Scripta Neophilologica Posnaniensia, Tom II, strony: 49 – 64
Wydział Neofilologii, UAM Poznań, 2000

PROZODIA W SYSTEMACH DIALOGOWYCH XXI WIEKU

GRAŻYNA DEMENKO

I. Podstawowe problemy automatycznej analizy mowy

Od początku rozwoju techniki komputerowej człowiek starał się stworzyć jak najefektywniejsze narzędzia ułatwiające komunikację z maszyną; od pierwszych nieskomplikowanych klawiatur, drukarek i monitorów, aż do coraz bardziej złożonych – dialogowo-kontekstowych, nawigowanych przez ekrany dotykowe systemów cyfrowych. Szczytem osiągnięć konstrukcyjnych był oczywiście automat porozumiewający się ze swoim twórcą w jego naturalnym języku; maszyna, która mówi, słyszy i reaguje na polecenia. Zadziwiał inteligencją w filmach „science fiction" system komputera HAL-9000 z „Odysei kosmicznej", prezentował doskonałą, syntetyczną angielszczyznę C3PO z „Gwiezdnych wojen", podczas gdy rzeczywiste systemy rozpoznawania i syntezy mowy nie pozwalały na uzyskanie efektów nawet zbliżonych do filmowych. Okazało się, że najbardziej naturalny dla człowieka sposób komunikacji jest jednym z najtrudniejszych zadań techniki. Od ponad pięćdziesięciu lat w wielu ośrodkach naukowych całego świata (przy bardzo znacznych nakładach finansowych) prowadzone są niestrudzenie badania nad językiem naturalnym, opracowywane są algorytmy analizy dźwięków mowy naturalnej, rozpoznawania a także syntezy. Problematyka łączności słownej w układach technicznych obejmuje szeroki zakres zagadnień z różnych dziedzin wiedzy związanych z analizą fonetyczno-akustycznych cech mowy, jej rozpoznawaniem, syntezą oraz transmisją. Automatyczne przetwarzanie dźwiękowej postaci języka stanowi przedmiot badań wielu dyscyplin naukowych, takich jak: technologia mowy, fonetyka, lingwistyka komputerowa, psycholingwistyka, informatyka, telekomunikacja, foniatria i audiologia.

W okresie obecnego szybkiego rozwoju techniki cyfrowej oraz postępu prac w zakresie analizy i przetwarzania języka naturalnego istnieje realna szansa, że rozpoznawanie i synteza mowy będą podstawowymi środkami komunikacji w komputerowych systemach dialogowych. Mowa jest bardzo wydajnym rodzajem komunikacji – szybkość mówienia jest

kilkakrotnie wyższa niż z pisania. Oczywiście, ludzie znacznie lepiej funkcjonują w komunikacji słownej niż maszyny, są bardziej elastyczni, pewni i skuteczni. Z drugiej jednak strony, dialogowe systemy komputerowe mogą się bardzo dobrze sprawdzać w praktyce (np. automaty w 24 godzinnej służbie informacyjnej).

W Tabeli 1 zilustrowano poprawność rozpoznawania mowy przez człowieka i komputer dla słownika składającego się z ograniczonej liczby jednostek oraz dla spontanicznej konwersacji przeprowadzonej na materiale językowym składającym się z 2000 słów.

Materiał	Wielkość słownika	Procent błędów popełniany przez maszynę	Procent błędów popełniany przez człowieka
Czytane cyfry	10	0,7	0,01
Czytane litery	26	5,0	1,6
Czytane zdania	1000	17,0	2,0
Spontaniczna konwersacja	2000	43,0	4,0

Tabela 1. Poprawność rozpoznawania mowy przez człowieka i komputer. Opracowanie własne na podstawie: Flexible, robust, and efficient human speech processing versus present-day speech technology, *Louis Pols (1999), Proceedings of the XIVth International Congress of Phonetic Sciences, s. 9-12.*

Wysoki procent błędów popełniany w rozpoznawaniu mowy spontanicznej przez maszynę (43%) jest nie do zaakceptowania. Z zestawienia wynika, że praktyczne implementacje rozpoznawania mowy spontanicznej są ciągle ograniczone. Mimo wielu badań prowadzonych intensywnie w ciągu ostatnich kilkudziesięciu lat relacje między sygnałem akustycznym i strukturą języka nie zostały w pełni ustalone. Złożoność problematyki zarówno na etapie wytwarzania, percepcji, jak i analizy akustycznej mowy wynika – niezależnie od tego, czy układem rozpoznającym jest mózg człowieka, czy komputer – z jej specyficznych własności. Powstałe w procesie artykulacji zespoły dźwięków są nośnikami różnorodnych informacji językowych, paralingwistycznych oraz pozajęzykowych. Na zróżnicowania akustycznych obrazów sygnału mowy wpływa szereg interaktywnych źródeł zmienności. Do najważniejszych z nich należą: (a) uwarunkowania językowe – struktura segmentalna, suprasegmentalna oraz zjawiska koartykulacyjne, (b) cechy indywidualne mówców – cechy fizyczne (np. płeć i wiek), uwarunkowania psychologiczne (np. stosunek do treści wypowiedzi bądź rozmówcy), styl, dialekt, (c) akustyczne warunki toru przenoszącego informacje – przesłuchy zrozumiałe oraz szumy otoczenia.

Mimo wielu lat badań i setek eksperymentów nie udało się wyjaśnić, w jaki sposób zmienność ta, która przysparza tak wiele trudności w automatycznym rozpoznawaniu mowy, eliminowana jest niezwykle skutecznie w procesie percepcji. „Ludzki" – naturalny sposób rozpoznawania mowy polega na odebraniu i rejestracji sygnału poprzez zmysł słuchu i odfiltrowaniu niepotrzebnych w danej chwili sygnałów. Skomplikowane procesy analizy głosu zachodzące w mózgu człowieka, pozwalają na wybiórcze traktowanie

sygnałów; niepożądane są odrzucane nawet wtedy, gdy ich amplituda jest większa od sygnałów oczekiwanych. Analiza automatyczna sygnału mowy nie pozwala niestety na taki sposób odbioru. We wszelkich komputerowych systemach dialogowych ważne jest, by rozpoznawana mowa była najważniejszym –najgłośniejszym sygnałem otoczenia.

