STUDIA INFORMATICA 2013
Volume 34 Number 2B (113)

Mariusz MĄSIOR, Magdalena IGRAS, Mariusz ZIÓŁKO, Stanisław KACPRZAK
AGH Akademia Górniczo-Hutnicza w Krakowie, Katedra Elektroniki

BAZA DANYCH NAGRAŃ MOWY DLA ANALIZY PORÓWNAWCZEJ RÓŻNOJĘZYCZNYCH FONEMÓW

Streszczenie. Artykuł prezentuje system gromadzenia, archiwizacji i akustycznej analizy wielojęzycznych próbek mowy. Głównym celem badań jest analiza porównawcza fonemów dla kilkuset języków i stworzenie drzewa genealogicznego języków świata. Opisana została implementacja systemu, jako bazy danych z portalem internetowym. Przedstawiono informacje dotyczące zawartości i formy bazy, perspektyw rozwoju i zastosowań w lingwistyce komputerowej.

Słowa kluczowe: różnojęzyczne nagrania mowy, analiza mowy

DATABASE OF SPEECH RECORDINGS FOR COMPARATIVE ANALYSIS OF MULTI-LANGUAGE PHONEMES

Summary. The paper presents a system of collecting and analyzing multi-language speech samples for research on characteristics of phonemes in several hundred world languages. We describe the implementation: database and webpage. The content and form of the database and applications for development of the new methods of speech analysis are presented.

Keywords: multi-language speech recordings, speech analysis

1. Wprowadzenie

Zespół Przetwarzania Sygnałów Katedry Elektroniki Akademii Górniczo-Hutniczej w Krakowie zajmuje się zagadnieniami teoretycznymi i zastosowaniem teorii falek w rozpoznawaniu mowy, kompresji sygnałów i transmultipleksacji.

Jednym z zagadnień bieżących prac jest zastosowanie metod opartych na dekompozycji falkowej do analizy fonemów różnojęzycznych, w celu stworzenia drzewa genealogicznego współczesnych języków świata. Projekt został zainspirowany hipotezą Atkinsona [1], sugerującą, że ekspansja języków nastąpiła z Afryki. Zakładamy, że modele używane do budowania

drzewa ewolucji człowieka na podstawie genów, mogą zostać zastosowane do tego samego celu na podstawie różnorodności (w tym zarówno ilości, jak i jakości) fonemów występujących w ludzkiej mowie. Znane w genetyce ewolucyjnej zjawisko efektu założyciela może zachodzić również dla języków i ich werbalnej formy, sprawiając, że wraz z oddalaniem się od punktu wyjściowego, różnorodność fonemów maleje.

Języki ewoluują pośród używających ich populacji, zmieniając liczbę i rodzaj fonemów. Nasze badania pozwolą uzyskać wiedzę o tym, jakie fonemy są używane w poszczególnych częściach świata, jakie są ich cechy charakterystyczne oraz stopień wzajemnego podobieństwa.

Podstawą i determinantą realizacji opisanych badań jest zgromadzenie odpowiedniego materiału badawczego w postaci zbioru nagrań o możliwie najlepszej, a zarazem jednolitej jakości, obejmującego dużą liczbę języków w ich oryginalnym brzmieniu, reprezentowanych przez odpowiednio dużą liczbę mówców. Wymagania te stanowią o trudności wyzwania badawczego.

2. Przegląd baz nagrań różnojęzycznej mowy

Obecnie istnieje kilka baz nagrań mowy wielojęzycznej. Projekty i badania naukowe, w ramach których zebrano te bazy, służyły różnym celom. Część z tych baz może zostać wykorzystana w opisywanych badaniach. Poniżej znajduje się krótki przegląd największych z nich.

UCLA Phonetics Lab Archive

Jedna z większych baz danych nagrań różnojęzycznych na potrzeby badań fonetycznych została zgromadzona pod przewodnictwem Petera Ladefoged, przez Uniwersytet Kalifornijski (UCLA). Usystematyzowany wg trzyliterowych kodów standardu ISO 639-3 (ISO 2007), zbiór nagrań jest dostępny w Internecie [2], dla niektórych wraz z transkrypcjami na licencji Creative Commons.

