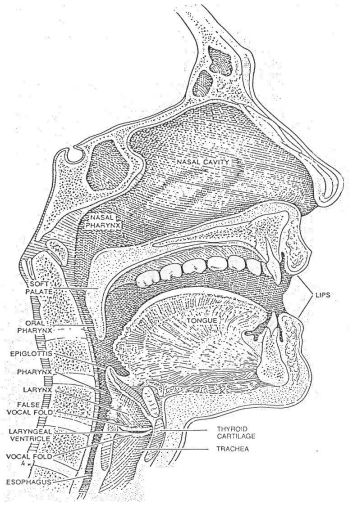
Polsko-Japońska Wyższa Szkoła
Technik Komputerowych

Przygotowanie bazy difonów języka polskiego dla realizacji syntezy mowy w systemie MBROLA



Krzysztof Szklanny

Praca magisterska napisana pod kierunkiem prof. dr hab. Krzysztofa Maraska

Konsultant: dr hab. Ryszard Gubrynowicz

Warszawa 2002

Moim Rodzicom

Serdeczne podziękowania chciałem złożyć Panu Krzysztofowi Maraskowi, który przez cały czas realizacji projektu wspierał mnie swoja wiedzą, podtrzymywał na duchu, odpowiadał na męczące pytania, a w chwilach zwątpienia dodawał otuchy. Dziękuję również Panu Ryszardowi Gubrynowiczowi, który w sytuacjach kryzysowych zawsze służył pomocą i udzielał wyjaśnień aż do „zmęczenia materiału” ☺. Również chciałem podziękować Panu Barisowi Bozkurtowi z zespołu MBROL-i, który wspierał mnie w moich dążeniach.

Wstęp

Praca ta jest związana ze stworzeniem nowej bazy difonów języka polskiego dla realizacji syntezy mowy w systemie MBROLA. System MBROLA powstał na Politechnice w Mons w Belgii. Autorami systemu są Thierry Dutoit i Vincent Pagel.

Stworzenie nowej difonowej bazy składało się z kilku etapów: wygenerowania korpusu difonów, przeprowadzenia nagrań, segmentacji i testowania, z których najtrudniejszym był etap segmentacji. Poprawność posegmentowanego korpusu została przetestowana przy użyciu słownika wyrazów zawierającego większość połączeń fonemów języka polskiego. Największym sukcesem jest akceptacja i umieszczenie bazy difonów na stronie internetowej MBROL-i (http://tcts.fpms.ac.be/synthesis/mbrola).

Do pracy dołączony jest dysk CD. Umieszczone w pracy oznaczenia (CD) wskazują na to, że materiały dotyczące tego zagadnienia zostały umieszczone na dysku CD. Między innymi znajdują się na nim animacje prezentujące zastosowania systemów syntezy mowy, przykładowe pliki dźwiękowe porównujące różne rodzaje syntezy mowy, jak i korpus, baza difonów oraz elektroniczna wersja pracy magisterskiej. Również na dysku CD znajdują się aplikacje potrzebne do uruchomienia oraz zapoznania się z projektem.

Bazę difonów opracowano w Polsko-Japońskiej Wyższej Szkole Technik Komputerowych.

Program ułatwiający przeglądanie dysku CD został napisany w Visual Basic-u.

1.Streszczenie

Głównym celem pracy było przygotowanie bazy difonów języka polskiego dla realizacji syntezy mowy w systemie MBROLA.

Synteza mowy jest procesem generowania mowy ludzkiej w sposób sztuczny. Im bardziej brzmi naturalnie i płynnie tym bardziej jest doskonała. Celem nowoczesnych projektów jest zapewnienie takiej jakości syntezy, by słuchający nie był w stanie odróżnić mowy syntetyzowanej od naturalnej.

Generalizując, istnieją cztery rodzaje syntezy mowy:

- Formantowa
- Artykulacyjna
- Konkatenacyjna
- Korpusowa

Zagadnienia dotyczące tworzenia bazy difonów są ściśle związane z konkatenacyjną syntezą mowy, która generuje mowę poprzez łączenie ze sobą elementów akustycznych powstałych z naturalnej mowy (fony, difony, trifony, sylaby).

Baza jednostek akustycznych stanowi "serce" każdego rodzaju konkatenacyjnej syntezy mowy.

W pracy przeanalizowano zagadnienia związane z generowaniem mowy naturalnej oraz tworzeniem jej w sposób sztuczny. Pierwszy rozdział pracy stanowi wprowadzenie w tematykę.

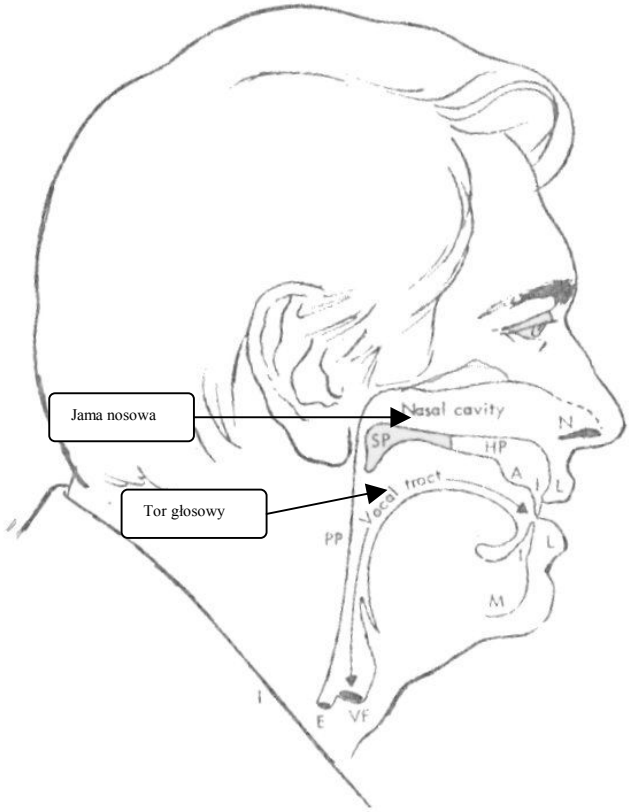
W drugim rozdziale zostały przedstawione zagadnienia związane z fonetyką akustyczną obrazującą sposób powstawania dźwięków u człowieka. Zaprezentowano historię fonetyki, budowę narządu człowieka oraz klasyfikacje dźwięków przez niego artykułowanych. Opisano również zagadnienia dotyczące organizacji wypowiedzi oraz transkrypcji fonetycznej.

W trzecim rozdziale zostały przedstawione niezbędne elementy i definicje związane z syntezą mowy. Począwszy od wprowadzenia w dziedzinę syntezy mowy poprzez opisanie budowy systemów syntezy mowy, rodzajów generowania sztucznej mowy oraz stosowanych algorytmów aż po zastosowania pełnych systemów TTS. System Text-to-speech (TTS) jest systemem odpowiadającym za konwersję w jakiejkolwiek formie wprowadzonego tekstu na mowę w postaci sztucznie generowanego sygnału.

W rozdziale czwartym został umieszczony opis przebiegu części praktycznej. Zostały przedstawione poszczególne etapy związane z przygotowaniem bazy difonów. Pierwszym etapem było przygotowanie korpusu, następnie przeprowadzenie nagrań. Najbardziej skomplikowanym etapem była realizacja procesu segmentacji, czyli wyodrębnienie difonów w nagranym korpusie. Etap ten wymagał dużej precyzji i dokładności. Efektywność i wkład pracy oceniłem testując korpus z uwzględnieniem wszystkich najczęściej występujących połączeń difonów w języku polskim. Najlepszym, dowodem na potwierdzenie jakości bazy difonów jest jej akceptacja i normalizacja na Politechnice w Mons przez zespół MBROL-i.

Od maja br. opracowana baza difonów znajduje się na stronie internetowej MBROL-i (http://tcts.fpms.ac.be/synthesis/mbrola) jako nowy model głosowy.

2. Akustyka mowy polskiej



Rysunek 2.1. Schemat narządu artykulacyjnego
(Źródło Gubrynowicz R. *PDA*)

2.1 Historia

Zainteresowanie mową sięga czasów starożytnych. Około VI wieku p.n.e indyjscy uczeni opracowali pierwsze, podstawowe reguły gramatyczne. Natomiast starożytni Grecy dokonali opisu narządów mowy, stworzyli klasyfikację dźwięków mowy i podstawy terminologii fonetycznej.

Zdecydowanie zainteresowanie procesem wytwarzania dźwięków mowy wzrosło pod koniec XVIII wieku, kiedy zaczęły powstawać pierwsze urządzenia sztucznie generujące dźwięk. (patrz 3.1 Początki syntezy mowy).

Fonetyka akustyczna narodziła się w pierwszych latach po zakończeniu II wojny światowej. Wtedy skonstruowany został spektrograf akustyczny. Przyrząd ten służący badaniom struktury akustycznej dźwięków mowy umożliwił uzyskanie zapisów na których przebiegi akustyczne charakterystyczne dla głosek układają się w plamy o pewnych zarysach. Zapisy uzyskane w ten sposób można odczytywać.

Terminem fonetyki akustycznej po raz pierwszy posłużył się fonetyk M. Joos w pracy „Acoustic phonetics” w 1948 roku.

2.2 Wprowadzenie w tematykę

Mowa jest podstawowym sposobem komunikacji. Zawiera informacje, które są wysyłane przez mówcę i odbierane przez słuchacza. Komunikacja ta odbywa się na kilku poziomach:

- Lingwistycznej
- Paralingwistycznej
- Extralingwitycznej

Podziału tego dokonał Laver w roku 1991.
(Źródło: Laver J. *POP*)

Warstwa lingwistyczna zawiera semantyczne informacje zakodowane w języku (zarówno gramatykę jak i fonologiczne jednostki) oraz fonetyczną reprezentację wypowiedzi. Generalizując warstwa lingwistyczna obejmuje informacje, które mamy do przekazania, to znaczy treść wypowiedzi.

Druga warstwa paralingwistyczna jest warstwą pozawerbalną i pozalingwistyczną. Zawiera informacje o aktualnym nastawieniu mówcy, jego stanie psychicznym i emocjonalnym. W przeciwieństwie do warstwy lingwistycznej nie da się jej jasno zrestrukturyzować..

Trzecia warstwa extralingwistyczna zawiera informacje pozwalające zidentyfikować mówcę takie jak: wiek, płeć, głos, oraz cechy osobnicze. Warstwa ta również zawiera informacje społeczne, kulturowe, nawykowe. Innymi słowy warstwa ta zawiera wszelkie informacje fizyczne i fizjologiczne wyróżniające daną osobę.

(Źródło: Marasek K. *EGG*)

Ilustracją tej treści jest poniższy schemat:

(Źródło: Gubrynowicz R. *PAF*)

Rysunek 2.2 Dziedziny wiedzy obejmujące komunikację werbalną

2.3 Podstawowe pojęcia

Celem lepszego zrozumienia zasady działania poszczególnych narządów mowy warto przybliżyć niektóre podstawowe pojęcia z zakresu fizyki. Należą do nich fala akustyczna, amplituda, ciśnienie akustyczne, natężenie i widmo dźwięku.

Podstawowym pojęciem jest „fala akustyczna". Fala akustyczna jest zaburzeniem rozchodzącym się w ośrodku sprężystym we wszystkich stanach skupienia materii w pełnym zakresie częstości drgań, jaki może wystąpić w przyrodzie. „Zaburzenie, o którym mowa wywołuje chwilowe zmiany gęstości ciśnienia i temperatury ośrodka. Cechą charakterystyczną fali akustycznej jest przenoszenie energii przez drgające cząstki".

(Źródło: Kleszewski Z. *PA*)

Podział fal akustycznych może być różny, rozważając częstotliwość fale dzielimy na:

- Infradźwięki
- Dźwięki słyszalne
- Ultra dźwięki
- Hiperdźwięki

Infradźwięki, czyli poddźwięki są to fale akustyczne o częstotliwości leżącej poniżej progu słyszalności, czyli poniżej 16 Hz.

Dźwięki słyszalne są to fale akustyczne o częstotliwościach z przedziału 16 Hz do 20 kHz. Zwykle częstotliwości około 16 kHz uznaje się za kres słyszalności.

Utradźwięki to fale akustyczne od 20 kHz do 1 Ghz. Górna granica jest wyznaczana przez techniczne możliwości wytwarzania fali.

Hiperdźwieki to fale o częstotliwości z zakresu od 10^9 do 10^{13} Hz. Są to częstotliwości bardzo wysokie odpowiadające częstości drgań atomów w sieci krystalicznej. Zakres częstotliwości jest przedmiotem zainteresowań fizyków zajmujących się badaniem ciał stałych i cieczy.

Innym kryterium podziału fal może być kierunek przesunięcia akustycznego w stosunku do kierunku propagacji fal.

W zależności od tego kierunku fale dzielimy na :

- Podłużne - gdy kierunek przemieszczenia jest równoległy do kierunku propagacji fali
- Poprzeczne, gdy kierunek przemieszczania jest prostopadły do kierunku propagacji fali

Fale podłużne mogą propagować w gazach cieczach i ciałach stałych. Natomiast fale poprzeczne, których propagacja powoduje zmianę kształtu ośrodka mogą propagować tylko w ośrodkach mających sprężystość postaci, czyli w ciałach stałych. Fale akustyczne w powietrzu są falami podłużnymi.

(Źródło: Kleszewski Z. *PA*)

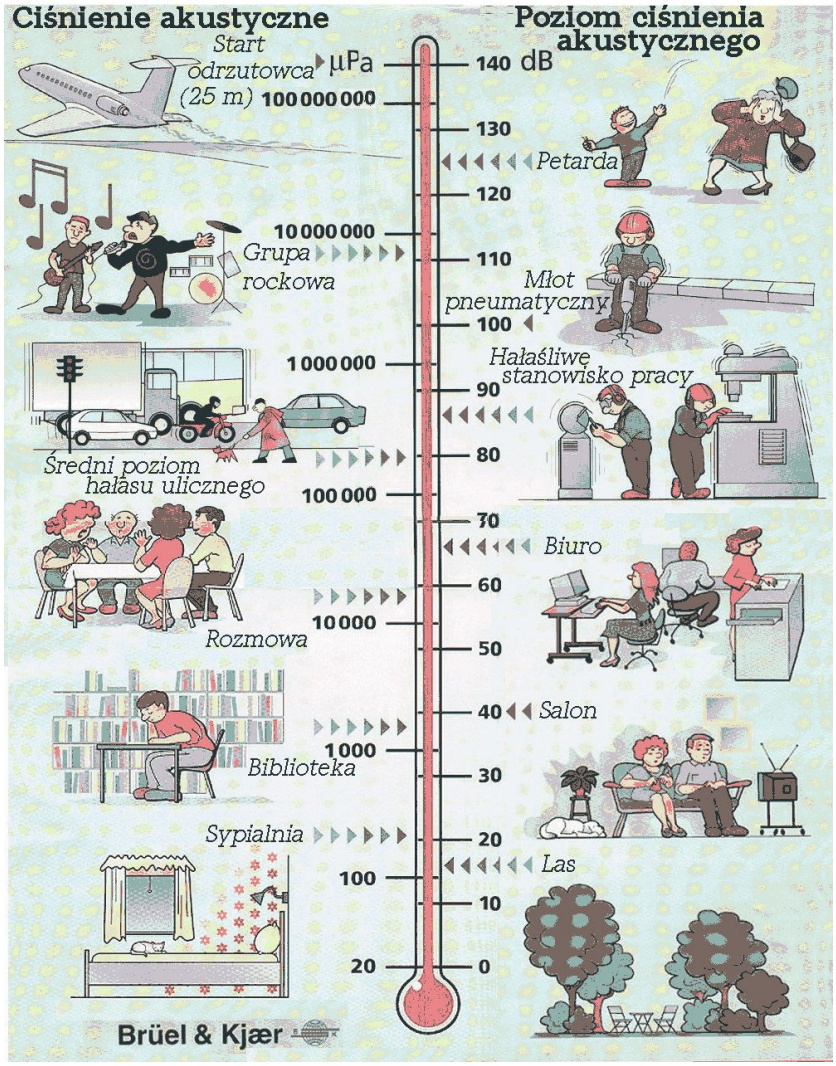
Fala akustyczna może się uginać przy opływaniu przedmiotów lub odbijać się od przedmiotów sztywnych. Wówczas mamy do czynienia z echem lub pogłosem.

Maksymalne oddalenie cząsteczki drgającej od jej położenia spoczynkowego nazywa się amplitudą.

Ciśnieniem akustycznym nazywamy różnicę pomiędzy normalnym ciśnieniem atmosferycznym[1] a ciśnieniem wytwarzającym się w każdym punkcie przestrzeni, w której rozchodzi się fala akustyczna.

[1] Ciśnienie atmosferyczne mierzy się w jednostkach zwanych barami.

Zależności między ciśnieniem akustycznym, mierzonym mikropaskalach, a jego poziomem, (decybele) obrazuje poniższy rysunek:



(Źródło Brüel i Kjaer)

Rysunek 2.3 Ciśnienie akustyczne i jego poziom

Natężenie albo moc dźwięku jest to ilość energii przepływająca w ciągu 1 sekundy przez 1 cm^2 powierzchni prostopadłej do kierunku rozchodzenia się fali głosowej.

Poziom natężenia dźwięku mierzy się w **belach**. Jednak jednostka ta jest bardzo duża, dlatego najczęściej używa się jednostki dziesięciokrotnie mniejszej – zwanej decybelem.

Poziom natężenia dźwięku podczas normalnej rozmowy wynosi około 60 decybeli (patrz rysunek 2.3), a przykładowo natężenie dźwięku startującego samolotu z odległości 100 m wynosi 130 dB – i jest to granica bólu. Niewiele mniejszy jest poziom dźwięku w przypadku zespołu hard-rockowego (w odległości 10-15 m).

Ciśnienie akustyczne jest wyrażane w tych samych jednostkach, co ciśnienie atmosferyczne, tj. w paskalach (N/m2).

Próg słyszalności wynosi 20 mPa, próg bólu 100 hPa. Normalne ciśnienie atmosferyczne wynosi około 1000 hPa. Poniższy wzór obrazuje przejście od skali liniowej do skali logarytmicznej.

$$L_{dB} = 10\log\left(\frac{p}{p_0}\right)^2 = 20\log\left(\frac{p}{p_0}\right)$$

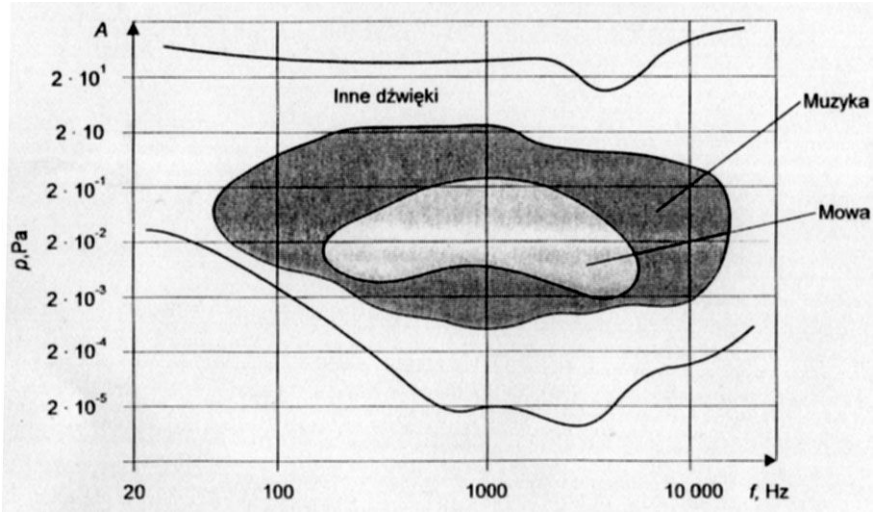
gdzie p to próg bólu, a p0 próg słyszalności

Decybele są wielkościami logarytmicznymi, dlatego nie mogą być dodawane bezpośrednio. Dwa źródła o poziomie 60 dB nie powodują wzrostu poziomu do 120 dB, lecz tylko o 3 decybele.

Zwykle jednak porównuje się amplitudy, dwukrotnie większa amplituda powoduje przyrost poziomu o 6 dB.

Widmo dźwięku zwane również wykresem spektralnym jest obrazem wartości częstości drgań i amplitudy.

Poniższy rysunek prezentuje zakresy częstotliwości wchodzące w skład mowy i muzyki oraz zakres słyszanych częstotliwości przez człowieka uwzględniających minimalny i maksymalny poziom słyszenia dźwięku.



Rysunek 2.4 Zakres częstotliwości mowy i muzyki.

(Źródło: Basztura Cz. *KSDA*)

Przedstawione powyżej zagadnienia są bardzo istotne dla opisu procesu artykulacji. Poza tym pozwolą one zrozumieć kwestie dotyczące mojej pracy praktycznej (Patrz 4.7 Problemy związane z segmentacją).

2.4 Budowa narządu mowy człowieka

Kolejnym etapem pozwalającym zrozumieć kwestie związane z procesem artykulacji jest przedstawienie budowy narządów mowy.

Narząd mowy człowieka składa się z trzech odcinków:

- Płuc wraz z tchawicą
- Krtani – odcinku fonacyjnego
- Nasady, na którą składają się jamy: gardłowa, ustna, nosowa

2.4.1 Płuca

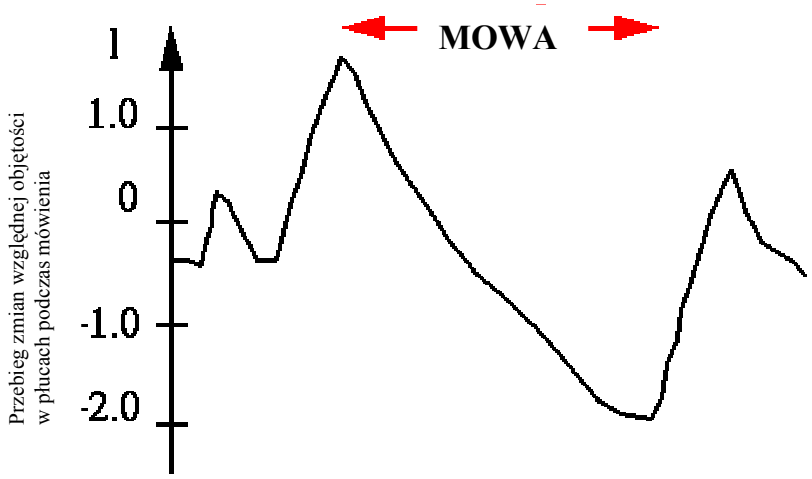
Płuca są pewnego rodzaju komorą ciśnieniową, z której wydobywa się powietrze wprawiające w drgania więzadła głosowe, co umożliwia powstawanie drgań w innych odcinkach kanału głosowego. Narząd ten mieści się w klatce piersiowej w dwu jamach opłucnowych.

Podczas wdechu powiększa się objętość jam opłucnowych, co z kolei powoduje powiększenie objętości pęcherzyków płucnych. Ciśnienie powietrza wewnątrz pęcherzyków spada i w ten sposób, poprzez napływ powietrza z zewnątrz, dochodzi do wyrównywania ciśnień.

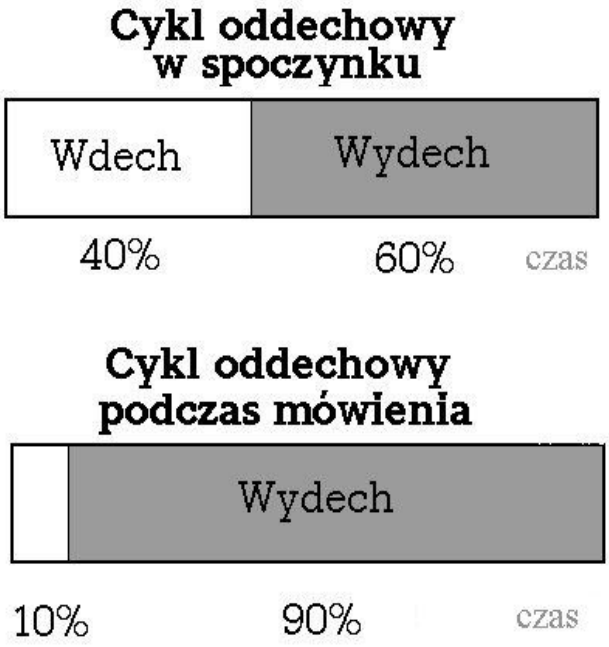
W trakcie wydechu natomiast zmniejsza się objętość jam opłucnowych, powodując zmniejszenie objętości płuc oraz wzrost ciśnienia w obrębie pęcherzyków płucnych. Powietrze ponownie na zasadzie wyrównywania ciśnień wydostaje się na zewnątrz.

Dorosły człowiek oddychając spokojnie nabiera do płuc około 0.5 litra powietrza. Podczas procesu mówienia, ilość powietrza pobieranego w czasie jednego oddechu wzrasta do około 2.5 litra. Wdech jest wtedy krótki i głęboki, wydech zaś długi i równomierny. Dorosły człowiek wykonuje w stanie spoczynku około 20 oddechów na minutę, przy czym najczęściej wdycha i wydycha powietrze przez nos.

Powyższy opis jest schematycznie przedstawiony na rysunkach 2.5 i 2.6



Rysunek 2.5 Inicjacja mowy



Rysunek 2.6 Cykl oddechowy człowieka

2.4.2 Krtań

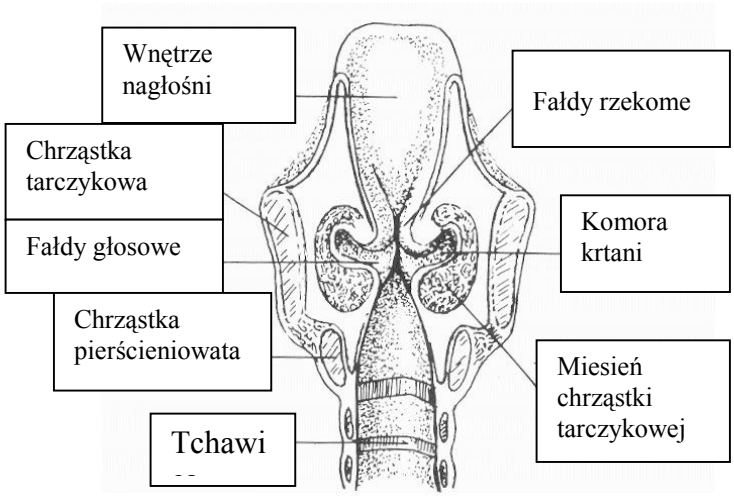
Kolejnym odcinkiem narządu mowy człowieka jest krtań. Krtań jest pewnym rodzajem puszki zbudowanej z czterech rodzajów chrząstek:

- Pierścieniowej
- Tarczowej
- Dwu chrząstek nalewkowych
- Nagłośniowej

Wnętrze krtani ma kształt rury wygiętej ku tyłowi. Wewnątrz krtani znajdują się dwie pary fałdów utworzonych przez mięśnie i wiązadła. Fałdy te leżą poziomo w poprzek krtani. Dolna para fałd nosi nazwę głosowych, fałdy górne zwane są fałdami kieszonek krtaniowych. Na brzegach fałd głosowych znajdują się więzadła głosowe.
W tyle krtani więzadła głosowe są przymocowane do wyrostków głosowych, które mogą się od siebie oddalać lub przybliżać. Jeśli są one od siebie oddalone, pomiędzy więzadłami głosowymi tworzy się otwór noszący nazwę głośni.

Zsunięte wiązadła głosowe mogą wibrować, czyli rozsuwać się i na chwilę zsuwać.
Częstotliwość wibracji dla głosu męskiego wynosi w mowie od około 80 Hz do około 160 Hz oraz od około 200 Hz do 400 Hz dla głosu kobiecego.

Wiązadła głosowe wibrują podczas wymawiania głosek dźwięcznych.
Ilustracją przytoczonej treści jest poniższy rysunek:



Rysunek 2.7 Głośnia wraz z fałdami głosowymi i tchawicą

Warto wspomnieć, że struktura anatomiczna krtani ma zasadniczy wpływ na częstotliwość drgań fałdów głosowych. Gdy masa fałdów jest mniejsza wówczas częstotliwość tonu podstawowego rośnie. Również napięcie fałdów głosowych wpływa na częstotliwość ich drgań. Przy zwiększeniu napięcia fałdów głosowych częstotliwość też ulega wzrostowi.

Żeby proces fonacji mógł się odbyć, fałdy głosowe muszą się zbliżyć na pewną krytyczną odległość. Wówczas przepływająca struga powietrza między fałdami wytwarza w szparze głośni (szpara między fałdami) podciśnienie, powodujące zbliżanie się fałdów głosowych i zamknięcie szpary głośni. W następnym cyklu parcie powietrza wychodzącego z płuc rozwiera fałdy głosowe. Mechanizm ten pojawia się cyklicznie do pierwotnego położenia (jest to tzw. efekt Bernouliego).

2.4.3 Nasada

Trzecim i ostatnim odcinkiem narządu mowy człowieka jest nasada.

„Nasada składa się z jam ponadkrtaniowych: nosowej, ustnej i gardłowej. Jamy te tworzą rozgałęziający się kanał, którego jeden człon - jama nosowa może zostać oddzielony od reszty nasady przez przywierające do tylnej jamy gardłowej podniebienie miękkie.”

(Źródło: Wierzchowska B. *OFJP*).

Jama nosowa składa się z dwóch kanałów rozgraniczonych przegrodą nosową zwaną blaszką kostną. Natomiast wąskie ujścia zewnętrzne jamy nosowej, noszą nazwę nozdrzy, zwanych również kanałami nosowymi. Kształt nozdrzy jest dość skomplikowany ze względu na występujące w nich małżowiny nosowe oraz zgrubienia kostne.

Jama nosowa przechodzi w nosową część jamy gardłowej.

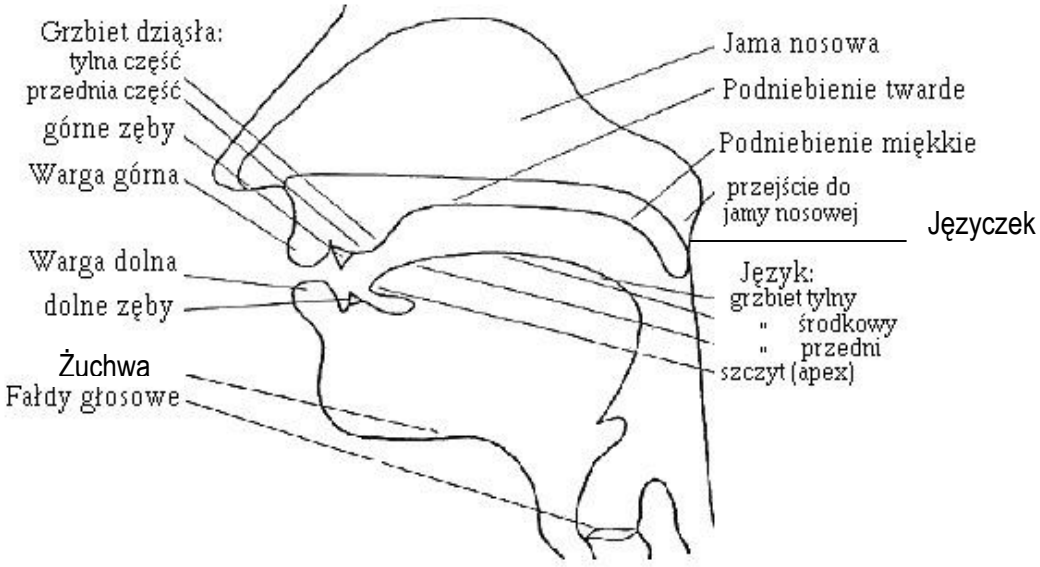
Jama ustna leży przed jamą gardłową oraz poniżej jamy nosowej. Jama ustna może przybierać różne kształty w zależności od położenia języka, ruchów warg, dolnej szczęki a także podniebienia miękkiego.