Kluczowym problemem automatycznej analizy mowy jest nieskończona różnorodność akustycznych obrazów sygnału (por. np. Jassem 1973). Analizuje się więc możliwość redukcji zmienności oraz interpretacji i modelowania zmienności sygnału. Przeprowadzenie kompleksowej analizy uwzględniającej oddziaływanie wielu interaktywnych źródeł zmienności wymaga obszernej bazy danych i pracochłonnych eksperymentów. Określenie tych źródeł i opisanie ich funkcjonowania jest zadaniem tak skomplikowanym, że istnieje pogląd sceptyczny, według którego sformułowanie odpowiednich algorytmów rozpoznawania oraz syntezy mowy wyłącznie na bazie teorii jest wątpliwe.

W związku z tym, zauważa się w ostatnich latach rozwiązania typowo techniczne. Powstają układy rozpoznawania oraz syntezy mowy oparte głównie na statystyczno-matematycznych algorytmach (np. sieciach neuronowych oraz procesach Markowa) umożliwiających uczenie systemów bez konieczności uwzględniania złożonych związków między językowymi a akustycznymi cechami sygnału. Dzisiejszy system rozpoznawania mowy posiada globalną statystyczną wiedzę i pracuje średnio dobrze. Jeżeli pojawia się nowa wypowiedź nie zawarta w zbiorze uczącym (np. mówca się przeziębi lub będzie mówił z obcym akcentem) to system popełnia błędy. Tego rodzaju opracowania nie zapewniają sformułowania uniwersalnych, poprawnie funkcjonujących algorytmów, niezależnych od doboru materiału językowego, głosu mówcy oraz akustycznych uwarunkowań otoczenia.

System taki nie uwzględnia wiedzy fonetycznej („nie zauważa", że we frazach pytających najczęściej kontur częstotliwości podstawowej jest wznoszący, że w szybkiej mowie głoski są krótsze a niektóre zredukowane, że nowa informacja jest akcentowana a końcowe sylaby frazy są wydłużane). W modelu difonowym czy trifonowym część tej wiedzy oczywiście jest zawarta w sposób probabilistyczny. Jeśli wziąć pod uwagę fakt, że wyniki rozpoznawania mowy przy zastosowaniu różnych algorytmów często dają podobne rezultaty, to można przypuszczać, że trudności w przetwarzaniu sygnału wynikają jednak nie tyle z nieadekwatności stosowanych metod, co z powodu nie uwzględniania w opisie inwariantów w zakresie poszczególnych typów informacji.

W latach 70. rozwijały się systemy oparte o wiedzę językową. Wyniki rozpoznawania mowy nie były najlepsze. Okazało się, że teoria lingwistyczna jest niedopasowana do bezpośredniego zastosowania, a reguły są niekompletne. W latach 90. zaczęły się więc rozwijać systemy wykorzystujące wiedzę statystyczną. Wykazano jednak, że oparcie algorytmów wyłącznie na cechach akustycznych sygnału mowy powoduje potrzebę tworzenia obszernych banków danych. Rozkłady prawdopodobieństw okazały się niekompletne. Przygotowanie zaś reprezentatywnej bazy danych dla potrzeb automatycznego uczenia systemu rozpoznawania mowy lub syntezy, możliwe w pewnym stopniu dla tekstów czytanych, wydaje się, w przypadku mowy spontanicznej problemem na razie nie do rozwiązania.

W ostatnich latach zaczęły rozwijać się systemy dialogowe przyszłości, wykorzystujące bazy danych, utworzone na podstawie wiedzy językowej. Systemy te mają bardzo duże

szanse na zapewnienie naturalności mowie syntetycznej i poprawności rozpoznawania mowy spontanicznej.

W nowoczesnych komputerowych systemach komunikacji słownej konieczne jest uwzględnienie informacji nie tylko segmentalnej, ale również informacji suprasegmentalnej w bardzo znacznym stopniu wykorzystywanej zarówno przez mówcę, jak i przez słuchacza.

II. Prozodia w technologii mowy

Podstawowe problemy związane z parametryzacją i modelowaniem struktur melodycznych poszczególnych języków nie są zadawalająco dobrze rozwiązane dla implementacji praktycznych. Duża liczba stosowanych technik w zakresie ekstrakcji, opisu cech suprasegmentalnych mowy oraz kwantytatywnych modeli intonacji opartych na różnego rodzaju manualnych transkrypcjach struktur melodycznych, świadczy o tym, że, jak dotychczas, nie jest opracowana odpowiednia metodologia badań w tej dziedzinie (por. np. Demenko 1999). Podobnie jak w przypadku cech segmentalnych mowy, również w przetwarzaniu suprasegmentaliów próbuje się w wielu opracowaniach, wykorzystujących automatyczne uczenie, pominąć metodologiczne problemy związane z niedostateczną wiedzą o interakcji różnych informacji zawartych w sygnale. Sposób „ślepego" uczenia układu nie wymaga ani obszernych eksperymentów ani manualnej transkrypcji złożonych struktur suprasegmentalnych. Tego rodzaju rozwiązanie nie prowadzi jednak do budowy uniwersalnego systemu; dlatego też w ostatnich latach w modelowaniu cech prozodycznych mowy coraz częściej wykorzystuje się analizę językową uwzględniającą informację syntaktyczną, semantyczną oraz pragmatyczną

Cechy suprasegmentalne odgrywają zasadniczą rolę w syntezie mowy. Modelowanie melodycznych struktur zwiększa zrozumiałość i w sposób decydujący wpływa na naturalność wypowiedzi. Trudno obecnie zaakceptować system dialogowy wytwarzający monotonną mowę. Do najczęściej wykorzystywanych w praktyce typów syntezy należą: metoda artykulacyjna, synteza modelująca wytwarzanie sygnału mowy, synteza formantowa, wykorzystująca bezpośrednio akustyczne cechy sygnału oraz konkatenacyjna polegająca na łączeniu krótkich segmentów sygnału w dłuższe jednostki (np. demisylab w sylaby, sylab w wyrazy itp.)

