Waldemar Maciejko
Biuro Badań Kryminalistycznych Agencji Bezpieczeństwa Wewnętrznego
w.maciejko@abw.gov.pl

Wpływ transmisji głosu z wykorzystaniem telefonii internetowej VoIP na skuteczność automatycznego systemu kryminalistycznej identyfikacji mówców opartego na metodzie EM-UBM-MAP

Streszczenie

W niniejszej pracy zbadano wpływ transmisji pakietowej na skuteczność systemu automatycznej identyfikacji mówców pracującego w oparciu o bayesowski klasyfikator LR. Modelowanie statystyczne w systemie prowadzone jest z wykorzystaniem algorytmów EM-GMM oraz MAP.

W badaniach założono, że z transmisją pakietową wiążą się dwa zjawiska: utrata pakietów oraz kodowanie sygnału. W badaniach wykorzystano niektóre kodeki audio standardu H.323 ITU-T. Zjawisko utraty pakietów przybliżono za pomocą dwustanowego modelu Gilberta. Wyniki badań przedstawiono w postaci charakterystyk Tippetta.

Słowa kluczowe Voice over Internet Protocol, kryminalistyczna identyfikacja mówców, model Gilberta, utrata pakietów, kompresja dźwięku, standard H.323

Wprowadzenie

Upowszechnienie dostępu do Internetu oraz rozwój technologii internetowych doprowadziły do powstania telefonii internetowej zwanej VoIP (*Voice over Internet Protocol*). Zaletą tej technologii jest obniżenie kosztów połączeń oraz możliwość równoległej transmisji innych danych. Na świecie rynek usług VoIP bardzo szybko się rozwija [1]. Szacuje się, że w Polsce już 49% użytkowników Internetu korzysta również z telefonii internetowej, przy czym najszerzej ta technologia wykorzystywana jest w łączności biznesowo-korporacyjnej. Dane te jednoznacznie wskazują na rosnącą rolę tego rodzaju usług telekomunikacyjnych. Rozwój telefonii VoIP jest konsekwencją zjawiska konwergencji sieci i usług. Ten postępujący proces definiowany jest jako zespolenie wszystkich funkcji kanałów transmisji głosu, obrazu oraz danych w jedną szerokopasmową strukturę opartą na protokole IP.

Wpływ zjawisk generowanych w kanale GSM i PSTN na skuteczność systemów automatycznego rozpoznawania mówców jest dobrze poznany. Przez ostatnie 20 lat prowadzono szereg prac badawczych z tego zakresu. Sprawdzono m.in. wpływ zakłóceń addytywnych imitowanych za pomocą pasmowego szumu białego i kolorowego [2]. Dodatkowo zweryfikowano wpływ zakłóceń nieliniowych powstających w mikrofonach [2, 3]. Uzyskane wyniki jednoznacznie wskazują, iż zakłócenia tego typu ujemnie wpływają na skuteczność systemu rozpoznawania. Badaniom poddano również wpływ metod kompresji sygnału mowy wykorzystywanych w telefonii komórkowej m.in. GSM 06.10, GSM 06.20 i GSM 06.60 oraz wpływ tych metod kodowania na cechy osobnicze głosu [4, 5]. Uzyskane wyniki pokazują, że badane zjawiska prowadzą do spadku skuteczności systemów rozpoznawania mówców. Stopień spadku skuteczności systemu jest uzależniony od zjawiska, które wpływa na sygnał mowy.

W literaturze kryminalistycznej brak jest szczegółowej analizy wpływu zjawisk generowanych w sieci na skuteczność metod kryminalistycznej identyfikacji mówców. Doświadczenie autora wskazuje, że nagrania rozmów za pośrednictwem protokołów internetowych są coraz częściej przedmiotem badań fonoskopijnych. Dotychczas prowadzone badania koncentrowały się przede wszystkim na możliwości transmisji danych głosowych, a wraz z nimi równoległej transmisji danych biometrycznych. W rozwiązaniu tym akwizycja sygnału mowy oraz modelowanie statystyczne odbywają się po stronie abonenta

[6]. Właściwy proces identyfikacji prowadzonej za pomocą systemu automatycznego w takich rozwiązaniach następuje po stronie usługodawcy na podstawie danych przesłanych wraz z sygnałem mowy. Tak więc badania koncentrują się na typowej aplikacji biometrycznej wykorzystywanej do autoryzacji aktywności zlecanych za pośrednictwem linii telefonicznej. Natomiast analiza fonoskopijna internetowych rozmów telefonicznych wymaga znajomości zjawisk występujących w sieci IP i ich wpływu na sygnał mowy i cechy osobnicze wraz z nim przenoszone.

Celem badań przedstawionych w niniejszej pracy było określenie wpływu zjawisk powstających w trakcie transmisji sygnału mowy za pośrednictwem telefonii internetowej na skuteczność automatycznego systemu kryminalistycznej identyfikacji mówców.

Transmisja głosu przez sieć IP

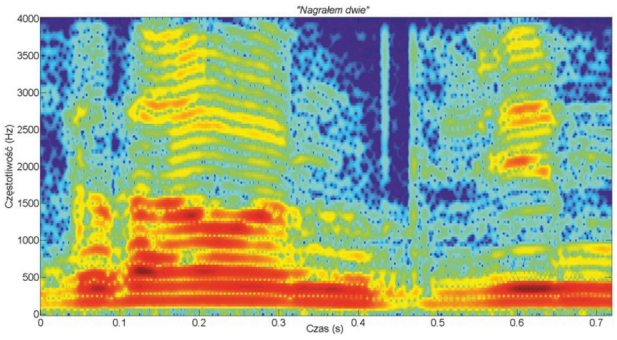
Publiczna telefonia komutowana (PSTN) oraz telefonia komórkowa (GSM) do zestawienia połączenia wykorzystują komutację łączy (*circuit switching*). Technika ta polega na zajmowaniu kolejnych odcinków drogi połączeniowej między dwoma stacjami na podstawie informacji sygnalizacyjnych wysyłanych przez stacje żądające połączenia. Po zestawieniu całej drogi połączeniowej ze stacji docelowej wysyłana jest informacja sygnalizacyjna o utworzeniu połączenia. Dopiero po tej fazie następuje transmisja danych. Utworzona droga pozostaje tylko do użytku stacji, które zainicjowały połączenie, aż do chwili rozłączenia [7].

