BIULETYN INSTYTUTU AUTOMATYKI I ROBOTYKI
NR 24, 2007

System dialogowy języka mówionego – przegląd problemów

Andrzej M. Wiśniewski

STRESZCZENIE: Przedstawiono strukturę systemu dialogowego języka mówionego. Scharakteryzowano pożądane własności składników funkcjonalnych systemu: urządzenia rozpoznawania mowy, procesora językowego, sterownika (menedżera) dialogu i syntezatora mowy. Scharakteryzowano przykładowe realizacje systemów dialogowych języka mówionego.

SŁOWA KLUCZOWE: system dialogowy, rozpoznawanie mowy, rozumienie mowy, synteza mowy

1. Wprowadzenie

Rośnie zapotrzebowanie na informację. Coraz więcej ludzi wykorzystuje Internet poszukując informacji dla celów edukacyjnych, finansowych, rozrywkowych czy do podejmowania decyzji. Coraz częściej ludzie są zainteresowani dostępem do informacji w ruchu (w każdej chwili, w dowolnym miejscu), poprzez telefon (stacjonarny, komórkowy czy internetowy). Wtedy tradycyjna klawiatura i myszka są niepraktyczne lub niedostępne. Wygodnym rozwiązaniem jest zastosowanie interfejsu głosowego, który zapewni użytkownikowi możliwość mówienia i słyszenia w języku naturalnym. Dotyczy to zwłaszcza małych, mieszczących się w dłoni urządzeń (iPod, palmtop) oraz dostępnych przez telefon portali głosowych, ale również komputerów przenośnych i stacjonarnych. Język mówiony jest atrakcyjny, ponieważ jest najbardziej naturalnym, najefektywniejszym i najtańszym sposobem komunikacji między ludźmi.

Dialog jest interakcją (wzajemnym oddziaływaniem, współdziałaniem) pomiędzy użytkownikiem i komputerem w osiągnięciu szczególnego celu (norma ISO 9241). Użytkownik jest osobą współdziałającą z komputerem. Jeżeli parę: akcja użytkownika i skojarzona z nią odpowiedź komputera (lub na odwrót), nazwiemy transakcją, wtedy dialog jest serią transakcji. Transakcja jest

najmniejszą jednostką interakcji człowiek – komputer.

W ostatniej dekadzie jesteśmy świadkami powstawania nowego rodzaju interfejsu człowiek - komputer, umożliwiającego użytkownikom komunikowanie z komputerem za pomocą dialogu (języka) mówionego. Na interfejs użytkownika składają się: sterowanie (umożliwia użytkownikowi tworzenie i przekazywanie poleceń i danych do komputera), zobrazowanie (umożliwia komputerowi zwracanie się, mówienie, do użytkownika) i dialog. Aby zapewnić skuteczny i wygodny dostęp do informacji, a także umożliwić ich wytwarzanie i przetwarzanie, interfejs łączy kilka technologii języka naturalnego.

2. System dialogowy

System dialogowy jest interfejsem systemu komputerowego, przeznaczonym do konwersacji z człowiekiem. System dialogowy wykorzystuje tekst, mowę, grafikę, sensory, stymulatory, gestykulację i inne sposoby komunikacji na wejściu i wyjściu interfejsu.

Celem systemu dialogowego jest ułatwić użytkownikowi realizację usługi, której sformułowanie (przeprowadzenie) za pomocą pojedynczego zdania może być niemożliwe. Typowy scenariusz realizacji usług w systemie dialogowym jest następujący:

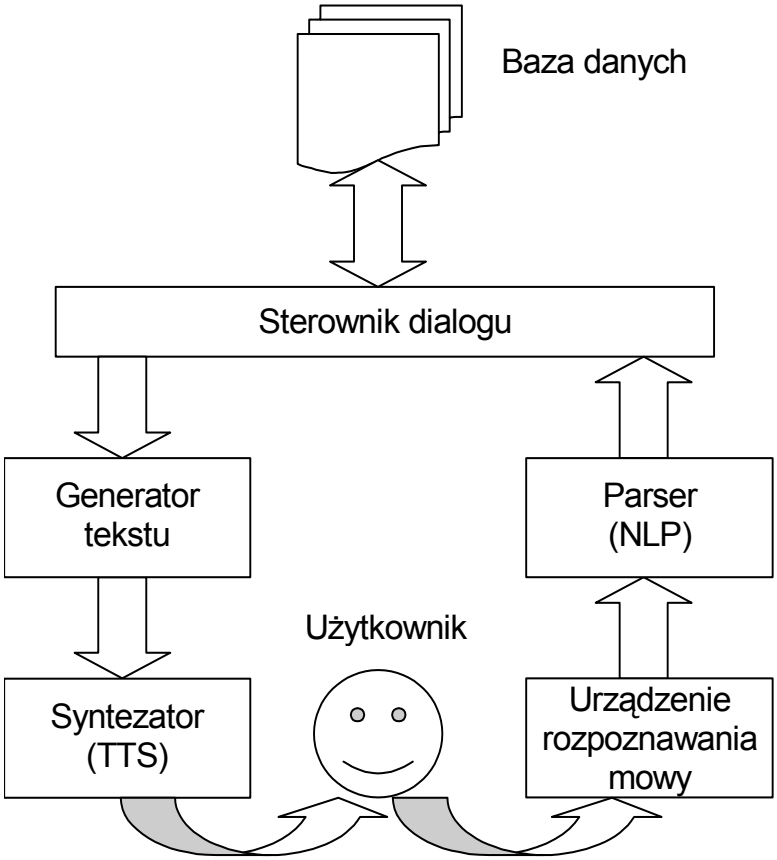
- użytkownik chce uzyskać informacje zawarte w bazie danych (np. rozkład jazdy pociągów, serwis bankowy) za pomocą telefonu,
- użytkownik, przy pomocy systemu dialogowego, dostarcza niezbędnych danych do wyszukania pożądanej informacji,
- system przejmuje kierowanie dialogiem, gdy pojawiają się niezrozumienia.

Architekturę typowego systemu dialogowego języka mówionego (spoken dialogue system, SDS) przedstawia rys. 1. Działanie SDS przebiega następująco:

- całością steruje sterownik dialogu, który umożliwia wymianę informacji z użytkownikiem, a tym samym dostęp do bazy danych i jej uaktualnianie,
- interakcja składa się z sekwencji transakcji (cyklów pytanie/odpowiedź), gdzie pytania są tak projektowane, aby ograniczyć odpowiedź do określonego zbioru informacji,
- odpowiedź użytkownika jest przetwarzana przez urządzenie

rozpoznawania mowy, którego wyjście (zwykle niejednoznaczne) jest przekształcane przez interpreter języka naturalnego – np. parser (natural language processing, NLP) - do postaci quasi-logicznej,

- sterownik, bazując na nowym wejściu, uaktualnia swój wewnętrzny stan i planuje następną akcję,
- postępowanie trwa, aż potrzeby użytkownika zostaną zaspokojone – wtedy interakcja jest przerywana.



Rys. 1. Architektura systemu dialogowego języka mówionego (SDS)

Stosuje się zamiennie następujące terminy: system dialogowy języka mówionego, interfejs konwersacyjny, system konwersacyjny.

Możliwe są różne modyfikacje przedstawionej architektury systemu

dialogowego, dostosowujące jego właściwości do potrzeb konkretnej aplikacji, np. uzupełnienie kanału głosowego na wyjściu interfejsu kanałem wizualnym (w przypadku, gdy wyjście ma również postać rysunków, tablic czy teksu wyświetlanego na ekranie monitora).

Kryteria projektowania SDS są różnorodne i zmieniają się, lecz podstawowym celem jest realizacja systemu, który umożliwi użytkownikowi szybką i dokładną realizację pożądanych zadań, w szczególności uzyskanie informacji. Aby osiągnąć ten cel, należy zaprojektować odpowiedni dialog, wiernie rozpoznawać mowę, zdefiniować miary zaufania do wyników rozpoznawania oraz generować istotne i dokładne prozodycznie wiadomości wyjściowe. Dialog powinien zapewnić inicjatywę zarówno użytkownikowi, jak i systemowi (mixed-initiative) i nie powinien ograniczać użytkownika do odpowiedzi na proste pytania systemu.

W procesie projektowania SDS istotne są następujące zadania:

- specyfikowanie dialogu i sterowanie jego przebiegiem,
- ograniczenie zakresu rozpoznawania wypowiedzi do dziedziny aplikacji i interpretacja wyjścia urządzenia rozpoznawania mowy,
- generowanie odpowiedzi właściwej kontekstowo (zgodnej z dotychczasowym przebiegiem dialogu).

