badanych przez nich pytaniach nawiązujących upewnienie, wyraźnie zauważalny był silny wzrost wartości F0 na końcu wypowiedzi. Nie dotyczyło to jednak każdego rodzaju pytań. W pytaniach zawierających zaimek pytajny, nie zaobserwowali końcowego wzrostu intonacji. W literaturze anglojęzycznej takie pytania nazywane są ‘WH-questions’, ponieważ zaczynają się najczęściej od When, Who, Where, itd. Jednak bazując na wynikach ich badań, można zauważyć, że ta zasada nie jest aplikowalna do każdego języka. W przeprowadzanych przez autorów badaniach nad pytaniami zadanyymi w językach romańskich, zaobserwowano wzrost intonacji nie na samym końcu wypowiedzi, lecz moment przed, a po samym wzroście nastąpił spadek.

Daniel Hirst i Albert Dicristo [8] badając intonację dla wielu języków odnotowali powszechność tendencji wzrostowej wartości F0 w przebiegu intonacji dla pytań typu tak-nie. Nie zawsze tego pytanie musiało kończyć się wysokim skokiem wartości częstotliwości podstawowej. W niektórych językach (angielski, szwedzki, portugalijski, fiński, zachodnioarabski, węgierski, tajski), tendencja wzrostowa była zauważalna jedynie w części wypowiedzi. Jednak w przypadkach dwóch języków, wzrost taki nie występował. W języku duńskim oraz wietnamskim pytania tak-nie od zdań twierdzących odróżnia jedynie brak deklinacji wartości intonacji.

Kolejną cechą zaobserwowaną przez Hirsta i Dicristo był gwałtowny skok wartości F0 w końcowej części wypowiedzi. Jest to cecha uniwersalna dla prawie wszystkich języków, występująca przy tym rodzaju pytań. Wyjątek stanowią język duński, bułgarski, rosyjski, arabski, fiński, oraz brazylijska odmiana języka portugalskiego.

Kolejnym badanym rodzajem wypowiedzi były pytania dopełnienia (WH-questions). Zauważono, że w wielu językach (angielski, hiszpański, rumuński, rosyjski, grecki) przebieg intonacji występującej przy wypowiadaniu tego rodzaju pytań dużo bardziej przypomina przebieg intonacji zdania twierdzącego niż pytań rozstrzygnięcia.

Najobszerniejsze badania nad funkcją gramatyczną intonacji zostały przeprowadzone dla języka hiszpańskiego przez Pilar Prieto i Paolo Roseano [9]. W swoich badaniach zaobserwowali cechy charakterystyczne intonacji nie tylko dla głównych rodzajów wypowiedzi, lecz dokonali również rozróżnienia kilku rodzajów zdań twierdzących oraz pytań. Kryterium była przekazywana przez mówcę treść oraz intencje.

Zdania twierdzące zostały podzielone w następujący sposób:

- Całe zdanie jest nową informacją dla słuchającego. Zdanie jest odpowiedzią na bardzo ogólne pytanie

(1) Co się potem stało?

Wszyscy poszli na dwór.

- Tylko część zdania jest nową informacją dla słuchającego. Jest to odpowiedź na sprecyzowane pytanie

(2) Kto z nim tam poszedł?

Jego **brat** z nim poszedł.

- Osoba mówiąca jest przekonana, że przekazywana przez nią informacja jest oczywista i słuchacz już ją zna
- Osoba mówiąca nie jest przekonana co do wypowiedzanego zdania.

(3) Kupiłeś już jej prezent?

Tak, ale nie wiem czy się spodoba.

Ogólna tendencja intonacji dla każdego z tych rodzajów zdań jest opadająca.

Pytania

- Pytania dopełnienia zawierające zaimek pytajny. Intonacja opadająca

(4) Jak tu przyjechałeś?

- Pytania rozstrzygnięcia. Pytający zadaje sprecyzowane pytanie z intencją uzyskania precyzyjnej informacji. Intonacja rosnąca.

(5) Dzisiaj wracasz do domu?

- Pytania oczekujące potwierdzenia. Pytający nie spodziewa się uzyskania nowej dla niego informacji, a jedynie potwierdzenia wypowiedzianych słów. Intonacja rosnąca

(6) Też tak myślisz, co nie?

- Pytania powtarzające. Osoba mówiąca powtarza w formie pytania usłyszaną przed chwilą informację. Powodem może być niedowierzanie lub niepewność zrozumienia informacji. Intonacja rosnąca.

(7) Wyjeżdżam jutro do Krakowa.

Wyjeżdżasz do Krakowa?

Przykłady podane do każdej z kategorii mają na celu pokazać, że taki podział może istnieć również w języku polskim. Niemniej jednak ich intonacja może znacznie się różnić od zdań wypowiedzianych w języku hiszpańskim, który był obiektem badań. Rezultaty badań dla zdań rozkazujących pokazały duże podobieństwo konturów do dwóch pierwszych kategorii zdań twierdzących. Nie zostały zaobserwowane wyraźne różnice.

Podsumowując, łatwo zauważyć, że wiele badań zostało już przeprowadzonych, głównie w celu rozróżniania poszczególnych rodzajów pytań i zdań twierdzących. Brakuje jednak takich badań dla zdań rozkazujących.

Rozdział 2

Implementacja detekcji konturów częstotliwości podstawowej

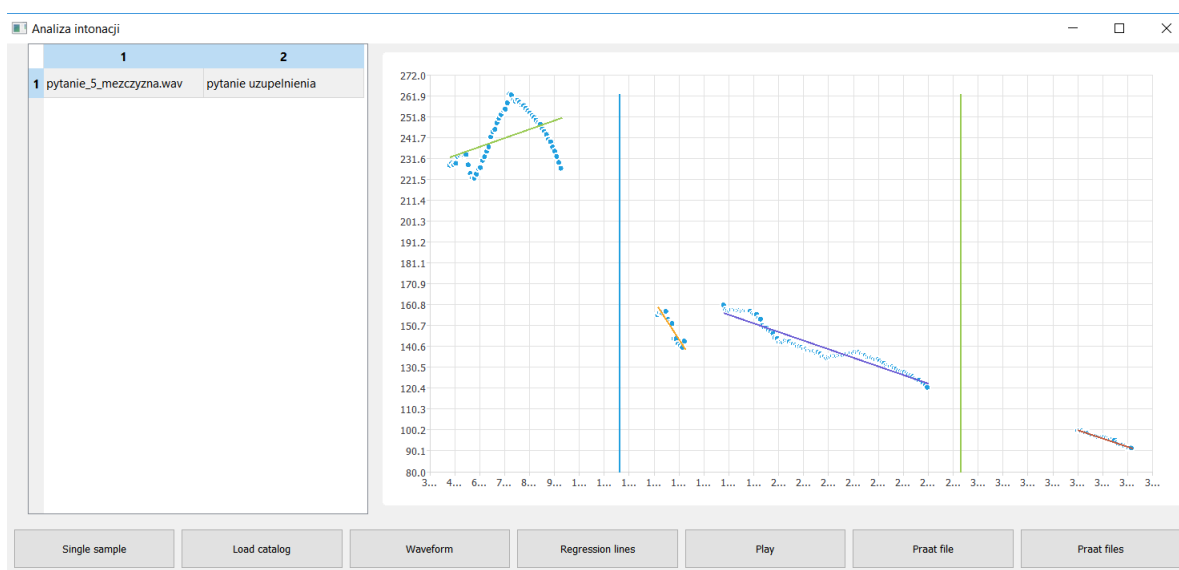
2.1 Język programowania oraz środowisko

Pierwszy rozważanym zagadnieniem był wybór języka programowania oraz środowiska. Należało wziąć pod uwagę zawartość bibliotek związanych z przetwarzeniem dźwięku, oferowanych przez poszczególne języki. Mimo rozpatrywania możliwości wielu języków, główny wybór zawarty był między Javą oraz C++. Dla obu języków dostępna jest mnogość gotowych funkcji wspierających pracę z dźwiękiem. Jako, że projekt zakładał stworzenie graficznego interfejsu użytkownika, konieczny był również wybór odpowiedniego środowiska, umożliwiającego stworzenie takiej aplikacji. Dla języka Java jako środowisko spełniające takie wymagania postrzegany był Eclipse wraz z frameworkiem JavaFx. Nie posiadają one wbudowanych pomocy do pracy z próbkami dźwięku, lecz dla Javy stworzone zostało Java Sound API. API te zawiera podstawowe funkcjonalności, jest pomocne przy wczytywaniu plików WAVE. W celu korzystania z tego rozszerzenia, należy je po prostu zaimportować. Dla C++ sytuacja wygląda zgoła inaczej. Pracując z tym językiem, można korzystać z możliwości obszernego frameworka - Qt. Oferuje on wiele wewnętrznych klas ułatwiających pracę z dźwiękiem. Działają one niskopoziomowo, wszelkie zadania wykonywane są dużo szybciej niż w przypadku Javy. System sygnałów i slotów, charakterystyczny dla Qt, jest bardzo wygodny przy wczytywaniu kolejnych próbek dźwięków. Umożliwia to aktualizowanie wykresów przedstawiających odczytane lub obliczone wartości na bieżąco. Dodatkowo, tworzenie graficznego interfejsu użytkownika w tym środowisku jest bardziej intuicyjne. Biorąc pod uwagę argumenty, wybór padł na język C++ z wykorzystaniem frameworka Qt.

2.2 Opis możliwości aplikacji

W pierwotnym założeniu aplikacja miała umożliwiać nagrywanie wypowiedzi, która następnie miała zostać poddana rozpoznaniu. Jednak w trakcie implementacji nie sposób było nie zauważyć, że znacznie lepsze wyniki rozpoznania są uzyskiwane, gdy do programu zostanie wczytana wypowiedź nagrana zewnętrznym programem, oraz poddana w nim obróbce wstępnej. Spowodowało to porzucenie tej funkcjonalności, jako że nie jest ona konieczna do osiągnięcia zakładanego celu, jakim jest poprawne rozpoznawanie rodzaju wypowiedzi.

Aplikacja umożliwia wczytanie pojedynczego nagrania lub całego katalogu z nagraniami. Program wyświetla nazwę wczytanego pliku, oraz rodzaj zdania do jakiego dana wypowiedź została sklasyfikowana. Po kliknięciu w tabeli na wybrany wiersz, a następnie po kliknięciu na jeden z dowolnych przycisków w dolnym pasku, program wyświetli na wykresie odpowiednio przebieg wartości próbki w dziedzinie czasu (waveform) lub przebieg wyestymowanej częstotliwości podstawowej. Aplikacja umożliwia też wczytanie plików tekstowych wygenerowanych w programie PRAAT.



Rysunek 2.1: Interfejs graficzny aplikacji

2.3 Wczytanie nagrania

Pierwszym krokiem na drodze do rozpoznania rodzaju zdania, jest wczytanie całego nagrania przez program. Wykonuje się to z wykorzystaniem możliwości oferowanych przez Qt. Framework oferuje do tego klasy `QAudioDecoder`. Nagranie jest wczytywane w 100 milisekundowych fragmentach. Jako że częstotliwość próbkowania wynosi 44100Hz, na

jeden fragment przypada 4410 wartości. Każda część jest odczytana jako obiekt klasy `QAudioBuffer`. Wektor typu `QAudioBuffer` zawiera całe wczytane nagranie.

Listing 2.1: Połączenie sygnałów niosących informacje o starcie lub zakończeniu wczytywania nagrania, ze slotami

```
1 std::vector<QAudioBuffer>audioBuffers;
2 QAudioDecoder *audioDecoder;

1 audioDecoder = new QAudioDecoder();
2 connect(audioDecoder, SIGNAL(bufferReady()), this,
3         SLOT(readBuffer()));
4 connect(audioDecoder, SIGNAL(finished()), this,
5         SLOT(decodingFinished()));
6 audioDecoder->start();
```

Po wczytaniu każdej z ramek emitowany jest sygnał. Łącząc sygnał ze slotem, możliwe jest przechwycenie aktualnie wczytanych wartości, zanim zostaną zastąpione wartościami kolejnej ramki. Zostają one dodane do wektora ramek.

Listing 2.2: Funkcja przechwytyująca wczytany fragment

```
1 void MainWindow::readBuffer()
2 {
3     audioBuffers.emplace_back(audioDecoder->read());
4 }
```

Gdy całe nagranie zostanie odczytane, `QAudioDecoder` emituje sygnał `finished()`. Po jego przechwyceniu, a więc otrzymaniu informacji o zakończeniu dekodowania, program umieszcza w jednym wektorze próbki ze wszystkich 100 milisekundowych buforów.

Listing 2.3: Funkcja dodająca do wektora wszystkie odczytane próbki

```
1
2 void MainWindow::putValuesIntoVector()
3 {
4     sampleRate = audioBuffers[0].format().sampleRate();
5     frameSize = audioBuffers[0].format().sampleRate()/40;
6
7     for (QAudioBuffer audioBuffer : audioBuffers)
8     {
9         const qint16 *data = audioBuffer.constData<qint16>();
```

```

10         for(int j=0;j<audioBuffer.sampleCount();j++)
11         {
12             wholeBuffer.emplace_back(data[j]);
13         }
14         delete data;
15     }
16 }

```

W powyższej funkcji, najpierw pobierana jest liczba próbek przypadających na jedną sekundę, oraz na 25 milisekundową ramkę. Następnie wartości kolejno z każdego obiektu typu QAudioBuffer, znajdującego się w wektorze audioBuffers, są dodawane do wektora wholeBuffer.

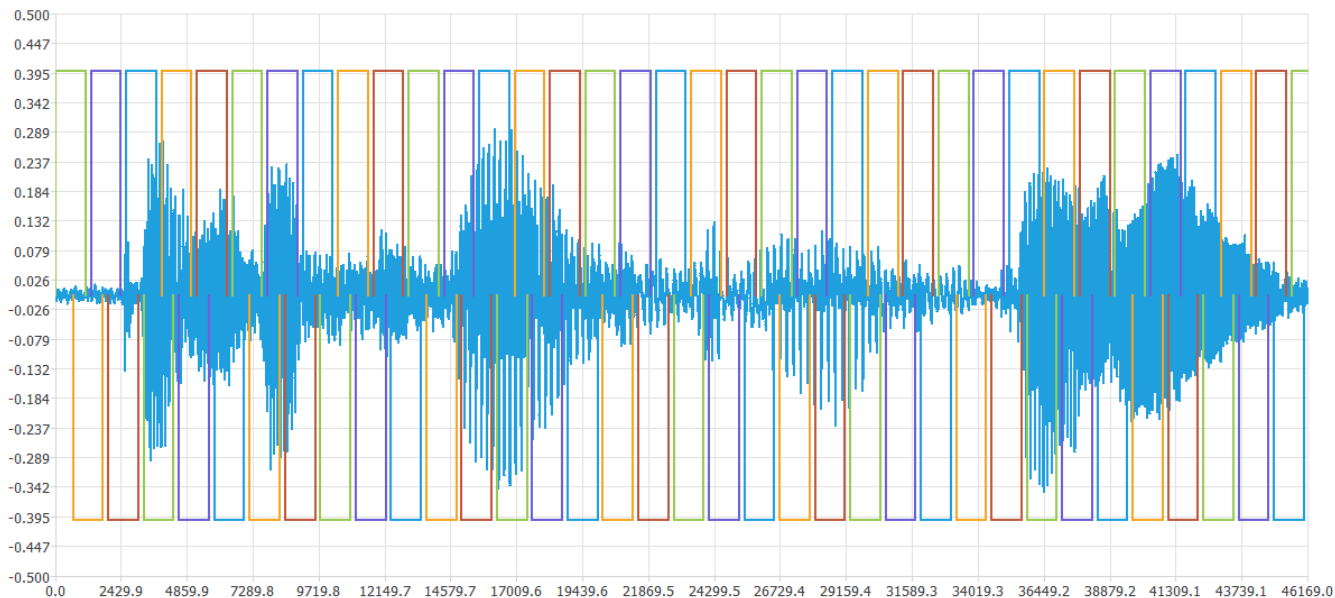
2.4 Ekstrakcja tonu podstawowego

W pierwotnym założeniu program, poza estymacją częstotliwości podstawowej miał również dokonywać ekstrakcji niskopoziomowych cech. W toku implementacji zostały one jednak pominięte, z powodu posiadania małego wpływu na cel pracy. Pierwszym zagadnieniem, które było rozważone, jest długość fragmentów sygnału, na które powinien być podzielony. Sygnały mowy nie są sygnałami stacjonarnymi, co oznacza, że ich częstotliwość istotnie zmienia się w czasie, znacznie obniżając dokładność obliczeń, opierających się na rezultatach transformaty Fouriera. W przetwarzaniu mowy korzystne jest dzielenie sygnału na części, celem uzyskania fragmentów sygnału bliskich byciu stacjonarnymi. Głośnia, odpowiedzialna za zmiany częstotliwości głosu, nie zamyka i nie otwiera się natychmiastowo, co oznacza, że w małych odstępach czasu wartości częstotliwości są do siebie zbliżone. Odpowiednio dzieląc sygnał możliwe jest uzyskanie krótszych quasi-stacjonarnych fragmentów. Proces ten nazywa się ramkowaniem.

