Zastosowanie MRROC++ do budowy układu sterowania robotem zdolnym do werbalnej komunikacji z człowiekiem*

Włodzimierz Kasprzak[1], Cezary Zieliński[1], Artur Janicki[2], Maciej Staniak[1]

Streszczenie

W artykule przedstawiono strukturę układu komunikacji werbalnej robota z człowiekiem zrealizowanego w systemie MRROC++. Omówiono podstawowe etapy analizy i syntezy sygnału mowy w zakresie dwustronnej komunikacji realizowanej za pomocą komend głosowych. Podano fonetyczne modele stosowane do reprezentacji rozpoznawanych i syntezowanych słów. W tym zakresie określono zasady tworzenia bazy *trzy-fonów* względnie *di-fonów*, czyli podstawowych elementów fonetycznej reprezentacji w obu zagadnieniach.

1. WSTĘP

Niniejsza praca stanowi element szerszego problemu dotyczącego analizy i syntezy obrazów i mowy na potrzeby systemu "inteligentnego asystenta operatora". Oczekujemy, że w niedalekiej przyszłości większość urządzeń technicznych obsługiwanych przez człowieka (np. roboty usługowe, urządzenia domowe) będzie posiadało interfejs użytkownika w postaci "widzącego i mówiącego" "asystenta" dysponującego pewnym zakresem inteligencji i autonomii.

Położenie operatora i jego gesty (okazywane głową lub ręką) [10] oraz jego mowa mają być (w sposób selektywny) automatycznie wykrywane [9] a następnie graficzny asystent (w postaci syntetyzowanej ludzkiej twarzy) ma komunikować człowiekowi reakcję systemu w postaci syntetyzowanej mowy i obrazie twarzy, wyrażającej różne stany emocjonalne. Celem niniejszej pracy jest analiza i synteza sygnału mowy, zasadniczo ograniczona do problemu komend głosowych, w zadaniach wymagających kontaktu z człowiekiem w wybranych zastosowaniach robotów usługowych.

2. STRUKTURA IMPLEMENTACJI W MRROC++

W systemie zbudowanym na bazie programowej struktury ramowej MRROC++ [14, 15, 16] za efektor uważa się każde urządzenie mogące oddziaływać na otoczenie, a

*Praca jest finansowana przez grant MiNI: 4T11A 003 25.

[1]Instytut Automatyki i Informatyki Stosowanej, Wydział Elektroniki i Technik Informacyjnych, Politechnika Warszawska, ul. Nowowiejska 15/19, 00–665 Warszawa, {W.Kasprzak,C.Zielinski,M.Staniak}@ia.pw.edu.pl

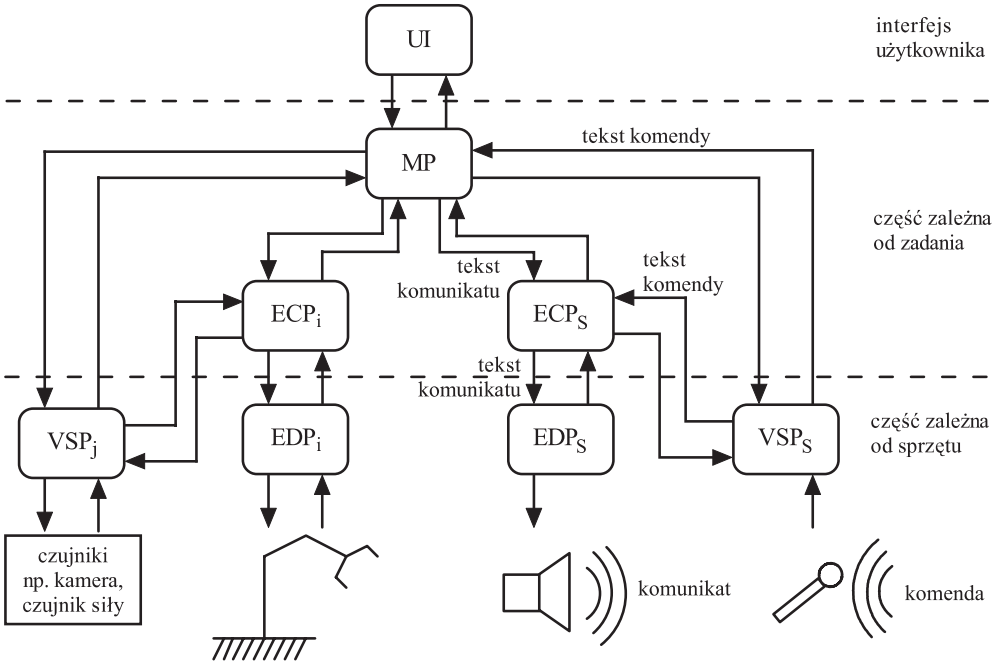
[2]Instytut Telekomunikacji, Politechnika Warszawska, ul. Nowowiejska 15/19, 00–665 Warszawa, A.Janicki@tele.pw.edu.pl

