INVESTIGATIONES LINGUISTICAE, VOL. XIII; POZNAŃ, DECEMBER 2006

Stylizacja tonów tajskich za pomocą Prosogramu

Stylisation of Thai tones using Prosogram

Marcin Włodarczak

Instytut Językoznawstwa, Uniwersytet im. Adama Mickiewicza
ul. Międzychodzka 5, 60-371 Poznań

Marcin.Wlodarczak@gazeta.pl

Abstract

The aim of this study is to establish whether stylisation of F0 contours based on d'Alessandro and Mertens's model of tonal perception can be successfully applied to lexical tones of Central Thai. The percentage of correct responses to the manipulated stimuli was found to be significantly lower than the results for natural tones reported in literature on the subject.

0. Wstęp

Niniejsza praca ma na celu ocenę przydatności zaproponowanej przez d'Alessandro i Mertensa metody stylizacji przebiegów częstotliwości podstawowej do opisu tonów leksykalnych centralnego dialektu języka tajskiego. Część 1. stanowi krótki przegląd najważniejszych metod opisu intonacji. Część 2. poświęcona jest wykorzystanemu tu modelowi d'Alessandro i Mertensa. W części 3. zebrane zostały podstawowe dane dotyczące percepcji tonów leksykalnych, wraz z opisem ich inwentarza w języku tajskim. Części 4. i 5. poświęcone zostały z kolei odpowiednio: opisowi przeprowadzonego eksperymentu oraz zestawieniu i omówieniu wyników.

1. Systemy stylizacji intonacji

Głównym fizycznym korelatem wysokości dźwięku jest częstotliwość podstawowa. Przebieg F_0 nie może być jednak uznany za jej w pełni adekwatną reprezentację, jako że zmiany częstotliwości podstawowej w mowie zależą od szeregu innych czynników, takich jak: wysokość własna samogłosek i spółgłosek, charakterystyka źródła dźwięku, siła fonacji (Pluciński 2003: 161). Czyni to wszelkie metody automatycznego czy też półautomatycznego opisu przebiegu częstotliwości podstawowej niezmiernie cennymi tak dla fonetyki deskryptywnej (automatyczna analiza i transkrypcja intonacji), jak i stosowanej (naturalnie brzmiąca mowa syntetyczna, systemy rozpoznawania mowy). Istniejące metody można podzielić na trzy grupy: anotację (ang. *labeling*), modelowanie i stylizacja (por. Fujisaki, Ohno, Wang 1998).

Pierwsza z nich polega na (ręcznej lub automatycznej) identyfikacji i symbolicznym zapisie kluczowych elementów konturu. Do systemów takich należą m.in. ToBI (Silverman *et al.* 1992), INTSINT (Hirst, Di Cristo 1998) oraz STEM-ML (Kochanski, Shih 2001). Ten ostatni stosowany

był do opisu tonów leksykalnych dialektów mandaryńskiego (Shih, Kochanski 2000) i kantońskiego (Lee, Kochanski, Shih, Li 2002) języka chińskiego.

Modelowanie opiera się na założeniu, że zmiany F_0 w mowie są uwarunkowane przez neuromotoryczne komendy, niosące znaczenie lingwistyczne i paralingwistyczne. Podejście to wykorzystano w modelu *Fujisakiego*. Przebieg F_0 jest tutaj sumą asymptotycznej wartości częstotliwości podstawowej (ang. *baseline value of fundamental frequency*), intonacji frazowej i struktury akcentowej, kształtowanych odpowiednio przez komendy frazowe i komendy akcentowe (Fujisaki, Ohno, Wang 1998: 1-2). System ten pozwala na oddzielenie od siebie obu rodzajów składowych (por. Pluciński 2003: 159). Stosowano go do opisu wielu języków[1], w tym języków tonalnych: chińskiego (Fujisaki, Ohno, Wang 1998; Mixdorff, Hu, Chen 2003), wietnamskiego (Dung, Mixdorff *et al.* 2004; Mixdorff 2003; Mixdorff, Hung *et al.* 2003) oraz tajskiego (Potisuk, Harper, Gandour 1995; Mixdorff, Luksaneeyanawin, Fujisaki, Charnvivit 2002; Mixdorff 2003; Mixdorff, Luksaneeyawin *et al.* 2003).

Trzecie z wymienionych podejść, stylizacja, polega na ekstrakcji elementów konturu istotnych z punktu widzenia komunikacji (Mertens 2005) czy też, w nieco ściślejszym ujęciu, zastąpieniu oryginalnego przebiegu F_0 prostszą funkcją liczbową, zachowującą jednak jego makroprozodyczną informację (Campione, Hirst, Veronis 2000). W zależności od tego, czy dana metoda uwzględnia charakterystykę percepcji, wyróżnić można stylizację akustyczną i percepcyjną (d'Alessandro, Mertens 1995)[2].

1.1. Stylizacja akustyczna

Stylizacja akustyczna opiera się na interpolacji punktów zwrotnych, będących miejscami, gdzie korelacja między kolejnymi wartościami F_0 a ich przybliżeniem spada poniżej wyznaczonej wartości, przy czym punkt zwrotny stanowi zarazem punkt docelowy poprzedzającego go fragmentu konturu. Możliwe jest przyjęcie apriorycznych parametrów funkcji interpolujących lub też aproksymacja za pomocą analizy regresji. Liczba punktów docelowych zależy od zastosowanej metody: wąskiej (dużo punktów docelowych) lub szerokiej (mało punktów docelowych) (por. Pluciński 2003: 159-161).

Kwestią sporną jest tu rodzaj interpolacji. I tak na przykład Hirst i Espesser kwestionują tezę 't Harta ('t Hart 1991), jakoby interpolacja liniowa, niebędąca zresztą według nich rozwiązaniem ekonomiczniejszym, była nieodróżnialna percepcyjne od aproksymacji za pomocą paraboli; nie podają jednak żadnych konkretnych argumentów na poparcie tego stwierdzenia[3] (Hirst, Espesser 1993). Uważają także, że aproksymacja za pomocą krzywych, jako dokładniejsza, pozwala w większym zakresie porównywać wyniki stylizacji z konturem oryginalnym. W tej samej pracy autorzy proponują system MOMEL (MOdélisation de MELodie) – stylizację opartą na aproksymacji za pomocą funkcji sklejanej drugiego stopnia. Skuteczność modelu oceniano poprzez porównywanie wizualne konturu oryginalnego i uzyskanego w wyniku stylizacji, obliczanie średniej odległości między nimi oraz nieformalne odsłuchy.

Innym interesującym podejściem jest wyznaczanie punktów zwrotnych na postawie dyskretnej transformaty falkowej (ang. *discrete wavelet transform*) (Wang, Narayanan 2005), w której sygnał zostaje rozłożony na dwie składowe: aproksymację (niskoczęstotliwościowe składowe sygnału) i detal (wysokoczęstotliwościowe składowe sygnału) za pomocą komplementarnych filtrów: dolno- i górnopasmowego. Procedura ta powtarzana jest dla składowej niskoczęstotliwościowej na każdym kolejnym poziomie dekompozycji, dając tzw. drzewo dekompozycji falkowej (por. Rak, Makowski 2006; Polikar 2001). W omawianej pracy zastosowano dekompozycję pięciopoziomową.

Główną wadą stylizacji akustycznej jest nieidentyfikowanie zmian niepercypowanych, jak również uśrednianie zdarzeń percypowanych oddzielnie (Pluciński 2003: 161).

---

[1] M.in. japońskiego, angielskiego, niemieckiego, greckiego, koreańskiego i hiszpańskiego.

