INVESTIGATIONES LINGUISTICAE VOL. XXXII, 2015

AL. NIEPODLEGŁOŚCI 4, 60-874, POZNAŃ – POLAND

Percepcja audytywna, właściwości akustyczne oraz cechy dystrybucyjne sylab w języku polskim

Auditive perception, acoustic and distributional properties of syllables in Polish

Daniel Śledziński

INSTYTUT JĘZYKOZNAWSTWA, UNIWERSYTET IM. ADAMA MICKIEWICZA
AL. NIEPODLEGŁOŚCI 4, 61-874 POZNAŃ

fon@amu.edu.pl

Abstract

This paper presents experiments concerning properties of selected CV syllables. Acoustic speech signal related to particular syllables was analyzed using artificial neural networks. The goal of the analyses was to investigate whether realizations of particular syllables retain acoustic features distinctive of these syllables. Aditionally, a perception test aiming at identification of the same syllable set was carried out. In the test we analyzed to which degree it is possible to identify syllables isolated from the linguistic context. The paper also discusses results on distributional properties of syllables which indicate that such properties may play a significant role in speech perception.

1. Wprowadzenie

W artykule omówiono wyniki doświadczeń związanych z właściwościami wybranych sylab typu CV. Eksperymenty prowadzono na dwóch płaszczyznach: badano akustyczny sygnał mowy związany z poszczególnymi sylabami oraz wykonano test percepcyjny polegający na identyfikacji tego samego zbioru sylab. Wykorzystano sztuczne sieci neuronowe dla badania sygnału akustycznego – wykonywały one zadanie klasyfikacyjne dychotomiczne. Celem tych eksperymentów było sprawdzenie czy realizacje poszczególnych sylab zachowują wyróżniające cechy akustyczne typowe tylko dla tych sylab. Natomiast przy pomocy testu percepcyjnego sprawdzano w jakim stopniu możliwa jest identyfikacja realizacji sylab wyciętych z kontekstu. Można powiedzieć, że takie porównanie stanowi konsolidację dwóch kierunków badań. Pierwszy z nich związany jest z możliwością uży-

cia sylaby jako podstawowej jednostki w systemach automatycznego rozpoznawania mowy (Kopecek 1999). Drugi kierunek obejmuje badania nad percepcją mowy – rola sylaby w percepcji mowy to temat licznych badań (Dupoux 1993), (Massaro 1974), (Mehler 1981), (Segui 1984). Również wpływ własności akustycznych segmentów mowy na ich percepcję był wielokrotnie badany, jednak w badaniach tych brano pod uwagę określone parametry sygnału mowy takich jak iloczas, VOT, częstotliwość podstawową (Hirihara, Kato 1992), (Liberman i inni 1952), (Lisker, Abramson 1967), (Ohde 1984), (Raphael, Dorman 1980), (Raphael i inni 1975).

W opisanych badaniach użyto parametrów akustycznych wykorzystywanych w systemach automatycznego rozpoznawania mowy – parametrów które odzwierciedlają cechy spektralne właściwe dla poszczególnych dźwięków mowy. Dzięki temu można wyciągać wnioski dotyczące właściwości sygnału akustycznego związanego z poszczególnymi sylabami – cech wyróżniających właściwych dla tych sylab. Równoległe badania percepcji umożliwiają konfrontacje własności sylab na płaszczyźnie akustycznej oraz percepcyjnej.

W artykule omówiono również wyniki dotyczące łączliwości sylab. Wyniki te wskazują, że cechy dystrybucyjne sylab mogą pełnić istotną rolę w przetwarzaniu zstępującym w percepcji mowy. Połączenie wszystkich wymienionych kierunków badań może być przydatne dla dyskusji dotyczącej udziału przetwarzania wstępującego oraz przetwarzania zstępującego w percepcji mowy.

2. Materiał badawczy

Omówione w artykule eksperymenty wymagały przygotowania obszernych zbiorów nagrań. Nagrano zarówno głosy męskie, jak i głosy żeńskie (udział obu płci był jednakowy). Wiek nagrywanych osób był zróżnicowany – byli to zarówno studenci, jaki i nauczyciele akademiccy. Utworzone zbiory obejmują:

- nagrania wyrazów izolowanych – zawierają łącznie około 26000 realizacji sylab. Nagrano 30 osób. Każda osoba przeczytała 332 wyrazy, zatem zbiór zawiera łącznie około 10000 nagranych wyrazów;
- nagrania mowy 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 mowy ciągłej zawierają łącznie około 18000 realizacji sylab. Zatem wszystkie nagrania (wyrazów izolowanych oraz mowy ciągłej) obejmują około 42000 realizacji sylab.

Zarówno wyrazy, jak i zdania zostały dobrane w ten sposób, żeby występowało w nich jak najwięcej określonych sylab typy CV (sylab złożonych z jednej spółgłoski i samogłoski). Nagrany materiał dźwiękowy wymagał dalszej obróbki. Nagrania zostały pocięte – każdy wyraz izolowany oraz każde zdanie zostało umieszczone w osobnym pliku dźwiękowym. Następnie naniesiono granice czasowe sylab na sygnał akustyczny – był to naj-

bardziej pracochłonny etap badań. Transkrypcja na poziomie sylab z informacjami o granicach czasowych kolejnych segmentów była zapisywana do oddzielnych plików (dla każdego pliku dźwiękowego został utworzony jeden plik z takimi informacjami). Na końcu akustyczny sygnał mowy poddano parametryzacji. Obliczano zestaw 12 wartości współczynników cepstralnych w siedmiu, dziesięciu lub czternastu punktach czasowych poszczególnych realizacji sylab. Dzięki temu akustyczna postać każdej nagranej sylaby była reprezentowana przez zestaw odpowiednio 84, 120 lub 168 wartości liczbowych. Obliczone w ten sposób wartości były zapisywane do oddzielnych plików (podobnie jak informacje o granicach czasowych segmentów).

Dla badań związanych z cechami dystrybucyjnymi sylab posłużono się korpusem tekstowym złożonym z dwóch milionów wyrazów. Teksty w korpusie pochodzą z różnych źródeł – z artykułów prasowych, z tekstów zamieszczonych w Internecie oraz z książek. Nie odnoszono się do znaczenia zdań, fraz oraz poszczególnych wyrazów. Przed przystąpieniem do badań wyrazy w korpusie zostały wymieszane oraz przetranskrybowane fonematycznie (Demenko i inni 2003).

Wprowadzono również podział na sylaby oparty na własnych rozwiązań umownych – dla języka polskiego istnieje kilka propozycji sylabizacji wyrazów i żadne z tych rozwiązań nie jest doskonałe ze względu na występujące licznie grupy spółgłoskowe o strukturze niespotykanej w innych językach. Również dostępne definicje sylaby nie umożliwiają przeprowadzenia podziału jednoznacznego (Trask 1996), (Ladefoged 1975). Przyjęte przez autora rozwiązanie oparte jest na przesłankach praktycznych związanych z efektywnym podziałem sygnału mowy na sylaby.