Bez względu na stosowany typ syntezy elementów segmentalnych mowy modelowanie intonacji ważne jest z kilku zasadniczych powodów.

1. Intonacja wpływa na zrozumiałość mowy. Spełnia funkcję segmentacyjną wypowiedzi i ułatwia słuchaczowi wyodrębnianie z ciągłego sygnału mowy przekazywanych przez mówcę poszczególnych informacji.
2. Błędy w budowie segmentalnej są przez słuchacza w większym stopniu tolerowane niż błędy w strukturze suprasegmentalnej wypowiedzi. Niewłaściwe miejsce wystąpienia akcentu, bądź nieprawidłowy typ akcentu może całkowicie zmienić sens wypowiedzi lub wywołać wrażenie nienaturalności. Lepszym w końcu rozwiązaniem w syntezie jest modelowanie monotonnej intonacji niż nieodpowiednie odwzorowywanie cech melodycznych wypowiedzi.

3. Dla uzyskania mowy wysokiej jakości niezbędne jest poprawne kształtowanie cech prozodycznych. Słuchacze z trudem akceptują mowę monotonną, ponieważ wymaga ona od nich dużo większej koncentracji uwagi niż odbiór wypowiedzi naturalnych.

Tradycyjnie najlepiej rozwinięta została synteza z reguł, zwykle stosowana do sterowania zmianami wysokości tonu w układach typu *text to speech*, w których dokonuje się automatycznie konwersji tekstu ortograficznego na odpowiedni sygnał akustyczny. Istnieje co najmniej kilkadziesiąt algorytmów teoretycznych i implementacji praktycznych sterowania intonacją w mowie czytanej opracowanych dla różnych języków. Do najciekawszych rozwiązań należą systemy: INVOVOX – system syntezy *text-to-speech* opracowany dla języków: angielskiego, niemieckiego, francuskiego, hiszpańskiego, szwedzkiego i włoskiego, DECTALK – system przetwarzania znaków ASCII w naturalnie brzmiącą mowę (posiada możliwość wytworzenia 4 typów głosu kobiecego, 4 głosów męskich i 1 dziecięcego), HADIFIX – synteza konkatenacyjna dla języka niemieckiego, MBROLA – system syntezy wysokiej jakości opartej na difonach z przeznaczeniem dla wielu języków (np. angielskiego, hiszpańskiego, włoskiego i holenderskiego).

Problematyka związana z modelowaniem intonacji dla syntezy mowy obejmuje następujące podstawowe zagadnienia.

1. Sterowanie sekwencją tonów (ustalenie kolejności akcentów, typu akcentu oraz synchronizacji czasowej zmian tonu względem własności segmentalnych wypowiedzi).
2. Uwydatnienie intonacyjne.

   Dotyczy ono podkreślania intonacyjnego szczególnie istotnych dla mówcy fragmentów zdania; może być także związane z modelowaniem informacji paralingwistycznych. Zagadnienie uwzględniania w syntezie mowy informacji paralingwistycznych oraz pozajęzykowych stanowi aktualnie na świecie ważny problem (por. np. Sagisaka et al. 1997). Jego rozwiązanie jest niezbędne dla uzyskania syntezy wysokiej jakości.
3. Globalne cechy intonacji.

   Nowoczesne układy syntezy wymagają również opracowania modelowania różnych zakresów zmian częstotliwości podstawowej, rejestrów, oraz normalizacji percepcyjnej i akustycznej konturu intonacyjnego w obrębie frazy.

Problem sterowania częstotliwością podstawową w syntezie mowy polskiej nie jest w sposób zadowalający rozwiązany. Nieliczne opracowania z tej dziedziny obejmują swym zakresem głównie wypowiedzi izolowane i dostarczają tylko fragmentarycznych wskazówek które mogą być implementowane w syntezie. Dla języka polskiego w systemie syntezy *text-to-speech* („Kubuś") zastosowano uproszczony model sterowania intonacją.[1] Założono, że program realizujący kształtowanie konturów intonacyjnych powinien

[1] System syntezy *text-to-speech* („Kubuś") opracowano w Zakładzie Fonetyki Akustycznej Instytutu Podstawowych Problemów Techniki w Poznaniu na początku lat 90.

uwzględniać następujące rodzaje informacji: (a) dane opisujące zdanie: liczbę fraz, strukturę frazy, pozycję frazy, granicę frazy, (b) dane opisujące frazę: liczbę akcentów, pozycję akcentu, c) długość frazy oraz (d) dane opisujące sylabę (por. Demenko et al. 1993).

Na ryc. 1 zilustrowano oscylogram oraz przebieg zmian wysokości tonu w przykładowej, syntetycznej wypowiedzi: *Proszę powtórzyć za mną: w Szczebrzeszynie chrząszcz brzmi w trzcinie, stół z powyłamywanymi nogami i wyindywidualizowaliśmy się z rozentuzjazmowanego tłumu.* Zakres zmian częstotliwości podstawowej wynosi oktawę.

Zastosowany system sterowania intonacją umożliwia modelowanie 4 typów akcentu rdzennego: HL (wysokiego opadającego), ML (niskiego opadającego), LH (wysokiego rosnącego i LM (niskiego rosnącego).

Przyjęto możliwość sterowania zmiennością częstotliwości podstawowej według modelu Fujisaki[2]. Model ten zakłada superpozycję składowej frazowej (określającej deklinację) i składowych akcentowych wyznaczonych dla poszczególnych sylab akcentowanych. Zmiany wysokości tonu aproksymowano sumą funkcji reprezentujących składową frazową oraz składowe akcentowe.

