Waldemar Maciejko

Zastosowanie automatycznego rozpoznawania mówców w kryminalistyce

Wprowadzenie

Rozpoznawanie przez człowieka znanych mu osób na podstawie ich głosu jest rzeczą naturalną. Powszechność tego zjawiska powoduje, iż człowiek świadomie nigdy nie analizuje cech głosu, które wpłynęły na proces percepcji. Próba zautomatyzowania tej czynności uświadamia nam, jak skomplikowany jest to proces. Współczesne aplikacje automatycznego rozpoznawania mówców są systemami informatycznymi, wykorzystującymi wiedzę z dziedziny elektroakustyki, akustyki słuchu i mowy, statystki oraz rachunku prawdopodobieństwa.

Podział systemów automatycznych

Rozpoznawanie mówcy jest pojęciem szerokim, które obejmuje m.in. identyfikację oraz weryfikację. W procesie identyfikacji tożsamość nie jest wstępnie deklarowana a mówca, którego głos podlega badaniu, może być już uprzednio zarejestrowany w systemie (tzw. identyfikacja w zbiorze zamkniętym) lub jest kimś zupełnie nie znanym dla systemu (identyfikacja w zbiorze otwartym). Zadanie weryfikacji natomiast polega na rozstrzygnięciu, czy badana wypowiedź należy do mówcy o deklarowanej tożsamości.

Systemy automatycznego rozpoznawania mówców można podzielić również na zależne i niezależne od tekstu. Zależność od tekstu oznacza, że w trakcie próby rozpoznania wymaga się, aby osoba rozpoznawana wypowiedziała słowa, które wystąpiły w sekwencji uczącej (wzorcowej). Jeżeli natomiast w wypowiedzi znajdują się dowolne słowa (stawia się jedynie wymagania co do długości wypowiedzi oraz jakości nagrania), to mówimy o systemach niezależnych od tekstu. Rozpoznanie zależne od tekstu wymaga użycia bardzo złożonych obliczeniowo algorytmów, przy czym skuteczność tych dwóch różnych metod jest zbliżona [Reynolds, D. A., 1995]. W niniejszej pracy skupiono się na systemach niezależnych od tekstu, ponieważ są najbardziej rozpowszechnione oraz - zgodnie z danymi publikowanymi przez NIST[1)] – są najbardziej efektywne pod względem osiąganych rezultatów przy minimalnej złożoności obliczeniowej[2)].

Automatyczne rozpoznawanie osób znajduje zastosowanie w systemach chroniących dostęp do zastrzeżonych usług, realizacji operacji finansowych w banku za pośrednictwem telefonu, kontroli dostępu do systemów zarządzania, systemów dowodzenia i chronionych stref w budynkach itp.

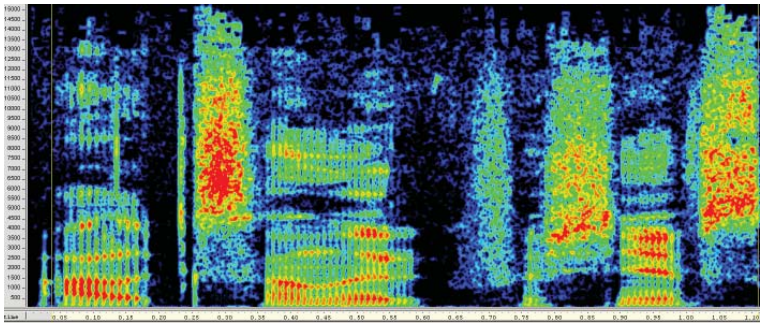
[1)] Narodowy Instytut Standaryzacji i Technologii (ang. *National Institute of Standards and Technology – NIST*) – w ramach NIST prowadzone są coroczne badania międzylaboratoryjne producentów systemów automatycznego rozpoznawania mówców. Celem tych badań jest określenie najlepszej metody identyfikacji poprzez porównanie wyników osiągniętych przez systemy poszczególnych laboratoriów [itl.nist.gov].

[2)] Złożoność obliczeniowa – odpowiada na pytanie; jak czas wykonania algorytmu (przez komputer) będzie rósł wraz ze wzrostem ilości danych wejściowych. Pojęcie to określa szybkość wykonania algorytmu [Szwabiński, J., 2006].

