2.23 — akustyka mowy

P. Łobacz, N. Mikołajczak, J. Wysocka

PSYCHOFONETYCZNE PODSTAWY SEGMENTACJI SYGNAŁU MOWY

48/1990

WARSZAWA 1990

ISSN 0208-5658

Praca wpłynęła do Redakcji dnia 6 listopada 1989 r.

Na prawach rękopisu

Instytut Podstawowych Problemów Techniki PAN
Nakład 120 egz. Ark.wyd.1.0 Ark.druk.1,25
Oddano do drukarni w grudniu 1990 r.
Nr zamówienia 16/91

Warszawska Drukarnia Naukowa, Warszawa,
ul.Śniadeckich 8

Piotra Łobacz
Nawoja Mikołajczak
Jolanta Wysocka
Instytut Językoznawstwa UAM

PSYCHOFONETYCZNE PODSTAWY SEGMENTACJI SYGNAŁU MOWY.[1]

Streszczenie.

W pracy porównano wyniki automatycznej segmentacji fonetycznej tekstu uzyskane w Zakładzie Fonetyki Akustycznej IPPT ze sposobami analizy fonetycznej w procesie przetwarzania sygnału mowy przez człowieka. Przeprowadzone testy psychofonetyczne z 25 osobową grupą słuchaczy pozwoliły na podtrzymanie hipotezy, że w procesie percepcji nadchodząca informacja akustyczna jest segmentowana dopiero na poziomie analizy fonologicznej. Analiza automatyczna operująca jednostkami o podobnej funkcji językowej zakłada wielosegmentalność niektórych fonemów - operuje jednostkami niższego rzędu. Ustalenie liczby segmentów danej wypowiedzi w przetwarzaniu naturalnym zależy w znacznym stopniu od świadomości bazy fonologicznej użytkownika języka.

I. Założenia ogólne doświadczenia percepcyjnego.

Osiągnięcie wyraźnych sukcesów w prowadzonych, na przestrzeni ostatnich trzydziestu lat, szeroko zakrojonych pracach nad automatycznym rozumieniem mowy ciągłej w dowolnym języku w znacznej mierze zależy od pokonania problemów związanych z segmentacją bieżącego tekstu. W automatycznej (komputerowej) analizie składniowej i semantycznej zapisu ortograficznego segmentacja tekstu nie stanowi praktycznie żadnego problemu. W wypadku mowy segmentacja jest kwestią zasadniczą, związane z nią problemy nie zostały rozwiązane do końca mimo bardzo wielu, mniej lub bardziej zadowalających prób. Istnieją natomiast udane rozwiązania automatyzacji przetworzenia zapisu ortograficznego na fonetyczny, w tym także dla tzw. normatywnej odmiany wymowy współczesnej polszczyzny [1]. Nie opracowano dotychczas metod transformacji mowy naturalnej na automatyczną reprezentację fonetyczną.

W funkcjonujących na świecie urządzeniach typu ARM o charakterze wdrożeniowym rzadko stosuje się bardziej drobnoziarnistą segmentację niż wyraz, czy prosta fraza. W automa-

tach funkcjonujących jako narzędzia badawcze wykorzystuje się różnorodne kryteria segmentacyjne bieżącego tekstu: analizę mikrosegmentalną co 10 ms [12], segmentację akustyczno-fonetyczną [17], podział odpowiadający dwóm kolejnym fonom [10], segmentację sylabiczną [18] oraz globalną - wyrazową [11]. Obok tych podstawowych segmentów w procesie rozpoznawania wyróżnia się także inne większe jednostki. Konstruowane są systemy interakcyjne, które na kilku etapach przetwarzania zakładają wieloznaczność granic międzysegmentalnych ustalonych dla określonego poziomu [19].

W latach osiemdziesiątych w Zakładzie Fonetyki Akustycznej prowadzone były systematyczne prace nad segmentacją sygnału mowy na elementy o rozciągłości głoski, reprezentującej fonem jako lingwistyczną jednostkę abstrakcyjną (por. np. [3, 2, 9]). Głównym celem tych prac, prowadzonych także obecnie, jest podzielenie kontinuum mowy na takie elementy, których dystynktywny zbiór byłby minimalny (rzędu kilkudziesięciu jednostek). Segmentacja tego rodzaju stanowi wstępny, ale praktycznie niezbędny etap automatycznej interpretacji dowolnego tekstu mówionego w języku polskim.