więc zarówno manipulator jak i głośnik. Natomiast receptorem jest dowolne urządzenie zbierające dane o stanie środowiska, a więc także mikrofon. Do sterowania efektorem tworzona jest para procesów *ECP*, *EDP* (Effector Control Process oraz Effector Driver Process), natomiast agregacją danych z jednego lub wielu receptorów zajmują się procesy *VSP* (Virtual Sensor Process).

Proces *EDP* zajmuje się bezpośrednim sterowaniem efektora, stąd jego kod zależy od użytego sprzętu, natomiast proces *ECP* zawiera program realizacji zadania powierzonego efektorowi do realizacji, a więc w dużej mierze zależy jedynie od klasy efektora, a nie jego typu (konkretnego sprzętu).

Proces *MP* (Master Process) jest koordynatorem systemu zawierającego wiele efektorów, natomiast proces *UI* (User Interface Process) służy głównie projektantowi do testowania systemu. Daje mu możliwość uruchamiania, zatrzymywania i chwilowego wstrzymywania procesów. Dodatkowo prezentuje na ekranie monitora komunikaty o stanie systemu, a w szczególności o wykrytych awariach i błędach w czasie jego działania.

W skład systemu przedstawionego na rys. 1 wchodzi para procesów EDP_i/ECP_i sterująca manipulatorem oraz para procesów EDP_s/ECP_s dokonująca syntezy mowy i sterowania głośnikiem.



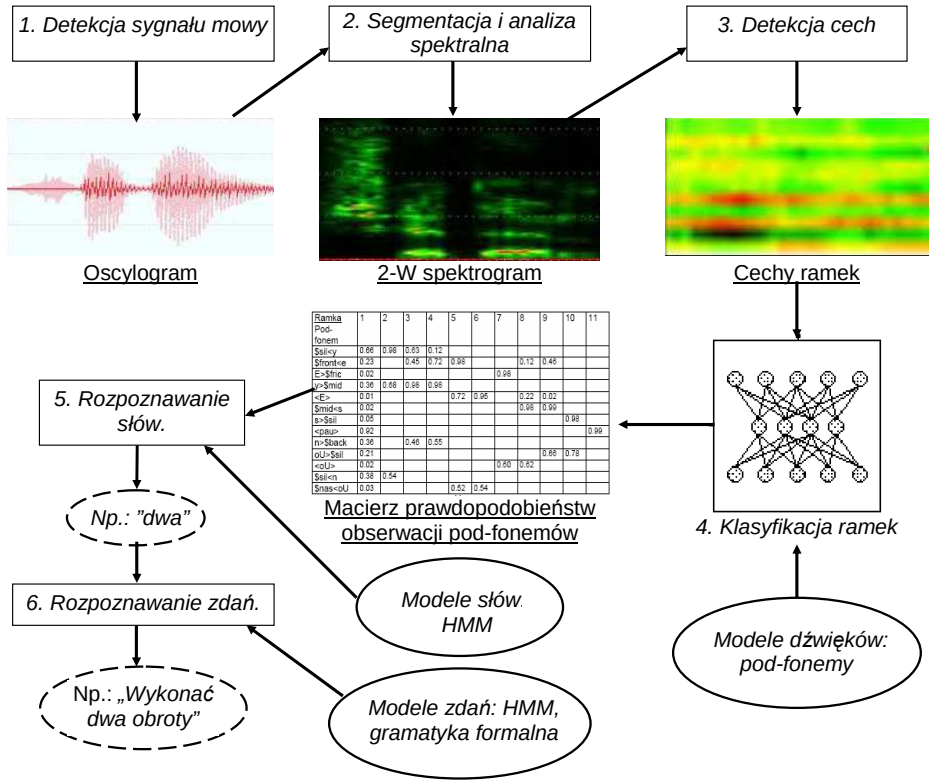
Rys. 1. Struktura systemu komunikacji dźwiękowej - podział na procesy

Proces VSP_s obsługuje mikrofon i dokonuje rozpoznania komend głosowych wydawanych przez operatora robotowi. Tekst komunikatu wraz z jego pożądaną prozodią przesyłany jest do EDP_s przez ECP_s.

