INVESTIGATIONES LINGUISTICAE, VOL. XVI; POZNAŃ, DECEMBER 2008

Fonetyczno-akustyczna analiza struktury sylaby w języku polskim na potrzeby technologii mowy

Acoustic-phonetic analysis of syllable in Polish for use in speech technology

Daniel Śledziński

Instytut Językoznawstwa, Uniwersytet im. Adama Mickiewicza
al. Niepodległości 4, 61-874 Poznań

danielsl@poczta.onet.pl, fon@amu.edu.pl

Abstract

This paper presents results of investigations concerning the role of syllable as the main unit for automatic speech recognition and speech synthesis for Polish. Research reveals that the acoustic properties of the particular syllable make it easy detectable in the speech signal. The artificial neural networks were used for classification tasks. Also auditory test showed that use of syllables is good solution for speech synthesis. The article presents parts of author`s doctor thesis: 'Acoustic-phonetic analysis of syllable in Polish for use in speech technology'.

1 Wprowadzenie

Sylaba jest pojęciem trudnym do zdefiniowania. W *A Dictionary of Phonetics and Phonology* (Trask, 1996, s. 345) znajduje się następujący zapis dotyczący sylaby: „(...) pomimo że użytkownicy danego języka najczęściej bez trudu potrafią określić, z ilu sylab składa się dane słowo lub wypowiedź, pomimo że systemy piśmiennicze oparte na sylabach funkcjonują od tysięcy lat (...), termin ten jest niezwykle trudny do zdefiniowania". Ladefoged stwierdził, że „ten termin nigdy nie został zdefiniowany" (Ladefoged, 1975, s. 217).

Pomimo niejasnego statusu lingwistycznego, sylaba jest jednostką, którą można wykorzystać dla celów praktycznych. Przeprowadzone badania dotyczyły potencjalnej możliwości zastosowania sylaby jako jednostki, na której mogą zostać oparte systemy rozpoznawania oraz syntezy mowy polskiej.

W publikacji przedstawiono fragmenty rozprawy autora: *Fonetyczno-akustyczna analiza struktury sylaby w języku polskim na potrzeby technologii mowy*. W skróconej formie przedstawiono wyniki badań. W pierwszej kolejności skoncentrowano się przeglądzie dostępnych definicji terminu sylaba. Zaproponowano również własną koncepcję przydatną dla celów praktycznych. Następnie przedstawiono rozważania dotyczące wykorzystania sylaby w systemach rozpoznawania mowy oraz syntezy mowy.

Zaprezentowano też wyniki obszernych badań a także metody, które były użyte w tych badaniach.

2 Pojęcie sylaby

Sylaba jest pojęciem, którego nie da się zdefiniować w sposób wyczerpujący i uniwersalny. Można wskazać przynajmniej na dwie przyczyny tego stanu rzeczy. Pierwsza związana jest z faktem, że w wielu przypadkach trudno jest określić w sposób jednoznaczny granice pomiędzy poszczególnymi sylabami, trudno jest też wyznaczyć kryteria, na podstawie których takie granice można by wyznaczać (można kierować się na przykład kryteriami morfologicznymi lub kryteriami fonetyczno-akustycznymi). Problem związany z wyznaczaniem granic między sylabami w języku polskim dotyczy przede wszystkim zbitek spółgłoskowych.

Druga przyczyna, która sprawia, że podanie uniwersalnej definicji terminu sylaba jest praktycznie niemożliwe, związana jest z różnicami występującymi pomiędzy językami. Różnice te sprawiają, że percepcja sylabiczności przez użytkowników języka może być różna w różnych systemach językowych (Clark, Yallop, 1995, s.68). Dlatego użytkownicy wielu języków, jako ośrodki sylab percypują tylko samogłoski (dotyczy to także języka polskiego), w innych językach mogą to być również spółgłoski sonorne (na przykład w języku angielskim czy czeskim). Istnieją też języki, których użytkownicy słyszą spółgłoski szczelinowe jako ośrodki sylab (na przykład użytkownicy języka Bella Coola) (Rogers, 1990, s 89). Inne różnice pomiędzy językami dotyczą właściwości fonotaktycznych sylab. Na przykład język japoński dopuszcza nagłos sylaby złożony wyłącznie z pojedynczej spółgłoski, natomiast w języku polskim mogą być to wieloelementowe zbitki spółgłoskowe o bardzo zróżnicowanej strukturze (na przykład w wyrazach: *pstrąg, wskroś*)

Przedstawione wyżej przyczyny uniemożliwiają podanie pełnej i uniwersalnej definicji tej jednostki – definicji na podstawie której można by było wyznaczyć wszystkie granice sylab w sposób jednoznaczny. W praktyce wszystkie definicje sylaby do pewnego stopnia opisują jej strukturę lub odnoszą się do zjawisk, które zachodzą w czasie wymawiania tych jednostek. Nigdy jednak nie są to definicje pełne i uniwersalne.

Dla zastosowań praktycznych, problemy związane z brakiem możliwości podania wyczerpującej definicji sylaby (oraz wynikającego z tej definicji jednoznacznego podziału wszystkich wyrazów na sylaby) można rozwiązać poprzez przyjmowanie pewnych ustaleń umownych. Takie podejście jest zgodne z przyjętą tezą i związane jest z praktycznym wykorzystaniem tej jednostki.

W dalszym ciągu przytoczono kilka definicji pochodzących z literatury oraz sformułowano założenia i koncepcję sylaby jako jednostki w konkretnych rozwiązaniach technicznych.

2.1 Definicje fonetyczne

Definicje fonetyczne odnoszą się do zjawisk o charakterze fizycznym, zachodzących w czasie artykulacji poszczególnych sylab. Niektóre publikacje zawierają informacje odnoszące się zarówno do płaszczyzny fonologicznej, jak i fonetycznej.

W *Encyklopedii językoznawstwa ogólnego* (Michowska, 1993, s. 585) zamieszczono następujące informacje dotyczące sylaby fonetycznej:

> Odcinek wypowiedzi stanowiący jedność ekspiracyjną, ruchową i akustyczną, posiadający jedno maksimum donośności, który potencjalnie może być fonetycznie samodzielną wypowiedzią. Brak dotąd w pełni zadowalającej fonetycznej definicji sylaby.

Dalej zwrócono uwagę na to, że sylaba była definiowana jako odcinek wypowiedzi między dwoma minimami: siły ekspiracji, energii artykulacyjnej, rozwarcia narządów artykulacyjnych oraz donośności.

Podejście fonetyczne prezentuje też Wierzchowska w książce *Wymowa polska* (Wierzchowska, 1971, s. 214-216). Oto jak definiuje ona sylabę:

> Odcinek mowy, zawarty między momentami jednoczesnych zmian w: układzie narządów mowy, ciśnieniu powietrza w tchawicy, natężeniu przebiegu akustycznego i jego donośności, nosi nazwę sylaby. Na ośrodek sylaby przypada maksimum rozwarcia narządów mowy, z czym wiąże się najniższe ciśnienie powietrza w tchawicy, najwyższe natężenie i największa donośność dźwięku. Na krańcach sylaby stopień zbliżenia narządów mowy jest znacznie większy, wzrasta też ciśnienie subglotalne, maleje za to natężenie i donośność dźwięku.

The Oxford Dictionary of English Grammar (Chalker, Weiner 1994, s. 387) zawiera następujące informacje dotyczące sylaby fonetycznej:

> Definition of the syllable in the universally valid phonetic terms has proved difficult, whether based on the auditory feature of prominence or on the articulatory feature of 'pulse'. In the prominence theory, some sounds which are more prominent then others form the core of a syllable, with the less prominent sounds at the syllabic boundaries. (…) In the pulse theory, it is claimed that the number of syllables correspond to the number of chest 'pulses', with vowel sounds again being central to the syllable(…).
>
> Różne podejścia do sylaby fonetycznej zawarto w *A Dictionary of Phonetics and Phonology*: There have been various attempts to define syllable phonetically: as a single respiratory movement (the chest-pulse theory), as the single opening and closing of the vocal tract, as a single peak of prominence in the soundstream resulting from the combination of stress, pitch, length and intrinsic sonority (the prominence theory).

Zestawienie fonetycznych teorii sylaby przedstawiono w książce Beats-and-Binding Phonology (Dziubalska-Kołaczyk, 2002, s. 44):

> (a) respiratory theory: the syllable is 'a sound-group produced with a single respiratory impulse(…);
>
> (b) acoustic theory: acoustic sonority (Shallfülle) – 'a portmanteau term' for voicing, aperture, expiratory force, pitch, muscular energy (of consonants), duration (of vowels), penetration (of fricatives); impressionistically, it is a measure of the audibility of sounds(…);
>
> (c) articulatory theory: (…) the syllable consists of explosion (i.e., sound(s) of increasing aperture) and implosion (i.e., sound(s) of decreasing aperture)(…);
>
> (d) motor theory: (…) the syllable is constituted by a ballistic movement of the intercostals muscles(…)."

2.2 Definicje fonologiczne

Istotą definicji fonologicznych jest to, że opisują one strukturę sylaby przy użyciu abstrakcyjnych klas dźwięków (samogłosek oraz spółgłosek). W niniejszym podrozdziale przedstawiono przegląd definicji fonologicznych.

W *Encyklopedii językoznawstwa ogólnego* zamieszczono następującą definicję sylaby:

> Podstawową cechą strukturalną sylaby jest kontrast pomiędzy jej składnikami: obligatoryjnym ośrodkiem (szczytem) i fakultatywnymi marginaliami.

Dalej wskazano na rodzaje dźwięków, które mogą pełnić funkcję ośrodka sylaby oraz marginaliów:

> Ośrodkiem sylaby jest najczęściej samogłoska, może nim być jednak również spółgłoska płynna lub nosowa, rzadko spółgłoska trąca. (…) Marginalia sylaby to nagłosowa grupa spółgłoskowa zwana następem sylaby oraz wygłosowa grupa spółgłoskowa zwana zestępem sylaby. Sylaby pozbawione wygłosowej grupy spółgłoskowej (tzn. zakończone elementem

> wokalicznym) nazywa się sylabami otwartymi, sylaby zakończone na spółgłoskę – sylabami zamkniętymi.

