Izabela Krzosek-Piwowarczyk, Olga Komosa
Waldemar Maciejko

Kryminalistyczna identyfikacja mówcy maskującego głos

Wstęp

Kryminalistyczna analiza wypowiedzi nieznanego mówcy, a przede wszystkim jego identyfikacja, wymaga wskazania w badanym materiale określonych cech. Mogą one stanowić na przykład opis sposobu realizacji dźwiękowej poszczególnych jednostek fonetycznych, wyniki pomiarów cech fizycznych lub opisową analizę spektrograficzną. Cechy te są użyteczne w analizie kryminalistycznej pod warunkiem, że niosą określone informacje o mówcy. Mogą one wyróżniać go spośród innych osób, wskazywać na region geograficzny, z którego pochodzi, informować o stanie emocjonalnym. Założeniem reprezentatywności analizy tego typu jest prawdziwość obserwowanych cech, to znaczy pewność, że nie zostały zniekształcone na przykład przez kanał telekomunikacyjny, urządzenie rejestrujące, warunki akustyczne panujące w trakcie rejestracji itp. Inną przyczyną zniekształceń może być maskowanie, czyli zamierzone działanie, którego wynikiem jest zmodyfikowane brzmienie głosu lub sposobu artykulacji w celu ukrycia cech osobniczych przekazywanych wraz z mową. Sprawcy przeróżnych przestępstw podejmują próby ukrycia cech osobniczych. Do przestępstw tych należą m.in. wyłudzenia środków pieniężnych za pośrednictwem linii telefonicznej, groźby karalne, ciężkie przestępstwa przeciwko życiu i zdrowiu, na przykład porwania dla okupu, oraz przestępstwa o charakterze terrorystycznym.

W literaturze przeprowadzono próbę usystematyzowania technik, którymi najczęściej posługują się mówcy w celu ukrycia cech osobniczych. Metody maskowania można podzielić na elektroniczne oraz nieelektroniczne. Maskowanie nieelektroniczne to wymuszanie nienaturalnej pracy narządu mowy i w efekcie zniekształcenie mowy. W tabeli poniżej przedstawiono nieelektroniczne techniki maskowania głosu [1–2].

Wpływ nieelektronicznych metod maskowania głosu stał się przedmiotem wielu analiz. W wyniku badań stwierdzono, że mówcy najchętniej modyfikują pracę krtani, manipulują intensywnością (głośnością) mowy, próbują ją nawet ubezdźwięczniać; zdarzają się również przypadki maskowania fonematycznego [3]. W wyniku analizy spektrograficznej stwierdzono, że modyfikowanie brzmienia głosu przez wymuszanie określonej artykulacji powoduje przesunięcie częstotliwości formantowych samogłosek, przy czym najmniejszym zmianom ulega formant pierwszy. W niektórych przypadkach maskowanie powoduje tłumienie wyższych formantów [4, 5]. Badania pokazały również, że maskowanie z wykorzystaniem metod nieelektronicznych powoduje wzrost współczynnika EER systemu automatycznego rozpoznawania mówców [3]. Istotnym zagadnieniem, które również doczekało się wielu opracowań, jest imitacja (naśladownictwo) mowy. Badania pokazały, że imitator jest w stanie tak manipulować pracą krtani oraz kanałem artykulacyjnym, że w pewnym zakresie parametry

Tabela 1

Nieelektroniczne techniki maskowania głosu [1–2]

Non-electronic techniques of voice disquise [1–2]

Technika maskowania	Przykładowy efekt
Fonacyjna	podniesiony lub obniżony ton krtaniowy, chrypka, szept, itp.
Fonematyczna	użycie nienaturalnego dla mówcy dialektu, imitacja wpływu języka obcego, imitacja wady wymowy, imitacja cech innej osoby, itp.
Prozodyczna	zmiana intonacji, inne usytuowanie akcentów, zmiana tempa wypowiedzi, itp.
Deformująca	tzw. mowa nosowa (poprzez zatkanie otworów nosowych), zniekształcenie (poprzez blokowanie ust), wypowiadanie się z przedmiotem w ustach itp.

Źródło (tab. 1–11): opracowanie własne

akustyczne jego głosu, w tym ton krtaniowy, mogą się zbliżyć do parametrów mówcy naśladowanego. Imitatorzy chętnie podrabiają zwłaszcza nawyki językowe mówców naśladowanych; jednak odtworzenie częstotliwości formantowych lub parametrów cepstralnych nie jest możliwe [6, 7]. Jedną z form maskowania głosu jest również prowadzenie rozmowy pod wpływem środków intoksykujących, na przykład alkoholu etylowego. Spożycie alkoholu powoduje m.in. wzrost częstotliwości tonu krtaniowego w funkcji ilości spożytego etanolu, spowolnienie tempa mowy, wzrost iloczasów oraz spadek intensywności mowy [8, 9].

Należy podkreślić, że zjawisko maskowania nieelektronicznego jest trudne do analizy laboratoryjnej ze względu na problem powtarzalności otrzymywanych wyników oraz pozyskanie reprezentatywnego materiału do badań. W rzeczywistych sprawach kryminalistycznych mówcy, ze względu na wymuszoną pracę aparatu mowy, nie są konsekwentni w sposobie modyfikowania głosu i mowy. Istotne z punktu widzenia identyfikacji może być powracanie w sposób nieświadomy do naturalnej artykulacji [10]. Praktyka pokazuje, że niektóre sposoby wpływania na pracę narządu mowy mogą być trudne do wykrycia. Dlatego też fundamentalną kwestią jest w ogóle stwierdzenie zaistnienia próby maskowania [5].

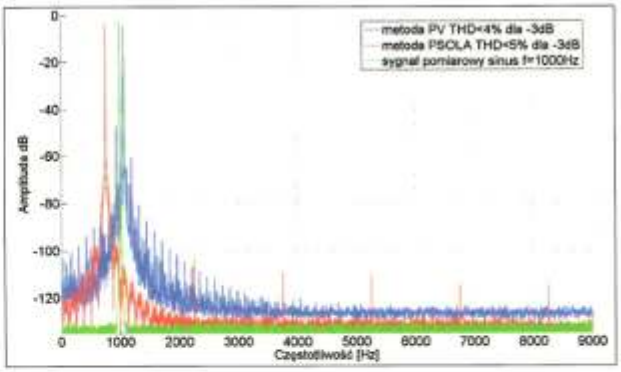
Elektroniczne metody maskowania są możliwe dzięki wykorzystaniu specjalnego oprogramowania, które za pomocą algorytmów z dziedziny cyfrowego przetwarzania sygnałów może dowolnie modyfikować brzmienie głosu. Wiele tych metod jest złożonych obliczeniowo, co ogranicza ich wykorzystanie w czasie rzeczywistym, inne natomiast często pogarszają zrozumiałość mowy do nieakceptowanego stopnia.

W dotychczasowej literaturze problem elektronicznego maskowania nie doczekał się zbyt wielu opracowań, choć staje się coraz istotniejszy. Wiadomo, że elektroniczna modyfikacja tonu krtaniowego może być skutecznym narzędziem maskowania głosu w stosunku do słuchaczy naiwnych już przy przesunięciu częstotliwości podstawowej o 4 półtony w dół lub górę na osi częstotliwości [10]. Słuchacze niebędący ekspertami w dziedzinie podczas identyfikacji fonoskopijnej najczęściej posługują się jedynie niektórymi parametrami zdradzającymi cechy osobnicze, takimi jak barwa, intonacja głosu czy szybkość artykulacji [11]. Kwestia wpływu tego sposobu modyfikacji tonu krtaniowego na cechy językowe oraz spektralne cechy osobnicze, a tym samym skuteczność systemu automatycznego rozpoznawania mówców, pozostaje otwarta. Ton krtaniowy jest parametrem, który za pomocą algorytmów DSP łatwo poddaje się kontroli, dlatego jest najczęściej wykorzystywany do modyfikacji przez dostępne na rynku narzędzia [12].

W niniejszej pracy przeanalizowano wpływ metod maskowania głosu z wykorzystaniem modyfikacji tonu krtaniowego na niektóre cechy językowe oraz na skuteczność automatycznego systemu kryminalistycznej identyfikacji mówców wyrażoną jako rzetelność testu ilorazu wiarygodności.

Ton krtaniowy i metody jego modyfikacji

Głównymi narządami aparatu głosowego człowieka są: płuca, tchawica, krtań, gardło, jama nosowa oraz usta. Powietrze, wydobywając się z płuc, wprawia fałdy głosowe krtani w szybki ruch, co powoduje wytwarzanie dźwięku zwanego tonem krtaniowym, zawierającym harmoniczne z dużego zakresu częstotliwości. Jego widmo jest modyfikowane następnie przez kanał głosowy, który pełni rolę układu filtrów [13]. Częstotliwość tonu krtaniowego to cecha, która w łatwy sposób może zostać poddana analizie z wykorzystaniem metod DSP. Programy komputerowe dostępne na rynku wykorzystują zwykle jedną z dwóch metod analizy, przetwarzania i syntezy: synchronizacji tonu metodą nakładania z dodawaniem (PSOLA) oraz wokodera fazowego (PV).