[2] Terminy *stylizacja* i *modelowanie* będą odtąd używane zamiennie.

[3] „Since quadratic spline function gives a closer approximation to real F_0 curves than do straight lines, it is, in our opinion, **quite possible** that these differences will be appreciable **under certain circumstances**. Even though subjects may claim that they are unable to distinguish certain stimuli **it is well known that under certain circumstances** these stimuli may give rise to different reactions" (Hirst, Espesser 1993: 78; podkr. moje – M.W.).

1.2. Stylizacja percepcyjna

Postępowanie w opracowanym początkowo dla języka holenderskiego modelu IPO opiera się na założeniu, że przebieg częstotliwości podstawowej może być skutecznie aproksymowany przy pomocy zsynchronizowanych z nagłosem, standaryzowanych linii prostych. Te prototypowe odcinki stanowią podstawowe jednostki intonacyjne *danego* języka. W założeniu twórców sygnał oryginalny i zresytnetyzowany powinny być od siebie nieodróżnialne, w praktyce wymóg ten nie był jednak realizowany (por. Campione, Hirst, Veronis 2000). W systemie tym przyjmuje się milczące założenie dotyczące normalnego tempa mowy. Metoda nie wymaga uprzedniej segmentacji sygnału na sylaby, opiera się jednak jedynie na konturach F_0, nie uwzględniając mechanizmów percepcji intonacji (por. Pluciński 2003: 161-162).

2. Automatyczna stylizacja percepcyjna

W zaproponowanym przez d'Alessandro i Mertensa percepcyjnym modelu intonacji (d'Alessandro, Mertens 1995) zmiany wysokości głosu są aproksymowane za pomocą prostoliniowych segmentów tonalnych[4] wyznaczanych na podstawie progu *glissanda* (ang. *glissando threshold*) i różnicowego progu *glissanda (ang. differential glissando threshold)*. Wyrażany w półtonach na sekundę (ST/s) *próg glissanda* (G_{tr}) odpowiada najmniejszej dostrzegalnej zmianie F_0 i pozwala podzielić tony[5] na dynamiczne (wartość zmian F_0 przekracza próg *glissanda*) i statyczne (wartość zmian nie przekracza progu *glissanda*). Z kolei *różnicowy próg glissanda* to „najmniejsza dostrzegalna różnica w nachyleniu (zboczu) konturu konieczna do rozróżnienia dwóch kolejnych glissand" (Pluciński 2003: 162). W omawianej pracy różnicowy próg *glissanda* jest zdefiniowany jako różnica wyrażonych w półtonach na sekundę współczynników nachylenia segmentów tonalnych. Przyjmuje on wartości dodatnie dla krzywych wypukłych i wartości ujemne dla krzywych wklęsłych. Co więcej, jego wielkość jest wprost proporcjonalna do współczynnika nachylenia, niezależnie od jego znaku. Ustalono, że tak zdefiniowany różnicowy próg glissanda dla zmian wysokości głosu przyjmuje wartości z przedziału <12, 40>, brak jednak danych eksperymentalnych dotyczących jego dokładnej wartości.

Postępowanie w modelu sprowadza się do pięciu następujących kroków:

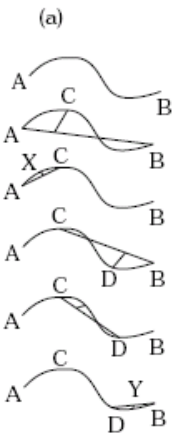
1 Określenie parametrów fizycznych (F_0, intensywność, detekcja fragmentów dźwięcznych i bezdźwięcznych, etc.)
2 Wstępna segmentacja sygnału akustycznego (oraz przebiegu F_0) na jednostki długości sylaby
3 Integracja percepcyjna chwilowych zmian F_0 lub wygładzanie przebiegu F_0 bazujące na odpowiednich własnościach percepcji słuchowej
4 Wtórna segmentacja przebiegu F_0 na segmenty tonalne i integracja zmian wysokości głosu zgodnie z progiem *glissanda* (detekcja statycznych lub dynamicznych segmentów tonalnych) i różnicowego progu *glissanda* (detekcja rosnących lub opadających segmentów tonalnych). Na tym etapie przetwarzania do segmentów tonalnych przypisywane są wartości docelowe.
5 Kategoryzacja segmentów tonalnych w ramach danego systemu językowego

Na szczególną uwagę zasługuje krok przedostatni, tj. właściwa stylizacja konturu F_0. Autorzy przyjmują założenie, że możliwe jest przybliżenie każdego przebiegu wysokości głosu za pomocą segmentów tonalnych, tak na poziomie sylaby, jak i całej wypowiedzi (d'Alessandro, Mertens 1995: 263). Aby jednak stylizacja taka była możliwa, konieczna jest segmentacja tonów złożonych (np. rosnąco-opadających). Polega ona na znalezieniu w przebiegu punktów zwrotnych „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" (Pluciński 2003:163). Analiza

[4] *Segment tonalny* odpowiada fragmentowi wypowiedzi, dla którego „the perceived pitch shows a uniform slope (either level, rising or falling)" (d'Alessandro, Mertens 1995: 263).

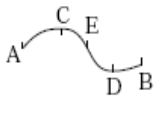
[5] Autorzy używają terminu *tone* w znaczeniu „the pitch object perceived for a stretch of speech corresponding to a phonetic syllable" (d'Alessandro, Mertens 1995: 263).

powtarzana jest rekursywnie do momentu, gdy współczynnik *glissanda*[6] całego segmentu zawartego w oknie analizy jest mniejszy od progu *glissanda* lub gdy różnica w miejscu potencjalnej granicy segmentu tonalnego spada poniżej 1 półtonu. Procedurę tę przedstawia schematycznie Ryc. 1.



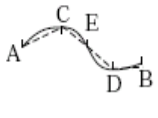
Analysis interval	Maximun difference at	Segmentation	New turning point
–	–	–	–
A-B	C	yes	C
A-C	X	no, too small	–
C-B	D	yes	D
C-D	E	yes	E
D-B	Y	no, not audible	–

Rycina 1: Segmentacja tonów złożonych poprzez wyznaczanie punktów zwrotnych (d'Alessandro, Mertens 1995:271)



Compare parts	*g*2-*g*1 > DGT	Decision
AC and CE	yes	keep AC
CE and ED	no	merge CE and ED
(CE+ED) and DB	yes	keep CD
DB	–	keep DB

Merge step



(b)

Tonal segments	*g* > GT	slope
AC	yes	Dynamic
CD	yes	Dynamic
DB	no	static

Stylization step

Rycina 2: Łączenie potencjalnych segmentów tonalnych w oparciu o różnicowy próg glissanda *oraz interpolacja liniowa punktów zwrotnych (d'Alessandro, Mertens 1995:271)*

Każdy z segmentów jest potencjalny, ponieważ dwa sąsiadujące z sobą segmenty mogą zostać połączone, jeśli różnica ich współczynników nachylenia spadnie poniżej wartości różnicowego progu *glissanda*. Ostatnim etapem jest interpolacja liniowa punktów zwrotnych. Procedurę tę ilustruje Ryc. 2.

Aby ocenić skuteczność modelu autorzy przeprowadzili test percepcyjny, w którym słuchaczom zaprezentowano pary sygnałów złożone z naturalnych wypowiedzi francuskich oraz tych samych wypowiedzi poddanych stylizacji przy różnych wartościach progu *glissanda* i różnicowego progu *glissanda*, a następnie zresyntetyzowanych za pomocą metody TD-PSOLA

[6] Współczynnik glissanda to szybkość zmian częstotliwości podstawowej, wyrażona w ST/s (d'Alessandro, Mertens 1995: 264).