Jama gardłowa jest rurą o długości około 7 cm. Rozciąga się ona od wejścia krtani do podstawy czaszki.

W obrębie kanału utworzonego poprzez jamę ustną i gardłową znajdują się:

- Narządy ruchome:
  - Język
  - Wargi
  - Podniebienie miękkie (języczek)
  - Żuchwa
- Narządy nieruchome:
  - Zęby
  - Dziąsła
  - Podniebienie twarde
  - Tylna ścianka jamy gardłowej

Poniżej rysunek obrazuje podstawowe elementy układu artykulacyjnego.



(Źródło: Gubrynowicz R. *PAF*)

Rysunek 2.8 Podstawowe elementy układu artykulacyjnego

Przedstawienie budowy narządu mowy człowieka, pozwoli na zapoznanie się z procesem artykulacji, czyli tworzenia dźwięków przez człowieka.

2.5 Artykulacja

Poznanie budowy narządu mowy umożliwia zrozumienie jego funkcjonowania. Z kolei analiza procesu artykulacji, czyli prześledzenie drogi powstawania dźwięków, pozwoli zrozumieć omówioną w następnym rozdziale artykulacyjną syntezę mowy.

Płuca dostarczają powietrze do procesu artykulacji. "Oskrzela i tchawica prowadzą dostarczony strumień [powietrza] do krtani, w której drgające struny głosowe są źródłem dźwięku dla dźwięcznych fragmentów mowy."

(Źródło Tadeusiewicz R. *SM*)

Dźwięk ten jest następnie modulowany przez język, podniebienie zęby i wargi. Podczas modulacji ważną rolę odgrywają ruchy żuchwy i policzków. Rezonanse powstające głównie w krtani, tchawicy i jamie ustnej mają wpływ na kształtowanie dźwięku oraz widma sygnału krtaniowego.

Przepływ powietrza, wprawia w drgania struny głosowe. W ten sposób powstaje dźwięk zwany tonem podstawowym lub tonem krtaniowym. Ton podstawowy charakteryzuje się bogatym widmem.

„Wynikowe widmo określonej głoski dźwięcznej powstaje jako nałożenie charakterystyki traktu głosowego, w której poszczególne rezonanse zaznaczone są w postaci maksimów charakterystyki częstotliwościowej na widmo tonu krtaniowego. Rezultatem tego jest powstanie widma o kształcie zależnym od konfiguracji narządów mowy w chwili artykulacji danej głoski, odmienne dla każdej głoski i umożliwiające jej identyfikację.”

(Źródło: Tadeusiewicz R. *SM*)

Ton podstawowy zmienia swoją częstotliwość, co jest podstawowym czynnikiem kształtującym intonacje wypowiedzi, a zarazem formującym melodię głosu.

Zakres zmian tonu krtaniowego zależy od:

- Płci - głosy kobiece mają z reguły dwukrotnie większą częstotliwość tonu krtaniowego niż głosy męskie
- Wieku - głosy dziecięce są znacznie wyższe niż głosy osób dorosłych
- Cech osobniczych

Drgania strun głosowych powodują powstanie tonu krtaniowego. Są to drgania bierne. Oznacza to, że powietrze przetłaczane przez szparę głośni, czyli szczelinę między fałdami błony śluzowej, nazwanymi strunami głosowymi, wprawia je w drgania na skutek dynamicznego oddziaływania strumienia powietrza i elastycznych fałdów".
(Źródło Tadeusiewicz R.*SM*)

W ten sposób proces generacji drgań głosowych w krtani jest precyzyjnie kontrolowanym procesem powstawania dźwięków. Zaś, intonacja i modulacja głosu, które zależą od pracy tych mięśni pozwalają na identyfikację osoby mówiącej.
(Uzupełnieniem wiadomości na temat tonu krtaniowego jest rozdział 2.7.6 Melodia.)

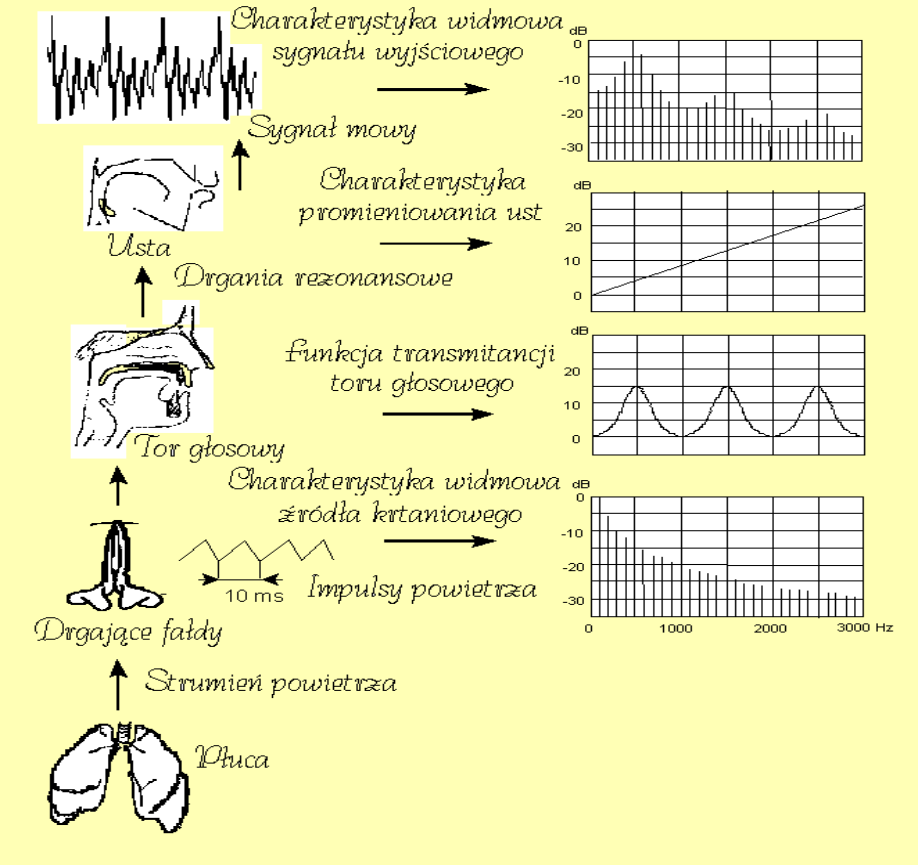
Z punktu widzenia przetwarzania sygnałów powstawanie mowy odbywa się w dwóch etapach. Pierwszym etapem jest inicjacja dźwięku, drugim zaś filtrowanie. Rozróżnienie pomiędzy tymi etapami można zrozumieć odwołując się do modelu generowania mowy zaproponowanego przez Fanta.
(Źródło: Fant G. *AToSP*)

Podstawowym założeniem tego modelu jest powstawanie sygnału źródłowego na poziomie głośni, następnie filtrowanie liniowo na poziomie toru głosowego. Efektem tego jest emitowanie na zewnątrz dźwięku, w postaci mowy. Model ten zakłada, że sygnał źródłowy i filtr są niezależne od siebie. Ostatnie badania dowiodły jednak zachodzenie pewnych interakcji między torem głosowym a źródłem głośni.

Obecnie teoria Fanta jest używana w opisie struktury ludzkiego głosu, szczególnie dla artykulacji samogłosek.

Z punktu widzenia lingwistyczno-fonetycznego generowanie mowy jest postrzegane jako proces złożony z kolejnych etapów: przygotowanie narządów mowy do procesu artykulacji, fonacja, artykulacja i organizacja procesu prozodycznego.

Poniżej zaprezentowano schemat formowania akustycznego sygnału mowy w narządzie artykulacyjnym.



(Źródło Gubrynowicz R. *PAF*)

Rysunek 2.9 Formowanie akustycznego sygnału mowy w narządzie artykulacyjnym i jego cechy widmowe - pobudzenie krtaniowe

2.6. Transkrypcja fonetyczna wypowiedzi języka polskiego

Podczas realizacji wyjścia akustycznego opanowanie transkrypcji fonetycznej wydaje się być nieodzowne. Szczególnie podczas procesu testowania posegmentowanego korpusu.

W języku polskim te same znaki ortograficzne lub jednakowe ich sekwencje mogą odpowiadać różnym dźwiękom mowy np. vur - „wór", fturnI- „wtórny". Natomiast różne znaki ortograficzne mogą odpowiadać tym samym dźwiękom mowy np. awto - „auto", daw-„dał".

Poniżej prezentuję reguły służące zamianie tekstu ortograficznego na fonetyczny. Przekształcenie to nazywamy transkrypcją fonetyczną. Kody dźwięków zapisano w kodzie SAMPA. (http://www.phon.ucl.ac.uk/home/sampa/polish.htm)

Literom samogłoskowym „y,e,a,o" odpowiadają fonemy (Patrz 3.1 Początki syntezy mowy oraz 3.12.1 Jednostki akustyczne) /I,e,a,o/. Litery „u" i „ó" nie sygnalizują różnic w wymowie.
Literę „i' przed literą spółgłoskową wymawia się jako samogłoskę /i/
Literę „i' przed samogłoską wymawia się jako:

- /j/ po zwartych, nosowej /m/, trących /f,v,x/, i głoskach /l,r/
- /i/ na końcu wyrazu
- podwójne „ii" po zwartych, nosowej /m/, trących /f,v/, głoskach /l,r/ i literze „ch" wymawia się jako /ji/

Następujące grupy spółgłoska-samogłoska /i/ odpowiadają następującym fonemom:

- „si" – /s'/
- „ci" - /ts'/
- „zi" – /z'/
- „dzi" - /dz'/
- „ni" - /n'/ (wyjątek „Dania" – /dan'ja/, ale /dan'a/)

Samogłoski nosowe „ę,ą” wymawia się jako:

- /e~,o~/ na końcu wyrazu
- /em,om/ przed /p,b/
- /en,on/ przed /t,d,ts,tS,dz,dZ/
- /en’,on’/ przed /ts’,dz’/
- /eN,oN/ przed /k,g/
- /e,o/ przed /l,w/ np. „wziąłem” – w czasie przeszłym

Głoski zwarte (/b,d,g/), zwarto-trące (/dz,dz’,dZ/) i trące (/v,z,z’,Z/) wymówione przed głoskami bezdźwięcznymi, przerwą(w wygłosie) stają się bezdźwięcznymi i ich wymowa jest dokładna, jak ich bezdźwięcznych odpowiedników, tj. /p,t,k/, /ts,ts’,tS/ czy /f,s,s’,S/. To samo występuje u zbiegu wyrazów wymówionych bez przerwy.
O ubezdźwięcznieniu lub udźwięcznieniu całej sekwencji powyższych spółgłosek o różnym typie pobudzenia decyduje w zasadzie ostatnia w sekwencji głoska – np. /lidZba/ - „liczba”, /Zat_SI/ -„rzadszy”.
Od powyższej zasady jest wyjątek, gdy przed literą „w” lub sekwencją „rz” stoi głoska bezdźwięczna. Cała sekwencja staje się bezdźwięczna. np. /kfjat/ - „kwiat”, /SfatSka/- „szwaczka”.
W języku polskim występują pewne nieregularności w wymowie „trz”, „drz”, „dż”, „dz” w obrębie wyrazu np. /tSSex/ - „trzech”, ale /tSex/ - „Czech”, /vodze/ - „wodze”, /od_zef/ - „odzew”.
Spółgłoski bezdźwięczne przed końcówką czasownikową „–my” pozostają bezdźwięczne np.. /kupmI/ - „kupmy”.
(Źródło Gubrynowicz R. *PAF*)

Wyżej wymienione zasady umożliwiają konwersję tekstu ortograficznego na fonetyczny. Podczas realizacji projektu posługiwałem się w fazie testów tym właśnie zapisem. Poniżej prezentuję podział głosek wraz z odpowiadającymi im znakami w kodzie SAMPA. (http://www.phon.ucl.ac.uk/home/sampa/polish.htm)

2.6.1 Samogłoski

System samogłosek w języku polskim składa się z 8 fonemów. Symbole ze znakiem:”~” oznaczają nazalizację[2].

Poniższa tabela przedstawia sposób reprezentacji samogłosek w transkrypcji fonetycznej.

Symbol SAMPA	Pisownia ortograficzna	Transkrypcja fonetyczna
i	pit	Pit
I	typ	TIp
e	test	Test
a	pat	Pat
o	pot	Pot
u	puk	Puk
e~	gęś	ge~s'
o~	wąs	vo~s

Tabela 2.1 Transkrypcja samogłosek w języku polskim

[2] Nazalizacja to dodatkowa nosowa artykulacja głoski; unosowienie.

2.6.2 Spółgłoski

System spółgłosek w języku polskim składa się 29 fonemów. Symbol ' oznacza palatalizację[3]. Poniższa tabela przedstawia symbole dla spółgłosek w reprezentacji fonetycznej.

Symbol SAMPA	Pisownia Ortograficzna	Transkrypcja fonetyczna
p	pik	pik
b	bit	bit
t	test	test
D	dym	dIm
k	kit	kit
g	gen	gen
f	fan	fan
v	wilk	vilk
s	syk	sIk
z	zbir	zbir
S	szyk	SIk
Z	żyto	ZIto
s'	świt	s'vit
z'	źle	z'le
x	hymn	xImn
ts	cyk	tsIk
dz	dzwon	dzvon
tS	czyn	tSIn
dZ	dżem	dZem
ts'	ćma	ts'ma
dz'	dźwig	dz'vik
m	mysz	mIS
n	nasz	naS
n'	koń	kon'
N	pęk	peNk
l	luk	luk
r	ryk	rIk
w	łyk	wIk
j	jak	Jak

Tabela 2.2 Transkrypcja samogłosek w języku polskim

[3] Palatalizacja to fonetyczne zmiękczenie spółgłoski twardej pod wpływem sąsiadującej z nią samogłoski (najczęściej przedniej).

2.7 Przykład transkrypcji fonetycznej (SAMPA)

Poniżej zamieszczam przykładowy tekst ortograficzny zapisany kodzie fonetycznym.

Konwersja tekstu na mowę otwiera nowe możliwości niedostępne w tradycyjnych systemach głosowych. Usługi katalogowe, informatory turystyczne, tematyczne serwisy informacyjne czy portale głosowe, to tylko nieliczne zastosowania tej technologii.

konversja tekstu na move~ otfjera nove moZlivos'ts'i n'edoste~pne f tradItsIjnIx sIstemax gwosovIx uswugi katalogove informatorI turIstItSne tematItSne servisI informatsIjne to tIlko n'elitSne zastosovan'a tej texnologji

(Źródło Gubrynowicz R. *PAF*)

SAMPA jest międzynarodowym sposobem zapisu głosek różnych języków świata. Istnieje również alfabet IPA, popularny wśród fonetyków, jednak nie jest używany w mojej pracy, ponieważ nie można zapisać jego symboli stosując standardowy zapis 8-bitowy ASCII.

2.8 Klasyfikacja dźwięków mowy

Wiemy, że artykulacja jest procesem generowania dźwięków mowy. Z uwagi na udział narządów biorących udział w formowaniu głosek możemy sklasyfikować artykulację dźwięków mowy.

2.8.1 Klasyfikacja akustyczna

Jest to jeden z rodzajów klasyfikacji. Drugim rodzajem klasyfikacji jest podział dźwięków mowy w zależności od charakteru składających się na nie przebiegów akustycznych. Podział ten jest podziałem akustycznym.

W podziale akustycznym wyróżnia się:

- Rezonanty
- Głoski plozywne - wybuchowe
- Frykaty
- Afrykaty
- Nosowe
- Ustne

(Źródło Wierzchowska, B.*OFJP*)

Głoski, których przebiegi akustyczne wykazują regularność lub mają przebieg tzw. quasi-periodyczny nazywa się rezonantami. Należą do nich: „a” „o” „u” „e~” „m” „n” „l „j” „v” „i” „I” „e” „o~”

Inną grupę stanowią głoski wybuchowe (plozywne). Odpowiadają im krótkie nieregularne przebiegi akustyczne. Do głosek plozywnych należą: „p” „t” „k” „g” „b” „d”.

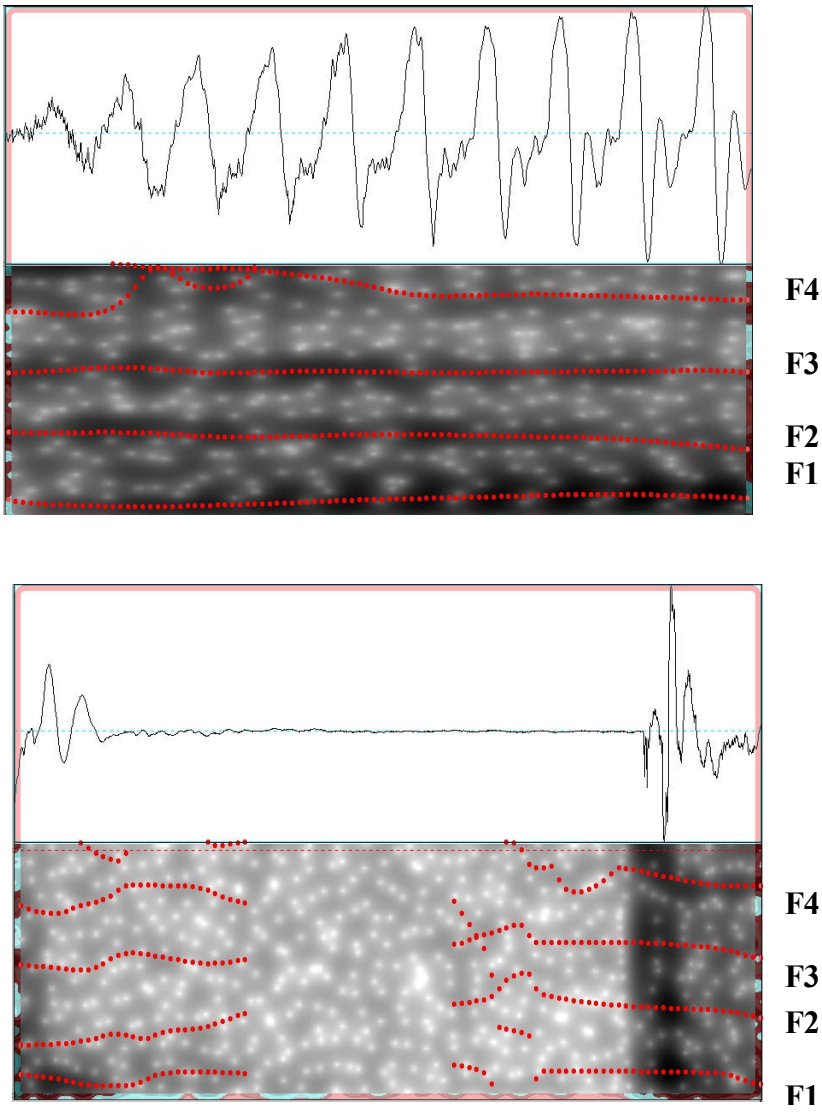
Głoski frykatywne składają się z przebiegów nieregularnych zwanych frykcjami Są to:”f” „s” „s’” „S” .

Afrykaty to głoski o przebiegu nieregularnym, których frykcje poprzedzone są słabym impulsem. Należą do nich: „ts”, „ts’” „tS” .

Również w klasyfikacji akustycznej wyróżnia się podział na głoski ustne i nosowe.

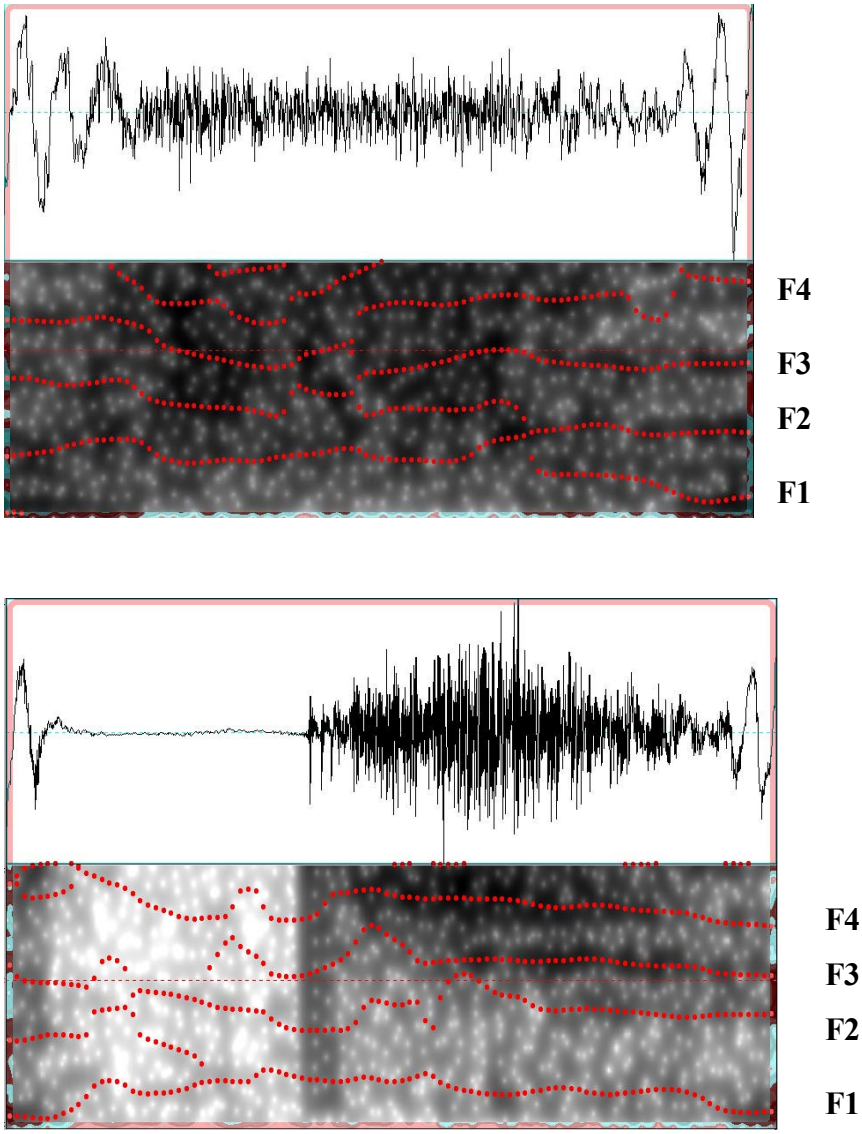
W widmie głosek nosowych można zaobserwować silne tłumienie składowych o wyższych częstotliwościach oraz antyformanty o częstotliwości około 900 Hz.

Samogłoski nosowe w języku polskim maja realizację dyftongiczną w przeciwieństwie do innych języków. Oznacza to, że otwarcie nosowe nie jest zsynchronizowane z otwarciem ustnym. Początkowo samogłoska nosowa zaczyna się od samogłoski ustnej, po której następuje płynne otwarcie kanału nosowego i przejście do artykulacji spółgłoski nosowej (n), co może okazać się kłopotliwe przy konkatenacyjnej syntezie mowy.(Patrz 4.6 Reguły w procesie segmentacji, rysunek 4.7)



Rysunek 2.10 Przykłady głoski regularnej(e) i wybuchowej (p) wraz ze spektrogramem[4] i analizą formantową (Patrz 4.5.1 Analiza formantowa). Na rysunku u góry widać charakterystyczny dla tych głosek przebieg regularny. U dołu widoczny charakterystyczny krótki i nagły impuls. Po prawej stronie każdego rysunku zaznaczono formant pierwszy (F1), drugi(F2), trzeci(F3) i czwarty(F4).

[4] Spektrogram jest zapisem przedstawiającym zmiany amplitudy w funkcji czasu dla poszczególnych częstotliwości.



Rysunek 2.11 Przykład frykaty i afrykaty wraz ze spektrogramem i analizą formantową. Na rysunku u góry głoska „S” wraz z charakterystycznym dla niej przebiegiem nieregularnym. U dołu głoska „ts’”. Afrykaty wyróżniają się występowaniem słabego impulsu poprzedzającego przebieg szumowy.

2.8.2 Klasyfikacja genetyczna - artykulacyjna

Innym rodzajem klasyfikacji jest klasyfikacja genetyczna. Polega ona na określeniu mechanizmów wytwarzania dźwięków w płaszczyźnie artykulacyjnej. Podstawowym podziałem w klasyfikacji genetycznej jest podział na spółgłoski i samogłoski.

Samogłoski to dźwięki, przy których wytwarzaniu powstaje w środkowej płaszczyźnie narządów mowy kanał bez silnych zwężeń.

Do spółgłosek zaliczamy głoski z wargowym, przedniojęzykowym, środkowojęzykowym oraz tylnojęzykowym miejscem styku artykulatorów.

Wyróżnia się również podział dźwięków ze względu na:

- zachowanie się więzadeł głosowych w czasie wytwarzania dźwięku
- stopień zbliżenia narządów mowy
- miejsce artykulacji głoski
- położenie podniebienia miękkiego
- artykulacje modyfikującą zasadniczą artykulację spółgłoski

Warto przyjrzeć się bliżej wymienionym kategoriom.

Z uwagi na zachowanie się więzadeł głosowych głoski dzielą się na dźwięczne i bezdźwięczne. Głoski dźwięczne powstają wówczas, gdy więzadła głosowe są zsunięte i wibrują. Głoski bezdźwięczne wymawiane są przy głośni rozsuniętej.

Podczas wymawiania głosek bezdźwięcznych narządy wytwarzające zwarcia stykają się na większej przestrzeni niż przy wymawianiu głosek dźwięcznych, a ruchy artykulacyjne trwają przy głoskach bezdźwięcznych nieco dłużej niż przy odpowiednich głoskach dźwięcznych.

Ze względu na stopień zbliżenia narządów mowy wyróżnia się:

- Spółgłoski zwarto-wybuchowe
- Głoski zwarto-szczelinowe
- Głoski szczelinowe
- Spółgłoski otwarte

Zwarcie narządów mowy polega na całkowitym zamknięciu kanału głosowego. Szczeliną zaś nazywamy przewężenie w określonym miejscu kanału głosowego.

Ze względu na miejsce artykulacji spółgłoski dzielimy na:

- Dwuwargowe
- Wargowo-zębowe
- Przednio-językowe zębowe
- Przedniojęzykowe-dziąsłowe
- Środkowojęzykowe
- Tylnojęzykowe-welarne

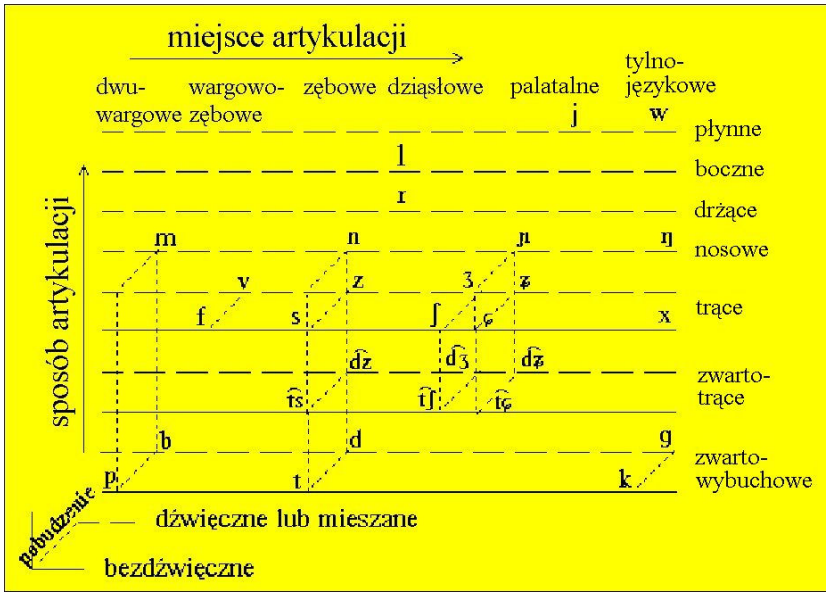
Podział ten oznacza powstawanie głosek z uwagi na lokalizację charakterystycznego dla spółgłoski zwarcia lub szczeliny. Lokalizacja zwarcia lub szczeliny ma miejsce w obrębie kanału głosowego.

Wyróżniamy również podział spółgłosek ze względu na położenie podniebienia miękkiego. Podział ten charakteryzuje głoski ustne i nosowe.

Ostatnim podziałem spółgłosek jest podział uwzględniający artykulacje dodatkowe. Zalicza się do nich:

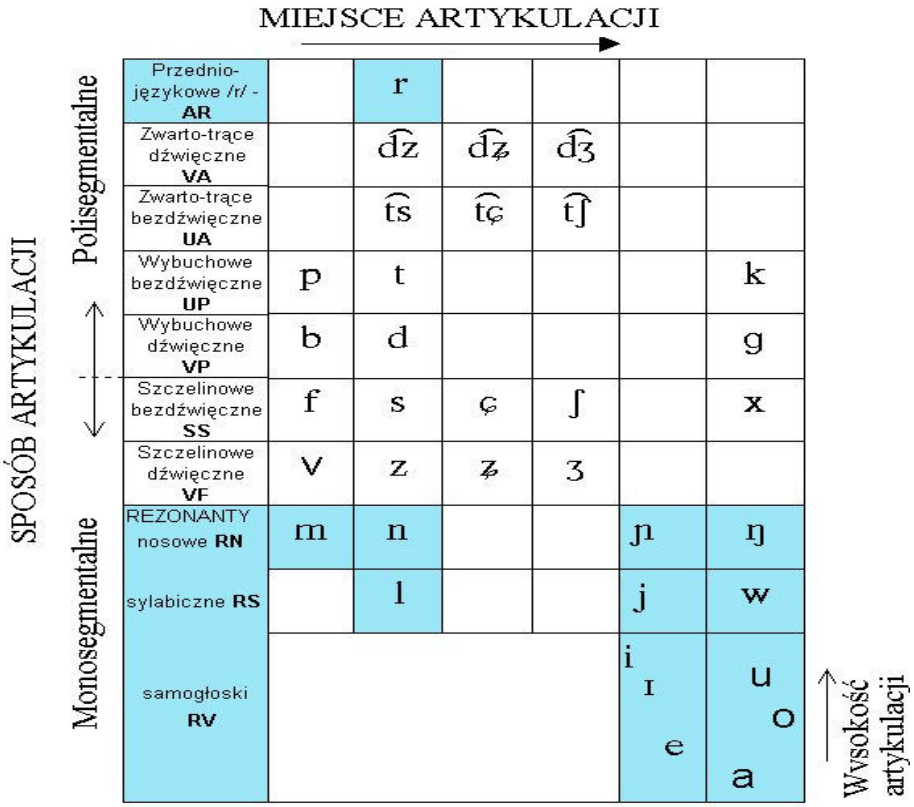
- Labializację, czyli zaokrąglenie wargowe
- Delabializację, czyli spłaszczenie warg
- Palatalizację
- Welaryzację czyli wzniesienie tylnej części języka
- Cerebralizację, czyli artykulację polegającą na wzniesieniu czubka języka i cofnięciu go

Poniższe schematy obrazują opis artykulacyjny dźwięków mowy.