W definicji tej wskazano również na inny sposób opisywania struktury sylaby:

> (…) w wielu opracowaniach przyjmuje się dwudzielną strukturę sylaby z rozróżnieniem następu i rymu sylaby, przy czym rym zawiera obligatoryjny ośrodek sylaby i fakultatywny zestęp.

W podręczniku *Gramatyka polska* (Strutyński, 2002, s.63) zawarto rzeczową definicję sylaby fonologicznej (ze szczególnym uwzględnieniem specyfiki języka polskiego):

> (…)składa się na nią ciąg głosek, których ośrodkiem w języku polskim jest zawsze samogłoska. Sylaba może się składać z jednej samogłoskim np. a-le, i-dę, o-ko; z połączenia samogłoski ze spółgłoską, np. a-le, i-dę, o-ko lub z połączenia samogłoski z grupą spółgłosek, np. sto-pa, strax. Sylaba może być otwarta, czyli równa samogłosce lub mająca w wygłosie samogłoskę, np. o-ko-li-ca, lub zamknięta, czyli zakończona spółgłoską lub grupą spółgłosek, np. las, most (…) Sylaba nie pokrywa się z morfemem (tj. rdzeniem, przedrostkiem, przyrostkiem, końcówką fleksyjną). W związku z tym nie jest ona stałym segmentem wyrazu, np. (…) most ale mo-stek".

Definicja zawiera dwie bardzo istotne informacje – informację mówiącą o tym, że w języku polskim w roli ośrodka sylaby występuje tylko samogłoska oraz drugą informację mówiącą o tym, że granice sylab nie pokrywają się z granicami morfemów.

W *A Dictionary of Phonetics and Phonology* (Trask, 1996, s. 345) umieszczono następujące informacje dotyczące sylaby:

> A fundamental but elusive phonological unit typically consisting of a short sequence of segments, most typically single vowel or diphthong possibly precedes by one or more consonants.

Przedstawiono też drugą możliwość fonologicznego opisu struktury sylaby:

> It is now usual to subdivide the syllable into an onset and rhyme, with the rhyme further divided into a nucleus (or peak) and coda.

Definicję fonologiczną przedstawiono również w *The Oxford Dictionary of English Grammar* (Chalker, Weiner 1994, s. 387):

> A unit of pronunciation forming the whole or a part of a word, and having one vowel (or syllabic consonant) phoneme, often with one or more consonants before or after.

W podręczniku *Kognitywne podstawy języka i językoznawstwa* (Tabakowska, 2001, s 165-166) również zamieszczono kilka ważnych informacji, jednak dotyczą one zarówno płaszczyzny fonologicznej, jak i fonetycznej:

> Można myśleć o sylabach jako jednostkach, które posiadają ośrodek (centrum, szczyt samogłoskowy o maksymalnej donośności), który może być otoczony przez elementy o mniejszej sonorności (spółgłoski). W ciągu mowy występują zatem przemienne elementy o większej i mniejszej donośności. (…) Polski jest językiem spółgłoskowym (tzn. językiem z przewagą fonemów spółgłoskowych), o bardzo dużych możliwościach kombinatorycznych w obrębie grup spółgłoskowych i nawet czterech spółgłoskach na początku sylab (drgnąć).

2.3 Analiza definicji wcześniejszych

W przytoczonej literaturze definiowanie sylab na płaszczyźnie fonetycznej polega na odwołaniu się do zjawisk o naturze fizycznej, które zachodzą w czasie ich artykulacji. Definicje fonetyczne sylaby są w literaturze spotykane rzadziej. Podanie takiej definicji jest zadaniem niezwykle trudnym. W praktyce istnieje kilka teorii odnoszących się do cech fonetycznych (fizycznych) sylab. Każda z tych teorii odzwierciedla w pewnym stopniu jakiś fragment rzeczywistości, jednak z pewnością żadna z nich nie jest w pełni wyczerpująca i uniwersalna. Na podstawie przytoczonych wcześniej opracowań można przedstawić podsumowanie kilku podstawowych podejść (teorii) związanych z

fonetycznym definiowaniem sylaby. Podejścia te związane są z następującymi zjawiskami fizycznymi:

a) ekspiracją (ang. chest-pulse theory) – sylaba jest odcinkiem mowy ulokowanym pomiędzy dwiema przerwami ciągłości wydechu,
b) stopniem rozwarcia narządów artykulacyjnych – sylaba definiowana jest jako pojedyncze rozwarcie (eksplozja) i zwarcie (implozja) traktu głosowego, na ośrodek sylaby przypada maksimum rozwarcia narządów artykulacyjnych,
c) donośnością akustyczną dźwięków – (ang. the prominence-theory) – sylaba definiowana jest poprzez ulokowanie jej pomiędzy kolejnymi minimami donośności akustycznej, na ośrodki sylaby przypadają dźwięki o największej donośności,
d) napięciem mięśniowym (energią artykulacyjną) – sylaba definiowana jest poprzez ulokowanie jej pomiędzy kolejnymi maksimami napięcia mięśniowego, na ośrodki sylaby przypadają minima napięcia mięśniowego.

W literaturze spotykane są dwa podejścia związane z fonologicznym opisem struktury sylaby. Pierwsze podejście zakłada występowanie:

a) obligatoryjnego ośrodka sylaby – w języku polskim w roli ośrodka sylaby może występować tylko samogłoska (z bardzo nielicznymi wyjątkami), natomiast w innych językach mogą to być również spółgłoski sonorne lub nawet spółgłoski szczelinowe. Dla języka polskiego zakłada się, że dany wyraz ma taką samą liczbę samogłosek oraz sylab.
b) fakultatywnych marginaliów – są to dźwięki należące do sylaby i znajdujące się przed lub za ośrodkiem sylaby. Dźwięki znajdujące się przed ośrodkiem sylaby nazywane są nagłosem sylaby (lub następem sylaby), natomiast dźwięki znajdujące się za ośrodkiem sylaby nazywane są wygłosem sylaby (lub zestępem sylaby). Istnienie nagłosu oraz wygłosu nie jest warunkiem niezbędnym dla istnienia sylaby – sylaba może składać się z samego ośrodka. Zarówno nagłos, jak i wygłos sylaby mogą składać się z jednej lub z większej liczby spółgłosek. Każdy język ma swoje reguły i ograniczenia dotyczące dopuszczalnej liczby spółgłosek w nagłosie oraz w wygłosie sylaby. Na przykład niektóre języki dopuszczają tylko jedną spółgłoskę w nagłosie sylaby oraz nie dopuszczają żadnej spółgłoski w wygłosie (brak wygłosu). Natomiast język polski dopuszcza wyjątkowo dużą liczbę spółgłosek zarówno w nagłosie, jak i w wygłosie. Wyniki obszernych analiz tego zagadnienia zostały przedstawione w publikacji *Fonotaktyczna analiza mówionego tekstu polskiego* (Łobacz, Jassem, 1974, s. 179-197). Połączenie większej liczby spółgłosek nazywane jest grupą spółgłoskową lub zbitką spółgłoskową. Jeżeli sylaba nie posiada wygłosu (kończy się samogłoską), to jest to sylaba otwarta, natomiast jeżeli sylaba zakończona jest jedną lub większą liczbą spółgłosek, to jest to sylaba zamknięta.

Inny sposób fonologicznego opisu struktury sylaby zakłada jej podział na:

a) nagłos – jest to fakultatywna spółgłoska lub grupa spółgłoskowa stojąca przed samogłoską,
b) rym – obejmuje on obligatoryjny ośrodek oraz fakultatywny wygłos. Wyrazy zakończone takim samym rymem rymują się. Również przy tym sposobie opisu struktury sylaby zakłada się podział na sylaby zamknięte (rym jest złożony z ośrodka oraz z wygłosu) oraz sylaby otwarte (brak wygłosu w rymie).

Przedstawione informacje dotyczą definicji ustalonych na płaszczyźnie fonologicznej. Warto wspomnieć, że fonologiczne definiowanie sylaby może mieć związek nie tylko z jej cechami strukturalnymi, ale również z jej funkcjami prozodycznymi. Wierzchowska w *Wymowie polskiej* (Wierzchowska, 1971, s. 216) definiuje akcent w następujący sposób:

> W podręcznikach fonetyki, a także w gramatykach, przez akcent rozumie się zwykle uwydatnienie pewnych sylab w wyrazach lub dłuższych wypowiedziach. W pierwszym przypadku mówi się o akcencie wyrazowym, w drugim – o akcencie zdaniowym. Uwydatnianie sylab akcentowanych dokonuje się przez zwiększenie ich donośności, przez przedłużenie ich, a także zmienianie wysokości muzycznej ich tonu podstawowego.

W istocie rzeczy sylaba często pełni kluczową rolę w modelowaniu intonacji fraz, jednak zagadnienie to cechuje się pewną odrębnością, dlatego nie było ono analizowane.

2.4 Założenia i definicja sylaby z punktu widzenia prowadzonych badań

Dalsze rozważania mają na celu ustalenie definicji sylaby dla potrzeb obecnych oraz przyszłych badań.

