STUDIA INFORMATICA 2003
Volume 24 Number 4 (56)

Piotr FABIAN
Politechnika Śląska, Instytut Informatyki

WYBÓR JEDNOSTEK ELEMENTARNYCH DLA SYSTEMU SYNTEZY MOWY[1]

Streszczenie. Główne metody syntezy mowy to metody parametryczne z interpolacją parametrów i konkatenacyjne z zestawianiem wypowiedzi w wybranej dziedzinie z fragmentów istniejących nagrań. Zestawianie daje tym lepsze efekty, im dłuższe są jednostki w odpowiednich kontekstach. Zgromadzenie odpowiednio dużej bazy elementarnych nagrań (polifonów) pozwala zastosować drugą metodę do syntezy mowy w języku polskim. Jakość istniejących syntezatorów dla mniej popularnych języków, np. polskiego, jest znacznie niższa niż uzyskana dla najczęściej badanego języka angielskiego. Przedstawiona koncepcja automatycznego budowania bazy polifonów pozwala na szybką optymalizację bazy jednostek fonetyczno-akustycznych do celów syntezy mowy.

Słowa kluczowe: synteza mowy, synteza konkatenacyjna, polifony

BASIC UNITS SELECTION FOR A SPEECH SYNTHESIS SYSTEM

Summary. Two common methods of speech synthesis are parametric synthesis and concatenation of basic speech units. Concatenation sticks speech units together in selected domain. The quality of the speech synthesis grows with the length of basic speech units in the vocabulary: one of possible solutions would be ideally to record a large corpus of continuous speech. Collecting a set of elementary speech units, like polyphones, makes possible to use the second method for the Polish language. Speech synthesis is not a new problem, there are many commercial products. But the quality of them for less popular languages, like Polish, is much worse than for the most popular English. The presented approach makes possible a fast optimization of a speech units database for speech synthesis.

Keywords: speech synthesis, concatenative speech synthesis, polyphones

[1]Praca wykonana w Instytucie Informatyki Pol. Śl. w ramach badań BW-486/RAu2/2002.

1. Wprowadzenie

Jedną z barier stosowania komputerów w nowych dziedzinach jest sposób komunikacji z maszyną. Zwykle stosowane rozwiązania w postaci konsoli tekstowej lub graficznej z klawiaturą i myszką w niektórych zastosowaniach są niewygodne, zbyt wolne lub niemożliwe do zastosowania. Naturalnym rozwiązaniem uzupełniającym lub zastępującym komunikację z maszyną stanie się prawdopodobnie w niedalekiej przyszłości wykorzystanie systemów rozpoznawania i syntezy mowy, przynajmniej w takich zastosowaniach, jak: przetwarzanie tekstów, systemy dialogowe, sterowanie aplikacjami. Ciągły wzrost szybkości dostępnych obecnie tanich komputerów klasy PC stwarza możliwość realizacji takiego interfejsu. Niniejsze opracowanie przedstawia metodę budowy bazy nagrań elementarnych jednostek mowy, która może być zastosowana do syntezy mowy. Omówiono także przykładową realizację syntezatora konkatenacyjnego stosującego polifony, tzn. fonemy z uwzględnieniem kontekstu, w jakim występują.

2. Rozpoznawanie

W wyniku prac prowadzonych przez autora nad rozpoznawaniem mowy powstał model statystyczny dla języka polskiego, pozwalający rozpoznawać wypowiedzi w języku polskim przy zastosowaniu metody ukrytych modeli Markowa (ang. *Hidden Markov Model*, HMM) [1].

Badania nad rozpoznawaniem były prowadzone na stosunkowo dużym zbiorze nagrań Corpora, dostępnym dzięki uprzejmości dra Stefana Grocholewskiego z Politechniki Poznańskiej [4]. Baza Corpora zawiera ponad 16 tysięcy nagrań, tj. wypowiedzi o długości rzędu kilku sekund, pochodzących od 44 różnych mówców. Duży rozmiar bazy nagrań pozwolił m.in. na zbadanie poprawności rozpoznawania dla różnych mówców. Zgromadzenie nagrań i możliwość ich automatycznego rozpoznania, opisana m.in. w [9], pozwoliły zbudować bazę polifonów przeznaczoną do zastosowania w syntezie mowy.