(Źródło: Gubrynowicz R. *PAF*)

Rysunek 2.12 Opis artykulacyjny dźwięków mowy – spółgłoski



(Źródło IPA i Gubrynowicz R. PAF)

Rysunek 2.13 Opis fonetyczny głosek polskich

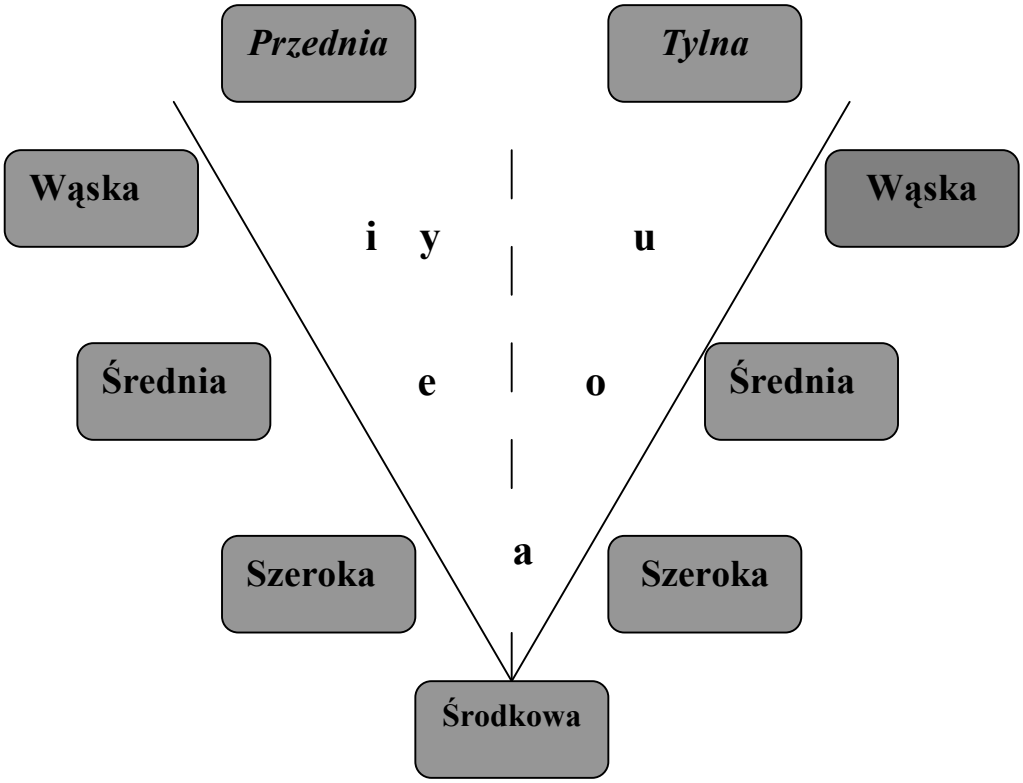
2.8.3 Klasyfikacja samogłosek

Wyżej wymienione podziały dotyczyły głównego podziału klasyfikacji artykulacyjnej spółgłosek. Teraz zajmę się systematyzacją drugiej grupy – samogłosek.

Pierwszy podział samogłosek został stworzony pod koniec XVIII wieku przez Hellwaga. Klasyfikacja ta nazywana jest trójkątem samogłoskowym.

Samogłoska „a", która znajduje w wierzchołku trójkąta jest podstawą klasyfikacji. Na jednym ramieniu trójkąta rozmieszczone zostały samogłoski szeregu przedniego na drugim samogłoski szeregu tylnego. Samogłoski szeregu środkowego umieszczone są na linii dzielącej trójkąt na połowy.

Poniżej na rysunku znajduje się schemat Hellwaga.



Rysunek 2.14 Schemat Hellwaga

Wartym uwagi jest opracowany przez Bella pod koniec XIX wieku prostokąt artykulacyjny. W prostokącie tym na linii poziomej znajdują się samogłoski zależne od poziomego ruchu języka. Na linii pionowej zaś zależne od ruchu pionowego języka.

Schemat ten obrazuje poniższa tabela.

	Przednie	*Środkowe*	*Tylne*
Wysokie	**i, y**		**u**
Średnie	**e**		**o**
Niskie		**a**	

Rysunek 2.15 Schemat Bella

Kolejnym równie interesującym podziałem jest klasyfikacja samogłosek Benniego. Jest to zmodyfikowana wersja prostokąta Bella. Dodatkowo zostało wprowadzonych pięć stopni głębokości w jamie ustnej, co pozwala różnicować samogłoski ze względu na przesunięcia języka.

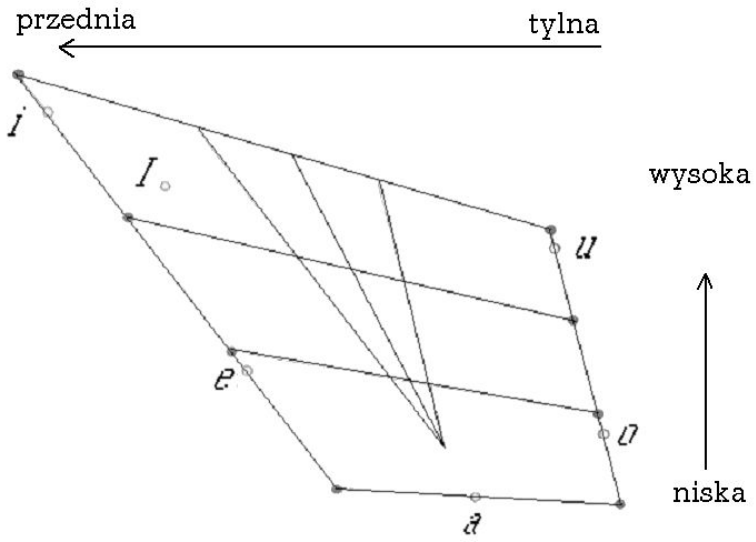
Poniższy schemat obrazuje tą klasyfikację.

Otwarcie	**Stopnie głębokości w jamie ustnej**				
	1 *przód* *2*		*3* *tył* *4*		*5*
Wąskie	***i***	***y***			***u***
Średnie		***e***		***o***	
Szerokie			***a***		

Rysunek 2.16 Schemat Benniego

Ostatnio dość często stosowanym podziałem jest czworobok samogłoskowy, opracowany przez angielskiego fonetyka D. Jonesa.

Badania rentgenograficzne pozwoliły na wyznaczenie najbardziej wzniesionych punktów grzbietu języka i przyporządkowanie im odpowiednich samogłosek. Na schemacie znajduje się czworokąt samogłoskowy:



Rysunek 2.17 Czworokąt samogłoskowy

Dopiero później powstał bardziej dokładny system klasyfikacji samogłosek, w którym bierze się pod uwagę:

- Poziome ruchy języka
- Pionowe ruchy języka
- Stopień obniżenia dolnej szczęki
- Układ warg
- Położenie podniebienia miękkiego

2.8.4 Ujednolicenie klasyfikacji dźwięków mowy

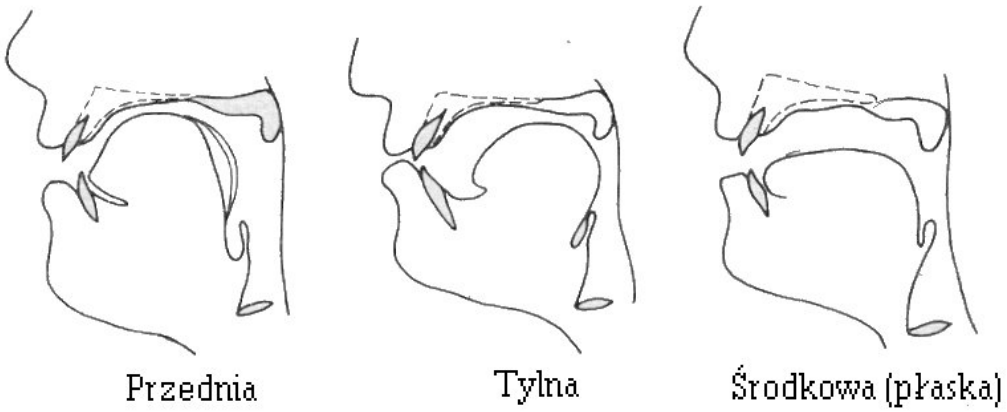
Fonetyczna klasyfikacja samogłosek jest dokonywana na podstawie innych kryteriów niż klasyfikacja spółgłosek. W przypadku samogłosek uwzględnia się położenia masy języka. Decyduje ono o kształcie kanału głosowego, rozkładzie formantów. W opisie spółgłosek bierze się pod uwagę stopień zbliżenia narządów mowy oraz miejsce powstawania dźwięków mowy.

Trzeba zdawać sobie sprawę, że tak skomplikowany podział jest niewygodny. Dlatego stosuje się podział spółgłosek i samogłosek z uwagi na układ masy języka oraz położenie drugiego formantu.

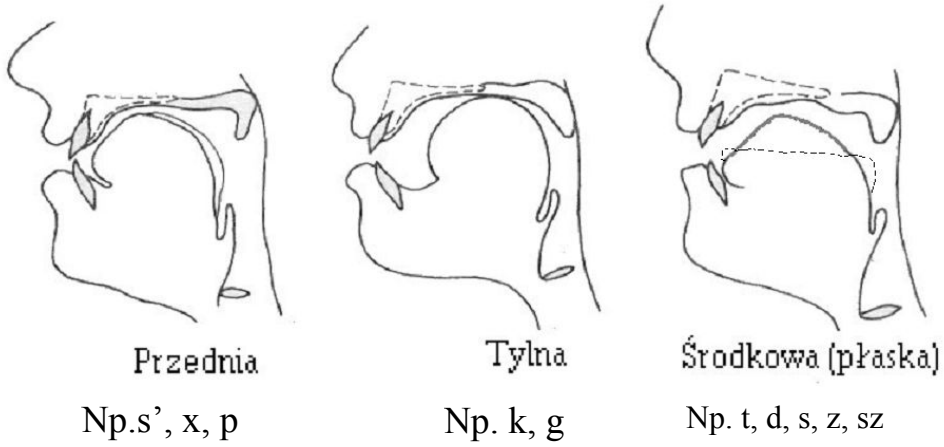
W klasyfikacji tej wyróżnia się:

- Położenie przednie języka
- Położenie tylne języka
- Położenie środkowe języka

Poniższy podział obrazują schematy:



Rysunek 2.18 Klasyfikacja samogłosek z uwagi na położenie masy języka



(Źródło: Gubrynowicz R. *PAF*)

Rysunek 2.19 Klasyfikacja spółgłosek z uwagi na położenie masy języka

Omówienie zagadnienia procesu artykulacji oraz sklasyfikowanie dźwięków mowy pozwala orientować się w cechach charakterystycznych głosek. Informacje te są bardzo potrzebne podczas tworzenia systemu syntezy mowy. W mojej pracy praktycznej wiadomości te były wykorzystane podczas procesu segmentacji.

2.9 Fonetyczna organizacja wypowiedzi

Stworzenie dobrej jakości syntezatora mowy nie jest procesem łatwym. Należy pamiętać, że dobry syntezator to taki, którego mowa będzie, płynna, zrozumiała i naturalna.

Żeby te wymagania były spełnione należy odnieść się do języka naturalnego i zdefiniować podstawowe pojęcia mówiące o organizacji wypowiedzi, czyli poruszające problemy związane z elementami mającymi wpływ na jakość mowy naturalnej. Przez język naturalny rozumiem każdy język powstały na drodze naturalnej ewolucji stworzonej przez człowieka (polski, angielski).

Zagadnienia te sprowadzają się do omówienia podstawowych problemów organizacji wypowiedzi języka naturalnego. Należą do nich: iloczas, akcent, koartykulacja, melodia. Omówienie ich pozwoli zrozumieć trudności, jakie napotyka twórca syntezy mowy.

2.9.1 Iloczas

Czas trwania wypowiedzi zależy przede wszystkim od:

- Tempa mówienia
- Długości wypowiedzi
- Sposobu artykulacji

Tempo mówienia zależy od rodzaju oraz charakteru wypowiedzi. Liczba głosek przypadających na 1 sekundę znajduje się w zakresie od 5 do 25. Przy czym dolna wartość obejmuje bardzo wolny sposób mówienia, podczas gdy górna wartość stanowi granicę zrozumiałość wypowiedzi.

Czas trwania głoski zależy również od długości wypowiedzi. Dźwięki które są wypowiadane w dłuższych frazach trwają nieco krócej niż gdy są wypowiadane w dłuższych frazach.

Czas trwania głoski związany jest również ze sposobem artykulacji. W języku polskim najdłuższe są głoski otwarte nosowe. Nieco krócej trwają głoski ustne podczas gdy spółgłoski nosowe są najkrótszymi głoskami. Bezpośrednio z czasem trwania głosek związany jest iloczas, który określa czas trwania głoski. Wyróżnia się dwa rodzaje iloczasu:

- Iloczas bezwzględny
- Iloczas względny

Iloczas bezwzględny opisuje czas trwania głoski w wypowiedzi, natomiast iloczas względny stanowi stosunek czasu trwania głosek w stosunku do innych głosek.
Generalnie przyjmuje się, że im bardziej skomplikowana artykulacja, tym czas trwania głoski jest dłuższy.

Iloczas jest pojęciem bardzo ważnym. Dotyczy on szczególnie procesu segmentacji i czasu trwania poszczególnych segmentów.

Również ważnym zagadnieniem są fazy wypowiedzi, które mają wpływ na charakterystykę głosek.

2.9.2 Fazy wypowiedzi

Podczas wypowiedzi wyróżnia się trzy fazy:

- początek czyli nagłos
- środkową część wypowiedzi czyli śródgłos
- końcową fazę wypowiedzi czyli wygłos

Nagłos wypowiedzi zazwyczaj rozpoczyna się przygotowaniem narządów mowy do artykulacji. Charakterystycznym elementem są występujące ruchy podniebienia miękkiego lub dolnej szczęki. Ruchy te można zaobserwować w przypadku wymawiania głosek zwarto-wybuchowych (p,b). Nagłos wypowiedzi zazwyczaj wymawiany jest bardzo starannie.

Dźwięki wypowiadane w śródgłosie różnią się nieco od dźwięków nagłosu i wygłosu.

Podczas wygłosu ruchy narządów artykulacyjnych są precyzyjnie i wolniejsze. Również następuje obniżenie tonu podstawowego w wyniku zwolnionej pracy więzadeł głosowych (obniżone ciśnienie podgłośniowe).

2.9.3 Koartykulacja

Podczas mowy często można zaobserwować ruchy narządów mowy podczas przechodzenia z jednej głoski do drugiej. Efekt akustyczny towarzyszący temu procesowi nazywa się przejściem tranzjentowym. Zdarza się, że podczas artykulacji głoski ruchy narządów mowy przypominają ruchy charakterystyczne dla głosek znajdujących się w sąsiedztwie. Proces ten nazywa się koartykulacją. Podczas przygotowania korpusu należało mieć na uwadze zjawisko koartykulacji. Mogło ono spowodować zniekształcenia w otrzymanym sygnale. Dlatego zrozumienie zagadnień fonetyki akustycznej jest bardzo ważne.

Bezpośrednio z zagadnieniem koartykulacji związane jest pojęcie upodobnień.

2.9.4 Upodobnienia

Proces koartykulacji z czasem doprowadził do zmian w obrębie zakresie form wyrazowych. Upodobnienia również zwane asymilacją dzieli się na :

- Upodobnienia wewnątrzwyrazowe
- Upodobnienia międzywyrazowe

Upodobnienia wewnątrzwyrazowe dzielą się na upodobnienia wsteczne i postępowe.

Upodobnienia dzieli się również pod względem miejsca artykulacji, dźwięczności oraz stopnia zbliżenia narządów mowy.

Upodobnienia pod względem miejsca artykulacji zachodzą „w takich wypadkach, kiedy zwarcia lub szczeliny właściwe sąsiadującym ze sobą głoskom, wytwarzane niegdyś w różnych miejscach kanału głosowego, są obecnie wytwarzane w tym samym miejscu. Upodobnienie to zachodzi np. w wyrazie *Pan Bóg* wymawianym *Pam Buk*."
(Źródło: Wierzchowska B.*OFJP*)

Jeżeli grupa spółgłoskowa składała się z głosek dźwięcznych i bezdźwięcznych, a dziś składa się z głosek bezdźwięcznych lub tylko dźwięcznych to mówimy o upodobnieniu pod względem dźwięczności. Dobrym przykładem jest dziś wymawiany wyraz „bapka" a kiedyś „babka".

Z upodobnieniem pod względem zbliżenia narządów mamy do czynienia gdy „w jakiejś formie zamiast głoski zwartowybuchowej zaczyna się wymawiać głoskę zwarto-szczelinową np. jak w wyrazach *dżewo, tszeba.*
(Źródło: Wierzchowska B.*OFJP*)

Upodobnienie międzywyrazowe zachodzą na pograniczach form wyrazowych. Upodobnienia te mogą zachodzić pod względem dźwięczności, miejsca artykulacji oraz stopnia zbliżenia narządów mowy.

2.9.5 Akcent

Oprócz czynników charakterystycznych dla danego języka takich jak zjawisko koartykulacji czy też połączenia dźwięków, ważnym elementem jest zróżnicowanie dynamiczne wypowiedzi. Zjawisko to określa się mianem akcentu.

Akcent jest również wyróżnieniem pewnych sylab w wyrazach bądź też w wypowiedziach. Przez akcent określa się zwiększenie donośności, zmianę wysokości tonu podstawowego lub przedłużenie czasu trwania sylaby.

W języku polskim akcentowana jest przeważnie przedostatnia sylaba (patrz 3.12.1 Jednostki akustyczne), jednak nie stanowi to 100% reguły. Istnieje wiele wyjątków dotyczących na ogół wyrazów obcego pochodzenia np. matem'atyka. W takich wyrazach akcent pada na trzecią sylabę od końca. Natomiast w wypowiedziach przez akcent określa się jedną z bardziej wyróżnionych sylab wypowiedzi. Sylaba ta jest przeważnie przedostatnią sylabą zdania bądź wypowiedzi. Akcent ten powoduje, że dany fragment wypowiedzi uzyskuje na ogół dodatkowe wzmocnienie i wydłużenie. (Patrz 3.4.1 Generowanie prozodii)

W języku polskim akcent pełni również funkcję ekspresywną, która jest odzwierciedleniem stanu psychicznego. Wyraża ona również nastawienie mówiącego do wypowiadanej treści. Czynnik ekspresywności jest bardzo silnie powiązany z przebiegiem melodii wypowiedzi.

2.9.6 Melodia

O wysokości muzycznej wypowiedzi decyduje ton podstawowy. Ton podstawowy, jak wiadomo, zależy od ilości zwarć więzadeł głosowych na sekundę. Wahania tonu podstawowego w obrębie wypowiedzi przeważnie nie przekraczają oktawy.

Wzrost wysokości tonu podstawowego przeważnie ma miejsce w sylabach akcentowanych nieco głośniej. W zdaniach oznajmujących oraz w zdaniach pytających ton dotyczy ostatniej sylaby i jest on względnie wysoki. W zdaniach wykrzyknikowych oraz rozkazujących opada w ostatnich sylabach.

W języku polskim zmiany tonu podstawowego nie powodują różnic znaczeniowych wyrazów. Przebieg zmian zależności tonu podstawowego nosi nazwę melodii zasadniczej.

Wyróżnia się cztery podstawowe rodzaje melodii:

- Rosnąca niska
- Rosnąca wysoka
- Opadająca niska
- Opadająca wysoka
- Równa niska
- Równa wysoka

Melodie opadająca niska i równa niska są charakterystyczne dla zdań oznajmujących. Melodia wysoka równa i wysoka rosnąca jest charakterystyczna dla zdań złożonych, dla drugiej części wypowiedzi. Melodia rosnąca niska występuje w zdaniach pytających.

Charakterystyka melodii jest ściśle powiązana z modelowaniem prozodii w systemach syntezy mowy. Zagadnienie to pozwala uzyskać głos zbliżony do naturalnego. Więcej wiadomości dotyczących tego zagadnienie znajduje się w rozdziale trzecim.

2.10 Podsumowanie

Podstawowym zadaniem tego rozdziału było wprowadzanie czytelnika w zagadnienia związane z fonetyką akustyczną obrazującą sposób powstawania dźwięków u człowieka. Przedstawiłem historię fonetyki, budowę narządu człowieka oraz klasyfikację dźwięków przez niego artykułowanych. W dalszej części opisuję zagadnienia dotyczące organizacji wypowiedzi oraz transkrypcji fonetycznej. Są to podstawy z którymi należy się zapoznać podczas tworzenia systemu syntezy mowy. Mają one niezwykle duże znaczenie dla procesu segmentacji.

Informacje przedstawione w następnym rozdziale ułatwią znalezienie rozwiązań służących uzyskaniu najlepszej jakości syntezy mowy.

3. Synteza mowy

3.1 Początki syntezy mowy

Synteza mowy jest procesem generowania mowy ludzkiej w sposób sztuczny. Im bardziej jest ona naturalna i płynna tym bardziej jest doskonała. Celem nowoczesnych projektów jest zapewnienie takiej jakości syntezy, by słuchający nie był w stanie odróżnić mowy syntetyzowanej od naturalnej mowy. Takie są dzisiejsze dążenia, zobaczmy jednak, jakie były zamysły i próby pierwszych fonetyków.

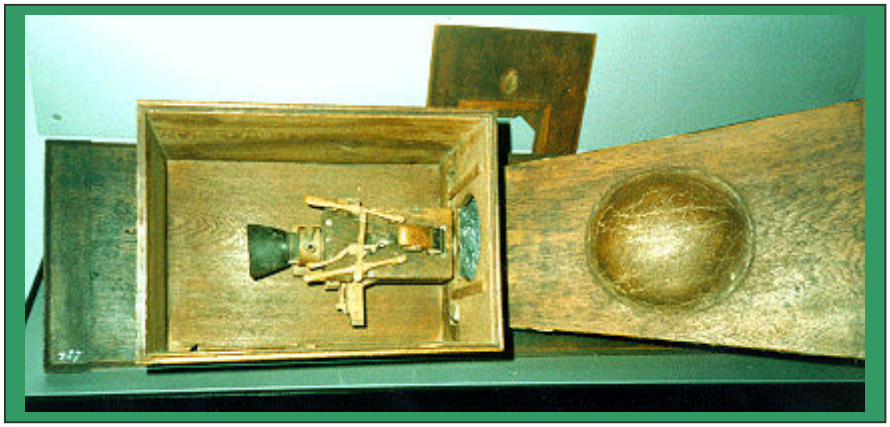
Fundamentalną próbą stworzenia ludzkiej mowy był eksperyment profesora fizjologii Ch. G. Kratzensteina. Profesor próbował wyjaśnić różnice w barwie dźwięków „a”„o” ”u” „i” „e”. W 1773 skonstruował piszczałki zbliżone do organowych, potrafiące syntezować te dźwięki.

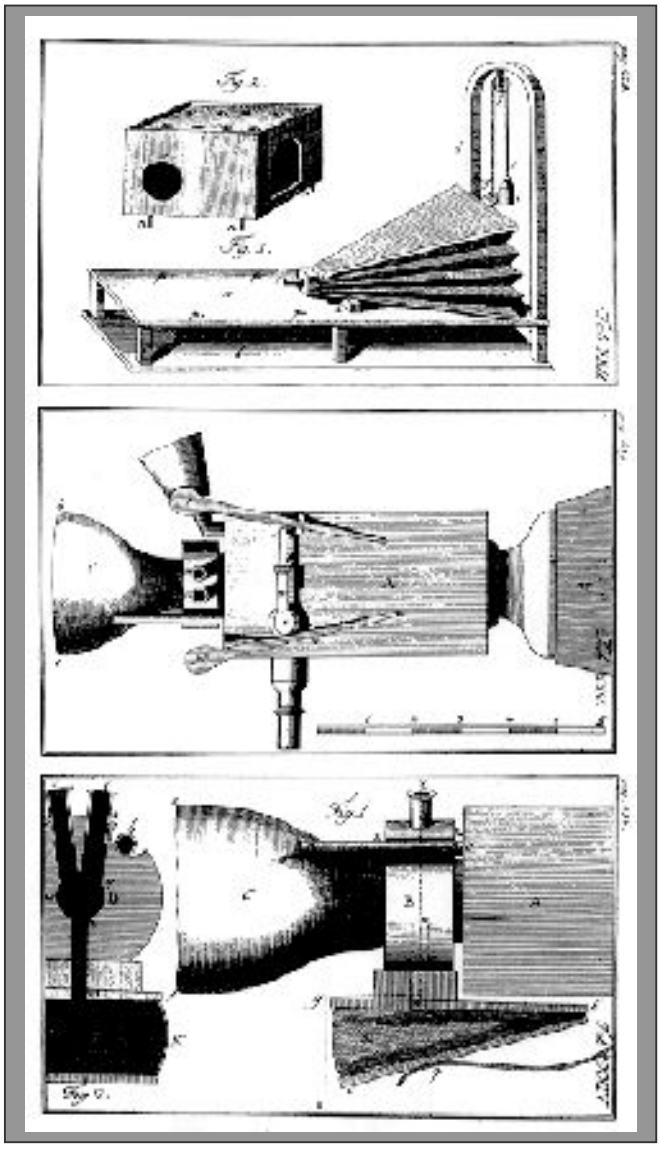
W tym samym czasie Wolfgang von Kempelen zaczął konstruować własną, mówiącą maszynę. Model von Kempelena składał się z miechów odpowiadających płucom, dziurek zamiast nosa oraz systemu pomocniczych mechanizmów. Maszyna von Kempelena umożliwiała kontrolę tonacji a powstający w wyniku jej działania głos – brzmiał wyraźnie i dostatecznie głośno, jak głos dziecka lub dorosłego człowieka. Maszyna von Kempelena umożliwiała generowanie nie tylko słów, ale i krótkich zdań.

W swojej książce „Mechanismus der menschlichen Sprache. Beschreibung einer sprechender Maschine” (Mechanizm ludzkiego języka. Opis mówiącej maszyny)
von Kempelen umieścił opis mówiącej maszyny. Opisał również podstawowe zasady działania narządów mowy. Jednak największym osiągnięciem autora było określenie roli narządów ponadkrtaniowych w procesie generowania dźwięku.

Wspomnę jeszcze, że do dziś von Kempelen jest uważany za pierwszego fonetyka.

Poniżej znajdują się rysunki z mechanizmem von Kempelena.





Rysunek 3.1 Syntezator Von Kempelena (od góry: zrekonstruowany model von Kempelena widok z góry, niżej po lewej schemat budowy, po prawej zrekonstruowany model widok z przodu)

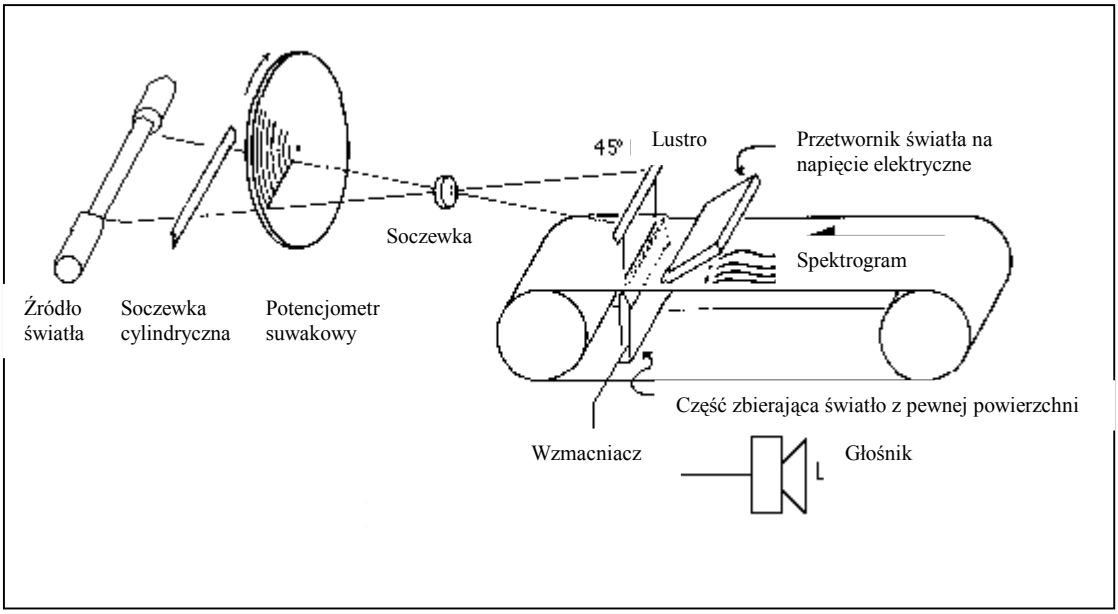
W XIX wieku nie odnotowano dużego postępu w dziedzinie syntezy mowy. Zostało skonstruowanych kilka maszyn syntetyzujących ludzki głos, jednak nie przyczyniły się one do znacznego rozwoju syntezy mowy.

W 1835 roku została stworzona maszyna przez Josepha Fabera. Maszyna ta zawierała sztuczny język oraz jamę gardłową i umożliwiała generowanie melodii w formie śpiewania. Wynalazek Fabera był obsługiwany przy pomocy klawiatury i pedałów. W roku 1846 w Londynie „Euphonia" – taką nosiła nazwę – maszyna „zaśpiewała" „ God Save the Queen".

W 1936 roku powstała VODER. Była to pierwsza maszyna, która wykorzystywała elektryczność. Urządzenie skonstruowane przez Homera Dudleya posiadało jedną dużą wadę. Do poprawnego działania wymagany był długi czas treningu. Urządzenie to zostało zaprezentowane publiczności w 1939 roku podczas „World Fair" (Światowe Targi).

W latach 50-tych dwudziestego wieku powstał mechanizm do syntezy mowy, opierający się na zupełnie innej technologii. Urządzenie działało jak odwrotny spektrogram. Lampa kierowała promieniście strumień promieni na obracający się dysk, na którym znajdowało się 50 koncentrycznie ułożonych ścieżek z podstawową częstotliwością 120 Hz. Światło padające na spektrogram odpowiednio zaczerniało ścieżki. Stopień zaczernienia odpowiadał mocy sygnału. Opisane urządzenie generowało monotonną mowę.

Poniżej znajduje się ilustracja tego wynalazku:



Rysunek 3.2 Urządzenie oparte na zasadzie działanie odwrotnej do spektrogramu

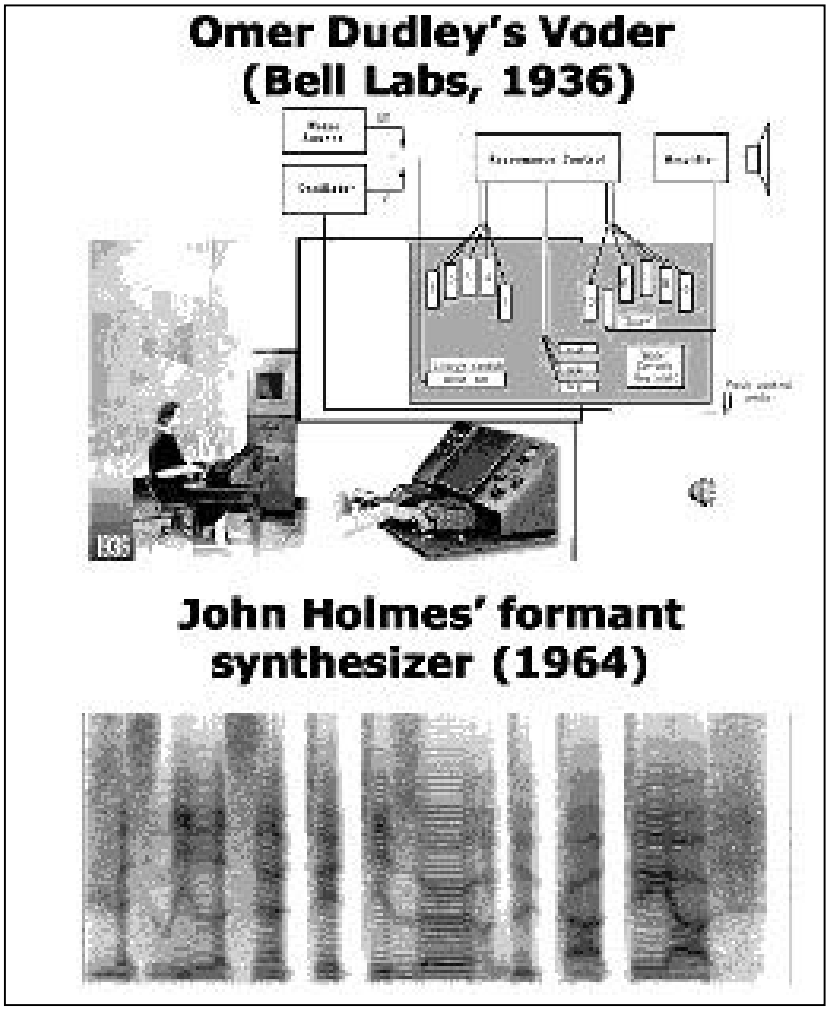
Na początku lat siedemdziesiątych zaczęły powstawać pierwsze komputerowe syntezatory mowy ludzkiej. Wczesne systemy wykorzystywały fonemy, których konkatenacja umożliwiła uzyskanie ciągłości mowy. Jednak fonem jest jednostką akustyczną, pozbawioną tranzjentu – czyli przejścia pomiędzy jednym a drugim elementem akustycznym, co powoduje, że uzyskana synteza będzie zawsze nienaturalna i nieciągła. Dlatego w latach późniejszych zaczęła rozwijać się technologia difonowej syntezy mowy[5].