System dialogowy charakteryzują następujące własności:

- pracuje w ograniczonej znaczeniowo dziedzinie - ograniczony słownik (najwyżej kilka tysięcy słów, zwykle około tysiąca),
- przeznaczony jest do pracy z użytkownikami nieprzygotowanymi (a więc rozpoznający mowę ciągłą, rozumiejący mowę spontaniczną i równoważniki zdań, radzący sobie z fragmentami słów, zjawiskami pozalingwistycznymi, czy przerwami wypełnionymi dźwiękami bez znaczenia, typu: mmm, aaa),
- zapewnia ograniczoną swobodę dialogu (użytkownik nie jest całkowicie swobodny: formułowane zdania mogą być zbyt długie i złożone, mogą przekraczać możliwości rozumienia systemu) - sterowanie przejmowane jest przez system, gdy pojawiają się kłopoty ze zrozumieniem,
- umożliwia naturalną interakcję - użytkownik może odwoływać się do informacji, która pojawiła się w dialogu wcześniej i realizacja życzenia musi brać pod uwagę wszystkie dotąd zebrane informacje,
- dostarcza sposobów pokonania trudności - zachęca do używania krótkich wypowiedzi, aby zmniejszyć ryzyko błędów rozpoznawania, oferuje sposoby wznowienia rozmowy po błędach rozumienia.

Interakcję w systemie dialogowym języka mówionego powinny

cechować:

- niezależność od mówcy,
- stosowanie mowy ciągłej (menu ze słowami izolowanymi jest zwykle niepraktyczne),
- stosowanie swobodnego i naturalnego języka (od przypadkowych użytkowników trudno wymagać stosowania prawidłowej syntaktyki),
- zapewnienie zarządzania dialogiem (sterowanie dialogiem musi być tak zaprojektowanie, aby pogodzić swobodę użytkownika z koniecznością zachowania kontroli systemu).

3. Struktura systemu dialogowego języka mówionego

Strukturę funkcjonalną systemu dialogowego języka mówionego przedstawia rys. 2. Oprócz elementów składowych systemu pokazano dziedzinowy zakres wiedzy wykorzystywanej podczas tworzenia systemu dialogowego, jak również główne modele konstruowane na potrzeby kolejnych etapów przetwarzania danych w systemie. Poniżej omówiono główne własności elementów funkcjonalnych systemu dialogowego.

3.1. Własności systemu rozpoznawania mowy

Rozpoznawanie mowy, będące elementem wstępnym i bardzo istotnym dla wszystkich kolejnych działań oraz dla jakości pracy całego systemu dialogowego, powinno cechować się:

- niezależnością od mówcy,
- możliwością rozpoznawania mowy ciągłej (spontanicznej),
- określonym precyzyjnie słownikiem rozpoznawanych słów (w zasadzie powinien zawierać wszystkie słowa, których może użyć użytkownik),
- umiejętnością reakcji na nieznane słowo lub zdarzenie nielingwistyczne (kaszel, niepewność, przerwy, powtórzenia).

Współcześnie w automatycznym rozpoznawaniu mowy stosowane są podejścia, określane jako:

akustyczno – fonetyczne (acoustic-phonetic approach),

rozpoznawania wzorców (pattern-recognition, template-based approach).

Metoda akustyczno-fonetyczna automatycznego rozpoznawania mowy

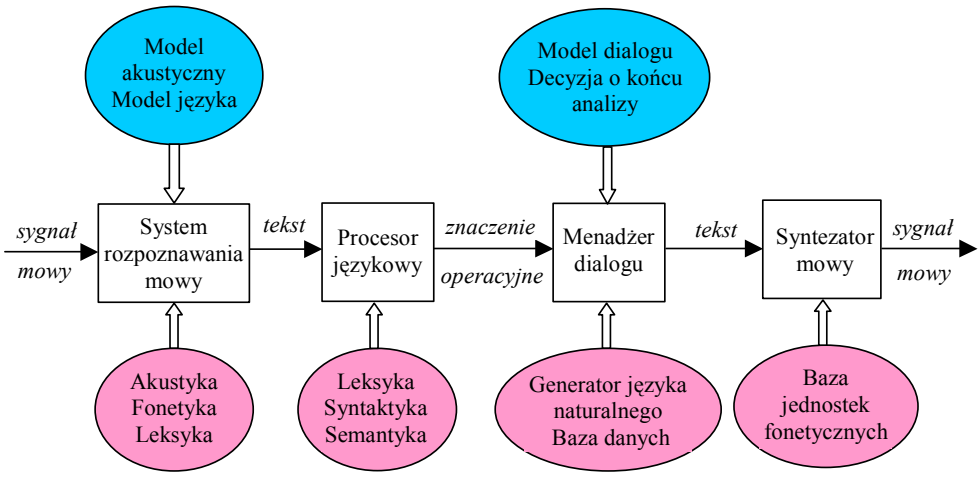
bazuje na założeniu, że:

- istnieje skończona liczba dźwięków (symboli dźwiękowych) języka mówionego,
- dźwięki są w pełni rozróżnialne poprzez zbiór charakterystyk akustycznych, które są wynikiem badań akustyczno – fonetycznych nad sygnałem mowy.

Pierwsze założenie jest spełnione: każdy dźwięk jest generowany przy określonej konfiguracji traktu głosowego. Co prawda liczba możliwych konfiguracji traktu głosowego jest nieograniczona, lecz ze względu na możliwości percepcji sygnału mowy przez człowieka, liczba rozpoznawanych dźwięków mowy w każdym znanym języku naturalnym jest skończona.

Sygnał mowy jest sekwencją dźwięków (jednostek akustycznych), które są realizacją fizyczną indeksowanych unikalną nazwą jednostek fonetycznych. Rozróżnialność dźwięków jest trudnym do spełnienia wymaganiem, ponieważ sygnał mowy charakteryzuje się dużą zmiennością związaną z mówcą, wpływem kanału transmisji oraz kontekstem (sąsiedztwem innych dźwięków).

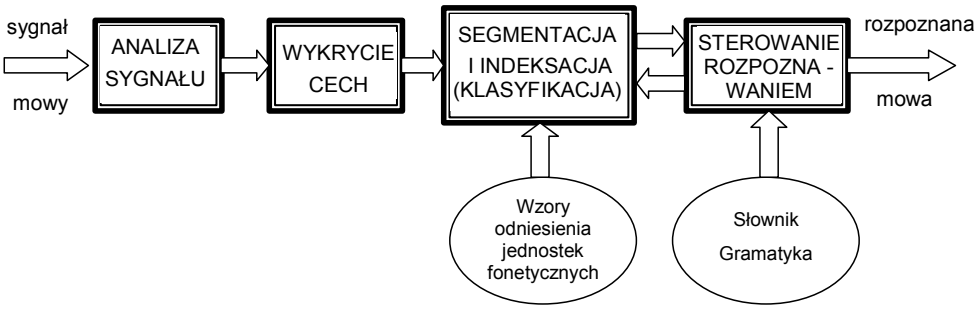
W rozpoznawaniu akustyczno – fonetycznym najczęściej stosuje się najmniejszą jednostkę fonetyczną – fonem, traktowany jako zespół cech dystynktywnych (jego realizacją fizyczną jest głoska, czyli dźwięk mowy). Stosowana też bywa sylaba, w której zasadniczą rolę odgrywa samogłoska.



Rys. 2. Struktura funkcjonalna systemu dialogowego języka mówionego

Na rys. 3 przedstawiono główne zadania realizowane w procesie

rozpoznawania mowy metodami akustyczno-fonetycznymi. Wynikiem analizy sygnału mowy (najczęściej stosowana jest analiza widmowa) jest wykrycie cech akustycznych umożliwiających rozpoznanie jednostek fonetycznych. Rozpoznawanie polega na sekwencyjnym dekodowaniu segmentów sygnału mowy na podstawie charakterystyk akustycznych tego sygnału i znanych związków między tymi charakterystykami i jednostkami fonetycznymi. Charakterystyki akustyczne sygnału mowy najczęściej mają związek ze sposobem wytwarzania mowy przez człowieka, w szczególności z modelem typu pobudzenie – filtr.



Rys. 3. Rozpoznawanie mowy metodą akustyczno-fonetyczną

W rozpoznawaniu mowy najczęściej wykorzystywane są następujące charakterystyki akustyczne związane z:

- pobudzeniem:
  - częstotliwość tonu podstawowego,
  - energia sygnału,
  - obecność w pobudzeniu sygnału okresowego i/lub przypadkowego, oznaczająca dźwięczność lub bezdźwięczność fonemów,
- filtrem (traktem głosowym):
  - częstotliwości formantowe, zwykle pierwsze trzy, będące maksimami lokalnymi amplitudowej charakterystyki częstotliwościowej traktu głosowego,
  - obecność w transmitancji traktu głosowego zer charakterystycznych dla dźwięków nosowych, czyli nosowość fonemu,
  - stosunek energii składowych wysoko- i niskoczęstotliwościowych.