3. Metody badań

3.1. Badania akustycznych właściwości sylab

W badaniach wykorzystano sztuczne sieci neuronowe, które wykonywały zadanie o charakterze klasyfikacyjnym dychotomicznym (użyto pakietu Statistica Naural Networks firmy Statsoft). Aby sieć neuronowa mogła klasyfikować dane, najpierw trzeba przeprowadzić trening tej sieci przy użyciu algorytmu uczącego. Uczenie perceptronu wielowarstwowego (najczęściej wykorzystywanego rodzaju sieci neuronowej) polega na odnalezieniu optymalnej kombinacji wag neuronów – takiej, przy której klasyfikacja jest najbardziej skuteczna. Dla treningu sieci trzeba dysponować odpowiednio dużą liczbą przypadków uczących, z których każdy obejmuje dane niezależne – podawane na wejściu sieci neuronowej oraz odpowiadającą tym danym informację o przynależności do określonej klasy. Na wstępie wszystkie dane dzielone się na trzy zbiory: uczący, walidacyjny oraz testowy. Atutem sztucznych sieci neuronowych jest to, że wytrenowana sieć potrafi klasyfikować przypadki podobne do tych, który były użyte w czasie treningu, ale nie koniecznie identyczne. Trzeba zaznaczyć że wszystkie informacje potrzebne do wytrenowania sieci zawarte są w danych uczących.

Natomiast brak możliwości uzyskania poprawnej klasyfikacji danych może świadczyć o tym że poszczególne klasy przypadków uczących nie posiadają wystarczających cech wyróżniających te klasy (danych wejściowych dla sieci neuronowych). W opisywanych badaniach oprócz perceptronów wielowarstwowych (MLP), sprawdzono działanie sieci probabilistycznych (PNN), sieci o radialnych funkcjach bazowych (RBF) oraz sieci liniowych (Linear). Bardziej szczegółowe informacje dotyczące funkcjonowania poszczególnych rodzajów sieci neuronowych można znaleźć w bogatej literaturze dotyczącej tego zagadnienia (Tadeusiewicz 1998, 2001).

W omawianych badaniach każda realizacja sylaby była reprezentowana przez zestaw 84, 120 lub 168 wartości liczbowych (por. rozdz. 2). Takie zestawy wartości stanowiły zmienne niezależne, które były podawane na wejściach sieci neuronowych. Klasyfikacja dychotomiczna zakłada przypisywania danych niezależnych do jednej z dwóch kategorii. 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 w czasie treningu poszczególnych sieci neuronowych 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 (innych od sylaby rozpoznawanej). Przyjęcie takich założeń pozwoliło na traktowanie sieci neuronowych jako detektorów potrafiących wykrywać w sygnale akustycznym określone struktury sylab i – a więc potrafiących odpowiednio reagować tylko wtedy, kiedy one rzeczywiście wystąpiły.

Przedstawiona wyżej metoda związana jest z następującym pytaniem dotyczącym sylab: 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 reagować na 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. Omówioną metodę badawczą zastosowano zarówno w odniesieniu do sylab pobranych z wyrazów izolowanych, jak do sylab pochodzących z mowy ciągłej.

3.2. Badania percepcji sylab

W ramach opisywanych badań przeprowadzono test percepcyjny. Polegał on na identyfikacji sylab wyciętych ze zdań (z mowy ciągłej). Kolejne bodźce (zapisy dźwiękowe sylab) były odtwarzane a zadaniem osób uczestniczących w badaniu było wpisanie na klawiaturze komputera sylaby, którą usłyszeli. Była możliwość ponownego odtworzenia bodźca. W tym eksperymencie uczestniczyło 30 osób – każda z badanych osób była poddana testowi percepcyjnemu dwukrotnie a pojedynczy test dla jednej osoby składał się ze 100 bodźców. Zatem w sumie wykorzystano 6000 bodźców i trzeba podkreślić, że wszystkie bodźce użyte w teście były różne – zatem były to różne realizacje sylab – odczytywane przez różne osoby. Jest to czynnik istotny, ponieważ powtarzanie tego samego bodźca (konkretnej realizacji danej sylaby nagranej przez daną osobę) stwarzałoby zagrożenie, że wyniki odzwierciedlałyby cechy właściwe tylko dla tej konkretnej realizacji sylaby – mogłyby one wynikać z indywidualnych cech artykulacyjnych mówcy a także przyspieszonej, zwolnionej lub niewyraźnej artykulacji.

Podstawowe pytanie związane z badaniem percepcyjnym jest rozwinięciem pytania postawionego dla badań związanych z właściwościami akustycznymi sylab (omówionych w podrozdziale 3.1). Pytanie to można sformułować następująco: czy wyjęte z kontekstu realizacje danej sylaby posiadają cechy akustyczne umożliwiające identyfikację tych sylab na drodze percepcji audytywnej.

3.3. Badania cech dystrybucyjnych sylab i fonemów

Badania cech dystrybucyjnych oparto na dwumilionowym korpusie tekstowym. Porównano wyniki dotyczące łączliwości sylab z wynikami dotyczącymi łączliwości fonemów. Z właściwości dystrybucyjnych sylab wynikają duże możliwości określania wyrazów – był to przedmiot kolejnych badań omówionych w artykule.

4. Wyniki badań

4.1. Porównanie wyników klasyfikacji dla sylab pobieranych z wyrazów izolowanych oraz mowy ciągłej

Przeanalizowano 43 często występujące sylaby pod kątem możliwości ich wykrywania przez sztuczne sieci neuronowe. Konkretne realizacje tych sylab były pobierane (wycinane) z nagranych wyrazów izolowanych. Wytrenowano 172 sieci neuronowe (dla 43 sylab sprawdzono działanie czterech rodzajów sieci neuronowych wymienionych w rozdz. 3.1). W tabeli 1 zamieszczono wyniki czterech przykładowych doświadczeń, które dotyczą sylaby /mo/. Tabela zawiera informacje o rodzaju i strukturze sieci, osiągniętej jakości dla zbiorów: uczącego, walidacyjnego oraz testowego oraz macierz pomyłek. Analizując informacje umieszczone w macierzy pomyłek

można zauważyć, że perceptron wielowarstwowy dobrze spełniał rolę detektora reagującego na sylabę /mo/ (w podanym przykładzie 385 z 389 realizacji sylaby /mo/ zostało wykrytych przez perceptron wielowarstwowy). Jednak 375 z 4976 realizacji sylab losowych zostało zaklasyfikowanych jako sylaba /mo/ – wynika z tego, akustyczna struktura niektórych sylab może być zbliżona do realizacji sylaby /mo/.

