Scripta Neophilologica Posnaniensia, Tom V, strony: 153 – 191
Wydział Neofilologii, UAM Poznań, 2003

MODELOWANIE ZMIAN CECH PROZODYCZNYCH NA POTRZEBY SYNTEZY MOWY

ANDRZEJ PLUCIŃSKI

1. Wstęp – znaczenie prozodii w komunikacji językowej i w syntezie

Obecny poziom syntezy mowy z tekstu osiągnął stan, w którym trudno jest niekiedy odróżnić ją od mowy naturalnej. Dotychczasowe wysiłki skupiały się przede wszystkim na poprawnym odwzorowaniu artykulacji poszczególnych głosek. Równolegle niemalże podejmowano próby wprowadzenia intonacji i akcentów do syntezowanego sygnału. Mowa bez tych elementów jest monotonna i na dłuższą metę męcząca dla słuchacza. Naturalność brzmienia wymaga poprawnego oddania cech prozodycznych sygnału mowy, a w szczególności wspomnianego akcentowania i intonacji. Zadanie to jest bardzo trudne do wykonania, ponieważ w parametrach tych zawarta jest niekiedy istotna treść semantyczna, której przekazanie wymaga rozległej informacji pozajęzykowej, trudnej do uzyskania z samego tylko tekstu. Do pełnego rozwiązania tego problemu konieczna okazuje się głębsza analiza tekstu – syntaktyczna, semantyczna, a nawet pragmatyczna.

W świecie dobrze została opanowana analiza syntaktyczna tekstu. Analiza semantyczna i pragmatyczna leżą wciąż poza możliwościami współczesnych systemów. Możliwe jest jednak zadowalające wytworzenie intonacji i akcentów leksykalnych właśnie tylko na podstawie analizy syntaktycznej. Podstawą jest kilka elementarnych spostrzeżeń dotyczących zachowania się czytającego oraz zaawansowana analiza psychoakustyczna dotycząca percepcji dźwięków i lingwistyczna dotycząca reguł realizacji zjawisk prozodycznych.

Analiza psychoakustyczna ukazuje możliwości uproszczenia konturów melodycznych, a na bazie analizy lingwistycznej tworzone są modele intonacji uwzględniające hierarchię fraz prozodycznych i rozkład akcentów w wypowiedzi.

Prozodia sygnału mowy jest realizowana poprzez cechy prozodyczne, do których zaliczają się:

- ton podstawowy (krtaniowy, *F0*),
- czasy realizacji poszczególnych fragmentów,
- poziom natężenia dźwięku.

Zdarzenia prozodyczne

Parametry prozodyczne w syntezatorze są realizowane przez generator zdarzeń prozodycznych. Ponieważ zdarzenia prozodyczne dotyczą sylab lub ich grup, więc mówi się, że tworzą zjawiska suprasegmentalne. Podobnie jak inne własności sygnału mowy, zdarzenia prozodyczne mogą być rozpatrywane na różnych poziomach reprezentacji: akustycznym, percepcyjnym i lingwistycznym.

Cechy akustycznego poziomu opisu mogą być mierzone za pomocą odpowiedniego hardware'u. Wartości cech poziomu percepcyjnego mogą być obliczane na podstawie wiedzy z psychoakustyki. Poziom lingwistyczny zaś reprezentuje prozodię wypowiedzi jako sekwencje abstrakcyjnych jednostek – znaków, czy symboli.

Reprezentacje lingwistyczne nie mogą być mierzone. Mogą być weryfikowane zaledwie na poziomie opisu. Jednakże możliwe jest konstruowanie systemów syntezy mowy z tekstu (SMT), które generują transkrypcję prozodii danych wypowiedzi zgodnie z pewnym modelem lingwistycznym.

Współdziałanie cech prozodycznych w sylabie wywołuje wrażenie przycisku (uwydatnienia). Uwydatnienie to, subiektywnie odczuwane też jako zwiększenie „siły", dotyczy głośności i dynamicznych właściwości oraz tonu i czasu trwania sylaby (Sluijter et al., 1995). Parametry akustyczne związane z prozodią podlegają krótkim, nie percypowanym wahaniom. Są to zdarzenia mikroprozodyczne. Ich obecność ma mały wpływ na naturalność brzmienia, dlatego pomija się je zarówno w opisie jak i w syntezie.

Znaczenie cech

Wśród zjawisk prozodycznych najbardziej oczywista jest zmienność tonu (rzędu pięciu półtonów). Zmienność ta ma największe znaczenie spośród wszystkich cech prozodycznych. Dlatego też na jej modelach głównie skupiamy naszą

uwagę. Mniej uwagi poświęca się modelom iloczasu, ponieważ są one mniej krytyczne dla naturalności brzmienia niż modele intonacji. Najmniejsze znaczenie ma energia. Okazuje się, że parametr ten jest dla każdego fonemu w przybliżeniu niezmienny w wypowiedziach niezbyt mocno nacechowanych emocjonalnie, dlatego można go traktować jako parametr właściwy fonemów (nie oznacza to niezmiennej głośności, bo to wrażenie zależy także od iloczasu).

Prozodia w komunikacji językowej

Cechy prozodyczne mają specyficzne funkcje w procesie komunikacji. Najłatwiej dostrzeganym efektem ich realizacji jest akcent zdaniowy. Prozodia wypełnia też inne, mniej oczywiste ogólniejsze funkcje, mianowicie cechy prozodyczne:

(1) dzielą łańcuch mowy na grupy sylab (dają przyczynek do grupowania ich w większe jednostki),
(2) ustalają porządek hierarchiczny pomiędzy takimi grupami, wskazując że są one w jakiś sposób połączone[1],
(3) sygnalizują: – koniec zdania oznajmującego poprzez spadek *F0*,
– koniec zdania pytającego poprzez wzrost *F0*.

Struktura prozodyczna jest skompletowana (zakończona), jeśli napotkany zostanie końcowy (najniższy) ton sekwencji, a zamknięta struktura jest deklaracją. Jeśli tak się nie stanie, a wypowiedź się zakończy, to struktura jest niedokończona lub zostawiona otwarta wskazująca na związek z tym, co ma nastąpić. Te aspekty intonacji klasyfikujemy jako lingwistyczne (funkcjonalne). Są one częścią struktury języka tak samo, jak morfologia czy syntaktyka i są specyficzne dla języka.

Prozodia wyraża także aspekty emocjonalne, np. w mowie ludzi zdenerwowanych następują szybsze zmiany tonu i mają większy zakres (większą dynamikę). Jednak, podczas gdy zakres zmienności tonu podstawowego może podlegać czynnikom emocjonalnym, to podstawowe funkcjonalne kształty tonu nie. Stany emocjonalne nie zmieniają kodu lingwistycznego, oddziałują zaledwie na jego realizację. Dlatego te aspekty łącznie z innymi parametrami takimi, jak jakość głosu, są zwane niekiedy paralingwistycznymi.

---

1 Zdanie *John says Peter is a layer* można wymówić jako *John, says Peter, is a layer* albo jako *John says: Peter is a layer*, o czym decyduje intonacja (nie mamy tu na myśli czytania, bo tu byłaby odpowiednia interpunkcja, lecz swobodne mówienie).

Syntaktyka i semantyka a prozodia

Zasadniczym zadaniem generatora prozodycznego jest wyprowadzenie akustycznej reprezentacji prozodii z poziomów modeli lingwistycznych biorąc pod uwagę leksykon, syntaktykę, semantykę i pragmatykę. Początkowo przyjmowano, że intonacja, jaką stosujemy w zdaniu, jest tylko funkcją interpunkcji. Znaki interpunkcyjne są jednak zaledwie zgrubnymi znakami pokazującymi strukturę języka. Intonacja jest echem tej struktury, której analiza akustyczna pozwala percypować wszystkie jej subtelności.

Wielkie znaczenie dla syntezy mowy z tekstu ma możliwość opisu prozodii na podstawie stosunkowo prostej analizy syntaktycznej. Jest intuicyjnie oczywiste, że syntaktyka ma duże znaczenie dla prozodii. Wskazują na to proste doświadczenia z czytaniem tekstu, w którym dokonano zgodnych z zasadami syntaktyki podstawień prowadzących do utraty logicznego sensu jego zdań, ale zachowujących oryginalną fleksję i słowa funkcyjne. Pokazują one, że w większości przypadków możliwe jest poprawne oddanie intonacji przez czytającego bez rozumienia czytanego tekstu. To sugeruje, że zadowalające efekty można osiągnąć poprzestając na powiązaniu intonacji ze strukturą syntaktyczną wypowiedzi. Powiązanie struktury intonacyjnej ze strukturą składniową nie jest jednak proste, ponieważ granice tych struktur nie zawsze się pokrywają. Struktura intonacyjna jest też bardziej „płaska" niż struktura składniowa.

Poza zasięgiem nie tylko analizy syntaktycznej, ale i semantycznej pozostaje wiele sytuacji, które wynikają z intencji czytającego, jego chęci zaznaczenia swej osobowości czy upodobań estetycznych. Należą one do tzw. pragmatyki.

2. Modele intonacji

Generowanie cech prozodycznych na potrzeby syntezy wymaga uprzedniego utworzenia modelu dotyczącego ich kształtowania w mowie naturalnej. Zadaniem modeli jest m.in. pominięcie wahań wynikających ze zmian mikroprozodycznych. Wahania te eliminuje się w dwóch etapach postępowania: na etapie pomiaru *F0* i na etapie modelowania akustycznego, bądź „stylizacji" percepcyjnej konturów *F0*. Na etapie pomiaru stosuje się uśrednianie ważone, np. wg wzoru (d'Alessandro & Castelengo, 1994):

$$\hat{F}(t) = \frac{\int_0^t e^{-\alpha(t-\tau)} F(\tau)\, d\tau}{\int_0^t e^{-\alpha(t-\tau)}\, d\tau},$$

gdzie: F to wartości *F0*, a współczynnik α, znaleziony metodą najmniejszych kwadratów, jest równy 22.

Badania pokazały, że kontur tonu w dłuższych wypowiedziach może być sprowadzony do sekwencji elementarnych konturów lub nawet do statycznych tonów docelowych związanych z sylabami lub ich częściami. To podejście pozwala sprowadzić opis intonacji do sekwencji tonów: niski|wysoki czy rosnący|malejący. Podejście to jest najbliższe intuicyjnemu, językoznawczemu opisowi sygnału mowy i znacznie ułatwia proces stylizacji.

Opracowano pewną liczbę formalizmów metod transkrypcji intonacji. Mogą one być klasyfikowane jako akustyczne, percepcyjne i lingwistyczne. Parametry ich są obliczane metodą resyntezy na podstawie porównania wyników przewidywań płynących z modeli z parametrami naturalnego sygnału mowy. Obliczenia są wykonywane bądź wprost metodą najmniejszych kwadratów (model *Fujisakiego*), bądź poprzez ustalanie punktów zwrotnych po każdym przekroczeniu zadanych a priori progów odchyleń (stylizacja akustyczna) lub określonych na podstawie przesłanek psychoakustycznych (stylizacja percepcyjna).

Szczegółowe omówienie poszczególnych modeli zaczniemy od modeli akustycznych, po czym omówione zostaną modele percepcyjnych i lingwistycznych.

2.1. Modele akustyczne

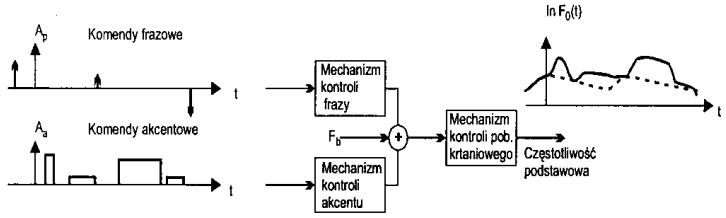
W modelach akustycznych stosowano dwa podejścia:

(1) *Fujisakiego* i
(2) „stylizację akustyczną".