Pierwszym standardem definiującym techniczne wymagania dla zapewnienia usług łączności audio w czasie rzeczywistym w sieciach pakietowych była rekomendacja Międzynarodowej Unii Telekomunikacyjnej numer H.323. Zgodnie z tą rekomendacją dźwięk przesyłany jest według reguł protokołów internetowych [8]. Podstawową jednostką informacji jest wymieniany między stacjami końcowymi pakiet (zwany też datagramem RTP), który stanowi zapis binarny o skończonej długości i składa się z nagłówka oraz obszaru danych [7, 9]. Nagłówek to informacje protokołów poszczególnych warstw modelu OSI (*Open Systems Interconnection*), natomiast obszar danych pakietu to dane reprezentujące sygnał dźwiękowy [10]. Proces formowania sygnału mowy w pakiety rozpoczyna się od kompresji dźwięku. Kolejnym etapem jest pakietyzacja, w trakcie której skompresowany sygnał mowy dzielony jest na segmenty, które zapisywane są w obszarze danych pakietu [11]. Parametrem definiującym czas trwania sygnału mowy reprezentowanym przez obszar danych w pojedynczym pakiecie, jest interwał pakietyzacji [11, 12]. Standardowy i zalecany interwał pakietyzacji wynosi 20 ms [11, 12]. Zgodnie rekomendacją H.323 dźwięk strumieniowany zakodowany jest za pomocą protokołu transmisji czasu rzeczywistego RTP (*Real Time Protocol*), który zapewnia informacje o czasie nadania pakietu [8, 10, 13]. RTP do transportu wykorzystuje protokół UDP (*User Datagram Protocol*), który jest protokołem bezpołączeniowym i nie udostępnia żadnych mechanizmów kontroli dostarczenia danych oraz kolejności otrzymania pakietów [8, 9]. Zaletą protokołu UDP jest brak narzutu dodatkowych danych związanych z mechanizmami kontroli oraz duża szybkość transmisji, co w przypadku przesyłania rozmowy prowadzonej w czasie rzeczywistym jest kluczowe [10]. UDP wykorzystuje protokół IP, który zapewnia sumy kontrolne oraz dane adresowe odbiorcy [10].

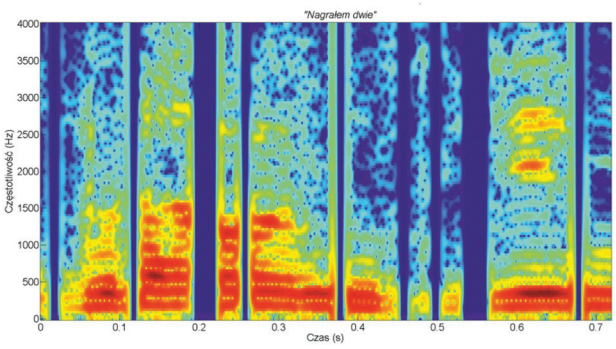
W telefonii VoIP, wykorzystującej komutację pakietów, inaczej zatem niż w przypadku komutacji łączy, kanał transmisyjny między stacjami końcowymi zajmowany jest tylko podczas przesyłania pakietów, po czym automatycznie jest zwalniany. Transmisja informacji głosowej za pomocą pakietów wiąże się z występowaniem niekorzystnych zjawisk, takich jak opóźnienia oraz utrata pakietów. Na opóźnienie wpływa: czas przesyłania danych między punktami końcowymi, czas przetwarzania informacji w urządzeniach sieciowych (np. w ruterach) oraz czas pakietyzacji danych dźwiękowych [9].

Utrata oraz ukrywanie utraty pakietów

Brak mechanizmów kontroli dostarczania danych w protokole UDP powoduje, że datagramy, które nie zostaną dostarczone do stacji końcowej, nie ulegają retransmisji i są tracone [8]. Innym powodem utraty pakietów jest odrzucanie tych, które straciły swoje znaczenie, ze względu na ich zbyt duże opóźnienie [14]. Efekt utraty pakietów obrazuje spektrogram wypowiedzi „nagrałem dwie” przedstawiony na ryci-



Ryc. 1. Spektrogram wypowiedzi „nagrałem dwie”, zastosowany kodek ITU-T G729 6,4 kb/s, prawdopodobieństwo utraty pakietów 0, czas trwania pakietu 20 ms.



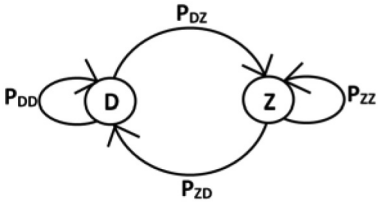
Ryc. 2. Spektrogram wypowiedzi „nagrałem dwie", zastosowany kodek ITU-T G729 11,8 kb/s, prawdopodobieństwo utraty pakietu $P_Z = 0{,}25$, interwał pakietyzacji jest równy 20 ms.

nie 2. Tą samą wypowiedź, niezniekształconą przedstawiono na rycinie 1.

Zjawisko utraty informacji może przyjmować dwie różne postaci. Jedną z nich jest przypadkowa utrata pojedynczych pakietów [9]. W drugim przypadku tracona może być grupa pakietów kolejno po sobie następujących, tworząc tzw. paczki błędów [9, 15, 16, 17]. Parametrem określającym liczbę utraconych kolejno po sobie następujących pakietów jest *BL* (*Burst Length*) w literaturze określany jako G_{min} [14]. Ze względu na zmieniające się warunki w sieci IP, wynikające np. z okresów zwiększonego natężenia ruchu w sieci, długości paczek błędów mogą się zmieniać w czasie [15]. Dlatego wprowadzono również parametr średniej długości paczek błędów (*Mean Burst Lost Length* – MBLL), który określa średnią liczbę pakietów w utraconej paczce [15, 18].

Badania nad oceną degradacji jakości transmisji sygnału mowy w sieci IP na skutek utraty pakietów mogą być prowadzone pod warunkiem stworzenia modelu matematycznego, który pozwoliłby w sposób kontrolowany wprowadzać te zakłócenia do sygnału. W literaturze przedmiotu zaproponowano kilka modeli matematycznych zjawiska utraty pakietów, np. model Bernoullego oraz bardziej złożony model oparty na szeregach Markowa [16, 18]. Według danych literaturowych stosunkowo prostym, dostarczającym dokładnego przybliżenia zjawiska utraty pakietów w paczkach, jest model Gilberta. Zakłada on, że stan bieżącego pakietu jest uzależniony jedynie od stanu pakietu go poprzedzającego [16, 17, 18]. Model ten został zaimplementowany w badaniach opisanych w niniejszej pracy. Jego ideę przedstawiono na rycinie 3 [19].

Jeżeli przyjmiemy, że X jest zmienną losową reprezentującą rezultat transmisji pakietu w chwili n, to sekwencja zdarzeń X_n, gdzie $n \in N$, stanowi dyskretny szereg Markowa. Rezultat transmisji może przyjmo-



Ryc. 3. Model Gilberta.