Kolejnym ważnym osiągnięciem było stworzenie formantowego syntezatora mowy w 1964 roku przez Johna Holmesa. Działanie tego modelu opiera się wykorzystaniu odpowiednich filtrów. Na wejściu filtru podawany jest sygnał elektryczny będący tonem harmonicznym. Sygnał ten występuje w mowie w częściach akcentowanych lub w szumie. Filtr pełni rolę rezonatora toru głosowego.

[5] Zagadnienie to zostanie omówione w dalszej części pracy

W formantowej syntezie mowy rozwinęły się dwie technologie. Pierwsza polega na wykorzystaniu rezonansu w celu generowania formantów poszczególnych głosek. Druga opiera się na symulowaniu artykulacji za pomocą większej liczby połączeń (filtrów). Każde z nich odpowiada za generowanie sygnału krótkiej sekcji toru głosowego.

Poniżej prezentuję opisane schematy syntezatorów z początku XX wieku.



Rysunek 3.3 Pierwsze syntezatory z początku XX wieku (od góry syntezator Homera Dudleya oraz poniżej pierwszy formatowy syntezator mowy)

Trzeba zaznaczyć, iż pierwsze projekty, związane z syntezą mowy nie były traktowane poważnie, ponieważ głównym ich zastosowaniem była rozrywka. Z czasem jednak poszerzono ich zastosowanie, jak i wiedzę na temat syntezy mowy.

Von Kempelen jako pierwszy zbadał mechanizm powstawania ludzkiej mowy natomiast celem projektu Homera Dudleya było stworzenie urządzenia ograniczającego pasma częstotliwości (tzw. Voice Coder) potrzebnego do realizacji rozmowy telefonicznej przez jedną linię.

3.2 Konwersja tekstu na mowę

Moduł konwersji tekstu na mowę (Text-to-speech system – TTS system) odpowiada za translację tekstu wprowadzonego do komputera albo przez operatora komputera albo bądź też przez system Optical Character Recognition (OCR). Zadaniem tego modułu jest przeczytanie każdego w jakiejkolwiek formie wprowadzonego tekstu. Znacznie prostszym jest Voice Response System, czyli system, który generuje jedynie słowa, czy też pojedyncze frazy z jakiejś dziedziny (np. informuje pasażerów o odjazdach pociągów). Skala możliwości systemu TTS jest znacznie większa, gdyż generuje pełen zakres słów.

Jak się można domyślać, nie jest możliwe stworzenie i nagranie wszystkich form i wszystkich słów, dla danego języka. Stąd też system TTS definiuje się jako system automatycznego generowania mowy z transkrypcją fonetyczną oraz modułami odpowiedzialnymi za prozodie i intonacje. Wydaje się, że taki system nie jest trudny do zrealizowania i wyuczenia.

Warto przytoczyć tu bardziej zrozumiałą analogię, odnoszącą się do czynności czytania. Człowiek z łatwością porusza się w świecie informacji pisanych i czytanie nie sprawia mu problemu. Ale jeśli tylko wróci pamięcią do czasów nabywania zdolności czytelniczych, przypomni sobie jak wielką trudnością było opanowanie tej sztuki na samym początku.

Musimy pamiętać, że komputer jest maszyną, która nie nauczy się niczego, o ile nie zostanie zaprogramowana w odpowiedni sposób. Myślę, że nie będzie wielkim błędem, stwierdzenie, że tak jak nauczenie czytania – zaprogramowanie tego zagadnienia na komputerze nie będzie łatwe. Komputer jest tylko mechanizmem pozbawionym inteligencji, a człowiek musi się zmierzyć z rozwiązaniem każdego problemu w sposób algorytmiczny. Dlatego liczba dobrze działających i wykorzystywanych systemów syntezy mowy jest niewielka. Skoro tak trudnym zadaniem jest zaprogramowanie i stworzenie nowego systemu syntezy mowy, warto sobie zadać pytanie, dlaczego tyle uwagi się jej poświęca. Otóż musimy pamiętać o ogromnych możliwościach i licznych zastosowaniach tej dziedziny multimediów.

Synteza mowy może być bardzo pożytecznym narzędziem przede wszystkim w edukacji jako pomoc w nauce języków obcych. Jednak taki system nie został zaimplementowany do dziś, a wynika to z jakości mowy, jaką realizują poszczególne systemy syntezy mowy.

Ostatnio coraz bardziej powszechne stają się wirtualne uniwersytety. Oferują naukę via Internet, tzw. e-learning. Więcej wiadomości na ten temat zawarłem na końcu tego rozdziału. Omówiłem tam podstawowe zalety systemów syntezy mowy. Można zauważyć, że wymienione zagadnienia mobilizują i zarazem są dużym wyzwaniem dla informatyków. Pierwsze próby takich systemów są uwieńczone sukcesem.

3.3 Budowa systemu TTS

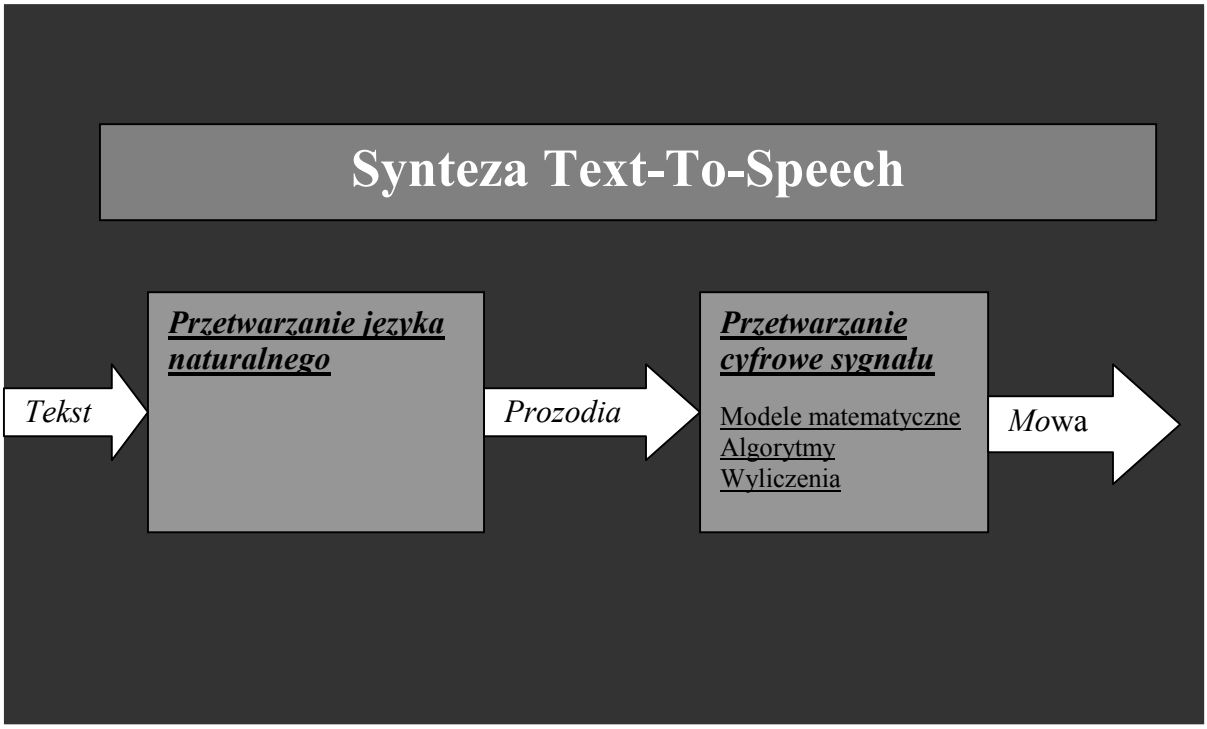
Wprowadzenie do omówienia systemu TTS rozpocząłem od przedstawienia jego funkcji. Teraz zajmę się opisem jego budowy.

System TTS składa się z dwóch podstawowych elementów:

Pierwszym elementem jest moduł NLP (Natural Language Processing), który jest odpowiedzialny za przetwarzanie języka naturalnego.

Drugi element stanowi moduł przetwarzania cyfrowego sygnału - DSP (Digital Signal Processing).

Poniżej umieściłem schemat funkcjonalny systemu konwersji tekstu na mowę. Na schemacie widać dwa odrębne moduły wchodzące w skład systemu TTS.



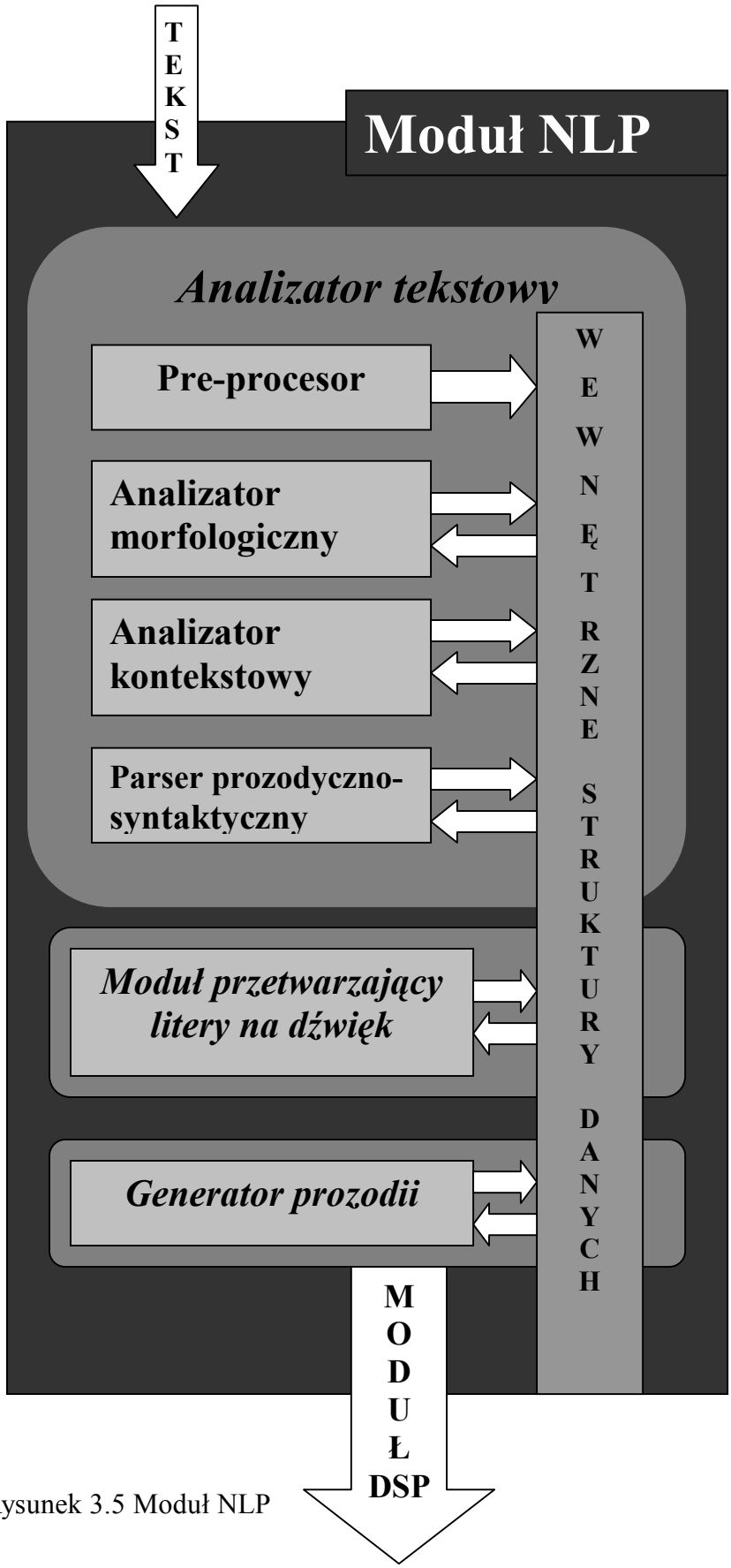
Rysunek 3.4 Ogólny schemat systemu TTS..

Zanim przejdę do szczegółowego opisu poszczególnych modułów, przedstawię ich funkcje. Dokładny opis pozwoli zrozumieć zasadę działania systemu konwersji tekstu na mowę oraz zobrazuje nakład pracy, jaki należy włożyć w celu stworzenia pełnego systemu syntezy mowy.

Celem NLP jest przekształcenie tekstu na zapis fonetyczny[6]. Moduł NLP jest również odpowiedzialny za wygenerowanie odpowiedniej intonacji i prozodii[7] tekstu.
Z uwagi na niejednoznaczności języka naturalnego oraz skomplikowane zależności międzywyrazowe stworzenie całego modułu NLP jest zadaniem trudnym.
Dodatkowe problemy związane są z uzyskaniem naturalnej prozodii, która w dużej mierze zależy od składni, ale ma również wiele wspólnego z semantyką i pragmatyką. Obecnie jednak, z powodu trudności znalezienia jednoznacznej kategorii przynależności słowa do kategorii semantycznej, systemy TTS skupiają się w głównej mierze na składni. Prowadzone są badania nad semantyką i pragmatyką, jednak dotychczasowe rezultaty nie są jeszcze wystarczające do praktycznej implementacji w systemach TTS.

[6] Zapis fonetyczny jest zapisem słów wypowiedzi, przy użyciu podziału fonetycznego głosek.

7 Termin prozodia odnosi się do pewnych właściwości sygnału mowy, które można usłyszeć w postaci zmiany głośności, długości sylab, sposobie intonacji. Patrz 3.4.1Generowanie prozodii.



Rysunek 3.5 Moduł NLP

3.4 Moduł NLP

Moduł NLP składa się z następujących elementów:

- Pre-procesor: (normalizator tekstu), jego zadaniem jest podział zdań na wyrazy. Proces podziału jest dość skomplikowany, z uwagi na dużą liczbę skrótów występujących w polskim języku. Moduł ten wydziela z tekstu skróty, liczby, idiomy, akronimy i rozwija je do pełnego tekstu. Pewnym problemem jest rozpoznawanie końca zdania. Zauważmy, że często po skrótach stawiany jest znak kropki, co nie zawsze oznacza koniec zdania. Po przetworzeniu dane są przechowywane w wewnętrznym module struktur danych.
- Analizator morfologiczny – jest odpowiedzialny za wyznaczenie części mowy dla każdego ze słów (rzeczownik, przymiotnik). Słowa te są rozbijane na morfemy, poprzez zastosowanie gramatyk regularnych, wykorzystanie słownika, tematu wyrazów i afiksów, (czyli przedrostków i przyrostków). Zadania analizatora morfologicznego sprowadzają się do zmniejszenia słownika oraz ustalenia części mowy.

  Innymi słowy zadaniem analizatora morfologicznego jest zorganizowanie znormalizowanych danych z preprocesora. Dane te to słowa, z przypisaną im w danym fragmencie wypowiedzi funkcją syntaktyczną.
- Analizator kontekstowy – zadaniem analizatora kontekstowego jest ograniczenie znaczenia poszczególnych słów. Ograniczenie to odbywa się na podstawie zbadania kontekstu słów (części mowy) znajdujących się w sąsiedztwie. Stosuje się tutaj metodę n-gramów, która opisuje syntaktyczne zależności pomiędzy słowami, na zasadzie badania prawdopodobieństw w skończonych przejściach automatu. Służą do tego modele Markova lub wielowarstwowe sieci perceptronowe. Użycie sieci neuronowych sprowadza się do odkrycia reguł rządzących kontekstem zdaniowym. Stosuje się również metody lokalnych niestochastycznych gramatyk.

- Parser syntaktyczno-prozodyczny jest odpowiedzialny za utworzenie prozodii i intonacji dla poszczególnych sekwencji fonemów. Parser ten bada jednocześnie pozostałe wyrażenia, które nie zostały zakwalifikowane do żadnej z kategorii. Następnie stara się znaleźć podobne do nich struktury tekstowe, których elementy prozodyczne będą najbardziej prawdopodobne i zbliżone do siebie.
- Moduł *letter to sound* – moduł jest odpowiedzialny za konwersję głosek na mowę oraz za utworzenie transkrypcji fonetycznej dla istniejących słów.

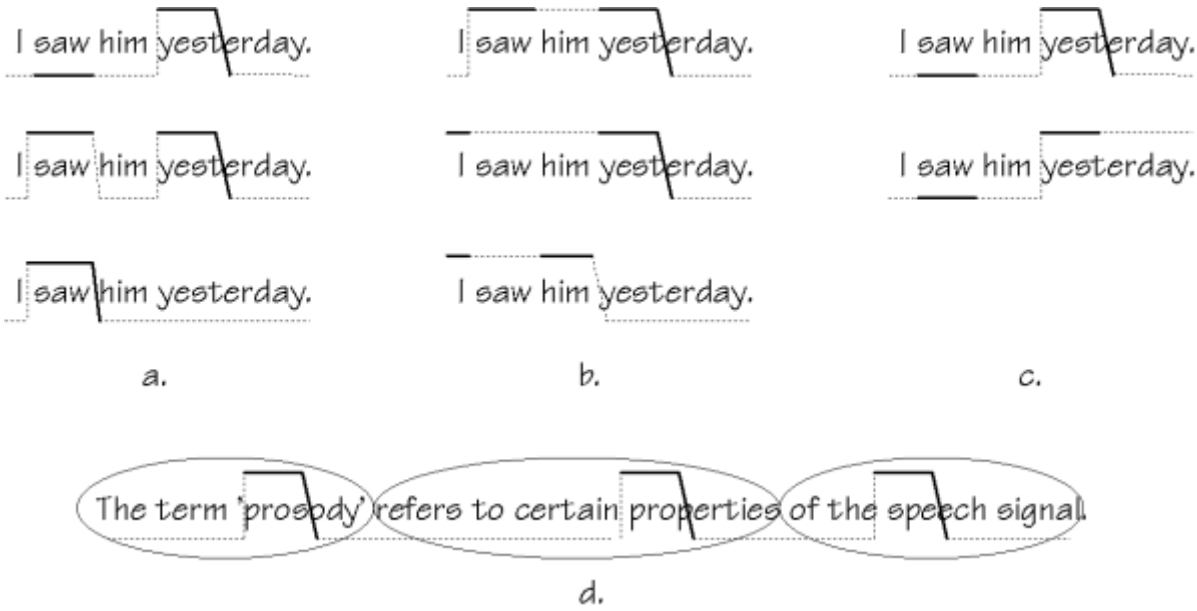
Powstaje jednak kilka problemów, dość istotnych z punktu widzenia realizacji tego modułu. Mianowicie: słownik wymowy obejmuje tylko podstawowe słowa, bez morfologicznych kombinacji to znaczy nie uwzględniający rodzaju, przypadku, liczby.

Istnieje wiele słów o podwójnym znaczeniu i takiej samej pisowni. Pojawia się problem homografów - słów o różnej wymowie i takiej samej pisowni w zależności od części mowy, jaką reprezentują. Dodatkowo trudno sobie wyobrazić istnienie wszystkich słów w słowniku.

3.4.1 Generowanie prozodii

Termin prozodia odnosi się do pewnych właściwości sygnału mowy, które można usłyszeć w postaci zmiany głośności, długości sylab, sposobie intonacji.
Cechy prozodyczne odgrywają duże znaczenie w komunikacji językowej. Odpowiednie zaakcentowanie sylaby zmienia znaczenie całej wypowiedzi.

Poniżej znajduje się rysunek obrazujący różnice w znaczeniu wypowiedzi.



Rysunek 3.6 Prozodia angielskiego zdania „I saw him yesterday"- widziałem go wczoraj.

W zależności od intonacji, w każdym przypadku różne cechy są wyszczególnione.
W przykładzie "a" koncentrujemy się na podawanej informacji
W przykładzie „b" na związku pomiędzy słowami widziałem wczoraj (I saw yesterday)i widziałem jego (I saw him)
W przykładzie „c" wyrażamy koniec lub kontynuację czynności.
W przykładzie „d" mamy segmentację zdania na poszczególne sylaby.

Kiedy model syntaktyczno-prozodyczny jest przygotowany, można przystąpić do precyzyjnego określenia czasu trwania poszczególnych fonemów oraz intonacji, w postaci wygenerowania sygnału. Projektowanie modelu prozodycznego nie jest zadaniem łatwym, jednak niezbędnym dla każdego systemu mowy. Powstała mowa będzie zrozumiała, jednakże bez zaprogramowania odpowiednich cech emocjonalnych, będzie brzmiała sztucznie a przecież dobry system syntezy mowy odznacza się dużą naturalnością głosu.

Model NLP realizuje szereg procesów związanych z przekształceniem tekstu oraz ukształtowaniem gotowej wypowiedzi wraz z intonacją. Kolejnym etapem jest przekształcenie tych danych na sygnał w modelu DSP.

3.5 Moduł DSP

Z procesem artykulacji ludzkiej związana jest praca mięśni twarzy oraz wytwarzanie sygnału o odpowiedniej częstotliwości. W języku komputerowym, za sztuczne wygenerowanie sygnału na kształt procesu artykulacji ludzkiej odpowiada moduł DSP.

Symulacja artykulacji jest możliwa na dwa sposoby:

Pierwszy zwany formantową syntezą mowy polega na jej generowaniu za pomocą parametrów.

Drugi zwany konkatenacyjną syntezą mowy polega na łączeniu jednostek akustycznych.

W tym rozdziale opiszę obie filozofie syntezy mowy oraz wskażę, jakie są między nimi różnice i co z tego wynika. W ten sposób łatwiej będzie zrozumieć tendencje na drodze realizacji projektów związanych z syntezą mowy.

3.6 Systemy syntezy mowy polskiej

Istnieje kilka systemów syntezy mowy polskiej.

Są to :

- MBROLA
- FESTIVAL
- RealSpeak firmy Lernout&Hauspie (obecnie ScanSoft)
- Elan
- Syntalk, Spiker

3.7 MBROLA

MBROLA jest międzynarodowym projektem, który został zapoczątkowany przez Tierriego Dutoit w 1995 roku w Belgii. Projekt ten jest realizacją syntezy mowy z tekstu fonetycznego na mowę.
Skompilowany kod źródłowy jest dostępny na 21 platform sprzętowych.

Projekt ma na celu propagowanie syntezy mowy oraz wzbudzenie zainteresowania tą dziedziną szczególnie w kręgach akademickich. Założeniem MBROL-i jest utworzenie systemu syntezy mowy obejmującego jak najwięcej języków świata.

System ten jest udostępniany nieodpłatnie pod warunkiem, że nie jest wykorzystywany do celów komercyjnych i militarnych.

Żeby uruchomić MBROL-ę potrzebna jest posegmentowana difonowa baza danych. Po zakończeniu segmentacji proces przygotowania oraz normalizacji bazy jest wykonywany przez zespół MBROL-i za darmo, pod warunkiem niewykorzystywania komercyjnego oraz wojskowego.(CD)

MBROLA „nie potrafi” czytać tekstu, do tego potrzebny jest moduł „Euler”. Moduł ten odpowiada za prozodię oraz pozwala zamienia tekst fonetyczny na tekst ortograficzny. Euler powstał w 1999 roku. Obecnie napisany jest na trzy platformy (Linux, Windows, Mac).

Obecnie na liście baz difonów znajduje się 30 różnych języków świata. Oto one:

- American English (us1 female voice)
- American English (us2 male voice)
- American English (us3 male voice)
- Arabic (ar1 male voice)
- Arabic (ar2 male voice)
- Brazilian Portuguese (br1 male voice)
- Breton (bz1 female voice)
- British English (en1 male voice)
- Croatian (cr1 male voice)
- Czech (cz1 female voice)
- Czech (cz2 male voice)
- Dutch (nl2 male voice)
- Dutch (nl3 female voice)
- Estonian (ee1 male voice)
- French(fr1 male voice)
- French(fr2 female voice)
- French(fr3 male voice)
- French(fr4 female voice)
- French(fr5 male voice)
- French(fr6 male voice)
- French(fr7 male voice)
- German (de1 female voice)
- German (de2 male voice)
- German (de3 female voice)
- Greek (gr1 male voice)
- Greek (gr2 male voice)
- Korean (hn1 (hanmal) male voice)
- Hebrew (hb1 male voice)
- Indonesian (id1 male voice)
- Hindi (in1 male voice)

- Hindi (in2 female voice)
- Italian (it1 male voice)
- Italian (it2 female voice)
- Italian (it3 male voice)
- Italian (it4 female voice)
- Japanese (jp1 male voice)
- Japanese (jp2 female voice)
- Portuguese (European) (pt1 female voice)
- Romanian (ro1 male voice)
- Spanish (es1 male voice)
- Spanish (es2 male voice)
- Spanish Mexican (mx1 male voice)
- Swedish (sw1 male voice)
- Swedish (sw2 female voice)
- Telugu (tl1 female voice)
- Turkish (tr1 male voice)
- Turkish (tr2 female voice)
- i od dziś dnia język Polski (pl1 female voice)

Generalizując wymienione projekty MBROLA, EULER, są ważną próbą stworzenia wielojęzykowej syntezy mowy. Obecnie systemy te są rozwijane, lecz jeszcze wiele brakuje do ich ukończenia. Niemniej jednak, co roku przybywa około 10-ciu nowych baz difonów, z czego połowa powstaje w nowych językach. Projekt „Euler" jest ciągle w fazie rozwoju.

3.8 Festival

Festival jest wielojęzycznym systemem syntezy mowy. System ten został stworzony na Uniwersytecie w Edynburgu w Centrum Rozwoju Technologii Mowy (The Centre for Speech Technology Research).

System Festival jest udostępniany za darmo.

System ten został stworzony przez:

Alana W Blacka (CMU)
Roba Clarka (CSTR)
Richarda Caleya (CSTR)
Paula Taylora

System został napisany w języku C++ i generalnie jest przeznaczony dla środowiska Unix. Umożliwia on syntetyzowanie mowy w różnych językach, w tym mowy polskiej w oparciu o konkatenacyjną syntezę mowy z wykorzystaniem difonów. Algorytm syntezowania mowy opiera się na analizie LPC oraz PSOL-i. Ważną cechą Festival-a jest możliwość współpracy z systemem MBROL-i.

Moduły umożliwiające syntezę mowy polskiej zostały stworzone przez Dominikę Oliver na Uniwersytecie w Edynburgu.

Podczas testów bazy difonów oraz generowania przykładów, które są umieszczone na dysku CD, wykorzystałem możliwość współpracy tych systemów i podłączyłem bazę difonów do Festivala. W ten sposób uzyskałem dodatkowo automatycznie generowany czas trwania poszczególnych głosek oraz elementów prozodycznych.

Festival jest cały czas udoskonalany. Prowadzi się prace nad wdrożeniem i realizacją nowych koncepcji.

Środki finansowe przeznaczone na rozwój systemu pochodzą z różnych źródeł głównym sponsorem rozwoju systemu Festival są: The Engineering and Physical Science Research Council (EPSRC grant GR/K54229), Sun Microsystems, AT&T Labs -- Research, and BT Labs.

Wyżej wymienione systemy były systemami nie komercyjnymi. Do komercyjnych systemów TTS umożliwiających generowanie mowy polskiej należą:

- SynTalk
- System RealSpeak firmy Lernout&Haupie[8]
- Elan

3.9 SynTalk

SynTalk jest pierwszym komercyjnym system konkatenacyjnej syntezy mowy polskiej zbudowanym w Polsce.

System ten używa fonemów jako jednostek akustycznych. Ma wbudowany pełny leksykon wymowy. SynTalk został stworzony przez firmę NeuroSoft.

Obecnie działa na dwóch platformach sprzętowych:

- Linux
- Windows

3.10 System RealSpeak firmy Lernout&Hauspie

System RealSpeak jest pełnym systemem TTS oferującym konwersję tekstu na mowę z użyciem zarówno męskich jak i żeńskich głosów. Wysoka zrozumiałość oraz naturalność generowanej mowy świadczy o licznych zastosowaniach. (Patrz 3.16 Zastosowania systemów syntezy mowy).

Postępujący rozwój technologii systemu RealSpeak sprawia, że jest to między innymi jeden z najlepszych systemów syntezy mowy w Europie, generujących syntetyczną mowę języka polskiego.

[8] Obecnie firma Lernout&Hauspie została przejęta przez firmę ScanSoft, która posiada prawa własności systemu RealSpeak .

Zostało to osiągnięte dzięki dobremu oszacowaniu funkcji kosztu konkatenacji oraz dużej bazie jednostek akustycznych (około 200 MB). Synteza mowy w tym systemie opiera się na wykorzystaniu metody korpusowej.

System ScanSoft RealSpeak™ jest dostępny dla 19 języków:

- Brazilian and European Portuguese
- Cantonese and Mandarin Chinese
- Castilian & Mexican Spanish
- Danish
- Dutch & Belgian Dutch
- French
- German
- Italian
- Japanese
- Korean
- Norwegian
- Polish
- Swedish
- UK & US English

System ScanSoft RealSpeak™ jest dostępny na następujące platformy sprzętowe:

- Linux® Red Hat
- Sun Solaris™
- IBM® AIX
- Microsoft® Windows® 95/98/Me/2000, Microsoft® Windows NT® and Microsoft® Windows® CE

(CD)

3.11 Elan

Elan jest komercyjnym systemem syntezy mowy opierającym się na konkatenacyjnych technikach łączenia difonów. Użycie difonów pozwoliło uzyskać bardzo dobrą jakość syntezy mowy przy zastosowaniu małego nakładu obliczeniowego. System ten również wykorzystuje technikę konkatenacji non-uniform, ma to na celu zapewnienie najlepszego współczynnika pomiędzy jakością oraz nakładami obliczeniowymi.

Elan rozwinął bardzo dobrze własną technikę kodowania i kompresji sygnału. Technika ta umożliwia generowanie wysokiej jakości bardzo naturalnie brzmiącej mowy.

Obecnie dostępnych jest 11 języków oraz 20 modeli głosowych.
Są to:

- US English
- UK English
- Brazilian Portuguese
- French
- German
- Polish
- Spanish
- Latin American Spanish
- Russian
- Italian
- Dutch

Elan jest firmą preferującą długoterminowe badania dotyczące syntezy mowy. Obecnie prowadzi prace badawcze nad inteligentnymi systemami TTS. Mowa tutaj o generowaniu emocji w systemach TTS oraz avatarach.(CD)

3.12 Rodzaje syntezy mowy

3.12.1 Jednostki akustyczne

Jak już wyżej wspomniałem dwa modele syntezy mowy są oparte na jednostkach akustycznych. Są to: metoda korpusowa i konkatenacyjna synteza mowa. Chciałbym rozwinąć ten temat, ponieważ od doboru odpowiednich jednostek zależy jakość generowanej mowy i nakład pracy, jaki trzeba włożyć by uzyskać satysfakcjonujące efekty.