Automatyczne rozpoznawanie mówców zajęło szczególnie ważne miejsce w kryminalistyce i sądownictwie. W przypadku, gdy treści zarejestrowanych wypowiedzi stanowią naruszenie prawa (np. groźby karalne) lub jest mowa o przestępczym działaniu (np. kradzież, zamach terrorystyczny), nagranie takie może stanowić dowód w sprawie, o ile podejrzany o te działania zostanie lub nie zostanie na podstawie głosu zidentyfikowany.

Metody rozpoznawania mówców

Pierwsze próby rozpoznawania głosów w inny sposób, niż za pomocą słuchu, prowadzono w latach sześćdziesiątych. Metody te bazowały na subiektywnym porównaniu obrazów wypowiedzi, tak zwanych spektrogramów (rys.1). Każdy spektrogram zawiera dużą ilość użytecznych informacji w układzie współrzędnych czas – częstotliwość – amplituda [Kersta, L.G., 1962].



Rys. 1. Spektrogram wypowiedzi „Aleksander”

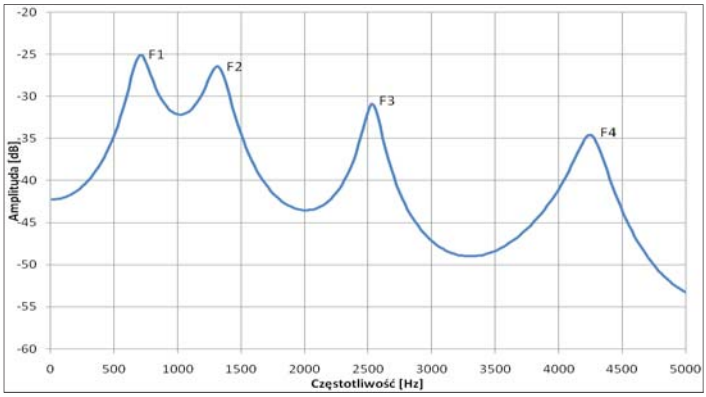
Możliwość zidentyfikowania mówcy jedynie na podstawie spektrogramu szybko jednak podważono. Jednocześnie prowadzono badania nad możliwością zastosowania parametrów takich, jak ton krtaniowy oraz częstotliwości formantowe. Ton krtaniowy jest to częstotliwość drgań wiązadeł głosowych. Natomiast częstotliwości formantowe są to wartości maksymalne w widmie samogłosek, powstałe na skutek formowania dźwięku przez układ artykulacyjny człowieka (rys. 2).

Badaniami podstawowych parametrów w dziedzinie częstotliwości, jakimi są ton krtaniowy oraz częstotliwości formantowe, zajął się między innymi Wiktor Jassem z Polskiej Akademii Nauk. Wyniki opublikował w 1973 r. jednoznacznie stwierdzając, iż częstotliwości formantowe są parametrami, które z bardzo dużym prawdopodobieństwem różnicują mówców [Jassem W., 1973].

Częstotliwości formantowe oraz ton krtaniowy stosowane są szeroko do dnia dzisiejszego. Pojawienie się szybkich komputerów pozwoliło na powszechne wykorzystanie dyskretnych metod analizy sygnału (realizowalnych jedynie przy pomocy maszyn cyfrowych). W roku 1963 amerykańscy naukowcy B.P. Bogert, M.J.R. Healy oraz J.W. Tukey zaproponowali tzw. cepstralną analizę sygnału [Bogert B.P., 1963]. Teoretyczne założenie analizy cepstralnej opiera się na możliwości zamiany operacji mnożenia sygnałów w ich sumowanie i w efekcie rozdzielenie[3)]. Dzięki temu parametry anali-

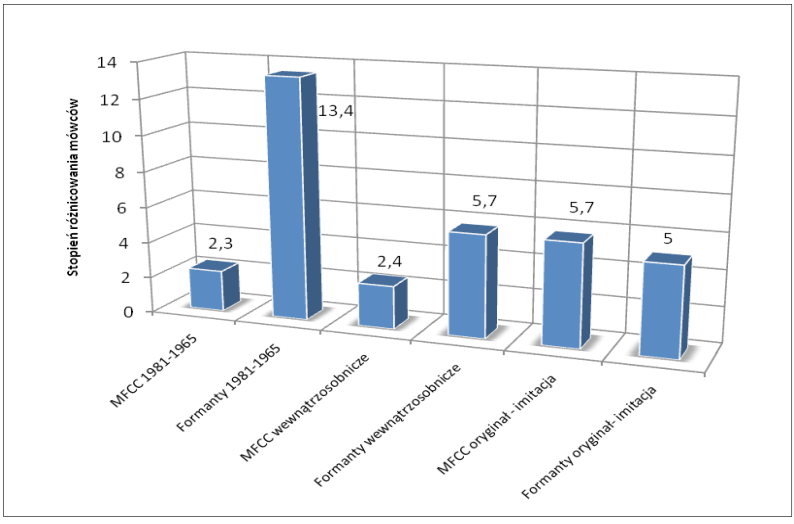
[3)] Chodzi o własność logarytmu: logarytm z iloczynu liczb to suma logarytmów tych liczb.

