Praca powinna być cytowana jako:
Szklanny, K., 2009. Optymalizacja funkcji kosztu w korpusowej syntezie mowy polskiej. Rozprawa doktorska. Polsko-Japońska Wyższa Szkoła Technik Komputerowych.

Polsko-Japońska Wyższa Szkoła Technik Komputerowych

Optymalizacja funkcji kosztu w korpusowej syntezie mowy polskiej

Krzysztof Szklanny

Rozprawa doktorska

Opiekun naukowy:

Dr hab. Krzysztof Marasek

Warszawa, wrzesień 2009

Mojej Mamie

i Mojemu świętej pamięci Tacie

Podziękowania

Szczególne podziękowania kieruję do mojego promotora prof. Krzysztofa Maraska, za opiekę naukową, cierpliwość, wsparcie oraz pomoc okazaną mi w trakcie realizacji tej pracy.

Bardzo chciałbym podziękować Dominice Oliver, która pozwoliła na wykorzystanie modułów z jej pracy doktorskiej. Mimo dużej odległości, współpraca zaowocowała kilkoma wspólnymi artykułami na konferencjach krajowych i międzynarodowych.

Specjalne podziękowania kieruję do Nickolaya Shymreva, za wsparcie merytoryczne oraz programistyczne w środowisku Festival.

Chciałbym również podziękować Łukaszowi Brockiemu oraz Danijelowi Koržinkowi za okazaną pomoc przy optymalizacji funkcji kosztu i wsparcie w dziedzinie algorytmów ewolucyjnych.

Serdeczne podziękowania kieruje do Michała Wójtowskiego, który włączył się czynnie w pracę realizowanego syntezatora a jego pomoc zaowocowała ukończeniem przez niego pracy magisterskiej. Współpraca ta była dla mnie nieocenionym dopingiem.

Bardzo chciałbym podziękować moim bliskim Mamie, Wujkowi, Pawłowi za wspieranie w trudnych momentach realizacji tej pracy.

Chciałbym podziękować mojej Monice za miłość oraz za to, że nigdy nie zwątpiła w pomyślne zakończenie tej pracy.

Badania przedstawione w pracy zostały zrealizowane w ramach grantu promotorskiego nr 0641/T02/2006/31 przyznanego przez Ministra Nauki i Szkolnictwa Wyższego

Wprowadzenie

Technologie głosowe są na świecie rozwijane co najmniej od połowy lat 70-tych. Ich główną zaletą jest możliwość stworzenia głosowej interakcji między użytkownikiem a komputerem.

Text-to-speech system jest modułem konwersji tekstu na mowę. Wykorzystuje się tą technologię do generowania dźwiękowej postaci danych tekstowych. Dzięki temu można tworzyć portale głosowe, czy też aplikacje z głosowym interfejsem. Celem nowoczesnych projektów jest zapewnienie takiej jakości syntezy, by słuchający nie był w stanie odróżnić mowy syntetyzowanej od naturalnej (Turing 1950). Z oczywistych powodów nie jest możliwe stworzenie i nagranie wszystkich form i wszystkich słów dla danego języka, stąd konieczność syntezowania mowy. System TTS definiuje się jako system automatycznego generowania mowy z tekstu ortograficznego, z modułem transkrypcji fonetycznej oraz modułami odpowiedzialnymi za prozodię i intonację.

Istnieje kilka metod generowania syntetycznej mowy. Obecnie stosowane są dwie technologie. Pierwsza, zwana regułową syntezą mowy, polega na jej generowaniu poprzez układ symulujący ludzki aparat mowy o zmiennych parametrach. Druga, zwana konkatenacyjną syntezą mowy polega na łączeniu jednostek akustycznych wybieranych z bazy nagrań głosu naturalnego. Synteza korpusowa jest szczególnym rodzajem syntezy konkatenacyjnej (Szklanny i wsp. 2008).

W syntezie korpusowej baza językowa jest znacznie większa i zawiera posegmentowane wypowiedzi, na segmenty akustyczne o różnej rozciągłości (np. głoski, difony, trifony, sylaby, wyrazy, całe zdania). Ta sama jednostka występuje wielokrotnie. Chcąc wygenerować zadaną wypowiedź dobierane są takie jednostki, które minimalizują wartość globalnej funkcji kosztu. Funkcja ta zwykle składa się z dwóch części: kosztu doboru jednostki oraz kosztu konkatenacji. Według badań przeprowadzonych dla języka angielskiego (Clark i wsp. 2007) wynika, iż w przypadku kosztu doboru jednostki najistotniejszym parametrem jest akcent. Waga akcentu powinna być jak największa. O ile w

języku angielskim dominuje akcent melodyczny, o tyle w polskim jest pewna swoboda w jego realizacji – może mieć on formę melodyczną lub dynamiczną. Istotne znaczenie zajmuje pozycja w frazie. Zatem obecnie zasadniczym problemem w syntezie mowy nie jest stworzenie mowy zrozumiałej, a uzyskanie jej jakości powszechnie akceptowalnej. O tym decyduje poprawna wymowa i właściwe akcentowanie.

Funkcja kosztu konkatenacji wyznacza jakość połączenia na podstawie czasu trwania jednostek akustycznych tworzących łączone fragmenty, ich intonacji, konturu widma oraz energii. Na ogół modyfikacje prozodyczne sygnału nie są konieczne (w przypadku syntezy korpusowej), co przekłada się na dużą naturalność brzmienia generowanej mowy. Metoda selekcji jednostek (ang. *unit selection*) jest najbardziej efektywną i popularną metodą syntezy konkatenacyjnej.

Głównym celem pracy było zoptymalizowanie funkcji kosztu w korpusowej syntezie mowy dla języka polskiego. W celu realizacji tego zadania należało przygotować kompletny system syntezy korpusowej. Proces ten obejmował etap przygotowania korpusu, realizacje nagrań, segmentację bazy językowej. Jakość segmentacji została zweryfikowana w prototypowym syntezatorze. Następnie przygotowano nowy głos w środowisku Festival wykorzystując nagraną bazę akustyczną. Praca ta zawierała realizację nowych modułów, jak i dostosowanie już istniejących do wymogów syntezy mowy polskiej. W ten sposób powstał kompletny system korpusowej syntezy mowy. Następnie zoptymalizowano funkcję kosztu wykorzystując do tego algorytm ewolucyjny. Efekt badań został potwierdzony percepcyjnym testem jakości syntetycznej mowy typu MOS (ITU 1996) (ang. *mean opinion score*).

W pracy zostały postawione trzy tezy:

- funkcję kosztu można optymalizować za pomocą metod heurystycznych. Jedną z metod optymalizacji jest metoda oparta na algorytmie ewolucyjnym
- optymalizacja funkcji kosztu ma istotny wpływ na poprawienie jakości syntezy korpusowej
- wybór odpowiedniego mówcy oraz jakość bazy akustycznej ma bardzo duży wpływ na finalną jakość generowanej mowy

Pierwszy rozdział pracy jest wprowadzeniem do opisu sygnału mowy. Przedstawiono w nim budowę narządu mowy oraz specyfikę języka polskiego. Omówiona została klasyfikacja dźwięków mowy. W dalszej części opisano reguły transkrypcji fonetycznej. Następnie przedstawiono modele opisu prozodii języka polskiego. W końcowej części rozdziału przedstawiono rodzaje jednostek akustycznych używanych w syntezie mowy oraz podstawowe modele opisu prozodii.

W drugim rozdziale zaprezentowano historię syntetyzatorów mowy. Opisano podstawowe rodzaje syntezy, a także dokonano analizy działania systemu TTS oraz jego poszczególnych modułów. System TTS (według (Dutoit 1997, Taylor 2009)) definiuje się jako automatyczny proces generowania mowy od momentu transkrypcji zdania aż po jego wypowiedzenie.

Rozdział trzeci stanowi wprowadzenie do jednej z najważniejszej funkcji w korpusowym syntezatorze mowy - funkcji kosztu.

W rozdziale czwartym przedstawiono szereg zadań, które twórca systemu korpusowej syntezy mowy musi rozwiązywać. Opisano sposób tworzenia korpusu, rejestracji nagrań oraz ich segmentacji. Przedstawiono również automatyczną metodę korekty posegmentowanych nagrań.

W systemach korpusowych istnieje kilka sposobów optymalizacji funkcji kosztu. Pierwszy z nich polega na intuicyjnym dobieraniu parametrów oraz przeprowadzaniu kontrolnych testów percepcyjnych, które mają umożliwić wyznaczenie najlepszych pod względem percepcyjnym współczynników wag. Drugim sposobem jest metoda automatyczna polegająca na trenowaniu poszczególnych wag kosztu doboru jednostki. W rozdziale piątym opisano strukturę i sposób działania algorytmu ewolucyjnego. Przedstawiono strategię (μ+λ) (Michalewicz 2004) wykorzystaną w procesie optymalizacji funkcji kosztu oraz sposób przeprowadzenia badań optymalizacyjnych.

W rozdziale szóstym zinterpretowano oraz dokonano analizy wyników badań. Wyniki tego testu wskazują, iż strategie ewolucyjne są skuteczne w procesie optymalizacyjnym i wygenerowane parametry dla funkcji kosztu potwierdziły to w badaniach testowych.

Rozdział siódmy zawiera opis testu percepcyjnego MOS, którego

wyniki potwierdziły skuteczność wykonanych badań optymalizacyjnych dzięki, którym uzyskano lepszą jakość syntetycznej mowy polskiej. Przedmiotem badań testowych jest porównanie 3 różnych funkcji kosztu, ocenia jakość sygnału syntezy mowy uzyskanej na drodze resyntezy, oraz nagrań pochodzących z bazy akustycznej.

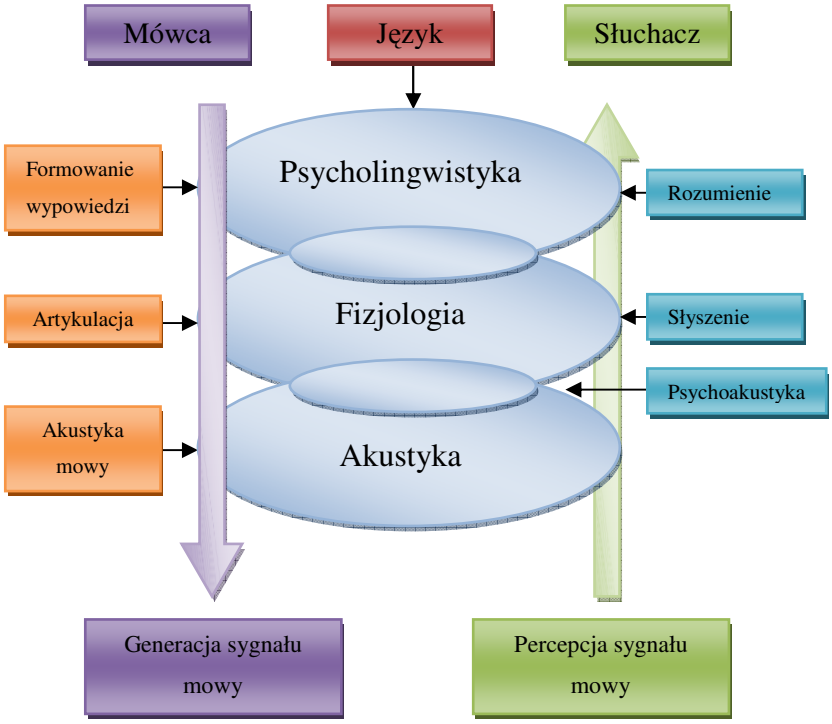
1 Sygnał mowy i jego opis fonetyczny

”Mowa jest jednym z wielu sposobów przekazywania informacji. Specyfiką mowy jest to, że ma postać dźwiękową. Jest zawsze kodowana w postaci ciągu dźwięków o określonych charakterystykach. Kod jest specyficzny dla danego języka, co powoduje, że każdy język ma określony dla siebie zbiór dźwięków mowy.” (Gubrynowicz 2004)

Badanie oraz analiza fal dźwiękowych generowanych przez ludzki narząd mowy w celu komunikacji z otoczeniem, jest domeną fonetyki akustycznej. Jest to techniczny dział nauki o języku, jakim jest lingwistyka. Fonetykę dzieli się na działy według obszarów badań, które niejednokrotnie przenikają się z innymi dziedzinami wiedzy jak np. fizjologią czy akustyką:

- fonetyka akustyczna - zajmuje się badaniem cech fizycznych (akustycznych) dźwięków mowy
- fonetyka artykulacyjna - zajmuje się sposobem wytwarzania dźwięków przez narządy mowy, czyli artykulacją
- fonetyka audytywna - zajmuje się analizą percepcji tychże dźwięków,
- fonetyka psycholingwistyczna - zajmuje się rozumieniem i formowaniem wypowiedzi
- fonetyka percepcyjna -zajmuje się percepcją dźwięków mowy

Rysunek 1.1 przedstawia dziedziny wiedzy związane z mową



Rys. 1.1 Dziedziny wiedzy obejmujące komunikację werbalną. Na podstawie (Gubrynowicz 2004)

Mowa jako podstawowy sposób komunikacji zawiera informacje, które są wysyłane przez mówcę i odbierane przez słuchacza. Komunikacja ta odbywa się na trzech poziomach (Laver 1994):

- lingwistycznym
- paralingwistyczym
- extralingwistycznym

Warstwa lingwistyczna zawiera informacje semantyczne oraz dotyczące struktury wypowiedzi (zarówno gramatykę jak i fonologiczne jednostki) i fonetyczną reprezentację wypowiedzi. Warstwa lingwistyczna obejmuje informacje, które są przekazywane, czyli treść wypowiedzi.

Warstwa paralingwistyczna jest strukturą werbalną i pozalingwistyczną. Zawiera informacje o aktualnym nastawieniu mówcy, jego stanie psychicznym i emocjonalnym. W przeciwieństwie do warstwy lingwistycznej nie da się jej jednoznacznie posegmentować. (Laver 1994) zdefiniował *setting*, który może być dowolnej długości, np. całego zdania lub tylko jego fragmentem np. pojedynczym segmentem. *Setting* współdzieli cechy kolejnych segmentów i

sylab dając wrażenie charakterystycznych cech mówcy lub jego zachowania podczas rozmowy. *Settings* są bardzo użyteczne podczas opisu jakości głosu człowieka, dzięki możliwościom w opisywaniu podobieństw podczas generowaniu dłuższych fragmentów sygnału mowy. *Settings* są używane na każdym poziomie opisu generowania sygnału mowy.

Trzecia warstwa, extra lingwistyczna, 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.(Marasek 1997)

1.1 Powstawanie sygnału mowy

Narząd mowy człowieka składa się z trzech części:

- płuc wraz z tchawicą
- krtani – odcinku fonacyjnego
- nasady, na którą składają się jamy: gardłowa, ustna, nosowa

1.1.1 Płuca

Płuca są pewnego rodzaju komorą ciśnieniową, z której wydobywa się powietrze wprawiające w drgania fałdy 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.(Stevens 1998)

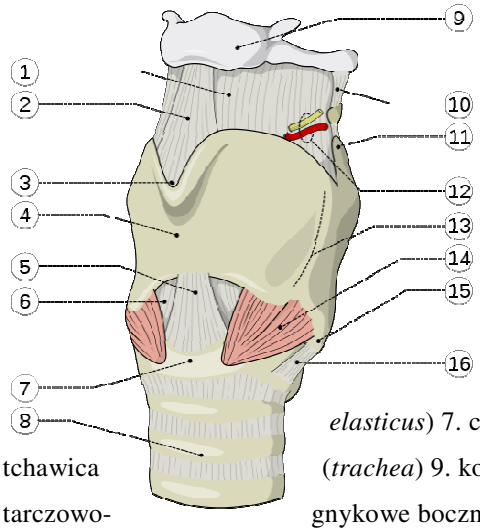
1.1.2 Krtań

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 fałdów nosi nazwę głosowych, fałdy górne zwane są fałdami kieszonek krtaniowych. Na brzegach fałdów głosowych znajdują się wiązadła głosowe. (Stevens 1998) W tyle krtani wiązadła głosowe są przymocowane do wyrostków głosowych, które mogą się od siebie oddalać lub przybliżać. Jeśli są one od siebie oddalone, pomiędzy wiązadłami głosowymi tworzy się szpara noszącą nazwę głośni. Zsunięte wiązadła głosowe mogą wibrować, czyli rozwierać się i na chwilę zwierać. Częstotliwość wibracji dla głosu męskiego wynosi średnio w mowie od około 80 Hz do około 160 Hz oraz od około 180 Hz do 250 Hz dla głosu kobiecego.

Wiązadła głosowe wibrują podczas wymawiania głosek dźwięcznych. Ilustracją przytoczonej treści jest poniższy rysunek 1.2:



1.błona tarczowo-gnykowa (*membrana thyrohyoidea*) 2. wiązadło tarczowo-gnykowe środkowe (*ligamentum thyrohyoideum medianum*) 3. wcięcie krtaniowe (*incisura laryngea*) 4. chrząstka tarczowata (*cartilago thyroidea*) 5. wiązadło pierścienno-tarczowe środkowe (*ligamentu cricothyroideum medianum*) 6. stożek sprężysty (*conus elasticus*) 7. chrząstka pierścieniowata (*cartilago cricoidea*) 8. tchawica (*trachea*) 9. kość gnykowa (*os hyoideum*) 10. wiązadło tarczowo-gnykowe boczne (*ligamentu thyrohyoideum laterale*)11. róg górny chrząstki tarczowatej (*cornu superius cartilaginis thyroideae*) 12. nerw (żółty) i tętnica (czerwona) krtaniowa górna (*nervus et arteria laryngeales superiores*) 13. kresa skośna (*linea obliqua*) 14. mięsień pierścienno-tarczowy (*musculus cricothyroideus*) 15. róg dolny chrząstki tarczowatej (*cornu inferius cartilaginis thyroideae*)16. staw pierścienno-tarczowy (*articulatio cricothyroidea*)

Rys. 1.2 Wiązadła i mięśnie zewnętrzne krtani (widok przednio-boczny) (Wikipedia 2009 http://pl.wikipedia.org/wiki/Plik:Larynx_external_base.svg)

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.(Stevens 1998).

Wzór 1 przedstawia sposób obliczania drgań fałdów głosowych. (Sonninen 1956)

$$F_0 = \frac{1}{2\Pi}\sqrt{\frac{(K + K^*)}{m}} \qquad (1)$$

gdzie :

m – masa fałdów

K – sztywność (napięcie) fałdów

K^* - sztywność aerodynamiczna

Żeby proces fonacji mógł się odbyć, fałdy głosowe muszą się zbliżyć się do siebie 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. Ruch ten odbywa się cyklicznie do pierwotnego położenia (jest to tzw. efekt Bernoulliego).

1.1.3 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 oddzielona od reszty nasady przez przywierające do tylnej jamy gardłowej podniebienie miękkie.” (Wierzchowska 1967)

Jama nosowa składa się z dwóch kanałów rozgraniczonych przegrodą nosową zwaną blaszką kostną. 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 w przybliżeniu 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

Rysunek 1.3 obrazuje podstawowe elementy układu artykulacyjnego.

1. zewnętrzna część górnej wargi 2. wewnętrzna część górnej wargi 3. górne zęby, 4. dziąsła, 5. za dziąsłami, 6. przednia cześć podniebienia twardego, 7. podniebienie twarde, 8. podniebienie miękkie, 9. języczek, 10. gardło, 11. krtań
Narząd artykulacji:
11. wiązadła głosowe, 12. nagłośnia, 13. korzeń języka, 14. tył języka, 15. środek języka, 16., 17., 18. przód języka

Rys. 1.3 Podstawowe elementy układu artykulacyjnego (Gubrynowicz 2004)

1.2 Proces artykulacji

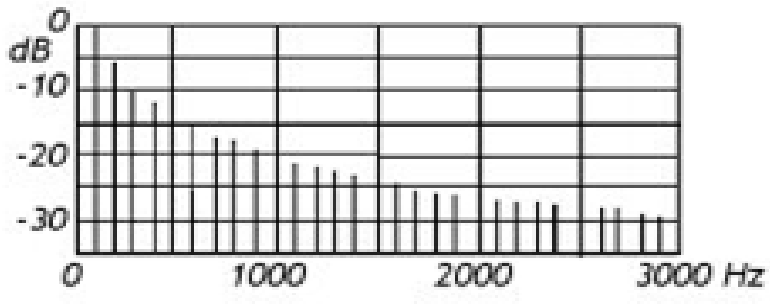
Poprzez poznanie budowy narządu mowy możliwe staje się zrozumienie jego funkcjonowania. Z kolei analiza procesu artykulacji, czyli prześledzenie drogi powstawania dźwięków, pozwala zrozumieć sposób działania artykulacyjnej syntezy mowy.

"Oskrzela i tchawica prowadzą dostarczony strumień [powietrza] do krtani, w której drgające struny głosowe są źródłem dźwięku dla dźwięcznych fragmentów mowy." (Tadeusiewicz 1988)

Dźwięk ten jest następnie formowany przez język, podniebienie, zęby i wargi, tworzące swoistego rodzaju układ akustyczny o zmiennych rezonansach. Podczas tego procesu ważną rolę odgrywają również ruchy żuchwy i w pewnym stopniu policzków.

Przepływ powietrza wprawia w drgania fałdy głosowe. W ten sposób

powstaje dźwięk zwany tonem podstawowym lub tonem krtaniowym. Ton podstawowy jest dźwiękiem harmonicznym o obwiedni opadającej w stosunku 6-12 dB na oktawę. (Rysunek 1.4)



Rys. 1.4 Widmo pobudzenia krtaniowego(Gubrynowicz 2004)

Ton podstawowy zmienia swoją częstotliwość, co jest podstawowym czynnikiem kształtującym intonację wypowiedzi, a zarazem decydującym o percepcji melodii głosu.

Zakres zmian tonu krtaniowego zależy od:

- płci - głosy kobiece mają z reguły 1,5-2-krotnie większą częstotliwość tonu krtaniowego niż głosy męskie
- wieku - głosy dziecięce są znacznie wyższe niż głosy osób dorosłych
- cech osobniczych

Powietrze wychodzące z tchawicy, przechodzące między fałdami głosowymi pobudza je do drgań zgodnie ze zjawiskiem Bernoulliego. Są to drgania bierne. Oznacza to, że powietrze przetłaczane przez szparę głośni, czyli szczelinę utworzoną między fałdami błony śluzowej, nazywanymi często strunami głosowymi, wprawia je w drgania na skutek dynamicznego oddziaływania strumienia powietrza na elastyczne fałdy". (Tadeusiewicz 1988)

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.

1.3 Specyfika języka polskiego

1.3.1 Klasyfikacja dźwięków mowy

Dźwięki mowy klasyfikuje się z uwagi na charakter przebiegów akustycznych oraz miejsce ich powstawania.

1.3.2 Klasyfikacja akustyczna

W podziale akustycznym wyróżnia się:

- rezonanty
- głoski zwarte (wybuchowe)
- głoski trące
- głoski zwarto-trące
- nosowe
- ustne (Wierzchowska 1967)

Głoski, których przebiegi akustyczne wykazują regularność (powtarzalność w czasie) 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~/ (zapis w kodzie fonetycznym SAMPA).

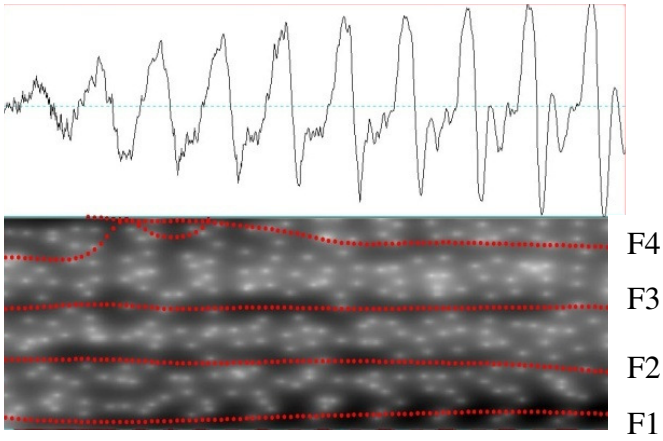
Inną grupę stanowią głoski wybuchowe (zwarte). Odpowiadają im krótkie nieregularne przebiegi akustyczne (impulsy). Segment zwarcia może mieć pobudzenie dźwięczne lub bezdźwięczne. Do głosek wybuchowych o pobudzeniu dźwięcznym należą: /g/ /b/ /d/, zaś do głosek o pobudzeniu bezdźwięcznym /p/ /t/ /k/.

Głoski trące składają się z przebiegów nieregularnych zwanych niekiedy frykcjami. Są to: /f/ /s/ /s'/ /S/.

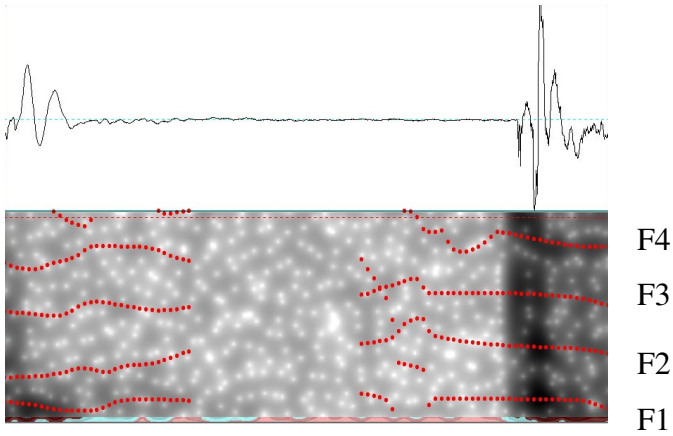
Afrykaty (zwarto-trące) są głoskami o przebiegu nieregularnym, których frykcje poprzedzone są słabym impulsem. Należą do nich: /ts/, /ts'/ /tS/ .

W kolejnej grupie, głosek nosowych, można zaobserwować silne tłumienie składowych o wyższych częstotliwościach oraz występowanie tzw. antyformantów (lokalne minima energii w widmie sygnału) głównie w zakresie częstotliwości od 900 do 2500 Hz.

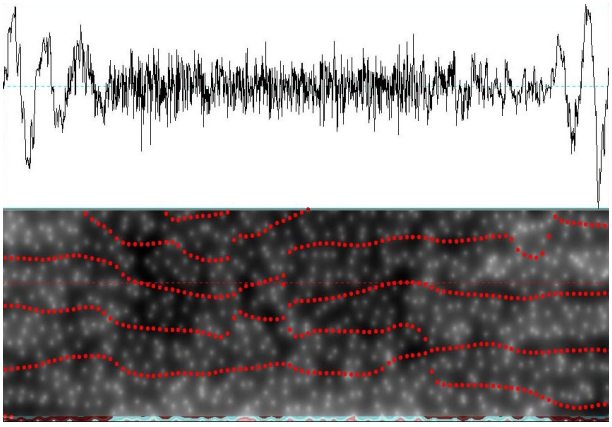
Samogłoski nosowe w języku polskim mają zazwyczaj realizację dyftongiczną. Oznacza to, że otwarcie nosowe nie jest zsynchronizowane z otwarciem ustnym. Początkowo samogłoska nosowa zaczyna się od samogłoski ustnej, po której następuje płynne otwarcie kanału nosowego i przejście do artykulacji spółgłoski nosowej (n). Taka realizacja spółgłosek nosowych może okazać się kłopotliwa przy konkatenacyjnym łączeniu ze sobą dźwięków mowy. Rysunek 1.5 przedstawia przebieg głoski o pobudzeniu dźwięcznym. Rysunek 1.6 przedstawia przebieg głoski wybuchowej. Po prawej stronie zaznaczono kolejne przebiegi formantowe pierwszy (F1), drugi (F2), trzeci (F3) i czwarty (F4). Rysunek 1.7 oraz 1.8 przedstawiają głoski o przebiegu nieregularnym.



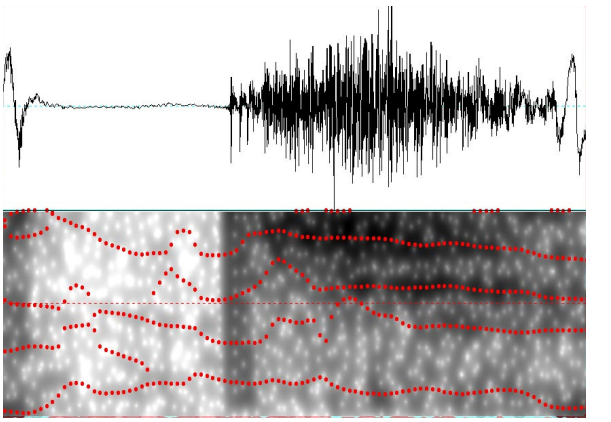
Rys. 1.5 Przykłady głoski regularnej /e/ wraz ze spektrogramem i analizą formantową



Rys. 1.6 Przykłady głoski wybuchowej /p/ wraz ze spektrogramem i analizą formantową



Rys. 1.7 Przykład głoski trącej /S/wraz ze spektrogramem i analizą formantową.



Rys. 1.8 Przykład afrykaty /ts/ wraz ze spektrogramem i analizą formantową.

Na rysunku 1.7 przedstawiona jest głoska /S/ wraz z charakterystycznym dla niej przebiegiem nieregularnym. Na rysunku 1.8 głoska /ts'/. Afrykaty wyróżniają się występowaniem słabego impulsu poprzedzającego przebieg szumowy.

1.3.3 Klasyfikacja genetyczna

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 lub zwężenia artykulatorów w torze głosowym.

Wyróżnia się również podział dźwięków ze względu na:

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

Z uwagi na zachowanie się wiązadeł głosowych głoski dzielą się na dźwięczne i bezdźwięczne. Głoski dźwięczne powstają wówczas, gdy wiązadła głosowe są zsunięte i wibrują. Głoski bezdźwięczne wymawiane są przy wiązadłach rozsuniętych.

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 odpowiadającym im głoskom dźwięcznych, np. /t – d/.(Wierzchowska 1980)

Ze względu na stopień zbliżenia narządów mowy wyróżnia się:

- spółgłoski zwarto-wybuchowe
- głoski zwarto-szczelinowe
- głoski szczelinowe
- spółgłoski otwarte

Zwarcie narządów mowy powoduje całkowite zamknięcie toru głosowego. Szczeliną zaś nazywamy przewężenie utworzone w określonym miejscu toru głosowego, powodujące znaczne zwiększenie przepływu strumienia powietrza i powstanie na ogół turbulencji.

Ze względu na miejsce artykulacji spółgłoski dzielimy na:

- dwuwargowe
- wargowo-zębowe
- przednio-językowe zębowe
- przedniojęzykowe-dziąsłowe
- środkowojęzykowe
- tylnojęzykowe-welarne

Podział ten umożliwia jednoznaczną klasyfikację głosek z uwagi na lokalizację charakterystycznego dla spółgłosek zwarcia lub szczeliny.

Wyróżnia się również podział spółgłosek oraz samogłosek ze względu na położenie podniebienia miękkiego. Podział ten charakteryzuje głoski ustne i nosowe.

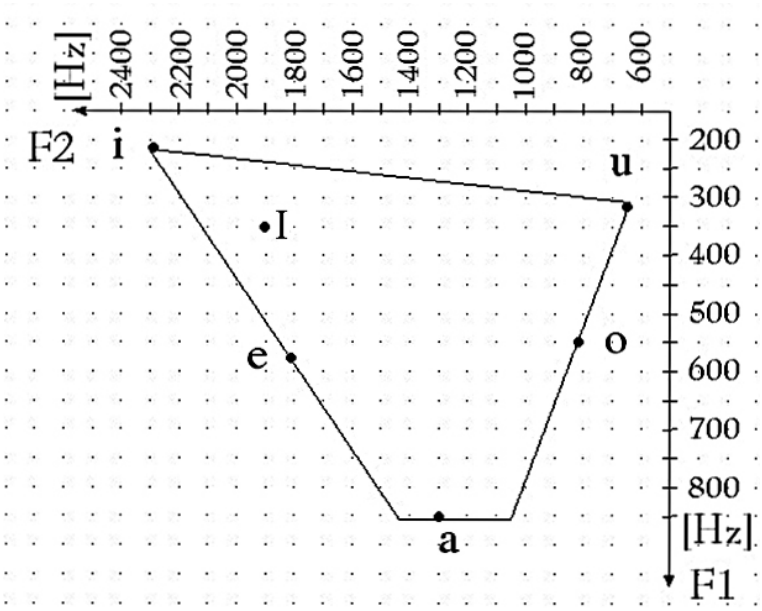
Ostatnim podziałem spółgłosek jest podział uwzględniający artykulacje dodatkowe. Zalicza się do nich:

- labializację, czyli zaokrąglenie wargowe
- delabializację, czyli spłaszczenie warg
- palatalizację
- welaryzację, czyli wzniesienie tylnej części języka
- retrofleksję, czyli artykulację polegającą na wzniesieniu czubka języka i cofnięciu go.

1.3.4 Klasyfikacja samogłosek

Przedstawione podziały dotyczyły głównej klasyfikacji artykulacyjnej spółgłosek. Poniżej opisano krótko klasyfikację samogłosek na podstawie czworoboku samogłoskowego, opracowanego przez angielskiego fonetyka Daniela. Jonesa.(Jones 1918) (Rysunek 1.9)

Badania rentgenograficzne pozwoliły na wyznaczenie najbardziej wzniesionych punktów grzbietu języka i przyporządkowanie poszczególnym konfiguracjom toru głosowego odpowiednich samogłosek.



Rys. 1.9 Czworobok artykulacyjny w płaszczyźnie F1- F2

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

1.3.5 Uproszczenie klasyfikacji dźwięków mowy

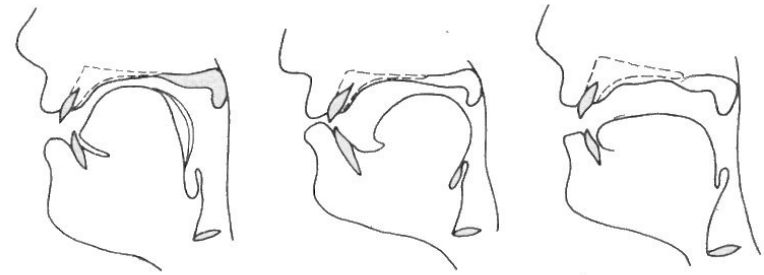
Fonetyczna klasyfikacja samogłosek jest dokonywana na podstawie innych kryteriów niż klasyfikacja spółgłosek. W przypadku samogłosek uwzględnia się położenie masy języka. Decyduje on o kształcie kanału głosowego, rozkładzie formantów. W opisie spółgłosek bierze się pod uwagę stopień zbliżenia narządów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

Tak skomplikowany podział jest niewygodny. Dlatego stosuje się podział spółgłosek i samogłosek z uwagi na układ masy języka oraz częstotliwość drugiego formantu.

W klasyfikacji tej wyróżnia się:

- położenie przednie masy języka
- położenie tylne masy języka
- położenie środkowe masy języka

Poniższy podział obrazują schematy:



Przednia np. i Tylna np. u Środkowa np. a

Rys. 1.10 Klasyfikacja samogłosek z uwagi na położenie masy języka(Borden i wsp. 1994)

Przednia np. s', x, p Tylna np. k, g Środkowa np. t, d, s, z, sz

Rys. 1.11 Klasyfikacja spółgłosek z uwagi na położenie masy języka (Borden i wsp. 1994)

Omówienie zagadnienia procesu artykulacji oraz sklasyfikowanie dźwięków mowy pozwala orientować się w cechach charakterystycznych głosek. Informacje te są niezbędne do realizacji korpusowej syntezy mowy. W pracy informacje te zostały wykorzystane do realizacji procesu segmentacji.

1.3.6 Fonetyczna organizacja wypowiedzi

Stworzenie dobrej jakości syntezatora mowy jest trudnym zadaniem. Chcąc spełnić wymagania naturalności brzmienia mowy oraz uniknąć błędów konkatenacji należy odnieść się do języka naturalnego i zdefiniować podstawowe pojęcia mówiące o organizacji wypowiedzi. Przez język

naturalny rozumie się każdy język powstały na drodze naturalnej ewolucji człowieka (polski, angielski, itp).

Zagadnienia te sprowadzają się do omówienia podstawowych problemów organizacji dźwiękowej wypowiedzi języka naturalnego. Należą do nich: koartykulacja, iloczas, akcent, melodia. Omówienie ich pozwoli zrozumieć trudności, jakie należało pokonać podczas realizacji syntezatora mowy. Zrozumienie tych pojęć opiera się na definicjach segmentalnych jednostek mowy, takich jak głoska, sylaba czy fraza. Głoskę językoznawcy (Wierzchowska 1967) definiują jako najmniejszą, niepodzielną cząstkę formy dźwiękowej języka. Definicja ta jest zawsze związana z konkretnym językiem, a określenie niepodzielność należy rozumieć umownie, ponieważ szereg głosek ma strukturę podzielną w przebiegu czasowym (np. tak zwane spółgłoski polisegmentalne – zwarcie + plozja w /p/). Wymiennie z pojęciem głoski używa się terminu segment, określany jako wycinek ciągu dźwiękowego między dwoma określonymi punktami zmiany w sygnale mowy. Punkty te są tak dobierane, że zamiana segmentu na inny (czy usunięcie) pociąga za sobą zmianę (lub utratę) znaczenia niesionego przez sygnał mowy. Sylaba jest fonetyczno-fonologiczną jednostką słowa jak i jednym z bardziej spornych zagadnień w fonetyce. Według Leonce Roudeta (Roudet 1947) sylaba jest odcinkiem mowy, na którego środkową część przypadają: minimum ciśnienia powietrza w tchawicy, maksimum otwarcia narządów mowy oraz maksimum głośności. Na jego zaś krańcach - (początku i końcu) odwrotnie: maksimum ciśnienia powietrza w tchawicy, maksimum zbliżenia narządów mowy oraz minimum głośności. Fraza natomiast jest jednostką frazeologiczną, zawierająca podmiot i orzeczenie. Również definiuje się ją jako pewien zamknięty człon rytmiczny wypowiedzi. (Wierzchowska 1980)

1.3.7 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 przygotowują się do

artykulacji następnej głoski. Proces ten nazywa się koartykulacją.

Bezpośrednio z zagadnieniem koartykulacji związane jest pojęcie upodobnień.

1.3.8 Upodobnienia

Koartykulacja prowadzi do częściowego (niekiedy całkowitego) zacierania się różnic pomiędzy sąsiadującymi ze sobą dźwiękami i tym samym do tzw. upodobnień. Powodują one zmianę ich postaci dźwiękowej. Upodobnienia obejmujące grupy głosek i połączone z redukcją (częściową, lub całkowitą) pewnych dźwięków tworzących te grupy, nazywane są „uproszczeniami" (Gubrynowicz 2004). Dzieli się je 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*." (Wierzchowska 1967)

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"*. (Wierzchowska 1967)

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

związane z redukcjami częściowymi lub całkowitymi oraz antycypacją czy też podtrzymywaniem (przedłużeniem)

1.3.9 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 zakresie wynosi przeciętnie od 5 do 25, przy czym dolna wartość obejmuje bardzo wolny sposób mówienia, podczas gdy górna wartość stanowi granicę zrozumiałości wypowiedzi.

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

Czas trwania głoski (iloczas) związany jest również ze sposobem artykulacji. Nieco krócej trwają głoski ustne a spółgłoski nosowe są najkrótszymi głoskami. Iloczas trwania głoski jest zawiązany z czasem jej artykulacji i jest użyteczny przy określaniu zmiany iloczasu głoski odpowiednio do otaczającej jej dźwięków mowy. Odpowiedni dobór iloczasów ma wpływ na percepcję wypowiedzi zarówno pod względem zrozumiałości, jak i jej brzmienia.

Wyróżnia się dwa rodzaje iloczasu:

- iloczas bezwzględny
- iloczas względny

Iloczas bezwzględny opisuje czas trwania głoski w wypowiedzi, oraz pozwala określić jego tempo, natomiast iloczas względny stanowi stosunek czasu trwania głosek w stosunku do innych głosek oraz ma pływ na percepcję rytmu wypowiedzi. Generalnie przyjmuje się, że im bardziej skomplikowana artykulacja, tym czas trwania głoski jest dłuższy. Również ważnym zagadnieniem są fazy wypowiedzi (np. czy głoska jest wypowiedziana w

nagłosie, wygłosie itp.), które mają wpływ na charakterystykę czasową (i nie tylko) głosek.

1.3.10 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 mniej precyzyjnie i wolniejsze. Również następuje obniżenie tonu podstawowego w wyniku zwolnionej pracy wiązadeł głosowych (wskutek malejącego ciśnienia podgłośniowego).

1.3.11 Akcent

Oprócz czynników charakterystycznych dla danego języka takich jak zjawisko koartykulacji czy też połączenia dźwięków, ważnym elementem jest zróżnicowanie dynamiczne oraz melodyczne wypowiedzi. Zjawisko to określa się mianem akcentu. Jest to proces uwydatniający wybrane segmenty w sygnale mowy ciągłej, np. sylab w wyrazach lub wyrazów w zdaniach. Uwydatnienie sylaby akcentowanej może polegać na silniejszym, a zarazem głośniejszym jej wypowiedzeniu, na bardziej precyzyjnym jej wymówieniu, co może spowodować jej wydłużenie czasu trwania. Może też wystąpić tylko podwyższenie (niekiedy obniżenie) częstotliwości pobudzenia krtaniowego.
W zależności od tego, który z tych czynników przeważa, akcent jest

określany jako:

- dynamiczny – gdy czynnikiem dominującym w płaszczyźnie akustycznej są chwilowe zmiany intensywności
- rytmiczny– gdy o wrażeniu akcentu decydują zmiany iloczasów sylab,
- melodyczny – gdy akcentowanie sylaby jest realizowane poprzez chwilową zmianę wysokości głosu

Dla języka polskiego przyjmuje się, że akcent jest zazwyczaj dynamiczny, choć jest to dyskusyjne. (Łukaszewicz i wsp. 2008, Gubrynowicz 2004) W języku polskim akcentowana jest przeważnie przedostatnia sylaba, 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 zwany akcentem frazowym w przeciwieństwie do wyrazowego powoduje, że dany fragment wypowiedzi uzyskuje na ogół dodatkowe wzmocnienie i wydłużenie. 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 i pewnym stopniu z przebiegiem zmian głośności.

1.3.12 Melodia

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

W zdaniach oznajmujących wysokość melodii jest niska a trend jej jest opadający. Wzrost wysokości tonu podstawowego przeważnie ma miejsce w sylabach akcentowanych 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, ale zmieniają funkcję zdania. Przebieg zmian zależności tonu podstawowego nosi nazwę melodii zasadniczej.

Wyróżnia się cztery podstawowe rodzaje konturów melodii:

- rosnąca niska
- rosnąca wysoka
- opadająca niska
- opadająca wysoka
- równa niska
- równa wysoka (Wierzchowska 1967, Steffen-Batogowa 1996,Gubrynowicz 2004)

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

Charakterystyka melodii jest ściśle powiązana z modelowaniem prozodii w systemach syntezy mowy i od dobrego odwzorowania konturu melodycznego zależy uzyskanie głosu syntetycznego zbliżonego do naturalnego.

1.4 Transkrypcja fonetyczna

Opis sygnału mowy wymaga nadania etykiet poszczególnym jego segmentom. Tekst ortograficzny nie pozwala na jednoznacznie określenie wymowy i nie jest dobrym sposobem jej reprezentacji. Te same znaki ortograficzne mogą odpowiadać różnym dźwiękom, podczas gdy ten sam dźwięk może odpowiadać różnym znakom. Przykładem może być litera /v/ w wyrazach /waga/ i /wtórny/, w pierwszym wypadku czytana jest jako /v/, w drugim jako /f/. Inne przykłady to litery /u/ i /ł/ w wyrazach /auto/ i /głóg/, obie czytane jako /ł/. Mniej oczywiste są różnice w wymowie litery /n/, np. w wyrazach /niewiadomo/ i /gong/, bo nie ma zmiany cechy dźwięczności na bezdźwięczność. W celu ujednolicenia zapisu wymowy oraz jego

jednoznaczności opracowany został szeroko stosowany międzynarodowy alfabet fonetyczny IPA (*International Phonetic Alphabet*), zawierający reprezentację dźwięków mowy wszystkich języków. Pewną wadą kodu IPA jest fakt, iż zawiera on znaki diakrytyczne nieistniejące w standardowym kodzie ASCII. Wygodniejszy do stosowania komputerowego jest alfabet SAMPA (*Speech Assessment Methods Phonetic Alphabet*) Wells 1997). Jest on w pełni kompatybilny z ASCII. Opracowywane równolegle były i wciąż są niezależne notacje dla 24 języków.

Proces przekształcania tekstu ortograficznego na kod fonetyczny opiera się o określone reguły i nazywa się transkrypcją fonetyczną. Opracowanie reguł transkrypcji fonetycznej w kodzie SAMPA dla języka polskiego jest niezbędne w procesie segmentacji sygnału mowy, będącej celem cząstkowym niniejszego projektu. Poniżej przedstawiono tabele ogólnych odwzorowań znaków ortograficznych (odpowiadających im fonemów) na kod SAMPA dla języka polskiego. Dodatkowo opisane zostały reguły precyzujące odstępstwa i wyjątki specyficzne dla języka polskiego, w głównej mierze zależne od otoczenia danego znaku.

1.4.1 Samogłoski

System samogłosek w języku polskim składa się z 8 fonemów. Symbole ze znakiem:/~/ oznaczają nazalizację.

