Tomasz ŚLIWAK-ORLICKI[1], Krzysztof GÓRSKI[2]

Akademia Wojsk Lądowych, Katedra Zarządzania Innowacyjnymi Projektami (1,2)
ORCID: 1. 0009-0002-8974-7380; 2. 0000-0001-8671-380X

doi:10.15199/48.2023.09.44

Rozpoznawanie głosu i identyfikacja mówcy: przegląd wybranych metod rozpoznawania cech biometrycznych mowy

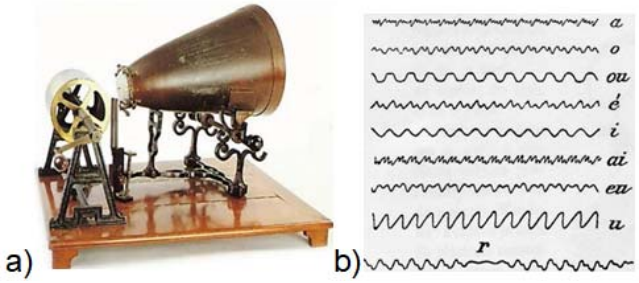
Streszczenie. *W niniejszej pracy przedstawiono ogólnie rozwój technologii rozpoznawania mowy, począwszy od pierwszych eksperymentów XIX wieku, aż po współczesne osiągnięcia w tej dziedzinie. Przeanalizowano przekształcenia technologiczne na przestrzeni ostatnich lat, omówiono kluczowe odkrycia oraz najważniejsze wydarzenia, które odegrały istotną rolę w rozwoju tej dziedziny, wskazując jednocześnie wybrane procesy wspomagające skuteczność rozpoznawania mowy pod kątem identyfikacji biometrycznej. Przedstawiono w zarysie charakterystyczne cechy wymowy dla języka polskiego.*

Abstract. *. This paper presents a general overview of the development of speech recognition technology, from the first experiments of the 19th century to modern developments in this field. It analyses technological transformations over the past years, discusses key discoveries and key events that have played a significant role in the development of this field, while highlighting selected processes that support the effectiveness of speech recognition in terms of biometric identification. The characteristic features of pronunciation for the Polish language are outlined..* ***Voice recognition and speaker identification: a review of selected methods for recognising biometric features of speech.***

Słowa kluczowe: ASR, cechy biometryczne głosu, głębokie uczenie, sieci neuronowe,
Keywords: automatic speech recognition (ASR), biometric voice identyfication, deep learning, neural networks

Wprowadzenie

Rozpoznawanie głosu i identyfikacja mówców to dziedzina nauki zajmująca się identyfikacją osób na podstawie ich głosu. Jest to ważny obszar badań z praktycznymi zastosowaniami w dziedzinach takich jak biometria, transkrypcja mowy czy analiza emocji. Niewątpliwie, początki rozpoznawania mowy (ASR - Automatic Speech Recognition) mają długą i bogatą historię, sięgającą XIX wieku. Jednym z najważniejszych kroków w kierunku zrozumienia tego, jak mowa może być analizowana i przedstawiana graficznie, to wynalezienie przez É.-L. Scott'a de Martinville'a urządzenia – 'fonoautografu' w 1857 roku. (rysunek 1). Dzięki temu wynalazkowi, powstało wiele prac dotyczących analizy głosu i mowy, jak prace H.Schneebeli, który dokonał pierwszej analizy dźwięków mowy na podstawie twierdzeń Fouriera w 1878 roku. Prace te stały się fundamentem dla dalszych badań nad modelowaniem oraz procesami przetwarzania sygnałów mowy.

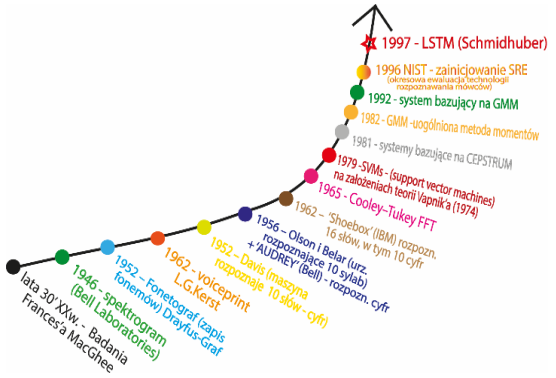


Rys.1. a). *Phonoautograph* Scott'a de Martinville z 1857 roku [1]; b). ślady dźwięków mowy z fonoautografu (badania H.Schneebeli'ego, 1878 rok) [2].

