STUDIA INFORMATICA 2003
Volume 24 Number 1 (52)

Piotr FABIAN
Politechnika Śląska, Instytut Informatyki

EFEKTYWNOŚĆ ROZPOZNAWANIA MOWY W ZALEŻNOŚCI OD ROZMIARU DANYCH WZORCOWYCH

Streszczenie. W artykule przedstawiono zastosowanie biblioteki HTK do budowy sieci probabilistycznej przeznaczonej do rozpoznawania wypowiedzi w języku polskim. Opisano konstrukcję sieci na poziomie fonemów, transkrypcji fonetycznej i gramatyki języka. Opisano metodę automatycznego tworzenia sieci dla różnych zbiorów wzorcowych i wyniki rozpoznawania dla zbiorów o różnej liczbie elementów.

Słowa kluczowe: mowa, rozpoznawanie, HTK.

DEPENDENCE OF THE SPEECH RECOGNITION ACCURACY ON THE TRAINING DICTIONARY SIZE

Summary. This article presents an approach of building a speech recognizer for the Polish language using the HTK toolkit. All levels of the constructed recognition network have been described: phoneme models, transcription and language grammar. A method of automated building of such networks for different training sets has been presented, with results of recognition for training sets of varying sizes.

Keywords: speech, recognition, HTK.

1. Wstęp

Najczęściej obecnie stosowaną metodą rozpoznawania mowy jest modelowane języka przez sieć probabilistyczną (ukryte modele Markowa, ang. *Hidden Markov Model*, HMM). Konfiguracja takiej sieci i jej parametry zależą od realizowanego zadania rozpoznawania. Ukryte modele Markowa znajdują zastosowanie w rozpoznawaniu różnego rodzaju wzorców, nie tylko mowy [1]. Ze względu na uniwersalność metody ukrytych modeli Markowa,

powstały programy narzędziowe wspomagające tworzenie takich sieci probabilistycznych, wyznaczanie ich parametrów i zastosowanie w rozpoznawaniu. Jedną z bibliotek tego rodzaju jest HTK, zestaw narzędzi rozwijany od 1989 r. w Cambridge University, obecnie przez firmę Entropic.

Niniejsze opracowanie przedstawia zastosowanie tej biblioteki do konstrukcji sieci przystosowanej do rozpoznawania wypowiedzi w języku polskim i osiągnięte wyniki rozpoznawania przy zastosowaniu różnych zbiorów danych wzorcowych.

2. Biblioteka HTK

Biblioteka HTK jest implementacją znanych algorytmów estymacji parametrów modeli Markowa (Baum-Welch, programy *HERest* i *HRest*) i znajdowania optymalnej sekwencji stanów (Viterbi, program *HVite*). Pomaga wyznaczyć wstępne parametry modeli HMM (programy *HCompV* i *HInit*). Pozwala w wygodny sposób budować sieć Markowa opisując jej strukturę w postaci plików tekstowych. Cała sieć wraz z parametrami również może być zapisana jako plik tekstowy, ewentualnie równoważny plik binarny. Dostępne są narzędzia wykonujące operacje edycyjne na takich plikach: usuwanie modeli, łączenie stanów, tworzenie makr przeznaczonych do użycia w opisach różnych modeli, przeglądanie plików binarnych itp. W skład biblioteki wchodzi też program wyznaczający wybrane cechy sygnału mowy. Jedynym narzędziem graficznym jest prosta przeglądarka plików z ciągami próbek, pozwalająca zaznaczać podciągi i nadawać im etykiety (nazwy).

Programy narzędziowe wchodzące w skład HTK mogą być używane w różnych środowiskach, m.in. MS Windows i Unix/Linux. W przedstawianych badaniach stosowana była wersja 3.0 biblioteki, uruchamiana w systemie Linux.

3. Elementy modelu dla języka polskiego

Model probabilistyczny utworzony na potrzeby eksperymentu przyjmuje, że węzeł sieci modelu Markowa opisuje fragment fonemu w określonym kontekście. Pojedynczy fonem, w przypadku uwzględnienia sąsiedztwa lewego i prawego nazywany trifonem, reprezentowany jest jako ciąg trzech stanów (węzłów) modelu Markowa emitujących oraz dodatkowych dwóch stanów nie generujących żadnych wartości. Fonemy uwzględniające kontekst ogólnie nazywane są polifonami. Słowa są zestawiane z trifonów, pełne wypowiedzi są ciągami słów. Poszczególne poziomy mogą być definiowane niezależnie. Ostatecznie

powstaje sieć z węzłami emitującymi wektory cech, dostosowana do rozpoznawania określonego rodzaju wypowiedzi.