Proces EDP_s zajmuje się przekształceniem tekstu komunikatu do postaci dźwię-

kowej oraz zlecaniem jego odtworzenia karcie muzycznej komputera. EDP_s w trakcie swej inicjalizacji ustawia parametry syntezy głosu takie jak: częstotliwość próbkowania, rozdzielczość, liczba kanałów. Są one niezbędne karcie muzycznej do prawidłowej syntezy dźwięku. Zwrotnie ECP_s otrzymuje od EDP_s informację o zakończeniu wygłaszania komunikatu. Synteza mowy doprowadza do wypełnienia odpowiedniego bufora komunikacyjnego kolejnymi próbkami dźwięku. Synteza mowy dla komunikatu trwającego około 5s zajmuje około 100ms, nie jest to więc czas, który wymaga zrównoleglenia syntezy i mówienia.

Zadaniem VSP_s jest rozpoznanie wydawanych komend głosowych oraz wysłanie ich w formie tekstowej do ECP_s. Proces VSP_s posiada wątek sprawdzający, czy odebrane dane zasługują na analizę, a ponadto, w którym momencie rozpoczął mówienie (wydawanie komendy) operator. Proces ECP_s uzyskawszy tekst usłyszanej komendy od VSP_s, wysyła go zarówno do EDP_s, jak i do MP w celu zlecenia wykonania komendy przez ECP_i, a w konsekwencji jej realizacji przez manipulator.



Rys. 2. Struktura procesu rozpoznawania mowy

3. ANALIZA KOMEND GŁOSOWYCH

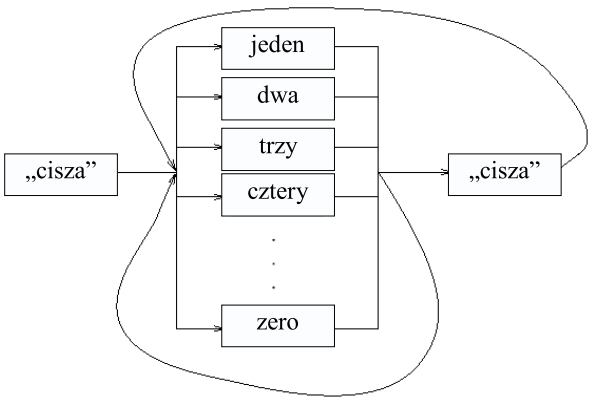
3.1. Modele HMM w procesie rozpoznawania mowy

Jak podano na rys. 2 proces analizy sygnału mowy w zakresie rozpoznawana komend głosowych może zostać przedstawiony jako potok etapów przetwarzania: (1) akwizycji sygnału dźwiękowego i detekcji sygnału użytecznego mowy w sygnale dźwiękowym, (2) przekształcenia ramek sygnału w dziedzinę częstotliwości, (3) detekcji cech ramek sygnału, (4) klasyfikacji cech w terminach jednostek fonetycznych, (5) rozpoznawania pojedynczych słów, (6) rozpoznawania zdań. Początkowe etapy 1-4 analizy sygnału mowy zostały opisane w towarzyszącym tej pracy artykule [11]. Teraz opiszemy jedynie modele fonetyczne potrzebne etapom 4-6 systemu analizy mowy.

Model fonetyczny dla systemu rozpoznawania komend przyjmuje postać jednej 3-warstwowej struktury HMM ("ukryty model Markowa"). Sama struktura modelu stworzona zostaje "ręcznie" i zależy od przyjętego słownika. Jednak macierze wag prawdopodobieństw przejść i wyjść oraz warunkowe a priori prawdopodobieństwa klasyfikacji wektorów cech, $p(\mathbf{c}|\Omega_k), k = 1, ..., K$, są określane w procesie uczenia.

3.2. Warstwa sekwencji słów

Na pierwszym poziomie *sekwencji słów* określane są sekwencje słów dozwolone w naszym języku. W przypadku rozpoznawania pojedynczych słów podłożowa gramatyka składni tego języka jest prosta. Np. jej reguły produkcji możemy zapisać jako:
<Słowo> ::= jeden | dwa | trzy | cztery | pięć | sześć | siedem | osiem | dziewięć | zero |...;
<zdanie> ::= /cisza/ [<Słowo> /cisza/];
Regułom gramatyki odpowiada struktura najwyższej warstwy modelu HMM (rys. 3).



Rys. 3. Przykład warstwy sekwencji słów w modelu HMM dla rozpoznawania komend

3.3. Warstwa pojedynczego słowa

W pośredniej warstwie *wypowiedzi pojedynczego słowa* model HMM przedstawia różne fonetyczne transkrypcje wymowy słowa. Każde słowo jest reprezentowane w

postaci *lewo-prawego* modelu HMM odpowiadającego sekwencji *fonemów*.