Tabela nr 1.: Wyniki klasyfikacji dychotomicznej dla sylaby mo.

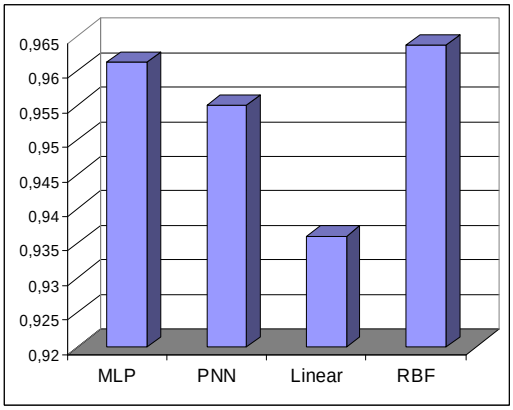
RODZAJ I STRUKT. SIECI	U	W	T	ROZP. JAKO:	random	mo
MLP 120: 120-120-1:1	0,934	0,920	0,931	random	4601	4
				mo	375	385
PNN 120: 120-2683-2:1	1,000	0,983	0,983	random	4942	12
				mo	34	377
RBF 90:90-129-1:1	0,983	0,975	0,968	random	4880	25
				mo	96	364
Linear 112:112-1:1	0,972	0,968	0,961	random	4833	29
				mo	143	360

Najlepszą średnią jakość sieci (w zbiorze testowym) uzyskano dla sieci probabilistycznych. Nieco gorzej wypadły perceptrony wielowarstwowe (94%) oraz sieci liniowe (93%). Jednak z punktu widzenia założeń przyjętych dla 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 (jest to odsetek realizacji określonej sylaby, który dana sieć zdołała wykryć). Średnią wartość parametru wykrywalność dla sylab pochodzących z wyrazów izolowanych przedstawiono na rysunku 1.

Następne badanie dotyczyło sylab pobranych z mowy ciągłej (przetestowano działanie 180 sieci neuronowych). Dla każdej sylaby sprawdzono trzy warianty sieci probabilistycznych – jeden wariant bez macierzy strat oraz dwa warianty 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).

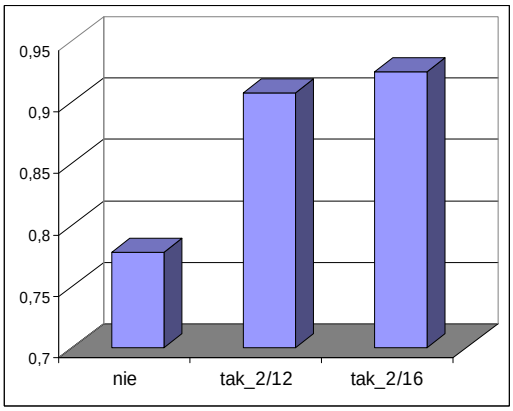
Z rysunku 2 wynika, że średnia wykrywalność w wersji bez macierzy strat wyniosła około 77%, jednak zastosowanie macierzy strat znacznie polepszyło ten rezultat. Z drugiej strony zastosowanie macierzy strat nieznacznie pogorszyło jakość sieci (z 99% do około 98%). Wynika to z faktu zwiększenia liczby sylab ze zbioru sylab losowych, klasyfikowanych do klasy sylaby, która miała być wykrywana przez daną sieć.

Rysunek nr 1.: Średnia wykrywalność sylab pobranych z wyrazów izolowanych (zbiór testowy)



Z przedstawionych danych wynika, że średnia wykrywalność dla sylab wycinanych z wyrazów izolowanych wyniosła około 95% (przy użyciu sieci probabilistycznych bez zaimplementowanej macierzy strat), natomiast analogiczna wartość dla sylab pobieranych z mowy ciągłej wyniosła około 77%. Dane te świadczą o tym, że ilość wyróżniających (relewantnych) cech akustycznych zawartych w poszczególnych segmentach (sylabach) zmniejsza się w mowie ciągłej (w stosunku do wyrazów izolowanych).

Rysunek nr 2.: Średnia wykrywalność pobranych z mowy ciągłej (zbiór testowy)



4.2. Wyniki badań percepcyjnych

W tabeli 2 zamieszczono wyniki uzyskane z testu percepcyjnego. Każda osoba biorąca udział w teście odpowiadała na zestaw losowych bodźców, dlatego wyniki pogrupowano według konkretnych sylab. Następnie w ramach wyników dotyczących poszczególnych sylab obliczono ile zostało udzielonych odpowiedzi bezbłędnych oraz błędnych. Wśród odpowiedzi błędnych wyróżniono:

- odpowiedzi, które zawierały błędnie rozpoznany ośrodek sylaby,
- odpowiedzi z błędnie podanym nagłosem sylaby,
- odpowiedzi w których nagłos oraz ośrodek sylaby zostały podane prawidłowo, jednak dodany został wygłos sylaby,
- inne odpowiedzi – czyli odpowiedzi w które zawierają co najmniej dwa odstępstwa od odpowiedzi bezbłędnych (na przykład błędnie podany nagłos oraz dodany wygłos sylaby).

Tabela nr 2.: Wyniki uzyskane z badań percepcyjnych.

LP.	SYLABA	DOBRA IDENTY-FIKACJA	BŁĘDNIE ROZP. OŚRO-DEK SY-LABY	BŁĘDNIE ROZP. NAGŁOS SYLABY	ROZP. JAKO SYLABA ZAMKN.	INNE	ŚREDNIA LICZBA POWT.
1	/la/	61,5%	0,0%	15,0%	8,0%	15,5%	1,45
2	/Sa/	44,0%	13,5%	0,5%	16,0%	26,0%	1,53
3	/na/	68,0%	10,0%	0,5%	0,0%	10,5%	1,43
4	/by/	73,5%	5,0%	9,5%	5,0%	7,0%	1,48
5	/ma/	65,5%	4,0%	6,5%	17,5%	6,5%	1,52
6	/to/	71,0%	11,0%	3,0%	4,0%	11,0%	1,50
7	/va/	67,6%	8,0%	1,5%	18,0%	5,0%	1,58
8	/vy/	82,0%	3,0%	0,5%	7,5%	7,0%	1,32
9	/ka/	74,5%	3,0%	3,0%	12,5%	7,0%	1,41
10	/wa/	40,5%	7,5%	9,5%	6,5%	36,0%	1,89
11	/po/	52,0%	9,0%	18,0%	7,5%	13,5%	1,74
12	/go/	52,0%	8,5%	7,0%	12,5%	20,0%	1,80
13	/ra/	49,0%	4,5%	10,5%	9,0%	27,0%	1,88
14	/ko/	58,0%	3,5%	4,0%	26,5%	8,0%	1,46
15	/n`e/	45,5%	6,5%	0,5%	4,0%	43,5%	1,44
16	/je/	28,5%	19,0%	0,0%	2,0%	50,5%	2,02
17	/mo/	57,0%	11,0%	1,5%	17,5%	13,0%	1,74
18	/no/	65,5%	12,0%	3,0%	6,0%	14,0%	1,48
19	/sa/	40,5%	14,0%	1,0%	21,5%	23,0%	1,63
20	/ry/	70,0%	10,0%	2,0%	4,0%	14,0%	1,54
21	/ta/	42,5%	21,5%	6,5%	12,0%	17,5%	1,67
22	/c`e/	55,5%	15,5%	6,5%	2,5%	20,0%	1,40
	Średnia	57,5%	9,1%	5,0%	10,0%	18,0%	