W pracy [4] autor stwierdza, że brak do dzisiaj adekwatnej akustycznej definicji fonemu stanowi prawdopodobną przeszkodę uniemożliwiającą rozwiązanie problemu automatycznej segmentacji. Interdyscyplinarnemu (w tym także akustycznemu) opracowaniu definicji podstawowej jednostki fonologicznej poświęcono bardzo wiele prac na świecie i w Polsce, ostatnio np. [8]. Nie rozwiązana do końca została interpretacja skutków zmienności wynikających z różnic międzyosobniczych, lingwistycznych i koartykulacyjnych, zachodzących równocześnie we wszystkich parametrach akustycznych. Interakcjom osobniczych i lingwistycznych źródeł zmienności poświęcono pracę [7]. Brak jest dotychczas ontologiczno-logicznego opracowania klasyfikującego konkretne akustyczne segmenty w sygnale mowy jako abstrakcyjne jednostki lingwistyczne. Znaczna większość prac formalizujących pojęcie fonemu przyjmuje za punkt wyjścia cechy artykulacyjne, jako bardziej jednoznacznie interpretowalne.

Dotychczas opublikowane wyniki automatycznej segmentacji

fonetycznej wybranych 6-11 fonemowych 18 wyrazów znaczących, wymówionych przez kilkanaście osób, zbioru 34 pięciofonemowych logatomów, czy 444 najczęściej występujących połączeń dwufonemowych w języku polskim, wypowiedzianych kilkakrotnie przez 8 osób (por.[2, 3, 4]), opierają się na generalnym założeniu, że istnieje możliwość jednoznacznej interpretacji fonetyczno-fonologicznej, występującej w cyfrowym obrazie sygnału mowy typu "visible speech". Jako granicę międzysegmentalną przyjmuje się lokalną zgodność kierunków zmian w obrazie widma dynamicznego. Źródła tych metod nawiązują bezpośrednio do możliwości wizualnej segmentacji zapisu spektrograficznego, prowadzącej do odczytania zawartego w nim tekstu. Wyniki tych prac, jakkolwiek zadowalające, nie doprowadziły do całkowitej jednoznacznej segmentacji według przyjętych zwyczajowo kryteriów fonetycznych. Nasunęły natomiast autorowi pomysł opracowania dla języka polskiego fonokodu - czyli systemu umożliwiającego utworzenie zbioru komend jednoznacznie interpretowanych przez układ ARM, opartego na występowaniu najbardziej dystynktywnych segmentalnie elementów [5]. Wyrazy poddane rozpoznaniu nie stanowiły jednostek leksykalnych żadnego naturalnego języka, wobec czego metoda ta nie może mieć praktycznego zastosowania w Speech Understanding Systems, nawiązujących w swych rozwiązaniach technicznych do sposobu odbioru informacji językowej przez człowieka.

Do chwili obecnej nie zostało wyczerpująco opracowane fundamentalne, dla omawianej problematyki, zagadnienie tzw. inwariantów w sygnale mowy. W latach osiemdziesiątych, w kilku światowych laboratoriach, nastąpił kolejny renesans badań nad możliwościami odczytywania wypowiedzi określonego języka ze spektrogramów dynamicznych (np.[20]). W Zakładzie Fonetyki Akustycznej IPPT prace dotyczące czytania spektrogramów nastawione były głównie na rewalidację osób niesłyszących (por.np. [14,15,16]). Doświadczenia z nauką czytania obrazów typu "visible speech" wykazały istotną różnicę pomiędzy dekodowaniem spektrogramów dynamicznych a odczytaniem zapisów ortograficznych w dowolnym piśmie alfabetycznym. W nauce czytania tradycyjnego zapisu ortograficznego przechodzi się od etapu rozpoznawania segmentalnego do rozpoznawania globalnego całych wyra-

zów, a potem dłuższych fragmentów tekstu. Z nabraniem wprawy częste wyrazy kilkuliterowe są wzrokowo odbierane całościowo, wyrazy długie (kilkunastoliterowe) są w procesie dekodowania składane z dwóch, trzech bloków literowych. Odczytywanie spektrogramów izolowanych wyrazów, w początkowych fazach treningu, przebiega w sposób częściowo segmentalny, polegający na jednoznacznym rozpoznaniu kilku najbardziej charakterystycznych segmentów, bądź zaklasyfikowaniu ich do jakiejś kilkuelementowej grupy głosek, a częściowo globalny, ponieważ segmenty nierozpoznane zostają odgadnięte dzięki językowej wiedzy użytkownika języka lub uprzednio zapamiętanego charakterystycznego wzoru w jaki ułożyły się wizualne cechy spektrogramu. Znaczna wprawa i większe przygotowanie fonetyczne pozwalają na dekodowanie spektrogramów segment po segmencie, przy czym uzyskane rezultaty są dalekie od jednoznacznej interpretacji. Dla wyłącznie segmentalnego czytania spektrogramów osiągnięto poprawność w granicach 50-64% (por.[20]). Natomiast w automatycznej segmentacji ograniczonego zbioru wypowiedzi w języku polskim osiągnięto wysoką poprawność interpretacji - w granicach 80-94% (por.np.[2,9]). Różnice te wyraźnie wskazują, że w wizualnej postaci sygnału mowy znajduje się znacznie więcej informacji fonetycznej dającej się ująć w formalne reguły segmentacyjne niż to ma miejsce w interpretacji wzrokowej.

