mgr inż. Krzysztof Szklanny

Katedra Multimediów,

PJWSTK

Warszawa, 24 października 2005

Autoreferat planowanej rozprawy doktorskiej

Przygotowanie funkcji kosztu dla realizacji korpusowej syntezy mowy dla języka polskiego

Moduł konwersji tekstu na mowę (Text-to-speech system) odpowiada za odczytanie głosem tekstu wprowadzonego do komputera. Zadaniem tego modułu jest wygenerowanie dźwiękowej postaci każdego wprowadzonego tekstu. Celem nowoczesnych projektów jest zapewnienie takiej jakości syntezy, by słuchający nie był w stanie odróżnić mowy syntetyzowanej od naturalnej.

Z oczywistych powodów nie jest możliwe stworzenie i nagranie wszystkich form i wszystkich słów dla danego języka, stąd koniecznośc syntetyzowania dźwięku System TTS definiuje się jako system automatycznego generowania mowy z transkrypcją fonetyczną oraz modułami odpowiedzialnymi za prozodię i intonację.

Każdy system TTS składa się z :

- modułu NLP (*Natural Language Processing*), który jest odpowiedzialny za przetwarzanie języka naturalnego.
- modułu przetwarzania cyfrowego sygnału - DSP (*Digital Signal Processing*), odpowiedzialnego za generację akustycznego sygnału mowy.

Celem NLP jest przekształcenie tekstu na zapis fonetyczny[1]. Moduł NLP jest również odpowiedzialny za wygenerowanie odpowiedniej prozodii dla analizowanego tekstu.

Z uwagi na niejednoznaczności języka naturalnego oraz skomplikowane zależności międzywyrazowe stworzenie modułu NLP jest zadaniem trudnym.

Dodatkowe problemy związane są z uzyskaniem naturalnej prozodii, która w dużej mierze zależy od semantyki wypowiedzi, ale ma również wiele wspólnego ze składnią i pragmatyką. Obecnie jednak, z powodu trudności znalezienia jednoznacznej kategorii przynależności słowa do kategorii semantycznej, systemy TTS skupiają się w głównej mierze na składni.

Prowadzone są badania nad semantyką i pragmatyką, jednak dotychczasowe rezultaty nie są jeszcze wystarczające do praktycznej implementacji w systemach TTS.

---

[1] Zapis fonetyczny jest zapisem słów wypowiedzi, przy użyciu fonetycznego zapisu głosek.

Moduł NLP składa się z następujących podmodułów:

- Pre-procesor: (normalizator tekstu), jego zadaniem jest podział zdań na wyrazy. Proces podziału jest dość skomplikowany, z uwagi liczne skróty i interpunkcję języka polskiego. Moduł ten wydziela z tekstu skróty, liczby, idiomy, akronimy i rozwija je do pełnego tekstu. Problem stanowi także rozpoznawanie końca zdania. Zauważmy, że często po skrótach stawiany jest znak kropki, co nie zawsze oznacza koniec zdania. Po przetworzeniu dane są przechowywane w wewnętrznym module struktur danych.
- Tagger– jest odpowiedzialny za wyznaczenie części mowy dla każdego ze słów (rzeczownik, przymiotnik). Zadania taggera sprowadzają się do zmniejszenia słownika oraz ustalenia części mowy.
- Analizator kontekstowy – zadaniem analizatora kontekstowego jest ograniczenie znaczenia poszczególnych słów. Ograniczenie to odbywa się na podstawie zbadania kontekstu słów (części mowy) znajdujących się w sąsiedztwie. Stosuje się tutaj metodę n-gramów, która opisuje syntaktyczne zależności pomiędzy słowami, na zasadzie badania prawdopodobieństw w skończonych przejściach automatu. Służą do tego modele Markova lub wielowarstwowe sieci perceptronowe. Użycie sieci neuronowych sprowadza się do odkrycia reguł rządzących kontekstem zdaniowym. Stosuje się również metody lokalnych niestochastycznych gramatyk. Analiza ta jest potrzebna od określania intonacji wypowiedzi, zakładają się bowiem, że pewne typy wyrazów są akcentowane a inne nie. Im analiza gramatyczna jest dokładniejsza, tym wierniejsza będzie intonacja syntetycznej mowy.
- Parser syntaktyczno-prozodyczny jest odpowiedzialny za utworzenie prozodii i intonacji dla poszczególnych sekwencji fonemów. Parser ten bada jednocześnie pozostałe wyrażenia, które nie zostały zakwalifikowane do żadnej z kategorii. Następnie stara się znaleźć podobne do nich struktury tekstowe, których elementy prozodyczne będą najbardziej prawdopodobne i zbliżone do siebie. Następnie generuje opis symboliczny przebiegu konturu intonacyjnego i iloczasów, które następnie są wykorzystane do modyfikacji własności akustycznych wybranych modeli głosek
- Moduł *letter to sound* –za utworzenie transkrypcji fonetycznej dla istniejących słów.