Trzeba rozważyć, jakie praktyczne konsekwencje niesie wykorzystanie definicji fonetycznej. Wcześniej wspomniano, że nie istnieje pełna i uniwersalna definicja fonetyczna sylaby. Nawet, jeżeli ograniczyć się do jednego języka, to podanie takiej definicji prawdopodobnie nie będzie możliwe. Poza tym posługiwanie się definicją fonetyczną uniemożliwiłoby segmentacje sygnału mowy na sylaby. Taka segmentacja musiałaby zawierać informacje o granicach czasowych sylab, na przykład „w minimach donośności akustycznej". W praktyce jest to niewykonalne. Kolejny problem wynikający z zastosowania definicji fonetycznej polega na braku możliwości przyjmowaniu pewnych założeń dotyczących struktury sylaby. Przedstawione problemy wykluczają użycie wyłącznie fonetycznej definicji. Natomiast przyjęcie definicji fonologicznej umożliwia segmentację sygnału mowy na sylaby (przy uwzględnieniu pewnych umownych założeń). Poza tym istnieje możliwość określania struktury poszczególnych sylab. Trzeba jednak pamiętać, że konieczne jest stosowanie określonych reguł podziału. Niestety tych reguł nie da się w pełni oprzeć na przesłankach lingwistycznych – trzeba stosować pewne rozwiązania umowne. Przyjęta definicja wygląda następująco:

(C) V (C)

gdzie V jest obligatoryjnym ośrodkiem sylaby (najczęściej samogłoską) natomiast (C) to fakultatywny nagłos oraz fakultatywny wygłos sylaby. Nagłos oraz wygłos może obejmować jedną spółgłoskę lub grupę spółgłosek (zbitkę spółgłoskową). Poza tym przyjęto następujące założenia dotyczące wyodrębniania sylab:

a) granice między wyrazami stanowią zawsze granice sylab,
b) dla dwóch bezpośrednio następujących po sobie ośrodków sylab stosuje się zawsze następującą regułę podziału: V|V,
c) przy połączeniu sylaby otwartej z sylabą o pojedynczym nagłosie stosuje się zawsze następującą regułę podziału: V|CV,
d) jeżeli pomiędzy kolejnymi ośrodkami sylab znajduje się więcej niż jedna spółgłoska, to stosuje się odpowiednią regułę podziału. Stworzona przez autora baza zbitek spółgłoskowych obejmuje propozycje reguł podziału dla ponad 2500 zbitek spółgłoskowych występujących w języku polskim. Wiele reguł ma charakter umowny, jednak większość z nich opiera się na konkretnych przesłankach. Przy opracowywania tego zestawu reguł autor kierował się praktycznymi potrzebami związanymi z segmentacją sygnału mowy na sylaby.

3 Analiza przydatności jednostek językowych w systemach ARM

Większość przeprowadzonych badań dotyczyła zastosowania sylaby w systemach automatycznego rozpoznawania mowy. Przed przystąpieniem do omówienia wyników tych eksperymentów warto przedstawić rozwiązania alternatywne i wskazać na potencjalne wady i zalety każdego nich. Trzeba jednak zaznaczyć, że zaprezentowane informacje nie

opierają się na wynikach konkretnych badań i mają charakter ogólnych rozważań. Twierdzenia te warto poddawać dyskusji.

3.1 Analiza przydatności innych jednostek

W dalszym ciągu rozdziału scharakteryzowano różne jednostki językowe oraz jednostki quasi językowe pod kątem ich potencjalnych wad oraz zalet dla systemu automatycznego rozpoznawania mowy polskiej. Charakterystyka ta może stanowić pewien punkt odniesienia dla rozważań dotyczących przydatności sylab w systemach ARM (rozważania te zawarto w podrozdziale 3.2).

3.1.1 Fonem

Liczba fonemów w języku polskim jest niewielka – jest ich około 40, zatem potencjalna liczba modeli odzwierciedlających cechy akustyczne tych jednostek również nie jest duża. Fonem jest abstrakcyjną klasą dźwięków. Pomiędzy różnymi realizacjami danego fonemu mogą występować znaczne różnice. Według książki *Mowa a nauka o łączności* (Jassem, 1974, s.128) mogą być to różnice:

- przypadkowe,
- osobnicze,
- kontekstowe,
- stylistyczne.

W mowie ciągłej występują liczne zniekształcenia i pominięcia konkretnych realizacji fonemów – realizacje te mogą się od siebie znacznie różnić, mogą też ulegać przeróżnym deformacjom lub być zupełnie pomijane. Niektóre głoski mają też skłonność do „zlewania się" z głoskami sąsiednimi (dotyczy to na przykład połączeń aproksymantów z samogłoskami) – wynikają z tego trudności związane z wyznaczaniem granic między głoskami. Można je dostrzec analizując spektrograficzny zapis mowy ciągłej.

Kolejny potencjalny problem związany z wykorzystaniem fonemów polega na tym, że pojedyncza realizacja fonemu może po prostu okazać się zbyt krótka, aby parametry akustyczne związane z tym krótkim odcinkiem niosły istotne i relewantne informacje.

Następna niedogodność polega na tym, że pojedyncza realizacja fonemu nie stanowi w zasadzie żadnego punktu odniesienia w sensie semantycznym – wykrycie pojedynczego fonemu nie umożliwia, nawet w przybliżeniu, określenia kategorii semantycznej. Poza tym istotne są właściwości dystrybucyjne tych jednostek (z punktu widzenia potrzeb rozpoznawania mowy) – dla pojedynczego fonemu liczba możliwych kontekstów jest bardzo duża.

3.1.2 Alofon

Istnieje możliwość utworzenia modeli dla poszczególnych wariantów pozycyjnych oraz wariantu podstawowego danego fonemu (dla alofonów danego fonemu). W zasadzie podejście takie posiada większość wad wyżej opisanego podejścia fonemicznego – przede wszystkim różnorodność akustyczną poszczególnych realizacji danego alofonu. Poza tym występuje problem „zlewania się" się głosek, krótki czas trwania i brak powiązania z semantyką oraz brak sprzyjających właściwości dystrybucyjnych.

Oczywiście podejście to uwzględnia różnice kontekstowe, które warunkują istnienie poszczególnych wariantów pozycyjnych danego fonemu. Trzeba jednak pamiętać, że warianty te są opisywane na płaszczyźnie artykulacyjnej.

3.1.3 Difon

Według niektórych definicji difon jest fragmentem mowy obejmującym odcinek od środka jednej głoski do środka kolejnej głoski. Teoretyczna liczba difonów w języku polskim to około 1400, jednak badania przeprowadzone przez autora pracy na wielomilionowych

korpusach tekstowych wykazały, że jest ich maksymalnie około 900 (difonów wewnątrzwyrazowych). Wady i zalety wykorzystania difonów są raczej trudne do określenia. Od razu uwypukla się problem związany z segmentacją tekstu na tak ujęte jednostki – jeżeli granica położona jest w środku realizacji głoski, to jej wyznaczenie może okazać się problematyczne.

Natomiast difon traktowany jako para dwóch głosek może okazać się całkiem przydatną jednostką dla systemu ARM, jednak dla ostatecznego stwierdzenia tej przydatności niezbędne jest przeprowadzenie odpowiednich badań.

3.1.4 Trifon

W technologii mowy trifon jest najczęściej traktowany jak fonem w określonym kontekście innych fonemów. Wynika z tego, że trifony obejmują wszystkie warianty pozycyjne i podstawowe fonemów. Potencjalna liczba trifonów dla języka polskiego przekracza 40000. Jednak badania autora przeprowadzone na wielomilionowych korpusach tekstowych wykazały, że w wyrazach języka polskiego występuje około 11000 różnych połączeń trifonowych. Uwzględnianie wszystkich możliwych kontekstów danego fonemu wydaje się być zabiegiem nieuzasadnionym (w szczególności dla spółgłosek właściwych). Definicja trifonu sprawia, że modele muszą być tworzone dla pojedynczych fonemów (występujących w określonych kontekstach). Trifon posiada zatem wszystkie wady, które dotyczą fonemów i alofonów. Natomiast jeżeli trifon traktować jako 3 fonemy (całe), to tak ujęte jednostki mogą okazać się przydatne dla systemów ARM.

3.1.5 Morfem

Użycie morfemów leksykalnych i fleksyjnych wydaję się być ciekawą alternatywą dla sylab. Jednostki takie posiadają silne powiązanie z semantyką. Poza tym realizacje morfemów leksykalnych są najczęściej długie (w sensie iloczasu). Oprócz tego segmentacja nagrań na morfemy jest zadaniem prostszym niż wykonanie takiej czynności w odniesieniu do głosek lub nawet w odniesieniu do sylab. Trzeba zaznaczyć, że rozwiązanie to nie koliduje ze złożonym systemem fleksyjnym języka polskiego, ponieważ można budować modele dla morfemów leksykalnych oraz fleksyjnych. Jednak dla ostatecznego określenia przydatności morfemów dla systemów rozpoznawania mowy konieczne jest przeprowadzenie odpowiednich badań.

3.1.6 Wyraz

Ponieważ język polski jest językiem fleksyjnym, przy czym bardzo złożonym pod tym względem (sumaryczna liczba form fleksyjnych wyrazów języka polskiego sięga kilku milionów), to wykorzystanie wyrazu w systemach rozpoznawania mowy raczej nie może być brane pod uwagę (chyba, że w odległej przyszłości). Możliwe i uzasadnione jest jednak budowanie modeli dla pojedynczych, często używanych wyrazów.

3.1.7 Inne jednostki

Kolejne rozwiązanie (które należy traktować raczej jako propozycję do rozważenia), polega na użyciu jednostek (fragmentów sygnału), które nie mają określonego statusu lingwistycznego, ale mogą być nośnikami pewnych osobliwych cech akustycznych. Może to być na przykład zwarcie, zmiana z dźwięczności na szum. Mogą to być również cechy możliwe do zamodelowania wyłącznie przy użyciu sztucznych sieci neuronowych. Poza tym jednostki takie mogą na siebie nachodzić – akustyczny sygnał mowy może być więc traktowany jako ciąg graniczących ze sobą, nachodzących na siebie lub zawierających się fragmentów sygnału, które są nośnikami pewnych relewantnych cech akustycznych.

3.2 Analiza przydatności sylab dla systemów ARM

Potencjalna przydatność sylaby dla wykorzystania w systemach automatycznego rozpoznawania mowy była przedmiotem opisywanych w tym artykule badań.