Cechy akustyczne zwykle wyznaczane są przez równoległy układ detektorów, a ich liczba powinna zapewnić jednoznaczne rozróżnienie wszystkich fonemów (stąd cechy te nazywa się wyróżniającymi lub dystynktywnymi). Najważniejszy i najtrudniejszy jest etap segmentacji i indeksacji, łącznie zwany klasyfikacją (ang. odpowiednio: segmentation, labelling, annotation):

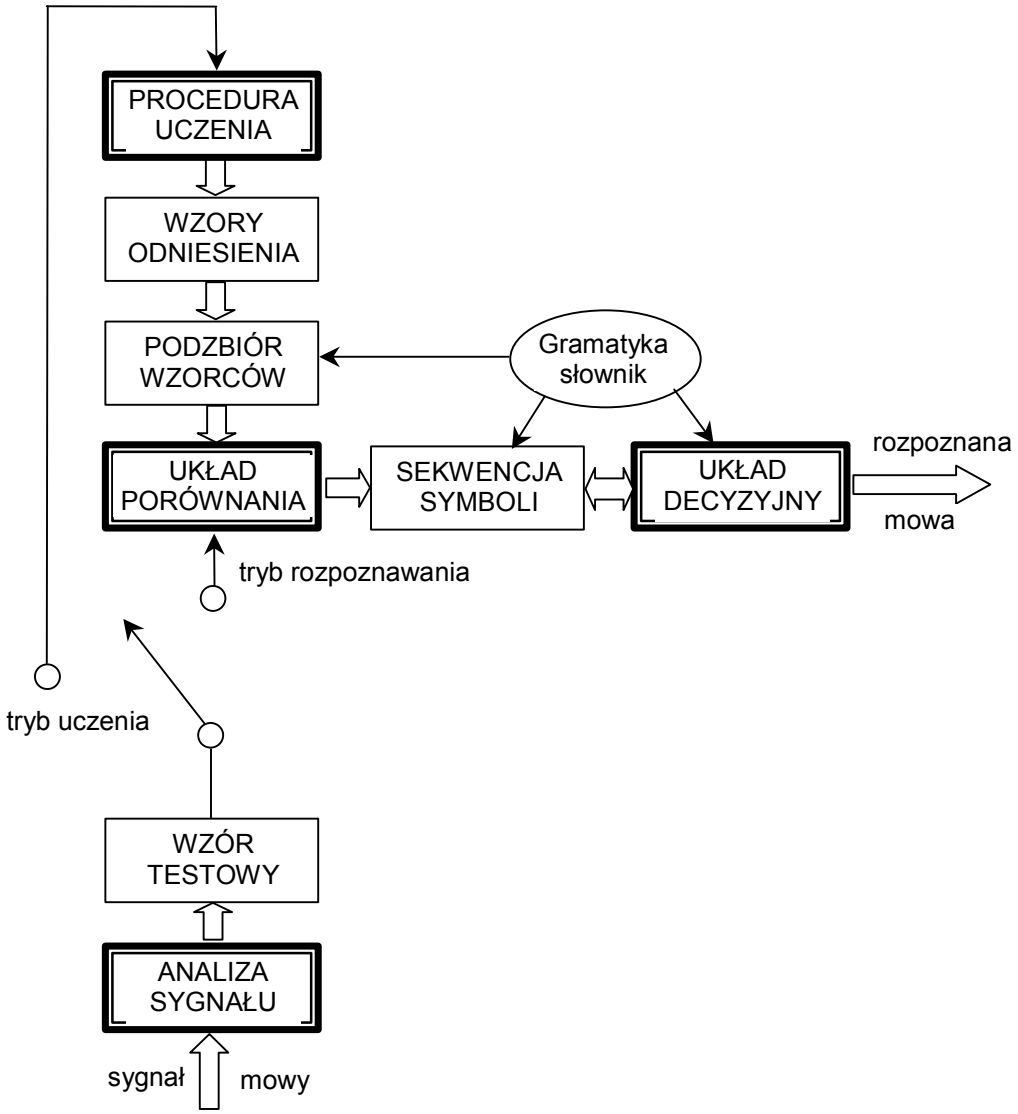
- najpierw wyszukiwane są fragmenty (segmenty) sygnału mowy, w których jego cechy akustyczne są stałe lub zmieniają się niewiele,
- następnie przypisuje się tym segmentom zgodnie z wyznaczonymi cechami akustycznymi jeden lub więcej indeksów (symboli fonetycznych). Wykorzystuje się tutaj eksperymentalnie wyznaczone wzory odniesienia (reference pattern) dla wszystkich rozpoznawanych jednostek fonetycznych. Wzory odniesienia najczęściej mają postać wiedzy o występowaniu lub braku jakichś cech albo wartości progowych lub wzajemnych zależności (proporcji) zmierzonych wcześniej cech akustycznych.

Aby prawidłowo rozpoznać mowę stosowany jest jeszcze jeden krok - sterowanie rozpoznawaniem - w którym do wyznaczenia końcowego wyniku wykorzystuje się wiedzę o ograniczeniach realizowanego zadania rozpoznawania mowy (słowa muszą pochodzić ze słownika właściwego dla pragmatyki systemu, ciągi słów powinny spełniać reguły syntaktyki i semantyki właściwe dla gramatyki języka).

Metody akustyczno-fonetyczne są interesującą ideą – umożliwiają rozpoznawanie sygnału mowy bez konieczności wcześniejszego tworzenia modeli akustycznych rozpoznawanych jednostek fonetycznych. Jednak, mimo ponad 50 lat ich rozwijania, są trudne do praktycznej realizacji i wymagają jeszcze rozległych badań oraz głębszego zrozumienia problemów.

Metoda rozpoznawania wzorców w rozpoznawaniu mowy wykorzystuje wzory (próbki), będące najczęściej obserwacjami pozyskiwanymi z segmentów sygnału mowy (ramek), które wydzielane są oknem o stałej długości. W przeciwieństwie do metody akustyczno – fonetycznej, nie wyznacza się charakterystyk akustycznych związanych ze sposobem wytwarzania sygnału mowy, jak również nie wydziela się z sygnału mowy segmentów o zróżnicowanej długości, odpowiadających fonemom.

Strukturę systemów rozpoznawania mowy metodą rozpoznawania wzorców zilustrowano na rys. 4.



Rys. 4. Rozpoznawanie mowy metodą rozpoznawania wzorców

Charakterystyczne dla tej metody rozpoznawania są dwa tryby pracy:

- „tryb uczenia" (treningowy), w którym ze zbiorów wzorów testowych (test pattern), pozyskanych z wypowiedzi uczących, tworzy się wzory odniesienia, czyli wzorce (reference pattern), reprezentujące jednostki

(symbole) fonetyczne,

- „tryb rozpoznawania", w którym pozyskany z rozpoznawanej wypowiedzi wzór testowy (lub ich sekwencję) porównuje się z każdym wzorem odniesienia, czyli wzorcem.

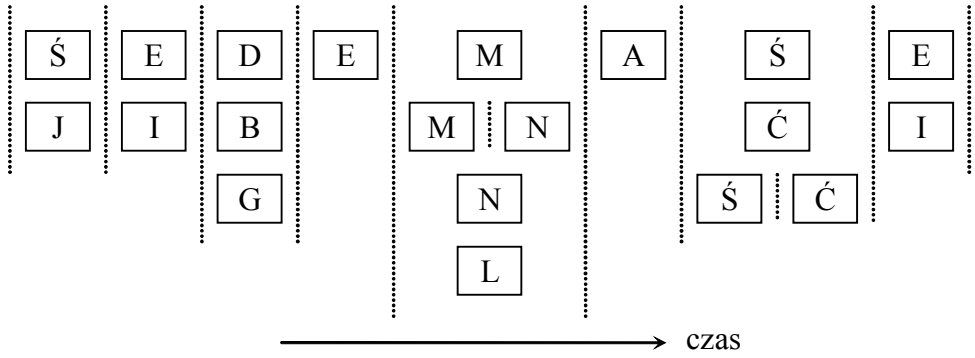
Wzorce mogą mieć postać szablonu (template) lub modelu statystycznego (statistical model). Podobieństwo wzoru testowego do wzorców w postaci modelu statystycznego (np. dla powszechnie stosowanych ukrytych modeli Markowa, HMM) jest określane najczęściej jako prawdopodobieństwo wygenerowania tego wzoru przez modele.

Liczebność zbioru wzorców w każdym miejscu rozpoznawanej wypowiedzi może być zmniejszana, np. przez zastosowanie reguł prostej gramatyki o skończonej liczbie stanów do rozpoznawania ciągów jednostek fonetycznych.