Tabela 1.1 przedstawia sposób reprezentacji samogłosek w transkrypcji fonetycznej

1.4.2 Spółgłoski

System spółgłosek w języku polskim składa się 29 fonemów. Symbol /'/ oznacza palatalizację. Palatalizacja jest to fonetyczne zmiękczenie spółgłoski twardej pod wpływem sąsiadującej z nią samogłoski (najczęściej przedniej). Tabele 1.2, 1.3, 1.4, 1.5 przedstawiają symbole dla spółgłosek w reprezentacji fonetycznej

Symbol ortograficzny	Symbol SAMPA	Np. w wyrazie
i	i	bit /bit/
y	I	byk /bIk/
e	e	bek /bek/
a	a	bak /bak/
o	o	bok /bok/
u	u	buk /buk/
ę	e~	tę /te~/
ą	o~	tą /to~/

Tabela 1.1 Transkrypcja fonetyczna samogłosek SAMPA (Gubrynowicz 2004, Wells 1997).

Symbol	Symbol SAMPA	Np. w wyrazie
f	f	fakt /fakt/
w	v	waga /vaga/
s	s	syk /sIk/
z	z	zbir /zbir/
sz	S	szyk /SIk/
ż	Z	żyto /ZIto/
ś	s’	świt /s’fit/
ź	z’	źle /z’le/
h, ch	x	hak /xak/

Tabela 1.2 Transkrypcja fonetyczna spółgłosek trących(Gubrynowicz 2004, Wells 1997).

Symbol	Symbol SAMPA	Np. w wyrazie
p	p	puk /puk/
b	b	bat /bat/
t	t	test /test/
d	d	dym /dIm/
k	k	kat /kat/
g	g	gen /gen/

Tabela 1.3 Transkrypcja fonetyczna spółgłosek zwartych, czyli plozyjnych(Gubrynowicz 2004, Wells 1997*).*

Symbol	Symbol SAMPA	Np. w wyrazie
m	m	mysz /mIS/
n	n	nasz /naS/
ń	n’	koń /kon’/
n(k,g)	N	bank /baNk/*
ł	w	łyk /wIk/
j	j	jak /jak/
l	l	luk /luk/
r	r	ryk /rIk/

* Spółgłoska nosowa /N/ występuje w języku polskim tylko przed spółgłoskami /k, g/.

Tabela 1.4 Transkrypcja spółgłosek zwanych sonorantami lub rezonantami(Gubrynowicz 2004, Wells 1997*).*

Symbol ortograficzny	Symbol SAMPA	Np. w wyrazie
c	ts	coś /tsos’/
dz	dz	dzwon /dzvon/
cz	tS	czapka /tSapka/
dż	dZ	dżem /dZem/
ć	ts’	ćwicz /ts’fitS/
dź	dz’	dźwiga /dz’viga/

Tabela 1.5 Transkrypcja fonetyczna spółgłosek zwarto-trących (Gubrynowicz 2004, Wells 1997).

Powyższe tabele określają jedynie odwzorowania symboli i wymagają uściślenia dodatkowymi regułami, które przedstawiono poniżej (zgodnie z Gubrynowicz 2004).

Literom samogłoskowym /y,e,a,o/ odpowiadają fonemy /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 pauzy między nimi.

O ubezdźwięcznieniu lub udźwięcznieniu całej sekwencji spółgłosek zwartych, zwarto-trących oraz trących 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/. Spółgłoski bezdźwięczne przed końcówką czasownikową /my/ także pozostają bezdźwięczne, np. /kupmy/ -> /kupmy/

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 /j/, /l/, /w/ (przymknięte) wymówione w środku dłuższych sekwencji spółgłoskowych, wymawiane są tak słabo, że często ulegają całkowitej redukcji, a ich otoczenie najczęściej staje się bezdźwięczne. Np. /jabłko/ -> /japko/, /rzemieślnik/ -> /Zemjes'n'ik/. (Gubrynowicz 2004)

Omówiona reprezentacja fonetyczna została wykorzystana podczas segmentacji korpusu. Pewne modyfikacje tego zapisu okazały się konieczne. Związane były one z wymogami syntezatora i systemem Festival, a także wymową autora nagrań. Modyfikacje te zostały opisane w rozdziale 4.

1.5 Modele opisu prozodii

Termin prozodia odnosi się do pewnych właściwości sygnału mowy, które można usłyszeć w postaci zmiany głośności, długości sylab, i intonacji. Cechy prozodyczne odgrywają duże znaczenie w komunikacji językowej. Odpowiednie zaakcentowanie sylaby może zmienić znaczenie całej wypowiedzi. Istnieje kilka modeli opisu cech prozodycznych. W niniejszym podrozdziale zostaną przedstawione te, które były dotychczas używane dla języka polskiego oraz zostały zaimplementowane w środowisku Festival.

1.5.1 ToBI – Tones and Break Indices

System ToBI wziął swój początek od reguł stworzonych przez Janet Pierrehumbert. (Pierrehumbert 1980, 1983). ToBI został zdefiniowany w celu anotacji amerykańsko-angielskiej melodii, następnie został przystosowany do innych języków.(Wagner 2004, Grice i wsp. 2002, Venditti 1997) ToBI oferuje dobrze zdefiniowaną fonologię intonacji dla posegmentowanej mowy, jest

jednym z bardziej rozpowszechnionych standardów. System ToBI nie posiada mechanizmu pozwalającego na automatyczne uzyskanie etykiet opisujących zmiany w konturze melodycznym jednak zostały stworzone pewne systemy regułowe pozwalające uzyskać łatwiejszą anotację. (Anderson i wsp. 1984). Powstały narzędzia pozwalające na automatyzacje procesu anotacji stworzone przez twórców systemu Festival Alana Blacka oraz Andrew Hunta (Black i wsp. 1996).

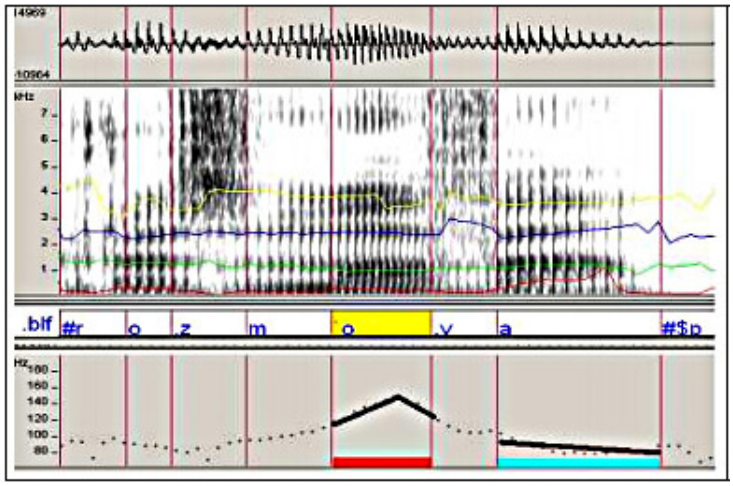
Opis ToBI zawiera opis tylko najważniejszych z lingwistycznego punktu widzenia przebiegu zmian F0, tak więc dla sylab nieakcentowanych wysokość tonu jest interpolowana z sylab akcentowanych. Z punktu widzenia syntezy mowy ToBI pozwala opisać istniejący kontur intonacyjny, nie podaje jednak gotowych reguł, jak z istniejącego opisu wygenerować zmiany F0 w syntezowanej wypowiedzi. Zwykle do tego celu wykorzystuje się napisane ręcznie reguły. Opis ToBI niekoniecznie może być też dobrze dostosowany do danego języka np. dla dialektu mandaryńskiego będącego językiem tonalnym.

W systemie ToBI wyróżnia się następujące znaczniki typów akcentu: H*, H+!H, L*, L*+H, L+H*. W dużej mierze zależy również od specyfiki języka. Dla języka polskiego powstały dodatkowe reguły, które zostały opracowane przez Prof. Grażynę Demenko oraz Dr Agnieszkę Wagner. (Demenko 1999, Wagner 2004). Symbol gwiazdki oznacza sylabę akcentowaną, natomiast % koniec frazy. Akcent frazowy oznacza się jako H- i L-. Granice tonów oznacza się jako L-L%, L-H%, H-L%, H-H%. H+!H znacznik akcentowanej sylaby (!H) używany gdy poprzednia sylaba jest nieakcentowana i posiada wysoką wartość F0. Pauzy oznacza się numerem 1,3,4 oraz 2, która jest zarezerwowana dla specjalnych przypadków (Silverman i wsp. 1992).

1.5.2 ToBI dla języka polskiego

W 1999 zaproponowany został system anotacji intonacyjnej przez (Demenko 1999), oparty na tzw. szkole brytyjskiej i pracach Wiktora Jassema. System ten jest krokiem w kierunku upowszechnienia i dostosowania do specyfiki danego języka, a zarazem umożliwienia prac porównawczych między językiem polskim i innymi językami. (Karpiński 2001). W systemie został wykorzystany fakt związania akcentów z tzw. fokusem zdania (jego

najważniejszym elementem). Podobnie jak dla systemu ToBI nie ma jednoznacznej relacji pomiędzy opisem symbolicznym a generacją konturu F0, co jest niezbędne dla procesu syntezy mowy. (Marasek 2003 B). Dla języka polskiego nie powstała w pełni funkcjonalna wersja systemu ToBI. (Karpiński 2001)



*Rys. 1.12 Przebieg czasowy, spektrogram i przebieg intonacji wraz z opisem dla PToBI L H*L melodia rosnąco-opadająca.(Demenko, Wagner 2007)*

1.5.3 INTSINT

Model INTSINT (Hirst 1994) jest systemem symbolicznego kodowania określającym przebieg konturu F0. W przeciwieństwie do systemu ToBI symbole do anotacji są takie same dla każdego języka. Punktu pomiaru mierzone są co 30 ms. Następnie każdy z tonów absolutnych jest flagowany jako TOP, MID lub BOTTOM w zależności od przedziału wysokości głosu określanego dla każdego mówcy osobno. Tony względne nieiteracyjne HIGHER, SAME, LOWER określane jako referencje do poprzedniego punktu docelowego (docelowej wysokości głosu). Tony względne iteracyjne UPSTEPPED, DOWNSTEPPED różnią się od nieiteracyjnych interwałami F0, które są większe. (Hałupka 2004)

1.5.4 Momel

System stylizacji konturu melodycznego Momel został zaproponowany przez Daniela Hirsta (Hirst 1994) w 1983, a następnie zautomatyzowany w roku 1993. Metoda stylizacji konturu polega na zamianie oryginalnego konturu F0 przez uproszczoną, ciągłą funkcję numeryczną. W ten sposób modelowany jest makroprozodyczny komponent F0 za pomocą funkcji kwadratowej *spline*, która w wyniku daje ciągły kontur. Segmenty nieakcentowane są interpolowane, dzięki temu kontur jest nie tylko ciągły na całym przebiegu, ale również pozbawiony jest dużych krzywizn.

Zaletą użycia kwadratowej funkcji *spline* jest uzyskanie krzywej bliższej do naturalnego przebiegu F0 niż estymowanej za pomocą krótkich odcinków prostych, i nie wprowadzającej znacznych zniekształceń (Oliver 2007). System wyznaczania przebiegu F0 dla języka polskiego, z wykorzystaniem stylizacji Momel, został stworzony oraz zaimplementowany w meta systemie Festival przez (Oliver 2007). Moduł ten został wykorzystany w prezentowanej pracy.

System pozwala na generowanie automatycznej intonacji przy wykorzystaniu parametrów pozyskanych na podstawie konturu F0 oraz klas akcentów w drodze klastrowania. Predykcja F0 wyznaczana jest na podstawie drzew klasyfikacyjnych oraz regresyjnych. Moduł liniowej regresji pozwala na predykcję wartości F0.

1.6 Klasyfikacja segmentów sygnału mowy o różnej rozciągłości.

Analiza mowy wymaga, by w sygnale będącym ciągłą sekwencją dźwięków, wyodrębnić charakterystyczne, segmenty o stosunkowo niewielkiej rozciągłości, o zróżnicowanej strukturze akustycznej. Wyróżnia się kilka takich klas jednostek wykorzystywanych najczęściej w analizie mowy na potrzeby segmentacji, syntezy i jej rozpoznawania:

- głoski

- alofony
- difony
- trifony
- półsylaby
- sylaby

(Wierzchowska1980) pisze: „W strukturze postaci dźwiękowej języka polskiego szczególna rola przypada głoskom, które są najkrótszymi elementami dźwiękowymi pełniącymi funkcję dystynktywną.” Zatem głoski są najprostszymi elementami dźwiękowymi mowy rozróżnianymi słuchowo przez użytkowników danego języka i odróżniającymi od siebie formy językowe mające różne wartości semantyczne. Inwentarz głosek, zarówno jak i każdej innej jednostki akustycznej zależny jest od języka. W pracy przyjęto, iż w języku polskim występuje 37 fonemów (alfabet fonetyczny SAMPA), co odpowiada liczbie rozróżnianych głosek, a każdy z 37 fonemów jest zbiorem cech dystynktywnych odpowiadającej mu głoski. Dlatego w innych opracowaniach liczba te może się różnić. Cytując (Wierzchowska1980): „Stosując kazańsko-praską procedurę wyróżniania fonemów, dla języka polskiego ustala się inwentarz fonemów obejmujący 41 pozycji.” Przytoczona odmienna liczba fonemów wynika z faktu, iż kazańsko-praska procedura nie wyróżnia fonemu /i/ jako osobnego fonemu w przypadku jego występowania po spółgłosce, wyróżnia natomiast jako osobne fonemy te, które są zmiękczone przez następujący po nich fonem /i/ (np. zbitka fonemów /pi/ uznawana jest za osobny alofon - /p’/).

Zgodnie ze stanowiskiem (Wierzchowska 1980) należy zauważyć, iż: *„Poszczególne realizacje tych samych głosek, nawet wymawianych w takim samym kontekście fonetycznym, w tych samych formach wyrazowych, nie są nigdy zupełnie takie same; w różnych wykonaniach tych samych ruchów artykulacyjnych obserwuje się zawsze pewien naturalny rozrzut charakteryzujący wszelkie, najbardziej nawet zautomatyzowane czynności człowieka.”* Oznacza to, iż głoska stanowi pewne uogólnienie dostatecznie podobnych dźwięków, będących jego realizacjami. Konkretne realizacje fonemów także mogą być rozróżniane i nazywa się je alofonami.

Difon jest jednostką akustyczną, która zawiera przejście (tranzjent) pomiędzy dwoma kolejnymi głoskami. Rozpoczyna się w połowie jednego głoski (tzw. części stacjonarnej), a kończy w połowie następnego (także w części stacjonarnej). Difon jest często stosowaną jednostką w systemach syntezy mowy. Jego zastosowanie umożliwia uzyskanie większej naturalności brzmienia mowy, niż w przypadku systemów opartych na konkatenacji głosek, ponieważ w przypadku konkatenacji difonów połączenie fragmentów mowy ma miejsce w części stacjonarnej głosek, która nie ulega „zniekształceniom" związanym z płynnymi ruchami narządów artykulacyjnych (koartykulacją). Dlatego głoski wycięte z nagrań i umieszczone w innym kontekście, często wnoszą zniekształcenia podyktowane upodobnieniami na ich krańcach, wynikającymi z ich oryginalnego otoczenia. Łączenie difonów w słowa następuje na stosunkowo stabilnych częściach segmentu, co wpływa na korzystne brzmienie. Dużą zaletą konkatenacyjnej syntezy mowy z zastosowaniem difonów jest mały nakład pamięci potrzebny do przeprowadzenia odpowiednich obliczeń. Ponadto, rozmiar bazy danych jest stosunkowo niewielki (1444 difony dla języka polskiego przy wyróżnieniu 37 fonemów oraz dodatkowego znacznika ciszy). Granice difonów są łatwiejsze do wyznaczenia, niż granice głosek, gdyż wspomniane części stacjonarne głosek, ulegają w znacznie mniejszym stopniu koartykulacji i są najbardziej charakterystycznymi elementami głosek. Dodatkowo granice difonów mogą być dość arbitralnie wyznaczane, w stosunkowo szerokim przedziale czasowym, np. 20% czasu trwania części stacjonarnej. Dlatego segmentacja nagrań z wykorzystaniem difonów jest łatwiejsza niż w przypadku głosek, dla których niejednokrotnie trudno, czy wręcz niemożliwe jest dokładnie określić początek i koniec danej głoski (np. dla spółgłosek płynnych, takich jak /j/).

Kolejną klasą segmentów akustycznych są trifony. Są to jednostki z określonym kontekstem lewo i prawostronnym. Oznacza to, iż trifony dzielą fonemy na grupy alofonów ze względu na ich lewe i prawe sąsiedztwo, modelując w ten sposób zależność głosek od ich kontekstu (koartykulację). Dlatego trifony dobrze nadają się do syntezy mowy i pozwalają uzyskać dość naturalne brzmienie. Warto zauważyć, że choć trifony stanowią dobrą alternatywę dla difonów, wymagają oczywiście znacznie większej bazy

akustycznej. W praktyce używa się najczęściej około 4000 występujących trifonów w danym języku. Należy dodać, że proces segmentacji trifonów dostarcza wielu problemów. (Bozkurt i wsp. 2003).

Definicja sylaby jest w fonetyce zagadnieniem spornym. Jak pisze (Wierzchowska1980) *„Problem sylaby rozpatrywany bywa bądź ze stanowiska artykulacyjnego, bądź ze stanowiska audytywnego, bądź w obu tych aspektach jednocześnie."* Poniżej przedstawiono cytat tej samej pozycji definiujący sylabę z artykulacyjnego punktu widzenia i percepcyjnego (akustycznego) jednocześnie, gdyż ta właśnie definicja wydaje się najpełniejsza. Jest to definicja tzw. sylaby fonetycznej.

„Fonetycy, którzy opisują sylabę w obu aspektach, tj. i w aspekcie artykulacyjnym, i w aspekcie akustycznym, kładą nacisk na jednoczesność zmian w układzie narządów mowy, ciśnieniu powietrza w tchawicy oraz donośności dźwięków (postrzegalności słuchowej). Ośrodkami sylab są te odcinki ciągu mownego, na które przypada maksymalne rozwarcie kanału głosowego i maksymalna donośność; na pograniczach i na krańcach sylab donośność dźwięków mowy jest najniższa, stopień zaś zbliżenia narządów mowy - największy."(Wierzchowska1980)

Podobnie kwestię sylaby fonetycznej ujmuje także (Roudet 1947). 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. Definicję sylaby podano w 1.3.6. Należy dodać, iż segmentacja sylab jest względnie łatwa.

Przytoczone jednostki akustyczne są fundamentalne dla opisanych w kolejnych rozdziałach pracy zagadnień: segmentacji nagranych wypowiedzi, rozpoznawania mowy oraz jej konkatenacyjnej syntezy, w tym w najefektywniejszej jej wersji, to jest metody korpusowej.

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

Tabela 1.6 Porównanie akustycznych jednostek mowy i jakości syntezy mowy przez nie generowanych

1.6.1 Podsumowanie

W rozdziale przedstawione zostały ogólne zagadnienia związane z fonetyką akustyczną obrazującą sposób opisu dźwięków człowieka mowy w płaszczyźnie artykulacyjnej. Przedstawiono budowę narządu człowieka oraz klasyfikację dźwięków przez niego artykułowanych. W dalszej części opisane zostały zagadnienia dotyczące organizacji wypowiedzi oraz transkrypcji fonetycznej. Opisano zostały podstawowe akustyczne jednostki segmentalne w języku polskim oraz przedstawiono wpływ ich doboru na jakość syntezy mowy. W końcowej części tego rozdziału opisano podstawowe modele opisu prozodii.

2 Metody syntezy mowy i ich realizacje dla różnych języków

Rozdział ten jest poświęcony prezentacji rozwoju syntetyzatorów mowy, opisowi podstawowych metod syntezy, a także analizie działania systemu TTS (*Text-to-speech*) oraz jego poszczególnych modułów. Według (Dutoit 1997, Taylor 2009) system TTS definiuje się jako automatyczny proces generowania mowy od momentu wprowadzenia na wejście systemu transkrypcji fonetycznej wypowiedzi aż po jej wypowiedzenie wygenerowanie w postaci akustycznego sygnału mowy.

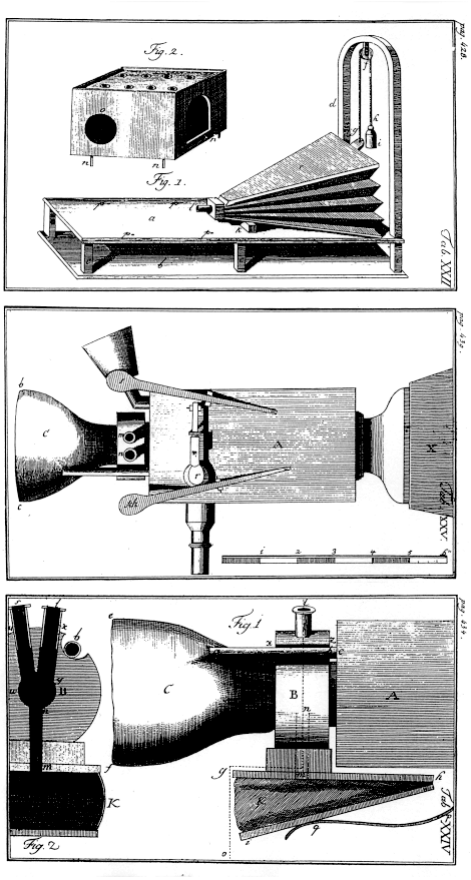
2.1 Rys historyczny

Pierwsze eksperymenty związane z próbą generowania mowy syntetycznej sięgają XVIII wieku. Fundamentalną próbą stworzenia mowy podobnej do ludzkiej był eksperyment profesora fizjologii Christina Kratzensteina, który podjął próbę wyjaśnienia różnic 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 miecha odpowiadających płucom, sterowanych za pomocą prawego przedramienia. Na środkowym i dolnym rysunku 2.1 znajduje się miech z dźwigniami oraz konstrukcja ust wykonanych z gumy wraz z odpowiednikiem nosa. Dwa nozdrza należało przykryć palcami chcąc uzyskać głoskę nosową. Maszyna von Kempelena umożliwiała ręczną kontrolę artykulacji i intonacji, 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 książce „*Mechanismus der menschlichen Sprache. Beschreibung*

einer sprechender Maschine" (*Mechanizm ludzkiego języka. Opis mówiącej maszyny*) von Kempelen umieścił opis mówiącej maszyny, poddając również analizie podstawowe zasady działania narządów mowy. Jednak największym osiągnięciem autora było określenie roli narządów ponadkrtaniowych w procesie generowania dźwięku.



Rys. 2.1Maszyna mówiąca von Kempelena.
(http://www.ling.su.se/staff/hartmut/kemplne.htm)

W 1835 roku została stworzona mówiąca maszyna przez Josepha Fabera, która zawierała sztuczny język, jamę gardłową oraz umożliwiała generowanie melodii w formie śpiewania. Wynalazek Fabera był obsługiwany przy pomocy klawiatury i pedałów. W roku 1846 w Londynie „Euphonia" – taką nosiła nazwę – „zaśpiewała" *„God Save the Queen"*. W 1936 roku powstał VODER, pierwsza maszyna, która wykorzystywała elektryczność. Urządzenie skonstruowane przez Homera Dudleya posiadało jednak jedną dużą wadę. Do poprawnego działania wymagany był długi czas nauki operatora oraz zapoznanie się z jej funkcjonowaniem. Sygnał dźwiękowy był rozdzielany na kilka pasm częstotliwościowych, a następnie przepuszczany

przez szereg filtrów. F0 było kontrolowane za pomocą pedału, palce kontrolowały wzmocnienia poszczególnych pasm, syczenie regulowane było za pomocą nadgarstka. Dodatkowe trzy przyciski kontrolowały pobudzenie wybranych filtrów w celu osiągnięcia głosek plozywnych. Urządzenie to zostało zaprezentowane publiczności w 1939 roku podczas „*World Fair*" (*Światowe Targi*) w Nowym Jorku.

Kolejnym ważnym osiągnięciem było stworzenie formantowego syntezatora mowy (Rozdział 2.2.2) przez Johna Holmesa. Działanie tego modelu opierało się o wykorzystanie odpowiednich filtrów. Na wejściach filtrów podawany był sygnał elektryczny będący tonem harmonicznym. Filtry pełniły rolę rezonatorów toru głosowego.

2.2 Metody syntezy akustycznej

Rodzaje syntezatorów dzieli się ze względu na sposób formowania sygnału mowy. Wyróżnia się: syntezę regułową (synteza formantowa, artykulacyjna) oraz syntezę konkatenacyjną (jej odmianą jest synteza korpusowa).

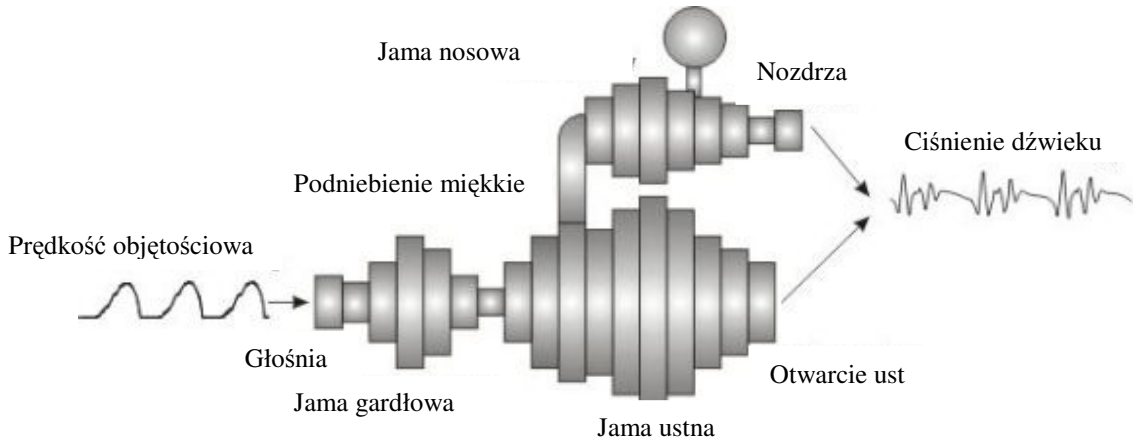
2.2.1 Synteza artykulacyjna

Artykulacyjna synteza mowy polega na modelowaniu rzeczywistego narządu artykulacyjnego. Model ten wymaga dynamicznego modelu toru głosowego, który pozwala na symulacje ruchu artykulatorów podczas procesu generowania mowy. (Wagner 2008)

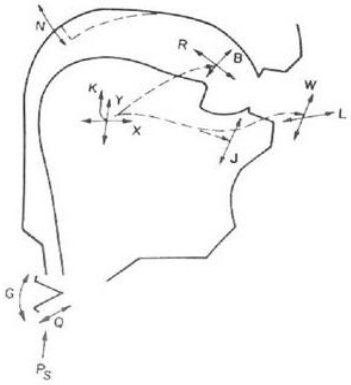
W modelu artykulacyjnym można kontrolować następujące parametry: szerokość otworu ust oraz ich wysunięcie, pozycję, wysokość języka jak i szerokość otworu głośni, napięcie strun głosowych oraz ciśnienie w płucach. (Lemmetty 1999)

Obecnie, z uwagi na skomplikowaną budowę oraz liczne problemy związane z konstrukcją, a także dużą złożoność obliczeniową i matematyczną przedstawionego modelu, synteza artykulacyjna nie jest rozpowszechniona, jednak używa się jej tam gdzie występuje konieczność

kontroli wszystkich cech głosu np. w syntezie emocjonalnej. Rysunek 2.2 przedstawia schemat toru głosowego zbudowanego w oparciu o model składający się z odcinków rur cylindrycznych.



Rys. 2.2 Przykładowy model toru głosowego zbudowany (na podstawie przekrojów) w oparciu o odcinki rur cylindrycznych
(http://wwwicg.informatik.uni-rostock.de/~piet/speak_main.html)



L- oś stopnia wysunięcia warg
W- oś stopnia otwarcia warg
J- oś położenia szczęki dolnej
X- pozioma oś położenia masy języka
Y- pionowa oś położenia masy języka
K- (0,1) zamknięcie toru
N- otwarcie wlotu do nosa
B- podniesienie czubka języka
R- przesunięcie czubka język

Rys. 2.3 Uproszczone modelowanie ruchów artykulacyjnych (Gubrynowicz. 2004, Stevens 1998)

2.2.2 Synteza regułowa

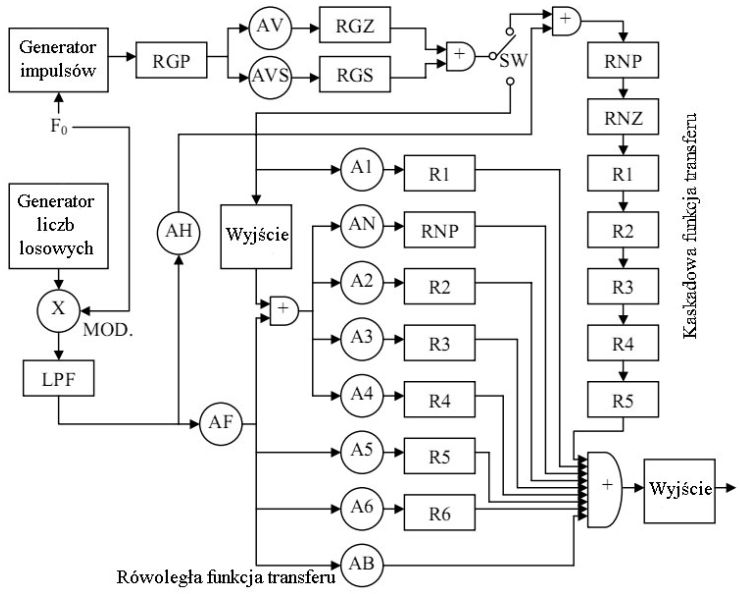
Kolejnym rodzajem syntezy jest synteza formantowa. Polega na modelowaniu funkcji przenoszenia toru głosowego w dziedzinie częstotliwości za pomocą filtrów cyfrowych. Filtry te połączone są ze sobą szeregowo i/lub równolegle i generują dźwięk o charakterystycznej barwie. Sygnał ten

odzwierciedla charakterystyczne formanty głosek. Do wygenerowania zrozumiałej mowy potrzebne jest odzwierciedlenie trzech formantów. Uzyskanie pięciu formantów pozwala na wygenerowanie dostatecznej, jakości mowy. Każdy formant jest modelowany za pomocą częstotliwości formantowej oraz pasma rezonansu. (Wagner2008).

Schemat elektronicznego formantowego syntezatora mowy został zaproponowany w 1979 przez Dennisa Klatta. Jest to pierwsza komputerowa symulacja powstawania mowy, składająca się z rezonatorów połączonych ze sobą równolegle bądź kaskadowo. W tej implementacji możliwe jest uzyskanie głosu męskiego bądź żeńskiego. Syntezator składa się z dwóch źródeł pobudzenia, pierwszy jest przeznaczony do syntezy samogłosek i charakteryzuje się opadaniem obwiedni widma 12 dB na oktawę. Drugi zaś, jest wykorzystywany do syntezy głosek trących, jest harmoniczny z opadaniem -24 dB. Przełącznik SW pozwala wysyłać oba sygnały do modułu równoległego bądź kaskadowego. Rysunek 2.4 przedstawia schemat syntezatora formantowego Dennisa Klatta. (Klatt 1987) Poniżej znajdują się objaśnienia do rysunku 2.4 Spośród 40 parametrów 34 z nich może być zmieniane dynamicznie – parametry te oznaczono literą „V", pozostałe posiadają oznaczenie „C".

DU	30	5000	Czas trwania zdania (ms)
NWS	1	20	Czas przerwy na skasowanie parametrów (ms)
SR C	5000	20000	Częstotliwość próbkowania (Hz)
NF C	1	6	Numery formantów w module kaskadowym
SW C	0	1	Kaskadowe/Równoległe pobudzenie toru głosowego przez
G0 C	0	80	Wzmocnienie sygnału (dB)
F0 V	0	500	Częstotliwość podstawowa (Hz)
AVS V	0	80	Amplituda sygnału przeznaczonego do akcentowania głosek (dB)
FGP V	0	600	Częstotliwość rezonatora "RGP"
BGP V	50	2000	Pasmo rezonatora "RGP"
FGZ V	0	5000	Częstotliwość antyrezonatora podgłośniowego "RGZ"
BGZ V	100	9000	Pasmo antyrezonatora podgłośniowego "RGZ"
BGS V	100	1000	Pasmo rezonatora głośni "RGS"
AH V	0	80	Amplituda aspiracji (dB)
AF V	0	80	Amplituda sygnału do generowania głosek trących (dB)
F1 V	180	1300	Częstotliwość 1 formantu (Hz)
B1 V	30	1000	Pasmo rezonansu 1 formantu (Hz)
F2 V	550	3000	Częstotliwość 2 formantu (Hz)
B2 V	40	1000	Pasmo rezonansu 2 formantu (Hz)
F3 V	1200	4800	Częstotliwość 3 formantu (Hz)

B3 V	60	1000	Pasmo rezonansu3 formantu (Hz)
F4 V	2400	4990	Częstotliwość 4 formantu (Hz)
B4 V	100	1000	Pasmo rezonansu4 formantu (Hz)
F5 V	3000	6000	Częstotliwość 5 formantu (Hz)
B5 V	100	1500	Pasmo rezonansu5 formantu (Hz)
F6 V	4000	6500	Częstotliwość 6 formantu (Hz)
B6 V	100	4000	Pasmo rezonansu6 formantu (Hz)
FNP V	180	700	Częstotliwość nosowego zero (Hz)
BNP V	40	1000	Pasmo nosowego zero (Hz)
FNZ V	180	800	Częstotliwość nosowego zero (Hz)
BNZ V	40	1000	Pasmo nosowe (Hz)
AN V	0	80	Amplituda nosowego formantu (dB)
A1 V	0	80	Amplituda 1 formantu (dB)
A2 V	0	80	Amplituda 2 formantu (dB)
A3 V	0	80	Amplituda 3 formantu (dB)
A4 V	0	80	Amplituda 4 formantu (dB)
A5 V	0	80	Amplituda 5 formantu (dB)
A6 V	0	80	Amplituda 6 formantu (dB)
AB V	0	80	Amplituda ścieżki “bypass” (dB)



Rys. 2.4 Schemat formantowego syntezatora mowy Dennisa Klatta.(Klatt 1987)

Należy dodać, że system ten został wykorzystany w syntezie komercyjnej Telesensory Systems Inc. Do dziś zaproponowany model przez Klatta jest wykorzystywany w syntezatorach formantowych.

Synteza formantowa ma nadal swoje zastosowanie szczególnie dla osób niedowidzących. Jej główną zaletą jest szybkość generowania sygnału mowy oraz niewielka moc obliczeniowa, jaka jest potrzebna do wygenerowania sygnału mowy. Wadą systemów formantowych jest brak naturalności.

2.2.3 Synteza konkatenacyjna

Model tej syntezy mowy, rozwijany od lat 70, zyskał dużą popularność z uwagi na możliwość generowania w stosunkowo prosty sposób bardzo naturalnej, dobrze brzmiącej i zrozumiałej mowy.

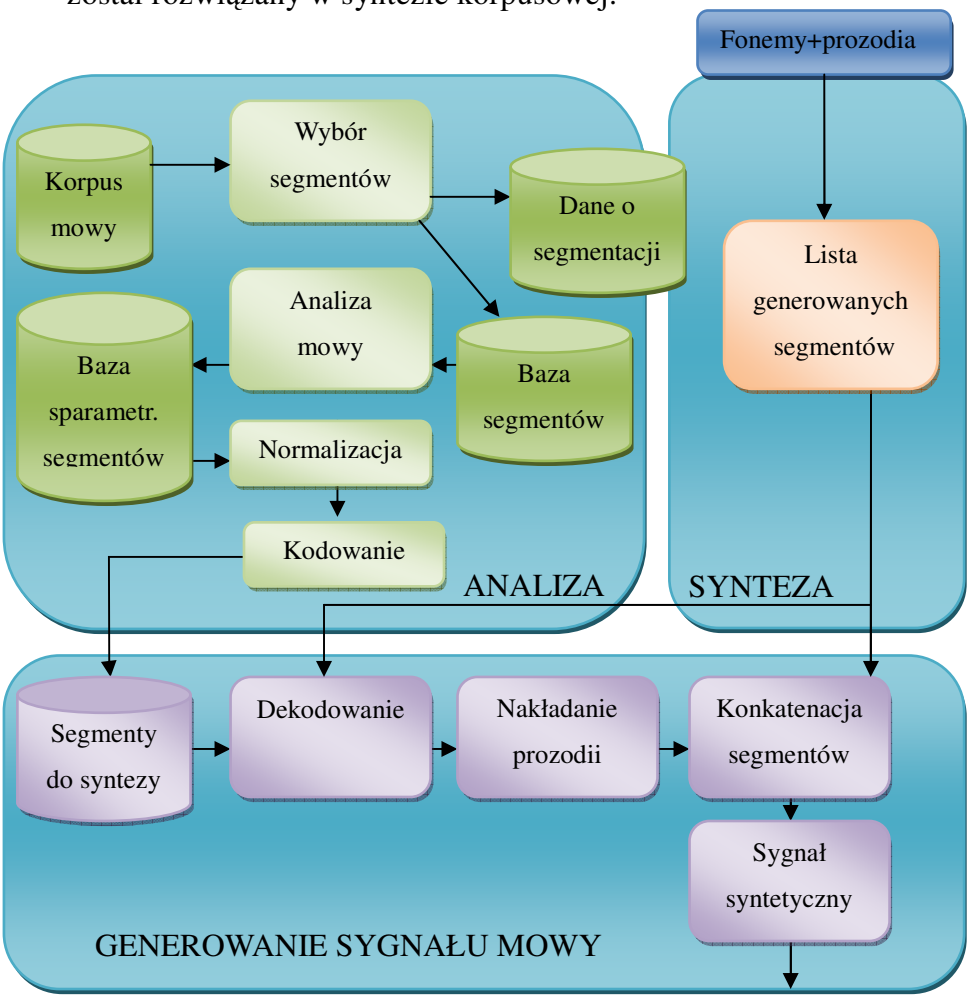
Pierwsze syntezatory generowały mowę słabej jakości i nienaturalnie brzmiącą. Głównym powodem był brak modelu intonacyjnego. Synteza mowy konkatenacyjnej generuje mowę poprzez łączenie ze sobą segmentów akustycznych powstałych z naturalnej mowy (głoski, difony, trifony, sylaby). Dużą zaletą tego rodzaju syntezy jest niewielki rozmiar bazy danych. Im mniejszy rozmiar bazy, tym szybciej będzie syntetyzowana mowa oraz wymagania sprzętowe będą mniejsze.

Konkatenacja mowy oparta na łączeniu wyrazów jest bardzo niepraktyczna z powodu ilości wyrazów, jakie należałoby przechowywać w pamięci. Poza tym nagrywanie korpusu składającego się z oddzielnych słów nie do końca ma sens, ponieważ brakuje tu przejścia naturalnego pomiędzy jednym a drugim słowem. Przyjęcie jako jednostki segmentalnej głoski nie jest dobrym rozwiązaniem z uwagi na brak możliwości symulowania koartykulacji. Kolejną wadą zastosowania głosek jest brak tranzjentów. Tranzjent definiuje się jako krótkotrwałą zmianę wewnętrznej struktury segmentu podczas przejścia z jednej głoski do drugiej. Wpływa on na wyrazistość i rozpoznanie dźwięku, barwy. Stosując jednostki tranzjentowe (difony) w syntezie można uzyskać większą naturalność mowy, ponieważ tranzjent stanowi reprezentację akustyczną procesu koartykulacji.

Konkatenacja sylab daje dość dobre rezultaty, jednak z uwagi na ich ilość (np. w języku angielskim około 160000, podczas gdy jest tylko 40 fonemów) nie jest rozsądnym rozwiązaniem. Bardzo często stosowana jest konkatenacja difonów. Difon jest jednostką akustyczną, która zawiera przejście między jedną a drugą głoską. To płynne połączenie pozwala na uzyskanie dobrej jakości syntezy mowy przy wykorzystaniu korpusu zawierającego około 1500 jednostek. (Szklanny 2002, Oliver 1998).

Dobór jednostek akustycznych jest jednym z istotniejszych problemów w tej metodzie syntezy. Dłuższe jednostki pozwolą na uzyskanie

bardziej naturalnej mowy. Ważnym problemem jest możliwość kształtowania prozodii syntezowanego przebiegu, czyli problemu czasu trwania poszczególnych jednostek segmentalnych oraz ich intonacji. Problem ten został rozwiązany w syntezie korpusowej.



Rys. 2.5 Schemat syntezy konkatenacyjnej.(na podstawie Gubrynowicz 2004)

2.2.4 Synteza korpusowa

Zamiast tworzyć bazę zawierającą tylko pojedyncze wystąpienie akustycznej jednostki segmentalnej, przygotowuje się specjalny korpus zawierający wiele wystąpień danej jednostki w różnych kontekstach oraz wykorzystuje się różnej długości jednostki akustyczne. Często dzięki temu unika się wielu sztucznych połączeń segmentów i generowana mowa jest zbliżona do naturalnej.

Najistotniejszym elementem odpowiedzialnym za selekcję segmentów

jest funkcja kosztu. Funkcja ta składa się z kosztu doboru jednostki *(target-cost)* oraz kosztu konkatenacji *(join-cost).*

Wzór 2 definiuje funkcję kosztu.

$$d(\Theta,T)=\sum_{j=1}^{N} d_u(\theta_j,T)+\sum_{j=1}^{N-1} d_t(\theta_j,\theta_{j+1}) \tag{2}$$

gdzie :

$$d_u(\theta_j,T)$$

stanowi koszt doboru jednostki, a

$$d_t(\theta_j,\theta_{j+1})$$

oznacza koszt połączenia elementów θ_j,θ_{j+1}

Optymalny ciąg segmentów jest wyznaczony jako:

$$\hat{\Theta}=\arg\min_{\Theta} d(\Theta,T) \tag{3}$$

(Huang i wsp. 2001)

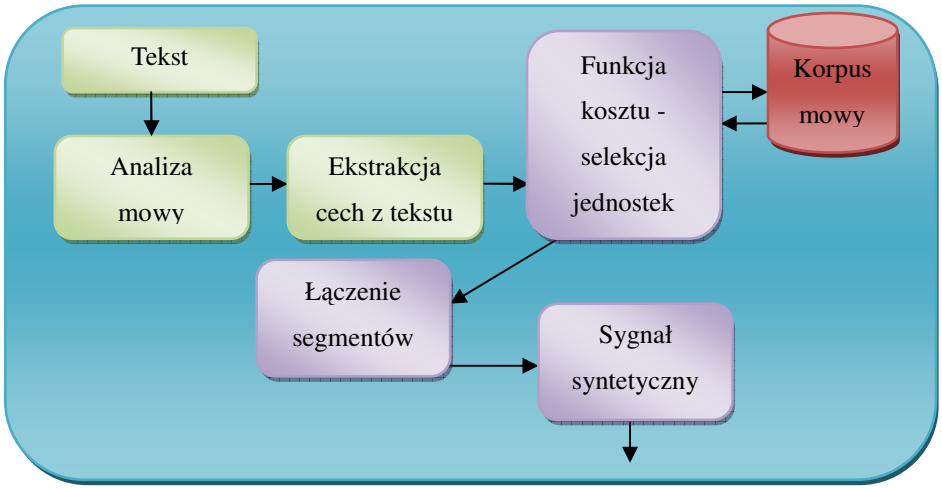
Im wartość zwracana przez funkcję kosztu będzie mniejsza, tym większe jest prawdopodobieństwo na otrzymanie bardziej naturalnie brzmiącego zdania. Jeśli wszystkie koszty konkatenacji będą jednakowe, to ciąg o najmniejszej ilości elementów będzie miał najniższy koszt. W praktyce często pociąga to za sobą wybranie jak najdłuższych jednostek. Należy jednak zwrócić uwagę, że wybór najdłuższych jednostek niesie ze sobą pewne zagrożenie. Istnieje bowiem małe, ale jednak istotne prawdopodobieństwo, że wybór najdłuższych jednostek nie będzie zgodny prozodycznie ze zdaniem, które ma być zsyntezowane. Dlatego zaprojektowanie funkcji kosztu pod względem generowania jak najdłuższych jednostek nie jest rozwiązaniem optymalnym. (Marasek2003 B).

Funkcja kosztu zależy od wielu czynników takich jak koszt akcentu, melodii, pozycji segmentu w słowie/wyrazie/frazie, dlatego baza akustyczna powinna być pod kątem tych czynników opisana. Bardzo często stosuje się techniki optymalizujące wyszukiwanie, np. poprzez ograniczanie przestrzeni wyszukiwania. Konstrukcja funkcji kosztu jest zadaniem trudnym, ponieważ

znalezienie wag parametrów oraz zależności między nimi ma pozwolić na wyszukanie takich jednostek, których łączenie ze sobą, pozwoli na uzyskanie naturalnej mowy.

Mało jest publikacji na ten temat, zwykle firmy zajmujące się tworzeniem systemów syntezy mowy funkcję kosztu traktują jako najbardziej strzeżoną informację. (Black i wsp. 1996)

Rysunek 2.6 przedstawia schemat działania korpusowego syntezatora mowy. W przeciwieństwie do syntezy regułowej na poziomie modułu DSP *(Digital Signal Processing)* zazwyczaj nie występuje modyfikacja sygnału. Modelowanie prozodyczne zostało zastąpione modułem wyszukiwania odpowiednich jednostek *(funkcja kosztu)*.



Rys. 2.6 Schemat syntezatora korpusowego

2.3 Przegląd korpusowych syntezatorów mowy dla języka polskiego

2.3.1 RealSpeak

Firma Lernout&Hauspie stworzyła pierwszy system korpusowej syntezy mowy dla języka polskiego nazwany RealSpeak w 2003 roku. System RealSpeak jest skalowalny systemem. To znaczy w zależności od platformy sprzętowej zmienia się rozmiar bazy akustycznej i dostosowuje do możliwości systemowych urządzenia i zajmuje odpowiednio od 8 do 100 MB.