(Time Domain Pitch Synchronous Overlapp-Add)[7]. W przypadku każdej pary zadaniem słuchaczy było podjęcie decyzji, czy oba sygnały są jednakowe, przy czym nie byli oni proszeni o zwracanie szczególnej uwagi na intonacje. Wypowiedzi stylizowane przy progu *glissanda* równym 0,16 ST/s i różnicowym progu *glissanda* równym 0,20 ST/s okazały się nieodróżnialne od wypowiedzi naturalnych w 67,72% przypadków[8]. Jak jednak zauważają autorzy: „Many subjects reported that they could distinguish signals on the basis of changes in some aspects of sound quality rather than on the basis of intonation" (d'Alessandro, Mertens 1995: 282), wynik ten można więc uznać za wysoki. System ten stosowano także do modelowania konturów intonacyjnych języka koreańskiego (Ratajszczak 2005). W tym wypadku posłużono się procedurą dyskryminacji A B X, zaś grupa odsłuchowa składała się z dwóch dwunastoosobowych zespołów: pierwszy stanowili rodzimi mówcy języka koreańskiego, drugi – rodzimi mówcy języka polskiego nieznający języka koreańskiego. Materiał badawczy składał się z naturalnych wypowiedzi koreańskich[9], wypowiedzi poddanych stylizacji przy progu *glissanda* równym 0,16 ST/s oraz wypowiedzi, w których dokonano wyraźnych modyfikacji w zakresie przebiegu F_0[10]. Wbrew oczekiwaniom, słuchacze koreańscy z wysoką skutecznością identyfikowali wypowiedzi naturalne oraz poddane stylizacji (76,85% poprawnych odpowiedzi); dużo większe trudności mieli z tym natomiast słuchacze polscy (55,56% poprawnych odpowiedzi).

3. Tony leksykalne języka tajskiego

Niniejsza praca stawia sobie za cel ocenę przydatności tej metody w badaniach tonów leksykalnych języka tajskiego. Wydaje się to możliwe, gdyż, po pierwsze, model ten abstrahuje od cech konkretnych języków[11] oraz, po drugie, opiera się na własnościach układu percepcyjnego człowieka i jako taki powinien poprawnie opisywać wszelkie zjawiska makroprozodyczne oparte na zmianach częstotliwości podstawowej.

Język tajski, oficjalny język Królestwa Tajlandii, posiada 5 dystynktywnych tonów leksykalnych (Abramson 1962: 9), w językoznawstwie zachodnim określanych tradycyjnymi nazwami: ton średni (ang. *mid tone*), ton niski (ang. *low tone*), ton opadający (ang. *falling tone*), ton wysoki (ang. *high tone*) i ton rosnący (ang. *rising tone*)[12]. Tony opadający i rosnący klasyfikowane są jako konturowe, pozostałe zaś – jako rejestrowe.

Kluczowym dla niniejszej pracy problemem jest pytanie, czy sama częstotliwość podstawowa umożliwia poprawną identyfikację tonów tego języka. Badania takie zostały przeprowadzone przez Abramsona (Abramson 1975). W swoich testach użył on syntetycznej sylaby [kʰaː] (ze stałą amplitudą), na którą nałożył pochodzące z wcześniejszej pracy (Abramson 1962) uśrednione kontury, tu przedstawione na Ryc. 3.

[7] W rzeczywistości, aby zapewnić jednakową jakość dźwięku, zresyntetyzowano także sygnał niepoddany uprzednio stylizacji.

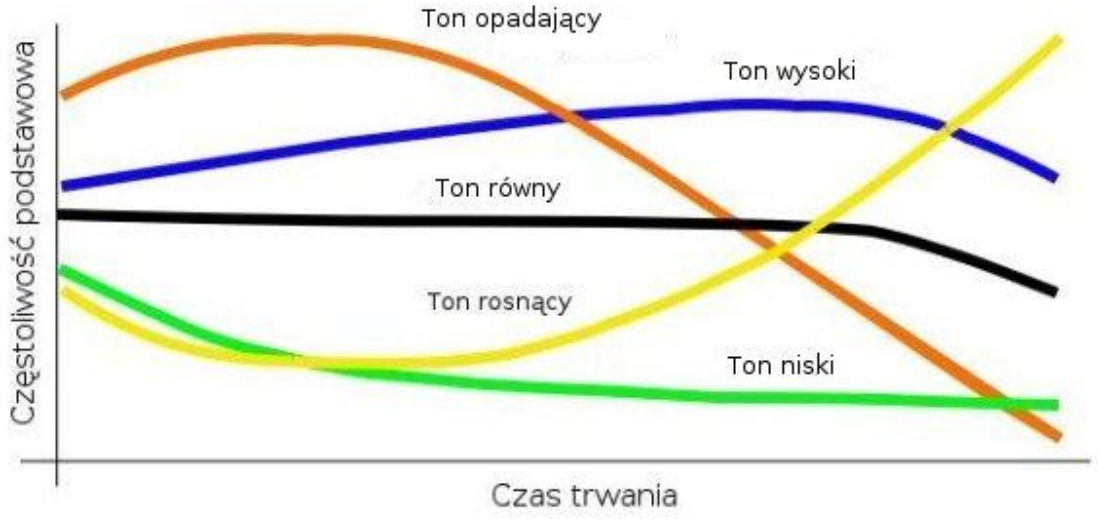
[8] Dla porównania, pary nieróżniących się do siebie, naturalnych sygnałów uznano za jednakowe w 89,76% przypadków. Dla większości słuchaczy hipoteza zerowa stwierdzająca brak różnic między tymi wynikami została odrzucona na poziomie istotności równym 0,01.

[9] Inaczej niż w teście d'Alessandro i Mertensa, wypowiedzi naturalne nie zostały zresyntetyzowane.

[10] Wprowadzenie ostatniego rodzaju wypowiedzi miało zapobiegać sytuacjom, w których „słuchacze *na siłę* szukaliby różnic w prezentowanych wypowiedziach lub uznaliby w trakcie testu, że wszystkie 3 bodźce jednej kolejki A, B oraz X brzmią tak samo" (Ratajszczak 2005: 40).

[11] Należy tu jednak przytoczyć uwagę autorów modelu, którzy zastrzegli, że „The model was tested for one language only, i.e. contemporary French as spoken in France; and this could be seen as a limitation. It is clear that segmental and suprasegmental properties of French may favour a certain approach which could be less successful for other languages. For instance, syllabic decomposition is an important feature of French (when compared to English, say), and the set of possible pitch movements is rather limited (again when compared to English)" (d'Alessandro, Mertens 1995: 286).

[12] W niniejszej pracy będą one oznaczane odpowiednio jako: t1, t2, t3, t4 i t5.



Rycina 3: Uśrednione kontury tonów języka tajskiego (na podstawie Abramson 1962: 126)

Uzyskany wynik (92,8% poprawnych odpowiedzi w porównaniu z 98,6% dla sygnałów naturalnych) wskazuje, że F_0 istotnie stanowi wystarczającą wskazówkę dla natywnych mówców języka tajskiego. Wynik ten uległ dalszej poprawie (wzrósł do 96,1% poprawnych odpowiedzi), gdy do użytych w poprzednim doświadczeniu sylab dodano charakterystyczne dla danego tonu zmiany amplitudy. Sam Abramson tłumaczył to faktem, że „changes in the contraction of certain laryngeal muscles and in subglottal air pressure can separately or together produce variations in the fundamental frequency of the voice. These mechanisms are also available for controlling intensity of phonation and thus variations in the overall amplitude of the speech signal. To a certain extent, then, the two acoustic features, F_0 and amplitude, may co-vary" (Abramson 1975: 5), wydaje się jednak, że ustalenie dokładnego związku między przebiegiem F0 a zmianami amplitudy oraz wpływu tych ostatnich na percepcję tonów wymagałoby oddzielnych badań.