3. Metody syntezy mowy

Próby generacji sztucznego sygnału mowy podjęto już w XIX w. metodami elektrycznymi, a jeszcze wcześniej mechanicznymi. Efekty nie były zadowalające, chociaż badania nad analizą mowy doprowadziły w XIX w. m.in. do powstania telefonu. Dopiero

pojawienie się komputerów pozwoliło na konstrukcję systemów odpowiedniej jakości, znajdujących zastosowanie w praktyce.

Wyróżnić można dwie klasy metod generacji mowy:

- artykulacyjną,
- konkatenacyjną.

Algorytmy realizujące pierwszą metodę naśladują sposób działania toru głosowego człowieka, tzn. symulują zjawiska fizyczne zachodzące w poszczególnych elementach tego toru. Metoda ta wymaga dokładnej znajomości budowy toru głosowego, aby symulacja zachowania tego toru jak najwierniej oddała brzmienie głosu. Na uzyskanie dźwięków przypominających mowę pozwalają już uproszczone modele toru głosowego, ale uzyskanie wysokiej jakości syntezy jest tu stosunkowo trudne. Za odrębną (trzecią) klasę syntezatorów uznawane są czasem syntezatory formantowe. Jednak ich działanie jest zbliżone do syntezatorów artykulacyjnych, jedynie modelowanie nie odzwierciedla dokładnie budowy toru głosowego, ale jego wpływ na widmo generowanego sygnału. Metody formantowe dają głos brzmiący nieco sztucznie. Syntezator formantowy zawiera generator tonu krtaniowego i szumu oraz sekwencyjne połączenie kilku filtrów kształtujących charakterystykę częstotliwościową generowanego sygnału tak, aby przekazać informację o częstotliwościach formantów i szerokości ich pasm. Zależnie od sposobu implementacji takiego syntezatora uwzględniane są częstotliwości rezonansowe toru głosowego lub również takie, dla których następuje tłumienie sygnału.

Drugą klasą syntezatorów są syntezatory konkatenacyjne [5]. Tutaj zwykle odtwarzane są wcześniej nagrane próbki mowy, zestawiane sekwencyjnie w dziedzinie czasu. Od doboru podstawowych jednostek mowy i sposobu ich zestawiania zależy jakość syntezy. Dalsza część rozważań dotyczy syntezatorów tej klasy.

Podstawowe elementy, z których zestawiane są wypowiedzi, mogą być różnej długości: od pojedynczych fonemów do całych zdań. Zaletą krótkich elementów są małe wymagania dotyczące bazy nagrań, wadą natomiast problemy z uwzględnianiem wariantów artykulacyjnych. Zaletą długich elementów, np. całych słów, jest wysoka jakość uzyskiwanej syntezy, wadą natomiast konieczność zgromadzenia dużej bazy nagrań i małe możliwości wpływu na brzmienie syntezowanej mowy. Pośrednie długości to np. bifony, trifony, sylaby. Systemy stosujące bardzo długie elementy często nie są nawet nazywane syntezatorami, jak np. oprogramowanie odtwarzające w sieciach telefonicznych komunikaty dla abonentów.

Konstrukcja syntezatora konkatenacyjnego wymaga więc kompromisu między jakością syntezy i związaną z tym koniecznością zebrania odpowiednio dużej liczby nagrań wzorcowych a realnymi możliwościami zgromadzenia takich nagrań.

4. Rozmiar bazy nagrań

Przyjmując słownik fonemów składający się z k elementów, baza nagrań dla syntezatora konkatenacyjnego stosującego pojedyncze fonemy powinna zawierać co najmniej k nagrań elementarnych. W takim przypadku długość kontekstu $m = 1$. Dłuższymi elementarnymi jednostkami mogą być bifony ($m = 2$), trifony ($m = 3$) i ogólnie polifony.

Oznaczając przez m zadaną długość kontekstu, w ogólnym przypadku baza nagrań powinna zawierać k^m nagrań, aby umożliwić syntezę dowolnego ciągu fonemów z uwzględnieniem kontekstu. Przyjmując dla języka polskiego liczbę elementów słownika $k = 37$ (taką wartość stosowano przy rozponawaniu mowy), uzyskujemy dla bifonów 1369-elementową bazę nagrań. Dla trifonów liczba elementów przekracza 50000, dla polifonów odpowiednio więcej. Dopiero trifony dają możliwość uwzględnienia obustronnego kontekstu danego fonemu.