Ryc. 1. Zniekształcenia nieliniowe wprowadzane przez wykorzystanie metod modyfikacji PSOLA oraz PV, parametry FFT: okno Hanninga, N = 16384
Fig. 1. *Non-linear signal distortion by pitch shifting with PSOLA and PV methods, analysis parameters: 16384 points FFT, Hanning window*
Źródło (ryc. 1–3): autorzy

Metoda PSOLA

Zgodnie z metodą PSOLA (*pitch synchronized overlap-add method*) sygnał ulega dekompozycji na serię elementarnych składowych reprezentujących dźwięczne elementy mowy. Istnieje kilka odmian tego algorytmu. Najczęściej wykorzystywanym jest TD-PSOLA (*time domain*). Przetwarza on sygnał w dwóch etapach: analizy oraz syntezy [14]. W trakcie analizy naturalny sygnał mowy jest poddany ramkowaniu z krokiem równym okresowi segmentu lokalnie periodycznego, zgodnie z zależnością (1)

$$x_m(n) = h_m(t_m - n) \cdot x(n) \qquad (1),$$

gdzie $x(n)$ jest dyskretnym sygnałem mowy, h_m oknem Hanninga lokowanym w chwili t_m. W przypadku gdy sygnał nie wykazuje periodyczności, ramka ma stałą długość. Warunkiem poprawności działania tego algorytmu jest właściwe wyznaczenie składowej periodycznej. Zwykle do tego celu wykorzystywana jest krótkoterminowa funkcja autokorelacji. Na etapie syntezy ramki poddane są

superpozycji z dowolnie wybraną zakładką. Ramki mogą być również pomijane lub poddane repetycji. W efekcie okres periodycznego sygnału tonu podstawowego może się wydłużyć lub ulec skróceniu [15]. W algorytmie tym punkty łączenia ramek w trakcie syntezy są zmienne – wyznaczane są dla wartości maksimum funkcji autokorelacji dwóch sąsiednich ramek. Takie podejście zapewnia ciągłość fazy sygnału. Zniekształcenia pojawiające się w wyniku działania algorytmu TD-PSOLA mogą być wynikiem błędnego wyznaczenia okresu sygnału periodycznego, na przykład na skutek występowania silnych składowych harmonicznych. Drugą przyczyną zniekształceń może być niedokładna estymacja częstotliwości podstawowej, co z kolei może wynikać z ograniczonej częstotliwości próbkowania lub szumu zakłócającego. Efekt błędów pracy algorytmu TD-PSOLA w postaci zniekształceń nieliniowych przedstawiono na rycinie 1.

Metoda wokodera fazowego

Rozdzielczość FFT jest ograniczona i zdefiniowana stosunkiem częstotliwości próbkowania do rozmiaru okna analizy. Rozdzielczość ta jest zwykle niewystarczająca do dokładnej estymacji składowych o niskich częstotliwościach. Jednym ze sposobów poprawy rozdzielczości częstotliwościowej bez zmiany rozmiaru okna jest wykorzystanie wokodera fazowego (*phase vocoder* – PV), który zwiększa dokładność estymacji częstotliwości, wykorzystując informację o fazie widma. Algorytm wokodera fazowego w pierwszym kroku oblicza transformatę STFT sygnału. Ze względu na to, że transformata ta operuje na skończonej rozdzielczości, daje prawidłowe wyniki jedynie wówczas, gdy ciąg danych wejściowych zawiera energię rozłożoną dokładnie przy częstotliwościach, dla których dokonujemy analizy, tj. określonych w wyrażeniu (2)

$$f_{analizy} = \frac{m \cdot f_s}{n} \qquad (2),$$

gdzie $m, n \in N$ oraz f_s to częstotliwość próbkowania. Następnie w poszczególnych ramkach lokalizowane są składowe główne. Obserwując fazę tych składowych w dwóch różnych ramkach z wykorzystaniem wyrażenia (3), określa się częstotliwość f_n

$$f_n = \frac{(\theta_2 - \theta_1) + 2 \cdot \pi \cdot n}{2 \cdot \pi \cdot \Delta t} \qquad (3),$$

gdzie $\Delta t = t_2 - t_1$ oraz θ_1, θ_2 to fazy tej samej składowej w chwilach t_1 oraz t_2, $n \in N$. Wynikiem wyrażenia (3) jest szereg wartości. Częstotliwość z szeregu, która jest najbliżej składowej wyznaczonej z wykorzystaniem STFT, jest dokładną wartością składowej podstawowej. Wyznaczony w ten sposób ton krtaniowy może zostać zmodyfikowany przez przesunięcie o dowolną wartość na osi częstotliwości [16]. W trakcie procesu resyntezy mogą powstawać zniekształcenia o charakterze fazowym oraz nieliniowym. Stopień powstających zniekształceń jest uzależniony od metody wykorzystanej do resyntezy. Zniekształcenia powstałe w wyniku zastosowania wokodera fazowego przedstawiono na rycinie 1.

Cel badań

Identyfikacja kryminalistyczna mówcy wymaga ekstrakcji cech osobniczych przenoszonych z sygnałem mowy. Tymczasem metody PSOLA oraz PV w trakcie resyntezy sygnału generują zniekształcenia, które muszą wpływać na obserwowane cechy. W ramach pracy zbadano wpływ zniekształceń wprowadzanych przez algorytmy modyfikacji tonu krtaniowego na językowe cechy osobnicze oraz skuteczność automatycznego systemu kryminalistycznej identyfikacji mówców wyrażoną za pomocą charakterystyk Tippetta.

Analiza wypowiedzi zniekształconych za pomocą metod modyfikujących częstotliwość tonu krtaniowego może wymagać przywrócenia pierwotnej częstotliwości tonu. Warunkiem przeprowadzenia odwrotnego przekształcenia jest znajomość wartości częstotliwości, o którą dokonano modyfikacji. Jeżeli w trakcie rzeczywistej analizy kryminalistycznej ekspert zaobserwuje w tle sygnał o znanej częstotliwości, który uległ modyfikacji wraz z maskowanym sygnałem mowy, może na jego podstawie określić nieznany zakres częstotliwości, o jaki zmodyfikowano ton podstawowy. Odwrócenie maskowania na podstawie sygnału referencyjnego wymusza dwukrotne wykorzystanie modyfikatora: pierwotne przez sprawcę w celu dokonania maskowania oraz wtórne przez eksperta prowadzącego badania w celu jego odwrócenia. Przypadek ten został przeanalizowany w niniejszej pracy. W badaniach założono, że maskowania pierwotne oraz wtórne zostały przeprowadzone tymi samymi (PSOLA) oraz różnymi metodami (PSOLA oraz PV).

Metody kryminalistycznej identyfikacji mówców

Metoda językowa

Analiza językowa w badaniach identyfikacyjnych polega na wyekstrahowaniu z badanych wypowiedzi (dowodowych i porównawczych) dystynktywnych cech lingwistycznych. Założenia takiego podejścia uwzględniają badanie przejawów językowej aktywności człowieka. Efektem badań identyfikacyjnych powyższą metodą jest pokazanie i przedstawienie zbieżności bądź rozbieżności w opisanych zakresach. Pozwala to na stwierdzenie podobieństwa albo jego braku w warstwie językowej materiału badawczego. Indywidualne zespoły cech stwierdza się przez analizę zjawisk językowych, takich jak: sposób

artykulacji, morfologia, zasób słownikowy, składnia, prozodia, błędy językowe (fonetyczne – artykulacyjne, frazeologiczne, leksykalne, fleksyjne, składniowe) oraz wady wymowy, o ile jakość badanego materiału pozwala na ich identyfikację.

W trakcie badań analizie poddano wypowiedzi pięciu mówców (M1, M2, M3, M4, M5) i zbadano wpływ zastosowanych modyfikacji dźwięku na niektóre zaobserwowane cechy językowe. W niniejszej pracy ze względu na charakter i stopień modyfikacji nagrań w zakresie analizy językowej z oczywistych względów odniesiono się jedynie do warstwy fonetycznej wypowiedzi. Zastosowane modyfikacje nie wpływają bowiem na zakres słownictwa, łączliwość syntaktyczną czy akcentowanie. Badaniom poddano zjawiska językowe o bardzo różnej etymologii. Są wśród nich zarówno uproszczenia artykulacyjne, mogące mieć źródło w dialektyzmach, jak i wady wymowy czy błędy językowe. W artykule analizie poddano jedynie dziewięć wybranych cech związanych z artykulacją. Są to: rotacyzm, sygmatyzm, uproszczenia grup spółgłoskowych *strz, zdrz, trz, drz*, uproszczenia grup spółgłoskowych powstałych w wyniku usunięcia składnika miękkiego, denazalizacja: synchroniczna realizacja samogłoski *ą* w wygłosie oraz asynchroniczna wymowa samogłoski nosowej *ą* w wygłosie, synchroniczna wymowa samogłoski nosowej *ę*, miękka (mazowiecka) wymowa spółgłosek tylnojęzykowych, fonetyczne błędy językowe. Poniżej podano krótką charakterystykę ww. zjawisk.

