Kamil KAMIŃSKI[1], Andrzej P. DOBROWOLSKI[1], Ewelina MAJDA-ZDANCEWICZ[1]

Wojskowa Akademia Techniczna, Wydział Elektroniki (1)

doi:10.15199/48.2015.10.43

Identyfikacja głosowa w otwartym zbiorze mówców

Streszczenie. *W artykule zaprezentowano wyniki badań systemu automatycznego rozpoznawania mówcy, przeprowadzane z wykorzystaniem komercyjnej bazy głosów TIMIT. Głównym celem badań było rozszerzenie funkcjonalności systemu rozpoznawania mówcy poprzez dodanie układu progowego, a tym samym umożliwienie identyfikacji w otwartym zbiorze mówców. Przedstawiono różne warianty zastosowanego układu progowego oraz dokonano próby wzbogacenia wektora cech dystynktywnych o różnicę częstotliwości podstawowej wyznaczanej dwiema różnymi metodami.*

Abstract*. In the article there are presented the test results of the automatic speaker recognition system, conducted while using the commercial voice basis TIMIT. The main purpose of the test was to extend the functionality of the speaker recognition system by adding the threshold based system, and consequently to enable the identification in the open set of speakers. There are presented different application variants of the threshold based system and there is an attempt to enrich the vector of distinctive features with the fundamental frequency difference determined with two different methods. (**Voice identification in the open set of speakers)***

Słowa kluczowe: sygnał mowy, rozpoznawanie mówcy, modele mieszanin gaussowskich, uniwersalny model głosu
Keywords: speech signal, speaker recognition, Gaussian mixtures models, universal background model

Wstęp

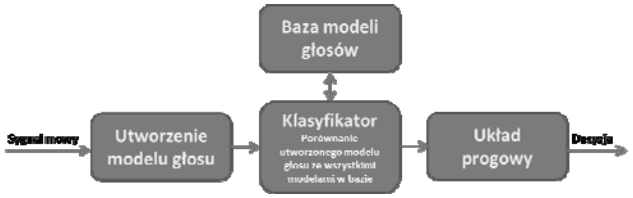
Systemy biometryczne należą do szybko rozwijającej się dziedziny wiedzy, która daje początek rentownym gałęziom przemysłu. Na całym świecie widoczna jest tendencja do uzupełninia, a nawet zastępowania klasycznych metod dostępowych przez systemy biometryczne. Dzieje się tak, gdyż konwencjonalne metody uwierzytelniania i autoryzacji, jak np. kody dostępu, w przeciwieństwie do biometryk, mogą być w łatwy sposób skradzione bądź zgubione. Do najczęściej używanych identyfikatorów biometrycznych należą odcisk palca, obraz twarzy, geometria dłoni, obraz tęczówki, podpis oraz głos [1].

Niniejszy artykuł przedstawia opis zaimplementowanego przez autorów systemu automatycznego rozpoznawania mówcy (*ASR* – ang. *Automatic Speaker Recognition*), który ze względu na swoją specyfikę będzie mógł służyć do identyfikacji rozmówcy telefonicznego. Jednakże rozpatrując systemy *ASR* należy zwrócić uwagę na dwie zasadniczo różniące się procedury: weryfikację oraz identyfikację.

Weryfikacja mówcy stanowi proces decyzyjny pozwalający na określenie czy mówca danej wypowiedzi jest osobą, za którą się podaje. Wynikiem weryfikacji jest potwierdzenie bądź odrzucenie deklarowanej przez mówcę tożsamości, poprzez określenie skali podobieństwa próbki jego głosu z modelem głosu mówcy deklarowanej tożsamości.

Znacznie trudniejszym zadaniem jest identyfikacja głosowa w określonym zbiorze mówców, ponieważ realizuje ona ustalanie tożsamości głosu mówcy wyłącznie na podstawie biometryki. Procedura ta wymaga porównania próbki głosu ze wszystkimi modelami głosów zgromadzonymi w bazie. W następnej kolejności mówca rozpatrywanego nagrania głosowego zostaje utożsamiony z mówcą charakteryzującym się najbardziej zbliżonym modelem głosu.