Powstaje jednak kilka problemów, dość istotnych z punktu widzenia realizacji tego modułu. Mianowicie: słownik wymowy obejmuje tylko podstawowe słowa, bez morfologicznych kombinacji, to znaczy nie uwzględniający rodzaju, przypadku, liczby.

Istnieje wiele słów o podwójnym znaczeniu i takiej samej pisowni. Dodatkowo trudno sobie wyobrazić istnienie wszystkich słów w słowniku.

Moduł cyfrowego przetwarzania sygnału może być realizowany na dwa sposoby:

Pierwszy, zwany regułową syntezą mowy, polega na jej generowaniu poprzez układ symulujący ludzki aparat mowny o zmiennych parametrach.

Drugi, zwany konkatenacyjną syntezą mowy polega na łączeniu jednostek akustycznych wyboeranych z bazy nagran głosu naturalnego.

Obecnie najbardziej rozpowszechniona jest ta druga metoda Model takiej syntezy mowy, rozwijany od lat 70, zyskał dużą popularność z uwagi na możliwość generowania bardzo naturalnej, dobrze brzmiącej i zrozumiałej mowy w stosunkowo prosty sposób.

. Konkatenacyjna synteza mowy generuje sygnał akustyczny poprzez sklejanie ze sobą elementów akustycznych powstałych z naturalnej mowy (fony, difony, trifony, sylaby).Zaletą syntezy difonowej jest niewielki rozmiar bazy danych, z uwagi na mała ilość jednostek akustycznych. Im mniejszy rozmiar bazy, tym szybciej będzie syntetyzowana mowa oraz wymagania sprzętowe będą mniejsze.

Konkatenacja sylab daje dość dobre rezultaty, jednak z uwagi na ich ilość (np. w języku angielskim, – około 160000 podczas gdy jest tylko 40 fonemów) też wydaje się być nie najlepszym rozwiązaniem. Bardzo często używana jest konkatenacja difonów, która umożliwia dobrą jakość syntezy mowy przy wykorzystaniu korpusu zawierającego około 1500 jednostek. Taką syntezę zrealizowałem w ramach mojej pracy magisterskiej

Konkatenacyjna syntezy mowy ma również pewne problemy. Należą do nich:

- Problem wyboru jednostek akustycznych,
- Konkatenacja jednostek nagranych w różnych kontekstach.
- Modyfikacja prozodii, czyli problem intonacji i czasu trwania.
- Problem kompresji nagranych segmentów.

Stosunkowo nowym rozwiązaniem jest metoda korpusowa (*unit selection*). Jest to zmodyfikowana postać konkatenacyjnej syntezy mowy. Metoda korpusowa zakłada, że korpus jest dużo większy, tak, że zawiera wiele instancji danej jednostki akustycznej. W korpusie mogą występować również inne jednostki akustyczne np. sylaby i trifony oraz całe wyrazy. W zależności od kontekstu wybierana jest jednostka najbardziej pasująca. Największym problem jest stworzenie odpowiednich reguł do realizacji funkcji kosztu. Funkcję kosztu rozbija się z reguły na dwie składowe: koszt konkatenacji i koszt doboru jednostki (*unit cost*):

Jeśli wszystkie koszty konkatenacji będą jednakowe, to ciąg o najmniejszej ilości elementów będzie miał najniższy koszt, a więc faworyzowany będzie wybór jak najdłuższych jednostek. Powinno to pozytywnie wpływać na jakość syntetycznej mowy.

Generalnie, jeśli funkcja kosztu zależy od wielu czynników i baza danych zawiera wiele elementów, stosuje się zaawansowane techniki optymalizujące przeszukiwanie, np. algorytm Viterbiego.

Elementy funkcji kosztu szacuje się metodami empirycznymi – przyjmując bądź pewne stałe wartości progowe lub szacując je dla danego zbioru segmentów.
Koszt konkatenacji elementów wyniesie zero jeśli następują one po sobie w bazie danych. Jeśli nie, zwykle zależy on od czynników związanych z koartykulacją i F0. Podobnie jest z kosztem doboru jednostki – zależy on od kontekstu koartykulacyjnego i ciągłości tonu podstawowego.

