ROZPOZNAWANIE WIEKU I PŁCI NA PODSTAWIE ANALIZY GŁOSU

AGE AND GENDER RECOGNITION BASED ON ANALYSIS OF VOICE

Justyna Gabryś*, Grzegorz Gil, Piotr Kiszka

Akademia Górniczo-Hutnicza im S. Staszica, Wydział Elektroniki, Automatyki, Informatyki i Inżynierii Biomedycznej, Katedra Automatyki i Inżynierii Biomedycznej, 30-059 Kraków, al. Mickiewicza 30

* e-mail: justyna.gabrys@gmail.com

STRESZCZENIE

Metody automatycznego rozpoznawania wieku i płci pozwalają na rozpoznanie cech osoby mówiącej tylko na podstawie nagrania jej wypowiedzi. Mowa ludzka, poza werbalnym komunikatem, niesie ze sobą informacje dotyczące osoby mówiącej. Nagranie mowy osoby pozwala na wyodrębnienie takich informacji, jak jej płeć, wiek, a także emocje. Zaprezentowano przegląd metod rozpoznawania wieku i płci osób na podstawie ich mowy oraz wykonano implementację i przetestowano połączenie metod wyznaczania parametrów MFCC (współczynniki analizy cepstralnej w skali mel (*Mel-frequency Cepstral Coefficients*) i wysokości tonu głosu f_0 oraz algorytmu SVM (metoda wektorów nośnych - *Support Vector Machines*) do klasyfikacji próbek głosowych. Testy zaimplementowanego rozwiązania pozwalają stwierdzić, że metoda jest skuteczna w większości przypadków testowych.

Słowa kluczowe: automatyczne rozpoznawanie mowy, wiek, płeć, współczynniki MFCC, klasyfikacja mówcy, maszyna wektorów nośnych

ABSTRACT

Methods for automatic recognition of the age and gender characteristics allow the identification of the person only on the basis of recording of this person speech. Human speech, beyond verbal communication, gives an information about the speaking person. Speech recording allows the identification personal characteristics such as gender, age, and the emotions. The paper presents an overview of methods of age and gender recognition of people based on their speech. A combination of methods for determining the parameters MFCC (*Mel-frequency Cepstral Coefficients*) and pitch of voice (f_0) and SVM (*Support Vector Machines*) algorithm for the classification of voice samples is implanted and tested. It was demonstrated that the method is effective in the majority of test cases.

Keywords: automatic speech recognition, age, gender, MFCC coefficients, classification of speaker, support vector machine

1. Wprowadzenie

Celem pracy jest przegląd metod rozpoznawania wieku i płci na podstawie mowy. Dokonano również implementacji jednej z tych metod. Mowa ludzka, poza werbalnym komunikatem, niesie ze sobą informacje dotyczące osoby mówiącej. Nagranie mowy osoby pozwala na wyodrębnienie informacji związanych z płcią, wiekiem czy emocjami. Systemy rozpoznawania wieku i płci użytkownika mają na celu klasyfikować próbki mowy do wcześniej zdefiniowanych klas.

2. Badania literaturowe

W rozpoznawaniu wieku i płci osoby mówiącej ważna jest wiedza czy glos zależy od wieku i płci. Analizuje się takie charakterystyczne cechy akustyczne, jak ton podstawowy (f_0) – częstotliwość, intensywność głosu – głośność, częstotliwości, widma, czas trwania dźwięku i przerw – rytm mowy [1]. Na podstawie tych podstawowych informacji oraz ich wariacji, np. zmiany tonu podstawowego i amplitudy, zostało opracowanych wiele metod automatycznego rozpoznawania wieku i płci.

