Przygotowanie bazy difonów języka polskiego dla realizacji syntezy mowy w systemie MBROLA

MBROLA.Creating the Polish diphone database for speech synthesis.

Krzysztof Szklanny

Polsko-Japońska Wyższa Szkoła Technik Komputerowych

Ul. Koszykowa 86
02-008 Warszawa
kszklanny@pjwstk.edu.pl

STRESZCZENIE

Głównym celem pracy było przygotowanie bazy difonów języka polskiego dla realizacji syntezy mowy w systemie MBROLA. Zagadnienia dotyczące tworzenia bazy difonów są ściśle związane z konkatenacyjną syntezą mowy, która generuje mowę poprzez łączenie ze sobą elementów akustycznych powstałych z naturalnej mowy takich jak: fony, difony, trifony, sylaby. W celu realizacji takiej bazy należało przygotować korpus difonów., następnie przeprowadzić nagrania. Najtrudniejszym etapem była realizacja procesu segmentacji, czyli wyodrębnienia difonów w nagranym korpusie. Jakość bazy została zweryfikowana specjalnie do tego celu przygotowanym testem sześciuset słów zawierającym najczęściej występujące połączenia fonemów w języku polskim. Ostatnim etapem była normalizacja bazy przez zespół MBROL-i. W obecnej chwili system działa dla danych wejściowych podanych w transkrypcji fonetycznej. Należy dodać iż jest to pierwsza baza taka baza powstała w Polsce. Od maja 2002 roku opracowana baza difonów znajduje się na stronie internetowej MBROL-i (http://tcts.fpms.ac.be/synthesis/mbrola) jako nowy model głosowy.

ABSTRACT

The goal of work was to create the Polish diphone database. Additional aim of the project was to obtain high quality speech synthesis for the Polish language. The whole process included several stages. First of all I had to prepare a phoneme list and create the corpus of diphones. Then next was to make the recordings of the corpus and the segmentation of diphones. One of the last processes was to test the database then export it and sent it to the MBROLA team who conducted the normalization process.

1. Realizacja bazy difonów

Pracę rozpoczęto od przygotowania korpusu. Najbardziej istotnym elementem podczas przygotowania korpusu było znalezienie odpowiedniego kontekstu dla difonów.

W korpusie musiały się znaleźć wszystkie możliwe połączenia głosek. W wygenerowanych wyrazach difon nie mógł stanowić sylaby akcentowanej, ponieważ mogło to mieć wpływ na zniekształcenia nagranego w późniejszym etapie sygnału. Otoczenie difonu nie mogło wpływać na jego koartykulację, co wbrew pozorem nie jest takim łatwym zadaniem.

Następujące reguły zostały uwzględnione:

- Dla grupy samogłoskowej-samogłoskowej: b+difon+nany np: baenany
- Dla grup samogłoskowo-spółgłoskowych :d+difon+bany np. didbany lub d+difon+pany np. difpany
- Dla grup spółgłoskowo-spółgłoskowych: a+difon+anany np. apsanany
- Dla grup spółgłoskowo-samogłoskowych: a+difon+nany np. afanany lub a+difon+banany np. azubanany
- Dla difonu w postaci cisza+głoska zastosowano kontekst: difon+ana jeśli difon był formacją spółgłoskową lub difon+bana jeśli difon był formacją samogloskową
- Dla difonu w postaci głoska + cisza kontekst był następujący: bana+difon – jeśli głoska była spółgłoskową lub ban+difon – jeśli głoska była samogłoskową

Podczas przygotowania korpusu napotkano bardzo wiele sytuacji, w których nie można było zastosować powyższych reguł. Sytuacje te przeważnie dotyczyły grup spółgłoskowo-spółgłoskowych i samogłoskowo-samogłoskowych. Np. difon w postaci „ii" miał kontekst „aniimadwo". W takich przypadkach kierowano się częściej intuicją niż konkretnymi regułami.

1.1. Nagrania

Warunkiem koniecznym realizacji nagrań było znalezienie profesjonalnego mówcy. Osobą taką była studentka trzeciego roku szkoły Teatralnej im. Zelwerowicza w Warszawie.