Model Fujisakiego (superpozycyjny)

Model *Fujisakiego* (1992, wcześniejsza wersja Ohman, 1967) oparty jest na założeniu, że krzywe intonacyjne, mimo iż ciągłe w czasie i częstotliwości, pochodzą od zdarzeń dyskretnych (rys. 1). Fujisaki (Ljungqvist & Fujisaki, 1993) odróżnia dwa typy zdarzeń:

- komendy frazowe i
- komendy akcentowe.



Rys. 1. Schemat generowania konturów *F0* wg modelu *Fujisakiego* (Ljungqvist & Fujisaki, 1993);
----- intonacja frazowa,
——— suma: intonacja frazowa + akcentowa

Komendy te, modelowane jako funkcje impulsowe, sterują krytycznie tłumione filtry drugiego rzędu, których wyjścia są sumowane celem uzyskania wartości *F0*. Równanie owego filtru, to:

$$lnF_0(t) = F_b + \sum_{i=1}^{I} A_{p_i} G_p(t - T_{0i}) + \sum_{j=1}^{J} A_{aj} \left\{ G_a(t - T_{1j}) - G_a(t - T_{2j}) \right\}$$

gdzie:

$$G_p(z) = \begin{cases} \alpha^2 z \exp(-\alpha z) & \text{dla } z \geq 0 \\ 0 & \text{dla } z < 0 \end{cases}$$

odpowiedź impulsowa mechanizmu kontroli fraz,

$$G_a(z) = \begin{cases} \min\left[1 - (1 + \beta z)\exp(-\beta z), \gamma\right] & \text{dla } z \geq 0 \\ 0 & \text{dla } z < 0 \end{cases}$$

funkcja schodkowa odpowiedzi mechanizmu kontroli akcentu,

natomiast:

F_b – to asymptotyczna wartość *F0* przy nieobecności składników akcentu,

I – liczba komend frazowych,

A_{pi} – wartość *i*-tej komendy frazowej,

G_p – reprezentuje odpowiedź impulsową mechanizmu kontroli fraz,

T_{0i} – to położenie w czasie *i*-tej komendy frazowej,

J – liczba komend akcentowych,

A_{aj} – amplituda *j*-tej komendy akcentowej,

G_a – schodkowa funkcja odpowiedzi mechanizmu kontroli akcentu,

T_{1j} – start (*onset*) *j*-tej komendy akcentowej,

T_{2j} – koniec *j*-tej komendy akcentowej,

α – częstotliwość kątowa mechanizmu kontroli fraz,

β – częstotliwość kątowa mechanizmu kontroli akcentu,

γ – maksymalny (*ceiling*) poziom składowej akcentu (= 0,9).

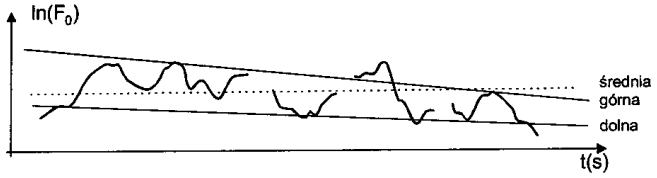
α, β są stałe w obrębie jednej wypowiedzi, ale mogą się zmieniać od wypowiedzi do wypowiedzi a także od osoby do osoby.

Korzystając z tej formuły można, po dopasowaniu jej parametrów do danego konturu *F0*, rozdzielić go na składowe dotyczące melodii frazy i akcentu. Podejście to wykorzystuje Möbius (Möbius et al., 1993).

Algorytm automatycznie analizujący kontury *F0* wg tego modelu przedstawił Geoffrois (1993). Algorytm ten dopasowuje na bieżąco parametry do obserwowanego konturu metodą najmniejszych kwadratów. W postępowaniu swym algorytm ten bada rozbieżność pomiędzy konturem generowanym przez model o aktualnych parametrach, a konturem rzeczywistym. Jeśli rozbieżność staje się istotna, to aktualizuje parametry modelu. Chwile, w których do tego dochodzi, wyznaczają rozkład czasowy komend, a efekty dopasowania określają amplitudy komend. Przydatność tego postępowania potwierdza też Rossi (2002) dla j. włoskiego.

Stylizacja akustyczna

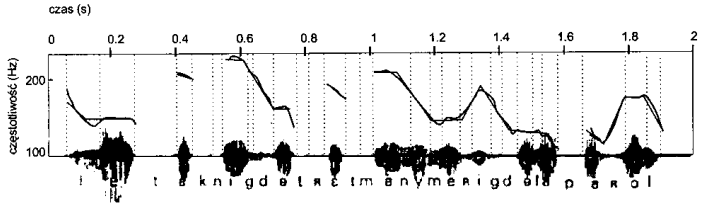
Innym sposobem prezentacji informacji wbudowanej w przebiegi *F0* jest ujawnienie inwariantów akustycznych, które mogą one zawierać, poprzez wyliczenie linii nachylenia (deklinacji) i aproksymowanie krzywych sekwencją punktów docelowych. Ustalono, że w wielu językach w obrębie wypowiedzi *F0* ma tendencję opadającą – rys. 2. Linie deklinacji w danej wypowiedzi oblicza się poprzez dopasowanie prostych do punktów minimów i do punktów maksimów (w skali log) oraz do ich obu otrzymując 3 linie: podstawową, średnią i szczytową (Vaissierre, 1983).



Rys. 2. Linie deklinacji (BeaugendreI 1995)

Obliczenie linii deklinacji jest utrudnione, ponieważ mówiący często „resetuje" przebieg po osiągnięciu szczególnie niskich wartości (zwykle po pauzie oddzielającej główne składniki frazy (t'Hart et al., 1991)). Obserwuje się też bardzo niskie *F0* na początku i końcu wypowiedzi, które przesuwają linie deklinacji. Linia szczytowa też nie zawsze rysuje się wyraźnie wskutek realizacji akcentu. Lepsze estymaty można uzyskać po „stylizacji percepcyjnej".

Niezależnie od obliczeń prostych deklinacji, kontury *F0* przybliża się wyznaczając szereg punktów docelowych i łącząc je liniami prostymi lub wyższego rzędu (t'Hart 1991, stwierdził, że wystarczy interpolacja prostoliniowa). Takie przybliżanie *F0* określa się mianem stylizacji akustycznej.



Rys. 3. Akustyczna stylizacja prostoliniowa. (wypowiedź w j. francuskim: *Les techniques de traitement numérique de la parole...*)

Parametry funkcji interpolujących są albo narzucane *a priori* albo przybliżane regresją liniową. Pozycje punktów docelowych są otrzymywane **poprzez obliczanie korelacji** pomiędzy kolejnymi wartościami *F0* a przybliżeniem i wstawianie punktu zwrotnego za każdym razem, kiedy korelacja spada poniżej wyspecyfikowanej wartości (Scheffers, 1988). Punkt zwrotny, to punkt „docelowy" dla poprzedzającego go odcinka konturu.

Stylizacja może być:

- szeroka – mało punktów docelowych (np. Pierrehumbert 1980),
- wąska – dużo punktów docelowych (3 punkty na każdą samogłoskę, np. Larreur, 1989),

Stwierdzono, że sposób interpolacji nie jest tak istotny, jak owe punkty docelowe.

2.2. Percepcyjne modele intonacji

Wadą przedstawienia akustycznego jest niepewność co do tego, czy:

(a) nie uśrednia ono zdarzeń percypowanych oddzielnie,
(b) nie udostępnia nie percypowanych szczegółów.

Percypowany ton, to nie to samo co *F0*. Wzorce *F0* zależą od:

- wysokości właściwej samogłosek i spółgłosek,
- charakterystyki źródła,
- głośności,
- siły fonacji.

Zaproponowano dwa modele:

- stylizacji za pomocą elementów zbioru konturów elementarnych – IPO, oraz
- stylizacji automatycznej wykorzystującej tzw. „progi *glissando*".

2.2.1. Model IPO

Idea stylizowania konturów *F0* powstała w połowie lat 60. w IPO (*Instytut Badań nad Percepcją Eindhoven*). Stylizacja konturu jest oparta na założeniu, że kontur tonu może być reprezentowany sekwencją linii prostych.

Postępowanie oparte jest na metodzie analizy przez syntezę. Wypowiedzi są analizowane i resyntezowane po zastosowaniu stylizacji konturów tonu podstawowego. Stylizacja polega na przybliżaniu liniami prostymi logarytmu konturu *F0*. Stylizowany kontur przybliża się sekwencją wzorcowych odcinków branych ze zbioru wzorców standardowych. Każda ze stylizowanych linii ma swoje nachylenie, czas trwania i pozycję w sylabie[2].

2 Początkowo czyniono to metodą prób i błędów oceniając wynik na podstawie resyntezy. Później posługiwano się zbiorem prototypowych odcinków linii prostych, które traktowano jako podstawowe jednostki intonacyjne. Odcinki te były synchronizowane z nagłosem

W podejściu IPO definiuje się formalną gramatykę umożliwiającą grupowanie konturów *F0* w podzbiory zwane standaryzowanymi konturami *F0*. Zarówno zbiór standaryzowanych konturów *F0* jak tworzące je gramatyki są specyficzne dla języka (t'Hart, et al. 1991, Collier, 1991). Zarzuca się jej głównie to, że opiera się tylko na przebiegach *F0*, a nie odwołuje do mechanizmów percepcji intonacji.

2.2.2. Model z tonami docelowymi

Hirst, Nicolas i Espesser (1991) zaproponowali stylizację z wartościami docelowymi tonu zamiast (standaryzowanych odcinków) linii prostych, stosując ją do segmentów w pełni sonornych. Zgodność (z oryginałem) sprawdzali jednak tylko wizualnie.

2.2.3. Automatyczna stylizacja percepcyjna

Glissando

Szybkość zmiany *F0* określa tzw. współczynnik *glissando* mierzony w półtonach / sekundę ($= 12 \log_2(F_{[Hz]})/T_{[s]}$).

Dostrzegalną różnicę określa się mianem „progu *glissando*"[3]. Próg ten zmienia się wraz z długością pobudzenia i rozkłada się wokół krzywej G_{tr}

$$G_{tr} = \frac{0{,}16}{T^2},$$

gdzie T jest długością tonu. Przy T równym w przybliżeniu długości sylaby współczynnik ten zawiera się w przedziale

$$\ln(G_{tr}) \pm \ln(2) \qquad \text{dla } T \approx \text{długość sylaby.}$$

W zastosowaniu do sygnału mowy segment dźwięczny określa się jako dynamiczny, jeśli zmiany przekraczają próg *glissando* i jako statyczny, jeśli nie. Różnicowy próg glissando zmian tonu, to najmniejsza różnica w nachyleniu (zbocza konturu) konieczna do rozróżnienia dwóch kolejnych glissando. t'Hart stosuje współczynnik g_1/g_2, gdzie g_1 i g_2 to nachylenia zboczy wyrażone w Hz/s. Znaleziono, że minimalny stosunek, to $2 \div 10$ w zależności od długości tonów.

samogłosek i były przeznaczone dla normalnego tempa mowy. Resynteza pozwalała ocenić równoważność stylizacji z konturem oryginalnym.

3 Seargent & Harris, 1962; Pollack, 1968; Rossi, 1971, 1978; Shouten, 1985; Mertens, 1987, 1989; t'Hart, 1976, 1990.

Stylizacja

Celem stylizacji jest znalezienie istotnych dla percepcji części i własności konturu. Przedstawiamy tu jedną z najlepiej uzasadnionych metod postępowania wg d'Alessandro i Mertensa (Mertens, 1987; d'Alessandro & Mertens, 1995). Autorzy ci opierają się na hipotezie mówiącej, że nie wszystkie zmiany tonu są postrzegane. O postrzegalności zmian decyduje tzw. „próg *glissando*" wyrażany w półtonach/s.

