Waldemar Maciejko

Biometryczne rozpoznawanie mówców w kryminalistyce

Wstęp

Pomiar dowolnej wielkości fizycznej jest obciążony błędem wynikającym z dokładności aparatury pomiarowej oraz metodyki pomiaru. Do oceny i eliminacji tych błędów w klasycznej metrologii wykorzystywana jest analiza statystyczna. Pomiar wielkości fizycznych opisujących anatomiczne i behawioralne cechy ludzkie jest znacznie bardziej złożony, ponieważ już samo badane zjawisko jest niepewne. Oprócz czynników wymienionych wyżej na wynik pomiaru ma wpływ szereg dodatkowych zjawisk. Każda badana wielkość jest wypadkową uwarunkowań indywidualnych i grupowych. Dlatego zaobserwowana u jednostki cecha musi być rozpatrywana w kontekście badań tej cechy w populacji. Analiza statystyczna w pomiarach biometrycznych, inaczej niż w klasycznej metrologii, jest narzędziem opisującym relację, jaka zachodzi pomiędzy cechą obserwowaną u jednostki a tą samą cechą zmierzoną u zbiorowości[1].

Współcześnie dorobek biometrii służy budowie wykorzystujących anatomiczne lub behawioralne cechy ludzkie zautomatyzowanych systemów rozpoznawania osób. Jedną z technik biometrycznych jest automatyczne rozpoznawanie osób na podstawie mowy. Ten rodzaj rozpoznawania osób znalazł zastosowanie tam, gdzie ważna jest weryfikacja tożsamości na odległość z wykorzystaniem telefonii, czyli głównie w instytucjach finansowych. Inną dziedziną jest dyscyplina kryminalistyki – fonoskopia.

Wykorzystanie systemów automatycznego rozpoznawania mówców (w skrócie ARM) w kryminalistyce przez wiele lat napotykało poważne trudności. Przykładem może tu być projekt AUROS rozwijany w latach 70. ubiegłego wieku w Niemczech Zachodnich oraz system SASIS budowany w USA. Oba tworzone były na potrzeby badań kryminalistycznych i zostały porzucone ze względu na niesatysfakcjonujące wyniki[2]. Głównymi przeszkodami była zła jakość nagrań oraz krótki czas ich trwania. Pojawienie się szybkich komputerów oraz opracowanie nowych metod parametryzacji sygnału i modelowania statystycznego cech osobniczych umożliwiły wdrożenie nowej generacji systemów. Dotychczasowe trudności związane głównie z jakością nagrań w dużym stopniu zostały przezwyciężone.

Krokiem milowym w rozwoju systemów automatycznego rozpoznania mówców była zmiana podejścia do procesu obliczania cech osobniczych. Metody wykorzystywane obecnie, bazujące na liniowym kodowaniu predykcyjnym, modelowaniu perceptualnym oraz analizie homeomorficznej opracowywano głównie w latach 60. i 70. ubiegłego wieku. Wówczas rozwijane systemy ARM opierały się głównie na niskopoziomowych cechach widmowych. Systemy współczesne natomiast na podstawie obserwacji cech widmowych z użyciem metod modelowania statystycznego dokonują opisu jednostek fonetycznych mowy. Według danych literaturowych najczęściej proces modelowania statystycznego prowadzony jest z wykorzystaniem metody GMM[3].

Celem niniejszego opracowania jest przedstawienie algorytmów obliczania widmowych cech głosu oraz metody ich modelowania statystycznego opartego na mieszaninie modeli normalnych. Prezentowane wyniki badań przeprowadzono z wykorzystaniem bazy mówców polskojęzycznych. Przy tej okazji przedstawiono również metodologię oceny pracy systemu opracowaną przez NIST (National Institute of Standards and Technology).

Automatyczny system rozpoznawania mówców

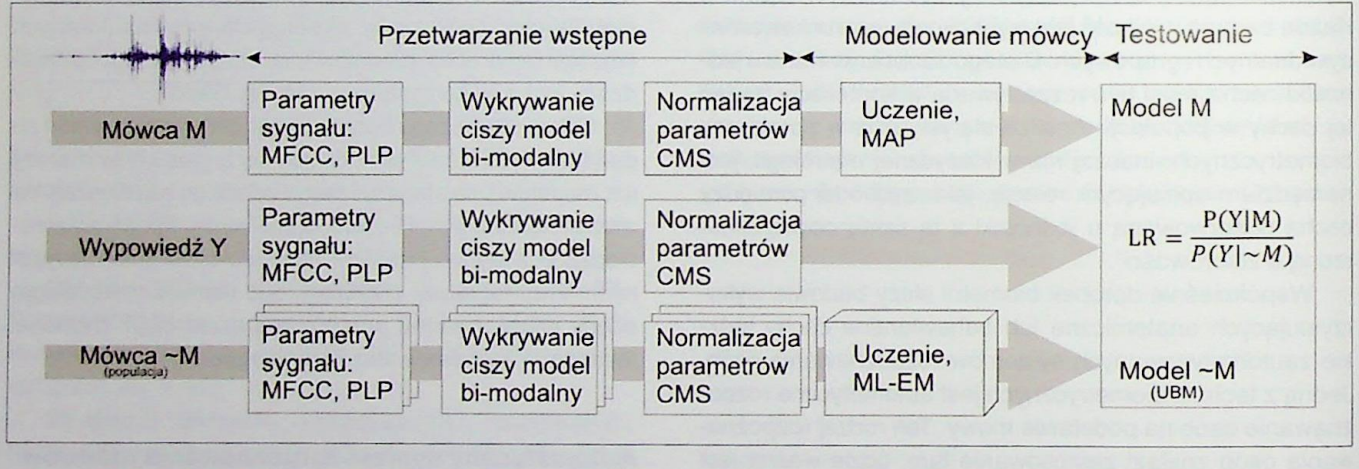
Działanie systemu automatycznego rozpoznania mówców składa się z trzech głównych etapów, są to: przetwarzanie wstępne, modelowanie i testowanie. Proces powyższy przebiega równolegle i synchronicznie na trzech źródłach sygnału: wypowiedzi dowodowej (określanej w publikacji jako wypowiedzi testowa – na rycinie 1 to wypowiedź Y), wypowiedzi porównawczej (inaczej wypowiedź wzorcowa – na rycinie 1 to wypowiedź mówcy M) oraz wypowiedziach mówców z bazy populacyjnej (na rycinie 1 to wypowiedź $\sim M$). Struktura takiego systemu przedstawiona jest na rycinie 1.