wać jeden z dwóch stanów oznaczonych jako: *D* (*pakiet dostarczony*) oraz *Z* (*pakiet utracony*). Prawdopodobieństwo przejścia do stanu *Z*, gdy stanem poprzednim był również stan *Z*, zdefiniowane jest następująco: $P_{zz}(X_{n+1} = Z \mid X_n = Z) = clp$. Zmienną losową X cechuje rozkład geometryczny [15, 16, 18]. Zatem prawdopodobieństwo, że *BL* następujących po sobie pakietów (paczka) zostanie utracone zdefiniowane jest następująco (1) [15, 16]:

$$P(BS) = clp^{BS-1}(1-clp), \quad BS \geq 0 \tag{1}$$

Średnia długość paczki błędów (MBLL) stanowi wartość oczekiwaną szeregu geometrycznego i wynosi (2) [15, 16]:

$$MBBL = \frac{1}{1-clp} \tag{2}$$

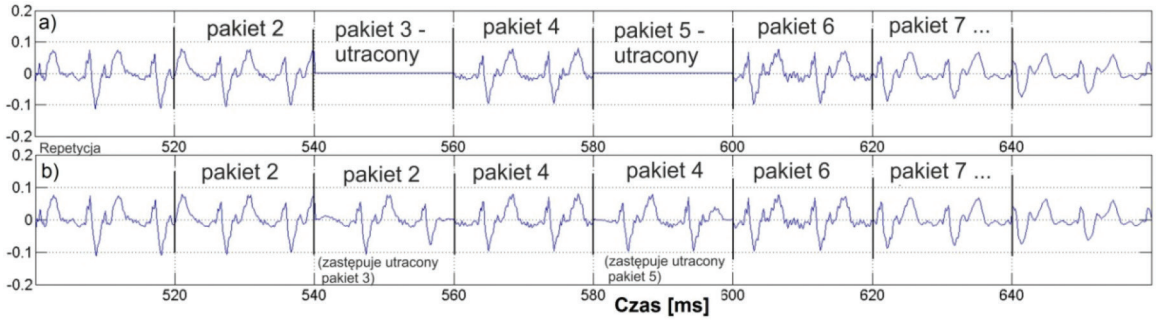
Prawdopodobieństwo P_z wystąpienia stanu *Z* pakietu nazywane jest prawdopodobieństwem utraty pakietów i zdefiniowane jest następująco (3) [16]:

$$\begin{aligned} P_Z = {} & P_D\left(X_n = D\right) \cdot P_{DZ}\left(X_{n+1} = Z \mid X_n = D\right) + \\ & + P_Z \cdot P_{ZZ}\left(X_{n+1} = Z \mid X_n = Z\right) \end{aligned} \tag{3},$$

gdzie $P_D(X_n = D)$ to prawdopodobieństwo tego, że pakiet znajdzie się w stanie *D* oraz $P_{DZ}(X_{n+1} = Z \mid X_n = D)$ to prawdopodobieństwo wystąpienia stanu *Z*, gdy pakiet poprzedni znajdował się w stanie *D*. Wielkość P_Z to podstawowy parametr definiujący stopień utraty pakietów. Jest on również definiowany jako stosunek liczby pakietów utraconych do liczby pakietów wysłanych w danej jednostce czasu (*Packet Loss Ratio* – PLR) [14, 18]. Ostatecznie na podstawie wyrażeń (2) oraz (3) P_{DZ} wynosi (4) [18]:

$$P_{DZ}(X_{n+1} = Z \,||\, X_n = D) = \frac{P_z}{MBLL \cdot (1-P_z)} \tag{4}$$

Utrata pakietów może powodować odczuwalne, chwilowe zaniki sygnału. W celu zamaskowania efektu tego zjawiska opracowano specjalne metody ukrywania utraty pakietów (*Packet Loss Concealment* – PLC). Obecnie wykorzystywanych jest wiele metod PLC. Najbardziej złożone obliczeniowo generują syntetyczną informację głosową na podstawie ostatnich



Ryc. 4. Efekt utraty pakietów bez zastosowania techniki PLC (a), ukrycie przez repetycję (b).

dostarczonych pakietów. Najprostszą metodą, która może być stosowana w czasie rzeczywistym jest repetycja pakietów [12, 20]. Jej sposób działania przedstawiono na rycinie 4.

Metoda ta opiera się na założeniu, że informacje zawarte w sąsiadujących ze sobą ramkach niewiele się różnią. Ukrycie przez repetycję polega na zastępowaniu utraconego pakietu przez ostatni otrzymany pakiet [12, 20]. Wykorzystanie tej techniki daje dobre rezultaty przy minimalnej złożoności obliczeniowej [20]. Ze względu na powszechność wykorzystania metody repetycji i jej prostotę efekt jej działania był przedmiotem badań, których wynik prezentowany jest w niniejszej pracy.

Kodeki

Zadaniem kompresji danych dźwiękowych jest redukcja liczby bitów potrzebnych do wiernego odwzorowania sygnału mowy w celu jego przesłania na odległość, a następnie odtworzenia. Proces ten jest realizowany z wykorzystaniem kodera oraz dekodera (*ang*. *COder-DECoder* – CODEC). Podstawowym celem kompresji sygnału mowy w telefonii VoIP jest redukcja strumienia informacji, dzięki czemu potrzebne pasmo może być wielokrotnie mniejsze. Standard ITU-T H.323 dopuszcza użycie wielu kodeków dźwięku. W celu oceny wpływu algorytmów kompresji sygnału mowy na skuteczność systemu automatycznej identyfikacji mówców wybrano trzy kodeki o różnej przepływności, których parametry, na podstawie rekomendacji ITU-T, przedstawiono w tabeli 1.

System automatycznej identyfikacji mówców

W pracy analizowany jest system identyfikacji mówców, który wykorzystuje 34 cechy widmowe (tworząc tzw. wektory 34-wymiarowe), na które składa się: 16 melowych współczynników cepstralnych (*mel frequency cepstral coefficients* – MFCC), 16 dynamicznych, melowych współczynników cepstralnych (ΔMFCC) oraz dwa współczynniki energetyczne: logarytm energii sygnału (*E*) oraz współczynnik dynamiczny logarytmu energii sygnału (Δ*E*). Jeden wektor 34-wymiarowy MFCC obliczany jest z ramki o czasie 20 ms, która w trakcie analizy przesuwana jest z krokiem równym 10 ms. Proces obliczania współczynników MFCC szeroko opisano w literaturze (zob. [24, 25]). Kolejny krok to detekcja sygnału mowy (*Voice Activity Detection* – VAD). W analizowanym systemie wykorzystano kryterium energetyczne. Metoda ta polega na porównaniu logarytmu energii sygnału w poszczególnych ramkach. Na podstawie analizy statystycznej

Tabela 1

Kodeki wykorzystane w trakcie badań

Standard	Algorytm kompresji	Ramka wg. standardu [ms]	Współczynnik kompresji	Przepływność [kb/s]
G.711 [21]	PCM a-Law Modulacja impulsowo-kodowa	0,125	1 : 1 bezstratny	64,0
G.729 [22]	CS-ACELP Sprzężona algebraiczna pobudzana kodem predykcja liniowa	10	8 : 1 stratny	11,8
G.723.1 [23]	MP-MLQ Modulacja wieloimpulsowa maksymalnego prawdopodobieństwa kwantyzacji	30	10 : 1 stratny	6,3

prowadzonej na przestrzeni całej wypowiedzi ustanawiany jest próg eliminacji. Jeżeli logarytm energii sygnału w danej ramce przekracza ustanowiony próg, wektor reprezentowany przez taką wartość *E* etykietowany jest jako „sygnał mowy" i jest wykorzystywany do dalszej analizy. Krytycznym elementem tej metody jest ustanowienie progu eliminacji. Stosuje się do tego celu mieszaninę modeli normalnych 2. rzędu, zbudowaną na podstawie współczynnika *E.* Wspomniane kryterium ustanawia się na podstawie zasady detekcji Bayesa [25, 26]. W kolejnym etapie realizowana jest standaryzacja cepstrum (*Cepstral Mean Variance Normalization* – CMVN). Algorytm ten ma na celu minimalizowanie wpływu zakłóceń addytywnych przez korygowanie wartości wektorów cech [25]. Zgodnie z tą metodą korygowanie współczynników MFCC następuje przez odjęcie od wartości danego współczynnika obliczonego w danej ramce sygnału współczynnika uśrednionego po całej wypowiedzi. Metodę CMVN opisano szeroko w literaturze (zob. [25, 27]).