Postępowanie w metodzie percepcyjnej stylizacji zawiera się w trzech krokach:

(1) Segmentacja fonetyczna i sylabiczna (akustycznego sygnału mowy).
(2) Określenie i wygładzenie krzywych *F0* za pomocą ważonego uśredniania czasowego punktów widzianych przez ruchome okienko. Operacja ta jest usprawiedliwiana analogicznymi właściwościami słuchu (integracja).
(3) Stylizacja konturów *F0* w obrębie sylab. Kontur *F0* sylaby jest rozkładany najpierw na ciąg segmentów tonalnych na podstawie progu *glissando* zwykłego i różnicowego. Po tym kroku następuje stylizacja, w której docelowe punkty *F0* są wiązane z segmentami tonalnymi.

Metoda ta była intensywnie badana dla języka francuskiego. Donosi się o zachowaniu po resyntezie wysokiej jakości percepcyjnej sygnału. Przeprowadzono także badania porównawcze z modelami lingwistycznymi dla j. francuskiego i duńskiego. Metodę stosowano też dla celów automatycznego rozpoznawania intonacji (Mertens, 1989).

Postępowanie

A. Segmentacja konturu

Segmentację konturu oparto na 2 progach percypowanych: *glissando* i *glissando* różnicowym.

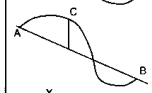
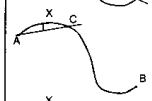
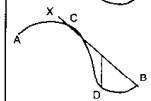
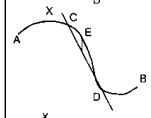
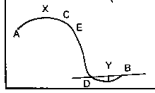
Segmentacja konturu sylabicznego polegała na lokalizowaniu punktu istotnej zmiany w jego przebiegu – tzw. punktów „zwrotnych". Punkty zwrotne były lokalizowane poprzez dopasowywanie linii prostej do punktów widzianych przez okno czasowe i poprzez obliczanie różnic pomiędzy dopasowaną linią a wartościami tonu. Punkt najbardziej odstający obierano za punkt zwrotny i za potencjalną granicę segmentu tonalnego, po czym rekurencyjnie ponawiano analizę dla podzielonego okna dotąd, dokąd nie osiągnięto punktu granicznego

określonego przez współczynnik *glissando* lub jeśli różnica w potencjalnym punkcie zwrotnym spadła poniżej 1 półtonu[4].

B. Przypisanie percypowanych tonów i stylizacja

Stylizacja konturu *F0* sprowadzała się do interpolacji liniowej pomiędzy kolejnymi punktami wykonywanej w skali liniowej (wyrażonej w Hz)[5]. Kierując się różnicowym progiem *glissando*, łączono sąsiadujące odcinki w jeden (interpolacja punktów krańcowych z pominięciem punktu styku) jeśli różnica *glissando* była podprogowa. Następnie tak wyłonione punkty łączono odcinkami prostoliniowymi.

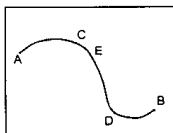
(a) Punkty zwrotne – umieszczane tam, gdzie odchyłka jest największa

	analizowany interwał	maksymalna różnica w punkcie	segmentacja g>GT	nowy punkt zwrotny
	–	–	–	–
	A–B	C	tak	C
	A–C	X	nie, zbyt mała odchyłka	–
	C–B	D	tak	D
	C–D	E	tak	E
	D–B	Y	nie	–

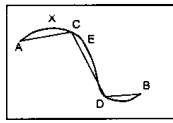
4 Łatwo jednak zauważyć, że w tym postępowaniu nie jest istotny wybór okna i jednostki. Prawdopodobnie równie dobre rezultaty osiągnięto by przy analizie znacznie dłuższych konturów, np. w zdaniu (po interpolacji odcinków bezdźwięcznych).

5 Co po przejściu do skali logarytmicznej musiało powodować zaokrąglenie interpolujących odcinków, łagodzące ostrość złącz.

(b) Łączenie – sąsiednie odcinki są łączone, jeśli różnica glissando nie przekracza pewnego progu

	porównaj części	g_2-g_1>DGT	decyzja
	AC z CE	tak	zachowaj AC
	CE z ED	nie	połącz CE i ED
	(CE + ED) z DB	tak	zachowaj CD
	DB	–	zachowaj DB

(c) Stylizacja – zastąpić oryginalny kontur *F0* przybliżeniem odcinkami prostymi

	Segmenty tonalne	g>GT	zbocze
	AC	tak	dynamiczny
	CD	tak	dynamiczny
	DB	nie	statyczny

Rys. 4. Algorytm automatycznej stylizacji wg Alessandro & Mertens (1995).
(a) rekurencyjna segmentacja konturu, wyznaczenie punktów zwrotnych,
(b) porównanie i łączenie segmentów tonalnych,
(c) stylizacja.

Etap łączenia korzysta z różnicowego progu *glissando* (DGT), a na etapie wyzna-czania punktów zwrotnych i stylizacji stosowany jest próg *glissando* (DG).

Wada Estymaty progów postrzegalności zmian *F0* są wciąż mało wiarygodne.

2.3. Lingwistyczne modele intonacji

Modele lingwistyczne operują skrajnie ograniczoną, dwupoziomową klasyfikacją wysokości tonów – wysokie|niskie, a konturów na rosnące|malejące. Bogatsza jest klasyfikacja sylab – klasyfikuje się je według położenia w obrębie frazy na akcentowane i nieakcentowane.

Dwa główne podejścia, to:

- analiza konturu tonu (Crystal 1967, Halliday 1969),
- analiza sekwencji tonów:
  - model *Pierrehumbert*,
  - model *Mertensa*.

2.3.1. Teoria konturu tonu (tonetyka)

Wg niektórych autorów *intonacja wypowiedzi powstaje z sekwencji elementarnych konturów wziętych z ograniczonego zbioru zaprogramowanego w umyśle człowieka i w przybliżeniu realizowanego przez jego aparat głosowy (w akcie mowy). Każdy kontur elementarny, określany mianem tonu, jest widziany jako podstawowa jednostka intonacji, która nie może być już rozłożona na mniejsze części.*

Zastosowanie tej zasady doprowadziło do teorii zwanej „tonetyką" („szkoła brytyjska"; Crystal, 1969; Halliday, 1967; O'Connor & Arnold, 1973). Według tej „szkoły", w wypowiedziach wyróżnia się grupy tonów. Każda z nich składa się z co najwyżej czterech części: wstępnej , początkowej, centralnej i końcowej (*prehead, head, nucleus, tail*)[6]. Jedyną obowiązkową częścią jest część centralna, która przenosi akcent główny na swej pierwszej sylabie. Tonetyka wiąże sylaby części centralnej (*nucleus*) z ograniczoną liczbą tonów i pewnych ich kombinacji określanych jako złożone tony części centralnej (*compound nuclear tones*).

Proponuje się pięć tonów prostych i dwa złożone (Halliday 1969) lub siedem tonów prostych i tylko jeden złożony (O'Connor). Końcówka składa się z dodatkowych sylab, na których ruchy tonu się kończą. Na część początkową (*head*) składają się sylaby poprzedzające pierwszą akcentowaną z wyłączeniem pierwszej sylaby części centralnej. Sylaby akcentowane części początkowej (akcentowane, gdy część ta występuje w izolacji) otrzymują akcent sekundarny. W ostatniej części (*tail*) następuje wygaszenie ruchów intonacji.

Crystal rozróżnia cztery tony proste:

- opadające,
- rosnące,
- opadająco-rosnące,
- rosnąco-opadające

i cztery tony złożone:

- rosnący + opadający lub opadająco-rosnący + opadający,
- opadający + rosnący lub rosnąco-opadający + rosnący.

Teoria ta pozostaje w silnej zależności od języka i nie jest w stanie uchwycić struktury hierarchicznej prozodii (Ladd,1986). Podejście to adoptuje dla j. polskiego Demenko (1999).

6 Jassem (1984), akcenty tonalne: anakruza, akcent preiktyczny, iktyczny, postiktyczny.

2.3.2. Sekwencyjna teoria tonu

Model *Pierrehumbert*

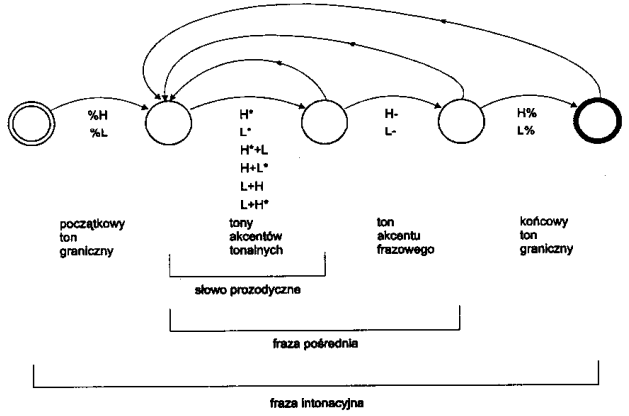
Janet Pierrehumbert przedstawiła w pracach z lat 1980, 1981 i 1988 tzw. ekwencyjną teorię tonu. Teoria ta opisuje krzywe melodii w kategoriach tonów względnych. Tony te są definiowane jako fonologiczne abstrakcje dla punktów docelowych otrzymanych po szerokiej, (czyli z małą ilością punktów) stylizacji akustycznej.

Pierrehumbert rozróżnia dwa tony: wysoki (H) i niski (L). Sekwencje H i L są ograniczone gramatyką skończenie-stanową, która z kolei rozróżnia cztery kategorie tonów na bazie ich własności dystrybucyjnych:

- wstępne tony graniczne (inicjujące) (*% initial boundry tones*),
- tony akcentu tonalnego (** pitch accent tones*),
- tony akcentu frazowego (*- phrase accent tones*),
- końcowe tony graniczne (*final boundry tones %*).

Gramatyka ta jawnie wprowadza trójpoziomowy hierarchiczny opis intonacji – wyróżniając:

- słowa prozodyczne,
- frazy pośrednie,
- frazy intonacyjne.



Rys. 5. Gramatyka skończenie stanowa sekwencji tonów H/L (stany oznaczają sylaby).

Wypowiedzi są rozkładane na słowa prozodyczne, które mają jeden tylko akcent tonalny. Sam akcent tonalny pojawia się jako albo pojedynczy ton (H* lub L*) albo jako dwa tony (H* + L, H + L*, L* + H, L + H*), gdzie * oznacza ton związany z sylabą akcentowaną. Słowa prozodyczne składają się na frazy pośrednie, które same tworzą frazy intonacyjne. Podstawa tonu (*offset pitch*) fraz pośrednich jest kontrolowana przez tonalny akcent frazowy (oznaczany przez -). Ton nagłosowy i wygłosowy fraz intonacyjnych są narzucone przez początkowy i końcowy ton graniczny (%). Teoria ta została głębiej sformalizowana przez Silvermana (1992) w systemie transkrypcji *ToBI* (***T**ones and **B**reak **I**ndicies*).

Implementowano też inne modyfikacje rozróżniające więcej poziomów elementarnych, np. Mertens (1990, 2002) dla j. francuskiego i Mertens (1989) dla j. duńskiego.

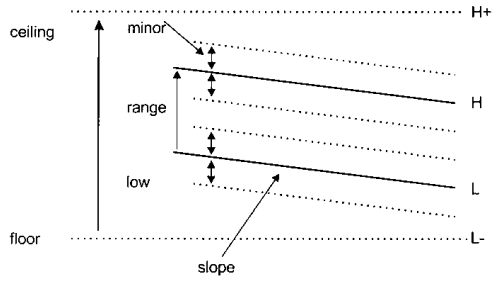
Model *Mertensa*

Mertens (2002) w modelowaniu *F0* wyróżnia parametry globalne i lokalne. Do globalnych zalicza:

- zakres zmian tonu jako cechę osobniczą,
- aktualnie używany zakres zmian tonu,
- wielkość deklinacji,
- tempo mowy,
- zmiany rytmu.