Wyróżniamy następujące jednostki akustyczne:

- Głoski (fonemy)
- Difony
- Sekwencje fonemów
- Półsylaby
- Sylaby

Difon – zaczyna się w drugiej połowie fonu i kończy w pierwszej połowie następnego fonu. Toteż dużą zaletą difonu jest przejście tranzjentowe pomiędzy dwoma fonemami. Różnica między difonem a fonem jest więc taka, że czas trwania difonu jest dłuższy i jego granice łatwiej znaleźć niż w przypadku fonemu. Łączenie difonów w słowa następuje na części stabilnej jednostki, co wpływa na korzystne brzmienie. Dużą zaletą konkatenacyjnej syntezy mowy z zastosowaniem difonów jest mały nakład pamięci potrzebny do przeprowadzenia odpowiednich obliczeń.

W swojej pracy praktycznej użyłem tej jednostki, ponieważ właśnie difon przy poprawnym przeprowadzeniu procesu segmentacji, daje dobrą jakość syntezy mowy, zdecydowanie lepszą niż fon. Wygenerowanie difonów jest niezwykle czasochłonne i wymaga dużego nakładu pracy, jednak jest to możliwe, czego najlepszym dowodem jest samodzielnie opracowana część praktyczna mojej pracy.

Sekwencje fonemów są dowolnymi jednak dopuszczalnymi w obrębie danego języka. Podstawową sekwencją fonemów jest sylaba.

Sylaba jest fonetyczno-fonologiczną jednostką słowa jak i jednym z bardziej spornych zagadnień w fonetyce. Według L.Roudeta sylaba jest odcinkiem mowy, na którego ośrodek przypadają: minimum ciśnienia powietrza w tchawicy, maksimum otwarcia narządów mowy oraz maksimum donośności [9]. Na krańcach zaś – odwrotnie: maksimum ciśnienia powietrza w tchawicy, maksimum zbliżenia narządów mowy oraz minimum donośności.

Na sylabę nie wpływa sąsiedztwo głosek w otoczeniu, których się znajduje. Segmentacja sylab jest względnie łatwa, jednak wymaga ponad 150000 sylab, (w języku japońskim około kilkuset) celem uzyskania optymalnych podstaw dla syntezy mowy – co dla jednego magistranta wydaje się nie wykonalne ☺.

Poniżej prezentuję tabelę porównującą jakość możliwej do uzyskania syntezy mowy w zależności od użytej jednostki akustycznej.

ELEMENT	*LICZBA*	*OPIS*	*TRANZJENT*	*JAKOŚĆ SYNTEZY MOWY*
Fon	40-60	Jednostka mowy	Nie	Słaba
Sekwencja fonemów	Około 450	Ciąg spółgłosek lub samogłosek	Częściowy	Słaba
Difon	1500-3000	Fragment z przejściem tranzjentowym od połowy jednego fonemu do połowy drugiego	Tak	Dobra
Sylaba	Około 150000	Fonetyczno-fonologiczna jednostka mowy	Tak	Bardzo dobra

Tabela 3.1 Porównanie akustycznych jednostek mowy i jakości syntezy mowy przez nie generowanych

[9] Donośność dźwięku jest wielkością wrażenia słuchowego, odbieranego przy słuchaniu tego dźwięku. (Źródło: Wierzchowska B. 1967)

3.13 Wymagania

Wybór jednostki akustycznej do syntezy mowy ma miejsce na samym początku projektu. Drugim ważnym faktem jest uzmysłowienie sobie, jakie elementy muszą być spełnione by projekt zakończył się sukcesem. Generalnie, poniższe wnioski dotyczą przeprowadzenia nagrań z określonymi jednostkami akustycznymi.

Chcąc otrzymać dobrą jakość syntezy mowy należy pamiętać o spełnieniu poniższych warunków:

- Podczas przeprowadzania nagrania, każdy wyraz z daną jednostką syntezy mowy musi być wymawiany z jednakową głośnością.
- Jednostki mowy powinny być wymawiane w sposób monotonny
- Artykulacja musi być naturalna i bardzo czysta
- Prędkość mówienia powinna być umiarkowana
- Korpus warto nagrać podczas jednej sesji
- Rekomendowane jest nagrywanie sygnału wysokiej jakości np. częstotliwość próbkowania conajmniej 20 kHz oraz 16 bit rozdzielczości.

3.14 Metody syntezy mowy

Przedstawienie zagadnienia jednostek akustycznych służyło jako wstęp do zrozumienia syntezy mowy, teraz przejdę do konkretnych rodzajów syntezy mowy.

Opisane wcześniej jednostki akustyczne są tak naprawdę ściśle powiązane z dwoma rodzajami syntezy mowy: konkatenacyjną oraz korpusową.

Generalizując, istnieją cztery rodzaje syntezy mowy:

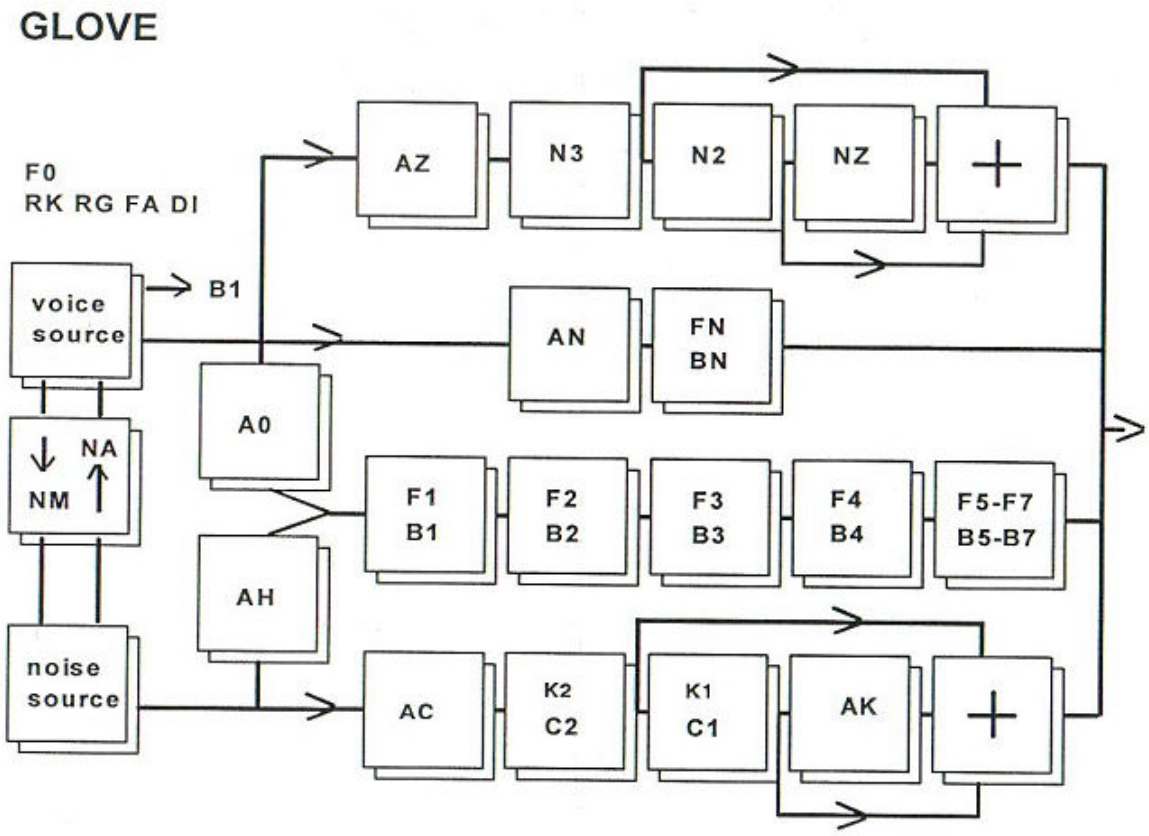
- Formantowa
- Artykulacyjna
- Konkatenacyjna
- Korpusowa

3.14.1 Formantowa synteza mowy

Formantowa synteza mowy generuje najgorszą jakość mowy, ponieważ nie pozwalają na to możliwości formatowego syntezatora mowy. Model tego syntezatora sprowadza się do zaprojektowania odpowiednich filtrów cyfrowych generujących dźwięk o charakterystycznych dla głosek częstotliwościach.

Na przykład samogłoskę możemy wygenerować przepuszczając sygnał przez odpowiedni filtr, który generuje odpowiedniej częstotliwości sygnał. Sygnał ten odzwierciedla charakterystyczne formanty głoski. Generowanie odpowiednich głosek odbywa się wedle pewnych istniejących już reguł, np. autorstwa Dennisa Klatta. Omawiana synteza nazywana jest też syntezą „by rule".

Poniżej, na rysunku, znajduje się charakterystyczny układ filtrów, czyli formantowy syntezator mowy.



(Źródło: Gubrynowicz R. *PAF*)

Rysunek 3.7 Formantowy syntezator mowy według Dennisa Klatta.

W rozdziale „Początki syntezy mowy" opisałem zasadę działania formantowego syntezatora mowy. Na rysunku powyżej kolejne „pudełeczka" z napisami są ciągami filtrów modyfikujących sygnał. Wprowadzenie oznaczeń ułatwi zrozumienie działania tego syntezatora.

- F0 oznacza częstotliwość podstawową wymowy
- AV oznacza poziom natężenia generowanego dźwięku
- F1 jest pierwszym formantem mieszczącym się w przedziale 200-1300 Hz
- F2 jest drugim formantem mieszczącym się w przedziale 550-3000 Hz
- F3 jest trzecim formantem mieszczącym się w przedziale 1200- 4999 Hz
- F4 jest czwartym formantem mieszczącym się w przedziale 1200- 4999 Hz
- F5 jest piątym formantem mieszczącym się w przedziale 1200- 4999 Hz
- F6 jest szóstym formantem mieszczącym się w przedziale 1200- 4999 Hz

- B1 oznacza pasmo rezonansu pierwszego formantu w przedziale 40-1000 Hz
- B2 oznacza pasmo rezonansu drugiego formantu w przedziale 40-1000 Hz
- B3 oznacza pasmo rezonansu trzeciego formantu w przedziale 40-1000 Hz
- B4 oznacza pasmo rezonansu czwartego formantu w przedziale 40-1000 Hz
- B5 oznacza pasmo rezonansu piątego formantu w przedziale 40-1000 Hz
- B6 oznacza pasmo rezonansu szóstego formantu w przedziale 40-1000 Hz
- FN oznacza częstotliwość nosowego mieszczącego się w zakresie 248-528 Hz
- BN częstotliwość nosowego zero w przedziale 40-1000 Hz

3.14.2 Artykulacyjna synteza mowy

Innym rodzajem syntezy mowy, opartym również na generowaniu mowy za pomocą reguł jest model artykulacyjny. Do modelowania głoski służy około 60 parametrów.

Model artykulacyjny schematem przypomina budowę ludzkiego toru głosowego, przy czym jego odpowiednikiem nie jest aplikacja, a analog elektromagnetyczny. Obecnie z uwagi na skomplikowaną budowę oraz liczne problemy związane z analogiem elektromagnetycznym synteza artykulacyjna ma znaczenie symboliczne i nie jest rozpowszechniona.

3.14.3 Konkatenacyjna synteza mowy

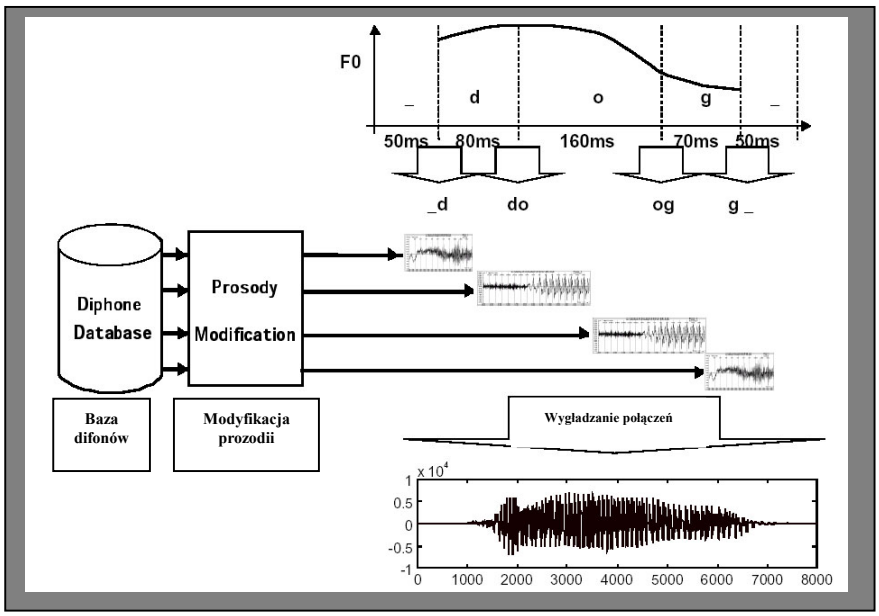
Obecnie najbardziej rozpowszechnioną metodą jest konkatenacyjna synteza mowy. Model tej syntezy mowy, rozwijany od lat 70, zyskał dużą popularność z uwagi na możliwość generowania bardzo naturalnej, dobrze brzmiącej i zrozumiałej mowy w prosty sposób.

Pierwsze syntezatory generowały mowę słabej jakości, gdyż nie brzmiała naturalnie i nie była zbytnio zrozumiała. Postęp w dziedzinie technologii umożliwił uzyskanie lepszych efektów. Synteza mowy konkatenacyjnej generuje mowę poprzez sklejanie ze sobą elementów akustycznych powstałych z naturalnej mowy (fony, difony, trifony, sylaby). Dużą zaletą tego rodzaju syntezy jest niewielki rozmiar bazy danych, z uwagi na mała objętość jednostek akustycznych. Im mniejszy rozmiar bazy, tym szybciej będzie syntetyzowana mowa oraz wymagania sprzętowe będą mniejsze.

Jest oczywiste, że konkatenacja mowy oparta na słowach jest bardzo niepraktyczna z powodu ilości wyrazów, jakie należy rozpatrzyć. Poza tym nagrywanie korpusu słów nie do końca ma sens, ponieważ brakuje tu przejścia naturalnego pomiędzy jednym a drugim słowem. Konkatenacja sylab daje dość dobre rezultaty, jednak z uwagi na ich ilość (np. w języku angielskim, – 160000 podczas gdy jest tylko 40 fonemów) też wydaje się być nie najlepszym rozwiązaniem. Bardzo często używana jest konkatenacja difonów, która umożliwia dobrą jakość syntezy mowy przy wykorzystaniu korpusu zawierającego około 1500 jednostek. Wydaje się to być zadaniem wartym realizacji.
Konkatenacyjna syntezy mowy posiada również swoje wady. Należą do nich:

- Problem wyboru jednostek akustycznych,
- Konkatenacja jednostek nagranych w różnych kontekstach.
- Modyfikacja prozodii, czyli problem intonacji i czasu trwania.
- Problem kompresji nagranych segmentów.

Dziś syntezatory mowy konkatenacyjnej generują bardzo wysokiej jakości mowę. Dlatego stała się ona zainteresowaniem takich aplikacji jak serwisy telefoniczne, edukacja komputerowa czy też mówiące zabawki. (CD)
Poniżej na rysunku prezentuje proces tworzenia słowa w konkatenacyjnej syntezie mowy.



Rysunek 3.8 Konkatenacja słowa „Dog”

3.14.4 Metoda korpusowa

Stosunkowo nowym rozwiązaniem jest metoda korpusowa (*unit selection*). Jest to zmodyfikowana postać konkatenacyjnej syntezy mowy. Wyjaśnię to na przykładzie korpusu difonów.

W moim korpusie każdy z difonów był reprezentowany tylko w jeden sposób. Natomiast metoda korpusowa zakłada, że korpus jest dużo większy, tak, że zawiera po kilka instancji danego difonu. W korpusie mogą występować również inne jednostki akustyczne np. sylaby i trifony oraz całe wyrazy.

W korpusie takim jeden ten sam difon może wystąpić 10 lub nawet 100 razy. W celu wygenerowania mowy obliczana jest funkcja kosztu. Funkcja ta polega na obliczeniu które połączenie z wszystkich możliwych pozwoli uzyskać najlepszą jakość mowy.

Na przykład system może wygenerować następujące zdanie: ”Ala ma kota”. Zdanie to jest generowane tylko za pomocą gotowych wyrazów, ponieważ według funkcji kosztu, w ten sposób powstanie najbardziej naturalne sformułowanie.

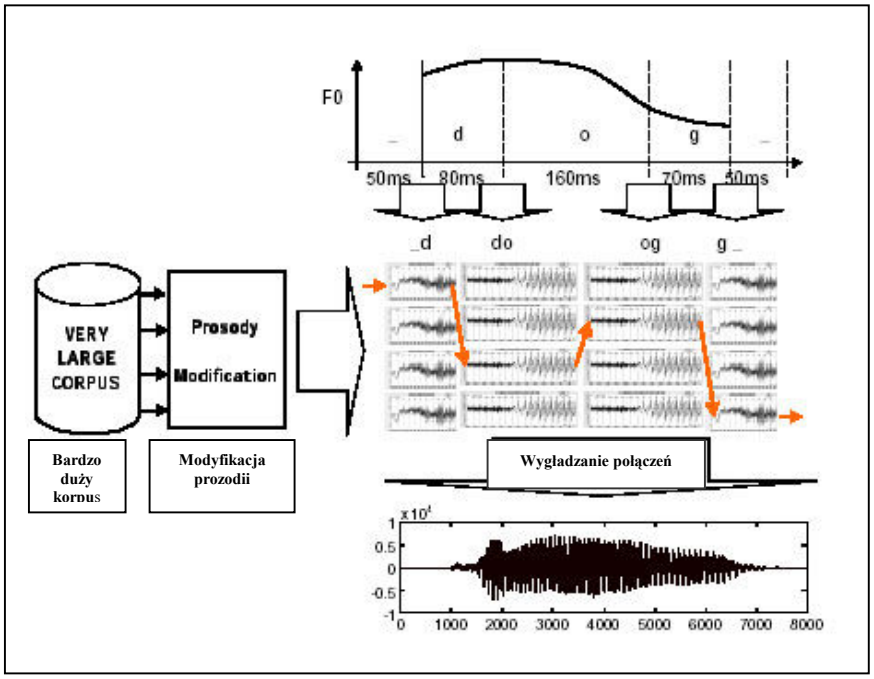
Natomiast pojawia się pewien problem, kiedy chcemy wygenerować zdanie: „Ala ma dużo kotów”. Okazuje, że w korpusie nie istnieje wyraz „kotów”. Ale istnieją odpowiednie difony i trifony za pomocą, których można wygenerować ten wyraz. Zadaniem funkcji kosztu jest wyliczenie, w jaki sposób należy utworzyć wyraz i jakich użyć jednostek akustycznych by brzmiał on najbardziej naturalnie.

Podsumowując, funkcja kosztu jest funkcją oszacowującą. Jej działanie sprowadza się do wyliczenia różnych możliwych sposobów wygenerowania danej wypowiedzi, przy użyciu różnych jednostek akustycznych znajdujących się w korpusie. Funkcja oszacowuje i porównuje zarazem, która wypowiedź będzie brzmiała najlepiej. Funkcja uwzględnia różne czasy trwania poszczególnych segmentów oraz ich intonację. Obecnie tym rodzajem syntezy mowy z uwagi na rozmiar korpusu, koszty przygotowania i oszacowania funkcji zajmują się tylko duże firmy. Między innymi AT&T, SpeechWorks oraz ScanSoft. (CD)

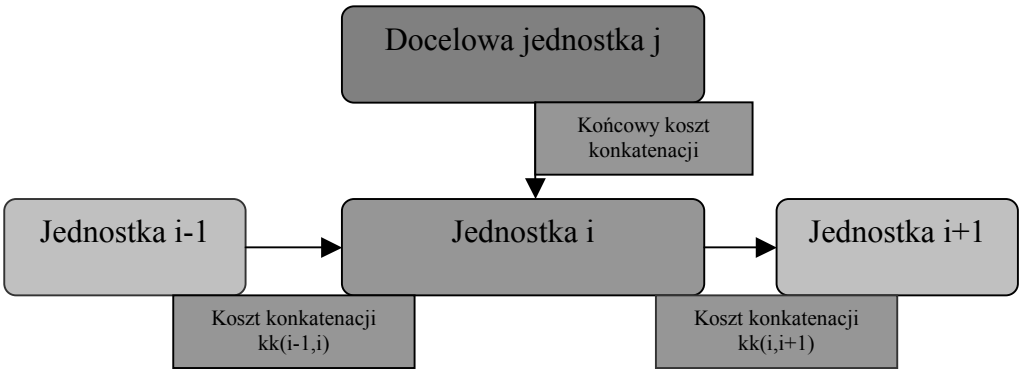
Warto wspomnieć, że funkcja kosztu jest również zaimplementowana w bezpłatnym systemie Festival.

Wydaje się, że właśnie ta technika ma szanse rozwinąć się w przyszłości. Obecnie są prowadzone badania nad udoskonaleniem funkcji estymacji. Celem jest utworzenie takiej funkcji, która wybierze najbardziej zbliżoną do mowy ludzkiej sekwencję jednostek akustycznych. Praktycznie jedynym problemem jest rozmiar korpusu, który wynosi około 200 MB oraz złożoność procesu obliczeniowego. System taki z uwagi na wymagania sprzętowe obecnie jest wykorzystywany w portalach głosowych.

Poniższe rysunki pokazują sposób generowania słowa „dog", przy użyciu korpusowej syntezy mowy oraz sposób obliczani funkcji kosztu.



Rysunek 3.9 Sposób generowania słów w korpusowej syntezie mowy



Rysunek 3.10 Funkcja kosztu

3.15 Algorytmy syntezowania mowy

Mowa jest najczęściej generowana poprzez użycie difonów w konkatenacyjnej syntezie mowy. Na poziomie syntezy cechy prozodyczne, które zanikają w procesie segmentacji muszą być zmodyfikowane by głos brzmiał naturalnie. Dodatkowo podczas łączenia poszczególnych jednostek akustycznych ze sobą należy mieć pewność, że ilość nieciągłości powstałych przy generowaniu mowy ciągłej będzie minimalna.

Za wyżej wymienione funkcje odpowiada algorytm syntezy mowy. W dużej mierze od niego zależy jakość wygenerowanej mowy.

Istnieje kilka algorytmów syntezowania mowy:

- LP-Linear Prediction. Zaletą modelu predykcji liniowej jest wydajny sposób kodowania mowy. Jednak jakość generowanej mowy jest niska.
- TD-PSOLA – (Time-Domain Pitch-Synchronous – Overlap-Add jeden z bardziej popularnych metod. Metodę tą stosuje się do modyfikacji intonacji. Zapewnia ona dobrą jakość generowanej mowy, choć posiada pewne ograniczenia związane z jej nieparametryczną strukturą. TD-PSOLA jest szczególnie dobrym algorytmem łączenia ze sobą sklejek, kiedy mamy do czynienia z małą modyfikacją jednostek akustycznych. W szczególności chodzi tutaj o nieciągłości widmowe na poziomie

granic jednostek akustycznych. Staje się to problemem, kiedy mamy do czynienia z małą bazą danych, na przykład bazą difonów. Model ten sprawdza się, gdy jednostki są ekstrahowane z bardzo dużego korpusu, a prozodia ich jest bardzo zbliżona do prozodii końcowej. Dodatkowo model ten wymaga bardzo dokładnego wskazania tonacji. MBROLA jest bardziej tolerancyjnym algorytmem podczas modyfikacji bazy jednostek akustycznych. Umożliwia ona dodatkowo proces wygładzania na łączonych granicach.

- MBROLA zapewnia bardzo dobrą jakość łączenia segmentów, przy jednocześnie niskim nakładzie kosztów obliczeniowych. Dodatkowo jest porównywana z modelami harmonicznymi. Resynetza w MBROLI dotyczy ramek tylko akcentowanych, natomiast nie akcentowane są kopiowane. W ten sposób eliminujemy niezgodności intonacji, co zapewnia pełną zgodność fazy podczas procesu konkatenacji.

  Resynteza w MBROLI przynosi dodatkowe korzyści. Jedną bardzo ważną zaletą jest zachowanie stałej intonacji, co pozwala zaznaczyć wysokość tonu tzw. pitch mark. Wiąże się to z czasem, jaki jest potrzebny na przygotowanie bazy danych.

  Algorytm ten posiada też pewne wady. Podczas resyntezy może nastąpić zniekształcenie fazy. Czasem powstają nieciągłości fazy, linowe wygładzenie może wprowadzić czasem nieco sztuczny dźwięk.

  Wydaje się że z wyżej wymienionych modeli MBROLA generuje najlepszą jakość mowy, mimo zniekształceń wprowadzanych na akcentowanych częściach.
- Modele harmoniczne – to modele parametryczne, pozwalające uniknąć problemów takich jak w wyżej wymienionych algorytmach. Ich wadą jest stosunkowo duża moc obliczeniowa, którą wykorzystują.

W porównaniu do metod TD-PSOLA i MBROLA, dużo lepszą jakość dla syntezy mowy dają metody hybrydowe, nie mniej jednak wymagają one znacznie większych nakładów obliczeniowych.

3.16 Zastosowanie systemów syntezy mowy

W niniejszym rozdziale opisałem cały proces tworzenia systemu syntezy mowy. Myślę, że warto teraz powiedzieć czemu to tak naprawdę służy i co z tego wynika.

Synteza mowy ma coraz większe zastosowanie i obejmuje coraz więcej dziedzin. Przede wszystkim należy wymienić tutaj edukacje w postaci wirtualnych uniwersytetów, liczne instytucje wirtualne, w których mowa nie stanowi języka naturalnego, lecz sztucznie generowany głos.

Kolejną dziedziną zastosowania syntezy mowy jest telekomunikacja. Większość rozmów, około 70%, jakie przeprowadzamy dzwoniąc do różnych serwisów informacyjnych nie wymaga dużej interaktywności. Stąd też wynika duże zainteresowanie tą dziedziną. AT&T zbudowała kilka systemów, mających zastosowanie w telekomunikacji. Jednym z nich jest informowanie o danych personalnych dzwoniącego przed odebraniem połączenia. Inny system opierał się o technologie czytania elektronicznych listów przez telefon. Systemy te dają dobrą jakość syntezy mowy, dlatego znalazły zastosowania i są dosyć popularne.

Mówiące książki i zabawki – to kolejna dziedzina, w której można zastosować syntezę mowy. Niemniej jednak realizowane syntezy mowy przez „grające cuda” jest zbyt mała żeby można ja było wykorzystać do celów edukacyjnych.

Synteza mowy w niedalekiej przyszłości z pewnością będzie miała zastosowanie przy kontrolowaniu urządzeń samochodowych takich jak klimatyzacja, radio, elektroniczna mapa. Niezbędna pomoc w postaci korzystania z Internetu podczas podróży oraz mówiący system nawigacyjny informacje o korkach drogowych czy też informacje o stanie poszczególnych urządzeń samochodu to tylko nieliczne zastosowanie tej technologii

Synteza mowy będzie miała również duże zastosowania w dziedzinie zasobów ludzkich. Dzięki syntezie mowy ludzie niewidomi mają dostęp do wiadomości tekstowych.

Mutimedia – komunikacja werbalna między człowiekiem a komputerem. Niezbędnym krokiem jest uzyskanie wysokiej jakości syntezy mowy, która to jest elementem koniecznym aby spełniło się marzenie człowieka o przebrnięciu testu Turinga[10].

Informacje głosowe – czasami informacja głosowa jest bardziej efektywna od informacji tekstowej. Szczególnie, jeśli myślimy o krótkiej informacji: alarmy, uwagi. Portale głosowe są tego najlepszym przykładem. Zadaniem portali głosowych jest symulowanie interakcji głosowej z użytkownikiem. Portale głosowe są wyposażone w wyrafinowane mechanizmy interakcji z użytkownikiem, których podstawą jest rozpoznawanie oraz konwersja tekstowej informacji pobranej z bazy danych do postaci dźwiękowej.

Portal głosowy jest nie tylko wymyślnym systemem do prowadzenia konwersacji z komputerem, lecz przede wszystkim stanowi bazę danych, czyli zasób ważnych informacji dla potencjalnych klientów serwisu. Informacje te przechowywane są w postaci tekstowej na serwerach baz danych, skąd pobierane są przez skrypty, zlokalizowane na serwerach WWW, obsługujące zapytania SQL. Wyselekcjonowane wiadomości konwertowane są do postaci dźwiękowej przez przeglądarkę głosową i emitowane.

Technologia IVP (Internet Voice Portal), mimo że jest jeszcze bardzo młoda, przeżywa w USA swój rozkwit. Pojawiło się szereg bogatych serwisów informacyjnych zarówno udostępniających własne zasoby, jak i korzystających z zasobów Internetu. Część z nich umożliwia także realizację podstawowej usługi internetowej, czyli dostępu do poczty elektronicznej. Portale te są powszechnie dostępne na terenie całych Stanów Zjednoczonych, a korzystanie z nich jest bezpłatne.

[10] Test Turinga (angielskie Turing test), eksperyment definiujący "maszynę myślącą", zaproponowany przez A. Turinga. W myśl testu Turinga maszynę można uznać za naśladującą dostatecznie dobrze procesy myślowe, jeśli człowiek prowadzący z nią dialog (nie poinformowany o tym, że rozmawia z maszyną), nie będzie w stanie odróżnić rozmowy z maszyną od rozmowy z drugim człowiekiem.

(Źródło: Słownik Encyklopedyczny - Informatyka")

Do najbardziej popularnych portali głosowych w Stanach Zjednoczonych należą:

- TellMe
- ShopTalk
- TeleSurf
- BeVocal
- AudioPoint

(Źródło Domalewski, W. *PC*)

Zastosowania Text-to-speech Systems są bardzo liczne. Zainteresowanie nimi rośnie a jakość syntezy mowy jest coraz lepsza. Systemy mają na celu nie tylko ułatwienie człowiekowi komunikacji w rzeczywistości, są także pomocą dla osób niewidomych. Służą jednocześnie celom edukacyjnym i stanowią wielkie dążenia do realizacji prawdziwej werbalnej komunikacji z komputerem.

Synteza mowy ma również zastosowanie w:
PIM – osobistym zarządzaniu informacjami – Palmtopy
GPS
Grach mobilnych
Systemach nawigacyjnych
Mówiących avatarach (Patrz 3.17 Avatary)
Czytaniu maili
(CD)

3.17 Avatary

Niewątpliwie ciekawostką jest zastosowanie syntezy mowy w dziedzinie sztucznej inteligencji. Avatary są pewnego rodzaju personifikacją systemu komputerowego. Wyglądem przypominają głowę człowieka lub jego cała sylwetkę. Są w pełni zaprogramowanymi obiektami mającymi odzwierciedlać ruchy, mimikę oraz gesty człowieka.

Językiem porozumiewania się avatarów jest synteza mowy. Obecne dążenia skupiają się na tworzeniu postaci wraz z elementami sztucznej inteligencji. Avatar ma w pewnym stopniu odzwierciedlać człowieka.

W przyszłości być może avatary w niektórych programach telewizyjnych będą zastępowały spikerów (obecnie występują w reklamach).

Poniżej znajduje się rysunek avatara:

Rysunek 3.11 Avatar, Babel Technologies

3.18 Podsumowanie

W tym rozdziale przedstawiłem niezbędne elementy i definicje związane z syntezą mowy. Począwszy od wprowadzenia w dziedzinę syntezy mowy poprzez opisanie budowy systemów syntezy mowy, rodzajów generowania sztucznej mowy oraz stosowanych algorytmów aż po zastosowania pełnych systemów TTS. Podane informacje obrazują jak ogromną i przyszłościową dziedziną multimediów jest synteza mowy.