Projekt Endangered Languages

Projekt Zagrożone Języki został utworzony wraz z Towarzystwem Różnorodności Językowej (ang. *Alliance for Linguistic Diversity*) przez Institute for Language Information and Technology (ILIT) w Eastern Michigan University, przy wsparciu Google 3. Celem projektu jest zgromadzenie dokumentacji zanikających języków całego świata w formie dokumentów tekstowych, nagrań, filmów i informacji o gramatyce i fonetyce języków. Narzędzia dostępne na stronie internetowej projektu umożliwiają udział społeczności i organizacji we współtworzeniu bazy wiedzy o językach świata. Obecnie baza obejmuje 3 176 języków.

Projekt Global Recordings Network

Projekt GRN (ang. *Global Recordings Network*) ma swoje korzenie w 1939 r. i od tego czasu zgromadził nagrania 6 000 języków [4]. Ma on podłoże ewangelizacyjne – jego misją jest udostępnianie materiałów audiowizualnych (modlitwy, opowieści, gospel) wszystkim

społecznościom, również tym, dla których niszowy język stanowił barierę w poznawaniu religii chrześcijańskiej.

3. Konstrukcja bazy danych próbek mowy

Powyżej opisane bazy danych często mają znaczne zasoby nagrań mowy, jednak większość z nich jest nieprzydatna dla automatycznej analizy. Zawierają one dużo wtrąceń w innych językach, szumy, zakłócenia i muzykę. Wiele z nich jest bardzo złej jakości. Z tego powodu należy zgromadzić nagrania od nowa, bezpośrednio od mówców i ewentualnie zaadaptować części istniejących baz po odpowiedniej i dokładnej weryfikacji.

3.1. Opis bazy danych

Jednym z największych kompendiów wiedzy dotyczących języków świata jest Ethnologue [5]. Tworzona baza nagrań mowy w dużym stopniu wzoruje się na tych zasobach. Zapożyczona została część struktury relacyjnej bazy danych opisujących języki, ich nazwy i lokalizacje. Dane zawarte w tej części bazy są publicznie dostępne, a także opisane międzynarodowym standardem ISO 639.

W bazie danych opisanych jest 7 333 różnych języków, jeśli wyeliminujemy języki wymarłe i sztuczne, wciąż pozostaje niemal 6 500 języków, które mogą zostać poddane analizie. Liczba ta w znacznym stopniu wpływa na metody gromadzenia danych oraz możliwości ich kontroli pod względem poprawności i przydatności w badaniach.

Na rys. 1 znajduje się podstawowa struktura bazy danych, wykorzystana do gromadzenia nagrań mowy (tabel oraz relacji między nimi). Sama struktura bazy danych w części dotyczącej języków, ich alternatywnych nazw i krajów, w których są używane, jest rozwinięciem bazy danych kodów języków ISO 639-3 [5].