Zakłada się, że parametry globalne mogą modulować kontury tonu dla symulowania emocji i stylów mówienia. Poziomy tonu obejmują (rys 6):

- skrajny zakres zmian: **L–** = minimum, **H+** = maksimum,
- zakres zmian głównych **L**, **H**,
- zakresy podwyższeń i obniżeń granic zakresu zmian głównych **L** i **H**.



Rys. 6. Siatka tonów podstawowych wg Mertensa (2002)

Mertens rozróżnia dwa typy akcentów:

- inicjalny – AI i
- finalny – AF

oraz trzy domeny prozodyczne:

- grupę akcentową – SG (***stress group***; słowo akcentowane + klityki, na pewno akcentowana w izolacji, a w zdaniu niekoniecznie),
- grupę intonacyjną – IG (***intonation group***; wymagana obecność sylaby akcentowanej finalnie (AF)),
- pakiet intonacyjny – IP (*intonation package*; sekwencja grup intonacyjnych (IG)).

Grupa akcentowa, jeśli jest rzeczywiście akcentowana, to na ostatniej pełnej sylabie, która pokrywa się z akcentem typu finalnego (AF) grupy intonacyjnej (IG). Jeśli nie jest akcentowana, to będzie zawarta w pewnej części grupy intonacyjnej.

Grupa intonacyjna składa się z jednej lub więcej powiązanych syntaktycznie grup akcentowych.

Sekwencja grup intonacyjnych tworzy pakiet intonacyjny. Pakiety intonacyjne są zhierarchizowane. Budowa grupy intonacyjnej podlega pewnej gramatyce określonej formułą

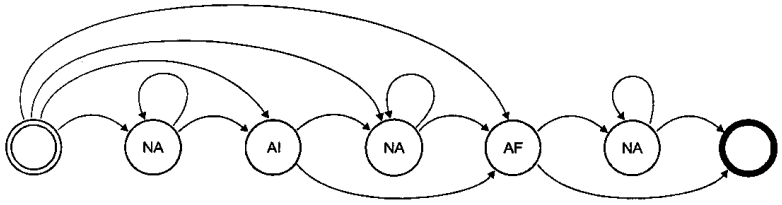
$$IG = \left[\{NA\}AI\right]\{NA\}AF\{NA\}^{7},$$

gdzie nawiasy [] i {} zawierają elementy opcjonalne, a NA oznacza sylabę nieak-centowaną. Formuła ta mówi, że grupa intonacyjna musi zawierać sylabę z akcentem typu AF. Sylaba ta może być poprzedzana pewną liczbą sylab nieakcentowanych (NA), po których może wystąpić sylaba z akcentem typu AI i pewna liczba sylab nieakcentowanych. Po sylabie z akcentem typu AF może nastąpić pewna liczba sylab nieakcentowanych. Formułę powyższą przedstawia graf pokazany na rys. 7.

Każdej sylabie w grupie intonacyjnej (czyli każdemu stanowi powyższego automatu) jest przypisana (podobnie jak u Pierrehumberta) pewna sekwencja tonów docelo-wych zapisanych w kategoriach określonych siatką z rys. 6. Sekwencje te dane są a priori[8].

7 Mertens zapisuje to formułą *IG* = [[*NA*]*AI*][*NA*]*AF*[*NA*], co jednak nie pokrywa się z opisem w tekście. Wg notacji BNF zero- lub wielokrotne wystąpienie zapisuje się za pomocą nawiasów {}. Nawiasy [] oznaczają zaś zero- lub jedno-krotne wystąpienie (w korespondencji prywatnej autor zgodził się z tą uwagą).

8 Autor nie publikuje ich list. Implementacja tego modelu (mingus) pozwala użytkownikowi manipulować tymi parametrami.



Rys. 7. Gramatyka grupy intonacyjnej wg Mertensa; NA – sylaby nieakcentowane, AI – sylaby z akcentem inicjalnym, AF – sylaby z akcentem finalnym

3. Przejście od symbolicznej do akustycznej reprezentacji prozodii

Ostatnim krokiem na drodze ku zaopatrzeniu wypowiedzi we właściwe cechy prozodyczne jest wytworzenie akustycznej reprezentacji jej intonacji i iloczasów jej fonemów oraz pauz.

Przedtem należy określić jej organizację syntaktyczno-prozodyczną, tj. hierarchiczne uporządkowanie grup sylab i oznakowanie ich poziomami akcentu.

3.1. Generowanie *F0*

Przekład opisu syntaktyczno-prozodycznego na postać fizyczną sygnału obejmuje dwa kroki:

(1) Ustalenie zależności pomiędzy parametrami modelu a opisem syntaktyczno-prozodycznym.
(2) Wygenerowanie końcowej krzywej melodii reprezentującej model.

To, jak będzie ostatecznie generowana *F0*, zależy od typu wybranego modelu prozodycznego (czy jest to model akustyczny, percepcyjny czy lingwistyczny).

3.1.1. Generowanie *F0* wg modelu *Fujisakiego*

Model *Fujisakiego* był szeroko wykorzystywany w systemach SMT (Ljungqwist & Fujisaki, 1993; Möbius, et al. 1993; Hirai et al., 1994).

Analiza powiązania rozkładu komend w czasie i ich amplitud z cechami lingwistycznymi wypowiedzi jest przeprowadzana za pomocą narzędzi statystyki (Möbius et al., 1993; Hirai, 1994). To prowadzi do (wieloczynnikowego) modelu kontroli, który wytwarza wartości parametrów (a stąd krzywe melodii dla tekstu danej wypowiedzi), jeśli dostarczy mu się cech syntaktyczno-prozodycznych (iloczasy syntezowanego sygnału muszą być więc obliczane najpierw).

3.1.2. Generowanie *F0* jako sekwencji stylizowanych konturów

Stosuje się zasadniczo dwie strategie. **W pierwszej strategii**, bazującej na pewnym modelu lingwistycznym, wybierany jest a priori zbiór segmentalnych cech prozodycznych, dla których jest organizowane hierarchiczne drzewo decyzyjne lub stosowany jest perceptron.

Drzewo decyzyjne – z każdym liściem tego drzewa są związane kontury melodyczne otrzymane albo przez uśrednianie informacji melodycznej dostarczonej przez każdą jej reprezentację w bazie danych (Auberge, 1991) albo przez arbitralny wybór jednej z nich (lub kilku, jak w Larreur et al., 1989).

PRZYKŁAD. Larreur et al. (1989) na podstawie wielkiego korpusu danych mierzy:

- iloczas sylab,
- iloczas fonemów,
- początkową, środkową i końcową częstotliwość *F0* samogłosek.

Dane te wiąże z informacją segmentalną taką, jak:

- poprzedzający, bieżący i następujący fonem,
- typ bieżącej sylaby (otwarta lub zamknięta),
- typ bieżącego słowa (funkcyjne lub znaczące),
- pozycja bieżącej sylaby wewnątrz słowa,
- pozycja bieżącego fonemu wewnątrz sylaby,
- pozycja bieżącego słowa wewnątrz hierarchii syntaktyczno-prozodycznej.

Perceptron stosowali Sagisaka (1990), Sagisaka & Kaiki (1992) oraz Taylor (1995).

PRZYKŁAD. Sagisaka (1990) modeluje globalny kształt *F0* fraz pośrednich (*minor*) trzema wartościami *F0* (start, szczyt, koniec) jako funkcją informacji syntaktyczno-prozodycznej, takiej jak:

- akcentuacja,
- długość frazy podrzędnej,
- całkowita liczba akcentowanych podrzędnych fraz poprzedzających w bieżącej intonacyjnej frazie nadrzędnej,
- długość poprzedzających i następujących fraz podrzędnych (*minor*) i
- lokalna struktura frazy.

Parametry kontrolne są wiązane z wartościami *F0* za pomocą trójwarstwowego perceptronu.

W metodzie IPO (Balestri et al., 1993; Meyer et al., 1993; Terken, 1993; Beaugendre, 1995) bada się zależności pomiędzy standaryzowanymi wzorcami, a hierarchią syntaktyczno-prozodyczną (Colier, 1991). Wynikający algorytm kontrolny jest złożony z niewielkiej liczby reguł, a jego organizacja zależy od analizowanego języka.

Druga strategia polega na tworzeniu przetwornika (transducer) pomiędzy dostępną informacją syntaktyczno-prozodyczną a obserwowanymi stylizowanymi krzywymi, bez wprowadzania jakiegokolwiek arbitralnego modelu pośredniczącego.

PRZYKŁAD. Traber (1993) uzyskał taką kontrolę *F0* stosując rekurencyjną SN. Przetwornikowi temu prezentowano dwa wejściowe strumienie:

- strumień symboli kodujących akcent, granice frazy i słów,
- strumień właściwości fonetycznych każdej sylaby.

Te strumienie były prezentowane sieci poprzez ruchome okno. Sieć była uczona wiązać

- stylizowane wzorce melodyczne (reakcja, czyli wyjście)
- z symbolami centralnymi okna (pobudzenie).

3.1.3. Generowanie *F0* poprzez sekwencje tonów

Teoria sekwencji tonów jawnie ogranicza dozwolone ruchy *F0* poprzez wymóg zgodności z pewną gramatyką. Modele wynikające z tej teorii jawnie wprowadzają hierarchiczne poziomy w opisie intonacji (frazy pośrednie i intonacyjne (Pierrehumbert, 1980), grupy akcentowe i intonacyjne (Mertens, 1993, 2002)). Nie wyjaśniają jednak, kiedy taka czy inna (akceptowalna) sekwencja jest używana. Użycie takich modeli wymaga zatem wytworzenia opisu intonacji zdań w kategoriach sekwencji tonów jako funkcji:

- ich modalności (oznajmujące, rozkazujące, pytające),
- liczby syntaktyczno-prozodycznych komponentów,
- długości,
- struktury akcentu (która sekwencja tonów powinna być użyta aby wyrazić akcent emfatyczny czy zdaniowy itd.).

Ponieważ wymaga to głębokiego wglądu w teorię lingwistyczną segmentalnych i suprasegmentalnych aspektów mowy, więc ta analiza jest zawsze wykonywana przez lingwistę.

- Pierrehumbert (1981) np. dostarcza opisu neutralnych, deklaratywnych (oznajmujących) intonacji j. angielskiego: akcenty tonalne (*pitch*) to **H**, końcowe tony frazowe to na ogół **L-L** dla fraz terminalnych, a **L-H** dla nieterminalnych. Tony **H** i **L** są dodatkowo wiązane z poziomami akcentu (od 1 do 5) określonymi na podstawie analizy syntaktycznej zdania (Anderson & Pierrehumbert, 1984).
- Podobny opis przedstawia dla j. szwedzkiego Bruce & Grandström (1993) uwzględniając dodatkowo akcent zdaniowy. Korzystają oni także z poziomów uwydatnienia (od 1 do 9).
- Relacje takie (chociaż dużo bardziej rozwinięte), określa także Mertens (1993) dla j. francuskiego.

Określenie sekwencji tonów dla wypowiedzi nie jest jednakże równoważne uzyskaniu jej krzywej *F0*. Tony należy transformować na numeryczne wartości *F0*. Można stosować wiele różnych podejść aby osiągnąć ten cel: od obliczeń średnich konturów tonów odpowiadających tonom w kontekście, do rozwijania systemów opartych na regułach aby wytwarzać docelowe *F0* i interpolować pomiędzy nimi stosując metody statystyczne lub sztuczne sieci neuronowe.

Intonacja wg Mertensa

Mertens stosuje dwa poziomy reprezentacji intonacji:

- tonalny, wiążący sekwencje tonów docelowych z poszczególnymi sylabami,
- fonetyczno-akustyczny, w którym tony docelowe są związane z pozycją w sylabach.