Wyniki badań z dziedziny psychoakustyki wskazują, że postrzeganie wysokości tonu przez człowieka jest uwarunkowane przede wszystkim częstotliwością tego tonu oraz jego poziomem. Stwierdzono również, że najmniejsza postrzegana różnica częstotliwości występuje dla wartości około 1 kHz i dla dźwięków o poziomie z zakresu 60–70

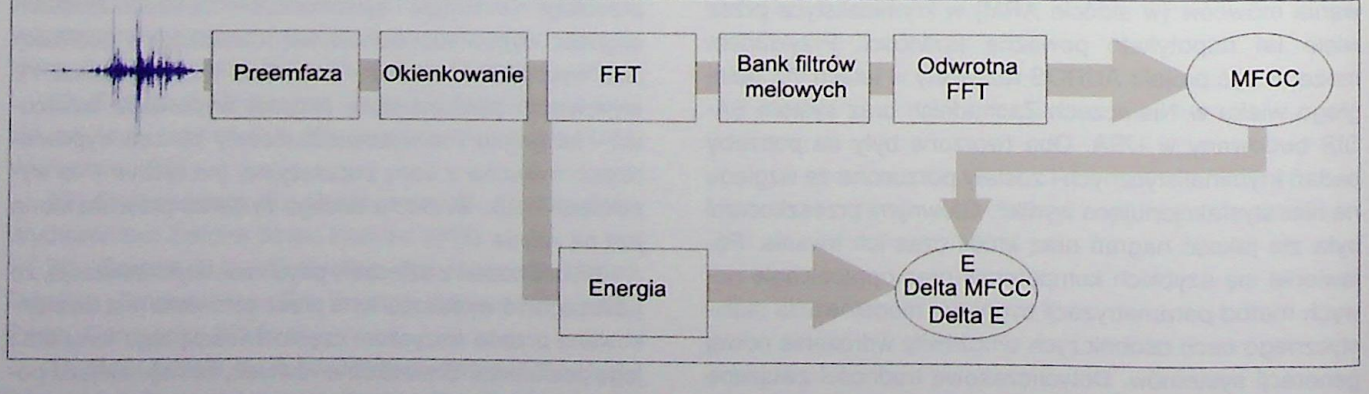
dB SPL[4]. Wyniki tych badań doprowadziły do stworzenia subiektywnej skali wysokości, tzw. mel-skali. Między mel-skalą a skalą częstotliwościową istnieje liniowa zależność do około 1000 Hz, powyżej natomiast tej wartości zależność przechodzi w logarytmiczną. Parametry widmowe, na których opiera się przedmiotowy system ARM, bazują właśnie na mel-skali. Obliczane są one według schematu przedstawionego na rycinie 2.

Bazą parametrów MFCC (*Mel Frequency Cepstral Coefficients*) jest dyskretne widmo sygnału. Widmo poddawane jest filtracji za pomocą filtrów pasmowo-przepustowych, których częstotliwości środkowe są równomiernie rozłożone na melowej osi częstotliwości (tzw. bank filtrów). Ten etap jest rodzajem ważenia charakterystyki widmowej, której efektem jest modyfikowanie rozdzielczości zależnie od zakresu częstotliwości[5]. Następnie wektory obserwacji z wszystkich pasm poddawane są logarytmowaniu i obliczona zostaje dyskretna odwrotna transformata Fouriera. To ostatnie przekształcenie to tzw. transformata cepstralna. Dodatkowa informacja obejmująca dynamikę zmian sygnału w czasie reprezentowana jest przez tzw. współczynniki delta. W ten sposób z każdej ramki otrzymywany jest wektor cech złożony z 16 współczynników MFCC, 16 ΔMFCC, energii E oraz ΔE. W sumie daje to 34 cechy widmowe w jednym wektorze obserwacji.

Wyliczone zgodnie z ryciną 2 parametry widmowe stanowią bazę dalszych obliczeń. Następnym etapem jest identyfikacja i eliminacja wektorów obserwacji, które nie niosą informacji osobniczej, a które zostały obliczone na fragmentach nagrań, gdzie brak jest sygnału mowy. Pauzy są nieodłącznym elementem każdej wypowiedzi. Są to zawieszenia głosu między dwoma kolejnymi dźwiękami w strumieniu fonetycznym. Pauzy są też elementem składni oraz podkreślają intonacyjne rozczłonkowanie ciągu wypowiedzeniowego[6]. System ARM samoczynnie potrafi rozpoznać sygnał nieniosący informacji osobniczej, jednakże cały proces ekstrakcji i przygotowania danych widmowych przebiega przy założeniu, że nagranie poddane analizie zawiera wypowiedź pochodzącą od jednego mówcy. Zgodnie z ryciną 1 w przypadku wypowiedzi dowodowej jest to mówca Y, a porównawczej mówca M. Tak więc podobnie jak w przypadku klasycznej analizy fonoskopijnej badania z wykorzystaniem systemu ARM muszą zostać poprzedzone identyfikacją mówcy



Ryc. 1. Schemat pracy systemu
Fig. 1. *Main modules of automatic speaker recognition*
Źródło (ryc. 1–11): autor



Ryc. 2. Proces obliczania parametrów widmowych MFCC
Fig. 2. *Block diagram of processing steps for extracting spectral MFC coefficients*

w obrębie materiału. W niniejszych badaniach do detekcji ramek mowy wykorzystano kryterium energetyczne. Metoda ta polega na porównaniu energii poszczególnych ramek. Na podstawie analizy statystycznej prowadzonej na przestrzeni całego nagrania ustanowiony jest próg eliminacji. Ramki, reprezentowane przez współczynnik energetyczny, którego wartość znajduje się poniżej progu, są etykietowane i pomijane w dalszej analizie. Założeniem tej metody jest istotna różnica charakterystyk energetycznych pomiędzy sygnałem ciszy a sygnałem mowy. Energia sekwencji próbek $\{s_n, n \in N\}$ obliczana jest ze wzoru:

$$E_i = \log \sum_{n=1}^{N} s_n^2 \quad (1),$$

gdzie s_n to wartość sygnału w danej chwili. Najbardziej złożonym etapem jest ustalenie progu eliminacji. Jedno z podejść wykorzystuje do tego celu analizę statystyczną. Metoda ta polega na zamodelowaniu dwumodalnego rozkładu gęstości współczynników energetycznych obliczonych na wszystkich wektorach obserwacji (ryc. 3). Następnie weryfikowany jest warunek:

$$p(E_i \mid PDF_{mowa}) < p(E_i \mid PDF_{cisza}) \quad (2).$$

Jeżeli warunek jest prawdziwy wektor zostaje zaetykietowany jako cisza i pominięty w dalszej analizie.