Korpus został nagrany w programie Mobile Recording Studio firmy Sony w postaci plików Raw z częstotliwością próbkowania 32 kHz. Taka częstotliwość zapewnia wysoką, dla syntezy mowy, jakość generowanego sygnału.

Do nagrań użyto 4 mikrofonów (mikrofon tzw. „Close-talking" oraz 3 mikrofony zbierające sygnał z otoczenia umieszczone w odległości nie większej niż jeden metr od mówiącego).

W procesie segmentacji wykorzystano nagrania przeprowadzone na mikrofonie Sennheiser ME 104 „close-talk". Okazało się że sygnał pochodzący z mikrofonu „close-talk" zawiera najmniej zniekształceń. Pomimo występujących przydechów, oraz mlaśnięć mówcy zagwarantował z wszystkich mikrofonów najlepszą jakość sygnału i wprowadził najmniejszą ilość zniekształceń w sygnale.

Nagrania zostały przeprowadzone w studiu Polsko-Japońskiej Wyższej Szkoły Technik Komputerowych. Jest to profesjonalne studio, specjalnie wygłuszone materiałem uniemożliwiającym powstawanie znacznych odbić fali. Studio zostało skonstruowane w celu realizacji projektów takich jak synteza, rozpoznawanie mowy, tworzenie zaawansowanych ścieżek audio.

Po przeprowadzeniu nagrań następnym etapem a zarazem stanowiącym główną część projektu była segmentacja wcześniej przygotowanego korpusu.

1.2. Segmentacja

Proces segmentacji był realizowany w Praacie, programie wspomagającym pracę fonetyków. Program ten jest dostępny za darmo na stronie www.praat.org

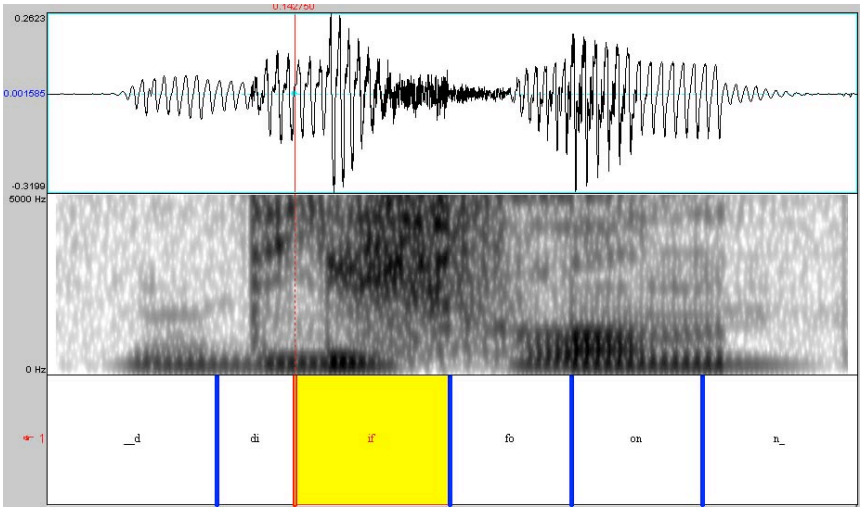
Segmentacja polega na wyznaczeniu granic jednostki akustycznej – difonu. Obejmuje zawsze początek jednostki akustycznej, środek, czyli moment przejścia pomiędzy jednym a drugim fonemem oraz koniec – czyli prawą granicę będącą końcem difonu.

Problemy, jakie mogą się pojawić podczas segmentacji mogą być związane z:

Amplitudą - nieciągłość amplitudy pojawia się, kiedy koniec difonu i początek bezpośrednio po nim następującego są zupełnie inne. Podczas łączenia takich elementów akustycznych pojawiają się trzaski w generowanym sygnale.

Energią - nieciągłość energii oznacza, że difon wraz z kontekstem został wymówiony ze zwiększoną energią niż następujący lub poprzedzający go.

Przesunięciem fazy - bardzo ważnym elementem segmentacji jest wystrzeganie się od poniższego błędu. Nieciągłość fazy pojawia się, jeśli granica difonu nie znajduje się na początku okresu krtaniowego fonemu. Efektem ustawienia niepoprawnych granic jest słyszalny trzask w generowanym sygnale.