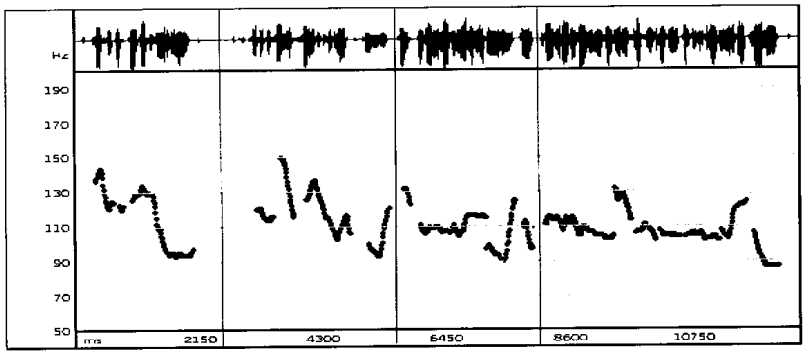
[2] Funkcję G_{pi}, sterującą frazą opisano zależnością:

$$G_{pi}(t) = K_{pi}\alpha_i t \exp(-\alpha_i t)$$

Funkcję G_{aj}, sterującą frazą opisano zależnością:

$$G_{aj}(t) = K_{aj}(1-(1+\beta_j \exp(-\beta_j t)))$$

gdzie: K_{aj}, K_{pi} – oznaczają współczynniki wzmocnienia,
α_i, β_j – współczynniki tłumienia,
i, j – numer kolejnego akcentu,
t – czas.



Ryc. 1. Oscylogram oraz intonogram wypowiedzi syntetycznej: *Proszę powtórzyć za mną: w Szczebrzeszynie chrząszcz brzmi w trzcinie, stół z powyłamywanymi nogami* oraz *wyindywidualizowaliśmy się z rozentuzjazmowanego tłumu.* Granice między frazami oznaczono kursorami

W systemach rozpoznawania mowy prozodia nie jest niezbędna, jednak jej uwzględnienie może zwiększyć efektywność pracy systemu, skrócić czas obliczeń oraz ułatwić korektę błędów.

Systemy rozpoznawania mowy dzieli się na systemy rozpoznające krótkie wypowiedzi (z większego lub mniejszego słownika) i mowę ciągłą (teksty czytane oraz wypowiedzi spontaniczne). W rozpoznawaniu pojedynczych wypowiedzi wykorzystanie suprasegmentaliów koncentruje się głównie na ustaleniu dla danego wyrazu wzorca akcentowego oraz na aspektach pozajęzykowych – np. wykrywaniu patologii w głosie.

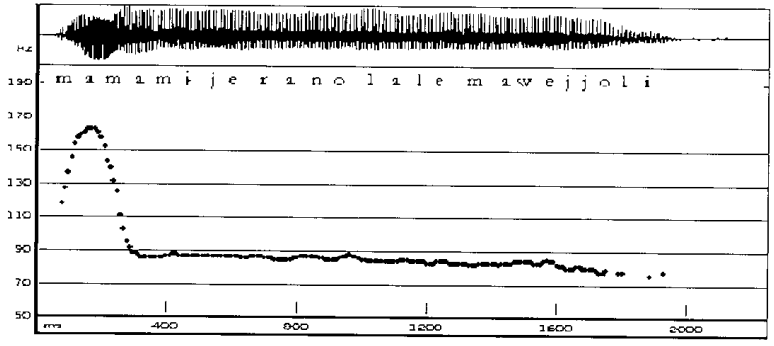
W systemach rozpoznawania mowy ciągłej cechy suprasegmentalne jako źródło informacji stają się bardziej istotne, ale ich ekstrakcja jest trudniejsza i może być obarczona wieloma błędami. Weryfikacja akcentu rdzennego w obrębie frazy i odnalezienie najistotniejszych fragmentów wypowiedzi pozwala na ograniczenie czasu przeszukiwania leksykonu. Paralingwistyczne i pozajęzykowe aspekty suprasegmentaliów odgrywają w tym przypadku drugorzędną rolę, jeśli pominąć zadanie szybkiej adaptacji systemu i konieczność wstępnego opracowania sygnału, (np. mowy z chrypką) lub identyfikację głosu.

Szczególnie ważne jest wykorzystanie cech suprasegmentalnych w rozpoznawaniu mowy dla języków tonalnych i tonicznych. W językach nietonicznych takich jak polski, angielski, niemiecki, francuski udział intonacji w przekazywaniu informacji polega na tym, że sygnalizuje ona pewne stany emocjonalne mówcy, jego stosunek do treści wypowiedzi lub do słuchacza. W językach tonicznych jak np. szwedzki, norweski oraz tonalnych jak np. chiński i wietnamski intonacja spełnia funkcję podwójną. W językach tych identyczne fonematycznie wypowiedzi o różnej dystynktywnie intonacji mogą stanowić odrębne części mowy. Różnice dystynktywne w intonacji tych języków występują na tle takich samych sekwencji fonemów i mają związek ze znaczeniem leksykalnym wyrazów (por. np. Hirst 1998).

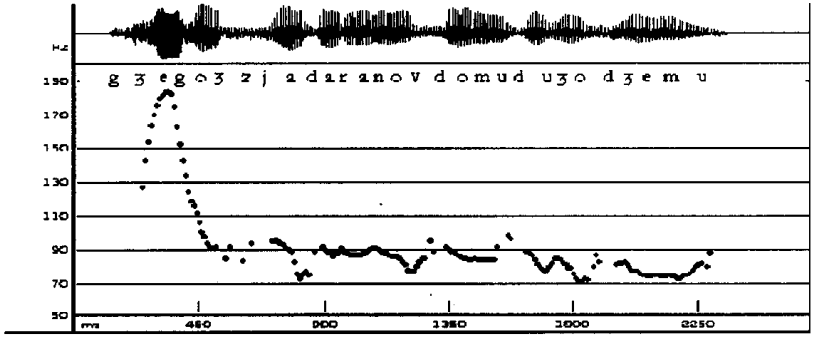
Suprasegmentalne cechy sygnału mogą być uwzględniane na różnych poziomach analizy:

(a) fonetycznym – badanie efektów koartykulacyjnych, specyficznych wartości częstotliwości podstawowej poszczególnych samogłosek,
(b) składniowym – określenie granic frazy, struktury syntaktycznej wypowiedzi,
(c) pozajęzykowym – wykrycie emocji lub patologii w głosie mówcy.