Sygnał przesyłany przez określony kanał transmisji to splot dwóch składowych: sygnału oryginalnego i odpowiedzi kanału. Takie zjawiska jak echo kanału i inne zakłócenia addytywne powstające w trakcie transmisji, modyfikując parametry widmowe, wpływają negatywnie na efektywność pracy systemu. Jedną z metod pozwalającą usunąć z wektorów obserwacji składową dodaną przez kanał jest technika CMS (*Cepstral Mean Subtraction*). Zgodnie z tą metodą obliczenie nowych parametrów następuje przez odjęcie od wartości danego współczynnika obliczonego w danej ramce sygnału (c_k) współczynnika uśrednionego po całej wypowiedzi[7]. Wyrażenie to przedstawiono na równaniu (3):

$$c_{k,CMS}(n) = c_k(n) - \sum_{i=1}^{N} [c_k(i)] \quad (3).$$

Z czasem metoda ta została zmodyfikowana i uzupełniona o normalizację współczynników, natomiast operacja uśredniania jest prowadzona jedynie na ramkach reprezentujących wypowiedzi. Oznacza to, że zmodyfikowana operacja CMS musi być poprzedzona algorytmem detekcji sygnału mowy.

Opisane wyżej działania to wstępny proces przetwarzania parametrów widmowych. Zgodnie z ryciną 1 kolejny etap to modelowanie cech osobniczych mówcy. Prezentowana tutaj metoda wykorzystuje mieszaninę modeli normalnych (*Gaussian Mixture Models*). Modele mieszanin rozkładów Gaussa opierają się na założeniu, że gęstość prawdopodobieństwa wystąpienia pewnej cechy zdefiniowanej w przestrzeni wielowymiarowej może zostać przybliżona funkcją będącą ważoną sumą skończonej liczby rozkładów normalnych. Prawdopodobieństwo wystąpienia zatem danej cechy zdefiniowane jest następująco:

$$p(x \mid \theta) = \sum_{i=1}^{R} \omega_i p_i(x) \quad (4),$$

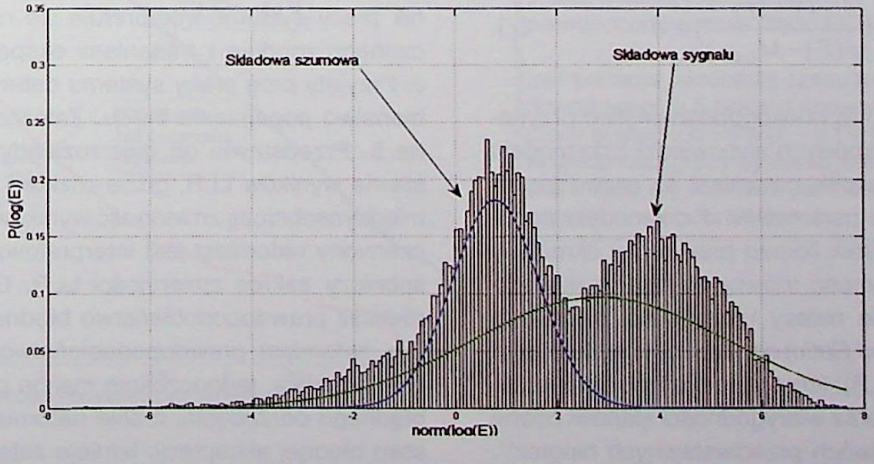
gdzie R to liczba rozkładów Gaussa (tzw. rząd modelu) w mieszaninie, ω_i to waga każdego komponentu mieszaniny spełniająca następujący warunek:

$$\sum_{i=1}^{R} \omega_i = 1 \quad (5).$$

Czynnik $p_i(x)$ z wyrażenia (4) obliczany jest ze wzoru:

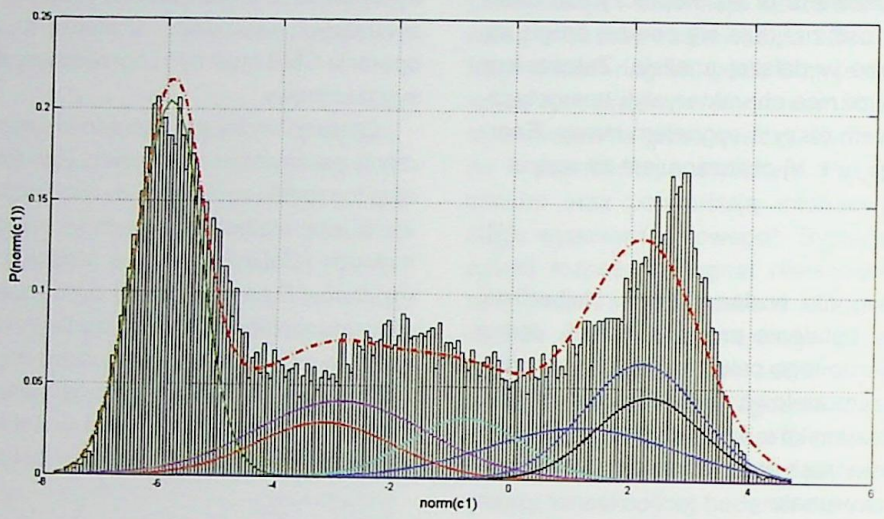
$$p_i(x) = \frac{1}{(2 \cdot \pi)^{D/2} |\Sigma_i|^{1/2}} \cdot \exp\left\{-\frac{1}{2} \cdot (x - \mu_i)'(\Sigma_i)^{-1}(x - \mu_i)\right\} \quad (6),$$

gdzie D to liczba elementów wektora, μ_i jest wektorem wartości średnich o wymiarach D x 1, Σ_i jest macierzą kowariancji o wymiarach D x D. Reasumując, model



Ryc. 3. Rozkład dwumodalny (rząd modelu równy 2) obliczony na około 3-minutowym nagraniu
Fig. 3. *Bi-gaussian modeling (order equals 2) of the energy distribution calculated on 3-minutes utterance*

mówcy opisany za pomocą GMM zdefiniowany jest przez następujący zbiór parametrów mieszaniny rozkładów: $\theta = \{\omega_i, \mu_i, \Sigma_i\}$ gdzie $i = 1, ..., R$[8]. Model graficzny stworzony na bazie obliczonych parametrów rozkładu przedstawiono na rycinie 4.