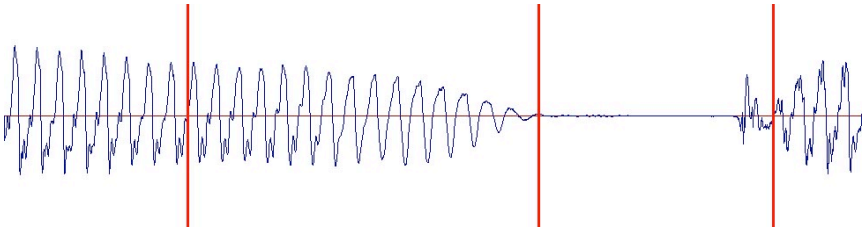
Rysunek 1.1 Proces segmentacji.

Poniżej przedstawiono pewne reguły, które zastosowano podczas procesu segmentacji i, które w znacznej mierze wpłynęły na sukces projektu.

Długość difonu nie powinna przekraczać 70 ms oraz, dodatkowo w difon nie powinien mieć więcej niż pięć okresów krtaniowych z lewej strony i pięć okresów krtaniowych z prawej.

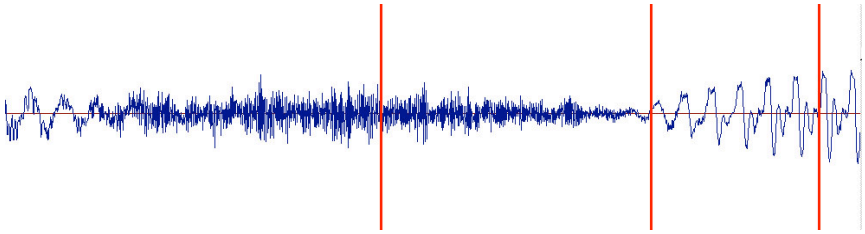
W przypadku samogłosek założenie to było możliwe do realizacji. Wynika to oczywiście ze struktury samogłoski i jej krótkiego czasu trwania. Jednak podczas segmentacji spółgłosek zdarzały się sytuacje, kiedy długość difonu wynosiła ponad 100 ms.

Poniższy rysunek prezentuje difon „e~p", którego czas trwania przekracza 130 ms.



Rysunek .1.2 Difon „e~p". Czas trwania ponad 130 ms

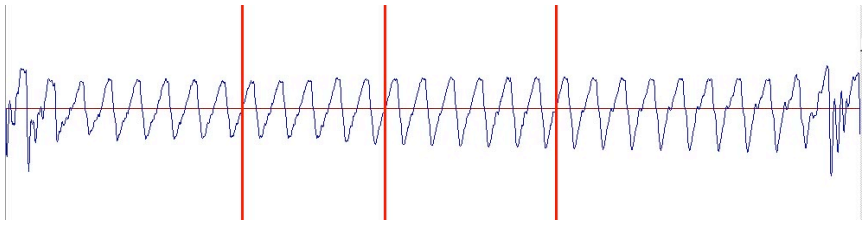
Zdarzały się sytuacje, w których ustawienie granic było trywialne. Miało to miejsce podczas segmentacji głosek szczelinowych. Jak widać na poniższym rysunku dokładnie można określić przejście pomiędzy jednym a drugim fonemem.



Rysunek.1.3 Granice w difonie „S-e”

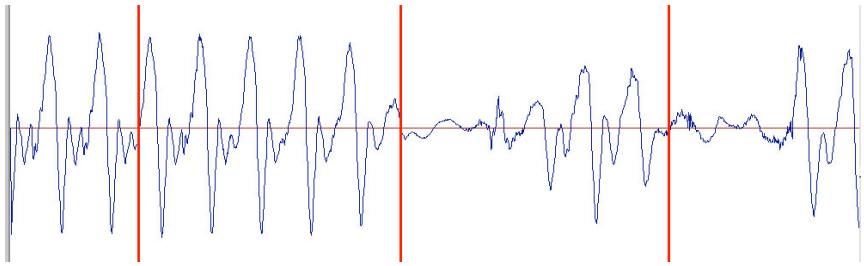
W przypadku głosek półotwartych zarówno ustnych jak i nosowych zasady podziału były bardzo intuicyjne i określenie granic było bardzo trudne.