Analizowany system identyfikacji mówców umożliwia porównywanie wypowiedzi różniących się pod względem treści. Oparty jest na metodzie modelowania parametrycznego GMM (*Gaussian Mixture Models*). Zgodnie z założeniem tej metody współczynniki widmowe MFCC, zdefiniowane w przestrzeni wielowymiarowej, mogą zostać przybliżone funkcją będącą ważoną sumą skończonej liczby *M* rozkładów normalnych opisanych z wykorzystaniem parametrów statystycznych: wektora wartości średniej $\vec{\mu}$, macierzy kowariancji Σ oraz wektora wag $\vec{\omega}$. Wielkość *M* definiuje tzw. rząd modelu. Komponenty mieszaniny modeli normalnych stanowią opis krótkoterminowej zmienności ludzkiego głosu [24, 28, 29]. Parametry rozkładu oblicza się z wykorzystaniem iteracyjnego algorytmu EM (ang. *Expectation Maximization*). Metoda ta opiera się na szacowaniu parametrów mieszaniny w dwóch krokach – estymacji i maksymalizacji.

W pierwszym kroku (*E* – estymacji) obliczane jest prawdopodobieństwo a posteriori przynależności wektora $\vec{x}_n$ (*D*-wymiarowy wektor cech widmowych MFCC obliczony w chwili *n, n* = 1, *…, N*) do mieszaniny *i,* zdefiniowanej za pomocą parametrów $\lambda_i(\vec{\mu}_i, \Sigma_i, \vec{\omega}_i)$, gdzie *i* = 1, *…, M* to rząd modelu (5) [21]:

$$p_i(i \mid \vec{x}_n, \lambda_i^t) = \frac{\omega_i^t \cdot p_i(\vec{x}_n, \lambda_i^t)}{\sum_{k=1}^{K} \omega_k \cdot p_i(\vec{x}_n, \lambda_i^t)} \tag{5}$$

W równaniu (5) *t* to indeks iteracji, natomiast $p(\vec{x}_n, \lambda_i^t)$ to *D*-wymiarowa funkcja gęstości komponentu, którą zdefiniowano w równaniu (6) [24, 28]:

$$p_i(\vec{x}, \lambda_i^t) = \frac{1}{(2\cdot\pi)^{D/2}\left|\Sigma_i^t\right|^{-1/2}} \cdot \cdot \exp\left\{-\frac{1}{2}\cdot(\vec{x}_n - \vec{\mu}_i^t)'(\Sigma_i^t)^{-1}(\vec{x}_n - \vec{\mu}_i^t)\right\} \tag{6}$$

W kolejnym kroku (*M* – maksymalizacji) obliczane są aktualne wartości parametrów rozkładu: wektor średnich, macierz kowariancji oraz wektor wag [24].

Wagi komponentów:

$$\vec{\omega}_i^{t+1} = \frac{1}{N}\Sigma_{n=1}^{N} p_i(i \mid \vec{x}_n, \lambda_i^t) \tag{7}$$

Średnia

$$\vec{\mu}_i^{t+1} = \frac{\Sigma_{n=1}^{N} p_i(i \mid \vec{x}_n, \lambda_i^t)\cdot \vec{x}_n}{\Sigma_{n=1}^{N} p_i(i \mid \vec{x}_n, \lambda_i^t)} \tag{8}$$

Wariancja:

$$\Sigma_i^{t+1} = \frac{\Sigma_{n=1}^{N} p_i(i \mid \vec{x}_n, \lambda_i^t)(\vec{x}_n - \vec{\mu}_i^t)(\vec{x}_n - \vec{\mu}_i^t)'}{\Sigma_{n=1}^{N} p_i(i \mid \vec{x}_n, \lambda_i^t)} \tag{9}$$

Pełną iterację zamyka szacowanie nowej wartości prawdopodobieństwa a posteriori na podstawie wyrażenia (5) z wykorzystaniem parametrów oszacowanych za pomocą wyrażeń (7), (8) oraz (9). Jeżeli zachodzi warunek konwergencji $p_i(i|\vec{x}_n, \lambda_i^t) > p_i(i|\vec{x}_n, \lambda_i^{t+1})$ iteracja zostaje zakończona [24, 25, 28, 29].

Algorytm *EM* wymaga dużej ilości danych wejściowych, aby otrzymany model w sposób reprezentatywny opisywał wypowiedź. Najczęściej wykorzystywany jest on do określenia parametrów modelu populacyjnego, tzw. modelu UBM (*Universal Background Model*). Dobór mówców do populacji musi się odbywać z uwzględnieniem wypowiedzi, które będą poddane później analizie kryminalistycznej. Tak więc dobór struktury populacyjnej odbywa się ze względu na płeć, wiek, język, którym posługują się mówcy, oraz technikę rejestracji, jaka została użyta w celu utrwalenia głosów stanowiących przedmiot analizy [28, 30].

Do określenia parametrów mieszaniny reprezentującej cechy osobnicze mówcy z nagrania porównawczego wykorzystywany jest iteracyjny algorytm MAP (*Maximum a posteriori*). Metodą tą skutecznie oszacowuje się parametry modelu λ_M mówcy przy minimalnej ilości danych uczących. W pierwszym kroku obliczane jest prawdopodobieństwo $p_i(i|\vec{x}_n, \lambda_{M,i}^t)$ na podstawie wyrażenia (5), przy czym parametrami wejściowymi $\lambda_{M,i}(\vec{\mu}_i, \Sigma_i, \vec{\omega}_i)$ są wartości modelu UBM. Następnie z wyrażenia (7) obliczany jest wektor wag $\vec{\omega}_i$. Pomocnicza wartość średnia obliczana jest z następującego wyrażenia (10) [28]:

$$\hat{\vec{\mu}}_i = \frac{1}{\vec{\omega}_i}\Sigma_{n=1}^{N} p_i(i \mid \vec{x}_n, \lambda_{M,i}^t)x_n \tag{10},$$

gdzie $p_i(i|\vec{x}_n, \lambda_{M,i}^t)$ to prawdopodobieństwo a poste-

riori przynależności wektora $\vec{x}_n$ (*D*-wymiarowy wektor cech widmowych MFCC obliczony w chwili *n*) do mieszaniny *i,* zdefiniowanej za pomocą parametrów $\lambda_{M,i}(\vec{\mu}_i,\Sigma_i,\vec{\omega}_i)$, gdzie *i* = 1, …, *M* [25, 27, 28]. Do adaptacji modelu mówcy wykorzystywana jest jedynie wartość średnia parametrów UBM. Możliwe jest również wykorzystanie pozostałych parametrów, tj. kowariancji oraz wektora wag, jednak zwiększa to znacznie złożoność obliczeniową algorytmu i zwiększa prawdopodobieństwo zwrócenia błędu przez system [30, 31]. Ostatecznie wartość średnia modelu λ_M obliczana jest na podstawie wyrażenia (11) [28]:

$$\vec{\mu}_i = \alpha_i \hat{\vec{\mu}}_i + (1-\alpha_i)\cdot\mu_i \quad (11),$$

gdzie współczynnik α jest zdefiniowany jako:

$$\alpha_i = \frac{\vec{\omega}_i}{\vec{\omega}_i + r} \quad (12),$$

gdzie współczynnik *r* decyduje o stopniu adaptacji [28].