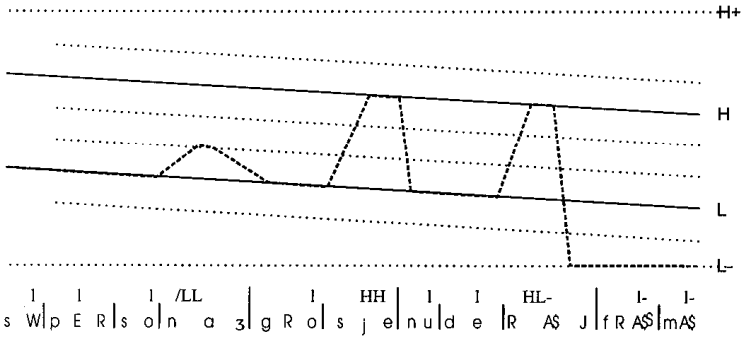
Poziom percepcyjny

Najpierw jest formułowany opis w dziedzinie percepcyjnej. Następnie, na podstawie wykonanej uprzednio analizy syntaktycznej i morfologicznej, wyznacza się grupy akcentowe, i:

(1) grupy akcentowe zalicza się do grup intonacyjnych biorąc pod uwagę liczbę sylab i zależności syntaktyczne (oraz ewentualnie tempo mowy),
(2) na podstawie relacji syntaktycznych[9] każda grupa intonacyjna otrzymuje poziom granicy prozodycznej (czy jest to granica frazy nadrzędnej czy podrzędnej).

9 Autor nie określa bliżej tych relacji.

Kontur *F0* jest tworzony poprzez łączenie liniami prostymi tonów docelowych. Tony docelowe są podawane dla samogłosek. Są to poziomy tonu osiągane w określonych stadiach realizacji dźwięku. Ich liczba zmienia się odpowiednio do kształtu konturu. Dla **HL-** np., są to trzy punkty (osiągane po zrealizowaniu samogłoski w: 33, 50 i 100%), dla **HH** zaś dwa (osiągane po zrealizowaniu samogłoski w 50 i 80%) – patrz tabela 1.



Rys. 8. Kontur *F0* wygenerowany wg modelu Mertensa (2002)

TABELA 1. Stopień realizacji samogłoski, po którym osiągany jest ton docelowy z danej sekwencji

sekwencja tonów	% realizacji, cel
HL–	(33,H), (50,H), (100,L–)
HH	(50,H), (80,H)

Po określeniu rozkładu czasowego tonów i ich kategorii obliczane są częstotliwości absolutne w Hz. Wynik pokazuje rysunek 8.

3.2. Generowanie iloczasów

Sygnał mowy ujawnia zawiłe wzorce czasowe, które powinien naśladować syntezator, aby uzyskać naturalne brzmienie mowy. W określaniu stosunków czasowych wypowiedzi pomocna jest jednostka zwana „dostrzegalną różnicą" iloczasu określoną na 20 ms (Bartkova & Sorin, 1987).

Jednostki iloczasu

Panuje zgoda co do tego, że percepcyjne studia nad długością powinny być oparte na uprzedniej segmentacji na jednostki iloczasowe. Celem określenia takich jednostek przywoływana jest stara zasada *izochronii* na rzecz analizy z sylabą lub stopą[10] jako jednostką (Lehiste, 1977) oparta na interakcji pomiędzy aparatem głosowym a systemem percepcyjnym mówiącego. Według tej zasady, *mówiący nieświadomie używałby pewnego wewnętrznego zegara do rozlokowania segmentów mowy (synchronizowanych z sylabami w językach ze stałym akcentem i ze stopami w językach z ruchomym akcentem)*. Zmiana tempa mowy po prostu implikowałaby zmianę szybkości biegu wewnętrznego zegara[11] (tu jednak nie ma ogólnej zgody, por. Wenk & Wioland, 1982).

Modele

Wiele współczesnych systemów SMT jest opartych na fonetycznych elementach jako ostatecznych jednostkach mowy. Systemy iloczasowe oparte na sylabie zawierają więc pewien algorytm do wyprowadzania długości segmentów z długości sylab (Campbell & Isard, 1991). Campbell wszystkie iloczasy segmentalne w ramach sylaby oblicza wg formuły

$$Dur_i = \exp(\mu_i + k\sigma_i),$$

w której Dur_i jest długością i-tego fonemu w sylabie, a μ i σ, to średnia i odchylenie standardowe logarytmów długości w dużym korpusie. Wychodząc od długości sylaby i biorąc pod uwagę μ i σ jej składowych segmentalnych, ta formuła dostarcza wartości k, skąd można wyliczyć aktualne długości segmentów. W praktyce okazuje się, że stosowanie zasady izochronii prowadzi do zbyt regularnej mowy. Należy zatem przyjmować, że języki naturalne ujawniają tendencję do izochronii. Większość systemów SMT skupia się aktualnie na iloczasie segmentów fonetycznych. van Santen (1993) pisze, że powinny być przy tym uwzględniane efekty granic sylab, słów i fraz.

Bartkova & Sorin (1987) analizowali kilka korpusów j. francuskiego, aby badać osobniczo niezależne iloczasy właściwe i modyfikacje samogłoski w sylabie akcentowanej zachodzące pod wpływem:

- następujących spółgłosek,
- typu słowa (funkcyjne/główne),

10 Stopa jest jednostką rytmiczną, która zawiera tylko jedną sylabę akcentowaną.
11 Często mówi się o niezależności części subfonetycznych od tempa mowy.

- lokalizacji segmentu w słowie,
- odległości od granicy frazy głównej i podrzędnej itp.,
- zgrupowań spółgłosek w różnych pozycjach słowa / frazy prozodycznej / zdania.

To doprowadziło ich do ustalenia prostej reguły multyplikatywnej
Iloczas samogłoski = $ID*V_i*m_c$,
Iloczas spółgłoski = $ID*C_{ij}$,
gdzie:

ID	– to iloczasy właściwe,
V_i, C_{ij}	– współczynniki wydłużające,
i	– pozycja segmentu wewnątrz słowa,
j	– odnosi się do klasy spółgłosek, a
m_c	– odzwierciedla wpływ spółgłoski lub półsamogłoski na poprzedzającą samogłoskę.

Wszystkie wartości współczynników (w sumie 63) były obliczone poprzez uśrednianie danych otrzymanych po zastosowaniu tego modelu do dużego korpusu.

W nowszych podejściach proponowane są bardzo ogólne modele, w których uwzględnia się wielką liczbę możliwych czynników kontrolnych (Kaiki et al., 1990 – generalny model addytywny, Riley, 1992 – CART-y, Campbell, 1992 – SN), a ich parametry są automatycznie wyprowadzane przez związany algorytm (standardowa metoda najmniejszych kwadratów, minimalizacja entropii, BP odpowiednio). Szczególnie godne uwagi są CART-y Riley'a oparte na imponującym zbiorze czynników:

- kontekst segmentalny, (3 lewe, 3 prawe[12]),
- akcent (3 poziomy),
- pozycja leksykalna (liczba segmentów od początku i od końca słowa),
- liczba samogłosek w słowie,
- pozycja we frazie (liczba segmentów od początku i od końca frazy).

Iloczas wg Mertensa

Długości sylab Mertens uzależnia od lokalizacji w grupie intonacyjnej i tempa mowy. Są one uzależnione od pozycji w grupie intonacyjnej. Podstawowe długości podaje tabela 2.

12 Aby ograniczyć różnorodność opisu Riley koduje segmenty czterema cechami fonetycznymi związanymi ze sposobem i miejscem artykulacji.

TABELA 2. Długość sylaby w grupie intonacyjnej

pozycja w grupie intonacyjnej	długość sylaby w ms	mnożnik
akcent finalny (AF)	200	1÷2,4 w zależności od tonu
penultima	152	–
inne	131	–

Ostateczne wartości iloczasu są określane za pomocą „czynnika z" (Mertens, 2001).

4. Parsing syntaktyczno-prozodyczny

Jak już powiedziano, granice fraz prozodycznych można wiązać ze strukturą syntaktyczną wypowiedzi. Z powodu braku analizy semantyczno-pragmatycznej jesteśmy skazani w systemach SMT na intonację neutralną. Nie wyraża ona emocji czy akcentu zdaniowego.

Stosując parsing syntaktyczny jako podstawę wyjściową do określania granic fraz prozodycznych opierano się na spostrzeżeniu mówiącym, że *Kiedy czytamy zdanie, to jesteśmy w stanie rozpocząć jego wymowę, daleko wcześniej zanim osiągniemy jego koniec. Musi zatem być tak, że organizujemy zdanie w proste grupy prozodyczne, które mogą być otrzymane z lokalnej analizy na poziomie frazowym.*

W uzasadnieniu wymienia się też pogląd, że czytający różnicuje prozodię z dwóch powodów:

- aby podzielić wypowiedź na krótsze części, aby (w rezultacie) ułatwić słuchaczowi parsing,
- aby wskazać pewne znaczenie poprzez akcent zdaniowy lub przez podkreślenie słów, które niosą nową informację, czy też aby wskazać, czy po pewnej grupie intonacyjnej następuje kolejna.

Powszechnie przyjmuje się, że struktura prozodyczna jest dużo mniej złożona, niż związana z nią struktura syntaktyczna wypowiedzi (jest bardziej płaska). Zatem w strukturze syntaktycznej muszą istnieć poziomy, które są mniej istotne dla prozodii. Pominięcie ich nie spowoduje istotnej degradacji konturu prozodycznego. Równoległe były rozwijane dwa podejścia do zadania wyznaczania granic prozodycznych:

- za pomocą gramatyk heurystycznych,
- za pomocą gramatyk wyprowadzanych automatycznie na podstawie odpowiednio opisanego tekstu (próby uczącej).

Ograniczamy się tutaj do bliższego omówienia po jednym tylko przykładzie dla każdego z wymienionych podejść. Szczególnie godną uwagi jest metoda automatyczna oparta na technice CART. Jest to unikalna metoda analityczna, która operuje na mieszanych zbiorach zmiennych niezależnych – jakościowych i ilościowych jednocześnie. Wynikiem jej działania jest zorganizowana w formie drzewa decyzyjnego lista współrzędnych płaskowyży funkcji prawdopodobieństwa wartości zmiennej zależnej w wielowymiarowej przestrzeni zmiennych niezależnych.

4.1. Heurystyki

Początkowo bazowano na przekonaniu, że prozodia może być organizowana wyłącznie na podstawie interpunkcji i rozkładów takich znaków jak nawiasy. Okazuje się to jednak niewystarczające. Nie wszystkie bowiem frazy prozodyczne ani nawet granice zdań są zaznaczone interpunkcją. Często są to spójniki albo w ogóle nic. Poza tym nie wszystkie znaki interpunkcyjne są tutaj istotne. W konsekwencji uznano, że podział zdania na frazy syntaktyczno-prozodyczne ma miejsce także na granicach słów leksykalnych i gramatycznych (funkcyjnych). Przyjmuje się, że te ostatnie rozgraniczają frazy. Rozwiązanie takie stosuje Larreur et al. (1989), Keller et al. (1993). Ogólniejsze podejście implementuje algorytm określany mianem *chinks'n'chunks*.

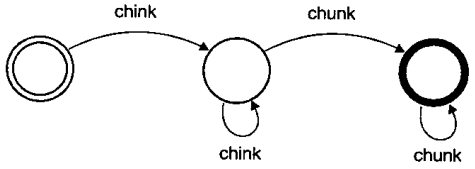
Algorytm chinks'n'chunks

Liberman i Church (1992) opisali bardzo prosty algorytm określany mianem *chinks'n'chunks*, w którym frazy prozodyczne są określane za pomocą reguły

a (minor) prosodic phrase = sequence of chinks followed by a sequence of chunks

((mała) fraza prozodyczna = sekwencja *chink* i następująca po niej sekwencja *chunk*).