Ryc. 4. Model GMM rzędu równego 8 dla pojedynczego współczynnika MFCC obliczonego na 20-sekundowej wypowiedzi mówcy posługującego się językiem polskim

Fig. 4. *Gaussian Mixture Model, order equals 8 of the MFC coefficient calculated on 20 seconds utterance of Polish speaker*

Na podstawie empirycznych obserwacji stwierdzono, że komponenty mieszaniny zbudowane na podstawie długookresowych obserwacji widma sygnału reprezentują jednostki fonetyczne, które określane są jako klasy akustyczne. Pojedyncze rozkłady reprezentują samogłoski i spółgłoski, a ich kombinacja jest reprezentacją układu artykulacyjnego mówcy. Ponieważ modelowane klasy nie są w żaden sposób oznaczane, mówi się o niejawnych klasach lub modelach[9].

Ostatnim elementem pracy systemu jest identyfikacja. Polega ona na ocenie ilorazu dwóch prawdopodobieństw według następującej zależności:

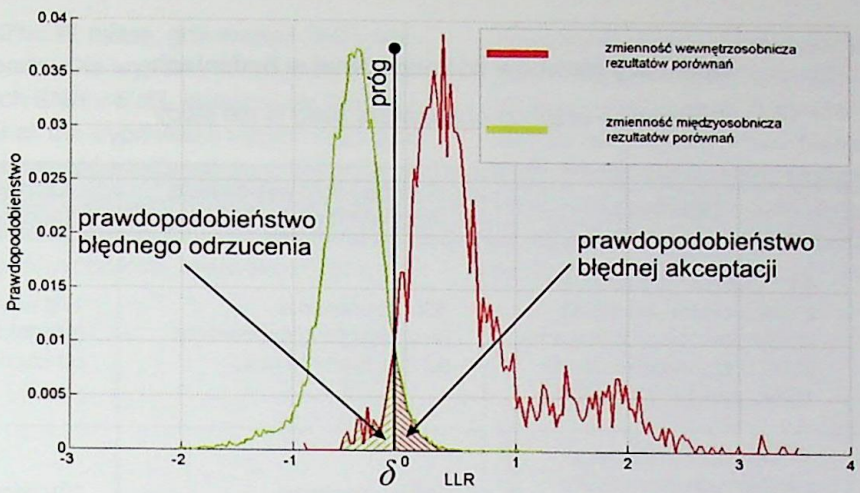
$$akceptacja \Rightarrow LLR = \log\left(\frac{p(Y \mid M)}{p(Y \mid \sim M_{UBM})}\right) \geq \delta \qquad (7).$$

Licznik wyrażenia (7) to prawdopodobieństwo przynależności parametrów widmowych wypowiedzi Y do modelu GMM mówcy M. Mianownik natomiast to prawdopodobieństwo przynależności parametrów Y do modelu populacyjnego $\sim M$. Jeżeli wynik ilorazu przekroczy określony wcześniej próg δ, tożsamość mówcy zostaje zaakceptowana. Wyrażenia (7) nie należy utożsamiać z ilorazem wiarygodności (*Likelihood Ratio*) oznaczanym w literaturze również przez LR (lub LLR), którego formalizacja matematyczna jest podobna[10]. Iloraz wiarygodności stanowi ocenę prawdziwości jednej z dwóch przeciwstawnych hipotez[11]. W niniejszym opracowaniu autor, posługując się skrótem LLR, odnosi się do wartości zdefiniowanej w wyrażeniu (7).

Model populacyjny, którego parametry są wykorzystywane do oszacowania wartości mianownika wyrażenia (7), stanowi estymację parametrów populacji na podstawie próby z tej populacji. Model taki w skrócie nazywany jest UBM (*Universal Background Model*). Kryterium wyboru odpowiedniej próby z populacji jest jednocześnie konkretną aplikacją systemu. Dotyczy to przede wszystkim systemów pracujących na potrzeby analizy kryminalistycznej. Tutaj istotny jest dobór struktury modelu populacyjnego nie tylko ze względu na płeć, wiek oraz język, którym posługują się mówcy, ale również technikę rejestracji, jaka została użyta w celu utrwalenia głosów będących przedmiotem analizy.

Wyniki badań

Miarą stopnia zgodności cech osobniczych mówców jest wartość uzyskana na podstawie wyrażenia (7). Wynik pracy systemu interpretuje się na podstawie wyznaczonego zgodnie z badaniami eksperymentalnego progu δ. Przyjęty próg pracy systemu determinuje prawdopodobieństwo popełnienia błędu. Zależność tę obrazuje rycina 5. Przedstawia on dwa rozkłady utworzone na podstawie wyników LLR, gdzie rozkład zielony reprezentuje międzyosobniczą zmienność wyników porównań. Rozkład czerwony natomiast jest interpretowany jako wewnątrzosobniczy zakres zmienności LLR. Gdy δ rośnie, rośnie również prawdopodobieństwo błędnego odrzucenia, maleje natomiast prawdopodobieństwo błędnej akceptacji. Gdy δ maleje, jednocześnie maleje prawdopodobieństwo błędnego odrzucenia, rośnie natomiast prawdopodobieństwo błędnej akceptacji. Istnieje zatem funkcja pomiędzy progiem δ a prawdopodobieństwem błędnego odrzucenia i prawdopodobieństwem błędnej akceptacji.