W systemie brakuje elementów paralingwistycznych i dlatego zdarza

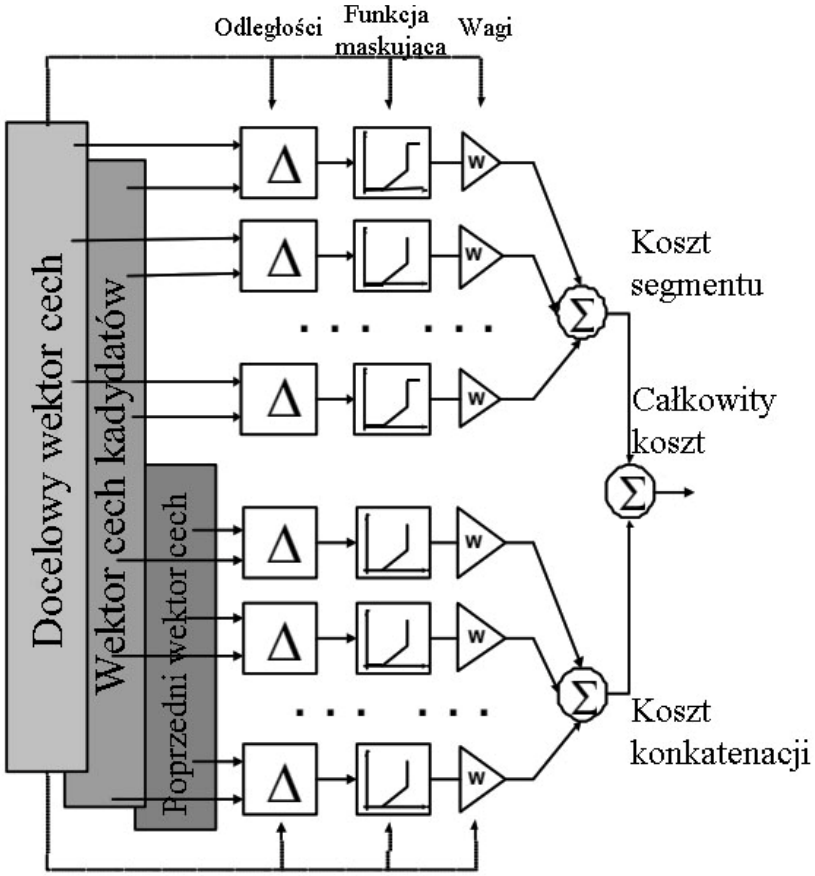
się, że syntezowana nie zawsze jest zbliżona do mowy naturalnej.

Synteza w systemie RealSpeak polega na wybieraniu odpowiednich difonów z bazy akustycznej a następnie łączeniu ich sekwencji ze sobą. Możliwe jest, że do konkatenacji zostanie wybrana jednostka dłuższa, gdy difony będą znajdowały się obok siebie w bazie i stanowiły sekwencję zdania docelowego. Funkcja kosztu wyznacza najbardziej pasujące jednostki pod względem kryteriów lingwistycznych. Minimalizowana jest również ilość słyszalnych zniekształceń w oparciu o różnice spektralne pomiędzy łączonymi kandydatami.

Do wyszukiwania optymalnych jednostek użyto metody dynamicznego programowania (Bellman 1954). Idea metody programowania dynamicznego polega na tym, że zadanie optymalizacyjne z N zmiennymi decyzyjnymi rozkłada sie na N zadań optymalizujących z jedną zmienną każde. Poszczególne etapy tego procesu są łączone ze sobą w oparciu o proces rekurencyjny. Według Bellmana procedura optymalizacyjna ma te własność, że niezależnie od początkowego stanu i początkowej decyzji pozostałe decyzje musza stanowić procedurę optymalizacyjną ze względu na stan wynikający z pierwszej decyzji.

Schemat funkcji kosztu przedstawiono na rysunku 2.7. Sposób łączenia jednostek opiera się na 3 głównych wektorach cech. Są to: wektor cech kandydata, wektor cech docelowego difonu oraz wektor cech poprzedniego difonu. Każdy z tych wektorów jest opisany przez: pozycję w sylabie, pozycję we frazie, wartość F0, czas trwania jednostki, wartość współczynnika ciągłości spektrum w difonie na granicy głosek, oraz kontekst fonetyczny.

Symbole delty na rysunku 2.7 reprezentują miary odległości pomiędzy wartościami tej samej cechy. Wynikiem działania jest znalezienie podobieństwa pomiędzy kandydatem proponowanym a docelowym dla kosztu doboru jednostki oraz stopień zgodności między dwoma kandydatami dla kosztu konkatenacji.



Rys. 2.7 Schemat funkcji kosztu w systemie L&H (Coorman i wsp. 2000)

Podobieństwo to jest oszacowywane za pomocą trzech funkcji:

- symbolicznej miary odległości – miara to może przyjąć wartość 0 jeśli porównywane odległości są identyczne lub 1 w przeciwnym wypadku. Zazwyczaj stosuje się bardziej zaawansowane metody określenia odległości, opisujące więcej stopni dopasowania, niż tylko wartość boolowską. Na przykład do głoskę /p/ z lewym kontekstem /b/ będzie lepiej konkatenować niż /p/ z sąsiedztwem /s/. Wynikiem takiej operacji będzie liczba rzeczywista z przedziału (0,1)
- skalarnej miary odległości – w której wylicza się bezwzględną wartość funkcji np. czasu trwania głoski, dla której skrócenie będzie bardziej karane niż jej wydłużenie
- odległości między wektorami (dla wszystkich cech pomiędzy wszystkimi segmentami) – jest to funkcja oszacowująca nieciągłości spektralne pomiędzy wektorami cech. Ponieważ stworzenie macierzy

jest bardzo złożone obliczeniowo stosuje się techniki optymalizacyjne takie jak kwantyzacje wektorową *(VQ)*(Gersho i wsp. 1991)

Dodatkowo w systemie zastosowano dwie metody ograniczenia wyszukiwanej bazy. Pierwsza metoda tzw. progowania przezroczystego (ang. *Transparency Threshold*) (Coorman i wsp. 2000) określa pewną dopuszczalna różnicę wartości F0 między łączonymi ze sobą jednostkami, poniżej której ucho ludzkie nie usłyszy nieciągłości. Funkcja ta wyszukuje i mapuje te wszystkie jednostki oznaczając je wartością 0, co pozwala uniknąć wyliczania wielu drobnych kosztów w finalnej sekwencji zdania. Druga funkcja *(Quality Threshold)* działa odwrotnie. Mapuje wszystkie jednostki, których połączenie ze sobą spowoduje słyszalne zniekształcenia. Obie metody umożliwiają znacznie zredukować czas potrzebny na wyszukiwanie optymalnych jednostek.

W celu optymalizacji jakości syntezy mowy zaimplementowano moduł realizujący koartykulację dla spółgłoski /r/, ponieważ jej realizacja silnie zależy od typu sąsiadujących z nią głosek, a także od położenia jej w wyrazie lub we frazie.

W przypadku sonorantów zwiększono wagę pozycji w sylabie funkcji kosztu. Dla jednostek znajdujących się w ostatniej sylabie wyszukiwane są jednostki o znacznie dłuższym czasie trwania. System reaguje również na rodzaj zdania (pytające, oznajmujące, wykrzyknikowe) i optymalizuje wyszukiwanie jednostek w zależności od niego wymaganego przebiegu konturu melodycznego.

Firma Lernout&Hauspie została przejęta przez ScanSoft, a obecnie należy do firmy Nuance.

2.3.2 Loquendo

Loquendo jest włoską firmą zajmującą się przetwarzaniem sygnału mowy. W swojej ofercie posiada zarówno wielojęzyczne systemy syntezy jak i rozpoznawania mowy. Firma kładzie głównie nacisk na tworzenie syntezy wysokiej jakości głosów naturalnych i ekspresyjnych. Prowadzi obecnie szczegółowe badania nad odwzorowaniem emocji w syntezie mowy.

Dla języka polskiego zostały stworzone w bazie danych dwa głosy, żeński – Zosia oraz męski Krzysztof.

Każdy głos może być syntezowany z częstotliwością próbkowania 48 kHz. W projektowanych systemach oferowany jest moduł o nazwie Language Guesser, który umożliwia identyfikację języka oraz zastosowanie odpowiednich reguł językowych. Dzięki temu obcojęzyczne słowa zostaną właściwie wypowiedziane. Dodatkowo system umożliwia odczytywanie e-maili oraz smsów.

Firma udostępnia oprogramowanie, dzięki któremu można zmodyfikować tembr głosu, jego wysokość oraz szybkość czytania.

W każdym z głosów zawarte są elementy paralingwistyczne takie jak: kichanie, śmiech, kaszel itp. W ofercie posiada następujące głosy: Amerykańsko-angielski, kanadyjsko-francuski, brazylijsko–portugalski, amerykańsko- hiszpański, argentyńsko-hiszpański, chilijsko-hiszpański, meksykańsko-hiszpański, brytyjsko-angielski, hiszpański (kastylijski, kataloński, galicyjski), holenderski, francuski, niemiecki, grecki, włoski, polski, portugalski, szwedzki, turecki, rosyjski, fiński, duński, mandaryński. Informacje dotyczące funkcji kosztu, która jest stosowana w tym systemie, nie zostały opublikowane.

2.3.3 Acapela

Acapela Group udostępnia systemy TTS w 25 językach, oferując ponad 50 głosów. Posiada trzy główne technologie – syntezę korpusową, syntezę konkatenacyjną w oparciu o difony oraz systemy ASR. Acapela oferuje głos Ani dla języka polskiego. Wadą systemu jest dość specyficzny francuski akcent głosu.

System Acapeli posiada moduł intonacji zbliżony do naturalnej, obsługuje elementy paralingwistyczne. Umożliwia syntezę z częstotliwością próbkowania dźwięku do 22 kHz.

Acapela jest dostępna również na platformy mobilne takie jak: Linux embedded, Symbian, Windows mobile. Informacje dotyczące funkcji kosztu nie zostały nigdzie opublikowane.

2.3.4 BOSS

Pierwszym systemem syntezy mowy korpusowej stworzonym w Polsce był system stworzony w ramach współpracy pomiędzy Uniwersytetem Adama Mickiewicza oraz IKP *(Institut für Kommunikationsforschung und Phonetik)* w Bonn. Projekt na Uniwersytecie w Poznaniu był koordynowany przez prof. Grażynę Demenko. TTS został zrealizowany w systemie BOSS *(Bonn Open Synthesis System)* (Klabbers i wsp. 2001 B,Demenko i wsp. 2007, 2008).

Baza akustyczna zawiera około jednej godziny tekstów czytanych, w tym monologów, dialogów, a także wiadomości prasowych. W bazie znajduje się w wersji podstawowej około 1200 difonów (występujących w kontekście samogłoski /a/) a w wersji rozszerzonej około 4000 difonów w różnych kontekstach. Dodatkowo baza zawiera około 6000 trifonów typu CVC. (Janicki 2004). Nagrano również frazy z najczęściej występującymi strukturami spółgłoskowymi. Bazę wzbogacono o 2320 zdań z 6000 najczęściej występujących słów. Do bazy zostały dodane zakończenia wypowiedzi, nieakcentowane wyrazy funkcyjne, liczebniki oraz zbitki spółgłoskowe.

W systemie zaimplementowano dwa moduły języka polskiego:

- moduł predykcji iloczasów
- moduł funkcji kosztu

W (Demenko 2008 i wsp. B) opisano sposób działania tej funkcji. Składa się ona z kilku części. Pierwszy moduł oblicza bezwzględną różnicę pomiędzy czasem trwania segmentu oszacowanym na podstawie drzew CART, a rzeczywistym czasem trwania poszczególnego kandydata. Drugi moduł wylicza różnicę typu boolowskiego pomiędzy przewidywanym a aktualnym typem akcentu (waga 10). Trzeci moduł wylicza rozbieżności pomiędzy rodzajem frazy (pytanie lub stwierdzenia, wzrastająca lub opadająca intonacja). Ostatnim elementem jest lokalizacja frazy w zdaniu (waga 20).

System łączy ze sobą prozodycznie podobne wyrazy, jeśli ich brakuje, wyszukuje odpowiednio sylaby, a następnie głoski. System nie modyfikuje prozodii uzyskanego w ten sposób sygnału. W systemie BOSS realizacja funkcji kosztu sprowadza się do wyboru odpowiednich jednostek na podstawie

cech suprasegmentalnych kontekstu fonetycznego i położenia jednostki akustycznej we frazie. Na etapie łączenia tworzony jest zbiór wszystkich dopuszczalnych kandydatów. Jednostki akustyczne reprezentowane są jako węzły grafu. Linie łączące wierzchołki reprezentują możliwe przejścia między głoskami. Ścieżka przez graf jest tworzona w taki sposób, by koszt znalezienia najbardziej odpowiednich jednostek był minimalny. Do wyliczenia zniekształceń spektralnych używa się odległości euklidesowej bazującej na współczynnikach MFCC, która nie jest najlepszym rozwiązaniem w wyznaczaniu funkcji kosztu konkatenacji, ponieważ nie jest dobrze skorelowana z percepcją mowy (Klabbers i wsp. 2001). Dlatego planowany jest dalszy rozwój systemu oraz wprowadzenie symetrycznej miary Kullback-Leiblera *(SKL)*. W systemie zaimplementowano również moduł obliczający bezwzględną różnicę wartości F0 pomiędzy segmentami. Zamiana tekstu ortograficznego na fonetyczny oparta jest na zmodyfikowanym kodowaniu SAMPA. (Demenko i wsp. 2007, 2008 B).

Z przeprowadzonych badań (Demenko i wsp. 2008 B), wynika że pozycja sylaby w słowie jest szczególnie istotna. Na jakość syntezy mają wpływ również:

- modelowanie rytmiczne wypowiedzi
- sąsiedztwo segmentu
- pozycja sylaby w zdaniu
- rodzaj sekwencji spółgłoskowej
- typ kontekstu spółgłoskowego poprzedzającego samogłoskę
- iloczas jednostki (długość wypowiedzi może być modyfikowana przez segmentalne jaki suprasegmentalne cechy)

2.3.5 Ivosoftware

Firma Ivosoftware powstała w 2001 roku. Jej celem jest opracowywanie i wprowadzanie na rynek produktów nowej technologii głosowej. Pierwszym komercyjnym produktem IVO Software był syntezator mowy - Spiker 1.0. Był to syntezator konkatenacyjny.

W czerwcu 2006 roku zaprezentowano Expressivo Demo Release

Candidate 1. Jest to syntezator korpusowej syntezy mowy połączony z technologią odczytywania e-maili, kanałów RSS oraz artykułów. Program posiada możliwość tworzenia audiobooków z dowolnej postaci tekstowej oraz podkładania głosu lektora do filmów. Obecnie dostępne są 4 głosy: Jacek, Maja, Jan, Ewa.

Firma stworzyła cyfrowego lektora w postaci programu Expressivo. Jest to program komputerowy czytający książki cyfrowe, wiadomości, artykuły, kanały RSS. Program pozwala na tworzenie audiobooków, oraz na oglądanie filmów z lektorem. (Ivo Software , 2008).

Proces syntezy sprowadza się do analizy przetworzenia tekstu oraz ekstrakcji cech lingwistycznych dla modelu funkcji kosztu. Następnie generowany jest kontur F0, poczym następuje konkatenacja odpowiednich jednostek z bazy akustycznej. IVONA wykorzystuje algorytm USLTM (Unit Selection algorithm with Limited Time-scale Modifications). Jest on oparty na funkcji kosztu i odpowiedzialny za wyszukiwanie optymalnych jednostek do konkatenacji. Umożliwia modyfikację iloczasu wybranych jednostek. Algorytm ten składa się z dwóch części: kosztu konkatenacji oraz kosztu doboru jednostki (wzór 4).

$$Koszt\ (u) = koszt\ doboru\ jednostki\ (u) + koszt\ konkatenacji\ (u) \quad (4)$$

gdzie *(u)* oznacza jednostki z bazy akustycznej.

Funkcja doboru jednostki(Kaszczuk i wsp. 2007) wyszukuje najlepsze jednostki na podstawie około 40 cech w wektorze wyekstrahowanych z tekstu takich jak pozycja jednostki w sylabie, w słowie i zdaniu oraz kontekst fonetyczny, akcent frazowy, akcent wyrazowy.

Koszt konkatenacji wyliczany jest na podstawie takich parametrów jak:

- melodia
- moc sygnału,
- sygnał harmoniczny/szumowy,
- długość poszczególnej akustycznej jednostki segmentalnej
- współczynniki cepstrum znormalizowane do 16 punktowej krzywej interpolowanej funkcją *spline*

Pozostałe spośród 40 cech nigdzie nie zostały opublikowane.

Do przeszukiwania bazy wybrany został algorytm efektywnego

programowania dynamicznego, który umożliwia znajdowania najlepszych kandydatów w czasie rzeczywistym.

W przypadku pojawienia się różnic czasów trwania poszczególnych jednostek akustycznych odbiegających od wartości oczekiwanych uzyskanych z modelu intonacyjnego, modyfikowany jest ich czas. Funkcja modyfikująca działa w dziedzinie czasu z synchronizacją F0 i pozwala wyeliminować wiele zniekształceń.

Jednostki akustyczne łączne są w dziedzinie czasu przy zastosowaniu algorytmu OLA (OverLap and Add).Są to jedyne informacje opublikowane przez twórców systemu. (Kaszczuk i wsp. 2007)

2.3.6 Podsumowanie polskich systemów korpusowej syntezy mowy

Testując opisane systemy za wyjątkiem systemu BOSS (brak możliwości odsłuchania syntetycznej mowy) można stwierdzić, iż generują one naprawdę naturalną mowę. Krokiem milowym w syntezie dla języka polskiego był Realspeak, który był pierwszym systemem korpusowej syntezy mowy dla języka polskiego. Jest on stosowany do dnia dzisiejszego przez m.in. osoby niedowidzące. Jak już wspomniano brakuje w nim elementów paralingwistycznych, co powoduje pewną nienaturalność generowanej mowy. System Loquendo jest pozbawiony tej wady. Jednak zarówno Loquendo jak i Acapela nie są zbyt rozpowszechnione na polskim rynku.

System IVONA jest najbardziej rozpowszechniony na rynku polskim. Gwarantuje bardzo dobrą jakość syntetycznej mowy. W publikacji (Kaszczuk i wsp. 2007) omówiono tylko architekturę systemu i podstawy związane z funkcją kosztu. Najwięcej konkretnych informacji o funkcji kosztu i architekturze systemu umieszczono w (Demenko 2008 B).

W publikacjach opisanych systemów korpusowych pominięto wiele zasadniczych informacji. Przede wszystkim brakuje danych na temat sposobu optymalizacji funkcji kosztu oraz przydatności zastosowania heurystycznych metod, co stanowi tezę niniejszej pracy. Pominięto także wyniki badań dotyczące wpływu procesu optymalizacji na jakość syntetycznej mowy.

Ważnym problemem w realizacji korpusowego systemu jest właściwy wybór mówcy oraz sposób rejestracji bazy akustycznej, ponieważ od niej zależy końcowa jakość syntetycznej mowy. Odpowiedzi na postawione pytania są istotne z punktu widzenia projektowania systemu korpusowego dlatego zostały podjęte w niniejszej pracy.

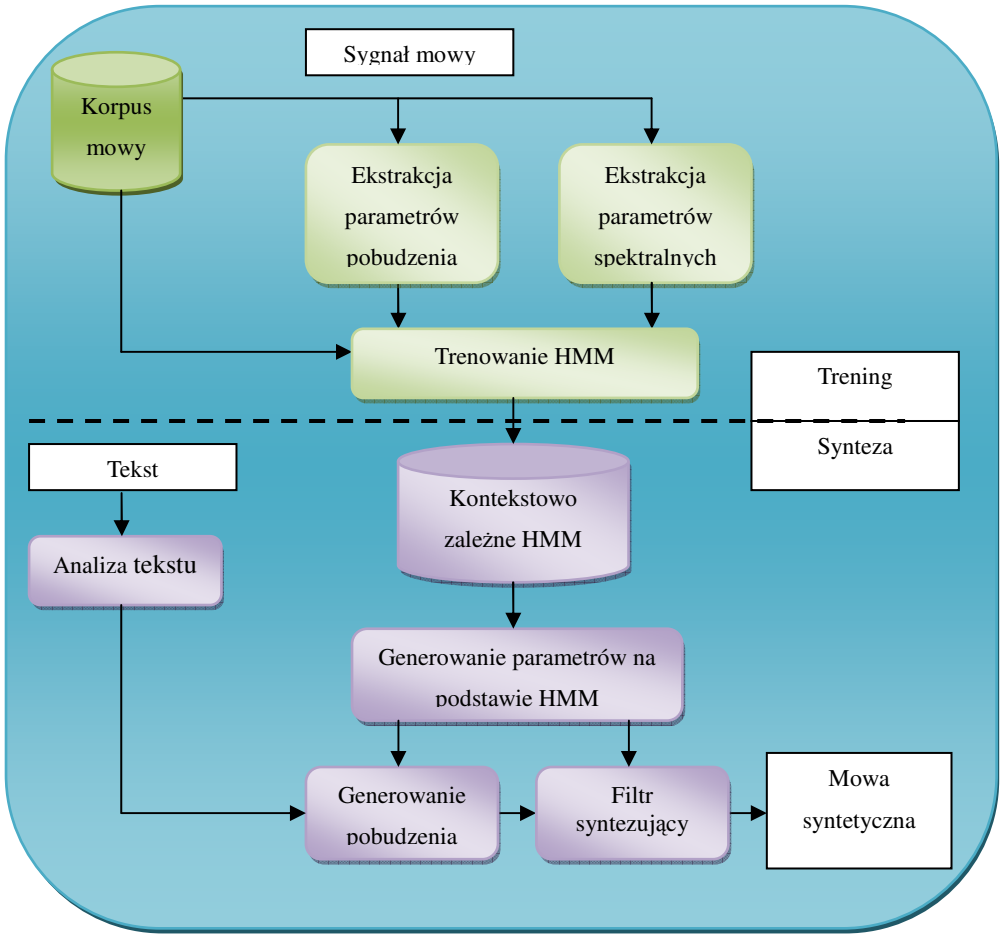
2.3.7 Synteza statystyczna (HTS)

Synteza HTS jest stosunkowo nowym rozwiązaniem, (Tokuda i wsp. 2002) wykorzystuje ukryte modele Markowa (*HMM-based speech synthesis system).* Jest to rozwiązanie w pewnym sensie zbliżone do metody konkatenacyjnej. Jednak w omawianym przypadku w procesie syntezy nie wykorzystuje się fragmentów mowy naturalnej, lecz kontekstowo zależne HMM. (Tokuda i wsp. 2002). Modele te są łączone odpowiednio do przetwarzanego tekstu, a wygenerowane przez nie wektory cech (obserwacje) są podstawą do syntezy mowy realizowanej przez odpowiedni filtr. Należy zaznaczyć, iż osobno modelowane są parametry dotyczące widma (lub cepstrum) i parametry dotyczące tonu krtaniowego (F0, dźwięczność). Dzięki rozdzieleniu tych przebiegów w dość łatwy sposób można modelować emocje, czy też zupełnie zmieniać charakterystykę głosu, wykorzystując techniki adaptacji modeli, opracowane na potrzeby rozpoznawania mowy. Jest to duża zaleta w porównaniu z metodą korpusową, gdzie modyfikacja głosu jest znacznie utrudniona. Istotną zaletą metody statystycznej jest też niewielki rozmiar wykorzystywanych w trakcie syntezy danych (np. 1MB pomijając moduły analizy tekstu) oraz duża szybkość działania. Oceniając jakość generowanej mowy należy zauważyć, że brzmi ona dobrze, choć mniej naturalnie niż w przypadku metody korpusowej. Metodę statystyczną cechuje wysoka zrozumiałość, stabilność i dobre modelowanie cech prozodycznych. W przeciwieństwie do syntezy konkatenacyjnej istnieje prostsza możliwość modyfikacji sposobu mówienia, czy też dodawania nowych dialektów. (Wójtowski 2007)

Interesującym podejściem w syntezie HTS jest wytrenowanie modeli na dużej bazie akustycznej, a następnie ich adaptacja do konkretnego mówcy. Takie podejście znacznie ułatwia tworzenie nowego syntetyzatora (Wagner

2008).

Synteza HTS została zaimplementowana w systemie Festival. Ogólny schemat syntezy mowy HTS przedstawiono na rysunku 2.8 (Tokuda i wsp. 2002)



Rys. 2.8 Schemat syntezy statystycznej na podstawie(Tokuda i wsp. 2002)

2.4 NLP na potrzeby syntezy mowy

W systemie TTS pierwszym ważnym elementem jest konwersja wprowadzonego tekstu na reprezentację lingwistyczną, odpowiedzialny jest za nią moduł NLP (przetwarzania języka naturalnego - ang.natural language processing). Realizuje on również i określa odpowiednią intonację i prozodię dla syntetyzowanego tekstu. W korpusowej syntezie mowy, moduł jest

odzwierciedleniem funkcji kosztu doboru jednostki.

Stworzenie modułu NLP jest zadaniem trudnym z uwagi na liczne niejednoznaczności języka naturalnego oraz skomplikowane zależności międzywyrazowe. 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 (Dutoit 1997, Taylor 2009).

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. Często po skrótach stawiany jest znak kropki, co nie zawsze oznacza koniec zdania np. „W 1973 r. kupiłem pierwszy telewizor" – skrót /r./ nie oznacza końca zdania, dodatkowo liczba 1973 powinno zostać rozwinięta do pełnych słów, oraz odpowiednio odmieniano (*tysiąc dziewięćset siedemdziesiątym trzecim).*W przypadku skrótów często pojawiają się dwuznaczności. Słownik zawierający ich rozwinięcia nie wystarczy, potrzebna jest dodatkowa analiza poprzedzającego słowa.*„On ważył 120 kg"* – gdzie *kg* jest rzeczownikiem w liczbie mnogiej.

Analizator morfologiczny *(POS)* 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, (przedrostków i przyrostków). Zadania analizatora morfologicznego sprowadzają się do zmniejszenia słownika oraz ustalenia części mowy.

Analizator kontekstowy ogranicza znaczenia poszczególnych słów. Ograniczenie to odbywa się na podstawie zbadania kontekstu słów (części mowy) znajdujących się w sąsiedztwie. Stosuje się tutaj metodę n-gramów

(Kupiec 1992, Willemse i wsp. 1992), która opisuje syntaktyczne zależności pomiędzy słowami, na zasadzie badania prawdopodobieństw w skończonych przejściach automatu. Służą do tego modele Markova lub wielowarstwowe sieci perceptronowe (Benello 1989). Użycie sieci neuronowych sprowadza się do stworzenia reguł rządzących kontekstem zdaniowym. Stosuje się również metody lokalnych niestochastycznych gramatyk tworzonych przez ekspertów lingwistyki lub pozyskiwanych z danych treningowych za pomocą drzew CART (Sproat i wsp. 1992, Yarowsky 1994, Dutoit 1997)

Parser syntaktyczno-prozodyczny jest odpowiedzialny za modelowanie 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. Po przygotowaniu modelu syntaktyczno-prozodycznego określa się precyzyjny czas trwania poszczególnych fonemów, głośności, wartości F0, oraz długości sylaby. Jednym z ważniejszych elementów w przygotowaniu modelu prozodii jest fokus zdania. Określa on właściwości F0, które wyróżniają sylabę lub grupy sylab z całej wypowiedzi. Do predykcji informacji prozodycznych stosuje się drzewa CART (Hirschberg 1991)

Projektowanie modelu prozodycznego nie jest zadaniem łatwym, jednak pożądanym dla każdego systemu mowy (Dutoit 1997). W korpusowej syntezie mowy mimo, iż bazuje ona na naturalności nagrywanego korpusu, coraz częściej przygotowuje się dodatkowo taki moduł.

W zdaniu „Ja nigdy nie powiedziałem, że ona ukradła moje pieniądze" w zależności od tego jakie słowo jest akcentowane przez mówiącego, zarówno w języku angielskim, jak i polskim zdanie to może mieć kilka odrębnych znaczeń:

„**On** nigdy nie stwierdził, że Jarek nie skończył studiów." – ktoś inny to stwierdził, ale nie on.

„On **nigdy** nie stwierdził, że Jarek nie skończył studiów." – on po prostu nigdy nie stwierdził tego.

„On nigdy **nie stwierdził**, że Jarek nie skończył studiów." – mogłem zasugerować to w inny sposób, ale nigdy nie powiedziałem tego wprost

„On nigdy nie stwierdził, że **Jarek** nie skończył studiów." -on nigdy nie stwierdził, że to właśnie Jarek nie skończył studiów, chodziło oczywiście o kogoś innego

„On nigdy nie stwierdził, że Jarek **nie skończył** studiów." - on nigdy nie stwierdził tego. Jarek je skończył tylko nie pokazał nikomu dyplomu

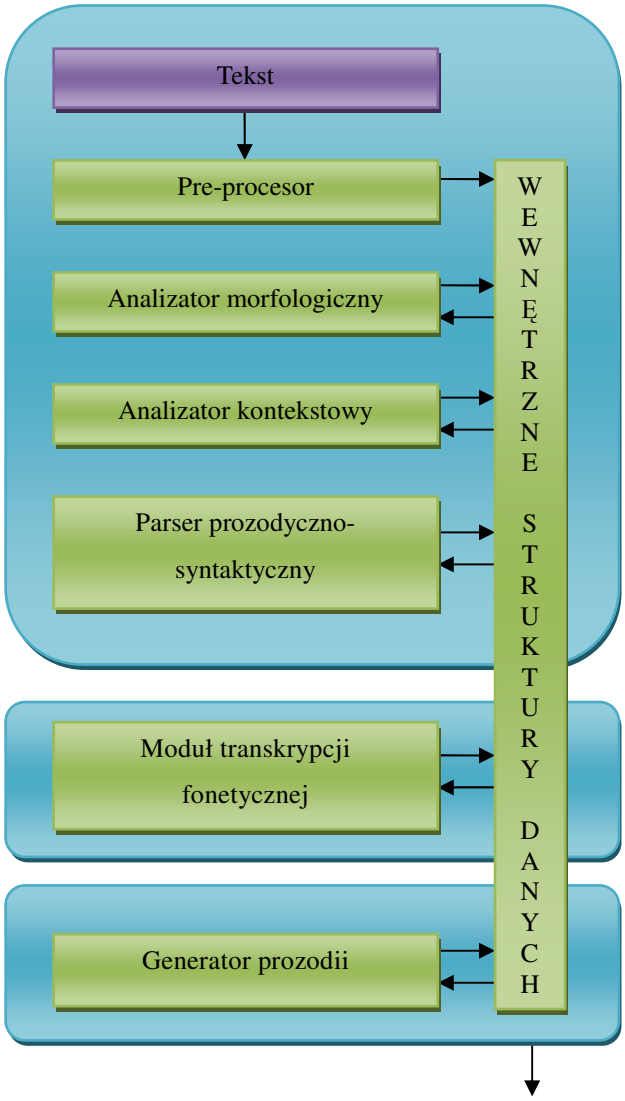
„**On** nigdy nie stwierdził, że Jarek nie skończył **studiów**." –on nie powiedział, że chodziło o studia. Jarek nie skończył kursu prawa jazdy!

Moduł *letter-to-sound* jest odpowiedzialny za utworzenie transkrypcji fonetycznej dla istniejących słów. Proponuje się (Dutoit 1997) realizację modułu w oparciu o słownik lub metodę regułową. Metoda słownikowa zawiera dużą ilość informacji fonologicznej w leksykonie. Chcąc zachować „rozsądny" rozmiar słownika przechowywane informacje ogranicza się do morfemów i morfo-fonologicznych reguł, które są stosowane podczas składania ich w słowa. Do transkrypcji morfemów, których nie ma w słowniku, stosowane są oddzielnie stworzone reguły. Następnie używane są reguły post-fonetyczne dla realizacji procesu koartykulacji. Stosunkowo innym rozwiązaniem jest metoda regułowa, które polega na stosowaniu reguł konwersji tekstu ortograficznego na fonetyczny. Do słów, które są umieszczone w słowniku wyjątków stosowane są oddzielne reguły transkrypcji (Dutoit 1997).

Przy realizacji modułu NLP powstaje kilka dość istotnych problemów. Słownik wymowy obejmuje tylko podstawowe słowa, bez morfologicznych kombinacji, to znaczy, nie uwzględnia on rodzaju, przypadku, liczby. Kolejnym problemem jest kwestia wyrazów obcojęzycznych, które już są lub stają się integralną częścią języka polskiego. 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. (np. *cis* /ts's/ to drzewo lub krzew, a *cis* /tsis/ to termin muzyczny).

Moduł NLP realizuje szereg procesów związanych z przekształceniem tekstu oraz ukształtowaniem gotowej wypowiedzi wraz z intonacją. Tak przygotowane dane trafiają w syntezie konkatenacyjnej do modułu DSP *(Digital Signal Processing),* a w korpusowej syntezie mowy do funkcji kosztu.

Rysunek 2.9 przedstawia schemat modułu NLP.



Rys. 2. 9 Moduł NLP

2.5 Festival

Obecnie synteza konkatenacyjna nie jest już tak popularna jak to miało miejsce jeszcze kilka lat temu. Główną wadą syntezy konktenacyjnej w Festivalu był brak naturalności brzmienia generowanej mowy. W celu osiągnięcia naturalności zastosowano rożne algorytmy mające poprawić jakość syntezy. Główne próby poprawy jakości były związane ze zwiększaniem długości jednostki akustycznej (sylaba) oraz z odpowiednim projektowaniem i balansowaniem korpusu. Algorytmy korpusowe pozwalają na osiągnięcie

dobrej jakości syntezy, jednak synteza ta również może generować bardzo słabej jakości mowę, szczególnie gdy korpus nie jest kompletny lub funkcja kosztu działa w sposób przypadkowy (Black i wsp. 2006).

Festival jest obecnie jedną z najlepiej rozwiniętych platform do realizacji systemów syntezy mowy. Jest to modułowy system i zapewnia wszystkie moduły potrzebne do tworzenia nowych głosów w syntezie konkatenacyjnej jak i korpusowej (Clark i wsp. 2004). System ten jednak jest trudny w wykorzystaniu oraz posiada sporo błędów (system uniwersytecki).

2.5.1 Rodzaje syntezy Unit-Selection w Festivalu

Festival oferuje kilka rodzajów algorytmów syntezy typu unit-selection:

- Clustergen
- Clunit
- Multisyn

Synteza typu Clustergen zwana również parametryczną, polega na trenowaniu modeli za pomocą algorytmu analizy skupień i wykorzystywaniu ich do syntezy w środowisku Festival. Wykorzystuje ona cechę odwracalnej parametryzacji sygnału (analiza/synteza). W analizie wykorzystuje się współczynniki MELCEP, *(melowe współczynniki cepstrum)* natomiast w resyntezie używa się algorytmu aproksymującego spektrum za pomocą zlogarytmizowanych współczynników cepstralnych w skali melowej - MLSA. Wymienione współczynniki są poddawane algorytmowi analizy skupień przy zastosowaniu drzew CART. Tworzone są w ten sposób klastry jednostek akustycznie podobnych przy wykorzystaniu informacji lingwistycznych takich jak kontekst fonetyczny, cechy prozodyczne, informacja o F0 oraz dodatkowe informacje o akcencie.

Proces klastrowania jest dodatkowo optymalizowany poprzez minimalizowanie sumy odchylenia standardowego dla każdej z cech dla MCEP, pomnożonych przez ilość próbek w klastrze (Black 2006 B).

Dla docelowego zdania które ma być zsyntezowane wyszukuje się

odpowiednie klastry, co stanowi odpowiedniki kosztu doboru jednostki w syntezie Multisyn. Następnie wyliczany jest koszt konkatenacji dla poprzedniej i następnej głoski. W etapie końcowym algorytm Viterbiego wyszukuje najlepszą sekwencję. W praktyce sekwencja głosek jest generowana, a następnie algorytm MLSA konwertuje wektor z cechami spektralnymi na reprezentację LPC. (Kominek i wsp. 2006)

Zaletą syntezy parametrycznej w porównaniu do syntezy typu Multisyn jest możliwość zastosowania mniejszej oraz mniej dokładnie posegmentowanej bazy danych. Zdecydowanie jakość mowy, uzyskana za pomocą algorytmu Multisyn pozostaje na wyższym poziomie niż w syntezie parametrycznej. Należy dodać, że proces resyntezy w syntezie parametrycznej wymaga zastosowania wokodera, co niestety może powodować dość nienaturalne brzmienie. W syntezie powinna być zastosowana również kontrola prozodii. Rozmiar syntezatora typu Clustergen jest zdecydowanie mniejszy wynosi około 2 MB (200 MB dla Multisyn (Black i wsp. 2006 A)).

Kolejnym algorytmem syntezy jest synteza typu Clunit. W syntezie tej sygnał mowy jest przetwarzany przez 12 cepstralnych wektorów, które są zorganizowane w bardzo wiele klastrów (każdy klaster posiada około 20-40 segmentów). Sygnał ten jest poddawany procesowi treningu przy zastosowaniu drzew CART. W wyniku tego procesu powstaje drzewo decyzyjne, które jest używane w procesie syntezy. Synteza polega na predykcji klastra identyfikującego każdy fonem ze zdania, które ma być zsyntezowane. Następnie używa się algorytmu Viterbiego w celu znalezienia jednostek, które minimalizują koszt konkatenacji. Do łączenia używa się wygładzania między-ramkowego. (Kominek i wsp. 2006). Zatem funkcja kosztu doboru została zastą-piona przez klasyfikator drzew regresyjnych. Podstawową jednostką w syntezie jest głoska, w syntezie Multisyn wykorzystuje się difony.

Wadą stosowania głosek jest problem ich konkatenacji ze sobą (brak kontekstu). Jednak zdecydowanie łatwiej opisać cechy dystynktywne fonemów niż difonów, w przypadku których potrzebna jest dwukrotnie większa ilość cech. Dodatkowo powstaje problem niejednoznaczności ich określania np. akcentowanie difonu może być częściowe, gdzie w przypadku fonemu, określono akcent wartością tak/nie. Dlatego synteza Clunit zawiera dużą ilość

technik optymalizacyjnych uwzględniających poprzedni i następny fonem. Mimo wszystko często nie daje się uniknąć złych połączeń (Clark i wsp. 2007). W przeciwieństwie do algorytmu Multisyn, Clunit używa cech lingwistycznych, żeby przewidzieć cechy akustyczne następnych segmentów. Algorytm Multisyn oznacza jednostki jako złe lub dobre na podstawie ich połączenia w kontekście lingwistycznym. W syntezie Clunit używany jest tylko pojedynczy wektor wartości do opisania wielu właściwości akustycznych głoski. Stanowi to główny problemem tej syntezy. W porównaniu do syntezy Clustergen różnica polega na procesie ekstrakcji cech, które powstają dla każdego segmentu, co znacznie zmniejsza ich ilość (Clunit) a w Clustergen dla każdego wektora cech (Black 2006B, Kominek i wsp. 2006).

2.5.2 Algorytm Multisyn

Najnowszym algorytmem syntezy korpusowej zaimplementowanym w Festivalu oraz pozwalającym na uzyskanie najbardziej naturalnej mowy jest algorytm Multisyn (Clark i wsp. 2007). Developerzy metasystemu przygotowali również narzędzia pomocne do budowania nowego głosu. Dla języka angielskiego budowa nowego głosu jest znacznie ułatwiona. Istnieją moduły lingwistyczne takie jak: POS, GTP oraz wiele narzędzi ułatwiających pracę z sygnałem np. wbudowany aligner (Rozdział 4.3.1). Istnieją narzędzia umożliwiające poprawną ekstrakcję pitchmarków *(maksimum wychylenia amplitudy dla poszczególnych okresów krtaniowych)*. (Rozdział 4.4)

Algorytm Multisyn według (Clark i wsp. 2005) został zaprojektowany w taki sposób by można było w łatwy sposób dodawać nowe głosy. Twórcy dążyli do stworzenia systemu posiadającego większą funkcjonalność niż tylko eksperymentalnego. Niestety, w rzeczywistości obsługa Festivala oraz tworzenie nowych modułów jest trudne. Deweloperzy Festivala przyznają, iż wciąż ma on wiele błędów, a korzystanie z niego jest utrudnione. Brakuje dobrej dokumentacji oraz wskazówek co do tworzenia głosów. Z drugiej strony jest to praktycznie jedyne środowisko posiadające tak duże możliwości. Wiele komercyjnych systemów TTS swój początek zawdzięcza Festivalowi, np. syntezatory firm Nuance, AT&T i Cepstral.

W przeciwieństwie do innych algorytmów typu korpusowego synteza z

zastosowaniem algorytmu Multisyn jest zdefiniowana jako program wyszukiwania poprawnych jednostek akustycznych, czyli w tym przypadku difonów. Celem algorytmu Multisyn jest znalezienie takich jednostek, które przy konkatenacji będą brzmiały jak najlepiej i w przeciwieństwie do syntezy difonowej nie będą wymagały dalszego przetwarzania sygnału.

W tym celu tworzy się zdanie za pomocą struktur językowych będące wynikiem działania modułów przetwarzania języka naturalnego. Jako wynik działania powstaje sekwencja jednostek uwzględniających akcent, części mowy oraz preferowane wartości F0. Następnie wyszukuje się odpowiednie jednostki w bazie akustycznej.

Najlepiej pasujące jednostki są oszacowane na podstawie dwóch kosztów – kosztu doboru jednostki oraz kosztu konkatenacji. Koszt doboru jednostki wyszukuje te segmenty, które będą najbardziej spełniały cechy lingwistyczne zdania docelowego.

Koszt konkatenacji pozwala ocenić jak bardzo pasują do siebie dwie jednostki, jeśli nie występują one bezpośrednio po sobie w bazie akustycznej.

Funkcje te działają na zasadzie zwiększania kosztu odpowiednich jednostek: im koszt większy, tym mniejsza szansa na wybór jednostki. Im wartości bliższe zeru tym jednostka bliższa naturalnej (Clark i wsp. 2005). Zatem z bazy wybierana jest taka sekwencja, która minimalizuje funkcję kosztu doboru jednostki oraz kosztu konkatenacji. Jednostka jest wybierana poprzez wyszukiwanie w algorytmie Viterbiego.

Dużą przewagą syntezy korpusowej nad syntezą konkatenacyjną w postaci difonowej jest mniejsza ilość parametrów potrzebnych do zsyntezowania zdania, w głównej mierze wykorzystuje sie naturalne brzmienie długość jak i prozodię zarejestrowana przez mówcę.

Dlatego w algorytmie Multisyn nie zmienia się czasu trwania poszczególnych jednostek akustycznych oraz ich tonu podstawowego. Funkcja kosztu doboru jednostki pozwoliła na wyeliminowanie modelu prozodycznego z syntezy. Posiadając dobrze zdefiniowaną funkcję doboru jednostki oraz dużą bazę zawierającą wiele różnych kontekstów można uzyskać naturalne brzmienie w syntetycznej mowie.

Podstawową jednostką w algorytmie Mulitsyn jest difon. Uzyskanie

informacji o difonach na podstawie transkrypcji jest dość łatwe. Festival dzieli każdy fonem na połowę i odpowiednio modyfikuje czasy trwania granic difonów. Zastąpienie fonemu dłuższą jednostką pozwala na zmniejszenie czasu wyszukiwania w bazie.

Należy dodać, że ograniczanie do jednego rodzaju jednostki nie wyklucza użycia pozostałych jednostek akustycznych. Nie mniej jednak z uwagi na rosnący koszt obliczeniowy nie jest zalecane stosowanie wielu jednostek jednocześnie. Rozwiązanie to zaproponowano w (Clark i wsp. 2007).

2.5.3 Tworzenie struktury zdaniowej *(utterance)* w systemie Festival

W celu uzyskania syntetycznego zdania konstruuje się strukturę, w której sekwencja fonemów jest zamieniana na difony. Następnie dodawane są informacje o frazach i częściach mowy. Tak przygotowana struktura jest poddana algorytmowi wyszukiwania. Celem jest znalezienie kandydatów o jak najmniejszych wartościach funkcji kosztu.

Jeśli baza zawiera wszystkie difony w licznych kontekstach, wtedy istnieje duża szansa na znalezienie zawsze poprawnych jednostek. W praktyce jednak sytuacja jest niemożliwa do uzyskania, ponieważ korpus rozrósł by się do bardzo dużych rozmiarów oraz mógłby przestać być reprezentatywnym zbiorem danego języka.

Zatem podczas syntezy powstaną sytuacje, że będzie brakować pewnych jednostek akustycznych. W takich sytuacjach wykorzystuje się moduł Back-off. Moduł ten zawiera listę reguł zastępujących określone sekwencje fonemów innymi dostępnymi w bazie. Należy jednak zwrócić uwagę, na sposób zamiany sekwencji, np.: w wyrazie /niedźwiedź/ - /n’ e dz’ v j e ts’/. Jeśli brakuje difonu /dz’-v/ i zastąpiony zostanie difonem /dz’-f/ , to może się pojawić problem w miejscu konkatenacji difonu /f-j/. Niestety, Festival nie pozwala na uwzględnianie kontekstu zastąpionego difonu.

Ponieważ ilość jednostek które muszą być przeszukane jest bardzo duża, optymalizuje się proces ich wyszukiwania poprzez zawężanie przestrzeni wyszukiwań. Do tego służy preselekcja. Jej głównym celem jest ograniczenie

ilości przeszukiwanych segmentów, które pasują do zdania syntezowanego. Rezultatem działania modułów lingwistycznych jest wynik w postaci odpowiednio dobranych fonemów, które zostają zamienione na listę difonowych kandydatów. Jeśli w funkcji kosztu najważniejszym elementem będzie akcent, zastosowanie preselekcji pozwoli po znalezieniu optymalnego segmentu zaprzestania przeszukiwania pozostałego zbioru.