Pewną rolę w poprawnej identyfikacji tonów odgrywa również informacja o charakterystycznym dla danego mówcy interwale tonalnym. Dotyczy to szczególnie tonów średniego i równego, wykazujących najmniejszą zmienność częstotliwości podstawowej, przy czym błędna identyfikacja tonu równego jako niskiego zdarza się częściej niż sytuacja odwrotna. Bowiem chociaż oba te tony charakteryzują się spadkiem częstotliwości podstawowej (por. Ryc. 3), jest on mniej gwałtowny dla tonu średniego i w pewnych przypadkach „the downdrift of the mid tone [...] may be enough to make some listeners uncertain and cause them to assign the only possible other choice, namely the low tone" (Abramson 1976: 9).

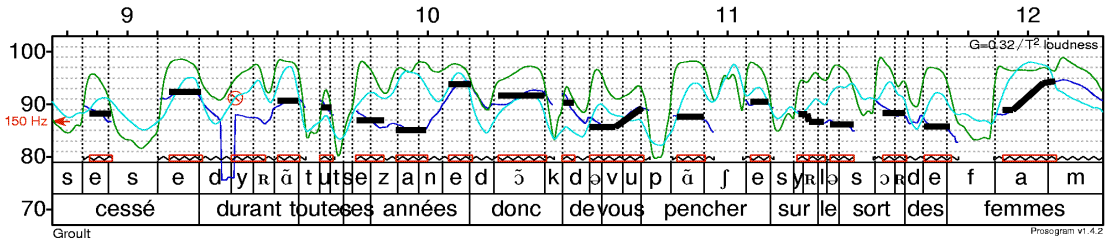
Co ważne w kontekście niniejszej pracy, ten sam badacz w innym miejscu (Abramson 1975) sprawdzał rozpoznawalność tonów tajskich po zastąpieniu konturów oryginalnych konturami o przebiegu prostoliniowym. Do testów użyto syntetycznych sylab, na które nałożono 16 płaskich konturów tonalnych z zakresu 92-152 Hz różniących się każdorazowo o 4 Hz. Sylaby te zostały w niemal 100% zidentyfikowane jako tony rejestrowe (98% odpowiedzi): ton wysoki rozpoznawano najczęściej przy wartości F_0 równej 152 Hz (87,7 % odpowiedzi), ton średni – przy 116 Hz (73% odpowiedzi), a ton niski – przy 92 Hz (90,1% odpowiedzi). W przypadku żadnej sylaby nie było więc pełnej zgodności wśród słuchaczy – nawet sylaby uzyskujące najwyższe wyniki w danej kategorii przypisywano do pozostałych kategorii rejestrowych.

W pracy *The Thai Tonal Space* (Abramson 1997) autor przedstawił wyniki trzech analogicznych doświadczeń z konturami prostoliniowymi, tym razem jednak wartość F_0 nie była stała, ale zmieniała się w obrębie sylaby. W pierwszym teście użyto 16 konturów o stałej wartości początkowej wynoszącej 106 Hz i wartościach końcowych z przedziału od 90 do 152 Hz, tak jak w poprzednim eksperymencie różniących się każdorazowo o 4 Hz. Zgodnie z przewidywaniami,

procent identyfikacji tonu średniego okazał się nieznaczny (maksymalnie 39% odpowiedzi dla wartości końcowej równej 106 Hz). Potwierdziło się także przypuszczenie, że dolne wartości docelowe są zbyt niskie, a spadki zbyt powolne dla tonu opadającego – kategoria ta w ogóle nie pojawiła się wśród odpowiedzi. Z drugiej jednak strony, najwyższe wartości końcowe okazały się wystarczające, aby wywołać u słuchaczy wrażenie tonu rosnącego (maksymalnie 64% odpowiedzi, 2 sylaby uzyskały wynik powyżej 50%), rezultat ten jest jednak niższy niż w przypadku tonów wysokiego (7 sylab z wynikami wyższymi od 50%) i niskiego (5 sylab powyżej 50% i maksimum wynoszące 90% odpowiedzi). Drugi eksperyment różnił się od pierwszego częstotliwością początkową, wynoszącą tu 90 Hz, wartości końcowe nie uległy zmianie. Wyniki potwierdziły trzy z czterech hipotez: (1) Wartość początkowa jest zbyt niska dla tonu średniego (maksymalnie 10% odpowiedzi, dla wyższych wartości docelowych liczba odpowiedzi spadła do 0); (2) Silniejszy wzrost wartości F_0 spowodował w porównaniu z poprzednim testem większą liczbę identyfikacji tonu rosnącego; (3) Dwie dolne wartości końcowe były identyfikowane przez uczestników głównie jako ton niski, jednak nieco wyższy wynik w teście poprzednim zdaje się sugerować, że nieznaczny spadek wzmaga percepcyjną wyrazistość tego tonu. Wbrew oczekiwaniom, jeden z konturów został rozpoznany jako ton wysoki aż w 40% przypadków. W ostatnim eksperymencie ustalono stałą wartość końcową (152 Hz), zaś wartość początkową zmieniano w zakresie 90-152 Hz o 4 Hz. Jak się spodziewano, słuchacze rozpoznali jedynie tony wysoki i rosnący, przy czym częściej identyfikowanym tonem był ton wysoki (wystąpiła także zaniedbywana ilość rozpoznań tonu niskiego).

4. Eksperyment

W niniejszej pracy wyniki wyżej wymienionych badań posłużyły za podstawę porównań dla rozpoznawalności tonów, których kontury poddano stylizacji percepcyjnej w Prosogramie – implementacji opisanego w punkcie 2. modelu d'Alessandro i Mertensa (Mertens 2004). Prosogram[13] jest makropoleceniem Praata (Boersma, Weenink 2006), darmowego programu do analizy mowy. Przykładowy prozogram przedstawia Ryc. 4.

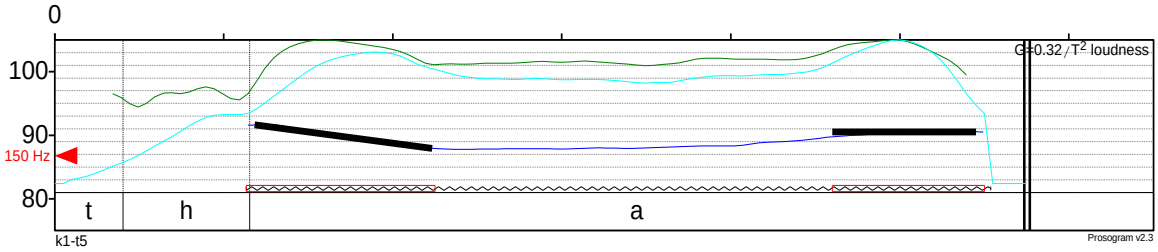


Rycina 4: Prozogram wypowiedzi francuskiej „Cessé durant toute ses années donc de vous pencher sur le sort des femmes" (segmentacja automatyczna). Linie niebieskie oznaczają oryginalny kontur F_0, czarne – kontur uzyskany w wyniku stylizacji percepcyjnej, zielone – intensywność dźwięku, a linie seledynowe – głośność dźwięku (Mertens, 2005: http://bach.arts.kuleuven.be/pmertens/prosogram/)

