XX Jubileuszowe Seminarium

ZASTOSOWANIE KOMPUTERÓW W NAUCE I TECHNICE' 2010
Oddział Gdański PTETiS
Referat nr 20

SYNTEZER MOWY UWZGLĘDNIAJĄCY PROZODIĘ WYPOWIEDZI

Kuba ŁOPATKA, Andrzej CZYŻEWSKI

Politechnika Gdańska, ul. G. Narutowicza 11/12,
80-952 Gdańsk, Wydział Elektroniki, Telekomunikacji i Informatyki, Katedra Systemów Multimedialnych
Tel: 058 347 6332 e-mail: {klopatka,andcz}@sound.eti.pg.gda.pl

Streszczenie: Przedstawiono system syntezy mowy polskiej uwzględniający w sposób automatyczny prozodię, tj. profil intonacyjny, tempo i akcenty wypowiedzi. Zastosowano syntezę konkatenacyjną z wykorzystaniem jednostek mowy zawierających przejścia między dwoma głoskami – difonów. Opisano poszczególne moduły wchodzące w skład syntetyzera: przetwarzanie tekstu, bazę jednostek mowy oraz algorytmy związane z tworzeniem syntetyzowanego sygnału. Przeprowadzono testy subiektywne potwierdzające wysoką zrozumiałość generowanej mowy i skuteczność modyfikacji prozodycznych. Przedstawiono możliwość zastosowania opisanego systemu w aplikacjach edukacyjnych lub terapeutycznych oraz interfejsach multimodalnych przeznaczonych dla osób niepełnosprawnych.

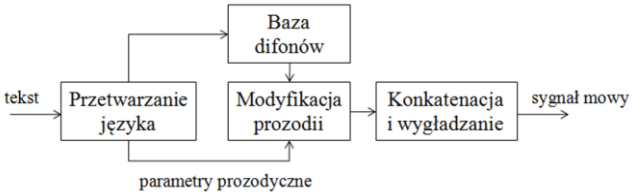
Słowa kluczowe: synteza mowy, prozodia, PSOLA.

1. WPROWADZENIE

Jednym z czynników, mających największy wpływ na jakość brzmienia mowy generowanej przez system syntezy mowy jest odwzorowanie prozodii wypowiedzi [1]. W naturalnej mowie obecne są akcenty wyrazowe, zdaniowe oraz intonacja. Te zjawiska fonetyczne wiążą się ze zmiennością takich cech mowy jak wysokość tonu podstawowego, czas trwania głosek i dynamika. Jest więc możliwe odwzorowanie prozodii przy pomocy odpowiednich narzędzi algorytmicznych. Celem niniejszej pracy jest stworzenie systemu, który uwzględnia prozodię wypowiedzi w sposób automatyczny. W referacie przedstawiono kolejne kroki tworzenia syntetyzera, ze szczególnym uwzględnieniem tych elementów, które odgrywają kluczową rolę w procesie kształtowania akcentu, intonacji i czasu trwania – tj. modułu analizy tekstu i algorytmów przetwarzania sygnałów. W rozdziale dotyczącym eksperymentów przedstawiono wyniki testów odsłuchowych, w których ocenie podlegały zrozumiałość generowanej mowy i skuteczność modyfikacji prozodycznych. W podsumowaniu opisano praktyczne zastosowanie przedstawionego systemu.

2. OPIS SYSTEMU TTS

Schemat opisywanego systemu zamieniającego tekst na mowę (ang. TTS – *Text To Speech*) przedstawiony został na rysunku 1. Poszczególne elementy systemu są opisane w następnych podpunktach.



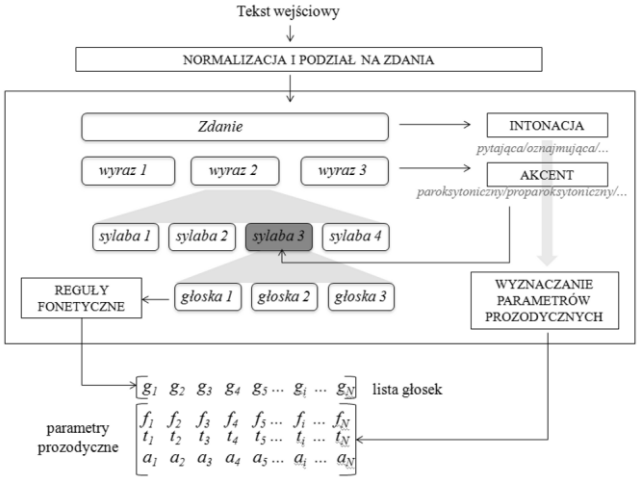
Rys. 1. Ogólny schemat blokowy systemu syntezy mowy