3 Realizacje funkcji kosztu w wybranych systemach syntezy mowy

3.1 Koszt doboru jednostki

Funkcja kosztu doboru jednostki składa się z nienumerycznych wartości (np. lewy prawy i kontekst, pozycja w słowie, pozycja we frazie, POS, natomiast koszt konkatenacji bazuje na parametrach numerycznych takich jak: energia sygnału, F0, nieciągłość spektralna na granicach łączonych ze sobą fragmentów. Najprostszą funkcję kosztu można zdefiniować w sposób binarny. Jeśli jednostka proponowana do konkatenacji ma taką samą wartość jak jednostka poszukiwana, wtedy koszt wynosi 0, w przeciwnym wypadku koszt wynosi 1. Jeśli poszukiwana jest jednostka znajdująca się w pozycji akcentu zdaniowego, a do wyboru jest jednostka nie posiadająca takiego akcentu, ale z akcentem wyrazowym, to jednostka ta otrzyma domyślnie wartość 1 (Clark i wsp. 2007).

Należy podkreślić, iż konstrukcja funkcji doboru jednostki jest mocno skorelowana z charakterystyką danego języka. Projektowanie modułów tej funkcji w większości przypadków sprowadza się do stworzenia modułów opisujących dany język (Rozdział 4). Bardzo istotny jest tu sposób ustalania współczynników wag, który może odbywać się w drodze automatycznej lub manualnej na podstawie wyniku testów percepcyjnych. (Hunt i wsp. 1996). W pracy (Black i wsp. 1996) proponowana metoda polega wybraniu jednego ze zdań z bazy, a następnie jego usunięciu. W kolejnym kroku syntezuje się za pomocą pozostałych jednostek usunięte zdanie. Ramki sygnału ze współczynnikami cepstrum są porównywane z ramkami zdania usuniętego z bazy. Następnie wylicza się w przestrzeni cepstralnej odległości między ramkami w tych zdaniach w oparciu o odległość euklidesowa pomiędzy cepstralnymi wektorami. Metoda ta, choć bardzo złożona obliczeniowo zapewnia lepszą jakość syntezy niż synteza, w której parametry były oszacowywane manualnie (Black i wsp. 1996).

Należy jednak zaznaczyć, że konstrukcja funkcji kosztu oparta na stałych parametrach kosztu doboru jednostki może jednak powodować pominięcie istotnych z punktu widzenia lingwistycznego elementów językowych. Dlatego firmy zajmujące się tworzeniem takich systemów (Coorman i wsp. 2000) tworzą dodatkowe reguły które są skorelowane z funkcją kosztu. Są to pewnego rodzaju „sztuczki", które w istotnym stopniu poprawiają jakość syntezy, choć nie są bezpośrednio związane z funkcją kosztu. Na przykład, w języku polskim w końcowej części zdania ostatnia sylaba jest na ogół wydłużona względem pozostałych. Zatem podczas wyszukiwania jednostek do syntezy ostatniej sylaby w zdaniu powinno się dodatkowo uwzględnić przy konkatenacji, z jakiej części zdania, a nie wyrazu, pochodzi wybierana sylaba. Wagi funkcji kosztu mogą być zmienne w czasie w zależności od dodatkowych reguł związanych z podstawowymi cechami lingwistycznymi przetwarzanego zdania. Mogą być zwiększane lub zmniejszane, w przypadku, gdy kontekst wpływa w istotny sposób na postać dźwiękową jednej z głosek w poszukiwanej sylabie. Na przykład, postać dźwiękowa głoski /r/ silnie zależy od otaczających ją głosek i miejsca w frazie. W skutek tego dochodzi do modyfikacji jej struktury akustycznej. Podczas wyszukiwania jednostek w sąsiedztwie /r/ ważne będzie więc uwzględnienie dodatkowej cechy, która zwiększy wagę kontekstu oraz położenia względem sąsiednich głosek. Efektywnym rozwiązaniem w zaawansowanych systemach jest zastosowanie reguł opartych na logice rozmytej (IVO 2005). W tym celu tworzy się reguły określające sposób dopasowania danej jednostki (dobrze, źle, bardzo źle). W praktyce przypisuje się słowu „bardzo źle" wartość 1, „źle" 0,75 bardzo dobrze wartość 0. Wymieniona wyżej reguła może mieć również swoje odzwierciedlenie w funkcji kosztu konkatenacji, gdzie priorytet będzie dany jednostkom o odpowiednim kontekście i dopiero przekazany zgodnie z nim następuje proces wyszukiwania jednostek minimalizujących wagę, energię, F0 oraz nieciągłości sygnału (IVO 2005).

3.2 Koszt konkatenacji

Koszt konkatenacji określa stopień dopasowania dwóch jednostek ze

sobą. Idealna funkcja kosztu wskaże takie jednostki z bazy, dla których spektralne nieciągłości będą na tyle znikome, że uzyska się płynną i naturalną mowę syntetyczną. Wartość funkcji kosztu będzie równa zeru. Funkcja kosztu konkatenacji zawiera jeszcze moduł wyliczania kosztu F0 oraz energii sygnału. Koszt sumaryczny jest kompromisem pomiędzy kosztem konkatenacji a kosztem doboru jednostki. Sposób wyszukiwania optymalnych jednostek odbywa się za pomocą algorytmu Viterbiego (Wzór 5) (Viterbi 1967), zazwyczaj w nieco zmodyfikowanej postaci. Podstawą wyszukiwania jest struktura danych *„trellis"* wszystkich kandydatów, utworzona przez ścieżki między nimi. Algorytm Viterbiego przeszukuje od lewej do prawej strony kratę wyliczając koszty częściowe, co stanowi sumę sekwencji kosztu doboru jednostki oraz sekwencji kosztu konkatenacji. Następnie zapamiętywana jest optymalna ścieżka o najmniejszym dotychczasowym koszcie. Gdy wyszukiwanie jest zakończone wybierana jest finalna ścieżka, o najmniejszym koszcie (Rozdział 2.2.4).

$$v_j(t) = MAX_i[v_i(t-1)a_{ij}b_j(y_t)] \quad (5)$$

W pracy (Vepa 2004) porównano 3 funkcje kosztu wyznaczone w oparciu o 3 typy odległości. W pierwszej użyto odległości mahalanobisa wyliczonej w przestrzeni częstotliwości widma liniowego(LSF). W drugiej funkcji kosztu uwzględniono odległość mahalanobisa w analizie centroidów (MCA) dla 7 ramek sygnału, 3 ramki sygnału z każdej strony plus jedna w miejscu łączenia. Ostatnia bazuje na filtrach Kalmana i współczynnikach LSF. Z cytowanych badań nie wynika jednoznacznie, który sposób wyliczania nieciągłości spektralnych jest najlepszy. We wspomnianej pracy wybrano funkcje LSF oraz MCA. Trudno jest wskazać optymalną funkcję kosztu. Być może na wyniki badań wpłynął stosunkowo ograniczony zbiór testowy – małe zbiory samogłosek w izolowanych wyrazach. Nie można też wykluczyć, że koszt konkatenacji jest podobnie, jak koszt doboru jednostki, zależny od języka, mówcy, czy akcentu.

W pracy (Bjørkan i wsp. 2005) porównano 5 różnych odległości spektralnych oraz różnic w przebiegu konturu F0 na podstawie dwóch samogłosek (/a:/ oraz /e:/) języka norweskiego. Samogłoska była umieszczona w testowym słowie, to zaś znajdowało się w krótkim zdaniu. Z

przeprowadzonych badań dla symetrycznej odległości Kullbacka-Leiblera *SKL - Symmetrical Kullback-Leibler distance)* na współczynnikach LPC, euklidesowej odległości na cepstralnych współczynnikach uzyskanych z LPC *(CEP)*, średniej opartej na niesymetrycznym podobieństwie współczynników LPC *(LR – likelihood ratio)*, euklidesowej odległości pomiędzy 13 Mel-cepstralnymi wektorami *(MFCC)* oraz zsynchronizowanych ze sobą współczynników F0 *(MPSC)* wynika, że współczynniki *MFCC, LR, CEP* były najlepszymi parametrami, podczas gdy *SKL* był zdecydowanie gorszy a *MPSC* najgorszy. Nie zaobserwowano dużych różnic w jakości syntezowanej mowy podczas stosowania najlepszych współczynników.

Z dotychczasowych badań nie wynika jasno, który sposób wyliczenia nieciągłości widmowych jest najlepszy. (Klabbers i wsp. 1998, 2001) zbadali wpływ rożnych odległości na pięciu samogłoskach dla języka holenderskiego. Z przeprowadzonego eksperymentu wynika, że symetryczna odległość Kullbacka-Leiblera na znormalizowanych współczynnikach LPC jest najlepszym wskaźnikiem spośród sześciu (Klabbers i wsp. 1998, 2001):

- odległości euklidesowej pomiędzy formantami F1 i F2 (D_{FED})
- odległości euklidesowej pomiędzy melowymi współczynnikami cepstrum (D_{MFCC})
- średnią niesymetryczną współczynników podobieństwa wyliczonych na podstawie energii znormalizowanych widmowych współczynników LPC (D_{LR}) (Gray i wsp. 1976)
- symetrycznej odległości Kullbacka-Leiblera (D_{SKL}),używanego w statystyce
- częściowej głośności (D_{PL}),
- odległości między średnio-kwadratowymi widmami logarytmicznymi (D_{MSLSD})

(Wouters i wsp. 1998) udowodnili, że odległość euklidesowa oparta na skali melowej cepstralnych współczynników LPC jest dobrym parametrem w oszacowaniu nieciągłości spektralnych. Zwrócili oni również uwagę na fakt, iż nieliniowa skala częstotliwości jest w miarę zgodna ze skalą percepcji wysokości dźwięków (np. melowa lub w barkach) oraz lepiej odwzorowuje słuchową percepcję nieciągłości spektralnych niż skala liniowa *(Prawo*

Webera-Fechnera) (Stevens 1998). Percepcyjnie przyrosty wysokości (subiektywne) są proporcjonalne do przyrostów częstotliwości wyrażonych w skali logarytmicznej.

Podsumowując wyniki powyższych badań trudno o wybór najlepszej odległości. Jednak cytowane wyniki badań (Vepa 2004, Vepa i wsp. 2006, Bjørkan i wsp. 2005) potwierdzają, że zaimplementowana funkcja kosztu konkatenacji w systemie Festival oparta na odległości Mahalanobisa wyznaczanej na współczynnikach MFCC jest dobrym estymatorem w ocenie efektywności funkcji kosztu konkatenacji. Uzyskane w pracy wyniki potwierdzają korzyści ze stosowania tej odległości również dla języka polskiego skali nieliniowej (Rozdział 5).

(Klabbers i wsp. 2004, Vepa 2004 Wouters i wsp. 1998, Bjørkan i wsp. 2005) przeprowadzili badania dotyczące wpływu różnych miar odległości akustycznych oraz cech sygnału akustycznego i ich percepcji na jakość syntetyzowanej mowy. Z przeprowadzonych badań wynika, że najlepsza korelacja pomiędzy kosztem akustycznym oraz percepcją ekspertów lingwistycznych nie przekracza 0,66 co jest niezadowalającym wynikiem z naukowego punktu widzenia. Można jednak uznać, iż ten obszar ten został zbadany dokładnie. Opublikowane wyniki prezentują tabele 3.1 i 3.2.

Odległość/wsp.	Euklidesowa	Bezwzględna	Mahalanobisa
MFCC	0,6	0,64	0,66
MFCC + Δ	0,55	0,55	0,50
LSF	0,63	0,64	0,64
LSF + Δ	0,63	0,64	0,58
Formant	0,59	0,58	0,55
Formant + Δ	0,46	0,46	0,62

Tabela 3.1 Prezentuje korelację perceptualnego dopasowania poszczególnych segmentów na podstawie różnych odległości akustycznych oraz parametryzacji sygnału. (na podstawie Klabbers i wsp. 2004, Vepa 2004 Wouters i wsp. 1998, Bjørkan i wsp. 2005)

Do wyliczenia korelacji zbudowano bazę składająca się ze słowa wzorcowego oraz zmodyfikowanej wersji słowa. Słowo referencyjne było zsyntezowane przez syntezator difonowy, słowo zmodyfikowane posiadało zmieniony difon lub pół difonu pochodzącego z innego kontekstu fonetycznego. Każde słowo składało się z jednej sylaby. 15 ekspertów

otrzymało 25 par słów do oceny, każdy z nich oceniał te same słowa jednak podane w innej kolejności. Odległość perceptualna została zdefiniowana jako średnia ocen odpowiedzi ekspertów dla każdej pary słów w skali 5 stopniowej od 0 do 4. W celu weryfikacji testu, przestudiowano 38 par słów, w których żaden z segmentów nie został zmieniony. Większość odpowiedzi była 0 (nie ma różnicy) lub 1, 1,5 % odpowiedzi była oznaczona jako 2 i jedna odpowiedź 3. Oceniano 3 kategorie słów zawierające centralnie położone segmenty /aa/ /ae/ /iy/ /uw/.

Test obiektywny polegał na wyliczeniu miary odległości między oryginalnym a zamienionym segmentem. (Liczona jest widmowa nieciągłość między nowo zastąpionym segmentem a słowem wzorcowym, która ma swój koszt). Celem badania było znalezienie miary, która potrafi przewidzieć za pomocą odległości obiektywnych zamiany percepcyjne. W wyniku wyliczono korelację między obiektywnymi odległościami oraz percepcyjnymi. (Tabela 3.2) (Wouters i wsp. 1998).

Skala	**Liniowa**		**Melowa**	
Odległość/wsp.	Euk	Mah	Euk	Mah
CEP	0,48	0,53	0,64	0,64
LSF	0,34	0,5	0,58	0,57

Tabela 3.2 Prezentuje korelację perceptualnego dopasowania poszczególnych segmentów na podstawie skali liniowej oraz nieliniowej z uwzględnieniem dwóch odległości: euklidesowej oraz mahalanobisa (Wouters i wsp. 1998).

Z powyższych badań wynika również, że nie można jednoznacznie określić uniwersalnej, optymalnej funkcji kosztu. Może się ona bowiem różnić w zależności od języka, a w przypadku tego samego języka będzie zależna od mówcy (kobieta, mężczyzna) jak również może być różna między dwoma mówcami tego samego języka oraz tej samej płci. Dodatkowym problemem jest interpretacja badań odsłuchowych, które mogą być określane w bardzo ogólny sposób i przez to prowadzić do pewnych nieporozumień lub też niejednoznaczności w interpretacji wyników. Jeśli istnieje potrzeba określenia podczas testów odsłuchowych, czy dany segment w mowie jest akcentowany, czy też nie, a różnica percepcyjna jest niewielka, to dwóch różnych lingwistów może określić segment niejednoznacznie, co może doprowadzić w

konsekwencji do błędów i złej konkatenacji w mowie syntetycznej.

3.3 Funkcja kosztu w systemie syntezy Festival

W meta systemie Festival wynik funkcji kosztu jest wyliczany na podstawie średniej ważonej kosztu konkatenacji funkcji oraz kosztu doboru jednostki.

Koszt konkatenacji definiuje następujące parametry:

- koszt F0,
- koszt energii sygnału,
- koszt nieciągłości spektralnej

Koszt doboru jednostki składa się z następujących składowych:

- koszt akcentu,
- koszt lewego i prawego kontekstu,
- koszt niewłaściwego doboru melodii,
- koszt pozycji w sylabie,
- koszt pozycji w słowie,
- koszt pozycji we frazie,
- koszt Part-Of-Speech

Funkcja ta jest wagowo znormalizowaną sumą wymienionych komponentów. Algorytm Multisyn wykorzystuje jedynie metodę nakładania sygnału OverLap-Add na granicach poszczególnych jednostek. W algorytmie Multisyn nie zachodzi modyfikacja spektralna, amplitudowa, czy też interpolacja F0 w konkatenowanych granicach (Clark i wsp. 2005).

Zoptymalizowana przez autorów systemu Festival funkcja kosztu jest zdefiniowana z następującymi wagami:

- koszt pozycji we frazie, ustawiony domyślnie na wartość 15
- koszt akcentu ustawiony domyślnie na wartość 10
- koszt niewłaściwego doboru melodii 25
- koszt POS ustawiony domyślnie na wartość 6. Moduł POS zawiera 5 dostępnych kategorii: rzeczownik, czasownik, modyfikatory, funkcja słowa, pozostałe

- koszt pozycji sylaby w słowie, domyślnie ustalone kategorie to początkowa, środkowa, końcowa oraz pomiędzy słowami, wartość ustawiona domyślnie na 5
- pozycja słowa ustawiona domyślnie na wartość 5. Dostępne są trzy pozycje: początkowa, środkowa oraz końcowa.
- lewy kontekst ustawiony domyślnie na wartość 4
- prawy kontekst ustawiony domyślnie na wartość 3

Wartości powyższe zostały dobrane heurystycznie przez twórców systemu Festival. (Clark i wsp. 2007), jednak wartości te okazały się nieoptymalne dla języka polskiego (Rozdział 6).

W innych systemach korpusowych syntezatorów dobór parametrów funkcji kosztu szacuje się metodami heurystycznymi lub metodami samouczącymi się. (Hunt i wsp. 1996, Clark i wsp. 2007, Hamdi i wsp. 2006). Ta ostatnia nie została jednak zaimplementowana w Festivalu, ponieważ według (Clark i wsp. 2007) nie przynosiła ona znaczącej poprawy. Według badań (Hunt i wsp. 1996) dotyczących percepcji ludzkiego słuchu wynika, że najistotniejsze jest zminimalizowanie ilości słyszalnych zniekształceń do minimum. Można to uzyskać przez minimalizowanie wartości funkcji doboru jednostki oraz kosztu konkatenacji. W pracy (Clark i wsp. 2007) wskazano, iż w algorytmie Multisyn brakuje modułu, który radził by sobie w sytuacjach, w których pożądana jest inna intonacja niż neutralna, co w wynika z trudności konstrukcji predyktora akcentu. (Clark i wsp. 2007) Brakuje również implementacji doboru ostatniego difonu na końcu frazy. Jeśli w tym przypadku wybrany segment zostanie wybrany ze środka wyrazu, syntezowana mowa będzie brzmieć nienaturalnie.

Pewnym problemem jest ograniczona możliwość modyfikacji prozodii. W algorytmie Multisyn pozostaje ona na poziomie symbolicznym. Z założenia nie modyfikuje się ani czasu trwania głoski ani F0. (Clark i wsp. 2006) Istnieje jednak możliwość dopisania własnych modułów i stworzenia hybrydy algorytmu Multisyn.

Funkcja kosztu konkatenacji używa trzech ważonych komponentów: spektrum, F0 i logarytmu energii. Spektrum i logarytm energii są wyliczane na podstawie plików MFCC. Kontur F0 ekstrahowany jest w wyniku działania algorytmu ESPS zaimplementowanego w środowisku Festival (Richmond i wsp.

2007). Nieciągłości spektralne oszacowywane są przy użyciu odległości Mahalanobisa pomiędzy dwoma wektorami zawierającymi 12 melowych współczynników cepstralnych (znormalizowanych na podstawie średnich i wariancji) z każdej strony miejsca łączenia. Dodatkowe dwa współczynniki (F0 i energia) używane są do estymacji zniekształceń energii oraz F0. (Clark i wsp. 2005). Algorytm kosztu konkatenacji działa następująco: łącząc ze sobą sąsiadujące segmenty o widmie harmonicznym pod uwagę brana jest wielkość różnicy energii, F0 oraz nieciągłości spektralnych w punkcie łączenia. Natomiast koszt łączenia segmentów szumowych jest automatycznie ustalany na zero, natomiast koszt łączenia segmentu dźwięcznego z sygnałem szumowym karany jest maksymalnie. Wszystkie te składowe kosztu są normalizowane do przedziału <0, 1>

4 Przygotowanie akustycznej bazy danych dla korpusowej syntezy mowy języka polskiego

We wprowadzeniu do pracy opisano wiele problemów, które twórca systemu korpusowej syntezy mowy musi rozwiązywać. Realizacja takiego systemu jest zadaniem trudnym i wymaga zaprojektowania wielu modułów, przeprowadzenia szeregu badań dla danego języka. W tej pracy stworzenie pełnego systemu syntezy mowy sprowadziło się do realizacji następujących etapów:

- stworzenia korpusu
- nagrania korpusu
- automatycznej i manualnej korekty segmentacji korpusu
- opisu nowego głosu w Festivalu polegającego na wybraniu metody syntezy korpusowej. Festival oferuje kilka technologii (Rozdział 2.5.1).
- stworzenia modułów lingwistycznych dla języka polskiego, ewentualnie ich modyfikacji na podstawie istniejących już dla innych języków
- zbudowania struktur zdaniowych opisujących lingwistyczne zależności w nagranym korpusie
- ekstrakcji pitchmarków (Rozdział 4.4)
- ekstrakcji konturów F0 z bazy akustycznej
- parametryzacji bazy akustycznej :
    - MFCC
    - LPC
- projektowania funkcji kosztu – kosztu konkatenacji oraz kosztu doboru jednostki za pomocą algorytmu ewolucyjnego
- testowania i korekty stworzonego głosu syntetycznego

W tym rozdziale został przedstawiony sposób realizacji korpusu a

następnie jego nagrania. Opisana została metoda automatycznej segmentacji bazy akustycznej oraz zautomatyzowany sposób korekty błędów. Przedstawiono metodę testowania bazy akustycznej oraz wnioski związane z pierwszą wersją stworzonego głosu w środowisku syntezy korpusowej Festival.

4.1 Przygotowanie korpusu

Naturalność mowy w korpusowej syntezie zależy od bardzo wielu czynników. Jednym z ważniejszych etapów podczas przygotowania nowego głosu syntetycznego jest przygotowanie korpusu tekstowego. Następną istotną kwestią jest właściwy dobór mówcy czytającego przygotowane teksty. Ostatnim etapem realizacji bazy akustycznej jest realizacja dźwiękowej bazy korpusowej oraz odpowiednia segmentacja i etykietyzacja odcinków nagranych sygnałów. Autor uważa, że od właściwej realizacji wymienionych etapów zależy w głównej mierze jakość generowanej mowy syntetycznej. Dlatego należy poświęcić odpowiednią ilość czasu na zbadanie zagadnienia tworzenia korpusu dla wybranego języka syntezy, selekcji tekstów oraz problemu, tzw. zrównoważenia korpusu. Proces balansowania polega na wyodrębnieniu z bardzo dużego zbioru tekstowego pewnej liczby zdań spełniających w największym stopniu zadane wejściowe kryteria (Rozdział 4.1.5). Im zbiór tekstowy jest liczniejszy tym większa szansa na spełnienie założonych kryteriów.

Dobrze zbalansowany korpus będzie stanowił reprezentacyjną bazę językową. Po przekształceniu go do postaci dźwiękowej, będzie tworzył odpowiednią bazę do zaprojektowania funkcji kosztu doboru jednostki oraz kosztu konkatenacji dla syntezy korpusowej.

Ostatnio zaczyna podkreślać się wagę wpływu przygotowania oraz realizacji dźwiękowej bazy akustycznej na jakość uzyskiwanej mowy syntetycznej (Clark i wsp. 2007, Kominek i wsp. 2003, Ellbogen i wsp. 2004, Bozkurt i wsp. 1997).

Ręczne projektowanie korpusu praktycznie stosuje się tylko w syntezie typu *Limited Domain Speech Synthesis*. Synteza ta polega na łączeniu ze sobą słów, całych fraz lub nawet zdań z niewielkiego korpusu (około kilkadziesiąt

wypowiedzi) w celu wygenerowania zdania z określonej dziedziny np.: informacja o odjazdach pociągów, autobusów, zegarynka itp. Udowodniono, że można uzyskać bardzo wysoką jakość syntezy mowy poprzez stworzenie korpusu złożonego ze zdań, które docelowo będą syntezowane (Black i wsp. 2000). Jednak projektując system korpusowej syntezy mowy oczekuje się od niego uniwersalności. Nie jest to możliwe, ponieważ oznaczałoby to stworzenie korpusu zawierającego każde słowo danego języka. Próbuje się ograniczać przestrzeń poszukiwań do zdań będących fonetycznie zrównoważonymi, to znaczy, że częstotliwość występowania fonemów w bazie jest zbliżona do częstotliwości ich występowania w mowie naturalnej, danego języka i do dziedziny, z której będą syntezowane zdania. (Black i wsp. 2000)

Obecnie istnieje wiele narzędzi ułatwiających projektowanie korpusów. Zazwyczaj narzędzia te bazują na tzw. algorytmie zachłannym - *greedy algorithm*. Działanie tego algorytmu sprowadza się do iteracyjnego wyodrębnienia z bardzo dużego zbioru tekstowego pewnej liczby zdań spełniających w największym stopniu zadane wejściowe kryteria takie jak ilość segmentów składających się na długość zdania, ilość segmentów w korpusie. Im zbiór tekstowy jest liczniejszy tym większa szansa na spełnienie założonych kryteriów.

Głównym celem pierwszego etapu pracy nad korpusową syntezą mowy było przygotowanie reprezentatywnego dla języka polskiego korpusu w oparciu o podstawowe jednostki akustyczne głoski, difony i trifony.

4.1.1 Wykorzystane zbiory tekstowe

Do realizacji korpusu zostały użyte materiały z wystąpień sejmowych posłów oraz teksty zawierające prasowe recenzje filmowe. Udowodniono, (Kominek i wsp. 2003), iż do realizacji korpusu powinno używać się możliwie jak największego zbioru tekstowego.

Według (Kominek i wsp. 2003) projektowanie korpusu sprowadza się do przeprowadzania następujących etapów:

- wyboru technologii syntezy mowy, w tym przypadku jest to korpusowa synteza mowy

- wyboru docelowej dziedziny
- wyboru źródła tekstów
- automatycznej selekcji bogatych fonetycznie tekstów
- sprawdzeniu i usunięciu z korpusu zbędnych zdań
- nagraniu zdań i ewentualnym usunięciu tych, które nie kwalifikują się z uwagi na trudność wymowy (Kominek i wsp. 2003)

Zdecydowano się na rozwinięcie etapu selekcji. Tworzony korpus został trzykrotnie zbalansowany. Zazwyczaj balansowanie przeprowadza się tylko raz. Udowodniono, (Oliver i wsp. 2006), iż wielokrotne balansowanie w znacznej mierze poprawia kompletność korpusu, w ten sposób uzyskuje się korpus bardziej reprezentatywny dla danego języka. Następnie przeprowadzono jego optymalizację porównując ilość występujących segmentów o różnej długości podczas zmniejszania ilości zdań. W korpusie zostały wprowadzone zdania zawierające terminy obcojęzyczne, dzięki temu uzyskano możliwość syntezy wyrazów obcego pochodzenia. W nagraniach starano się zachować wymowę oryginalną dla danego języka.

W celu realizacji korpusu wykorzystano trzy zbiory tekstów dużych rozmiarów[1]:

- zbiór z przemówieniami sejmowymi (korpus sejmowy)
- zbiór z recenzjami filmów, płyt (korpus z recenzjami)
- zbiór zawierający teksty z gazet (korpus gazetowy)
- zbiór wyrazów z rzadko występującymi fonemami w języku polskim (lista rzadkich wyrazów)

Zbiór z przemówieniami sejmowymi zawiera 5778460 zdań, co odpowiada około 300 MB tekstu. Plik zbioru z recenzjami ma rozmiar 15MB i zawiera 21135 zdań. Dodatkowo korpus uzupełniono listą wyrazów ze stosunkowo rzadko występującymi fonemami w języku polskim[2]. Lista ta była wcześniej użyta w przygotowaniu bazy akustycznej dla projektu europejskiego Speecon. (Marasek i wsp. 2004)

1 Autor użył dodatkowego korpusu zawierającego również teksty z gazet. Korpus z uwagi na mały rozmiar 3692 zdań, został pominięty w finalnym balansowaniu.

2 Autor dr hab. Ryszard Gubrynowicz

Tabela 4.1 przedstawia rozkład względnej częstotliwości występowania poszczególnych fonemów w korpusie sejmowym oraz w korpusie z recenzjami prasowymi. Pomimo dużej różnicy w rozmiarach tych korpusów rozkłady statystyczne występowania fonemów są zbliżone. Na podstawie tej analizy korpus sejmowy został wybrany do dalszej selekcji. Korpus sejmowy powstał z zapisu wypowiedzi w postaci dźwiękowej. W tym zapisie pojawiły się tagi i meta dane, które zostały usunięte. Korpus z recenzjami został pozyskany w automatyczny sposób ze stron internetowych, w których występują dodatkowe znaczniki oznaczające rysunki, oraz dodatkowe adnotacje. Te dane zostały usunięte, a istniejące skróty zostały rozwinięte do pełnych wyrazów. Posłużono się metodą automatyczną w postaci słownika skrótów (Marasek i wsp. 2004).

Fonem	Korpus Sejmowy	Korpus z recenzjami
a	8,88	9,04
b	1,32	1,3
d	2,47	2,27
dZ	0,03	0,02
dz	0,25	0,25
dz'	0,63	0,73
e	9,2	9,68
e~	0,08	0,09
f	1,74	1,92
g	1,45	1,59
i	3,91	4,06
I	4,1	4,07
j	4,24	4,6
k	3,26	3,46
l	2,13	2,42
m	2,99	3,18
N	0,16	0,12
n	4,47	4,58
n'	2,45	1,97
o	8,59	8,11
o~	0,56	0,54
p	3,28	2,78
r	3,5	3,92
S	1,71	1,39
s	3,29	3,45
s'	1,3	1,41
t	4,3	4,57
tS	1,08	1,04
ts	1,4	1,47
ts'	1,07	1,03
u	3,35	3,18
v	3,44	3,02
w	1,69	1,4
x	1,28	1,26
Z	0,99	0,97
z	1,81	1,61
z'	0,17	0,16

Tabela. 4.1 Rozkład względnej częstotliwości występowania poszczególnych fonemów w korpusie sejmowym oraz w korpusie z recenzjami gazetowymi

Kolejnym etapem było przygotowanie transkrypcji fonetycznej dla wybranego korpusu tekstowego.

Przeanalizowano dwie metody zautomatyzowanej transkrypcji:

- regułową – uzyskaną za pomocą modułu transkrypcji fonetycznej w

Festivalu[3] (Black i wsp. 1998)

- metodę automatyczną opartą na drzewach decyzyjnych C5.0 (Marasek 2003 B)

W wyniku porównania obydwu metod okazało się, że lepszym i znacznie efektywniejszym rozwiązaniem jest zastosowanie metody drzew decyzyjnych, dzięki której uzyskuje się znacznie większą dokładność w transkrypcji fonetycznej. Alfabet SAMPA został wybrany do zapisu transkrypcji fonetycznej języka polskiego (Wells 1997). Niestety, decyzja o wykorzystaniu nie-Festivalowego modułu transkrypcji fonetycznej miała istotny wpływ w późniejszym etapie na przebieg tworzenia systemu i dołożyła autorowi dodatkowej pracy przy tworzenia nowego głosu syntezy.

4.1.2 Transkrypcja fonetyczna wypowiedzi języka polskiego

Okazało się, że czas wykonania transkrypcji fonetycznej dla tak dużego zbioru tekstów (5778460 zdań) zajmie on ponad 20 dni. Dlatego korpus został rozbity na kilkanaście podkorpusów i dla każdego z nich została wykonana transkrypcja fonetyczna.[4] Na podstawie wygenerowanej transkrypcji fonematycznej wykonano transkrypcję przy użyciu difonów oraz trifonów, dzięki czemu proces ten przebiegł znacznie szybciej. W celu realizacji wykorzystano skrypt napisany w Perlu. Weryfikacja finalnej wersji korpusu oraz manualna korekta została opisana w rozdziale 4.1.8.

Rysunek 4.1 zawiera przykładowe zdanie w korpusie sejmowym w trzech zapisach: ortograficznym (1a), fonematycznym (1b), difonowym (1c) oraz trifonowym (1d).

---

3 Autorem modułu jest dr Dominika Oliver z Uniwersytetu w Saarbrucken

4 W celu realizacji procesu transkrypcji została zarezerwowane laboratorium komputerowe na okres 2 dni. Łącznie zostało użytych 14 komputerów, każdy z procesorem Intel Pentium IV 2,0Ghz oraz 512 Mb RAM-u, w laboratorium Polsko-Japońskiej Wyższej Szkole technik Komputerowych.

1a. jeśli chodzi o utrzymanie infrastruktury szacuje się potrzeby roczne

1b. #j e s' l i x o dz' i o u t S I m a n' e i n f r a s t r u k t u r I S a ts u j e s' e~ p o t S e b I r o tS n e#

1c. #j je es' s'l li ix xo odz' dz'i io ou ut tS SI Im ma an' n'e ei in nf fr ra as st tr ru uk kt tu ur rI IS Sa ats tsu uj je es' s'e~ e~p po ot tS Se eb bI Ir ro otS tSn ne e#

1d. #je jes' es'l s'li lix ixo xodz' odz'i dz'io iou out utS tSI SIm Ima man' an'e n'ei ein inf nfr fra ras ast str tru ruk ukt ktu tur urI rIS ISa Sats atsu tsuj uje jes' es'e~ s'e~p e~po pot otS tSe Seb ebI bIr Iro rotS otSn tSne ne#

Rys. 4.1 Zapis zdania w korpusie w transkrypcji fonematycznej, difonowej oraz trifonowej. Znak /#/ oznacza ciszę

Kolejnym ważnym etapem było przeprowadzenie analizy rozkładu częstotliwości poszczególnych jednostek akustycznych korpusu. Niestety, okazało się to niemożliwe, ponieważ dopuszczalny maksymalny rozmiar korpusu akceptowany przez program CorpusCrt[5] nie może przekroczyć 20 MB. Problemu tego nie udało się rozwiązać przez odpowiednią modyfikację struktury programu. Ostatecznie korpus został podzielony na 12 podkorpusów o maksymalnym rozmiarze 20 MB i dla każdego z nich przeprowadzono analizy. Każdy podkorpus zawiera więc około 22000 zdań.

4.1.3 Algorytm zachłanny w programie CorpusCrt

Proces równoważenia fonematycznego korpusu polega na wybraniu pewnej ilości zdań spełniających w najbardziej precyzyjny sposób zadane kryteria. Metodę tą stosuje się w celu uzyskania tzw. bogatych fonetycznie zdań, co jest bardzo istotne dla jakości generowanej syntetycznie mowy, a także umożliwia minimalizację rozmiarów korpusu.

Na przykład: korpus wejściowy M wraz z transkrypcją fonematyczną oraz difonową lub trifonową może zostać podzielony na N korpusów zawierających X zdań powtarzających się lub niepowtarzających. Dodatkowo możliwe jest określenie minimalnego progu określającego ilość powtórzeń segmentów występujących w nowym korpusie oraz ilość zdań. Ustalenie tych kryteriów dla realizacji korpusu do syntezy mowy jest szczególnie istotne (Black i wsp.

5 http://gps-tsc.upc.es/veu/personal/sesma/download/linux/corpuscrt.zip

2000).

W celu uzyskania zbalansowanego korpusu zastosowany został program CorpusCrt. Program ten został napisany przez (Bailador 1998) na Politechnice Uniwersytetu Katalonii i jest rozpowszechniany jako freeware.

Wybór zdań do korpusu został dokonany w oparciu o fonemy, difony oraz trifony.[6] Udowodniono, że difony oraz trifony są jednostkami, które mogą być w łatwy sposób konkatenowane ze sobą. Łatwość konkatenacji wynika z tranzjentu występującymi pomiędzy nimi oraz łączenia ich na odcinkach quasistacjonarnych. (Szklanny 2003, Van Santen i wsp. 1997, Beutnagel i wsp. 1997, Black i wsp. 2001, Möbius i wsp. 2001). Niestety, stworzenie korpusu opartego na trifonach staje się praktycznie niemożliwe z uwagi na rozmiar bazy akustycznej oraz wzrost złożoności obliczeniowej w wyszukiwaniu segmentów. (Villaseñor-Pineda i wsp. 2003). Stosuje się pewne uproszczenie. Podczas projektowania korpusu używa się najczęściej występujących trifonów. Udowodniono, że w praktyce wykorzystuje się około 4000 trifonów, które występują więcej niż 100 razy. Takie rozwiązanie zostało zastosowane w (Bozkurt i wsp. 2003).

Stwierdzono również (Clark i wsp. 2005), że korpus o większej liczbie głosek, w przypadku głosu korpusowego Niny zawierającego globalnie 175 000 w stosunku 36 000 jednostek dla głosu Rms2 umożliwia osiągnięcie bardziej naturalnie brzmiącej mowy syntetycznej. W systemie stworzonym przez autora uzyskano liczebność 143000 jednostek, co w praktyce oznacza dłuższy czas syntezy zdania, ale jednocześnie pozwala na osiągnięcie bardziej naturalnej syntezy poprzez zwielokrotnienie ilości głosek występujących w różnych kontekstach.(Clark 2005 i wsp. , Taylor i wsp. 1998)

4.1.4 Pierwsze balansowanie korpusu

Korpus sejmowy został podzielony na 12 podkorpusów po 20 MB każdy a następnie zastosowano program CorpusCrt (Bailador 1998). Każdy podkorpus zawierał około 189000 fonemów. Okazało się, że dla każdego z tych podkorpusów częstotliwości występowania fonemów są bardzo zbliżone

6 Kryteria podano w 4.1.5

do siebie. Tabela 4.2 ilustruje procentowy udział poszczególnych fonemów dla dwóch dowolnie wybranych podkorpusów.

Fonem	I korpus %	II korpus %
a	9,28	9,29
b	1,39	1,37
d	2,31	2,3
dZ	0,02	0,02
dz	0,24	0,25
dz'	0,49	0,49
e	9,27	9,3
e~	0,61	0,63
f	1,45	1,51
g	1,44	1,44
i	3,86	3,84
I	4,08	4,08
j	4,38	4,43
k	3,36	3,42
l	1,99	2,01
m	2,92	3
N	0,23	0,22
n	4,36	4,38
n'	2,8	2,73
o	9,03	8,8
o~	0,57	0,59
p	3,88	3,96
r	3,82	3,78
S	1,58	1,56
s	4,04	4,05
s'	1,11	1,1
t	4,72	4,88
tS	1,06	1,02
ts	1,37	1,45
ts'	1,06	1,1
u	3,42	3,5
v	4,03	3,86
w	1,69	1,68
x	1,14	1,02
Z	1,26	1,24
z	1,86	1,81
z'	0,07	0,07

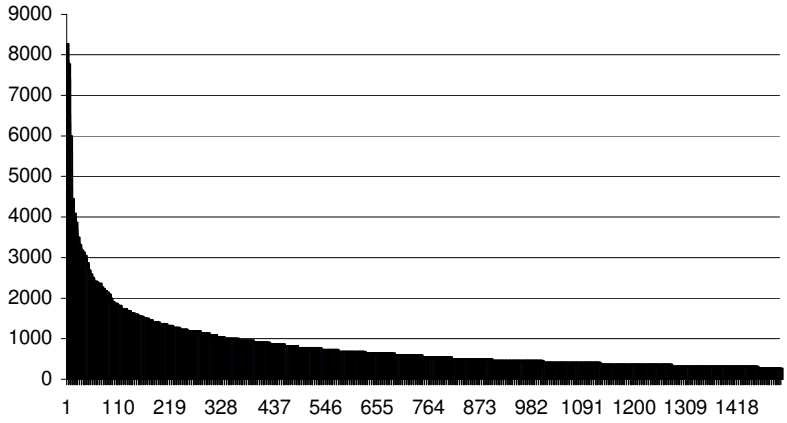
Tabela 4.2 Porównanie rozkładu częstotliwości występowania fonemów w dwóch korpusach sejmowych

Kolejnym eksperymentem było stworzenie podobnych statystyk dla trzech dużych korpusów. Pierwszy korpus zawierający 21135 zdań z gazet, drugi zawierający 3692 zdań, również z gazet oraz trzeci z recenzji - 19733 zdania. Łącznie podkorpusy zawierały 9236725 fonemów. Tabela 4.2 jest ilustracją powyższego zestawienia. Tabela 4.3 przedstawia procentowy udział poszczególnych fonemów dla trzech dowolnie wybranych podkorpusów sejmowych wraz z korpusem z recenzjami gazetowymi.

Fonem	Recenzje	Sejm I	Sejm II	Sejm III
a	9,04	9,8	9,57	9,82
b	1,3	1,43	1,45	1,44
d	2,28	2,06	2,13	2,09
dZ	0,02	0,06	0,04	0,09
dz	0,25	0,28	0,25	0,28
dz'	0,73	0,61	0,61	0,55
e	9,68	9,55	9,55	9,35
e~	0,1	0,87	0,8	0,87
f	1,93	1,45	1,53	1,48
g	1,6	1,17	1,26	1,24
i	4,06	3,54	3,83	3,57
I	4,08	4,04	3,99	3,93
j	4,6	4,25	4,3	4,36
k	3,46	3,47	3,24	3,59
l	2,42	1,89	1,92	1,88
m	3,18	3,28	3,12	3,26
N	0,13	0,21	0,21	0,22
n	4,59	3,82	4,02	3,57
n'	1,97	2,83	2,79	2,8
o	8,12	8,64	8,51	8,88
o~	0,54	0,56	0,64	0,63
p	2,79	4,33	4,08	4,32
r	3,92	3,84	3,61	3,79
S	1,39	1,8	1,59	1,83
s	3,45	4,51	4,37	4,57
s'	1,42	1,08	1,07	1,03
t	4,57	4,73	5,13	4,5
tS	1,04	1	1,12	1,01
ts	1,48	1,26	1,39	1,31
ts'	1,04	1,07	1,13	1,13
u	3,18	3,1	3,36	3,21
v	3,02	3,69	3,79	3,65
w	1,41	2,16	1,91	2,25
x	1,26	0,91	0,93	0,89
Z	0,97	1,2	1,21	1,1
z	1,62	1,61	1,64	1,59
z'	0,17	0,11	0,1	0,11

Tabela 4.3 Porównanie rozkładu częstotliwości występowania fonemów w trzech korpusach sejmowych oraz zestawienie z korpusem z recenzjami gazetowymi.

Do realizacji korpusu tekstowego w ostatecznej postaci został wykorzystany korpus z tekstami przemówień sejmowych oraz połączony korpus z recenzji prasowych oraz innych tekstów z gazet. Pierwsza analiza rozkładu częstotliwościowego wykazuje, że we wszystkich podkorpusach można wyodrębnić blisko 400 najczęściej występujących trifonów, a każdy z nich pojawia się, co najmniej 1000 razy w pierwotnym korpusie. Rysunek 4.2 ilustruje częstotliwościowy rozkład 1500 najczęściej występujących trifonów we wszystkich podkorpusach.



Rys. 4.2 Najczęściej występujące trifony we wszystkich podkorpusach. Oś pionowa oznacza ilość wystąpień, pozioma liczbę trifonów.

4.1.5 Powtórne równoważenie korpusu

12 korpusów utworzonych z głównego zbioru tekstowego zostało zbalansowanych zgodnie z niżej podanymi regułami. Każdy z nich zawiera 2500 zdań. Konieczność ponownego równoważenia wynika z potrzeby zwiększenia ilości rzadko występujących segmentów oraz optymalizacji zdań już wybranych.

Przyjęto następujące założenia:

- minimalna długość fonetyczna zdania to 30 znaków[7]
- maksymalna długość fonetyczna zdania to 80 znaków
- korpus powinien zawierać około 2500 zdań
- w korpusie każdy fonem powinien wystąpić co najmniej 40 razy
- każdy difon powinien wystąpić co najmniej 4 razy
- każdy trifon powinien wystąpić co najmniej 3 razy, to wymaganie jest dostępne tylko dla najczęściej występujących trifonów

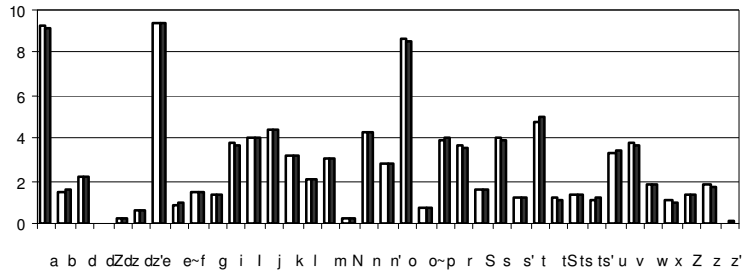
Założenia te zostały określone na podstawie prac: (Bozkurt i wsp. 2003, Clark i wsp. 2007).

Istotnym założeniem było maksymalne zwiększenie częstości wystąpienia podstawowej jednostki - difonu w korpusie.

Rysunek 4.3 zawiera porównanie pomiędzy dwoma korpusami utworzonymi w drugim etapie balansowania. Rozkład częstotliwości

7 Bez znaków interpunkcyjnych

występowania poszczególnych fonemów jest bardzo zbliżony. Pierwszy korpus zawierał 43218 zdań, a drugi 38137 zdań.



Rys. 4.3 Wykres zawiera porównanie rozkładu poszczególnych fonemów dwoma obu korpusach utworzonymi w drugim etapie balansowania. Oś pionowa zawiera względną częstotliwość występowania fonemów.