Sylaba pozbawiona jest wielu wad, które posiada głoska. Pojedyncze głoski mogą być pomijane, natomiast sylaby artykułowane są prawie zawsze, chociaż one też mogą ulegać zniekształceniom. Jednak wiele z tych deformacji odbywa się w sposób na tyle regularny, że można je opisać i uwzględniać przy tworzeniu odpowiednich modeli. Dotyczy to przede wszystkim często występujących w języku polskim redukcji zbitek spółgłoskowych – na przykład wyraz */fSystko/* jest częściej wymawiany jako */FSysko/*. Poza tym sylaby są jednostkami dłuższymi niż pojedyncze głoski, można zatem przypuszczać, że poszczególne realizacje tych jednostek zawierają relewantne cechy akustyczne (z twierdzeniem tym związane były przeprowadzone badania). Oprócz tego sylaby mają pozytywne właściwości semantyczne oraz dystrybucyjne (z punktu widzenia potrzeb systemów ARM). Często wykrycie pojedynczej sylaby umożliwia określenie nielicznego zbioru potencjalnych kategorii semantycznych. Właściwości dystrybucyjne sylab sprawiają, że wykrycie pojedynczej sylaby umożliwia wyznaczenie zbioru sylab (najczęściej bardzo nielicznego), z którymi może ona graniczyć. Kolejna istotna zaleta sylab ma związek z przygotowywaniem dźwiękowych zbiorów nagrań – wystarczy, żeby baza zawierała granice czasowe pomiędzy kolejnymi sylabami, a nie głoskami. Dzięki temu, granic tych jest znacznie mniej a wiele kłopotliwych połączeń (np. aproksymant-samogłoska) jest pomijanych, ponieważ znajdują się one wewnątrz sylaby. Należy wspomnieć również o tym, że wiele procesów fonologicznych (np. palatalizacja) odbywa się wewnątrz sylaby. Modele tworzone dla sylab obejmują też przejścia (transienty) występujące wewnątrz sylab.

Podstawową wadą omawianego podejścia jest liczba sylab – jest ich wiele tysięcy, z drugiej strony najważniejszych i najczęściej występujących sylab typu CV jest w języku polskim tylko około 150, a często używanych – połowa tej liczby.

Inna niedogodność, związana z użyciem sylab, wynika z braku możliwości jednoznacznego określenia granic między poszczególnymi sylabami (problem ten został opisany wcześniej). Z drugiej strony można to uznać za zaletę – jeżeli istnieje kilka możliwości wyznaczenia granicy pomiędzy sylabami, to można przyjąć granicę dogodniejszą (umowną) – na przykład z punktu widzenia łatwości segmentacji sygnału.

4 Sylaba w systemach syntezy mowy

Przeprowadzono szereg eksperymentów związanych z potencjalną możliwością wykorzystania sylab w syntezie mowy polskiej. Analizowane podejście zakładało łączenie całych sylab. Założenie to może wydawać się oczywiste, jednak w literaturze spotykane są też podejścia nieco odmienne – również odnoszące do wykorzystania sylaby w syntezie mowy. W publikacji (Łobacz, 2001, s. 81-112) przedstawiono następujące rozwiązania (stanowią one odniesienie do publikacji Peterson oraz Siverstena):

- pół sylaby – jednostka rozciągające się od początku sylaby do jej centralnej części,
- sylaby – zaproponowane podejście polega na wyznaczaniu granic pomiędzy sylabami w stadium ustalonym spółgłoski. Rozwiązanie to może okazać się przydatne (przynajmniej w niektórych sytuacjach), jednak mniej praktyczne – wykonanie segmentacji takich jednostek byłoby zdecydowanie trudniejsze.

Przydatność tych rozwiązań dla syntezy mowy polskiej mogłaby zostać przeanalizowana w oddzielnych badaniach.

4.1 Potencjalne zalety wynikające z użycia sylab

Należy zwrócić uwagę na kilka faktów, które mogą stanowić istotne argumenty przemawiające za użyciem sylaby jako podstawowej jednostki dla konkatenacyjnej syntezy mowy języka polskiego:

a) w przeciwieństwie do innych jednostek – na przykład difonów, trifonów albo morfemów- sylaba jest naturalną jednostką artykulacyjną. Można tutaj odnieść się do fonetycznych definicji sylaby – według niektórych definicji sylaba jest związana z pojedynczym wydechem lub z pojedynczym wzrostem i spadkiem energii akustycznej,
b) sylaba jest podstawową jednostką służącą do opisywania i modelowania cech prozodycznych – jest to zagadnienie niezwykle istotne dla konkatenacyjnej syntezy mowy. Jeden z etapów działania tych systemów związany jest z modelowaniem intonacji we frazach złożonych z połączonych jednostek. Zastosowanie sylaby jako podstawowej jednostki syntezy umożliwia łatwe i przejrzyste modelowanie cech prozodycznych frazy już na etapie doboru jednostek do łączenia (ang. unit selection). Jest to istotne, ponieważ zbyt duże modyfikacje przebiegu częstotliwości podstawowej są mocno słyszalne,
c) baza nagrań związana z danym systemem syntezy mowy musi przechowywać granice czasowe jednostek, które są łączone. Aby jakość syntezy była dobra, granice te powinny być wyznaczone dobrze. Najczęściej wstępną segmentację wykonuje się metodą automatyczną. Jednak wyznaczone w ten sposób granice nie są dokładne – występują częste przesunięcia (czasami znaczne). Poza tym zdarza się, że jakiś dodatkowy czynnik, na przykład chrząknięcie, spowoduje, że wykonana segmentacja będzie błędna. Dlatego niezbędne jest przeprowadzanie ręcznej korekty automatycznie segmentowanego tekstu. Tutaj uwypukla się kolejna zaleta wykorzystania sylaby. Dzięki takiemu podejściu niezbędna jest ręczna korekta granic czasowych tylko i wyłącznie w odniesieniu do granic pomiędzy poszczególnymi sylabami – nie ma konieczności korygowania granic fonemów znajdujących się wewnątrz sylab. Dzięki temu pominięta zostaje znaczna ilością granic, których nie da się wyznaczyć w sposób jednoznaczny.

Przedstawione argumenty przekonują, że system syntezy mowy oparty na sylabach może być rozwiązaniem optymalnym (w stosunku do wariantów użycia innych jednostek) oraz że warto prowadzić badania w tym zakresie. Wymieniając potencjalne korzyści takiego rozwiązania, należy również określić wady lub ewentualne trudności, które mogą mu towarzyszyć.

4.2 Potencjalne problemy związane z użyciem sylab

Podstawowy problem związany z łączeniem fragmentów sygnału akustycznego polega na tym, że takie syntetyczne połączenia mogą być słyszalne. Poniżej wymieniono rodzaje połączeń, jakie mogą wystąpić w analizowanym modelu syntezy mowy oraz problemy, które mogą być związane z tymi połączeniami:

a) Połączenie typu samogłoska-samogłoska (V_V) – istnieją argumenty, które przemawiają za tym, aby tego typu połączeń w ogóle nie rozdzielać. Przede wszystkim występują one stosunkowo rzadko, poza tym liczba możliwych kombinacji nie jest duża. Nawet jeżeli następujące po sobie samogłoski są wymówione wyraźnie, to jednoznaczne wyznaczenie granicy pomiędzy nimi nie jest możliwe. Jeżeli artykulacja jest niestaranna, to zamiast dwóch samogłosek, może zostać wypowiedziana jedna wydłużona samogłoska. W czasie jej artykulacji położenie masy języka może stanowić stadium pośrednie pomiędzy tymi dwiema samogłoskami (z ewentualnym niewielkim przesunięciem). Zjawisko to można zaobserwować wypowiadając w pośpiechu np.

słowo aeroplan. Biorąc pod uwagę te argumenty, autor zasugerował, aby tego typu połączenia nie były rozdzielane, dlatego nie zostały one uwzględnione w badaniach.

b) Połączenia typu samogłoska-spółgłoska (V_C) – w rozważanym modelu syntezy mowy tego typu połączenie występuje przy łączeniu sylaby otwartej (zakończonej samogłoską) z kolejną sylabą (rozpoczynającą się od spółgłoski). Samogłoski są bardzo podatne na zjawisko koartykulacji – masa języka najczęściej znacznie zmienia swoje położenie w miarę oddalania się od części ustalonej samogłoski. Kierunek tych zmian jest uzależniony od miejsca artykulacji sąsiedniej spółgłoski, a więc jest on różny dla różnych spółgłosek (do tego zjawiska odnosi się teoria lokusa). Powstaje zatem pytanie o uniwersalność sylab jako jednostek dla syntezy mowy (w połączeniach V_C). Pytanie to dotyczy tego, czy pobierane sylaby otwarte z jednego kontekstu (kontekstu, jaki stanowi spółgłoska należąca do następnej sylaby), mogą być bez ograniczeń używane w innych kontekstach. Jeżeli pewne ograniczenia występują, to kolejne pytanie może dotyczyć ich zasięgu. Aby odpowiedzieć na te pytania, zostały przeprowadzone obszerne doświadczenia.

c) Połączenia typu spółgłoska-spółgłoska (C_C) – ten rodzaj połączeń występuje przy łączeniu sylaby zamkniętej (zakończonej spółgłoską) z kolejną sylabą (rozpoczynającą się od spółgłoski). Można przypuszczać, że połączenia tego typu mają dobrą jakość, szczególnie w przypadku połączeń spółgłosek właściwych. Aby nie opierać końcowych wniosków na przypuszczeniach, przeprowadzono również badania audytywne, które dotyczyły syntetycznych połączeń typu C_C.

5 Metody obliczeniowe, narzędzia i bazy danych

Przed przystąpieniem do omawiania wyników doświadczeń warto przedstawić środki techniczne, które zostały użyte. Scharakteryzowano krótko metody obliczeniowe, narzędzia informatyczne (programy komputerowe) oraz bazy danych. Niektóre programy zostały utworzone przez autora, natomiast jeżeli chodzi o bazy nagrań, to wszystkie one zostały stworzone specjalnie na potrzeby tych badań.