Zadanie niniejszej pracy polegało na przeprowadzeniu wstępnych eksperymentów psychofonetycznych dotyczących segmentacji fonetycznej jednostek leksykalnych w naturalnym procesie przetwarzania informacji językowej. Zasadniczą kwestię stanowiło sprawdzenie, czy na podstawie testów odsłuchowych można ustalić w jaki sposób działa mechanizm segmentacyjny na poziomie niższym niż wyraz czy sylaba, jaką ten mechanizm generuje jednostkę i czy jednostka taka jest w przybliżeniu zgodna definicyjnie z segmentem fonetyczno-akustycznym typu plozji, afrykacji, raptownego ugięcia formantowego itp. czy też odpowiada bardziej lingwistycznemu pojęciu głoski czy fonemu, bądź wreszcie ma charakter mieszany, podobnie jak w automatycznej segmentacji obrazów spektrograficznych (por. np.[2]). Zakłada się, że każdy użytkownik języka może dokonywać klasyfikacji segmentów traktowanych jako zjawiska fizyczne

bądź klasy abstrakcji na głoski lub fonemy, choć nie potrafi zapewne ustalić kryteriów takiej klasyfikacji. W poprawnych opisach lingwistycznych klasyfikacja fonematyczna uwzględnia kryteria fonetycznej dystynktywności i podobieństwa i w tym zakresie istnieją analogie z wizualnym rozpoznawaniem spektrogramów komputerowych.

II. Konstrukcja i przebieg testu

W naturalnej sytuacji komunikacyjnej odbiorca nie dokonuje dekodowania kolejno wszystkich jednostek fonetycznych w danej wypowiedzi, a tylko te z nich, które występują w określonych fragmentach tekstu, szczególnie przy granicach wyrazowych i morfologicznych. Potwierdziły to różnorodne eksperymenty psycholingwistyczne wywodzące się z teorii interakcyjnych sposobów przetwarzania informacji językowej (np. [6,13]). W celu zbadania kryteriów i potrzeb segmentacyjnych użytkownika języka na poziomie fonetyczno-fonologicznym, materiał językowy tak dobrano, by maksymalnie zredukować przetwarzanie na poziomie analizy leksykalnej i składniowej przy zachowaniu naturalnej substancji językowej. Uznano, że warunki takie zostaną dobrze spełnione gdy materiał językowy zawierał będzie krótkie, najczęściej jednowyrazowe wypowiedzi konkretnego języka, ale takiego, który jest całkowicie nieznany słuchaczom, wymuszając na nich wyłącznie subgramatyczną interpretację.

Każdy test odsłuchowy odwołuje się do pamięci użytkownika języka. Poszczególne procedury testowe wymagają wykorzystania różnych jej typów: pamięci sensorycznej, magazynującej napływające informacje na przeciąg około 250 ms, operacyjnej pamięci krótkotrwałej, której górna granica czasowa nie została w psychologii jednoznacznie określona, ale najczęściej nie przekracza kilkudziesięciu sekund oraz pamięci długotrwałej, interpretowanej jako magazyn wiedzy poszczególnych osobników. Zaproponowana procedura badawcza odpowiadała powszechnie stosowanej technice rozpoznawania, ale tylko w ogólnych zarysach. W klasycznym ujęciu rozpoznawanie polega na porównaniu nadchodzącej informacji z wzorcami zmagazynowanymi w pamięci długotrwałej. Obecny eksperyment wymagał od uczestników nie tylko porównania bodźców z istniejącymi w pamięci informacjami, ale również jednoczesnej adaptacji ich wiedzy językowej do usły-

szanych sygnałów, których struktura fonetyczna była im w znacznej mierze nieznana.

Założono, że funkcjonowanie pamięci krótkotrwałej będzie można oszacować ustalając liczbę kolejnych poprawnie zapamiętanych jednostek segmentalnych, po trzykrotnym zaprezentowaniu poszczególnych bodźców do momentu pojawienia się pierwszego błędu. Odzwierciedleniem procesów zachodzących w pamięci długotrwałej miała być interpretacja sygnałów testowych, polegająca na określeniu wszystkich, dających się wyodrębnić w danym wyrazie, elementów fonetycznych.