Poniższy rysunek prezentuje sytuację krytyczną, kiedy należało znaleźć granice pomiędzy fonemami „n” i „m”



Rysunek 1.4 Difon „n-m” .Granice między fonemem „n” i „m”

Segmentacja głoski drżącej „r” również była dość ciekawym przypadkiem. Trudno było w tym przypadku odróżnić nagłos od wygłosu. W dodatku głoska ta charakteryzuje się bardzo krótkim czasem trwania. Najdłuższa głoska „r” trwała około kilkunastu milisekund a wiadomo, że do poprawnego brzmienia potrzeba kilkanaście do kilkudziesięciu milisekund kontekstu.



Rysunek 1.5 Segmentacja difonu „e-r”

Po przeprowadzeniu procesu segmentacji należało zsynchronizować dane. To znaczy dokonać ich exportu z Praata do Diphone Studio. Etap ten miał na celu przygotowanie danych do ostatniego, bardzo ważnego etapu - testowania.

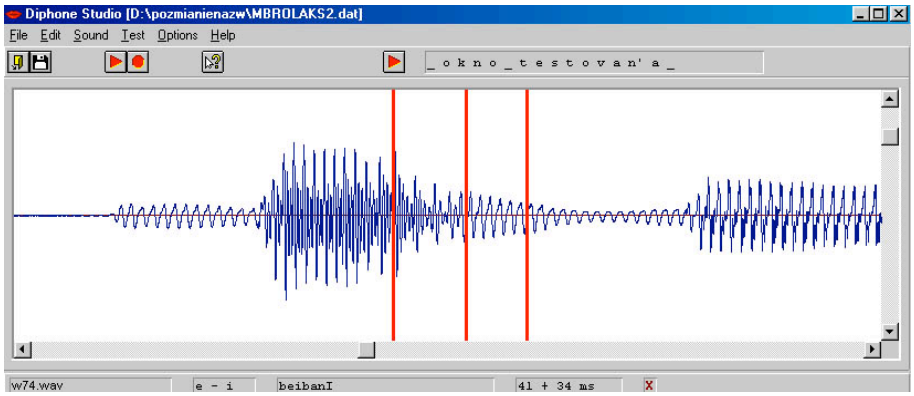
1.3 Testowanie

Do testowania danych użyto programu Diphone Studio. Program umożliwia generowanie mowy przy użyciu posegementowanego korpusu. Fragment wypowiedzi należy zapisać przy użyciu transkrypcji fonetycznej. Niejednokrotnie okazywało się, że granice difonu są źle ustawione, dlatego należało wprowadzić korektę.

W celu skorygowania i ustalenia czy granice difonów zostały poprawnie określone przetestowano cały korpus. W tym celu wykorzystano test składającym się z 600 słów. Test ten zawiera wszystkie połączenia, jakie mogą wystąpić w języku polskim. Słownik ten został stworzony przez pp. prof. Ryszarda Gubrynowicza oraz prof. Krzysztofa Maraska.

.

Poniżej znajduje się okno programu Diphone Studio pokazujące difon „e-i”.



Rysunek 1.6 Okno programu Diphone Studio

1.4 MBROLA - algorytm syntetyzowania mowy

MBROLA to nie tylko nazwa projektu, jest to również nazwa algorytmu syntetyzowania mowy. Zapewnia on bardzo dobrą jakość łączenia segmentów, przy jednocześnie niskim nakładzie kosztów obliczeniowych. Resynetza w MBROLI dotyczy ramek tylko akcentowanych, natomiast nie akcentowane są kopiowane. W ten sposób eliminujemy niezgodności intonacji, co zapewnia pełną zgodność fazy podczas procesu konkatenacji. Resynteza w MBROLI przynosi dodatkowe korzyści. Jedną bardzo ważną zaletą jest zachowanie stałej intonacji, co pozwala zaznaczyć wysokość tonu tzw. pitch mark. Wiąże się to z czasem, jaki jest potrzebny na przygotowanie