Korpus zawierający około 2500 zdań odpowiada około sześciu godzinom nagrań, co znacznie przekracza wymagania (Clark i wsp. 2007) stawiane syntezie korpusowej i wiąże się ze znacznym nakładem pracy (rejestracja korpusu, segmentacja). Zdecydowano się na większy nakład pracy, ponieważ wzrastają szanse na uzyskanie bardziej naturalnej mowy syntetycznej (Conkie 1999). Tym samym istnieje możliwość redukcji bazy dźwiękowej, dzięki temu można uzyskać krótszy czas syntezy (Clark i wsp. 2004) i jednocześnie uruchomić system na mniej wydajnych obliczeniowo platformach (np. telefon z platformą Android). Minimalna ilość zdań w systemie korpusowym nie tylko dla języka angielskiego powinna wynosić około 1000 (Clark i wsp. 2007). W głosach typu unit-selection tworzonych przez autorów systemu Festival znajduje się również około 1000 promptów (Clark i wsp. 2007, Conkie 1999). Stosując większe korpusy znacznie wydłuża się czas syntezy. Według (Clark i wsp. 2007) dla korpusu zawierającego 14000 głosek czas potrzebny na zsyntezowanie dwóch zdań zawierających około 150 fonemów wynosi 1,00 sek., dla korpusu zawierającego 36000 1,96 sek a dla korpusu zawierającego 175 000 fonemów 9,25 sek. Autor tworząc system eksperymentalny zdecydował się na stworzenie korpusu pozwalającego na uzyskanie jak najlepszej jakości syntezy mowy kosztem czasu wykonywania syntezy.(Clark i wsp. 2004)

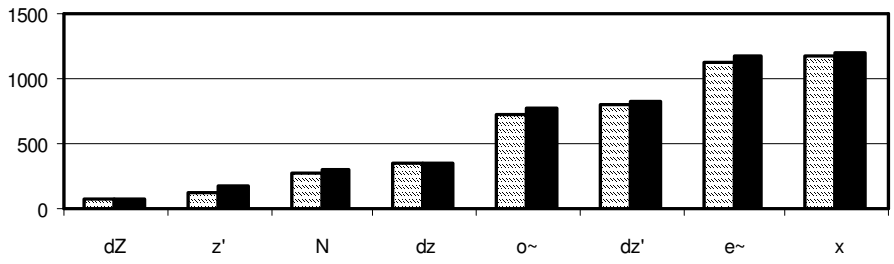
Ostateczne wyniki prezentowane w rozdziale 6 pokazują, że duża ilość czasu poświęcona na przygotowanie korpusu przyniosła pozytywne rezultaty.

4.1.6 Trzecie balansowanie

Wygenerowane korpusy po drugim etapie balansowania zawierały około 1100 różnych difonów. W przypadku trifonów, w korpusach znajdują się te, które najczęściej występują w języku polskim.

Kolejnym etapem tworzenia korpusu było połączenie dwunastu korpusów w jedną całość oraz ponowne przeprowadzenie procesu jego optymalizacji. Decyzja o realizacji tego etapu była motywowana uzyskaniem bardziej zoptymalizowanego i reprezentatywnego dla języka polskiego korpusu. W wyniku uzyskano większą częstotliwość występowania w korpusie najrzadszych w mowie polskiej fonemów. Na przykład fonem /dZ/ występujący około 55 razy w każdym z podkorpusów, po kolejnym balansowaniu pojawia się 87 razy. Korzyść z występowania większej ilości rzadkich fonemów w korpusie została zasygnalizowana w (Beutnagel i wsp. 1999). Z cytowanego artykułu wynika, że dostatecznie częste występowanie w korpusie rzadkich jednostek jest korzystne dla jakości syntezy mowy. Wynika to z faktu, że choć częstość ich występowania jest mała, to jednak istnieje duże prawdopodobieństwo, że w pojedynczym zdaniu wystąpi jeden z rzadkich fonemów. Jeśli jego realizacja w korpusie nie będzie pasowała pod względem akustycznym wtedy jakość syntezowanej wypowiedzi pogorszy się w znaczny sposób (Beutnagel i wsp. 1999, Möbius 2001, Bozkurt i wsp. 2003).

Rysunek 4.4 ilustruje rozkłady najrzadziej występujących fonemów po pierwszym oraz po drugim procesie balansowania korpusu.



Rys. 4.4 Porównanie rozkładu rzadkich fonemów w korpusie po I i II etapie balansowania. Oś pionowa reprezentuje ilość wystąpień.

Dodatkowo po drugim balansowaniu uzyskano w przypadku fonemów:

- dłuższe zdania, ilość fonemów w zdaniu zwiększyła się z 58,3916 do

59,3256 fonemów

- większą ich liczbę (145979 w korpusie po pierwszym balansowaniu 148314 w korpusie po drugim balansowaniu)
- większą bezwzględną liczbę występowania fonemów z 3945,38 do 4008,49

W przypadku difonów:

- uzyskano dłuższe zdania w korpusie. Średnia długość zdania zwiększyła się o jeden fonem (z 59,3916 do 60,3256 fonemów)
- zwiększyła się całkowita liczba difonów z 148479 do 150814
- zmniejszyła się liczba difonów występujących mniej niż 4 razy, z 175 do 68 difonów
- najważniejszym odnotowanym faktem jest zwiększenie się liczby różnych difonów występujących w korpusie z 1096 do 1196

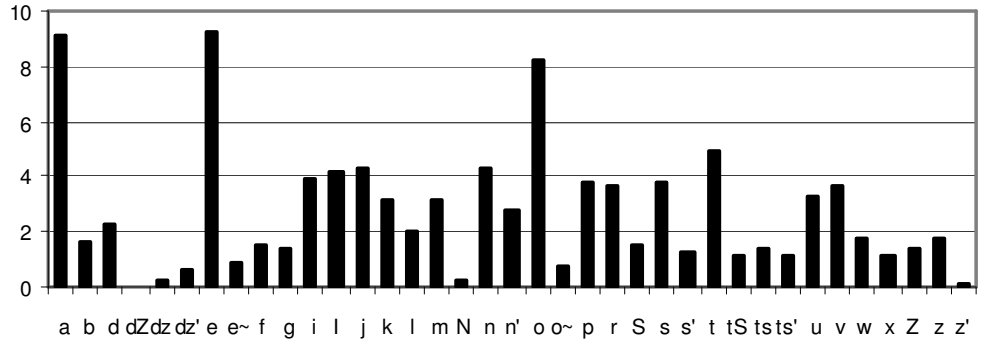
W przypadku trifonów:

- uzyskano dłuższe zdania w korpusu (z 58,3916 do 59,3256 trifonów)
- zwiększyła się całkowita liczbę trifonów z 145979 do 148314
- najważniejszym odnotowanym faktem jest zwiększenie się liczby różnych trifonów występujących w korpusie z 11524 do 13832

Lepsze rezultaty wynikają z balansowania ze sobą korpusów zrównoważonych fonetycznie w pierwszym etapie ich przetwarzania.

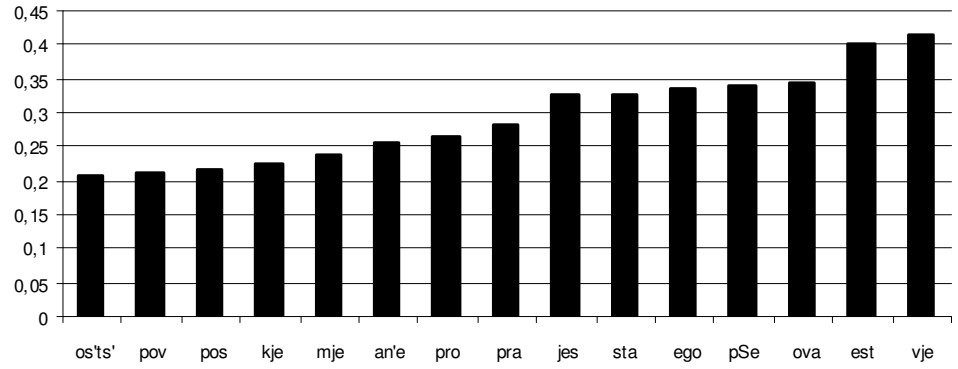
4.1.7 Końcowy etap przetwarzania korpusu

Utworzony w ten sposób korpus zawiera 2500 zdań. Zdania te wybrano z wypowiedzi sejmowych. Rysunek 4.5 przedstawia rozkład częstotliwości występowania w zaprojektowanym korpusie.



Rys. 4.5 Rozkład statystyczny 15 najczęściej występujących trifonów. Reprezentują one 4,4 % wszystkich trifonów występujących w korpusie.

Na rysunku 4.6 przedstawiono rozkład statystyczny 15 najczęściej występujących difonów w korpusie. Reprezentują one około 14,8% wszystkich difonów w korpusie.



Rys. 4.6 Rozkład statystyczny 15 najczęściej występujących difonów w korpusie.

Dwukrotne fonetyczne równoważenie pozwoliło uzyskać bogaty fonetycznie korpus. Tak powstały korpus zawiera jednak tylko teksty z przemówień sejmowych. Dlatego, zdecydowano by zweryfikować otrzymany rozkład częstotliwości występowania trifonów w drugim co do wielkości korpusie zawierającym teksty z recenzjami prasowymi. Okazało się, że w korpusie z recenzjami jest o ponad 2000 więcej trifonów niż w korpusie sejmowym. Wynika to z faktu występowania wyrazów obcojęzycznych. Autor zdecydował, żeby wykonać kolejny etap balansowania.

Obecnie otrzymany korpus oraz zbalansowany korpus z recenzjami zostały poddane etapowi kolejnej selekcji. Korpus z recenzjami zawierał 21135 zdań, 1202 różne difony oraz 20904 trifony. Balansowanie odbyło się zgodnie z wcześniej ustalonymi kryteriami i objęło dwa korpusy po 2500 zdań każdy.

Dla połączonych korpusów wyliczono rozkłady statystyczne. Poniżej zaprezentowano otrzymane wyniki:

- ilość zdań w korpusie 5000
- średnia długość wypowiedzi 59,183 fonemów
- sumaryczna liczba fonemów 295913
- liczba różnych difonów 1277
- liczba różnych trifonów 18052

Korpus ten został zbalansowany do 2500 zdań.

Do tak stworzonego korpusu dodane zostały wyrazy z rzadko występującymi fonemami w języku polskim. Lista ta zawiera ponad 300 wyrazów. Umieszczono ją w załączniku 2. Poniżej znajduje się kilka przykładów wyrazów ze stworzonej listy (Marasek i wsp. 2004)

- nozdrza
- dżul
- dżudo
- dżem
- dżungla
- dżuma
- dżip

Statystyczny rozkład korpusu z dodanymi na końcu korpusu wyrazami prezentuje się następująco:

- liczba zdań 2804
- średnia długość zdania 53,9586 fonemów
- liczba difonów 1280
- liczba trifonów 15916

4.1.8 Ręczna korekta fonetyczna i ortograficzna

Wygenerowana automatycznie transkrypcja fonetyczna zawierała błędy, zwłaszcza w wyrazach obcojęzycznych. W celu korekty błędów został stworzony słownik, który zawiera ponad 11000 słów z uwzględnionymi wariantami wymowy. Ich transkrypcja została poprawiona manualnie,

transkrypcję wyrazów obcojęzycznych poprawiono starając się odwzorować wymowę jak najbardziej zbliżoną do języka oryginalnego, ale z wykorzystaniem fonemów języka polskiego.

Kolejnym etapem była weryfikacja korpusu i usunięcie zdań, w których brakowało czasownika, lub które były zbyt krótkie (dolna granica długości zdań wynosi 40 fonemów). Poprawiono zdania, które kończyły się w niewłaściwy sposób. Na przykład, następujące zdanie:

Jak pan ją ocenia jako celnik, jako osoba która nadzoruje sprawy celne i sprawy graniczne wcale.

zdanie zostało skorygowane do postaci:

Jak pan ją ocenia jako celnik, jako osoba która nadzoruje sprawy celne i sprawy graniczne?

Ostatnim etapem przygotowania korpusu było przeprowadzenie testu mającego na celu sprawdzenie możliwości ograniczenia ilości promptów przy zachowaniu proporcjonalnie podobnej ilości jednostek akustycznych.

4.1.9 Etap testowania

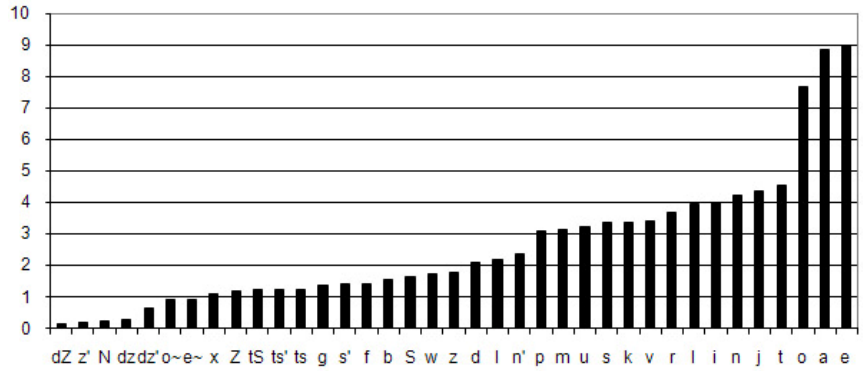
Korpus zawierający ok. 2800 zdań został kolejny raz zbalansowany i w wyniku tego utworzonych zostało 11 korpusów o różnej ilości zdań (promptów). Celem tego etapu było sprawdzenie, czy istnieje możliwość ograniczenia rozmiaru korpusu przy zachowaniu podobnej liczby segmentów.

Tabela 4.4 prezentuje otrzymane wyniki.

Ilość zdań	**2600**	**2150**	**2150v2**	**2150v3**	**2100**	**2000**	**1900**	**1800**	**1700**	**1500**	**1000**
Lb. difonów	1200	1200	1200	1200	1200	1200	1200	1200	1200	1200	1200
Lb. trifonów	14860	14819	14803	14804	14752	14582	14383	14119	13885	13344	11404
Lb. trifonów< 3	7198	7620	7616	7577	7694	7869	8044	8207	8388	8796	10094
Liczba difonów < 5	142	154	154	154	158	164	166	169	180	201	278

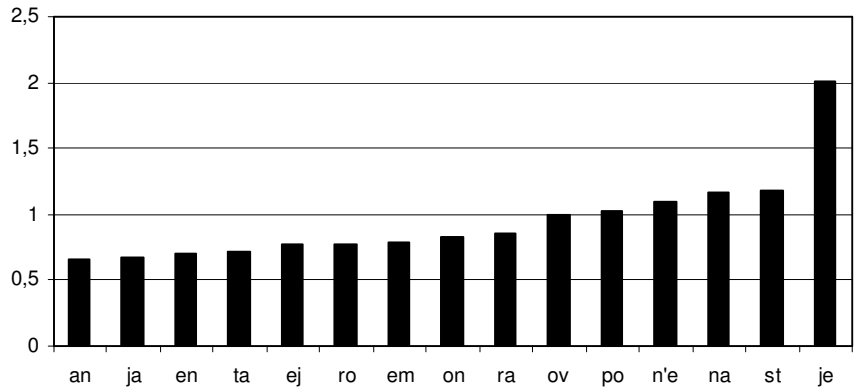
Tabela 4.4 Tabela przedstawia porównanie ilości wystąpień fonemów, difonów i trifonów w 11 korpusach o różnej wielkości w ostatnim etapie balansowania.

Ostatecznie do nagrań został wybrany korpus zawierający 2150 zdań. Korpus ten zawierał jedynie 38 trifonów mniej niż korpus z 2600 promptami. Rysunek 4.7 prezentuje względny rozkład częstotliwości występowania poszczególnych fonemów w ostatecznej wersji korpusu.



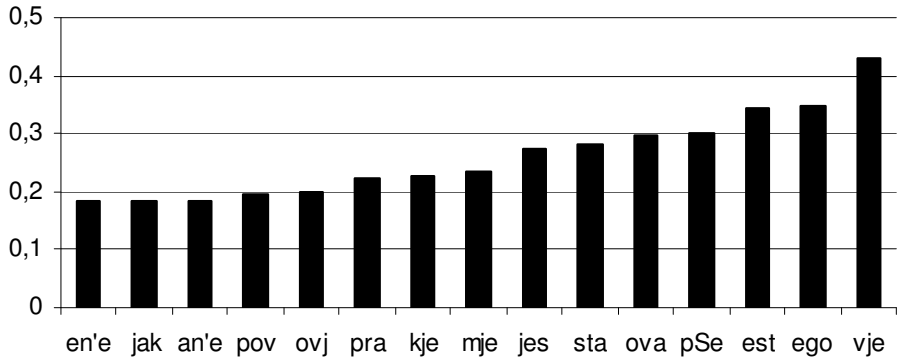
Rys. 4.7 Rozkład statystyczny fonemów w ostatecznej wersji korpusu.

Rysunek 4.8 przedstawia 15 najczęściej występujących difonów. Stanowią one 14,224% wszystkich difonów w korpusie.



Rys. 4.8 15 najczęściej występujących jednostek o długości difonu w korpusie.

Rysunek 4.9 prezentuje 15 najczęściej występujących trifonów. Reprezentują one 4,09 % wszystkich trifonów w korpusie.



Rys. 4.9 15 najczęściej występujących jednostek o długości trifonu w korpusie

Struktura tego korpusu ma następującą postać :

- liczba zdań 2150
- średnia długość zdania 60,5553
- łączna liczba fonemów 130194
- liczba różnych difonów 1200
- liczba różnych trifonów 14505

Baza akustyczna została zweryfikowana pod kątem prozodycznym. Wykonano stylizację Insint (Hirst 1999), będącą systemem anotacji wzorców prozodycznych. (Rozdział 1.5.3) Dodatkowo przeprowadzono prozodyczną anotację dla przerw między frazami w zdaniach. Przeprowadzone badania wskazują ze stworzony korpus jest prozodycznie bogaty. Tabela 4.5 przedstawia rozkład procentowy poszczególnych etykiet w systemie Insint (Oliver i wsp. 2006). Wszystkie etykiety występują w przedziale częstotliwości 9,2 % do 20, 2%.

		Ilość	Procent (%)
Etykieta	B	3 353	9,2%
	D	6 196	17,1%
	H	5 846	16,1%
	L	7 325	20,2%
	M	2 362	6,5%
	S	3 161	8,7%
	T	3 470	9,6%
	U	4 598	12,7%
Ogółem		36 311	100,0%

Tabela 4.5 Procentowy rozkład etykiet w systemie Insint

4.2 Realizacja bazy akustycznej

W poprzednim podrozdziale przedstawiony został sposób tworzenia korpusu, który stanowi „silnik" syntezatora. Kolejnym bardzo ważnym etapem jest właściwa realizacja nagrań korpusu. Udowodniono(Clark i wsp. 2004), iż realizacja bazy akustycznej w cichym pomieszczeniu oraz przy użyciu średniej jakości przetwornika analogowo-cyfrowego przynosi jedynie dostateczne efekty. Na ogół rekomenduje się realizację nagrań w komorze bezechowej z użyciem studyjnego sprzętu.

4.2.1 Realizacja nagrań

Podczas tworzenia systemu syntezy mowy należy poświęcić

odpowiednio dużo czasu i wysiłku na właściwe przygotowanie bazy akustycznej. Mówcą powinna być osoba znającą transkrypcję fonetyczną języka polskiego, o charakterystycznymi energicznym głosie. Każde zdanie powinno być wymówione jak najdokładniej, bez nadmiernych emocji.

Ponieważ system tworzony w ramach pracy doktorskiej jest systemem eksperymentalnym, autor dołożył wszelkich starań, aby jakość nagrań była jak najlepsza. Udowodniono, (Janicki 2004, Klabbers i wsp. 2004) iż sygnał mowy pozbawiony zakłóceń, pogłosu jest znaczniej łatwiej przetwarzać na dalszych etapach tworzenia korpus. Ważne jest by w nagraniach nie występowały niepożądane przydechy, mlaśnięcia, inne elementy paralingwistyczne oraz szumy. Można uniknąć ich poprzez zastosowanie odpowiedniego mikrofonu oraz jego ustawienie.

Równie istotny jest wybór odpowiedniej osoby mówiącej, która powinna się charakteryzować czystym głosem czyli pozbawionym chrypki, nosowania oraz umiejętnością utrzymania stałego F0 podczas czytania dużych ilości tekstu. Jeśli lektor posiada nienaturalną barwę głosu to w syntetycznie brzmiącej mowie będzie brzmiał on znacznie gorzej (Kominek i wsp. 2003). Głos lektora radiowego nie sprawdza się w systemach korpusowej syntezy mowy. Osoby takie są przyzwyczajone do radiowej, przesadnej intonacji, co utrudnia podczas procesu syntezy jej dostosowanie do melodii generowanego zdania. Do realizacji nagrań zastosowano opracowany korpus zawierający 2150 zdań wraz z rzadkimi wyrazami, zapisanymi w postaci ortograficznej w pliku tekstowym. W każdej linii znajduje się jeden prompt zdaniowy. Na początku każdego zdania znajduje się jego identyfikator w postaci sxxxx, gdzie xxxx oznacza kolejny numer zdania. Kolejny numer zdania oddzielony jest dwukropkiem i tabulacją od zdania zakończonego kropką, znakiem zapytania, bądź wykrzyknikiem. Przyjęty zapis czterocyfrowy pozwolił uniknąć kłopotów z kolejnością wyświetlania plików. W dalszych etapach prac, by ułatwić identyfikację, każde ze zdań zapisywane było w osobnym pliku o nazwie takiej samej jak numer zdania.

Poniżej znajdują się przykładowe zdania umieszczone w korpusie:

s0029: czy wtedy wolno już oferować wyroki lub publikować opinie co do rozstrzygnięcia ?

s0030: czy samochód z silnikiem iveco przejechać już tyle kilometrów ?

s0031: bo przecież proszę zwrócić uwagę czy jeśli chodzi o nasz przemysł dziś coś się eksportuje ?

s1633: nie chcemy nowej żelaznej kurtyny pomiędzy europą a azją

s1635: ważną częścią tych działań byłby też nurt edukacyjny i możliwość wspierania inicjatyw lokalnych

s1636: rząd proponuje by wytwórców nie wykorzystujących surowców wtórnych karać karą aresztu lub grzywny

s1637: pieniądze te daje się zaś gminom w których jasno świeci słońce rosną grzyby i jest czysta woda

s1638: najogólniej mówiąc powiedzenie jak cię widzą tak cię piszą wydaje się najlepiej oddawać sens sprawy

Nagrania były realizowane przez autora w studio nagraniowym w Polsko-Japońskiej Wyższej Szkole Technik Komputerowych, o dość przeciętnych własnościach akustycznych. Korpus został nagrany z częstotliwością próbkowania 48 kHz oraz 16 bitową rozdzielczością w formacie RAW. Sygnał o takiej częstotliwości daje się bezstratnie przepróbkować do 16 kHz, co jest pewnego rodzaju standardem w głosach syntetycznych. Do nagrań został użyty mikrofon dynamiczny Rode NT 1000. Dodatkowo użyty został pop-filter zainstalowany pomiędzy mówcą a mikrofonem. Dzięki temu uzyskano mniejszą moc strugi powietrza uderzającą w membranę mikrofonu podczas artykulacji głosek zwartych takich jak p, b, d, t, k , g. (Kominek i wsp. 2003)

Korpus został nagrany przy zastosowaniu programu Mobile Recording Studio firmy Sony. Do nagrań użyto laptopa oraz zewnętrznej karty dźwiękowej M-Audio Transit. Karta ta zapewniła wysoką jakość rejestrowanego sygnału. Interfejs został połączony za pomocą kabla optycznego z przedwzmacniaczem DIGIPORT. Umożliwiło to zminimalizowanie zakłóceń, które mogłyby powstać podczas rejestracji analogowego sygnału przy użyciu kabla typu chinch lub jack. Do weryfikacji nagrań stosowano słuchawki Beyerdynamic DT 231 PRO.

Sesja nagraniowa trwała około miesiąca i odbywała się w trudnych warunkach, podczas roku akademickiego. Studio znajduje się w pobliżu laboratoriów, co niestety miało również wpływ na jakość rejestrowanych nagrań. Przed każdą sesją odsłuchiwano poprzednie nagrania w celu uzyskania podobnej intonacji oraz sposobu mówienia. Lektor starał się wymawiać każde zdanie jak najdokładniej zgodnie z wyświetlana transkrypcją fonetyczną. Jeśli wyświetlona transkrypcja fonetyczna była niepoprawna lub inna od

fonetycznej realizacji wypowiedzianego zdania, zmiana była odnotowywana. Ma to duży wpływ na wpływ na zgodność z wymową kanoniczną. W ten sposób uzyskano korpus z dokładną transkrypcją fonetyczną. Dość istotne jest rejestrowanie sygnału z naturalną intonacją oraz czytanie w umiarkowanym tempie, przy czym stwierdzenie to jest prawdziwe dla większości języków. (Louw i wsp. 2005)

Dużą uwagę przyłożono do eliminacji wszelkiego rodzaju zakłóceń, szumów, stuków oraz innych zniekształceń harmonicznych. Podczas rejestracji sygnału stwierdzono obecność stałych składowych o częstotliwościach ok. 400-500 Hz. Częstotliwości te były przenoszone poprzez przewody wentylacyjne laboratoriów uczelni. Podjęto próbę wyeliminowania istniejących zniekształceń, ponieważ wiadomo, że im mniejsza jest manipulacja sygnałem dźwiękowym (np. automatyczna redukcja szumów, poprawa dynamiki sygnału itp.) tym lepsza jakość generowanej mowy syntetycznej.(Van Santen i wsp. 1997) Podjęte działania nie przyniosły oczekiwanych rezultatów. Z tego powodu około 25% nagrań musiało zostać powtórzonych, ze względu na występujące częstotliwości oraz dodatkowe zniekształcenia typu DC.

Nagrany korpus zawiera zdania zarówno pytające, wykrzyknikowe oraz pojedyncze wyrazy. Zdania pytające oraz wykrzyknikowe autor starał się przeczytać z przesadną intonacją. Założenie takie było podyktowana chęcią realizacji bazy akustycznej, która również będzie umożliwiała generowanie wypowiedzi pytających. W praktyce okazało się, iż 76 zdań pytających oraz 13 zdań wykrzyknikowych to zbyt mało, aby stworzyć odpowiedni model intonacyjny wypowiedzi języka polskiego. Z tych powodów w finalnej wersji korpusu zrezygnowano z tych promptów, czyli korpus został zmniejszony o 89 zdań. Lista wyrazów z rzadko występującymi fonemami została nagrana podczas osobnej sesji, w celu uzyskania jak najlepszej jakości. Wyrazy te zostały zaintonowane w neutralny sposób.

4.3 Segmentacja sygnału bazy akustycznej

4.3.1 Automatyczna segmentacja nagrań

Proces segmentacji powinien być przeprowadzany z jak największą uwagą i

dokładnością. Od jego poprawności zależy przydatność wynikowej bazy akustycznej, a w opisywanym przypadku - jakość mowy generowanej przez syntezator. (Oliver i wsp. 2006). Jeśli granica głoski zostanie niedokładnie wyznaczona i jest przesunięta w czasie, to w rezultacie może nie tylko spowodować wybranie złej jednostki do syntezy (włącznie z niewłaściwym akcentem), lecz również błędne wyliczenie funkcji kosztu doboru jednostki (Kominek i wsp. 2004). Duża baza akustyczna wymusza jednak przynajmniej częściową automatyzację tego zadania. Polega ona przede wszystkim na wstępnym dopasowaniu granic sygnału i transkrypcji fonetycznej (*alignment*), przy pomocy narzędzi opartych o mechanizmy rozpoznawania mowy.

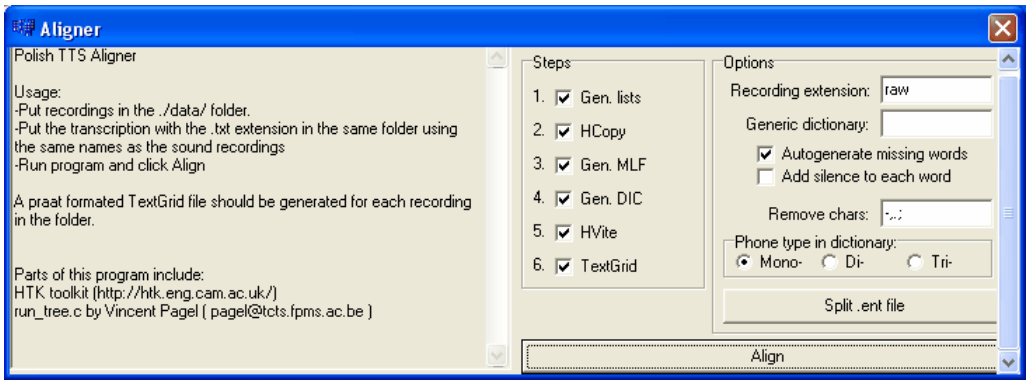
Według (Clark i wsp. 2007) automatyczny proces segmentacji jest ze względów praktycznych najbardziej optymalnym rozwiązaniem. Wynika to z faktu, że metody ręcznej segmentacji nie gwarantują spójności przebiegu procesu segmentacji oraz wyznaczania granic (Szklanny 2003).

Nadal jednak istnieje potrzeba poprawiania wyników segmentacji uzyskiwanych metodami automatycznymi. Jedna z nich została zaprezentowana w pracy (Richmond i wsp. 2007). Polega ona na stworzeniu modułu wykorzystującym regularne wyrażenia dla post-leksykalnych reguł transkrypcji fonetycznej.

Do automatycznej segmentacji stworzonej bazy akustycznej wykorzystano zmodyfikowaną wersję programu Aligner (Marasek 2003 B, Koržinek i wsp. 2007). Program ten powstał w oparciu ogólnie dostępne oprogramowania HTK[8], będącego zestawem współdziałających ze sobą narzędzi, umożliwiających przetwarzanie, oraz rozpoznawanie mowy, z wykorzystaniem niejawnych modeli Markowa (HMM) (Young i wsp. 2001). Elementy pakietu HTK stanowią użyteczne moduły dla programów wykorzystujących mechanizmy rozpoznawania mowy, obejmujący cały zakres przetwarzania związanego z tą dziedziną zagadnień. Danymi wejściowymi dla programu Aligner są pliki dźwiękowe z wraz z ich ortograficznym zapisem. Wynikiem jego działania są pliki o rozszerzeniu Textgrid, w formacie programu Praat, opisujące granice segmentów akustycznych. Aligner może wygenerować transkrypcję fonetyczną wypowiedzi w kodzie SAMPA

8 http://htk.eng.cam.ac.uk/, 12-2008

automatycznie, może również korzystać z zewnętrznego słownika odwzorowań, stworzonego ręcznie. Transkrypcja, zarówno jak i segmentacja może być dokonana w oparciu o jednostki segmentalne o rozciągłości fonemów, difonów lub trifonów. Ponadto, użytkownik Alignera może zadecydować, które etapy procesu mają być wykonane (etapy parametryzacji, transkrypcja, segmentacja). Okno programu Aligner przedstawiono na rysunku 4.10.



Rys. 4.10 Okno programu Aligner.

Należy zauważyć, iż poprawność segmentacji realizowanej przez Aligner, wynika z zastosowanych modeli Markowa (HMM). W tym celu zostało stworzonych 5 różnych zestawów modeli HMM (Szklanny i wsp. 2008).

Podczas realizacji pierwszej wersji segmentacji pominięto proces generowania transkrypcji fonetycznej. Automatycznie generowana segmentacja wymagałaby dodatkowej weryfikacji i korekty, gdyż wymowa konkretnego mówcy w pewnych przypadkach może odbiegać od reguł języka polskiego. Ponadto reguły te z definicji byłyby błędne w przypadku słów obcojęzycznych, których stosunkowo duża ilość znalazła się w korpusie. Dlatego wykorzystano słownik z transkrypcją fonetyczną, wraz z korpusem językowym. Słownik został wygenerowany w sposób automatyczny, a następnie zweryfikowany i poprawiony manualnie. Uwzględniono reguły udźwięczniania i ubezdźwięcznienia głosek znajdujących się na stykach wyrazów. Transkrypcja całego korpusu została zweryfikowana manualnie podczas realizacji korekty segmentacji za pomocą skryptu korygującego (Rozdział 4.3.5).

4.3.2 Wybór modeli HMM oraz jednostki akustycznej

Sygnał mowy jest splotem funkcji pobudzenia i odpowiedzi impulsowej kanału głosowego. Analiza cepstralna pozwala na rozdzielenia tych dwóch przebiegów (Huang 2001). Jest ona wynikiem odwrotnej transformaty Fouriera, obliczonej dla widma amplitudowego sygnału, poddanego wcześniej operacji logarytmowania. Splot sygnałów przekształcony zostaje w sumę, a składowe addytywne są rozdzielone za pomocą filtracji liniowej. Wynikiem analizy są współczynniki cepstralne, podawane najczęściej w skali melowej. Początkowe współczynniki określają ogólny charakter widma. Stanowią wektor parametrów opisujących jego obwiednię. Dodatkowo ułatwiają estymację częstotliwości formantowych (współrzędne lokalnych maksimów wygładzonego cepstrum). Pozostałe (wysokie) współczynniki mogą służyć do stwierdzenia czy istnieje, oraz ewentualnego określenia częstotliwości tonu krtaniowego (dla głosek dźwięcznych). Typowa w literaturze liczba współczynników opisujących każdą ramkę sygnału wynosi 39. Pierwszym współczynnikiem jest logarytm poziomu energii. Kolejne, to wektor 12 współczynników MFCC, opisujących charakterystykę cepstrum w skali melowej (*MFCC – Mel Frequency Cepstral Coefficients*). Pozostałe, to pochodne pierwszego i drugiego stopnia obliczone zarówno dla energii jak i 12 podstawowych współczynników, obrazujące dynamiczne zmiany w sygnale. (E, 12 MFCC, ΔE, 12 ΔMFCC, ΔΔE, 12 ΔΔMFCC) (Young i wsp. 2001, Huang i wsp. 2001).

5 zestawów modeli zostało stworzonych i wytrenowanych z wykorzystaniem pakietu HTK (*Hidden Markov Model Toolkit*). Każdy z model został sparametryzowany w przestrzeni 39 współczynników mel-cepstralnych. (Young i wsp. 2001, Black i wsp. 2006) Różnice między wyznaczonymi zestawami modeli dotyczyły reprezentowanego rodzaju segmentów akustycznych oraz rozmiaru i cyklu pobierania kolejnych ramek czasowych sygnału. Cztery z wyznaczonych zestawów modelowały głoski, piąty natomiast difony. Modeli trifonów nie stworzono z uwagi na skomplikowany i czasochłonny proces treningu. Wszystkie dostarczone komplety reprezentujące fonemy zawierały 38 modeli HMM. Dodatkowy model HMM, wykorzystywany przez program

Aligner, reprezentował ciszę. Każdy z modeli składał się z trzech stanów, reprezentujących nagłos, śródgłos oraz wygłos danego fonemu (Zhang i wsp. 2004). W pierwszym z zestawów wykorzystano okno analizy o szerokości 5 ms, pobierane z przesunięciem czasowym co 1ms. W drugim, ramka miała rozmiar 15 ms i była pobierana co 5 ms, natomiast w trzecim odpowiednio ramkę 25 ms z przesunięciem 10ms. Każdy z 3 wymienionych zestawów został wytrenowany na podstawie 585 fonetycznie dobranych nagrań wypowiedzi 40 mówców, zawierających pojedyncze wyrazy i zdania (Oliver i wsp. 2006). Aby podnieść poziom dokładności rozpoznawania, do każdego stanu wstępnie wytrenowanych modeli HMM dodano mikstury Gaussowskie, po czym ponownie estymowano parametry modeli. Ten proces we wszystkich przypadkach powtórzono 3 razy. Na zakończenie przeprowadzono ostateczną estymację parametrów wszystkich modeli HMM na podstawie zarejestrowanych nagrań 40 mówców. Czwarty zestaw ukrytych modeli Markowa reprezentujących fonemy powstał w wyniku dodatkowej estymacji trzeciego zestawu na posegmentowanej bazie akustycznej za pomocą trzeciego zestawu modeli. Ostatni zestaw HMM wykorzystujący difony wytrenowany i estymowany (Williams 1995) został na podstawie bazy danych Speecon (Marasek i wsp. 2004), która zawiera zarejestrowane wypowiedzi 600 mówców. W przypadku modeli difonów trening w całości odbywał się na podstawie nieposegmentowanych nagrań.

Test rozpoznawalności został przeprowadzony z wykorzystaniem procedury HResults będącej integralną częścią pakietu HTK. Program ten umożliwia porównanie sekwencji rozpoznanych jednostek mowy z transkrypcją nagrania zapisaną w pliku oraz określa jaki procent zdań (całych wypowiedzi) i pojedynczych wyrazów zostało poprawnie rozpoznanych. W przypadku zdań (wypowiedzi) poziom rozpoznawalności jest wyrażony w procentach jako stosunek zdań dla których rozpoznana sekwencja znaków odpowiadała transkrypcji fonetycznej, do wszystkich zdań. W przypadku pojedynczych wyrazów program porównując rozpoznaną sekwencję jednostek z transkrypcją fonetyczną oblicza odległość między tymi dwoma ciągami znaków (odległość Levenshteina) (Young i wsp. 2001).Odległość ta wyrażona jest w liczbie operacji wstawienia, usunięcia lub zamiany symboli (jednostek,

znaków), które należy wykonać, aby rozpoznana sekwencja jednostek była identyczna z transkrypcją fonetyczną danego nagrania. Wyniki rozpoznawalności przygotowanych modeli, na podstawie 125 wypowiedzi, o tematyce związanej z dziedziną informatyki, przedstawiono w tabeli 4.6.

Modele HMM	Rozpoznane słowa (%)	Rozpoznane zdania (%)
(fonemy) ramka 5ms, pobierana co 1ms	38,14	33,06
(fonemy) ramka 15ms, pobierana co 5ms	71,47	55,65
(fonemy) ramka 25ms, pobierana co 10ms	93,27	89,52
(fonemy) ramka 25ms, pobierana co 10ms, estym. na b.ak.	92,95	89,52
(difony) ramka 25ms, pobierana co 10ms	71,79	53,23

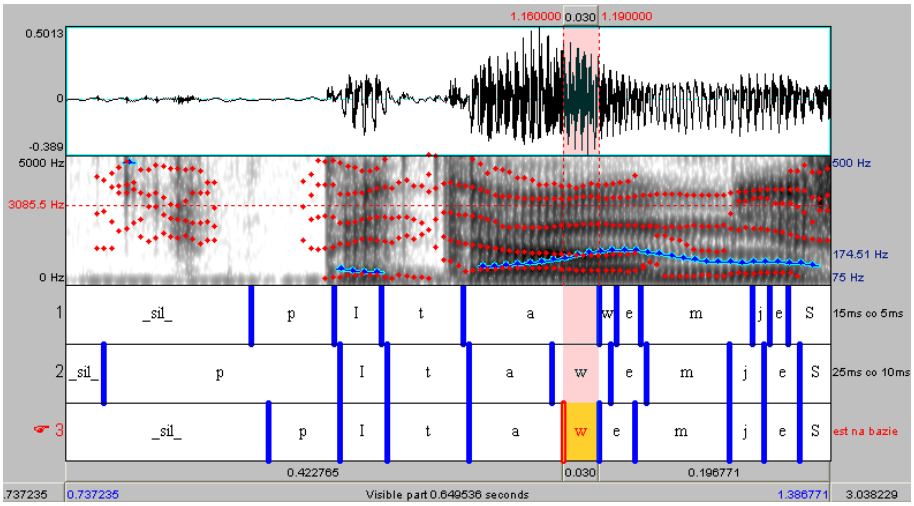
Tabela 4.6 Porównanie poziomu rozpoznawalności różnych modeli HMM.(Szklanny i wsp. 2008)

Poziom rozpoznawalności danego zestawu modeli HMM nie określa poprawności wyznaczanych przez niego granic jednostek akustycznych. Poziom rozpoznawalności niższy o 0,32 % od najlepszego zestawu modeli, dzięki adaptacji do konkretnego mówcy i warunków nagrań, pozwolił na uzyskanie mniejszej ilość błędów w segmentacji (Oliver i wsp. 2006).

Dowodem na to jest zrealizowane porównanie automatycznej segmentacji wybranych fragmentów bazy akustycznej, wygenerowanej z wykorzystaniem różnych zestawów modeli HMM. W teście tym postanowiono pominąć pierwszy zestaw o ramce 5 ms, pobieranej co milisekundę, ze względu na zbyt niski poziom rozpoznawalności (38,14%).

Porównanie poprawności segmentacji różnych zestawów modeli Markova przeprowadzono w programie Praat (Boersma 2001). Umożliwia on umieszczenie w jednym oknie programu kilku wersji (warstw, poziomów) segmentacji wraz z przebiegiem czasowym i częstotliwościowym (spektrogramem) analizowanej wypowiedzi. Porównano segmentację wygenerowaną przez różne modele fonemów i wybrano najlepsze z nich. Następnie, aby obiektywnie porównać wybrane wcześniej modele fonemów z modelami difonów, wymagane było przekonwertowanie segmentacji opartej na difonach do postaci opartej na głoskach. Dlatego, aby przekonwertować difony na głoski, wystarczyło przeciąć każdy z nich w połowie i usunąć granice między sąsiadującymi ze sobą difonami. Powstałe w ten sposób przekształcenie difonów na głoski nie zawsze jest idealne, jednak było wystarczające na potrzeby porównania wyników automatycznej segmentacji z

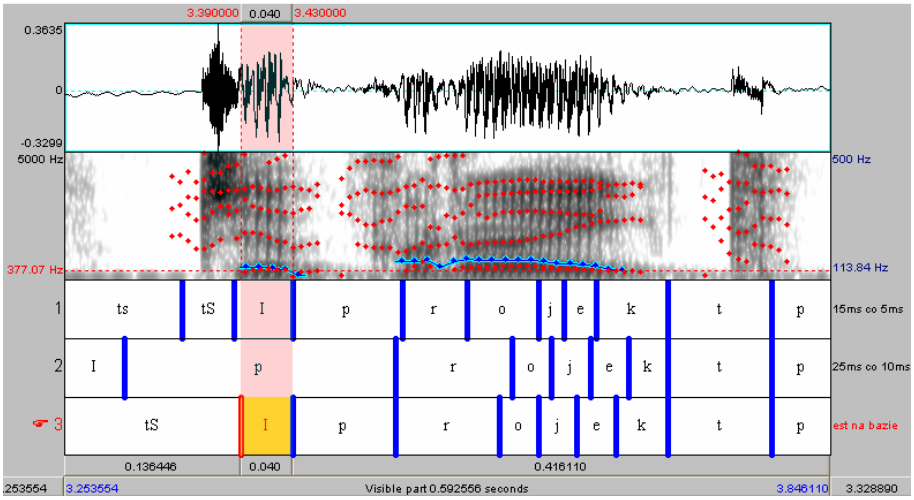
wykorzystaniem obu tych jednostek sygnału akustycznego. Rysunki 4.11 i 4.12 przedstawiają fragmenty wyników segmentacji uzyskanych dla różnych zestawów modeli fonemów. Na obu rysunkach kolejne wersje segmentacji to zaczynając od góry: model o ramce 15 ms i kroku 5 ms, model o ramce 25 ms i kroku 10 ms oraz te sam model o ramce 25 ms, pobieranej, co 10ms, dodatkowo estymowane na bazie akustycznej. (Williams 1995)



Rys. 4.11 Porównanie segmentacji opartej na modelach (HMM) fonemów.

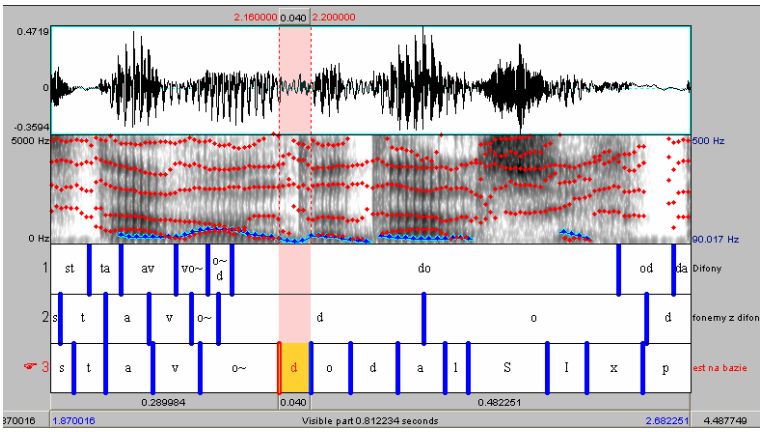
Zgodnie z wynikami testu rozpoznawalności, najmniejszą dokładność w określaniu granic głosek uzyskano przy zastosowaniu modeli o ramce 15 ms, pobieranej co 5 ms, pomimo ich teoretycznie największej precyzji (rzędu 5 ms). Zdarzało się, iż dana głoska w całości znajdowała się poza wyznaczonymi granicami (Szklanny i wsp. 2008).Wybór pomiędzy pozostałymi dwoma zestawami nie był tak oczywisty. Na przykład, trudne w segmentacji okazały się głoski płynne (/l/, /r/, /j/, /w/), których granice czasem nie są wyraźne i nawet ręcznie trudno jest je wyznaczyć. Dość często i niezależnie od zestawu modeli HMM, głoski te były błędnie oznaczone (szczególnie w połączeniu z samogłoskami), dlatego trudno było wskazać zestaw modeli, który pozwalał najlepiej wyznaczać ich granice. W wielu sytuacjach podstawowe modele wykazywały większą precyzję, niż te po dodatkowej estymacji. Powodem była w pełni automatyczna segmentacja materiału wykorzystanego do dodatkowych obliczeń. Brak adaptacji do konkretnego mówcy, powodował jednak częstsze występowanie poważnych błędów niż miało to miejsce w przypadku drugiego rozważanego zestawu. Problemem były np. sytuacje, w których szum związany

z oddechem mówcy uznawany był za początek wyrazu i całe głoski wyznaczane były w miejscu ciszy, co pociągało za sobą dalsze kolejne błędy segmentacji. Dlatego pomimo dokładności rozpoznawania niższej o 0,32%, od uzyskanej dla modeli podstawowych, do dalszych prac nad segmentacją wybrano modele estymowane dodatkowo na bazie akustycznej (ramka 25 ms, krok 10 ms). Ich dużą zaletą była największa spośród wszystkich zestawów modeli duża przewidywalność i systematyczność popełnianych błędów, co ułatwiało w wielu przypadkach ich wyszukiwanie i ich korektę (Szklanny i wsp. 2008).