Rotacyzm to wadliwy sposób wymowy głoski *r*. Jest najczęściej występującym i najbardziej różnorodnym pod względem formy zniekształceniem. Prawidłowe polskie *r* jest dźwiękiem wibracyjnym. Narządem artykułującym jest przód języka, który zbliżając się do wałka dziąsłowego aż do wytworzenia małej szczeliny lub krótkotrwałego zwarcia, wykonuje kilka drgań. Bardzo często określa się głoskę *r* jako fizjologicznie trudną. Po stronie artykulacyjnej o trudnościach tych stanowią: wibracja, złożony układ języka oraz krótki czas trwania zmieniających się artykulacji. W literaturze medycznej oraz w pracach logopedycznych wyodrębnia się 3 grupy form wadliwej wymowy głoski *r*: opuszczenie głoski *r*, zastępowanie *r* inną głoską danego języka (np. *l, j*), produkowanie głoski niewystępującej w systemie fonetycznym danego języka (rotacyzm właściwy) [17]. Stwierdzenie występowania konkretnej realizacji danej głoski w wypowiedzi mówcy możliwe jest po wykonaniu na przykład rentgenogramu ilustrującego miejsce artykulacji i wystąpienie zwarcia narządów mowy. Eksperci fonoskopii dysponują w swojej pracy jedynie nagraniem, bez obserwacji mówcy, nie mówiąc już o możliwości dokładniejszego sprawdzenia stopnia zwarcia czy miejsca artykulacji. Dlatego też w wielu przypadkach jedynym wnioskiem z analizy mowy osoby z rotacyzmem jest stwierdzenie jego obecności, sporadycznie wskazanie istnienia wibracji czy zastąpienia przez inną głoskę. Bardzo często jednak zniekształcenia wynikające z charakterystyki dostarczanych nagrań pozwalają jedynie na zaobserwowanie wadliwej, niestandardowej realizacji głoski. W trakcie badań zostały poddane analizie wypowiedzi dwóch mówców realizujących wadliwie głoskę *r*. W wypowiedzi M2 słyszalna jest wibracja w trakcie artykulacji głoski, jest ona jednak krótsza, przypominająca *r* jednouderzeniowe; w nagraniu mówcy M5 ma miejsce realizacja wokaliczna, polegająca na zastępowaniu głoski dźwiękiem zbliżonym do *y*.

Seplenienie (sygmatyzm) to nieprawidłowa artykulacja głosek dentalizowanych. Stwierdzenie występowania sygmatyzmu i określenie jego charakteru, tak jak w przypadku rotacyzmu, możliwe jest jedynie po badaniu logopedycznym. Na podstawie nagrania, bez możliwości obserwacji sposobu artykulacji głosek, nie można wypowiedzieć się na temat obecności konkretnej dysfunkcji, można mówić jedynie o innej realizacji fonemu. Nie do końca też można stwierdzić, czy odbierane audytywnie przez eksperta fonoskopii seplenienie nie jest zniekształceniem wprowadzanym przez aparaturę rejestrującą. Analizowany przykład zawiera wypowiedzi osoby, z którą autorzy mieli okazję się zetknąć i dzięki temu wiedzą o występowaniu u mówcy sygmatyzmu lateralnego w obrębie głosek szeregu szumiącego.

Kolejną cechą językową analizowaną w niniejszych badaniach jest uproszczenie grupy spółgłoskowej *strz, zdrz, trz, drz*. Stanowi ono wynik upodobnienia fonetycznego (asymilacji) polegającego na dostosowaniu artykulacji danej głoski do wymowy głosek sąsiednich. Grupy spółgłoskowe *trz, drz, strz, zdrz* na dużym obszarze Polski, powszechnie w Małopolsce i Wielkopolsce, są wymawiane jako *cz, dż, szcz, żdż*, na przykład *czeba, dżewo, szczelać, żdżemnąć sie* = trzeba, drzewo, strzelać, zdrzemnąć się.

Inną cechą poddaną analizom jest uproszczenie grup spółgłoskowych, powstałych w wyniku usunięcia składnika miękkiego. Taka sytuacja, dotycząca zredukowania grupy spółgłoskowej powstałej z rozłożenia spółgłoski *m'* przez opuszczenie miękkiego jej składnika zaszła w końcówce N. lmn. *-amy < -amni // -amji < -ami* (gdzie uległ morfologizacji) oraz w gwarowych odpowiednikach form zaimkowych *mi, mię*, por. *tykamy, wołamy, żelaznymy pługamy, sochamy, za toramy, nawalili my* = nawalili mi, *oczamy, osłamy, cepamy, saniamy, ziarkamy, szrubamy, tamy deskamy, widłamy, uszamy* [18].

Denazalizacja (odnosowienie) jest to zjawisko przejawiające się wymową samogłosek nosowych jako odpowiadających im samogłosek ustnych, czyli *ę* jako *e*, *ą* jako *o*. Polega zatem na utracie przez samogłoski nosowe rezonansu nosowego. Denazalizacja jest kolejnym rodzajem uproszczenia fonetycznego wynikającego ze skomplikowanej artykulacji, tym razem samogłosek nosowych. Składają się na nią bowiem połączone ruchy warg, języka i podniebienia miękkiego, które opuszczając się, umożliwia przejście powietrza przez jamę nosową. Jednoczesne

wykonanie tych wszystkich ruchów stanowi pewną trudność i powoduje dążność do uproszczenia ich wymowy, które obserwować można w historii wszystkich języków słowiańskich. Uproszczenia te szły w dwóch kierunkach: zaniku samogłosek nosowych lub realizacji dyftongicznej (dwufonemowej) [19].

W niniejszym artykule skupiono się na realizacjach samogłosek nosowych w pozycji wygłosowej. W wygłosie nosowość *ę* zanika na całym terenie Polski: *robie, siedze*, rzadko natomiast zanika nosowość samogłoski tylnej *-ą*, na przykład *s koso, chodzo, robio* = z kosą, chodzą, robią. Wymowa taka występuje głównie na pograniczu wschodnim [20].

Innym przykładem denazalizacji samogłosek jest asynchroniczna (rozłożona) artykulacja samogłoski tylnej *-ą* w wygłosie, na przykład *robiom, siedzom, chodzom, tom drogom* = robią, siedzą, chodzą, tą drogą.

Analizowany materiał zawiera również błędne realizacje wyrazowe. Różnego rodzaju błędy językowe pojawiające się w badanym materiale i charakteryzujące się pewną powtarzalnością często stanowią o dystynktywności mówców. W niniejszym artykule analizie poddana została wymowa wyrazów: *dzisiaj, tutaj*, jako *dzisiej, tutej*. Wymowa *dzisiej* nie znajduje uznania w słownikach poprawnej polszczyzny, choć jest artykulacyjnie uzasadniona (między wysokimi głoskami [ś] i [j] samogłoska [a] ulega podwyższeniu, co może ją upodobnić do [e]). W potocznej wymowie słychać po prostu *dzisiej*, co nie jest naganne. Inny czynnik, który może skłaniać do błędnej wymowy, to zakończenie przymiotnika *dzisiejszy*. Wymowa *tutej* jest również potoczna. Forma taka powstała prawdopodobnie przez analogię do słowa *dzisiaj*. Upodobnieniu wymowy *tutej* do *dzisiej* sprzyja dodatkowo identyczne zakończenie obu wyrazów [21].

Metoda automatyczna

Do identyfikacji mówców wykorzystywane są również niektóre widmowe parametry sygnału mowy. Ich wartość oraz zakres zmienności stanowią informację osobniczą. W klasycznej analizie pomiarowej najczęściej wykorzystywany jest ton krtaniowy oraz częstotliwości formantowe samogłosek. Formanty samogłoskowe pod względem ich częstotliwości, amplitudy, dobroci są ściśle zależne od budowy fizjologicznej narządu mowy. Specyfiką klasycznej metody pomiarowej opartej na analizie formantów jest badanie jedynie wybranych segmentów mowy – niektórych samogłosek. Automatyczne rozpoznawanie mówców (ARM) opiera się na podobnym założeniu jak klasyczna metoda pomiarowa, tzn. niektóre fizyczne parametry widmowe opisujące sygnał mowy niosą informację osobniczą. W odróżnieniu od klasycznej metody pomiarowej, w przypadku ARM ekstrakcja cech dokonywana jest z całej wypowiedzi. Wykorzystany w niniejszym opracowaniu system automatycznego rozpoznawania mówców od strony matematycznej oraz technicznej przedstawiono w pracy *Biometryczne rozpoznawanie mówców w kryminalistyce* [22]. Ocena zgodności cech porównywanych głosów może zostać wyrażona za pomocą ilorazu wiarygodności (*likelihood ratio*). Metody szacowania tej wartości w odniesieniu do kryminalistycznej identyfikacji mówców szeroko opisano w literaturze. W niniejszej pracy posłużono się metodą zaprezentowaną w publikacji *Forensic speaker recognition based on a Bayesian framework and Gaussian mixture modeling* [23]. Do oszacowania zmienności wewnątrzosobniczej i międzyosobniczej wykorzystano populację czterdziestu czterech mówców polskojęzycznych.

Do oceny rzetelności wyniku w postaci ilorazu wiarygodności wykorzystano charakterystyki Tippetta [23]. Na podstawie charakterystyk określono dwa parametry:

1) prawdopodobieństwo otrzymania wartości LR > 1 oznaczane jako $P_{LRHd>1}$, gdy autorem wypowiedzi porównawczej jest przypadkowa osoba (prawdopodobieństwo uznania nieprawdziwego dowodu na korzyść oskarżenia),
2) prawdopodobieństwo otrzymania wartości LR < 1, oznaczane jako $P_{LRHp<1}$, gdy autorem wypowiedzi porównawczej jest autor wypowiedzi dowodowej (prawdopodobieństwo uznania nieprawdziwego dowodu na korzyść obrony).

Wielkości $P_{LRHp<1}$ oraz $P_{LRHd>1}$ opisują prawdopodobieństwo wystąpienia błędnej wartości LR. Charakterystyki, na podstawie których wyznaczono te parametry, opisują reprezentatywność testu LR dla poszczególnych wariantów modyfikowanego sygnału i wraz z charakterystykami Tippetta stanowią opis skuteczności metody badawczej w tych wariantach.