Przypisanie wzoru testowego (jednego lub częściej ich sekwencji) do określonego wzoru odniesienia stanowi wynik rozpoznawania w pierwszym jego etapie. W drugim etapie sekwencji jednostek fonetycznych przypisywany jest wyraz ze słownika wyrazów rozpoznawanych. Zastosowanie różnych reguł (ograniczeń) gramatycznych pozwala na zmniejszenie niepewności w procesie przekształcania rozpoznanej sekwencji symboli fonetycznych w wyraz. Jakość rozpoznawania mierzona jest wskaźnikiem dopasowania, który może mieć sens prawdopodobieństwa. Zwykle rozpoznanie jest niejednoznaczne, gdyż segmentowi sygnału mowy może być przypisany więcej niż jeden symbol fonetyczny. Wówczas wynikiem etapu rozpoznawania jest nie pojedyncza sekwencja, lecz sieć symboli fonetycznych z przypisanymi wartościami wskaźnika dopasowania.

Przykładowy wynik automatycznego rozpoznawania liczb dwucyfrowych przedstawiony został na rys. 5 (zastosowano symbole przyjętej transkrypcji fonetycznej języka polskiego). Jest to sekwencja czasowa zbiorów fonemów rozpoznanych z różną jakością (symbole umieszczone wyżej mają większy wskaźnik dopasowania do rozpoznawanego sygnału mowy).

Jednym z możliwych rozwiązań w analizowanym przykładzie jest słowo ŚEDEMNAŚĆE (siedemnaście w transkrypcji gramatycznej). Innym możliwym rozwiązaniem jest słowo JEDENAŚĆE (jedenaście). Oznacza to, że wynik rozpoznania mowy w przykładzie jest niejednoznaczny, chociaż pierwszy z nich jest bardziej prawdopodobny (lepiej dopasowany do sygnału wejściowego).



Rys. 5. Przykładowy wynik automatycznego rozpoznawania

W ogólnym przypadku wyjście urządzenia automatycznego rozpoznawania mowy może przybierać jedną z poniższych postaci:

- pojedyncze zdanie;
- lista N najlepszych zdań (najlepiej dopasowanych do sygnału wejściowego): jest to wskazane, gdy okaże się, że z powodu błędów rozpoznania zdanie najlepsze jest niegramatyczne, liczba N może być duża;
- krata słów: lista słów ważonych wskaźnikiem dopasowania, zwykle charakteryzuje się dużą redundancją i w efekcie długim czasem pracy procesora językowego;
- tzw. graf słów: rozwiązanie pośrednie (grupa N najlepszych zdań, w których wspólne części są połączone w celu utworzenia grafu) – daje to takie same możliwości jak lista N najlepszych zdań, lecz pozwala na przyśpieszenie procesu.

Kluczem do sukcesu w tej metodzie rozpoznawania jest proces porównywania wzorów testowych i wzorów odniesienia. Dość wcześnie zaczęto stosować technikę zwaną liniową normalizacją czasową, która pozwoliła przezwyciężyć trudności związane ze zmiennością czasu trwania wymawianych słów. Długości wzorów były normalizowane do standardowego czasu trwania drogą wydłużania (skracania) przez zastosowanie wyznaczonego rozszerzenia (kompresji) skali czasu równomiernie dla całej próbki. Porównanie otrzymanych w ten sposób wzorów o stałej długości polega na obliczeniu odległości euklidesowej między tymi wzorami.

Metoda rozpoznawania wzorców wykorzystująca jako wzorce (wzory odniesienia) modele statystyczne jest chętnie stosowana z powodu prostoty, odporności na zakłócenia ze strony środowiska oraz niezależności na zmiany

słownictwa, zbioru charakterystyk, algorytmów porównywania i reguł decyzyjnych. Liczne jej aplikacje pokazały wysoką skuteczność w realizacji zadania automatycznego rozpoznawania mowy.

3.2. Własności procesora językowego (modułu przetwarzania języka naturalnego, modułu rozumienia)

Procesor językowy dostarcza reprezentacji znaczenia operacyjnego rozpoznanej frazy. Na obecnym etapie rozwoju umożliwia rozumienie ograniczone do podzbioru języka naturalnego i dla określonej dziedziny aplikacji (pragmatyka). Przyczyną największych trudności w przetwarzaniu języka naturalnego jest brak ogólnego sposobu:

a) definiowania rozwiązywanego problemu (w wyniku tego trudno ocenić wyniki przetwarzania języka naturalnego (NLP) w różnych aplikacjach systemów),
b) automatycznego pozyskiwania informacji potrzebnej do efektywnej pracy z nowymi aplikacjami dziedzinowymi, nowymi słowami, nowymi znaczeniami słów, nowymi strukturami gramatycznymi.

W zaawansowanych systemach dialogowych przetwarzanie języka spełnia podwójną rolę:

- umożliwia zrozumienie wejścia mówionego (interpretację łańcuchów słów wyznaczonych przez system rozpoznawania mowy);
- jest dodatkowym źródłem wiedzy (ograniczeń), które - przez odrzucenie łańcuchów słów bezsensownych oraz określenie łańcuchów słów sensownych – poprawia zarówno rozpoznawanie jak i rozumienie.

W procesie przetwarzania języka naturalnego wykorzystuje się wiedzę lingwistyczną, a w szczególności syntaktykę i semantykę. Istniejące rozwiązania systemów dialogowych wyraźnie rozdzielają reprezentację syntaktyczną i semantyczną języka. Przyczynami takiego postępowania jest większa łatwość reprezentacji (wyboru najodpowiedniejszego formalizmu można dokonać oddzielnie) oraz możliwość zmian, uaktualniania, a także adaptacji dla innych dziedzin i języków.

Tradycyjnie analiza języka naturalnego jest sterowana syntaktyką - wykonywana jest pełna analiza syntaktyczna, która usiłuje wyjaśnić rolę wszystkich słów w wypowiedzi. Takie podejście, gdy pojawiają się nieznane słowa, nowe konstrukcje językowe, błędy rozpoznawania i zdarzenia charakterystyczne dla mowy spontanicznej, rzadko kończy się sukcesem. Stąd próby analizy sterowanej semantyką w dialogach mówionych w ograniczonej

dziedzinie. Trwają prace nad łącznym wykorzystaniem wiedzy syntaktycznej i semantycznej już na etapie automatycznego rozpoznawania mowy, gdyż panuje przekonanie, że jednoczesne zastosowanie wielu ograniczeń może zwiększyć efektywność (zmniejszyć czasochłonność i poprawić jakość) rozpoznawania, a tym samym rozumienia języka.

Przetwarzanie syntaktyczne (rozbiór gramatyczny, analiza zdania, parsowanie) jest najbardziej dojrzałym obszarem NLP i polega na rozpoznaniu struktury gramatycznej zdania, umożliwiając jednocześnie:

- sprawdzenie, czy fraza wejściowa jest prawidłowo sformułowana,
- uproszczenie procesu określania znaczenia (rozumienia),
- pomoc w wykryciu nowych i niezwykłych znaczeń.

Dotychczas sformułowano i zastosowano różne formalizmy syntaktyczne, jednak wszystkie dostarczają niekompletnego opisu zjawisk występujących w języku naturalnym. Dla języka mówionego stosuje się modyfikacje metod zastosowanych dla języka pisanego: trzeba uwzględnić fakt, że sekwencja słów wyznaczonych przez urządzenie rozpoznające może zawierać błędy (wynik rozpoznawania w postaci kraty lub grafu wprowadza alternatywy do przetwarzania językowego).

Każda metoda analizy jest efektywna dla zdań prostych i krótkich. Szczególnych trudności przysparza rozbiór gramatyczny zdań spontanicznych. Typowe wypowiedź w mowie spontanicznej może wyglądać następująco:

Zatem – chciałbym wiedzieć mhm – pociąg, który wyjeżdża o czwartej z Poznania, o której, tak, o której przyjeżdża on do Warszawy.

Powyższy przykład pozwala na następujące wnioski:

- rzeczywiste zdania są złożone: niezbędna jest rozległa wiedza do przedstawienia ich struktury gramatycznej,
- istotna informacja jest przekazywana w ‘wyspach’ („o czwartej”, „z Poznania”,...), złożoność syntaktyczna głównie leży w przestrzeni między wyspami, w nieistotnych semantycznie segmentach zdania.