W literaturze opisano kilka rozwiązań i potencjalnych zastosowań. Sposoby automatycznego rozpoznawania mowy (ang. *automatic speech recognition*, ASR) ze względu na wiek i płeć osoby mówiącej przez telefon, zostały zaprezentowane w pracach [2, 3]. Należą do nich m.in. system PPR (ang. *Parallel Phone Recognizer*), powstały na podstawie systemu automatycznego rozpoznawania języka, system oparty na sieci bayesowskiej i brzmieniowych właściwościach mowy (prozodia), system wykorzystujący analizę predykcji liniowej, mieszany model Gaussa oparty na współczynnikach MFCC (ang. *Mel-frequency Cepstral Coefficients*), rozpoznający wiek i płeć oddzielnie. Ich docelowym zastosowaniem jest identyfikacja problematycznych rozmów, np. zdenerwowanych, niezadowolonych klientów na linii, co pozwala na przełączenie takiej osoby do odpowiedniego pracownika. Adaptacyjne dialogi telefoniczne pozwalają na dopasowanie takich składowych rozmowy jak: stopień automatyzacji, kolejność odtwarzania, rodzaj muzyki odtwarzanej w czasie czekania na konsultanta, przykładowo możliwe byłoby dopasowanie bloku reklamowego w zależności od wieku i płci osoby dzwoniącej. Automatyczne rozpoznawanie stanu emocjonalnego mówcy na podstawie jego mowy ma bardzo szerokie zastosowania [4]. Oszacowanie wieku oraz płci na podstawie nagranego głosu oparte na modelu mieszanym Gaussa (ang. *Gaussian Mixture Models*, GMMs) [5], gdzie proces rozpoznawania wieku i płci składa się z uczenia modelu na zestawie grup osób mówiących i klasyfikacji na podstawie nauczonego modelu. Kolejnym przykładem jest system GMM/SVM-supervector (ang. *Gaussian Mixture Model combined with Support Vector Machine*) rozpoznający wiek i płeć do zastosowań w elastycznym module samochodowym, służącym do komunikacji z samochodem, który bierze pod uwagę preferencje grupy docelowej użytkowników samochodu pod względem ich wieku i płci [6]. Projekt wdrożenia tego systemu został zainspirowany rosnącą potrzebą przenoszenia cyfrowego otoczenia do i z samochodu. Stopień współpracy systemu samochodowego uzależniony jest od potrzeb i oczekiwań indywidualnego użytkownika. Autorzy założyli, że najlepszą metodą określenia, z jakim użytkownikiem samochód ma do czynienia jest nieinwazyjny system rozpoznawania wieku i płci, np. wersja systemu dla osób młodszych, które potrzebują stałego dostępu do usług mobilnych i komunikacji. Natomiast dla osób starszych, słabo widzących, system wspomagający jazdę z wyświetlanymi znakami i ostrzeżeniami dla poprawy bezpieczeństwa.

W ramach praktycznej realizacji do klasyfikacji próbek głosowych, ze względu na popularność rozwiązania, wybrano połączenie metod wyznaczania parametrów MFCC i wysokości tonu głosu f_0 oraz algorytm SVM.

3. Metody realizacji rozwiązania

Rozwiązanie zaimplementowano w środowisku do obliczeń numerycznych Matlab. Do wyznaczenia parametrów MFCC wykorzystano funkcje dostępne w pakiecie rastamat [7]. Klasyfikację danych testowych umożliwiły funkcje implementujące algorytm SVM zawarte w pakiecie Bioinformatics Toolbox. Wbudowana funkcjonalność spełniała podstawowe założenia algorytmu, więc ograniczona

była do klasyfikacji dwuklasowej, co było wystarczające do rozpoznania płci mówcy, jednak nie wystarczało do przyporządkowania mówcy do wielu klas wiekowych. W celu pokonania tego ograniczenia wykorzystano rozwiązanie rozszerzające algorytm do użytku wieloklasowego [8].

3.1 Opis wykorzystanej bazy danych

Do implementacji rozwiązania i jego testowania użyto nagrań z bazy ELSDSR (ang. *English Language Speech Database for Speaker Recognition*) [9]. Większość mówców to pracownicy i studenci wydziału Informatyki i Modelowania Matematycznego na Technicznym Uniwersytecie Duńskim. Językiem bazy jest język angielski, a mówcy pochodzą z Danii (21 osób), Islandii (1 osoba) i Kanady (1 osoba). Baza zawiera nagrania wiadomości głosowych 23 osób w wieku od 24 do 63 lat, w tym 13 mężczyzn i 10 kobiet. Średnia wieku kobiet wyniosła 40,6 lat, a mężczyzn 31,3 lat. Urządzenie nagrywające to MARANTZ PMD670, częstotliwość próbkowania16 kHZ, bit rate: 16, format zapisu *.wav. W bazie znajduje się 161 nagranych wypowiedzi, przeznaczonych do treningu klasyfikatorów mowy i 46 do ich testowania. Każda próbka głosowa w bazie posiada identyfikator składający się z czterech liter, pierwsza oznacza płeć: F – kobieta, M – mężczyzna, pozostałe trzy są unikalne dla danego mówcy.