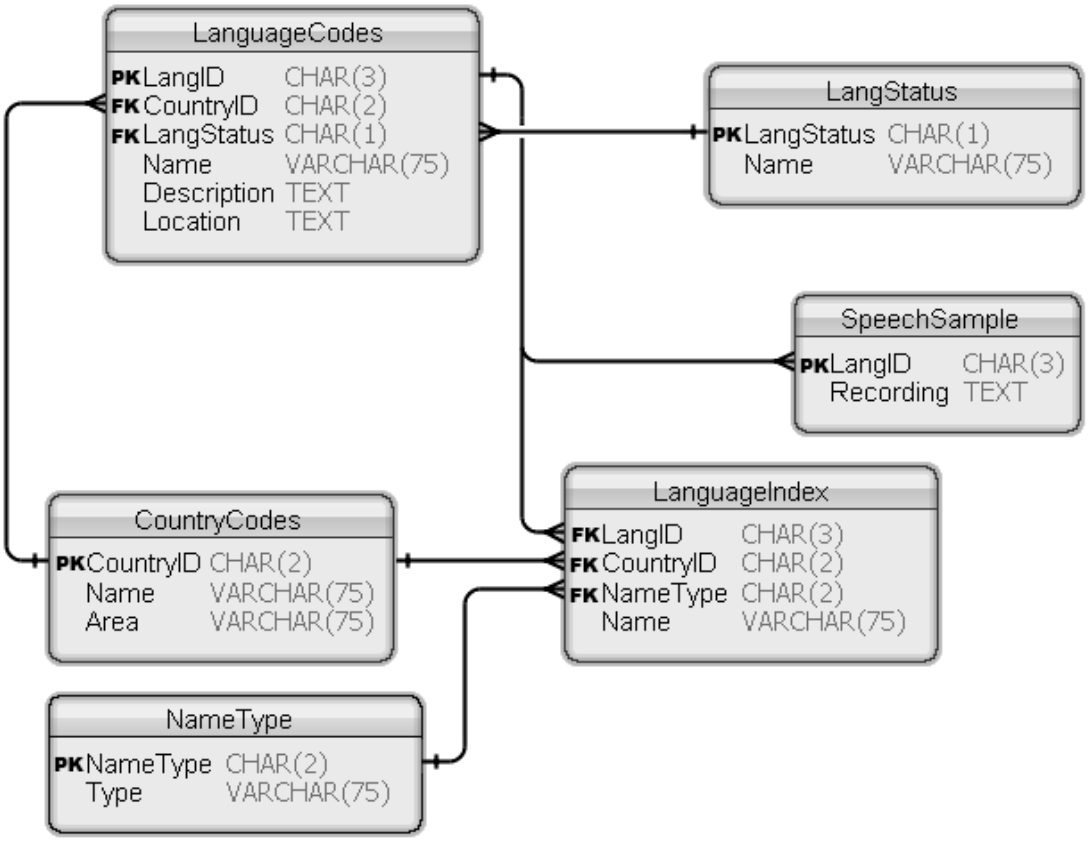
Zgodnie z normą ISO 639, wszystkim językom (zgromadzonym w tabeli LanguageCodes) jest przydzielony jeden z poniższych statusów (LangStatus):

- Living – język używany,
- Nearly extinct – język prawie wymarły,
- Extinct – język wymarły,
- Second language only – język nieużywany jako podstawowy.

Pewnym problemem, który musi być wzięty pod uwagę przy kolekcjonowaniu nagrań mowy, jest fakt, że funkcjonuje ponad 55 000 nazw języków. Dla niektórych z nich istnieje nawet kilkadziesiąt równoprawnych określeń. Baza danych musi zawierać wszystkie te na-

zwy, aby wykluczyć możliwe niespójności (tabela LanguageIndex). Określa się następujące typy nazw języków (NameType):

- Language – podstawowa nazwa języka,
- Language Alternate – nazwa alternatywna,
- Dialect – podstawowa nazwa dialektu,
- Dialect Alternate – alternatywna nazwa dialektu,
- Language Pejorative, Dialect Pejorative – określenia pejoratywne na język lub dialekt.



Rys. 1. Struktura bazy danych wielojęzycznych próbek mowy
Fig. 1. Structure of the multi-language speech samples database

Baza jest wciąż rozwijana i rozbudowywana o kolejne atrybuty języków (dokładne lokalizacje geograficzne obszarów stosowania, informacje lingwistyczne) oraz nagrań (charakterystyka mówcy: wiek, płeć, pochodzenie i dane z analizy akustycznej).

3.2. Portal internetowy bazy nagrań

Zgodnie z założeniami, badania mają się skupić na porównywaniu dużej liczby nagrań mowy. Ze względu na ogromną liczbę języków mówionych, które należałoby poddać analizie, nie da się całości bazy zebrać bezpośrednio od osób władających tymi językami. Dlatego

głównym narzędziem do budowy bazy nagrań będzie strona internetowa, funkcjonująca wg paradygmatu Web 2.0 6.

Portal dostępny pod adresem http://speechsamples.agh.edu.pl, umożliwi wygodne zbieranie i automatyzację zarządzania bazą danych. W przyszłości posłuży także jako podstawa do implementacji całości funkcjonalności akustycznej analizy mowy i prezentacji wyników badań.

Serwis został wykonany na podstawie systemu CMS TYPO3, który jest jednym z najbardziej rozwiniętych i zaawansowanych systemów zarządzania treścią. Do silnika TYPO3 zostały dopisane komponenty, które umożliwią nie tylko dodawanie nagrań na stronie, ale także ułatwią zarządzanie całością zgromadzonych danych.