Wyniki badań

Badania przeprowadzono z wykorzystaniem bazy mówców polskojęzycznych, których swobodne wypowiedzi zostały zarejestrowane za pośrednictwem telefonu GSM. Wypowiedzi zmodyfikowano przez zastosowanie trzech algorytmów. Dwa z nich wykorzystują metodę PSOLA; będą one oznaczone w tekście jako PSOLA(A) oraz PSOLA(B). Trzeci algorytm wykorzystuje metodę wokodera fazowego – PV. Wyniki badań przedstawiono dla systemu automatycznego rozpoznawania mówców w postaci charakterystyk Tippetta, natomiast dla metody językowej w postaci tabel zawierających opis obserwowanych cech językowych pięciu wybranych mówców. Poszczególne warianty analizowanego materiału testowego opisano, uwzględniając dwa parametry: zakres względnego przesunięcia częstotliwości podstawowej wyrażonej w półtonach oraz dla analizy językowej średnie wartości tonów krtaniowych przed modyfikacją oraz po modyfikacji (oznaczenie T_{μ}) zmierzonych dla każdego z mówców.

Wpływ maskowania z wykorzystaniem metody PSOLA

W celu określenia wpływu modyfikowania głosu z wykorzystaniem metody nakładania z dodawaniem na cechy językowe oraz rzetelność testu LR przeanalizowano następujące warianty:

- **wariant 1** – „mowa naturalna" – wypowiedzi oryginalne bez modyfikacji;
- **wariant 2** – „maskowanie" – wypowiedzi oryginalne, zmodyfikowane za pomocą algorytmu PSOLA(A) przez obniżenie tonu krtaniowego o 5 półtonów;
- **wariant 3** – „po korekcji" – wypowiedzi zmodyfikowane jak w wariancie 2 ponownie zmodyfikowano za pomocą tego samego algorytmu, ale innej implementacji (PSOLA[B]) w celu odwrócenia maskowania, przez podniesienie tonu krtaniowego o 5 półtonów w górę; w efekcie uzyskano ton krtaniowy, zbliżony do oryginalnego;
- **wariant 4** – „maskowanie" – wypowiedzi oryginalne, zmodyfikowane za pomocą algorytmu PSOLA(A) przez podniesienie tonu krtaniowego o 5 półtonów;
- **wariant 5** – „po korekcji" – wypowiedzi zmodyfikowane jak w wariancie 4 ponownie zmodyfikowano za pomocą tego samego algorytmu, ale innej implementacji (PSOLA[B]) w celu odwrócenia maskowania przez obniżenie tonu krtaniowego o 5 półtonów w dół; w efekcie uzyskano ton krtaniowy, zbliżony do oryginalnego.

A. Metoda językowa

Wyniki analiz językowych dla 5 mówców przedstawiono w tabelach poniżej. Bardziej szczegółowe opisy dotyczące każdego z analizowanych przypadków znajdują się pod odpowiednimi tabelami.

Tabela 2

Analiza językowa wypowiedzi mówcy 1, które zmodyfikowano metodą PSOLA
Language analysis of speaker 1 utterance, which has been modified with PSOLA method

Obserwowane realizacje	Wariant 1 (mowa naturalna)	Wariant 2 (maskowanie)	Wariant 3 (po korekcji)	Wariant 4 (maskowanie)	Wariant 5 (po korekcji)
Mówca 1 (M1) $T_{\mu,\ oryginalne}$ = 135 Hz		$T_{\mu,\ oryginalne}$ = 135 Hz $T_{\mu,\ końcowe}$ = 107 Hz	$T_{\mu,\ początkowe}$ = 107 Hz $T_{\mu,\ końcowe}$ = 130 Hz	$T_{\mu,\ oryginalne}$ = 135 Hz $T_{\mu,\ końcowe}$ = 175 Hz	$T_{\mu,\ początkowe}$ = 175 Hz $T_{\mu,\ końcowe}$ = 132 Hz
„praktycznie", „szkołę", „oborze"	Sygmatyzm lateralny szeregu szumiącego *sz, rz, cz*	Seplenienie boczne niezauważalne	Sygmatyzm lateralny szeregu szumiącego *sz, rz, cz*	Seplenienie boczne niezauważalne	Sygmatyzm lateralny szeregu szumiącego *sz, rz, cz*
„agroturystyka" „kwaterowcy"	Wymowa zgodna z normą	Wymowa zgodna z normą	Wymowa zgodna z normą	W związku z pojawieniem się zniekształcenia o charakterze drgania zauważalna jest niestandardowa realizacja *r*	W związku z pojawieniem się zniekształcenia o charakterze drgania zauważalna jest niestandardowa realizacja *r*

Zarówno seplenienie, jak i rotacyzm w wyżej wymienionym nagraniu pojawiają się tylko w określonych wariantach modyfikacji oryginalnego nagrania. Niestandardowa realizacja szeregu szumiącego zanika pod wpływem modyfikacji (wariant 2 i 4) i pojawia się w nagraniach będących odwróceniem maskowania (wariant 3 i 5). Realizacja głoski *r*, która pierwotnie jest zgodna z normą, ulega zmianie w sytuacji opisanej w wariancie 4 i 5.

Tabela 3

Analiza językowa wypowiedzi mówcy 2, które zmodyfikowano metodą PSOLA
Language analysis of speaker 2 utterance, which has been modified with PSOLA method

Obserwowane realizacje	Wariant 1 (mowa naturalna)	Wariant 2 (maskowanie)	Wariant 3 (po korekcji)	Wariant 4 (maskowanie)	Wariant 5 (po korekcji)
Mówca 2 (M2) $T_{\mu,\ oryginalne}$ = 231 Hz		$T_{\mu,\ oryginalne}$ = 231 Hz $T_{\mu,\ końcowe}$ = 179 Hz	$T_{\mu,\ początkowe}$ = 179 Hz $T_{\mu,\ końcowe}$ = 224 Hz	$T_{\mu,\ oryginalne}$ = 231 Hz $T_{\mu,\ końcowe}$ = 299 Hz	$T_{\mu,\ początkowe}$ = 299 Hz $T_{\mu,\ końcowe}$ = 226 Hz
„kwaterowcy" „porykiwania" „Kraków"	Rotacyzm właściwy ze słyszalnym zwarciem	Rotacyzm obecny, ale mniej jaskrawy	Rotacyzm właściwy ze słyszalnym zwarciem	W związku z pojawieniem się zniekształcenia o charakterze drgania, rotacyzm bardziej jaskrawy	W związku z pojawieniem się zniekształcenia o charakterze drgania, rotacyzm bardziej jaskrawy

Rotacyzm, zaobserwowany w nagraniu oryginalnym, utrzymuje się mimo maskowania. Zmienia się jedynie stopień nasilenia zwarcia, w zależności od wariantu maskowania.

Tabela 4

Analiza językowa wypowiedzi mówcy 3, które zmodyfikowano metodą PSOLA

Language analysis of speaker 3 utterance, which has been modified with PSOLA method

Obserwowane realizacje	Wariant 1 (mowa naturalna)	Wariant 2 (maskowanie)	Wariant 3 (po korekcji)	Wariant 4 (maskowanie)	Wariant 5 (po korekcji)
Mówca 3 (M3) $T_{\mu,\ oryginalne} = 217$ Hz		$T_{\mu,\ oryginalne} = 217$ Hz $T_{\mu,\ końcowe} = 168$ Hz	$T_{\mu,\ początkowe} = 168$ Hz $T_{\mu,\ końcowe} = 210$ Hz	$T_{\mu,\ oryginalne} = 217$ Hz $T_{\mu,\ końcowe} = 289$ Hz	$T_{\mu,\ początkowe} = 289$ Hz $T_{\mu,\ końcowe} = 212$ Hz
„poczeba", „dżewa" „potczymywać"/ „podczymywać" „szczyże", „czeba"	Uproszczenie grupy spółgłoskowej	Uproszczenie grupy spółgłoskowej	Uproszczenie grupy spółgłoskowej	Uproszczenie grupy spółgłoskowej	Uproszczenie grupy spółgłoskowej
„piętnasta", „mówię" „kontroluję"	Synchroniczna wymowa samogłoski nosowej *ę*	Synchroniczna wymowa samogłoski nosowej *ę*	Synchroniczna wymowa samogłoski nosowej *ę*	Synchroniczna wymowa samogłoski nosowej *ę*	Synchroniczna wymowa samogłoski nosowej *ę*
„wesołom", „powodujom" „lubiom" „uderzeniowom"	Asynchroniczna wymowa samogłoski nosowej *ą* w wygłosie	Asynchroniczna wymowa samogłoski nosowej *ą* w wygłosie	Asynchroniczna wymowa samogłoski nosowej *ą* w wygłosie	Asynchroniczna wymowa samogłoski nosowej *ą* w wygłosie	Asynchroniczna wymowa samogłoski nosowej *ą* w wygłosie
„atmosfera", „zrobiłam" „kreseczki"	Wymowa zgodna z normą	Wymowa zgodna z normą	Wymowa zgodna z normą	W związku z pojawieniem się zniekształcenia o charakterze drgania zauważalna jest niestandardowa realizacja *r*	Wymowa zgodna z normą

Zarówno zastosowane maskowania, jak i jego odwrócenie nie spowodowało zmian w zjawiskach językowych dotyczących uproszczeń spółgłoskowych i sposobu realizacji samogłosek nosowych. W wypowiedzi maskowanej programem PSOLA(A) (wariant 4) pojawia się natomiast dodatkowe drganie głoski *r*, które zanika podczas powrotu do wersji oryginalnej.