Wnioski sugerują zastosowanie analizy częściowej, aby zwiększyć odporność algorytmów na zakłócenia. W pełnej analizie musi być analizowane całe zdanie, zatem może być potrzebna obszerna wiedza (szczególnie do modelowania niegramatyczności w wejściu mówionym). Gdy pełna analiza całego zdania nie jest możliwa, analizuje się pewne segmenty zdania w nadziei, że zawierają istotną informację dla jego prawidłowego zrozumienia (określenia znaczenia operacyjnego w ograniczonej dziedzinie). Częściowa analiza może znacznie przyśpieszyć prawidłowe rozumienie zdań dla ograniczonej wiedzy lingwistycznej. Takie podejście może być przyczyną błędnej interpretacji złożonych konstrukcji językowych, lecz jednocześnie umożliwia analizę

wypowiedzi spontanicznych. Stosuje się różne implementacje tej koncepcji: można albo uruchomić częściową analizę, gdy pełna skończyła się fiaskiem, albo stosować częściową analizę od początku procesu NLP, a następnie zastosować dodatkowy mechanizm do sklejania znaczeń poszczególnych fraz wypowiedzi w celu przeprowadzenia pełnej analizy znaczeniowej.

Przetwarzanie semantyczne ma na celu określenie znaczenia analizowanego zdania. Opracowano wiele języków reprezentacji znaczeniowej, jednak brak jest języka jednolitego dla wszystkich zakresów NLP. Trudności powoduje fakt, że znaczenie operacyjne wypowiedzi zależy od pragmatyki aplikacji, w szczególności od kontekstu oraz od celu do osiągnięcia.

Najmniej rozpoznanym i najtrudniejszym aspektem NLP jest modelowanie kontekstu i jego wykorzystanie. Kontekst nie jest czasowo zlokalizowany (jak w sygnale mowy), jest szeroki i niezwykle silny, może sięgać odległych słów wypowiedzianych i takich, które dopiero będą wypowiedziane. Kontekst może obejmować zakres wielu zdań, akapitów, nawet dokumentów.

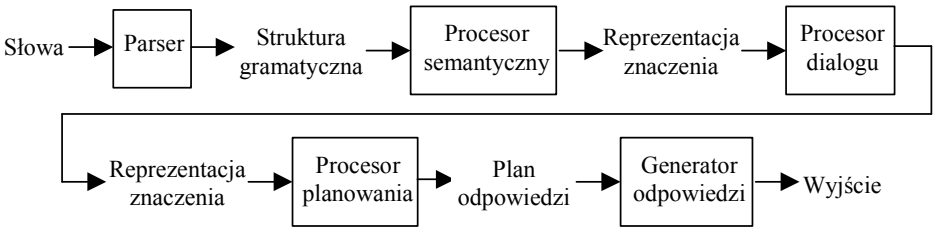
Określenie znaczenia operacyjnego wymaga określenia odniesień zaimków, zrozumienia zdań eliptycznych, fałszywych początków wypowiedzi, błędów, nieklasycznych postaci języka. Znaczenie operacyjne zależy od wielu innych zjawisk językowych, nawet właściwie dotąd formalnie nie scharakteryzowanych.

Znaczenie operacyjnie zmienia się wraz z kolejnymi wejściami, zależy od przebiegu dialogu. Oznacza to, że może być potrzebna zmiana stanu dialogu, żeby późniejsze wejście użytkownika było rozumiane w kontekście odpowiedzi wcześniej udzielonej użytkownikowi. Tego typu sprzężenia są bardzo ważne dla przetwarzania języka naturalnego, ponieważ rzeczywisty język rzadko bywa izolowanymi zdaniami.

Wynikiem powyższych rozważań jest widzenie procesu przetwarzania języka naturalnego jako sekwencji operacji, wykonywanych na ciągu słów, będących wyjściem urządzenia automatycznego rozpoznawania mowy (rys. 6). Nie ma zgody, czy planowanie i generowanie odpowiedzi są częścią przetwarzania języka, czy też częścią następnego procesu: sterowania dialogiem.

Komunikację werbalną między ludźmi, która jest procesem dwukierunkowym dotyczącym aktywnych uczestników, nazywa się dyskursem. Wzajemne zrozumienie osiąga się poprzez bezpośrednie i pośrednie oddziaływania słowne, wymianę, wyjaśnienia i okoliczności wynikające z pragmatyki. Zdolność analizowania dyskursu umożliwia systemowi

dialogowemu zrozumienie wypowiedzi w kontekście poprzednich interakcji. Aby komunikacja była efektywna, system musi umieć poradzić sobie z takimi zjawiskami, jak odniesienia anaforyczne (anafora to zaimek wskazujący zapobiegający powtarzaniu podmiotu z poprzedniego zdania), umożliwiającymi użytkownikowi odnoszenie się do przedmiotu rozmowy. Efektywny system dialogowy powinien także umieć radzić sobie z elipsami (elipsa to opuszczenie w zdaniu wyrazu lub wyrazów, domyślnych w szerszym kontekście) i fragmentami zdań, aby użytkownik nie musiał formułować każdego zapytania w pełnym brzmieniu. Możliwość dziedziczenia informacji z poprzednich wypowiedzi jest szczególnie pomocna w obliczu błędów rozpoznawania. Użytkownik może zadać złożone, wymagające kilku atrybutów (wyróżników) pytanie - urządzenie rozpoznawania może nie zrozumieć pojedynczego słowa, np. numeru lotu lub czasu przylotu. Jeśli istnieje dobry model kontekstowy, użytkownik może wypowiedzieć potem krótką frazę korekcyjną, a system będzie potrafił zamienić tylko źle zrozumiane słowo, zapobiegając konieczności powtarzania całej wypowiedzi i zmniejszając ryzyko kolejnych błędów rozpoznawania.



Rys. 6. Przetwarzanie języka naturalnego jako ciąg operacji

Praktyczne realizacje procesu przetwarzania języka naturalnego są zwykle uproszczeniami problemu: nie każdy system NLP zawiera (lub potrzebuje) wszystkie wymienione wyżej składniki. Istnieją systemy, które:

- rezygnują z parsera i określają znaczenie bez informacji syntaktycznych,
- łączą przetwarzanie syntaktyczne i semantyczne w jeden proces,
- nie wymagają wykorzystywania kontekstu,
- eliminują generator odpowiedzi w aplikacjach o kilku możliwych wyjściach,
- rezygnują w całości z tej struktury i przechodzą od rozpoznanych słów do znaczenia operacyjnego (system ekspertowy), wyznaczając znaczenie bez szczegółowej analizy językowej na jakimkolwiek poziomie.

Postęp w badaniach systemów NLP będzie chyba polegał na uczeniu i ewaluacji (podobnie jak w przetwarzaniu sygnału mowy) - jest to trudne ze względu na liczbę składników i ich różnorodne charakterystyki we/wy. Osiągnięcia ostatnich lat polegają na:

- badaniach wykorzystujących odpowiednio przygotowane zasoby językowe - a nie przykłady i intuicję;
- próbach pomiaru pokrycia i efektywności systemów NLP;
- próbach zastosowania wiedzy analitycznej i statystycznej.

Największą barierą w zastosowaniach procesów NLP jest ich mała podatność na zastosowanie w nowych dziedzinach (możliwość konfigurowania systemu NLP dla nowej, określonej aplikacji).

3.3. Własności menadżera dialogu (sterownika dialogu, jądra systemu)

Zadaniem menedżera dialogu jest zapewnienie współpracy systemu dialogowego (cooperative agent) z użytkownikiem poprzez maksymalne upodobnienie dialogu między systemem i użytkownikiem do dialogu między ludźmi. Sterowanie dialogiem polega na:

- interpretacji znaczenia operacyjnego wypowiedzi w oparciu o model dialogu (interakcji) i w kontekście dotychczasowych wypowiedzi;
- decydowaniu o dalszej akcji: żądać kolejnych danych, odszukać informację, zainicjować na nowo błędnie przebiegający dialog;
- generowaniu fraz języka naturalnego (budowa generatora nie jest tak złożona, jak pozostałych składników systemu dialogowego).

Projektując sterowanie dialogiem, przyjmuje się minimalne wymaganie: system współpracuje z użytkownikiem. Interakcja powinna być wygodna, wyczerpująca i zrozumiała. Zorientowane zadaniowo systemy dialogowe w wypełnianiu swej roli są porównywane z człowiekiem. Dąży się do rozszerzenia interakcji w kierunku:

- przejmowania inicjatywy przez użytkownika,
- używania zwrotów anaforycznych,
- używania wyrażeń eliptycznych,
- przejmowania odpowiedzialności za przeprowadzenie użytkownika poprzez zadanie,
- radzenia sobie z problemami pojawiającymi się w dialogu.

Zwykle przy projektowaniu menedżera dialogu wykorzystywane jest doświadczenie uzyskane w dialogu między ludźmi w tej samej lub podobnej dziedzinie. Obserwacje zachowania rozmówców w słownym dialogu między ludźmi nie są wystarczającą bazą do projektowania menedżera dialogu – trzeba wziąć pod uwagę fakt, że ludzie zachowują się odmiennie, gdy interakcja dotyczy komputera, a nie człowieka.