Fon (lub artykułowany dźwięk) - najmniejszy element w sygnale mowy. "Fon jest dźwiękiem mowy rozumianym jako fizyczne zdarzenie niezależne od miejsca i języka mówcy" (słownik Webstera). Fon może mieć różne implementacje określone przez: ton, czas trwania, intonację. IPA (Int. Phonetic Association) zdefiniowała ok. 100 fonów, z których przynajmniej 40 potrzebnych jest do transkrypcji wyrazów.

Fonemy (głoski) to podstawowe kategorie dźwięków mowy (fonów). Dwa dźwięki należą do tej samej kategorii (fonemu) jeśli są sobie wystarczająco bliskie - takie sformułowanie może powodować różne podejścia fonetyków. Liczba fonemów waha się między 30 a 60 w zależności od języka i dialektu mówcy. W języku amerykańskim (US-English) wyróżnia się najczęściej 41 fonemów. W języku polskim spotkać można większe zróżnicowanie - przyjmiemy tu 37 podstawowych fonemów.

Fonemy będziemy reprezentować za pomocą alfabetu *Worldbet* (J.L. Hieronymous et al., AT&T Bell Laboratories, 1994) [13] ujmując ich symbole w nawiasy typu "slash", np. /y/, /E/, /s/. W celu transkrypcji słów języka polskiego musimy rozszerzyć uznany alfabet *WorldBet* o następujące fonemy: /ci/ - afrykat bezdźwięczny zazębowy , obok istniejącego już /tS/ ("cz"). /si/ - spirant bezdźwięczny zazębowy, obok istniejącego już /S/ ("sz"). /zi/ - spirant dźwięczny zazębowy, obok istniejącego już /Z/ ("ż"). /dzi/ - afrykat dźwięczny zazębowy, obok istniejącego już /dZ/ ("dż"). /eN/, /oN/ - dwa dyftongi ("ę", "ą").

Samogłoski

1. Monoftongi - jedno-głoskowe samogłoski (może ich być 7): "a" - /a/, "e" - /E/, "i" - /i/, krótkie "i" - /I/, "o" - /o/, "u" - /u/, "y" - /Y/.
2. Dyftongi - dwie głoski wymawiane łącznie - zmiana jakości dźwięku między przednią a tylnią częścią (2): "ą" - /oN/, "ę" - /eN/.

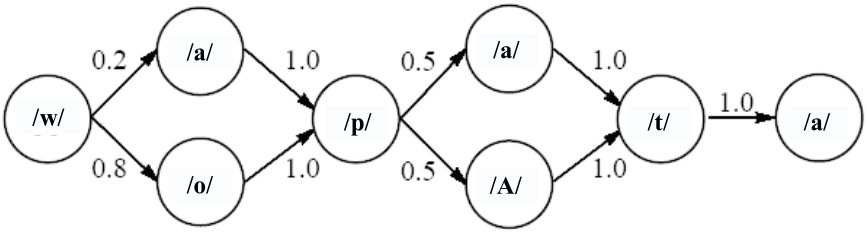
Spółgłoski

3. Ustne - pół-samogłoski i sonanty. Dźwięki pośrednie między samo- i spółgłoskami - tylko nieznaczne przeszkody w trakcie głosowym. (4): półsamogłoski - "ł" - /w/, "j" - /y/, sonanty - "l" - /l/, "r" - /r/.
4. Nosowe. Zablokowanie przepływu powietrza przez jamę ustną ale jednoczesne obniżenie podniebienia miękkiego umożliwia słaby przepływ przez nos. (3) : "m" - /m/, "n" - /n/, "ń" - /ni/.
5. Spiranty. Świszczące dźwięki tworzone w wyniku zbliżania do siebie elementów artykulacji dźwięku powodujące zawirowania przepływu powietrza. (9) : "f" - /f/, "s" - /s/, "w" - /v/, "z" - /z/, "h" - /x/, "ś" - /si/, "ź" - /zi/, "sz" - /S/, "ż" - /Z/.
6. Zwarte. Wybuchowy charakter dźwięku spowodowany najpierw zupełnym zamknięciem traktu wymowy a następnie raptownym zwolnieniem zamknięcia. (6): /p/, /t/, /k/, (bezdźwięczne), /b/, /d/, /g/ (dźwięczne).
7. Afrykaty. Głoski zwarte powodujące świst podczas zwalniania. (6): głoski "c" - /ts/, "ć" - /ci/, "cz" - /tS/, "dz" - /dz/, "dź" - /dzi/, "dż" - /dZ/.