Tabela 5

Analiza językowa wypowiedzi mówcy 4, które zmodyfikowano metodą PSOLA

Language analysis of speaker 4 utterance, which has been modified with PSOLA method

Obserwowane realizacje	Wariant 1 (mowa naturalna)	Wariant 2 (maskowanie)	Wariant 3 (po korekcji)	Wariant 4 (maskowanie)	Wariant 5 (po korekcji)
Mówca 4 (M4) $T_{\mu,\ oryginalne} = 166$ Hz		$T_{\mu,\ oryginalne} = 166$ Hz $T_{\mu,\ końcowe} = 141$ Hz	$T_{\mu,\ początkowe} = 141$ Hz $T_{\mu,\ końcowe} = 161$ Hz	$T_{\mu,\ oryginalne} = 166$ Hz $T_{\mu,\ końcowe} = 208$ Hz	$T_{\mu,\ początkowe} = 208$ Hz $T_{\mu,\ końcowe} = 164$ Hz
„telefonamy" „z kropkamy" „z wamy" „takimy ludźmy"	Redukcja grupy spółgłoskowej w wyniku usunięcia składnika miękkiego	Redukcja grupy spółgłoskowej w wyniku usunięcia składnika miękkiego	Redukcja grupy spółgłoskowej w wyniku usunięcia składnika miękkiego	Redukcja grupy spółgłoskowej w wyniku usunięcia składnika miękkiego	Redukcja grupy spółgłoskowej w wyniku usunięcia składnika miękkiego
„tutej" „dzisiej"	Potoczna, nieznormalizowana wymowa	Potoczna, nieznormalizowana wymowa	Potoczna, nieznormalizowana wymowa	Słabsza artykulacja samogłoski *e* – na pograniczu z *a*	Potoczna, nieznormalizowana wymowa
„chiba" „okiej"	Miękka (mazowiecka) wymowa spółgłosek tylnojęzykowych	Słabiej słyszalne zmiękczenie *k*; wymowa zmiękczonej głoski *ch* bez zmian	Miękka (mazowiecka) wymowa spółgłosek tylnojęzykowych	Bardziej miękka wymowa spółgłoski *k*; wymowa zmiękczonej głoski *ch* bez zmian	Miękka (mazowiecka) wymowa spółgłosek tylnojęzykowych

Tabela 5 cd.

Analiza językowa wypowiedzi mówcy 4, które zmodyfikowano metodą PSOLA

Language analysis of speaker 4 utterance, which has been modified with PSOLA method

Obserwowane realizacje	Wariant 1 (mowa naturalna)	Wariant 2 (maskowanie)	Wariant 3 (po korekcji)	Wariant 4 (maskowanie)	Wariant 5 (po korekcji)
Mówca 4 (M4) $T_{\mu,\ oryginalne}$ = 166 Hz		$T_{\mu,\ oryginalne}$ = 166 Hz $T_{\mu,\ końcowe}$ = 141 Hz	$T_{\mu,\ początkowe}$ = 141 Hz $T_{\mu,\ końcowe}$ = 161 Hz	$T_{\mu,\ oryginalne}$ = 166 Hz $T_{\mu,\ końcowe}$ = 208 Hz	$T_{\mu,\ początkowe}$ = 208 Hz $T_{\mu,\ końcowe}$ = 164 Hz
„znajo", „so" „wchodzo" „tako", „końcówko"	Denazalizacja	Denazalizacja	Denazalizacja	Denazalizacja	Denazalizacja
„różnych" „numerem' „robione"	Wymowa zgodna z normą	Wymowa zgodna z normą	Wymowa zgodna z normą	W związku z pojawieniem się zniekształcenia o charakterze drgania zauważalna jest niestandardowa realizacja *r*	W związku z pojawieniem się zniekształcenia o charakterze drgania zauważalna jest niestandardowa realizacja *r*

Brzmienie wyrazów, w których doszło do redukcji grupy spółgłoskowej w wyniku usunięcia składnika miękkiego, nie zmieniło się pod wpływem przeprowadzonych modyfikacji. Błędy fonetyczne dotyczące wymowy *dzisiej, tutej* ulegają zatarciu w przypadku maskowania programem PSOLA(A) w wariancie 4. Po odwróceniu maskowania błędne brzmienie staje się ponownie słyszalne. Miękka realizacja spółgłoski tylnojęzykowej *ch* pozostaje bez zmian we wszystkich wariantach modyfikacji nagrania. Niestabilnie natomiast zachowuje się głoska *k'*, której zmiękczenie słabnie pod wpływem modyfikacji programem PSOLA(A) w wariancie 2, a wyostrza się po zastosowaniu maskowania programem PSOLA(A) w wariancie 4. Denazalizacja głoski *ą* w wygłosie utrzymuje się w każdym wariancie. W nagraniu maskowanym programem PSOLA(A) w wariancie 4 oraz w wersji po korekcji pojawia się charakterystyczne drżenie głoski *r*, mogące sprawiać wrażenie rotacyzmu u mówcy.

Tabela 6

Analiza językowa wypowiedzi mówcy 5, które zmodyfikowano metodą PSOLA

Language analysis of speaker 5 utterance, which has been modified with PSOLA method

Obserwowane realizacje	Wariant 1 (mowa naturalna)	Wariant 2 (maskowanie)	Wariant 3 (po korekcji)	Wariant 4 (maskowanie)	Wariant 5 (po korekcji)
Mówca 5 (M5) $T_{\mu,\ oryginalne}$ = 135 Hz		$T_{\mu,\ oryginalne}$ = 135 Hz $T_{\mu,\ końcowe}$ = 110 Hz	$T_{\mu,\ początkowe}$ = 110 Hz $T_{\mu,\ końcowe}$ = 131 Hz	$T_{\mu,\ oryginalne}$ = 135 Hz $T_{\mu,\ końcowe}$ = 174 Hz	$T_{\mu,\ początkowe}$ = 174 Hz $T_{\mu,\ końcowe}$ = 133 Hz
„problem" „razy" „rower" „pracy" „góry" „szwagier"	Rotacyzm, przejawiający się zastępowaniem głoski *r* dźwiękiem zbliżonym do *y*, bez słyszalnego zwarcia	Rotacyzm, przejawiający się zastępowaniem głoski *r* dźwiękiem zbliżonym do *y*, bez słyszalnego zwarcia	Rotacyzm, przejawiający się zastępowaniem głoski *r* dźwiękiem zbliżonym do *y*, bez słyszalnego zwarcia	Rotacyzm, przejawiający się zastępowaniem głoski *r* dźwiękiem zbliżonym do *y*, bez słyszalnego zwarcia	Rotacyzm, przejawiający się zastępowaniem głoski *r* dźwiękiem zbliżonym do *y*, bez słyszalnego zwarcia

Rotacyzm utrzymuje się w każdej wersji nagrania.

B. Metoda automatyczna

Rycina 2 przedstawia wpływ modyfikacji metodą PSOLA na rzetelność testu LR.

Para dystrybuant niebieskich (ciągła oraz przerywana) przedstawia rzetelność testu LR dla systemu automatycznego rozpoznawania mówców wówczas, gdy analizowane wypowiedzi nie są maskowane (wariant 1). W tym przypadku prawdopodobieństwo uznania błędnego dowodu jest niewielkie i równe $P_{LRHd>1} = 0,07$ oraz $P_{LRHp<1} = 0,04$. Maskowanie (warianty 2 oraz 4, dystrybuanty czerwone oraz czarne) uniemożliwia skuteczne prowadzenie identyfikacji z wykorzystaniem metod automatycznych. Otrzymane LR w każdym analizowanym przypadku osiąga wartość większą od 1. Odwrócona modyfikacja (warianty 3 oraz 5, dystrybuanty różowe oraz zielone) ma duży wpływ na rzetelność testu LR. Prawdopodobieństwa uznania błędnego dowodu wzrosły w obydwu wariantach w stosunku do mowy niemaskowanej o tę samą wartość i wynoszą $P_{LRHd>1} = 0,33$ oraz $P_{LRHp<1} = 0,09$.