Najczęstszą aplikacją systemu dialogowego jest dostarczanie użytkownikom przez telefon informacji o konkretnych usługach. W typowych informacyjnych dialogach usługowych (information service dialogues) wyróżnia się następujące fazy:

1. Otwarcie dialogu,
2. Sformułowanie życzenia,
3. Sformułowanie odpowiedzi,
4. Zakończenie dialogu.

Otwarcie i zamknięcie nie zależą od dziedziny zastosowania i są podobne dla większości dialogów języka mówionego.

Do rozpoczęcia dialogu między ludźmi, przed sformułowaniem życzenia, rozmówcy zwykle stosują wyrazy uprzejmości (*Dzień dobry*, *Witam*, *Czy mogłaby mi pani pomóc*?) lub oznaki wahania (chrząknięcia, *mhm*). Jako zakończenie dialogu stosowana jest wymiana podziękowań (*Dziękuję Panu*, *Dziękuję bardzo*, *Dziękuję*), a następnie wymiana pozdrowień (*Do widzenia*), która kończy dialog.

W dialogu człowiek – komputer otwarcie jest podobne, pojawienie się wyrazów uprzejmości zależy od „uprzejmości" systemu. Zamknięcie może być prostsze: rozmówca odkłada słuchawkę telefonu.

Sformułowanie życzenia i sformułowanie odpowiedzi są zależne od zadania, czyli zdeterminowane przez strukturę tego zadania (identyfikacja życzenia rozmówcy, uzyskanie odpowiedniej informacji przez przeszukanie bazy danych i wydanie żądanej informacji). Realizacja zadania może wymagać kilku kroków pośrednich:

- potwierdzenia, aby uniknąć pomyłki,
- naprawy, gdy doszło do pomyłki,
- doprecyzowania szczegółów itp.

Są to zjawiska w zasadzie wspólne dla wszystkich dialogów.

Jest wiele sposobów implementacji zarządzania dialogiem. Wiele systemów do opisu przebiegu dialogu wykorzystuje języki skryptowe jako ogólny mechanizm. Inne przedstawiają dialog jako graf obiektów lub modułów dialogowych. Kolejnym aspektem implementacji systemów dialogowych jest

zmiana aktywnego słownika lub możliwości rozumienia dialogu w zależności od jego stanu. Niektóre systemy są zbudowane tak, aby umożliwić użytkownikowi zadawanie dowolnych pytań w dowolnym miejscu dialogu, czyli cały słownik jest aktywny przez cały czas. Inne systemy ograniczają słownik i/lub język, który jest akceptowany w określonych miejscach dialogu. Trudność polega na pogodzeniu potrzeby rosnącej swobody użytkownika (elastyczności w reakcji na zapytanie lub odpowiedź systemu) i rosnącej dokładności rozumienia systemu (drogą ograniczeń na dopuszczalne wejście użytkownika).

3.4. Własności syntezatora sygnału mowy

Generatorem mowy syntetycznej (syntezatorem mowy) nazywa się urządzenie (obecnie komputerowe) do zamiany tekstu w postaci symbolicznej na mowę (text to speech, TTS). Tekst może być wprowadzony z klawiatury, wczytany z pliku w postaci sformatowanej, odczytany za pomocą systemu rozpoznawania pisma (OCR), bądź też utworzony w procesie planowania i generowania odpowiedzi przez sterownik dialogu. Urządzenie powinno umożliwiać automatyczne wytwarzanie zdań zbudowanych z dowolnych słów określonego języka. Najczęściej syntezę sygnału mowy uzyskuje się drogą modelowania dynamiki traktu głosowego podczas artykulacji wypowiedzi (synteza artykulacyjna) lub modelowania bezpośrednio samego sygnału mowy (generowanie sygnału o charakterystykach akustycznych takich samych jak sygnału mowy).

Syntezatory artykulacyjne bazują na reprezentacji traktu głosowego. Początkowo syntetyzatory tego typu wykorzystywały szereg dynamicznie sterowanych filtrów analogowych (Rosen 1958, Dennis 1962), nowoczesne systemy są modelowane na komputerach cyfrowych (Ladefoged 1978, Scully i Clark 1986). Informacją wejściową dla takich systemów są wartości wielu parametrów reprezentujących położenie (pozycję) poszczególnych części traktu głosowego (artykulatorów). Parametry te określają kształt traktu głosowego i są wyznaczane dla jednakowych odcinków, zwykle o długości 0,5 cm, a cały trakt jest modelowany jako ciąg cylindrów (rur prostych). Aby dokonać syntezy sygnału mowy ta złożona rura jest pobudzana przez impulsy quasiokresowe o kształcie określonym przez Rosenberga (1970) lub Fanta (1985).

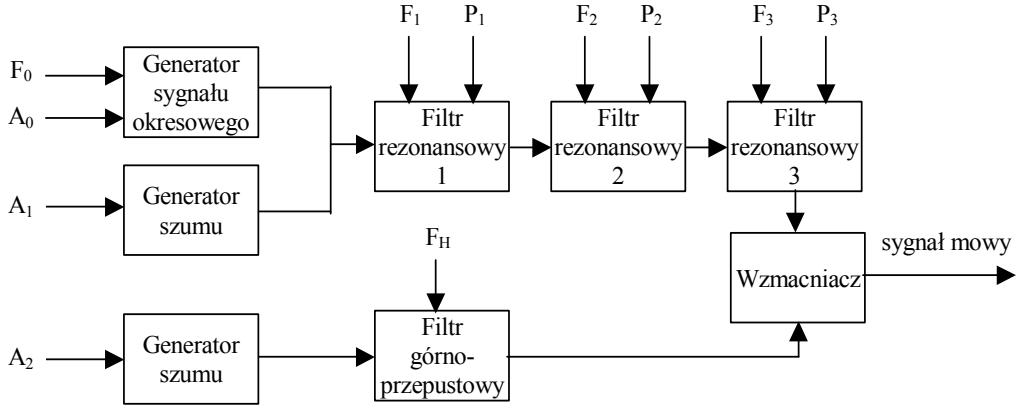
Sygnał emitowany przez usta można wyznaczyć jako rozwiązanie równania ciśnienia fali dźwiękowej wzdłuż traktu głosowego (równania Webstera). W celu wygenerowania ciągu fonemów należy zapewnić zmianę w czasie wartości parametrów artykulacyjnych. Wadą metody jest jej złożoność i w konsekwencji duża ilość obliczeń.

Najczęściej synteza artykulacyjna występuje w dwu postaciach: jako synteza formantowa i synteza z predykcją liniową.

Synteza formantowa wykorzystuje model pobudzenie – filtr. Trakt głosowy człowieka modelowany jest za pomocą zestawu filtrów rezonansowych, które kształtują jego przybliżoną częstotliwościową charakterystykę amplitudową. Częstotliwości rezonansowe tych filtrów są równe częstotliwościom formantów, które charakteryzują kolejne fragmenty sygnału mowy syntezowanej wypowiedzi. Do wygenerowania zrozumiałej mowy wystarczy znajomość trajektorii pierwszych trzech formantów, do wygenerowania wysokiej jakości sygnału mowy: trajektorie czterech lub pięciu formantów.

Wyróżnia się dwie metody łączenia filtrów rezonansowych:

- w syntezatorze równoległym: sygnał pobudzenia podawany jest na wszystkie rezonatory równolegle; wyjścia, każdy z odpowiednim wzmocnieniem, są sumowane,
- w syntezatorze kaskadowym rezonatory łączone są szeregowo (rys. 7).



A_1,A_2, A_0 – skalowanie amplitudy
F_0 – częstotliwość tonu podstawowego
F_H – częstotliwość odcięcia filtru górnoprzepustowego
F1, F2, F3 – częstotliwości formantowe
P1, P2, P3 – szerokość pasma filtrów formantowych

Rys. 7. Przykład syntezatora kaskadowego

Synteza z predykcją liniową również wykorzystuje model pobudzenie -

filtr. Sygnałem pobudzenia jest sygnał szczątkowy predykcji liniowej (błąd predykcji), zaś filtrem - model traktu głosowego, będący układem dynamicznym o transmitancji, której bieguny są wyznaczane za pomocą współczynników predykcji liniowej.