5.1 Metody obliczeniowe

W badaniach wykorzystano następujące metody obliczeniowe (Francuz, 2007, s. 3-576):

a) Testy parametryczne. Bardzo często stosowanym testem parametrycznym jest test t dla prób niezależnych. Został on użyty w niektórych analizach dotyczących działania sieci neuronowych. Umożliwia on ocenę różnicy pomiędzy średnimi wartościami pewnej cechy w dwóch różnych grupach. Testy parametryczne cechują się większą mocą statystyczną niż testy nieparametryczne – posługując się nimi można wykryć bardziej subtelne różnice pomiędzy grupami. Jednak aby korzystać z tego rodzaju testów, muszą zostać spełnione pewne warunki wstępne. Przede wszystkim rozkład cech w obydwu grupach musi być normalny.

b) Testy nieparametryczne. Jeżeli warunki wstępne dotyczące użycia testów parametrycznych nie są spełnione, wtedy trzeba korzystać z testów nieparametrycznych. Praktycznie wszystkie testy parametryczne mają swoje odpowiedniki nieparametryczne. Istnieje kilka testów nieparametrycznych, które stanowią odpowiedniki dla testu t dla prób niezależnych. Ta grupa testów obejmuje: test serii Walda-Wolfowitza, test dla dwóch prób Kołmogorowa-Smirnowa oraz test U Manna-Whitneya.

c) Analiza wariancji. Kolejną metodą użytą w pracy jest analiza wariancji (ANOVA – ang. Analysis of Variance). Najczęściej jest ona używana do porównywania średniej

wartość danej cechy w kilku podgrupach danej grupy, dlatego należy wskazać zmienną grupującą.

d) Analiza skupień. Celem analizy skupień jest ułożenie obiektów w grupy w taki sposób, aby stopień powiązania obiektów z obiektami należącymi do tej samej grupy był jak największy, a z obiektami z pozostałych grup jak najmniejszy. Metoda ta wykrywa struktury w danych (skupienia) bez wyjaśniania, dlaczego one występują.
e) Sieci neuronowe. Zastosowanie sieci neuronowych daje zupełnie nowe możliwości (w stosunku do tradycyjnych metod statystycznych). Dzięki nim możliwe stało się modelowanie zależności o charakterze nieliniowym, co więcej – możliwe jest modelowanie zmiennych wielowymiarowych. Sieci neuronowe mają zdolność samodzielnego uczenia się na podstawie pewnego zbioru danych uczących (informacje niezbędne do wytrenowania sieci tkwią w samych danych). Charakterystyczną cechą sieci neuronowych jest to, że ich struktura i działanie przypomina ludzki mózg (jego pewne fragmenty), co prawda jest to bardzo uproszczony model, zachowana została jednak istota działania biologicznego systemu nerwowego. Sieci neuronowe podczas wykonywania zadań klasyfikacyjnych potrafią poprawnie klasyfikować przypadki podobne lub zbliżone do przypadków użytych do uczenia sieci. Właściwość ta nazywana jest generalizacją, ma ona szczególne znaczenie w takich dziedzinach jak rozpoznawanie tekstu lub rozpoznawanie mowy. W badaniach wykorzystano następujące rodzaje sieci neuronowych: perceptrony wielowarstwowe (MLP), sieci o radialnych funkcjach bazowych (RBF), probalistyczne sieci neuronowe (PNN), sieci liniowe. (Tadeusiewicz, 1998, s. 1-164), (Tadeusiewicz, 2007, s. 9-426)

5.2 Narzędzia

Dla realizacji celów pracy wykorzystano kilka aplikacji. Były to zarówno programy autorskie (napisane przez autora publikacji), jak i programy innych autorów. Najważniejsze aplikacje, które wykorzystano w badaniach to:

- program służący do przeprowadzania zaawansowanych analiz fonetyczno-akustycznych (program autorski),
- program służący do rozbudowanych analiz fonostatystycznych (program autorski),
- inne niewielkie programy autorskie dla realizacji poszczególnych zadań, na przykład dla: konwersji transkrypcji fonemicznej na sylabiczną, konwersji formatu plików, zliczania jednostek językowych,
- programy służące do wykonywania operacji na plikach dźwiękowych zawierających zapis mowy ludzkiej. Aplikacje WaveSurfer oraz Praat umożliwiają prezentację graficzną głosu, segmentację sygnału mowy oraz obliczanie ciągów parametrów akustycznych. Program Wave Lab jest narzędziem firmy Steinberg przeznaczonym do wykonywania operacji na zbiorach plików dźwiękowych (np. konwersji nagrań z płyt oraz konwersji częstotliwości próbkowania plików dźwiękowych),
- program służący do konwersji tekstów ortograficznych na transkrypcję SAMPA (Demenko, Wypych, Baranowska, 2003, s. 79-95). Program bazuje na regułach konwersji opracowanych przez Marię Steffen-Batogową,
- Statistica – rozbudowany program służący do obliczeń statystycznych. W badaniach wykorzystano możliwości związane z przeprowadzaniem testów parametrycznych, testów nieparametrycznych oraz tworzeniem wykresów. Wykorzystano również dodatkowy moduł programu – Sieci Neuronowe.

5.3 Bazy danych

Dla realizacji celów pracy autor stworzył obszerne zbiory nagrań:

- Nagrania wyrazów izolowanych – zostały one utworzone na potrzeby eksperymentów związanych z potencjalną możliwością wykorzystania sylaby jako jednostki dla automatycznego rozpoznawania mowy. Zawierają one łącznie około 26000 realizacji sylab. Nagranych zostało 30 osób (podział płci jest równy). Każda osoba przeczytała 332 wyrazy, zatem w bazie danych znajduje się łącznie około 10000 wyrazów,
- Nagrania mowy ciągłej – utworzono je z myślą o eksperymentach związanych z rozpoznawaniem sylab w mowie ciągłej. W nagraniach wzięło udział 70 osób. Każda osoba przeczytała 20 zdań, zatem nagrano łącznie prawie 1400 zdań (kilka nagrań zostało odrzuconych ze względów technicznych). Nagrania obejmują w sumie około 18000 realizacji sylab. Wynika z tego, że razem z wcześniej opisywaną bazą wyrazów izolowanych, na potrzeby eksperymentów związanych z rozpoznawaniem mowy nagrano około 42000 sylab,
- Nagrania logatomów – przeprowadzono obszerne badania dotyczące zjawiska koartykulacji. Dotyczyły one wpływu spółgłosek na samogłoski. Badania te miały związek z syntezą mowy opartą na sylabach,
- Nagrania wyrazów dla doświadczeń audytywnych – część przeprowadzonych badań związana była z wykorzystaniem sylab w syntezie mowy. Badania te miały charakter audytywny. Przygotowano nagranie około pięciuset wyrazów. Część z nich została pocięta na fragmenty (sylaby). Potem sylaby te zostały ponownie połączone (przeprowadzono resyntezę),
- Oprócz nagrań sygnału akustycznego przygotowano również obszerne zbiory tekstowe. Najważniejszy z nich to baza zbitek spółgłoskowych języka polskiego. Zawiera ona dokładne statystyki występowania oraz informacje o możliwych kontekstach dla wszystkich zbitek spółgłoskowych języka polskiego. Poza tym zawiera ona reguły podziału na sylaby (dla niemal wszystkich zbitek spółgłoskowych języka polskiego).

6 Badania związane z wykorzystaniem sylab w systemach ARM

W dalszym ciągu przedstawiono wyniki doświadczeń związanych z potencjalną możliwością wykorzystania sylab języka polskiego w systemach automatycznego rozpoznawania mowy. W badaniach wykorzystano sztuczne sieci neuronowe (sprawdzono działanie ponad 600 sieci). Ze względu na znaczną obszerność eksperymenty podzielono na cztery oddzielne serie.

6.1 Metody badań

Zgromadzony materiał badawczy umożliwił przeprowadzenie obszernych badań. Wykorzystano w nich sztuczne sieci neuronowe, które wykonywały zadania o charakterze klasyfikacyjnym dychotomicznym. Jako zmienne niezależne użyto zestaw 12 wartości współczynników cepstralnych, które były pobierane z 7, 10 lub 14 punktów czasowych poszczególnych realizacji sylab. Dzięki temu akustyczna postać tych sylab była reprezentowana przez zestaw odpowiednio 84, 120 lub 168 wartości liczbowych. Wartości te były podawane na wejściach sieci neuronowych. Klasyfikacja dychotomiczna zakłada istnienie dwóch możliwości klasyfikacji danych. W przeprowadzonych eksperymentach dane wejściowe mogły być przyporządkowywane do konkretnej sylaby (którą musiała rozpoznawać dana sieć neuronowa) lub do zbioru sylab losowych (wszystkich innych, różnych od tej sylaby, która miała być rozpoznawana). Uwzględniono jednak pewien

czynnik dodatkowy, który sprawił, że stopień trudności omawianego zadania był znacznie większy niż przy typowej klasyfikacji dychotomicznej. Utrudnienie to polegało tym, że liczba przypadków należących do klasy sylab losowych była znacznie większa od liczby realizacji tej sylaby, która miała być rozpoznawana (dotyczy to zbiorów: uczącego, walidacyjnego oraz testowego). Przykładowo 300 realizacji sylaby, która miała być rozpoznawana, mogło być wymieszanych z piętnastoma tysiącami realizacji sylab losowych. Przyjęcie takich założeń pozwoliło na traktowanie sieci neuronowych jako detektorów potrafiących wykrywać w sygnale akustycznym określone struktury – a więc potrafiących odpowiednio reagować tylko wtedy, kiedy one rzeczywiście wystąpiły.