Cisza - /sil/ i zwarcia krtaniowe /tc/, /hc/, /pc/ poprzedzające "t", "h" i "p".

Ewentualne alternatywne sposoby artykulacji słowa tworzą alternatywne ścieżki w modelu HMM. Np. jeśli dopuszczamy 4 alternatywne sposoby artykulacji słowa "łopata" to (rys. 4):

$p(\text{"/w//o//p//a//t//a/"}|\text{"łopata"}) = 0.4, \quad p(\text{"/w//o//p//A//t//a/"}|\text{"łopata"}) = 0.4,$
$p(\text{"/w//a//p//a//t//a/"}|\text{"łopata"}) = 0.1, \quad p(\text{"/w//a//p//A//t//a/"}|\text{"łopata"}) = 0.1.$



Rys. 4. Przykładowy fragment warstwy transkrypcji pojedynczego słowa w modelu HMM

Alternatywnym sposobem na reprezentację zmienności artykulacji jest ograniczenie struktury modelu transkrypcji słowa - do jednej możliwej ścieżki przejścia - z jednoczesnym zwiększeniem liczby symboli wyjściowych modelu HMM, akceptowanych (wyprowadzanych) w każdym stanie HMM.
Przykład. Przyjmiemy podstawowe transkrypcje cyfr mówionych w alfabecie *World-Bet* poszerzonym o /eN/, /dzi/, /si/, /ci/: Jeden → /y/ /E/ /d/ /E/ /n/; Dwa → /d/ /v/ /a/; Trzy → /t/ /S/ /Y/; Cztery → /tS/ /t/ /E/ /r/ /Y/; Pięć → /p/ /i/ /eN/ /ci/; Sześć → /S/ /E/ /si/ /ci/; Siedem → /si/ /E/ /d/ /E/ /m/; Osiem → /o/ /si/ /E/ /m/; Dziewięć → /dzi/ /E/ /v/ /i/ /eN/ /ci/; Zero → /z/ /E/ /r/ /o/.

3.4. Trzy-fonowa reprezentacja fonemu

Efektem wspólnej artykulacji sąsiadujących ze sobą fonemów w słowie jest potrzeba zastosowania reprezentacji kontekstowej dla fonemów.

W reprezentacji trzyfonowej dzielimy każdą głoskę na jedną, dwie lub trzy części, w zależności od tego jak duży może być na nią wpływ sąsiednich dźwięków. Np. głoska /E/ zostanie podzielona na trzy części. Głoska /n/ jest w miarę niezależna od sąsiednich fonów - dzielimy /n/ na dwie części lub nawet utrzymujemy jej 1-częściową reprezentację.

Gdyby założyć maksymalną reprezentację głosek w postaci trzech części, to liczba lewych i prawych kontekstów odpowiadałaby liczbie fonemów języka. Dla 40 fonemów potrzebowalibyśmy 64.000 trzy-częściowych modeli. Gdyby wykluczyć konteksty, które w danym języku (lub w rozpatrywanym zbiorze słów) nie zachodzą nigdy to liczba możliwych modeli zmniejszyłaby się do ok. 5.000 (lub jeszcze bardziej dla ograniczonego słownictwa).

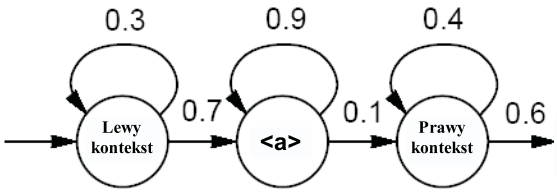
Wyróżniając grupy kontekstowe dla fonemów - zamiast rozpatrywać wpływ indywidualnych głosek rozpatrujemy wpływ grup głosek na wymawianą głoskę. Wyróżnimy osiem grup kontekstowych - w wyniku otrzymamy kilkaset (ok. 600) trzy-częściowych modeli fonemów. Nazwę grupy poprzedza znak "$":

- $Przednie - przednie samogłoski i zbliżone spektralnie ustne spółgłoski;
- $Centralne - centralne samogłoski i zbliżone spektralnie ustne spółgłoski;
- $Tylnie - tylne samogłoski i zbliżone spektralnie ustne spółgłoski;
- $Sil - cisza, przerwa, zwarcie krtaniowe;
- $Nosowe - nosowe głoski;
- $Retro - głoski typu "retroflex" - kolorowane tylnym "r";

- $Frykatywy - spiranty i afrykaty;
- $Inne - głoski zwarte i pozostałe ustne.