zy cepstralnej okazały się istotnie odporne na zakłócenia w kanale transmisji. Okazało się również, że są to wartościowe parametry, jeżeli chodzi o rozpoznawanie mówców.



Rys. 2. Widmo głoski /a/ we frazie /aleks/ otrzymane w wyniku zastosowania LPC[4]. Widoczne maksyma to częstotliwości formantowe.

Wykres na rys. 3 przedstawia porównanie możliwości różnicowania osób poprzez parametry cepstralne (MFCC) oraz formanty. Do badań wykorzystano trzy bazy głosów. Pierwsza baza zawierała głosy tych samych mówców zarejestrowane w roku 1965 oraz 1981. Kolejna składała się z głosów tych samych mówców nagrywanych co kilkadziesiąt dni na przestrzeni około 6 miesięcy. Za pomocą tej bazy oceniono średnie wahania wewnątrzosobnicze. Trzecia baza to wypowiedzi mówców oryginalnych oraz ich imitatorów. Porównań dokonano na tych samych wypowiedziach. Jako stopień różnicowania mówców przyjęto odległość Bhattacharyya [Malegaonkar A., 2008].



Rys. 3. Porównanie możliwości różnicowania głosów poprzez częstotliwości formantowe oraz parametry MFCC [Maciejko W., 2005].

[4] LPC – liniowe kodowanie predykcyjne to metoda stosowana m.in. do obliczania widma sygnału (praktycznych zastosowań tej metody jest wiele, np. kompresja dźwięku, kodowanie głosu do celów transmisji GSM itp.)

Pierwsza baza posłużyła do oceny, jak wielkim zmianom ulegną parametry na długiej przestrzeni czasowej (16 lat). Ze względu na to, że porównywano wypowiedzi tych samych osób, spodziewać się należy, iż zaobserwowane różnice pomiędzy głosami będą niewielkie.

W przypadku parametrów cepstralnych wynoszą one średnio 2,3[5], tymczasem dla formantów aż 13,4. Odnosząc to do obserwowanych średnich zmian wewnątrzosobniczych, opierając się na formantach, stwierdzonoby, iż analizie poddano głosy różnych mówców. Zbadano również, na ile wyćwiczony imitator jest w stanie naśladować widmowe cechy głosu. Również w tym przypadku parametry cepstralne okazały się skuteczniejsze.

Dzięki temu, że obliczenia MFCC dokonuje się z całych wypowiedzi (formanty realizowane są w obrębie głosek dźwięcznych), dużo łatwiejsze stało się zautomatyzowanie operacji obliczenia cech osobniczych mówcy. Analiza cepstralna generuje „obraz" mówcy w postaci setek kolumn liczb. Pojawił się zatem kolejny problem: w jaki sposób dokonać porównania dwóch modeli mówców, zakodowanych w ogromnej ilości danych. Tak więc na przestrzeni ostatnich 20 lat rozwijano algorytmy, które na podstawie nagranych głosów nie tylko obliczą cechy osobnicze, ale również będą samodzielnie wnioskować o tożsamości mówców. W te badania zaangażowało się wiele uznanych laboratoriów (takich jak MIT Lincoln Laboratory, Bell Telephone Laboratories). Przetestowano wiele metod modelowania oraz wnioskowania, takich jak sieci neuronowe, niejawne modele Markowa, metody nieparametryczne (kwantyzacja wektorowa, najbliższy sąsiad, najbliższa średnia). Dziś, ze względu na bardzo dużą efektywność, w centrum uwagi badaczy znalazły się tzw. metody modelowania parametrycznego, w których wynikiem rozpoznania jest prawdopodobieństwo, że dana wypowiedź została wyartykułowana przez określoną osobę. Podstawą tych algorytmów jest podejście bayesowskie. Można powiedzieć, że dobrze znany wzór Bayesa stał się podstawą do rozwoju teorii i algorytmów różnych form wnioskowania probabilistycznego [Cichosz P. (2000)].