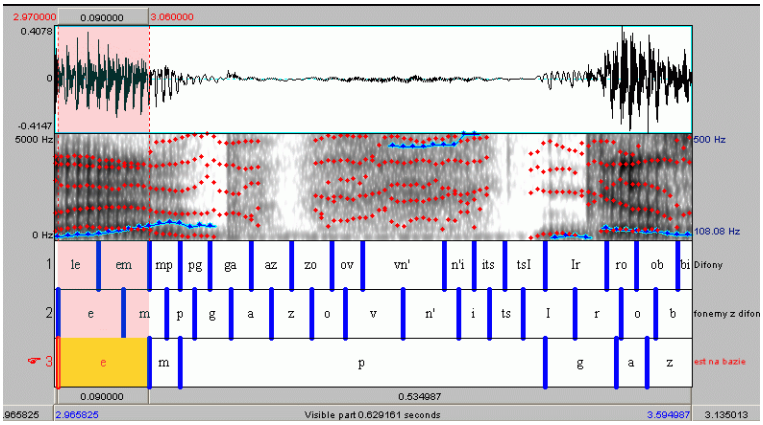


Rys. 4.12 Porównanie modeli HMM opartych na głoskach. Rysunek obrazuje niewłaściwe wykrywanie granic w głoskach z przydechem na początku (zwarto-trące i plozyjne bezdźwięczne)

Dokładność segmentacji na difony okazała się niższa, niż w przypadku drugiego porównywanego kompletu zestawu modeli fonemów. Głównymi przyczynami niedokładności segmentacji mogły być: zbyt mała baza danych treningowych, trening w całości na podstawie nieposegmentowanych nagrań, czy też brak adaptacji do konkretnego mówcy i warunków nagrań. Przykładowe porównanie wyników segmentacji dla modeli głosek z modelami difonów obrazują rysunki 4.13 i 4.14. Na nich przedstawiono kolejne wyniki segmentacji: (od góry) dla modeli difonów, difonów przekonwertowanych na głoski oraz głoskach estymowanych na bazie. Oba rysunki pokazują skrajne błędy, pojawiły się podczas segmentacji. Rysunek 4.14 dobrze wskazuje jak silnie wpływa niekorzystny stosunek sygnału do szumu na dokładność segmentacji.



Rys. 4.13 Porównanie modeli głosek z modelami difonów dla głosek wybuchowych. Pierwsza warstwa (od góry) pokazuje sposób segmentacji na modelach difonów, kolejno difonów przekonwertowanych na głoski, oraz głosek estymowanych na bazie.



Rys. 4.14 Porównanie modeli głosek z modelami difonów, przy niekorzystnym stosunku sygnału do szumu.

4.3.3 Korekta wyników automatycznej segmentacji

Po realizacji automatycznej segmentacji nagrań, należało przeprowadzić korektę błędów. Korekta polegała na zidentyfikowaniu, wyszukaniu oraz poprawieniu wszelkich cyklicznie powtarzających się jednoznacznych błędów. Ze względu na rozmiar bazy akustycznej, postanowiono znacznie zautomatyzować proces wyszukiwania błędów. W tym celu napisano skrypty, których działanie polegało na obliczeniu czasu trwania każdej głoski w nagraniach, wyliczeniu globalnych średnich i odchylenia standardowego dla różnych głosek oraz wyszukaniu i wypisaniu wystąpień, których czas trwania znacząco odbiegał od wyliczonych wcześniej średnich

(2x odchylenie standardowe) (Oliver i wsp. 2006, Kominek i wsp. 2004). Metoda ta okazała się w miarę skuteczna w wyszukiwaniu istotnych błędów w segmentacji, jak i błędnej transkrypcji fonetycznej. W wyniku działania skryptów otrzymano listę ok. 4500 wykrytych głosek o nienaturalnym czasie trwania. Kilka przykładowych wpisów przedstawiono poniżej. Kolejne kolumny to od lewej: numer nagrania, symbol fonemu oraz miejsce (podane w sekundach od początku trwania pliku dźwiękowego) danego fonemu w nagraniu.

Przykładowe wpisy wygenerowanej listy błędów:

...

s1000.phones	*k*	*6,21000 dur*
s1002.phones	*s'*	*5,31000 dur*
s1003.phones	*p*	*2,02000 dur*
s1003.phones	*dz'*	*2,83000 dur*

...

Należy zaznaczyć, iż większość plików dźwiękowych zawierała przynajmniej jeden fonem, który znalazł się na liście wygenerowanej przez skrypty (Oliver i wsp. 2006). Procedura korekty polegała na ręcznej weryfikacji i poprawieniu zarówno transkrypcji fonetycznej, jak i segmentacji całego nagrania, a przede wszystkim granic głoski wskazanej przez skrypty.

4.3.4 Ręczna korekta błędów automatycznej segmentacji

W procesie segmentacji dużej bazy akustycznej przeznaczonej na potrzeby korpusowej syntezy mowy istotna jest konsekwencja w wyznaczaniu granic segmentów. W algorytmie korpusowej syntezy mowy wykorzystuje się różnej długości segmenty sygnału akustycznego. Dokładność ich łączenia ze sobą ma decydujący wpływ na jakość generowanej mowy, dlatego tak istotna jest ręczna lub możliwie dokładna automatyczna korekta zaobserwowanych błędów w segmentacji oraz transkrypcji (Adell i wsp. 2004).

Podczas ręcznej korekty błędów starano się zidentyfikować charakterystyczne błędy powstające w procesie automatycznej segmentacji. Działania te miały na celu opracowanie listy błędów, a także zbadanie

możliwości wprowadzenia półautomatycznej korekty niektórych z tych błędów. Potrzeba ograniczenia liczby nanoszonych poprawek wynikała z dużej czasochłonności dokładnej weryfikacji i korekty wypowiedzi wskazanych przez skrypty (ok. 1400, czyli ok. 2/3 wszystkich promptów). Skrócenie czasu nanoszenia poprawek dla jednego nagrania o minutę pozwoliło na oszczędność ok. 1400 minut co przekłada się na kilka dni pracy. Należy też zauważyć, że bardzo dokładna korekta wszystkich błędów przyniosłaby umiarkowane korzyści, biorąc pod uwagę, iż 1/3 nagrań nie wymagała poprawek. Ze względu na brak podobieństwa poprawionej części nagrań segmentacji z wersją wygenerowaną automatycznie, jakość syntezy mowy byłaby zmienna i zależna od wykorzystanych nagrań. Dlatego postanowiono podczas nanoszenia ręcznych poprawek zachować pewną zgodność z automatycznie wygenerowaną segmentacją, co w wielu przypadkach oznaczało większą tolerancję na przybliżony niekiedy charakter wyznaczonych granic, a czasem nawet pozostawienie błędu. Zauważone zostały problemy związane z głoskami zwartymi /p/, /t/, /k/ (wybuchowe bezdźwięczne), występujące między wyrazami, które w automatycznej segmentacji prawie zawsze zaczynały się zbyt wcześnie, od końca poprzedniej głoski. Ręcznie poprawiane były jedynie w skrajnych przypadkach.

Największa część błędów wskazanych przez skrypty dotyczyła głosek znajdujących się na końcach wypowiedzi lub dłuższych przerw między wyrazami, które nie były rozpoznawane i były dołączane do sąsiadujących głosek, powodując na ogół ich nienaturalne wydłużenie. Nie są to istotne błędy z punktu widzenia zastosowania, a w przypadku ciszy dołączanej do ostatnich głosek wypowiedzi, były wręcz mało istotne.

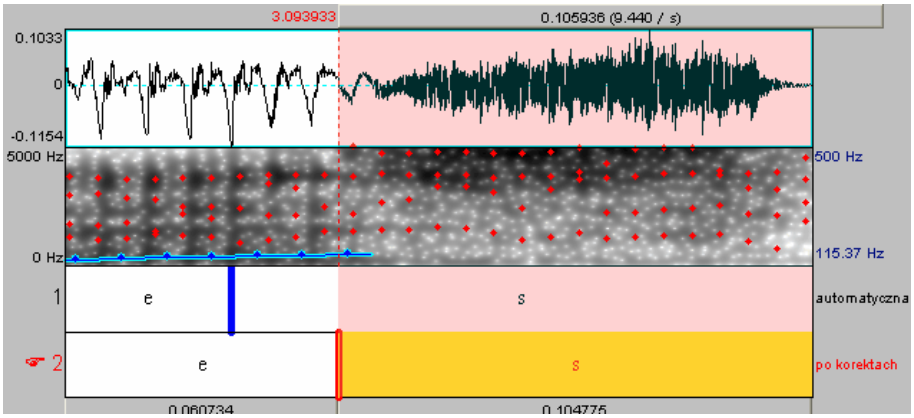
Zaobserwowano również, że transkrypcja fonetyczna zawierała dość dużo rozbieżności w stosunku do wymowy w nagraniach, w dużej mierze dotyczących obcojęzycznych słów, imion czy nazw własnych, których dość duża liczba znalazła się w korpusie. Zlokalizowano także przypadki różnej wymowy tych samych wyrazów wymagające poprawienia ich transkrypcji. Przykładem może być wyraz /mógłby/, który wymawiany był zarówno /mugwbI/, jak i /mugby/ (obie formy są dopuszczalne, druga dość częsta zredukowana forma). Błędy lub rozbieżności dotyczące transkrypcji

fonetycznej poprawiane były zgodnie z wymową, poprawki nanoszono także w słowniku transkrypcji. Najczęstsze błędy automatycznej segmentacji przedstawia tabela 4.7.

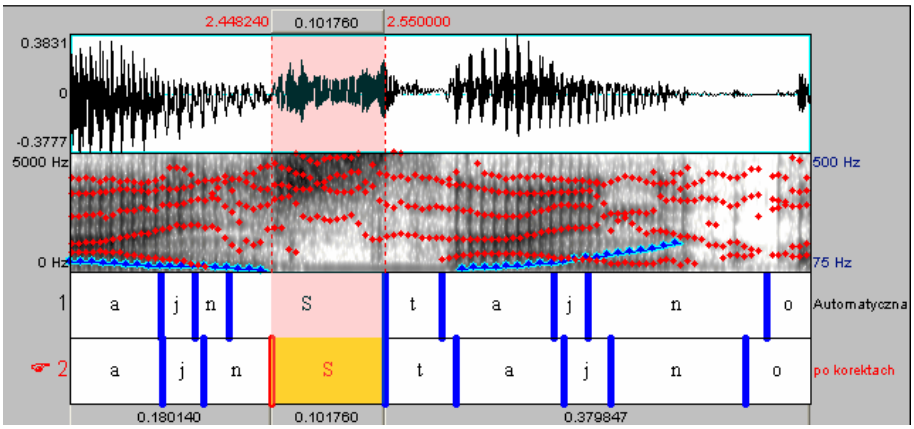
Fonemy których dotyczy	Krótki opis typowych błędów
/p/, /t/, /k/ (wybuchowe bezdźwięczne)	- zaczynają się od części poprzedniego fonemu, a nie od początku ciszy - zbyt wczesne zakończenie (np. brak plozji) - niekiedy wyznaczone granice obejmują koniec poprzedniego fonemu + cisza
/b/, /d/, /g/ (wybuchowe dźwięczne)	- za krótki segment (bez zwarcia dźwięcznego)
wybuchowe bezdźwięczne w połączeniu z trącymi (np. /pS/, /t S/, /ks'/, /ps/)	- niekiedy do spółgłoski wybuchowej dołączany jest fragment spółgłoski trącej
samogłoska w połączeniu z: /s/, /S/, /Z/, /z/, /s'/ (trące)	- część samogłoski często jest przydzielana do następującej po niej spółgłoski
/ts'/, /tS/, /ts/ (zwarto-trące), także w połączeniu z samogłoskami	- w wielu przypadkach granica początkowa segmentu wypada końcowej części poprzedniego fonemu - w połączeniu z samogłoskami koniec fonemu znajdował się w początkowym fragmencie samogłoski
/l/, /w/, /r/, oraz /v/	- często zbyt krótkie segmenty
dwa takie same fonemy jeden po drugim (głoski podwójne – geminaty) np. /rannI/	- jeśli nie były wyraźnie rozdzielone przez mówcę, prawie w całości oznaczane są jako jeden fonem,
ostatni fonem w nagraniu	- na ogół do fonemu dołączany jest fragment ciszy
dłuższe przerwy międzywyrazowe	- cała cisza przydzielana jest do sąsiadujących fonemów

Tabela 4.7 Najczęstsze błędy automatycznej segmentacji.

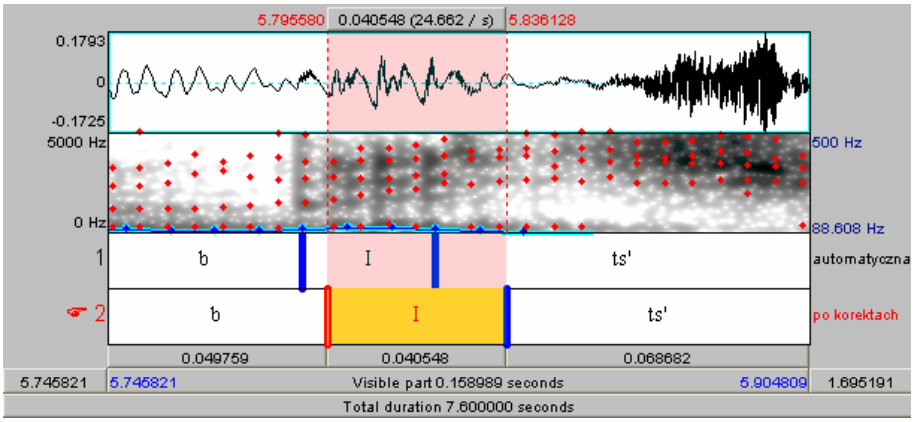
Przykłady kilku błędów oraz ręcznych korekt przedstawiono na rysunkach 4.15, 4.16 oraz 4.17. Na rysunkach tych pierwsza od góry warstwa opisu przedstawia opis wygenerowany automatycznie, natomiast dolna uwzględnia wprowadzone korekty.



Rys. 4.15 Przykład korekty częstego błędu automatycznej segmentacji – samogłoska /e/ w połączeniu z trącą /s/.



Rys. 4.16: Przykład ręcznych korekt automatycznej segmentacji.



Rys. 4.17 Inny przykład ręcznych korekt.

Na podstawie listy wynotowanych błędów został opracowany skrypt korygujący.

4.3.5 Opracowanie skryptu korygującego oraz weryfikacja jego działania

Ze względu na wymogi stawiane przez syntezę konkatenacyjną, granice wszystkich głosek powinny zostać wyznaczone dokładnie w miejscu dodatniego przejścia sygnału przez zero (z wartości ujemnych na dodatnie) (Szklanny 2002). W przeciwnym przypadku mogą powstawać trzaski w syntezowanej mowie, wynikające z braku ciągłości zmian amplitudy. Automatyczna segmentacja nie spełniała tego wymogu ze względu na wykorzystane modele HMM, a także brak odpowiednich mechanizmów w programie Aligner. (Koržinek i wsp. 2007). Dokładność automatycznie

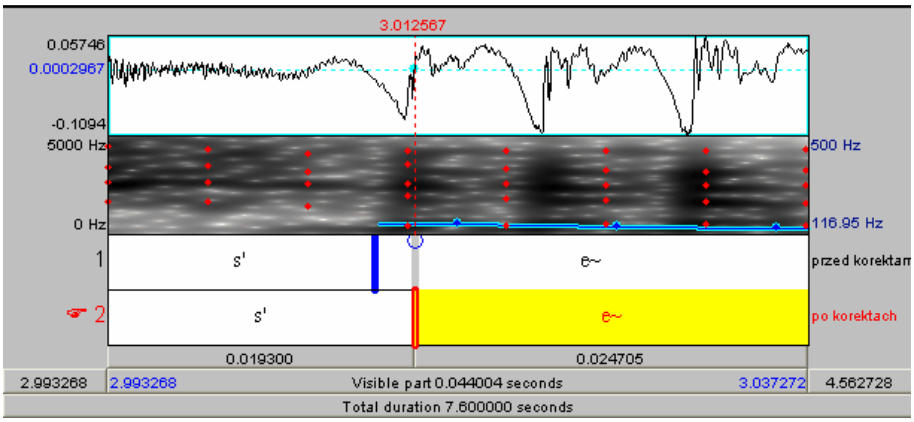
wyznaczonych granic wynosi 10ms i wynika ona z częstotliwości pobierania kolejnych ramek sygnału podczas jego parametryzacji.

Opracowano algorytm zapewniający przesunięcie granic głosek do najbliższych dodatnich przejść sygnału przez zero oraz weryfikację wprowadzanych zmian. Za podstawowe kryterium weryfikacji wprowadzanych przez skrypt korekt uznano odległość, o jaką skrypt przesuwał granice danej głoski względem ich poprzedniego położenia. Jeśli przesunięcie było większe niż 50 milisekund przypadek był raportowany, w skrajnych sytuacjach dodatkowo pozostawiana była dawna granica. Automatyczna korekta dotyczyła tylko głosek /p/, /t/, /k/. Są to jedne z najbardziej istotnych błędów na liście, gdyż zdarzały się przypadki wspomnianych głosek, które zawierały tylko koniec poprzedniej głoski wraz z fragmentem ciszy.

Należy zauważyć, że plozje bezdźwięczne poprzedzone są bardzo dużym spadkiem energii (cisza) tuż przed ich początkiem, po którym następuje duży skok energii (część wybuchowa - impuls) zakończony niekiedy krótkim szumem zwłaszcza w przypadku głoski /k/. Na podstawie śledzenia zmian poziomu energii dobrane zostały takie wartości parametrów, które zapewniły minimalizację ryzyka zbyt późnego wykrycia początku głoski oraz jej końca. Algorytm sprawdzał wartość energii w trzech miejscach, w przejściu w zerze, 0,0001 sekundy przed oraz 0,0015 sekundy za przecięciem w zerze. Jeśli wartość energii wyznaczona przez program Praat przed przejściem jest większa niż 0,002 Pa^2 lub wartość energii za punktem przecięcia jest większa niż 0,006 Pa^2 oznaczało to że granica jest prawidłowa, w przeciwnym wypadku następowała jej korekta.

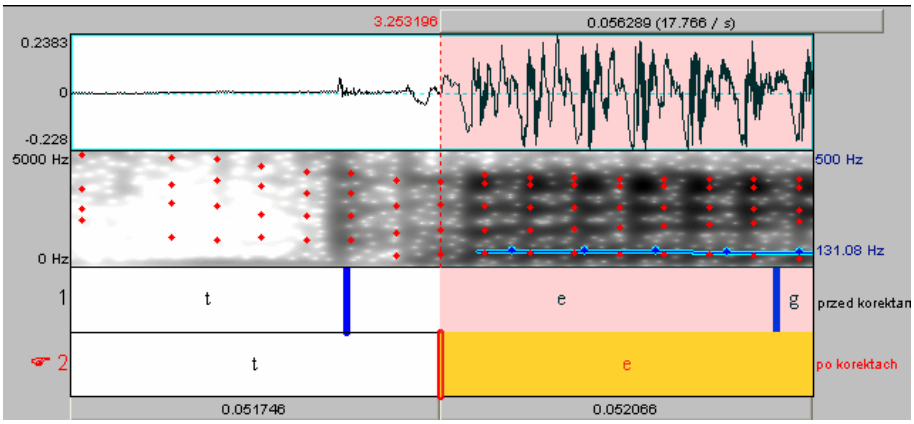
Problemem były tutaj duża ilość i zróżnicowanie nagrań, przekładające się na mnogość kontekstów, w jakich występowały głoski, ich różny czas trwania oraz poziom energii.

Na rysunku 4.18 przedstawiono przykład przesunięcia przez skrypt granicy głoski do dodatniego przejścia sygnału przez zero. Natomiast przykład wprowadzonej przez skrypt korekty głoski /t/ przedstawiono na rysunku 4.20.

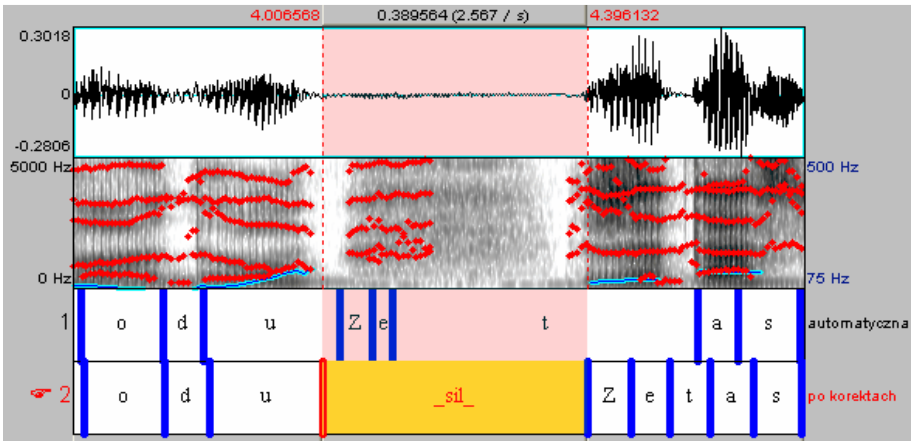


Rys. 4.18 Przykład przesunięcia granicy do dodatniego przejścia przez zero.

Rysunek 4.19 przedstawia porównanie pierwszej automatycznie wygenerowanej segmentacji (góra) z wersją uwzględniającą opisane wszystkie etapy korekty (dół). Rysunek z 4.20 ilustruje wprowadzone korekty.



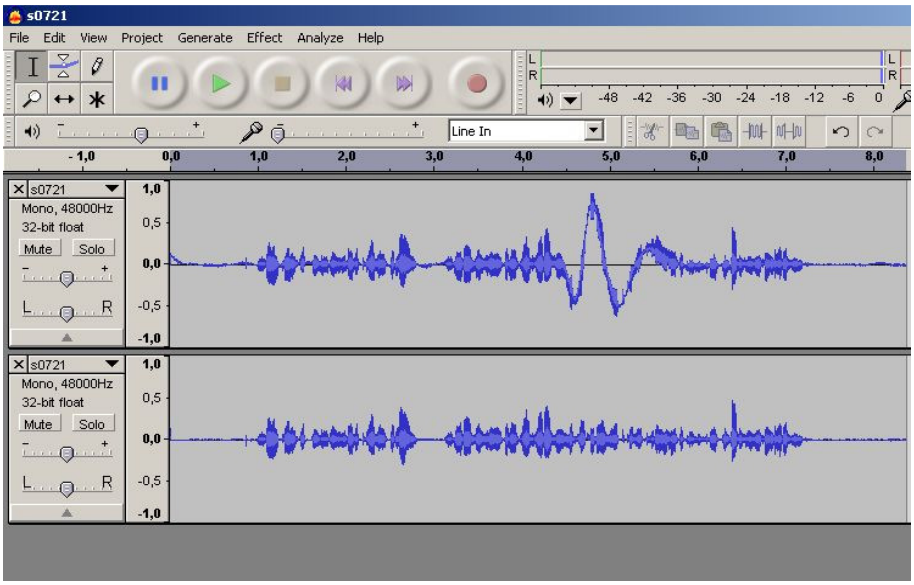
Rys. 4.19 Przykład korekty wprowadzonej przez skrypt.



Rys. 4.20 Porównanie automatycznej segmentacji oraz wersji po korektach.

Całość korpusu została poprawiona w związku z nałożeniem się

niskich częstotliwości (50 Hz) na sygnał. Zakłócenia objawiały się zmiennym, oscylującym przesunięciem sygnału względem zera, powodując miejscami brak jego przejścia przez zero. Z pomocą skryptu korygującego sporządzono listę nagrań, w których one występowały. Przytoczone zakłócenia usunięto za pomocą filtracji górnoprzepustowej (*high-pass filter*) z częstotliwością graniczną 50Hz. Do tego zadania wykorzystano program Audacity[9] rozpowszechniony na licencji freeware. Rysunek 4.21 ilustruje okno programu Audacity zawierające przebieg sygnału z zakłóceniami oraz przebieg tego samego sygnału poddanego filtracji.



Rys.4.21 Filtracja zakłóceń sieci elektrycznej w programie Audacity.

W ten sposób usunięto wszelkie istotne błędy powstałe w wyniku automatycznej segmentacji. Weryfikację bazy przeprowadzono w testowym syntezatorze.

4.3.6 Wstępna weryfikacja segmentacji w testowym syntezatorze

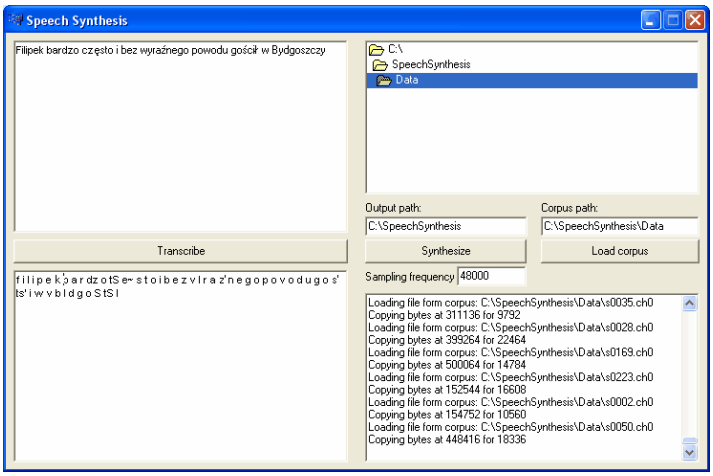
Weryfikację bazy akustycznej przeprowadzono w syntezatorze stworzonym przez studentów Polsko-Japońskiej Wyższej Szkoły w ramach prowadzonych zajęć z Podstaw Fonetyki Akustycznej (autorzy: Danijel Koržinek oraz Łukasz Brocki). Okno syntezatora przedstawiono na rysunku

9 http://audacity.sourceforge.net/, 12-2008

4.22.Wyznacznikiem dokładności segmentacji i etykietyzacji, stanowiącej warunek zakończenia etapu korekt, była liczba (10) błędnych głosek, zniekształceń, trzasków i innych zakłóceń w mowie syntetycznej, których powodem mogła być błędna segmentacja.

Procedura weryfikacji polegała na generowaniu oraz odsłuchu, czasem też analizy spektrogramu pojedynczych wypowiedzi, generowanych z tekstów pobranych z różnych stron internetowych. Dodatkowo (głównie w celach porównawczych) wygenerowano 50 zdań, z korpusu testowego (Rozdział 5.4). Korpus ten powstał przede wszystkim na potrzeby końcowych testów.

Wygenerowano również około 30 zdań testowych przed wprowadzonymi poprawkami oraz po poprawkach. Zdania były odsłuchiwane przez dwóch ekspertów lingwistycznych. Głównym celem testu była weryfikacja wprowadzonych zmian oraz ich wpływu na jakość generowanej mowy. W 90% generowana mowa lepsza niż przed wprowadzonymi poprawkami.



Rys. 4.22 Okno testowego syntezatora.

Należy zauważyć, że wykorzystany syntezator nie powstał z myślą o testowanej bazie akustycznej. By wygenerować zadaną wypowiedź, funkcja kosztu wybierała i łączyła możliwie najdłuższe fragmenty zarejestrowanej mowy. Różnice w F0 pomiędzy łączonymi grupami jednostek akustycznych nie były brane pod uwagę. Ze względu na dużą ilość nagrań w testowanej bazie akustycznej, a także zawarcie w niej nawet zdań pytających, przy wspomnianym kryterium doboru istniało duże ryzyko połączenia w procesie syntezy skrajnie różnych pod kątem czasu trwania czy intonacji fragmentów

mowy, np. fragmentów zdania oznajmującego i pytającego. Oznaczało to, iż pewne zniekształcenia musiały się pojawiać niezależnie od poprawności segmentacji, utrudniając także jej weryfikację (Wójtowski 2007).

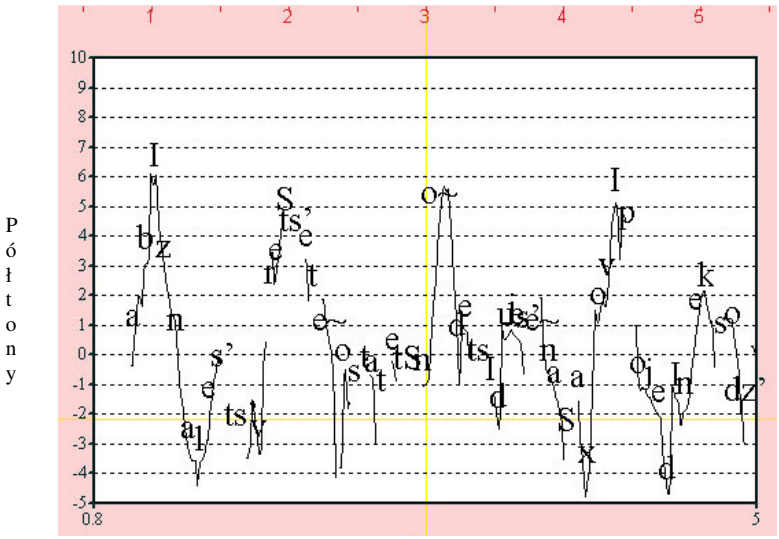
4.4 Poprawa jakości głosu w prototypowym głosie Multisyn w środowisku Festival

Po przygotowaniu bazy akustycznej wykonano szereg czynności w środowisku Festival pozwalających na uruchomienie w nim prototypowego głosu. Zmodyfikowano moduły lingwistyczne języka polskiego (Oliver 1998). Przygotowano struktury zdaniowe opisujących lingwistyczne zależności w nagranym korpusie. W kolejnym etapie wyekstrahowano z sygnału kontur F0 następnie przygotowano sygnał z opisem pitchmarków oraz sparametryzowano bazę akustyczną (LPC i MFCC).

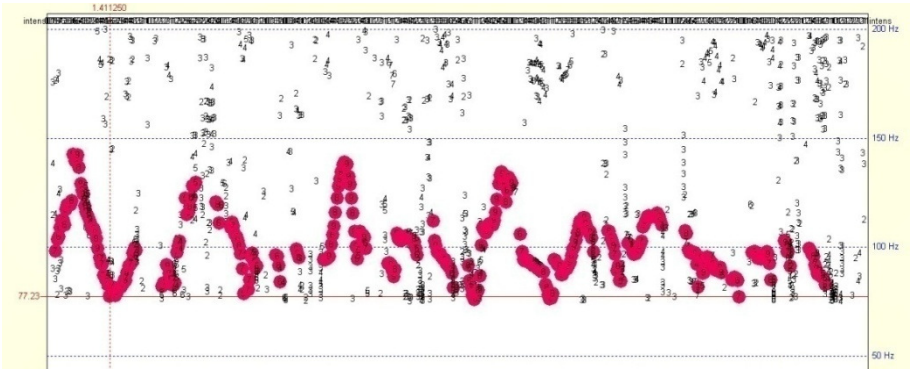
Pierwsza wersja przygotowanego głosu nie brzmiała naturalnie. Udowodniono, że złe brzmienie w głosie korpusowym jest związane z błędami w segmentacji oraz niewłaściwym śledzeniem i umieszczaniem maksimów amplitudy pobudzenia krtaniowego a to jest niezbędne do ustalenia prawidłowej melodii wypowiedzi (Black i wsp. 2006). Jeśli w sygnale mowy występuje zbyt wiele segmentów, które nie posiadają wspomnianego maksimum oznacza to, że moduł umieszczający je posiadał niewłaściwe parametry, bądź mówca mówił zbyt szybko i nastąpiła częściowa lub całkowita redukcja głosek. Jeśli źle zostały wyznaczone kontury F0, wówczas algorytm synchronizujący je podczas łączenia nie jest w stanie wybrać właściwych jednostek segmentalnych. Jeśli mówca mówi zbyt szybko, wówczas dokładność segmentacji maleje, ponieważ przewidywana z tekstu sekwencja wystąpienia określonych głosek jest znacznie mniej prawdopodobna, niż ta która została wypowiedziana. W wyniku tego następuje desynchronizacja sygnału z tekstem ortograficznym oraz jego transkrypcją fonetyczną. W pierwszym prototypowym głosie nie zwrócono uwagi na wpływ tempa mowy na dokładność segmentacji i etykietyzacji, dlatego jakość mowy syntetycznej była początkowo niezadowalająca.

Dość istotnym problemem jaki ma wpływ na jakość syntetycznej mowy

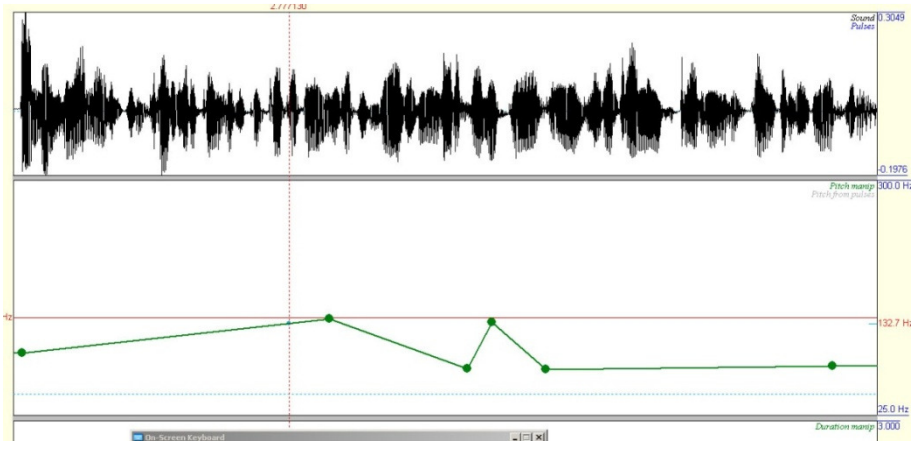
jest dobór odpowiedniego mówcy. Powinien charakteryzować się niezbyt ekspresyjnym, energicznym głosem. Równie ważny jest prawidłowy sposób akcentowania wypowiedzi. Niestety, w zarejestrowanej bazie zaobserwowano duże fluktuacje F0 oraz akcentowanie niewłaściwych sylab. To znacznie utrudnia realizację płynnej syntetycznej mowy. Z własnych badań wynika, że różnice na akcentowanej sylabie w postaci 3-4 półtonów są już słyszalne. Na rysunku 4.24 przedstawiono zmiany F0 dla zdania oznajmującego: „Aby znaleźć wreszcie tę ostateczną decyduje się na szachowy pojedynek z odzianą w czarną opończę śmiercią.” Można zaobserwować znaczne wahania już dla pierwszej sylaby /aby/ wynoszące około 5 półtonów. Dla wyrazu „ostateczną” jest podobnie, skok wynosi około 5 półtonów. Dodatkowo akcent został przesunięty na ostatnią sylabę (powinna być akcentowana sylaba /te/) co w konsekwencji doprowadziło do zdominowania akcentu rytmicznego przez akcent melodyczny. Mówca charakteryzuje się wymową melodyjną co w konsekwencji prowadzi do opisanych zmian. W konsekwencji utrudnia to łącznie ze sobą jednostek akustycznych we właściwych miejscach i syntetyczna mowa również posiada fluktuacje, które w znacznym stopniu wpływają na jakość brzmienia. Pewnego rodzaju rozwiązaniem tego problemu mogłoby być zrównoważenie melodyczne bazy akustycznej do ok. 6-8 półtonów. Sytuację przed równoważeniem przedstawia rysunek 4.23, 4.24, po zrównoważeniu 4.25. Niestety, rozwiązanie takie miałoby również swoje przełożenie na jakość syntezy, ponieważ głos byłby pozbawiony naturalności brzmienia. Przypominałoby ono brzmienie syntezatorów opartych na difonach. Z tego też powodu zrezygnowano z możliwości zrównoważenia melodycznego bazy. Autor uważa, że najlepszym rozwiązaniem tego problemu jest znalezienie profesjonalnego mówcy, potrafiącym w poprawny sposób akcentować zdania oraz wypowiadać je bez dodatkowych emocji.



Rys. 4.23 Fragment kontur melodyczny zdania „Aby znaleźć wreszcie tę ostateczną decyduje się na szachowy pojedynek z odzianą w czarną opończę śmiercią."



Rys. 4.24 Przedział zmian F0 dla zdania oznajmującego.



Rys. 4.25 Uproszczony kontur melodyczny w, którym usunięto z oryginalnego przebiegu lokalne zmiany nie większe niż 8 półtonów.

Kolejny problem, który może powodować gorszą jakość

syntetyzowanej mowy wynika z braku występowania w korpusie wszystkich możliwych w mowie polskiej difonów. Może również okazać się, że z powodu zbyt szybkiej wymowy mówcy niektóre difony są redukowane, przez co nie występują w korpusie, lub też są zamieniane na inne, zwykle częściej występujące. Jeśli syntezator nie może znaleźć difonu o odpowiedniej strukturze lista odpowiednich alternatyw jest wybierana na podstawie modułu Back-off. (Rozdział 2.5.3)

Pierwsza wersja głosu nie wykorzystywała modułu tagującego części mowy *(POS)* (Rozdział 2.4). Według (Clark i wsp. 2007) moduł ten może pomóc w uzyskaniu naturalnej syntezy, nie mniej jednak nie jest on wymagany w Festivalu przy pierwszych próbach nowego głosu.

Opisane badania przy zastosowaniu algorytmu genetycznego przyniosły lepsze efekty, a mowa syntetyczna brzmi bardziej naturalnie (Rozdział 6).

Przykłady syntetycznej mowy z różnych etapów pracy nad syntezatorem autor umieścił na płycie DVD dołączonej do pracy.

5 Optymalizacja funkcji kosztu w systemie syntezy mowy

W poniższym rozdziale został opisany sposób działania algorytmu ewolucyjnego. Następnie przedstawiony został algorytm ewolucyjny zrealizowany do optymalizacji funkcji kosztu. W końcowej części rozdziału opisana została metoda optymalizacji oraz zaprezentowane zostaną jej wyniki.

W systemach korpusowych istnieje kilka sposobów optymalizacji funkcji kosztu. Pierwszy ze sposobów polega na intuicyjnym dobieraniu parametrów oraz przeprowadzaniu testów percepcyjnych, które mają wyznaczyć najlepsze pod względem percepcyjnym współczynniki. Drugim sposobem jest metoda automatyczna, polegająca na trenowaniu poszczególnych wag kosztu doboru jednostki. Metoda te nie została zaimplementowana w algorytmie Multisyn, ponieważ według (Clark i wsp. 2007) zysk związany z metodami heurystycznymi w stosunku do metod automatycznych czy też manualnych nie przynosi istotnej poprawy jakości. Z drugiej strony, w cytowanej pracy parametry optymalnej funkcji kosztu w środowisku Festival zostały jednak ustalone heurystycznie. Warto podkreślić, że nie istnieją w literaturze jednoznaczne stwierdzenia, które by dyskwalifikowały metody heurystyczne oraz wskazywały na ich niską użyteczność.

Jedną z ostatnio podejmowanych heurystycznych technik stosowanych do oszacowania parametrów funkcji kosztu są metody oparte na sieciach neuronowych. Na przykład taka próba została podjęta dla języka arabskiego przez (Hamdi 2006).

Inną automatyczną metodą, zaproponowaną przez (Hunt i wsp. 1996) jest znajdowanie optymalnych wag w przestrzeni możliwych wyszukiwań. Nagrywane jest 10 zdań wzorcowych, następnie syntezuje się te same zdania z wszystkimi możliwymi wartościami wag i porównuje się ze sobą. W (Hunt i

wsp. 1996) testowanych było od 3 do 5 wag. (Zdania wzorcowe są porównywane z około 100000 zdań zsyntezowanych). Celem jest znalezienie takich wag, dla których zsyntezowane zdania będą różniły się jak najmniej od mowy naturalnej. Do porównania jakości syntezy stosowane są algorytmy, które pozwolą odzwierciedlić jak największe perceptualne podobieństwo między zdaniami syntezowanymi a naturalnymi. Miarą obiektywnej odległości dla poszczególnych ramek sygnału są współczynniki w przestrzeni cepstralnej Metoda ta wymaga jednak bardzo długiego czasu treningu. Według (Hunt i wsp. 1996)do wytrenowania 3-5 parametrów potrzebne jest ponad 150 godzin dla bazy danych zawierającej 40 000 jednostek akustycznych, na stacji Sun SPARC Station 20.

Kolejną metodą uzyskania automatycznych wag jest wyszukiwanie osobno wagi dla kosztu doboru jednostki oraz kosztu konkatenacji. Do porównania zdań wzorcowych i syntezowanych stosuje się linową kombinację współczynników cepstralnych i różnicę w energii sygnału w miejscu konkatenacji. Można również dodatkowo zastosować różnicę w F0 w miejscu łączonych segmentów. Wagi kosztu doboru jednostki uzyskiwane są przez zastosowanie funkcji wyliczającej obiektywnej odległości oraz wielokrotnej liniowej regresji. Proces treningu sprowadza się do wyliczenia różnic akustycznych pomiędzy jednostką docelową a wszystkimi wystąpieniami tego samej głoski w bazie oraz do wybrania n-najlepszych (np. 20) kandydatów. Następnie określa się koszty składowe dla jednostki docelowej i najlepszych kandydatów. Zbiera się obiektywne odległości pod-kosztów dla wszystkich docelowych segmentów oraz n-najlepszych kandydatów. Stosuje się liniową regresję do predykcji obiektywnych odległości poprzez liniowe ważenie docelowych pod-kosztów. Następnie stosuje się wagi uzyskane na drodze liniowej regresji jako wagi dla docelowych pod-kosztów dla wybranego zbioru głosek. (Hunt i wsp. 1996)

Celem treningu jest uzyskanie wag dla docelowych składowych pod-kosztów. W ten sposób wybrane zostaną segmenty podobne do tych, które byłyby wskazane przez funkcję kosztu. Zaletą tej metody jest wyznaczanie wag dla różnych klas głosek pogrupowanych pod względem fonematycznym oraz prozodycznym przy zdecydowanie większej wydajności w stosunku do

metody poprzednio opisanej (ok. 10 h na tej samej platformie sprzętowej). Dodatkowo możliwe jest wytrenowanie większej ilości wag w zależności od potrzeby (w poprzedniej metodzie tylko 3-5).(Hunt i wsp. 1996)

Zastosowanie algorytmów ewolucyjnych w syntezie mowy opisano w (Rozdziale 5.2)

5.1 Algorytm ewolucyjny

Algorytm ewolucyjny jest rodzajem algorytmu przeszukującego przestrzeń alternatywnych rozwiązań problemu w celu określenia najlepszych rozwiązań. Algorytm ewolucyjny ma swoje zastosowanie przy poszukiwaniu rozwiązania problemu, którego nie da się rozwiązać w linowy sposób.

Schemat działania algorytmu ewolucyjnego jest następujący:

- generowanie losowej populacji
- oszacowanie funkcji przystosowania
- tworzenie kolejnych populacji do momentu spełnienia warunki końcowego, przez zastosowanie operatorów genetycznych takich jak:
    - selekcja
    - krzyżowanie
    - mutacja
    - elitarność
    - reprodukcja

Standardowy typ algorytmu ewolucyjnego przeszukuje dużą przestrzeń równolegle w wielu miejscach, następuje to w wyniku generowania populacji startowej o losowych parametrach. W drugim kroku algorytmu ewolucyjnego dla każdego rozwiązania oszacowywana jest funkcja *fitness* (przystosowania). Funkcja ta jest miarą przystosowania, dzięki której pewne osobniki mają szanse na przeżycie do następnej generacji. Następnie tworzona jest kolejna generacja poprzez zastosowanie modyfikatorów genetycznych selekcji, krzyżowania, mutacji oraz elitarności. Operator selekcji wybiera chromosomy, które będą brały udział w tworzeniu potomków na następnego pokolenia. Wybór ten odbywa się zgodnie z zasadą naturalnej selekcji, zatem największą

szansę na wybranie mają chromosomy o największej wartości funkcji przystosowania.

Wyróżnia się cztery metody selekcji osobników:

- metoda ruletki - prawdopodobieństwo wybrania osobnika jest równe jego wartości przystosowania,
- selekcja blokowa - usunięciu z populacji N najgorszych osobników i podstawieniu w ich miejsce osobników najsilniejszych,
- selekcja turniejowa - metoda niedeterministyczna, polegającą na wybraniu losowych N osobników, tworzących turniej. M (określona liczba) zwycięzców przechodzą do populacji potomnej. Etapy są powtarzane, aż do uzyskania przez populację określonego rozmiaru.
- metody rankingowe – tworzy się ranking osobników na podstawie funkcji oceny osobnika, prawdopodobieństwo wybrania osobnika jest zależne, od jego pozycji w rankingu, przy czym pierwsi na liście maja największa szanse na reprodukcję.

Operator krzyżowania rekombinuje materiał genetyczny dwóch osobników z populacji z takim samym lub losowym prawdopodobieństwem dla każdego z nich, generując w ten sposób jednego osobnika dla nowej generacji.

Operator mutacji modyfikuje pewne cechy osobnicze w celu wygenerowania nowych rozwiązań oraz przeszukiwania przestrzeni dotychczas nie przeszukanej. Jest to losowy proces zmiany pojedynczego genu w chromosomie, polegający na transpozycji, lub dodaniu wartości genu do losowego zaburzenia lub negacji każdego z genów z niewielkim prawdopodobieństwem.

Operator elitarności zapewnia, że najzdolniejsze osobniki każdej generacji zawsze rozmnażają się w następnej generacji.