Dzięki stronie internetowej możliwe jest gromadzenie i dodawanie nie tylko próbek mowy, ale również informacji o językach. Dane te, początkowo zbierane i uzupełniane przez twórców portalu na bazie dostępnej literatury, będą docelowo uzupełniane także przez chętnych do współpracy z całego świata.

Ilość danych niezbędnych do zebrania bardzo utrudnia, nawet dużemu zespołowi, ich gromadzenie i analizę, dlatego niezbędne jest zbudowanie społeczności internetowej, która zechce wziąć udział w gromadzeniu danych. Należy zachęcić przedstawicieli wielu narodów do nagrań próbek mowy i zamieszczenia ich na portalu. Z tego powodu serwis i jego działanie zostaną zintegrowane z portalami społecznościowymi (Facebook, Google+, ResearchGate, Twitter itp.).

4. Automatyczna weryfikacja zbieranych danych

Zebrane nagrania muszą podlegać weryfikacji, czyli ocenie użyteczności nagrania w dalszych analizach akustycznych. Na podstawie tej oceny będzie musiała zostać podjęta decyzja o ewentualnym odrzuceniu nagrania, jeśli jego parametry nie rokują poprawnych wyników dalszego przetwarzania.

Nagrania powinny być jak najlepszej jakości. Przy ocenie jakości bierzemy pod uwagę aspekty techniczne cyfrowego sygnału (częstotliwość próbkowania, rozdzielczość bitową, rodzaj lub brak kompresji) oraz jego zawartość (np. zakłócenia, muzyka w tle). Ze względu na charakterystykę sygnału mowy minimalna częstotliwość próbkowania wymagana do jego reprezentacji to 8 000 Hz. W praktyce zalecane jest stosowanie częstotliwości próbkowania o wartości 16 000 Hz. Taka wartość pozwala na redukcję błędu rozpoznawania nawet o 20% w stosunku analizy sygnału z próbkowaniem 8 000 Hz [7].

Nagrania nie mogą być kompresowane. Użycie metod kompresji stratnej powodowałoby degradację sygnału w istotnych dla analizy pasmach częstotliwościowych, natomiast metody kompresji bezstratnej powodowałyby duży nakład obliczeń związanych z dekompresją.

Przy automatycznym gromadzeniu i adaptacji istniejących baz nagrań istotna jest możliwość oceny, czy sygnał jest sygnałem mowy, muzyką lub połączeniem mowy i muzyki. Takie nagrania nie tylko uniemożliwiłby poprawną analizę, ale mogłyby w znacznym stopniu utrudnić właściwą segmentację i identyfikację fonemów dla danego języka.

Sygnał powinien charakteryzować się jak największym stosunkiem sygnału do szumu (SNR), który ma ogromny wpływ na wyniki rozpoznawania i poprawnej klasyfikacji fonemów. Jak pokazano w literaturze [8], skuteczność rozpoznawania głosek przy SNR rzędu −10 dB spada z 80% aż do 28%.

Ważnym aspektem jest też, aby sygnał zawierał jak najmniej fragmentów ciszy. Usunięcie takich, zbędnych fragmentów pozwoli na zmniejszenie zapotrzebowania pamięciowego bazy oraz wpłynie na szybkość analizy sygnału.

Ze względu na ilość zbieranych danych oczywisty jest fakt, że weryfikacja nagrań nie może być dokonywana ręcznie, a konieczna jest budowa systemu automatycznej kontroli. O ile sprawdzenie częstotliwości próbkowania czy rozdzielczości bitowej jest trywialne, pewnych trudności dostarcza ocena, czy sygnał nie był kompresowany, oraz klasyfikacja jego zawartości (czy nagranie jest mową).