W przypadku rzeczywistego języka nie wszystkie sekwencje fonemów są dopuszczalne lub nie występują w wypowiedziach, które mają być syntezowane. Zmniejsza to objętość bazy nagrań, ale mimo to liczba nagrań wciąż pozostaje duża.

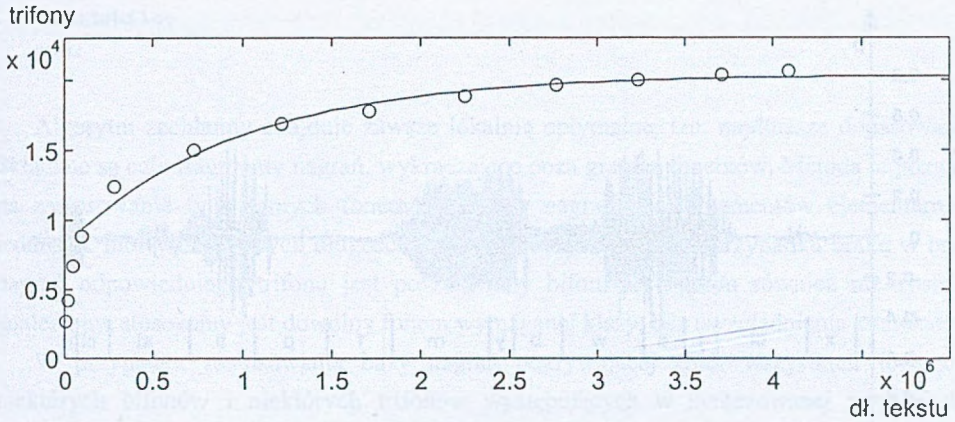
Tabela 1 przedstawia zależność liczby różnych trifonów od długości tekstu dla wybranego tekstu w języku polskim, „Biblii tysiąclecia". Tekst Biblii został podzielony na zdania i poddany automatycznej transkrypcji fonetycznej. Wynikowe ciągi fonemów pogrupowano w trifony i usunięto ich powtórzenia.

Tabela 1

Liczba różnych trifonów w funkcji długości tekstu

Długość tekstu	Liczba trifonów
10 KB	2703
50 KB	6700
500 KB	13740
4000 KB	20553

Zależność liczby różnych trifonów od długości tekstu zbliżona jest do funkcji $y(t) = a - be^{-ct}$, gdzie t jest długością tekstu. Początkowo szybko rośnie wraz ze wzrostem długości tekstu, dalej wolniej. Oszacowanie parametrów tej funkcji metodą najmniejszych kwadratów daje a=20334; do tej wartości dąży $y(t)$ dla $t \to \infty$. W polskich tekstach można więc spodziewać się około 20000 trifonów. Na rys. 1 przestawiono omawianą zależność.



Rys. 1. Liczba różnych trifonów w funkcji długości tekstu
Fig. 1. The dependence of unique triphones on the text length

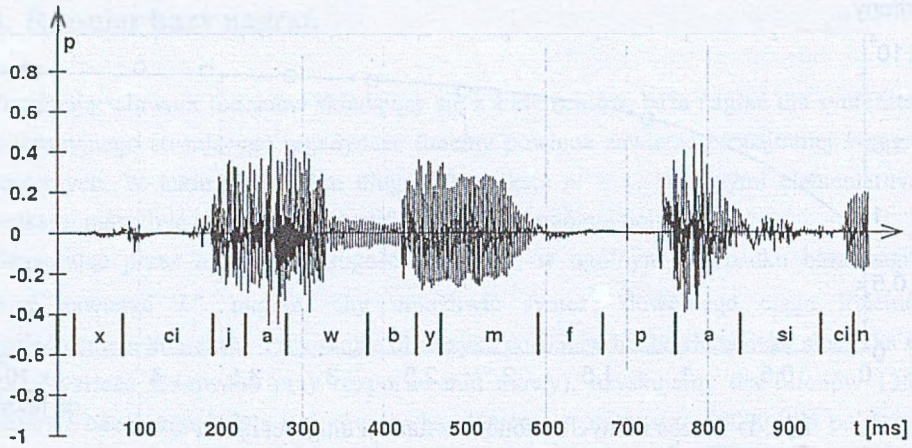
Ręczne przygotowanie tak dużej bazy nagrań jest kłopotliwe i czasochłonne. Opisany wcześniej system rozpoznawania mowy daje jednak możliwość precyzyjnego określenia granic czasowych poszczególnych fonemów w rozpoznawanych nagraniach. Dlatego może zostać zastosowany do automatycznej budowy bazy nagrań.