Myślę, że to wprowadzenie ułatwi zrozumienie następnego rozdziału, a także podkreśli istotę projektu. Mowa o przygotowanym przeze mnie wyjściu akustycznym dla języka polskiego, o którym traktuje moja praca praktyczna. W następnym rozdziale przejdę do jej omówienia.

4. Przygotowanie bazy difonów

4.1 Wstęp

Celem mojej pracy było uzyskanie wyjścia akustycznego dla języka polskiego. Proces tworzenia składał się z kilku etapów. W pierwszym etapie należało przygotować listę fonemów oraz korpus difonów.

Kolejnym etapem było przeprowadzenie nagrań i segmentacja difonów. Etap ten był najbardziej czasochłonny. Wymagał dużej precyzyjności ostrożności, a także przeprowadzenia bardzo dużej ilości testów, które pozwoliłyby stwierdzić i ocenić jakość zaznaczonych granic. Etapem końcowym było wyeksportowanie bazy difonów do formatu plików Raw oraz wysłanie ich do dalszego przetworzenia do zespołu MBROL-i na Politechnice w Mons.

W tym rozdziale znajduje się opis poszczególnych procesów.

4.2 Przygotowanie i utworzenie listy fonemów

W języku polskim istnieje 37 fonemów. Ilość możliwych połączeń w celu uzyskania wszystkich możliwych słów wynosi 1369.

Przyjąłem założenie, że w moim korpusie znajdą się wszystkie kombinacje fonemów nawet takie, których sekwencje nie występują w języku polskim. Jest to metoda bezpieczna, mimo iż nakład pracy jest większy. Albowiem nigdy nie dojdzie sytuacji, kiedy system stanie przed słowem lub jakimś skrótem, którego nie będzie w stanie wygenerować.

Korpus liczy 1443 słowa. Różnica ta wynika z konieczności dodania elementu ciszy na początku i końcu słowa, co stanowi dodatkowo 74 difony.
Wyliczenie to wynika z :
Ilość możliwych połączeń = ilość fonemów * ilość fonemów + 37 fonemów z ciszą na początku + 37 fonemów z ciszą na końcu = 37*37+ 74=1443
(CD)

Przygotowanie tej wersji korpusu polegało na utworzeniu połączeń każdego fonemu z każdym. Etap ten został przeprowadzony automatycznie na podstawie listy fonemów w programie Diphone Studio[11]. Dalsza obróbka została przeprowadzona w programie Windows Notepad

4.3 Przygotowanie korpusu

Najbardziej istotnym elementem podczas przygotowania korpusu było znalezienie odpowiedniego kontekstu dla difonów. Przyjąłem następujące reguły:

W korpusie musiały się znaleźć wszystkie możliwe połączenia głosek. W wygenerowanych wyrazach difon nie mógł stanowić sylaby akcentowanej, ponieważ mogło to mieć wpływ na zniekształcenia nagranego w późniejszym etapie sygnału. Otoczenie difonu nie mogło wpływać na jego koartykulację, co wbrew pozorem nie jest takim łatwym zadaniem.

[11] Diphone Studio jest programem służącym do przeprowadzenia segmentacji danych. Program ten jest dostępny za darmo pod warunkiem nie wykorzystywania komercyjnego.

Dla grupy samogłoskowej-samogłoskowej[12]:

- b+difon+nany" np: baenany

Dla grup samogłoskowo-spółgłoskowych

- D+difon+bany np. didbany
- D+difon+pany np. difpany

Dla grup spółgłoskowo-spółgłoskowych:

- a+difon+anany np. apsanany

Dla grup spółgłoskowo-samogłoskowych:

- a+difon+nany np. afanany
- a+difon+banany np. azubanany

Dla difonu w postaci cisza+głoska użyłem kontekstu:

- difon+ana jeśli difon był formacja spółgłoskową
- difon+bana jeśli difon był formacją samogloskową

Dla difonu w postaci głoska + cisza kontekst był następujący

- bana+difon – jeśli głoska była spółgłoskową
- ban+difon – jeśli głoska była samogłoskową

Podczas przygotowania korpusu napotkałem bardzo wiele sytuacji, w których nie można było zastosować powyższych reguł. Sytuacje te przeważnie dotyczyły grup spółgłoskowo-spółgłoskowych i samogłoskowo-samogłoskowych. Np. difon w postaci „ii" miał kontekst „aniimadwo". W takich przypadkach kierowałem się częściej intuicją niż konkretnymi regułami.

[12] Przez formację samogłoskową rozumiem difon w postaci np. „a+e"

Korpus można utworzyć w dowolnym edytorze tekstowym. Mój korpus został utworzony w Notepadzie. Korpus ma format 8-bitowego pliku ASCII i zawiera następujące informacje:
Pierwsza linia korpusu zawiera informacje poprzedzoną wykrzyknikiem o częstotliwości próbkowania, z jaką zostały przeprowadzone nagrania.
Następne linie zawierają kolejno:
Nazwę difonu, kontekst (w kodzie SAMPA), numer pliku, w którym znajduje się sygnał z nagranym kontekstem oraz granice przedstawione w numerach próbek. Pierwsza oznacza początek, druga środek, a trzecia koniec difonu.

Poniżej znajdują się przykładowe linie z ostatecznej wersji korpusu:

!16000

i	**i**	**w0.wav**	**ani imadwo**	**41658**	**42577**	**43144**
i	**I**	**w1.wav**	**bi InanI**	**42773**	**43626**	**44339**
i	**e**	**w2.wav**	**biebanI**	**30567**	**31103**	**31638**
i	**a**	**w3.wav**	**biabanI**	**26871**	**27512**	**28075**

Rysunek 4.1 Pierwsze linie korpusu wraz z kontekstem (CD)

4.4 Nagrania

Kolejnym bardzo istotnym elementem, było przeprowadzenie nagrań. Etap ten był najważniejszym elementem całego projektu. Jakość syntezy mowy wyłącznie zależała od sposobu mówienia, dokładności wymawiania, zachowania stałej głośności oraz monotonnego mówienia. Nagrania zostały przeprowadzone podczas jednej sesji.

W celu realizacji powyższych wymagań musiałem znaleźć osobę o czystym, naturalnie brzmiącym głosie, jednocześnie znającą rewelacyjnie fonetykę języka polskiego.

Muszę przyznać, że napotkałem duże trudności w poszukiwaniu takiej osoby. Poszukiwania rozpocząłem od skontaktowania się panią Górską, zastępcą rektora Akademii Teatralnej im. Zelwerowicza w Warszawie p. Jana Englerta. Pani Górska

zaproponowała mi do realizacji przedstawionego projektu Annę Dereszowską studentkę trzeciego roku szkoły Teatralnej w Warszawie. Okazało się, że był to bardzo dobry wybór, co można usłyszeć w wygenerowanych przeze mnie plikach przykładowych.

Korpus został nagrany w programie Mobile Recording Studio firmy Sony w postaci plików Raw z częstotliwością próbkowania 32 kHz. Taka częstotliwość zapewnia optymalną, dla syntezy mowy, jakość generowanego sygnału.
Do nagrań użyłem 4 mikrofonów:

Mikrofon tzw. „Close-talking" oraz 3 mikrofony zbierające sygnał z otoczenia umieszczone w odległości nie większej niż jeden metr od mówiącego.

W procesie segmentacji wykorzystałem nagrania przeprowadzone na mikrofonie „close-talking". Mikrofon ten zagwarantował najlepszą jakość dźwięku i wprowadził najmniejszą ilość zniekształceń w sygnale.

Nagrania zostały przeprowadzone w studiu Polsko-Japońskiej Wyższej Szkoły Technik Komputerowych.

4.5 Segmentacja

Po przeprowadzeniu nagrań następnym etapem a zarazem stanowiącym główną część mojego projektu była segmentacja wcześniej przygotowanego korpusu.

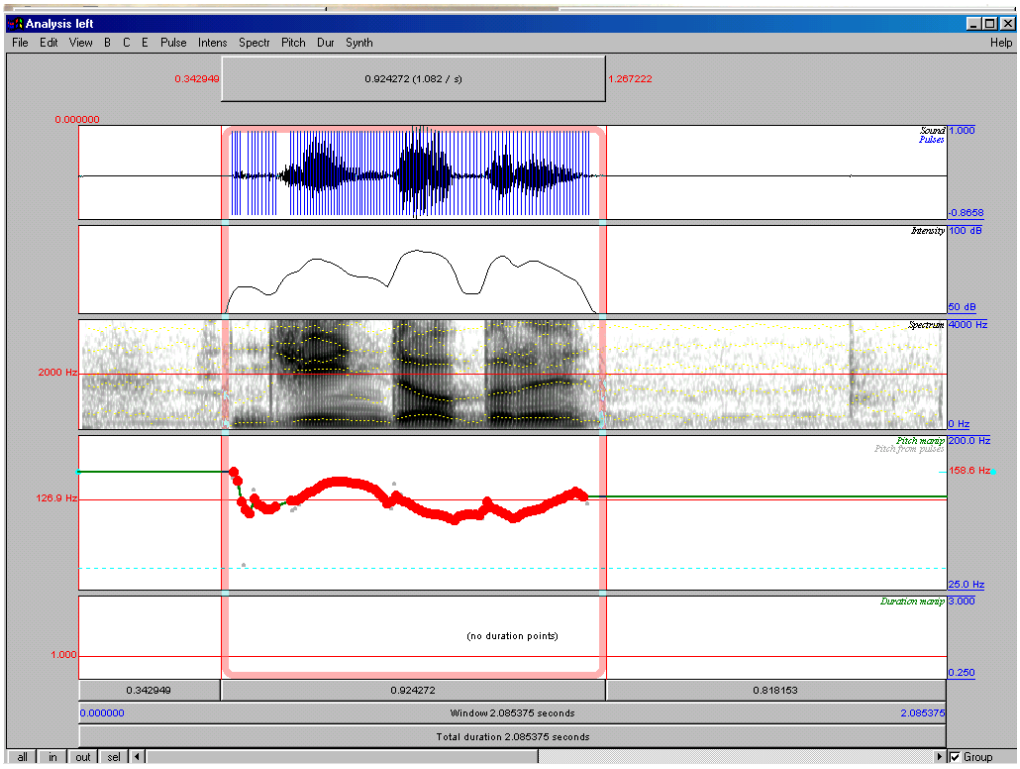
Proces segmentacji jest zagadnieniem podstawowym w obrębie syntezy mowy. Należy dobrze nie tylko przygotować nagrania, ale również dużą ilość czasu poświęcić na studiowanie fonetyki i reprezentacji akustycznej głosek. Ułatwia to później pracę nad segmentacją.

Segmentacja nie jest procesem trywialnym. W celu jej realizacji należy być obeznanym z podstawowymi parametrami dźwięku używanymi podczas tego procesu.

Proces segmentacji realizowałem w Praacie[13], programie wspomagającym pracę fonetyków.

Poniżej znajduje się okno tego programu przedstawiające:

- Sygnał w dziedzinie czasu
- Energię sygnału
- Spektrogram
- Analizę formantową
- F0
- Oraz miejsce do opisania poszczególnych elementów

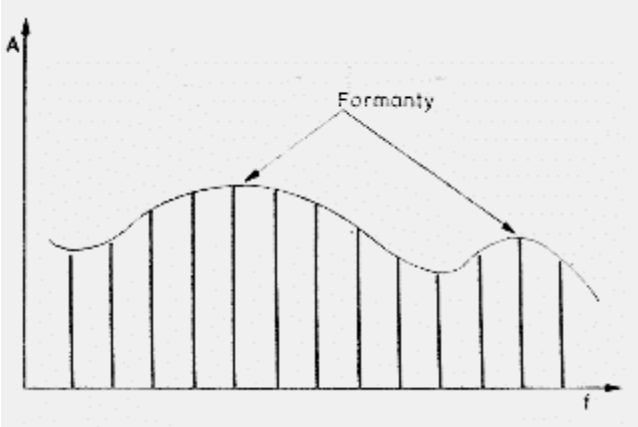


Rysunek 4.2 Podstawowe parametry stosowane podczas procesu segmentacji

[13] Program ten jest dostępny za darmo po wcześniejszym zarejestrowaniu się na stronie www.praat.org

4.5.1 Analiza formantowa

Jedną z podstawowych metod, stosowanych podczas procesu segmentacji, to znaczy wyznaczania granic difonów, jest posłużenie się spektrogramem z analizą formantową.



Rysunek 4.3 Ilustracja pojęcia formantu.

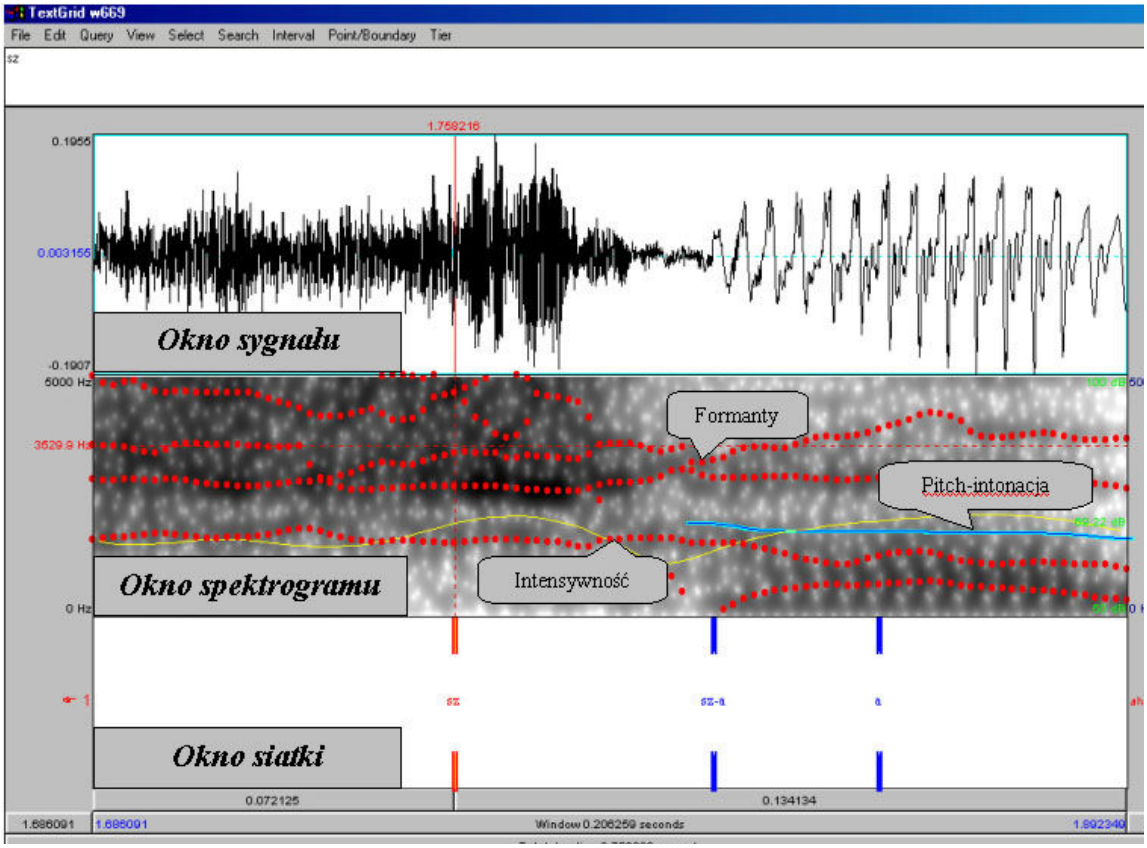
Widmo sygnału ma charakter dyskretny, co na rysunku obrazują prążki. Maksimum obwiedni widma prążków nazywa się formantem. A częstotliwość, przy której występuje owo maksimum, nosi nazwę częstotliwości formantowej.

Segmentacja korpusu jest zadaniem trudnym wymagającym bardzo dużej precyzji, wiedzy i znajomości fonetyki. Polega ona na poprawnym ustawieniu granic jednostek akustycznych.

Granica obejmuje zawsze początek jednostki akustycznej, środek, czyli moment przejścia pomiędzy jednym a drugim fonemem oraz koniec – czyli prawą granicę będącą końcem difonu. Podczas segmentacji powstają bardzo często problemy natury technicznej. Jak należy ustawić granice difonu, żeby zsyntetyzowany wyraz brzmiał dobrze.

W zasadzie trudno jest podać obowiązujące w sposób jednoznaczny reguły segmentacji sygnału na kolejne difony, ponieważ generalna zasada segmentacji difonowej sygnału mowy polegająca na wyodrębnianiu z sygnału końcowej części stacjonarnej jednej głoski, wraz z przejściem do następnej i dołączeniem początkowej części stacjonarnej drugiej głoski, nie zawsze jest możliwa. Reguły segmentacji muszą być dostosowane do rzeczywistej postaci difonów. Będzie o tym mowa w dalszej części pracy.

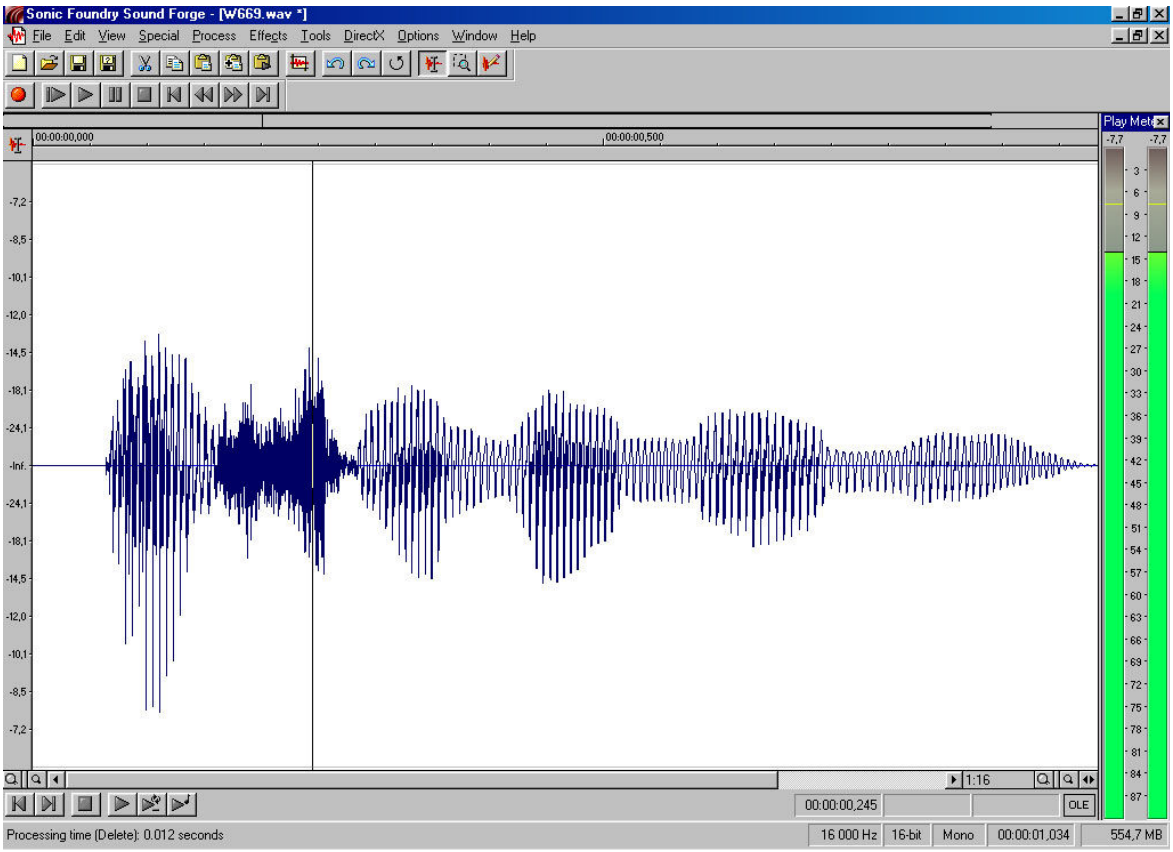
Poniżej znajduje się okno systemu Praat oraz przykładowa segmentacja difonu



Rysunek 4.4 Segmentacja difonu

Podczas procesu segmentacji zdarzało się odkryć, że difon był źle nagrany. Sprowadzało się to do usunięcia najczęściej zbyt długiej ciszy lub pewnych fragmentów sygnału zniekształcających brzmienie difonu. Tak było na przykład z difonem „v-s”. Po fonemie „v” słychać było dość często fonem „I”. Przy konkatenacji odpowiedniego słowa z tym kontekstem brzmiało to niezbyt dobrze. Dlatego operacje na sygnale czasem były konieczne. Do tego celu użyłem programu Sound Forge w wersji 5.0.

Poniżej znajduje się okno tego programu oraz przykładowa operacja na sygnale.

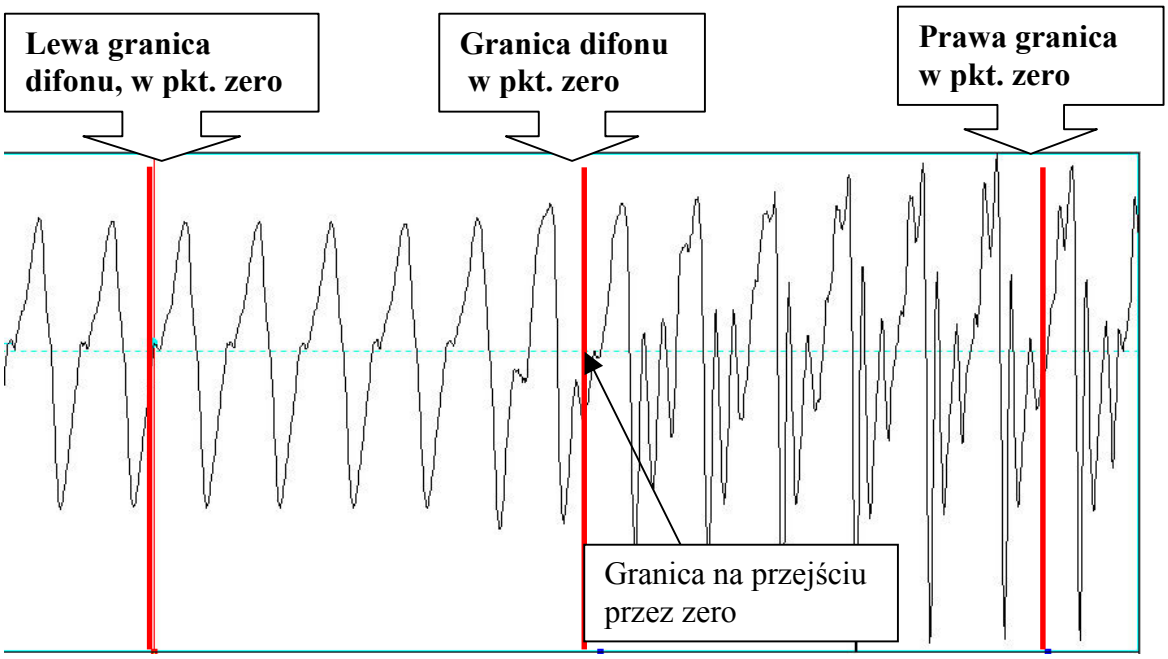


Rysunek 4.5 Okno programu Sound Forge

4.6 Reguły w procesie segmentacji

Proces segmentacji powinien być przeprowadzony z wielką uwagą i dokładnością, ponieważ jakość generowanej mowy zależy od jego poprawności.
Podczas ustawiania granic difonu należy zawsze pamiętać o tym żeby, początek zaczynał się na początku okresu krtaniowego i był zaznaczony w przejściu przez „zero". W ten sposób uniknie się trzasków podczas generowania mowy.

Poniższy rysunek prezentuje sposób przeprowadzenia segmentacji w przejściu przez „zero".

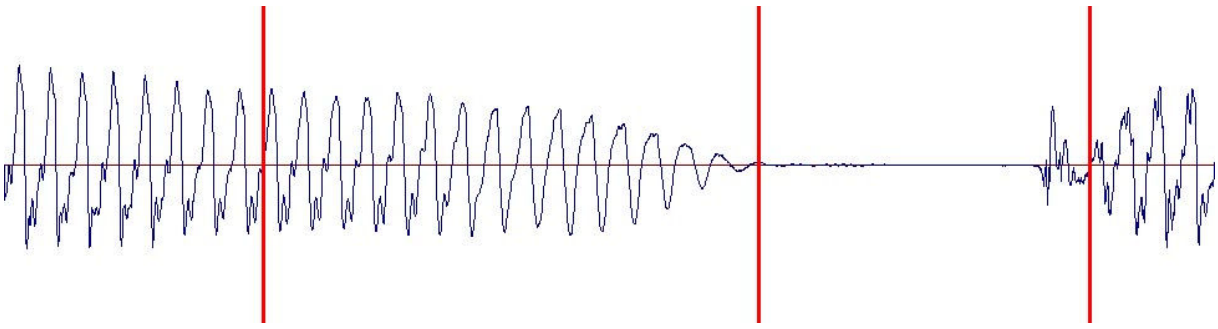


Rysunek 4.6 Podstawowe reguły, które zastosowałem podczas procesu segmentacji

Przedstawię teraz pewne reguły, które stosowałem podczas tego procesu i jednocześnie, które w znacznej mierze wpłynęły na sukces projektu.
Wychodziłem z założenia, że długość difonu nie powinna przekraczać 70 ms oraz, żeby difon nie był dłuższy niż pięć okresów krtaniowych z lewej strony i pięć okresów krtaniowych z prawej.

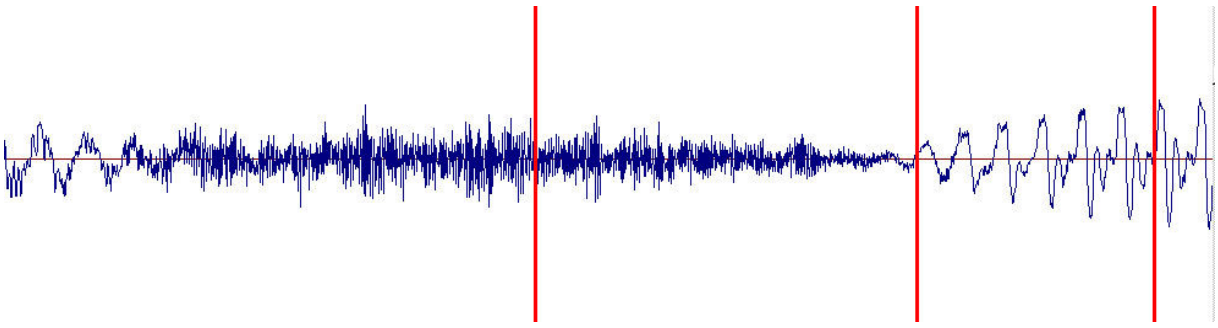
W przypadku samogłosek założenie to było możliwe do realizacji. Wynika to oczywiście ze struktury samogłoski i jej krótkiego czasu trwania. Jednak podczas segmentacji spółgłosek zdarzały się sytuacje, kiedy długość difonu wynosiła ponad 100 ms.

Poniższy rysunek prezentuje difon „e~p", którego czas trwania przekracza 130 ms.



Rysunek 4.7 Difon „e~p". Czas trwania ponad 130 ms

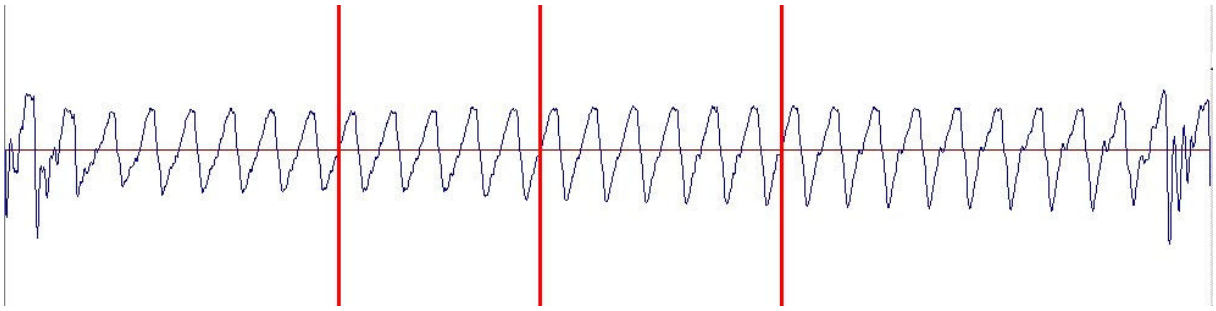
Zdarzały się sytuacje, w których ustawienie granic było banalne. Miało to miejsce podczas segmentacji głosek szczelinowych. Jak widać na poniższym rysunku dokładnie można określić przejście pomiędzy jednym a drugim fonemem.



Rysunek 4.8 Granice w difonie „S-e"

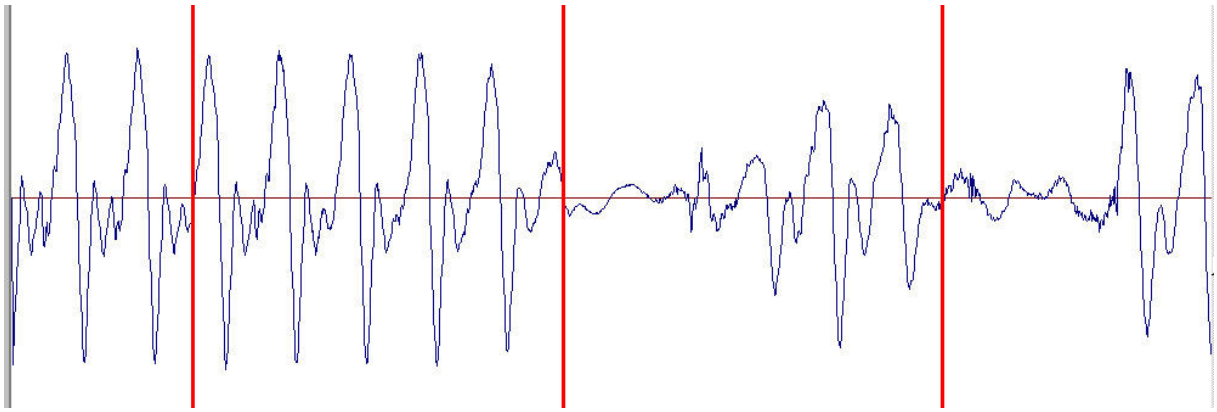
W przypadku głosek półotwartych zarówno ustnych jak i nosowych zasady podziału były bardzo intuicyjne i określenie granic było bardzo trudne.

Poniższy rysunek prezentuje sytuację krytyczną, kiedy należało znaleźć granice pomiędzy fonemami „n” i „m”



Rysunek 4.9 Difon „n-m” .Granice miedzy fonemem „n” i „m”.

Segmentacja głoski drżącej „r” również była dość ciekawym przypadkiem. Trudno było w tym przypadku odróżnić nagłos od wygłosu. W dodatku głoska ta charakteryzuje się bardzo krótkim czasem trwania. Najdłuższa głoska „r” trwała około kilkunastu milisekund a wiadomo, że do poprawnego brzmienia potrzeba kilkanaście do kilkudziesięciu milisekund kontekstu.(CD)



Rysunek 4.10 Segmentacja difonu „e-r”

4.7 Problemy związane z segmentacją

Segmentacja ręczna jest procesem bardzo długim i trudnym. Dlatego pojawiły się mechanizmy automatyczne pozwalające zaznaczać granice. Niestety do tej pory mechanizmy te nie zyskały dużej popularności z uwagi na ich małą skuteczność, w szczególności ze względu na niewstarczającą dokładność oszacowania granic difonu (conajmniej 10 ms).