Z danych zawartych w tabeli 2 wynika, że osoby biorące udział w teście percepcyjnym nie były w stanie poprawnie zidentyfikować wszystkich realizacji sylab wyciętych z mowy ciągłej. Średni odsetek odpowiedzi bezbłędnych (biorąc pod uwagę wszystkie odpowiedzi udzielone przez wszystkie badane osoby) wyniósł 57,5%. Pozostałe odpowiedzi nie były poprawne, jednak ich odstępstwa od poprawności były zróżnicowane – dużo udzielonych odpowiedzi było podobnych do odpowiedzi poprawnych. Średni od-

setek odpowiedzi, które zawierały błędnie rozpoznany ośrodek sylaby wyniósł 9,1%, natomiast średni odsetek odpowiedzi z błędnie rozpoznanym nagłosem sylaby wyniósł 5%. Średni odsetek odpowiedzi z dodanym wygłosem sylaby wyniósł dokładnie 10% – takie odpowiedzi mogą wynikać ze zjawiska koartykulacji oraz prawostronnego ugięcia formantów w ośrodkach sylab typu CV – można tutaj dostrzec analogię do znanego eksperymentu Coopera, który dotyczył jednak lewostronnego ugięcia formantów w ośrodku sylaby oraz możliwości identyfikacji sylaby tylko na podstawie tego ugięcia (Cooper i inni 1952). Stosunkowo dużo (średnio 18%) udzielonych odpowiedzi zawierało więcej niż jedno odstępstwo od odpowiedzi poprawnych. W tej grupie niektóre odpowiedzi miały pewne cechy wspólne odpowiedziami poprawnymi, natomiast niektóre bardzo się różniły. W tabeli 3 zaprezentowano szczegółowe wyniki testu percepcyjnego dla kilku pozycji (sylab) zamieszczonych w tabeli 2. Dane te dają pogląd na udział i rodzaje błędów popełnianych przez osoby biorące udział w eksperymencie.

Przeglądając dane zawarte w tabeli 2 można zauważyć, że identyfikacja niektórych sylab wyciętych z kontekstu sprawiała mniej problemów (na przykład sylab złożonych ze spółgłoski nosowej oraz samogłoski). Zaobserwowano również sylaby, których identyfikacja była mniej skuteczna (na przykład sylab złożonych z półsamogłoski oraz samogłoski). Poza tym zaobserwowano zależność między odsetkiem odpowiedzi poprawnych (dla poszczególnych sylab) a średnią liczbą powtórzeń danego bodźca – cechy te są silnie skorelowane ujemnie (współczynnik korelacji wyniósł −0.71). Zatem liczba powtórzeń bodźców odzwierciedla możliwości identyfikacji sylab wyciętych z kontekstu.

Tabela nr 3.: Przykładowe szczegółowe wyniki uzyskane z testu percepcyjnego.

ODPOWIEDŹ[1]	LICZBA	ODSETEK	ODPOWIEDZI OSÓB BIORĄCYCH UDZIAŁ W TEŚCIE
Sylaba /la/, średnia liczba powtórzeń: 1,45			
A	123	61,5%	la
B	0	0,0%	-
C	30	15,0%	wa, na, ba, ra, ja, da, ta, ła
D	16	8,0%	lap, lam, lat, lab, lad, lam
E	31	15,5%	nad, nal, no, ze, ne, ała, ola, ne, dla, ala, to, a, laty, bla, lot, rab, els, lam, dla, we, we, not, nad, ula, ły, aa

1 W kolumnie ODPOWIEDŹ tabeli nr 3 użyto liter, które mają następujące znaczenie: A – poprawna odpowiedź, B – błędnie rozpoznany ośrodek sylaby, C – błędnie rozpoznany nagłos sylaby, D – dodany wygłos sylaby (odpowiedź zawiera wygłos którego nie było), D – inne odpowiedzi.

Sylaba /Sa/, średnia liczba powtórzeń: 1,53			
A	88	44,0%	sza
B	27	13,5%	sze, szo, szy
C	1	0,5%	fa
D	32	16,0%	szak, szan, shat, szaj
E	52	26,0%	psza, szej, sha, szek, sua, prza, prze, ksza, cza, sia, sa, rza, szwe
Sylaba /na/, średnia liczba powtórzeń: 1,43			
A	136	68,0%	na
B	20	10,0%	no
C	3	0,5%	ma
D	0	0,0%	-
E	21	10,5%	pek, nie, ten, pan, mol, nia, oła, wań, ten, noł, nieu, nod, fla, pno, nio, nie
Sylaba /ma/, średnia liczba powtórzeń: 1,52			
A	131	65,5%	ma
B	8	4,0%	mo, my
C	13	6,5%	ba, wa, la, ła, na
D	35	17,5%	mat, mag, man, mal, mad, mat, mak
E	13	6,5%	hmak, ama, ly, mua, kma, mel, moj, ema, młot, łat, nad
Sylaba /to/, średnia liczba powtórzeń: 1,50			
A	142	71,0%	to
B	22	11,0%	ty, tu, te, ty, tw, ta
C	6	3,0%	do, co, fo
D	8	4,0%	toj, tob, top, tok
E	22	11,0%	są, tam, kot, fy, two, taj, tak, tyn, ky, tso, foj, fe, cop, tut, t, pot, tup, dał, dop, toi
Sylaba /va/, średnia liczba powtórzeń: 1,58			
A	135	67,5%	wa
B	16	8,0%	wy, we, wo, wę
C	3	1,5%	ła, ma
D	36	18,0%	wap, war, wań, wam, wan, waj, wal, wat, wał
E	10	5,0%	dwa, włu, owa, wop, woj, ot, ala

4.3. Porównanie wyników klasyfikacji dychotomicznej i wyników testu percepcyjnego

W tabeli 4 zamieszczono wyniki badań percepcyjnych (odsetek bezbłędnej identyfikacji) oraz analogiczne wyniki badań, w których użyto sztucznych sieci neuronowych. Sieci neuronowe wykonywały zadanie klasyfikacyjne na tym samym materiale badawczym, który został użyty w badaniu percepcyjnym (na sylabach wyciętych z mowy ciągłej). W kolumnie trzeciej zamieszczono odsetek odpowiedzi bezbłędnych dla poszczególnych

sylab (w teście percepcyjnym). Kolumna czwarta i piąta dotyczy odsetku poprawnie wykrytych realizacji sylab przez sieci probabilistyczne (w wersji bez macierzy strat oraz z macierzą strat). Widać, że niektóre wyniki są zbliżone (na przykład dla sylab: /na/, /vy/, /ka/). Zdarzało się również, że wyniki testu percepcyjnego były lepsze (sylaby /la/, /ma/), jednak zdecydowanie częściej wyniki uzyskane przez sieci neuronowe były lepsze. Fakt ten został potwierdzony przez wartości średnie – w teście percepcyjnym odsetek bezbłędnych identyfikacji wyniósł 57,5%, natomiast dla badań przeprowadzonych przy użyciu sieci probabilistycznych – 75%.