Przedstawiona wyżej metoda badań związana jest z następującym pytaniem dotyczącym sylab języka polskiego: Czy poszczególne realizacje danej sylaby posiadają cechy akustyczne właściwe realizacjom tylko tej sylaby i niespotykane w realizacjach innych sylab? (można też mówić o koincydencji cech akustycznych właściwych tylko i wyłącznie dla realizacji danej sylaby). W dalszym ciągu artykułu przedstawiono wyniki badań, dzięki którym można odpowiedzieć na to pytanie. Autor wyszedł z założenia, że jeżeli sztuczne sieci neuronowe nauczą się odpowiednio klasyfikować realizacje określonej sylaby (wymieszane ze znacznie większą liczbą sylab losowych), to na tej podstawie można wnioskować o koincydencji cech akustycznych właściwych tylko realizacjom tej sylaby.

Wytrenowano i sprawdzono działanie prawie 600 sztucznych sieci neuronowych – zastosowano sieci liniowe, sieci probalistyczne, perceptrony wielowarstwowe oraz sieci o radialnych funkcjach bazowych. Eksperymenty podzielono na cztery serie różniące się pewnymi założeniami. Proces przygotowywania danych dla sieci neuronowych był przeprowadzany przy użyciu oprogramowania stworzonego przez autora artykułu. Do uczenia oraz testowania sieci użyto programu Statistica (moduł Sieci Neuronowe).

6.2 Wyniki badań

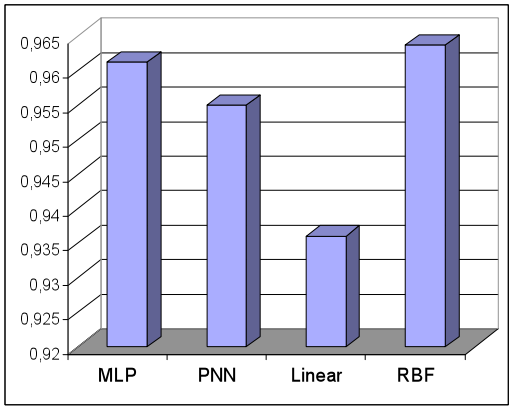
Doświadczenia przeprowadzono w czterech seriach. Pierwsza seria miała charakter wstępny i dotyczyła wyrazów izolowanych. W drugiej serii wykorzystano mniejszą liczbę sylab, jednak przeprowadzone badania były znacznie bardziej szczegółowe. W trzeciej serii sprawdzono możliwość dwustopniowego wykrywania sylab. Czwarta seria doświadczeń była przeprowadzona na nagraniach mowy ciągłej. W dalszym ciągu przedstawiono w skrócie założenia i wyniki każdej serii doświadczeń.

Pierwsza seria doświadczeń miała charakter wstępny. Zostało przeanalizowanych łącznie 43 popularnych sylab języka polskiego pod kątem ich zdolności do wykrywania przez sztuczne sieci neuronowe. Sprawdzono działanie czterech (wymienionych wcześniej) rodzajów sieci. W sumie wytrenowano 172 sieci neuronowe. W tabeli 6.1 zamieszczono wyniki czterech przykładowych doświadczeń. Tabela zawiera informacje o rodzaju i strukturze sieci, osiągniętej jakości dla zbiorów: uczącego, walidacyjnego oraz testowego oraz macierz pomyłek. Analizując informacje umieszczone w macierzach pomyłek można zauważyć, że sieci neuronowe dobrze spełniają rolę detektorów reagujących na określone sylaby.

Tabela 6.1 Przykładowe zestawienie wyników doświadczeń w pierwszej serii eksperymentów (dotyczy rozpoznawalności sylaby v+y przez 4 rodzaje sieci neuronowych)

Rodzaj i strukt. sieci	U	W	T	rozpozn. jako:	random	v+y
MLP 120: 120-120-1:1	0,970	0,956	0,946	random	4765	**10**
				v+y	214	**616**
PNN 120:120 -2803-2-2:1	1,00	0,969	0,967	random	4894	**4**
				v+y	85	**622**
RBF 87:87-159-1:1	0,955	0,958	0,951	random	4736	**12**
				v+y	243	**614**
Linear 120:120-1:1	0,941	0,942	0,931	random	4700	**63**
				v+y	279	**563**

Ogólnie, najlepszą średnią jakość sieci (w zbiorze testowym) uzyskano dla sieci probalistycznych. Nieco gorzej wypadły perceptrony wielowarstwowe (94%) oraz sieci liniowe (93%). Jednak z punktu widzenia założeń do tych eksperymentów większe znaczenie ma zastosowany parametr wykrywalność. Parametr ten dotyczy tylko i wyłącznie tej sylaby, której realizacje miały być wykrywane przez daną sieć neuronową. Średnią wartość parametru wykrywalność przedstawiono na rysunku 6.1



Rys. 6.1 Średnia wykrywalność (zbiór testowy)

Druga seria doświadczeń dotyczyła niewielkiej liczby sylab, jednak przeprowadzone badania były znacznie bardziej szczegółowe. Uwzględniono różne liczby punktów, z których były pobierane zestawy wartości parametrów akustycznych (współczynników cepstralnych). Poza tym przetestowano różne warianty liczebności zbioru sylab losowych – mógł on zawierać 5000, 10000 lub 15000 elementów. Obliczono też macierze pomyłek osobno dla zbiorów: uczącego, walidacyjnego oraz testowego.

W ramach tej części doświadczeń przeprowadzono szereg jednoczynnikowych analiz wariancji. Sprawdzano wpływ wymienionych wyżej czynników na osiąganą średnią jakość sieci neuronowych oraz średnią wartość parametru wykrywalność (wszystkie analizy dotyczyły zbiorów testowych przypadków). Jedna z ważniejszych analiz dotyczyła wpływu liczby realizacji sylab w zbiorze sylab losowych na osiągniętą wartość parametru wykrywalność. Zwiększanie liczby sylab w zbiorze sylab losowych nie spowodowało spadku średniej wartości tego parametru (który utrzymywał się na poziomie 94%) – jest to wynik zaskakujący, a jednocześnie bardzo pozytywny. Tendencja ta została potwierdzona przez wyniki testu post-hoc dla analizy wariancji (dot. perceptronów wielowarstwowych).

W trzeciej serii doświadczeń sprawdzono możliwość dwustopniowego wykrywania sylab. W pierwszej kolejności były ustalane pewne klasy sylab o podobnej strukturze (na przykład klasa sylab CV o identycznym nagłosie oraz różnych ośrodkach). Na potrzeby tych eksperymentów wprowadzono następujące nazewnictwo: sieci pierwotne oraz sieci wtórne. Sieci pierwotne miały wykrywać realizacje sylab pewnej określonej klasy, natomiast sieci wtórne decydowały o tym, do której konkretnie sylaby przyporządkować wykryte dane akustyczne. W roli sieci pierwotnych użyto sieci probalistycznych, natomiast zadanie sieci wtórnych spełniały perceptrony wielowarstwowe. Działanie sieci pierwotnych sprawdzano w wersji bez macierzy strat oraz z macierzą strat. Tabela 6.2 przedstawia macierz pomyłek uzyskaną z działania przykładowej sieci wtórnej.

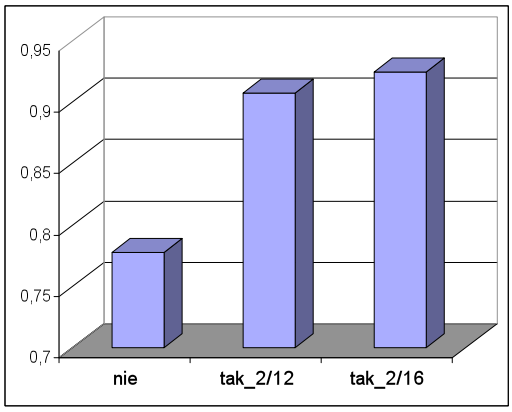
Tabela 6.2 Przykładowa macierz pomyłek (dla sieci wtórnej)

	v+a	v+i	v+y	v+o
v+a	**153**	0	0	1
v+i	0	**62**	2	0
v+y	1	4	**150**	1
v+o	2	0	0	**55**

Czwartą serię eksperymentów przeprowadzono na nagraniach mowy ciągłej (przetestowano działanie 180 sieci neuronowych). Dla każdej analizowanej sylaby sprawdzono trzy warianty sieci probalistycznych – jeden bez macierzy strat oraz dwa z macierzą strat. Czynnik straty ustalony w tych dwóch wariantach na wartości 12 oraz 16 dotyczył błędu polegającego na niewykryciu sylaby, która musiała być wykryta przez daną sieć (błąd polegający na zaklasyfikowaniu realizacji tej sylaby do zbioru sylab losowych).

Z rysunku 6.2 wynika, że wartość parametru wykrywalność w wersji bez macierzy strat wyniosła niecałe 80%, jednak zastosowanie macierzy strat znacznie polepszyło ten rezultat. Z drugiej strony zastosowanie macierzy strat nieznacznie pogorszyło jakość sieci (z 99% do około 98%). Wynika to z faktu zwiększenia liczby sylab ze zbioru sylab losowych, klasyfikowanych do klasy sylaby, która miała być wykrywana.

Wyniki badań wykazały, że poszczególne sylaby były wykrywane nawet wtedy, gdy stosunkowo niewielka liczba ich realizacji (na przykład 200), była wymieszana ze znacznie większym zbiorem sylab losowych (na przykład zbiorem obejmującym realizacje 15000 sylab). Wskaźnik przedstawiający odsetek wykrytych sylab prawie za każdym razem przekraczał 90% (wynik ten dotyczył zbiorów testowych przypadków). W ocenie autora uzyskany rezultat jest pozytywny. Mając na uwadze właściwości dystrybucyjne sylab, uzyskany wynik ma jeszcze większą wartość. Dzięki tym właściwościom, dla danej sylaby można z dużym prawdopodobieństwem wyznaczyć zbiór potencjalnych sylab (najczęściej niezbyt liczny), z którymi może ona graniczyć.