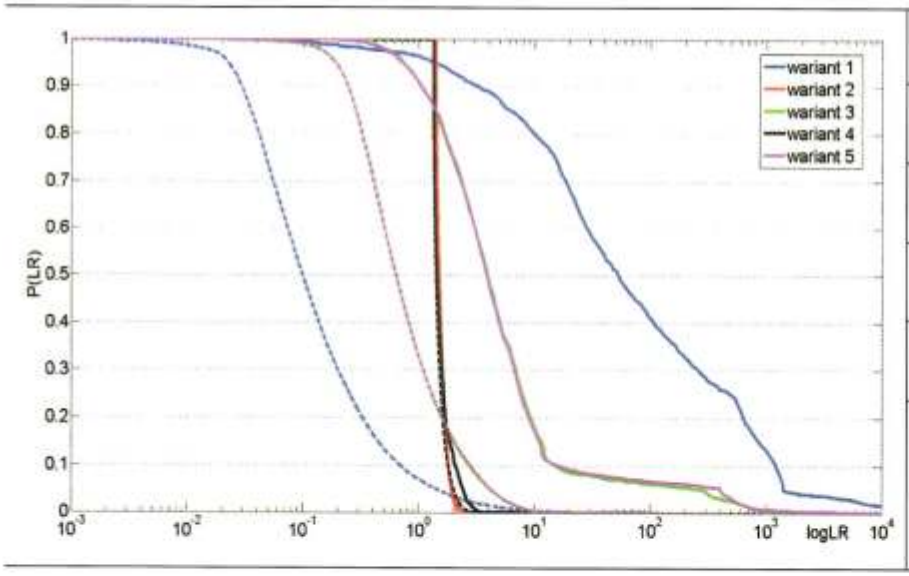
Wpływ maskowania z wykorzystaniem metod PSOLA oraz PV

W celu określenia wpływu modyfikowania głosu z wykorzystaniem metody nakładania z dodawaniem oraz wokodera fazowego na cechy językowe oraz rzetelność testu LR, przeanalizowano następujące warianty:

- **wariant 1** – „mowa naturalna" – wypowiedzi oryginalne bez modyfikacji;
- **wariant 6** – „maskowanie" – wypowiedzi oryginalne, zmodyfikowane za pomocą algorytmu PSOLA(A) przez obniżenie tonu krtaniowego o 5 półtonów;
- **wariant 7** – „po korekcji" – wypowiedzi zmodyfikowane jak w wariancie 6 ponownie zmodyfikowano za pomocą metody wokodera fazowego w celu odwrócenia maskowania przez podniesienie tonu krtaniowego o 5 półtonów w górę; w efekcie uzyskano ton krtaniowy, zbliżony do oryginalnego;
- **wariant 8** – „maskowanie" – wypowiedzi oryginalne zmodyfikowane za pomocą algorytmu PSOLA(A) przez podniesienie tonu krtaniowego o 5 półtonów;
- **wariant 9** – „po korekcji" – wypowiedzi zmodyfikowane jak w wariancie 8 ponownie zmodyfikowano za pomocą algorytmu wokodera fazowego w celu odwrócenia maskowania przez obniżenie tonu krtaniowego o 5 półtonów w dół; w efekcie uzyskano ton krtaniowy, zbliżony do oryginalnego.

A. Metoda językowa

W kolejnym etapie badań sprawdzono, jak na cechy językowe wpływa modyfikowanie tonu krtaniowego z wykorzystaniem wokodera fazowego oraz metody PSOLA.



Numer wariantu	$P_{LRHd>1}$	$P_{LRHp<1}$
Wariant 1 (mowa naturalna)	0,07	0,04
Wariant 2 (maskowanie)	1,00	0,00
Wariant 3 (po korekcji)	0,33	0,09
Wariant 4 (maskowanie)	1,00	0,00
Wariant 5 (po korekcji)	0,33	0,09

Ryc. 2. Wpływ maskowania głosu z wykorzystaniem metody PSOLA na rzetelność testu LR
Fig. 2 *Effect of speech disguised with PSOLA method on reliability of likelihood ratio test*

Tabela 7

Analiza językowa wypowiedzi mówcy 1, które zmodyfikowano metodą PSOLA oraz PV

Language analysis of speaker 1 utterance, which has been modified with PSOLA and PV methods

Obserwowane realizacje	Wariant 1 (mowa naturalna)	Wariant 6 (maskowanie)	Wariant 7 (po korekcji)	Wariant 8 (maskowanie)	Wariant 9 (po korekcji)
Mówca 1 (M1) $T_{\mu,\ oryginalne} = 135$ Hz		$T_{\mu,\ oryginalne} = 135$ Hz $T_{\mu,\ końcowe} = 106$ Hz	$T_{\mu,\ początkowe} = 106$ Hz $T_{\mu,\ końcowe} = 139$ Hz	$T_{\mu,\ oryginalne} = 135$ Hz $T_{\mu,\ końcowe} = 178$ Hz	$T_{\mu,\ początkowe} = 178$ Hz $T_{\mu,\ końcowe} = 141$ Hz
„praktycznie" *„szkołę"* *„oborze"*	Sygmatyzm lateralny szeregu szumiącego *sz, rz, cz*	Seplenienie boczne niezauważalne	Sygmatyzm lateralny szeregu szumiącego *sz, rz, cz*	Seplenienie boczne niezauważalne	Sygmatyzm lateralny szeregu szumiącego *sz, rz, cz*

Seplenienie pojawia się tylko w określonych wariantach modyfikacji oryginalnego nagrania. Niestandardowa realizacja szeregu szumiącego zanika pod wpływem modyfikacji (wariant 6 oraz 8) i pojawia się w nagraniach będących odwróceniem maskowania (wariant 7 oraz 9).

Tabela 8

Analiza językowa wypowiedzi mówcy 2, które zmodyfikowano metodą PSOLA oraz PV

Language analysis of speaker 2 utterance, which has been modified with PSOLA and PV methods

Obserwowane realizacje	Wariant 1 (mowa naturalna)	Wariant 6 (maskowanie)	Wariant 7 (po korekcji)	Wariant 8 (maskowanie)	Wariant 9 (po korekcji)
Mówca 2 (M2) $T_{\mu,\ oryginalne} = 231$ Hz		$T_{\mu,\ oryginalne} = 231$ Hz $T_{\mu,\ końcowe} = 175$ Hz	$T_{\mu,\ początkowe} = 175$ Hz $T_{\mu,\ końcowe} = 237$ Hz	$T_{\mu,\ oryginalne} = 231$ Hz $T_{\mu,\ końcowe} = 304$ Hz	$T_{\mu,\ początkowe} = 304$ Hz $T_{\mu,\ końcowe} = 242$ Hz
„kwaterowcy" *„porykiwania"* *„Kraków"*	Rotacyzm właściwy ze słyszalnym zwarciem	Rotacyzm obecny, ale mniej jaskrawy	Rotacyzm obecny, ale mniej jaskrawy	Rotacyzm obecny, ale mniej jaskrawy	W związku z pojawieniem się zniekształcenia o charakterze drgania rotacyzm bardziej jaskrawy

Rotacyzm zaobserwowany w nagraniu oryginalnym utrzymuje się mimo maskowania. Zmienia się jedynie stopień nasilenia zwarcia w zależności od wariantu maskowania.

Tabela 9

Analiza językowa wypowiedzi mówcy 3, które zmodyfikowano metodą PSOLA oraz PV

Language analysis of speaker 3 utterance, which has been modified with PSOLA and PV methods

Obserwowane realizacje	Wariant 1 (mowa naturalna)	Wariant 6 (maskowanie)	Wariant 7 (po korekcji)	Wariant 8 (maskowanie)	Wariant 9 (po korekcji)
Mówca 3 (M3) $T_{\mu,\ oryginalne} = 217$ Hz		$T_{\mu,\ oryginalne} = 217$ Hz $T_{\mu,\ końcowe} = 164$ Hz	$T_{\mu,\ początkowe} = 164$ Hz $T_{\mu,\ końcowe} = 223$ Hz	$T_{\mu,\ oryginalne} = 217$ Hz $T_{\mu,\ końcowe} = 287$ Hz	$T_{\mu,\ początkowe} = 287$ Hz $T_{\mu,\ końcowe} = 226$ Hz
„poczeba" *„dżewa"* *„potczymywać"/ „podczymywać"* *„szczyże", „czeba"*	Uproszczenie grupy spółgłoskowej	Uproszczenie grupy spółgłoskowej	Uproszczenie grupy spółgłoskowej	Uproszczenie grupy spółgłoskowej	Uproszczenie grupy spółgłoskowej

Tabela 9 cd.

Analiza językowa wypowiedzi mówcy 3, które zmodyfikowano metodą PSOLA oraz PV

Language analysis of speaker 3 utterance, which has been modified with PSOLA and PV methods

Obserwowane realizacje	Wariant 1 (mowa naturalna)	Wariant 6 (maskowanie)	Wariant 7 (po korekcji)	Wariant 8 (maskowanie)	Wariant 9 (po korekcji)
„piętnasta" „mówię" „kontroluję się"	Synchroniczna wymowa nosówek	Synchroniczna wymowa nosówek	Synchroniczna wymowa nosówek	Synchroniczna wymowa nosówek	Synchroniczna wymowa nosówek
„wesołom" „powodujom" „lubiom" „uderzeniowom"	Asynchroniczna wymowa nosówek w wygłosie	Asynchroniczna wymowa nosówek w wygłosie	Asynchroniczna wymowa nosówek w wygłosie	Asynchroniczna wymowa nosówek w wygłosie	Asynchroniczna wymowa nosówek w wygłosie

Zastosowane maskowania i korekcje nie spowodowały zmian w zjawiskach językowych dotyczących uproszczeń grup spółgłoskowych i sposobu realizacji samogłosek.