II.1. Materiał językowy.

Sporządzono dwie listy wyrazowe. Jedna zawierała 38 wyrazów krymsko-tatarskich, druga 36 wyrazów niderlandzkich o wyraźnie zróżnicowanych długościach. Według lingwistycznej interpretacji fonologicznej wyrazy na obu listach zawierały od dwóch do dziesięciu fonemów. Jednostki leksykalne określonej długości występowały w materiale kilkakrotnie (najczęściej czterokrotnie, sporadycznie dwu lub sześciokrotnie). Wyrazy nagrane zostały na taśmę magnetofonową w warunkach studyjnych przez rodowitych mówców. Przygotowanie taśm testowych polegało na: (1) przypadkowym uporządkowaniu wyrazów ze względu na ich długość, (2) trzykrotnym następującym po sobie, powtórzeniu każdego z nich w równych odstępach czasu, wynoszących około 20s. Materiał doświadczalny dla każdego z języków przygotowany został osobno.

II.2. Słuchacze.

Osobami biorącymi udział w odsłuchach byli studenci z różnych kierunków studiów oraz młodzi asystenci bez przygotowania fonetycznego. Materiał obu języków zaprezentowano tej samej grupie słuchaczy na oddzielnych sesjach. 28 osób wzięło udział w sesji poświęconej językowi niderlandzkiemu. W sesji odsłuchowej języka krymsko-tatarskiego uczestniczyło ogółem 25 osób.

II.3. Przebieg doświadczenia.

Testy percepcyjne przeprowadzono odrębnie dla dwóch grup słuchaczy. Dla pierwszej, 20 osobowej grupy słuchaczy przed pierwszą sesją testową zorganizowano około 40 minutowy wykład dotyczący sposobu zapisu różnych dźwięków mowy w tradycyjnej

ortografii. Odwoływano się głównie do szkolnej wiedzy na temat różnic występujących pomiędzy wymową a pisownią w języku polskim. Przygotowanie drugiej - 8 osobowej grupy - odbywało się w trzech etapach. W czasie pierwszego z nich, trwającego około godziny, zapoznano uczestników testu z zasadami międznarodowej transkrypcji fonetycznej (IPA). Dwa półgodzinne etapy następne polegały na zapisie w transkrypcji różnych połączeń dźwiękowych prezentowanych w studio. Oba szkolenia wstępne miały na celu maksymalne wyeliminowanie zapisów materiału testowego, odzwierciedlających specyficzne, polskie zasady ortografii, niekoniecznie zgodne z dokonywaną przez słuchaczy segmentacją. Również ankiety odsłuchowe przygotowano tak, by dodatkowo usunąć istniejące nawyki ortograficzne. Słuchacze, każdy wyodrębniony przez siebie segment fonetyczny, zapisywali w oddzielnej kratce, bez względu na to, ile liter wykorzystali na określenie danego segmentu. Liczba kolejnych zapisanych kratek była odbiciem subiektywnej długości każdego wyrazu.

Wszyscy słuchacze przed pierwszym testem otrzymali następującą instrukcję: "Usłyszycie wyrazy nieznanego sobie języka. Każdy wyraz zostanie powtórzony trzykrotnie. Pierwsza prezentacja ma jedynie na celu osłuchanie się, po drugiej prezentacji ma nastąpić zapis w kwestionariuszu, trzecia replikacja służy celom kontrolnym."

III. Wyniki doświadczenia psychofonetycznego.

III.1. Porównanie grup słuchaczy.

Porównanie to miało na celu określenie wpływu stopnia przygotowania fonetycznego na sposób dokonywania segmentacji poszczególnych wyrazów. Dla języka krymsko-tatarskiego, w grupie (1) - mniej przygotowanej pod względem fonetycznym - przypadało przeciętnie 15 błędów na osobę, w grupie (2) - osób po dodatkowym treningu - liczba błędów wynosiła 16. Analogiczne dane dla języka niderlandzkiego wynoszą: w grupie (1) 17 błędów na osobę, w grupie (2) 20 błędów. Chcąc ustalić ewentualną istotność różnicy między grupowej w teście języka niderlandzkiego obliczono kryterium χ^2. Wartość χ^2 okazała się nieistotna na poziomie $\alpha = 0.05$, wobec czego w dalszych procedurach obliczeniowych obie grupy traktowano łącznie.

Przeciętna liczba błędów na osobę wynosiła dla języka

rach obliczeniowych obie grupy traktowano łącznie.

Przeciętna liczba błędów na osobę wynosiła dla języka krymsko-tatarskiego 15.6, a dla języka niderlandzkiego 17.6.

III.2. Liczba błędów a długość wyrazu.