Modelowanie efektów mikroprozodycznych jest szczególnie istotne dla syntezy. Na ryc. 2. przedstawiono przykłady przebiegów częstotliwości podstawowej w wypowiedzi: *mama myje rano lalę małej Joli* oraz w wypowiedzi *Grzegorz zjada rano w domu dużo dżemu* (z akcentem rdzennym na pierwszej sylabie wypowiedzi). W przeprowadzonym teście percepcyjnym słuchacze ocenili przebiegi intonacyjne w obu przykładach jako takie same. Wizualna ocena oraz analiza akustyczna wykazuje względnie gładki kontur intonacyjny w wypowiedzi *mama myje rano lalę małej Joli* (spółgłoski nosowe i płynne) natomiast w wypowiedzi *Grzegorz zjada rano w domu dużo dżemu* zauważalne są duże perturbacje częstotliwości podstawowej występujące głównie na spółgłoskach zwartych i zwartotrących. Mikroprozodia nie wpływa na słuchowy odbiór akcentu, przyczynia się natomiast do wrażenia naturalności sygnału. W większości języków różnice mikroprozodyczne dochodzą do 1,5 półtonu.



Ryc. 2a. Oscylogram oraz intonogram wypowiedzi: *mama myje rano lalę małej Joli.*



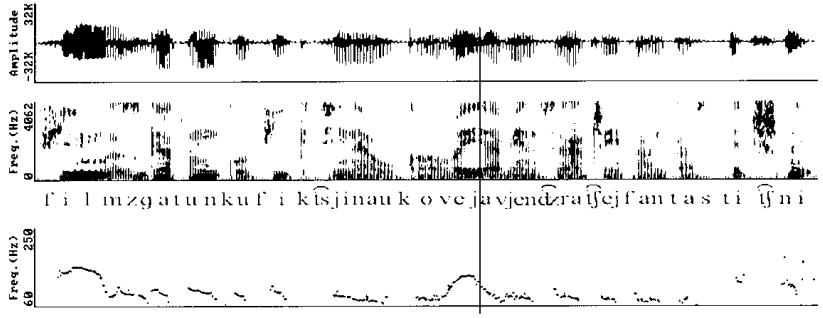
Ryc. 2b. Oscylogram oraz intonogram wypowiedzi: *Grzegorz zjada rano w domu dużo dżemu.*

Dla systemów syntezy mowy z tekstu ***text-to-speech*** analiza struktur składniowych i intonacyjnych języka ma fundamentalne znaczenie. Na podstawie dowolnego tekstu ortograficznego system musi wygenerować sygnał mowy o akceptowalnych cechach prozodycznych. Automatyczna analiza gramatyczna jest konieczna w celu dokonania podziału wypowiedzi na frazy oraz w celu właściwego rozmieszczenia pauz. Nie wszystkie bowiem znaki interpunkcyjne określają granice międzyfrazowe i nie wszystkie granice międzyfrazowe są związane są znakami interpunkcyjnymi.

Jednym z najpopularniejszych formalizmów mających na celu reprezentację informacji składniowej oraz semantycznej dla potrzeb automatycznej analizy mowy jest gramatyka HPSG (Head – driven Phrase Structure Grammar). Prozodyczna informacja określona symbolem PSCB (Prosodic Syntactic Clause Boundary) oraz informacja składniowa i semantyczna wykorzystywana jest w aktualnie na świecie rozwijanych systemach dialogowych (por. np. Hess et al. 1997). Dla języka polskiego jak dotąd brak jest szerszych opracowań w tym zakresie.

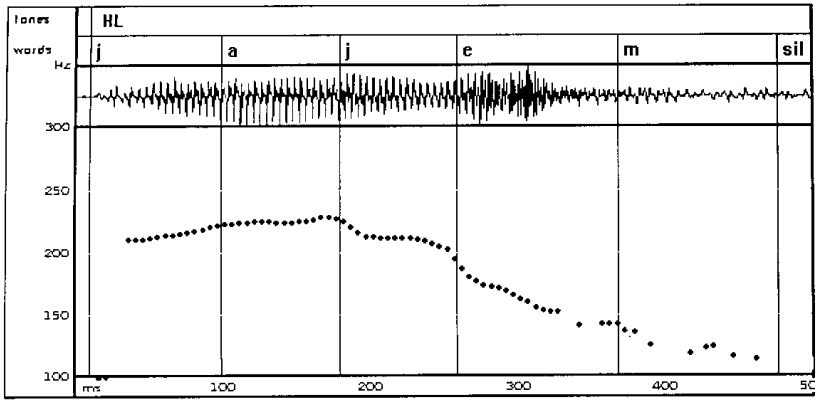
Określenie granic frazy w systemach rozpoznawania mowy odgrywa coraz większą rolę. Rozpoznawanie mowy dokonuje się głównie na płaszczyźnie segmentalnej. Jednakże poprawne rozpoznawanie mowy związane jest również z funkcjonowaniem cech suprasegmentalnych, które pozwalają dokonać podziału tekstu na sekwencje wyrazów stanowiących spójną całość pod względem syntaktycznym lub semantycznym. W tekście pisanym funkcję taką pełnią znaki interpunkcyjne, których obecność pozwala odbiorcy przeprowadzać podział tekstu na jednostki informacji zgodnie z intencją nadawcy. W tekście mówionym wyodrębnianie jednostek zwanych frazami jest osiągane poprzez realizację określonych konturów intonacyjnych, strukturę rytmiczną wypowiedzi (rozkład akcentów, zjawiska iloczasowe) oraz obecność pauz.