Jako ciekawostkę, można wymienić w tym miejscu urządzenie - "RADIO REX" - pierwszą zabawkę uruchamianą głosem, patentu H.C. Berger'a z 1913 roku. Uruchomienie zabawki następowało poprzez wprawienie w drgania metalowego stroika efektem rezonansu z tonem krtaniowym (F0) samogłoski "E" (około 500 Hz) podczas wymawiania słowa: "REX". Choć pierwsze oficjalne próby identyfikacji osoby na podstawie analizy głosu, to lata trzydzieste XX wieku (F.MacGhee), to na poważne efekty rozwoju nowoczesnych technologii rozpoznawania mowy trzeba było poczekać [3]. Dopiero lata 50. i 60. XX wieku i gwałtowny rozwój technologii, dającej wsparcie wykonawcze dla opracowywanych koncepcji, wygenerowały silny impuls inicjujący badania i powstawanie nowych metod rozpoznawania głosu i identyfikacji mówcy. I tak: już w roku 1960 G. Fant opracowuje fizjologiczny model ludzkiego głosu, który tworzy podstawy do analizy mowy, a w roku 1962, zostaje opracowany: 'Voiceprint Identyfication' - sonogram głosu - elektroniczne tłumaczenie dźwięku na graficzną reprezentację, który możemy nazwać pierwszym systemem ASR - automatycznego rozpoznawania mówcy (L.G.Kersta, Bell Laboratories). Jednak rozwój technologii ASR (rysunek 2) nastąpił dopiero w trwającej obecnie epoce gwałtownej ewolucji systemów sztucznej inteligencji (AI). Początkowo, skupiano się głównie na rozpoznawaniu pojedynczych fonemów i na zwiększaniu zdolności (głównie ilości) rozpoznawania przez systemy wyrazów, oraz na modyfikowaniu i kombinowaniu metod, poprawiając sprawność i wydajność systemów ASR [4,5]. Krokiem milowym stało się wykorzystanie metod przetwarzania języka naturalnego NLP (ang. neuro-linguistic programming) [6]. NLP pozwala na analizę i interpretację tekstu w sposób podobny do ludzkiego myślenia, co umożliwia przetwarzanie mowy na pismo oraz identyfikacje charakterystycznych cech, takich jak długość i rodzaj zdania, czy użycie słów kluczowych. Końcówka XX wieku to przełomowy moment dla ASR, gdyż prezentacja przełomowej metody LSTM (ang. Long Short-Term Memory) rekurencyjnej sieci neuronowej (RNN), w 1997 roku, umożliwiła rozwój systemów end-to-end (E2E) [7]. Obecnie, to właśnie systemy E2E są wykorzystywane do ASR, gdyż uczą się jednocześnie na wszystkich modułach, mapując ramki akustyczne na fonemy w jednym kroku, a to rewolucyjna zmiana paradygmatu ASR z cech cepstralnych, na cechy dyskryminacyjne pozyskiwane bezpośrednio z surowej mowy.

W dalszej części niniejszego artykułu nakreślono wybrane metody ASR, od tradycyjnych, jak modele statystyczne HMM (ang. Hidden Markov Model), GMM-UBM (ang. Gaussian Mixture Model - Universal Background Model), i-vector (GMM z analizą czynnikową - Factor Analisys), po nowoczesne modele głębokiego uczenia (ang. DeepLerning) [8-10]. Wybór przedstawionych tu metod ekstrakcji cech mowy i klasyfikacji mówców został dokonany, biorąc pod uwagę ogólną efektywność tych technik oraz ich potencjalne zastosowanie w systemach

ASR i biometrii głosu, w kontekście specyficznych cech języka polskiego. W niniejszej pracy skupimy się na Deep speaker embeddings i metodzie x-vector, oraz jej wykorzystaniu do przeprowadzenia badań, w kontekście rozpoznawania cech biometrycznych w wymowie w języku polskim. Z uwagi na to, że większość wyżej wymienionych metod została już bardzo szczegółowo opisana w literaturze, nie będziemy wnikliwie ich tu omawiać. Zachęcamy za to, do zapoznania się z proponowaną literaturą źródłową. Omówimy także niektóre etapy procesu przetwarzania i analizy sygnału mowy oraz wybrane cechy charakterystyczne głosu dla mowy. Przedstawimy także zarys specyfiki wymowy języka polskiego, przybliżając kilka jego unikalnych cech, jak na przykład różne rodzaje szumów czy dźwięki nosowe, które mogą stanowić wyzwanie dla systemów ASR.



Rys.2. Rozwój technologii ASR w XX wieku.