Podsumowując, funkcja kosztu jest funkcją oszacowującą. Jej działanie sprowadza się do wyliczenia różnych możliwych sposobów wygenerowania danej wypowiedzi, przy użyciu różnych jednostek akustycznych znajdujących się w korpusie. Funkcja oszacowuje i porównuje zarazem, która wypowiedź będzie brzmiała najlepiej. Funkcja uwzględnia różne czasy trwania poszczególnych segmentów oraz ich intonację.

Konstrukcja funkcji kosztu jest jednym z najważniejszych problemów do rozwiązania w implementacji syntezy korpusowej. Mało jest publikacji na ten temat, zwykle stanowi ona dobrze strzeżoną tajemnicę firmy zajmującej się syntezą mowy. Zagadnienie to nie było nigdy analizowane dla języka polskiego

Innym problemem przy tworzeniu korpusowej syntezy mowy jest przygotowanie korpusu. Z zasady syntezy korpusowej wynika, że w bazie danych powinny znajdować się jak najdłuższe segmenty w odpowiednich kontekstach prozodycznych charakterystyczne dla domeny syntezy, tj. tekstów które będą syntetyzowane. Należy zatem nagrać i przechowywać segmenty wybrane z naturalnych wypowiedzi (słów, zdań). Ważne jest także dokładne określenie i przeanalizowanie tekstów jakie będą przedmiotem syntezy. Oczywiście, im więcej segmentów dostępnych będzie w bazie danych, tym naturalniejsza będzie generowana mowa. Jak dobrać jednak ich minimalną ilość zapewniającą dobrą jakość? Zwykle stosuje się iteratywną metodę doboru Greedy'ego: z bazy danych w każdym kroku usuwa tą się jednostkę której usunięcie spowoduje najmniejszy wzrost całkowitego kosztu. Zwykle w tym celu tworzy się pewien testowy zbiór tekstów na którym analizuje się koszty doboru segmentów.

Realizacja systemu korpusowej syntezy mowy

Moim celem jest stworzenie w pełni funkcjonalnego systemu korpusowej syntezy mowy języka polskiego. Dotychczas określiłem domenę syntezy. Obejmuje ona wypowiedzi na tematy polityczne oraz zawiera różnego rodzaju teksty gazetowe. Bardzo istotnym elementem podczas realizacji korpusowej syntezy mowy jest stworzenie modułu akustycznego. Jego realizacja wymaga dokonania wyboru odpowiednich tekstów oraz stworzenia słownika nagrań. Następnie należy stworzyć minimalny zbiór zdań 'bogatych' pod względem fonetycznym, żeby zminimalizować rozmiar korpusu. Mój korpus składa się z 2150 zdań, zawierających większość difonów, najczęściej występujące trifony oraz często występujące wyrazy języka polskiego. Konstrukcję modułu akustycznego można uznać za ukończoną. Korpus został przygotowany i nagrany w studio. Chcąc

zapewnić większą do bazy naturalność syntezy dograne zostały elementy paralingwistyczne (śmiech, oddech, kaszel, zająknięcia). Korpus został posegmentowany przy użyciu specjalnie do tego stworzonych narzędzi. Należy jeszcze sprawdzić jakość segmentacji i dokonać korekty ręcznej. Następnie należy skoncentrować uwagę na przygotowaniu modułu lingwistycznego oraz funkcji kosztu. Oczywiście w pełni funkcjonujący system należy odpowiednio przetestować i przeprowadzić badania percepcyjne generowanej mowy.
System będzie zaimplementowany w środowisku Festival[2]. Festival dostępny jest na zasadzie *GPL/GNU public license.*

Podsumowanie

Stworzenie systemu korpusowej syntezy mowy nie jest zadaniem trywialnym. Istnieje komercyjny system syntezy korpusowej dla języka polskiego stworzony przez firmę Lernout & Hauspie[3]. W Polsce został stworzony hybrydowy system syntezy mowy przez firmę Ivo Software. System ten nie jest stricte system korpusowym. Trwają badania nad implementacją systemu korpusowej syntezy mowy na Uniwersytecie Adama Mickiewicza w Poznaniu.

Uważam, że istnieje potrzeba stworzenia systemu syntezy mowy korpusowej dla języka polskiego. Stworzenie pełnej aplikacji wiąże się z przeprowadzaniem dokładnych badań na temat funkcji kosztu i dokładnego zbadania tego zagadnienia. Istnieje bardzo mało publikacji, które mogłyby pomóc w stworzeniu dobrej funkcji kosztu w związku z tym moja praca ma wymiar nie tylko teoretyczny ale i praktyczny.

[2] http://www.cstr.ed.ac.uk/projects/festival
[3] Firma ta została wykupiona przez Scansoft.