Przykład. Model głoski /a/ w słowie "tak":
- $Inne<a ("a" w kontekście poprzedzającej ją głoski zwartej lub pozostałej),
- <a> ("a" środkowe, bez kontekstu),
- a>$Inne ("a" w kontekście następującej głoski zwartej lub pozostałej).

Wychodząc z transkrypcji słów określamy zbiór potrzebnych fonemów a następnie podział każdego fonemu na (maksimum) trzy pod-fonemy. Odpowiada temu standardowy "lewo-'prawy" model HMM o pięciu stanach (3 stany obserwacji, jeden stan wejściowy i jeden wyjściowy) (rys. 5).



Rys. 5. Fragment warstwy pojedynczego fonemu w modelu HMM

3.5. Zintegrowany model HMM

Zestawiamy elementarne modele HMM dla fonemów w modele HMM dla słów. Modele HMM słów wstawiamy w łączny model HMM systemu rozpoznawania sekwencji słów. W zintegrowanym modelu każdy stan (poza początkowym i końcowym) jest indeksowany trzema wielkościami: indeksem słowa w słowniku, indeksem fonemu w artykulacji słowa, indeksem pod-fonemu w dekompozycji fonemu.

Stosujemy metodę przeszukiwania Viterbiego dla detekcji najlepszej ścieżki przejścia w lewo-prawym modelu HMM, gdy istnieje sekwencja wektorów cech $(\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_n)$. Ta metoda jest odmianą programowania dynamicznego, jednak jej reguła decyzyjna może być wyrażona w terminach reguły Bayesa. Algorytm przeszukiwania Viterbiego poszukuje dla kolejnej ramki najlepszego rozszerzenia każdej dotychczasowej ścieżki $(S_1, S_2, ..., S_t)$ przejścia w modelu HMM, wyznaczając iteracyjnie prawdopodobieństwa warunkowe wszystkich możliwych ścieżek modelu:
$P(\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_{t+1})|(S_1, S_2, ..., S_{t+1}) = P(\mathbf{c}_{t+1}|S) \cdot \max_S [P(S|S_t) \cdot P(\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_t)|(S_1, S_2, ..., S_t)]$

4. SYNTEZA KOMEND GŁOSOWYCH

4.1. Konkatenacyjna synteza mowy

Konkatenacyjna synteza mowy opiera się na łączeniu jednostek akustycznych, którymi mogą być między innymi wyrazy, sylaby, difony i polifony, a także fonemy i mikrofonemy. Tu oparto się na konkatenacji difonów. Synteza mowy na podstawie tekstu wymaga przeprowadzenia kolejno następujących operacji. (1) Przetwarzanie wstępne

tekstu. (2) Transkrypcja fonetyczna tekstu. (3) Opracowanie bazy difonów dla języka polskiego. (4) Wyodrębnienie wzorców prozodycznych i sterowania parametrami prozodycznymi (czas trwania głosek, energia, częstotliwość tonu krtaniowego). (5) Synteza akustyczna.

Przetwarzanie wstępne tekstu ma na celu zastąpienie ciągów znaków alfanumerycznych podawanych na wejście syntezera przez znaki literowe. Obejmuje ono między innymi cyfry, skróty, litery greckie, znaki specjalne, jednostki miary, tytuły. Po przetworzeniu wstępnym tekstu otrzymuje się ciąg liter przedzielonych spacjami i znakami interpunkcyjnymi.

Zadaniem *transkrypcji fonetycznej* jest zastąpienie liter fonemami. Przetwarzane są polskie litery (wyłącznie małe - duże są zamieniane na małe): *a, ą, b, c, ć, d, e, ę, f, g, h, i, j, k, l, ł, m, n, ń, o, ó, p, r, s, ś, t, u, w, y, z, ż, ź*.

Przyjęto zbiór 37 fonemów dla języka polskiego (w zapisie alfabetycznym SAMPA): */e/, /a/, /o/, /j/, /t/, /I/, /n/, /i/, /r/, /m/, /v/, /u/, /p/, /s/, /k/, /n'/, /d/, /w/, /l/, /S/, /z/, /s'/, /ts/, /f/, /g/, /b/, /ts'/, /Z/, /tS/, /x/, /dz/, /N/, /k'/, /z'/, /dz'/, /g'/, /dZ/*.