3.2 Ekstrakcja cech

W opisywanym rozwiązaniu wykorzystano wektor cech złożony z 28 parametrów pochodzących z MFCC oraz wysokości tonu głosu f_0. MFCC są popularnymi cechami wykorzystywanymi w systemach rozpoznawania mowy. Wyliczane są jako spektrum z dyskretnej transformacji cosinusowej sygnału będącego logarytmicznym spektrum częstotliwości wyrażonym w skali melowej. Wektor MFCC złożony został z 13 statycznych parametrów. Wyznaczano je w ramkach o szerokości 32 ms z krokiem wynoszącym 16 ms, następnie – w celu ograniczenia ilości danych – dla kolejnych parametrów obliczono wartość średnią i odchylenie standardowe ze wszystkich ramek. W rezultacie otrzymano wektor 26 cech wynikających z parametrów MFCC opisujących dany sygnał mowy. Wysokość tonu głosu wyznaczona została metodą wykorzystującą stosunek subharmonicznych do harmonicznych (ang. *Subharmonic-to-Harmonic Ratio*, SHR) [10]. Algorytm SHR opiera się na analizie widma sygnału mowy – wybierając szczyty reprezentujące efekt nakładania się harmonicznych i subharmonicznych próbuje dokonać dekompozycji efektów pochodzących od różnych częstotliwości i oceny, która z nich jest najlepszym kandydatem na częstotliwość podstawową tonu głosu. Wartości f_0 wyznaczono w ramkach o szerokości 40 ms z krokiem 10 ms. Jako zakres poszukiwań częstotliwości tonu głosu przyjęto wartość domyślną od 50 do 550 Hz. Z zebranych wartości obliczono wartość średnią i odchylenie standardowe ze wszystkich ramek. W rezultacie otrzymano dwie dodatkowe cechy opisujące dany sygnał mowy.

3.3 Klasyfikacja danych

W celu klasyfikacji danych wykorzystano maszynę wektorów nośnych (ang. *Support Vector Machine*, SVM). Jest to klasyfikator pozwalającym na podział danych wejściowych na określone w cyklu nauki kategorie. Pozwala na zbudowanie modelu z jak najmniejszej liczby wektorów danych treningowych, tj. wektorów nośnych. Algorytm SVM w najprostszej formie pozwala na znalezienie płaszczyzny rozdzielającej z maksymalnym marginesem punkty należące do dwóch klas. To podejście zostało zastosowane w przypadku wyznaczenia płci osoby mówiącej. Większą liczbę klas wprowadzono w trakcie nauki klasyfikatora, pozwalającej na rozpoznanie wieku mówcy. Ze względu na rozpiętość wieku od 24 do 63 lat zastosowano podział na pięć klas, każda obejmująca dekadę (np. 35–44 lat). Uzyskano w ten sposób 5 klas ponumerowanych od 2 do 6. Każdy z mówców został przyporządkowany do odpowiadającego mu przedziału. Przedziały te mają za zadanie rozmycie klasyfikacji, przy małej ilości danych uczących. Oznacza to, że algorytm ma za zadanie określić czy wprowadzanie zbyt małych przedziałów klasyfikujących spowodowałoby zbyt duże błędy przy rozpoznaniu wieku. Najlepsze wyniki klasyfikacji uzyskano, stosując podczas nauki maszyny wektorów nośnych funkcję kernela w postaci wielomianu 3 rzędu, a także funkcję kwadratową.

Pozostałe testowane opcje, których użycie wpływało na gorszą dokładność klasyfikacji to: funkcja liniowa, gaussowska radialna funkcja bazowa, wielowarstwowy perceptron. Jako metodę wyznaczającą hiperpłaszczyznę klasyfikującą dane użyto metody najmniejszych kwadratów.

Baza podzielona jest na dwa zestawy: 154 próbki uczące i 44 próbki testowe. W obu kategoriach znajdują się próbki dla każdego z nagranych mówców. Klasyfikacje wypowiedzi ze względu na płeć, jak i na wiek przeprowadzono dla dwóch przypadków testowych. W pierwszym wykorzystano wszystkie próbki z zestawu uczącego do nauki maszyny wektorów nośnych i klasyfikację wszystkich próbek z zestawu testowego. Przypadek ten pozwala na sprawdzenie, jak użyty algorytm zakwalifikuje mówców, których wypowiedzi (inne niż podczas testowania) zostały użyte podczas nauki klasyfikatora. W drugim przypadku dla każdego mówcy wykorzystano wszystkie próbki z zestawu uczącego, oprócz próbek danego mówcy, do nauki maszyny wektorów nośnych. Klasyfikacja próbek danego mówcy z zestawu testowego ma za zadanie sprawdzenie, jak użyty algorytm zakwalifikuje osobę, której wypowiedzi nie były użyte podczas nauki klasyfikatora.