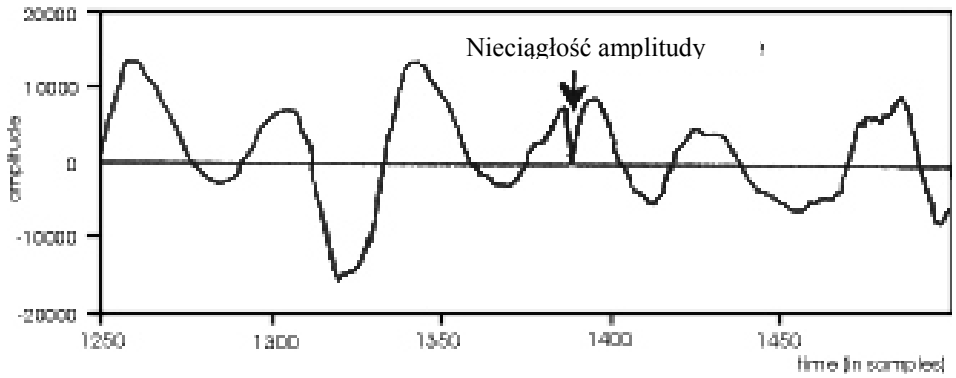
Ponieważ segmentacja jest bardzo ważnym elementem podczas tworzenia systemu TTS, przedstawię problemy, jakie napotkałem podczas jej przeprowadzania.

Problemy, jakie mogą się pojawić podczas segmentacji mogą być związane z:

- Amplitudą
- Energią
- Częstotliwością
- Fazą

Nieciągłość amplitudy pojawia się, kiedy koniec difonu i początek bezpośrednio po nim następującego są zupełnie inne. Podczas łączenia takich elementów akustycznych pojawiają się trzaski w generowanym sygnale.

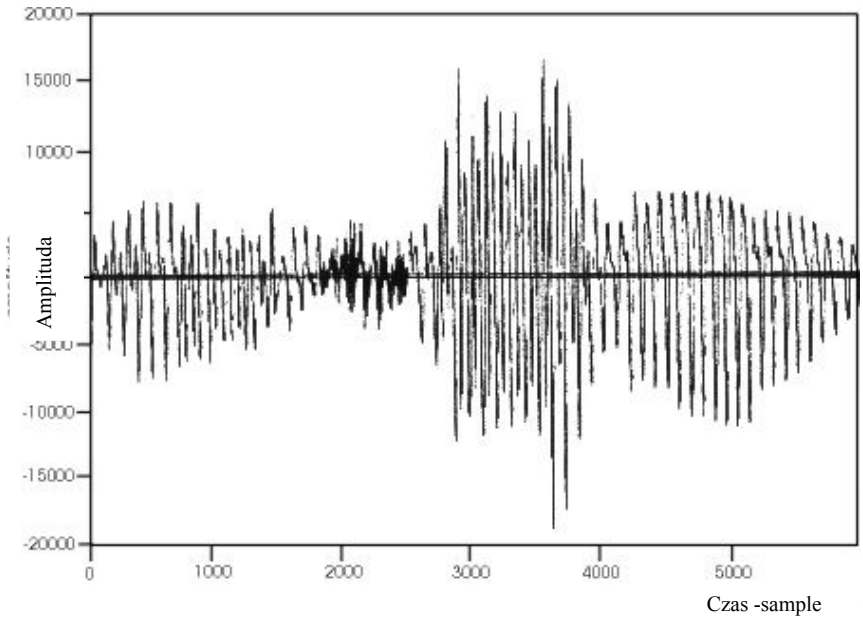
Poniższy rysunek obrazuje opisaną sytuację



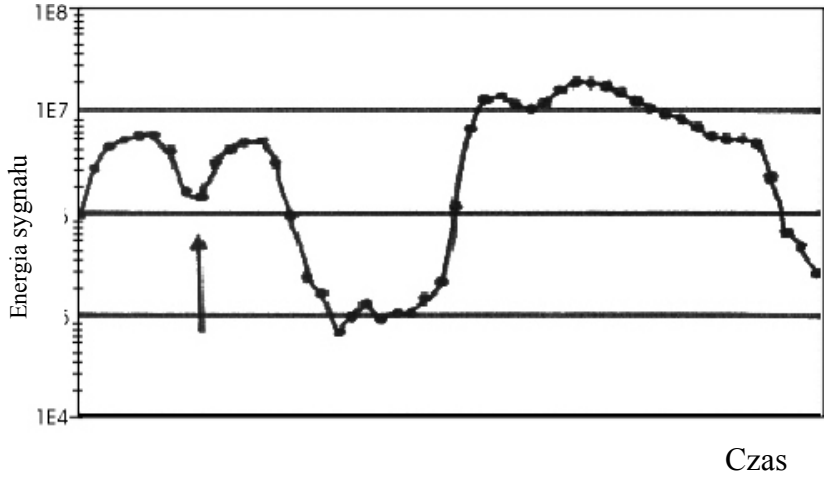
Rysunek 4.11 Brak ciągłości w amplitudzie.

Nieciągłość energii oznacza, że difon wraz z kontekstem został wymówiony ze zwiększoną energią niż następujący lub poprzedzający go.

Poniższe schematy obrazuje tą sytuację.



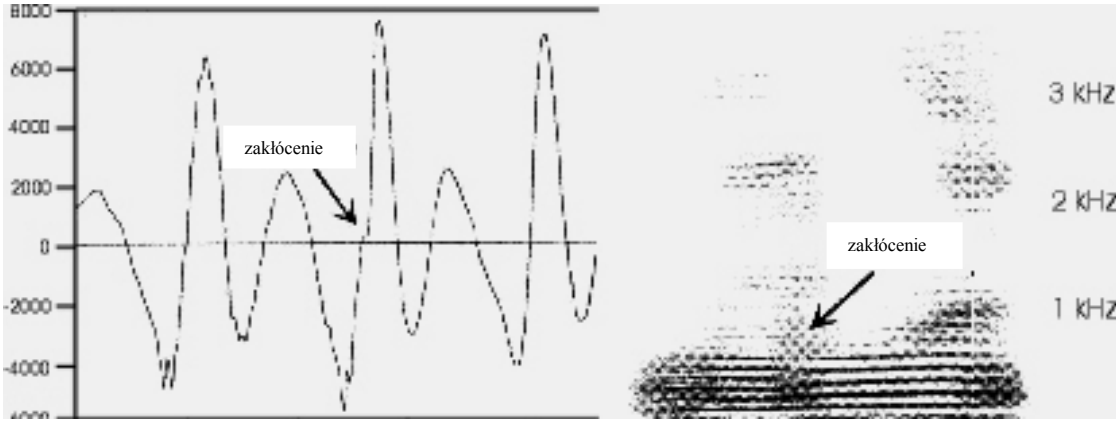
Rysunek 4.12 Nieciągłość energii w dziedzinie czasu



Rysunek 4.13 Nieciągłość energii

Kolejnym problemem jest przesunięcie fazy. Jest ona słyszana bardzo krótko, jednak wystarczająco długo, żeby usłyszeć trzask.

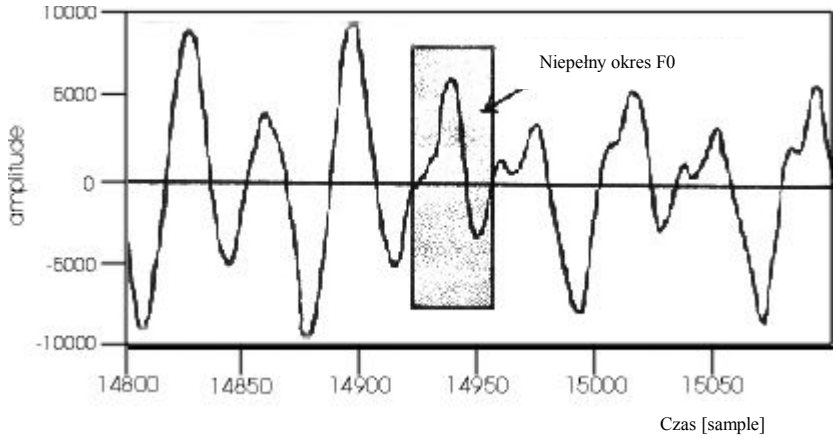
Rysunek poniżej przedstawia tę sytuację.



Rysunek 4.14 Nieciągłość sygnału w dziedzinie czasu i częstotliwości

Bardzo ważnym elementem segmentacji jest wystrzeganie się od poniższego błędu. Nieciągłość fazy pojawia się, jeśli granica difonu nie znajduje się na początku okresu krtaniowego fonemu. Efektem ustawienia niepoprawnych granic jest słyszalny trzask w generowanym sygnale.

Poniższy schemat odzwierciedla tę sytuację.



Rysunek 4.15 W skutek wtrącenia niepełnego okresu powstaje w sygnale skok fazy.

Wyżej wymienione błędy mogą mieć kolosalny wpływ na jakość posegmentowanych jednostek akustycznych. Można uniknąć wymienionych problemów, przeprowadzając staranną segmentację. Jest to element podstawowy a zarazem konieczny w celu osiągnięcia zadowalającego efektu końcowego.

4.8 Charakterystyka klas głosek

W celu lepszego zobrazowania, problemów jakie mogą się pojawić podczas segmentacji prezentuję poniżej tabele charakteryzująca poszczególne klasy głosek.

Typ głoski	Przykład	Dziedzina czasu	Dziedzina częstotliwości
Samogłoski	/a/ /e/ /i/	- quasi-periodyczne - nośnik dużej energii	- od 3 do 6 widocznych formantów - wyraźne maksima intonacji (peak pitch) - główna energia sygnału zawiera się w niskich częstotliwościach
Akcentowane wybuchowe	/b/ /d/ /g/	- nagły skok amplitudy	- formant w dolnych częstotliwościach - krótkotrwały skok energii we wszystkim pasmach częstotliwości
Nieakcentowane wybuchowe	/p/ /t/ /k/	- nagły i wysoki skok amplitudy	- brak struktury formantów - energia skoncentrowana w wysokich częstotliwościach
Akcentowane frykaty	/v/ /z/ /j/	- szeroki quasi-niezmienny obszar - nośnik małej ilości energii	- formanty - pierwszy formant podobny do formantu samogłoskowego - energia skoncentrowana w wysokich częstotliwościach
Nieakcentowane frykaty	/f/ /s/	- charakterem podobna do szumu - nośnik małej ilości energii	- szerokie spektrum
Głoski nosowe	/m/ /n/	- quasi-periodyczne - podobne do samogłosek - mniejsza ilość energii niż w samogłoskach	- minimum w zasięgu spektrum - formanty podobnie jak w samogłoskach
Spółgłoski płynne	/l/ /r/	- brak wyraźnego, niezmiennego początku -mało widoczne przejście da sąsiadującej samogłoski	- pierwszy formant o mniejszej częstotliwości

Tabela 4.1 Charakterystyka poszczególnych klas głosek

4.9 Skrypty

W celu ułatwienia pracy związanej z segmentacją napisałem dwa skrypty w języku skryptowym Praata. Pomogły one konwertować nagrane pliki korpusu i przygotować je do bezpośredniej segmentacji.

Skrypt znajdujący się poniżej wykonywał szereg operacji. Najpierw otwierał odpowiedni plik. Nazwa pliku potrzebnego do otwarcia znajdowała się w zmiennej „nazwa_pliku". Wczytywany plik był plikiem nagranym o częstotliwości 32 kHz. Jednak w nagłówku pliku znajdowała się informacja, że jest to plik o częstotliwości 16 kHz. Należało odpowiednio zmienić tą informację oraz przekonwertować plik do częstotliwości 16 kHz. Taka częstotliwość jest wymagana w MBROL-i. Po resamplingu plik był otwierany w nowym oknie edycyjnym.

Poniżej znajduje się kod programu:

```
form supply_arguments
sentence name  nazwa_pliku
sentence dir d:\zpoprzedniejwersji\Praat\powtorzone\
endform
Read Sound from raw 16-bit Little Endian file... 'dir$''name$'.ch1
select Sound 'name$'
Override sample rate... 32000
Resample... 16000 50
Write to WAV file... d:\'name$'_16000.wav
#Read from file... 'dir$''name$'.nist
To TextGrid... ahmbanany ahmbanany
select Sound 'name$'_16000
plus TextGrid 'name$'_16000
Edit
```

4.10 Edycja posegmentowanego korpusu

Kolejnym ułatwieniem, jakie zastosowałem było przygotowanie skryptu do reedycji difonu. Często zdarzało się, że pliki z zaznaczonymi granicami w Praacie należało poddać dodatkowej edycji w celu wprowadzenia korekt. W tym celu napisałem skrypt, który wczytywał plik tekstowy wraz z odpowiadającym plikiem dźwiękowym oraz otwierał okno edycji.
Poniżej znajduje się kod skryptu:

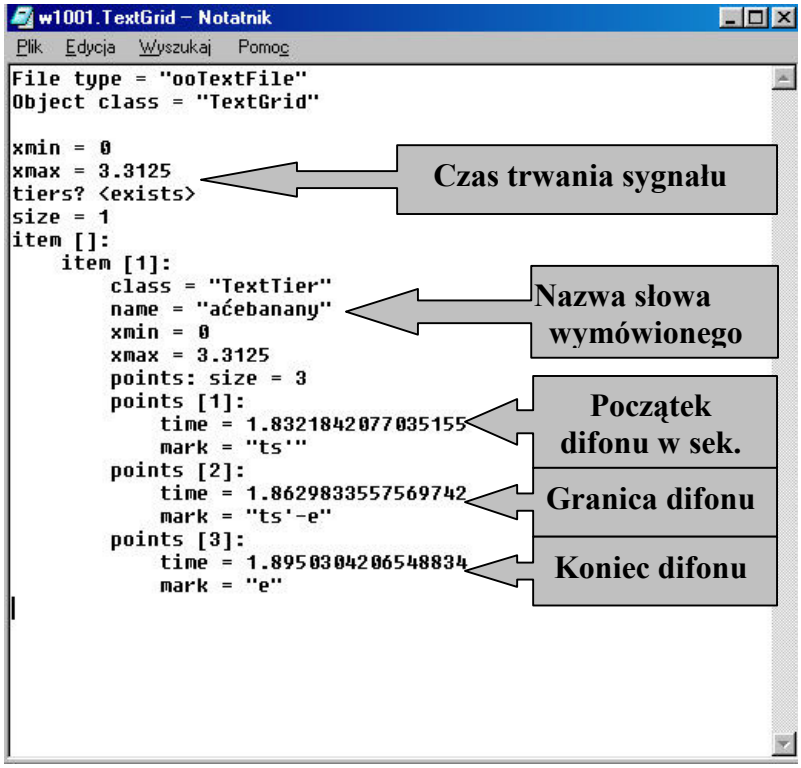
```
form supply_arguments
 sentence name  NAZAWPLIKU
 sentence dir d:\praat\ZROBIONE\
endform
Read from file... 'dir$''name$'.TextGrid
Read from file... 'dir$''name$'.wav
select TextGrid 'name$'
plus Sound 'name$'
Edit
```

4.11 Export danych - Konwersja Visual Basic

Po przeprowadzeniu procesu segmentacji należało zsynchronizować dane. To znaczy dokonać ich exportu z Praata do Diphone Studio[14]. Etap ten miał na celu przygotowanie danych do ostatniego, bardzo ważnego etapu, - testowania.
Pliki utworzone w Praacie o rozszerzeniu *.TextGrid zawierają podstawowe wiadomości na temat obrabianego pliku. Znajduje się tam czas trwania, maksimum, minimum sygnału oraz granice podane w sekundach.

[14] Opis programu Diphone Studio znajduje się w rozdziale 4.12

Poniżej przedstawiono przykładowy plik z danymi o difonie.



Rysunek 4.16 Struktura pliku z danymi opisującymi difon

Moim zadaniem było odszukanie danych czasowych z odpowiedniego miejsca z każdego z 1443 plików a następnie zapisanie ich w próbkach do pliku zawierającego listę difonów. Kod napisanego przez mnie programu w Visual Basicu znajduje się w dodatku C. Program ten zczytuje odpowiednie linie, w których znajduje się czasowo przedstawiona granica, następnie zamienia ją na próbki i wpisuje do określonego miejsca w pliku z listą difonów.

4.12 Diphone Studio

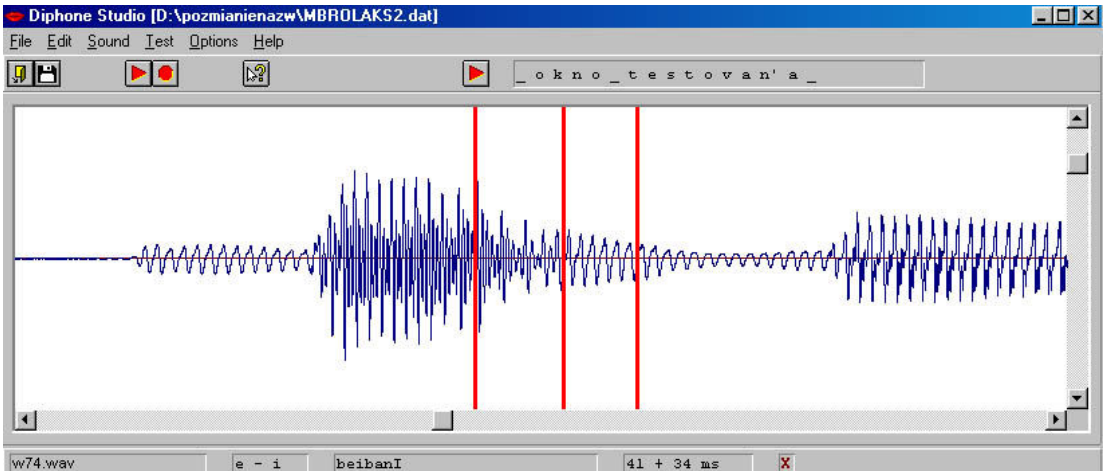
Po przekonwertowaniu danych pozostał ostatni etap przetestowania posegementowanego korpusu. Do testowania danych używałem programu Diphone Studio.(CD) Diphone Studio jest narzędziem służącym do pracy nad korpusem difonów. Program ten posiada dużo zalet, niemniej jednak ma też swoje wady.

Program umożliwia generowanie mowy przy użyciu posegementowanego korpusu. Fragment wypowiedzi należy zapisać przy użyciu transkrypcji fonetycznej. Diphone Studio generuje żądany sygnał, niestety nie daje możliwości obejrzenia go pod spektrogramem oraz zapisania do pliku *.wav. Utrudnia to testowanie danych. Brak powyższych modułów powoduje, że program nie nadaje się do przeprowadzenia segmentacji, z uwagi na utrudnienia w przeglądaniu sygnału z dużą dokładnością, oraz problemy związane z nieintuicyjną obsługą programu.

Diphone Studio daje możliwość przeprowadzenia następujących etapów:

- Nagrywania korpusu
- Segmentacji korpusu
- Preliminarny evaluation czyli testowania
- Post-processingu

Poniżej znajduje się okno programu Diphone Studio pokazujące difon „e-i”.



Rysunek 4.17 Okno programu Diphone Studio

Pad edycyjny pokazuje sygnał zawierający słowo „beibany”, nagrane żeńskim głosem.

Poziome linie oznaczają granice difonu „ei”:

Pierwszy marker reprezentuje początek difonu. Drugi marker reprezentuje granice pomiędzy dwoma fonemami w difonie. Trzeci marker reprezentuje koniec difonu.

Markery ustawia się poprzez przyciśnięcie lewego klawisza myszy, w odpowiednim miejscu sygnału.

Pionowa linia skrolująca pozwala nam przesunąć się do odpowiedniego miejsca sygnału.

Pozioma linia skrolująca pozwala powiększać dany fragment sygnału uzyskując przy tym większą liczbę detali.

Na dole znajdują się dane dotyczące aktualnie otwartego pliku:

- Nazwa pliku
- Nazwa difonu
- Słowo z korpusu, w którym istnieje dany difon
- Czas trwania poszczególnych części difonu zgodnie z zaznaczonymi markerami

Znak „X” oznaczający włączenie opcji znajdywania najbliższego przecięcia z zerem sygnału.

4.13 Testowanie

W celu skorygowania i ustalenia czy granice difonów zostały poprawnie określone musiałem przetestować cały korpus. W tym celu posłużyłem się testem składającym się z 600 słów. Test ten zawiera wszystkie połączenia, jakie mogą wystąpić w języku polskim. Słownik ten został stworzony przez pp. Ryszarda Gubrynowicza oraz Krzysztofa Maraska. Umieściłem go w dodatku A.

Podczas testowania i operacji bezpośredniej na sygnale, Diphone Studio daje możliwość bezpośredniego odsłuchu tworzonych sekwencji wyrazów. Niejednokrotnie okazywało się, że granice difonu są źle ustawione, dlatego należało wprowadzić korektę.

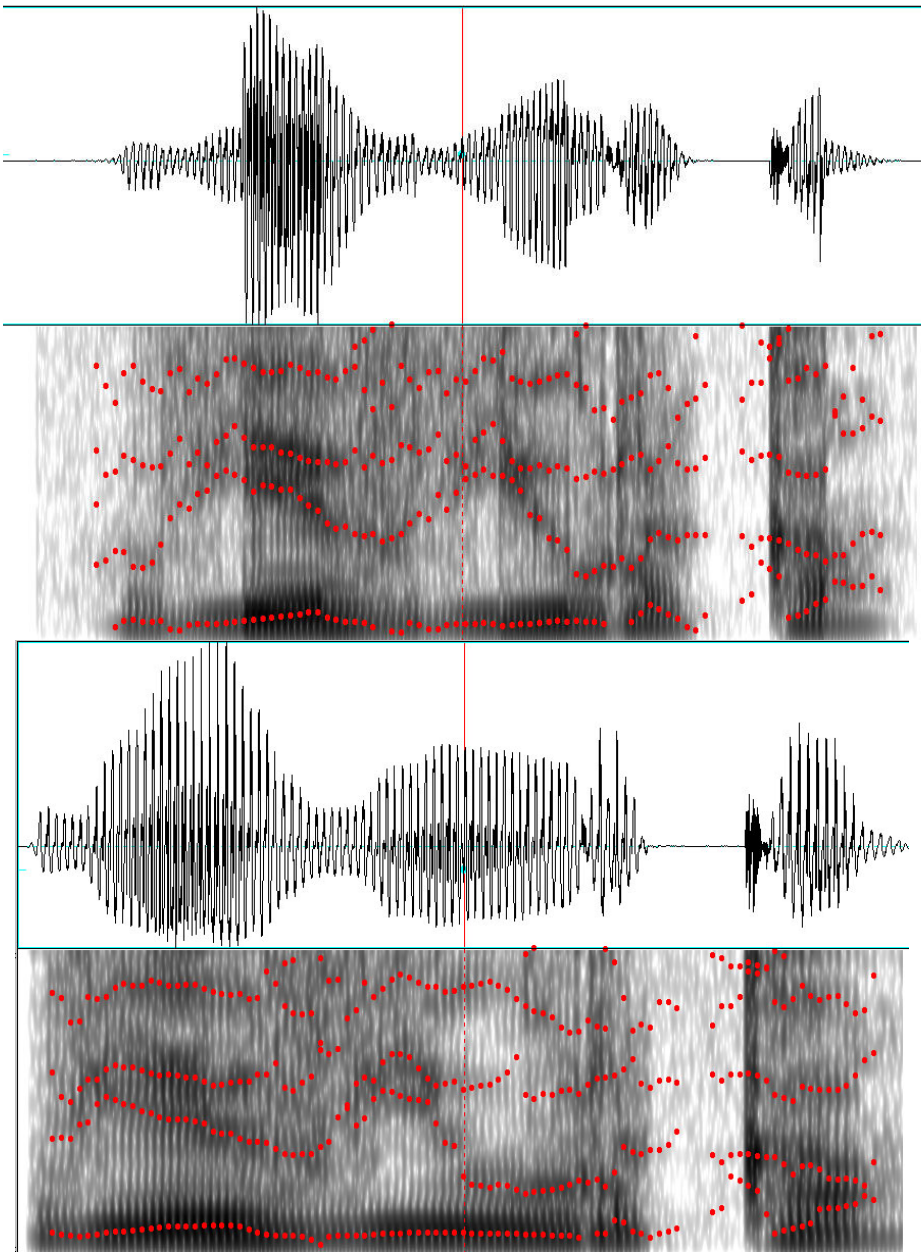
Diphone Studio umożliwia wprowadzenie transkrypcji fonetycznej danego wyrazu a następnie wygenerowanie go. Jeśli występowały jakieś trzaski szukałem granic poszczególnych difonów i ewentualnie je przesuwałem. Jeśli nie dawało to satysfakcjonującego efektu generowałem plik *.wav, który zawierał sygnał z wymawianym wyrazem i sprawdzałem jego poprawność w spektrogramie. Algorytm ten umożliwił przeprowadzenie dokładnej segmentacji i był bardzo dobrym testem na poprawność przeprowadzonej segmentacji. W ten sposób udało mi się zlikwidować większą ilość problemów.(CD)

Etap testowania danych był bardzo czasochłonny, jednocześnie stanowił ostatnią cześć pracy, którą należało przygotować.

4.14 Normalizacja bazy difonów

Tak otrzymany korpus difonów należało zapisać do formatu wymaganego przez MBROL-ę oraz wysłać do Belgii, gdzie zespół MBROL-i dokonał odpowiednich operacji na korpusie. Normalizacja bazy polegała na zmodyfikowaniu danych, usunięciu wszelkich artefaktów sygnału powodujących trzaski oraz zapisanie ich w zmodyfikowanej postaci do jednego pliku.

Obraz normalizacji, która została przeprowadzona w Belgii jest przedstawiony na rysunkach poniżej. Diametralne różnice z trzaskami zostały usunięte. Na pierwszym rysunku znajduje się spektrogram oraz sygnał w dziedzinie czasu słowa „wiewiórka" przed normalizacją, na drugim ten sam wyraz po procesie normalizacji.
Można zauważyć zastosowany proces wygładzania sygnału.



Rysunek 4.18 Sygnał nieznormalizowany i znormalizowany

4.15 Podsumowanie

W rozdziale tym umieściłem przebieg mojej praktycznej pracy. Opisałem poszczególne etapy związane z przygotowaniem wyjścia akustycznego.

Pracę mogę podzielić na cztery etapy:

Pierwszy był związany z przygotowaniem korpusu. Musiałem przygotować listę fonemów, oraz stworzyć dla nich odpowiedni kontekst. Następnym etapem było przeprowadzenie nagrań. Kolejnym etapem było wyodrębnienie difonów w nagranym korpusie, oraz wyeksportowanie do formatu akceptującego przez MBROL-ę. Ostatnim etapem był test korpusu oraz jego normalizacja.

Efekt, jaki uzyskałem jest satysfakcjonujący. Muszę przyznać, że praca okazała się bardzo interesującym i ambitnym projektem, o czym świadczy efekt, jaki uzyskałem. Podsumuję teraz wszelkie czynności związane z tworzeniem projektu.

Zakończenie

Celem pracy było stworzenie syntezy mowy polskiej opartej na bazie difonów języka polskiego dla realizacji syntezy mowy w systemie MBROLA.

Warunkiem tego było stworzenie bazy difonów dla języka polskiego w taki sposób by jakość syntezy była możliwie jak najlepsza.

Moduł akustyczny powinien mieć naturalne i zrozumiałe brzmienie, żeby był w pełni funkcjonalny w aplikacjach wykorzystujących syntezę mowy. Naturalność brzmienia jest warunkiem koniecznym do przyjęcia jej do powszechnych zastosowań.

Mając na myśli aplikacje należy tutaj wymienić dziedzinę edukacji w postaci wirtualnych uniwersytetów, portale głosowe, bezwzrokowe przekazywanie informacji. Również należy wspomnieć o syntezie mowy jako pomocy dla ludzi z zaburzeniami mowy.

Praca składała się z kilku etapów. Pierwszym było przygotowanie korpusu by można go było wykorzystać do projektu MBROL-i, co stanowiło dodatkową pracę.

Następnie należało przeprowadzić nagrania. Najbardziej skomplikowanym etapem była realizacja procesu segmentacji. Etap ten wymagał dużej precyzyjności i dokładności. Ostatni etap będący podsumowaniem całej pracy stanowił przetestowanie korpusu przy uwzględnieniu wszystkich najczęściej występujących połączeń difonów w języku polskim. Sfinalizowaniem pracy była normalizacja bazy difonów na Politechnice w Mons przez zespół MROL-i.

Baza difonów została zweryfikowana na zbiorze wyrazów zawierających najczęściej występujące połączenia w języku polskim. Dodatkowo został użyty test generowania przypadkowych zdań, co wbrew pozorom nie było zadaniem łatwym. Wyniki testu świadczą o dobrej jakości bazy difonów.

Opisana akustyka mowy polskiej oraz obszernie potraktowane zagadnienie syntezy mowy w pracy teoretycznej pozwala zorientować się w problemach, z jakimi spotykał się autor pracy. Podczas realizacji bazy difonów największym problemem było poprawne ustawienie granic jednostek akustycznych w procesie segmentacji.

Przygotowanie wyjścia akustycznego jest bardzo ważnym etapem w realizacji pełnego systemu TTS, ponieważ od niego zależy jakość generowanej mowy. Etap ten

został przygotowany, dlatego praca daje możliwość kontynuacji do uzyskania w pełni systemu TTS.

W obecnej chwili system działa dla danych wejściowych podanych w transkrypcji fonetycznej. Kolejnym etapem z pewnością będzie przygotowanie modelu przetwarzania języka naturalnego włącznie z generowaniem prozodii.

Baza difonów jest dostępna od maja br. i znajduje się na stronie internetowej MBROL-i (http://tcts.fpms.ac.be/synthesis/mbrola) jako nowy model głosowy języka polskiego w postaci difonowej.