Fonemy w powyższej liście zostały uszeregowane pod względem częstości ich występowania. Wprowadzono modyfikację zapisu SAMPA, umożliwiającą co najwyżej dwuznakowy zapis każdego fonemu: */ts'/* zamieniono na */Ts/* i */dz'/* zamieniono na */Dz/*. W słowniku wyjątków zapisuje się słowa, których transkrypcja odbiega od zaimplementowanych reguł. Zasady transkrypcji oparto na pracy [2], poddając je pewnym uproszczeniom [9].

W pracy wykorzystano *bazę difonów* przygotowaną dla syntezera Radek [4]. Rozróżniane są następujące typy sylab (V- samogłoska, C - spółgłoska inna niż zwarta, P - spółgłoska zwarta): *V, CV, PV, PPVC, PPVP, CPV, C...CV, VC...C, CVC, PVP, PVC, C...CVC...C*. Akcent wyrazowy jest położony na przedostatniej sylabie, wyjątki umieszczono w słowniku (np. f"izIka) lub zaprogramowano - np. *samogłoska* + -liśmy, np. zrobiliśmy.

4.2. Sterowanie parametrami prozodycznymi

Podstawowe parametry prozodyczne mające wpływ na precepcję mowy to: (1) *okres* (lub częstotliwość) tonu krtaniowego (ang. *pitch*), czyli okres (lub częstotliwość) drgania strun głosowych w artykulacji mowy dźwięcznej; (2) *iloczas* - czas trwania poszczególnych głosek i przerw między nimi; (3) *głośność* poszczególnych fragmentów mowy.

W syntezie mowy z tekstu najczęściej korzysta się ze wzorców (fonemów, allofonów, difonów) mających określone wartości parametrów prozodycznych, zbliżone do ich wartości średnich. Bezpośrednia konkatenacja tych wzorców dałaby w efekcie mowę o monotonnym brzmieniu i spłaszczonej intonacji. Z tego względu niezbędna jest modyfikacja wartości parametrów prozodycznych. Ma to na celu: uzyskanie pożądanej intonacji (np. akcent, pytanie, wyrażenie emocji), wydłużenie głoski (akcent, emocje) lub całej wypowiedzi, przyspieszenie tempa mówienia. Najczęściej modyfikacji poddawany jest pitch, a także iloczas. Modyfikacja głośności nie jest krytyczna i często się z niej rezygnuje.

4.3. Ekspresyjna synteza mowy

Sterowanie parametrami prozodycznymi służyć może nie tylko nadawaniu mowie naturalnego brzmienia, ale również może służyć przekazywaniu emocji. Synteza mowy, która uwzględnia aspekty przekazywania emocji, określana jest często mianem *ekspresyjnej syntezy mowy* (ang. expressive speech synthesis). Jak wiadomo stan emocjonalny mówcy wpływa na brzmienie jego głosu. Pod wpływem stresu rośnie napięcie mięśni kontrolujących tor głosowy, stąd też głos ma wyższą średnią częstotliwość podstawową. Smutek i znudzenie powodują rozluźnienie tych mięśni, stąd też ton głosu się obniża a głoski wydłużają, maleje też dokładność artykulacji.

Zazwyczaj wyróżnia się 6 podstawowych typów emocji [7]: smutek, strach, wstręt / niesmak, zadowolenie, zdziwienie i złość. Uznaje się, że pozostałe emocje są wynikiem kombinacji powyższych 6 typów. Realizuje się je poprzez odpowiednią modyfikację parametrów prozodycznych, zwłaszcza poprzez odpowiednie modelowanie intonacji i czasu trwania poszczególnych elementów wypowiedzi.

Smutek najczęściej oddaje się poprzez obniżenie częstotliwości podstawowej F0, zawężony zakres zmian F0, spowolnione tempo wypowiedzi oraz długie pauzy między frazami. *Strach* charakteryzuje się wysokim średnim F0 i dużą dynamiką zmian F0. W wypowiedzi często występują akcenty, tempo wypowiedzi jest przyspieszone. Wydłużone są pauzy między frazami. *Wstręt / niesmak* modeluje się poprzez spowolnienie tempa wypowiedzi i wstawianie pauz w obrębie fraz. *Zadowolenie* charakteryzuje się lekko obniżoną średnią wartością F0, natomiast dynamika zmian F0 jest duża. Wydłuża się przerwy i lekko zwalnia wypowiedź oraz realizuje częstsze akcentowanie. *Zdziwienie* modeluje się poprzez rozszerzanie zakresu zmian F0 w trakcie wypowiedzi, wysoka jest też zazwyczaj końcowa wartość F0. Tempo wypowiedzi jest lekko podwyższone. *Złość* można oddać poprzez wysokie wartości F0 przy końcu wypowiedzi, dużą dynamikę zmian F0, a także szybkie tempo wypowiedzi.