Poniżej przedstawiono zestawienie ilustrujące liczbę poprawnie zapisanych wyrazów dla każdej długośći wyrazu wyrażonej liczbą fonemów.

długość wyrazu w fonemach	liczba nadanych wyrazów		liczba poprawnie odebranych wyrazów	
	krymsko-tat.	niderl.	krymsko-tat.	niderl.
2	4	3	3	2
3	4	2	4	2
4	6	7	2	2
5	4	6	2	2
6	3	2	0	1
7	4	4	1	1
8	4	2	0	0
9	5	7	1	0
10	4	3	0	0

Z zestawienia wynika, że zachodzi tendencja zwiększania się liczby błędów wraz ze wzrostem długości wyrazów. W celu szczegółowego zbadania tej zależności wykonano analizę błędów przy uwzględnieniu następujących parametrów: długości wyrazów wyrażonej liczbą fonemów, liczby wyrazów dla określonej długości oraz liczby błędów potencjalnych, oznaczającej liczbę wszystkich fonemów dla danej kategorii wyrazowej pomnożonej przez liczbę słuchaczy. Każdą kategorię wyrazową scharakteryzowano za pomocą współczynnika określonego jako błąd względny (stosunek błędów realnych do potencjalnych w %). Wartości poszczególnych parametrów przedstawiono w Tablicy 1 dla języka krymsko-tatarskiego oraz w Tablicy 2 dla języka niderlandzkiego.

Z obliczeń przedstawionych w Tablicy 1 wynika, że zachodzi zależność pomiędzy wielkością błędu względnego a długością wyrazu dla języka krymsko-tatarskiego. Wyznaczony współczynnik korelacji r pomiędzy wartościami błędów względnych a liczbą segmentów w wyrazach wynosi r = +0.6, co oznacza słabą korelację dodatnią. Dla języka niderlandzkiego nie stwierdzono żadnej zależności.

długość wyrazu	liczba wyrazów danej długości	liczba błędów potencjalnych	liczba błędów realnych	wartość błędu względnego
2	4	200	13	6.5
3	4	300	2	0.7
4	6	600	18	3.0
5	4	500	17	3.4
6	3	450	37	8.0
7	4	700	63	9.0
8	4	800	72	9.0
9	5	1125	119	10.0
10	4	1000	48	5.0
Razem	38	5675	389	

Tablica 1. Liczebność błędów dla danych długości wyrazów. Język krymsko-tatarski.

długość wyrazu	liczba wyrazów danej długości	liczba błędów potencjalnych	liczba błędów realnych	wartość błędu względnego
2	3	168	2	1.1
3	2	168	2	1.1
4	7	784	76	9.7
5	6	840	37	4.4
6	2	336	27	8.0
7	4	784	66	8.4
8	2	448	27	6.0
9	7	1764	162	9.2
10	3	840	94	12.2
Razem	36	6132	493	

Tablica 2. Liczebność błędów dla danych długości wyrazów. Język niderlandzki.

W celu ustalenia zależności między językiem a poprawnością segmentacji wykonano test χ^2 z tablicą kontyngencji 2 x 2 :

	krymsko-tatarski	niderlandzki
poprawne	5286	5639
niepoprawne	389	493

Wartość statystyki χ^2 wynosi 5.99 i jest istotna na poziomie $\alpha = 0.025$, co oznacza, że rozkład zmiennej względem języków i poprawności segmenetacji jest niezrównoważony. Dla niderlandzkiego uzyskano mniejszą poprawność decyzji słuchaczy.

III.3. Klasyfikacja błędów segmentacyjnych.

W przypadku automatycznej segmentacji sygnału mowy zastosowano następujące relacje między uzyskanymi segmentami a lingwistycznymi jednostkami fonetycznymi tekstu poddanego analizie:

(1) segmentację właściwą, czyli zgodność jednostek fonetycznych z automatycznie uzyskanymi segmentami,

(2) segmentację dodatkową, oznaczającą większą liczbę segmentów niż jednostek fonetycznych oraz

(3) brak segmentacji, gdy jeden segment obejmuje więcej niż jedną jednostkę fonetyczną (por.[2]).

Z punktu widzenia apriorycznej wiedzy fonetycznej za błędne uznano w obecnym doświadczeniu tak wszelkie segmentacje dodatkowe jak i braki segmentacji. Dla obu języków braki segmentacyjne występowały częściej i wynosiły 61% wszystkich błędów dla języka krymsko-tatarskiego oraz odpowiednio 78% dla języka niderlandzkiego.