Tabela 10

Analiza językowa wypowiedzi mówcy 4, które zmodyfikowano metodą PSOLA oraz PV

Language analysis of speaker 4 utterance, which has been modified with PSOLA and PV methods

Obserwowane realizacje	Wariant 1 (mowa naturalna)	Wariant 6 (maskowanie)	Wariant 7 (po korekcji)	Wariant 8 (maskowanie)	Wariant 9 (po korekcji)
Mówca 4 (M4) $T_{\mu,\ oryginalne}$ = 166 Hz		$T_{\mu,\ oryginalne}$ = 166 Hz $T_{\mu,\ końcowe}$ = 140 Hz	$T_{\mu,\ początkowe}$ = 140 Hz $T_{\mu,\ końcowe}$ = 173 Hz	$T_{\mu,\ oryginalne}$ = 166 Hz $T_{\mu,\ końcowe}$ = 208 Hz	$T_{\mu,\ początkowe}$ = 208 Hz $T_{\mu,\ końcowe}$ = 174 Hz
„telefonamy" „z kropkamy" „z wamy" „takimy ludźmy"	Redukcja grupy spółgłoskowej w wyniku usunięcia składnika miękkiego	Redukcja grupy spółgłoskowej w wyniku usunięcia składnika miękkiego	Redukcja grupy spółgłoskowej w wyniku usunięcia składnika miękkiego	Redukcja grupy spółgłoskowej w wyniku usunięcia składnika miękkiego	Redukcja grupy spółgłoskowej w wyniku usunięcia składnika miękkiego
„tutej" „dzisiej"	Potoczna, nieznormalizowana wymowa	Potoczna, nieznormalizowana wymowa	Potoczna, nieznormalizowana wymowa	Słabsza artykulacja samogłoski *e* – na pograniczu z *a*	Potoczna, nieznormalizowana wymowa
„chiba" „okiej"	Miękka (mazowiecka) wymowa spółgłosek tylnojęzykowych	Słabiej słyszalne zmiękczenie *k*; wymowa zmiękczonej głoski *ch* bez zmian	Miękka (mazowiecka) wymowa spółgłosek tylnojęzykowych	Bardziej miękka wymowa spółgłoski *k*; wymowa zmiękczonej głoski *ch* bez zmian	Bardziej miękka wymowa spółgłoski *k*; wymowa zmiękczonej głoski *ch* bez zmian
„znajo", „so" „wchodzo" „tako" „końcówko"	Denazalizacja	Denazalizacja	Denazalizacja	Denazalizacja	Denazalizacja

Brzmienie wyrazów, w których doszło do redukcji grupy spółgłoskowej w wyniku usunięcia składnika miękkiego, nie zmieniło się pod wpływem zastosowanych programów. Błędy fonetyczne dotyczące wymowy *dzisiej*, *tutej* ulegają zatarciu w przypadku maskowania programem PSOLA(A) (wariant 8). W wariancie 9 błędne brzmienie staje się ponownie słyszalne. Miękka realizacja spółgłoski tylnojęzykowej *ch* pozostaje bez zmian we wszystkich wariantach modyfikowanego nagrania. Niestabilnie natomiast zachowuje się głoska *k'*, której zmiękczenie słabnie pod wpływem modyfikacji programem PSOLA(A) (wariant 6), a wyostrza się po zastosowaniu maskowania programem PSOLA(A) (wariant 8) i próbie powrotu do oryginału programem PV (wariant 9). Denazalizacja głoski *ą* w wygłosie utrzymuje się w każdym wariancie.

Tabela 11

Analiza językowa wypowiedzi mówcy 5, które zmodyfikowano metodą PSOLA oraz PV

Language analysis of speaker 5 utterance, which has been modified with PSOLA and PV methods

Obserwowane realizacje	Wariant 1 (mowa naturalna)	Wariant 6 (maskowanie)	Wariant 7 (po korekcji)	Wariant 8 (maskowanie)	Wariant 9 (po korekcji)
Mówca 5 (M5) $T_{\mu,\ oryginalne}$ = 135 Hz		$T_{\mu,\ oryginalne}$ = 135 Hz $T_{\mu,\ końcowe}$ = 109 Hz	$T_{\mu,\ początkowe}$ = 109 Hz $T_{\mu,\ końcowe}$ = 141 Hz	$T_{\mu,\ oryginalne}$ = 135 Hz $T_{\mu,\ końcowe}$ = 204 Hz	$T_{\mu,\ początkowe}$ = 204 Hz $T_{\mu,\ końcowe}$ = 144 Hz
„problem" *„razy"* *„rower"* *„pracy"* *„góry"* *„szwagier"*	Rotacyzm przejawiający się zastępowaniem głoski *r* dźwiękiem zbliżonym do *y*, bez słyszalnego zwarcia	Rotacyzm przejawiający się zastępowaniem głoski *r* dźwiękiem zbliżonym do *y*, bez słyszalnego zwarcia	Rotacyzm przejawiający się zastępowaniem głoski *r* dźwiękiem zbliżonym do *y*, bez słyszalnego zwarcia	Rotacyzm przejawiający się zastępowaniem głoski *r* dźwiękiem zbliżonym do *y*, bez słyszalnego zwarcia	Rotacyzm przejawiający się zastępowaniem głoski *r* dźwiękiem zbliżonym do *y*, bez słyszalnego zwarcia

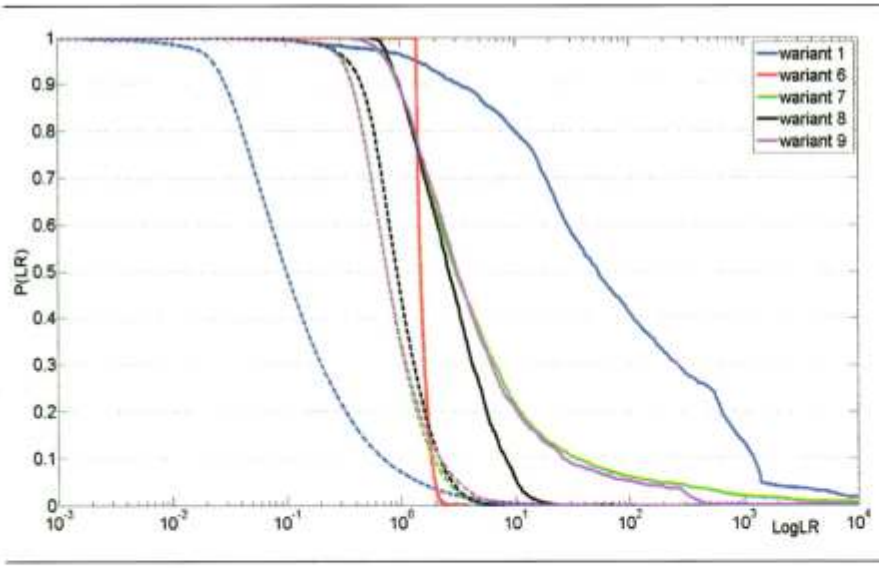
Rotacyzm utrzymuje się w każdej wersji nagrania.

B. Metoda automatyczna

Rycina 3 przedstawia wpływ modyfikacji metodą PSOLA. Wykorzystanie maskowania przez modyfikację tonu krtaniowego (warianty 6 oraz 8, dystrybuanty czarne oraz czerwone) uniemożliwia skuteczne prowadzenie identyfikacji z wykorzystaniem metod automatycznych. Otrzymane LR w każdym z analizowanych przypadków mowy maskowanej (warianty 6 oraz 8) osiąga wartość większą od 1. Odwrócenie modyfikacji z wykorzystaniem wokodera fazowego (warianty 7 oraz 9, dystrybuanty zielone oraz różowe), podobnie jak w przypadku maskowania metodą PSOLA, spowodowało istotny spadek rzetelności testu LR. Dla wariantu 7 otrzymano: $P_{LRHd>1} = 0{,}33$ oraz $P_{LRHp<1} = 0{,}12$ oraz dla wariantu 9: $P_{LRHd>1} = 0{,}34$ oraz $P_{LRHp<1} = 0{,}12$.

Wnioski

Pod względem językowym najbardziej stabilnymi, a tym samym najmniej podatnymi na zastosowane modyfikacje zjawiskami w warstwie fonetycznej, są w obserwowanych przykładach te, związane z wymową samogłosek nosowych oraz spółgłoski *ch'*, a także uproszczenia głosek mające swoją genezę w realizacji regionalnej. Na modyfikacje pod wpływem działania programów maskujących narażone są głównie spółgłoski charakteryzujące się silnym, słyszalnym zwarciem oraz spółgłoski dentalizowane. W analizowanych przykładach odnosi się to głównie do głoski *r*, realizowanej ze zwarciem, oraz do spółgłosek szeregu szumiącego (*sz, rz, cz*).



Numer wariantu	$P_{LRHd>1}$	$P_{LRHp<1}$
Wariant 1 (mowa naturalna)	0,07	0,04
Wariant 6 (maskowanie)	1,00	0,00
Wariant 7 (po korekcji)	0,33	0,12
Wariant 8 (maskowanie)	1,00	0,00
Wariant 9 (po korekcji)	0,34	0,12

Ryc. 3. Wpływ maskowania głosu z wykorzystaniem metod PSOLA oraz PV na rzetelność testu LR

Fig. 3. *Effect of speech disguised with PSOLA and PV methods on reliability of likelihood ratio test*

Fluktuacja natężenia cechy związanej z realizacją głoski *r* dotyczy nie tylko artykulacji wadliwej (mówca M2), w przypadku której w zależności od zastosowanej modyfikacji cecha ta ulega niewielkiemu zatarciu bądź wzmocnieniu, ale też pojawia się w realizacjach oryginalnie poprawnych (M1, M3, M4), a zauważalnych dopiero po zastosowaniu programu PSOLA(A), gdzie ton zmodyfikowano o 5 półtonów w górę, i odwróceniu tego maskowania. Związane jest to ze zniekształceniem w postaci dodatkowego drgania, wprowadzanym zastosowaną metodą. Głoska drżąca, w artykulacji której wibracja jest nieodzownym elementem normatywnej realizacji, ulega wyostrzeniu, a nawet pewnej deformacji. Co ciekawe, podobnych zmian nie zaobserwowano u mówcy M5, który realizuję tę głoskę wokalicznie, bez zwarcia, przez zastąpienie jej dźwiękiem pozasystemowym, zbliżonym do *y*. W tym wypadku rotacyzm obecny jest w identycznym stopniu we wszystkich modyfikacjach i nie zmienia swojej charakterystyki we wspomnianej sytuacji.