5. PODSUMOWANIE

Przedstawiono strukturę systemu do komunikacji głosowej z robotem zrealizowaną z wykorzystaniem programowej struktury ramowej MRROC++. W zakresie analizy mowy określono fonetyczny model słów języka polskiego i zrealizowano model dla systemu rozpoznawania komend głosowych. W zakresie syntezy mowy w szczególności zrealizowano modelowanie stanu emocji (normalny, powaga-ostrzeżenie, radość-zabawa, zapytanie-niepewność).

LITERATURA

[1] *The CSLU Speech Toolkit*. Centre for Spoken Language Understanding, Oregon Graduate Institute of Science and Technology, USA, 2000.

[2] M. Steffen-Batogowa. *Automatyzacja transkrypcji fonematycznej tekstów polskich*. Warszawa, PWN 1975.

[3] J. E. Cahn. *Generating Expression in Synthesized Speech*. Technical Report, Massachusetts Institute of Technology, Media Laboratory, Cambridge 1990.

[4] P. Dymarski, S. Kula, S. Kukliński. Difonowy syntezer tekstowy. In: *Krajowe Sympozjum Telekomunikacji, KST'95*. Bydgoszcz, Poland, 1995. s. 133-142.
[5] S. Grocholewski. *Statystyczne podstawy systemu ARM dla języka polskiego*. Rozprawy Nr. 362, Poznań, Wyd. Politechniki Poznańskiej, 2001.
[6] A. Janicki, S. Kula. Badanie wpływu modelowania intonacji na jakość mowy syntetyzowanej z tekstu. In: *Krajowe Sympozjum Telekomunikacji, KST 2004*. Bydgoszcz, Poland, 2004. s. 213-220.
[7] A. Janicki, P. Dymarski, S. Kula. Modelowanie stanów emocjonalnych w syntezerze testowym mowy polskiej. In: *Krajowe Sympozjum Telekomunikacji, KST 2005*. Bydgoszcz, Poland, 2005. s. 111-118.
[8] J. C. Junqua, J. P. Haton. *Robustness in automatic speech recognition*. Kluwer Academic Publications, Boston, 1996.
[9] W. Kasprzak (red.). *Detekcja sygnału mowy i cech osobniczych sygnału mowy na potrzeby rozpoznawania komend głosowych*. Raport IAiIS 04-15, Politechnika Warszawska, IAiIS, Warszawa, grudzień 2004.
[10] W. Kasprzak, P. Skrzyński, A. F. Okazaki. *Sekwencja metod do identyfikacji osoby na podstawie analizy obrazu dłoni*. Raport IAiIS 06-1, Politechnika Warszawska, IAiIS, Warszawa, styczeń 2006.
[11] W. Kasprzak, A. B. Kowalski. Analiza sygnału mowy sterowana danymi dla rozpoznawania komend głosowych. In: *9-ta Krajowa Konferencja Robotyki, 2006*, *w druku*.
[12] S. Kula et al. Prosody control in a diphone-based speech synthesis system for Polish. In: *Int. Workshop "Prosody 2000 - Speech recognition and synthesis"*. Kraków, Poland, 2000, s. 135-142.
[13] T. Lander, T. Carmell. *Structure of Spoken Language: Spectrogram Reading*, Centre for Spoken Language Understanding, Oregon Graduate Institute of Science and Technology, USA, February 1999.
[14] C. Zieliński, A. Rydzewski, W. Szynkiewicz. Multi-Robot System Controllers. In: *5th Int. Symp. on Methods and Models in Automation and Robotics MMAR'98*, Międzyzdroje, Poland, August 25-29, 1998, vol. 3, s. 795-800.
[15] C. Zieliński. The MRROC++ System. In: *1st Workshop on Robot Motion and Control, RoMoCo'99*. Kiekrz, Poland, June 28-29, 1999, s. 147-152.
[16] C. Zieliński et al. Mechatronic Design of Open-Structure Multi-Robot Controllers. *Mechatronics*, vol. 11, no. 8, November 2001, s. 987-1000.

APPLICATION OF MRROC++ FRAMEWORK TO ROBOT CONTROL WITH VERBAL HUMAN-ROBOT COMMUNICATION

The paper presents the structure of a MRROC++ based application which performs verbal communication between a human and a robot. It consists of a several control processes, one speech recognition- and one speech synthesis-process. The basic structure and functions of the word-recognition process and the sentence synthesis process are described. Special focus is on two model bases for Polish speech: the tri-phone model for speech analysis and the diphone model for the synthesis process. The problem of prosody-based control for speech synthesis is addressed.