5. Etykietowanie nagrań

Rysunek 2 przedstawia dokonane automatyczne etykietowanie wykonane dla przykładowego nagrania. Etykietowanie zostało wykonane przy zastosowaniu narzędzi z biblioteki HTK [7] i opracowanego przez autora modelu statystycznego wypowiedzi w języku polskim.

Zaznaczone granice między fonemami ustawiane są z rozdzielczością 10 ms. Oczywiście wskazanie precyzyjnej granicy między fonemami nie jest możliwe, gdyż parametry sygnału mowy zmieniają się w sposób ciągły. Granice dobrze jednak lokalizują fragmenty nagrania o cechach najbardziej zbliżonych do cech rozpoznanego fonemu.

W wyniku etykietowania zestawu 110 nagrań wypowiedzi dla 44 mówców z bazy Corpora oznaczono 129140 fonemów.



Rys. 2. Etykietowanie nagrania wypowiedzi z bazy nagrań Corpora
Fig. 2. Labeling of a waveform from the Corpora speech recordings

6. Dopasowanie elementarnych jednostek mowy

Duża liczba nagrań w bazie koniecznych nawet dla systemu stosującego krótki kontekst (trifony) sprawiła, że opracowany został algorytm pozwalający zbudować bazę składającą się z elementów o różnej długości i zastosować je do syntezy.

Algorytm ten jest algorytmem zachłannym, poszukującym wśród nagrań dostępnych w bazie najdłuższych ciągów fonemów, z których zestawiana jest syntezowana wypowiedź. Można opisać go następującym pseudokodem:

```
procedure dopasuj(s, n);
{ s jest łańcuchem zawierającym tekst do syntezy, n tekstami z bazy nagrań }
begin
  tranksrypcja_fonetyczna(s); { dokonanie transkrypcji s }
  transkrypcja_fonetyczna(n); { oraz n }
  while (długość(s)>0) do
    w := s;
    len := długość(w);
    while ((len>0) and (p=znajdź(prefiks(w,len), n)=BRAK)) do
      len := len - 1;
    endwhile;
    if (len=0) then
      błąd(BrakFonemu)
    endif;
    wyślij(p,len);
    usuń_z_lewej(s,len);
```

```
    endwhile;
end;
```

Algorytm zachłanny znajduje zawsze lokalnie optymalne, tzn. najdłuższe dopasowanie. Składane są całe fragmenty nagrań, wykraczające poza granice fonemów. Metoda ta pozwala na zastosowanie tych samych fonemów z bazy nagrań jako fragmentów elementarnych jednostek mowy dla różnych długości kontekstu. Przykładowo, w przypadku braku w bazie nagrań odpowiedniego trifonu jest poszukiwany bifon. Jeśli bifon również nie zostanie znaleziony, stosowany jest dowolny fonem wymaganej klasy, bez uwzględniania kontekstu.