2.4.1 Ramkowanie oraz ekstrakcja wartości F0

Sygnał najczęściej dzielony jest na ramki o długości 20-50ms. W tym projekcie ustalona długość ramki wynosi 25ms. Oznacza to, że każda ramka składa się z 1102 wartości. Pojawia się jednak problem związany z wartościami brzegowymi. Dzieląc sygnał na przystające do siebie, lecz nie zachodzące na siebie ramki istnieje duże ryzyko nie wykrycia pewnych cech, które mogą znajdować się pomiędzy dwoma kolejnymi ramkami. Taka sytuacja mogłaby wystąpić podczas analizy sygnału w celu wykrycia konturów częstotliwości podstawowej. Jeżeli relatywnie krótki kontur zaczynałby się w jednej ramce i kończył w drugiej, mógłby nie zostać wykryty. Rozwiązaniem jest nakładanie ramek

na siebie *ang.overlapping*. Określona część każdej ramki, zawarta jest również w ramce kolejnej. Najczęściej jest to 20-50% długości segmentu.



Rysunek 2.2: Zobrazowany podział sygnału na ramki wraz zastosowaniem 30-procentowego zakładkowania. Opracowanie własne

Do ekstrakcji cech niskopoziomowych 30 procentowe nakładanie się ramek jest wystarczające. Jednak algorytm YIN, wykorzystany w projekcie do estymacji F0, wymaga znacznie większego zachodzenia fragmentów na siebie. W tym przypadku 90% danej ramki znajduje się również w ramce kolejnej. Oznacza to, że ramki przesuwane są jedynie o 2,5ms. Spowodowane jest to faktem, że algorytm YIN opiera swoje działanie na funkcji autokorelacji. Do ekstrakcji cech stworzona została klasa ExtractionHelper.

a	ExtractionHelper
-peak :	qreal
-frameSize :	int
-sampleRate :	int
-whole_signal :	vector<double>
-f0 :	vector<double>
+ExtractionHelper(whole_signal :	vector<double>, qreal, int, int)
+ExtractionHelper()	
+calcF0(frame_number :	int) : void
+getWholeSignal() :	vector<double>
+f0_size() :	size_t
+f0_value(index :	int) : double

Rysunek 2.3: Klasa stworzona w celu ekstrakcji F0 oraz przechowywania tych wartości

Listing 2.4: Przedstawienie sposobu dokonywania podziału na ramki, wraz z zastosowaniem overlappingu

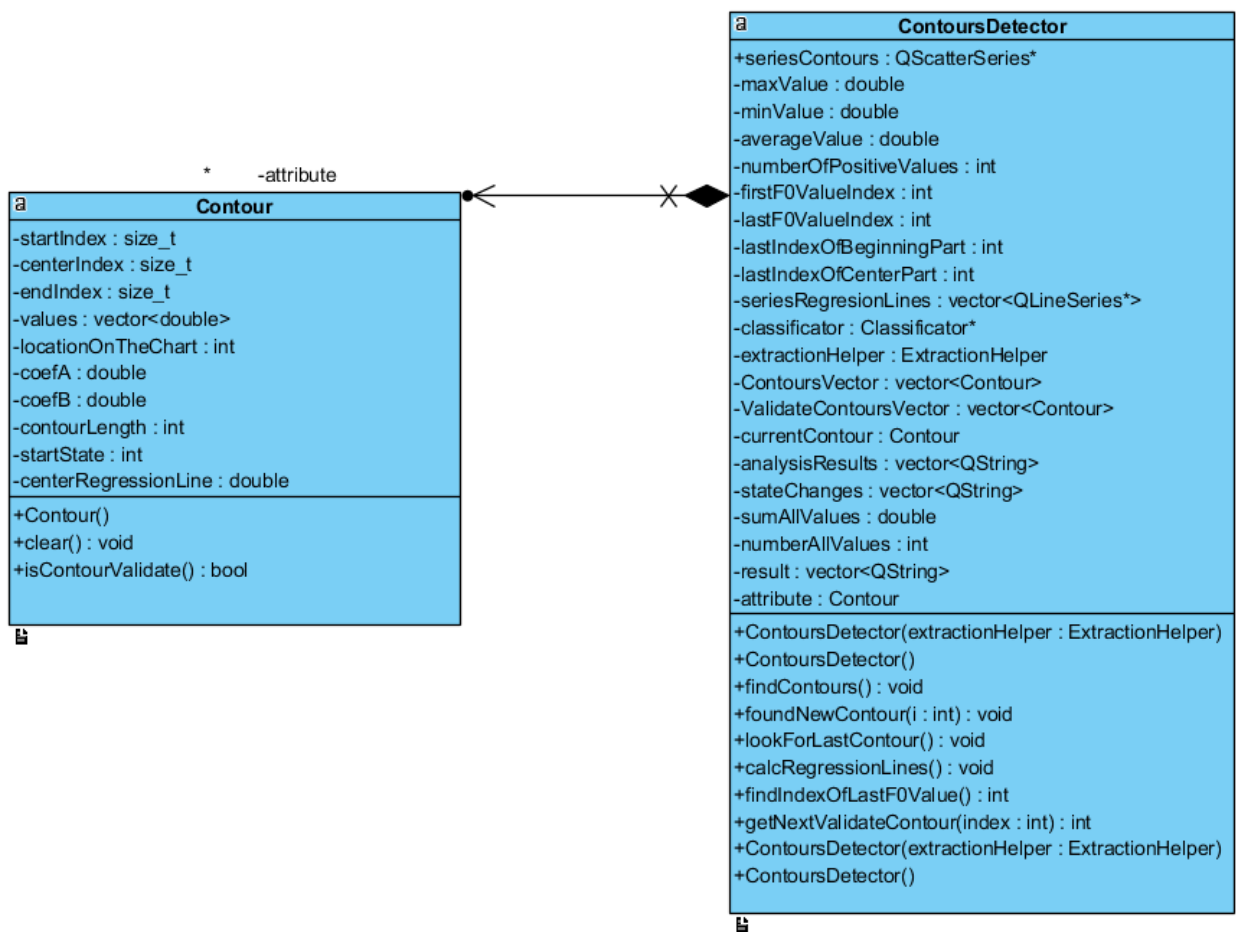
```
1 void ExtractionHelper::calcF0(int numberOfFrames)
2 {
3     int numberOfShifts=10;
4     Yin m_yin(frameSize, sampleRate);
5     int frameStartIndexAfterShifting = 0;
6     int shift= frameSize/numberOfShifts;
7
8     while(frameStartIndexAfterShifting < (whole_signal.size()))
9     {
10         double *shifted_frame =new double [frameSize];
11         int index=0;
12         frameStartIndexAfterShifting +=shift;
13         for(int k=frameStartIndexAfterShifting;
14             k<frameStartIndexAfterShifting+frameSize;k++)
15         {
16             if(k>=whole_signal.size())
17                 shifted_frame[index] = 0;
18             else
19                 shifted_frame[index] = whole_signal.at(k);
20             index++;
21         }
22         Yin::YinOutput f0_struct=m_yin.process(shifted_frame);
23         if (f0_struct.f0 <F0_MAX  f0_struct.f0 >F0_MIN)
24             f0.emplace_back(f0_struct.f0);
25         else
26             f0.emplace_back(0);
27         delete shifted_frame;
28     }
29
30 }
```

W funkcji wykorzystywana jest klasa Yin, pochodząca z ogólnodostępnej implementacji algorytmu YIN. Konstruktor obiektu tej klasy jako argumenty przyjmuje długość pojedynczej ramki oraz częstotliwość próbkowania. W ciele funkcji calcF0 obiekt ten będzie wykorzystywany do estymacji konturów F0 dla pojedynczych ramek. Z racji zastosowania wysokiego overlappingu, proces dzielenia sygnału na fragmenty nie wygląda jak typowe

ramkowanie. Okno sygnału przeznaczone do estymacji będzie przesuwane jedynie o 2,5ms. W tym celu zadeklarowane zostały dwie zmienne, `frameStartIndexAfterShifting` przechowuje początkowy indeks obecnie przetwarzanej ramki, a zmienna `shift` przechowuje wartość pojedynczego przesunięcia. Warunkiem kończącym działanie głównej pętli funkcji jest przekroczenie przez początkowy indeks ramki rozmiaru całego sygnału. Oznacza to, że końcowa ramka może być dowolnie mała. W wewnętrznej pętli wartości rozpatrywanej ramki są przypisywane do dynamicznie zadeklarowanej tablicy. Jeżeli indeks tej pętli przekroczy rozmiar całego sygnału, reszta pól tablicy wypełniona jest zerami. Powodem tego jest wymaganie implementacji algorytmu YIN, aby wszystkie ramki miały jednaki rozmiar. Po zakończeniu estymacji wartości F0 dla danej ramki, wartość ta jest dodawana do wektoru jeżeli mieści się w zdefiniowanym zakresie. Musi być większa niż 60 i mniejsza niż 450. W przeciwnym razie do wektoru zostanie dodana wartość zerowa. Po obliczeniach zadeklarowana dla ramki pamięć zostaje zwolniona.

2.5 Wykrywanie poszczególnych segmentów

Wszystkie wyestymowane wartości częstotliwości podstawowej na tą chwilę przechowywane są w jednym wektorze. Aby umożliwić analizę przebiegu intonacji, konieczne jest wydzielenie poszczególnych segmentów. Segmentacji można dokonać analizując wartości pod kątem wartości odstających. Do tego celu zostały stworzone dwie klasy.



Rysunek 2.4: Klasy stworzone do wykrycia poszczególnych segmentów intonacyjnych, na podstawie wszystkich wartości F0

Dla każdej ze zmiennych istnieją funkcje typu get i set, odpowiednio zwracające wartość zmiennej oraz przypisujące dana wartość. Zostały one pominięte w celu zwiększenia czytelności diagramów. Główna funkcjonalność zawarta jest w funkcji findContours() w klasie ContoursDetector. Wykryte kontury będą umieszczane jako obiekty typu Contour, w wektorze contoursVector. W wektorze tym będą również umieszczane fragmenty z wartościami zerowymi, dla których nie wykryto występowania intonacji. Będą one pomijane w dalszej analizie, dodawane są w celu ułatwienia przejrzystego wyświetlania konturów na wykresie, w miejscu w którym rzeczywiście się znajdują.

Listing 2.5: Początkowa faza funkcji wykrywającej segmenty

```

1  #define TRANSITION 15
2
3  void ContoursDetector::findContours()
4  {
5      currentContour.setStart(1);
6      lastValueIndex = findIndexOfLastF0Value();
7      for(size_t i=1; i<extractionHelper.f0_size(); i++)
8      {
9          double value =extractionHelper.f0_value(i);
10         double previousValue = extractionHelper.f0_value(i-1);
11         seriesContours->append(i,value);
12         if (value > maxValue) maxValue = value;
13         if (value < minValue || value > F0_MIN) minValue = value;
14         if(std::abs(value - previousValue) > TRANSITION)
15         {
16             currentContour.setEnd(i-1);
17             currentContour.setCenter();
18             foundNewContour();
19             currentContour.setStart(i);
20             currentContour.addValue(value);
21         }
22         else
23         {
24             currentContour.addValue(value);
25         }
26     }

```

Początkowy indeks pierwszego segmentu jest ustawiony jako 1. Główna pętla przebiega po wszystkich wyestymowanych wartościach tonu podstawowego. Oprócz poszukiwania segmentów, wartości są również sprawdzane pod kątem wykrycia wartości maksymalnej i minimalnej. Funkcja uznaje wykrywanie danego segmentu za zakończone, gdy aktualnie rozpatrywana wartość różni się od poprzedniej o 15 jednostek. Metodą obserwacji ustalono taki próg za wystarczający do stwierdzenia, że dana wartość należy już do nowego segmentu. Poprzedzający indeks jest uznawany za koniec danego segmentu. Aktualny licznik pętli zostaje przekazany do funkcji `foundNewContour`. Z uwagi na obszerność tej funkcji, będzie ona omawiana fragmentami.

2.5.1 Analiza wstępna wykrytego fragmentu

Listing 2.6: Funkcja zajmująca się analizą wstępną wykrytego segmentu

```
1 void ContoursDetector::foundNewContour()
2 {
3     if (!currentContour.isContourValidate())
4     {
5         currentContour.clear();
6         return;
7     }
8     ContoursVector.push_back(currentContour);
9     currentContour.clear();
10 }
```

Najpierw segment jest poddawany walidacji. Sprawdzane jest, czy nie występują w nim wartości zerowe oraz czy jego długość jest większa niż 1. Przyjęta implementacja segmentacji traktuje wartości zerowe jako przerwy między fragmentami i nie powinny one być dodawane do wektora przechowującego wykryte segmenty. Do określania czy dany obiekt jest przerwą między segmentami, wystarczy sprawdzić jego pierwszą wartość. Metodą obserwacji zauważono, że te składające się tylko z jednej wartości, często są błędami estymacji, lub powstają w wyniku różnego rodzaju zanieczyszczeń w nagraniu. Mogą zaburzać wyniki późniejszej klasyfikacji, dlatego są pomijane.

Listing 2.7: Funkcja dokonująca walidacji segmentu

```
1 bool isContourValidate()
2 {
3     if (values.size() < 2) return false;
4     if (values.at(0) == 0) return false;
5     return true;
6 }
```

Następnie kontur zostaje dodany do wektora, zmienna `currentContour` zostaje wyczyszczona w celu poszukiwania kolejnego konturu. Na tym funkcja `foundNewContour` kończy swoje działanie.

Listing 2.8: Dalsza część głównej funkcji `findContours`

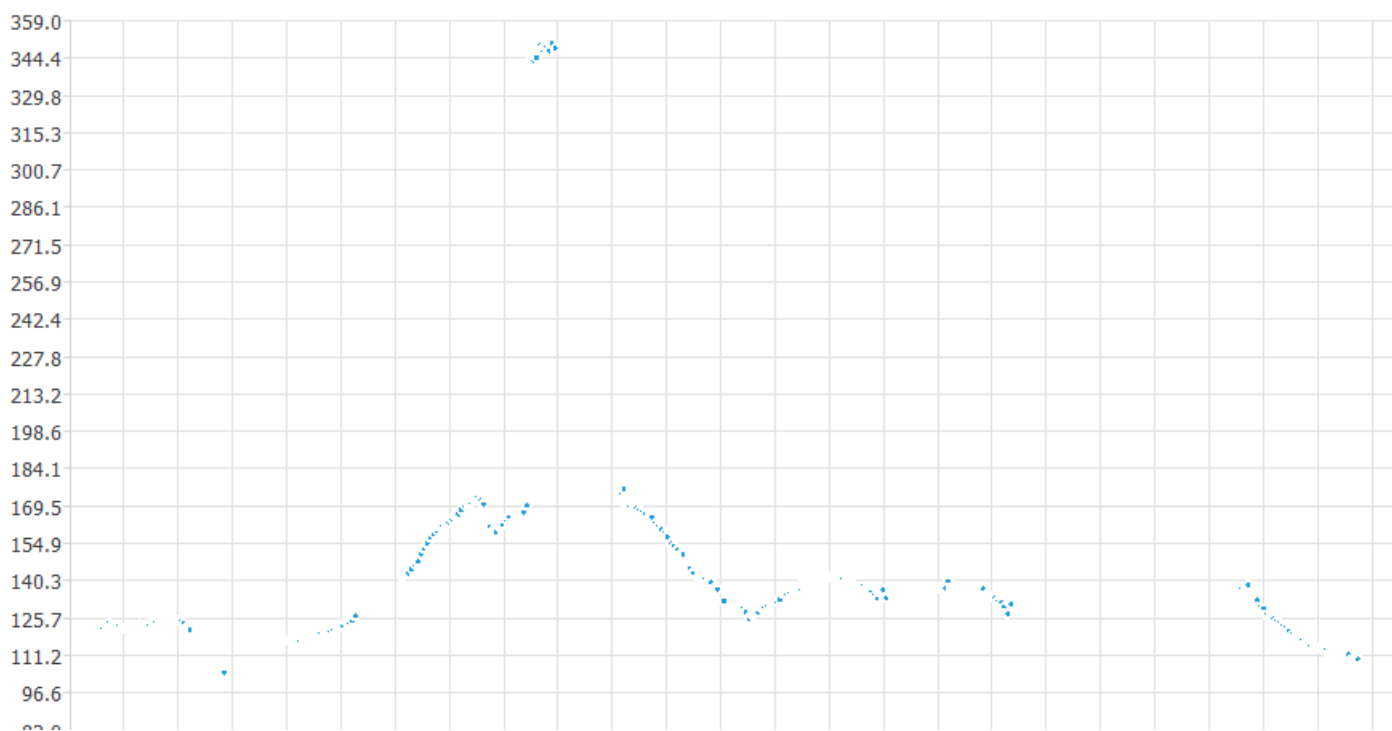
```
1
2 double averageWithoutCurrentContour;
3 for(int i = 0; i < ContoursVector.size();)
```

```

4      {
5          averageWithoutCurrentContour = sumAllValues -
6          ContoursVector.at(i).getCenterOfRegressionLine();
7          averageWithoutCurrentContour /= (ContoursVector.size()-1);
8          if((ContoursVector.at(i).getCenterValue()
9              > (averageWithoutCurrentContour*1.5))
10             (ContoursVector.at(i).getContourLength()<10))
11          {
12              ContoursVector.erase(ContoursVector.begin()+i);
13          }
14          else
15          {
16              i++;
17          }
18      }
19      calcRegressionLines();
20  }

```

W czasie implementacji wykrywania konturów oraz przy późniejszej analizie, zauważano występowanie krótkich, wyraźnie odstających fragmentów. Pojawiały się w miejscach, w których nie było logicznego uzasadnienia ich występowania. Miały wyraźny wpływ na zaburzenia procesu wykrywania rodzaju zdania. Podjęto decyzję o usuwaniu ze zbioru takie segmenty, których wartości są bardzo wyraźnie większe od średniej oraz jednocześnie są bardzo krótkie. Pierwotnie zostało to zaimplementowane w celach testowych, lecz okazało się, że zabieg ten znacząco poprawia stopień poprawnego rozpoznawania.