Tabela nr 4.: wyniki testu percepcyjnego oraz wyniki uzyskane przez sieci probabilistyczne.

LP.	SYLABA	POPRAWNA IDENTYFIKACJA	SIECI PROBABILISTYCZNE	SIECI PROBABILISTYCZNE 2/12
1	/la/	61,5%	32,1%	64,4%
2	/Sa/	44,0%	67,3%	80,0%
3	/na/	68,0%	71,4%	94,0%
4	/by/	73,5%	89,7%	90,9%
5	/ma/	65,5%	45,1%	71,4%
6	/to/	71,0%	87,7%	96,9%
7	/va/	67,6%	82,9%	93,1%
8	/vy/	82,0%	85,6%	96,4%
9	/ka/	74,5%	80,6%	97,8%
10	/wa/	40,5%	84,0%	92,5%
11	/po/	52,0%	86,4%	98,7%
12	/go/	52,0%	70,7%	92,0%
13	/ra/	49,0%	75,6%	88,0%
14	/ko/	58,0%	70,7%	87,3%
15	/n`e/	45,5%	72,3%	95,3%
16	/je/	28,5%	77,6%	87,5%
17	/mo/	57,0%	91,8%	91,9%
18	/no/	65,5%	56,7%	74,3%
19	/sa/	40,5%	88,2%	96,6%
20	/ry/	70,0%	69,4%	75,4%
21	/ta/	42,5%	80,0%	97,8%
22	/c`e/	55,5%	85,1%	96,4%
	Średnia	57,5%	75,0%	89,0%

4.4. Wyniki badań dotyczących właściwości dystrybucyjne sylab

Omówione w poprzednich podrozdziałach wyniki badań dotyczą właściwości akustycznych oraz percepcji realizacji sylab wyciętych z kontekstu (z akustycznego sygnału mowy). Badania te mają związek z próbą określenia roli sylaby w przetwarzaniu wstępującym w percepcji mowy oraz z potencjalną możliwością wykorzystania sylab w systemach automatycznego rozpoznawania mowy. Przeprowadzono również badania, których wyniki umożliwiają podjęcie dyskusji dotyczącej właściwości sylaby istotnych dla przetwarzania zstępującego. Te właściwości związane są z łączliwością wewnątrzwyrazową sylab oraz wynikającymi z tej łączliwości możliwościami określania wyrazów. Trzeba zaznaczyć może być to istotny, ale z pewnością nie jedyny istotny czynnik w przetwarzaniu zstępującym w percepcji mowy.

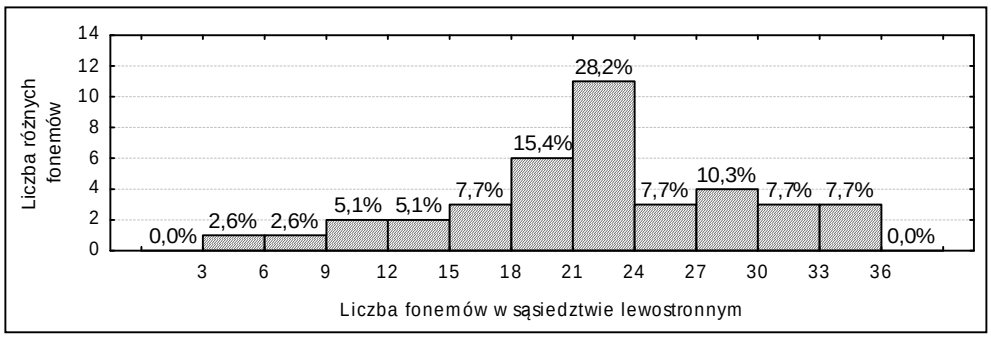
Badania przeprowadzono na podstawie korpusu tekstowego liczącego dwa miliony wyrazów. Przyjęta definicja wyrazu jest równoważna z wyrazem ortograficznym (wyrazem wyznaczonym przez spacje), dlatego w tych badaniach różne formy fleksyjne wyrazów były traktowane jako różne wyrazy [Jurkowski, 1999]. Przed przystąpieniem do badań wszystkie wyrazy zostały przetranskrybowane fonematycznie.

W pierwszej kolejności przedstawione zostały wyniki badań dotyczące łączliwości wewnątrzwyrazowej fonemów oraz sylab. W badaniach uwzględniono fonemy w celu porównania ich właściwości dystrybucyjnych z właściwościami dystrybucyjnymi sylab. Oddzielnie badano łączliwość lewostronną oraz łączliwość prawostronną jednostek.

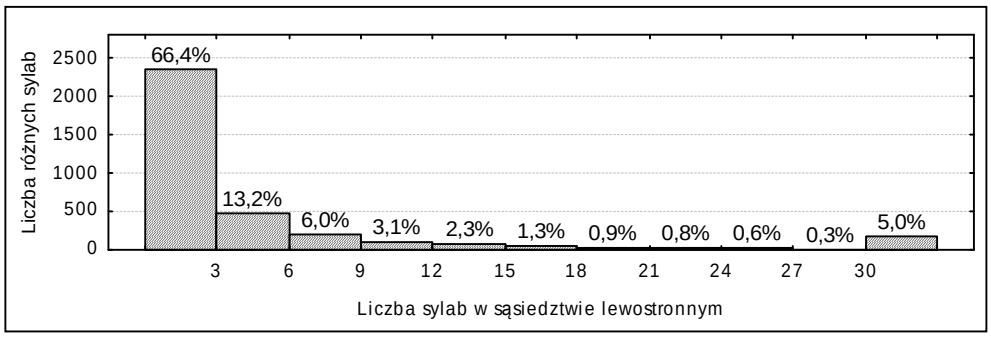
Z informacji zamieszczonych na rysunku 3 wynika, że dla większości fonemów liczba dopuszczalnych fonemów w sąsiedztwie lewostronnym mieści się w przedziale 18-30. Zaobserwowano 11 fonemów (28,2%), dla których liczba różnych fonemów w sąsiedztwie lewostronnym należy do przedziału 21-24. Na rysunku 4 zaprezentowano wyniki dotyczące sąsiedztwa lewostronnego sylab. Około 66% sylab miało w swoim lewostronnym sąsiedztwie nie więcej niż 3 różne sylaby.