Na ryc. 3 przedstawiono przykładowo oscylogram, spektrogram i intonogram wypowiedzi .. *film z gatunku finkcji naukowej, a więc raczej fantastyczny*, w której słuchacze wyznaczyli granicę po sylabie *wej* (w wyrazie *naukowej)*. Między sylabą *wej* i sylabą *a* (w wyrazie *a więc..*) brak jest jakiejkolwiek pauzy. Również w przebiegu intensywności obserwuje się małą zmienność. Na sylabie *ko* (w wyrazie *naukowej*) występuje spadek częstotliwości do F_{min} (90 Hz). Na samogłosce *e* w sylabie *wej* zauważa się wzrost częstotliwości (w zakresie 115 – 170 Hz), na spółgłosce *j* (w tej samej sylabie) występuje spadek częstotliwości (w zakresie 170 – 120 Hz). Na samogłosce *a* (po granicy frazowej) wartość częstotliwości zmienia się w zakresie 120 – 110 Hz. Obserwuje się tutaj typową granicę utworzoną przez przebieg parametru F_0, tzw. wzrost kontynuacyjny („continuation rise"). Spodziewać się więc można, że w przypadku braku pauzy istotną rolę w podziale wypowiedzi na frazy odgrywają również inne cechy akustyczne (np. iloczas).

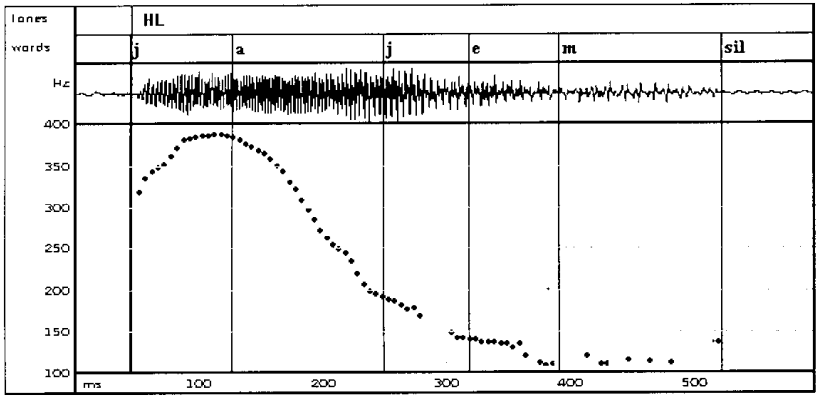


Ryc. 3. Oscylogram, spektrogram oraz intonogram wypowiedzi: *Film z gatunku fikcji naukowej a więc raczej fantastyczny*. Granice frazową oznaczono kursorem.

Szczegółowa analiza parametru F_0 wykazuje możliwość różnicowania znaczenia wyrazów na podstawie zmian wysokości tonu w obrębie samogłosek. Ryc. 4a oraz 4b ilustrują przykładowe przebiegi częstotliwości podstawowej dla pary wyrazów *jajem* oraz *ja jem*. I tak dla wypowiedzi *jajem* (ryc. 4a) spadek parametru F_0 na samogłosce *a* wynosi 170 Hz. Dla wypowiedzi *ja jem* (ryc. 4b) spadek częstotliwości podstawowej występuje na samogłosce *e* (100 Hz).



Ryc. 4a. Oscylogram oraz intonogram wypowiedzi *jajem.*



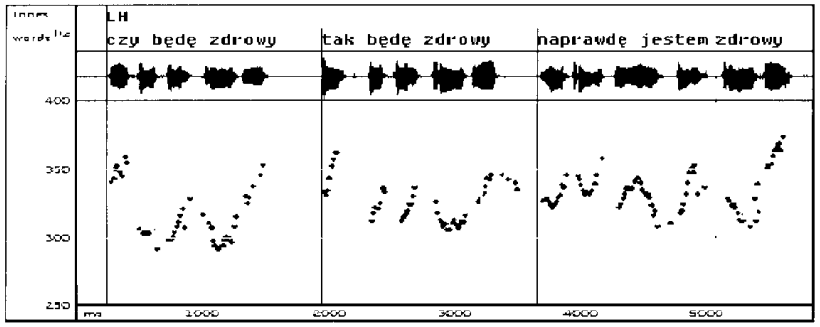
Ryc. 4b. Oscylogram, intonogram wypowiedzi *ja jem.*

Intensywnie wzrasta w ostatnich latach wykorzystanie analiz cech suprasegmentalnych mowy na poziomie pozajęzykowym w audiologii i foniatrii zarówno dla potrzeb diagnozy jak i rehabilitacji. Różnicowanie patologii będącej następstwem zmian przede wszystkim w masie albo napięciu fałdów głosowych, jest zasadniczym kryterium oddzielenia zmian

organicznych związanych najczęściej z przyrostem masy drgającego fałdu od zmian czynnościowych uwarunkowanych głównie zmianami napięcia i koordynacją drgań fałdów głosowych. O ile zmiany organiczne w obrębie głośni powodują zazwyczaj asymetryczny wzrost masy fałdu głosowego to zmiany czynnościowe wpływają na stopień jego napięcia. Czynnościowe zaburzenia głosu prowadzą do dyskoordynacji przede wszystkim fonacji i oddychania.

W audiologii analiza cech suprasegmentalnych ukierunkowana jest na rehabilitację. Interesująca jest analiza zaburzeń mowy u osób niedosłyszących, przeprowadzona po kilku latach utraty słuchu. Zależnie od rodzaju i stopnia uszkodzenia narządu słuchu mowa tych osób może charakteryzować się nietypowymi cechami melodycznymi, zawężonym zakresem zmian częstotliwości podstawowej, zaburzeniami cech segmentalnych wypowiedzi.