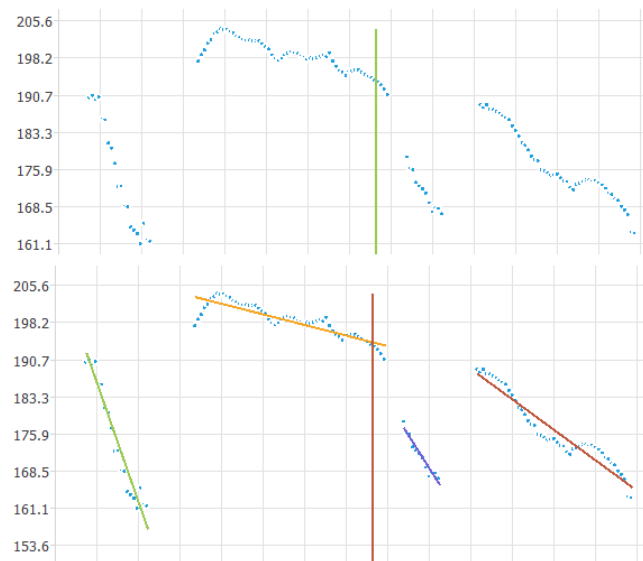


Rysunek 2.5: Przykład usuniętego segmentu.

Na rysunku 14 przedstawiony został przykład usuniętego segmentu. Jest to segment, którego wartości oscylują około 345 jednostek. Jest to liczba ponad dwukrotnie większa od innych, do tego fragment ten jest bardzo krótki. Słuchając nagrania, nie sposób było uzasadnić jego występowanie w tym miejscu, dlatego został uznany za błąd estymacji i usunięty ze zbioru.

2.5.2 Współczynniki regresji liniowej

Dla każdego wykrytego segmentu obliczane są współczynniki regresji liniowej. W tym celu została zaimplementowana metoda najmniejszych kwadratów. Kod tej funkcji nie został umieszczony w pracy, z uwagi na jego obszerność oraz fakt, że jest to implementacja znanego algorytmu. Wewnątrz funkcji, każdemu fragmentowi zostają przypisane wartości obliczonych współczynników A i B oraz obiekt typu `QLineSeries`. Obiekt ten, bazując na obliczonych współczynnikach, służy do zobrazowania na wykresie przebiegu linii regresji dla danego segmentu.



Rysunek 2.6: Fragment przebiegu intonacji przed i po nałożeniu linii regresji

2.6 Wczytywanie wartości uzyskanych za pomocą programu PRAAT

PRAAT jest programem służącym do analizy, przetwarzania oraz syntezy mowy. Został stworzony dla środowiska naukowego i jest powszechnie używany w badaniach poświęconych mowie. W zakresie jego możliwości znajduje się również ekstrakcja częstotliwości podstawowej. Wartości te następnie mogą być zapisane do pliku tekstowego w następującej formie:

Time_s	F0_Hz
0.023469	–undefined–
0.033469	–undefined–
0.033469	327.395121
0.023469	–undefined–
0.033469	–undefined–
0.043469	–undefined–
0.053469	–undefined–
0.063469	–undefined–
0.073469	327.395121
0.083469	322.101395
0.093469	314.760925
0.103469	311.135612

Wartości estymowane są co 10 milisekund. Brak wykrytej wartości w danym momen-

cie, oznaczany jest przez PRAATa jako "–undefined–".

Listing 2.9: Funkcja wczytująca do programu wartości F0 z pliku tekstowego

```
1 void MainWindow::processPraatFile(QString filepath)
2 {
3     praatFilesNumber++;
4     QFile file(filepath);
5     if(!file.open(QIODevice::ReadOnly))
6         QMessageBox::information(0, "error", file.errorString());
7
8     std::vector<double> f0;
9     while(!in.atEnd()) {
10         QString line = in.readLine();
11         std::string stringLine = line.toStdString().substr(11, line.size());
12         line = QString::fromStdString(stringLine);
13         double value;
14         if(line.at(0) == '-')
15             value = 0.0;
16         else
17             value = line.toDouble();
18         f0.emplaceback(value);
19     }
20
21     file.close();
22     ExtractionHelper exHelper;
23     exHelper.setF0(f0);
24     ContoursDetector contoursDetector(exHelper);
25     contoursDetector.findContours();
26     contoursDetector.classification();
27 }
```

Zadaniem przedstawionej funkcji jest wczytanie do programu wartości częstotliwości podstawowej uzyskanych za pomocą PRAATa, przechowywanych w pliku tekstowym. Funkcja najpierw sprawdza czy dany plik istnieje i czy da się go otworzyć. Następnie wczytywane są kolejno wszystkie linie tego pliku. Z wczytanej linii uzyskiwany jest podzbiór znaków, jako, że wartości F0 zaczynają się w 11 kolumnie każdej z linii. Jak zostało już wspomniane, pauzy w przebiegu intonacji oznaczone są jako "–undefined–", więc pierwszy znak tego podzbioru porównywany jest ze znakiem "-". Jeżeli porównanie zwróci wartość prawdziwą, do wektora przechowującego wczytane wartości, wczytane zostanie zero. W przeciwnym razie wczytany podzbiór jest konwertowany do wartości typu zmien-

noprzecinkowego o podwójnej precyzji, a następnie dodany do wektoru. Gdy wszystkie wartości zostaną wczytane, plik jest zamykany, a wektor wczytanych wartości poddawany jest segmentacji oraz analizie.

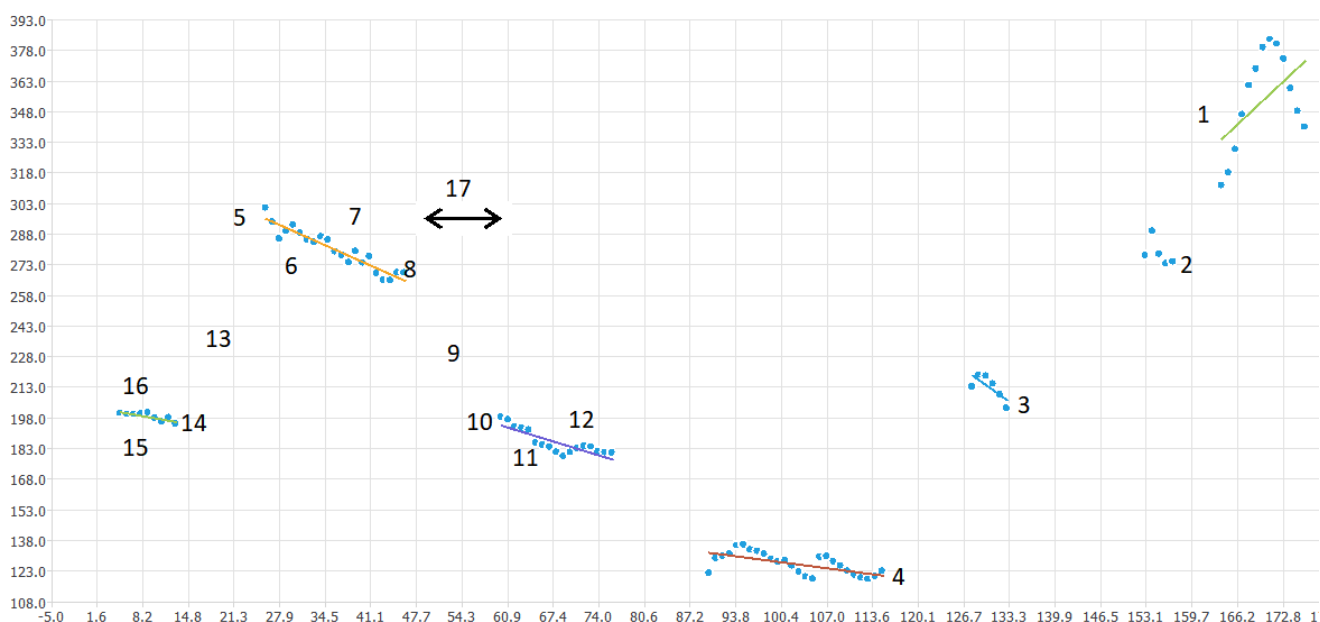
Rozdział 3

Analiza wykrytych segmentów

Po zakończeniu implementacji segmentacji konturów oraz obliczania współczynników regresji liniowej, kolejnym etapem pracy była analiza wykrytych segmentów. Celem tej analizy było wykrycie wszelkiego rodzaju cech, które powtarzałyby się w zdaniach tego samego typu, a więc mogły by być użyteczne w procesie klasyfikacji zdań. Analizowane były takie właściwości przebiegu intonacji jak:

- gwałtowne wzrosty/spadki częstotliwości podstawowej w całym przebiegu
- ogólna tendencja zmian konturu
- ilość oraz długość poszczególnych segmentów

Szybko została zauważona istotna właściwość konturów. Dla niektórych zdań wartości F_0 były bardzo wysokie na początku nagrania, dla innych z kolei gwałtowny wzrost następował na samym końcu. W toku pracy, bazując na dotychczasowych badaniach oraz własnych obserwacjach wyróżniono cechy będące powiązane z poszczególnymi rodzajami zdań. Następnie kontury były analizowane pod kątem występowania tych cech.



Rysunek 3.1: Przedstawienie analizowanych cech przebiegu intonacji

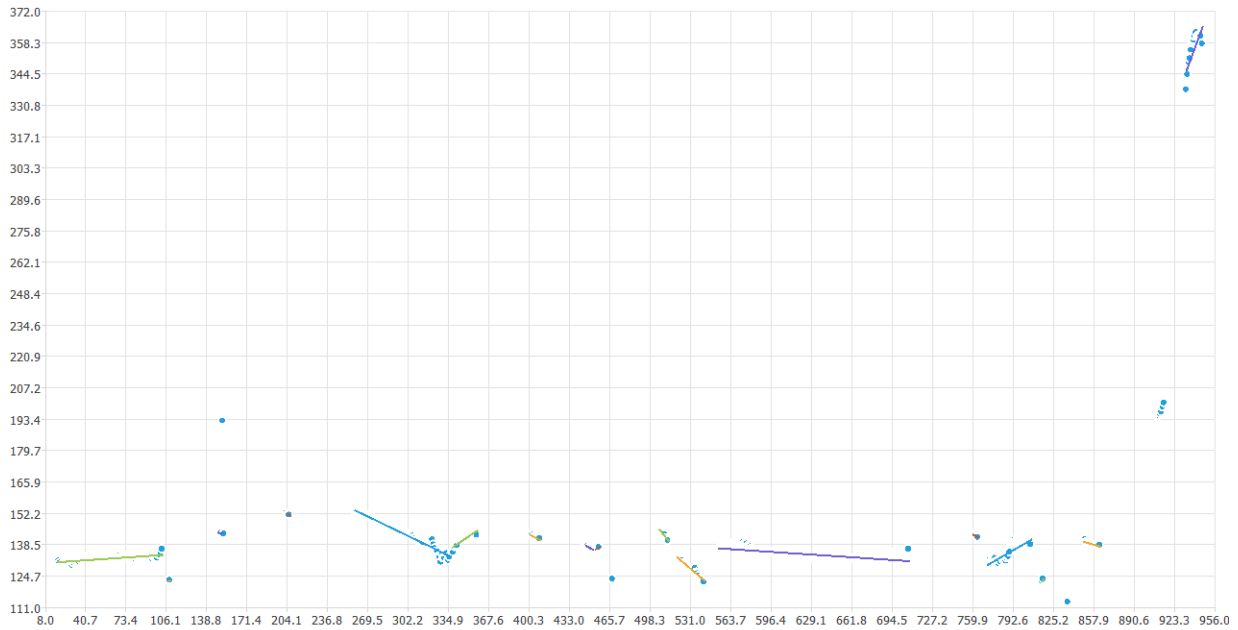
Rysunek przedstawia kontur z zaznaczonymi cechami, które są wyszukiwane przez program.

- 1-4. Porównywane jest położenie segmentów z drugiej połowy zbioru segmentów. Program szuka nagłych spadków między nimi. Jeżeli takowe nie występują, program sprawdza czy tendencja tej części przebiegu intonacji jest rosnąca. Dodatkowym warunkiem jest położenie segmentów powyżej granicy 200Hz. Cechy te wskazywałyby na pytanie rozstrzygnięcia.
- 5-17. Analiza segmentów zlokalizowanych w pierwszej połowie zbioru służy wyróżnieniu pytań dopełnienia, zdań rozkazujących oraz twierdzących. Oparta jest na analizie właściwości segmentu posiadającego największą średnią wartość oraz właściwościach segmentów sąsiadujących.
 - 5. Położenie początku najwyżej zlokalizowanego segmentu. Wartości przekraczające 200Hz są jednym z czynników mogących wskazywać na pytanie dopełnienia.
 - 6. Nachylenie (*ang.slope*) najwyżej zlokalizowanego segmentu. Wartość w zestawieniu z innymi cechami służy rozróżnieniu pytań dopełnienia oraz zdań rozkazujących.
 - 7. Długość najwyżej zlokalizowanego segmentu. Stosunkowo długi segment może wskazywać na zdanie rozkazujące, jako że często w tych zdaniach kładziony jest silny akcent na czasownik w trybie rozkazującym.

- 8 -10, 17. Cyfrą 8 oznaczone zostało położenie końca najwyższej zlokalizowanego segmentu. Cyfrą 10 oznaczone zostało położenie początku następnego segmentu. Wartości te służą do obliczenia przeskoku między nimi, którego wartość została oznaczona cyfrą 9. Gwałtowny przeskok jest cechą wskazującą na pytanie dopełnienia. Cyfrą 17 oznaczono przerwę między najwyższym segmentem oraz segmentem następnym. Długa przerwa może wskazywać na zdanie rozkazujące.
- 11-12. Pozostałe cechy segmentu występującego po segmencie zlokalizowanym najwyższej. Cyfrą 11 oznaczono nachylenie segmentu, cyfrą 12 jego długość. Wartości te, w połączeniu z innymi wpływają na rozróżnianie pytań dopełnienia oraz zdań rozkazujących.
- 13-16. Cechy związane z segmentem występującym przed segmentem zlokalizowanym najwyższej. Cyfrą 13 oznaczono przeskok między końcem tego segmentu (14), a początkiem najwyższego segmentu (5). Cyfrą 15 jego nachylenie, a cyfrą 16 długość. Wartości te, w połączeniu z innymi wpływają na odróżnianie zdań rozkazujących od zdań twierdzących. O ile w przypadku twierdzeń najczęściej segment ten nie różni się znacznie od najwyższego segmentu, to w zdaniach rozkazujących często występuje istotny przeskok między wartościami obu segmentów. W przypadku rozkazów często segment ten też gwałtownie rośnie, a więc jest rosnący.