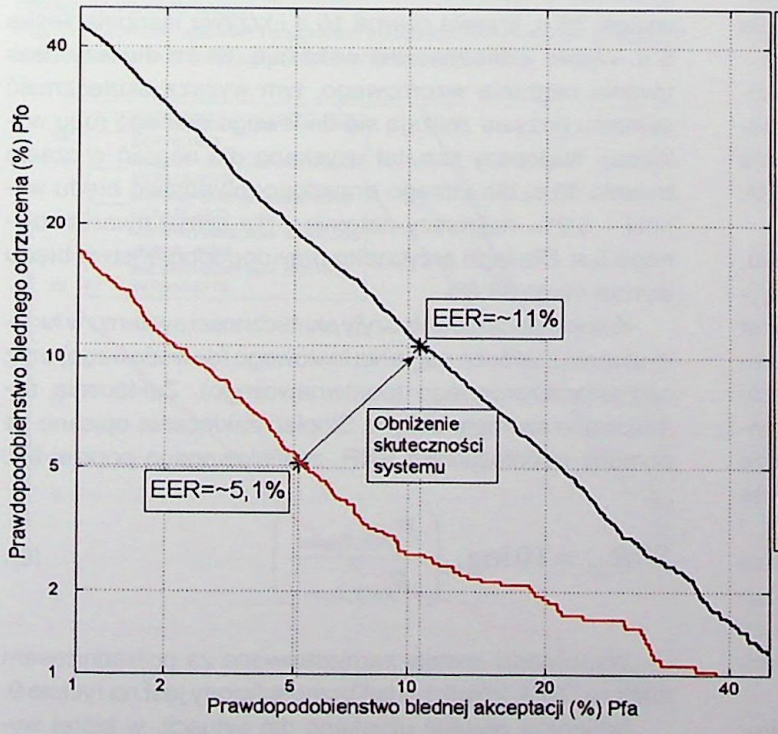
Ryc. 5. Rozkłady wartości logarytmu LR wyznaczonego zgodnie z wyrażaniem (7) dla wyników 1754 porównań pomiędzy tymi samymi mówcami (rozkład czerwony) oraz 25446 porównań pomiędzy różnymi mówcami (rozkład zielony)

Fig. 5. *Distributions of log* LR *scores calculated according to expression no. 7, red curve represents within score variability calculated for 1754 comparisons, green curve represents between score variability calculated for 25446 comparisons*

Funkcja taka prezentowana jest graficznie za pomocą charakterystyk DET (*Detection Error Tradeoff*). Charakterystyki te mówią o relacji, jaka zachodzi pomiędzy prawdopodobieństwem wystąpienia błędnej akceptacji (P_{fa}) i prawdopodobieństwem błędnego odrzucenia (P_{fo}). Dzięki nieliniowo wyskalowanym osiom rośnie rozdzielczość charakterystyki, co ułatwia porównanie różnych krzywych reprezentujących pracę w różnych warunkach lub różnych systemów[12].

Oprócz charakterystyk DET do oceny wykorzystano również współczynnik EER (*Equal Error Rate*). Jest to punkt na krzywej DET, dla którego wartości prawdopodobieństwa wystąpienia jednego z dwóch rodzajów błędów są sobie równe. Rycina 6 przedstawia dwie charakterystyki DET wraz z ich interpretacją.

Do badań wykorzystano bazę składającą się z 44 mówców polskojęzycznych. Jej opis zawarty jest w tabeli.



Porównanie skuteczności pracy systemu dla dwóch przypadków:

nagranie testowe (dowodowe) o czasie trwania około 30 sekund, nagranie wzorcowe (porównawcze) o czasie trwania około 30 sekund. Otrzymano rezultat prawdopodobieństwa błedu EER = ~ 5,1%.

nagranie testowe (dowodowe) o czasie trwania około 25 sekund, nagranie wzorcowe (porównawcze) o czasie trwania około 30 sekund. Otrzymano rezultat prawdopodobieństwa błędu EER = ~ 11%.

Interpretacja: skrócenie czasu trwania nagrania dowodowego o 5 sekund spowodowało pogorszenie pracy systemu o ~ 6% (z ~ 5,1% do ~ 11%). Pogorszenie skuteczności pracy systemu jest widoczne przez odsunięcie się charakterystki w kierunku prawego górnego rogu pola wykresu.

Ryc. 6. Charakterystyki DET wraz z ich interpretacją. Wyniki uzyskano na podstawie wypowiedzi 44 mówców polskojęzycznych komunikujących się za pośrednictwem telefonii GSM

Fig. 6. *Example of two DET curves with description. Test base consists of 44 Polish language speakers communicating with GSM*

Tabela

Opis bazy mówców wykorzystanej w badaniach

Description of group of apeakers used in the study

Badanie	Opis bazy	Liczba i opis wypowiedzi testowych (dowodowych)	Liczba i opis wypowiedzi wzorcowych (porównawczych)
Wpływ czasu trwania nagrania testowego oraz wpływ czasu trwania nagrania wzorcowego na skuteczność	44 mówców, rejestracja z wykorzystaniem GSM (częstotliwość próbkowania 8000 kHz, rozdzielczość 16 bit, zapis PCM WAV)	160 wypowiedzi (średnio ok. 4 wypowiedzi od każdego mówcy)	170 wypowiedzi (średnio ok. 4 wypowiedzi od każdego mówcy)
Wpływ jakości nagrania	44 mówców, rejestracja z wykorzystaniem GSM z dodaniem szumu białego (częstotliwość próbkowania 8000 kHz, rozdzielczość 16 bit, zapis PCM WAV)	160 wypowiedzi (średnio ok. 4 wypowiedzi od każdego mówcy)	170 wypowiedzi (średnio ok. 4 wypowiedzi od każdego mówcy)
Badanie wpływu transmisji	44 mówców, wypowiedzi każdego zarejestrowano trzema technikami GSM, PSTN oraz w warunkach pokojowych mik. pojemnościowy (częstotliwość próbkowania 44100 Hz, rozdzielczość 16 bit, zapis bezstratny PCM WAV).	GSM: 160 wypowiedzi (średnio ok. 4 wypowiedzi od każdego mówcy). PSTN: 213 wypowiedzi (średnio ok. 5 wypowiedzi od każdego mówcy). Mik. pojemnościowy: 312 wypowiedzi (średnio ok. 8–9 wypowiedzi od każdego mówcy)	GSM: 170 wypowiedzi (średnio ok. 4 wypowiedzi od każdego mówcy). PSTN: 249 wypowiedzi (średnio ok. 5 wypowiedzi od każdego mówcy). Mik. pojemnościowy: 312 wypowiedzi (średnio ok. 8–9 wypowiedzi od każdego mówcy)

W ramach eksperymentu zbadano wpływ czasu trwania, jakości oraz techniki transmisji na skuteczność systemu. W badaniach wykorzystano nagrania typowe dla kryminalistycznej identyfikacji mówców.