Etapy klasycznego procesu przetwarzania i analizy sygnału mowy

Przetwarzanie wstępne (ang. pre-processing), to zestaw kluczowych działań, przeprowadzanych w początkowej fazie procesu ASR, na zarejestrowanym sygnale mowy. W ramach tego etapu, sygnał mowy najpierw jest digitalizowany (próbkowanie i kwantyzacja), co pozwala na dalsze przetwarzanie sygnału przez komputer. Następnie jest poddawany oczyszczaniu i usuwane są z niego fragmenty niebędące mową (filtracja). Po tym etapie wykonywane są zadania wykrywania końca sygnału, oraz preemfaza. Następnie stosuje się różne metody ekstrakcji unikatowego wektora cech, na etapie parametryzacji (odwzorowania dowolnych obiektów za pomocą wektorów liczb o skończonej długości) sygnału mowy, odgrywającego kluczową rolę w systemach ASR. Celem generatora cech jest opisanie sygnału mowy za pomocą jak najmniejszej liczby deskryptorów, bez utraty istotnych informacji dla rozpoznawania mówcy. Parametryzacja powinna być odporna na warunki akustyczne, techniczne rejestracji oraz zawartość lingwistyczną nagranego materiału. Można wydzielić dwie grupy parametrów: czasowe i widmowe (formantowe, cepstralne i liniowe kodowanie predykcyjne LPC (ang. linear predictive coding)). Jednymi z najczęściej stosowanych parametrów, są tak zwane 'parametry mel-cepstralne' MFCC (ang. Mel Frequency Cepstral Coefficients) - szeroko stosowane w akustyce mowy oraz w kompresji sygnałów fonicznych. Powstają z cepstrum sygnału przedstawionego w skali melowej (mel-cepstrum). Metoda ta opiera się na skali perceptualnej, która oddaje sposób postrzegania dźwięku przez człowieka i pozwala na efektywną separację cech mowy. W odróżnieniu od MFCC, Dyskretna transformata falkowa DWT (ang. discrete wavelet transform), posiada dobrą rozdzielczość dla niskich częstotliwości, co daje lepszą lokalizację zjawisk przejściowych w dziedzinie czasu. Wspomnieć trzeba jeszcze o: falkach ortagonalnych (Daubechies, Haar) i różnej ilości poziomów dekompozycji. Nie można pominąć tu LPC (liniowe kodowanie predykcyjne -analiza w dziedzinie czasu – rezonansowa struktura toru głosowego), percepcyjnej predykcji liniowej PLP (lepsza skuteczność przy zakłóceniach) [11]. Należy powiedzieć także o technikach mieszanych, jak RASTA-PLP (ang. Relative Spectral Transform - Perceptual Linear Prediction), czy innych: vector quantisation, PCA (ang. principal component analysis), LDA (ang. linear discriminant analysis). W efekcie, cechy mowy, takie jak częstotliwość podstawowa, formanty czy kształt widma, są oddzielone i mogą być wykorzystane w procesie analizy sygnału, by uzyskać informacje istotne dla identyfikacji mówcy [12]. Podsumowując, ekstrakcja cech, jest stosowana w fazach szkoleniowej i testowej systemów rozpoznawania mówcy, ma na celu przekształcenie cyfrowego sygnału mowy na zbiory wektorów cech lub opisów liczbowych, które zawierają podstawowe charakterystyki głosu mówcy. Modelowanie mówcy (ang. speaker modelling), polega na generowaniu algorytmów rozpoznawania mówcy dla dopasowania cech głosu mówcy. Metody zawierające wzmocnione informacje specyficzne dla mówcy z kompresowanym wolumenem nazywane są modelami mówcy. W trakcie szkolenia lub rejestrowania, modele stanowe mówcy są generowane poprzez wykorzystanie konkretnych cech wyodrębnionych z głosu [13]. Istnieją różne klasyfikacje ASR, które służą do kategoryzowania różnych typów systemów rozpoznawania mówcy. Te klasyfikacje obejmują m.in. rozpoznawanie zależne od tekstu i niezależne od tekstu, weryfikację mówcy i identyfikację mówcy. Zależność od tekstu dotyczy słów kluczowych lub fraz kluczowych dla rozpoznania głosu, podczas gdy odmiana niezależna - jest bardziej elastyczna.

Identyfikacja – SI (ang. speaker identification), weryfikacja – SV (ang speaker verification) i diaryzacja – SD (speaker diarization).

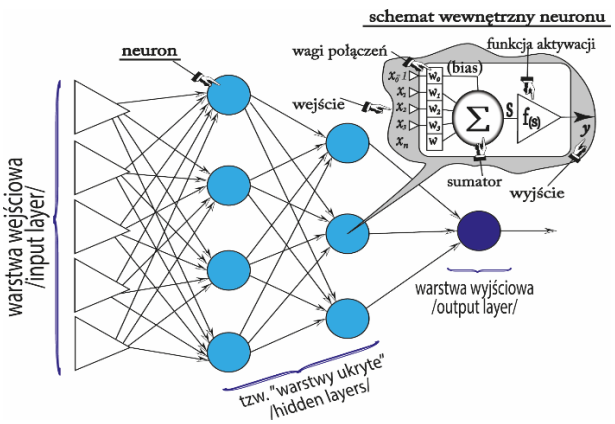
SI określa tożsamość anonimowego mówcy na podstawie wypowiedzi mówcy. SI znajduje dokładnego mówcę z zestawu rozpoznanych głosów na podstawie różnych wypowiedzi zawartych w bazie danych). Podejście, to 1:N dopasowanie, w którym określona wypowiedź jest porównywana z N-szablonami. SV zajmuje się uwierzytelnianiem głosu określonej tożsamości mówcy. Charakterystyki uzyskane przez system SI są porównywane z charakterystykami wszystkich mówców składających się na bazę modeli głosowych. W przypadku systemów SV uzyskane cechy są tylko połączone z cechami przechowywanymi przez mówcę, który twierdzi, że to on lub ona jest tożsamością. Jest to dopasowanie 1:1, w którym wypowiedź jednego mówcy jest porównywana z jednym szablonem. SD to podział głosu z wieloma ludźmi na jednorodne segmenty związane z każdą osobą. Jest to istotna część systemów rozpoznawania mówcy. Ma zastosowanie w wielu kluczowych dziedzinach, takich jak tworzenie napisów do filmów, zrozumienie treści jakichkolwiek rozmów, itp. Dzięki zrozumieniu tych różnych klasyfikacji i elementów systemu ASR, można projektować i rozwijać bardziej skuteczne i dokładniejsze modele rozpoznawania mówcy [14].