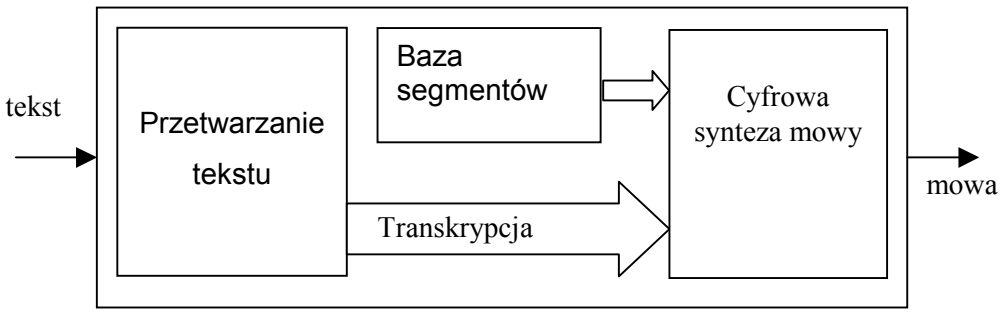
Syntezatory modelujące sygnał mowy wykorzystują konkatenację segmentów sygnału mowy odpowiadających wybranym:

- jednorodnym jednostkom fonetycznym, najczęściej difonom (stosowane ze względu na dokładność odtwarzania transjentów, które decydują o zrozumiałości sygnału mowy),
- zróżnicowanym jednostkom fonetycznym: fonemom, difonom i sylabom.

Przykładem syntezy konkatenacyjnej jest syntezator zbudowany przez France Telecom, wykorzystujący algorytm PSOLA (ang. *The Pitch Synchronous OverLap and Add*) i umożliwiający:

- płynne łączenie segmentów, odpowiadających jednostkom fonetycznym,
- zmianę wysokości dźwięku,
- zmianę długości (czasu trwania) poszczególnych segmentów.

Schemat generatora mowy syntetycznej przedstawia rys. 8. Urządzenie to, wykorzystując stworzoną wcześniej (w procesie analizy, na podstawie pozyskanego od lektora materiału dźwiękowego) bazę segmentów, dokonuje syntezy sygnału mowy.



Rys. 8. Generator mowy syntetycznej

Na proces syntezy składają się następujące czynności:

1. Wybór segmentów odpowiadających transkrypcji fonetycznej generowanego tekstu.

2. Ustalenie częstotliwości tonu podstawowego oraz czasu trwania generowanego fragmentu sygnału mowy (cechy prozodyczne).
3. Synteza fragmentów bezdźwięcznych poprzez skopiowanie danych z bazy segmentów; ewentualne ich powielenie, bądź skrócenie długości.
4. Synteza fragmentów dźwięcznych, w której uwzględniając okres częstotliwości tonu podstawowego należy:
    a. nałożyć na siebie (z właściwym rozłożeniem na osi czasu) segmenty dźwięczne z bazy,
    b. zsumować nałożone segmenty.

Doświadczenie pokazało, że synteza PSOLA zapewnia wyższą jakość generowanej mowy niż synteza z predykcją liniową.

Modyfikacją syntezy konkatenacyjnej jest synteza korpusowa (zasobowa), w której łączy się segmenty sygnału mowy o długości dobieranej każdorazowo dla przekształcanego tekstu. Kryterium doboru jest jakość generowanego sygnału (definiuje się wskaźniki jakości). Z zasobu mowy wybierane są różnorodne jednostki fonetyczne: difony, trifony, sylaby, wyrazy, frazy (grupy wyrazowe) czy nawet całe zdania. Jednostki fonetyczne występują w zasobie wielokrotnie w różnych kontekstach. Generowany sygnał mowy jest konkatenacją różnych jednostek fonetycznych. Istnieje wiele różnych możliwości złożenia pożądanego sygnału mowy. Dobór jednostek fonetycznych oceniany jest za pomocą funkcji kosztu (estymacji), uwzględniającej zarówno czas trwania poszczególnych fragmentów jak i cechy prozodyczne mowy. Proces obliczeniowy jest dość złożony.

Obecnie syntezą korpusową zajmuje się wiele firm (np.: AT&T, SpeechWorks, ScanSoft). Przygotowany dla języka angielskiego zasób mowy ma rozmiar ok. 200 MB. W Polsce syntezą korpusową zajmuje się firma IVO Software z Gdyni. Wydaje się, że właśnie ta technika ma szanse rozwinąć się w przyszłości. Obecnie są prowadzone badania nad udoskonaleniem zasobu mowy (aby pokrył wszystkie zjawiska fonetyczne w danym języku) i funkcji estymacji. Synteza korpusowa jest obecnie wykorzystywana w systemach dialogowych portali głosowych.

4. Wyniki dotychczasowych doświadczeń

Historia systemów dialogowych języka mówionego zaczęła się w końcu lat osiemdziesiątych. Wówczas rozpoczęły się, wspomagane przez dotacje rządowe, programy:

- Spoken Language System (SLS) Program realizowany przez Spoken

Language Systems Group (MIT Laboratory for Computer Science, Cambridge) w USA, wspierany przez Defense Advanced Research Projects Agency (DARPA, potem ARPA);

- Esprit SUNDIAL (speech understanding and dialog) w Europie.

Obydwa programy dotyczyły dostępu do bazy danych przy planowaniu podróży: lotniczych i kolejowych w systemie europejskim i tylko lotniczych w amerykańskim. Projekt europejski był wielojęzyczny: angielski, francuski, niemiecki i włoski. Wszystkie miały słownik ograniczony do kilku tysięcy słów. Obecnie tego typu systemy pracują w czasie rzeczywistym na standardowej stacji roboczej i komputerach typu PC bez dodatkowego osprzętu.

Program SLS był rozwijany przez wiele zespołów w dziedzinie informacji o podróżach lotniczych (Air Travel Information System, ATIS) – pozwalał uzyskiwać informacje o liniach lotniczych, rozkładach, transporcie naziemnym, zawarte w statycznej relacyjnej bazie danych. Wymaganie, aby wszystkie zespoły wykorzystywały tę sama bazę danych (zasób uczący zawiera 14000 spontanicznych wypowiedzi), umożliwiło porównywanie wyników ich prac w regularnych odstępach czasu i zapewniało stały rozwój wszystkich systemów. Na początku w 1989 r. akceptowanym wskaźnikiem była dokładność dla rozpoznawania mowy, już w trakcie dalszych prac opracowano wskaźnik rozumienia mowy zarówno dla wejścia głosowego, jak i pisanego. Do dzisiaj brakuje syntetycznego wskaźnika, który łączyłby ocenę zdolności systemu do efektywnego komunikowania się z użytkownikiem oraz zdolności rozumienia działań użytkownika. W momencie zakończenia programu (1995) najlepszy system rozpoznawał słowa z błędem 2,3%, zdania z błędem 15,2%. Dodatkowo błędy rozumienia były na poziomie 5,9%dla wejścia tekstowego i 8,9% dla wejścia mówionego.

Program SUNDIAL nie był regularnie oceniany, w przeciwieństwie do SLS jednak, jego celem było zbudowanie systemów, które mogły być publicznie zastosowane. Wynikiem prac, zakończonych w 1993 r., były opracowane mechanizmy sterowania dialogiem.

Potem podejmowane były różne sponsorowane programy w zakresie systemów dialogowych języka mówionego:

- ARISE (Automatic Railway Information Systems for Europe) – różniące się systemy informacji o rozkładach kolejowych, rozwijane w językach: holenderskim, francuskim i włoskim;
- Communicator, sponsorowany przez DARPA, w którym twórcy kładli nacisk na interakcje bazujące na dialogu wykorzystujące język pisany i mówiony.

Oprócz badań sponsorowanych w ramach wielkich programów rozwijane

były niezależne inicjatywy, na przykład: Berkeley Restaurant Project (informacja o restauracjach w Berkeley w Kalifornii), AutoRes (rozwijany przez AT&T, telefoniczny system wypożyczania samochodów), „How may I help you?" (system informacji i usług łączenia rozmów telefonicznych), WAXHOLM (system informacji o rozkładach promów oraz informacji turystycznej na wyspach wokół Sztokholmu), TRAINS (rozkład jazdy pociągów, University of Rochester).

Jednym z najważniejszych trendów w systemach dialogowych języka mówionego jest rosnąca liczba publicznie dostępnych realizacji. Takie systemy to nie tylko prototypy badawcze, lecz również produkty komercyjne wykorzystywane nie tylko w takich dziedzinach jak: centra informacji telefonicznych, ceny akcji giełdowych, rozkłady jazdy pociągów, rezerwacje miejsc w samolotach.

5. Uwagi końcowe

Coraz więcej centrali telefonicznych czy centrów kontaktowych dużych i średnich firm (Call Center, Contact Center) zastępuje operatorów portalami głosowymi (Voice Portal). Zadaniem portali głosowych jest umożliwienie interakcji głosowej z użytkownikiem. Portale głosowe są wyposażone w mechanizmy interakcji, których podstawą jest rozpoznawanie i rozumienie mowy oraz konwersja pobranej z bazy danych informacji tekstowej do postaci dźwiękowej.