Jedną z najchętniej stosownych metod modelowania parametrycznego mówców jest tzw. kombinacja modeli normalnych (ang. *gaussian mixture models* - GMM). Systemy rozpoznawania tego typu od kilku lat zapewniają uzyskanie najlepszych wyników spośród wszystkich metod rozpoznawania [Reynolds, 1995][6].

Kolejnym elementem systemu automatycznego podwyższającym skuteczność, niezbędnym w praktyce kryminalistycznej, jest moduł kompensujący cechy osobnicze oraz normalizujący kanał transmisji. W praktyce kryminalistycznej najczęściej porównuje się głosy nagrywane w różny sposób, często nieznany dla eksperta w trakcie prowadzenia badań. W celu zminimalizowania wpływu właściwości kanału transmisji na cechy osobnicze głosu, stosuje się jednocześnie wiele metod. Jedna z nich to specjalny rodzaj filtracji, która realizowana jest w oparciu o parametry cepstralne. Polega ona na odejmowaniu uśrednionej charakterystyki współczynników cepstrum, których wartości zmieniają się wolniej niż cepstrum sygnału mowy[7]. Inne popularne metody służące do normalizacji to stosowanie parametrów liniowego kodowania perceptualnego [Hermansky, H., 1994] oraz statystyczna kompensacja cech osobniczych. Ostatnia

[5] Odległość Bhattacharyya - jest miarą stosowaną w statystyce do oszacowania różnicy między dwoma rozkładami prawdopodobieństwa. W analizowanym przypadku zastosowano wzór na n – wymiarowy rozkład normalny. Odległość należy interpretować w ten sposób, że: im miara odległości jest większa, tym różnica pomiędzy głosami jest większa.

[6] Celem nie jest szczegółowe opisanie wspominanych metod, przekraczałoby to znacznie ramy niniejszej pracy. Zainteresowanych odsyłam do literatury wymienionej na końcu.

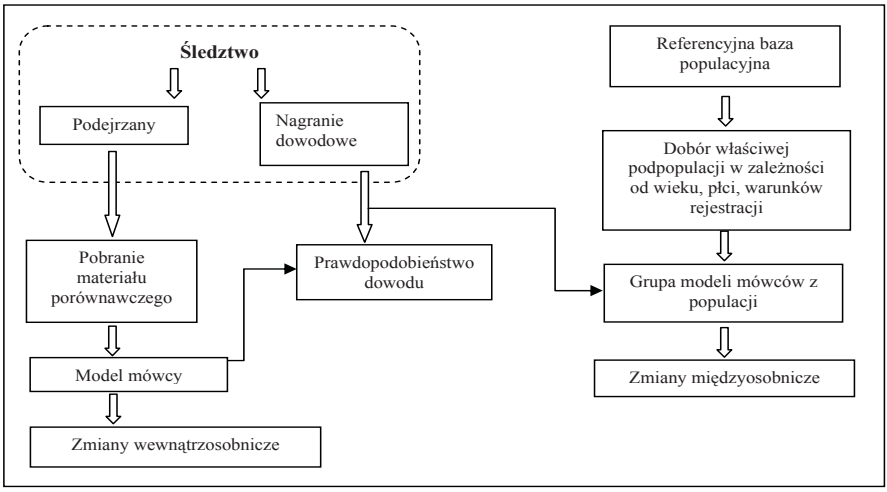
[7] Metoda CMS (ang. *Cepstral Mean Substraction*).

z wymienionych metod polega na empirycznej obserwacji tego, jak zmieniają się parametry osobnicze na grupie mówców referencyjnych (w praktyce około 10 osób) pod wpływem danego rodzaju transmisji, a następnie zmiana cech głosów zniekształconych o zaobserwowane wartości.

Automatyczne rozpoznawanie mówców w kryminalistyce

Metody automatycznego rozpoznawania mówców znalazły zastosowanie przede wszystkim w identyfikacji kryminalistycznej. Dzięki temu, iż badania przy ich pomocy są w pełni obiektywne oraz pozwalają na identyfikację na podstawie krótkich (kilkunastosekundowych) wypowiedzi, uzupełniają powszechnie stosowane metody językowe[8)].