Metody ASR, Modelowanie statystyczne. Modele HMM, GMM-UBM, i-vector.

Jednym z podstawowych podejść do rozpoznawania mówców jest modelowanie statystyczne, w tym HMM i GMM-UBM. W obu metodach, cechy charakterystyczne mowy mówcy są analizowane i wykorzystywane do stworzenia modelu statystycznego, który pozwala na identyfikację mówcy na podstawie próbek jego głosu. W metodzie HMM, w procesie treningowym, wypowiedź

mówcy jest dzielona na sekwencję stanów ukrytych, a następnie estymowane są parametry modelu HMM na podstawie sekwencji stanów i odpowiadających im cech mowy. Następnie, sekwencja stanów jest szacowana dla próbki mowy testowej i porównywana z sekwencjami stanów dla mówców referencyjnych w celu wyznaczenia mówcy. W metodzie GMM-UBM, najpierw tworzony jest model tła (UBM), który opisuje zmienność w danych mowy, następnie estymowane są modele GMM dla każdego mówcy referencyjnego. W procesie rozpoznawania, próbka mowy testowej jest porównywana z modelem tła i modelami GMM mówców referencyjnych, a mówca z najwyższym prawdopodobieństwem zostaje wybrany jako wynik. W obu metodach, cechy mowy są ekstrahowane z próbek mowy i przetwarzane w procesie treningowym i rozpoznawania. W celu zminimalizowania wpływu właściwości kanału transmisji na cechy osobnicze głosu, stosuje się różne metody normalizacji, w tym filtrację opartą na parametrach cepstralnych. Dodatkowo, w metodzie GMM-UBM, stosowane są również techniki wyrównywania kanału, takie jak MAP (łać. maximum a posteriori - podejście, które łączy informacje z danych obserwowanych, z wcześniejszymi przekonaniami o parametrach, aby znaleźć najlepsze oszacowanie wartości parametrów modelu) i NAP (ang. nuisance attribute projection - projekcja nieistotnych atrybutów). Metoda i-vector, wykorzystuje model generatywny GMM-UBM lub DNN (ang. deep neural network), do automatycznej ekstrakcji cech z sygnału mowy [8,15]. Metoda ta wymaga minimalnego zaangażowania użytkownika, ponieważ nie wymaga ręcznej definicji cech. Polega ona na tworzeniu niskowymiarowej reprezentacji cech mowy (i-wektory), która uwzględnia zarówno cechy indywidualne mówcy, jak i cechy wspólne dla różnych mówców. Co ważne, metoda ta, umożliwia skuteczne rozpoznawanie mówcy nawet w obecności szumów i innych zakłóceń sygnału wejściowego.

Era głębokich sieci neuronowych (rysunek 3) [16-18]. Neuronowe sieci konwolucyjne CNN (ang. convolutional neural network), czy rekurencyjne RNN (ang. recurrent neural network) i sieci neuronowe długiej krótkotrwałej pamięci (LSTM), są nowoczesnym podejściem do rozpoznawania mówców. Te techniki wykorzystują hierarchiczne struktury i adaptacyjne algorytmy uczenia się, co pozwala na skuteczne modelowanie cech mowy, klasyfikację mówców i radzenie sobie z szumami.



Rys.3. Ogólny zarys modelu działania sieci neuronowej - przykład sieci jednokierunkowej, wielowarstwowej MLP (ang. multi layer perceptron), z przykładowym schematem budowy neuronu.

Systemy End-to-end (E2E) ASR oferują jednolite podejście do automatycznego rozpoznawania mowy, eliminując potrzebę etapów ekstrakcji cech i modelowania mowy. Dwie popularne architektury E2E ASR to: LAS (ang. listen, attend and spell) oraz CTC (ang. connectionist temporal classification). LAS wykorzystuje rekurencyjne sieci neuronowe (RNN) do przetwarzania sekwencji sygnału mowy, a następnie stosuje mechanizm uwagi do dekodowania przetworzonej sekwencji CTC pozwala na równoczesne uczenie się i rozpoznawanie sekwencji, eliminując potrzebę segmentacji i etykietowania [19,20].

Deep Speaker Embeddings [21-23], systemy te polegają na wykorzystaniu głębokich sieci neuronowych do uczenia się reprezentacji wektorowych mówców na podstawie próbek głosu. Przykładem tego podejścia jest model x-vector, który wykorzystuje sieci neuronowe konwolucyjne (CNN) do przetwarzania sekwencji sygnału mowy i ekstrakcji cech. Metoda ta jest wydajna, gdyż pozwala na jednoczesne modelowanie wielu mówców.