Dla potrzeb rozpoznawania wypowiedzi w języku polskim został zdefiniowany hierarchiczny model sieci probabilistycznej. Poszczególne poziomy opisują fonemy, słowa i prostą gramatykę.

4. Opis fonemów, notacja SAMPA

Do opisu poziomu fonemów zastosowano zmodyfikowaną notację SAMPA (ang. *Speech Assessment Methods Phonetic Alphabet*) [2, 6], której opracowywanie rozpoczęto w ramach projektu ESPRIT 1541 w 1987 r. Wbrew nazwie nie jest to alfabet, gdyż fonemy opisywane są ciągami znaków, a nie pojedynczymi symbolami. Notacja ta zastępuje tradycyjną notację fonetyczną, w której stosowane są symbole spoza standardowego zestawu ASCII ze względu na przyjęcie zasady „jeden dźwięk – jeden znak" i brak odpowiednio dużej liczby znaków ASCII. W polskiej literaturze często stosowany jest międzynarodowy alfabet fonetyczny IPA (ang. *International Phonetic Alphabet*), opracowany i uaktualniany przez Międzynarodowe Towarzystwo Fonetyczne IPA (ang. *International Phonetic Association*). Dla języka polskiego wystarczający jest prawie 80-elementowy podzbiór IPA (tzw. transkrypcja słowiańska) [4].

W notacji SAMPA wybrane symbole alfabetu IPA zapisano jako ciągi znaków ASCII, dzięki czemu można je bez przeszkód stosować w systemach komputerowych. Początkowo zdefiniowano notację dla najczęściej używanych języków w Unii Europejskiej, język polski został dołączony w 1996 roku w ramach projektu Babel. Notacja SAMPROSA (ang. *SAM Prosodic Alphabet*) zawiera opisy pomocniczych cech, akcentów i intonacji. Notacja X-SAMPA [3] jest propozycją zapisu wszystkich znaków IPA w postaci ciągów znaków ASCII.

SAMPA dla języka polskiego przewiduje 37 fonemów, w tym 8 samogłosek i 29 spółgłosek. Tabela 1 przedstawia kilka przykładowych fonemów w przyjętej na potrzeby kodowania programu wersji TP1 notacji SAMPA i przykłady transkrypcji. Proponowana przez autora notacja TP1 obejmuje w podstawowej wersji te same fonemy, co SAMPA dla języka polskiego, z nieznacznymi różnicami ich zapisu. W TP1 stosowane są wyłącznie litery [a-z] i nie jest istotna ich wielkość. Ani SAMPA, ani TP1 nie są kodami przedrostkowymi. Nie jest to jednak wymagana właściwość, ponieważ kody fonemów podczas przetwarzania zawsze są od siebie oddzielane separatorami.

Podczas prac nad przygotowaniem transkrypcji fonetycznej nagrań został opracowany program dokonujący automatycznej konwersji zapisu ortograficznego na notację SAMPA i

TP1, z uwzględnieniem zjawisk zachodzących w języku polskim, m.in. zmiękczania głosek i utraty dźwięczności spółgłosek w niektórych przypadkach.

Tabela 1

Notacje Sampa i TP1 (wybrane fonemy)

Notacja		Przykład transkrypcji		
Sampa	TP1	Ortograficzna	Sampa	TP1
i	i	bit	b-i-t	b-i-t
I	y	byt	b-I-t	b-y-t
z	z	zen	z-e-n	z-e-n
S	sz	szal	S-a-l	sz-a-l
Z	rz	żak	Z-a-k	rz-a-k
s'	si	ślad	s'-l-a-t	si-l-a-t
z'	zi	źle	z'-l-e	zi-l-e
x	x	hak	x-a-k	x-a-k
ts	c	cud	ts-u-t	c-u-t
w	w	łza	w-z-a	w-z-a

5. Baza nagrań Corpora

Wyznaczenie parametrów sieci Markowa wymaga dużej liczby danych wzorcowych. W badaniach zastosowano bazę nagrań Corpora, opracowaną w Instytucie Informatyki Politechniki Poznańskiej i udostępnioną przez dr Stefana Grocholewskiego. Baza obejmuje nagrania pochodzące od 45 osób (28 mężczyzn, 11 kobiet i 6 dzieci), po 365 nagrań od każdej. Zestaw nagrań składa się ze 114 kilkuwyrazowych zdań oraz imion, cyfr, komend i alfabetu. Zdania dobrane są pod kątem pokrycia zbioru polskich fonemów i nie stanowią sensownych wypowiedzi, na podstawie których można budować model języka (np. *„ten górnik wyżył z ochłapu"*, *„czy upewnić się, czy łże"*, *„od ilu wolniej, ślepcze"*). Nagrań dokonano z częstotliwością próbkowania 16 kHz, rozdzielczością 16-bitową i zapisano w formacie WAVE PCM. Łącznie baza Corpora zawiera ponad 10 godzin nagrań. W użytej tu wersji podstawowej nie są dostępne transkrypcje.