W przypadku zastosowania bazy nagrań pokrywającej zbiór wszystkich fonemów, niektórych bifonów i niektórych trifonów występujących w syntezowanej wypowiedzi, wyniki działania algorytmu wyglądają następująco:

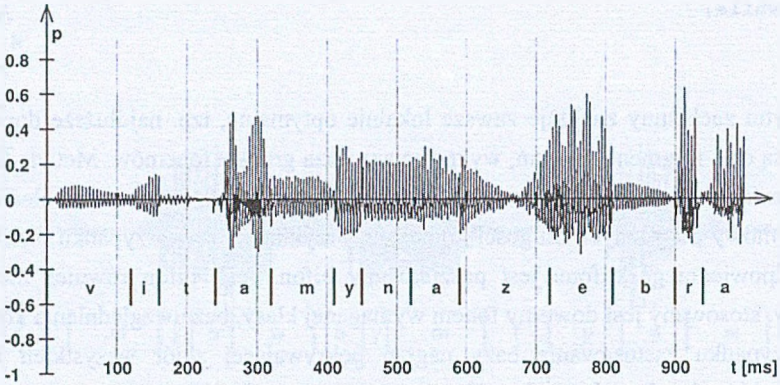
- Tekst do syntezy:
  witamy na zebraniu instytutu
- Automatyczna transkrypcja fonetyczna SAMPA (fonemy) [2]:
  `v i t a m y n a z e b r a n i u i n s t y t u t u`
- Automatyczna transkrypcja z zastosowaniem kontekstu (trifony):
  `v-i v+i-t i+t-a t+a-m a+m-y m+y-n y+n-a n+a-z a+z-e z+e-b e+b-r b+r-a r+a-ni a+ni-u ni+u-i u+i-n ...`
  Wyznaczona sekwencja nagrań z bazy (numery nagrań):
  333 334 176 456 123 102 1811 230 49 50 357 1873 603 1519 2 1747 2508 710 711 717 1630 147 148 2

Kolejne numery w tej sekwencji (np. 710, 711) oznaczają odnalezienie w bazie nagrań sekwencji o odpowiednio długim kontekście.

Fragmenty nagrań z bazy są zestawiane w dziedzinie czasu. W miejscu połączenia kolejnych fragmentów występuje zwykle nieciągłość przebiegu czasowego sygnału, dająca krótkie, szerokopasmowe zakłócenie. Możliwe jest złagodzenie tego efektu poprzez okienkowanie składanych nagrań, sprowadzające do zera wartość sygnału na granicach okienek lub przesuwanie granic fonemów do chwil, w których bieżąca wartość sygnału jest zerowa. To drugie rozwiązanie nie zapewnia całkowitej likwidacji zakłócenia, ale subiektywnie daje lepszy efekt niż okienkowanie.

Rysunek 3 przedstawia przykładowy przebieg czasowy sygnału mowy - wynik automatycznej syntezy dokonanej przy zastosowaniu utworzonej bazy nagrań.

Transkrypcja fonetyczna dokonywana jest automatycznie z zastosowaniem reguł fonetycznych dla języka polskiego [8]. Stosuje zmodyfikowaną notację SAMPA o mniejszej liczbie fonemów w stosunku do tzw. transkrypcji słowiańskiej [3]. W zamian stosowanie kontekstów pozwala uwzględnić różnice artykulacyjne.



Rys. 3. Wynik syntezy mowy (1 s)
Fig. 3. The result of speech synthesis (1 s)

7. Ocena jakości

W celu oceny jakości syntezowanych wypowiedzi porównano wyniki z wypowiedziami syntezowanymi przez inne dostępne na rynku syntezatory. Dla języka polskiego były to: „Spiker" (Ivo Software), „SynTalk" (Neurosoft), „Lektor" (Drive), „SM23" (Pol. Śl.) i „Speak" (Altix). Prawdopodobnie wszystkie są syntezatorami artykulacyjnymi, „Lektor" jest syntezatorem formantowym. Syntezatory generują zrozumiałą mowę, chociaż niektóre wymagają pewnego przyzwyczajenia. Najlepszą jakością wykazuje się „Spiker". Ocena jakości syntezy musi tu być subiektywna. Według autora, zrozumiałość syntezy konkatenacyjnej dokonywanej z zastosowaniem bazy zawierającej około 3000 fonemów z oznaczonym kontekstem jest porównywalna z wymienionymi syntezatorami, lepsza od najsłabszych z nich.

Wśród syntezatorów dla innych języków szczególnie wysoką jakością wyróżnia się „Next-Gen TTS" firmy AT&T [6]. Jest to syntezator o mieszanej architekturze opracowany tak starannie, że praktycznie nie da się odróżnić wyników syntezy od rzeczywistych nagrań mowy.