W pierwszej kolejności zbadano jak obniży się skuteczność systemu w miarę skracania trwania nagrania testowego (dowodowego). Badania te przeprowadzono dla wypowiedzi zarejestrowanych za pośrednictwem GSM. Wynik przedstawiony jest na rycinie 7.

Krzywa czerwona reprezentuje wyniki dla przypadku, w którym nagrania testowe (dowodowe) trwają 30 s, krzywa czarna 10 s, a jasnoniebieska 5 s. Najlepszy rezultat otrzymano dla wypowiedzi testowych (dowodowych) trwających około 30 s, dla których prawdopodobieństwo błędu (EER) wynosi najmniej i jest równe około 5,2%. Dla wypowiedzi o czasie trwania 10 s oraz 5 s otrzymano bardzo zbliżone wyniki, prawdopodobieństwo błędu (EER) równe jest około 11%.

W kolejnej serii pomiarowej sprawdzono wpływ czasu trwania nagrania wzorcowego (porównawczego) na skuteczność systemu ARM. Wypowiedzi zarejestrowano za pośrednictwem telefonii GSM. Wyniki tych badań przedstawiono na rycinie 8.

Badania przeprowadzono w czterech seriach pomiarowych. We wszystkich przypadkach czas trwania nagrania dowodowego to 30 s. Krzywa czerwona reprezentuje wyniki badań, w których czas trwania nagrania porównawczego wynosi 30 s. Krzywa żółta reprezentuje wyniki dla przypadku, w którym czas trwania nagrania porównawczego wynosi 15 s, krzywa czarna 10 s i krzywa jasnoniebieska 5 s. Wyniki jednoznacznie wskazują, że im dłuższy czas trwania nagrania wzorcowego, tym wyższa skuteczność systemu (krzywe zbliżają się do lewego dolnego rogu wykresu). Najlepszy rezultat uzyskano dla nagrań o czasie trwania 30 s, dla którego prawdopodobieństwo błędu wynosi ~ 5,2%, najgorszy natomiast dla czasu trwania równego 5 s. Dla tego przypadku prawdopodobieństwo błędu wynosi około 38,6%.

Kolejne badania dotyczyły skuteczności systemu w funkcji stopnia zakłóceń nagrania testowego (dowodowego) oraz nagrania wzorcowego (porównawczego). Zakłócenia symulowano szumem białym. Stopień zakłócenia opisano za pomocą współczynnika SNR, zdefiniowanego poniżej (8).

$$SNR_{dB} = 10\log_{10}\left(\frac{P_{RMS,Sygnal}}{P_{RMS,Szum}}\right) \quad (8).$$

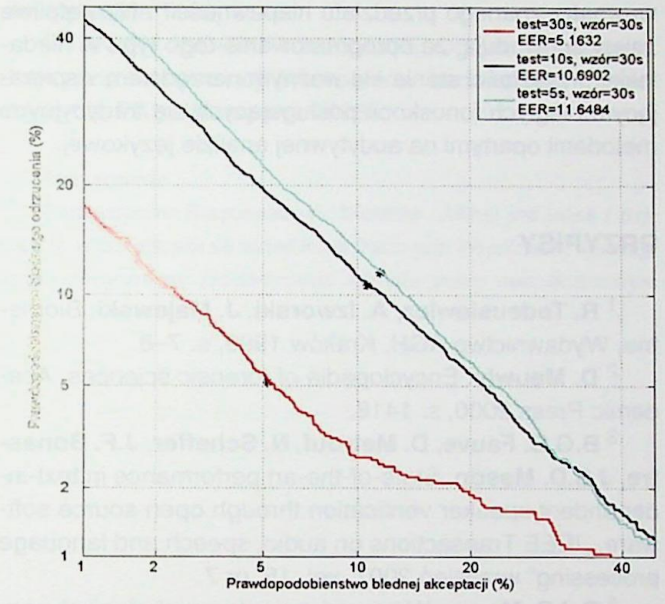
Wypowiedzi zostały zarejestrowane za pośrednictwem telefonii GSM. Wynik badań przedstawiony jest na rycinie 9.

Najlepszy rezultat uzyskano dla sytuacji, w której wypowiedź testową (dowodową) oraz wzorcową (porównawczą) charakteryzuje współczynnik SNR > 20 dB (krzywa czerwona), dla tego przypadku uzyskano prawdopodobień-

stwo błędu równe 5,2%. W miarę, gdy maleje SNR, skuteczność systemu spada. Dla wypowiedzi testowych (dowodowych), dla których SNR = 5 dB, skuteczność systemu spada do ~ 20%. Jednak dla wypowiedzi wzorcowych o takim samym SNR skuteczność spada już do ~ 42%.

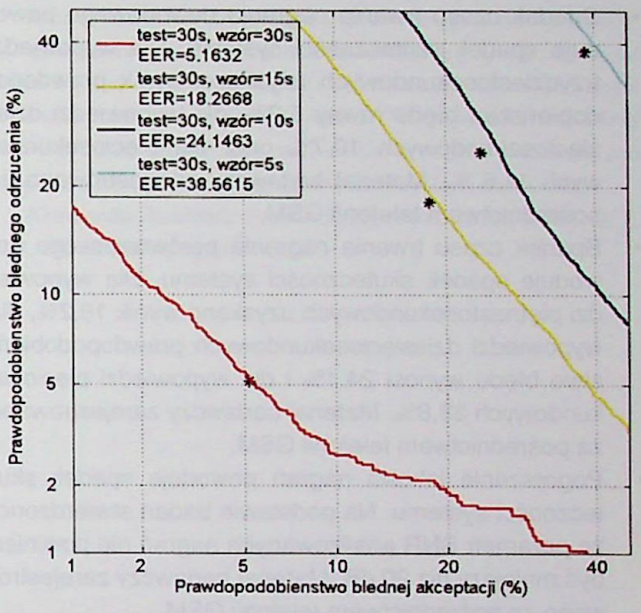
Ostatnim elementem zbadanym w ramach niniejszej pracy jest skuteczność systemu zależnie od toru elektroakustycznego. Przeanalizowano wypowiedzi zarejestrowane za pośrednictwem trzech różnych kanałów: telefonii stacjonarnej PSTN, telefonii komórkowej GSM oraz mikrofonu pojemnościowego – wypowiedzi bezpośrednie zarejestrowane w warunkach pokojowych. Czas trwania porównywanych wypowiedzi wynosił około 30 s, współczynnik SNR > 20 dB.