6. Słownik i model gramatyki

Słownik transkrypcji fonetycznej utworzony na potrzeby badań obejmuje zarówno wszystkie 652 słowa zawarte w bazie Corpora, jak i słowa, które mają być rozpoznawane, a

nie występują w bazie. W słowniku zastosowano notację TP1 dla transkrypcji fonetycznej. Tabela 2 przedstawia fragment tego słownika.

Należy zaznaczyć, że mimo mniejszej liczby fonemów w notacji SAMPA niż w transkrypcji słowiańskiej, docelowa transkrypcja zawiera bifony i trifony, co pozwala wyróżnić różne warianty artykulacyjne w liczbie większej niż przewiduje to transkrypcja słowiańska. Trifony uwzględniają kontekst, w jakim występuje dany fonem, i rozszerzają zbiór elementów stosowanych w transkrypcji do 37^3.

Tabela 2

Przykład automatycznie dokonanej transkrypcji fonetycznej w notacji TP1

RYBACTWO	r-y-b-a-c-t-f-o-sp	SAMĄ	s-a-m-on-sp
RYSZARD	r-y-sz-a-r-t-sp	SĄ	s-on-sp
RZĘSACH	rz-en-s-a-x-sp	SEBASTIAN	s-e-b-a-s-t-j-a-n-sp
SABINA	s-a-b-i-n-a-sp	SENSEM	s-e-n-s-e-m-sp

Model gramatyki opisuje dopuszczalne sekwencje słów, jakie mogą być generowane przez sieć probabilistyczną i w konsekwencji rozpoznawane. Model gramatyki dla pojedynczej wypowiedzi (np. zdania) reprezentowany jest przez graf skierowany z wyróżnionymi wierzchołkami początkowym i końcowym. Wierzchołkami grafu są słowa, krawędzie oznaczają możliwe następstwa słów. Z krawędzią związane jest prawdopodobieństwo przejścia. Graf może też zawierać dodatkowe, pomocnicze wierzchołki, z którymi nie są związane żadne słowa. Może również zawierać cykle.

Postać grafu zależy od rozpoznawanych sekwencji słów. W ogólnym przypadku może to być graf pełny, dopuszczający dowolne ich sekwencje (wtedy można przyjąć, że model gramatyki nie jest stosowany). W badaniach były stosowane grafy pełne oraz opisujące wypowiedzi z określonego zakresu tematycznego.

Tabela 3

Fragment opisu grafu definiującego gramatykę języka

I=0	W=SENT-END	J=0	S=3 E=0
I=4	W=ŻYŁ	J=1	S=4 E=0
I=5	W=ŻYĆ	J=2	S=5 E=0
I=6	W=ŻYCIE	J=3	S=6 E=0

Graf jest definiowany w postaci opisu tekstowego, którego przykład przedstawia tabela 3. Pierwsza część opisu zawiera listę wierzchołków, z którymi są związane słowa identyfikowane numerami. Druga część opisu grafu zawiera listę krawędzi opisywanych numerami początkowych (S) i końcowych (E) wierzchołków. Opcjonalnie oznaczane jest prawdopodobieństwo przejścia z wierzchołka S do E. Wartość ta stosowana jest na etapie

rozpoznawania mowy. W badaniach stosowane były gramatyki opisujące pytania o zadanej składni oraz gramatyki generowane automatycznie na podstawie danego zbiory wypowiedzi.

7. Badania

Badania miały na celu sprawdzenie zależności poprawności rozpoznawania wypowiedzi od rozmiaru zbioru danych wzorcowych, na podstawie których budowana była sieć probabilistyczna. Do budowy sieci i rozpoznawania zastosowano bibliotekę HTK, skompilowaną w systemie Linux. Ze względu na konieczność wielokrotnego tworzenia sieci dla różnych zbiorów danych wzorcowych, zaimplementowano zestaw skryptów automatycznie wybierających nagrania z bazy Corpora, transkrypcje nagrań i słów, dokonujących reestymacji parametrów i ostatecznie przeprowadzających test z nagraniami kontrolnymi. Dla danego zbioru uczącego (zestawu nagrań w określonym katalogu) przeprowadzane są w sposób automatyczny następujące operacje:

- utworzenie listy nagrań wzorcowych oraz ich transkrypcji fonetycznej TP1 na podstawie transkrypcji ortograficznej (przygotowanej wcześniej) i słownika transkrypcji fonetycznych,
- utworzenie listy słów występujących w nagraniach wzorcowych i wypowiedziach, które mają być rozpoznawane,
- utworzenie listy modeli fonemów niezbędnych do budowy sieci na podstawie danych wzorcowych, sprawdzenie pokrycia zbioru fonemów w wypowiedziach rozpoznawanych,
- opcjonalnie: wyznaczenie grafu opisującego gramatykę rozpoznawanych wypowiedzi na podstawie transkrypcji ortograficznej przykładowych wypowiedzi lub wygenerowanie grafu pełnego,
- wyznaczenie cech sygnału dla nagrań wzorcowych,
- utworzenie sieci o węzłach opisujących fonemy zgodnie z transkrypcją fonetyczną, inicjalizacja parametrów emisji dla węzłów tej sieci uśrednionymi parametrami (średnie i wariancje) wyznaczonymi dla wszystkich danych wzorcowych,
- reestymacja parametrów sieci,
- utworzenie transkrypcji nagrań wzorcowych oraz słów rozpoznawanych z uwzględnieniem trifonów zamiast pojedynczych fonemów (monofonów), powielenie modeli fonemów do odpowiednich trifonów (rozbudowa sieci), reestymacja parametrów rozbudowanej sieci,
- wyznaczenie cech sygnału dla nagrań rozpoznawanych,
- realizacja algorytmu Viterbiego dla nagrań rozpoznawanych,

- ocena poprawności rozpoznawania przez porównanie z poprawnymi transkrypcjami nagrań rozpoznawanych.

W zbiorze nagrań rozpoznawanych występowały również słowa, które nie znajdowały się w zbiorze nagrań wzorcowych. Sekwencje stanów dla takich słów są tworzone z trifonów o parametrach wyznaczanych na podstawie innych nagrań.

Tabela 4

Wyniki rozpoznawania dla słów spoza zbioru danych wzorcowych i z tego zbioru

Zestaw uczący	Zestaw testowy	Model gramatyki	Poprawność
4724 zdania bez cyfr (41 mówców)	67 ciągów 3-cyfr.	ciągi cyfr	45,1 %
		3-el. ciągi cyfr	54,3 %
450 cyfr (45 mówców) i 114 zdań (1 mówca)	j.w.	ciągi cyfr	50,3 %
		3-el. ciągi cyfr	50,9 %
j.w. i dodatkowo ciągi cyfr 3-el.	j.w.	ciągi cyfr	69,3 %
		3-el. ciągi cyfr	76,2 %

Tabela 4 przedstawia wyniki rozpoznawania dla zbioru uczącego pochodzącego z wybranych nagrań bazy Corpora i zbioru testowego nagranego przez autora. Zbiorem testowym były nagrania składające się z ciągów trzech cyfr. Jako cechy sygnału mowy zastosowano 12 współczynników mel-cepstrum oraz 12 współczynników regresji dla tych współczynników (łącznie 24 cechy). Poprawność rozpoznawania była wyznaczana jako $(N-I-S)/N$, gdzie N jest długością rozpoznawanej sekwencji, I liczbą wstawień, S liczbą zastąpień przy porównywaniu dwóch sekwencji słów metodą przekształcania sekwencji rozpoznanej we wzorcową w najmniejszej liczbie kroków. Uwzględniono dwa modele gramatyki języka: ciągi cyfr o dowolnej długości oraz trzyelementowe ciągi cyfr. Uzyskano względnie niską poprawność (54,3%) rozpoznawania przy zastosowaniu jako wzorcowych nagrań nie zawierających słów rozpoznawanych. W tym przypadku modele artykulacji słów (cyfr) zostały utworzone na podstawie fragmentów modeli innych słów, bez dostępu do fonemów we właściwych kontekstach. Sytuacja ta obrazuje zachowanie systemu przy rozbudowie słownika o słowa nie występujące w bazie nagrań. Podobne wyniki osiągnięto przy zastosowaniu jako nagrań wzorcowych jedynie izolowanych cyfr i rozpoznawaniu ciągów cyfr. Również w tym przypadku system nie dysponował modelami fonemów we wszystkich wymaganych kontekstach. Dołączenie do bazy nagrań wzorcowych ciągów cyfr poprawiło skuteczność rozpoznawania do ok. 75%.