8. Wnioski i podsumowanie

Przedstawiona metoda budowy bazy nagrań pozwoliła na szybkie zgromadzenie danych niezbędnych do syntezy mowy. Przygotowanie bazy obejmującej 3000 fonemów dla nowego mówcy zajmuje przy zastosowaniu tej metody kilkadziesiąt minut. Utworzona baza zawiera nagrania pełnych wypowiedzi, wyznaczone granice czasowe fonemów, ich opis wraz z

kontekstem (trifony) oraz procedurę udostępniającą wybrany ciąg fonemów. Dołączenie nowych elementów do bazy wymaga jedynie uzupełnienia transkrypcji fonetycznych i ponownej realizacji algorytmu dokonującego etykietowania. Jest to istotne dla zwiększania średniej długości kontekstu fonemów stosowanych w syntezie. Synteza konkatenacyjna daje najlepsze wyniki dla długich jednostek mowy, stąd dążenie do uzyskania obszernej bazy nagrań.

Łączenie elementarnych jednostek mowy w dziedzinie czasu z prostym wygładzaniem daje pewne zakłócenia. Prawdopodobnie mogą one zostać w większym stopniu wyeliminowane przez zastosowanie kombinowanych metod wygładzania w dziedzinie czasu i częstotliwości. W bieżącej wersji syntezator może znaleźć zastosowanie jako uzupełnienie opracowywanego w Instytucie Informatyki Politechniki Śląskiej systemu wizualizacji języka migowego.

LITERATURA

1. Rabiner L., Juang B. H.: Fundamentals of Speech Recognition, Prentice Hall, Englewood Cliffs 1993.
2. Wells J. C.: Computer-coding the IPA: a proposed extension of SAMPA, Department of Phonetics and Linguistics, University College London, 1995.
3. Nagórko A.: Zarys gramatyki polskiej. PWN. Warszawa 1996.
4. Grocholewski S.: Akustyczna baza danych dla języka polskiego CORPORA, Raport Instytutu Informatyki Politechniki Poznańskiej RB 12/97, Poznań 1997.
5. Beutnagel M., Conkie A., Syrdal A.: Diphone synthesis using unit selection, The 3rd ESCA/COCOSDA Workshop on Speech Synthesis, Paper F.2 (R52), Australia 1998.
6. Beutnagel M., Conkie A., Schroeter J., Stylianou Y., Syrdal A.: The AT&T Next-Gen TTS System, Joint Meeting of ASA, EAA and DAGA, Berlin 1999.
7. Young S., Kersha D., Odell J., Ollason D., Vatchev V., The HTK Book, Microsoft Corporation, 2000.
8. Ostaszewska D., Tambor J.: Fonetyka i fonologia współczesnego języka polskiego. PWN, Warszawa 2000.
9. Fabian P.: Efektywność rozpoznawania mowy dla języka polskiego w zależności od rozmiaru bazy próbek. Studia Informatica, Gliwice 2002.

Recenzent: Dr hab. inż. Grażyna Demenko

Wpłynęło do Redakcji 30 stycznia 2003 r.

Abstract

This article presents an approach to data preparation for a speech synthesis system. Speech synthesizers developed currently over the world usually use parametric methods or concatenation of basic speech units. The first class requires appropriate algorithms interpolating parameters across the generated utterance, while the second sticks speech units together in selected domain. The quality of the speech synthesis grows with the length of basic speech units in the vocabulary. It would be ideally to have all possible utterances prerecorded. Collecting a set of elementary speech units, like polyphones, makes possible to use the second method for the Polish language. The presented method automatically creates a set of basic speech units given a prerecorded set of utterances with transcriptions. It has been applied to the Corpora speech recordings, containing over 16000 utterances from 44 speakers. Over 129000 phonemes have been labeled, regarding the context. This database has been used for speech synthesis with speech units of varying lengths. The synthesized speech, although not comparable to most advanced parametric synthesizers, is intelligible and may be used as a part of a translation system developed in the CS division. The presented approach makes possible a fast build of a speech synthesizer with given properties.

Adres

Piotr FABIAN: Politechnika Śląska, Instytut Informatyki, ul. Akademicka 16,
44-101 Gliwice, Polska, piotrcf@star.iinf.polsl.gliwice.pl .