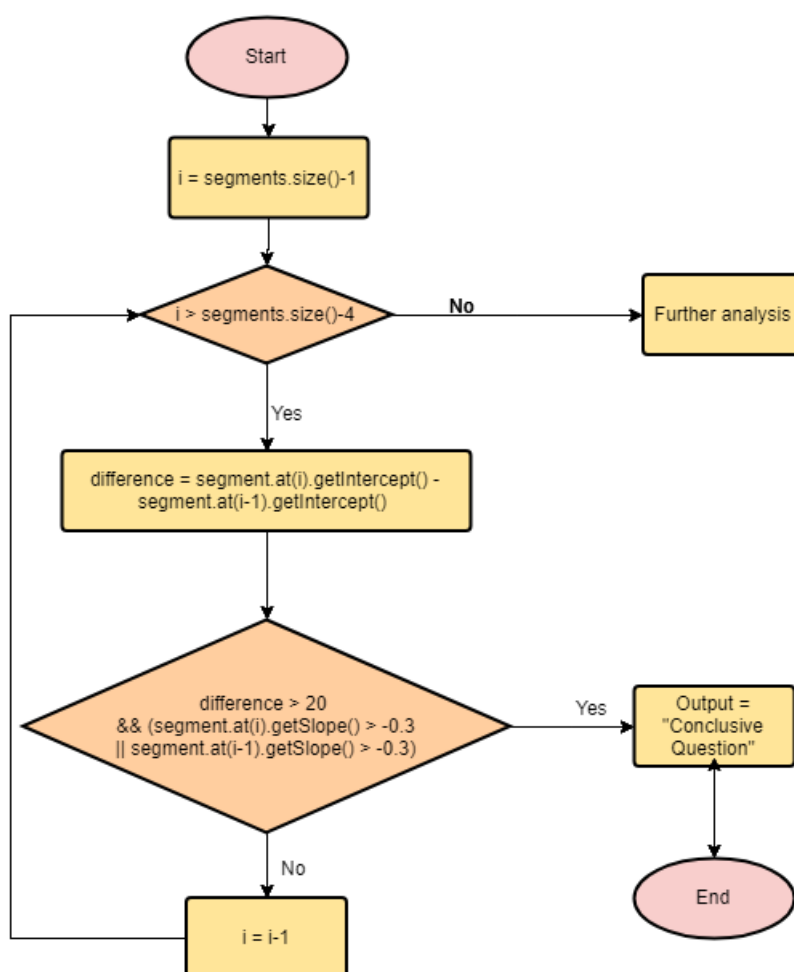
3.1 Pytania rozstrzygnięcia

W pierwszej kolejności nagrania były analizowane pod kątem rozpoznania ich jako pytania rozstrzygnięcia. Ten rodzaj wypowiedzi cechuje się silną antykadencją zlokalizowaną w końcowej części zdania. Jest to najbardziej charakterystyczny rodzaj zdania, dlatego program zaczyna klasyfikację od niego.



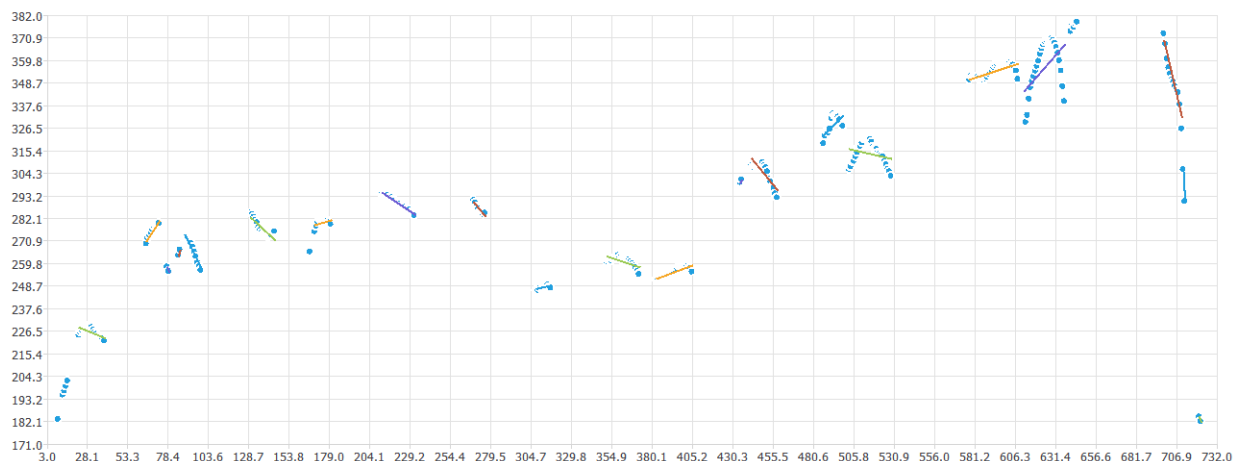
Rysunek 3.2: Pytanie rozstrzygnięcia zadane przez mężczyznę

Pytanie przedstawione na rysunku 10 brzmi 'Faktycznie jest gdzieś w Tobie taka pasja?'. Jest ono wypowiedziane przez mężczyznę. Nie ma w tej wypowiedzi żadnego słowa, które wyraźnie wskazywałoby na to, że jest to pytanie, a nie stwierdzenie. Jednak, dzięki nadaniu wypowiedzi odpowiedniej intonacji, możliwe jest rozpoznanie jej jako pytanie - zarówno przez program, jak i przez człowieka. W przedstawionym przykładzie, wzrost intonacji na końcu nagrania jest bardzo wyraźny, znacznie przekracza zakres typowych częstotliwości F0 uzyskiwanych w głosie męskim. Dla pytań rozstrzygnięcia dość charakterystyczne jest to, że ogólna tendencja intonacji wcale nie musi być rosnąca, zazwyczaj ten wzrost następuje gwałtownie, dla jednego lub kilku końcowych segmentów. W danym przykładzie, przed wystąpieniem akcentu intonacyjnego w ostatnim słowie, intonacja utrzymywała się na stałym poziomie. Nie może jednak być to uznane za regułę, co udowodni kolejny przykład.



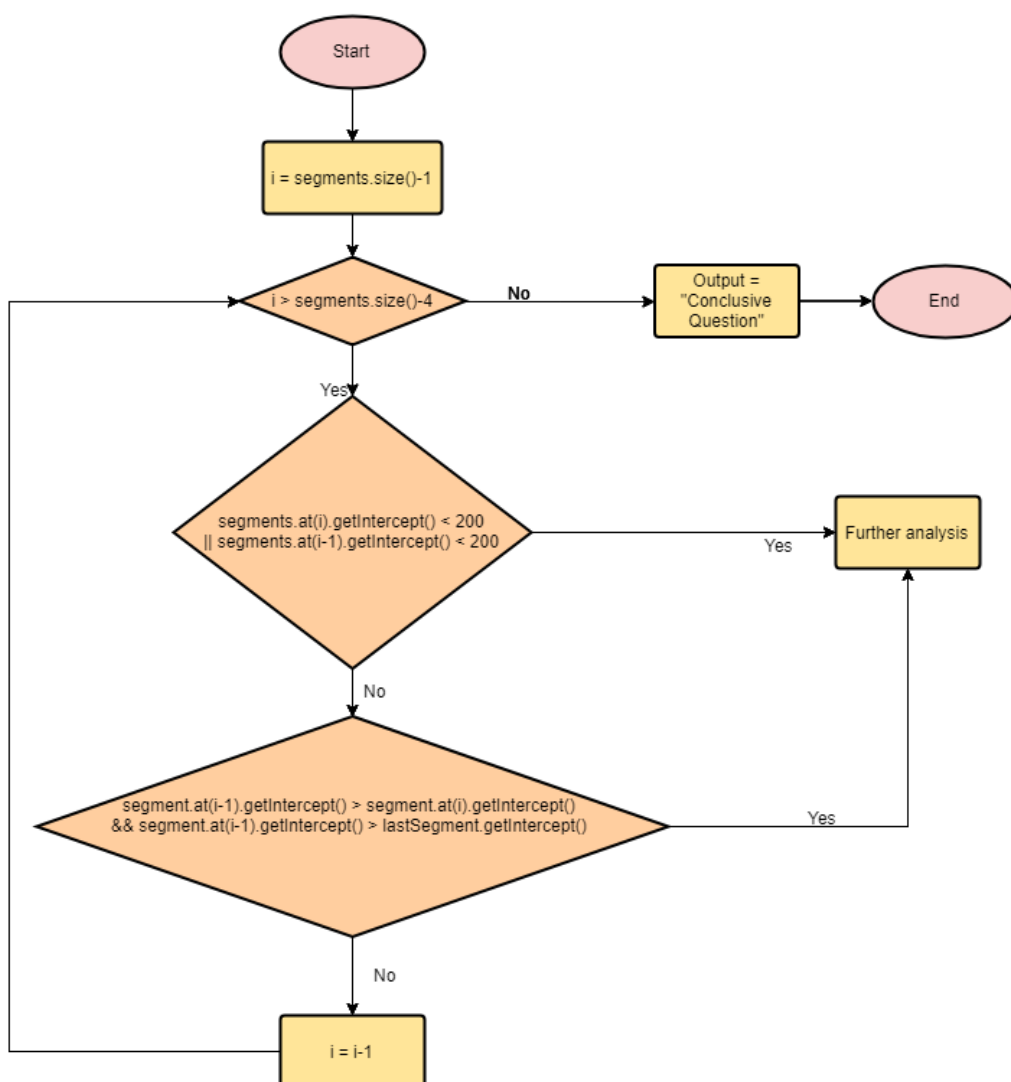
Rysunek 3.3: Schemat rozpoznawania pytań rozstrzygnięcia

Na rysunku 11 przedstawiony został algorytm rozpoznawania pytań rozstrzygnięcia, w przebiegu których nastąpił gwałtowny wzrost wartości F0 w ostatniej fazie wypowiedzi. Pod uwagę brane są trzy ostatnie kontury. Porównywane jest położenie (*ang.intercept*), oraz sprawdzane jest nachylenie (*ang.slope*) przypisanej do sąsiadujących segmentów linii regresji. Jeżeli różnica w położeniu segmentów jest znacząca, a do tego przynajmniej jeden z nich nie jest gwałtownie opadający, wypowiedź zostaje rozpoznana jako pytanie rozstrzygnięcia.



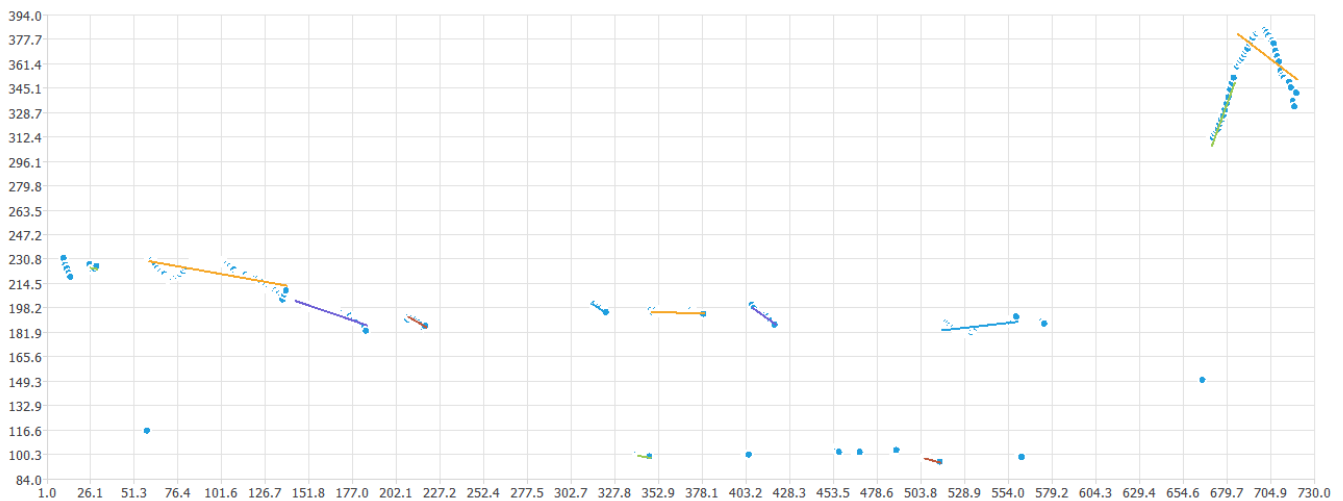
Rysunek 3.4: Pytanie rozstrzygnięcia z nacechowaniem emocjonalnym
, wypowiedziane przez mężczyznę

Pytanie, którego intonacja została przedstawiona na rysunku 12, brzmi "Może o to, że jest to rażąco niesprawiedliwe?" Zostało wypowiedziane przez mężczyznę. Nosi ono również znamiona pytania retorycznego. W tej wypowiedzi nie ma gwałtownego wzrostu intonacji na samym końcu nagrania, zamiast tego intonacja zauważalnie rośnie w trakcie całej wypowiedzi. Mimo, że jest to nagranie głosu męskiego, wyestymowane wartości F_0 znacznie wykraczają poza zakres typowy dla mężczyzn. Spowodowane jest to dużym zabarwieniem emocjonalnym wypowiedzi.



Rysunek 3.5: Schemat rozpoznawania pytań rozstrzygnięcia z tendencją rosnącą

Na rysunku 13 przedstawiony został algorytm rozpoznawania pytań rozstrzygnięcia, w przebiegu których nie wystąpił gwałtowny wzrost wartości F0 w ostatniej fazie wypowiedzi lecz ogólna tendencja intonacji była rosnąca. Ponownie pod uwagę brane są ostatnie trzy kontury. Jako że zaobserwowane zostało występowanie wysokich wartości F0 w przypadku tego rodzaju pytań, jako pierwsze sprawdzane jest położenie segmentów. Jeżeli uzyskane wartości dla któregoś z sąsiadujących segmentów jest mniejsza od 200, wypowiedź na pewno nie jest pytaniem rozstrzygnięcia. Jeżeli warunek jest spełniony, program porównuje położenie badanych segmentów. Jeżeli segment znajdujący się bliżej początku jest położony wyżej od porównywanego z nim segmentu oraz dodatkowo leży wyżej również od ostatniego segmentu, intonacja nie może być postrzegana jako rosnąca, a zatem wypowiedź nie jest pytaniem rozstrzygnięcia. Tym sposobem zostało rozpoznane pytanie przedstawione na rysunku 12.

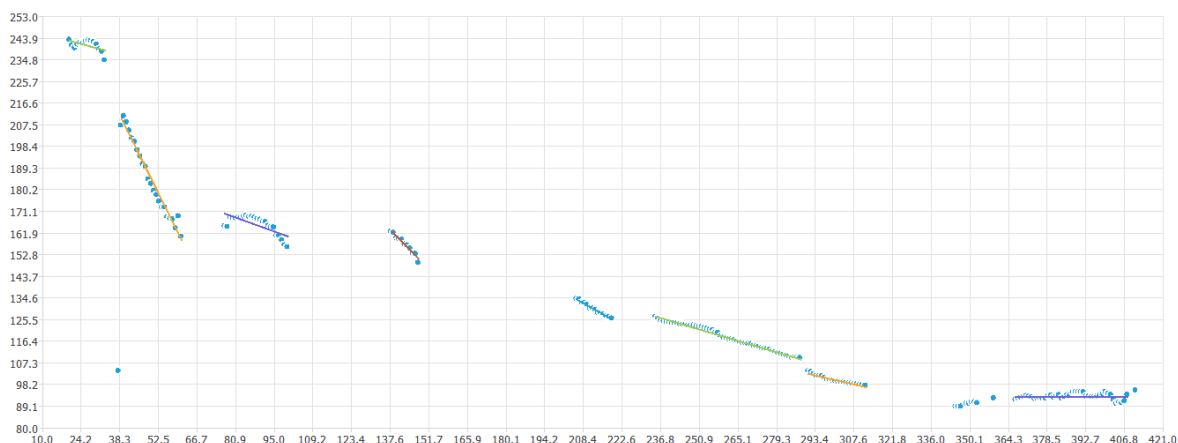


Rysunek 3.6: Pytanie rozstrzygnięcia zadane przez kobietę

Ostatnie z przedstawionych pytań rozstrzygnięcia zostało wypowiedziane przez kobietę. Brzmi ono "A czy Tobie marzy się kariera zagraniczna?". Przebieg intonacji jest zbliżony do przykładu przedstawionego na rysunku 10, tutaj również tendencja intonacji nie jest rosnąca, lecz następuje silny skok intonacji na końcu wypowiedzi.