Transformer-based ASR

Architektura Transformer została stworzona z myślą o tłumaczeniu maszynowym oraz przetwarzaniu języka naturalnego. W przeciwieństwie do rekurencyjnych sieci neuronowych (RNN), architektura Transformer nie korzysta z rekurencji, a z enkoderów i dekoderów, które są połączone za pomocą warstw samouwagi. Enkodery odpowiadają za generowanie reprezentacji wejściowych danych, podczas gdy dekodery służą do generowania wyjściowych danych. Każdy blok enkodera i dekodera składa się z wielu warstw, a każda warstwa ma wiele neuronów i przetwarza dane sekwencyjnie, wykorzystując informacje z poprzednich warstw (metoda autoagresywna) do wygenerowania wyniku przetwarzania, co może poprawić jakość generowanych reprezentacji sekwencji. W kontekście rozpoznawania mówców, modele Transformer-based ASR, pozwalają na efektywną ekstrakcję cech mowy i klasyfikację mówców, nawet w obecności szumów i przeszkód akustycznych [24].

Metody oceny wydajności

Wydajność automatycznych systemów rozpoznawania mówców jest powszechnie oceniana za pomocą współczynnika błędu równego EER (ang. equal error rate) oraz funkcji kosztu decyzji DCF (ang. decision cost function) [25]. Współczynnik błędu równego (EER) to algorytm systemu biometrycznego, który określa wartości progowe dla współczynników fałszywych akceptacji i fałszywych odrzuceń Wartość EER wskazuje, że proporcje fałszywych akceptacji są równe proporcjom fałszywych odrzuceń. Im niższa wartość EER, tym wyższa dokładność systemu biometrycznego. Alternatywnie, funkcja kosztu decyzji uwzględnia prawdopodobieństwa wystąpienia mówcy docelowego, proporcję mówców docelowych i nie docelowych. Funkcja kosztu detekcji to jednoczesny miernik dyskryminacji i kalibracji. Często minimalna wartość krzywej DCF nazywana jest minDCF.

Cechy charakterystyczne głosu ludzkiego

Każdy głos ludzki jest unikalny i niepowtarzalny, ponieważ jest wynikiem współbrzmienia wielu tonów składowych. Ton podstawowy powstaje w krtani, jednak podczas przepływu przez jamy gardłowe i nosowe, dodawane są do niego kolejne harmoniczne, co wpływa na barwę głosu. Tor głosowy u człowieka, działa bowiem jak układ filtrów o różnych częstotliwościach rezonansowych, zależnych od długości i kształtu całego toru [26]. Indywidualność głosu zależy więc w dużej mierze od intensywności i rozmieszczenia składowych dźwięków.
W skład ogólnej charakterystyki indywidualnej osoby mówiącej wchodzą:

- budowa fizyczna kanału głosowego (wpływa na pozycję, energię oraz kształt formantów głosowych)
- kształt jamy nosowej (brzmienie głosek nosowych)
- cechy behawioralne: idiolekt, dialekt, prozodia, nawyki mówcy [27].

Elementy te wpływają na właściwości spektralne mowy, szczególnie na tak zwane formanty, które są maksymalnymi wartościami amplitudy widma mowy. Formanty odgrywają kluczową rolę w rozpoznawaniu dźwięków mowy oraz identyfikacji mówcy.

Częstotliwości mowy

Najwyższa wartość częstotliwości, jaką człowiek może wytworzyć głosem, wynosi około 10 kHz, podczas gdy najniższa wartość - około 70 Hz. Możemy wyróżnić trzy główne grupy częstotliwości mowy:

1. Podstawowe – zakres około 80-300 Hz.
2. Samogłoskowe – dźwięki te, ponieważ zawierają większość energii i mocy głosu mówionego, mieszczą się w zakresie 250 Hz-2,8 kHz.
3. Spółgłoskowe – spółgłoski zajmują pasmo, z zakresu pomiędzy 1,5 kHz, a 4 kHz. Pomimo, że nie niosą ze sobą zbyt wiele energii, to są kluczowe dla zrozumiałości mowy.

Ad.1: Częstotliwość podstawowa, F0, to najniższa częstotliwość harmoniczna w dźwięku periodycznym. W mowie, F0 jest ściśle związana z wysokością głosu mówcy. Dla mężczyzn, F0 zwykle wynosi od 80 do 180 Hz, dla kobiet od 165 do 255 Hz, a dla dzieci od 250 do 300 Hz. F0 jest często zwana formantem krtaniowym.