2.1. Przetwarzanie języka

Zadaniem systemu TTS jest zamieniać dowolny tekst na mowę. W związku z tym pierwszy blok odpowiada za przetwarzanie danych językowych, które są podane w formie tekstowej. Koncepcja przetwarzania tekstu w systemie przedstawiona jest na rysunku 2. W skład przetwarzania języka wchodzą następujące operacje:

- normalizacja tekstu, tj. zamiana znaków nieliterowych na odpowiadającą im informację słowną;
- rozbiór morfologiczny wypowiedzi – podział na zdania, wyrazy, sylaby;
- analiza fonetyczna – uwzględnienie reguł fonetycznych występujących w języku, wykorzystanie słownika dla uwzględnienia wyjątków;
- analiza prozodyczna.

Recenzent: Dr inż. Mirosław Rojewski – Wydział Elektroniki, Telekomunikacji i Informatyki Politechnika Gdańska



Rys. 2. Schemat koncepcyjny przetwarzania tekstu

Zadaniem analizy prozodycznej jest wyznaczenie parametrów prozodycznych każdej głoski występującej w syntetyzowanej wypowiedzi. Ostatecznym wynikiem przetwarzania językowego jest lista głosek i skojarzone z nią wektory parametrów: częstotliwości podstawowej (*f*), czasu trwania (*t*) i amplitudy (*a*). Na problem wyznaczania prozodii wypowiedzi składają się zagadnienia wyznaczania akcentu wyrazowego i intonacji zdania. Na rysunku 3 przedstawiono schemat algorytmu wyznaczania akcentu wyrazowego. Punktem wyjścia tego algorytmu jest podział zdania na wyrazy i podział wyrazów na sylaby. Domyślnie akcentowana jest druga sylaba od końca (akcent regularny – paroksytoniczny). Istnieją jednak wyrazy, które są akcentowane inaczej niż paroksytonicznie (np. słowa pochodzenia obcego). W związku z tym każde słowo jest sprawdzane pod kątem wystąpienia akcentu wyjątkowego. Oprócz tego wykrywane są cząstki ruchome typu *–by, -byś*, które odpowiadają za przesunięcie akcentu na trzecią lub czwartą sylabę od końca.

Rys. 3. Schemat algorytmu wyznaczania akcentu wyrazowego

Końcowym etapem algorytmu kształtowania akcentu jest przypisanie głoskom wchodzącym w skład analizowanej wypowiedzi odpowiednich parametrów sygnałowych (f_0,*t* ,*a*). Akcent wiąże się z modyfikacją wszystkich trzech ww. parametrów poprzez:

- podwyższenie albo obniżenie tonu (w zależności od miejsca w zdaniu);
- wydłużenie sylaby akcentowanej;
- zwiększenie dynamiki głosek wchodzących w skład akcentowanej sylaby.

Zagadnienie wyznaczania intonacji zdania można sprowadzić do przyporządkowania analizowanego zdania do podstawowych typów: np. zdanie oznajmujące, pytające, wykrzyknienie. Następnie wykorzystywane są wzorce wypowiedzi lektorskich ilustrujących odpowiednie kontury intonacyjne. W procesie kształtowania intonacji głoskom wchodzącym w skład zdania przypisywane są odpowiednie częstotliwości podstawowe, aby odwzorować wzorcową wypowiedź o danej intonacji.

2.2. Baza difonów

Współczesne systemy syntezy mowy są w większości oparte na metodzie konkatenacyjnej, tj. łączeniu syntetycznej wypowiedzi z mniejszych jednostek nagranych przez lektora. Długość wykorzystanych jednostek może być różna: od mikrofonemów, przez difony, trifony (odpowiednio złączenia dwóch i trzech fonemów) po sylaby i całe wyrazy. Przedstawiony syntetyzer wykorzystuje jednostki o stałej długości – difony. Punktem wyjścia syntezy jest baza 1369 difonów, które są wystarczające do wygenerowania dowolnej wypowiedzi w języku polskim. Baza difonów została nagrana w Katedrze Systemów Multimedialnych.