Analizowany system kryminalistycznej identyfikacji mówców wykorzystuje metodę modelowania *EM* w celu oszacowania parametrów mieszaniny modelu populacyjnego UBM. W badaniach przeanalizowano przypadek, w którym rząd modelu wynosi 128. Do oszacowania parametrów modelu mówcy wykorzystywany jest algorytm *MAP* adaptujący wartość średnią. Wartość współczynnika *r* = 16 modelu *MAP* dobrano na podstawie danych literaturowych [28].

Materiał użyty do badań

W pracy wykorzystano wypowiedzi spontaniczne w języku polskim, zarejestrowane w warunkach pomieszczenia mieszkalnego 38 mówców polskojęzycznych. Każdy z mówców udzielił po 10 wypowiedzi, których czas trwania wynosił około 30 sekund, co dało 380 plików dźwiękowych. Zapis dźwięku prowadzono w formacie PCM WAV z częstotliwością próbkowania 44,1 kHz oraz rozdzielczością 16 bitów. Następnie pliki dźwiękowe poddano konwersji z wykorzystaniem 3 kodeków audio opisanych wcześniej, co daje 3 zbiory nagrań po 380 plików dźwiękowych. Następnie do nagrań wprowadzono zakłócenia w postaci utraty pakietów. Proces ten przeprowadzono na każdym pliku dźwiękowym i składał się z następujących etapów: 1. podział nagrania na ramki o czasie 20 ms (interwał pakietyzacji), 2. wygenerowanie z wykorzystaniem modelu Gilberta, z zadanym parametrem $P_z \in \{0, 0{,}01, 0{,}1, 0{,}25\}$ oraz MBLL = 2 szeregu czasowego o wartościach 0 (pakiet utracony) lub 1 (pakiet dostarczony), 3. odwzorowanie szeregu czasowego w kolejnych ramkach nagrania.

Definicja skuteczności automatycznego systemu kryminalistycznej identyfikacji mówców

W systemie automatycznej identyfikacji mówców do interpretacji wyniku analizy wykorzystano iloraz wiarygodności (*likelihood ratio* – LR) mający swoje uzasadnienie w ilorazowej postaci twierdzenia Bayesa. Wielkość LR porównuje prawdopodobieństwo, że analizowane dwie wypowiedzi pochodzą od tej samej osoby z prawdopodobieństwem, że pochodzą od różnych osób [32, 33]. Iloraz wiarygodności można zdefiniować jako stosunek dwóch wartości prawdopodobieństw warunkowych [29, 33]:

$$LR = \frac{\mathrm{P}(zgodność \mid podejrzany\ jest\ mówcą\ w\ nagraniu\ dowodowym)}{\mathrm{P}(zgodność \mid podejrzany\ przypadkowo\ ma\ cechy\ zgodne\ z\ mówcą\ z\ wypowiedzi\ dowodowej)}$$

Metodykę szacowania ilorazu wiarygodności w odniesieniu do systemów automatycznej identyfikacji mówców opisano w literaturze (zob. [29]).

Wyniki badań przedstawiono za pomocą charakterystyk Tippetta, które są szeroko wykorzystywane w badaniach rzetelności wartości LR [33]. Interpretację charakterystyk Tippetta oraz metodę ich szacowania w odniesieniu do automatycznego systemu identyfikacji mówców przedstawiono w publikacji D. Meuwly'ego i A. Drygajły [29]. Na podstawie charakterystyk Tippetta wyznaczono dwie charakterystyczne wielkości prawdopodobieństwa [33]:

$\mathbf{P}_{LRHd>1}$ – prawdopodobieństwo otrzymania wartości LR > 1, gdy porównywane wypowiedzi pochodzą od dwóch różnych osób,

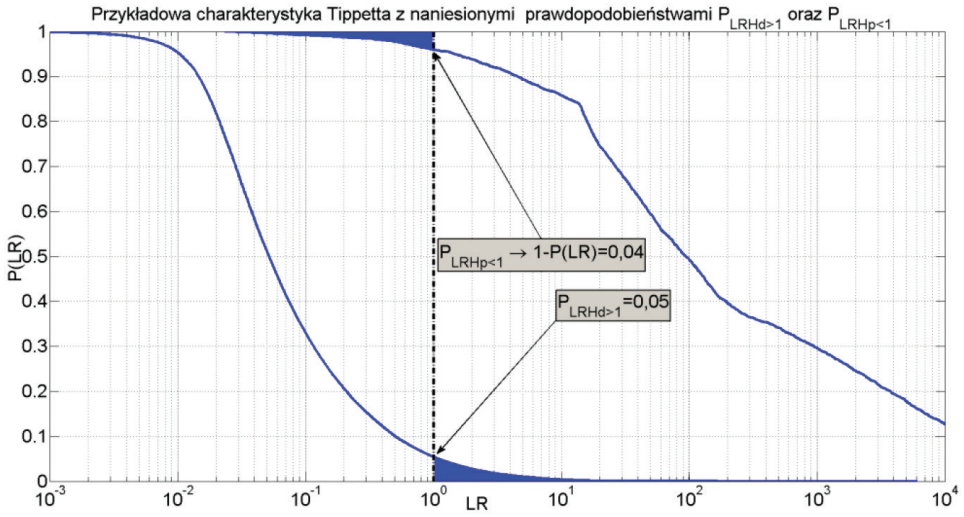
$\mathbf{P}_{LRHp<1}$ – prawdopodobieństwo otrzymania wartości LR < 1, gdy porównywane wypowiedzi pochodzą od tej samej osoby.

Ich interpretację graficzną przedstawiono na rycinie 5.

Wielkości $P_{LRHp<1}$ oraz $P_{LRHd>1}$ opisują skuteczność automatycznego systemu kryminalistycznej identyfikacji mówców. Stosując terminologię procesową, współczynnik $P_{LRHd>1}$ określa, na ile prawdopodobne jest uzyskanie ilorazu wiarygodności większego od 1, jeżeli podejrzany nie jest źródłem wypowiedzi dowodowej [33]. Natomiast współczynnik $P_{LRHp<1}$ określa, na ile prawdopodobne jest to, że iloraz wiarygodności będzie mniejszy od 1, jeżeli podejrzany jest źródłem wypowiedzi dowodowej [33].