Podobnie sytuacja przedstawia się w odniesieniu do głosek postpalatalnych *k'*, *ch'*, realizowanych miękko u mówcy M4. Fluktuacji podlega jedynie głoska *k'* charakteryzująca się zwarciem w trakcie artykulacji. Szczelinowa głoska *ch'* zachowuje swoją miękkość w identycznym niemal stopniu we wszystkich przypadkach.

Powyższe badania pozwoliły zwrócić uwagę na niestandardowe przejawianie się pewnych cech językowych pod wpływem działania określonych programów maskujących. W związku z niewielkim zakresem materiału poddanego analizom trudno we wnioskach wprowadzić uogólnienie dotyczące wszystkich zjawisk danego rodzaju czy danej głoski. Możliwe jest jedynie zauważenie prawidłowości w realizacjach konkretnej głoski obserwowanej w przykładowym nagraniu. Podsumowując poczynione obserwacje, należy stwierdzić, że analiza językowa mówców dotycząca artykulacyjnej warstwy wypowiedzi powinna ze szczególną ostrożnością uwzględniać anomalie w zakresie realizacji głosek detalizowanych szeregu szumiącego *sz*, *rz*, *cz*, zmiękczonej głoski zwarto-wybuchowej *k'* oraz drżącej *r* ze względu na ich podatność na modyfikacje dźwięku.

Badania skuteczności systemu automatycznego rozpoznawania mówców wykazały, że maskowanie głosu z wykorzystaniem elektronicznych metod opartych na modyfikacji tonu krtaniowego, takich jak PSOLA czy PV, jest skutecznym środkiem ukrywania cech osobniczych. Podstawowym warunkiem przeprowadzenia badań metodą pomiarową lub automatyczną wypowiedzi zmodyfikowanych jest odtworzenie oryginalnej wartości tonu krtaniowego. Odwrócenie maskowania na podstawie sygnału referencyjnego wymusza dwukrotne wykorzystanie modyfikatora: pierwotne przez sprawcę w celu dokonania maskowania oraz wtórne przez eksperta prowadzącego analizę. Badania wykazały istotny wpływ odwracania maskowania na skuteczność systemu. W wyniku tego prawdopodobieństwo uznania błędnego dowodu na korzyść oskarżenia ($P_{LRHd>1}$) wzrosło maksymalnie do 0,34, natomiast prawdopodobieństwo uznania błędnego dowodu na korzyść obrony ($P_{LRHp<1}$) wzrosło maksymalnie do 0,12. Spadek rzetelności testu LR najprawdopodobniej wynika ze zniekształceń fazowych i nieliniowych sygnału wprowadzanych przez metody modyfikacji oraz niedokładność detekcji tonu. Badania wykazały również, że nie jest istotne, jaką metodą modyfikacji posługuje się biegły, a jaką osoba dokonująca modyfikacji tonu. Zastosowanie tych samych, a następnie różnych metod, spowodowało zbliżony spadek rzetelności testu LR.

BIBLIOGRAFIA

1. R.D. Rodman: Speaker recognition of disguised voices, in Consortium on Speech Technology Conference on Speaker Recognition by Man and Machine: Directions for Forensic Applications COST250, 1998.

2. P. Perrot, G. Averano, G. Chollet: Voice disguise and automatic detection review and perspectives, Progress in Nonlinear Speech Processing Lecture Notes in Computer Science Volume 4391, 2007, s. 101–117.

3. S. Kajarekar, H. Bratt, E. Shriberg, R. De Leon: A study of intentional voice modifications for evading automatic speaker recognition, Speaker and Language Recognition Workshop, 2006. IEEE Odyssey 2006.

4. W. Enders, W. Bambach, G. Flösser: Voice spectrograms as a function of age, voice disguise and voice imitations J. Acoust. Soc. Am. Volume 49, Issue 6B, s. 1842–1848 (1971).

5. P. Perrot, G. Aversano, G. Chollet: Voice disguise and automatic detection: review and perspectives, Progress in Nonlinear Speech Processing Lecture Notes in Computer Science Volume 4391, 2007, s. 101–117.

6. W. Maciejko: Różnice wewnątrzosobnicze i międzyosobnicze w parametrach akustycznych mówców oryginalnych i ich imitatorów, praca dyplomowa, Politechnika Wrocławska, Wrocław 2005, niepublikowana.

7. A. Eriksson, P. Wretling: How flexible is the human voice? – a case study of mimicry, Fifth European Conference on Speech Communication and Technology, EUROSPEECH 1997, Rhodes, Greece, September 22–25, 1997. ISCA, 1997.

8. H. Hollien, G. DeJong, C. A. Martin, R. Schwartz, K. Liljegren: Effects of ethanol intoxication on speech suprasegmentals, The Journal of the Acoustical Society of America, Vol. 110, No. 6. (December 2001), s. 3198–3206.

9. H. Hollien, K. Liljegren, C. A. Martin, G. DeJong: Production of intoxication states by actors acoustic and temporal characteristics, J Forensic Sci. 2001 Jan;46(1), s. 68–73.

10. J. Clark, P. Foulkes: Identification of voices in electronically disguised speech International Journal of Speech Language and the Law, Vol. 14, No 2 (2007).

11. A. Alexander, F. Botti, D. Dessimoz, A. Drygajlo: The effect of mismatched recording conditions on human and automatic speaker recognition in forensic applications, Forensic Science International, Vol. 146, Supplement, s. 95–99, 2004.

12. H. Künzel: Effects of voice disguise on speaking fundamental frequency, Forensic Languages 7, s. 149–179.

13. C.J.B. Moore: Wprowadzenie do psychologii słyszenia, PWN, Warszawa 1999.

14. L. R. Rabiner, R. W. Schafer: Digital processing of speech signals Prentice-Hall, New Jersey 1978.

15. A. Moussa: Voice conversion using pitch shifting algorithm by time stretching with PSOLA and re-sampling, Journal of Electrical Engineering, Vol. 61, No. 1, 2010, s. 57–61.

16. W.A. Sethares: Rhythm and transforms, Springer Madison 2007.

17. J.T. Kania: Patologiczne artykulacje głoski *r*, Szkice logopedyczne, Polskie Towarzystwo Logopedyczne, Zarząd Główny, Lublin 2001, 243–525.

18. Pod redakcją H. Karaś: „Dialekty i gwary polskie" z dnia 31.12.2010 http://www.dialektologia.uw.edu.pl/index.php?l1=leksykon&lid=741.

19. B. Bartnicka-Dąbkowska: Podstawowe wiadomości z dialektologii polskiej z ćwiczeniami, pod red. B. Wieczorkiewicza, Państwowe Zakłady Wydawnictw Szkolnych, Warszawa 1959, 37–42.

20. Pod redakcją H. Karaś: Dialekty i gwary polskie z dnia 31.12.2010http://www.dialektologia.uw.edu.pl/index.php?l1=leksykon&lid=760.

21. M. Bańko [w]: http://poradnia.pwn.pl/lista.php?szukaj=dzisiej&kat=18, 02.01.2013.

22. W. Maciejko: Biometryczne rozpoznawanie mówców w kryminalistce, Problemy Kryminalistyki 275, Warszawa 2012.

23. D. Meuwly, A. Drygajlo: Forensic speaker recognition based on a bayesian framework and gausian mixture modelling, In 2001, A Speaker Odyssey: The Speaker Recognition Workshop, s. 145–150, 2001.

Streszczenie

Identyfikacja kryminalistyczna mówcy wymaga ekstrakcji cech osobniczych przenoszonych wraz z sygnałem mowy. Sprawcy przeróżnych przestępstw podejmują próby ukrycia tych cech. Jedna z najpopularniejszych technik maskowania polega na wykorzystaniu urządzenia modyfikującego częstotliwość tonu krtaniowego i jest oparta na metodach PSOLA lub PV. Metody te w trakcie resyntezy sygnału generują zniekształcenia, które muszą wpływać na obserwowane cechy. W ramach pracy zbadano wpływ zniekształceń wprowadzanych przez algorytmy modyfikacji tonu krtaniowego na językowe cechy osobnicze oraz skuteczność automatycznego systemu kryminalistycznej identyfikacji mówców wyrażoną za pomocą charakterystyk Tippetta.

Słowa kluczowe: *maskowanie, metoda językowa, automatyczne rozpoznawanie mówców, iloraz wiarygodności, rzetelność LR, PSOLA, wokoder fazowy*

Summary

Forensic speaker recognition is based on individual features which are conveyed with speech signal. Various crime offenders undertake attempts to disguise their individual features. One of the most common voice disguise method involves pitch shifting with PSOLA or PV methods. These methods distort speech signal during signal re-synthesis which has the influence on individual features. In hereby study, the Authors examined the effect of using pitch shifting algorithms on language individual features and effectiveness of forensic automatic speaker recognition which is assessed through Tippett plots.

Keywords: *voice disguise, language method, automatic speaker recognition, likelihood ratio, LR reliability, PSOLA, phase vocoder*