Rys. 6.2 Średnia wykrywalność (zbiór testowy) dla różnych wariantów macierzy strat

7 Badania związane z wykorzystaniem sylab w syntezie mowy

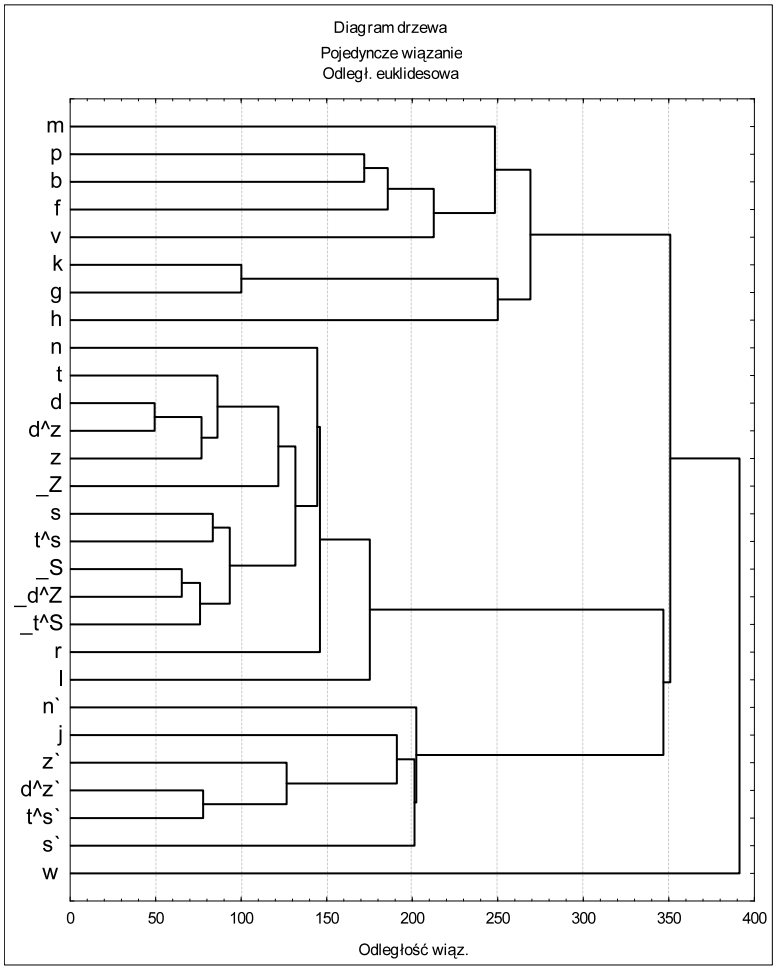
W rozdziale przedstawiono wyniki dwóch obszernych eksperymentów audytywnych dotyczących wykorzystania sylaby jako podstawowej jednostki dla syntezy mowy języka polskiego. Pierwszy eksperyment dotyczył syntetycznych połączeń typu V_C, natomiast drugi odnosił się do połączeń typu C_C. W analizowanym modelu syntezy mowy połączenie typu V_C dotyczy łączenia sylaby otwartej (zakończonej samogłoską) z kolejną sylabą (rozpoczynającej się od spółgłoski). Połączenie typu C_C dotyczy łączenia sylaby zamkniętej (zakończonej spółgłoską) z kolejną sylabą (rozpoczynającą się również od spółgłoski). Jednak przed omówieniem tych doświadczeń przedstawiono krótką informacje o innych doświadczeniach – związanych ze zjawiskiem koartykulacji.

7.1 Zjawisko koartykulacji

Koartykulacja jest zjawiskiem związanym ze wzajemnym oddziaływaniem na siebie dźwięków mowy ludzkiej. W mowie ciągłej koartykulacja występuje zawsze, we wszystkich językach i dotyczy ona wszystkich głosek. Koartykulacja, jak sama nazwa wskazuje, jest zjawiskiem odnoszącym się do aspektu artykulacyjnego, jednak wywiera ona wpływ również na cechy akustyczne sygnału mowy – co więcej, koartykulacja może być analizowana na płaszczyźnie fonetyczno-akustycznej poprzez pomiary odpowiednich parametrów (przede wszystkim wartości formantu drugiego).

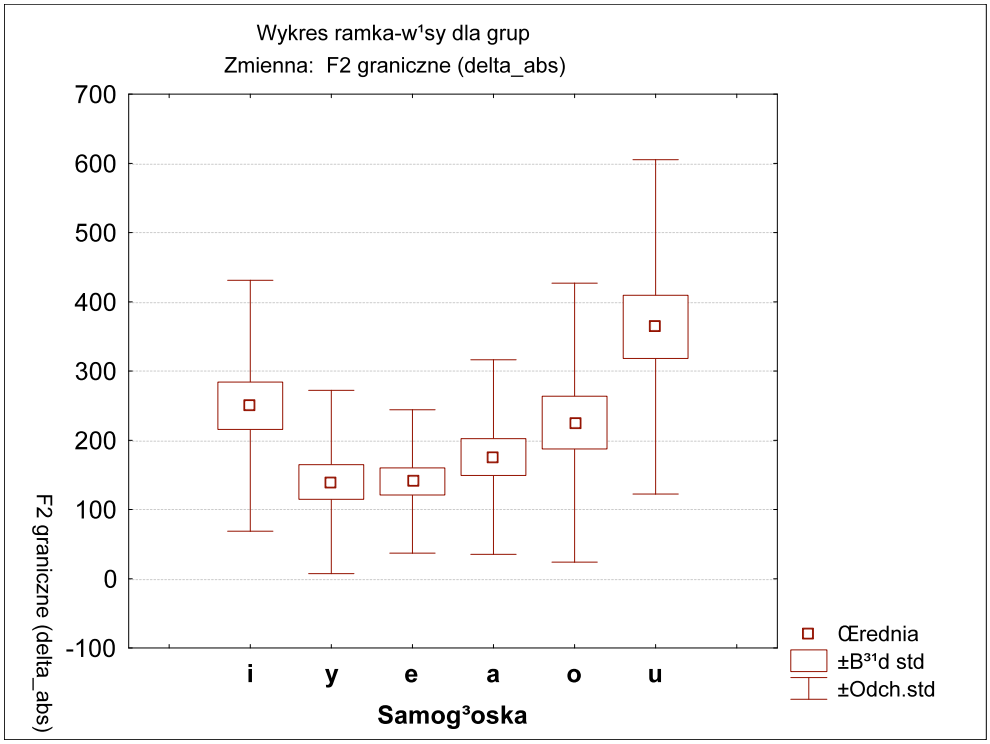
Autor przeprowadził wiele doświadczeń związanych z tym zjawiskiem – dotyczyły one wpływu różnych spółgłosek na poprzedzające samogłoski będące ośrodkami sylab otwartych. Wyniki tych doświadczeń były pomocne przy konstruowania testów percepcyjnych, będą one również istotne dla przyszłych eksperymentów.

Jedno z doświadczeń dotyczyło oddziaływania poszczególnych spółgłosek na wartości graniczne formantu drugiego poprzedzających je samogłosek (ośrodków sylab). Uzyskane wyniki analizy skupień wskazały, że spółgłoski o takim samym miejscu artykulacji wywierają podobny wpływ na samogłoski (należące do poprzedniej sylaby otwartej), a więc powodują, że samogłoski te osiągają podobne wartości graniczne formantu drugiego (rys. 7.1). Wyniki te stanowią potwierdzenie teorii lokusa. Mogą mieć one znaczenie dla syntezy mowy opartej na sylabach – mogą one sugerować, że dana sylaba użyta w pewnym kontekście powinna być pobrana z podobnego kontekstu. Z tym zagadnieniem były związane doświadczenia audytywne, których wyniki przedstawiono w dalszym ciągu tego rozdziału.



Rys. 7.1 Wynik analizy skupień (przeprowadzonej metodą aglomeracji) – podobieństwo oddziaływania poszczególnych spółgłosek na wartości graniczne formantu drugiego poprzedzających je samogłosek (ośrodków sylab)

Inne doświadczenie dotyczyło tego, jak mocno poszczególne samogłoski są podatne na wpływ spółgłosek. Porównywano wartości formantu drugiego w części ustalonej samogłoski oraz w miejscu połączenia z następującą spółgłoską (uwzględniono wszystkie spółgłoski). Rysunek 7.2 obrazuje średnie wartości różnic tych dwóch wartości (podano je w hercach).



Rys. 7.2 Podatność samogłosek na wpływ następujących spółgłosek

7.2 Doświadczenia percepcyjne

W dalszym ciągu przedstawiono wyniki dwóch obszernych eksperymentów audytywnych związanych z wykorzystaniem sylaby jako podstawowej jednostki syntezy mowy dla języka polskiego. Pierwszy eksperyment dotyczył syntetycznych połączeń typu V_C, natomiast drugi odnosił się do połączeń typu C_C.

7.2.1 Założenia dla doświadczeń percepcyjnych

W eksperymentach wykorzystano nagrania wyrazów. Nagrywana była osoba płci męskiej o bardzo dobrej wymowie. Na podstawie tych nagrań utworzono pewien zbiór wyrazów. Wyrazy te otrzymano poprzez połączenie dwóch części pochodzących z dwóch różnych nagrań (z dwóch różnych wyrazów). Otrzymane w ten sposób syntetyczne wyrazy były poddawane ocenie osób biorących udział w eksperymentach. Osoby te osłuchiwały kolejne pliki dźwiękowe. Każdemu plikowi odpowiadał zapis ortograficzny, przy czym w tym zapisie zaznaczono miejsce połączenia. Na tym miejscu osoby biorące udział w eksperymencie miały koncentrować swoją uwagę. Poza tym każdy połączony wyraz miał swój identyczny odpowiednik, który nie był łączony (było to nagrania całego wyrazu). Jednak osoby biorące udział w eksperymencie były błędnie informowane (zgodnie z zamierzeniem autora), że te wyrazy również były łączone (wskazano te same miejsca połączenia, co w odpowiednich wyrazach zmodyfikowanych). Użycie nagrań wyrazów pozornie zmodyfikowanych stanowiło punkt odniesienia dla ocen wyrazów łączonych. Jeżeli średnia ocena wystawiana danemu wyrazowi łączonemu była gorsza od średniej oceny wyrazu całego, to oznaczało to, że resynteza była wyczuwalna.