Portal głosowy jest nie tylko systemem do prowadzenia konwersacji z komputerem, lecz przede wszystkim stanowi bazę danych z informacjami dla potencjalnych klientów serwisu. Informacje te przechowywane są w postaci tekstowej na serwerach baz danych, skąd pobierane są przez skrypty, zlokalizowane na serwerach WWW, obsługujące zapytania, np. SQL. Wyselekcjonowane wiadomości konwertowane są do postaci dźwiękowej przez przeglądarkę głosową za pomocą syntezatora TTS.

Technologia IVP (Internet Voice Portal), mimo że jest jeszcze bardzo młoda, przeżywa swój rozkwit. Pojawiło się szereg bogatych serwisów informacyjnych zarówno udostępniających własne zasoby, jak i korzystających z zasobów Internetu. Część z nich umożliwia także realizację podstawowej usługi internetowej, czyli dostępu do poczty elektronicznej. Portale te są powszechnie dostępne na terenie całych Stanów Zjednoczonych, a korzystanie z nich jest bezpłatne.

Popularny staje się stale rozwijany język (standard) VoiceXML umożliwiający realizację systemów dialogowych języka mówionego.

Literatura

[1] Barnard E., Halberstadt A., Kotelly C., Phillips M.: *A Consistent Approach to Designing Spoken-Dialog Systems*, Proc. ASRU Workshop, Keystone, CO, 1999.

[2] Beutnagel M., Conkie A., Schroeter J., Stylianou Y., Syrdal A.: *The AT&T Next-Gen TTS System*, Proc. ASA, Berlin, 1999.

[3] Billi R., Canavesio R., Rullent C.: *Automation of Telecom Italia Directory Assistance Service: Field Trial Results*, Proc. IVTTA, 1998.

[4] Bobrow R., Ingria R., Stallard D.: *Syntactic and Semantic Knowledge in the DELPHI Uniffication Grammar*, Proc. DARPA Speech and Natural Language Workshop, 1990.

[5] Boves L., Os E.: *Applications of Speech Technology: Designing for Usability*, Proc. IEEE Worshop on ASR and Understanding, 1999.

[6] Cohen P., Johnson M., McGee D., Oviatt S., Clow J., Smith I.: *The Effeciency of Multimodal Interaction: A Case Study*, Proc. ICSLP, 1998.

[7] Cole, R. A., Mariani, J., Uszkoreit, H., Zaenen, A. and Zue, V. W. (Editorial Board), Varile, G. and Zampolli, A. (Managing Editors): *Survey of the State of the Art in Human Language Technology*, 1996. URL:http://www.cse.ogi.edu/CSLU/HLTsurvey/.

[8] Dal D.: *Practical Spoken Dialog Systems*, 2005.

[9] Dowding J., Gawron J., Appelt D., Bear J., Cherny L., Moore R., Moran D., Gemini: *A Natural Language System for Spoken Language Understanding*, Proc. ARPA Workshop on Human Language Technology, 1993.

[10] Flammia G.: *Discourse Segmentation of Spoken Dialogue: An Empirical Approach*, Ph.D. Thesis, MIT, 1998.

[11] Fant G., Liljencrants J., Lin Q.: *A Four-parameter Model of Glottal Flow*, STL-QPSR, 4, 1985.

[12] Fant G.: T*he LF-model Revisited. Transform and Frequency Domain Analysis*, STL-QPSR, 2-3, 1995.

[13] Glass J., Flammia G., Goodine D., Phillips M., Polifroni J., Sakai S., Seneff S., Zue V.: *Multilingual Spoken-Language Understanding in the MIT Voyager System*, Speech Communication, 17, 1995.

[14] Goddeau D.: *Using Probabilistic Shift-Reduce Parsing in Speech Recognition Systems*, Proc. ICSLP, 1992.

[15] Gorin A., Riccardi G., Wright J.: *How may I help you?,* Speech Communication, 23,1997.

[16] Hetherington L., Zue V.: *New words: Implications for Continuous Speech Recognition*, Proc. Eurospeech, 1991.

[17] Lippmann R.P.: *Speech Perception by Humans and Machines*, Speech Communication, 22(1), 1997.

[18] McDonald D. Bolc L. (Eds.): *Natural Language Generation Systems (Symbolic Computation Artificial Intelligence)*, Springer Verlag, Berlin, 1998.

[19] Miller S., Schwartz R., Bobrow R., Ingria R.: *Statistical Language Processing Using Hidden Understanding Models*, Proc. ARPA Speech and Natural Language Workshop, 1994.

[20] Moore R., Appelt D., Dowding J., Gawron J., Moran D.: *Combining Linguistic and Statistical Knowledge Sources in Natural-Language Processing for ATIS*, Proc. ARPA Spoken Language Systems Workshop, 1995.

[21] Nuance Communications, http://www.nuance.com

[22] Oh A.: *Stochastic Natural Language Generation for Spoken Dialog Systems*, M.S. Thesis, CMU, May 2000.

[23] Os E., Boves L., Lamel L., Baggia P.: *Overview of the ARISE project*, Proc. Eurospeech, 1999.

[24] Pao C., Schmid P., Glass J.: *Con_dence Scoring for Speech Understanding Systems*, Proc. ICSLP, 1998.

[25] Peckham J.: *A New Generation of Spoken Dialogue Systems: Results and Lessons from the SUNDIAL Project*, Proc. Eurospeech, 1993.

[26] Price P.: *Evaluation of Spoken Language Systems: the Atis Domain*, Proc. DARPA Speech and Natural Language Workshop, 1990.

[27] Rabiner L., Juang B-H.: *Fundamentals of speech recognition*, 1993.

[28] Reiter E., Dale R.: *Building Natural Language Generation Systems*, Cambridge University Press, Cambridge, 2000.

[29] Rosenberg A. E.: *Effect of Glottal Pulse Shape on the Quality of Natural Vowels*, Journal of The Acoustical Society of America vol. 49, 1970.

[30] Rosset S., Bennacef S., Lamel L.: *Design Strategies for Spoken Language Dialog Systems*, Proc. Eurospeech,1999.

[31] S. Seneff, Tina: *A natural language system for spoken language applications*, Computational Linguistics, 18(1), 1992.

[32] Seneff S., Goddeau D., Pao C., Polifroni J.: *Multimodal discourse modelling in a multi-user multi-domain environment*, Proc. ICSLP, 1996.

[33] Seneff S., Lau R., J. Polifroni: *Organization, Communication, and Control in the Galaxy-II Conversational System*, Proc. Eurospeech, 1999.

[34] Seneff S.: *Robust Parsing for Spoken Language Systems*, Proc. ICASSP, 1992.

[35] Souvignier V., Kellner A., Rueber B., Schramm H., Seide F.: *The Thoughtful Elephant: Strategies for Spoken Dialogue Systems*, IEEE Trans. SAP, 8(1), 2000.

[36] Stallard D., Bobrow R.: *Fragment Processing in the DELPHI System*, Proc. DARPA Speech and Natural Language Workshop, 1992.

[37] Sutton S., et al.: *Universal Speech Tools: The CSLU Toolkit*, Proc. ICSLP, 1998.

[38] Tatham M., Morton K.: *Developments in Speech Synthesis*, 2005.

[39] Van Kuppevelt J.C., Smith R.W.: *Current and New Directions in Discourse and Dialogue*, 2005.

[40] Ward W.: *The CMU Air Travel Information Service: Understanding Spontaneous Speech*, Proc. ARPA Workshop on Speech and Natural Language, 1990.

[41] Yi J., Glass J.: *Natural-Sounding Speech Synthesis Using Variable Length Units*, Proc. ICSLP, 1998.

[42] Young S., Bloothooft G.: *Corpus-based methods in Language and speech processing*, 1997.

[43] Zue V., Seneff S., Glass J., Polifroni J., Pao C., Hazen T., Hetherington L., Jupiter: *A Telephone-Based Conversational Interface forWeather Information*, IEEE Trans. SAP, 8(1), 2000.

[44] Zue V., Seneff S., Polifroni J., Phillips M., Pao C., Goddeau D., Glass J., Brill E., Pegasus: *A Spoken Language Interface for On-Line Air Travel Planning,* Speech Communication, 15, 1994.

[45] Zue Victor W., Glass James R.: *Conversational Interfaces: Advances and Challenges, Proceedings of the IEEE*, vol. 88, no. 8, 2000.

Spoken language dialogue system

ABSTRACT: In this paper, the structure of a spoken language dialogue system was described. The underlying human language technologies were described: automatic speech recognizer, natural language understanding, dialogue manager, and speech synthesizer. The recent progress in spoken dialogue systems and some of the ongoing research challenges were presented.

KEYWORDS: dialogue system, speech recognition, speech understanding, speech synthesis

Recenzent: prof. dr hab. inż. Włodzimierz Kwiatkowski

Praca wpłynęła do redakcji: 28.12.2007 r.