Operator reprodukcji określa w jaki sposób osobniki potomne zastąpią bieżące populację. Wyróżnia się kilka strategii zastępowania:

- (1+1)
- (1+λ)
- (μ,λ)

- (μ+λ)

Etapy dopasowywania oraz tworzenia nowej populacji są powtarzane aż do momentu uzyskania stabilnej populacji i znalezienia optymalnego rozwiązania (aż do spełnienia warunku stopu) (Hue 1997).

5.1.1 Strategie ewolucyjne

Wybór strategii ewolucyjnej ma wpływ na szybkość i jakość osiąganych wyników. Do oszacowania parametrów funkcji kosztu należało wybrać strategię elitarną oraz pozwalającą na wyszukanie najlepszego potomka zarówno spośród rodziców jak i dzieci, taką strategia jest (μ+λ).

5.1.2 Strategia (μ+ λ).

Strategia (μ+λ) jest uogólnieniem strategii (1+1). W strategii (1+1) wprowadzony został mechanizm adaptacji zasięgu mutacji. W strategii (1+1) przetwarzany jest tylko jeden chromosom bazowy X^t. W każdym kroku generowany jest nowy chromosom Y^t, który jest wynikiem mutacji X^t. Następnie wartości funkcji przystosowania w obu chromosomach są porównywane, a w kolejnej iteracji chromosomem X^{t+1} staje się ten, którego wartość funkcji przystosowania jest największa.

W strategii (μ+λ) dodatkowo dochodzi czynnik samodzielnej adaptacji zasięgu mutacji. W strategii tej zastosowano też operator krzyżowania. Przetwarzana jest startowa populacja zawierająca μ osobników. Następnie generuje się populację potomną zawierającą λ osobników. Reprodukcja przebiega w następujący sposób: Wielokrotnie powtarza się losowanie z powtórzeniami osobnika μ z populacji bazowej. W ten sposób zostanie utworzona populacja pomocnicza. Tą populację poddaje się operatorom krzyżowania i mutacji. Tak utworzona populacja jest łączona z populacją bazową. Nowa populacja jest tworzona z najlepszych osobników wybranych spośród (μ+λ) (Michalewicz 2004).

Uwzględniając istniejące strategie, autor zdecydował się wybrać

strategię (μ+λ), dokładnie (7+1) do oszacowania parametrów funkcji kosztu. Strategia (μ,λ) została odrzucona z uwagi na mały czynnik elitarności. Najbardziej elitarnymi strategiami są (1+1), (μ+λ). Elitarność oznacza fakt faworyzowania jednego, najlepszego osobnika, a zatem istnieje większa szansa na szybsze uzyskanie dobrego potomka – czyli estymowania parametrów funkcji kosztu, co przy skomplikowanym procesie odsłuchu nagrań, jest jedynym sensownym rozwiązaniem. Znak /+/ w strategii oznacza, że wybierany kandydat będzie zarówno z rodziców jak i z potomków. Strategia (1+1) została odrzucona z uwagi na konieczność długiego czasu oszacowywania i szukania właściwego rozwiązania, ponieważ w każdej iteracji z jednego osobnika powstaje tylko jeden nowy, w praktyce oznacza to konieczność przeprowadzenia kilkukrotnie większej ilości iteracji porównaniu do strategii (μ+λ).

5.2 Zastosowanie algorytmów ewolucyjnych w syntezie mowy.

Istnieje potrzeba szukania heurystycznych rozwiązań, które pozwolą na optymalizację parametrów funkcji kosztu w syntezie mowy. Wynika to z faktu stosowania dłuższych jednostek niż głoski, co zwiększa całkowitą ich liczbę (np. około 1200 difonów) i przez to utrudnia przeprowadzenie eksperymentów. Niestety istnieje bardzo mało publikacji zawierających opisy zastosowania algorytmów ewolucyjnych w syntezie mowy.

W pracy (Alías i wsp. 2003) zaprezentowano jedną z metod opartą na zastosowaniu tego typu algorytmów. Eksperyment został przeprowadzony na bazie 1520 katalońskich zdań czytanych przez profesjonalnego mówcę. Baza akustyczna zawiera około 10000 realizacji 33 głosek tego języka i do jej stworzenia nie został użyty algorytm zachłanny (Rozdział 4.1.3) W realizacji syntezy wykorzystuje się jednostki akustyczne o rozciągłości difonu, i/lub trifonu. W opisywanym eksperymencie przyjęto rozmiar populacji równy 200 jednostkom, ilość iteracji - 100, oraz prawdopodobieństwa p_c=0,3, p_m=0,003[10].

10 P_m – oznacza prawdopodobieństwo mutacji, a P_c – prawdopodobieństwo krzyżowania

Z cytowanego eksperymentu wynika, że jakość syntezy przy zastosowaniu algorytmu ewolucyjnego do oszacowania parametrów funkcji kosztu korzyści jest znacznie lepsza, niż w przypadku zastosowania, wcześniej wspomnianej, metody regresji liniowej. Warto podkreślić, że połączenie obu metod algorytmu ewolucyjnego oraz regresji liniowej przynosi gorsze efekty niż zastosowanie wyłącznie algorytmu ewolucyjnego.

Kolejną próbę optymalizacji funkcji kosztu przeprowadzono dla języka hinduskiego (Kumar 2004). Eksperyment dotyczył doboru jednostek do syntezowania zadanej sekwencji. Baza zawiera różne jednostki akustyczne (głoski, difony, sylaby). Każda z jednostek została opisana na wielu poziomach - poziomie prozodycznym i lingwistycznym. Opis ten zawiera informację o F0, czasie trwania, energii, kontekście fonetycznym pozycji głoski w sylabie oraz w wyrazie. Motywacją do zastosowania algorytmu ewolucyjnego była chęć optymalizacji czasu działania syntezatora podczas wyszukiwania najbardziej odpowiednich jednostek. Chcąc wygenerować sekwencję składającą się z 15 jednostek, poprzez wybór jednostek z bazy należy teoretycznie uwzględnić przynajmniej 10^{15} możliwych sekwencji, przy założeniu, że w bazie występuje jedynie 10 instancji każdej jednostki. Zazwyczaj jest ich w praktyce znacznie więcej. Koszt optymalizacji selekcji powinien dążyć do minimum, to w praktyce oznacza wyszukiwanie tych jednostek, które najbardziej spełniają zadane kryteria zdania wejściowego. Udowodniono w (Davis 1991), że algorytmy ewolucyjne znajdują swe zastosowanie w rozwiązywaniu i optymalizacji problemów wyszukiwania w dużych przestrzeniach poszukiwań. W wyniku przeprowadzonych eksperymentów okazało się, że zaproponowana strategia ewolucyjna jest tylko nieznacznie gorsza, niż algorytm lokalnej optymalizacji[11] stosowany również przez autorów. Stwierdzono, iż rozwijanie

11 Algorytm lokalnej optymalizacji operuje na pojedynczym stanie z pewnej przestrzeni i generuje kod do stanu sąsiedniego. Często znajduje akceptowalne rozwiązanie w dużej i nieskończonej przestrzeni rozwiązań gdy metody systematyczne zawodzą. Algorytm ten jest stosowany do optymalizacji zadań, w których celem jest znalezienie najlepszego stanu według funkcji celu. Do lokalnej optymalizacji można używać algorytmu zachłannego uwzględniającego specyfikę rozwiązywanego problemu lub ogólnego algorytmu lokalnych ulepszeń. Przykładem algorytmu lokalnej optymalizacji jest Hill-climbing, czyli algorytm wspinaczki i znajdowania szczytu. Schemat działania algorytmu jest następujący: idź w kierunku wzrastającej wartości funkcji heurystycznej, oceń stan początkowy, jeśli nie jest to cel wykonaj, utwórz nowy stan, jeśli nowy stan jest stanem celu zakończ, jeśli nowy stan jest bliżej stanu celu przyjmij go, w przeciwnym przypadku zignoruj go, jeśli nie można utworzyć nowych stanów zatrzymaj się.

tej techniki ma duże szanse na zoptymalizowanie szybkości generowania zadanych sekwencji oraz uzyskanie większej spójności, niż w przypadku stosowania algorytm lokalnej optymalizacji wbudowanego w system Festivala. W szczególności może dotyczyć sytuacji przypadku, gdy rozmiar bazy akustycznej, jak i rozmiar jednostek będzie coraz większy.

Interesujący eksperyment został przeprowadzony w przypadku syntezy mowy języka arabskiego (Hamdi i wsp. 2006). Wykorzystano w nim sieć neuronową, której wagi były optymalizowane przez algorytm genetyczny. Syntezator jest używany do konwersji arabskich niezwokalizowanych zdań na ciągi głosek. Metoda ta jest znacznie bardziej skuteczna niż synteza z reguł i pozwala na konwersję tekstu na mowę w czasie rzeczywistym.

W pracy (Tsao i wsp. 2001) udowodniono, że algorytmy genetyczne są efektywne w optymalizacji działania sieci neuronowej. Zastosowanie algorytmu genetycznego generującego rozwój populacji prowadzi do uzyskania osobników lepiej przystosowanych do danego środowiska, niż w przypadku osobników wytworzonych w toku naturalnej adaptacji.

W cytowanej już pracy (Hamdi i wsp. 2006), wykazano, że algorytmy genetyczne mogą być stosowane przy poprawianiu jakości mowy syntetycznej, ponieważ używają zakodowanych parametrów. Uwzględniana jest cała populacja a nie tylko pojedyncze osobniki. W algorytmie genetycznym używane są wartości funkcji, a nie jej pochodne. Stosowane są probabilistyczne i niedeterministyczne reguły przejść. Stosując tego typu algorytm genetyczny w zaimplementowanym środowisku Matlab uzyskano zoptymalizowanie działania sieci neuronowej, uzyskując około 95% poprawnie rozpoznanych słów dla bazy 400 słów. Podczas treningu sieci algorytm genetyczny umożliwił szybkie osiągnięcie globalnego optimum. Poprzez optymalizację algorytmem genetycznym uzyskano 380 poprawnie rozpoznanych słów. W przypadku rozpoznawania słów, przy zastosowaniu jedynie funkcji wstecznej propagacji uzyskano 336 poprawnie rozpoznanych słów. Stosowanie algorytmu genetycznego jest procesem bardzo czasochłonnym i złożonym pod względem obliczeniowym. Istnieje możliwość wykorzystania tak stworzonego syntezatora w czasie rzeczywistym, ponieważ czas potrzebny na wygenerowanie zdania jest znacznie krótszy, niż w

przypadku syntezy konkatenacyjnej czy tez systemów regułowych. (Hamdi i wsp. 2006)

5.3 Zastosowanie algorytmu ewolucyjnego do estymacji funkcji kosztu.

Uwzględniając opisane wyżej zalety algorytmów ewolucyjnych, autor zdecydował się zastosować tego typu algorytm do estymacji funkcji kosztu w korpusowej syntezie mowy. Warto zaznaczyć, że jak dotychczas nie została podjęta próba optymalizacji funkcji kosztu w algorytmie Multisyn w środowisku Festival, za pomocą algorytmu ewolucyjnego. Dotychczasowe próby optymalizacji tej funkcji dotyczyły jedynie algorytmu analizy skupień stosowanego w korpusowej syntezie mowy. W literaturze brak szczegółowych wyników badań związanych z optymalizacją funkcji kosztu, szczególnie dla syntezatorów korpusowych. Nie istnieją publikacje opisujące które parametry funkcji kosztu są najistotniejsze oraz jakie są zależności między nimi. Autor uważa, że jeśli istnieją zależności między parametrami kosztu doboru jednostki oraz kosztu konkatenacji, to w przypadku każdego tworzonego głosu w korpusowej syntezie mowy mogą one być inne. Funkcja kosztu w stworzonym i opisywanym systemie korpusowej syntezy mowy polskiej obejmuje 11 parametrów (Rozdział 3.3). W celu optymalizacji funkcji kosztu zaproponowano użycie algorytmu ewolucyjnego z zaimplementowaną strategią (μ+λ). W pierwszej iteracji algorytm generuje populację losową w postaci siedmiu osobników, z których każdy ma 11 cech, co oznacza, że optymalizowanych będzie 11 parametrów funkcji kosztu. Następnie dokonuje się syntezy zdania w 7 realizacjach w syntezatorze z parametrami wyznaczonymi przez algorytm ewolucyjny. Te zdania poddane zostają subiektywnej ocenie ekspertów. Zdanie, które zostało ocenione jako najlepsze pod względem jakości syntezy jest osobnikiem preferowanym do wygenerowania kolejnej populacji. Iteracja powtarzana jest 17 razy. Cechy poddane optymalizacji opisano w rozdziale 3.3. Należy szczególnie podkreślić, że autor przeprowadził estymację funkcji *fitness* manualnie, poprzez zliczanie głosów ekspertów. Na tej podstawie generowana była kolejna populacja.

Obecnie w algorytmach ewolucyjnych wykorzystuje się metody automatyczne, pozbawione udziału eksperta w teście przez co badania są znacznie mniej czasochłonne, ale przez to dają gorsze rezultaty.

5.4 Optymalizacja parametrów funkcji kosztu

Ważnym elementem optymalizacji funkcji kosztu było przygotowanie odpowiedniego korpusu testowego. Ustalono, że realizowany korpus powinien zawierać około 100 krótkich wypowiedzi i jednocześnie być maksymalnie zróżnicowany pod względem występujących w nim jednostek akustycznych. Korpus testowy opracowany został w programie CorpusCrt. Zdania dobrane zostały z trzech baz językowych, zawierających teksty z gazet o różnej tematyce, jednak bez wypowiedzi sejmowych oraz recenzji. Schemat tworzenia korpusu był analogiczny do opisanego w rozdziale 4. Przyjęto maksymalną długość dobieranych zdań równą 60 fonemów. Tabela 5.1. przedstawia jak zmieniała się średnia ilość wystąpień fonemów, a także ilość difonów i trifonów zależnie od przyjętej maksymalnej długości zdań.

Maksymalna długość zdań (liczba fonemów)	**Średnia ilość wystąpień dowolnego fonemu**	**Liczba difonów**	**Liczba trifonów**
100	25	925	3400
80	25	870	3200
60	**25**	**820**	**2850**
50	20	775	2500
40	15	730	2150

Tabela 5.1 Porównanie statystyk korpusu testowego zależnie od długości zdań.

Poniżej przedstawiono kilka zdań z korpusu testowego:

…

Ćwiczy się również ratunkową ewakuację poprzez wyrzutnie torpedowe.

Hydrotelefon ma kilka źródeł zasilania.

Okręt nabrał wody w ciągu dwóch minut i piętnastu sekund.

Są wybuchowi, w zdenerwowaniu krzyczą, a za chwilę już się śmieją.

Wszedł do niej jako stały ingredient.

Boksy z inkubatorami będące schronieniem nowonarodzonych.

…

Ze 100 zdań wybranych zostało 19. Zdania czasami były łączone po dwa w

celu otrzymania większej ilości trudnych elementów akustycznych. Szczególnie uwzględniono grupy spółgłoskowe. W załączniku 1 umieszczone zostały zdania wykorzystane w finalnym korpusie.

Zdania były syntezowane zgodnie z parametrami wygenerowanych osobników. Dalszy etap polegał uzyskaniu miarodajnych wyników poprzez ocenę odpowiedniej grupy ekspertów z dziedziny z fonetyki i lingwistyki. Test taki powinien być przeprowadzany dla wszystkich w podobnych warunkach akustycznych. Dlatego rozważano dwa sposoby przeprowadzenia testu oceny. Zebranie wszystkich ekspertów w studio dźwiękowym lub w specjalnie przygotowanej sali komputerowej, w przypadku syntezy w czasie rzeczywistym nie było to możliwe z powodu ilości czasu jaki musiałby być poświęcony na przygotowanie plików dźwiękowych do odsłuchu i oceny w środowisku Festival podczas każdej iteracji algorytmu. Przyjęto, że należy przygotować 20 iteracji algorytmu ewolucyjnego (17 iteracji + 3 podsumowujące). Zdecydowano, na przygotowanie testu ewaluacji funkcji kosztu w trybie on-line.

W teście wzięło udział 20 ekspertów językowych. Wśród nich znalazło się 3 ekspertów z dziedziny syntezy mowy, 3 fonetyków. Pozostałe osoby zajmują się lingwistyką na co dzień i są osłuchane z mową. Taki wybór osób do testów umożliwił uzyskanie miarodajnych wyników. Test był umieszczony na stronie internetowej (synteza.pjwstk.edu.pl) i dwa razy dziennie aktualizowany przez okres dwóch tygodni. Uczestnicy testu byli proszeni o wybór jednego najlepiej brzmiącego zdania spośród siedmiu pochodzących ze specjalnie przygotowanego korpusu testowego. Użytkownicy testu mogli dowolną ilość razy odsłuchiwać nagrania. Nagranie można było odsłuchać od dowolnego fragmentu dzięki zamieszczeniu plików dźwiękowych jako embedded objects. Pliki dźwiękowe były zapisane w formacie Microsoft Wave.

Kryterium wyboru najlepiej brzmiącego zdania – czyli osobnika w przeszukiwanej populacji, było znalezienie syntetycznego zdania z najmniejszą ilością błędów łączeniowych, prozodycznych, intonacyjnych. Największy wpływ miało prawidłowe łączenie sąsiadujących ze sobą elementów, niż dobre odtworzenie cech prozodycznych.

Tabela 5.2 prezentuje numer wygenerowanej populacji oraz numer

osobnika, który wygrał w danej sesji. Kolorem ciemno szarym zaznaczono sesje podsumowujące.

1	III
2	IV
3	VII
4	VI
5	IV
6	IV
7	V
8	VII
9	VII
10	VII
11	V
12	VII
13	III
14	II
15	VII
16	V
17	IV
18	III
19	II i V
20	IV

Tabela 5.2 Wyniki zwycięzców w poszczególnych sesjach

Tabela 5.3 przedstawia parametry poszczególnych osobników w pierwszej iteracji algorytmu. W pierwszej kolumnie zaznaczono „1" kandydata, który wygrał sesję. Kolejne kolumny oznaczają parametry funkcji kosztu. (Rozdział 3.3)

0	77,754	20,455	49,516	22,028	1,3224	22,652	22,287	52,384	28,291	3,2472	87,236
0	33,829	68,987	23,538	98,485	33,451	66,948	18,442	53,55	7,5543	31,292	38,335
1	97,025	98,611	20,446	90,952	52,436	96,497	84,371	91,688	42,133	31,62	9,128
0	6,1531	42,517	62,964	28,765	50,73	81,287	33,723	22,407	79,15	52,669	44,973
0	62,257	51,381	92,65	45,573	94,172	22,937	36,127	17,415	91,34	13,748	71,522
0	24,191	4,7965	67,136	26,53	19,181	68,756	93,497	67,144	75,681	98,895	72,563
0	76,544	27,681	59,951	52,849	83,626	88,504	33,759	60,29	39,466	8,2222	95,728

Tabela 5.3 Parametry poszczególnych osobników wygenerowanych w pierwszej iteracji.

Sesje 1-17 były sesjami podstawowymi to znaczy, że każda iteracja służyła wybraniu najlepszego potomka po to by w kolejnej uzyskać pokolenie jeszcze bardziej przystosowane, czyli posiadające bardziej zoptymalizowane

cechy funkcji kosztu.

Sesje 18-20 były sesjami podsumowującymi. W sesji 18 wybrano najlepszego kandydata spośród sesji 1-7. W sesji 19 wybrano najlepszego kandydata spośród sesji 8-14. W sesji 20 wybrano najlepszego kandydata spośród sesji 15, 16, 17, 18, oraz spośród dwóch kandydatów z sesji 19. W sesji 17 został wytypowany osobnik 4 (pierwotnie populacja 3). Potwierdzeniem wyboru tego kandydata jest sesja 20, w której znaleźli się najlepsi wygrani osobnicy z poszczególnych grup sesyjnych. W niej ponownie zwyciężył osobnik 4 z pierwotnej populacji 3.

W ten sposób autor dokonał estymacji parametrów funkcji kosztu. Ich interpretację oraz analizę przedstawiono w rozdziale 6.

6 Wyniki

Jak już wspomniano optymalizacja funkcji kosztu jest zadaniem trudnym. W rozdziale 5 opisano metody optymalizacji wag za pomocą metod automatycznych oraz heurystycznych. W niniejszym rozdziale przedstawione zostaną wyniki estymacji heurystycznej polegającej na zastosowaniu techniki ewolucyjnej, dzięki której brzmienie syntetycznej mowy uległo zasadniczej poprawie.

Z przeprowadzonych badań wynika, że najważniejszymi elementami w funkcji kosztu dla projektowanego głosu są:

- koszt pozycji w sylabie,
- koszt nieciągłości melodii,
- koszt prawego kontekstu,
- koszt nieciągłości energii,
- koszt akcentu,
- koszt złego doboru F0 dla kosztu doboru jednostki.

Uzyskanie optymalnych wyników w strategiach ewolucyjnych może wiązać się z koniecznością przeprowadzenia kilku tysięcy iteracji, a zatem proces optymalizacji powinien być maksymalnie zautomatyzowany w celu uniknięcia czasochłonnego udziału człowieka. W niniejszym badaniu nie było to możliwe, ponieważ nie można automatyzować procesu oceny nagrań realizowanych przez ekspertów językowych, dodatkowo tworzenie kolejnej populacji osobników (nagrań) jest uzależnione od wyniku wszystkich osób głosujących.

Jednak dzięki zastosowaniu strategii elitarnej, polegającej na wyborze wyłącznie jednego najlepszego osobnika, szybciej można uzyskać optymalne parametry. Strategia (μ+λ) jest strategia bardzo elitarną, typuje się w niej nie tylko jednego osobnika - zwycięzcę, lecz również wybiera się go spośród rodziców, jaki dzieci. W wyniku przeprowadzonych badań parametry najlepszej funkcji kosztu zostały uzyskane już w 3 z 20 sesji. Wadą tego rozwiązania jest większe prawdopodobieństwo uzyskania nie do końca -

optymalnych rozwiązań. Nie jest możliwe jednak realizowanie przez dłuższy okres czasu badań (zebranie grupy 20 ekspertów, dwukrotnie w ciągu dnia przez okres dwóch tygodni jest bardzo trudne).

Pewną alternatywą byłoby zastosowanie automatycznej metody oceny głosu takie jak *MUSHRA*, *PEAQ* uniezależniającej od eksperta i próba wprowadzenia większej ilość iteracji. Jednak jest mało prawdopodobne dzięki temu można było uzyskać wyniki porównywalne do metody z zastosowaniem ekspertów.

W tabeli 6.1 zaprezentowano parametry zwycięskich osobników w każdej z sesji. Pogrubioną czcionką przedstawiono parametry osobnika, który wygrał wszystkie iteracje i został ostatecznym zwycięzcą wszystkich sesji. Osobnik z trzeciej iteracji, został finalnym zwycięzcą dlatego pojawia się w tabeli w sesjach podsumowujących - 18 i 20.

	f0	energia	spektrum	akcent	l.kontekst	p.kontekst	złe f0 KDJ	pozycja w syl.	p. w słowie	p. wefra	POS
1	97	98,6	20,4	91	52,4	96,5	84,4	91,7	42,1	31,6	9,1
2	100	42,5	21,2	87,6	0	92,1	79,9	96,7	36,3	32,8	0
3	**98,4**	**81,5**	**24,8**	**77,8**	**3,8**	**97,9**	**77,1**	**100**	**46,6**	**29,4**	**0**
4	100	70	25,3	61,4	100	100	60,1	74,4	54,6	26,8	33,8
5	100	77,1	24,3	58	13,5	100	100	59,8	44,3	24,3	30
6	96,4	75	33,8	59,3	0	99,2	93,5	46,5	55,3	23,5	25
7	100	68	24,7	45,5	14,8	97,3	91,8	44,7	59,4	19,6	34,2
8	100	65,7	29,6	42,9	29,7	100	83,3	48,8	56	20,9	26,4
9	100	61,4	30,3	97,8	58,4	100	67,2	60,7	30,1	22,6	22,2
10	100	56,2	36,1	100	18,2	100	63,9	53,6	16	18,6	41,3
11	100	55,9	53,3	50,9	0	91,6	83,9	49,1	24,5	21	52,5
12	100	76,4	57,9	100	38,9	100	72,9	1,1	51,4	3,4	70,4
13	97,4	76,9	56,3	0	0	76,6	94,5	0	59,7	0	86,5
14	92,4	95,1	46	88,8	100	95	100	0	100	0	82,1
15	90,1	100	40,5	100	0	83	100	0	0	8	82,2
16	96,4	100	33	0	0	80	93,2	10,4	0	6,3	83,8
17	54,7	25	47,4	5,9	74,5	96,1	31,5	88,6	59,4	75,6	12,6
18	**98,4**	**81,5**	**24,8**	**77,8**	**3,8**	**97,9**	**77,1**	**100**	**46,6**	**29,4**	**0**
19	100	61,4	30,3	97,8	58,4	100	67,2	60,7	30,1	22,6	22,2
19v2	100	76,4	57,9	100	38,9	100	72,9	1,1	51,4	3,4	70,4
20	**98,4**	**81,5**	**24,8**	**77,8**	**3,8**	**97,9**	**77,1**	**100**	**46,6**	**29,4**	**0**
AV	95,4	72,076	35,58235	62,758	29,65882	94,42941	81,0117	48,5941	43,27	21,43	40,711
ME	100	75	33	61,4	14,8	97,3	83,9	49,1	46,6	21	33,8
STD	10,9	20,521	12,4811	35,077	35,26229	7,52826	17,9655	35,5289	24,47	17,56	30,116
MA	100	100	57,9	100	100	100	100	100	100	75,6	86,5
MIN	54,7	25	20,4	0	0	76,6	31,5	0	0	0	0

Tabela 6.1 Wartości zwycięskich osobników z każdej sesji

W tabeli 6.5 przedstawiono wartości parametrów każdego z 7 osobników w każdej sesji. Należy zwrócić uwagę, że w pierwszej iteracji algorytm zaproponował dużą rozbieżność wartość wagi F0. Mimo, iż

zastosowana strategia jest elitarna i wybrany zostaje osobnik z dużą wartością F0, to w 17 sesji można zaobserwować ponownie znaczne fluktuacje tego parametru, co daje możliwość jego ponownej estymacji, w przypadku gdyby okazał się nieistotnym parametrem. Po przeanalizowaniu tabeli 6.5 okazuje się, że dotyczy to wszystkich optymalizowanych parametrów, a zatem algorytm mimo zastosowanej strategii o dużej elitarności nie wpada w kolejnej sesji w wyznaczone uprzednio lokalne optimum.

Każdy parametr może osiągnąć wartość z przedziału <0,100>. Mimo 20 sesji jedynie 3 parametry podczas wszystkich iteracji nie osiągają minimalnej wartości: koszt F0 (5,9), koszt energii (4,8), oraz prawy kontekst (16,1). Maksymalnej wartości nie osiągają parametry: spektrum (92,7) oraz pozycja we frazie (98,9). Wartości tych parametrów stanowią kilka procent z przedziału możliwych wartości z wyjątkiem prawego kontekstu.

Optymalne parametry funkcji kosztu doboru jednostki dla języka polskiego można zestawić z wartościami parametrów dla języka angielskiego zaimplementowanych w Festivalu (Clark i wsp. 2007) (Tabela 6.2).Wszystkie parametry kosztu konkatenacji zostały ustawione domyślnie w Festivalu z wagą 1 (najmniejszą), dlatego w tabeli wstawiono w ich miejscu znak /?/, ponieważ nie jest możliwe wiarygodne odniesienie ich do odpowiednika funkcji konkatenacji dla języka polskiego. Istnieje podobieństwo w trzech kategoriach parametrów języka polskiego i angielskiego. Są to: koszt akcentu w słowie oraz w mniejszym stopniu koszt pozycji sylabie.

Nazwa parametru	Język polski	Język angielski
koszt pozycji w sylabie	I	V
koszt nieciągłości melodii	II	?
koszt prawego kontekstu,	III	VIII
koszt nieciągłości energii,	IV	?
koszt akcentu	V	III
koszt złego doboru F0 KDJ	VI	I
Koszt pozycji w słowie	VII	VI
Koszt pozycji we frazie	VIII	II
Koszt nieciągłości spektralnej	IX	?
Koszt lewego kontekstu	X	VII
POS	XI	IV

Tabela 6.2 Przedstawia porównanie parametrów funkcji kosztu w języku polskim oraz angielskim

Z pracy (Demenko i wsp. 2008 B) wynika, że koszt pozycji sylaby jest szczególnie istotny, co pokrywa się z rezultatami uzyskanymi przez autora.

Należy dodać, że wynik ten uzyskano dla całkowicie rożnych korpusów tekstowych. Należy podkreślić, że tematyka prac (Demenko i wsp. 2008 B) oraz autora zasadniczo się różni. W (Demenko i wsp. 2008 B) funkcję kosztu uszeregowano według pewnych kryteriów. Autor pracy opracował narzędzie heurystyczne oraz metodę pozwalająca na optymalizację parametrów funkcji kosztu przy zastosowaniu dowolnej bazy akustycznej. Metoda ta pozwala na wyliczenie dokładnych wartości parametrów funkcji kosztu i może być zastosowana do estymacji funkcji kosztu dla innych języków.

Interesujący wynik daje zestawienie wyników funkcji kosztu wyliczonej ze średnich każdego parametru dla każdego zwycięzcy z każdej sesji (tabela 6.3). W tabeli 6.4 znajduje się posortowane zestawienie względem priorytetów pomiędzy finalną wersją funkcji kosztu a funkcją otrzymaną ze średnich wartości. Z tabeli 6.4 wynika, że większość tych jest ze sobą skorelowana. Współczynnik korelacji parametrów funkcji koszt 0,79995.

F0	energia	spektrum	akcent	l.kontekst	p.kontekst	złe F0 dla KDJ	p.w syl.	p. w słowie	p.we frazie	POS
1	**4**	**9**	**5**	**10**	**2**	**3**	**6**	**7**	**11**	**8**
95,4	72,0	35,5	62,7	29,6	94,4	81,0	48,5	43,2	21,4	40,7

Tabela 6.3 Średnia wartość poszczególnych parametrów dla każdego ze zwycięzców każdej iteracji

Nazwa parametru		średnia wart.
koszt pozycji w sylabie	I	VI
koszt nieciągłości melodii	II	I
koszt prawego kontekstu,	III	II
koszt nieciągłości energii,	IV	IV
koszt akcentu	V	V
koszt złego doboru F0 KDJ	VI	III
Koszt pozycji w słowie	VII	VII
Koszt pozycji we frazie	VIII	XI
Koszt nieciągłości spektralnej	IX	IX
Koszt lewego kontekstu	X	X
POS	XI	VIII

Tabela 6.4 Przedstawia porównanie parametrów zoptymalizowanej funkcji kosztu w języku polskim oraz funkcji kosztu otrzymanej ze średnich wartości każdego zwycięzcy.

Pewną niewiadomą jest wartość parametru POS oraz jego mała wartość. Jest prawdopodobne, że moduł ten działał niewłaściwie i wskazywane przez niego losowe wartości nie są wiarygodne dla modułu POS w projektowanej funkcji kosztu języka polskiego.

Tabela 6.5 prezentuje wartości wszystkich osobników we wszystkich sesjach

Nr sesji	f0	energia	spektrum	akcent	l.kontekst	p.kontekst	złe f0 KDJ	pozycja w syl.	p. w słowie	p. we frazie	POS
I	77,8	20,5	49,5	22	1,3	22,7	22,3	52,4	28,3	3,2	87,2
	33,8	69	23,5	98,5	33,5	66,9	18,4	53,6	7,6	31,3	38,3
	97	98,6	20,4	91	52,4	96,5	84,4	91,7	42,1	31,6	9,1
	6,2	42,5	63	28,8	50,7	81,3	33,7	22,4	79,1	52,7	45
	62,3	51,4	92,7	45,6	94,2	22,9	36,1	17,4	91,3	13,7	71,5
	24,2	4,8	67,1	26,5	19,2	68,8	93,5	67,1	75,7	98,9	72,6
	76,5	27,7	60	52,8	83,6	88,5	33,8	60,3	39,5	8,2	95,7
II	99,3	100	18,8	100	67,8	94,9	83,3	80,2	39,2	32,1	16
	96,9	93,8	15,5	98,3	63,9	100	90,9	75,2	49	38,8	0
	98,7	100	13,7	68,7	76	100	86,2	71,5	42,1	35,3	14,5
	100	100	21,2	87,6	0	92,1	79,9	96,7	36,3	32,8	0
	96,9	98,8	21,1	73,6	0	96,2	81,8	85,4	42,8	31,5	15,2
	93,5	96	27,3	100	46,3	96,6	88,3	100	40,6	30,9	0
	99	94,1	23,6	100	48,6	96,7	99,3	80,7	19,3	31,6	40,4
III	100	100	28,3	70,4	44,2	88,8	63,1	89	28,5	38,1	4,1
	90,5	95,8	0,4	97,8	0	90,8	71	97,8	69,2	29,4	73,9
	100	100	16,2	100	0	95,2	85,1	99,4	29,2	32,6	9,2
	98,1	98,9	14,9	97,3	0	92,6	73,4	100	42,9	34	0
	98,1	99	18,6	100	0	94,7	95,3	100	28,9	34,4	10,5
	94,3	88,4	12,3	100	0	97,5	79,3	85,4	1,7	31,8	0
	98,4	81,5	24,8	77,8	3,8	97,9	77,1	100	46,6	29,4	0
IV	95,1	53,8	17	18,9	70,7	85,7	73,7	99,1	35,8	29,5	0
	97,9	82	33,9	89,6	17,4	100	77,6	100	52,9	28,5	0,5
	98,8	92,9	34,7	100	0	100	54,6	89,3	96,5	35,2	0
	100	90,9	0	100	42,6	96,7	64,9	100	29,6	30,4	2,7
	99,5	71,2	19,5	67,3	46,9	98,6	79,9	99,7	32,8	31	0
	100	70	25,3	61,4	100	100	60,1	74,4	54,6	26,8	33,8
	98	45,5	10,5	96,1	0	81,2	78,9	90,2	82,3	29,9	0
V	98,3	86,4	7,3	64,5	40,2	93,6	64,4	64,9	43,8	27,8	24,7
	100	66,9	29,1	45,3	100	99,3	64	74,2	34,8	24	52,2
	92,4	72,1	24,7	76,5	33,1	95,5	46,1	77,7	30,5	31,9	3,4
	100	77,1	24,3	58	13,5	100	100	59,8	44,3	24,3	30
	100	81,9	18,9	0	100	93	73,2	77,5	52,8	22,3	51,1
	100	64	37,3	0	100	79,2	61,4	86	50,1	15,6	37,2
	100	47,9	27,3	71,1	100	99,9	71,4	54	32,5	22,6	40,4
VI	99,7	100	24,9	69,5	56,2	100	100	75,1	62,9	23,2	5,1
	92,8	81,9	23,7	25,1	0	92,1	98,1	45,7	61,2	23,8	41,7
	100	38,9	19,3	75	0	100	100	39,4	40,6	25,9	0
	96,4	75	33,8	59,3	0	99,2	93,5	46,5	55,3	23,5	25
	100	81,3	22,8	19,1	60,8	100	91,8	57,6	51,2	25,9	48,5
	61,4	83,6	34	0	0	85,5	98,6	75,6	81,2	32,9	0
	97,2	85,7	10,7	100	0	96	59,3	74,6	28,2	23,6	13,5
VII	95,1	69,6	34,9	20	0	100	94,8	53,6	58,6	27,4	26,3
	91,4	71	33,8	16,4	42,2	97,1	100	37,6	51,1	24,1	24,9
	93,9	73,4	20	92,7	0	91,4	100	56,2	36,9	23,4	13,7
	87,5	72,7	32,4	100	100	100	78,2	41,8	49,2	26,4	27,7
	100	68	24,7	45,5	14,8	97,3	91,8	44,7	59,4	19,6	34,2
	93,8	85,1	34,8	0	42,5	99,2	100	34,4	56,2	20,3	17,5
	97,4	80,7	30,7	49,6	0	89,1	72,1	45,1	81,2	25,6	31,6
VIII	100	78	13,6	9,5	57,9	100	72,4	27,7	49,6	23,1	27,2
	100	67,4	27,1	45	54,6	97,7	85	46,2	70,3	21,2	26,9
	100	65,5	19	2,3	14,7	93,6	99,8	49,1	56	22,3	30
	99,1	61,3	17,1	48,8	42	95,6	95,9	38,3	42	17,8	28,4
	98,3	67,7	30,6	67,3	23,9	95,9	87,2	48,4	59,5	21,3	35
	100	70,1	23,7	61,1	11,6	96,6	100	35,5	75,4	20,7	26,9
	100	65,7	29,6	42,9	29,7	100	83,3	48,8	56	20,9	26,4
IX	100	64,4	30,7	57,3	27,7	89,6	75,9	42,8	61,7	22,9	22
	98,3	73,5	24,1	100	0	96,3	78,6	47,5	62	17,5	31,2
	99,5	66,4	31,7	49,9	100	94,5	79	40,6	48,1	22,2	26,6
	100	73,5	26,8	85,4	15	98,8	75,3	39,4	55	17,4	22
	96,4	55,5	16,2	29,8	100	97,9	98,9	50,8	71,1	24,2	30,9
	100	68,1	26,8	18,2	27,9	100	73	42,7	54,2	23,2	35,3
	100	61,4	30,3	97,8	58,4	100	67,2	60,7	30,1	22,6	22,2
X	99,6	59,5	19,9	62,5	67,8	98,5	63,3	48,5	38,4	24,3	33,4
	100	62,7	32,2	87,9	67,5	94,9	62,9	58,1	34,2	22	29,7
	96,4	61,2	32,6	100	100	91,5	75,6	57,8	41	24,1	53,2
	100	57,5	34,9	100	100	91,3	72,3	54,6	46,6	23,9	11,7
	100	66,4	33,3	0	18,2	100	65,9	65,2	36,8	22,8	14,1
	99	61,4	27,3	89,6	100	91	59,7	51,3	10,2	22,9	22,9
	100	56,2	36,1	100	18,2	100	63,9	53,6	16	18,6	41,3
XI	99,7	55,2	35,1	100	0	100	86,3	76,5	46,7	20,7	42,8
	100	68,5	42,7	100	0	97,1	61,3	53,6	1,9	14,7	48,8

	92	65,2	16	100	0	100	65,2	21,1	21,1	20	33,8
	97	83,1	38,8	100	0	100	71,6	40	0	10,4	22,2
	100	55,9	53,3	50,9	0	91,6	83,9	49,1	24,5	21	52,5
	99,4	57	52,8	100	17,4	95,7	65,8	49	15,7	17,9	43,7
	100	52,8	30,7	0	0	100	60,9	51,4	10,8	16,4	44,9
XII	93	28,5	28,4	0	0	62,5	68,4	25,8	0	22,4	57
	99,3	55,9	64,4	100	95,1	78,6	82,9	64,1	24,3	20,8	61,3
	99,9	62,2	55,8	100	100	93,9	94,2	57,9	0	20,5	45,3
	95,9	67,5	42,1	100	100	90,8	77,3	67,9	0	19,3	44,3
	93,4	59,4	68,5	100	0	100	98,4	54,2	41	14,8	38,5
	93,8	64,2	35,9	0	100	100	86,2	38,6	69,1	12,3	33,4
	100	76,4	57,9	100	38,9	100	72,9	1,1	51,4	3,4	70,4
XIII	100	80,6	58,9	100	0	98,7	72	21,1	48,6	5	47,2
	96,9	73,4	37	100	84,7	88	70,3	33	54,8	13,6	74,1
	97,4	76,9	56,3	0	0	76,6	94,5	0	59,7	0	86,5
	100	97	4,1	100	0	100	81,2	45,6	77,3	0,6	71,6
	100	72,5	18,1	100	0	100	100	21,2	22	0	13,3
	100	93,6	22,6	0	0	78,3	78,6	0	53,4	0	38,5
	100	91,9	59,8	100	100	65,2	45,2	20	69,8	20	100
XIV	96,8	77,9	65,5	0	0	89,4	98	9,3	58,4	0	92,1
	92,4	95,1	46	88,8	100	95	100	0	100	0	82,1
	100	47,2	84,9	0	100	16,1	0	0	34,2	0	98,8
	89,6	33,2	57,7	6,8	85,9	60,6	72	6,9	43,6	0	97,4
	98,7	100	21,3	0	0	100	89,1	0	51,5	10,2	67,4
	100	49,4	59,9	100	100	51,5	30,5	4,3	0	0	77,6
	100	92,1	50,5	0	0	55,6	100	0	33,7	19,4	100
XV	100	100	30,4	100	0	83,4	100	2,8	82,2	0	66,9
	87,6	62,1	33,8	100	0	100	58,2	0	100	12,4	84,5
	79,6	56,9	14,2	0	0	100	86,8	0	100	17,6	54,4
	100	100	45,1	100	100	78,7	100	0	100	8,7	89,6
	91,7	98	52,4	100	0	95	99,6	9,7	100	6,8	77
	75,2	100	44,7	100	100	96,5	100	3,2	100	0	87,9
	90,1	100	40,5	100	0	83	100	0	0	8	82,2
XVI	85,6	76	47,2	100	0	100	100	0	0	0	75,5
	95,5	100	33,1	100	0	94,2	78,9	16,8	19,9	10,1	82,9
	91,4	79,7	45,1	100	0	74	100	0	0	18,3	72
	86,1	100	46,9	100	91,4	100	98,6	7,1	19,2	0	86,4
	96,4	100	33	0	0	80	93,2	10,4	0	6,3	83,8
	100	75,8	43,9	100	0	31,6	100	6,4	0	6,6	79
	85,8	94,1	28,9	0	100	56	100	17,3	0	0	66,4
XVII	23,4	39,4	55,2	21	30	19,8	33	51,8	64,7	38,4	77,9
	72,3	27,8	51,2	93,3	48,5	40,6	70,6	3,5	92,8	92,5	84
	18,4	89,9	68,6	1,5	35,3	25,8	70,8	40,3	90,7	5,5	21,6
	54,7	25	47,4	5,9	74,5	96,1	31,5	88,6	59,4	75,6	12,6
	5,9	22,4	68,4	93,4	86,9	92,9	44,7	40,9	44,3	32,9	98,9
	74,5	97,6	79,7	4,5	22,7	74,2	42,7	31,9	18,8	54,9	99,6
	54,9	99,3	22,4	21,2	42,5	36,5	19,8	88,6	43,7	88,8	0,1
XVIII	97	98,6	20,4	91	52,4	96,5	84,4	91,7	42,1	31,6	9,1
	100	21,2	87,6	0	92,1	79,9	96,7	36,3	32,8	0	0
	98,4	81,5	24,8	77,8	3,8	97,9	77,1	100	46,6	29,4	0
	100	70	25,3	61,4	100	100	60,1	74,4	54,6	26,8	33,8
	100	77,1	24,3	58	13,5	100	100	59,8	44,3	24,3	30
	96,4	75	33,8	59,3	0	99,2	93,5	46,5	55,3	23,5	25
	100	68	24,7	45,5	14,8	97,3	91,8	44,7	59,4	19,6	34,2
XIX	100	65,7	29,6	42,9	29,7	100	83,3	48,8	56	20,9	26,4
	100	61,4	30,3	97,8	58,4	100	67,2	60,7	30,1	22,6	22,2
	100	56,2	36,1	100	18,2	100	63,9	53,6	16	18,6	41,3
	100	55,9	53,3	50,9	0	91,6	83,9	49,1	24,5	21	52,5
	100	76,4	57,9	100	38,9	100	72,9	1,1	51,4	3,4	70,4
	97,4	76,9	56,3	0	0	76,6	94,5	0	59,7	0	86,5
	92,4	95,1	46	88,8	100	95	100	0	100	0	82,1
XX	90,1	100	40,5	100	0	83	100	0	0	8	82,2
	96,4	100	33	0	0	80	93,2	10,4	0	6,3	83,8
	54,7	25	47,4	5,9	74,5	96,1	31,5	88,6	59,4	75,6	12,6
	98,4	81,5	24,8	77,8	3,8	97,9	77,1	100	46,6	29,4	0
	100	61,4	30,3	97,8	58,4	100	67,2	60,7	30,1	22,6	22,2
	100	76,4	57,9	100	38,9	100	72,9	1,1	51,4	3,4	70,4

Tabela 6.5 Wartości wszystkich osobników we wszystkich sesjach

W niniejszym rozdziale przedstawiono wyniki badań optymalizacji funkcji kosztu uzyskane przez zastosowanie algorytmu ewolucyjnego. Wyniki te wskazują na cechy najbardziej istotne w funkcji kosztu, określają relacje między nimi oraz opisują dokładne parametry tych cech. Uzyskane wyniki są podobne do wyników badań (Demenko i wsp. 2008 B), mimo zastosowania rożnych korpusów, co oznacza są one reprezentatywne dla języka polskiego. Dodatkowym dowodem na ich poprawność jest test *MOS* opisany w rozdziale 7.

7 Wnioski

Zastosowanie algorytmu ewolucyjnego w korpusowej syntezie mowy miało na celu uzyskanie zoptymalizowanych parametrów funkcji kosztu. Wyniki tego testu wskazują, iż strategie ewolucyjne przynoszą pożądane efekty a wygenerowane parametry dla funkcji kosztu potwierdziły to w ostatnich trzech iteracjach testu. Test MOS, przedstawiony w tym rozdziale, jest kolejnym dowodem na skuteczność wykonanych badań optymalizacyjnych. Test ten jest porównaniem 3 różnych funkcji kosztu, ocenia jakość sygnału syntezy mowy uzyskanej na drodze resyntezy, oraz nagrań pochodzących z bazy akustycznej.