W wyniku badań uzyskano współczynnik EER dla telefonii stacjonarnej (krzywa czerwona) równy około 6,6%, dla telefonii komórkowej 5,2% (krzywa niebieska) oraz dla



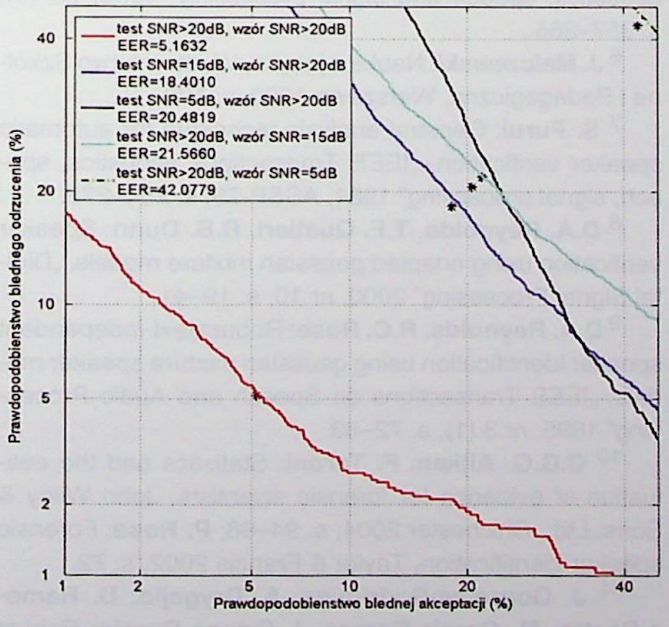
Ryc. 7. Ocena pracy systemu ARM dla malejącego czasu trwania nagrania testowego (dowodowego) przy stałym czasie trwania nagrania wzorcowego (porównawczego)

Fig. 7. *Automatic speaker recognition evaluation in function of duration of test utterance and constant duration of target utterance*



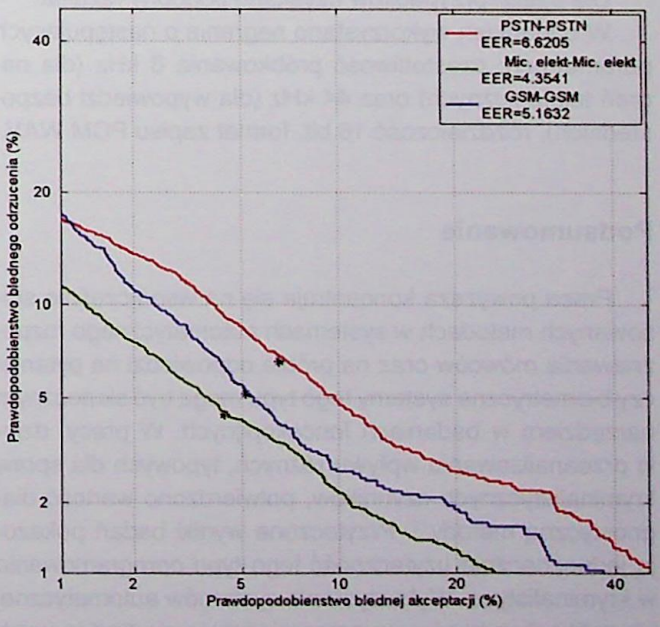
Ryc. 8. Badanie skuteczności pracy systemu ARM dla malejącego czasu trwania nagrania wzorcowego przy stałym czasie trwania nagrania testowego

Fig. 8. *Automatic speaker recognition evaluation in function of duration of target utterance and constant duration of test utterance*



Ryc. 9. Badanie skuteczności systemu ARM zależnie od jakości nagrań testowego i wzorcowego

Fig. 9. *Automatic speaker recognition evaluation in function of quality of target and test utterance*



Ryc. 10. Porównanie skuteczności systemu ARM dla wypowiedzi przesłanych różnymi metodami transmisji sygnału mowy

Fig. 10. *Comparison of automatic speaker recognition system for test and target utterances transmitted with 3 different methods*

mikrofonu pojemnościowego 4,4% (krzywa zielona). Wyniki te pokazują, że zbudowany system nie jest uzależniony od techniki rejestracji i transmisji dźwięku. We wszystkich analizowanych przypadkach rezultat był podobny.

Wnioski

Na podstawie badań stwierdzono:

- Spadek czasu trwania nagrania dowodowego powoduje spadek skuteczności systemu. Dla wypowiedzi trzydziestosekundowych uzyskano wynik prawdopodobieństwa błędu równy 5,2%, dla wypowiedzi dziesięciosekundowych 10,7% oraz dla pięciosekundowych 11,6 %. Materiał badawczy zarejestrowano za pośrednictwem telefonii GSM.
- Spadek czasu trwania nagrania porównawczego powoduje spadek skuteczności systemu. Dla wypowiedzi piętnastosekundowych uzyskano wynik 18,2%, dla wypowiedzi dziesięciosekundowych prawdopodobieństwo błędu wynosi 24,1% i dla wypowiedzi pięciosekundowych 38,8%. Materiał badawczy zarejestrowano za pośrednictwem telefonii GSM.
- Pogorszenie jakości nagrań powoduje spadek skuteczności systemu. Na podstawie badań stwierdzono, że parametr SNR analizowanych nagrań nie powinien być mniejszy niż 20 dB. Materiał badawczy zarejestrowano za pośrednictwem telefonii GSM.
- technika rejestracji nie wpływa na skuteczność systemu. Przebadano wypowiedzi zarejestrowane za pośrednictwem telefonii GSM, PSTN oraz bezpośrednio mikrofonu pojemnościowego w warunkach pokojowych. Dla trzech przypadków uzyskano podobny rezultat.