Wyniki badań

Użyty do badań materiał dźwiękowy został wykorzystany do sprawdzenia skuteczności systemu. Zbadano:



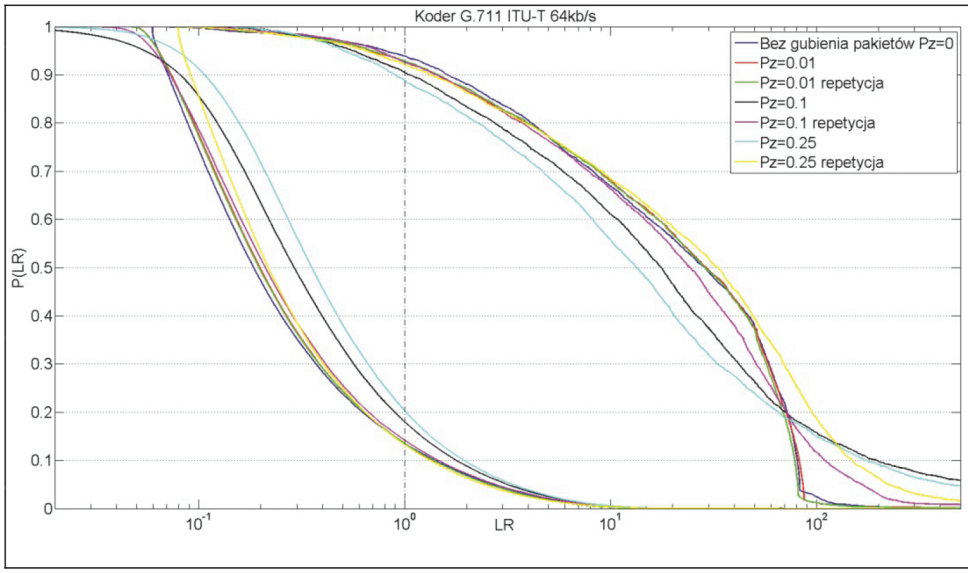
Ryc. 5. Przykład charakterystyki Tippetta.

1) wpływ kompresji dźwięku G.711 ITU-T 64 kb/s, G.729 ITU-T 11,8 kb/s oraz G.723 ITU-T 6,3 kb/s,
2) wpływ utraty pakietów z prawdopodobieństwem $P_z = 0{,}01, 0{,}1, 0{,}25$,
3) wpływ utraty pakietów z prawdopodobieństwem $P_z = 0{,}01, 0{,}1, 0{,}25$ z zastosowaniem repetycji do ukrywania utraty pakietów.

Uzyskane wartości $P_{LRHp<1}$ oraz $P_{LRHd>1}$ wraz z charakterystykami Tippetta przedstawiono na rycinach 6–8.

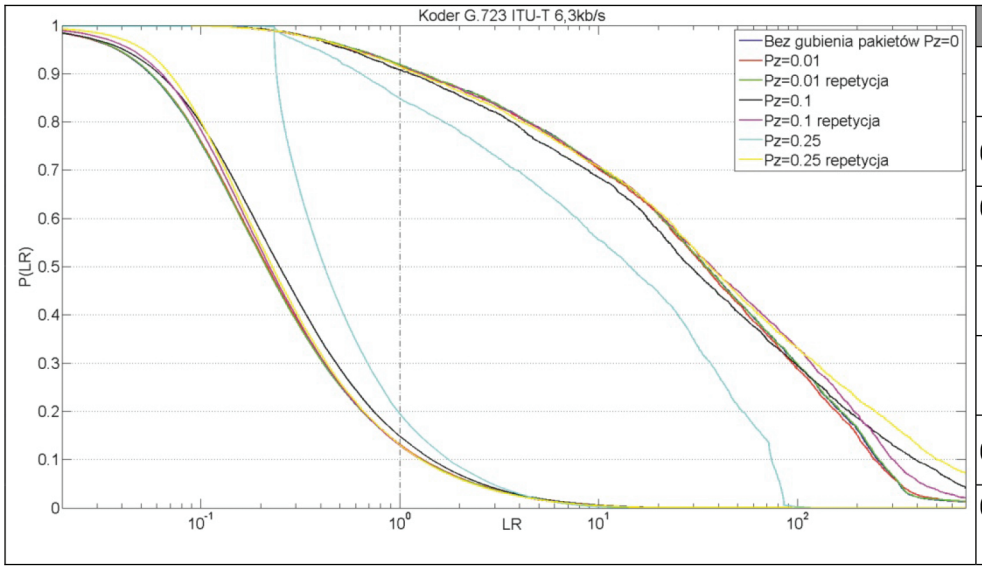
Wnioski

Skuteczność systemu dla różnych metod kompresji bez zjawiska utraty pakietów jest podobna: $P_{LRHd>1}$ od 0,13 do 0,14 oraz dla $P_{LRHp<1}$ od 0,06 do 0,08. Natomiast dla tych samych wypowiedzi zarejestrowanych w warunkach pomieszczenia mieszkalnego otrzymywane są rezultaty: $P_{LRHd>1}$ 0,5 oraz $P_{LRHp<1}$ 0,04 (ryc. 8). Podobne wyniki otrzymano dla tych samych wypowiedzi zarejestrowanych za pośrednictwem telefonii stacjonarnej (PSTN) oraz komórkowej (ryc. 8). Otrzymane charakterystyki pokazują zatem, że kodowanie audio według standardu ITU-T H.323 negatywnie wpływa na skuteczność automatycznej identyfikacji mówców. Zauważyć można również, że



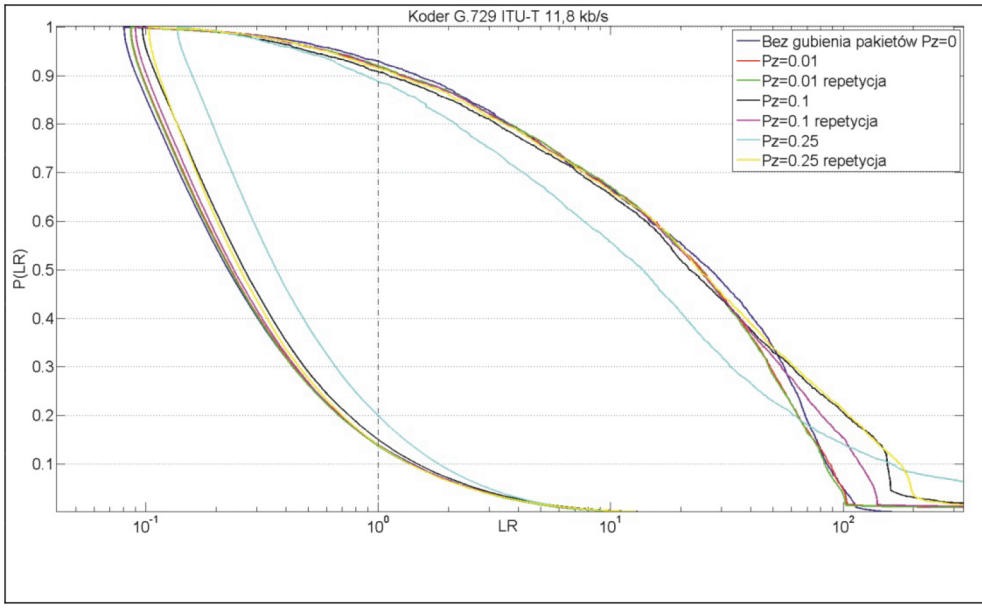
Pz	$P_{LRHd>1}$	$P_{LRHp<1}$
0	0,14	0,06
0,01	0,14	0,07
0,01 PLC	0,14	0,07
0,1	0,18	0,10
0,1 PLC	0,14	0,07
0,25	0,20	0,11
0,25 PLC	0,14	0,08

Ryc. 6. Wpływ transmisji głosu z wykorzystaniem VoIP – kodek audio G.711 64 kb/s na skuteczność automatycznej identyfikacji mówców dla zmiennego stopnia utraty pakietów od 0 do 0,25 oraz dla jednej metody ukrywania utraty pakietów.