Odpowiednia fuzja dwóch opisanych powyżej procedur pozwala na utworzenie systemu identyfikacji głosowej w tzw. otwartym zbiorze mówców. W jego pierwszym etapie realizowane jest porównywanie rozpatrywanego fragmentu mowy z modelami głosów z określonej bazy. Kolejnym etapem jest przyrównanie otrzymanego stopnia podobieństwa rozpoznawanego nagrania głosowego z najbardziej podobnym modelem głosu z bazy do określonej wartości progowej. Jeśli progowe kryterium nie zostanie spełnione następuje odrzucenie nagrania głosowego, co może świadczyć o wykryciu głosu nieznajdującego się w bazie lub głosu nadmiernie zniekształconego. Schemat działania takiego systemu został przedstawiony na rys 1.



Rys. 1. Schemat identyfikacji głosowej w otwartym zbiorze mówców.

W artykule przedstawiono kolejno poszczególne etapy działania systemu automatycznego rozpoznawania mówcy zaimplementowanego w środowisku *Matlab*. Ponadto przedstawiono sposób oceny klasyfikacji wykorzystany w testach niniejszego systemu, a także zamieszczono wyniki najnowszych badań prowadzonych przez autorów. Przeprowadzone eksperymenty w głównej mierze dotyczą wyboru optymalnego wariantu układu progowego oraz próby zastosowania różnicy częstotliwości podstawowej wyznaczanej dwiema niezależnymi metodami, jako cechy dystynktywnej.

Baza głosów

W celu zaprojektowania i sprawdzenia skuteczności działania systemu *ASR* niezbędnym elementem jest różnorodna baza głosów. Do tego celu mogą być wykorzystywane autorskie lub komercyjne bazy głosów. Te ostatnie pozwalają na porównanie projektowanego systemu *ASR* z już istniejącymi implementacjami innych autorów. W niniejszych badaniach wykorzystano bazę nagrań *TIMIT* stworzoną przez *MIT* (ang. *Massachusetts Institute of Technology*), *SRI* (ang. *Stanford Research Institute*) oraz *TI* (ang. *Texas Instruments*) [6]. W skład bazy wchodzą nagrania 630 mówców obojga płci, zarejestrowane z szybkością próbkowania 16 kS/s, przy zapisie jednokanałowym z 16-to bitową rozdzielczością amplitudową. Głos każdego z mówców został zapisany w 10 niezależnych nagraniach o długości około 3 s.

Dla potrzeb badań przedstawionych w artykule wykorzystano 100 kolejnych głosów męskich i 100 kolejnych głosów żeńskich. Głosy nie były w żaden sposób dodatkowo dobierane i selekcjonowane. Pozostałe nagrania głosowe posłużyły w procesie tworzenia uniwersalnego

modelu głosu (*UBM* – ang. *Universal Background Model*). Prezentowany system *ASR* został przystosowany do rozpoznawania nagrań głosowych próbkowanych z szybkością 8 kS/s, co pozwala na sprawdzenia działania jego skuteczności w warunkach zbliżonych do transmisji telefonicznej. To założenie wymagało przepróbkowania bazy głosów *TIMIT* do wspomnianej szybkości próbkowania. Przetworzone nagrania zostały w następnej kolejności scalone (niezależnie dla każdego mówcy), a następnie rozdzielone na segmenty uczące o długości 25 s oraz segmenty testowe o długości 5 s, co pozwoliło na zachowanie niezależności pomiędzy uczącymi i testowymi fragmentami nagrania.

Wstępne przetwarzanie sygnałów mowy

W celu zmniejszenia wpływu konfiguracji sprzętu nagrywającego na skuteczność poprawnej identyfikacji mówcy przeprowadzana jest filtracja, normalizacja oraz trzyetapowa selekcja ramek.