4. Wyniki testów zaimplementowanego rozwiązania

Wyniki testów przedstawiono w tabeli 1. Wpływ warunków nagraniowych i sprzętu nagrywającego został pominięty, poprzez testowanie algorytmu na próbkach pochodzących z tej samej bazy.

Tabela 1. Wynik klasyfikacji próbek testowych z bazy ELSDR. Numery grup wiekowych odpowiadają dziesięcioletnim przedziałom. Płeć: M – mężczyzna, K – kobieta. Czcionką pogrubioną oznaczono błędnie wyniki klasyfikacji

Dane			Klasyfikacja płci		Klasyfikacja wieku			
ID próbki w bazie ELSDR	Wiek	Płeć	I	II	I		II	
					Grupa	Błąd	Grupa	Błąd
FAML_Sr3	48	K	K	K	5	-	**3**	-2
FAML_Sr4			K	**M**	5	-	**3**	-2
FDHH_Sr25	28	K	K	K	3	-	**5**	2
FDHH_Sr26			K	K	3	-	**5**	2
FEAB_Sr5	58	K	K	K	6	-	**3**	-3
FEAB_Sr6			K	**M**	6	-	**3**	-3
FHRO_Sr31	26	K	K	K	**5**	2	**5**	2
FHRO_Sr32			K	K	3	-	**5**	2
FJAZ_Sr35	25	K	K	**M**	3	-	3	-
FJAZ_Sr36			K	**M**	3	-	3	-
FMEL_Sr21	38	K	K	**M**	4	-	**3**	-1
FMEL_Sr22			K	K	4	-	**3**	-1
FMEV_Sr10	46	K	K	K	**3**	-2	**3**	-2
FMEV_Sr9			K	K	5	-	**4**	-1
FSLJ_Sr33	24	K	K	K	2	-	**6**	4
FSLJ_Sr34			K	K	2	-	**6**	4
FTEJ_Sr13	50	K	K	**M**	5	-	**3**	-2
FTEJ_Sr14			K	**M**	5	-	**3**	-2
FUAN_Sr39	63	K	K	K	6	-	**2**	-4
FUAN_Sr40			K	K	6	-	**3**	-3
MASM_Sr11	27	M	M	M	3	-	**4**	1
MASM_Sr12			M	M	3	-	3	-

Dane			Klasyfikacja płci		Klasyfikacja wieku			
ID próbki w bazie ELSDR	Wiek	Płeć	I	II	I		II	
					Grupa	Błąd	Grupa	Błąd
MCBR_Sr23	26	M	M	M	3	-	3	-
MCBR_Sr24			M	M	3	-	3	-
MFKC_Sr43	47	M	M	M	5	-	5	-
MFKC_Sr44			M	M	5	-	5	-
MKBP_Sr19	30	M	M	**K**	**6**	3	3	-
MKBP_Sr20			M	**K**	**4**	1	**4**	1
MLKH_Sr37	47	M	M	M	5	-	**3**	-2
MLKH_Sr38			M	**K**	5	-	**3**	-2
MMLP_Sr27	27	M	M	M	**4**	1	**4**	1
MMLP_Sr28			M	M	**4**	1	**5**	2
MMNA_Sr15	26	M	M	M	3	-	3	-
MMNA_Sr16			M	M	3	-	3	-
MNHP_Sr1	28	M	M	M	3	-	3	-
MNHP_Sr2			M	**K**	3	-	**4**	1
MOEW_Sr41	37	M	M	M	4	-	**3**	-1
MOEW_Sr42			M	M	4	-	**3**	-1
MPRA_Sr29	29	M	M	M	3	-	**5**	2
MPRA_Sr30			M	M	3	-	3	-
MREM_Sr7	29	M	M	M	3	-	3	-
MREM_Sr8			M	M	3	-	3	-
MTLS_Sr17	28	M	M	M	3	-	3	-
MTLS_Sr18			M	M	3	-	3	-

W pierwszym przypadku testowym zaimplementowany algorytm wykazał 100% trafność wyników rozpoznawania płci, 86% wieku, przy średnim błędzie 0,14 i odchyleniu standardowym błędów równym 0,67. Oczywistym problemem tego przypadku testowego jest to, że dane uczące i testujące to wypowiedzi tych samych osób. W celu sprawdzenia, jak zaimplementowane rozwiązanie poradzi

sobie z próbkami głosowymi pobranymi od mówców, których dane wokalne nie były podawane na wejście algorytmu uczącego, przeprowadzono klasyfikacje dla drugiego przypadku.