Regułę tę reprezentuje automat (rys. 9)



Rys. 9. Automat reprezentujący regułę *chinks'n'chunks*

gdzie *chink* i *chunk* należą odpowiednio do zbiorów słów funkcyjnych (*chinks*) i znaczących (*chunks*) z wyjątkiem zaimków osobowych (takich jak *him, them*), które zaliczono do zbioru *chunks* oraz form czasowych czasowników (*tensed verb forms*, np. *produced*), które zaliczono do *chinks*. Takie podejście w większości przypadków funkcjonuje nieco lepiej niż prostsza dekompozycja na sekwencje słów funkcyjnych i głównych. Poniższy przykład pokazuje efekty dzielenia zdania na frazy prozodyczne.

Tabela 3. Porównanie granic przy zastosowaniu tylko słów funkcyjnych i algorytmu *chinks'n'chunks*

słowa funkcyjne/główne	chinks'n'chunks
I asked	*I asked them*
them if they were going home	*if they were going home*
to Idaho	*to Idaho*
and they said yes	*and they said yes*
and anticipated	*and anticipated one more stop*
one more stop	
before getting home	*before getting home*

I to rozwiązanie okazało się niewystarczające głównie dlatego, że – jak się okazało – frazy tworzą strukturę hierarchiczną, w której wyróżnia się frazy nadrzędne zawierające frazy podrzędne związane np. ze zdaniami wtrąconymi wymawianymi niższym tonem. Problem ten próbowano przezwyciężyć tworząc bardziej złożone heurystyki (Quene & Krager, 1989; Bailly, 1989; Frankenberger et al., 1994) i stosując gramatyki rozszerzane w kierunku gramatyk typu 1 i 0 w hierarchii Chomskiego (ATN[13]: w systemie MITALK, Allen et al., 1987; gramatyki ograniczające: Karlson, 1990; DCG i UG[14]: Traber, 1993; Lingström et al., 1993).

4.2. Metody automatyczne

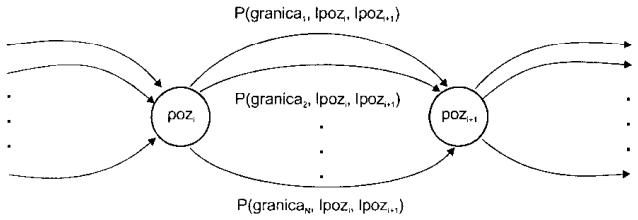
Ulepszanie heurystyk prowadzi w pewnym momencie do znacznych komplikacji reguł. Dalszy postęp staje się nieosiągalny dla eksperta. Nieodzowne do tego celu stają się duże teksty i silne reguły postępowania. Po wykonaniu parsingu syntaktycznego pozostaje jeszcze ustalenie granic fraz prozodycznych i ich

13 Augmented Transition Networks.
14 Definite Clause Grammars i Unification Grammars.

hierarchii (stosowne reguły postępowania można znaleźć np. w pracach Bachenko & Fitzpatrick, 1990; Traber,1993). Problem ten odpada w metodach automatycznych.

Pierwszym, prostym podejściem jest wyrażenie każdej sekwencji liter jako procesu Markova, w którym stany byłyby związane ze słowami, a łuki z a priori możliwymi granicami (Ljolje & Fallside 1986). W praktyce, ponieważ nawet bardzo duży korpus nie wystarczy aby wiarygodnie opracować prawdopodobieństwa pojawienia się granicy każdego typu pomiędzy każdą parą słów, słowa powinny być pogrupowane w klasy równoważności np. jako części mowy, co prowadzi do skojarzenia stanów z częściami mowy. Chociaż podejście to może być wysubtelnione poprzez zwiększenie kontekstu, to błąd nie może być pomniejszany w nieskończoność, ponieważ:

- Granice fraz prozodycznych nie mogą być ustalane niezależnie, bo długość frazy jest często przyjmowana jako wyznacznik położenia granicy, dlatego frazy prozodyczne mają w przybliżeniu tę samą długość (Bachenko, Fitzpatrick 1990).
- Proponuje się też, aby składowe granice syntaktyczne pokrywały się z granicami intonacyjnymi. Nie ma niestety odpowiedniości jeden do jeden pomiędzy lokalnymi sekwencjami części mowy i struktury syntaktycznej, do której one należą. To może być rozwiązane przez akceptowanie bardzo dużych sekwencji kontekstowych, ale oszacowanie odnośnych prawdopodobieństw stanie się szybko niemożliwe.
- Jest prawdopodobne, że lokalizacja granic prozodycznych zależy zarówno od długości sekwencji jak i od jej pozycji w wypowiedzi.



Rys. 10. Dozwolone przejścia pomiędzy dwoma stanami prostego łańcucha *Markova* dla obecności/nieobecności granic prozodycznych (wybranych z N możliwych) wewnątrz zdania (*sentence*). Słowa są widziane jako części mowy.

Streszczając, w strategii wyznaczania granic prozodycznych należy brać pod uwagę wiele wskaźników kontekstowych. Tak dużo, że trudno będzie określić

probabilistyczny automat skończenie stanowy uwzględniający je wszystkie równocześnie. Proces znajdowania drogi przypisywania granic prozodycznych może być widziany jako równoważny wyprowadzeniu wielopoziomowych reguł przepisywania (*MLRR*[15]).

Szczęśliwie zostały opracowane wysoce efektywne metody, które organizują heurystyki w drzewa decyzyjne otrzymywane techniką *klasyfikacji i regresji* (CART; Breiman et al., 1984; Bahl et al., 1989; Withgott et al., 1993) i umożliwiają wybór najistotniejszych czynników kontekstowych stosując algorytm zachłanny (taki, który podejmuje optymalną decyzję w każdym kroku nie zważając na kroki następne; drzewa jego są więc lokalnie optymalne).

CART-y pozwalają rozważać cechy zarówno kategorialne jak i ciągłe. W sposób naturalny dostarczają drzewiastej reprezentacji zbioru wyprowadzonych automatycznie MLRR-ów. Taka reprezentacja wyników pozwala na łatwą ich interpretację człowiekowi.

Zasadniczo CART-y opisują proces etykietowania za pomocą binarnych drzew decyzyjnych. W każdym węźle nieterminalnym stawiane jest pytanie wymagające odpowiedzi tak/nie o wartość wskaźnika kontekstowego i dla każdej odpowiedzi jest gałąź prowadząca do następnego pytania (rys. 1).

Pytania odnoszące się do wskaźników jakościowych dzielą ich wartości na dwa rozłączne podzbiory. Pytania odnoszące się do cech ciągłych dzielą zakres ich wartości na dwa podzakresy poprzez wstawienie punktu granicznego. Liście są związane z etykietą (tutaj z granicą prozodyczną). Aby zbudować takie drzewo, potrzeba zbioru trenującego złożonego z tokenów i związanych z nimi decyzjami. Z tego zbioru są wykorzystywane relacje pomiędzy cechami kontekstowymi do określenia predykatów dla drzewa decyzyjnego.

Kryterium rozszczepiania

Na początku wszystkie dane trenujące są przypisane pierwszemu węzłowi. Drzewo jest następnie budowane poprzez rekurencyjne rozszczepianie danych węzła rodzica na podzbiory, które formułują węzły potomne.

Każdy węzeł koduje rozkład danych uczących w danym kontekście (zbiorze kontekstów). Centralnym punktem CART-ów jest algorytm rozszczepiający węzły oparty na minimalizacji entropii danych uczących. Entropia jest miarą „losowości" danych:

$$H(L \mid node) = -\sum_{l \in L} P(l \mid node) \log_2 P(l \mid node),$$

15 MLRR (Multi Level Rewriting Rules), to transducer z warunkami formułowanymi na różnych poziomach abstrakcji (np. dla słów na poziomie części mowy):
$a \rightarrow b \mid l\_r \mid \text{poziom}_i : l_i\_r_i \mid \ldots \mid \text{poziom}_k : l_k\_r_k \mid$

gdzie L, to zbiór dopuszczalnych etykiet.

Ta wartość zmienia się od węzła do węzła, jako że na rozkład etykiet w węźle mają wpływ wszystkie wybory dokonane na wskaźnikach (cechach) kontekstowych począwszy od pierwszego węzła. Jeśli np. L zawiera 2^N równoprawdopodobnych etykiet na pozycji danego węzła, to formuła jw. redukuje się do

$$H(L \mid node) = -\log_2\left(2^{-N}\right) = N,$$

czyli do liczby bitów potrzebnych do zakodowania dowolnej etykiety w tym węźle. W przypadku nierównoprawdopodobnych etykiet $H(L)$ spada, co odpowiada redukcji „losowości" L i dowolna etykieta może być kodowana mniejszą liczbą bitów. Każde rozszczepienie w rodzicu – oparte na podziale c_i^j wartości czynnika kontekstowego c_i na dwa podzbiory C_1^j i C_2^j – wytwarza dwa węzły potomne, a wynikowa entropia średnia, to:

$$H(L \mid child_1, child_2) = H(L \mid child_1)P(child_1) + H(L \mid child_2)P(child_2),$$

gdzie $P(child_1)$ i $P(child_2)$ oznaczają prawdopodobieństwa „odwiedzin" węzłów potomnych, czyli prawdopodobieństwa, że c_i wpada do C_1^j i C_2^j odpowiednio przy danym węźle ojca[16]. Indeksy i oraz j odnoszą się do cech kontekstowych (i) i do podziałów ich wartości odpowiednio (j). Najlepszy podział $c_{i,best}^{j,best}$ to taki, który maksymalizuje różnicę pomiędzy entropiami przed i po rozszczepieniu. Ta różnica jest definiowana jako średnia wzajemna informacja $I\left(L, c_i^j\right)$ pomiędzy etykietą przewidywaną a (j-tą wartością) rozszczepiającą (i-tej cechy) c_i^j:

$$I\left(L, c_i^j\right) = H(L, parent) - H(L \mid child_1, child_2).$$

Może ona być otrzymana poprzez:

(1) zbadanie każdego czynnika kontekstowego c_i i znalezieniu podziału $c_i^{j,best}$, który maksymalizuje $I\left(L, c_i^j\right)$, i

16 Np. 2974/3379 dla lewego węzła pierwszego rozszczepienia na rys. 15.

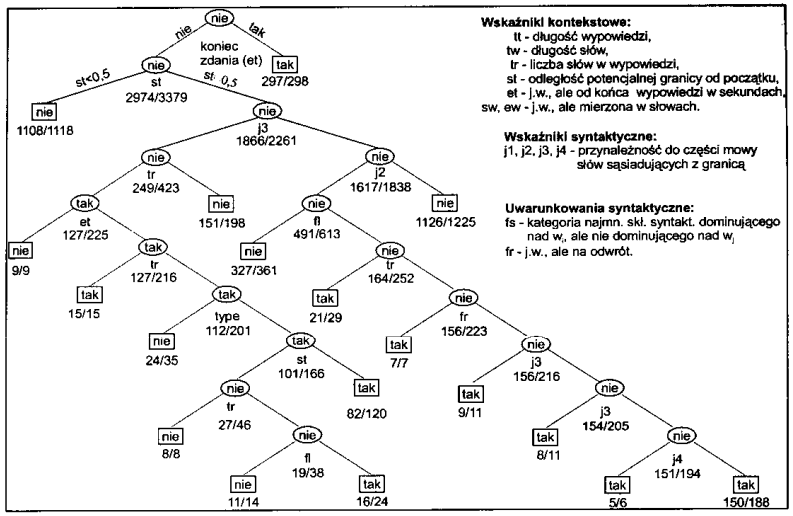
(2) następnie przez maksymalizację $I\left(L, c_i^{j,best}\right)$:

$$c_i^{j,best} = arg\ \max_j I\left(L, c_i^j\right),$$

$$c_{i,best}^{j,best} = arg\ \max_j I\left(L, c_i^{j,best}\right).$$

Algorytm rozszczepiający węzły jest stosowany iteracyjnie do węzłów potomnych i rozbudowywanie drzewa jest wstrzymywane, kiedy maksymalna średnia informacja wzajemna spada poniżej progu, kiedy to dalsza redukcja entropii jest uznawana za nieistotną. W praktyce entropie muszą być określone na podstawie danych uczących poprzez wyliczenie częstości względnych. Jest jasne, że to obciąża rozwój drzewa, jako że entropia skończonego zbioru próbek może być zawsze sprowadzona do dowolnie małej wartości poprzez powiększanie liczby liści. Stąd im niższy próg przerywający tym większe jest obciążenie. Na ogół takie drzewo decyzyjne będzie zdążać do osiągnięcia jak najwyższych wskaźników predykcji dla danej próby uczącej poprzez modelowanie wszystkich jej osobliwości – nastąpi przetrenowanie na niekorzyść zdolności do generalizacji. Węzły są więc rozszerzane wtedy tylko, kiedy liczba próbek, które one chwytają jest większa niż podany próg. Dane są dzielone na dwie części: trenującą i weryfikującą, celem sprawdzania dokładności spadku entropii określonej za pomocą danych trenujących. Kiedy estymata otrzymana z próbki weryfikującej spada poniżej danego progu, to rozbudowa drzewa jest zatrzymywana, nawet jeśli estymata otrzymana z danych trenujących sugeruje, że dalsze podziały dałyby istotną poprawę predykcji.