7.2.2 Analiza wyników doświadczeń percepcyjnych

Przeprowadzono dwa obszerne doświadczenia – pierwsze z nich dotyczyło syntetycznych połączeń typu V_C (połączenie samogłoski będącej ośrodkiem sylaby otwartej ze spółgłoską należącą do następnej sylaby), natomiast drugie doświadczenie dotyczyło

połączeń typu C_C (połączenie spółgłoski znajdującej się w wygłosie sylaby zamkniętej ze spółgłoską należącą do nagłosu następnej sylaby). Dla połączeń typu V_C uzyskano następujące wyniki:

- połączenie samogłoski (ośrodka sylaby) ze spółgłoską zwartą bezdźwięczną – odsetek par wyrazów, w których zaobserwowano gorszą jakość w połączeniach syntetycznych wyniósł 14,3% (w dwóch spośród czternastu par wyrazów);
- połączenie samogłoski ze spółgłoską zwartą dźwięczną – odsetek par, w których syntetyczne połączenie miało gorszą jakość wyniósł 9,1% (w jednej z jedenastu par). Wyniki te wskazują, że syntetyczne połączenia samogłosek (będących ośrodkami sylab otwartych) ze spółgłoskami zwartymi (bezdźwięcznymi i dźwięcznymi) są prawie zawsze niesłyszalne. Oczywistym wytłumaczeniem jest zwarcie poprzedzające plozję;
- połączenie samogłoski ze spółgłoską szczelinową bezdźwięczną – tutaj odsetek par z pogorszoną jakością w wyrazie zawierającym syntetyczne połączenie wyniósł 7,7% (w jednej spośród trzynastu par). Wynik ten wskazuje, że ewentualne różnice w wartościach granicznych formantu drugiego nie mają wpływu na pogorszenie jakość syntetycznego połączenia;
- połączenie samogłoski ze spółgłoską szczelinową dźwięczną – tutaj odsetek par w których zaobserwowano pogorszenie w wyrazie z syntetycznym połączeniem był znacznie większy – wyniósł on 64,3% (w dziewięciu spośród czternastu par). Pogorszenie jakości związane jest z połączeniem dwóch sygnałów dźwięcznych. Trzeba zaznaczyć w tych eksperymentach nie użyto żadnych algorytmów wygładzających (dopasowujących do siebie) łączone fragmenty dźwięczne – dlatego powstałe pogorszenia jakości mogą wynikać z różnicy w kształcie widma łączonych sygnałów dźwięcznych, różnej częstotliwości podstawowej lub niezachowaniu okresowości drgania. Trzeba jednak zaznaczyć, że zaobserwowane pogorszenie jakości w tym rodzaju połączeń nie było duże;
- połączenie samogłoski ze spółgłoską nosową – odsetek par z gorszymi połączeniami syntetycznymi wyniósł aż 88,9% (osiem z dziewięciu par). Oprócz opisanego problemu związanego z łączeniem dwóch sygnałów dźwięcznych, pogorszenie jakości może wynikać z zastępowania samogłoski unosowionej przez nieunosowioną;
- połączenie samogłoski z aproksymantem – tutaj odsetek par w których jakość dla syntetycznych połączeń była gorsza wyniósł 94,7% (w osiemnastu z dziewiętnastu par). Poza tym pogorszenie to było znacznie większe niż w innych analizowanych typach połączeń. Oprócz opisanych problemów wynikających z łączenia sygnałów dźwięcznych, można przypuszczać, że tego typu syntetyczne połączenia są bardzo wrażliwe na zmiany przebiegu formantu drugiego – wynika to z tego, że pomiędzy samogłoską a aproksymantem nie ma wyraźnej granicy, tylko łagodny obszar przejścia (związany ze stopniową zmianą ustawień narządów mowy).

Następne doświadczenie dotyczyło oceny syntetycznych połączeń typu C_C (spółgłoska-spółgłoska). Doświadczenie to zostało przygotowane oraz przeprowadzone w sposób podobny do doświadczenia dotyczącego połączeń typu V_C. Przeprowadzone eksperymenty wykazały, że łączenie spółgłosek przynosi bardzo dobre rezultaty. W prawie wszystkich przypadkach takie połączenia w ogóle nie były wyczuwalne dla osób biorących udział w badaniach. Pogorszenie jakości, najczęściej niezbyt duże, było słyszalne przy niektórych połączeniach aproksymantów. Jednak biorąc pod uwagę fakt, że tego typu połączenia są rzadkie, trzeba uznać, że połączenie typu C_C jest bardzo dobrym rozwiązaniem i rozwiązanie to może być wykorzystywane w syntezie mowy zawsze

wtedy, kiedy łączy się sylabę zamkniętą z kolejną sylabą (rozpoczynającą się od spółgłoski).

Uzyskane wyniki badań są obiecujące. Większość syntetycznych połączeń miała dobrą jakość (nieodróżnialną od połączeń niezmienionych). Największe problemy wystąpiły przy syntetycznych połączeniach sylab otwartych ze spółgłoskami nosowymi i z aproksymantami. Jednak z pewnością można szukać skutecznych rozwiązań – oprócz stosowania modyfikacji sygnału (algorytmów wygładzających), można na przykład stosować pary sylab, zawierające wrażliwe połączenia.

8 Podsumowanie

W artykule przedstawiono fragmenty rozprawy doktorskiej: *Fonetyczno-akustyczna analiza struktury sylaby w języku polskim na potrzeby technologii mowy*. Przytoczono najistotniejsze fragmenty z części teoretycznej pracy (dotyczące definicji sylaby). W skróconej formie opisano wyniki badań. Na podstawie uzyskanych wyników można śmiało stwierdzić, że sylaba jest niezwykle cenną jednostką dla celów praktycznych. Problemy dotyczące definiowania sylab można rozwiązać poprzez przyjmowanie pewnych ustaleń umownych. Na pewno warto kontynuować badania związane z wykorzystaniem tych jednostek dla aplikacji syntezy oraz rozpoznawania mowy języka polskiego. Według autora do najważniejszych osiągnięć związanych z niniejszą rozprawą należą:

- zaproponowanie własnej koncepcji sylaby przydatnej dla celów praktycznych,
- utworzenie reguł podziału na sylaby (dla ponad 2500 zbitek języka polskiego),
- stworzenie rozbudowanych baz nagrań (obejmujących kilkadziesiąt tysięcy plików),
- stworzenie kilku przydatnych programów,
- przeprowadzenie rozległych eksperymentów związanych z potencjalną możliwością wykorzystania sylaby w systemach rozpoznawania mowy dla języka polskiego,
- przeprowadzenie doświadczeń dotyczących koartykulacji w języku polskim,
- przeprowadzenie rozległych badań związanych z potencjalną możliwością wykorzystania sylaby w syntezie mowy.

Bibliografia:

Chalker Sylvia, Weiner Edmund. *The Oxford Dictionary of English Grammar*. Oxford University Press. New York, 1994. s. 387;
Clark John, Yallop Collin. *An Introduction to Phonetics and Phonology*. Blackwell Publishers. Oxford, 1995 (wyd. II). s. 67-68;
Demenko Grażyna, Wypych Mikołaj, Baranowska Emilia. *Implementation of Grapheme-to-Phoneme Rules and Extended SAMPA Alphabet in Polish Text-to-Speech Synthesis*. W: Speech and Language Technology, vol. 7. PTFon. Poznań 2003. s.79-95;
Dziubalska-Kołaczyk Katarzyna. *Beats-and-Binding Phonology*. Peter Lang GmbH. Frankfurt am Main, 2002. s. 39-51;
Francuz Piotr, Mackiewicz Robert. Liczby nie wiedzą skąd pochodzą. Wydawnictwo KUL. Lublin 2007. s.3-576;
Jassem Wiktor. *Mowa a nauka o łączności*. Państwowe Wydawnictwo Naukowe. Warszawa, 1974. s. 75-141;
Ladefoged Peter. *A course in phonetics*. Harcourt Brace Jovanovich. London 1975. s. 217;
Łobacz Piotra, Jassem Wiktor. *Fonotaktyczna analiza mówionego tekstu polskiego*. Biuletyn PTJ XXXII. 1974. s. 179-197;
Łobacz Piotra. *Badania fonostatyczne na potrzeby syntezy mowy*. W: Speech and Language Technology, vol. 6. PTFon. Poznań, 2002. s. 81-112;
Michowska Ewa, Wasielczyk Krystyna. *Encyklopedia językoznawstwa ogólnego*. Zakład Narodowy im. Ossolińskich – Wydawnictwo. Wrocław, 1999. s. 575;
Tabakowska Elżbieta (redakcja). *Kognitywne podstawy języka i językoznawstwa*. Towarzystwo Autorów i Wydawców Prac Naukowych UNIVERSITAS. Kraków, 2001. s 165-166;
Tadeusiewicz Ryszard. *Elementarne wprowadzenie do techniki sieci neuronowych z przykładowymi programami.* Akademicka Oficyna Wydawnicza. Warszawa, 1998. s. 1-164;
Tadeusiewicz Ryszard, Gąciarz Tomasz, Borowik Barbara, Leper Bartosz. *Odkrywanie właściwości sieci neuronowych*. Polska Akademia Umiejętności. Kraków, 2007. s. 9-426;
Trask R.L. *A Dictionary of Phonetics and Phonology*. Routlege. New York, 1996 (wyd.II). s. 327, 345 (tłumaczenie autora);
Rogers Henry. *The Sounds of Language.* Pearson Education. London, 2000. s. 89;
Strutyński Janusz. *Gramatyka polska*. Wydawnictwo Tomasz Strutyński. Kraków, 2002 (wyd. V). s. 28-63;
Wierzchowska Bożena. *Wymowa polska.* Państwowe Zakłady Wydawnictw Szkolnych. Warszawa, 1971 (wyd. II). s 102-197, 213-216;