Ad.2: Formanty - to szczególne częstotliwości dźwięków mowy, wzmacniane przez rezonans kształtu jamy ustnej podczas mówienia. Są one kluczowe dla rozpoznawania samogłosek. W mowie ludzkiej wyróżnia się zwykle trzy do pięciu głównych formantów (F1,…,F5). F1 i F2 są najważniejsze dla rozpoznawania samogłosek i zależą od ułożenia języka (dzielącego jamę ustną na dwie wnęki), podczas gdy F3 i wyższe - wpływają na barwę głosu. Szerokość pasma formantów odnosi się do szerokości zakresu częstotliwości wokół formantów, w którym następuje wzmocnienie sygnału. Wartości te są istotne dla analizy cech akustycznych mowy. Szerokość pasma formantów samogłoskowych ogółem u dorosłych zwykle wynosi od 269 Hz do 821 Hz dla F1, od 899 Hz do 2752 Hz dla F2.

Częstotliwości przejściowe są równie istotne w rozpoznawaniu spółgłosek i występują, gdy dźwięki mowy zmieniają się z jednego na drugi [28]. W analizie mowy często bada się "delta" cechy, które opisują szybkość zmiany sygnału mowy.

Zarys charakterystycznych cech wymowy dla języka polskiego

W językach świata ogólnie występuje od 11 do 75 fonemów. Ich liczba w zależności od języka różni się znacznie. Wszystkie języki zawierają oczywiście samogłoski i spółgłoski, z prostym systemem dźwiękowym składającym się z trzech samogłosek ({i}, {u}, {a}) i siedmiu spółgłosek ({s}, {p}, {t}, {k}, {m}, {n}, {r} lub {l}). To tak zwany system prymarny, który ma proporcję samogłosek do spółgłosek wynoszącą 3:7, co stanowi punkt odniesienia dla typologii fonologicznej języków świata. Niektóre języki, jak hawajski, prawie nie zmieniają systemu podstawowego, ale większość dodaje dodatkowe fonemy. Języki z więcej niż 30% samogłosek są nazywane samogłoskowymi, jak angielski, podczas gdy te z więcej niż 70% spółgłosek są określane jako spółgłoskowe. Wszystkie języki słowiańskie są językami spółgłoskowymi. Język polski, rozbudował system wokaliczny w porównaniu do systemu podstawowego, dodając samogłoski średnie: przednią {e} i tylną {o}. W porównaniu z językiem angielskim, polski system wokaliczny jest jednak skromniejszy [29]. System konsonantyczny za to, rozbudował się znacznie, dodając nowe typy spółgłosek, takie jak zwarto-szczelinowe (afrykaty) {c}, {č}, {ć}, {z}, {ż}, {ź} oraz szczelinowe (frykaty) [23]. Te spółgłoski tworzą charakterystyczne dla języka polskiego trzy szeregi spółgłosek artykułowanych dzięki ruchom języka w przedniej części jamy ustnej: zębowe ({s}, {z}, {c}, {Z}), dziąsłowe ({š}, {ž}, {č}, {ż}) i środkowojęzykowe ({ś}, {ź}, {ć}, {ź}). Ze względu na ich właściwości dźwiękowe, nazywane są odpowiednio spółgłoskami syczącymi, szumiącymi lub ciszącymi. Znane są również jako sybilanty. To one powodują, że w strumieniu mowy wyróżniają się syczące i – szczególnie w języku polskim – szeleszczące dźwięki, które mają charakter sygnałów szumowych [30]. Sygnały szumowe zajmują bardzo szerokie pasmo na skali częstotliwości, co sprawia, że potrafią skutecznie zamaskować występujące tuż po nich samogłoski czy spółgłoski „niesyczące” (jak np.{d}, {f}, {h}, {k}, {t} itd.) [31,32].

Wybór metody Deep Speaker Embeddings dla pracy nad rozpoznawaniem cech biometrycznych mówcy w języku polskim

W celach identyfikacji biometrycznej mówców w języku polskim, metoda Deep Speaker Embeddings, w szczególności model x-vector, wydaje się być szczególnie obiecująca. Głównym powodem tego wyboru jest zdolność modelu do jednoczesnego modelowania wielu mówców, co pozwala na efektywną identyfikację nawet wśród dużej liczby użytkowników. Aby dostosować model x-vector do specyfiki języka polskiego, można wprowadzić następujące modyfikacje. W pierwszym etapie, zebranie dużego zbioru danych mowy w języku polskim jest kluczowe. Powinien on obejmować różnorodne próbki mowy od mówców różnych grup wiekowych, płci, akcentów i dialektów. Dodatkowo, dane powinny być zebrane w różnych warunkach akustycznych, aby uwzględnić różne poziomy szumów tła i odległości od mikrofonu. Preprocessing: Język polski ma specyficzne fonemy i struktury, które wpływają na charakterystykę głosu mówcy. W związku z tym, konieczne jest odpowiednie przetwarzanie danych wejściowych, aby uwzględnić te specyficzne cechy. Można to osiągnąć poprzez zastosowanie odpowiednich algorytmów ekstrakcji cech, które będą uwzględniać cechy fonetyczne języka polskiego. Architektura modelu: Architektura modelu x-vector powinna być dostosowana do języka polskiego, uwzględniając specyfikę głosu mówców tego języka. Może to obejmować zmiany w liczbie warstw sieci neuronowej, liczbie neuronów w każdej warstwie, a także modyfikacje w funkcjach aktywacji lub technikach regularyzacji [33]. Trenowanie i walidacja modelu**:** Model x-vector powinien być trenowany na zebranych danych mowy w języku polskim. W celu osiągnięcia najlepszej wydajności, można zastosować techniki uczenia transferowego, w których wstępnie wytrenowany model jest dalej dopasowywany do danych specyficznych dla języka polskiego. Podczas procesu trenowania, ważne jest również przeprowadzenie walidacji krzyżowej, aby ocenić skuteczność modelu na danych niewidocznych. Optymalizacja i testowanie**:** Po wytrenowaniu modelu na danych języka polskiego, można przeprowadzić optymalizację hiperparametrów i pozostałych aspektów architektury modelu, aby uzyskać najlepszą możliwą wydajność. Proces optymalizacji może obejmować strojenie hiperparametrów: współczynnika uczenia, rozmiaru wsadu czy techniki regularyzacji. Po optymalizacji modelu, należy przeprowadzić testy na niezależnym zbiorze danych, aby ocenić jego ogólną skuteczność w identyfikacji mówców w języku polskim. Implementacja. Po wytrenowaniu i optymalizacji modelu x-