PRZYKŁAD. Techniki CART były z sukcesem stosowane do problemu syntaktyczno-prozodycznego frazowania w j. angielskim przez Wanga i Hirschberga (Wang & Hirschberg, 1991, 1992; Hirschberg, 1991). Ich korpus trenujący składał się z 298 zdań odręcznie zaetykietowanych etykietami wskazującymi położenie i typ granic fraz intonacyjnych (granicę główną, podrzędną lub zerową przypisano początkowo każdej parze sąsiadujących słów , później jednak zaniechano rozróżniania granic głównych i podrzędnych).



Rys. 11. Przykładowa budowa drzewa decyzyjnego w technice CART. Drzewo tak/nie jest przeznaczone dla przewidywania granic prozodycznych w tekście tylko na podstawie tekstu.

- Przy każdym węźle podano liczbę prawidłowo sklasyfikowanych prób.
- W każdym z nich wskazano tylko (najlepszy) czynnik kontekstowy; wartości dzielące zbiory wartości cech i wartości wynikłych podzbiorów za wyjątkiem dotyczących pierwszego rozszczepienia wg cechy *et* i wg cechy *st*, pominięto (Wang & Hirschberg, 1991).
- Wartości *tak|nie* w węzłach, to wartości etykiet (tutaj dwie: *jest_granica_frazy|nie_ma_jej*).
- Wybierane są najbardziej prawdopodobne wartości w podzbiorze reprezentowanym przez dany węzeł.

Uwagi

(1) „tak" może pojawić się w węźle nieterminalnym. Oznacza to tylko, że w różnych przypadkach znalezionych w próbie uczącej etykieta „tak" przeważała, co jednak nie musi skutkować wstawieniem granicy i przerwaniem dalszego badania warunków kontekstowych.

(2) Dane liczbowe, np. dla pierwszego rozszczepienia wg cechy st, należy czytać następująco: „Na 3379 napotkanych przypadków, dla których et = 'nie', znaleziono 2974 przypadki z etykietą 'nie ma granicy frazy prozodycznej'. Spośród owych 3379 przypadków 1118 spełniało warunek st<0,5. Wśród tych 1118 przypadków w 1108 znaleziono etykietę 'nie...'. Spośród rozważanych 3379 przypadków 2261 spełniało warunek st≥0,5, a wśród nich znalazło się 1866 przypadków również z etykietą 'nie...'. Zachodzi relacja 1118 + 2261 = 3379.

Wskaźniki kontekstowe zawierały informacje czasowe, takie jak[17]:

- długość wypowiedzi (*tt*),
- długość słów (*tw*),
- liczba słów w wypowiedzi (*tr*),
- odległość potencjalnej granicy od początku i od końca wypowiedzi mierzona w sekundach i w słowach (*st*, *et*, *sw*, *ew*).
- Brano pod uwagę także wskaźniki syntaktyczne zezwalając algorytmowi indukującemu drzewo sprawdzać części mowy w oknie obejmującym cztery słowa otaczające potencjalną granicę (*j1*, *j2*, *j3*, *j4*, to części mowy dla słów w_{i-1}, w_i, w_j,w_{j+1}) i
- uwarunkowania syntaktyczne; były to:
  - *fs* – kategoria najmniejszego składnika syntaktycznego dominującego nad w_i,
  - *fl* – kategoria największego składnika dominującego nad w_i, ale nie nad w_j,
  - *fr* – kategoria największego składnika dominującego nad w_j ale nie nad w_i.

Otrzymano drzewo, które poprawnie wykrywało 96% granic w tekście uczącym i 90% granic w tekście weryfikującym. Rozróżnienie fraz podrzędnych i nadrzędnych dawało drzewo rozpoznające poprawnie 82% granic w próbie weryfikującej.

17 Jak się okazało, cechy odnoszące się do końca wypowiedzi, których określenie wymaga znajomości całej wypowiedzi, okazały się nieistotne i nie pojawiły się w drzewie wynikowym.

5. Podsumowanie

Podsumowując, można powiedzieć o przedstawionych modelach intonacji co następuje:

Model *Fujisakiego* pozwala automatycznie wskazywać w akustycznym sygnale mowy sylaby akcentowane i granice fraz prozodycznych, co może mieć wielkie znaczenie dla procesu automatycznego rozpoznawania mowy. Ten aspekt ma mniejsze znaczenie dla syntezy mowy z tekstu, ponieważ dane takie – na potrzeby stworzenia bazy danych – można łatwo określić odręcznie na podstawie odsłuchów. Konfrontacja wyników obydwu tych podejść pozwoli natomiast określić zakresy zmian amplitud komend frazowych i akcentowych w powiązaniu z różnymi ich typami (akcent poboczny|główny, fraza nadrzędna|podrzędna).

Zastosowanie modelu *Fujisakiego* do przebiegów transformowanych do dziedziny percepcyjnej pozwoli uniezależnić te wyniki od czynników osobniczych.

Model *Fujisakiego* może być także zastosowany do obliczania bezwzględnych wartości *F0* w ostatnim etapie generowania syntetycznego sygnału mowy. Model ten, ze względu na własności uśredniające zastosowanego filtru, eliminuje zjawiska mikroprozodyczne w analizowanym sygnale.

Modele stosujące stylizację konturów również dostarczają danych istotnych dla kształtowania intonacji i akcentów. Również mogą być stosowane zarówno na etapie tworzenia bazy danych jak i na etapie generowania syntetycznego sygnału mowy. Nie umożliwiają jednak lokalizacji granic fraz prozodycznych i sylab akcentowanych.

W syntezie mowy z tekstu strukturę generowanej wypowiedzi określa się na podstawie analizy syntaktycznej tekstu. Powiązanie syntaktyki z prozodią nie jest jednoznaczne. Modele lingwistyczne – formułując reguły budowy grup intonacyjnych – pozwalają ograniczyć liczbę alternatywnych rozwiązań.

Z uwagi na jasno sprecyzowany model zmienności konturów tonu, model *Mertensa* może być stosowany na etapie obliczania wartości bezwzględnych częstotliwości podstawowej w syntezie mowy z tekstu.

Ogólnie o technikach postępowania można powiedzieć, że:

Ze względu na złożone uwarunkowania odręczne tworzenie reguł opisu intonacji na podstawie tekstu jest bardzo uciążliwe i nie pozwala uwzględnić wszystkich korelat dostarczanych przez tekst. Należy więc starać się stosować metody automatycznego kreowania reguł wiązania z tekstem zarówno cech intonacji jak i iloczasu. Największe możliwości dają tu techniki CART. Techniki te pozwalają nie tylko na łatwą eliminację cech nieskorelowanych, ale dają czytelne dla człowieka wyniki, przez co stwarzają możliwość wprowadzania odręcznych uogólnień przez eksperta.

Bibliografia

ALLEN J., HUNNICUT S. KLATT D. (1987). *From Text to Speech, the MITALK System.* Cambridge University Press, Cambridge.

ANDERSON M.D., PIERREHUMBERT J. (1984). *Synthesis by Rule of English Intonation Patterns.* Proceedings of the International Conference on Acoustics Speech and Signal Processing 84, pp. 2.8.1-2.8.4.

AUBERGE V. (1991). *La Synthèse de la Parole: des Règles aux Lexiques.* Ph.D. dissertation, Universitè Pierre Mendès France, Grenoble

BACHENKO J., FITZPATRICK E. (1990). *A Computational Grammar of Discourse-Neutral Prosodic Phrasing in English.* Computational Linguistics, n°16, September, pp. 155-167.

BAHL L. R., BROWN P. F., de SOUZA P. V., MERCER R. L., (1989). *A Tree-Based Statistical Language Model for Natural Speech Recognition.* IEEE Transaction on Acoustics, Speech, and Signal Processing, vol. 37, n°7, pp. 1001-1008.

BAILY G. (1989). *Integration of Rhytmic and Syntactic Constrains in a Model of Generation of French Prosody.* Speech Communication, vol. 8. pp. 137-146.

BALESTRI M., LAZZARETTO S., SALZA P. L., SANDRI R. (1993). *The CSELT System for Italian Text-to-Speech Synthesis.* Proceedings of Eurospeech 93, Berlin, pp. 2091-2094.

BARBOSA P. BAILY G. (1994). *Characterization of Rhytmic Patterns for Text-to-Speech Synthesis.* Speech Communication, vol. 15, n°1-2, pp. 127-139.

BARTKOVA K., SORIN C. (1987). *A Model of Segmental Duration for Speech Synthesis in French.* Speech Communication, vol. 16, , pp. 245-260.

BEAUGENDRE F. (1995). *Une Etude Perceptive de l'Intonation du Francais.* Ph.D. dissertation, LIMSI, Paris.

BÖHM A. (1992). *Maschinelle Sprachausgabe Deutschen und Englische Textes*" Ph. D. dissertation, Ruhr-Universität Bochum.

BREIMAN L., FRIEDMAN J.H., OLSHEN R.A., STONE C.J., (1984). *Classification and regresion* trees. Wadsworth & Brook, Monterey, CA.

BREIMAN L. (1996). *Born again trees.* ftp://ftp.stat.berkeley.edu/pub/users/breiman/Batrees.ps

BREIMAN L. (1996). *Out-of-bag estimation.* ftp://ftp.stat.berkeley.edu/pub/users/breiman/00Bestimation.ps.Z

BRUCE G., GRANDSTRÖM B. (1993). *Prosodic Modeling in Swedish Speech Synthesis.* Speech Communication, n°13, pp. 63-73.

CAMPBELL W. N., ISARD D. (1991). *Segment Durations in a Syllabe Frame.* Journal of Phonetics, 19, pp. 37-47.

COLIER R. (1991). *Multi-Language Intonation Synthesis.* Journal of Phonetics, vol. 19, pp. 61-73.

CRYSTAL D. (1969). *Prosodic Systems and Intonation of English.* Cambridge University Press, Cambridge

d'ALESSANDRO C., CASTELLENGO M. (1994). *The pitch of short-duration vibrato tones.* Journal of the Acoustical Society of America 95, 1617-1630.

d'ALLESANDRO C., MERTENS P. (1995). *Automatic Pitch Contour Stylization Using a Model of Tonal Perception.* Computer Speech and Language, 9, 3 pp. 257-288.

DEMENKO G. (1999). *Analiza cech suprasegmentalnych języka polskiego na potrzeby technologii mowy.* Wydawnictwo Naukowe UAM, Poznań.

FRANKENBERGER S., SCHNABEL B., ASSALI M., KOMMENDA M. (1994). *Prosodic Parsing Based on Parsing of Minimal Syntactic Structures.* Proceedings of the second ESCA/IEEE workshop on Speech Synthesis, New-Paltz, NY, pp. 143-146.