2.3. Modyfikacja prozodii

Aby nadać syntetyzowanej wypowiedzi wymagany przebieg prozodii, konieczna jest modyfikacja następujących parametrów sygnału mowy [1][2]:

- częstotliwości podstawowej (*f*),
- czasu trwania(*t*),
- amplitudy (*a*).

Docelowe wartości parametrów prozodycznych, skojarzone z kolejnymi głoskami, są wynikiem analizy prozodycznej, będącej częścią procesu przetwarzania tekstu. Modyfikacje parametrów *f* i *t* wykonywane są bezpośrednio na difonach przy pomocy algorytmów zmiany czasu trwania (ang. *timescale modification*) i tonu podstawowego (ang. *pitch-shifting*). Większość opisywanych w literaturze algorytmów poświęconych temu zastosowaniu dedykowana jest do przetwarzania skomplikowanych sygnałów o złożonym charakterze (np. muzyki) [3]. Sygnał mowy jest sygnałem o wiele mniej skomplikowanym i możliwe jest wykorzystanie metody o mniejszej złożoności. Wykorzystano zatem algorytm TD-PSOLA (ang. *Time-Domain Pitch-Synchronous OverLap and Add*) [4]. Jest to metoda działająca wyłącznie w dziedzinie czasu oparta na przetwarzaniu sygnału mowy zgodnie z okresem podstawowym. Algorytm resyntezy difonów TD-PSOLA składa się z dwóch etapów: syntezy i analizy.

W procesie analizy pozyskiwane są z sygnału ramki rozmieszczone dokładnie co jeden okres podstawowy, zgodnie ze wzorem:

$$x_m(n) = x(t_m + n) \cdot w(n) \qquad n = 1 \ldots W \tag{1}$$

gdzie: W – długość okna analizy (okno Hamminga), x(n) – sygnał analizowanego difonu, w(n) – okno Hamminga, t_m – początek m-tej ramki

Następnie, w procesie resyntezy, pobrane ramki sygnału są poddane superpozycji z nakładkowaniem, z innym odstępem niż w oryginalnym sygnale. Efektem tej modyfikacji jest zmiana okresu podstawowego sygnału, skutkująca podniesieniem (gdy ramki są rozłożone z większym zagęszczeniem niż w oryginalnym sygnale) lub obniżeniem tonu (gdy ramki rozłożone są rzadziej).

Algorytm TD-PSOLA pozwala na płynną modyfikację tonu podstawowego, co jest niezwykle ważne dla kształtowania konturu intonacyjnego wypowiedzi.

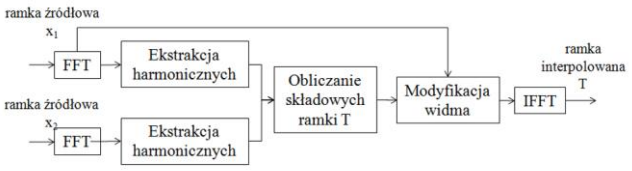
2.4. Konkatenacja i wygładzanie

Konkatenacja, czyli połączenie difonów ma istotny wpływ na jakość syntetyzowanego sygnału. Przy konkatenacji jednostek mowy możliwe jest ich niedopasowanie pod trzema względami:

- fazy, gdy faza końcowa difonu i faza początkowa difonu następnego nie są zgodne;
- tonu podstawowego, gdy sąsiadujące difony różnią się pod względem wysokości głosu, na jakiej zostały wypowiedzianie;
- obwiedni widma, gdy występują różnice w rozmieszczeniu formantów w sąsiednich difonach.

Konkatenacja difonów dokonywana jest za pomocą algorytmu PSOLA, co oznacza że difony dodawane są do siebie zgodnie z okresem podstawowym [4]. Gwarantuje to ciągłość tonu podstawowego. Ciągłość fazy jest również zagwarantowana, jeśli fazy początkowe i końcowe difonów zostały uwzględnione w procesie ekstrakcji segmentów z nagrania lektorskiego. Problemem pozostaje natomiast niedopasowanie obwiedni widmowej. Ze względu na długi czas trwania nagrania difonów, zdarza się, że barwa głosu lektora ulega wyraźnym zmianom dla różnych segmentów. Skutkuje to pogorszeniem jakości syntetyzowanego sygnału. Aby zredukować to zjawisko, konieczne jest zastosowanie wygładzania widmowego segmentów.