Ponieważ segmentacje słuchaczy wskazywały na występowanie błędów systematycznych i niesystematycznych, wprowadzono bardziej szczegółową klasyfikację. Brak segmentacji podzielono na dalsze kategorie: opuszczenia i ściągnięcia. Opuszczenie oznacza brak segmentu niezależnie od otaczającego go kontekstu fonetycznego, natomiast ściągnięcie polega na zastąpieniu, pod wpływem reguł fonotaktycznych rodzimego języka słuchaczy, dwóch kolejnych jednostek segmentem pojedynczym. Analogicznie podzielono na dwie kategorie: wstawienia i rozszczepienia wszystkie segmentacje dodatkowe.

Procentowy udział błędów poszczególnych kategorii dla obu

języków przedstawia się następująco:

	krymsko-tatarski	niderlandzki
opuszczenia	53.7	39.3
ściągnięcia	6.9	38.5
wstawienia	3.1	3.7
rozszczepienia	36.2	18.9

(1) Opuszczenia.

Dla obu języków stwierdzono najwięcej opuszczeń dotyczących krótkich, nieakcentowanych samogłosek. Tego typu błędy stanowią 71% wszystkich opuszczeń w języku niderlandzkim i 42% w języku krymsko-tatarskim. Najczęściej opuszczaną spółgłoską w języku krymsko-tatarskim był dźwięk /qʻ/, szczególnie w nagłosie i wygłosie wyrazów (21% opuszczeń), np. wyraz /qʻapayı/ zapisany został przez większość słuchaczy jako /apayı/. Należy wyjaśnić, że różnorodne postaci zapisu poszczególnych segmentów jakie wystąpiły na kwestionariuszach odsłuchowych dla przejrzystości ujednolicono i w całym tekście niniejszej pracy stosowano poprawną transkrypcję fonematyczną.

Żaden słuchacz nie dokonał podwójnej segmentacji w wypadku wystąpienia geminaty, np. wyraz krymsko-tatarski /gəŋılly/ rozsegmentowany został jako /gəŋıly/. Geminata ilustruje różnicę między analizą fonetyczną a fonologiczną. Na poziomie fonetycznym jest to jeden segment, najczęściej o wzdłużonym iloczasie. W języku niderlandzkim opuszczeniu ulegały wygłosowe spółgłoski (20% wszystkich opuszczeń). W wyrazie /ɔpzukən/ 25 osób nie usłyszało końcowego /n/.

(2) Ściągnięcia.

Błędy tego typu są charakterystyczne głównie dla języka niderlandzkiego. Dotyczą (a) uproszczenia grup spółgłoskowych /rh,yr/ (39% wszystkich ściągnięć), np. /hɛrha:l/ 23 osoby rozsegmentowały jako /hɛha:l/, (b) monoftongizacji dyftongów (32% wszystkich błędów tej kategorii), np. sygnał /bəlɑŋrɛ.ik/ odebrany został przez wszystkich jako /bəlɑŋrik/, (c) potraktowania dwóch kolejnych elementów /t + s/ jako afrykaty /ts/ (21% śiągnięć), oraz (d) łączeniu /ɔ + n/ jako /ɔ̃/ (11%).

Ta ostatnia tendencja występowała także w języku krymsko-tatarskim. Ponadto w tym języku uproszczeniu uległa nagłosowa grupa /ji/. W języku krymsko-tatarskim fonem /j/ ma spółgłos-

kowy charakter, jest artykułowany z wyraźną frykcją. Mimo to, żaden słuchacz go nie wyodrębnił w tym kontekscie. Połączenie /ji/ zapisywano jako /i/.

(3) Wstawienia.

Najrzadziej popełnianym dla obu języków błędem była segmentacja dodatkowa, polegająca na wstawieniu dodatkowego elementu. 5 osób w niderlandzkim wyrazie /tənmɪnstə/ zapisało na ostatniej pozycji dźwięk /r/. Segmentacja dodatkowa tego typu najczęściej występuje w wygłosie wyrazu (8 przypadków w krymsko-tatarskim i 9 w niderlandzkim).

(4) Rozszczepienia.

Błędy tej kategorii dwukrotnie częściej występują w języku krymsko-tatarskim niż w niderlandzkim. W pierwszym z nich dotyczą one najczęściej asynchronicznej miękkości spółgłosek i są związane z bardzo różnym pierwotnym miejscem artykulacji tych dźwięków. Zjawisko to występuje dla /b,m,n,t,ʤ,ʧ,l,ɕ,k/ i stanowi 67% wszystkich błędów tej kategorii. Asynchroniczną miękkość słuchacze oznaczali wstawiając samogłoskę /i/ lub /j/. Interpretacja wstawionego elementu /i/ jest niejasna. Nie wiadomo czy był to konsekwentnie wprowadzany znak na oznaczenie dodatkowego segmentu, czy silna zależność od polskiej ortografii.