vector, można go przykładowo zintegrować z istniejącymi systemami biometrycznymi, takimi jak systemy kontroli dostępu, identyfikacji głosowej czy monitoringu. Dzięki wykorzystaniu tego modelu, systemy te będą w stanie identyfikować mówców w języku polskim z większą precyzją i niezawodnością. Wprowadzenie powyższych modyfikacji do modelu x-vector pozwoli na stworzenie efektywnego narzędzia identyfikacji biometrycznej dla mówców języka polskiego .

Wnioski

Biometria głosu, jest ważnym obszarem badań, który ma wiele praktycznych zastosowań w dziedzinach takich jak transkrypcja mowy czy analiza emocji [34, 35]. W niniejszym artykule przedstawiliśmy różne etapy procesu przetwarzania i analizy sygnału mowy, zarys specyfiki mowy języka polskiego oraz wybrane cechy charakterystyczne głosu ludzkiego. Analizując specyfikę mowy w danym języku, można pod tym kątem wstępnie dobrać odpowiednią metodę biometrycznej analizy głosu. Dalsze badania w tej dziedzinie mogą koncentrować się na porównaniu wydajności tego modelu z innymi podejściami oraz na eksploracji możliwości zastosowania modelu do innych zadań, takich jak analiza emocji czy transkrypcja mowy.

***Autorzy**: mgr Tomasz Śliwak-Orlicki, E-mail: tomasz.sliwak-orlicki@awl.edu.pl, dr inż. Krzysztof Górski, E-mail: krzysztof.gorski@awl.edu.pl., Katedra Zarządzania Innowacyjnymi Projektami, ul. Czajkowskiego 109, 51-147 Wrocław.*