Na rysunku 4 przedstawiono poglądowy schemat automatycznego systemu rozpoznawania mówców. Jednym z elementów tego systemu jest populacyjna baza głosów, która wykorzystywana jest do oceny zmian międzyosobniczych. Taka baza zawiera głosy, które powinny być maksymalnie zbliżone do głosu nagrania dowodowego.



Rys. 4. Podstawowy schemat pracy automatycznego systemu rozpoznawania mówców [Gonzalez - Rodriguez J., 2003].

Na pierwszym etapie procesu porównania tworzone są modele głosów osoby podejrzanej oraz model populacyjny. Na model mówcy (lub model populacyjny), w przypadku wspomnianego wyżej algorytmu GMM, składają się wartości średnie, odchylenie standardowe oraz wagi poszczególnych parametrów. Na drugim etapie obliczane są prawdopodobieństwa wystąpienia cech głosu z nagrania dowodowego w modelu mówcy podejrzanego oraz populacji. Iloraz tych prawdopodobieństw to tzw. iloraz wiarygodności, oznaczany jako LR. Wartość LR mówi, ile razy prawdopodobieństwo tego, że mówca w nagraniu dowodowym i porównawczym to ta sama osoba jest większe od prawdopodobieństwa, że jest to inny mówca.

[8)] Celem analizy językowej jest wyszczególnienie zestawu cech i parametrów indywidualizujących poszczególne osoby.

Zakończenie

W ciągu ostatnich kilku lat można zaobserwować zintensyfikowane działania laboratoriów badawczych na całym świecie zmierzające do udoskonalenia metod identyfikacji osób na podstawie głosu. Przejawia się to m.in. w liczbie publikacji, ciągle rosnącej liczbie laboratoriów biorących udział w testach organizowanych przez NIST oraz, wreszcie, w pojawieniu się na rynku nowego oprogramowania, które wykorzystuje coraz doskonalsze rozwiązania.

Nad wdrożeniem lub doskonaleniem własnych metod automatycznej identyfikacji pracuje również wiele laboratoriów kryminalistycznych. Obserwując postęp w tej dziedzinie, wydaje się, że wkrótce tego typu metody badawcze zajmą ważne miejsce w procesie kryminalistycznej identyfikacji mówców.

Literatura:

1. D.A. Reynolds and R.C. Rose, *Robust text-independent speaker identification using Gaussian mixture speaker models*, IEEE Transactions on Speech and Audio Processing 1995, 3(1):72–83.
2. J. Szwabiński, *Metody numeryczne,* 2006.
3. L.G. Kersta, *Voiceprint Identification,* Nature 1962, vol. 196, pp. 1253–1257.
4. W. Jassem, *Podstawy fonetyki akustycznej,* IPPT PAN 1973.
5. B.P. Bogert, M.J.R. Healy, and J.W. Tukey, *The quefrency alanysis of time series for echoes: Cepstrum, pseudo-autocovariance, cross-cepstrum, and saphe cracking,* w: M. Rosenblatt, Ed., 1963, Time Series Analysis, ch. 15, s. 209–243.
6. W. Maciejko, *Różnice wewnątrzosobnicze i międzyosobnicze w parametrach mówców oryginalnych i ich imitatorów*, Politechnika Wrocławska 2005.
7. A. Malegaonkar, P. Ariyaeeinia, P. Sivakumaran, S. Pillay, *Discrimination effectivness of speech cepstral features,* Biometrics and Identity Management Volume 2008, 5372/2008.
8. H. Hermansky, *RASTA Processing of Speech*, IEEE Trans. on Speech, and Audio Proc, 1994, 2(4):578–589.
9. P. Cichosz, *Systemy uczące się,* WNT 2000
10. http://www.itl.nist.gov/iad/mig//tests/sre/.
11. J. Gonzalez-Rodriguez, J. Fierrez-Aguilar, J. Ortega-Garcia, *Forensic identification reporting using automatic speaker recognition systems,* ICASSP 2003.

ABSTRACT

Automatic Speaker Recognition (ASR) is among most extensively developed biometric techniques. The highly effective recognition methods can be successfully implemented in various fields, such as forensic sciences. This paper describes the fundamentals of automatic speaker recognition, including brief history of speech analysis research, aiming at speaker recognition as well as classification of the ASR systems. Presented are state – of – the – art recognition methods in connection with individual speaker features and with the classification techniques. The principal requirements of forensic speaker recognition were defined in order to characterize the theoretical model of an optimal automatic speaker recognition system.