W języku niderlandzkim najczęstszym błędem w tej klasie była dyftongizacja samogłosek długich (72% wszystkich rozszczepień), np. /ɣəstheːr/ 8 osób zapisało jako /ɣəsthjer/. Sporadycznie występowały również rozszczepienia /ɣ/ na dwa elementy /r + h/.

W obu językach nastąpiło rozszczepienie nosowej spółgłoski tylnojęzykowej /ŋ/ na połączenie /n + g/.

IV. Interpretacja wyników i wnioski końcowe

Przedstawiony powyżej eksperyment psychofonetyczny stanowił kolejną próbę wyjaśnienia sposobu przetwarzania sygnału mowy przez człowieka. Miał odpowiedzieć na pytanie, czy nadchodząca informacja językowa jest segmentowana percepcyjnie dopiero na poziomie analizy fonologicznej i czy w związku z tym przetwarzanie fonologiczne jest pierwszym etapem percepcji wyspecjalizowanym językowo, czy też segmentacja zachodzi na poziomie analizy fonetyczno-akustycznej.

Zastosowanie materiału językowego w postaci wypowiedzi nieznanego języka miało na celu wymuszenie na słuchaczach dokładnej analizy segmentalnej. Jednakże całościowy kształt dźwiękowy poszczególnych wyrazów także wywierał swój wpływ na decyzje badanych osób. Wszystkie sygnały testowe posiadały naturalny rytm, intonację i akcent. Cechy prozodyczne ułatwiały niewątpliwie zapamiętanie całego wyrazu, wobec czego procedura analityczna słuchaczy miała zarówno charakter wstępujący jak i zstępujący. Interakcyjność przetwarzania zachodziła w obecnym eksperymencie najczęściej między płaszczyzną niepełnej analizy wyrazowej a poziomem fonologicznym.

Braki segmentacyjne zdefiniowane jako ściągnięcia Złączenie /t + s/ w jeden fonem /ts/, wprowadzenie samogłoski nosowej w miejsce grupy fonemów /ɔ + n/, świadczą niewątpliwie o wykorzystaniu wiedzy fonologicznej dotyczącej rodzimego języka słuchaczy. Na podobne zjawisko wskazuje także dyftongizacja samogłosek długich w języku niderlandzkim. W języku polskim iloczasnie jest cechą fonologiczną i w związku z tym wzdłużenie samogłoski traktowane jest jako sekwencja dwóch dźwięków: niesylabicznego /j/ i samogłoski przedniej lub samogłoski tylnej i niesylabicznego /w/.

Trudności segmentacyjne mające postać rozszczepienia i dotyczące asynchronicznej miękkości spółgłosek nie dają się jednoznacznie wyjaśnić wpływem wiedzy fonologicznej badanych osób, choć w pewnym zakresie stanowią także odbicie procesów depalatalizacyjnych zachodzących w języku polskim. Wniosek jest o tyle uzasadniony, że w automatycznej segmentacji naturalnych wypowiedzi w języku polskim, po spółgłoskach miękkich pojawia się często dodatkowo segment typu /j/.

Rozszczepienie nosowego spółgłoskowego fonemu tylnojęzykowego /ŋ/ na dwa elementy /n + g/ może być interpretowane dwojako: brakiem umiejętności zapisu takiego dźwięku przez słuchaczy lub też silnym wpływem uwarunkowań fonotaktycznych. W języku polskim spółgłoska ta występuje wyłącznie przed /k,g/.

Tendencja do ściągnięcia w jeden fonem grupy spółgłoskowej /rh/ oraz /ɣh/ świadczy o wykorzystaniu wiedzy fonetycznej - /h/ i /ɣ/ są, nie tylko dla polskiego odbiorcy, bardzo

podobnymi dźwiękami uwularnymi. Redukcja geminat także ma podłoże wyłącznie fonetyczne.

Odstępstwa segmentacyjne polskich słuchaczy od lingwistycznej interpretacji fonologicznej obu języków posiadają różnorodny charakter. Wydaje się jednak, że wpływ fonologicznego poziomu przetwarzania i własnej wiedzy fonologicznej użytkownika języka jest najsilniejszy.

Przy przetwarzaniu akustycznego sygnału mowy na kod lingwistyczny, np. przy automatycznym rozpoznawaniu, czy transformacji sygnału na tekst mówiony (tzw. system speech to text), konieczne jest uwzględnienie segmentacji językowej. Przeprowadzone doświadczenie wskazało, że na poziomie fonologicznym niektóre zasady segmentacji mają charakter uniwersalny a inne są specyficzne. Przy automatycznej zamianie sygnału na tekst, forma docelowa musi składać się z dyskretnych symboli, np. liter. Automatyczna segmentacja odbywa się kolejno w płaszczyźnie akustycznej, fonetycznej i fonologicznej. Niektóre przejścia z jednego etapu na wyższy mają charakter ogólny i mogą być stosowane w różnych językach. Natomiast niektóre algorytmy segmentacji muszą odzwierciedlać zasady obowiązujące wyłącznie w danym języku.