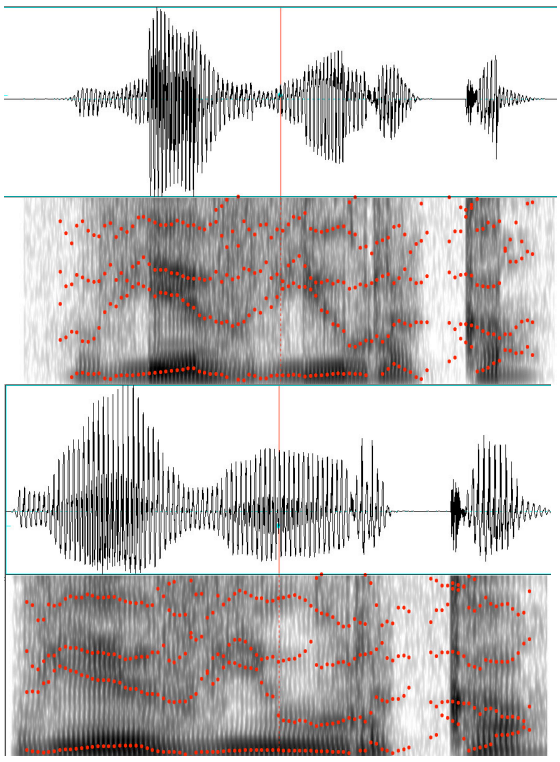
bazy danych. Algorytm MBROL-i prawdopodobnie generuje najlepszą jakość mowy, mimo zniekształceń wprowadzanych na akcentowanych częściach.

1.5 Normalizacja bazy difonów

Otrzymany korpus difonów należało zapisać do formatu wymaganego przez MBROL-ę oraz wysłać do Belgii, gdzie zespół MBROL-i dokonał odpowiednich operacji na korpusie. Normalizacja bazy polegała na zmodyfikowaniu danych, usunięciu wszelkich artefaktów sygnału powodujących trzaski oraz zapisanie ich w zmodyfikowanej postaci do jednego pliku.

Obraz normalizacji, która została przeprowadzona w Belgii jest przedstawiony na rysunkach poniżej. Diametralne różnice z trzaskami zostały usunięte. Na pierwszym rysunku znajduje się spektrogram oraz sygnał w dziedzinie czasu słowa „wiewiórka” przed normalizacją, na drugim ten sam wyraz po procesie normalizacji.

Można zauważyć zastosowany proces wygładzania sygnału.



Rysunek 1.7 Sygnał nieznormalizowany i znormalizowany

Podsumowanie

Celem pracy było stworzenie syntezy mowy polskiej opartej na bazie difonów języka polskiego dla realizacji syntezy mowy w systemie MBROLA.

Warunkiem tego było stworzenie bazy difonów dla języka polskiego w taki sposób by jakość syntezy była możliwie jak najlepsza.

Moduł akustyczny powinien mieć naturalne i zrozumiałe brzmienie, żeby był w pełni funkcjonalny w aplikacjach wykorzystujących syntezę mowy. Naturalność brzmienia jest warunkiem koniecznym do przyjęcia jej do powszechnych zastosowań.

Mając na myśli aplikacje należy tutaj wymienić dziedzinę edukacji w postaci wirtualnych uniwersytetów, portale głosowe, bezwzrokowe przekazywanie informacji. Również należy wspomnieć o syntezie mowy jako pomocy dla ludzi z zaburzeniami mowy.

Etap ten został przygotowany, dlatego praca daje możliwość kontynuacji do uzyskania w pełni systemu TTS .

W obecnej chwili system działa dla danych wejściowych podanych w transkrypcji fonetycznej. Kolejnym etapem z pewnością będzie przygotowanie modelu przetwarzania języka naturalnego włącznie z generowaniem prozodii.

Baza difonów jest dostępna od maja br. i znajduje się na stronie internetowej MBROL-i (http://tcts.fpms.ac.be/ synthesis/mbrola) jako nowy model głosowy języka polskiego w postaci difonowej. Na stronie również znajdują się przykładowe pliki z syntetycznym głosem.

BIBLIOGRAFIA

[1] Marasek, K. (1999) *Wykład Werbalna komunikacja z komputerem*

[2] Huang X.,Acero A..,Hon H. (1998) *Spoken Language Processing.*

[3] http://tcts.fpms.ac.be/synthesis/mbrola/mbruse.html

[4] http://www.ias.et.tu-dresden.de/kom/lehre/tutorial/ selection.htm

[5] http://tcts.fpms.ac.be/synthesis/mbrola.html