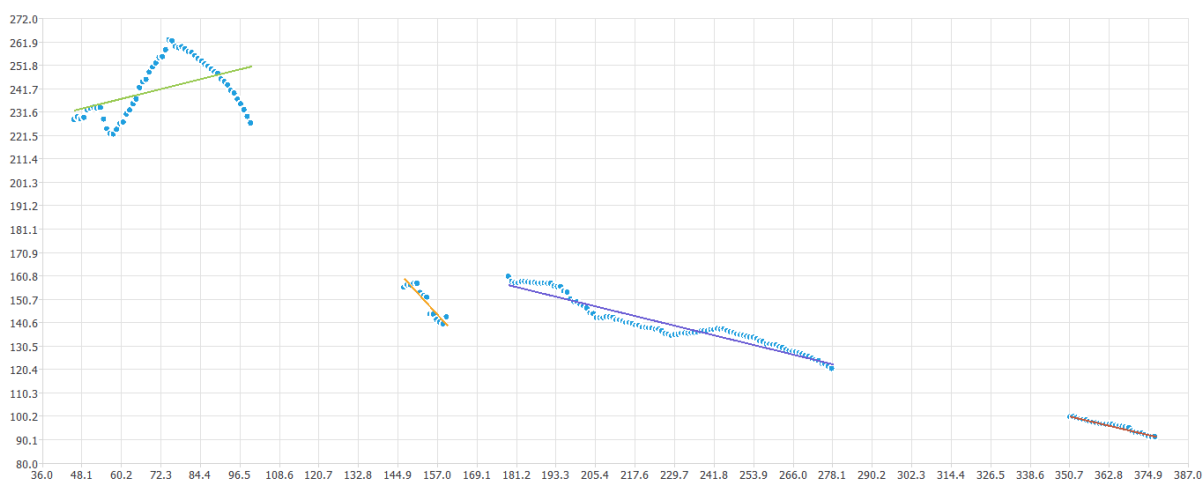
3.2 Pytania dopełnienia

Intonacja nadawana pytaniom dopełnienia całkowicie różni się od opisanej dla pytań rozstrzygnięcia. W tym przypadku zaobserwowany został brak jakiegokolwiek wzrostu intonacji w końcowej części wypowiedzi, oraz cały przebieg intonacji jest najczęściej opadający. Treść tych wypowiedzi jawnie wskazuje, że jest to pytanie, ponieważ na ich początku zawarty jest zaimek pytajny. Przykłady takich zaimków to "kto, dlaczego, który, czemu, jak, co". Charakterystyczny dla tego rodzaju pytań jest wysoko zaintonowany początek wypowiedzi, po którym następuje znaczny spadek estymowanych wartości.



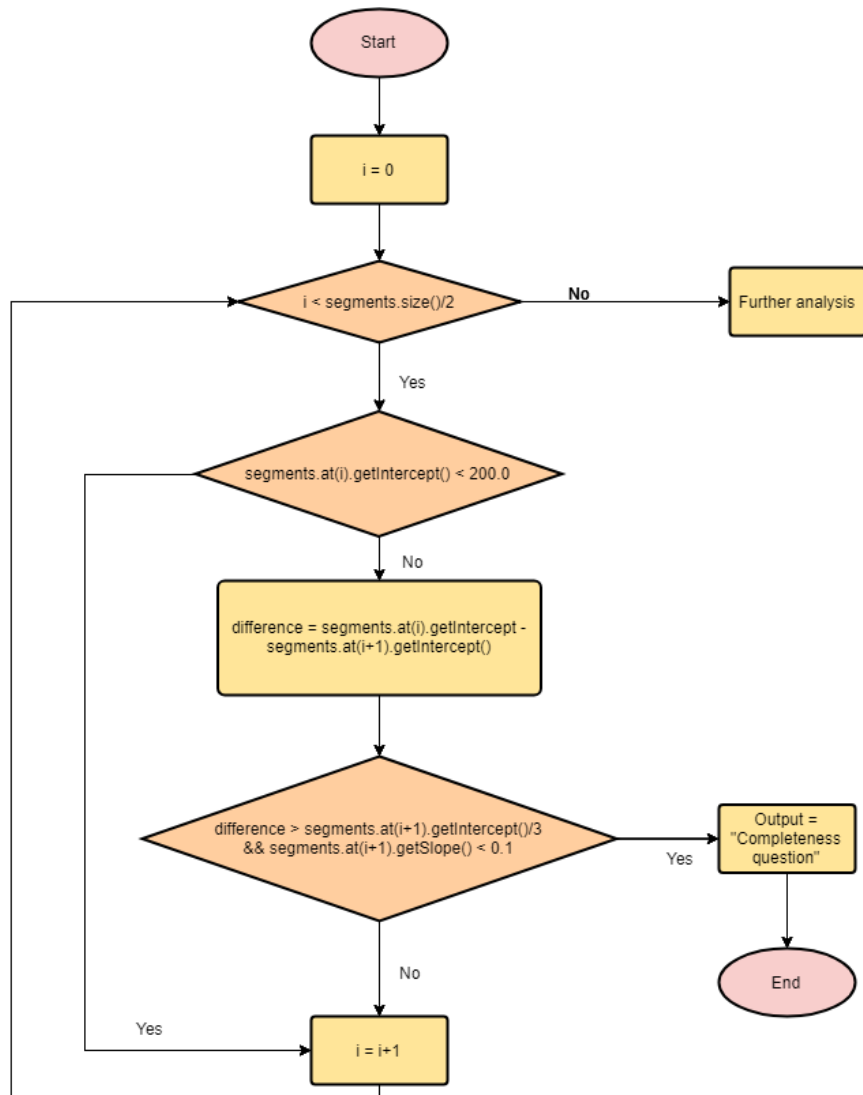
Rysunek 3.7: Pytanie dopełnienia zadane przez mężczyznę

Pytanie, którego intonacja została przedstawiona na rysunku 13, zostało wypowiedziane przez mężczyznę. Jego treść to "Jak zostać marynarzem?". Na pierwszy rzut oka zauważalny jest pierwszy segment, którego wartości górują nad resztą wykresu. W wypowiedzianym pytaniu, duży nacisk intonacyjny został nałożony na zaimek pytajny, po czym nastąpił znaczny spadek wartości F_0 . Cała intonacja jest wyraźnie opadająca. Jako, że zaimek występujący w tym zdaniu składa się jedynie z trzech liter, odpowiadający mu segment również jest krótki. Istotne jest również to, że wartości kolejnych segmentów są mniejsze aż o 80-140Hz w porównaniu do segmentu położonego najwyżej.



Rysunek 3.8: Pytanie dopełnienia zadane przez mężczyznę

Treść pytania przedstawionego na rysunku 16 brzmi "Czego Cię to nauczyło?". Również zostało wypowiedziane przez mężczyznę. W tym pytaniu także najwyżej położony segment jest znacznie większy od pozostałych. Nacisk położony na zaimek pytajny jest wyraźnie widoczny.

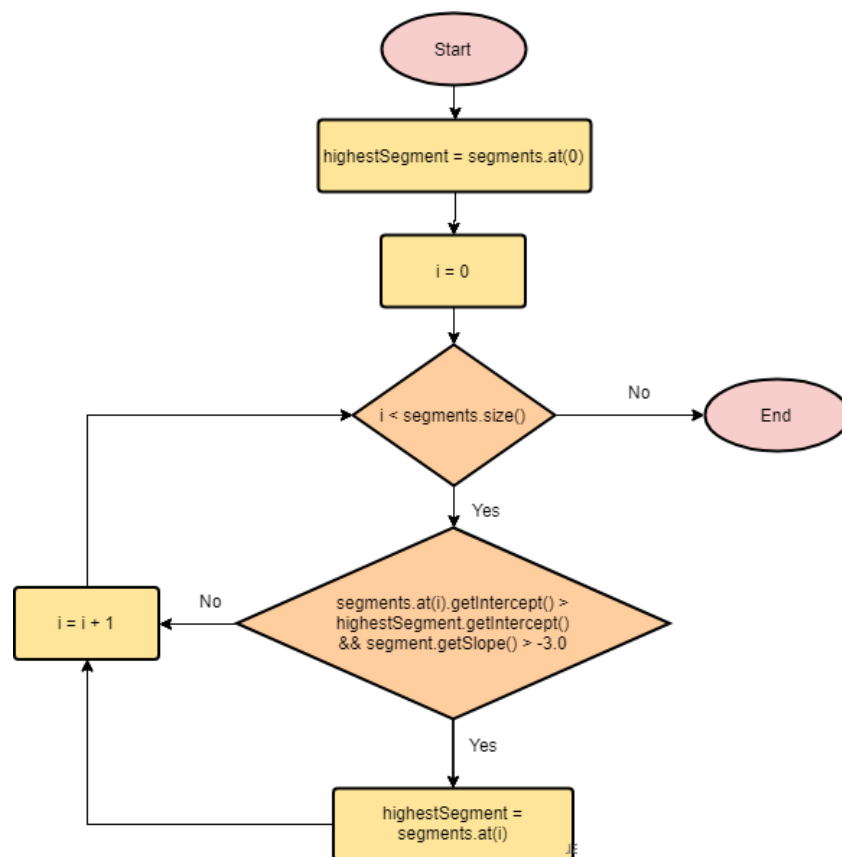


Rysunek 3.9: Schemat rozpoznawania pytań dopełnienia

Rysunek 17 przedstawia schemat rozpoznawania pytań dopełnienia. Ten rodzaj wypowiedzi charakteryzuje się wysokimi wartościami intonacji na początku wypowiedzi, zazwyczaj towarzyszy im późniejszy znaczny spadek wartości. Pod uwagę brana jest pierwsza połowa zbioru segmentów. Najpierw program sprawdza czy dany segment jest położony wystarczająco wysoko. W przeciwnym razie segment ten zostaje ominięty. Następnie obliczona zostaje różnica w położeniu między sąsiadującymi segmentami. Aby zdanie zostało sklasyfikowane jako pytanie dopełnienia, różnica między nimi musi być większa niż wartość położenia niżej znajdującego się segmentu, podzielona na trzy. Zauważono, że z racji dużych różnic między wartościami F0 uzyskiwanymi dla kobiet i mężczyzn, takie porównanie jest wydajniejsze niż porównywanie ze stałą wartością. Dodatkowym warunkiem klasyfikacji jest brak rosnącej tendencji niżej położonego segmentu.

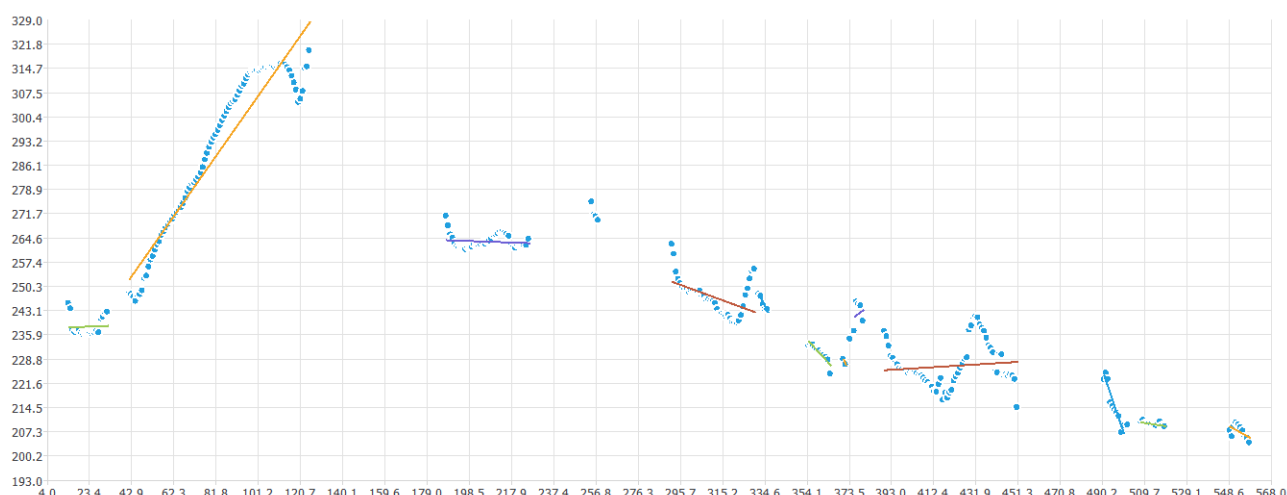
3.3 Zdania rozkazujące

Dla pytań dopełnienia charakterystyczna jest obecność krótkich segmentów o wysokich wartościach zlokalizowanych na początku przebiegu intonacji. W przypadku pytań rozstrzygnięcia, taki charakterystyczny segment obecny był na końcu wypowiedzi. W zdaniach nakazujących komus wykonanie jakiej czynności, kładziony jest silny akcent na czasownik. W skutek tego, najczęściej segmenty o największych wartościach F0 zlokalizowane są w centralnej części przebiegu intonacji, lub zaczynają się w początkowej części, kończąc w środkowej. Z racji nałożonego na nie akcentu, dość często są najdłuższymi segmentami w danej wypowiedzi. W tym przypadku funkcja gramatyczna intonacji przeplata się z funkcją podkreślającą znaczenie danego słowa. Zaprezentowane zostanie 5 różnych grup cech wskazujących na ten rodzaj zdania. Pierwszym krokiem jest znalezienie najwyżej położonego segmentu w pierwszej połowie zbioru.



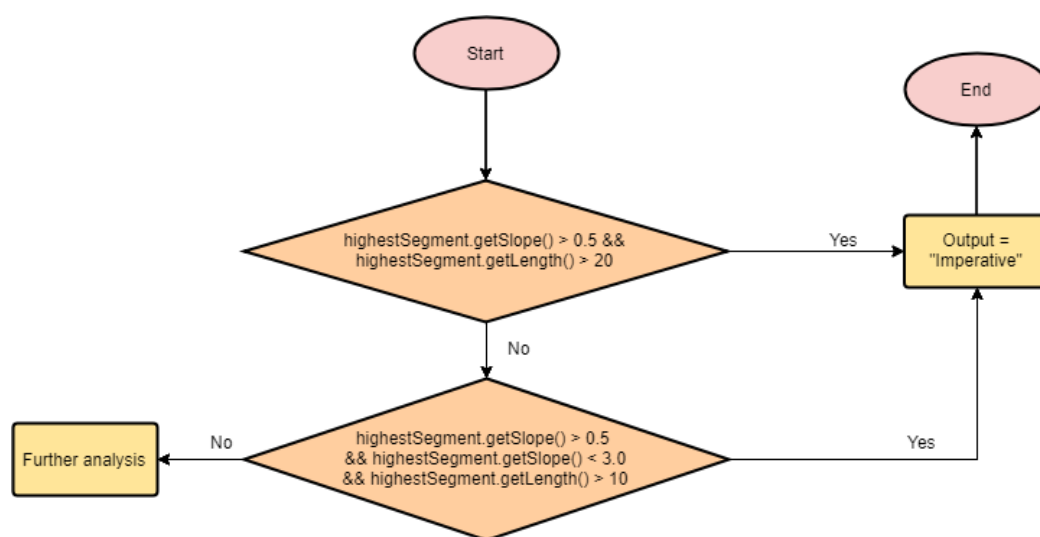
Rysunek 3.10: Schemat wykrywania najwyżej położonego segmentu

Aby dany segment został uznany za najwyżej zlokalizowany, oprócz położenia musi spełniać jeszcze jeden warunek. Nie może być to segment silnie, wręcz pionowo, opadający.