Wyniki dotyczące sąsiedztwa prawostronnego fonemów są bardziej rozproszone niż wyniki uzyskane dla sąsiedztwa lewostronnego tych jednostek. Z rysunku 5 wynika, że prawie 40% fonemów graniczy prawostronnie z liczbą fonemów należącą do przedziału 21-28. Prawie 16% fonemów graniczy prawostronnie z więcej niż 32 różnymi fonemami. Ponad 23% fonemów może graniczyć prawostronnie z najwyżej dwunastoma różnymi fonemami. Natomiast z badań dotyczących sylab wynika, że ponad 55% sylab może mieć w swoim sąsiedztwie prawostronnym najwyżej 4 różne sylaby.

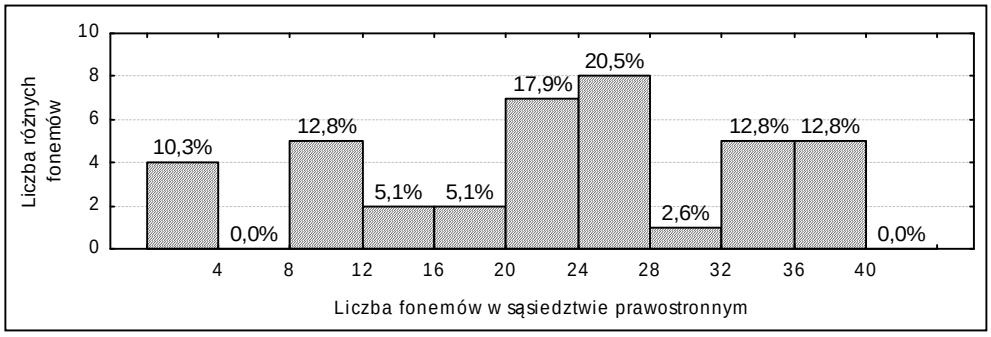
Rysunek nr 3.: Sąsiedztwo lewostronne fonemów.



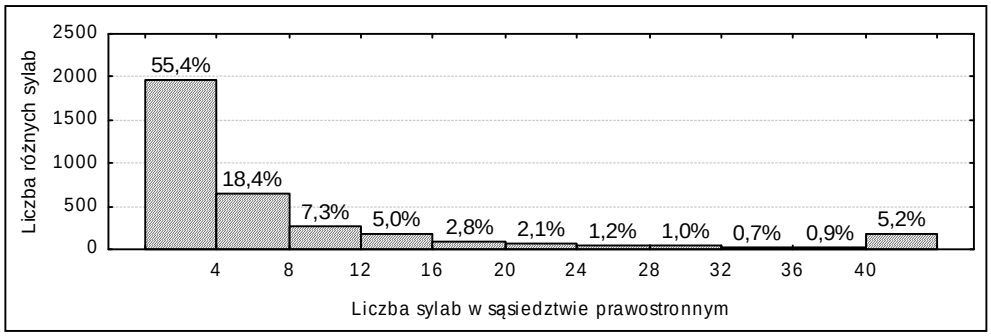
Rysunek nr 4.: Sąsiedztwo lewostronne sylab.



Rysunek nr 5.: Sąsiedztwo prawostronne fonemów.



Rysunek nr 6.: Sąsiedztwo prawostronne sylab.

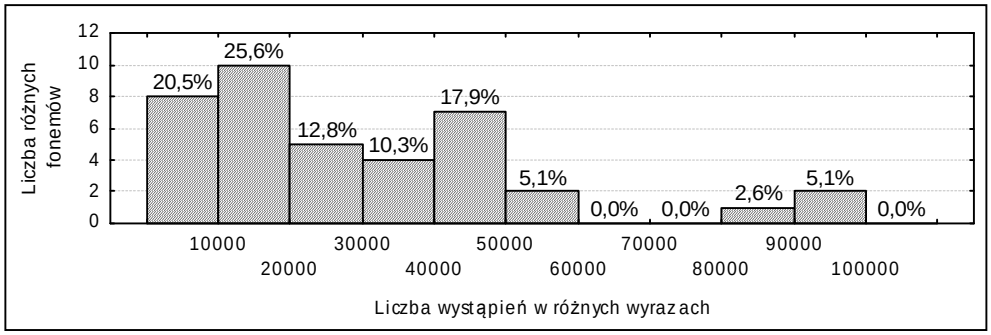


Przedstawione wyniki pokazują, że dla poszczególnych sylab istnieje niewielka liczba różnych sylab mogących znaleźć się w ich sąsiedztwie lewostronnym oraz prawostronnym (w obrębie wyrazu). Dzięki temu istnieje możliwość identyfikacji wyrazów na podstawie informacji o występowaniu poszczególnych sylab – obecność określonej sylaby wyznacza pewien skończony zbiór wyrazów, które mogą zawierać tę sylabę.

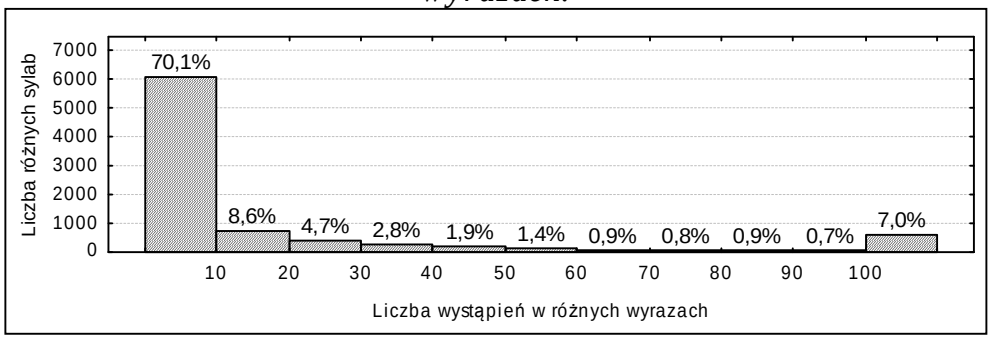
Z danych przedstawionych na kolejnym histogramie (rysunek 7) wynika, że obecność poszczególnych fonemów wyznacza bardzo dużo potencjalnych wyrazów, które mogą zawierać te fonemy. Na przykład 17,9% fonemów należy do liczby różnych wyrazów określonej przedziałem 40001-50000. Zatem możliwość identyfikacji wyrazu na podstawie pojedynczego fonemu jest bardzo niewielka – wynika to z oczywistego faktu istnienia niewielkiej liczby różnych fonemów.