Dodatek A - Słownik wyrazów użytych do testowania

abdykacja
absmak
Aida
alfa
auto
basniski
bąk
Bełchatów
bęcwał
bijak
bilardzista
biurko
boa
boisko
bóldżokeja
bronićniczego
bronićszansy
bryndza
bulwa
burłak
bzik
cążki
cebrzyk
cechmistrz
Cejlon
celta
cenięFelka
cenięMarka
centaur
cenzor
cewnik
cętka
chamsyn
chciwiec
cherlak
Chęciny
chińczyk
chłopcy
chłopczyk
choćby
chryzmat
chrząszcz
chuchrak
chwalba
chwilka
ciamajda
ciasnota
ciągdzbanów
ciernie
ciężar
ciśnięcie
ciżma
cło
cmok
kajmak
kajzerka
kaletnik
Kamczatka
kanwa
każdy
kąt
keson
kędy
kędzior
kilogram
klechda
kłamca
kmiot
knajpa
kocCesi
kocciężki
kocczarny
kocduży
kocHani
kocialba
kocica
kociołjasny
kocniani
kocróżowy
kocSabiny
kocsiny
kocszary
kocTomcia
koercja
kokpit
Kolchida
kołczan
kołduny
kołnierz
kołpak
kołpak
komtur
koncha
koniuszy
Konrad
kondzwonnika
kongniady
konHeleny
konjasny
konLecha
konNataszy
konPiotra
konRomana
konsiwy
konSzymona
konśukasza
konwyścigowy
konzrzędem
Kajfasz
rybdżungla
ryczałt
rydel
rydzdziwny
rydze
rydzgórski
rymarz
rysik
ryży
rzeźwy
sadzonka
sadzulec
sąniskie
sąruskie
sąsiad
serdak
sęp
sierpy
silos
siunia
skądże
składnia
słońce
słuchaćDoroty
słuchaćpana
służbdżungla
spółdzielnia
sroka
staramsięitak
stypendium
synbarona
syndziekana
synJanka
synMaćka
synPiotra
synsiostry
synśukasza
synZiuty
sytuacja
szajka
szeregzielny
szlif
szmugler
sznaucer
szparka
szron
sztych
szwaczipłótno
szwaczleniwy
szwaczRadek
ściągacz
śmie
śmieć

cnota
cofnięty
comber
córcia
cudeńka
cudzy
cyngiel
cytronada
czaszka
czciciel
czeremcha
człek
czoło
czujęrosół
czujnik
czyjż
ćwierćnuta
dajna
dajrękę
Dalmacja
darmozjad
dąbrowa
dążność
deizm
dębnik
diablę
dingo
dłońniani
dłońZiuty
dłużnik
dno
dojrzałość
domknięcie
dorsz
dosiebny
doustny
drewno
driada
druhna
dryl
drzazga
drzwi
duet
Dulcynea
dusznica
dwojga
dwóchczerwonych
dwóchnaszych
dwóchpanów
dwóchsilnych
dwóchsuchych
dwukropek
dyngus
dynia
dyscyplina
dzban
dziąsło
dzieńdziekana
dzieńmamy
dziesięćkrot
kończółty
kooptacja
kopciuch
kosićlubin
koszCesi
koszęjęczmień
koszistół
koszsiostry
koszstary
kościęsto
kotCesi
kotciężki
kotpana
kotsiostry
kozioł
koźle
krabmorski
kruganek
krzywda
krzywdzenie
krzywdziciel
książe
Kuczma
kupfabrykę
kupmebel
kutia
łajba
łaźnia
łącznik
łękotka
łękotka
łokcie
łośszary
majcher
małedżemy
małża
mamcia
mamdzban
mamdżem
mamzięcia
melba
mędrzec
miałdzban
miałdżem
miałdżonkę
MieczysławDzwonek
miećnotes
miećsiano
miećsowę
miedza
mierzwa
mizdrzyć
mizia
mleczko
morwa
mójdzień
mójdżem
mójwygląd
mysz
mżawka
tanidżem
tenżółty
teżdzielny
tęsknota
tik
tnący
Toruń
trafięnafront
trądzik
trencze
trik
triumwirat
trzpień
trzy
trzyaktowy
trzyosiowy
twójzięć
tychkotów
uchwyt
uczta
udrzeć
udział
ulżenie
uniknęli
upał
uziom
Wachock
wachta
walnąć
wańtuch
warkoczCesi
warkoczHani
warkoczsiostry
warkoczSoni
wąsik
wąsstarego
wąwóz
wąziutki
wążciemny
wbity
wchacie
wczasowicz
wdrzewie
wedługdzidzi
wełna
weźżabę
węch
węgiel
węszyć
węże
wgięcie
wibrator
widzająśniej
Widzew
widzęwelon
widziećCesię
widziećciało
widziećlepiej
widziećRadka
widzmały

dzwon
dźwignia
dżinsy
dżokej
dżudo
ekshalacja
Eljasz
Elka
encyklika
fajfer
fajtłapa
faktura
falszkil
fanfara
farba
farsa
farsidło
fartuch
fąfel
felczer
ferryt
feudał
figus
flądra
fluid
folga
fontanna
fosfor
fragment
Franio
frędzle
futbol
fuzle
gawron
gąszcz
gbur
Gdańsk
gdzie
gejsza
geodeta
gęślarz
gibkość
giełda
giemza
gilza
głośność
gnida
gonciarz
gorczyca
goździk
gracki
groźba
gwożdżenie
hacle
hakhelski
hańba
harfa
Helcia
hełmy
hetman
nadaldziękuje
nadaldzwoni
nafta
naszły
nędznik
nicwięcej
niemalpewny
nikczemnik
nimfa
obdzierganie
obgadanie
obgadanie
odbiórdżemu
oddźwięk
odsetka
odznaka
odżywka
ojciec
Ojców
okucie
olśnienie
omdlenie
omłoty
omszenie
osiemdziesiąt
pałłaty
palrupiecie
pandżokej
pańcia
pąs
pejcz
pełzanie
pętla
piarżysko
pichcenie
pieszczot
piłka
PiotrZięba
płótno
pnącz
pniak
pociotek
początek
poczęstunek
poczwara
podczaszy
poddasze
podgrzybek
podział
podziękować
pomącić
pomruk
pomyłka
poręczyciel
poszum
poszwa
pośpiech
pośrednik
poświst
potwierdzać
wiedząlepiej
wiedźma
wierząmocniej
wierzchołek
więź
więźdomowa
więźGustawa
więźmamy
więźwujka
władza
wmowie
wodze
wodzLechitów
Wołga
wońdaleka
wódzJan
wódzradosny
wódzzajadły
wódzzimny
wpływdżokeja
wsiąkanie
wsuwka
wśróddżonek
wychowawca
wydrzeć
wydział
wyewakuować
wyimek
wyrok
wyrzut
wytchnienie
wyuzdanie
wyziew
wyżreć
wzrok
wżer
zaakceptować
zaciąłsię
zaczęły
zaćma
zadźgać
zagłówek
zakwas
zamglenie
zAniąNowicką
Zaolzie
zapchlony
zawleczka
zawzięty
ząbki
ząbwielki
ząbwielki
zbieg
zdrzemnąć
zdziczenie
zegarynka
zewpolski
zewsząd
zeznanie
zęby

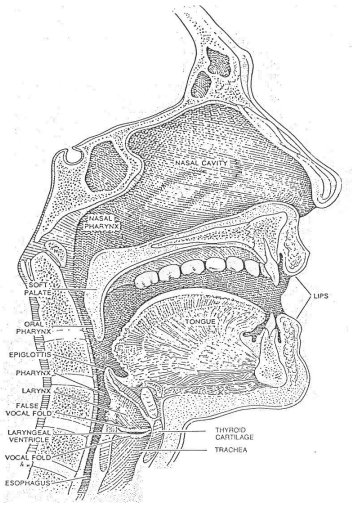
hiena
Hilda
hucpa
hufnal
huśtawka
hyslina
Idzi
Ignacy
impuls
infułat
inlet
inszość
ircha
iszjas
Jaśka
jedźzanim
jeździćwolniej
język
kabłąk
kabza
kadzidło
kaemy
pozór
pożoga
półcień
półidiota
półlitrówka
półrocze
półsen
półziemny
presja
proca
przedmieście
psiak
radża
rajski
rajza
rano
robićtosamo
rozczynnik
rozdzwonić
rógdziwny
rógdżonki
rybdzban
ziarno
ziąb
zięba
ziszczenie
Ziuta
zjakimśchłopem
złom
znicz
zohydzenie
zszywka
związek
zwiechąisłomą
zygzak
zziajany
zziajany
zżycie
źrebak
żądza
żiga
żleb
żłobek

(Źródło: Gubrynowicz R. i Marasek K.)

Dodatek B – Streszczenie pracy w języku angielskim

Polish-Japanese Institute of Information Technology

MBROLA.
Creating the Polish diphone database for speech synthesis.
Summary of Master's thesis.



Krzysztof Szklanny

Warsaw 2002

1. Preparing the Polish diphone database

The goal of my work was to create the Polish diphone database. Additional aim of the project was to obtain high quality speech synthesis for the Polish language. The whole process included several stages. First of all I had to prepare a phoneme list and create the corpus of diphones. Then next was to make the recordings of the corpus and the segmentation of diphones. One of the last processes was to test the database then export it and sent it to the MBROLA team who conducted the normalization process.

2. Preparing the corpus

The most important thing at this stage was to create such a context of nonsense words including diphones that, the diphone could not be the stressed syllab. Secondly it's neighborhood should not influence the co-articulation of the diphone.
Sometimes there were no rules that how to find the appropriate context of the diphone and that was the most difficult problem while creating the corpus.

The corpus includes such information as :
the name of the diphone, context of the diphone, number of wave file where it is placed, and the three numbers which are :

- The beginning
- The middle
- The end of the diphone

For example:

i i w0.wav ani imadwo 41658 42577 43144

These numbers are presented in samples. In order to get the real time of the diphone it is necessary to divide the number by 16000 which is the frequency of recorded corpus. Such a frequency allowed me to obtain the best quality of speech synthesis and preserve a good quality/size ratio.

As I mentioned, the next thing to do was to make recordings. These took place in the Polish-Japanese Institute of Information Technology in it's recording studio.
Four microphones were used. A so called close-talk microphone and three table stand microphones. In the segmentation process I used the close-talk microphone, which guaranteed the best quality signal and made no distortion.

The next stage was to prepare the segmentation process which was the most difficult and time-consuming stage.

3. Segmentation

For ***manual*** segmentation, the sound elements are taken from the speech material "by hand". Generally I used Praat[15] as a tool for segmentation. It has a built-in spectrogram and uses a graphic-acoustic display for marking the parts of the signal to be cut out and for acoustical control. During manual segmentation, considerable difficulties might occur. The main difficulties arise, if there are no sharp boundaries between the individual sounds (flowing sound transitions). In such cases, marking of the boundaries is often arbitrary.

Manual segmentation is very lavish and time consuming. That's why, ***automatic*** segmenting procedures have been developed. These procedures do not achieve the reliability of a good expert. However, the advantage of an automatic procedure is that it can make suggestions which facilitate the work for the user.

[15] Praat is a program for doing acoustic phonetic.

4. Characteristics of sound classes

For the segmentation, knowledge of the characteristics of the sound classes is necessary. The sound classes (subdivided according to the type of articulation) are specified in the table below with their characteristics in the time and frequency domain:

type of sound	example	time domain	frequency domain
vowels	/a/ /e/ /i/	- quasi-periodic - high energy	- 3 - 6 clearly visible formants - significant pitch peaks - main energy in the low frequency range
voiced plosives	/b/ /d/ /g/	- sudden and high rise of the amplitude	- formant at low frequencies - rise of the formants after closure opening
unvoiced plosives	/p/ /t/ /k/	- sudden and high rise of the amplitude - no signal before closure opening	- no formant structure - energy in high frequency area
Voiced fricatives	/v/ /z/ /j/	- broad quasi-stationary area - low energy	- existing formants - lower first formant compared to vowels - energy in high frequency area
unvoiced fricatives	/f/ /s/	- noise-like form - low energy	- broad spectrum
Nasals	/m/ /n/	- quasi-periodic - similar to vowels - lower energy than vowels	- minima within the spectrum - formants similar to vowels
liquids	/l/ /r/	- short or missing stationary area - weak transition to a following vowel	- lower first formant

Table 4.1 Characteristics of sound classes

5. Conventions

The segmentation of the sound elements should be executed very carefully because the quality of a speech synthesis system strongly depends on correct segmentation. The following fundamental conventions have to be considered.

Cutting must be done always in a positive zero crossover. Thus, errors at concatenation boundaries can be avoided or reduced when assembling the sound elements.

The cut has to be made at the beginning of a new period. The figure below shows this principle.

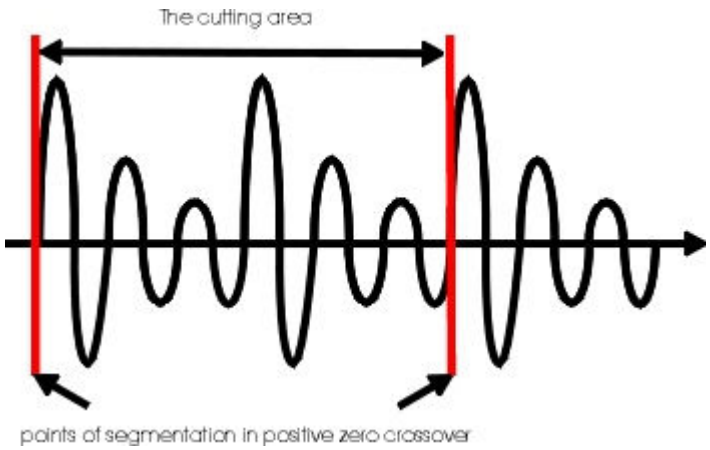


Figure 5.1 Conventions for the segmentation of sound elements

The segmentation of ***vowels***, ***nasals*** and ***laterals*** should be made in the zero crossover before the pitch mark (The figure 5.2).

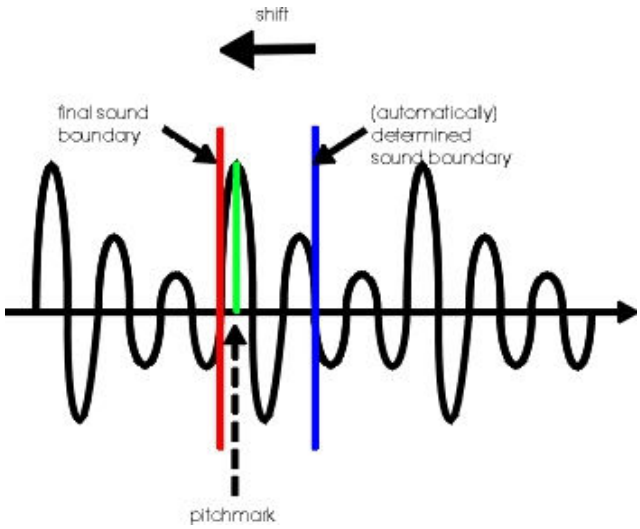


Figure 5.2: Conventions for the segmentation of vowels, nasals and laterals

The segmentation of ***fricatives***, ***affricates*** and ***intermittends*** should be selected as follows (The figure 5.3):

- If the boundary proposed by the system shows positive values, it has to be shifted to negative time direction.
- If the boundary has negative values, it has to be shifted in positive time direction to the next positive zero crossing.

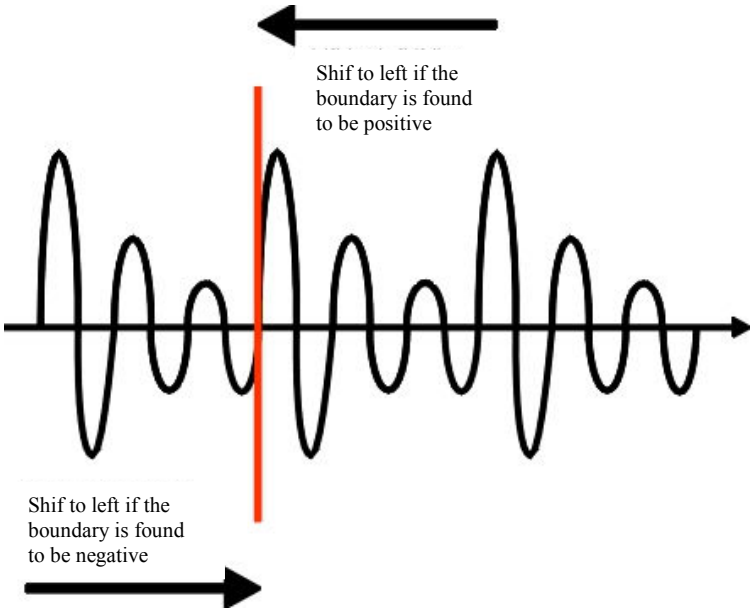


Figure 5.3: Conventions for the segmentation of fricatives, affricates and intermittends

Plosives are segmented similarly. The boundary should be shifted always temporally forward. The sudden rise of the amplitude at the start of the sound is particularly critical for the heard impression of the sound.

6. Problems

After the concatenation of the segmented sound elements the result is often unsatisfactory, because discontinuities at the sound boundaries cause clearly audible disturbances. Discontinuities can occur on the following items:

- amplitude

The ***discontinuities of amplitude*** are already visible in the time domain. They are produced, if the amplitudes at the end of a sound and at the beginning of the following sound are strongly different. It is clearly audible as “cracks”. The figure 6.1 shows a discontinuity of the amplitude in the time domain.

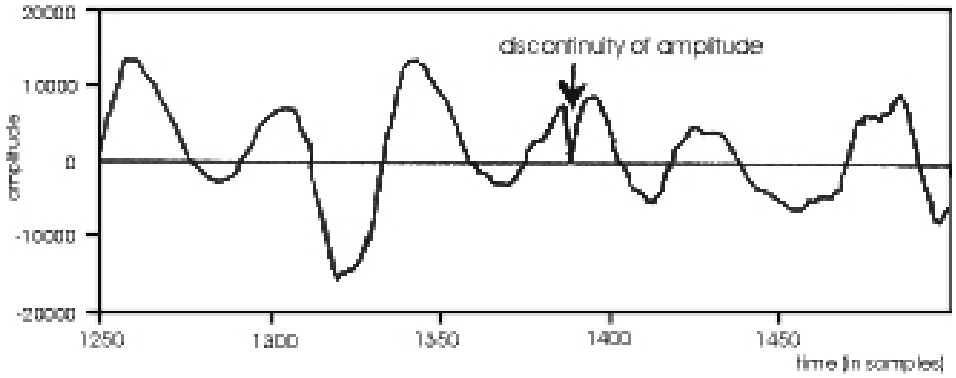


Figure 6.1: Discontinuity of the amplitude

- energy

The ***discontinuities of energy*** are produced by different volumes of the speech material. Great changes usually exist over time. The figure 6.2 shows a discontinuity of the energy in the time domain and the figure 6.3 the power over time.

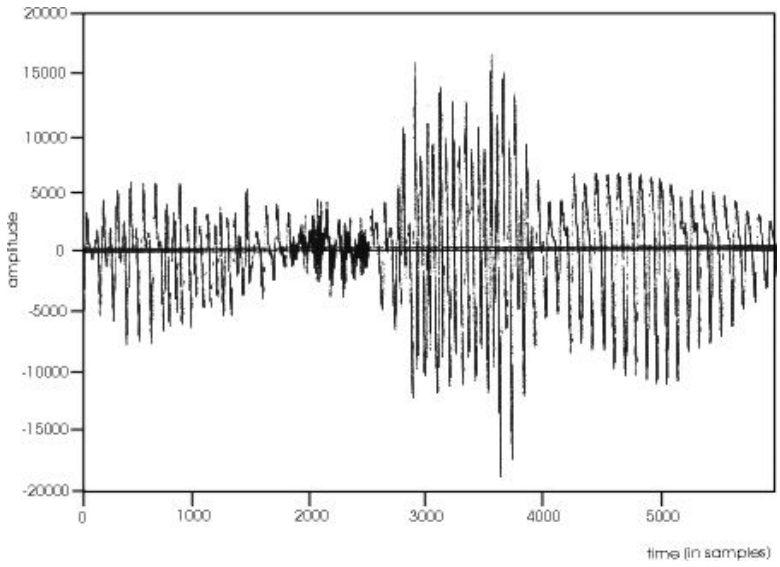


Figure 6.2: Discontinuity of the energy (time domain)

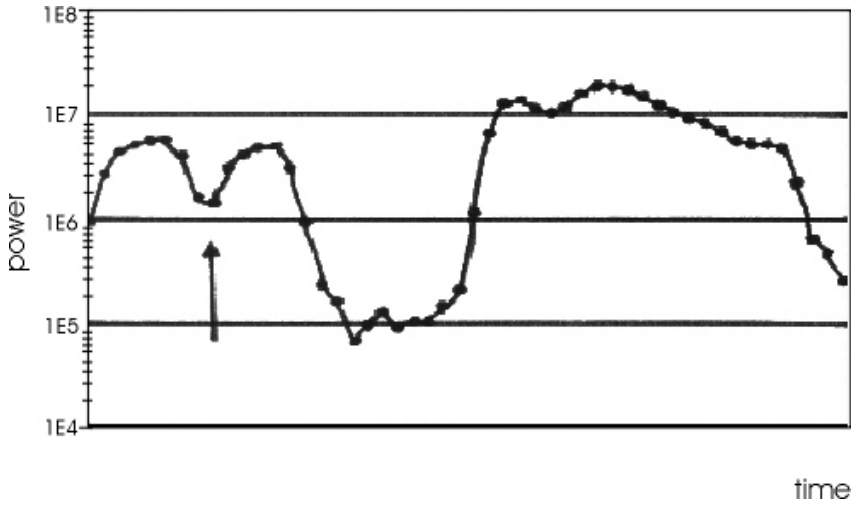


Figure 6.3: Discontinuity of the energy (power over time)

- frequency

The ***discontinuities of the frequency*** are very short, but they are clearly audible as cracks. The figure 6.4 shows an example.

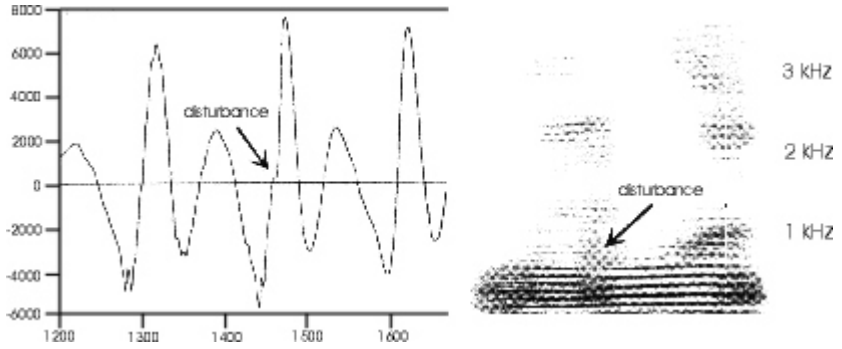


Figure 6.4: Discontinuity of the frequency (time and frequency domain)

- phase

Discontinuties of phases occur, if the boundaries are not set at the beginning of a new pitch period. They are also clearly audible as cracks. The figure 6.5 shows the effect of such a discontinuity ("additional" area).

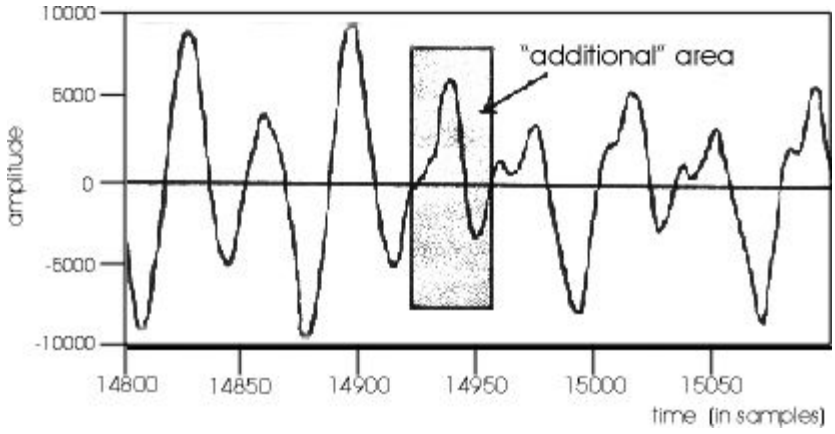


Figure 6.5 Discontinuity of the phase

These disturbances can be widely reduced or avoided by careful segmentation.

7. Testing

The last stage was to test the quality of the diphone database in the Diphone Studio program (the figure 7.1) and send the database to MBROLA in order to normalize it. So I used the test made by Krzysztof Marasek and Ryszard Gubrynowicz. It consists of all the possible connections that exist in the Polish language and then sent to Belgium.

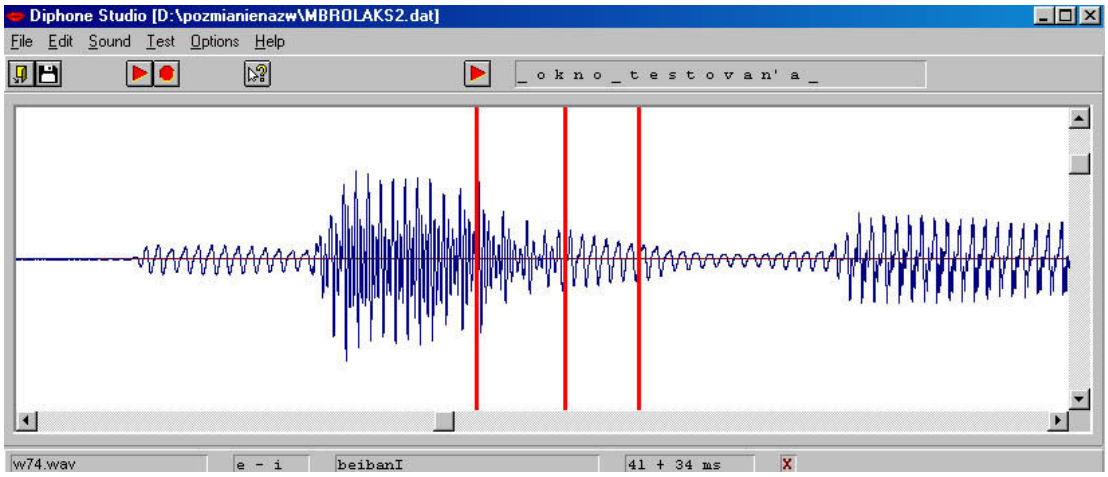


Figure 7.1 Diphone Studio window

8. Conclusions

The goal of my work was to create the Polish speech synthesis based on diphones. This is for the realization of speech synthesis for the MBROLA system. The requirement was to create the diphone database that generates speech synthesis in the best possible way.

The quality of speech synthesis must be intelligible, and it has to sound natural so that the acoustic model could be further used in public applications, which use speech synthesis. By public use I mean the using of the database for education, voice portals, Enhanced Eyes-Free Access to Critical Information While Driving, and also an aid to the people with speech disabilities.

There were a few stages that had to be done. First of all I had to prepare the corpus that could be used further in the MBROLA project, which was additional work.

Next I had to conduct the recordings. The most sophisticated stage was the realization of the segmentation stage. It required accuracy and precision. The last stage was to test the quality of speech synthesis by using the most popular connections of diphones in the Polish language. Finalization of the work was the normalization of the database by MBROLA in Polytechnic in Mons.

Now the speech synthesis system works for the input data as phonetic transcription. The next stage will be the creation of the prosodic model and natural language processing in order to create the full TTS system.

The polish database is available since May this year on the MBROLA web site (http://tcts.fpms.ac.be/synthesis/mbrola) as new voice model of polish language.

Dodatek C – kod źródłowy skryptów

Poniżej znajduje się kod źródłowy programu służącego do exportu danych:

```
Dim tab_lica(3) As String  'tablica o wart. od 0 do 3
Dim freefileint As Integer
Dim freefileint2 As Integer
Dim filename As String
Dim dupa As String
Dim znak As String
Dim licznik As Integer
Dim linia As String
Dim linia_bez_konca As String
Dim pierwszapoz As Integer
Dim drugapoz As Integer
Dim trzeciapoz As Integer
Dim x As String
x = Asc("    ")
freefileint = FreeFile ' daje kolejny numer pliku
filename = "c:\magisterka\w352a_16000.textgrid"
Open filename For Input As #freefileint  'freefileint zmienna ktora przechowuje nazwe pliku obecnie otw. pliku
licznik = 0
wiersz = 15 ' podać o jedna linie  mniej niz trzeba
Do While Not EOF(freefileint) And licznik < wiersz
    znak = Input(1, #freefileint)
    If Asc(znak) = 10 Then licznik = licznik + 1
Loop
tab_lica(1) = Input(24, #freefileint)
MsgBox tab_lica(1)
' druga wartosc
licznik = 0
wiersz = 3 ' podac o jedna linie  mniej niz trzeba
Do While Not EOF(freefileint) And licznik < wiersz
    znak = Input(1, #freefileint)
    If Asc(znak) = 10 Then licznik = licznik + 1
Loop
tab_lica(2) = Input(24, #freefileint)
MsgBox tab_lica(2)
' trzecia wartosc
licznik = 0
wiersz = 3 ' podac o jedna linie  mniej niz trzeba
Do While Not EOF(freefileint) And licznik < wiersz
    znak = Input(1, #freefileint)
    If Asc(znak) = 10 Then licznik = licznik + 1
Loop
tab_lica(3) = Input(24, #freefileint)
MsgBox tab_lica(3)
Close #freefileint
```

```
'input odczyt
'output zapis
filename = "c:\magisterka\ola.txt"
Open filename For Input As #freefileint
freefileint2 = FreeFile
Open "c:\magisterka\ola3.txt" For Output As #freefileint2
 ' podac o jedna linie  mniej niz trzeba
 znak = Input(1, #freefileint)
 Do While Not Asc(znak) = 10
    linia = linia & znak
    znak = Input(1, #freefileint)
 Loop
 linia_bez_konca = Mid(linia, 1, Len(linia) - 1)
 tab_lica(1) = Mid(tab_lica(1), 20, Len(tab_lica(1)))
 'zamiana na probki oraz zamian ze str na wartosc
 pierwszapoz = Val(tab_lica(1)) * 16000
 'MsgBox pierwszapoz
 tab_lica(2) = Mid(tab_lica(2), 20, Len(tab_lica(2)))
 drugapoz = Val(tab_lica(2)) * 16000
 'MsgBox drugapoz
 tab_lica(3) = Mid(tab_lica(3), 20, Len(tab_lica(3)))
 trzeciapoz = Val(tab_lica(3)) * 16000
 'MsgBox trzeciapoz
  linia_bez_konca = linia_bez_konca & Chr(32) & pierwszapoz & Chr(32) & drugapoz & Chr(32) & trzeciapoz
  Print #freefileint2, linia_bez_konca 'print wpisuje do pliku
  MsgBox linia_bez_konca
  wiersz = wiersz + 1
 Close #freefileint2
   End Sub
```

Bibliografia

Basztura, Cz. (1996). *Komputerowe systemy diagnostyki akustycznej*. Warszawa: PWN *(KSDA)*

Domalewski, W. *Numery z Internetem* PCKurier 21/2000) *(PC)*

Fant G. *Acoustic Theory of Speech Production,* The Hagues:Mounton (*AToSP*)

Gubrynowicz, R. Wykład *Podstawy Fonetyki Akustycznej (PAF)*

Kleszewski Z.(1995). *Podstawy akustyki.* Skrypt. Wydawnictwo Politechniki Łódzkiej *(PA)*

Laver J. *Principles of phonetics,* Oxford University Press., Oxford, UK (*POP*)

Marasek, K. Wykład *Werbalna komunikacja z komputerem (WKK)*

Marasek, K. *Egg and voice quality (EGG)*

Olivier, D. *Polish Text To Speech Synthesis M*.Sc. Speech and Language Processing

Płoski Z. *Słownik Encyklopedyczny - Informatyka* Wyd. Europa (1999).

Tadeusiewicz, R. *Sygnał mowy* (*SM)*

Wierzchowska, B. (1967). *Opis fonetyczny języka polskiego*. Warszawa: PWN *(OFJP)*

Xuedong Huang, Alejandr Acero, Hsiao-Wuen Hon (1999). *Spoken Language Processing*

http://www.phon.ucl.ac.uk/home/sampa/polish.htm *Alfabet Sampa*

http://www.rider.edu/users/suler/psycyber/psyav.html *The Psychology of Avatars and Graphical Space in Multimedia Chat Communities*

http://tcts.fpms.ac.be/synthesis/introtts.html *A Short Introduction to Text-to-Speech Synthesis*

http://tcts.fpms.ac.be/synthesis/mbrola/mbruse.html *Intonacja w MBROL-i*

http://xroads.virginia.edu/~HYPER/POE/kempelen.html *Dzieło von Kempelena*

http://www.ling.su.se/staff/hartmut/kemplne.htm *Biografia Von Kempelena*

www.mindspring.com/~dmaxey/ssshp/ss_home.htm *Początki syntezy mowy*

www.mindspring.com/~dmaxey/ssshp/ss_home.htm *Historia syntezy mowy*

http://www.research.att.com/projects/tts/ *strona AT&T Labs*

http://www.speechworks.com - *strona firmy Speechworks dotycząca syntezy mowy*

http://www.lhsl.com/realspeak/ - *strona firmy ScanSoft, dotycząca syntezy mowy Real Speak*

http://www.ias.et.tu-dresden.de/kom/lehre/tutorial/selection.htm - *Segmentacja korpusu*

http://tcts.fpms.ac.be/synthesis/mbrola.html - *strona MBROL-i*

Spis rysunków