Rysunek 3.11: Zdanie rozkazujące wypowiedziane przez kobietę

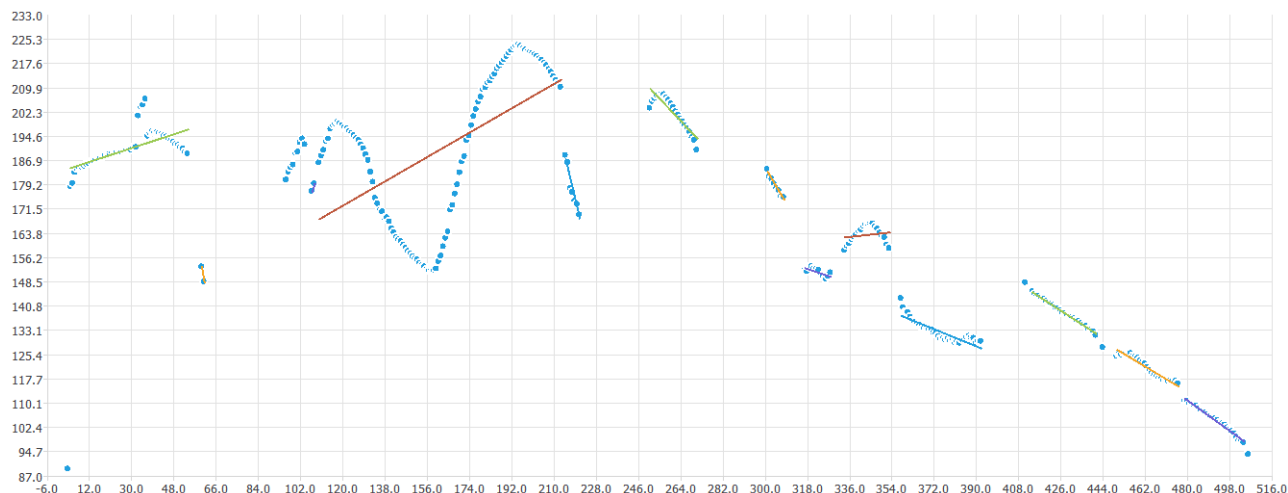
Zdanie przedstawione na rysunku 19 zostało wypowiedziane przez kobietę, a jego treść brzmi "Wyłącz w końcu ten telewizor". Najwyżej położony segment odpowiada czasownikowi oraz gwałtownie rośnie będąc jednocześnie najdłuższym segmentem w zbiorze. Są to cechy wskazujące na zdanie rozkazujące. Zostało wykryte przez fragment algorytmu, który został przedstawiony na rysunku 20.



Rysunek 3.12: Schemat wykrywania zdania rozkazującego zawierającego pierwszą grupę cech

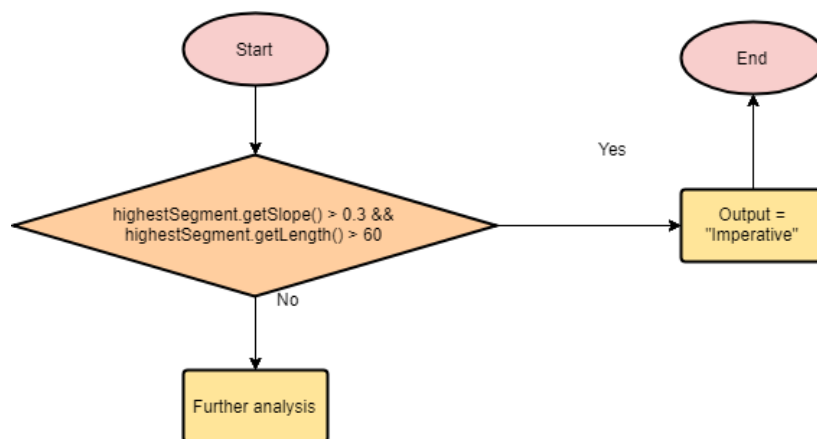
W tym fragmencie algorytmu, program podejmuje próbę rozpoznania zdania jako rozkazującego, bazując na nachyleniu oraz długości najwyżej położonego segmentu. Jeżeli długość danego segmentu jest większa niż 20 jednostek i jednocześnie segment jest silnie rosnący, właściwości te wskazują na zdanie rozkazujące. Jeżeli jednak jest krótszy, wciąż może wskazywać na zdanie rozkazujące, jeżeli nie jest ostro rosnący. Ta cecha jest od-

rzucana przez program, ponieważ zaobserwowano jej związek ze zdaniami twierdzącymi. Jeśli zdanie nie spełnia żadnego z tych warunków, jest poddawane dalszej analizie.



Rysunek 3.13: Zdanie rozkazujące wypowiedziane przez mężczyznę

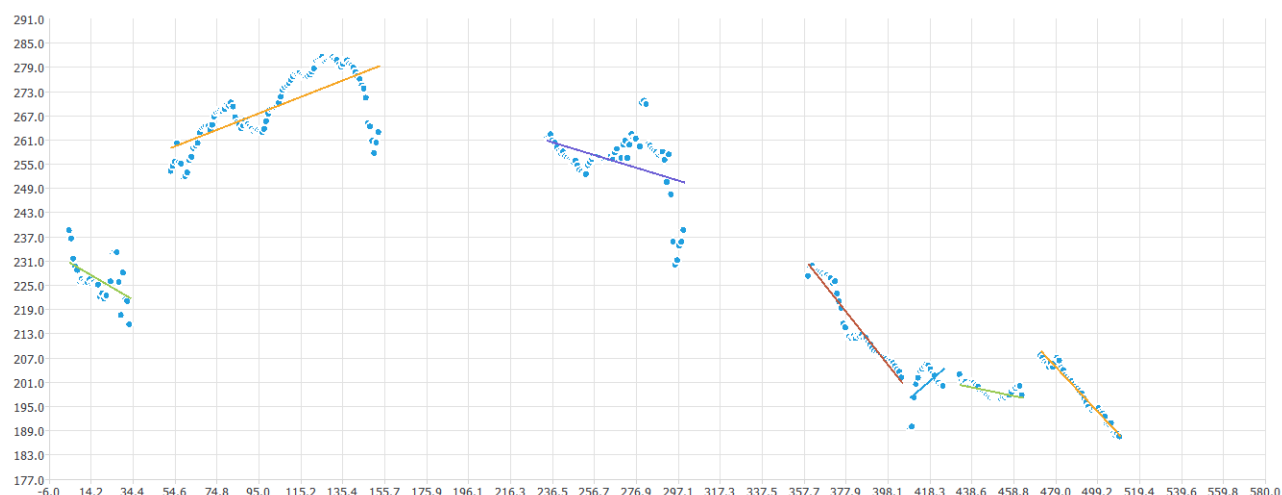
Zdanie przedstawione na rysunku 21 zostało wypowiedziane przez mężczyznę, a jego treść brzmi "Nie próbujcie tego w domu". Wyraźny akcent w tej wypowiedzi jest położony na drugie słowo, sylaby w czasowniku są wręcz przeciągnięte, a więc wykorzystany został również iloczys. Na skutek tego omawiany segment poza największymi wartościami, odznacza się również długością. Kolejną zauważalną cechą jest jego gwałtowny wzrost, zakończony krótkim spadkiem.



Rysunek 3.14: Schemat wykrywania zdania rozkazującego zawierającego drugą grupę cech

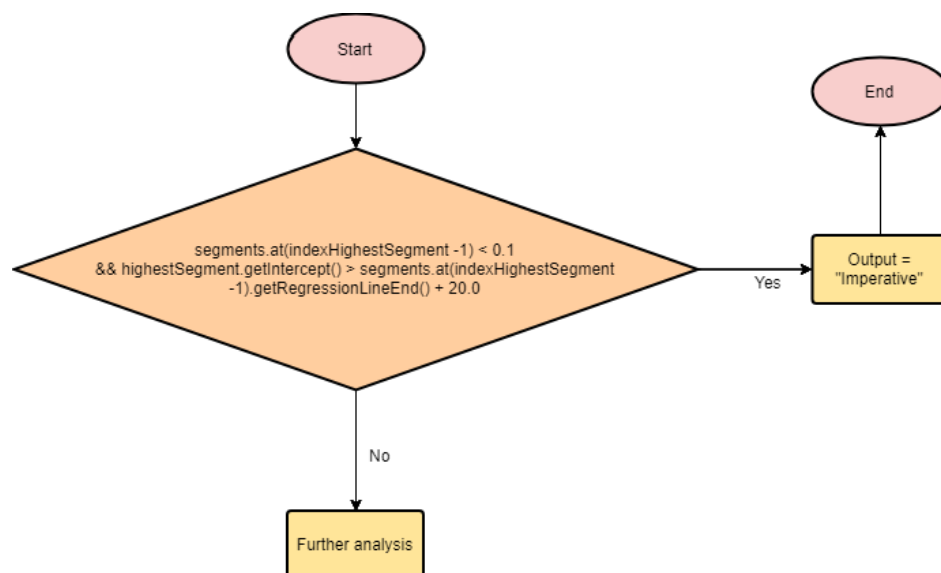
Zdanie może zostać rozpoznane jako rozkazujące, również gdy jego nachylenie jest słabo rosnące, ale za to jest to znacząco długi segment. Długość segmentu odzwierciedla nacisk położony na czasownik. Przykład przedstawiony na rysunku 21 został zidentyfikowany w ten sposób. Badany segment odpowiada naciskowi nałożonemu na słowo

”próbujcie”.



Rysunek 3.15: Zdanie rozkazujące wypowiedziane przez kobietę

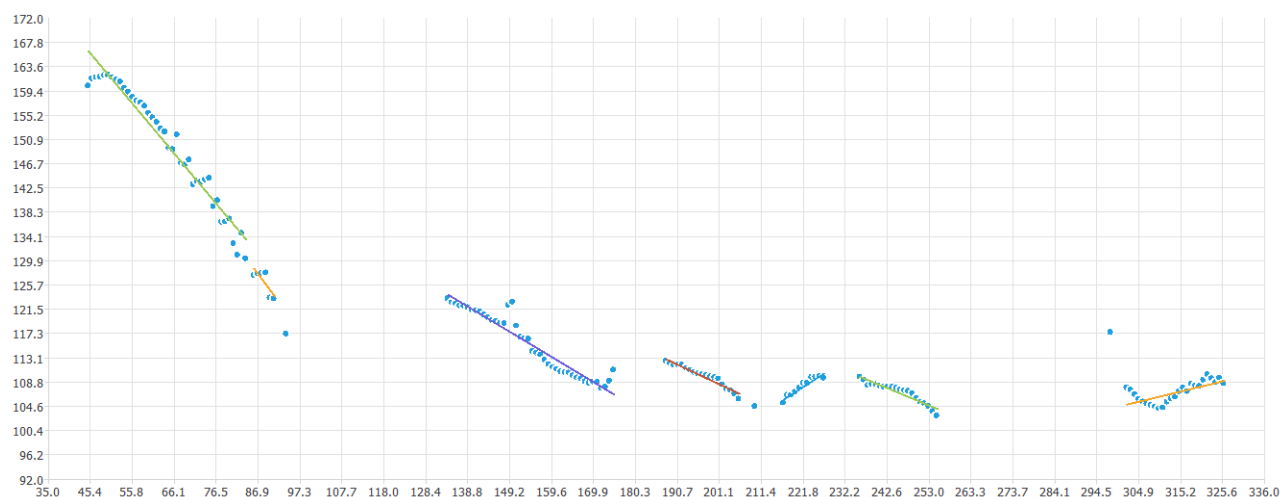
Zdanie przedstawione na rysunku 23 zostało wypowiedziane przez kobietę, a jego treść brzmi ”Wynieś śmieci jak wrócisz”. Akcent intonacyjny ponownie nałożony jest na czasownik w trybie rozkazującym, lecz tym razem jedynie na jego drugą oraz trzecią sylabę. Po pierwszej zauważalna jest krótka pauza. W tym przypadku najistotniejszy segment jest opadający, co pokazuje, że nie zawsze jest rosnący jak w poprzednich przykładach i nie może być ta właściwość traktowana jako powszechna cecha.



Rysunek 3.16: Schemat wykrywania zdania rozkazującego zawierającego trzecią grupę cech

Rysunek 24 przedstawia warunki, po których spełnieniu zdanie z rysunku 23 zostało sklasyfikowane jako zdanie rozkazujące. Ta gałąź algorytmu sprawdza czy segment po-

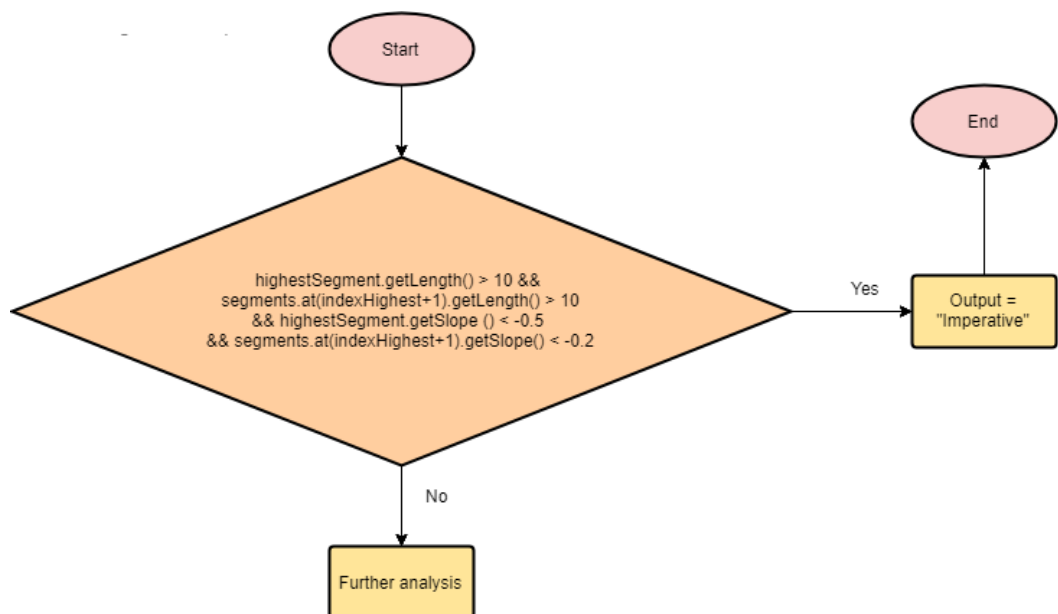
przedzający najwyższy zlokalizowany segment jest opadający. Jeżeli warunek jest spełniony, sprawdzana jest różnica między położeniem najwyższego segmentu z położeniem końca segmentu poprzedzającego. Duży przeskok w takim przypadku wskazuje na zdanie rozkazujące. Jeżeli dane warunki nie zostaną spełnione, program kontynuuje analizę zdania.



Rysunek 3.17: Zdanie rozkazujące wypowiedziane przez mężczyznę

Intonacja tego zdania przypomina przebieg intonacji występujący przy pytaniach dopełnienia. Jednak w tym przypadku różnica między najwyższym położonym segmentem, a następnym segmentem nie była wystarczająco duża, aby rozpoznać je jako pytanie. Kluczowe dla rozpoznania tego zdania jako zdania rozkazującego jest nachylenie obu segmentów. Jeżeli oba są dłuższe niż 10 jednostek oraz są to segmenty opadające, zdanie zostaje rozpoznane jako zdanie rozkazujące.

Warunek, po którego spełnieniu te zdanie zostało sklasyfikowane jako rozkazujące, przedstawiony jest na rysunku 26.

