Ryszard Gubrynowicz, dr hab.
Polsko-Japońska Wyższa Szkoła
Technik Komputerowych
ul. Koszykowa 86
02-008 Warszawa

Warszawa, 22 listopada 2009 r.

Recenzja pracy doktorskiej mgr Krzysztofa Szklannego
pt. „ *Optymalizacja funkcji kosztu w korpusowej syntezie mowy polskiej*"

1. Uwagi wstępne

Ogólnie wiadomo, że mowa jest jedną z najbardziej efektywnych form przekazywania informacji od i do człowieka. Stąd jest coraz szerzej stosowana w systemach komunikacji człowiek – komputer, a zwłaszcza w obecnie intensywnie rozwijanych systemach dialogowych. W przypadku bardziej rozbudowanych systemów, informacja jest przekazywana użytkownikowi w postaci mowy syntetycznej. Warto, podkreślić, że mowa w wielu przypadkach znakomicie uzupełnia inne formy komunikacji (przede wszystkim wizualnej), a w niektórych sytuacjach może być wyłącznym sposobem komunikowania, tak jak to ma miejsce w przypadku porozumiewania się z osobami niewidzącymi. Rozwój syntezatorów trwający nieprzerwanie od kilkudziesięciu lat doprowadził, że generacja zrozumiałej mowy syntetycznej nie jest już obecnie problemem, natomiast wciąż istotnym zagadnieniem jest generacja mowy syntetycznej o wysokiej naturalności brzmienia, będącym istotnym warunkiem powszechnej akceptacji przez użytkowników tego typu systemów.

Bogactwo informacji niesionych przez akustyczny sygnał mowy, ich wzajemne nakładanie się, utrudnia w znacznym stopniu jednoznaczne określenie, które parametry i cechy fizyczne tego sygnału decydują jednoznacznie o naturalności brzmienia mowy syntetycznej. Jednym ze sposobów ominięcia tej trudności jest bezpośrednie wykorzystanie i łączenie ze sobą odpowiednio wybranych segmentów uprzednio zarejestrowanego sygnału mowy naturalnej, stanowiącego na ogół zbiór właściwie dobranych wypowiedzi, nagranych w określony sposób przez odpowiednią osobę. Ta technika syntezy mowy jest podstawą wszystkich metod nazywanych korpusowymi. Jednym z istotnych zagadnień w tych metodach jest odpowiednia konstrukcja funkcji kosztu, która obok właściwie dobranej bazy danych, jest jednym z kluczowych problemów w praktycznej realizacji systemu syntezy korpusowej. Mało jest publikacji na ten temat (zwłaszcza dla języka polskiego) i zwykle stanowi ona pilnie strzeżoną tajemnicę ośrodków zajmujących się projektowaniem syntezatorów mowy. W publikacjach na ten temat podawane są na ogół tylko zdawkowe informacje.

2. Temat i zakres rozprawy

Problematyka badawcza opiniowanej rozprawy jest przede wszystkim związana z optymalizacją funkcji kosztu w korpusowej syntezie mowy dla języka polskiego. Funkcja ta składa się z dwóch części: kosztu doboru jednostki oraz kosztu konkatenacji. Koszt doboru jednostki jest funkcją, która optymalizuje fragmenty łączonej mowy poprzez wybieranie najbardziej pasujących do siebie pod względem lingwistycznym. Natomiast funkcja kosztu konkatenacji determinuje jakość tego połączenia na podstawie analizy iloczasów jednostek akustycznych tworzących łączone fragmenty, ich intonacji, konturu widma oraz energii.

W celu przebadania wpływu optymalizacji funkcji kosztu na jakość mowy syntetycznej mgr Krzysztof Szklanny przygotował kompletny system syntezy korpusowej. Proces ten

obejmował niezwykle pracochłonny etap przygotowania korpusu, realizację odpowiednich nagrań, segmentację i transkrypcję stworzonej przez siebie bazy językowej. Funkcję kosztu autor rozprawy zoptymalizował za pomocą metody heurystycznej, przez zastosowanie algorytmu ewolucyjnego z tak zwaną strategią (μ+λ). Jak wykazały badania percepcyjne, przeprowadzony proces optymalizacji funkcji kosztu znacznie poprawił jakość syntetycznej mowy, generowanej w środowisku Festival opracowanym przez uniwersytet w Edynburgu.