Większość opisywanych w literaturze algorytmów wygładzania widmowego opartych jest na interpolacji obwiedni LPC sygnału (ang. *Linear Predictive Coding* – liniowe kodowanie predykcyjne) lub metodzie DFW (*Dynamic Frequency Warping* – dynamiczne morfowanie częstotliwości) [5]. Proponowany algorytm jest oparty na modyfikacji składowych harmonicznych sygnału. Schemat blokowy algorytmu wygładzania widmowego został przedstawiony na rysunku 4.



Rys. 4. Schemat blokowy algorytmu wygładzania widmowego

Celem wygładzania widmowego jest wyznaczenie interpolowanej ramki połączenia sygnału (T), która zapewnia najlepsze dopasowanie dwóch ramek źródłowych (x_1, x_2), pochodzących z dwóch sąsiednich difonów. Aby zbadać strukturę widmową sygnałów źródłowych, obliczane jest widmo FFT ramek (ang. *Fast Fourier Transform* – szybkie przekształcenie Fouriera) i dokonywana jest ekstrakcja składowych sinusoidalnych. Następnie amplitudy kolejnych harmonicznych ramki docelowej są wyznaczane jako średnia geometryczna amplitud składowych harmonicznych ramek źródłowych z odpowiednimi wagami:

$$H_i^T = H_i^{x_1} \cdot \left[\left(H_i^{x_1} \right)^{w_1} \cdot \left(H_i^{x_2} \right)^{w_2} \right]^{\frac{1}{w_1 + w_2}} \qquad (2)$$

gdzie: H_i^T $H_i^{x_1}$, $H_i^{x_2}$ – kolejna harmoniczna sygnału odpowiednio ramki interpolowanej (T), ramki źródłowej x_1 i ramki źródłowej x_2, w_1 – waga ramki x_1, w_2 – waga ramki x_2

Po wyznaczeniu docelowych amplitud harmonicznych, dokonywana jest modyfikacja widma FFT ramki źródłowej x_1 w celu uzyskania struktury widma będącej najlepszym dopasowaniem źródłowych sygnałów.:

3. EKSPERYMENTY

W celu sprawdzenia jakości syntetyzowanej mowy przeprowadzono testy odsłuchowe [6][7]. Eksperci mieli za zadanie ocenić zrozumiałość i rozpoznać intonacje generowanych wypowiedzi.

3.1. Test zrozumiałości

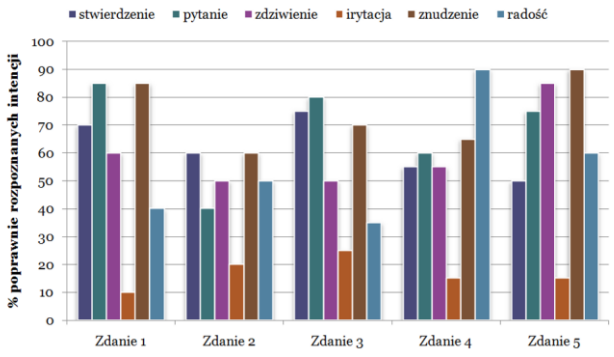
Aby zbadać zrozumiałość mowy generowanej przez syntezer przeprowadzono test zrozumiałości z wykorzystaniem logatomów i zdań [7]. W teście uczestniczyła grupa 12 ekspertów. W teście logatomowym wykorzystano 50 wyrazów pozbawionych znaczenia. W wyniku testu poprawnie zrozumianych zostało 73,16% logatomów. W teście zdaniowym wykorzystano 10 zdań. Wynik zrozumiałości wyniósł 97,65%. Na podstawie testu można stwierdzić, że zrozumiałość generowanej mowy jest wystarczająca dla typowych zastosowań systemu TTS.

3.2. Test modyfikacji prozodycznych

Modyfikacja prozodii wypowiedzi ma na celu nadanie zdaniu przebiegu parametrów prozodycznych charakterystycznych dla pewnej intencji mówcy. W tym celu przeprowadzono test zrozumiałości intencji syntetyzowanej wypowiedzi [6]. Jako materiał językowy wykorzystano 5 zdań. Każdemu ze zdań nadano w sposób syntetyczny prozodię mającą odwzorować 6 intencji:

- stwierdzenie,
- pytanie,
- zdziwienie,
- irytacja,
- znudzenie,
- radość.