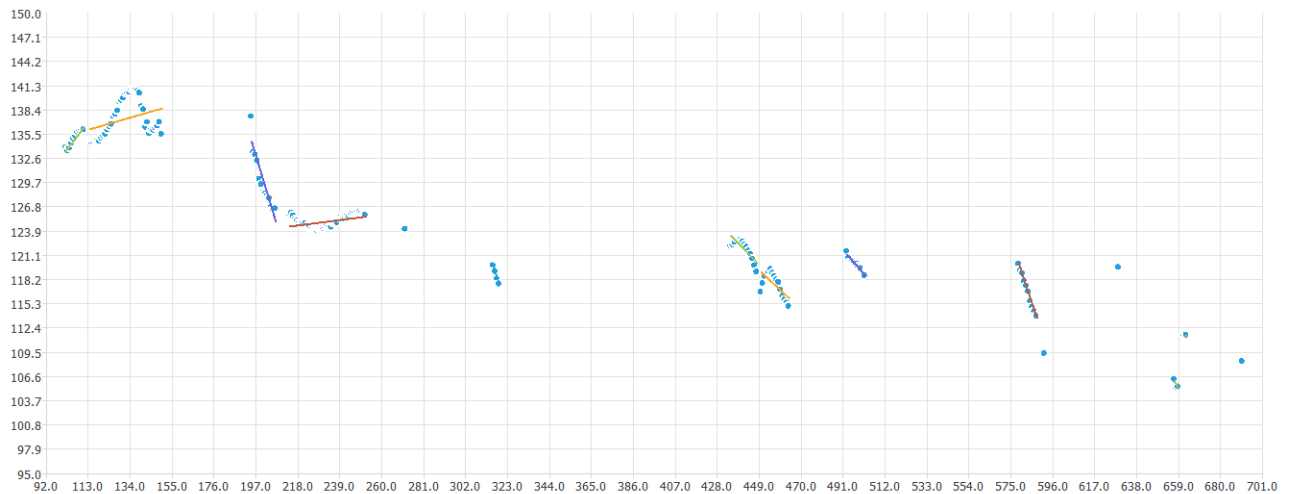


Rysunek 3.18: Schemat wykrywania zdania rozkazującego zawierającego czwartą grupę cech



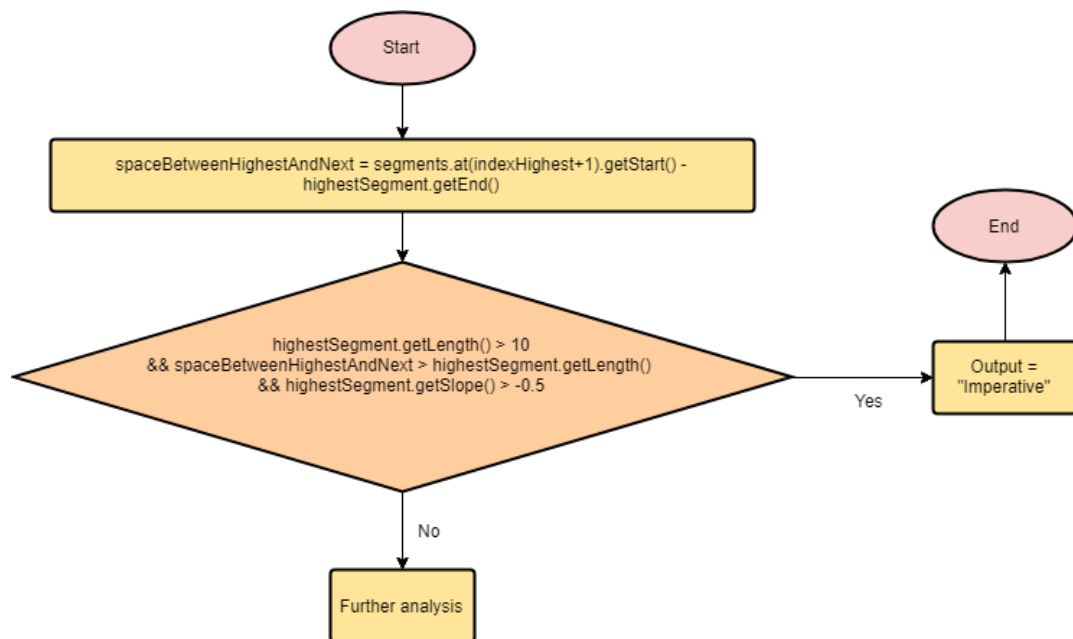
Rysunek 3.19: Zdanie rozkazujące wypowiedziane przez kobietę

Treść rozkazu przedstawionego na rysunku 27 brzmi "Zanieś zakupy do domu". Zdanie zostało wypowiedziane przez kobietę. Charakterystyczna jest obecność dość długiej przerwy zaraz po najbardziej znaczącym segmencie.



Rysunek 3.20: Zdanie rozkazujące wypowiedziane przez mężczyznę

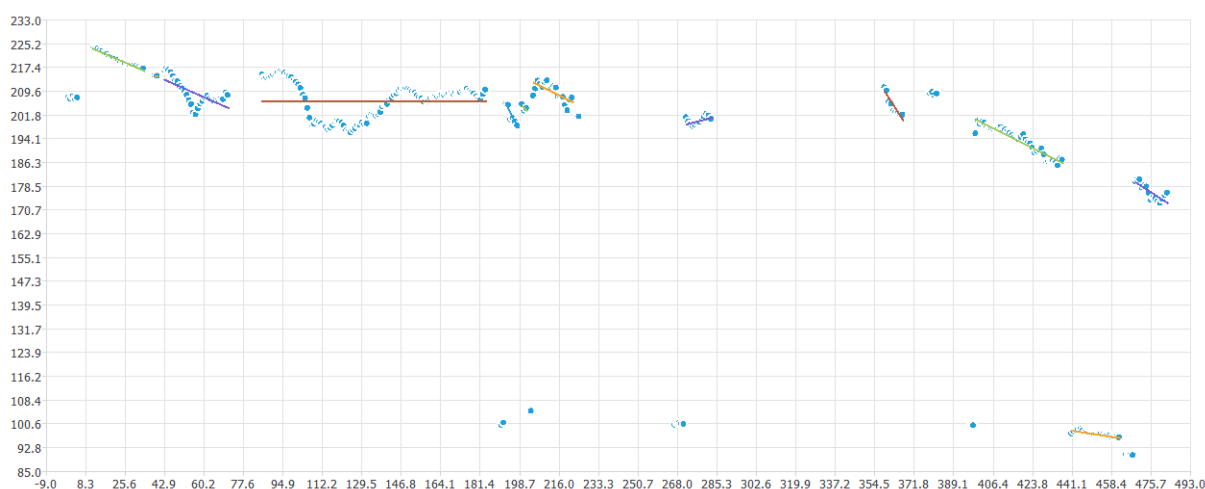
Zdanie przedstawione na rysunku 28 brzmi "Skończ najpierw sprzątać swój pokój". Zauważalne są te same cechy co w zdaniu poprzednim. Bazując na tych obserwacjach opracowany został ostatni warunek rozpoznania zdania rozkazującego, przedstawiony na rysunku 29. Zdanie zostaje uznane za rozkazujące, jeśli najwyżej położony segment jest wystarczająco długi, nie opada stromo, oraz przerwa między nim, a kolejnym segmentem jest większa niż długość najwyższego segmentu.



Rysunek 3.21: Schemat wykrywania zdania rozkazującego zawierającego pięć grupę cech

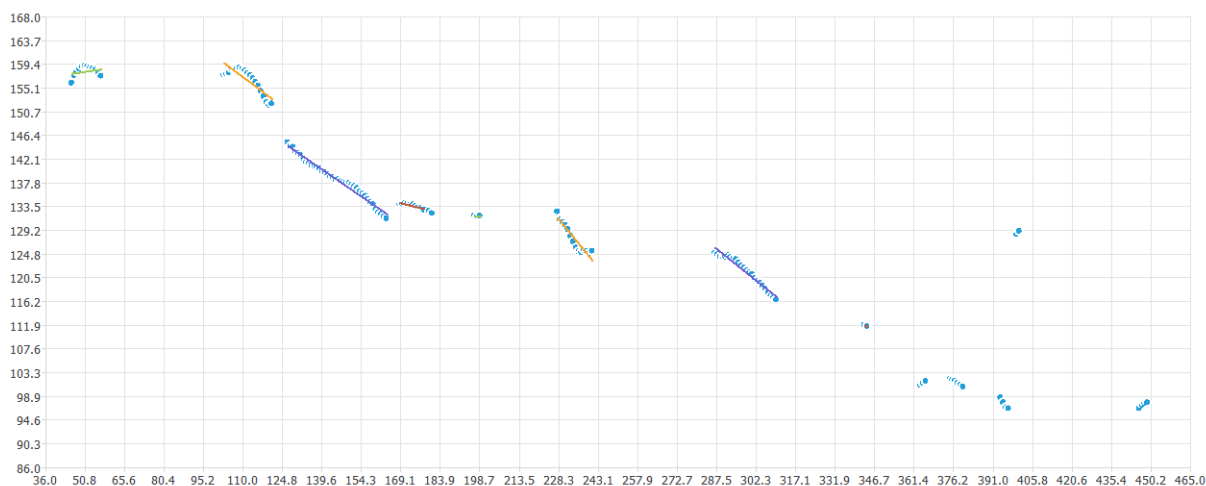
3.3.1 Zdania twierdzące

Podczas obserwacji intonacji zdań twierdzących zauważono stałą lub delikatnie opadającą tendencję. Nie występują w nich nagłe wzrosty lub spadki wartości F0 na początku wypowiedzi. Brak w nich również gwałtownych wzrostów lub wysokich wartości na końcu zdania, tak charakterystycznych dla pytań rozstrzygnięcia. W przypadku opadającego przebiegu intonacji, spadek wartości ma łagodny charakter. Wpływ na to ma fakt, że zdania te są zazwyczaj wypowiadane bez zabarwienia emocjonalnego, ich główną funkcją jest przekazanie informacji. W porównaniu do wcześniej analizowanych rodzajów zdań, w przypadku zdań twierdzących intonacja ma najmniejszy wpływ na postrzeganie tych zdań przez człowieka.



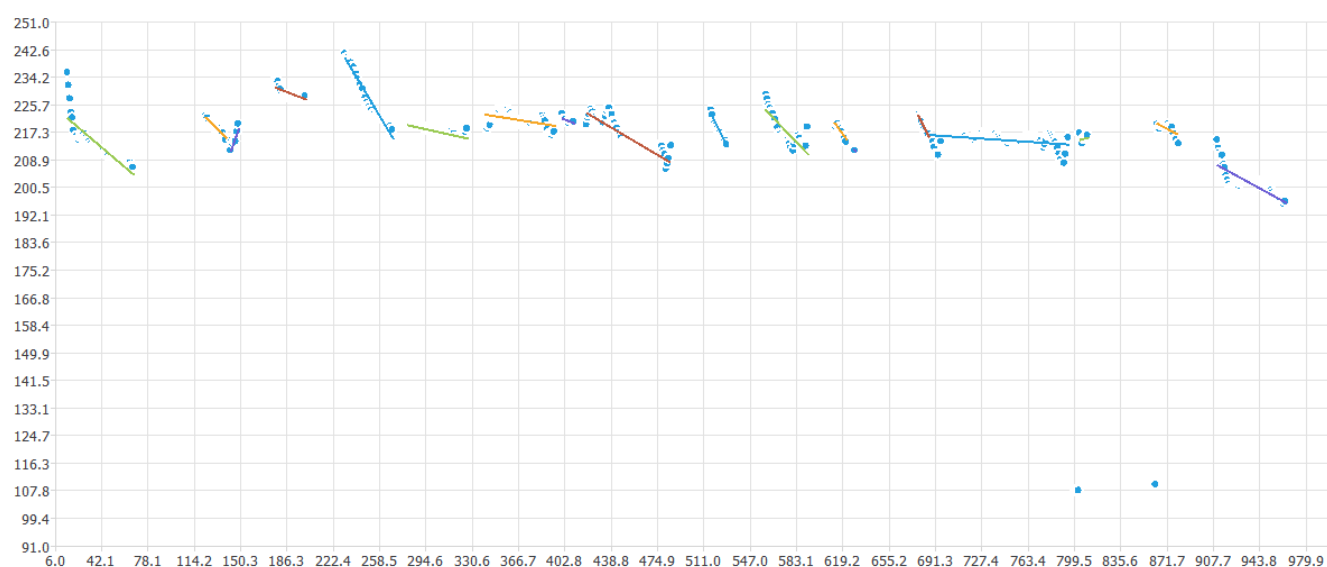
Rysunek 3.22: Zdanie twierdzące wypowiedziane przez kobietę

Zdanie przedstawione na rysunku 30 przedstawia zdanie twierdzące, którego intonacja przedstawia stałą tendencję. Zostało wypowiedziane przez kobietę, a jego treść brzmi "Zawsze byli bardzo wspierający". Średnie wartości kolejnych segmentów są bardzo do siebie zbliżone. Najwyżej zlokalizowany segment jest delikatnie opadający, dodatkowo brak większej przerwy między nim, a następnym segmentem.



Rysunek 3.23: Zdanie twierdzące wypowiedziane przez mężczyznę

Zdanie przedstawione na rysunku 31 zostało wypowiedziane przez mężczyznę, a jego treść brzmi "W zeszłym roku ten las wycięto". Największe wartości intonacji zostały zauważone dla słów "W zeszłym" lecz spadek wartości dla segmentów odpowiadających kolejnym słowom jest dość łagodny. Dopiero ostatnie słowo ma wyraźnie niższą intonację. Główną cechą zdań twierdzących pozwalających rozróżnić ten rodzaj wypowiedzi od pytań dopełnienia jest niewielka różnica między wartościami najwyższego segmentu, a otaczającymi segmentami. W przypadku pytań dopełnienia ta różnica potrafiła wynosić nawet 100Hz i znacznie wykraczać poza typowy zakres częstotliwości. Ten przebieg intonacji jest wręcz idealnym przykładem zdania twierdzącego, średnie wartości każdego segmentu są mniejszego od segmentu poprzedniego, przy tym nie występują gwałtownego spadki między nimi oraz wszystkie segmenty są opadające.



Rysunek 3.24: Zdanie twierdzące wypowiedziane przez kobietę

Przykład przedstawiony na rysunku 32, przedstawia przebieg intonacji zdania twierdzącego, wypowiedzianego przez kobietę, którego treść brzmi "Więc to też naraża ich na niebezpieczeństwo". W tym przykładzie przebieg intonacji również ma wyraźnie stałą tendencję. Nie sposób tutaj mówić o znaczącej różnicy między średnią wartością segmentu zlokalizowanego najwyżej, a otaczającymi go segmentami. Stała tendencja przebiegu intonacji jest jedną z cech wyraźnie wskazujących, że dana wypowiedź jest zdaniem twierdzącym.

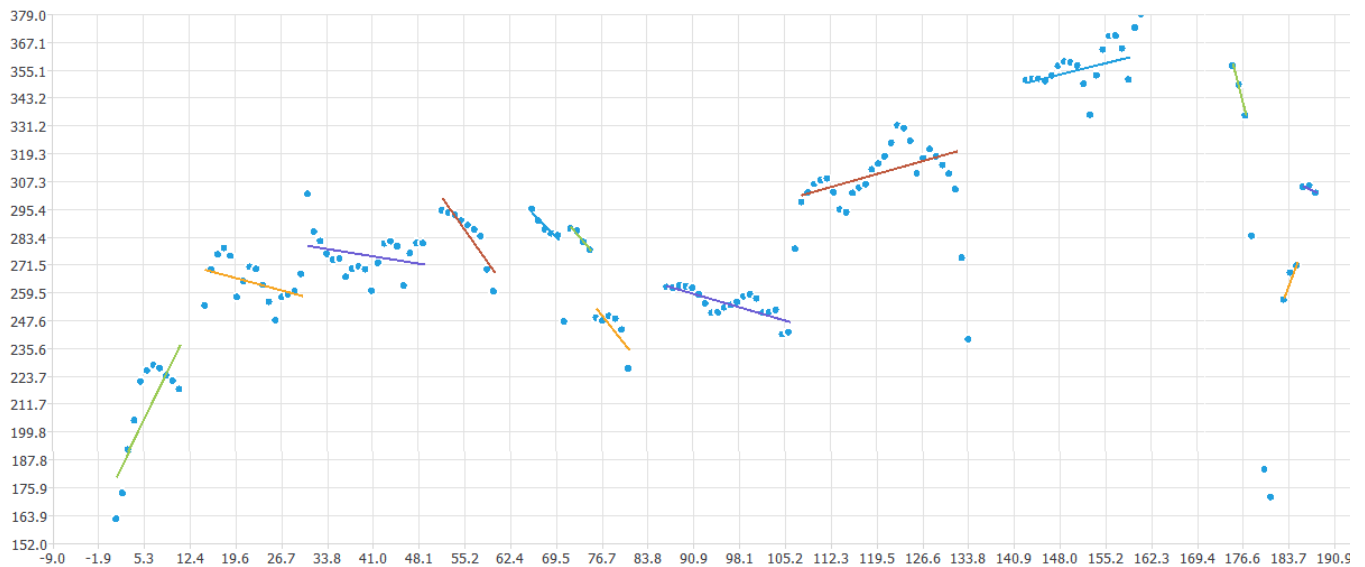
3.4 Porównanie wyników otrzymanych z wykorzystaniem YIN i Praata

Jednym z założeń pracy było porównanie wyników uzyskanych przez proponowaną metodę rozpoznawania, dla konturów wykrytych za pomocą dwóch różnych metod ekstrakcji F0. W tym celu dla każdego z nagrań znajdujących się w bazie uczącej przygotowano plik z wartościami F0 uzyskanymi przez PRAAT-a. Wartości te następnie są wczytywane do programu będącego efektem tej pracy, dzielone na segmenty oraz analizowane. Tendencje przebiegów intonacji uzyskane za pomocą PRAAT-a są takie same jak te uzyskane z użyciem algorytmu YIN. Różnice mogą być obserwowane jedynie w długościach poszczególnych segmentów, na co wpływ ma zastosowany sposób dzielenia intonacji na segmenty.

Poprawność klasyfikacji dla intonacji uzyskanej za pomocą PRAAT-a jest nieznacznie gorsza od tej uzyskanej dla segmentów opartych na algorytmie YIN. Większa ilość błędnych rozpoznań została zaobserwowana dla zdań rozkazujących. Warunki rozpoznania tych zdań są oparte w istotnym stopniu na wartościach nachylenia segmentów oraz odległościach między nimi. Wartości progowe były ustalone metodą obserwacyjną. Kontury uzyskane za pomocą PRAAT-a zachowują podobną tendencję, lecz inna metoda ekstrakcji powoduje, że w poszczególnych przypadkach parametry te istotnie się różnią.