Istotnym osiągnięciem pracy autora nad optymalizacją funkcji kosztu było określenie w sposób obiektywny istniejących relacji między poszczególnymi parametrami funkcji kosztu. Wykazał, że do najbardziej istotnych, mających wpływ na jakość generowanej mowy w języku polskim, należą takie parametry jak: pozycja głoski w sylabie, ciągłość melodii wypowiedzi, prawy kontekst (głoska następująca), ciągłość energii oraz akcent. Skuteczność wykonanych badań nad optymalizacją funkcji kosztu autor rozprawy zweryfikował przez przeprowadzenie subiektywnego testu MOS (Mean Opinion Score), w którym uczestniczyły osoby mające doświadczenie w ocenie jakości sygnałów audio. Ponadto, w pracy zbadano wpływ wyboru mówcy oraz sposobu tworzenia bazy akustycznej na jakość generowanej mowy syntetycznej.

3. Ogólna charakterystyka pracy

Praca składa się z 7 rozdziałów (rozdział zatytułowany „Wprowadzenie" bez numeracji, z oddzielną paginacją stron), w sumie o objętości 154 stron i dwóch załączników o objętości 4 stron. Wyniki pracy są rzeczowo udokumentowane i przedstawione w postaci 46 rysunków oraz 26 tabel. Cytowanych jest 125 pozycji bibliograficznych. Do pracy załączono płytę DVD z tekstem rozprawy, z przykładami mowy syntetycznej difonowej (system MBROLA) i korpusowej (system Multisyn). Ponadto, w dwa komplety procedur syntezy difonowej i korpusowej przeznaczone odpowiednio do środowiska UNIX i Windows. Wiele z tych procedur zostało opracowanych przez autora rozprawy.

W rozdziale wprowadzającym przedstawiono ogólną tematykę rozprawy, zakres przeprowadzonych badań oraz jej główne tezy.

W Rozdziale 1, zatytułowanym *„Sygnał mowy i jego opis fonetyczny"* przedstawiono podstawy działania narządu artykulacyjnego oraz ogólne zasady opisu fonetycznego dźwięków mowy polskiej opierając się zarówno na ich opisie artykulacyjnym, jak i akustycznym. Ponadto, omówiono zjawisko koartykulacji oraz metody opisu symbolicznego melodii mowy. Szczegółowo omówiono podstawowe segmenty akustycznego sygnału mowy stosowane przy konkatenacyjnej syntezie mowy, zwłaszcza difony oraz trifony. Szkoda, że przy tym nie dokonano próby scharakteryzowania poszczególnych dźwięków mowy polskiej pod kątem trudności jakie mogą wystąpić w procesie segmentacji potoku mowy na difony, powodując tym samym niejednoznaczną ich postać. W takich przypadkach, wydaje się że raczej korzystniejsze byłoby stosowanie trifonów.

W Rozdziale 2, zatytułowanym *„Metody syntezy mowy i ich realizacje dla różnych języków"*, w oparciu o dane z literatury przedstawiono zasadnicze metody syntezy mowy, z których obecnie tylko metoda korpusowa ma istotne znaczenie praktyczne

W Rozdziale 3 *„Realizacje funkcji kosztu w wybranych systemach syntezy mowy"*, opisano konstrukcję funkcji kosztu oraz dokonano ogólnego przeglądu metod ich realizacji w systemach syntezy mowy.

Rozdział 4 *„Przygotowanie akustycznej bazy danych dla korpusowej syntezy mowy języka polskiego"* jest bardzo istotną częścią rozprawy (41 stron) i stanowi wart podkreślenia wkład autora w rozwiązanie problemu poprawy jakości mowy syntetycznej, a zwłaszcza naturalności jej brzmienia. Szczegółowo omówiono dobór tekstów do tworzonego korpusu, a następnie jego optymalizacji pod kątem jego rozmiarów, jak i reprezentatywności dla mowy polskiej. Jednocześnie balansując wielokrotnie tworzony korpus pod kątem częstotliwości występowania fonemów i ich połączeń, weryfikowano czy stosunkowo rzadko występujące w mowie polskiej fonemy są dostatecznie często reprezentowane w tworzonym korpusie. Tak skorygowany fonetycznie korpus (zwany inaczej „bogatym fonetycznie") został zoptymalizowany do 2500 zdań i uzupełniony o listę ponad 300 wyrazów z rzadziej występującymi w mowie polskiej fonemami. Tak bardzo szczegółowo opracowany i statystycznie zweryfikowany korpus tekstowy stanowi istotny wkład autora w rozwój metod syntezy mowy polskiej. Z tego powodu uważam, że tytuł rozprawy nie w pełni odzwierciedla jej zakres tematyczny. Wydaje mi się, że bardziej właściwy tytuł byłby, na przykład, „Optymalizacja bazy danych i funkcji kosztu w korpusowej syntezie mowy polskiej". Również cennym wkładem tej części pracy jest szczegółowe omówienie warunków tworzenia bazy nagraniowej korpusu, jej automatycznej transkrypcji i segmentacji w oparciu o technikę HMM (niejawnych modeli Markowa), a także metod korekcji błędów segmentacji i anotacji oraz usuwania zakłóceń, co zostało bogato zilustrowane różnorodnymi przykładami. W literaturze na ogół zagadnienia te są omawiane bardzo powierzchownie, a jednak właściwa realizacja bazy nagraniowej ma tu decydujący wpływ na uzyskaną jakość mowy syntetycznej.