Jako źródło konturów intonacyjnych posłużyła referencyjna wypowiedź lektorska. W teście wzięło udział 30 ekspertów. Wyniki zrozumiałości intencji wypowiedzi przedstawiono na wykresie na rysunku 5. Najlepiej rozpoznawanymi intencjami są stwierdzenie, pytanie i znudzenie. Znacznie mniejszy procent poprawnie rozpoznanych intencji otrzymano dla irytacji i radości. Eksperyment dowodzi jednak, że dzięki opracowanym algorytmom możliwe jest nadanie syntetycznej wypowiedzi podstawowych typów intonacji.



Rys. 5. Procent poprawnie zrozumianych intencji wypowiedzi [6]

4. PODSUMOWANIE

Przedstawiony system jest zdolny generować syntetyczną mowę o wysokiej zrozumiałości, której prozodia skutecznie odwzorowuje naturalną prozodię wypowiedzi. Syntezowana mowa wzbogacona o prozodię umożliwia skuteczniejszą interakcję człowieka z komputerem niż wypowiedź pozbawiona akcentów i intonacji. Opisywany system znajduje zastosowanie w aplikacjach edukacyjnych bądź terapeutycznych oraz w interfejsach multimodalnych ułatwiających interakcję użytkownika z komputerem. Przykładem jest aplikacja wykorzystująca system śledzenia fiksacji wzroku *CyberOko* i wirtualną klawiaturę [7][8]. Aplikacja ta przeznaczona jest dla osób niepełnosprawnych, niezdolnych do poruszania kończynami i mówienia. Wykorzystanie systemu śledzenia fiksacji i wirtualnej klawiatury wyświetlanej na ekranie komputera do wprowadzania tekstu umożliwia takim osobom generację komunikatów głosowych z wykorzystaniem syntetyzera mowy. Daje to możliwość werbalnej interakcji niepełnosprawnego użytkownika z otoczeniem.

5. PODZIĘKOWANIA

Badania dofinansowane w ramach projektu POIG.01.03.01-22-017/08 pt.:„Opracowanie typoszeregu komputerowych interfejsów multimodalnych oraz ich wdrożenie w zastosowaniach edukacyjnych, medycznych, w obronności i w przemyśle”

6. BIBLIOGRAFIA

1. Dutoit T.: An introduction to Text-to-Speech synthesis, 129-170, Kluwer Academic Publishers, Dordrecht, 1997.
2. Johnson M.: Synthesis of English Intonation using explicit models of reading and spontaneous speech, 4th Int. Conf. on Spoken Language, 3, 1844-1847, 3-6.10.1996, Philadelphia.
3. Laroche J., Dolson M.: Improved phase vocoder time-scale modification of audio, IEEE Trans. on Speech and Aud. Proc., 7,3, New York.
4. Moulines E., Charpentier F., Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones, Speech Communication, 453-467, North-Holland.
5. D. Chappell, J. Hansen: A comparison of spectral smoothing methods for segment concatenation based speech synthesis, Speech Communication, 36, 343-374, North-Holland, 2002.
6. K. Łopatka, P. Suchomski, A. Czyżewski: Time-domain prosodic modification for Text-To-Speech synthesizer, IEEE Conf. on Signal processing algorithms, architectures, arrangements and applications SPA 2010, 73-77, 23-25.09.2010, Poznań.
7. A. Czyżewski, K. Łopatka, B. Kunka, R. Rybacki, B. Kostek: Speech synthesis controlled by eye gazing, 129th Convention of the Audio Engineering Society, 04-07.11.2010, San Francisco.
8. B. Kunka, B. Kostek, M. Kulesza, P. Szczuko, A. Czyzewski: Gaze-tracking-based audio-visual correlation analysis employing quality of experience methodology, Intelligent Decision Technologies, vol. 4, No. 3, pp. 217-227, 2010.

TEXT-TO-SPEECH SYNTHESIZER EMPLOYING AUTOMATIC PROSODIC MODIFICATION

Key-words: speech synthesis, prosody, PSOLA

Abstract: The paper presents a Text-To-Speech synthesizer of Polish language employing automatic prosodic modification. The method used for synthesizing the speech signal is concatenative synthesis using constant-length segments – diphones. The subsequent modules of the synthesizer are introduced. Employed language analysis and signal processing techniques are described. The synthesized speech yields high intelligibility and naturalness, which is proved by auditory tests. The proposed system can be used in educational and therapeutic applications or multimodal interfaces for disabled people.