Na ryc. 5 zilustrowano przykładowe wypowiedzi pacjentki po 3 letnim okresie utraty słuchu (powyżej 60dB). Wszystkie wypowiedzi zostały zrealizowane z intonacją rosnącą, niezależnie od typu frazy.



Ryc. 5. Oscylogram oraz intonogram wypowiedzi: *Czy będę zdrowy? Tak będę zdrowy. Naprawdę jestem zdrowy.* Granice między frazami oznaczono kursorami.

III. Systemy dialogowe XXI wieku

Dla praktycznych implementacji suprasegmentaliów w systemach dialogowych konieczne jest rozwiązanie podstawowych problemów metodologicznych oraz technicznych w zakresie:

b) wiarygodnej ekstrakcji parametrów suprasegmentalnych – głównie częstotliwości podstawowej,

c) kwantytatywnego opisu cech suprasegmentalnych oraz modelowania intonacji,

d) automatycznej transkrypcji struktur melodycznych,

e) integracji cech suprasegmentalnych z cechami segmentalnymi.

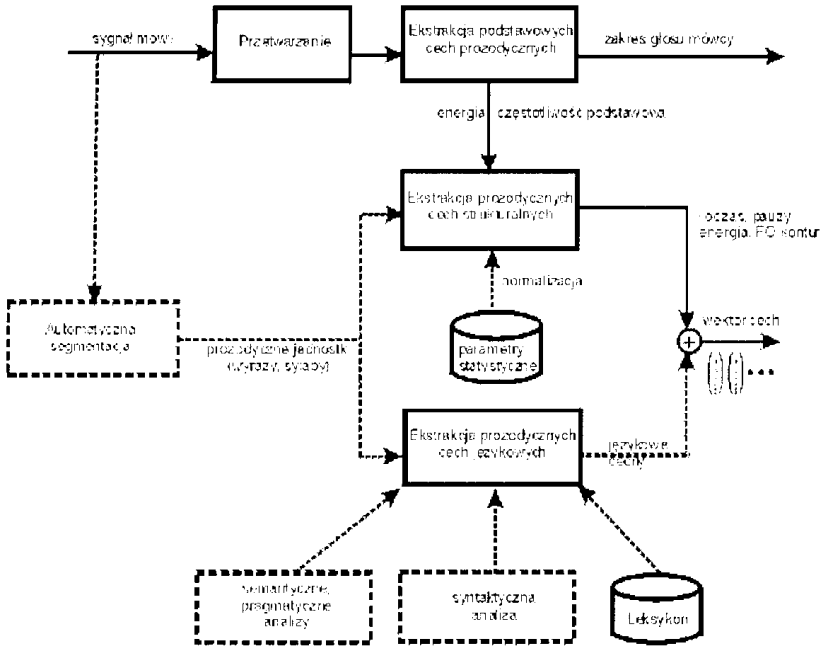
Najszerzej wykorzystano integrację cech segmentalnych i suprasegmentalnych w prototypowym, automatycznym systemie tłumaczenia tekstów –Verbmobil (por. np. Hess et al. 1997). Na podstawie 242 cech opisujących cechy suprasegmentalne sylaby (względem 6 poprzedzających i 6 następujących sylab), uzyskano dla mowy spontanicznej poprawność rozpoznawania akcentu i granic frazy w zakresie 82,5 – 91,7%.

Na ryc.6 przedstawiono moduł analizy prozodycznej zaprojektowany w systemie Verbmobil.

W bloku przetwarzania mowy przeprowadzana jest głównie filtracja (ograniczenie do określonego zakresu częstotliwości) sygnału akustycznego. Wyniki ekstrakcji podstawowych cech prozodycznych (częstotliwości podstawowej oraz energii sygnału) przesyłane są do bloku analizy strukturalnej, w którym formułowany jest wektor akustycznych cech prozodycznych. Parametry muszą cechować się łatwością pomiaru – kryterium to jest związane ze złożonością procedur pomiarowych (przykładowo: pomiar amplitudy, jest znacznie prostszy niż ekstrakcja formantu). Ważna jest również stabilność parametrów (zakresy zmienności mierzonych parametrów powinny mieścić się w określonych przedziałach zmienności) oraz odporność na zakłócenia (wnoszonych przez otoczenie, w którym rejestrowany jest sygnał mowy, jak i na zakłócenia wprowadzane przez tor transmisyjny, np. przez łącza telefoniczne od nadawcy do komputera).

Parametry pierwotne bez segmentacji sygnału nie stanowią bezpośrednio podstawy klasyfikacji. Segmentacja sygnału przeprowadzana jest w zakresie określenia granic samogłosek oraz sylab. W przypadku analizy suprasegmentaliów precyzyjne określenie granic samogłoski lub sylaby nie jest wymagane. Przeciętna dokładność pomiaru rzędu 10 – 15 ms zapewnia wystarczająco poprawną segmentację.

Na podstawie analiz semantycznych, pragmatycznych i syntaktycznych tworzony jest wektor cech językowych. W wyniku sumowania wektora cech językowych oraz cech akustycznych formułowany jest wyjściowy wektor cech reprezentujący struktury melodyczne. Wektory te, przesyłane są do bloku klasyfikacji, gdzie odpowiednie procedury dokonują ich porównania ze znajdującymi się w pamięci wzorcami, czyli obrazami posiadającymi uogólnione wartości i opisy danych rozpoznawanych struktur melodycznych. Klasy wzorcowe oraz inne dodatkowe dane, np. językowe, ułatwiające proces rozpoznawania tworzone są wcześniej, w trakcie procesu zwanego uczeniem. Wyniki klasyfikacji w postaci kodu lub ciągów kodowych mogą stać się danymi wejściowymi systemu dialogowego.



Ryc. 6. Moduł analizy prozodycznej w systemie Verbmobil.