Rozdział 5 *„Optymalizacja funkcji kosztu w systemie syntezy mowy"* przedstawia przyjętą przez autora rozprawy metodę optymalizacji funkcji kosztu opartą na oryginalnym zastosowaniu algorytmu ewolucyjnego. Warto podkreślić, że opracowana metoda ta może być zastosowana dla dowolnej nagranej bazy korpusowej. W oparciu o zastosowaną technikę ewolucyjną w Rozdziale 6 zatytułowanym *„Wyniki"* przedstawiono rezultaty badań nad określeniem wag poszczególnych elementów funkcji kosztu, mających istotny wpływ na brzmienie mowy syntetycznej. Wykazano, że chociaż wyniki te uzyskano dla określonego mówcy i korpusu mają one charakter uniwersalny dla języka polskiego, powołując się przy tym na inną pracę (*Demenko i in.,2008*), w której podano ogólne, podobne uszeregowanie parametrów funkcji kosztu.

W Rozdziale 7 (*„Wnioski")* przedstawiono wyniki testu percepcyjnego MOS (chyba ten podpunkt powinien był znaleźć się w rozdziale 6), zastosowanego do subiektywnej oceny jakości obejmującego nie tylko mowę syntetyczną, ale również jakość głosu lektora, który nagrał całą bazę korpusową. Otrzymane wyniki ocen subiektywnych mowy syntetycznej potwierdziły skuteczność przyjętej metodologii łączenia segmentów opartej na optymalizacji funkcji kosztu. We wnioskach podsumowano wady i zalety opracowanego systemu syntezy mowy polskiej osadzonego w środowisku Festival, który jednak narzuca szereg ograniczeń, które uniemożliwiły autorowi rozprawy uzyskanie w pełni satysfakcjonującej mowy syntetycznej, a zwłaszcza syntezy mowy w czasie rzeczywistym.

3. Uwagi szczegółowe

Jak już wcześniej wspomniałem tytuł rozprawy jest zbyt zawężający jej tematykę i przy ewentualnej publikacji rozprawy zalecałbym jego zmianę uwzględniającą wyniki otrzymane przy tworzeniu korpusu. Jest to tym bardziej uzasadnione, że sam autor we wprowadzeniu oprócz dwóch zasadniczych tez rozprawy wymienia, że istotny wpływ na brzmienie mowy syntetycznej ma jakość nagranej bazy korpusu. Pomimo, że praca pod względem

merytorycznym nie budzi wątpliwości, to w jej tekście jest niestety szereg niejasnych stwierdzeń, czy niekiedy niezbyt poprawnych sformułowań, z których zostaną wymienione te najważniejsze.

Na początku rozdziału 1 wymieniając obszary badań w fonetyce wymienia się dwukrotnie tę samą dziedzinę o różnych nazwach fonetyka audytywna – fonetyka percepcyjna. Omawiając w tym rozdziale głoski języka polskiego szkoda, że po punkcie 1.6 dotyczącym klasyfikacji segmentów mowy o różnej rozciągłości nie przedstawiono, które głoski są bardziej złożone pod kątem tworzenia difonów (dotyczy to szczególnie głosek polisegmentalnych, nazalizowanych oraz płynnych), bowiem ich w miarę płynne łączenie nie zawsze jest łatwe do uzyskania (autor tu wymienia tylko głoskę /j/). Być może w takich przypadkach, zalecane by było stosowanie segmentów o rozciągłości trifonu. Na str. 31 autor pisze, że segmentacja trifonów jest w wielu przypadkach bardzo złożona. Wydaje mi się, że nie bardziej, niż ma to miejsce w przypadku difonów. Brak w tym punkcie definicji półsylaby, a także nie ma w tabeli 1.6 oceny jakości syntezy mowy opartej na półsylabach.