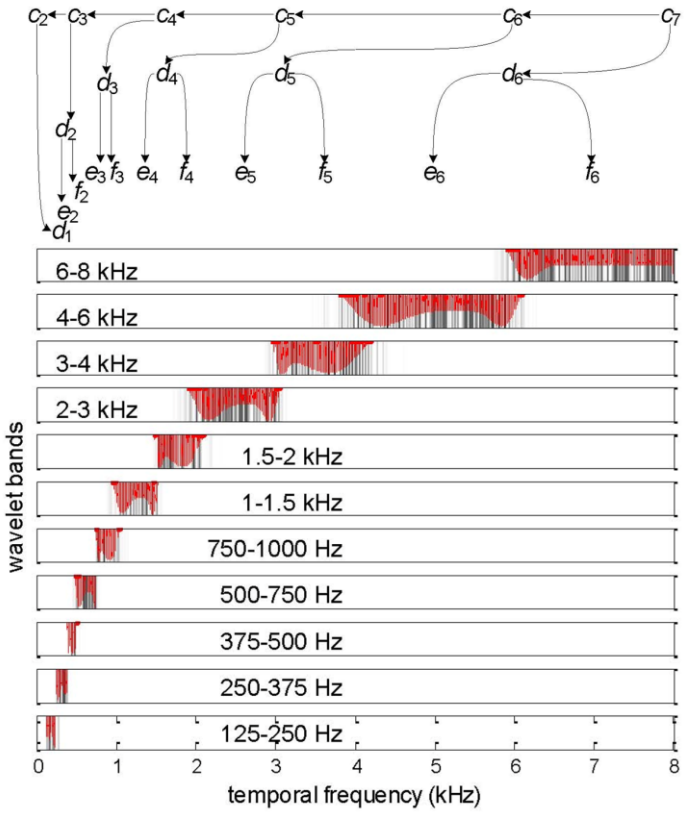
Do oceny, czy sygnał był kompresowany z użyciem metod stratnej kompresji, możemy badać jego energię w poszczególnych pasmach częstotliwościowych oraz cechy charakterystyki spektralnej. Do klasyfikacji sygnału na mowę, muzykę, ciszę czy szum możemy skorzystać z istniejących już algorytmów, które charakteryzują się dużą, ponad 90% skutecznością [9, 10]. Algorytmy te działają na bazie zestawu cech sygnału akustycznego, takich jak: liczba przejść przez zero, energia modulacji 4 Hz czy „spectral flux”. Cechy te wyliczone dla krótkich (16 – 32 ms) ramek sygnału i uśrednione dla długich (1 – 2 s) okien czasowych pozwalają na odrzucenie fragmentów sygnału niebędących mową.

5. Opis planowanych prac badawczych

Zgromadzone nagrania najpierw zostaną poddane segmentacji przy użyciu metod czasowo-częstotliwościowych, z powodzeniem stosowanych dotychczas w systemie automatycznego rozpoznawania mowy AGH [11]. Metoda opiera się na dyskretnej transformacji falkowej (DWT) w skali melowej (rys. 2), dobrze oddającej charakterystykę sygnału mowy i jego percepcji. W jej wyniku otrzymuje się wartości energii w poszczególnych pasmach częstotliwości. Dane te wykorzystuje algorytm segmentacji, realizujący podział sygnału mowy na pojedyncze fonemy.

Otrzymane w wyniku segmentacji fonemy zostaną poddane parametryzacji działającej na podstawie transformację falkowej Fouriera (WFT). Uzyskane wektory parametrów, odpowia-

dające pojedynczym fonemom, zostaną poddane algorytmom klasteryzacji z wykorzystaniem miary odległości Itakura-Saito [12].



Rys. 2. Drzewo dekompozycji sygnału oraz pasma częstotliwości otrzymane w wyniku dyskretnej transformacji falkowej o skali perceptualnej [11]

Fig. 2. Decomposition tree and frequency subbands obtained from discrete wavelet transform with perceptual scale [11]

Opracowanie wyników klasteryzacji pozwoli na zgromadzenie wiedzy na temat rodzaju i cech charakterystycznych fonemów w poszczególnych językach. Kolejnym krokiem będzie zastosowanie metod używanych w genetyce ewolucyjnej, do stworzenia taksonomii języków współczesnych.

Praca naukowa finansowana ze środków Narodowego Centrum Nauki jako projekt badawczy NCN 2011/03/B/ST7/00442.

BIBLIOGRAFIA

1. Atkinson Q. D.: Phonetic Diversity Supports a Serial Founder Effect Model of Language Expansion from Africa. Science, 32, 2011, s. 346÷348.
2. Strona internetowa bazy nagrań UCLA: http://archive.phonetics.ucla.edu/.