Od dzisiejszych systemów rozpoznawania mowy wymaga się całego spektrum działań; począwszy od prostych systemów, znajdujących już swoje realne zastosowania w życiu codziennym, po systemy skomplikowane, ekspertowe. Każdy system „speech recognition" wymaga uszczegółowienia celów, do jakich będzie wykorzystywany, zasad współpracy użytkownika z systemem, zakresu zadań (słownictwa, liczby mówców, dziedziny słownikowej) wymaganych do rozwiązywania przez system – proces ten nazywamy modelowaniem systemu. Obecnie intensywnie rozwijają się następujące dziedziny technologii mowy:

a) przetwarzanie mówionego języka (SLP spoken language processing),
b) automatyczne rozpoznawanie mowy (ASR automatic system recognition, input),
c) automatyczne wytwarzanie mowy (TTS synthesis, *text to speech*, output),
d) systemy rozumiejące mowę (SUS speech understanding system, input/output),

e) systemy dialogowe (SDS speech dialogue systems),
f) systemy analizy lub pozajęzykowego przetwarzania,
g) identyfikacja/weryfikacja mówców (forensic phonetics),
h) identyfikacja konkretnego języka.

Najbardziej obiecujące zastosowania znajdują systemy dialogowe w telekomunikacji. Szybki postęp prac w dziedzinie technologii mowy w ostatnich latach, to niezaprzeczalny dowód na to, że przyszłość automatycznego rozpoznawania mowy rysuje się bardzo wyraźnie. Fakt, że przeszkodą implementacyjną nie są już dzisiejsze technologie komputerowe otwiera coraz szerzej drogę do nowych, coraz bardziej skomplikowanych systemów. W dotychczasowych opracowaniach poświęconych technologii mowy można znaleźć wizję aplikacji o „zupełnie nieograniczonych słownikach, składni i semantyce", które już na progu XXI wieku będą mogły bez problemu rozumieć mowę bez względu na stosowany przez użytkownika język. Komputery pokładowe będą systemami realnej pomocy kierowcom, operatorom maszyn itd., telefonia będzie bardziej interaktywna, wspomagana przez automatyczne systemy głosowe a 24-godzinny serwis usługowy dostępny przez systemy komunikacji telefonicznej bądź internetowej będzie przypominał kontakt z prawdziwym, realnym operatorem.

XXI wiek rozpoczynamy wraz z komputerowymi systemami pomagającymi zdobywać wiedzę, zastępującymi lektorów języków obcych, znoszącymi bariery ludzkiej niepełnosprawności, a przede wszystkim – z systemami coraz bardziej prostego i naturalnego sposobu wymiany olbrzymiej ilości otaczających nas informacji.

IV. Współczesne systemy komercyjne

Ostatnie lata przyniosły znaczny rozwój popularnych komputerów klasy PC, a tym samym możliwość wejścia w życie technik opracowywanych do tej pory tylko na dużych systemach o wielkiej mocy obliczeniowej. Firma Dragon Inc. w najnowszym pakiecie Dragon Naturally Speaking zapewnia skuteczne rozpoznawanie mowy ciągłej już na komputerze klasy PC z procesorem 486 DX2/66, 16 MB RAM oraz 40 MB wolnej powierzchni dyskowej. Słownik pakietu zawiera w podstawowej wersji 60 tysięcy słów (pakiet dodatkowy rozszerza słownik do 120 tysięcy). Podobne wymagania sprzętowe posiadają także systemy firm IBM (IBM Via Voice) ze swoim 64-o tysięcznym słownikiem, Lernout & Hauspie (Voice Xpress Plus) czy Kurzweil (Kurzweil Voice). Trochę większe wymagania mają systemy bardziej nastawione na technologie badawcze, jak np. BBS (Baylar Biomedical Services Dallas), wymagające procesora minimum Pentium 200, 60 MB HDD oraz 48 MB RAM. Systemy: BBS uwzględniający cechy prozodyczne mowy oraz system Philips FreeSpeech'98 wymaga około 30-to minutowego treningu, po którym skuteczność działania wzrasta nawet do 95%. Niestety, nie istnieją jeszcze polskie rozwiązania rozpoznawania mowy; w laboratorium sopockiej firmy Drive powstał system Lektor 4.0, który służy przede wszystkim do syntezy dźwięku, ale potrafi rozpoznawać też krótkie komendy głosowe. Także i wrocławski Neurosoft pracuje nad modułem rozpoznawania dla pakietu SynTalk (projekt NeuroEar). Aktualnie prowadzony jest przez

Instytut Lingwistyki UAM oraz Politechnikę Poznańską projekt poświęcony rozpoznawaniu mowy, przeznaczony do sterowania głosem systemu Windows: *System rozpoznawania mowy dla komunikacji głosowej osób niewidomych z komputerem.*

Bibliografia

Demenko G., Nowak I., Imiołczyk J. (1993) *Analysis and Synthesis of Pitch Movements in a Read Polish Text*, Proceedings of the 3rd Eurospeech'93, Berlin, vol. 2, 797-800.

Demenko G. (1999) *Analiza cech suprasegmentalnych jezyka polskiego na potrzeby technologii mowy.* Wyd. UAM, Poznań.

Hess W., Batliner, A., Kiessling, A., Kompe, R., Nöth, E., Petzold, A., Reyelt, M., Strom, V. (1997) *Prosodic Modules for speech Recognition and Understanding in VERBMOBIL*, in Computing Prosody, Sagisaka, Y., Campbell, N., Higuchi, N. ed., Springer-Verlag New York, Inc., 361-381.

Hirst D., Di Cristo A. (1998) *Intonation Systems, A survey of Twenty Languages*, ed.by Daniel Hirst, Cambridge University Press.

Jassem W. (1973) *Podstawy fonetyki akustycznej*, PWN, Warszawa.

Pols L. (1999) *Flexible, robust, and efficient human speech processing versus present-day speech technology*, Proceedings of the XIVth International Congress of Phonetic Sciences, s. 9-12. Materiały ICPhS99, s. 9-12.

Sagisaka Y., Campbell N., Higuchi N. (1997) *Computing Prosody, Computational Models for Processing Spontaneous Speech*, Springer -Verlag, New York.