Pz	$P_{LRHd>1}$	$P_{LRHp<1}$
0	0,13	0,08
0,01	0,13	0,09
0,01 PLC	0,13	0,09
0,1	0,15	0,09
0,1 PLC	0,13	0,09
0,25	0,19	0,15
0,25 PLC	0,15	0,09

Ryc. 7. Wpływ transmisji głosu z wykorzystaniem VoIP – kodek audio G.723 6,3 kb/s na skuteczność automatycznej identyfikacji mówców dla zmiennego stopnia utraty pakietów od 0 do 0,25 oraz dla jednej metody ukrywania utraty pakietów.



Pz	$P_{LRHd>1}$	$P_{LRHp<1}$
0	0,14	0,07
0,01	0,14	0,08
0,01 PLC	0,14	0,08
0,1	0,15	0,09
0,1 PLC	0,14	0,08
0,25	0,20	0,11
0,25 PLC	0,14	0,08

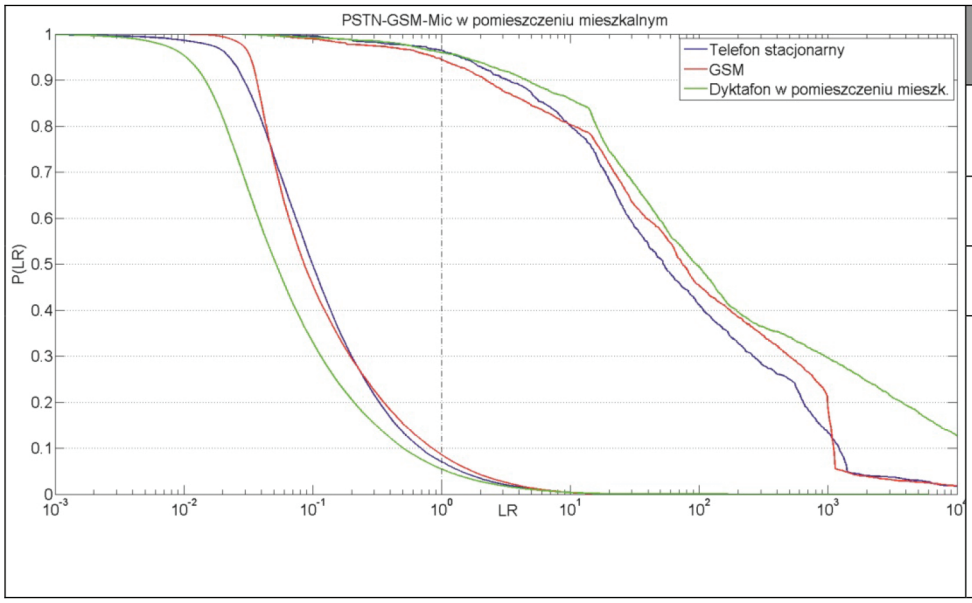
Ryc. 8. Wpływ transmisji głosu z wykorzystaniem VoIP – kodek audio G.729 11,8 kb/s na skuteczność automatycznej identyfikacji mówców dla zmiennego stopnia utraty pakietów od 0 do 0,25 oraz dla jednej metody ukrywania utraty pakietów.

wpływ przepływności poszczególnych kodeków jest pomijalny.

Charakterystyki przedstawione na rycinach 5, 6 oraz 7 wyraźnie pokazują, że decydującym zjawiskiem ujemnie wpływającym na skuteczność systemu identyfikacji mówców jest utrata pakietów, przy czym dla wartości P_z nieprzekraczającej 0,10 ten wpływ jest minimalny. Natomiast dla P_z równej 0,25 prawdpodobieństwa $P_{LRHd>1}$ oraz $P_{LRHp<1}$ wyraźnie rosną i wynoszą odpowiednio od 0,19 do 0,20 oraz od 0,09 do 0,11.

Zastosowanie mechanizmu PLC (ukrywania utraty pakietów) poprawia skuteczność systemu. W przedmiotowej pracy przeanalizowano jeden mechanizm – repetycję. Wynik skuteczności metody pokazuje, że dla P_z równej nawet 0,25 przez zastosowanie mechanizmu repetycji $P_{LRHd>1}$ oraz $P_{LRHp<1}$ maleją odpowiednio od 0,14 do 0,15 oraz od 0,08 do 0,09.

W pracy zbadano wpływ zjawisk powstających w trakcie transmisji sygnału mowy za pośrednictwem telefonii internetowej na skuteczność automatycznego systemu kryminalistycznej identyfikacji mówców opartego na



Pz	$P_{LRHd>1}$	$P_{LRHp<1}$
PSTN	0,07	0,04
GSM	0,09	0,06
Dyk.	0,05	0,04

Ryc. 9. Wpływ transmisji głosu z wykorzystaniem PSTN, GSM oraz rejestracji w warunkach pomieszczenia mieszkalnego na skuteczność automatycznej identyfikacji mówców.

metodzie EM-UBM-MAP. W wyniku przeprowadzonych badań stwierdzono, że decydującym zjawiskiem wpływającym na skuteczność analizowanego systemu jest liczba utraconych pakietów w stosunku do liczby pakietów wysłanych. W praktyce istnieje zatem konieczność oszacowania parametru P_z (liczba ramek utraconych do wszystkich ramek, z których składa się analizowane nagranie) przed przystąpieniem do analizy porównawczej. Wyjątkiem jest przypadek, gdy zastosowano repetycję jako metodę PLC, wtedy efekt utraty pakietów jest pomijalny. Wpływ rodzaju kodeka okazał się podobny dla wszystkich trzech przypadków wykorzystanych w badaniach (G.711 ITU-T 64 kb/s, G.723 ITU-T 6,3 kb/s, G.729 ITU-T 11,8 kb/s). Wniosek ten jest istotny, ponieważ w praktyce, dysponując jedynie nagraniem rozmowy przesłanej z wykorzystaniem telefonii VoIP, nie można ustalić, jakiego rodzaju kodek został wykorzystany.

W świetle przeprowadzonych badań należy stwierdzić, że automatyczny system kryminalistycznej identyfikacji mówców jest skutecznym narzędziem analizy porównawczej głosów, gdy przedmiotem badań są wypowiedzi przesłane z wykorzystaniem telefonii VoIP. Kwestią otwartą pozostają wpływ metod ukrywania utraty pakietów (PLC) przy zmiennej średniej długości paczki błędów (MBLL) oraz wpływ standardów kompresji wykorzystywanych w innych protokołach alternatywnych wobec H.323, np. również popularnym protokole SIP (ang. *Session Initiation Protocol*).