7.1 Ewaluacja systemu w teście MOS

MOS *(Mean Opinion Score)* jest subiektywną metodą stosowana do testowania jakości dźwięku (mowy) np. w telefonii czy też w systemach syntezy mowy (ITU 1996 , Viswanathan i wsp. 2005). Mierzone są dwie cechy sygnału mowy zrozumiałość oraz naturalność. Ocena podawana jest w skali od 1 do 5:

- 1 - zła
- 2 - słaba
- 3 - średnia
- 4 - dobra
- 5 – znakomita

Wynikiem testu jest średnia arytmetyczna poszczególnych ocen.

W przygotowanym teście wzięło udział 28 studentów studiów magisterskich na kierunku informatyka, znających zagadnienia dotyczące syntezy mowy, fonetyki języka polskiego, transkrypcji fonetycznej oraz posiadających wiedzę związaną przetwarzaniem języka naturalnego. Test został zrealizowany w tych samych warunkach odsłuchowych przy

zapewnieniu studentom identycznego sprzętu odsłuchującego w postaci słuchawek Philips HP1900.Test został podzielony na 5 części. W każdej z nich zostało przygotowanych 5 zdań do odsłuchu. Test został przygotowany w wersji on-line i zamieszczony pod adresem www.synteza.pjwstk.edu.pl/mos.html.

Pierwszą część testu stanowiły pliki z nagrane w ramach korpusu. Kryterium wyboru plików było znalezienie jak najbardziej bogatych fonetycznie i jednocześnie trudnych do wymówienia zdań. Ponieważ korpus był kilkukrotnie optymalizowany istniała pewność, iż wybrane zdania będą reprezentatywne dla języka polskiego.

Druga część testu polegała na resyntezie zdań z korpusu. Zsyntezowane zostały zdania, odpowiednio ze znakami przystankowymi. Do pierwszej i drugiej części testu zostały wybrane zdania 1-5 (Tabela 7.1), z korpusu opisanego w rozdziale 5.4

Do trzeciej, czwartej i piątej części testu zostało wybranych kolejnych 5 zdań, są to zdania 6-10 z tabeli 7.1.

Poniżej znajduję się zdania wybrane do syntezy:

Chyba najwyższy czas, by przestać szufladkować geograficznie scenę jazzową na my i oni.
Zapewne nawet jej nie znał, zwłaszcza że Wharton jest literacką gwiazdą chyba tylko u nas.
Może to był brat i siostra, jedno o przerażonych, szeroko otwartych oczach i otwartym pyszczku, jakby skomlało.
Pan poseł Potulski tak rozsmakował się w definicji lekceważki że przytaczał ją dwukrotnie
Proszę łaskawie jeszcze raz wcisnąć dowolny przycisk w urządzeniu do głosowania
Wystarczyło kilka chudszych lat i sny o potędze runęły, ponieważ firmy fonograficzne uwierzyły w swoją siłę i zaczęły kreować rynek według własnych wyobrażeń.
Sto lat minęło od pojawienia się na ulicach Warszawy pierwszych konnych tramwajów.
Wczoraj byłem na pogrzebie wielkiego boksera i wspaniałego człowieka, którego znałem od kilku lat i z którym byłem bardzo
Ustawa o zakładach fryzjerskich z dwutysięcznego czwartego roku wyraźnie mówi, że panie muszą być w odzieży ochronnej, którą da się wydezynfekować.
W Zimbabwe zaobserwowano dwa stare wypędzone ze stada lwy jeden z nich polując kiedyś na guźca utknął w norze.

Tabela 7.1 Korpus użyty do testu MOS

Trzecia część testu polegała na syntezie z domyślną funkcją kosztu w Festivalu, czwarta z najgorszymi ustawieniami wytypowanymi na etapie estymacji funkcji kosztu za pomocą algorytmu ewolucyjnego. Kryterium wyboru parametrów była jakość generowanego sygnału oraz sposób głosowania ekspertów. Piąta część testu zawierała pliki dźwiękowe uzyskane przez zastosowanie parametrów, wytypowanych jako najlepszą funkcję kosztu.

W tabeli 7.2 zaprezentowano sposób głosowania poszczególnych ekspertów dla każdego zdania oraz średnie wartości z każdej części testu. W pierwszej kolumnie „WAV" znajduje się ocena plików dźwiękowych w formacie wav nagranych przez autora bazy akustycznej, a zatem jest to ocena jakości głosu lektora. Średnia ocena 4,6 wskazuje, iż mówcy ocenili głos dość wysoko. Jednocześnie można przyjąć następujące założenie, że jest to maksymalna ocena jaką mógłby otrzymać idealny syntezator mowy skonstruowany na głosie autora. Sytuacja taka oczywiście w rzeczywistości nie jest możliwa do spełnienia. Druga kolumna zawiera ocenę resyntezy zdań, to znaczy syntezowane są bogate fonetycznie, trudne do wymówienia zdania z korpusu. Tak wygenerowany sygnał musi oznaczać utratę jakości. Eksperci ocenili średnio jakość syntezy na poziomie 3,793 co jest rezultatem dobrym. Trzecia kolumna reprezentuje oceny domyślnej funkcji kosztu w Festivalu, czwarta najgorszą funkcję kosztu otrzymaną w wyniku optymalizacji algorytmem ewolucyjnym, piąta najlepszą funkcję kosztu wyodrębnioną w procesie optymalizacyjnym.

Z porównania trzech funkcji kosztu wynika, że proces estymacji parametrów przyniósł oczekiwany efekt. Eksperci ocenili domyślną funkcje kosztu na 2,185, najgorszą funkcje kosztu uzyskana podczas estymacji parametrów na 1,97. Przy najlepszej funkcji kosztu wartość średniej wzrosła do 2,7111. Wynik tego testu potwierdza uzyskane rezultaty optymalizacji funkcji kosztu i oznacza, że funkcja ta nie tylko daje się estymować za pomocą algorytmu ewolucyjnego ale również proces ten przynosi satysfakcjonujące efekty i poprawia jakość mowy w syntezie korpusowej dla języka polskiego. Różnica pomiędzy wynikiem z parametrami domyślnej funkcji kosztu a estymowanymi za pomocą algorytmu ewolucyjnego oznacza, że wartości funkcji kosztu będą inne dla języka polskiego oraz inne dla angielskiego.

	WAV					Resynteza					Domyślna f. kosztu					Najgorsza f. kosztu					Najlepsza f. kosztu				
	4	4	4	4	4	4	4	4	4	4	1	1	1	1	1	2	2	2	2	2	3	3	2	3	3
	5	5	5	5	5	5	5	4	4	5	3	3	4	4	3	4	3	4	3	3	2	3	3	3	3
	4	4	2	5	4	5	3	1	2	5	1	2	2	2	1	1	1	1	1	2	2	2	3	3	3
	5	5	5	5	5	4	5	3	3	5	2	3	4	4	1	2	1	1	1	1	2	2	3	3	4
	5	4	5	5	5	5	4	3	2	4	2	3	3	2	2	2	2	2	3	2	2	3	4	3	3
	5	5	5	5	5	5	5	4	3	4	3	3	4	4	3	2	2	2	2	2	2	3	3	3	3
	5	5	5	5	5	5	5	4	5	5	3	3	2	2	2	2	2	2	2	2	2	2	2	2	2
	5	5	5	5	5	4	4	2	3	4	1	2	2	2	2	2	1	2	2	2	2	3	3	3	4
	3	4	5	5	3	3	3	2	3	4	1	2	3	3	2	1	3	3	4	5	2	4	4	5	4
	5	5	5	5	5	4	4	3	4	4	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2
	4	5	4	5	5	4	5	3	3	5	2	3	3	3	3	3	2	3	3	3	3	3	4	4	3
	4	4	3	5	5	3	4	3	3	5	2	3	4	3	2	1	1	3	3	3	2	3	3	3	3
	5	5	5	5	5	4	4	3	4	5	4	3	3	3	3	3	2	3	3	3	4	4	4	5	4
	5	5	5	5	5	5	5	4	4	5	2	3	3	3	3	2	2	3	3	4	1	3	3	4	4
	5	5	5	5	5	4	5	3	3	5	1	2	2	2	2	1	1	2	2	1	1	3	3	3	3
	5	5	5	5	5	4	5	4	4	5	2	3	3	3	2	2	1	2	2	2	2	3	3	3	2
	3	3	5	3	3	3	4	2	2	5	1	2	1	1	1	1	1	1	2	1	2	3	2	2	1
	4	5	3	3	5	4	5	2	2	5	2	1	2	2	1	1	1	2	3	3	2	3	3	4	3
	5	5	5	5	5	4	4	3	3	5	2	2	3	2	3	2	1	2	3	3	3	3	3	3	3
	5	5	4	4	5	4	2	2	3	5	1	2	3	2	2	2	1	2	3	3	3	3	3	3	2
	4	5	3	5	5	4	5	2	3	5	1	2	2	2	2	1	1	1	2	2	2	3	2	2	2
	5	5	5	5	5	5	5	3	4	5	2	3	3	3	2	2	2	2	2	2	2	3	3	3	2
	5	5	4	5	4	4	4	2	3	4	1	1	2	2	1	1	1	2	2	2	1	3	3	3	2
	4	5	4	5	4	3	4	2	2	4	1	1	2	2	2	1	1	1	2	2	2	2	3	3	2
	5	4	5	5	5	3	4	3	3	5	2	2	2	2	1	2	1	2	1	1	1	2	3	2	3
	4	5	4	4	4	3	4	2	3	3	2	3	3	2	2	2	2	2	2	2	1	2	3	2	2
	5	4	4	5	5	5	5	3	3	5	2	2	3	2	2	3	2	2	2	2	3	2	3	3	3
AVG	4,556	4,63	4,407	4,741	4,667	4,074	4,296	2,815	3,148	4,63	1,778	2,259	2,593	2,37	1,926	1,815	1,519	2,037	2,222	2,259	2,037	2,741	2,963	3,037	2,778
AVG (AVG)					4,6					3,793					2,185					1,97					2,711
MEDIANA	5	5	5	5	5	4	4	3	3	5	2	2	3	2	2	2	1	2	2	2	2	3	3	3	3
AVG(MEDIANY)					4,8					3,8					2,2					1,8					2,8
STDEV	0,641	0,565	0,844	0,594	0,62	0,73	0,775	0,834	0,77	0,565	0,801	0,764	0,888	0,839	0,73	0,786	0,643	0,759	0,751	0,944	0,759	0,656	0,587	0,808	0,801
AVG(STDEV)					0,653					0,735					0,804					0,777					0,722

Tabela 7.2 Sposób głosowania poszczególnych uczestników

Można próbować porównać uzyskane wartości do komercyjnego systemu korpusowej syntezy mowy IVONA stworzonego przez firmę IVOSOFTWARE i przedstawionego w konkursie Blizzard Challenge. Należy dodać, że oceny dotyczą systemu korpusowej mowy dla języka angielskiego. Nie istnieje jednak bezpośrednie porównanie lub zestawienie dla głosu polskiego. W teście MOS komercyjny system korpusowej syntezy mowy w ramach konkursu Blizzard Challenge w latach 2006-2007 uzyskał następujące oceny:

- *Blizzard 2007* — *3,9*
- *Blizzard 2006* — *3,6 (Kaszczuk i wsp. 2007)*
- *Autorski system resynteza* — *3,8*
- *Autorski system synteza* — *2,7*

Tabela 7.3 przedstawia oceny wszystkich systemów korpusowej syntezy mowy uzyskane w ramach konkursu Blizzard Challenge.

System	Średnio	Studenci angielscy	Ochotnicy	Eksperci	Studenci amerykańscy
A	3,8	3,4	3,6	4,2	3,4
B	3,0	2,7	2,9	3,1	3,1
C	3,2	3,0	2,9	3,4	2,9
D	2,6	2,2	2,3	2,9	2,1
E	3,0	3,0	2,8	3,1	2,5
F	1,5	1,6	1,4	1,4	1,7
G	1,4	1,5	1,4	1,4	1,2
H	3,2	3,0	3,0	3,4	3,1
LEKTOR	4,7	4,6	4,7	4,8	4,3
J	3,4	3,1	3,5	3,4	3,5
K	3,6	3,4	3,5	3,7	3,2
L	1,3	1,3	1,4	1,3	1,1
M	3,0	2,5	2,6	3,4	2,8
N	2,7	2,2	2,7	2,9	2,3
O	2,5	3,3	3,5	3,7	3,2
IVONA	3,9	3,6	3,8	4,1	3,7
Q	2,5	2,3	2,4	2,5	2,4

Tabela 7.3 Wyniku testu MOS w konkursie Blizzard Challenge 2007

Porównując uzyskane wartości z opracowanym samodzielnie systemem autor uważa, iż istnieje duże prawdopodobieństwo uzyskania podobnych wyników w wyniku poprawy pewnych słabych punktów systemu. W nowej wersji syntezatora należałoby uniezależnić się od środowiska Festival przepisując wiele modułów i tworząc własny system. W wyniku takiego przygotowania poprawić można moduł transkrypcji fonetycznej tworząc go od początku. W praktyce nie oznacza to rezygnacji całkowitej z Festivala. Wiele algorytmów w metasystemie jest gotowych i należy z nich skorzystać parametryzując sygnał dźwiękowy, czy też ekstrahując informacje lingwistyczne. Niestety, jak przyznają programiści Festivala, posiada on wiele błędów przez, które praca często jest utrudniona, dlatego jest on uważany za środowisko badawcze i eksperymentalne. Wykorzystanie Festivala jako systemu generującego syntezę w czasie rzeczywistym staje się niemożliwe. Kolejnym ważnym punktem, od którego będzie zależała jakość systemu będzie odpowiednio dobrany profesjonalny głos. Autor uważa, iż poprzez fakt doboru nieprofesjonalnego głosu napotkał później w realizacji syntezy wiele problemów. Problemy te nie zaistniałyby w takiej skali w przypadku doboru lepszego mówcy. Należy dodać, iż w wyniku testu MOS ocena głosu zapewne byłaby wyższa.

W początkowym etapie projektowania systemu największym problemem były znaczne fluktuacje F0 w syntezowanych zdaniach. W rozdziale 4.4 opisano problem nadmiernej intonacji oraz specyfiki wymowy mówcy. Problem został tylko częściowo rozwiązany poprzez dodanie modułu intonacyjnego. W rzeczywistości należałoby ponownie nagrać bazę akustyczną ze zmniejszoną prędkością mówienia oraz z mniejszymi fluktuacjami F0 w zdaniach.

Chcąc otrzymać większą naturalność syntezy należałoby również zmodyfikować korpus. Część korpusu dotyczącą przemówień sejmowych należałoby zastąpić tekstami gazetowymi lub tekstami zawierającymi różnego rodzaju wypowiedzi z życia codziennego, w ten sposób poprawie uległa by jakość zdań wymienionej dziedziny.

7.2 Wady i zalety opracowanego systemu

Do mocnych punktów systemu należy opracowana technologia. Pozwala ona na tym etapie prowadzić bardziej zaawansowane badania być może zmierzające do zrealizowania pełnego systemu korpusowej syntezy mowy działającego w trybie on-line oraz czasie rzeczywistym. Zauważano również, iż jakość segmentacji jest na poziomie wystarczająco dobrym. Opracowana została również innowacyjna technika poprawiająca jakość segmentacji. Najistotniejszym punktem jest jednak sposób doboru parametrów dla funkcji kosztu. Należy dodać, iż zaprezentowane w tej pracy parametry funkcji kosztu będą różne dla innych głosów w języku polskim. W przypadku głosu z mniejszymi fluktuacjami F0, parametr kosztu F0 mógłby się zmniejszyć. Technologia optymalizacji przez zastosowanie algorytmu ewolucyjnego pozwala nie tylko na dobranie właściwych parametrów, ale przede wszystkim na ich poprawne oszacowanie. Podczas pierwszych prób pracy z funkcją kosztu próbowano ustalać parametry kierując się pewnymi przesłankami związanymi ze specyfiką języka polskiego a zwłaszcza z realizacją akcentu. Niestety, nie przyniosło to pozytywnych rezultatów, a generowana mowa była znacznie gorszej jakości. (Przykłady dołączone do płyty DVD).

Do zrealizowania i optymalizacji funkcji kosztu niezbędne było stworzenie w pełni funkcjonującego systemu korpusowej syntezy mowy w środowisku Festival. Aplikacja została zaimplementowana w systemie unixowym w dystrybucji Debian, oraz skompilowana do środowiska Windows. Aplikacja jest dostępna pod adresem:

http://syntezamowy.pjwstk.edu.pl/korpus.html

Proces tworzenia systemu obejmował przygotowanie korpusu, nagrania oraz ich segmentację. Kolejnym etapem było dostosowanie istniejących oraz przygotowanie nowych modułów lingwistycznych. Tak przygotowana aplikacja pozwoliła na realizację optymalizacji funkcji kosztu za pomocą algorytmu ewolucyjnego. Zastosowana strategia ewolucyjna oraz przeprowadzone badania wskazują, iż funkcję kosztu można optymalizować za

pomocą metod heurystycznych, a proces optymalizacji funkcji kosztu ma wpływ na jakość syntezy korpusowej. Ważnym etapem pracy jest wybór odpowiedniego mówcy oraz jakość rejestracji bazy akustycznej. Proces ten ma duży wpływ na finalną jakość generowanej mowy. Wynikiem zakończonych badań jest w pełni działający korpusowy syntezator mowy generujący satysfakcjonującą i bliską naturalnej mowę.

Literatura

Adell J., Bonafonte A. (2004) *Towards phone segmentation for concatenative speech synthesis*,Proc. 5thISCA ITRW on Speech Synthesis (SSW5), Pittsburgh, PA, USA.

Alías F., Llorá X. (2003) *Evolutionary weight tuning based on diphone pairs for unit selection speech synthesis*, Proc. EuroSpeech, vol. 2, Geneve, Switzerland, pp. 1333–1336.

Anderson M., Pierrehumbert J., Liberman M. (1984) *Synthesis by rule of English intonation patterns*, Proc. ICASSP'84, pp. 2.8.1–2.8.4.

Bailador A. (1998) *CorpusCrt. Technical repor*t, Polytechnic University of Catalonia (UPC).

Bellman R. (1954) *The theory of dynamic programming*, Bulletin of the American Mathematical Society, **60**, 503-515, 1954.

Benello J., Mackie A.W., Anderson J., (1989) *Syntactic category disambiguation with neural networks*, Computer Speech and Language, n3, pp. 203-217.

Beutnagel M., Conkie A. (1999) *Interaction of units a unit selection database*, Proc. European Conference on Speech Communication and Technology, vol. 3, pp. 1063–1066.

Bjørkan I., Svendsen T., Farner S. (2005) *Comparing Spectral Distance Measures for Join Cost Optimization in Concatenative Speech Synthesis*, Proc. Interspeech pp. 2577-2580.

Black A., (2006 B) *CLUSTERGEN: A Statistical Parametric Synthesizer using Trajectory Modeling*, Interspeech 2006 - ICSLP, Pittsburgh, PA, pp. 1762-1765

Black A., (2006 C) *Statistical Parametric Speech Synthesis*, The Blizzard Challenge 2006 CMU Entry.

Black A., Bennett C., Kominek J., Langner B., Prahallad K., Toth A. (2008) *CMU Blizzard 2008: Optimally using a large database for unit selection synthesis*, Blizzard Challenge 2008, Brisbane, Australia.

Black A., Campbell N. (1995) *Optimising selection of units from speech databases for concatenative synthesis*, Proc. Eurospeech95, volume 1, pp. 581-584, Madrid, Spain.

Black A., Hunt A. (1996) *Generating F0 contours from ToBI labels using linear regression*, Proc. of ICSLP 96, Philadelphia, pp. 1385–1388.

Black A., Lenzo K.(2001) *Optimal data selection for unit selection synthesis, Proc.* 4th ISCA Workshop on Speech Synthesis, pp. 63–67.

Black A., Lenzo K. (2000) *Limited domain synthesis*, Speech Communication archive Volume 49, Issue 4 (April 2007) table of pp. 317-330 ISSN:0167-6393

Black A., Lenzo K. (2006) *Buliding Synthetic Voices* (1996-2006) dokumentacja system Festival, www.festvox.org.

Black A., Schultz T.(2006 A) *Speaker Clustering for Multilingual Synthesis,* Proc. of the ISCA Tutorial and Research Workshop on Multilingual Speech and Language Processing, Stellenbosch, South Africa, April 9-11, 2006.

Black A., Taylor P. (1998) *Festival Speech Synthesis System: system documentation,* Technical Report HCRC/TR–83, University of Edinburgh, Human Communication Research Centre.

Boersma P. (2001) *Praat, a system for doing phonetics by computer,* Glot international, 5(9/10):341–345.

Borden G. J., Harris K, Raphael L. (1994) *Speech Science Primer: Physiology, Acoustics, and Perception of Speech (Hardcover)*, Lippincott Williams & Wilkins

Bozkurt B., Dutoit T., Ozturk O., (2003) *Text Design For TTS Speech Corpus Building Using A Modified Greedy Selection*, Proc. Eurospeech, Geneva 2003, pp 277-280.

Briony J., Williams B. *Text-to-speech synthesis for Welsh and Welsh English,* Proc. 1113-1117, Eurospeech '95, Madrid, 1995

Clark R., Richmond K., King S. (2004) *Festival2 - Build your own general purpose unit selection speech synthesizer,* Proc. 5th ISCA Speech Synthesis Workshop pp. 173-178, 14th-16th June 2004, Carnegie Mellon University Pittsburgh.

Clark R., Richmond K., King S. (2005) *Multisyn Voices from ARCTIC Data for the Blizzard Challenge*, CSTR, The University of Edinburgh, Edinburgh, Proc. Interspeech 2005 pp. 101-104

Clark R., Richmond K., King S. (2007) *Multisyn: Open-domain unit selection for the Festival speech synthesis system,* Speech Communication, 49(4):317-330,.

Clark R., Richmond K., Strom V. (2006) *Multisyn voices for the Blizzard Challenge 2006,*Proc. Blizzard Challenge Workshop (terspeech Satellite), Pittsburgh, USA, September 2006.

Conkie A. (1999) *A robust unit selection system for speech synthesis*, The Journal of the Acoustical Society of America, Volume 105, Issue 2, February 1999, p.978

Coorman G., Fackrell J., Rutten P., Coile B. (2000) *Segment selection the L&h Realspeak laboratory TTS system*, Proc. ICSLP-2000, vol.2, 395-398.

Davis L. (1991) *Handbook of Genetic Algorithm*, Van Nostrand Reinhold, New York, 1991.

Demenko G, Möbius B, Klessa K (2008 B) *The design of Polish Speech Corpus for Unit Selection Speech Synthesis*, Speech and Language Technology Volume 11, pp. 85-92, Poznan 2008.

Demenko G., (1999) *Analiza cech suprasegmentalnych języka polskiego na potrzeby technologii mowy,* Poznań: Wydawnictwo Naukowe UAM.

Demenko G., Bachan J., Möbius B., Klessa K., Szymanski M., Grocholewski S. (2008) *Development and Evaluation of Polish Speech Corpus for Unit Selection Speech Synthesis Systems,* To appear in: Interspeech 2008 Proc. of Interspeech 2008 (Brisbane)

Demenko G., Klessa K., Szymański M., Bachan J. (2007) *The design of Polish speech corpora for speech synthesis in BOSS system,* Mat.XII Sympozjum Podstawowe Problemy Energoelektroniki, Elektromechaniki i Mechatroniki (PPEEm'2007), Wisła, Poland, pp. 253-258.

Demenko G., Wagner A. (2007) *Prosody annotation for unit selection text-to-speech synthesis,* Archives of acoustics, 32(1):.25-40

Donovan R. (1996) *Trainable Speech Synthesis,* PhD. Thesis. Cambridge University Engineering Department, England.

Dutoit T. (1994) *High Quality Text-To-Speech Synthesis: A Comparison of Four Candidate Algorithms*, Proc.ICASSP'94, Adelaide, Australia, 19-22 April 1994, vol. 1, pp. 565-568.

Dutoit T. (1997) *An introduction to Text-To-Speech Synthesis*, Kluwer Academic Publishers, Dordrecht, 320 pp., ISBN 0-7923-4498-7.

Ellbogen T., Steffen A., Schiel F. (2004) *The BITS Speech Synthesis Corpus for German*, Proc. of the IV. International Conference on Language Resources and Evaluation, pp 2091-2094.

Fant G. (1970) *Acoustic Theory of Speech Production*, The Hagues, Mounton.

Gersho A., Gray R. M. (1991) *Vector Quantization and Signal Compression,* (The Springer International Series in Engineering and Computer Science) (Hardcover)

Gray A.H., (1976) Markel J.D. *Distance Measures for Speech Processing*, IEEE Trans. on vol. 24, Issue 5 ASSP, pp. 380-391,Oct 1976

Gray H. (2008) *Gray's Anatomy: The Anatomical Basis of Clinical Practice*, 40th edition, 1576 pages, Churchill-Livingstone, Elsevier. ISBN 978-0-443-06684-9.

Grice M., Baumann S. (2002),*Deutsche Intonation und GToBI.* Linguistische Berichte 191. 267-298.

Gubrynowicz R. (2004) *Wykład Podstawy fonetyki akustycznej (PFA),*PJWSTK.

Hałupka A. (2004) *Intonation modelling for speech synthesis applications*, Poznań 2004, praca magisterska.

Hamdi R., Bedda M. (2006) *Arabic Speech Synthesis Using Optimized Neural Networks with Genetic Algorithms*, Asian Journal of information technology 5(7):686-690.

Hirschberg J. (1991) *Using text analysis to predict intonational boundaries*, Proc. EUROSPEECH-1991, 1275-1278.

Hirst D. (1994) *The symbolic coding of fundamental frequency curves : from acoustics to phonology*, Proc. of International Symposium on Prosody. Yokohama.

Hirst D. (1999) *The symbolic coding of segmental duration and tonal alignment: an extension to the INTSINT system,* Proc. of EUROSPEECH'99, pages 1639–1642.

Huang X., Acero A., Hon H. (2001) *Spoken Language Processing: A Guide to Theory*, Algorithm and System Development All of Microsoft Research, Redmond, Washington ISBN-10: 0130226165, ISBN-13:9780130226167, Wydawnictwo Prentice Hall 2001.

Hue X. (1997) *Genetic Algorithms for Optimization Background and Applications*, Edinburgh.

Hunt A., Black A. (1996) *Unit selection in a concatenative speech synthesis system using a large speech database,* Proc. of the ICASSP 1996, Atlanta, USA, Vol. 1, pp. 373–376.

IPA (1999) *Handbook of the international Phonetic Association : A Guide to the Use of the international Phonetic Alphabet* (Paperback) ISBN 0521 652367.

IVO (2005), *European Patent Office Description of EP1501075*, wniosek patentowy.

IVO Software (2008)http://www.ivo.pl/wydarzenia/wydarzenia.html

Janicki A. (2004) *Selected Methods of Quality Improvement In Concatenative Speech Synthesis For The Polish Language*, Rozprawa doktorska.

Karpiński M. (2001) *Intonacyjna baza danych dla języka polskiego, Sprawozdanie merytoryczne z przebiegu projektu badawczego KBN (H01 D 011 18) przy współpracy Wiktora Jassema i Janusza Kleśty*

Kaszczuk M., Osowski L. (2007) *The IVO Software Blizzard 2007 Entry: Improving Ivona Speech Synthesis System*.

Klabbers E., Stöber K., Veldhuis R., Wagner P., Breuer S. (2001 B) *Speech synthesis development made easy: The Bonn Open Synthesis System*, Eurospeech 2001, Aalborg,

Klabbers E., van Santen J. (2004) *Clustering of foot-based pitch contours in expressive speech*, Fifth ISCA ITRW on Speech Synthesis (SSW5), Pittsburgh, PA, USA.

Klabbers E., Veldhuis R. (1998) *On the reduction of concatenation artefacts in diphone synthesis,* Proc. ICSLP, vol. 6, (Sydney, Australia), pp. 1983–1986, 1998.

Klabbers E., Veldhuis R. (2001) *Reducing Audible Spectral Discontinuities,* IEEE Transactions on Speech and Audio Processing, vol. 9, nr. 1, January 2001, p39-51.

Klatt D.(1987) *Review of text-to-speech conversion for English*, J. Acoust. Soc. Am. 82, 1987.

Kominek J., Black A. (2003) *CMU ARCTIC databases for speech synthesis*,CMU-LTI-03-177 Ver. 0.95 Language Technologies institute School of Computer Science Carnegie Mellon University 5000 Forbes Ave., Pittsburgh, PA 15213 www.lti.cs.cmu.edu

Kominek J., Black A. (2004) *Impact of durational outlier removal from unit selection catalogs* Fifth ISCA ITRW on Speech Synthesis (SSW5), Pittsburgh, PA, USA.

Kominek J., Black A. (2006) *The Blizzard Challenge 2006 CMU Entry introducing hybrid trajectory-selection synthesis*.

Kopaliński W. (2000) *Słownik Wyrazów Obcych i Zwrotów Obcojęzycznych z Almanachem* 83-7227-582-3.

Koržinek D., Brocki L. (2007) *Grammar based automatic speech recognition system for the Polish language,* Recent Advances in Mechatronics 2007: 87-91.

Koza J.R., Rice J. P. (1991) *Genetic generation of both the weights and architecture for a network Neural Networks,* 1991 IEEE international conference pp: 397-044.

Kumar R. (2004) *Genetic Algorithm for Unit Selection based Speech Synthesis*, International Conference on Spoken Language Processing (Interspeech - ICSLP), October 2004, Jeju Korea.

Kupiec J. (1992) Robust part-of-speech tagging using a Hidden Markov Model, *Computer Speech and Language*, n6, pp. 225-242.

Laver J. (1994) *Principles of phonetics,* Oxford University Press., Oxford, UK.

Lemmetty S. (1999) *Review of Speech Synthesis Technology,* This Master's Thesis has been submitted for official examination for the degree of Master of Science in Espoo on March 30, 1999.

Louw J.A., Davel M., Barnard E. (2005) *A general-purpose IsiZulu Speech Synthesiser Human Language Technologies Research group*, Meraka Institute / University of Pretoria August 2005.

Marasek K. (1997) Electroglottographic description of voice quality. Arbeitspapiere des Instituts fur mascinele Sprachverarbeitung, Stuttgart, 3(2)

Marasek K. (2003 B) *Synteza mowy: przegląd technologii i zastosowań ze szczególnym uwzględnieniem języka polskiego*.

Marasek K. (2003) *LVCSR system for Polish, Archives of Acoustic*.

Marasek K., Gubrynowicz R. (2004) *Multi-level Annotation in SpeeCon Polish Speech Database,* inIMTCI, pages58–67.

Michalewicz Z. (2004) *Algorytmy genetyczne + struktury danych = programy ewolucyjne* wyd. Wydawnictwa Naukowo-Techniczne ISBN: 83-204-2881-5.

Möbius B. (2001) *Rare events and closed domains: Two delicate concepts in speech synthesis*, in 4th ISCA Workshop on Speech Synthesis, 2001, pp. 41–46.

Oliver D. (1998) *Polish Text to Speech Synthesis*, praca magisterska University of Edinburgh Department of Linguistics 1998.

Oliver D. (2007) *Modelling Polish intonation for Speech Synthesis*, Saarbrucken 2007, Phd.

Oliver D., Szklanny K. (2006) *Creation and analysis of a Polish speech database for use in unit selection synthesis*, LREC Genoa, Italy 2006.

Pierrehumbert J. (1980) *The Phonology and Phonetics of English intonation,* Phd dissertation, MIT. IULC edition.

Pierrehumbert J. (1983) *Automatic recognition of intonation patterns,* ACL Proc. of 21st Annual Meeting, pp.85-90.

Richmond K., Strom V., Clark R., Yamagishi J., Fitt S. (2007) *Festival Multisyn Voices for the 2007 Blizzard Challenge,* Centre for Speech Technology Research University of Edinburgh, Edinburgh, United Kingdom.

Rong-Wei Yi J. (2003) *Corpus-Based Unit Selection for Natural-Sounding Speech Synthesis*, Phd.

Roudet L. (1947) przekład T. Benni, *Zasady fonetyki ogólnej*, Warszawa, 1947.

Silverman K., Beckman M., Pitrelli J., Ostendorf M., Wightman C., Price P., Pierrehumbert J., Hirschberg, J. (1992) *TOBI: a Standard for Labeling English Prosody*.

Sonninen A. (1956), *The Role of the External Laryngeal Muscles in Length Adjustemnents of the Vocal Cords in Singing*, Acta Otto-Laryngol., Suppl. 156.

Sproat R., Hirschberg J., Yarowsky D. (1992) *A Corpus-based Synthesizer*, *Proc. ICSLP* Alberta, pp. 563-566.

Stevens K. (1998) *Acoustic phonetics*. Current studies in linguistics (No. 30). Cambridge, MA: MIT. ISBN 0-262-19404-X.

Szklanny K. (2002) *Przygotowanie bazy difonów języka polskiego dla realizacji syntezy mowy w systemie MBROLA*, praca magisterska, Warszawa.

Szklanny K. (2003) *Preparing the Polish diphone database for speech synthesis in MBROLA,* 50. Otwarte Seminarium z Akustyki Szczyrk, Poland.

Szklanny K. (2008) *Synteza mowy w E-learningu dla osób niepełnosprawnych,* Postępy e-edukacji, praca zbiorowa pod redakcja zespołu ośrodka kształcenia na odległość OKNO PW, str. 371-379, Warszawa 2008 ,Oficyna Wydawnicza Politechniki Warszawskiej ISBN 978-83-7207-795-0.

Szklanny K. Oliver D. (2005) *Corpus Creation for Polish Unit Selection Speech Synthesis*, Proc. of Speech Analysis, Synthesis and Recognition: Applications of Phonetics, SASR 2005, Cracow, Poland,

Szklanny K. Oliver D. (2005) *Preparing the Corpora for Unit Selection Speech Synthesis for Polish*, Proc. of One-day Meeting for Young Speech Researchers, (OMYSR 2005), UCL, London, 14 April, 2005, p. 9.

Szklanny K.(2004) *Zajęcia dydaktyczne WKK - Werbalna Komunikacja z Komputerem*, PJWSTK.

Szklanny K., Wójtowski M. (2008) *Automatic segmentation quality improvement for realization of unit selection,* Proc. of Human System interactions p.251-256 Digital Object Identifier 10.1109/HSI.2008.4581443, Krakow 2008, Poland.

Tadeusiewicz R. (1988) *Sygnał mowy*, Wyd. Komunikacji i Łączności, Warszawa 1988.

Taylor P., Black A., Caley R. (1998) *The architecture of the festival speech synthesis system,* in The Third ESCA Workshop in Speech Synthesis, pages 147-151, Jenolan Caves, Australia, 1998.

Taylor P, (2009) *Text-to-Speech Synthesis* Hardback (ISBN-13: 9780521899277) Cambridge, UK ; New York : Cambridge University Press, 2009.

ToBI http://www.ling.ohio-state.edu/~blodgett/TOBICLINIC/tobiclinic.html

Tokuda K., Zen H., Black A. (2002) *An HMM-based speech synthesis system applied to English*, IEEE SSW, 2002 - http://hts.sp.nitech.ac.jp/?Publications, 12-2007.

Tsao T.P., Chen G.C. S.H. (2001) *Short-term load forecasting using neural networks and evolutionary programming,* ID Proc. of the fifth Intl Power Engineering Conference, Singapore, pp:443-748

Turing A.M. (1950) *Computing machinery and intelligence* Mind, 59, 433-460.

Van Santen J., Buchsbaum A. (1997) *Methods for Optimal Text Selection,*Proc.5th Euro Conf on Speech Communication and Technology (EUROSPEECH-97), pages553–6, Rhodes, Greece.

Venditti, J.(1997), *Japanese ToBI Labelling Guidelines*, Ohio State University Working Papers in Linguistics, 50: 62-72.

Vepa J. (2004) *Join Cost for Unit Selection Speech Synthesis,* University of Edinburgh, 2004, Phd.

Vepa J., King S. (2006) Subjective evaluation of join cost and smoothing methods for unit selection speech synthesis, IEEE Transactions on Speech and Audio Processing, 14(5):1763-1771, September 2006.

Villaseñor-Pineda L. Montes-y-Gómez M. Vaufreydaz D. Serignat J (2004) *Experiments on the Construction of a Phonetically Balanced Corpus from the Web*, A. Gelbukh (Ed.): CICLing 2004, LNCS 2945, pp. 416–419, 2004.© Springer-Verlag Berlin Heidelberg 2004.

Villaseñor-Pineda L., Gómez M., Coutino M., Vaufreydaz D. (2003) *A Corpus Balancing Method for Language Model Construction,* in Computational Linguistics and intelligent Text Processing, 4th international Conference, CICLing, pages393–401,MexicoCity,Mexico.

Viswanathan M., Viswanathan M. (2005) Measuring speech quality for text-to-speech systems: development and assessment of a modified mean opinion score (MOS) scale, Computer Speech and Language 19, 55–83.

Viterbi A. J. (1967) Error bounds for convolutional codes and an asymptotically optimal decoding algorithm, IEEE Transactions on Information Processing, 13:260-269.

Wagner A. (2004) *A phonological model of intonation and intonation transcription system ToBI for Polish - a preliminary study,* Speech and Language Technology, vol.8, pp. 137-162

Wagner A. (2008) *A comprehensive model of intonation for application in speech synthesis*, Rozprawa doktorska.

Wells J.C. (1997) *SAMPA computer readable phonetic alphabet,* in Gibbon, D., Moore, R. and Winski, R. (eds.), 1997. Handbook of Standards and Resources for Spoken Language Systems. Berlin and New York: Mouton de Gruyter. Part IV, section B.

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

Wierzchowska B. (1980) *Fonetyka i fonologia języka polskiego*, Wrocław: Ossolineum.

Willemse R., Gulikers L. (1992) *Word class assignment in a Text-To-Speech system*, Proc.int. Conf. on Spoken Language Processing, Alberta, pp. 105-108

Williams B., (1995) Text-to-speech synthesis for Welsh and Welsh English, Proceedings, Eurospeech '95, Madrid, 1995.

Wouters J. Macon M. (1998) *Perceptual evaluation of distance measures for concatenative speech synthesis*, Proc. ICSLP, vol. 6, (Sydney, Australia), pp. 2747–2750, 1998.

Wójtowski M. (2007) *Segmentacja akustycznej bazy językowej na potrzeby realizacji korpusowej syntezy mowy w systemie Festival*, praca magisterska

Yarowsky D. (1994) *Homograph Disambiguation in Speech Synthesis'*, *Proceedings, 2nd ESCA/IEEE Workshop on Speech Synthesis*, New Paltz, NY.

Young S., Evermann G., Kershaw D., Moore G., Odell J., Ollason D., Valtchev V., Woodland P. (2001) *The HTK Book (for HTK Version 3.1)* 1995-1999 Microsoft Corporation. 2001-2002 Cambridge University Engineering Department. First published December 1995 Revised for HTK Version 3.1 December 2001.

Zhang J.,Toth A., Collins-Thompson K., Black A. (2004) *Prominence prediction for super-sentential prosodic modeling based on a new database*. in Fifth ISCA ITRW on Speech Synthesis (SSW5), Pittsburgh, PA, USA.

Spis rysunków

Spis tabel

Załącznik 1: Zdania użyte do estymacji funkcji kosztu

W Tadżykistanie padał bez przerwy rzęsisty deszcz i zamienił boisko w dżungli w grzęzawisko.

Szwagier i bratowa siedli mu na plecach i potężną dłonią przydusili łeb mężczyzny do ziemi.

Fala powietrza z dzwonka nad drzwiami wejściowymi, wdziera się do środka i strąca kwiatek z okna.

Irlandczycy przyswoili sobie angielski i świetnie na tym wyszli, bo są wybuchowi, w zdenerwowaniu krzyczą, a za chwilę już się śmieją.

Hydrotelefon ma kilka źródeł zasilania, a za unikalny wyrób trzeba słono płacić.

Okręt nabrał wody w ciągu dwóch minut i piętnastu sekund, dlatego ćwiczy się również ratunkową ewakuację poprzez wyrzutnie torpedowe.

Do dziś tamtejsi mieszkańcy czczą go jak narodowego bohatera, mimo że zrobił jakieś grube szachrajstwo, i przywłaszczył sobie majątek.

Słyszę nagle wrzask, z którego ciężko coś zrozumieć, to uciekło czterdzieści pensjonariuszek z poprawczaka.

Rzeczywiście rozwojowi pleśni zapobiega dodatek ziarenek gorczycy lub chrzanu.

Francuzi nie zapomnieli o skromnych, acz smakowitych detalach, niektóre z nich można zobaczyć w błękitnej lagunie.

Obszar badań nazwany inżynierią ziarna podobno dotyczył kwiatów, które przypominały dzwonki lub rurki.

Wczoraj wybuchł pożar, i zbudowanie dróg dojazdowych przez Rospudę przestało być problemem.

gęste włosy to skarb i trzeba o nie dbać, jednak zanim wybierzesz odżywkę sprawdź jej skład na etykiecie.

W Zimbabwe zaobserwowano dwa stare, wypędzone ze stada lwy, jeden z nich polując kiedyś na guźca utknął w norze.

Byłabym ostrożna w formułowaniu daleko idących wniosków, nawet duża firma przy ryzykownych przedsięwzięciach, też mogłaby paść.

W pobliskim sadzie gruszki jak dzbany i śliczne jabłka migały między strzechami.

okoliczni mieszkańcy skarżą się że do podzłotoryjskich kamieniołomów dzień i noc jeżdżą dziesiątki ciężarówek.

Robotnicy mają do dyspozycji ogromne przestrzenie, dlatego szybko budują sztuczne zbiorniki wodne i elektrownie.

Załącznik 2: Lista wyrazów z rzadko występującymi fonemami

różdżkarz	dżonkile	dodrzewiać
Odrzykoń	w katedrze	bezdrzewny
Nadodrzańskie	rozedrzeć	mądrze
dżin	zedrzeć	nadjeżdżający
dżambo-dżet	wedrzeć się	mądrzyć się
dżentelmen	odedrzeć	żongler
dżiu-dżitsu	Podrzeć	żonkoś
pidżama	drzewce	laryngofon
kartridże	odjeżdża	dyftonga
gadżet	wyjeżdżony	ingerować
menadżer	brydżysta	ranga
radża	brydżowy	kangur
nozdrza	drzemka	ewangelik
dżul	zdrzemnąć się	szylinga
dżudo	dżonka	stangret
dżem	mędrzec	koniunkcja
dżungla	drożdże	posłanka
dżuma	przedrzeźniać	konkordat
dżip	Kilimandżaro	konkury
dżinsy	wydrze	łowczanka
zadżumiony	wydrzyj	uczynkami
dżokej	zadrzeć	szlachcianka
modrzew	zadrzyj	szacunkowo
gwiżdże	podżegacz	piosenka
drzazga	nadrzeć	w pojedynku
móżdżek	mizdrzyć się	z frasunku
zadrzewienie	kindżał	świnka
wydrzeć	hidżra	panienka
udrzeć	gwożdżenie	rabunku
podrzędny	gnieżdżenie się	zaściankowy
Krowodrza	dżdżysto	cienka
jeżdżę	dżdżownica	punktak
dojeżdżali	dżamper	bez ożenku
miażdżyca	dżygitówka	kochanka
obedrzeć	drzewiej	zielonka

opiekunka
z pocałunkami
szynka
na ganku
kamionkowy
dunka
mango
rankami
jutrzenka
koronka
instynkt
kiszonka
w pojedynkę
powodzianka
krzemionka
łodzianka
barmanka
cyganka
wychowanka
konkluzja
wzmianka
powiedzonko
bez szwanku
sukienka
mrzonka
franków
ziemianka
okienko
podarunkami
synka
inkwizycja
strunka
polanka
godzinka
pisanka
kuzynka
olszynka
Kongo
Angola
na ringu
szwankować

onkolog
inkubator
w rysztunku
drink
studzenie
słodzenie
judzenie
zochydzenie
znudzenie
wędzarz
wędzonka
rozwodzenie
rozpogodzenie
krzywdzenie
księdzu
jędza
błądzenie
nawiedzony
biedzenie
radzenie
nieodzowny
siedzenie
powiedzony
twierdzenie
wodzący
uprzedzony
zasadzenie
sprowadza
wiedza
rodzaj
dzwon
gładzony
sprawdzone
sprawdzać
rozchodzenie
pobudzany
gromadzenie
przechadzać się
prowadzenie
urządzony
urządzenie

upośledzony
pochodzą
rodzeństwo
przeprowadzenie
poprzedzają
budzenie
uszkodzony
stwierdzono
oszczędzać
doradzać
przewodzenie
odzyskać
zwiedzać
odrodzenie
widzowie
śledzenie
uszkodzenie
władza
cudzysłów
kadzenie
grodzenie
ogrodzony
dzbanek
złudzenie
rydzowy
gawędzenie
swędzenie
niewidzenie
sadzawka
rdza
zrządzenie
nudzenie
zrodzony
dowodzenie
kładzenie
szydzenie
powodzenie
marudzenie
pędzenie
przędza
rdzeń

twierdza
bruździć
maźnięcię
moździerz
późny
w płaszczyźnie
wyraźny
źle
gwoździe
gwóźdź
mroźny
październik
rzeźba
źrenica
źródło
zwięźle
jeździec
ugryzienie
źdźbło
buzie
jezioro
wziernik
weźmij
wziąść
źrebak
kaźnia
więźba
koźlę
kozica
ździebełko
kuźnia
liźnięcie
zięba
łaźnia
groźny
nieziemski
przezierać
przeźrocze
ziewa
tuzin
uziemienie
użyźniać
więźniarka
woźny
ziarno
wyziew
ozimina
zielny
ziomek
zięć
znaleźne
zrazić
groźba
spóźnić się
cudzoziemiec
we frazie
w analizie
w wyrazie
raźny
narazić się
obraźliwy
ziółka
doraźny
zaraźliwy
trzeźwy
przekaźnik
gruźlica
na gazie
na głazie
w guzie
buzia
w optymiźmie
zipnąć
zima
ziąb
w wozie