Do testu użyto sylabę [tʰaː] w pięciu wariantach tonalnych wymówioną przez 8 natywnych użytkowników języka tajskiego (dialektu centralnego): 4 kobiety i 4 mężczyzn[14]. Początkowo planowano zastosować automatyczną metodę detekcji jąder sylabicznych na podstawie różnic głośności, okazało się to jednak niemożliwe, jako że powodowała ona błędy. W licznych przypadkach (szczególnie dla tonu rosnącego) program znajdował dwa maksima głośności w obrębie jednej samogłoski i dokonywał osobnej stylizacji każdego z tak wyznaczonych segmentów, pomijając przy tym fragment środkowy (Ryc. 5). Możliwe jest zapewne usunięcie tego

[13] Korzystano z wersji 2.30.
[14] Nagrania zostały wykonane przez dr. Janusza Kleśtę (Instytut Językoznawstwa, UAM).

błędu poprzez obniżenie odpowiedniej wartości progowej w kodzie źródłowym. W tej sytuacji zdecydowano się zastosować segmentację ręczną; do tego celu także wykorzystano Praata. Jeśli dana sylaba zawierała segment, w obrębie którego występowały aperiodyczne drgania fałdów głosowych (ang. *creaky voice*), odcinka tego nie obejmowano stylizacją[15]. W przypadku każdej z 40 sylab (8 mówców, 5 wariantów tonalnych) przeprowadzono stylizację przy czterech różnych wartościach progu *glissanda* (0.16 ST/s, 0,24 ST/s, 0,32 ST/s i 0,40 ST/s), a następnie zresyntetyzowano je (PSOLA) za pomocą Praata. W efekcie uzyskano 160 sygnałów[16], które następnie zrandomizowano i, aby zapobiec znużeniu słuchaczy, podzielono na 5 partii po 32 sylaby każda. Odstęp między sygnałami w każdej grupie ustalono na 4 sekundy, zaś między partiami następowała dłuższa, kilkunastosekundowa przerwa.



Rycina 5: Błędna detekcja jąder sylabicznych przy wyborze metody automatycznej

Grupę odsłuchową stanowiło troje natywnych mówców centralnego dialektu języka tajskiego (1 mężczyzna i 2 kobiety), obecnych lub byłych lektorów tego języka w Instytucie Językoznawstwa UAM. Badanie przeprowadzono w cichym pomieszczeniu z wykorzystaniem komputera PC wyposażonego w standardowe głośniki. Zadaniem słuchaczy było zaznaczenie rozpoznanego przez nich tonu na karcie odpowiedzi poprzez zakreślenie odpowiedniego znaku tonalnego. W celu uniknięcia nieporozumień związanych z różnymi sposobami notacji tonów zastosowano tradycyjne znaki używane w ortografii tajskiej; ton średni, który nie ma swojego ortograficznego symbolu, oznaczono jako półpauzę.

5. Wyniki

Ogółem poprawnie rozpoznanych zostało 71,25% tonów (342 z 480 odpowiedzi). Należy jednak pamiętać, że jest to wynik zbiorczy dla wszystkich czterech wartości progu *glissanda*. Rozpatrywany pod tym względem przedstawia się następująco (Tabela 1):

Tabela 1: Procent poprawnych odpowiedzi dla poszczególnych wartości progu glissanda

WARTOŚĆ PROGU *GLISSANDA*	PROCENT POPRAWNYCH ODPOWIEDZI	LICZBA ODPOWIEDZI
0,16	69,17% (83)	120
0,24	73,33% (88)	120
0,32	71,67 % (86)	120
0,40	70,83% (85)	120

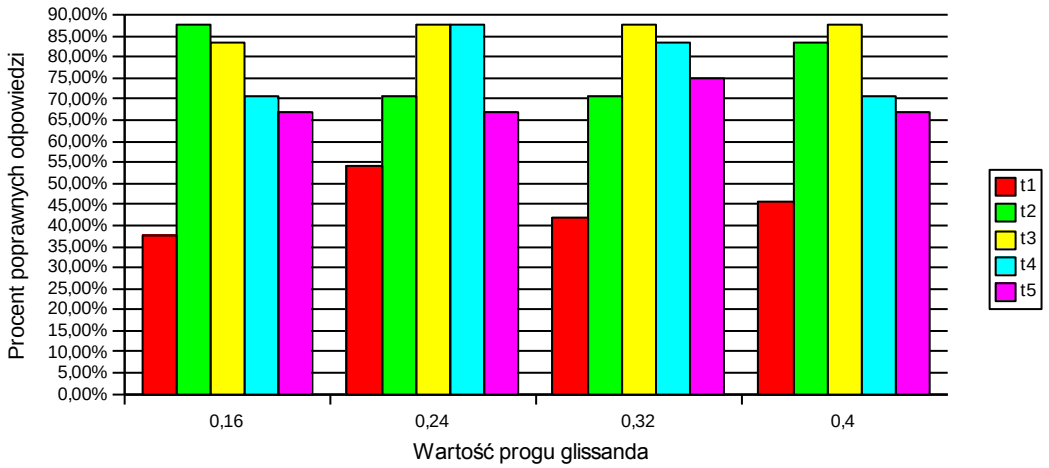
[15] Nagłe zmiany F_0 występujące w takich przypadkach prowadziły do generowania przez program błędów, natomiast wygładzanie konturów dawało wysoce nienaturalne efekty.

[16] Zrezygnowano z przeprowadzenia dodatkowego testu z sylabami naturalnymi, przyjmując, że rodzimi użytkownicy języka tajskiego są w stanie identyfikować tony tego języka z poprawnością bliską 100%. Hipotezę tę potwierdzają przytaczane wyżej wyniki, jak również, pośrednio, funkcja pełniona przez tony leksykalne w systemie językowym.

Wyniki te, bardzo zbliżone do wyniku ogólnego, zdają się sugerować, że wartość *progu glissanda* nie ma znaczącego wpływu na poprawną identyfikację tonów. W celu zweryfikowania tej hipotezy przeprowadzono test χ^2 na niezależność. Wartość empiryczna $\chi^2 = 0{,}53$ okazała się niższa od wartości krytycznej $\chi^2_{\alpha;\,df}$, która przy 3 stopniach swobody i poziomie istotności równym 0,05 wyniosła 7,82, nie było więc podstaw do odrzucenia hipotezy zerowej, stwierdzającej niezależność tych dwóch zmiennych. Uzyskane rezultaty są jednak wyraźnie niższe niż w testach z sylabami naturalnymi (98,6% poprawnych odpowiedzi), jak i z konturami uśrednionymi (92,8%) (Abramson 1957).

Najlepiej identyfikowanym tonem okazał się ton opadający – 83 poprawne odpowiedzi (86,46%). Jednakowy wynik uzyskały tony wysoki i niski – po 75 poprawnych odpowiedzi (78,13%). Trzeci rezultat uzyskał ton rosnący z 66 poprawnymi odpowiedziami (68,75%). Tylko 43 razy rozpoznano poprawnie ton równy (44,79%). (Zob. także Ryc. 7).

Wyniki zrelatywizowane do wartości progu *glissanda* przedstawia Ryc. 6.



Rycina 6: Procent poprawnej identyfikacji poszczególnych tonów w zależności od wartości progu glissanda

Jak widać, wyniki te co prawda różnią się znacznie od siebie (np. w przypadku tonu niskiego różnica dla dwóch pierwszych wartości progu *glissanda* wyniosła aż 16,67%), nie zaobserwowano jednak systematycznego spadku liczby poprawnych odpowiedzi wraz ze wzrostem wartości progu *glissanda*.