[1] Praca wykonana w ramach problemu CPBP 02.13,
na zamówienie Zakładu Fonetyki Akustycznej IPPT PAN.

BIBLIOGRAFIA

[1] BATOGOWA-STEFFEN, M.: Automatyzacja transkrypcji fonematycznej tekstów polskich, PWN, Warszawa, 1975.

[2] DOMAGAŁA, P.: Automatyzacja procesu segmentacji sygnału mowy w układzie analogowo-cyfrowym, Prace IPPT PAN, Nr 5, Warszawa, 1984.

[3] DOMAGAŁA, P.: Tworzenie wzorców jednostek segmentalnych sygnału mowy dla jej automatycznego rozpoznawania, Prace IPPT PAN, Nr 20, Warszawa, 1985.

[4] DOMAGAŁA, P.: Automatyczna segmentacja typowych ciągów głoskowych języka polskiego, w: Wizualizacja mowy i jej zastosowania (red. W. Jassem), IPPT PAN, Warszawa, s.43-64, 1987.

[5] DOMAGAŁA, P.: Segmentalne rozpoznawanie fonokodu do sterowania maszyną roboczą, Prace IPPT PAN, Nr 11, Warszawa, 1987.

[6] GROSJEAN,F.: Spoken word recognition process and the gating paradigm, Perception and Psychophysics, 28, s.267-283, 1980.

[7] JASSEM, W.: Vowel formant frequencies as linguistic and speaker-specific features of the speech signal, w: Language in global perspective (B.F.Elson, ed.), The Summer Institute of Linguistics, s.303-312, 1986.

[8] JASSEM, W.: The phoneme and the acoustical speech signal, referat: FASE Speech '88, Edinburgh, 1988.

[9] JASSEM, W., KUBZDELA, H., DOMAGAŁA, P.: Segmentacja sygnału mowy na podstawie zmian rozkładu energii w widmie, Prace IPPT PAN, Nr 13, Warszawa, 1983.

[10] KLATT, D.H.: The problems of variability in speech recognition and in models of speech perception, Abstracts of the Tenth International Congress of Phonetic Sciences, s.287-298, Utrecht, 1983.

[11] KUBZDELA, H.: Metoda globalnego rozpoznawania wyrazów na podstawie spektrogramów binarnych, Prace IPPT PAN, Nr 28, Warszawa, 1986.

[12] MARCUS, S.M.: From 'past history' to 'interactive activation' in speech recognition, IPO Annual Progress Report, 18, s.26-31, 1983.

[13] MARSLEN-WILSON, W.D.: Aspects of human speech understanding w: Computer speech processing (Fallside, F., Woods, W.A., eds), Prentice Hall International, s. 383-404, 1983.

[14] RICHTER, L.: Rozpoznawanie wzorców wizualnych 100 najczęstszych wyrazów polskich przez osoby z głębokimi upośledzeniami słuchu, Prace IPPT PAN, Nr 14, Warszawa, 1987.

[15] RICHTER, L.: Wizualne rozpoznawanie wybranych wyrazów w oparciu o informacje segmentalne zawarte w spektrogramach komputerowych, w: Wizualizacja mowy i jej zastosowania, (red. W.Jassem), IPPT PAN, s.135-180, Warszawa, 1987.

[16] RICHTER, L.: Wpływ liczebności zbioru słownikowego i długości haseł na wyniki wizualnego rozpoznawania mowy, Prace IPPT PAN, Nr 1, Warszawa, 1988.

[17] SCHWARTZ, R.M., ZUE, V.W.: Acoustic-phonetic recognition in BBN Speechlis, Proc. Int. Conf. Speech Communication, s.129-136, 1976.

[18] TANAKA, A., TOGAVA, F.: A study of the syllable oriented recognition of continuous speech, Speech Communication, 2, 207-210, 1983.

[19] WOLF, J.J., WOODS, W.A.: The HWIM speech understanding system, w: Trends in speech recognition, (W.A. Lea, ed.), s.316-339, New York, 1980.

[20] ZUE, V.W.: Acousti-phonetic knowledge representation: implications from spectrogram reading experiments, w: Automatic speech analysis and recognition, (J.P. Haton, ed.), s.101-120, Dordrecht, 1982.