Drugi eksperyment polegał na rozpoznawaniu cyfr przy zastosowaniu bazy nagrań uczących o zmiennym rozmiarze. Nagraniami wzorcowymi były nagrania cyfr (po 10 dla jednego mówcy), kontrolnymi również nagrania cyfr, ale pochodzące od 3 innych mówców.

Tabela 5

Zależność poprawności rozpoznawania od rozmiarów zbioru nagrań wzorcowych

Mówcy w zb. ucz.	1	2	3	4	5	6	7	8	9	10
Poprawność (%)	42,9	58,3	81,8	84,0	88,5	89,3	93,1	90,0	90,0	90,0
Mówcy w zb. ucz.	11	12	13	14	15	16	17	18	19	42
Poprawność (%)	90,0	90,0	93,1	90,0	90,0	90,0	93,1	90,0	89,7	89,6

Tabela 5 przedstawia wyniki rozpoznawania dla tego eksperymentu. Zbiór nagrań wzorcowych był sukcesywnie zwiększany o nagrania kolejnych mówców. Nie był stosowany model gramatyki języka. Poprawność rozpoznawania wyznaczana była jako stosunek poprawnych do wszystkich rozpoznań. Poprawność ta zwiększa się wraz z liczbą mówców, których nagrania uwzględnia zbiór danych wzorcowych, ale tylko do pewnej granicy. Już przy 6 mówcach poprawność osiąga ok. 90% i nie zmienia się w istotny sposób przy dołączaniu kolejnych mówców, do 42. Przyczyny stosunkowo niskiej poprawności to mała liczba nagrań pochodzących od poszczególnych mówców oraz model gramatyki nie uwzględniającej różnic w prawdopodobieństwach poszczególnych sekwencji słów. Późniejsze eksperymenty wykazały, że prawidłowo zbudowany opis gramatyki znacząco zwiększa poprawność rozpoznawania, do 98% w przypadku wypowiedzi z zadanego zakresu tematycznego. Jednak w przypadku dowolnych ciągów cyfr nie jest możliwe założenie różnic w prawdopodobieństwach przypisanych krawędziom grafu opisującego gramatykę.

8. Podsumowanie

Przedstawiono tu zastosowanie biblioteki HTK do rozpoznawania wypowiedzi w języku polskim. Opisano wszystkie elementy niezbędne do utworzenia sieci probabilistycznej: postać modeli poszczególnych fonemów, zestaw fonemów przeznaczony do transkrypcji fonetycznej w języku polskim, słownik transkrypcji, graf opisujący gramatykę języka. Opisano technikę automatycznego budowania sieci dla zadanego zbioru danych wzorcowych (uczących). Zestaw skryptów pozwalających na budowę takich sieci można pobrać ze strony WWW autora. Przedstawiono wyniki eksperymentów nad rozpoznawaniem mowy z zastosowaniem bazy danych Corpora. Baza ta zostanie wykorzystana do dalszych prób, m.in. z różnymi wektorami cech.

LITERATURA

1. Rabiner L., Juang B. H.: Fundamentals of Speech Recognition, Prentice Hall, 1993.

2. Fourcin A., Gibbon D.: Spoken language assessment in the European context. Literary & Linguistic Computing, vol.9, no.1/1994, London 1994.
3. Wells J. C.: Computer-coding the IPA: a proposed extension of SAMPA, Department of Phonetics and Linguistics, University College London, 1995.
4. Nagórko A.: Zarys gramatyki polskiej, PWN, Warszawa 1996.
5. Young S., Kershaw D., Odell J., Valtchev V., Woodland P.: The HTK Book for HTK Version 3.0, Microcoft Corporation, 2000.
6. SAMPA computer readable phonetic alphabet, strona http://www.phon.ucl.ac.uk/home/sampa/home.htm, Londyn 2000.

Recenzent: Dr hab. Zygmunt Vetulani, Prof. UAM

Wpłynęło do Redakcji 8 lutego 2002 r.

Abstract

This paper presents an approach to building a speech recognizer for the Polish language, using the HTK toolkit. This toolkit has been released for scientific research a year ago. All elements necessary for a Polish recognizer have been described: the set of Polish phonemes, dictionary with transcriptions, grammatical model of the language. A method of automatic building of recognizers for different sets of training data has been described, with scripts available on the author's web page. A Polish database Corpora, containing recordings of Polish native speakers has been described, as well as some experiments with recognizers based on sets of these recordings. The dependence of the accuracy of recognition on the number of different speakers in these experiments has been described.

Adres

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

Badania finansowane z funduszu *BW-444/RAu2/2001*