W przypadku trzech z czterech wartości progu *glissanda* potwierdziła się dominacja tonu opadającego (maksymalnie 87,5%). Rezultat ten jest jednak niższy od przytaczanych przez Abramsona wyników dla sylab naturalnych (99,1%) oraz „idealnych" konturów (97,8%)[17].

Jedynie przy progu *glissanda* równym 0,16 ST/s najlepszy wynik (także 87,5%) należy do tonu niskiego. Jest to zarazem jedyny przypadek w całym teście zbliżania się do wyników eksperymentu z konturami „idealnymi" (87,3% dla tonu niskiego); z drugiej jednak strony ton ten, obok tonu niskiego, był w badaniach Abramsona identyfikowany o wiele słabiej od pozostałych. Także przy najwyższej wartości progu *glissanda* ton niski był rozpoznawany z dużą poprawnością (83,33%), przy czym wyniki te nie znalazły odzwierciedlenia w pozostałych przypadkach.

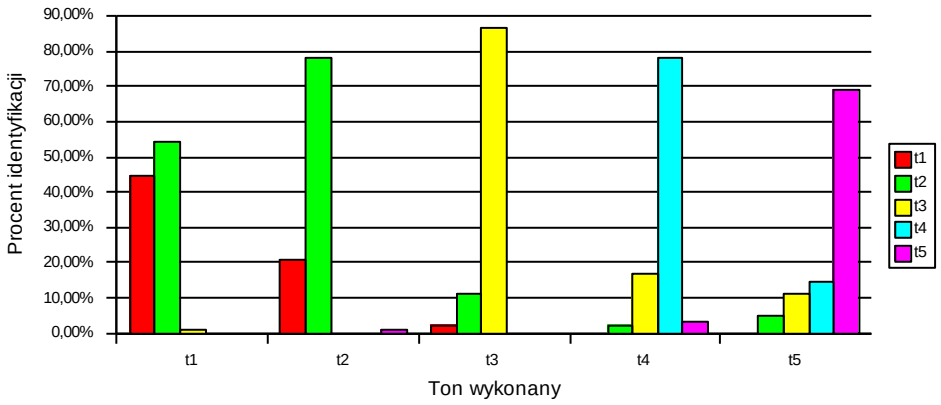
Ton równy osiągnął najniższy wynik w całym teście (37,50% przy progu *glissanda* równym 0,16 ST/s) oraz w poszczególnych kategoriach. Co ciekawe, jego najwyższy wynik, równy 54,15% dla progu *glissanda* równego 0,24 ST/s, jest niższy niż w teście, w którym jako materiał odsłuchowy posłużyły sylaby ze stałą wartością częstotliwości podstawowej (maksymalnie 73% poprawnych odpowiedzi dla $F_0 = 116$ Hz).

[17] Pomijamy tu rzecz jasna wynik uzyskany przy stałej częstotliwości podstawowej, który dla tego tonu wyniósł maksymalnie 0,3% poprawnych odpowiedzi przy F_0 równej 100 oraz 148 Hz.

Ton wysoki był wprawdzie identyfikowany z dość wysoką skutecznością przy środkowych wartościach progu *glissanda*, jednak dla wartości skrajnych procent poprawnej identyfikacji wyniósł jedynie 70,83%. Najwyższy wynik, 87,50% dla G_{tr} = 0,24 ST/s, jest niemal identyczny jak w przypadku sylab o stałej wartości F_0 równej 152 Hz (87,7%). Co więcej, sylaby z rosnącymi liniowo konturami F_0 były rozpoznawane lepiej, bo maksymalnie aż w 97% (F_0 początkowe równe 133 Hz i stałe F_0 końcowe równe 152 Hz). W pozostałych testach Abramsona maksymalny procent poprawnej identyfikacji tego tonu wyniósł odpowiednio: 97,7% („kontur idealne"), 75% (stałe F_0 początkowe równe 106 Hz, F_0 końcowe równe 124 Hz) oraz 40% (stała wartość początkowa równa 90 Hz i wartości końcowe z przedziału 100-120 Hz). Słaba rozpoznawalność w dwóch ostatnich przypadkach jest zresztą zrozumiała, jako że zarówno 106 Hz, jak i 90 Hz są zbyt niskimi wartościami początkowymi dla tonu wysokiego.

Dziwić może także dość niski wynik dla tonu rosnącego. Ton ten należy zwykle do najlepiej identyfikowanych nawet w przypadku mówców nienatywnych (Fedak 2005), tutaj natomiast procent poprawnej identyfikacji wyniósł 75% dla G_{tr} = 0,32 ST/s i 66,67% dla pozostałych wartości progu *glissanda*. Oprócz wyniku dla wartości początkowej F_0 równej 106 Hz (maksymalnie 64% przy końcowej wartości F_0 wynoszącej 152 Hz), rezultat ten był więc znacznie niższy od uzyskanych w pozostałych eksperymentach przeprowadzonych przez Abramsona (maksymalnie po 90% poprawnych odpowiedzi przy stałej częstotliwości początkowej równej 90 Hz i wartości końcowej równej 148 Hz oraz przy częstotliwości początkowej równej 90 Hz i stałej częstotliwości końcowej wynoszącej 152 Hz; 99,1% w przypadku „idealnych konturów")[18].

Ryc. 7 przedstawia rozkład odpowiedzi dla każdego z tonów. Są to wyniki zbiorcze dla wszystkich wartości progu *glissanda*.



Rycina 7: Rozkład odpowiedzi dla poszczególnych tonów (wyniki zbiorcze dla wszystkich wartości progu glissanda*)*

W wynikach tych znajduje odzwierciedlenie ogólna tendencja zaobserwowana wśród rodzimych mówców języka tajskiego, polegająca na myleniu tonu średniego z niskim i na odwrót (Abramson 1976, Shapiro [rok wydania nieznany]). Daje się także dostrzec wyraźną przewagę pomyłek pierwszego rodzaju, co pozostaje w zgodzie z przytaczaną powyżej uwagą Abramsona. (Abramson 1976: 9). Pomyłki te są wręcz o 9,38% częstsze od rozpoznań poprawnych. Jest to szczególnie widoczne u słuchaczki K2, która udzieliła jedynie 7 poprawnych odpowiedzi na 32 wystąpienia tonu równego w całym teście (21,88%), przy czym poprawna identyfikacja nie była w tym przypadku uzależniona od wartości progu *glissanda* (5 z błędnie rozpoznanych sylab zostało poddanych stylizacji przy G_{tr} = 0,24 ST/s, 6 – przy G_{tr} = 0,32 ST/s, 7 – przy G_{tr} = 0,40 ST/s i 7 – przy G_{tr} = 0,16 ST/s). Należy także zauważyć, że sylaby te były poprawnie identyfikowane przez pozostałych słuchaczy, którzy wykazali zresztą w kategorii tonu niskiego niemal całkowitą zgodność co do swoich odpowiedzi – jedynym wyjątkiem był błąd popełniony przez słuchacza M1,

[18] Podobnie jak w przypadku tonu opadającego abstrahujemy tu od wyniku dla stałej częstotliwości podstawowej (maksymalnie 0,1% poprawnych odpowiedzi przy F_0 = 100 Hz).

który rozpoznał ton opadający. Poprawność wyniosła w ich przypadku 56,25% i także nie była uzależniona od wartości progu *glissanda* (po 8 błędnie rozpoznanych sylab dla G_{tr} = 0,16 ST/s i G_{tr}= 0,32 ST/s oraz po 6 przypadków dla G_{tr} = 0,24 ST/s i G_{tr} = 0,40 ST/s). Zgodność wśród wszystkich słuchaczy wyniosła zaledwie 21,88% (7 z 32 sylab).