W badaniach wykorzystano nagrania o następujących parametrach: częstotliwość próbkowania 8 kHz (dla nagrań telefonicznych) oraz 44 kHz (dla wypowiedzi bezpośrednich), rozdzielczość 16 bit, format zapisu PCM WAV.

Podsumowanie

Praca powyższa koncentruje się na współcześnie stosowanych metodach w systemach automatycznego rozpoznawania mówców oraz na próbie odpowiedzi na pytanie: czy biometryczne systemy tego typu mogą być skutecznym narzędziem w badaniach fonoskopijnych. W pracy, dzięki przeanalizowaniu wpływu różnych, typowych dla spraw kryminalistycznych czynników, potwierdzono wartość diagnostyczną metody[13]. Przytoczone wyniki badań pokazują jednoznacznie użyteczność tego typu oprogramowania w kryminalistyce. Wykorzystanie systemów automatycznej identyfikacji mówców, w przeciwieństwie do tradycyjnych metod stosowanych dotychczas w fonoskopii opartych głównie na analizie językowo-audytywnej, ma wiele niewątpliwych zalet, są to m.in. obiektywizacja wyniku, możliwość analizy porównawczej wypowiedzi mówców obcojęzycznych oraz wypowiedzi zaledwie kilkusekundowych różniących się pod względem leksykalnym. Fundamentalna, ze względu na ocenę dowodu z opinii biegłego przez organ procesowy, jest znajomość prawdopodobieństwa błędu metody zastosowanej w badaniach. Ważne jest to również w kontekście normy PN-EN ISO/IEC 17025:2005, która zakłada, że wynik miarodajny to wynik, którego wartość rzeczywista z określonym prawdopodobieństwem znajduje się wewnątrz znanego przedziału niepewności[14]. Niewątpliwe zalety spowodują, że oprogramowanie tego typu w niedalekiej przyszłości stanie się ważnym narzędziem wspierającym biegłych fonoskopii posługujących się tradycyjnymi metodami opartymi na audytywnej analizie językowej.

PRZYPISY

[1] **R. Tadeusiewicz, A. Izworski, J. Majewski**: Biometria, Wydawnictwo AGH, Kraków 1993, s. 7–8.

[2] **D. Meuwly**: Encyclopedia of forensic sciences, Academic Press 2000, s. 1418.

[3] **B.G.B. Fauve, D. Matrouf, N. Scheffer, J.F. Bonastre, J.S.D. Mason**: State-of-the-art performance in text-independent speaker verification through open-source software, „IEEE Transactions on audio, speech and language processing" wrzesień 2007, vol. 15, nr 7.

[4] **C.J.B. Moore**: Wprowadzenie do psychologii słyszenia, PWN, Warszawa-Poznań 1999, s. 199.

[5] **S.B. Davis, P. Mermelstein**: Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, „IEEE Transactions on acoustics, speech and signal processing" 1980, 28 (4), s. 357–366.

[6] **J. Malczewski**: Nauka o języku, Wydawnictwo Szkolne i Pedagogiczne, Warszawa 1990, s. 199.

[7] **S. Furui**: Cepstral analysis technique for automatic speaker verification, „IEEE Transactions acoustics, speech, signal processing" 1981, ASSP-29, s. 254–272.

[8] **D.A. Reynolds, T.F. Quatieri, R.B. Dunn**: Speaker verification using adapted gaussian mixture models, „Digital Signal Processing" 2000, nr 10, s. 19–41.

[9] **D.A. Reynolds, R.C. Rose**: Robust text-independent speaker identification using gaussian mixture speaker models, „IEEE Transactions on Speech and Audio Processing" 1995, nr 3 (1), s. 72–83.

[10] **C.G.G. Aitken, F. Taroni**: Statistics and the evaluation of evidence for forensic scientists, John Wiley & Sons, Ltd., Chichester 2004, s. 94–98; **P. Rose**: Forensic speaker identification, Taylor & Francis 2002, s. 72.

[11] **J. Gonzales-Rodriguez, A. Drygajlo, D. Ramos-Castro, M. Garcia-Gomar, J. Ortega-Garcia**: Robust estimation, interpretation and assessment of likelihood ratios in forensic speaker recognition, „Computer Speech and Language" 2006, nr 20, s. 331–355.

[12] **A. Martin, G. Doddington, T. Kamm, M. Ordowski, M. Przybocki**: The DET curve in assessment of detection task performance, [w:] Proceedings of the 5th European Conference on Speech Communication and Technology, vol. 4, Rhodes, Greece 1997, s. 1895–1898.

[13] Termin wartości diagnostycznej (identyfikacyjnej) zdefiniowano w literaturze, patrz na przykład: **T. Widła**: Ocena dowodu z opinii biegłego, Prace naukowe Uniwersytetu Śląskiego w Katowicach nr 1309, Uniwersytet Śląski, Katowice 1992, s. 46.

[14] PN-EN ISO/IEC 17025:2005, Ogólne wymagania dotyczące kompetencji laboratoriów badawczych i wzorcujących, PKN, Warszawa 2005.

Streszczenie

Automatyczne Rozpoznawanie Mówców (ARM) jest jedną z najszybciej rozwijających się metod biometrycznych. Współczesne systemy, dzięki efektywnemu przetwarzaniu sygnału mowy oraz skutecznym metodom rozpoznawania, mogą być wykorzystane w wielu dziedzinach, m.in. identyfikacji kryminalistycznej. W pracy opisano krótką historię rozwoju ARM, przedstawiono system zgodny ze współczesnym stanem wiedzy w zakresie przetwarzania sygnału oraz metod rozpoznawania. Opisano również metody oceny pracy systemu.

Słowa kluczowe: *Automatyczne Rozpoznawanie Mówców, GMM, UBM, DET, EER, LR, wartość diagnostyczna metody*

Summary

Automatic Speaker Recognition (ARM) belongs to one of the most extensively developed biometric techniques. The highly effective signal processing and recognition methods can be successfully implemented in various fields, such as forensic science. This paper describes a brief history of ARM development, state-of-the-art signal processing and recognition methods, including techniques of systems assessment.

Keywords: *Automatic Speaker Recognition, GMM, UBM, DET, EER, LR, method's diagnostic value*