W rozdziale 2 chyba błędnie opisano działanie maszyny mówiącej von Kempelena. Na pewno usta były zamodelowane w postaci tulei wykonanej ze skóry, a nosową nie otrzymuje się przez przykrycie nozdrzy palcami. Podobnie, niezbyt niedokładnie został opisany pierwszy elektryczny syntezator VODER (w tym niezbyt szczęśliwe stwierdzenie „... syczenie regulowane było za pomocą nadgarstka."). Do obu powyższych opisów (i także syntezatora Fabera) brakuje odsyłaczy do literatury. Przy opisie syntezy konkatenacyjnej brak szerszego omówienia Rys. 2.5 (str. 40) i funkcje niektórych bloków z tego rysunku nie są dla czytelnika jasne. Podobnie jest z Rys. 2.6 (str. 42), w którym ponadto blok zatytułowany „analiza mowy" powinien chyba być określony jako „konwersja fonetyczna tekstu".

W opisie funkcji kosztu syntezatora mowy polskiej RealSpeak niezbyt jasne jest znaczenie funkcji maskującej. Ponadto, w opisie miary odległości (symbolicznej) niezrozumiałe jest zdanie „...głoskę /p/ z lewym kontekstem /b/ będzie lepiej konkatenować, niż /p/ z sąsiedztwem /s/" (?). Rys. 2.8 (schemat syntezy statystycznej) też nie został opisany.

W końcowym zdaniu podpunktu 4.1 (powinno być 4.1.1) autor stwierdza, że w swojej pracy nie wykorzystał modułu transkrypcji zrobiony przez Dominikę Oliver dla Festivala, tylko zdecydował się na opracowanie własnego (?) opartego na metodzie drzew decyzyjnych (?). Nie za bardzo jest jasne, czy moduł istniejący w Festiwalu był w ogóle możliwy do wykorzystania w opiniowanej pracy. W podpunkcie 4.1.7 Rys.4.5 i 4.6 mają błędne podpisy, a poza tym wykresy słupkowe w Rys. 4.5 i 4.7 powinny być uporządkowane według tej samej kolejności fonemów. Podpunkt 4.3 ma tytuł zbyt lakoniczny, lepiej gdyby był „Segmentacja i transkrypcja wypowiedzi zarejestrowanych w bazie korpusowej". Autor rozprawy, często utożsamia segmentację z jednoczesną transkrypcją fonetyczną wypowiedzi.

W podpunkcie 4.3.2 opisując modele HMM utworzone za pomocą pakietu HTK autor pisze, że w każdym modelu wyróżnione są trzy stany – nagłos, śródgłos i wygłos. Tych terminów bym nie stosował, bowiem w fonetyce mają określone znaczenie i dotyczą przede wszystkim fraz, a nie pojedynczych głosek. Ponadto, na tej samej stronie (str. 97) wprowadzono w opisie wytrenowanych modeli termin „mikstury Gaussowskie" bez jego wyjaśnienia. Dalej, nie za bardzo wiadomo, dlaczego w zdaniu dotyczącym ostatniego (5-ego) zestawu modeli HMM wytrenowanego w pracy dano odsyłacz do publikacji B. Williamsa.

Na początku podpunktu 4.4 jest stwierdzenie, że w pracy zmodyfikowano istniejące (w Festivalu ?) dla języka polskiego moduły lingwistyczne, bez podania dlaczego i jaki był zakres zmian. W tym samym podpunkcie autor rozważa problem wpływu zbyt dużych wahań częstotliwości podstawowej F0 w sygnale mowy na jakość syntezowanej mowy. Zaznacza słusznie, że nagrywane teksty nie powinny być wypowiadane z nadmiernie modulowaną częstotliwością F0, bowiem utrudnia to późniejsze dobre łączenie ze sobą segmentów. Niemniej, proponowane w pracy zawężenie przedziału zmian wysokości głosu do 6-8 półtonów (czego zresztą słusznie nie zrobiono) wydaje się być zbyt daleko idące, bowiem zmiany jej w zdaniach pytających, a także w pierwszej części zdań z uzupełnieniem są na ogół większe i sięgają blisko 12 półtonów (czyli jednej oktawy, co odpowiada dwukrotnemu wzrostowi częstotliwości).