Badania zostały przeprowadzone w ramach projektu badawczego nr 0023/RID3/2012/02 z dnia 30 lipca 2012 roku finansowanego ze środków NCBiR.

Źródła rycin i tabel

Ryciny 1–9: autor
Tabela: opracowanie własne

Bibliografia

1. Myers D.: VoIP services market getting boost from cloud telephony, hosted UC, SIP trunking, „Infonetics Research", 16 kwietnia 2013, [dostęp: 20 sierpnia 2013], http://www.infonetics.com/pr/2013/2H12-VoIP-UC-Services-Market-Highlights.asp
2. Reynolds D.A., Zissman M.A., Quatieri T.F., O'Leary G.C., Carlson B.A.: The effect of telephone transmission degradation on speaker recognition performance, „Acoustics, Speech, and Signal Processing" 1995, ICASSP-95.
3. Reynolds D.A.: The effects of handset variability on speaker recognition performance: experiments on the switchboard corpus, „Acoustics, Speech, and Signal Processing" 1996, ICASSP-96 Conference Proceedings.
4. Byrne C., Foulkes P.: The mobile phone effect on vowel formants, „International Journal of Speech Language and the Law" 2004, nr 1.
5. Besacier L., Grassi S., Dufaux A., Ansorge M., Pellandini F.: GSM speech coding and speaker recognition, ICASSP 2000.
6. Besacier L., Ariyaeeinia A.M., Mason J.S., F.Bonastre J., Mayorga P., Fredouille C., Meignier S., Siau J., Evans N.W.D., Auckenthaler R., Stapert R.: Voice biometrics over the internet in the framework of COST action 275, „EURASIP Journal on Applied Signal Processing" 2004, nr 4, s. 466–479, Hindawi Publishing Corporation.

7. Jajszczyk A.: Wstęp do telekomunikacji, Podręczniki akademickie, Warszawa 1998.
8. Seria H: Audiovisual and multimedia systems. Infrastructure of audiovisual services – systems and terminal equipment for audiovisual services. Packet-based multimedia communications systems, Rekomendacja ITU-T H.323.
9. Gartkiewicz M.: Usługa VoIP – czynniki wpływające na jakość transmitowanego głosu, „Telekomunikacja i Techniki Informacyjne" 2004, nr 1–2.
10. Davidson J., Peters J.: Voice over IP fundamentals, A systematic approach to understanding the basic of voice over IP, „Cisco Press" 2000.
11. ITU-T H.225.0 (12/2009), Seria H: Audiovisual and multimedia systems, Infrastructure of audiovisual services – Transmission multiplexing and synchronization. Call signaling protocols and media stream packetization for packet-based multimedia communication systems, Rekomendacja ITU-T H.255.0.
12. Szigeti T., Hattingh Ch.: Quality of Service Design Overview, „Cisco Press" 2004.
13. Gartkiewicz M.: Usługa VoIP – mechanizmy poprawy jakości, „Telekomunikacja i Techniki Informacyjne" 2004, nr 1–2.
14. ITU-T G.1020 (07/2006), Seria G: Transmission systems and media, digital systems and networks Quality of Service and performance – generic and user-related aspects. Performance parameter definitions for quality of speech and other voiceband applications utilizing IP networks, Rekomendacja ITU-T G.1020.
15. Jelassi S., Rubino G.: A study of artificial speech quality assessor of VoIP calls subject to limited bursty packet losses, „EURASIP Journal on Image and Video Processing" 2011, nr 9.
16. Sanneck H.: Packet loss recovery and control for voice transmission over the internet, Technischen Universität, praca doktorska, Berlin 2000.
17. Wesołowski K.: Podstawy cyfrowych systemów telekomunikacyjnych, WKiŁ, Warszawa 2006.
18. Mohamed S., Rubino G., Varela M.: Performance evaluation of real-time speech through a packet network: a random neural networks-based approach, „Performance Evaluation", nr 57 (2), s. 141–161.
19. Gilbert E.N.: Capacity of a burst-noise channel, „The Bell System Technical Journal", September 1960.
20. Mayorga P., Besacier L., Lamy R., Serignat J-F.: Audio packet loss over IP and speech recognition, Automatic Speech Recognition and Understanding, 2003, ASRU 03.2003 IEEE Workshop on, 30 Nov.–3 Dec. 2003, s. 607–612.
21. ITU-T G.711.0 (09/2009) Seria G: Transmission systems and media, digital systems and networks digital terminal equipments – coding of voice and audio signals. Lossless compression of G.711 pulse code modulation, Rekomendacja ITU-T G.711.0.
22. ITU-T G.729 (03/2012) Seria G: Transmission systems and media, digital systems and networks. Digital terminal equipments – coding of voice and audio signals. Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear prediction (CS-ACELP). Annex E CS-ACELP speech coding algorithm at 11.8 kbit/s, Rekomendacja ITU-T G.729.
23. ITU-T G.723.1 (05/2006), Seria G: Transmission systems and media, digital systems and networks, Digital terminal equipments – coding of analogue signals by methods other than PCM. Dual rate speech coder for multimedia communications transmitting at 5.3 and 6.3 kbit/s, Rekomendacja ITU-T G.723.1.
24. Reynolds D.A., Rose R.C.: Robust text-independent speaker identification using gaussian mixture speaker models, „IEEE Transactions on Speech and Audio Processing" 1995, nr 3 (1), s. 72–83.
25. Makowski R.: Automatyczne rozpoznawanie mowy – wybrane zagadnienia, Oficyna Wydawnicza Politechniki Wrocławskiej, Wrocław 2011.
26. Magrin-Chagnolleau I., Gravier G., Blouet R.: Overview of the 2000-ELISA consortium research activities, A Speaker Odyssey, czerwiec 2001, s. 67–72.
27. Viikki O., Laurila K.: Cepstral domain segmental feature vector normalization for noise robust speech recognition, „Speech Communication" 1998, nr 25, 133–147 Elsevier (1998).
28. Reynolds D.A., Quatieri T.F., Dunn R.B.: Speaker verification using adapted gaussian mixture models, „Digital Signal Processing" 2000, nr 10, s. 19–41.
29. Meuwly D., Drygajlo A.: Forensic speaker recognition based on a bayesian framework and gausian mixture modelling, In 2001, A Speaker Odyssey: The speaker recognition workshop 2001, s. 145–150.
30. Auckenthaler R., Carey M., Lloyd-Thomas H.: Score normalization for text-independent speaker verification system, „Digital Signal Processing" 2000, nr 10 (1–3), s. 42–54.
31. Barras C., Gauvain J.-L.: Feature and score normalization for speaker verification of cellular data. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, t. 2, s. 49–52, Hong Kong, China, 2003.
32. Zadora G., Trzcińska B.M.: The evidential value of transfer evidences – a hit-and-run accident, „Problems of Forensic Sciences" 2008, nr LXXIII, s. 70–81.
33. Gil P., Curron J., Neumann C., Kirkham A., Clayton T., Whitaker J., Lambert J.: Interpretacja złożonych profili DNA za pomocą modeli empirycznych i metoda mierzenia rzetelności tych modeli, „Problemy Kryminalistyki" 2008, nr 261.