Skuteczność rozpoznawania płci w drugim przypadku testowym, czyli w przypadku mówców, których próbki głosowe nie brały udziału w nauce klasyfikatora wyniosła 75%. Wyniki rozpoznawania wieku 36%. Średni błąd wyniósł 0,18 dekady, a odchylenie standardowe błędów 1,77.

5. Podsumowanie i możliwości rozwoju proponowanej metody

Testy zaimplementowanego rozwiązania pozwalają stwierdzić, że metoda klasyfikacji do grupy wiekowej jest skuteczna w większości przypadków testowych. Najmniej zadowalające wyniki uzyskano dla przyporządkowania wieloklasowego do grupy wiekowej.

Czynniki, które mogły wpłynąć na błędne klasyfikacje podczas testowania to mała ilość mówców w bazie, zbyt mała grupa próbek uczących, rozkład wieku mówców oraz zbyt mała rozpiętość wiekowa mówców. Język angielski, będący językiem wszystkich próbek w bazie, nie jest językiem ojczystym zdecydowanej większości mówców. Może to prawdopodobnie prowadzić do pewnych niejednorodności w próbkach nagranych przez tego samego mówcę.

Podział klasyfikacji wieku na dekady wydaje się wystarczający, a niekiedy wręcz nadmiarowy w aplikacjach użytkowych. Dobranie odpowiednich przedziałów powinno zależeć od przeznaczenia rozwiązania. Minimalizując ilość powstających klas zmniejszyć można prawdopodobieństwo złego przyporządkowania. Przykłady zastosowania algorytmu to m.in. rozpoznawanie cech mówcy w kryminalistyce, wykrywanie wieku i płci w celu dopasowania odpowiednich materiałów reklamowych, w branży motoryzacyjnej do określania domyślnych ustawień pojazdu.

Dalszy rozwój rozwiązania musiałby opierać się na normatywnych bazach dla danego języka, bądź nawet dialektu. W chwili obecnej zaimplementowane rozwiązanie działa w bardzo wąskim zakresie, najlepiej klasyfikując wykształconych Duńczyków w średnim wieku, mówiących po angielsku. Należałoby z grupy możliwych parametrów klasyfikujących wybrać te, które cechują się odpowiednio wysoką różnorodnością. Klasyfikator również wymaga dopracowania i rozbudowania, w celu zmniejszenia liczby błędnych dopasowań. Ilość grup wiekowych, do których klasyfikuje się mówców zależy od konkretnego zastosowania i zwykle nie jest duża.

LITERATURA

[1] S. Schötz: *A perceptual study of speaker age*, Working Papers 49, Department of Linguistics and Phonetics, Lund University, 2001.

[2] F. Metze, J. Ajmera, R. Englert et al: *Comparison of four approaches to age and gender recognition for telephone applications*, Acoustics, Speech and Signal Processing, 2007.

[3] T. Bocklet, A. Maier, J.G. Bauer: *Age and gender recognition for telephone applications based on GMM supervectors and support vector machines*, Acoustics, Speech and Signal Processing, 2008.

[4] T. Seehapoch, S. Wongthanavasu: *Speech emotion recognition using Support Vector Machines*, Knowledge and Smart Technology (KST), 2013.

[5] V. Hubeika: *Estimation of Gender and Age from Recorded Speech*, Proc. ACM Student Research competition 2006.

[6] M. Feld, F. Burkhardt, C. Muller: *Automatic Speaker Age and Gender Recognition in the Car for Tailoring Dialog and Mobile Services*, Proc. Interspeech 2010.

[7] P. Daniel, W. Ellis: *PLP and RASTA (and MFCC, and inversion) in Matlab*, http://www.ee.columbia.edu/~dpwe/resources/matlab/rastamat/, 2005.

[8] A. Mishra: *Multi Class Support Vector Machine - File Exchange – MATLAB Central*, http://www.mathworks.com/matlabcentral/fileexchange/33170-multi-class-support-vector-machine, 2012.

[9] L. Feng: *Speaker Recognition, Informatics and Mathematical Modelling*, Technical University of Denmark, DTU, 2004.

[10] X. Sun: *Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio*, Proc. of ICASSP2002, Orlando, Florida, May 13–17, 2002.

otrzymano / submitted: 11.01.2015
wersja poprawiona / revised version: 26.06.2015
zaakceptowano / accepted: 15.07.2015