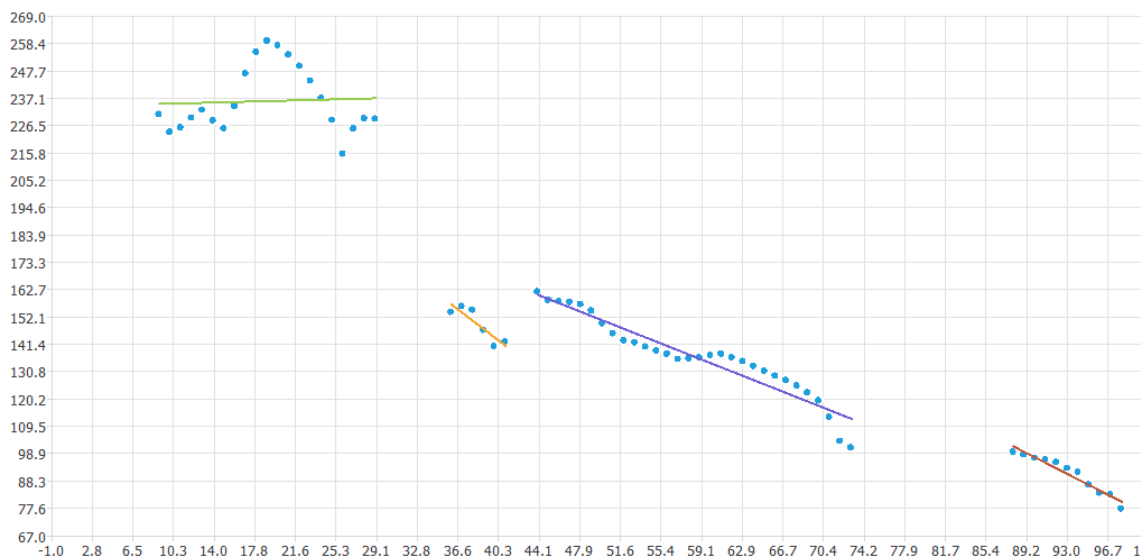
3.4.1 Pytania rozstrzygnięcia oraz dopełnienia

Metoda rozpoznawania pytań sprawdza się równie dobrze, bez względu na źródło pochodzenia wartości F0.



Rysunek 3.25: Intonacja pytania przedstawionego na rysunku ... lecz uzyskana za pomocą PRAAT-a

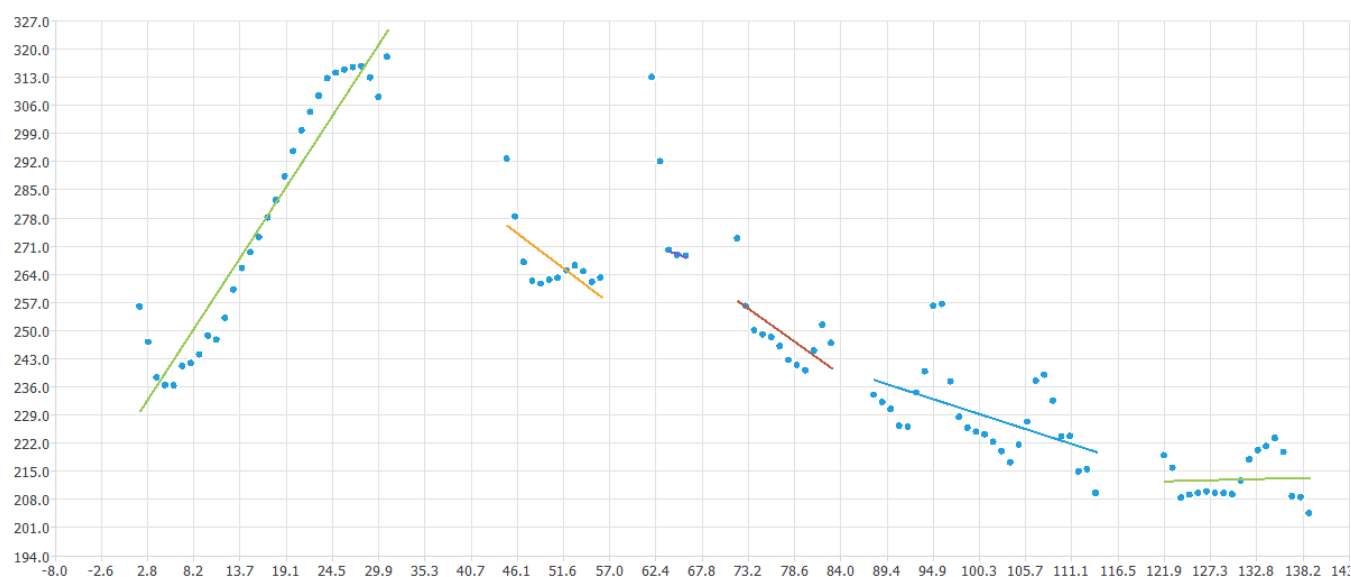
Rosnąca tendencja intonacji została wykryta również gdy uzyskane segmenty oparte były na wartościach F0 pochodzących z PRAAT-a.



Rysunek 3.26: Intonacja pytania przedstawionego na rysunku ... lecz uzyskana za pomocą PRAAT-a

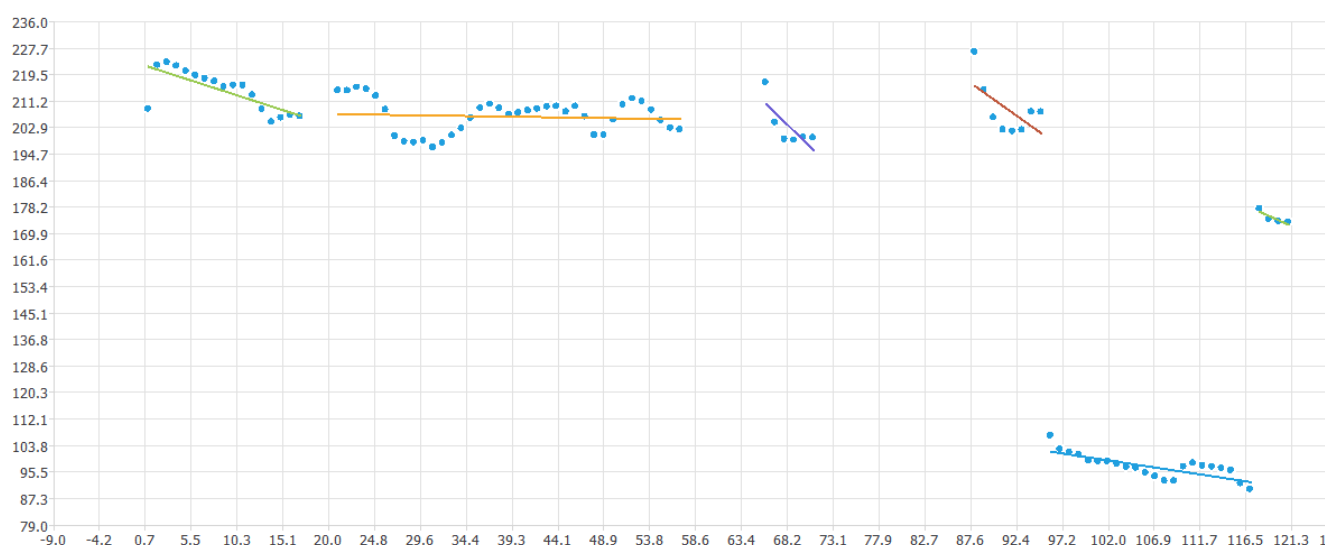
W przypadku pytań dopełnienia, gwałtowne spadki między najwyżej zlokalizowanym segmentem oraz następnym segmentem są wykrywane bez względu na sposób pochodzenia wartości F0.

3.4.2 Zdania rozkazujące oraz zdania twierdzące



Rysunek 3.27: Intonacja zdania rozkazującego przedstawionego na rysunku ... lecz użyta za pomocą PRAAT-a

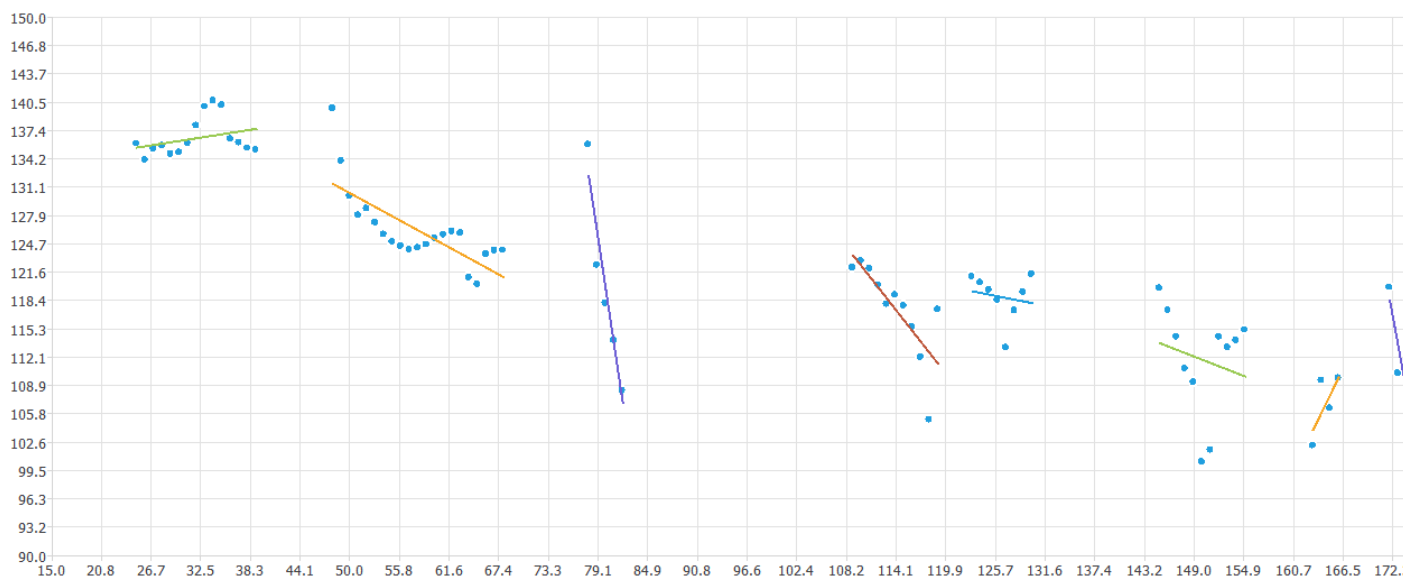
Silny nacisk położony na czasownik w trybie rozkazującym jest również odzwierciedlony w przebiegu intonacji uzyskanej za pomocą PRAAT-a. Zdanie zostało poprawnie rozpoznane przez program.



Rysunek 3.28: Intonacja zdania twierdzącego przedstawionego na rysunku ... lecz użyta za pomocą PRAAT-a

Przykład zdania twierdzącego również pokazuje podobieństwo tendencji przebiegów intonacji uzyskanych w różny sposób. Stała tendencja została zachowana.

3.4.3 Błędne rozpoznania



Rysunek 3.29: Intonacja zdania rozkazującego przedstawionego na rysunku ... lecz uzyskana za pomocą PRAAT-a

Zdanie rozkazujące, którego intonacja uzyskana za pomocą PRAAT-a została przedstawiona na rysunku ... zostało błędnie sklasyfikowane jako zdanie twierdzące. Intonacja tego samego zdania, lecz uzyskana za pomocą YIN, została przedstawiona na rysunku ... i w tym przypadku zostało sklasyfikowane poprawnie jako zdanie rozkazujące. Przyczyną błędnego rozpoznania należy szukać w założeniach podejścia opartego na długościach segmentów oraz odległościach między nimi. W przypadku wartości uzyskiwanych za pomocą PRAAT-a zarówno segmenty jak i przerwy między nimi są stosunkowo krótsze, powodując błędy w rozpoznawaniu.

Rozdział 4

Wnioski

Uzyskane rezultaty można rozpatrywać dwustopniowo. Program wykorzystując algorytm YIN do ekstrakcji intonacji uzyskuje bardzo dobrą skuteczność w rozpoznawaniu obu rodzajów pytań. Zaobserwowane zmiany w intonacji towarzyszącej pytaniom są wystarczające dla poprawnego rozpoznania w zdecydowanej większości przypadków. Potwierdza to dotychczasowe badania wykonane w tym zakresie. Trudniejszym zadaniem okazało się rozróżnianie zdań rozkazujących oraz twierdzących. Pewnym utrudnieniem była ubogość dotychczasowych badań wykonanych dla zdań rozkazujących. W toku pracy udało się jednak zaobserwować wiele cech różniących oba rodzaje zdań, których zastosowanie pozwoliło uzyskać zadowalające wyniki.

Opracowaną metodę opartą na zaobserwowanych cechach przetestowano również na intonacji zdań uzyskanej za pomocą innego algorytmu. Był to algorytm oparty na autokorelacji, używany przez program PRAAT. W przypadku rozpoznawania pytań, nie zaobserwowano różnic. Rozpoznawanie przebiega prawidłowo w takim samym stopniu. Rozpoznawanie zdań rozkazujących oraz twierdzących wskazało jednak wady w podejściu opartym na długościach segmentów oraz odległościach między nimi. Uzyskane rezultaty są nieznacznie, ale jednak gorsze od tych uzyskanych bazując na ekstrakcji opartej na algorytmie YIN.

Porównanie wyników opartych na analizie intonacji uzyskanej za pomocą różnych algorytmów pokazało możliwe drogi usprawnienia metody rozpoznawania. Podejście oparte na ustaleniu wartości progowych dla długości segmentów oraz odległości między nimi okazało się dość wadliwe. Pierwszym krokiem w celu poprawy uzyskiwanych wyników powinno być zwiększenie skalowalności tego rozwiązania. Nowe podejście nie powinno używać stałych jako wartości progowych. Kolejnym krokiem udoskonalenia metody powinno być zwiększenie bazy nagrań. Dotyczy to zwłaszcza zdań rozkazujących. Większa baza nagrań umożliwiłaby zwiększenie liczby cech odróżniających te zdania od zdań twierdzących.

Bibliografia

- [1] Terapia mowy [Online] [Dostęp 05.02.2019] <http://terapiamowy.com/index.php/niedowlad-krtani/>
- [2] Three parts of speech [Online] [Dostęp 10.02.2019] <https://uiowa.edu/voice-academy/three-parts-speech>
- [3] Fernando Trujillo, The Production of Speech Sounds, English Phonetics and Phonology
- [4] Zhaoyan Zhang, Mechanics of human voice production and control [Online] [Dostęp 01.03.2019] <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5412481/>
- [5] The prosody of speech: Melody and rythm, Sieb Nooteboom, 1997, Utrecht University
- [6] Thornbury, Scott.: (2006). An A-Z of ELT. Oxford: Macmillan Publishers, Ltd.
- [7] Martine Grice and Stefan Baumann, An introduction to intonation-functions and models, Ifl Phonetik and Universität Köln(2007)
- [8] Daniel Hirst Albert François di Cristo, A survey of intonation systems, 1998
- [9] Prieto, P., and Roseano, P. (2018). Prosody: Stress, Rhythm, and Intonation. In K. Geeslin (Ed.), The Cambridge Handbook of Spanish Linguistics (Cambridge Handbooks in Language and Linguistics, pp. 211-236)
- [10] Definition of phoneme, <https://www.britannica.com/topic/phoneme>, [Online][Dostęp 10.03.2019]
- [11] Bogusław Dunaj: Dwa dyskusyjne problemy polskiej fonologii. W: Prace językoznawcze 19. Studia polonistyczne. Alina Kowalska, Aleksander Wilkoń (red.). Katowice: Uniwersytet Śląski, 1991, s. 40–46, seria: Prace Naukowe Uniwersytetu Śląskiego w Katowicach.

- [12] Hollien, H., and Ship, T. (1972) “Speaking fundamental frequency and chronological age in males,” *J. Speech Hear. Res.* 15, 155–159
- [13] Pegoraro-Krook, M. 1. (1988). “Speaking fundamental frequency characteristics of normal Swedish subjects obtained by glottal frequency analysis,” *Folia Phoniatica* 40, 82–90
- [14] Gilbert, H.R., and Weismer, G.G. (1974). “The effects of smoking on the speaking fundamental frequency of adult women,” *J. Psycholing. Res.* 3, 225–231.
- [15] What is Formant? https://ec-concord.ied.edu.hk/phonetics_and_phonology/wordpress/learning_
[Online][Dostęp 10.04.2019]
- [16] Kawahara H., de Cheveigne A. YIN, a fundamental frequency estimator for speech and music, 2002