LITERATURA

[1] Źródło: www.teylersmuseum.nl /nl/ collectie/ instrumenten/fk-0275-phonautograph-after-leon-scott, dostęp z dnia 20.04.2023 r.
[2] B. Teston, A la poursuite de la trace du signal de parole, *Journées d'Etude sur la Parole (JEP)*, Jun 2006, 7-10.
[3] A. D. Yarmey, M.J.Yarmey, L. Todd; Frances McGehee (1912–2004: The First Earwitness Researcher, *Perceptual and Motor Skills*, 2008, 387-394.
[4] C. D. Shaver, J. M. Acken, A Brief Review of Speaker Recognition Technology, *Electrical and Computer Engineering Faculty,* 2016, 19320.
[5] M. M. Kabir, M. F. Mridha, J. Shin, I. Jahan and A. Q. Ohi, A Survey of Speaker Recognition: Fundamental Theories, Recognition Methods and Opportunities*, IEEE Access,* vol. 9, 2021, 79236-79263.
[6] D. Keith D. Foote, A Brief History of Natural Language Processing (NLP), *Dataversity,* 2019.
[7] J. Oruh, S. Viriri and A. Adegun, Long Short-Term Memory Recurrent Neural Network for Automatic Speech Recognition, *IEEE Access*, 10, 2022, 30069-30079.
[8] D. Sztahó, G. Szaszák, A. Beke, Deep Learning Methods in Speaker Recognition: A Review, *Period. Polytech. Elec. Eng. Comp. Sci.*, vol. 65, no. 4, 2021, 310–328.
[9] J. Chang and D. Wang, Robust speaker recognition based on DNN/i-vectors and speech separation, *IEEE ICASSP,* 2017, 5415-5419.
[10] A. Mobiny, M. Najarian, Text-Independent Speaker Verification Using Long Short-Term Memory Networks, *arXiv*, 2018.
[11] M. Katz, S. Krüger, M. Schafföner, E. Andelic, A.Wendemuth, "Speaker Identification and Verification Using Support Vector Machines and Sparse Kernel Logistic Regression, *Springer*, 2006, 176-184.
[12] E. Majda-Zdancewicz, A. P. Dobrowolski, Ocena przydatności wybranych cech sygnału mowy wyróżniających osoby ze zmianami neurodegradacyjnymi, *PE*, 11, 2019.
[13] E. Majda, A. P. Dobrowolski, B. L. Smólski, „Modelowanie i optymalizacja generatora cech dla systemu rozpoznawania mówcy", *Biuletyn WAT,* The Phonology of Polish *Vol. LXI, Nr 4, 2012*
[14] P. Walendowski, Zastosowanie sieci neuronowych typu SVM do rozpoznawania mowy, praca doktorska, *Politechnika Wrocławska*, 2008.
[15] J. Guo, N. Xu, K. Qian, Y. Shi, K. Xu, Y. Wu, A. Alwan, Deep neural network based i-vector mapping for speaker verification using short utterances", *arXiv,* 2018, 1810.07309.
[16] A. Géron, Hands-On Machine Learning with Scikit-Learn, Keras,and TensorFlow, *O'Reilly Media, Inc*., 2nd Edition, 2019.
[17] W. Yin, K. Kann, M. Yu, H. Schütze, Comparative study of CNN and RNN for natural language processing, *arXiv*, 2017, 1702.01923.
[18] S. Hochreiter and J. Schmidhuber, Long short-term memory, *Neural Computation*, 9, no. 8, 1997, 1735–1780.
[19] C. Feng Yeh, J. Mahadeokar, K. Kalgaonkar, Y. Wang, D. Le, M. Jain, K. Schubert, C. Fuegen, M. L. Seltzer, Transformertransducer: End-to-end speech recognition with self-attention, *ArXiv*, , 2019, 1910.12977.
[20] G, Pundak, T. N. Sainath, R. Prabhavalkar, A. Kannan, D. Zhao, Deep context: end-to-end contextual speech recognition, *arXiv*, 2018, 1808.02480.
[21] Q. Wang, K. Okabe, K. A. Lee, H. Yamamoto, T. Koshinaka, Attention Mechanism in Speaker Recognition: What Does It Learn in Deep Speaker Embedding?, *arXiv*, 2018, 1809.09311.
[22] S. Novoselov, A. Shulipa, I. Kremnev, A. Kozlov, V. Shchemelinin, On deep speaker embeddings for text-independent speaker recognition, *arXiv,* 2018, 1804.10080.
[23] D.Snyder, D.Garcia-Romero, G.Sell, D.Povey, S.Khudanpur, "X-Vectors: Robust Dnn Embeddings For Speaker Recognition, *IEEE,* Conerence (ICASSP), 2018.
[24] M.Orken, O.Dina, A.Keylan, A study of transformer-based end-to-end speech recognition system for Kazakh language. *Sci Rep*, 12, 2022, 8337.
[25] N. Brümmer, E. de Villiers, The BOSARIS Toolkit: Theory, Algorithms and Code for Surviving the New DCF, *ArXiv*, 2011, 1304.2865.
[26] R. Tadeusiewicz, Sygnał mowy, *WKiŁ,*Warszawa,1988.
[27] A. Wagner, J. Bachan, K. KLessa G. Demenko, Przegląd wybranych aspektów analizy prozodii mowy spontanicznej na potrzeby technologii mowy, *PF,* (LXVI) 2015, *271*–298.
[28] I. R. Titze, Principles of Voice Production, *Prentice Hall,* 1994.
[29] K. Johnson, P. Ladefoged, M. Lindau, Individual differences in vowel production, *J Acoust Soc Am,* 94, 1993, 701–714.
[30] M. Mela, V. Schulte, Jak piękny jest ludzki głos! Techniki wizualizacji, mierzenia i deskrypcji ludzkiego głosu, *Lingwistyka Stosowana,* 4/2016, 91–103.
[31] M. Kaniewska, Zespolona pulsacja chwilowa w analizie i konwersji głosu, Rozprawa doktorska, *Wydział Elektroniki, Telekomunikacji i Informatyki*, Politechnika Gdańska, Gdańsk, 2011.
[32] W. Jassem, Acoustic-phonetic variability of Polish vowels, *Department of Acoustic Phonetics,* Institute of Fundamental Technological Research, Polish Academy of Sciences, 1992, 61-704.
[33] D. Yin, M. Farajtabar, A. Li, N. Levine, A. Mott, Optimization and Generalization of Regularization-Based Continual Learning: a Loss Approximation Viewpoint", arXiv, 2006, 10974.
[34] H. N. M. Shah, M. Z. Ab Rashid, M. F. Abdollah, M. N. Kamarudin, Ch. K. Lin, Z. Kamis. Biometric Voice Recognition in Security System", *Indian Journal of Science and Technology,* 7(2), 2014, 104-112.
[35] D. Kamińska, A. Pelikant. Zastosowanie multimodalnej klasyfikacji w rozpoznawaniu stanów emocjonalnych na podstawie mowy spontanicznej, *Informatyka, Automatyka, Pomiary w Gospodarce i Ochronie Środowiska*, 2012, 36-39.