GEOFFROIS E. (1993). *A Pitch Contour Analysis Guided by Prosodic Events Detection.* Proceedings Eurospeech 93, Berlin, pp. 793-796.

HALLE M., KEYSER S. J. (1971). *English Stress: Its Form, its Growth, and its Role in Verse.* Harper and Row, New York.

HALLIDAY M. A. K. (1967). *Intonation and Grammar in British English.* Mouton, Berlin.

HIRAI T., IWAHASHI N., HIGUCHI N., SAGISAKA Y. (1994). *Automatic Extraction of F0 Control Parameters Using Statistical Analysis.* Proceedings of the Second ESCA/IEEE Workshop on Speech Synthesis, New-Paltz, NY, pp. 57-60.

HIROSE K., FUJISAKI H., KAWAI H. (1986). *Generation of Prosodic Symbols for Rule Synthesis of Connected Speech of Japanesse.* Proceedings of the International Conference on Acoustics Speech and Signal Processing 86, Tokyo, pp. 2415-2418.

HIRSCHBERG J. (1991). *Using Text Analysis to Predict Intonational Boundaries.* Proceedings of Eurospeech 91, Genova, pp. 1275-1278.

HIRST D.J., NICOLAS P., ESPESSER R. (1991). *Coding the F0 of a continuos text in French: an experimental approach.* Proceedings of the International Congress of Phonetic Sciences, Aix-en-Provence, 234-237.

HOUSE D. (1990). *Tonal Perception in Speech.* Lund University Press, Lund.

JASSEM W. (1984). *Isochrony in English speech,* in: Gibbon D. Richter H., eds. *Intonation, accent and rhythm.* de Gruyter, Berlin, 203-225.

KAIKI N., TAKEDA K., SAGISAKA Y. (1990). *Statistical Analysis for Segmentation Duration Rules in Japanese Text-to-Speech.* Proceedings of the International Conference on Spoken Language Processing 90, pp. 17-20.

KARLSSON K. (1990). *Constraint Grammars as a framework for Parsing Running Text.* Proceedings of the International Conference on Spoken Language Processing 90, pp. 17-20.

KELLER E., ZELLNER S., WERNER S., BLANCHOUD N. (1993). *The prediction of Prosodic Timing: Rules for Final Syllabe Lengthening in French.* Proceedings of the ESCA Workshop on Prosody, Lund, pp. 212-215.

KLATT D. H. (1976). *Linguistic use of Segmental Duration in English: Acoustic and Perceptual Evidence.* Journal of the Acoustical Society of America, vol. 59, pp. 1208-1221.

KLATT D. H. (1987). *Review of Text-to-Speech Conversion for English.* Journal of the Acoustical Society of America, vol. 82, 3, pp. 737-793.

LADD D. R. (1986). *IntonationalPhrasing: the Case for Recursive Prosodic Structure.* Phonology Yearbook 3, pp. 311-340.

LARREUR D., EMERARD F., MARTY F. (1989). *Linguistic and Prosodic Processing for a Text-to-Speech Synthesis System.* Proceedings of Eurospeech 89, Paris, pp. 510-513.

LEHISTE I. (1977). *Isochrony Reconsidered.* Journal of Phonetics, vol. 5, pp. 253-263.

LIBERMAN M. J., CHURCH K. W. (1992). *Text Analysis and Word Pronunciation in Text-to-Speech Synthesis,* in *Advances in Speech Signal Processing.* FURUI S., SONDHI M. M. eds., Dekker, New York, pp. 791-831.

LINDSTRÖM A., LJUNGQVIST M. GUSTAFSONN K. (1993). *A modular Architecture Supporting Multiple Hypotheses for Conversion of Text to Phonetic and Linguistic Entities.* Proceedings of Eurospeech 93, Berlin, pp. 1463-1466.

LJOLJE A. FALLSIDE F. (1986). *Synthesis of natural sounding pitch contours in isolated utterances using hidden Markov models.* IEEE Transactions on Acoustics, Speech and Signal Processing, 34, 1074-1079

LJUNGQVIST M., FUJISAKI H. (1993). *Generating Intonation for Swedish Text-to-Speech Conversion Using a Quantitative Model for the F0 Contour.* Proceedings of Eurospeech 93, Berlin, pp. 873-876.

MERTENS P. (1987). *Automatic segmentation of speech into syllables.* Proceedings of the European Conference on Speech Technology, Laver & Jack, Eds. Edinburgh: CEP Consultants, Vol. II, pp. 9-12.

MERTENS P. (1987). *L'intonation du français. De la dèscription linguistique: á la reconnaissance automatique.* Ph. D., Catholic University of Leuven.

MERTENS P. (1989). *Automatic Recognition of Intonation in French and Dutch.* Proceedings of Eurospeech 89, Paris, pp. 46-49.

MERTENS P. (1993). *Intonational Grouping, Boundaries, and Syntactic Structure in French.* Proceedings of the ESCA Workshop on Prosody, Lund, pp. 156-159.

MERTENS P.; GOLDMAN J. P.; WEHRLI E., GAUDINAT A. (2001). *La synthèse de l'intonation à partir de structures syntaxiques riches.* Traitement Automatique des Langues 42 (1), 142-195. http://bach.arts.kuleuven.ac.be/pmertens/

MERTENS P. (2002). *Synthesising elaborate intonation contours in text-to-speech for French.* Speech Prosody 2002, Aix-en-Provence, France, 11-13 April, 2002, ss. 4 http:/www.lpl.univ-aix.fr/sp2002

MEYER P., RÜHL H. W., KRÜGER R., KUGLER M. VOGTEN L. L. M., DIRKSEN A., BELHOULA K. (1993). *A Text-to-Speech Synthesier for the German Language.* Proceedings of Eurospeech 93, Berlin, pp. 877-890.

MÖBIUS B. PATZOLD M., HESS W. (1993). *Analysis of Synthesis of German F0 Contours by Means of Fujusaki's Model.* Speech Communication, vol. 13, pp. 53-61.

MONAGHAN A. I. C. (1990b). *Rhythm and Stress Shift.* Computer Speech and Language, n°4, pp. 71-78.

O'CONNOR J. D., ARNOLD G. F. (1973). *Intonation of Colloquial English, Longman.* New York

OHMAN S. (1967). *Word and sentence intonation: a Quantitative Model.* Quaterly Progress and Status Report, vol. 2, pp. 25-54.

O'SHAUGHNESSY D. (1984). *A Multispeaker Analysis of Durations in Read French Paragraphs.* Journal of the Acoustical Society of America, vol. 76, pp. 1664-1672.

O'SHAUGHNESSY D. (1987). *A Multispeaker Analysis of Duration in a Text-to-Speech System Using Only a Small Dictionary.* Proceedings of the International Conference on Acoustics, Speech, and Signal Processing 87, pp. 1430-1433.

PIERREHUMBERT J. (1980). *The Phonology and Phonetics of English Intonation.* PhD dissertation MIT, Indiana University Linguistics Club.

PIERREHUMBERT J. (1981). *Synthesizing Intonation.* Journal of the Acoustical Society of America, 70(4), pp. 985-9995.

PIERREHUMBERT J., Beckman M. (1988). *Japanese Tone Structure.* MIT Press, Cambridge, MA.

POLLACK I. (1968). *Detection of rate of change of auditory frequency.* Journal of Experimental Psychology 77, 535-541.

QUENE H., KRAGER R. (1989). *Automatic Accentuation and Prosodic Phrasing for Dutch Text-to-Speech Conversion.* Proceedings of Eurospeech 89, Paris, pp. 214-217.

RILEY M. D. (1992). *Tree-Based Modeling for Speech Synthesis*, in *Talking Machines: Theories, Models and Designs.* Baily G., Benoit C., eds., North Holland, pp. 256-273.

ROSSI M. (1971). *Le seuil de glissando ou seuil de perception des variations tonales pour les sons de la parole.* Phonetica 23, 1-33.

ROSSI M. (1978). *La perception des glissando descendans dans les contours prosodiques.* Phonetica 35, 11-40.

ROSSI M., PALMIERI F. CUTUNGO F. (2002). *A method for automatic extraction of Fujisaki-model parameters.* Prosody 2002, Aix-en-Provence, France, 11-13 April, 2002, ss. 4; http:/www.lpl.univ-aix.fr/sp2002

SAGISAKA Y. (1990). *On the prediction of Global F0 shape for Japanese Text-to-Speech.* Proceedings of the International Conference on Acoustics Speech and Signal Processing 90, pp. 325-328.

SAGISAKA Y. KAIKI N. (1992). *Optimization of Intonation Control Using Statistical F0 Resetting Characteristics.* Proceedings of the International Conference on Acoustics Speech and Signal Processing 92, vol. 2, pp. 49-52.

SCHEFFERS M. T. M. (1988). *Automatic stylization of F0 contours.* Proceedings of the Seventh FASE Symposium, Edinburgh, vol. 3, pp. 981-987.

SCHOUTEN H. E. M. (1985). *Identification and discrimination of sweep tones.* Perception and Psychopsychics 37, 369–376.

SEARGANT L, HARRIS J. D. (1962). *Sensitivity to unidirectional frequency modulation.* Journal of the Acoustical Society of America 34, pp. 1625-1628.

SILVERMAN K., BECKMAN N., PITRELLI J., OSTENDORF M., WHIGHTMAN C. PRICE P., PIERREHUMBERT J. HIRSCHBERG J. (1992). *ToBi: a standard for labelling English Prosody.* Proceedings of the International Conference on Spoken Language Processing, Alberta, pp. 867-870.

SLUIJTER A. M. C., SHATTUCK-HUFNAGEL S. STEVENS K. N., van HEUVEN V. (1995). *Supralaryngeal Resonance and Glottal Pulse Shape as Correlates of Stress and Accent in English.* Proceedings of the XIII International Congress on Phonetic Sciences, vol. 2, pp. 630-633.

t'HART J. (1976). *Psychoacoustic backgrounds of pitch contour stylization.* IPO—Annual Progress Report 11, Eindhoven, pp. 11-19.

t'HART J. (1991). *F0 stylization in Speech: Stright Lines versus Parabollas.* Journal of the Acoustical Society of America, 90(6), pp. 3368-3370.

t'HART J., COLLIER R., COHEN A. (1990). *A perceptual study of Intonation.* Cambridge University Press, Cambridge.

t'HART J., COLLIER R., COHEN A. (1991). *A perceptual Study of Intonation: an Experimental Phonetic Approach to Speech Melody.* Cambridge University Press, Cambridge.

TAYLOR P. (1995). *Using neural networks to locate pitch accents.* ESCA, Eurospeech'95, Proceedings, 1345-1348.

TERKEN J. (1993). *Synthesizing Natural Sounding Intonation for Dutch: Rules and Perceptual Evaluation.* Computer Speech and Language, vol. 7, pp. 27-48.

TRABER C. (1993). *Syntactic Processing and Prosody Control in the SVOX TTS System for German.* Berlin, vol. 3, pp. 2099-2102.

VAISSIERRE J. (1983). *Prosody: Models and Measurements* in: CUTLER A., LADD D. R., eds. Springer-Verlag, Berlin, pp. 53-66.

van SANTEN J. P. H. (1993). *Timing in Text-to-Speech Systems.* Proceedings of Eurospeech 93, Berlin, pp. 1397-1404.

WANG M. Q., HIRSCHBERG J. (1991). *Predicting Intonational Boundaries Automatically from Text: The ATIS Domain.* Proceedings of the DARPA Speech and Natural Language Workshop, pp. 378-383.

WANG M. Q., HIRSCHBERG J. (1992). *Automatic classification of intonational phrase boundaries.* Computer Speech and Language, 6, pp. 175-196.

WENK, B. J., WIOLAND F. (1982). *Is French Really Syllabe-Timed?* Journal of Phonetics, 10, pp. 193-216.

WITHGOTT M. M., CHEN F. R., (1993). *Computational Models of American Speech.* CSLI Lecture notes n°32, 143 pp.