3. Strona internetowa projektu Endangered Languages: http://www.endangeredlanguages.com/.
4. Strona internetowa projektu GRN: http://globalrecordings.net/.
5. Lewis M. P.: Ethnologue. Languages of the World. 16th Edition, SIL International, Dallas, Texas 2009, online version: http://www.ethnologue.com/.
6. O'Reilly T.: What Is Web 2.0. Design Patterns and Business Models for the Next Generation of Software. O'Reilly Media, 2009.
7. Huang X., Acero A., Hon H.: Spoken Language Processing: A Guide to Theory, Algorithm, and System Development. Prentice Hall PTR, Upper Saddle River, NJ, USA 2001.
8. Kollmeier B., Brand T., Meyer B.: Springer Handbook of Speech Processing: Perception of Speech and Sound. Springer, Berlin-Heidelberg2008.
9. Saunders J.: Real time discrimination of broadcast speech/music. Proc. 1996 ICASSP, 1996, s. 993÷996.
10. Scheirer E., Slaney M.: Construction and evaluation of a robust multifeature speech/music discriminator. Proc. IEEE Int. Conf. on Acoustic, Speech and Signal Processing, ICASSP, 1997, s. 1331÷1334.
11. Ziółko M., Gałka J., Ziołko B., Jadczyk T., Skurzok D., Mąsior M.: Automatic speech recognition system dedicated for Polish. Proceedings of Interspeech, Show and tell session, Florence 2011.
12. Iser B., Minker W., Schmidt G.: Bandwidth Extension of Speech Signals. Volume 13 of Lecture Notes in Electrical Engineering, Springer, 2008.

Wpłynęło do Redakcji 15 stycznia 2013 r.

Abstract

The paper presents research inspired by the Atkinson hypothesis that the modern languages expanded from Africa 1. The research aims to create a natural languages taxonomy based on the multi-linguistic phoneme analysis and methods used in evolutionary genetics. It will be enabled by obtaining the knowledge which phonemes are used in the different parts of the world, which are their characteristic features and by comparison of their similarities.

The task requires a large amount of multi-language speech recordings to be collected. It is impossible to gather it directly from the native speakers. Therefore, the construction of recordings database (Fig. 1.) will be based on Internet application. Details of speech samples

database implementation and the website associated with it are described. The website, made in Web 2.0 concept, was developed with TYPO3 CMS, which is one of the most advanced content management systems. The portal, available on URL: http://speechsamples.agh.edu.pl, enables convenient collecting and automatic management of languages and speech recordings database. The website provides gathering and administration functionality for speech samples as well as for all linguistic information concerning different languages. All this data will be uploaded by people all over the world, willing to cooperate with the research performers. In future, the portal will serve as a platform for whole acoustic processing implementation, speech analysis and research results presentation.

The methods of phonemes extraction and analysis are based on author's previous works connected with creating the automatic speech recognition system 11 (Fig. 2.). Paper presents a brief overview of existing multi-language speech databases (2, 3, 4). Moreover, it includes some considerations on speech recording quality issues and possibilities of automatic verification recording features affecting the quality of acoustic analysis. Algorithms for checking noise and interference level, compression presence and speech – non-speech discriminations are proposed.

Adresy

Mariusz MĄSIOR: AGH Akademia Górniczo-Hutnicza, Katedra Elektroniki,
al. Mickiewicza 30, 30-059 Kraków, Polska, masior@agh.edu.pl.
Magdalena IGRAS: AGH Akademia Górniczo-Hutnicza, Katedra Elektroniki,
al. Mickiewicza 30, 30-059 Kraków, Polska, migras@agh.edu.pl.
Mariusz ZIÓŁKO: AGH Akademia Górniczo-Hutnicza, Katedra Elektroniki,
al. Mickiewicza 30, 30-059 Kraków, Polska, ziolko@agh.edu.pl.
Stanisław KACPRZAK: AGH Akademia Górniczo-Hutnicza, Katedra Elektroniki,
al. Mickiewicza 30, 30-059 Kraków, Polska, skacprza@agh.edu.pl.