Ton niski – poza jednym wyjątkiem, kiedy zidentyfikowano go jako ton rosnący – był mylony z tonem średnim (po 7 błędów popełniono przy progu *glissanda* równym 0,24 i 0,32 ST/s, 4 błędy przy 0,40 ST/s i 2 błędy przy 0,16 ST/s). W tym przypadku słuchaczka K2 uzyskała z kolei wynik najlepszy, z jedynie dwoma błędnymi odpowiedziami. Zgodność wśród wszystkich słuchaczy wyniosła dla tego tonu 62,5% (20 sylab na 32); zaś dla słuchaczy M1 i K1 – 84,36% (27 sylab).

Ton opadający był przeważnie mylony z tonem niskim, były to jednak błędy stosunkowo rzadkie (2 błędy przy G_{tr} = 0,40 ST/s oraz po 3 pomyłki przy pozostałych wartościach progu *glissanda*), co znalazło odbicie w ogólnym wyniku tego tonu. Słuchacze jednakowo rozpoznali 30 z 32 sylab (93,75%).

O wiele wyraźniejsza jest błędna identyfikacja tonu wysokiego jako opadającego, szczególnie dla skrajnych wartości progu *glissanda.* Trzeba jednak zauważyć, że w 15 przypadkach na 16 błędy te popełnił słuchacz M1. Zgodność wśród wszystkich słuchaczy wyniosła więc jedynie 37,5% (12 z 32 sylab), słuchaczki K1 i K2 jednakowo zidentyfikowały aż 87,5% (28) sygnałów.

Ton rosnący mylono natomiast głównie z tonem opadającym i wysokim oraz, w mniejszym stopniu, z tonem niskim. Za pomyłki pierwszego rodzaju odpowiada ponownie słuchacz M1 (9 na 11 razy). Należy przy tym podkreślić, że zarówno w tym, jak i w poprzednim przypadku wszystkie udzielone przez niego błędne odpowiedzi pojawiły się w drugiej części testu (dla tonu wysokiego począwszy od 80. sygnału, dla tonu rosnącego – od sygnału 116.), można by więc przypisać je znużeniu. Z drugiej jednak strony analogiczne zjawisko nie wystąpiło u tej osoby w przypadku trzech pozostałych tonów, trudno zatem uznać je za reprezentatywne. Z kolei mylenie tonu rosnącego z wysokim jest najsilniej dostrzegalne u słuchaczki K2 (10 na 14 przypadków), nie zależy jednak od wartości progu *glissanda* (po 3 pomyłki przy G_{tr} = 0,24 ST/s i G_{tr} = 0,40 ST/s oraz po 2 przy G_{tr} = 0,16 ST/s i G_{tr} = 0,32 ST/s). Słuchacze byli zgodni w przypadku 13 z 32 sylab (40,63%).

Ciekawe wyniki przynosi porównanie powyższych danych z błędami popełnianymi przez słuchaczy w przeprowadzonych przez Abramsona testach z sylabami naturalnymi i wyidealizowanymi konturami. Główną różnicą między nimi, oczywiście oprócz większej ilości pomyłek w drugim przypadku, jest większe rozproszenie odpowiedzi. I tak, tony średni i niski są identyfikowane jako tony: średni, niski, opadający i wysoki. Tony opadający i rosnący są mylone ze wszystkimi pozostałymi, a ton wysoki ze wszystkimi oprócz niskiego. Pod tym względem wyniki uzyskane w niniejszej pracy są bardziej zbliżone do wyników dla sygnałów niepoddanych manipulacji. Może to świadczyć o tym, że zresyntetyzowane sylaby niosą wskazówki percepcyjne skutecznie zapobiegające myleniu z sobą pewnych tonów (np. niskiego z opadającym). Z drugiej jednak strony, te same wskazówki znacznie utrudniają odróżnianie od siebie innych tonów (niskiego i średniego, opadającego i niskiego). Innymi słowy, wywołane przez stylizację przesunięcia w przestrzeni percepcyjnej mogły spowodować zwiększenie dystansu między pewnymi tonami i jednocześnie jego zmniejszenie między tonami innymi. Zgodność ta może być jednak również spowodowana niewielkim rozmiarem grupy odsłuchowej.

Tak skonstruowany test nie pozwala oczywiście na bezpośrednie wnioskowanie o naturalności zresyntetyzowanych sygnałów. Biorąc jednak pod uwagę fakt, że natywni mówcy języka tajskiego są w stanie identyfikować naturalne tony tego języka z niemal stuprocentową poprawnością, wynik osiągnięty w przeprowadzonym przez nas badaniu może świadczyć o tym, że brzmienie użytych w nim sylab było dalekie od naturalności. Do podobnych wniosków doszli słuchacze po zakończeniu testu. Przypuszczali oni nawet, że być może mówcy używają innego niż centralny dialektu lub nawet że nie są w ogóle natywnymi użytkownikami języka tajskiego.

6. Podsumowanie

Testy percepcyjne przeprowadzone z udziałem tak niewielkiej grupy odsłuchowej nie mogą rzecz jasna rościć sobie prawa do reprezentatywności. Niniejsza praca powinna być więc uważana jedynie za przyczynek do badań zakrojonych na większą skalę. Uzyskane tu wyniki wskazują jednoznacznie, że stylizacja tonów tajskich za pomocą Prosogramu nie przynosi oczekiwanych rezultatów. Procent poprawnej identyfikacji zresyntetyzowanych sylab okazał się nie tylko niższy od wyników uzyskanych dla sylab naturalnych i uśrednionych konturów F_0, ale, co szczególnie istotne, w większości także dla sylab syntetycznych z prostoliniowym przebiegiem częstotliwości podstawowej, niezależnie od przyjętej wartości progu *glissanda*. Wśród niektórych słuchaczy zaobserwowano silne tendencje do mylenia pewnych tonów, z których przynajmniej część może okazać się reprezentatywna w przypadku większej grupy odsłuchowej.

Bibliografia:

Abramson, A. S. 1962. The vowels and tones of standard Thai: Acoustical measurements and experiments. w: *Indiana University Research Center in Anthropology, Folklore and Linguistics*. Bloomington. http://www.haskins.yale.edu/Reprints/HL0035.pdf [data dostępu: październik, 2006]

Abramson, A.S. 1975. The tones of central Thai: Some perceptual experiments. w: J.G. Harris, J.R. Chamberlain (Eds.) *Studies in Tai Linguistics. In honor of William J. Geddney*. Bankok: Central Institute of English Language. http://www.haskins.yale.edu/Reprints/HL0191a.pdf [data dostępu: październik, 2006]

Abramson, A. S. 1976. Thai tones as a reference system. w: T.W. Gething, J.G. Harris, P. Kullavanijaya (Eds.) *Thai linguistics in honor of Fang-Kuei Li*. Bangkok: Chulalongkorn University Press. http://www.haskins.yale.edu/Reprints/HL0215.pdf [data dostępu: październik, 2006]

Abramson, A.S. 1997. The Thai tonal space. w: A.S. Abramson (Ed.) *Southeast Asian Linguistic Studies in honour of Vichin Panupong*. Bangkok: Chulalongkorn University Press. http://www.haskins.yale.edu/Reprints/HL1074.pdf [data dostępu: październik, 2006]