Podpunkt 5.2 poświęcony przedstawieniu algorytmu ewolucyjnego powinien mieć umieszczony odsyłacz do monografii X. Hue na samym początku, a nie na końcu tego podrozdziału. W podpunkcie 5.2, na stronie 121 (samego dołu) stwierdza się, że stosowanie tego algorytmu jest procesem bardzo czasochłonnym, poczym w następnym zdaniu jest powiedziane, że tak stworzony syntezator może działać w czasie rzeczywistym, ponieważ czas potrzebny na wygenerowanie zdania jest znacznie krótszy, niż w przypadku syntezy konkatenacyjnej (?).

W rozdziale omawiającym otrzymane wyniki (Rozdział 6), moim zdaniem, jest zła kolejność tabeli 6.5, która powinna być umieszczona przed tabelą 6.1 lub bezpośrednio za nią, zgodnie z tekstem, w którym jest ona omawiana.

We wnioskach (Rozdział 7) jest niezbyt jasne jak dobierano najgorszą funkcję kosztu do generowania zdań w teście MOS. Ponadto, wyniki testu umieszczone w Tabeli 7.2 są nieczytelne. W zakończeniu (str. 139) autor rozprawy stwierdza, że „Podczas pierwszych prób pracy z funkcją kosztu próbowano ustalać parametry kierując się pewnymi przesłankami związanymi ze specyfiką języka polskiego, a zwłaszcza z realizacją akcentu." Nie za bardzo wiadomo o co chodzi, czy akcent w korpusie nie zawsze jest prawidłowy, czy też nie zawsze był właściwie anotowany ?

4. Ocena edytorska pracy

Praca jest napisana na ogół w przejrzysty sposób, od strony edytorskiej nie budzi wiele zastrzeżeń. Poza wymienionymi wcześniej, pewne zastrzeżenie może budzić częste stosowanie w tabelach szarego tła pogarszającego ich czytelność, a także ich małe rozmiary, bądź zbyt mała czcionka. Ponadto, często są używane w pracy terminy, które są zrozumiałe tylko dla osób zajmujących się syntezą, bądź automatycznym rozpoznawaniem mowy. Przykładowo, na str. 128 jest zdanie „Pewną alternatywą byłoby zastosowanie automatycznej metody oceny głosu takie jak MUSHRA, PEAQ uniezależniającej od eksperta....", które dla osób nie zajmujących się metodami subiektywnej oceny jakości dźwięku powinien być odsyłacz do zalecenia Międzynarodowej Unii Telekomunikacyjnej (ITU) BS-1534 (MUSHRA) i BS-1387 (PEAQ) (MUSHRA chyba nie jest metodą automatyczną jak PEAQ, tylko subiektywną). W wielu przypadkach brak jest odsyłaczy do literatury, co już wcześniej zostało w niektórych przypadkach zasygnalizowane. Natomiast, jeśli chodzi o wykaz literatury jest on wyczerpujący i starannie dobrany. Załączona do pracy płyta DVD stanowi jej cenne uzupełnienie, jednakże uważam, że warto dołączyć do pracy szczegółowy wykaz zawartości tej płyty, tym bardziej że duża część plików zapisana jest w formacie skompresowanym.

5. Podsumowanie

Po zapoznaniu się z przedłożoną pracą, uważam że rozprawa doktorska mgr Krzysztofa Szklannego jest napisana na bardzo wysokim poziomie naukowym i stanowi oryginalne rozwiązanie problemu naukowego. Zawiera ważne wyniki teoretyczne, które w istotny sposób rozszerzają wiedzę o znaczeniu funkcji kosztu w procesie optymalizacji korpusowej syntezy mowy. Teza ta została udowodniona na przykładzie syntezy mowy polskiej. Istotnym wkładem autora rozprawy, o dużym znaczeniu praktycznym, jest bardzo szczegółowa analiza tworzenia optymalnego korpusu mowy polskiej umożliwiającego syntezę o dużej naturalności brzmienia, przy jednoczesnym umożliwieniu działania syntezatora w czasie rzeczywistym, stworzonego w innym środowisku niż Festival. Uzyskane wyniki mogą prowadzić również, moim zdaniem, do bardzo ważnych zastosowań praktycznych w dziedzinie automatycznego rozpoznawania i syntezy mowy polskiej, i choćby z tego powodu bardzo zachęcam autora rozprawy do opublikowania jej w czasopiśmie naukowym.

W konkluzji stwierdzam, że opiniowana rozprawa spełnia wymogi stawiane przez odpowiednią ustawę o stopniach i tytułach naukowych dla prac doktorskich, a jej autor w pełni zasługuje na przyznanie mu stopnia naukowego doktora nauk technicznych w dyscyplinie Informatyka. Stawiam więc wniosek o dopuszczenie tej rozprawy do publicznej obrony.