Znacznie korzystniejsze wyniki uzyskano dla sylab (rysunek 8) – około 70% sylab jest częścią mniej niż 10 różnych wyrazów w korpusie złożonym z dwóch milionów wyrazów. Zatem sylaby dają duże możliwości określania wyrazów.

Rysunek nr 7.: Histogram obrazujący liczbę wystąpień fonemów w różnych wyrazach z korpusu.



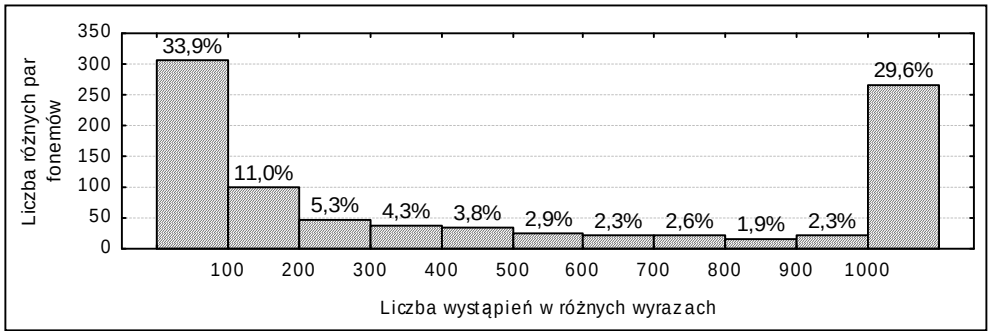
Rysunek nr 8.: Histogram obrazujący liczbę wystąpień sylab w różnych wyrazach.



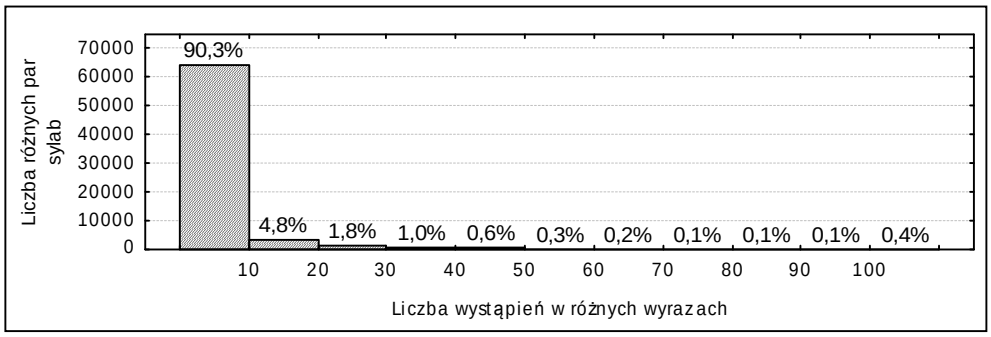
Badano również możliwości określania wyrazów na podstawie par fonemów oraz par sylab (dwóch fonemów lub dwóch sylab znajdujących się w bezpośrednim sąsiedztwie). Na rysunku 9 przedstawiono histogram z rezultatami dotyczącymi par fonemów. Wynika z nich, że większość par fonemów (ponad 65%) wystąpiło w więcej niż stu różnych wyrazach należących do korpusu tekstowego, natomiast prawie 30% par fonemów wystąpiło w więcej niż tysiącu różnych wyrazach. Zatem możliwość określania wyrazów na podstawie par jednostek nie jest duża.

Wyniki uzyskane dla sylab (dla par sylab) są znacznie bardziej korzystne. Z rysunku 10 wynika, że aż 90% par sylab wystąpiło w najwyżej dziesięciu różnych wyrazach w korpusie. Kolejny histogram (rysunek 11) również dotyczy sylab, jednak wyszczególniono na nim zakres 1-10 (liczby wystąpień w różnych wyrazach). Wynika z niego, że aż 53,4% par sylab wystąpiło tylko w jednym wyrazie w korpusie, 14,6% par sylab wystąpiło w dwóch różnych wyrazach a ponad 7% par sylab wystąpiło w trzech różnych wyrazach.

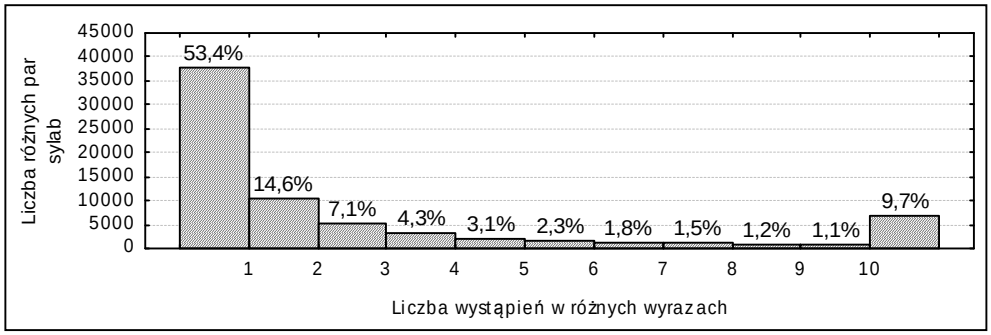
Rysunek nr 9.: Histogram obrazujący liczbę wystąpień par fonemów w różnych wyrazach.



Rysunek nr 10.: Histogram obrazujący liczbę wystąpień par sylab w różnych wyrazach (zakres 0–100).



Rysunek nr 11.: Histogram obrazujący liczbę wystąpień par sylab w różnych wyrazach (zakres 0–10).



Reasumując, sylaby są jednostkami znacznie bardziej wyróżniającymi niż fonemy i zaprezentowane wyniki ukazują skalę tego zjawiska – w badanym korpusie 70% sylab wystąpiło najwyżej w dziesięciu różnych wyrazach, natomiast około 75% par sylab wystąpiło w najwyżej trzech różnych wyrazach (w korpusie złożonym z dwóch milionów wyrazów). W tabeli 5 przedstawiono statystki opisowe dotyczące możliwości określania wyrazów na podstawie fonemów i sylab oraz par fonemów i par sylab. Uzyskane wartości średnie potwierdzają, że na podstawie sylab można wyznaczać wyrazy z największym prawdopodobieństwem. Statystyczna sylaba znalazła się średnio w prawie 60 różnych wyrazach w korpusie, natomiast statystyczna para sylab była częścią średnio 5,29 różnych wyrazów w korpusie. Biorąc pod uwagę fakt, że w tym badaniu różne formy fleksyjne wyrazu były uznawane za różne wyrazy (różne wyrazy ortograficzne), rzeczywista możliwość wyznaczania wyrazów na podstawie sylab jest jeszcze większa.