D'Alessandro, C., Mertens, P. 1995. Automatic pitch contour stylization using a model of tonal perception, *Computer Speech and Language*, 9(3), 257-288. http://bach.arts.kuleuven.be/pmertens/papers/csl1995.pdf [data dostępu: wrzesień, 2006]

Boersma, P., Weenink. D. 2006. *Praat: doing phonetics by computer* (Wersja: 4.4.30) [Program komputerowy]. http://www.praat.org [data dostępu: wrzesień, 2006]

Campione, E., Hirst, D., Veronis, J., Automatic Stylization and Symbolic Coding of F_0: Implementations of the INTSINT Model. w: A. Botinis (Ed.) *Intonation. Research and Applications*. Dordrecht: Kluwer. http://www.up.univ-mrs.fr/veronis/pdf/2000Campione.pdf [data dostępu: październik, 2006]

Dung, T.N., Mixdorff, H. *et al*. 2004. Fujisaki Model based F0 contours in Vietnamese TTS. w: *Proceedings of ICSLP2004*. Jeju. http://www.tfh-berlin.de/~mixdorff/thesis/files/dung_mixdorff_icslp2004.pdf [data dostępu: grudzień, 2006]

Fedak, A. 2006. *Percepcja tonów tajskich przez polskich słuchaczy*. Poznań. (Nieopublikowana praca magisterska napisana pod kierunkiem prof. dr hab. P. Łobacz).

Fujisaki, H., Ohno S., Wang, C. 1998. A command-response model for F_0 contour generation in multilingual speech synthesis, w: *Proceedings of the 3rd ESCA/COCOSDA International Workshop on Speech Synthesis*. Jenolan Caves. 299-304. http://www.slt.atr.co.jp/cocosda/jenolan/Proc/r51/r51.pdf [data dostępu: grudzień, 2006]

't Hart, J., 1991. F_0 stylization in speech: straight lines versus parabolas. *Journal of the Acoustical Society of America*, 6, 3368-3370.

Hirst, D., Espesser, R. 1993. Automatic modeling of fundamental frequency using a quadratic spline function, *Travaux de l'Institut de Phonétique d'Aix*, 15, 71-85. http://aune.lpl.univ-aix.fr/ ~hirst/articles/1993%20Hirst&Espesser.pdf [data dostępu: październik 2006]

Hirst, D., Di Cristo, A. 1998. A survey of intonation systems. w: D. Hirst, A. Di Cristo (Eds.) *Intonation Systems: A Survey of Twenty Languages*. Cambridge: Cambridge University Press. http://aune.lpl.univ-aix.fr/~hirst/articles/1998%20Hirst&DiCristo.pdf [data dostępu: listopad, 2006]

Lee, T., Kochanski, G., Shih, C., Li, Y. 2002. Modeling tones in continous Cantonese speech. w: *ICSLP 2002*. Denver. http://prosodies.org/papers/2002/stemml-cantonese.pdf [data dostępu: listopad, 2006]

Mertens, P. 2004. The Prosogram: Semi-Automatic Transcription of Prosody based on a Tonal Perception Model. w: B. Bel, I. Marlien (Eds.) *Proceedings of Speech Prosody 2004*. Nara. <http://bach.arts.kuleuven.be/pmertens/papers/sp2004.pdf> [data dostępu: grudzień 2006]

Mertens, P. 2005. *The Prosogram*. http://bach.arts.kuleuven.be/pmertens/prosogram [data dostępu: wrzesień 2006]

Mixdorff, H., Luksaneeyanawin, S., Fujisaki, H., Charnvivit, P. 2002. Perception of Tone and Vowel Quantity in Thai. w: *Proceedings of ICSLP2002*. Denver. http://www.tfh-berlin.de/~mixdorff/thesis/files/mixdorff_luksaneeyanawin_icslp2002.pdf [data dostępu: grudzień, 2006]

Mixdorff, H. 2003. Modeling Prosody in a Cross-language Perspective. w: *SASRTL Workshop*, Szczyrk. http://www.tfh-berlin.de/~mixdorff/thesis/files/mixdorff_sasrtl2003.pdf [data dostępu: grudzień, 2006]

Mixdorff, H., Hu, Y. and Chen, G. 2003. Towards the Automatic Extraction of Fujisaki Model Parameters for Mandarin. w: *Proceedings of Eurospeech 2003*. Geneva. http://www.tfh-berlin.de/~mixdorff/thesis/files/mixdorff_fujisaki_eurosp2003.pdf [data dostępu: grudzień, 2006]

Mixdorff, H., Hung, N. et al. 2003. Quantitative Analysis and Synthesis of Syllabic Tones in Vietnamese. w: *Proceedings of Eurospeech 2003*. Geneva. http://www.tfh-berlin.de/~mixdorff/thesis/files/mixdorff_bach_eurosp2003.pdf [data dostępu: grudzień, 2006]

Mixdorff, H., Luksaneeyawin, S. *et al*. 2003. Modeling Rhythmic Variation in Thai and its Application to Speech Synthesis. w: *Proceedings of ICPhS2003*. Barcelona. http://www.tfh-berlin.de/~mixdorff/thesis/files/mixdorff_luksaneeyanawin_icphs2003.pdf [data dostępu: grudzień, 2006]

Pluciński, A. 2003. Modelowanie zmian prozodycznych na potrzeby syntezy mowy. *Scripta Neophilologica Posnaniensia*, Tom V, 153-191.

Polikar, R. 2001. *The Wavelet Tutorial*. http://users.rowan.edu/~polikar/WAVELETS/WTtutorial.html [data dostępu: grudzień, 2006]

Potisuk, S., Harper, M., Gandour, J.T. 1999. The Classification of Thai Tones in Connected Speech using the Analysis by Synthesis Method, *IEEE Transactions on Speech and Audio Processing*, Vol. 7, No. 1, 91-102. ftp://ftp.ecn.purdue.edu/harper/papers/toneieee.pdf [data dostępu: listopad, 2006]

Rak, R.J., Majkowski, A. 2006. *Analiza czasowo-częstotliwościowa sygnałów.* <http://wazniak.mimuw.edu.pl/index.php?title=Laboratorium_wirtualne_1/Modu%C5%82_5_-_%C4%87wiczenie_5> [data dostępu: grudzień, 2006]

Ratajszczak, G. 2005. *Testowanie systemu do półautomatycznej analizy intonacji.* Poznań. (Nieopublikowana praca magisterska napisana pod kierunkiem prof. dr hab. P. Łobacz).

Shapiro, L. [brak roku wydania] *Perception of Thai tones by naive and native listeners of Thai*. http://grove.ufl.edu/~linclub/focus/shapiro.pdf [data dostępu: październik, 2006]

Shih, C., Kochanski, G. 2000. Chinese Tone Modeling with Stem-ML. w: *ICSLP 2000.* Pekin. http://www.prosodies.org/tutorial2002/papers/01232.pdf [data dostępu: listopad, 2006]

Shih, C., Kochanski, G. 2003. Prosody Modeling with Soft Templates. *Speech Communication*, 39, 3-4, 311-352. http://prosodies.org/papers/SpeechComm1_2001.pdf [data dostępu: listopad, 2006]

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., Hirschberg, J. 1992. ToBI: a standard for labelling English prosody. w: *Proceedings of ICSLP'92*. Banff. 867-870.

Wang, D. Narayanan, S. 2005. Piecewise Linear Stylization of Pitch Via Wavelet Analysis. w: *INTERSPEECH 2005*. Lisbon. 3277-3280. http://sail.usc.edu/publications/dagen_shri_euro_final.pdf [data dostępu: listopad, 2006]