Uzyskane wyniki sugerują bliższe przyjrzenie się tym zależnościom w badaniach przyszłych – w przypadku par fonemów można wziąć pod uwagę informację o tym, czy pary te są częścią sylab typu CV lub innych bardziej złożonych struktur sylab.

Tabela nr 5.: Statystyki opisowe dot. występowania jednostek i par jednostek w wyrazach

SEKWENCJA	N WAŻNYCH	ŚREDNIA	MINIMUM	MAKSIMUM	ODCH. STD.
Pojedynczy fonem	39	28513,85	370	91686	23409,94
Pojedyncza sylaba	8669	59,77	1	13073	395,80
Para fonemów	899	1264,56	1	22767	2402,12
Para sylab	70887	5,29	1	1520	20,74

5. Wnioski końcowe

Badania wykazały, że poszczególne sylaby były wykrywane przez sieci neuronowe nawet wtedy, gdy stosunkowo niewielka liczba ich realizacji była wymieszana ze znacznie większym zbiorem sylab losowych. Średnie wartość wskaźnika przedstawiającego odsetek wykrytych sylab w przypadku sylab pobranych z wyrazów izolowanych wyniosła około 95%, natomiast dla sylab wycinanych z mowy ciągłej było to około 77%. Średni odsetek bezbłędnie zidentyfikowanych sylab na drodze testu percepcyjnego wyniósł 57,5%.Wyniki badań sugerują, że poszczególne osoby biorące udział w teście percepcyjnym nie wykorzystywały w pełni cech wyróżniających związanych z poszczególnymi realizacjami sylab. Można dyskutować o przyczynach takiego stanu rzeczy – być może cechy te nie są dostępne percepcyjnie, a może przegrały z innymi, silniejszymi czynnikami przetwarzania zstępującego.

Uzyskane wyniki sugerują, że na obecnym etapie rozwoju systemy rozpoznawania mowy nie powinny zbyt dalece wzorować się na człowieku, bo działają wydajniej idąc własnymi ścieżkami. W przypadku percepcji mowy funkcjonuje złożony mechanizm obejmujący przetwarzanie wstępujące oraz przetwarzanie zstępujące. Percepcja bodźców w postaci wyciętych z kontekstu sylab nie jest zatem zgodna z istotą tego mechanizmu. Trzeba jednak zauważyć, że bardzo dużo odpowiedzi uzyskanych z testu percepcyjnego, pomimo tego, że były one błędne, było odpowiedziami podobnymi do oczekiwanych (na przykład sylaby ze zmienionym nagłosem lub ośrodkiem). Z tych rozważań wyłania się ostateczny wniosek dotyczący wyników badań: w przetwarzaniu wstępującym mogą być identyfikowane pewne wstępne (zbliżone) kształty segmentów (sylab), przy czym w tej identyfikacji nie jest wykorzystywana cała informacja zawarta w akustycznym sygnale mowy związanym z tymi segmentami. Te kształty zostają uściślone dopiero na drodze przetwarzania zstępującego.

Wyniki badania dotyczącego właściwości dystrybucyjnych sylab oraz możliwości określania wyrazów na podstawie informacji o obecności konkretnej sylaby lub pary sylab sugerują, że sylaba może odgrywać znaczącą rolę również w przetwarzaniu zstępującym.

Uzyskane wyniki w pełni uzasadniają kontynuowanie badań przy uwzględnieniu nowych szczegółów i aspektów. Aktualne wyniki stanowią dla

autora podstawę i punkt wyjścia dla badań przyszłych. Na podstawie uzyskanych wyników można postawić tezę dla tych badań: sylaba jest najbardziej wydatną formą wiązania przetwarzania wstępującego oraz przetwarzania zstępującego, zatem jest to podstawowa jednostka percepcji mowy.

Bibliografia

Cooper R.S., Delattre P.C., Liberman A.M., Borst J.M., Gerstman L.J. 1952. Some experiments on the perception of synthetic speech sounds. Journal of the Acoustical Society of America, 24(6). s. 597-606.

Jurkowski M., Karolak S., Laskowski R., Lewicki A.M., Polański K., Saloni Z. 1999. Encyklopedia językoznawstwa ogólnego. (wyd. II). Zakład Narodowy im. Ossolińskich. Wrocław. s. 644.

Kopecek I. 1999. Speech Recognition and Syllable Segments. Publisher Springer. Berlin /Heidelberg. s. 1-203.

Massaro D. W. 1974. Perceptual units in Speech Recognition. JEP, vol. 102. s. 199-208.

Demenko G., Wypych M., Baranowska E. 2003. 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. s.79-95.

Dupoux E. 1993. Prelexical processing: the syllabic hypothesis revisited. W: Cognitive models of speech processing: The second sperlonga meeting, G. T. M. Altmann and R. Shillcock, Eds. Hove East Sussex UK: LEA, s. 81-114.

Hirihara T., Kato H. 1992. The efect of F0 on vowel identification. 1992. W: Speech Perception, Production and Linguistic Structure. Burke. IOS Press.

Ladefoged P. 1975. A course in phonetics. Harcourt Brace Jovanovich. London. s. 217.

Liberman A.M., Delattre P.C., Cooper F.S. 1952. The role of selected stimulus variables in the perception of the unvoiced stop consonants. American Journal of Psychology, 65, s. 497-516.

Lisker L., Abramson A.S. 1967. Some effects of context on voice onset time in English stops. Language and Speech, 10. s 1-28.

Mehler. J. 1981. The role of syllables in speech processing: Infant and adult data. W: Philosophical Transactions of the Royal Society, vol. 295. s. 333–352.

Segui J. 1984. The syllable: A basic perceptual unit in speech perception? W: Attention and Performance X, H. Bouma and D. G. Bouwhuis, Eds. Hillsdale NJ: Erlbaum. s. 165-181.

Ohde R.N. 1984. Fundamental frequency as an acoustic correlate of stop consonant voicing. Journal of the Acoustical Society o America, 75. s. 224–230.

Raphael L.J., Dorman M.F. 1980. Silence as a cue to the perception of syllable-initial and syllable-final stop consonants. Journal of Phonetics, 8. s. 269-275.

Raphael L.J., Dorman M.F., Freeman F., Tobin C. 1975. Vowel and nasal duration as cues to voicing in word-final stop consonants: Spectrographic and perceptual studies. Journal of Speech & Hearing Research, 18. s. 389-400.

Tadeusiewicz R., Lula P. 2001. Statistica Neural Networks PL; Kurs użytkowania programu. StatSoft. Kraków. s. 5-86.

Tadeusiewicz R. 1998. Elementarne wprowadzenie do techniki sieci neuronowych z przykładowymi programami. Akademicka Oficyna Wydawnicza. Warszawa. s. 1-164.

Trask R.L. 1996. A Dictionary of Phonetics and Phonology. Routlege. New York (wyd.II). s. 327, 345.