Filtracji dokonano przy pomocy filtru pasmowo-przepustowego o skończonej odpowiedzi impulsowej, którego rząd oraz częstotliwość odcięcia poddano wcześniejszej optymalizacji [5]. W następnej kolejności przeprowadzana jest normalizacji sygnału względem jego maksymalnej wartości, co pozwala na zachowanie odpowiednich relacji energetycznych pomiędzy poszczególnymi fragmentami nagrania głosowego. Kolejnym realizowanym procesem jest segmentacja sygnału, która jest tożsama z operacją okienkowania. W prezentowanym systemie zastosowano okno Hamminga, co dodatkowo zminimalizowało przeciek częstotliwości. Długość okna czasowego oraz jego przesunięcie również były poddane wcześniejszym optymalizacjom [5].

Wstępne przetwarzanie sygnału mowy kończy się trójetapową selekcją ramek. W pierwszym etapie eliminowane są długotrwałe ciche fragmenty mowy, które nie spełniają wyznaczonego empirycznie kryterium mocy w ramce. W drugim etapie na podstawie drugiego maksimum funkcji autokorelacji, wyselekcjonowane zostają wyłącznie dźwięczne fragmenty mowy niosące najwięcej informacji o barwie tonu krtaniowego mówcy. W ostatnim etapie dzięki określeniu różnicy częstotliwości podstawowej (Δf_0) wyznaczonej metodą cepstralną i autokorelacyjną następuje usunięcie ramek nadmiernie zaszumionych. Jest to możliwe dzięki zróżnicowanej odporności wspomnianych metod na zaszumienie sygnału [2].

Generacja i selekcja cech dystynktywnych

Ze względu na nierekompensowalność błędów popełnionych na etapie generacji cech przyjmuje się, że stanowi on kluczowy elementy systemu biometrycznego. Sygnał mowy w bezpośredniej reprezentacji czasowej, ze względu na jego wysoką redundancję jest praktycznie niemożliwy do poddania bezpośredniej analizie przez w systemie *ASR*. Dlatego zostaje on przetransformowany do postaci częstotliwościowej, co jednak nadal nie pozwala uwidocznić różnic osobniczych. Powodem tego jest większa wrażliwość widma amplitudowego na zmiany treści zarejestrowanych wypowiedzi niż na zmiany mówców. Dlatego prezentowany system wykonuje dwa dalsze przekształcenia, które pozwalają na uzyskanie tzw. ważonych cech cepstralnych (1) oraz cech melcepstralnych (2) [2].

$$CC(i) = \sum_{k=-dp}^{dp} \mathcal{F}^{-1}\left\{ \log\left(\left| \mathcal{F}\left\{ s(t) \right\} \right| \right) \right\} \tag{1}$$

$$MFCC(i) = DCT\left\{ \log\left(\left| \mathcal{F}\left\{ s(t) \right\} \cdot Mel(i)_{bank} \right| \right) \right\} \tag{2}$$

gdzie: i – indeks cechy (i = 1,…,30), dp – zakres sumowania w obrębie i-tego maksimum cepstrum, $s(t)$ – ramka rzeczywistego sygnału mowy, Mel_{bank} – bank filtrów melowych.

W celu wyznaczenia ważonych cech cepstralnych w pierwszej kolejności należy zmienić multiplikatywny związek pomiędzy składową wolnozmienną wyrażającą artykulację, a składową pochodzącą od pobudzenia krtaniowego na addytywny, poprzez poddanie logarytmowaniu widma amplitudowego. W dalszej kolejności realizowana jest odwrotna transformacja Fouriera zlogarytmowanego widma do dziedziny tzw. pseudoczasu cepstralnego. Taka forma sygnału pozwala w prosty sposób odseparować informacje związane z treścią wypowiedzi znajdujące się w pobliżu zera od impulsów związanych z tonem krtaniowym usytuowanych w okolicach okresu sygnału krtaniowego i powtarzających co ten okres. Ostatnim przekształceniem związanym z wyznaczaniem ważonych cech cepstralnych jest wymnożenie cepstrum przez bank filtrów sumacyjnych pozwalających uwzględnić zarówna maksima amplitud, jak i wartości je okalające, gdyż zawierają one również cenne informacje o głosie mówcy.

W przypadku cech melcepstralnych widmo amplitudowe sygnału zostaje wymnożone przez bank filtrów sumacyjnych, których szerokość i położenie są związane ze skalą melową. Dzięki takiemu przekształceniu możliwe jest odwzorowanie nierównomiernej rozdzielczości częstotliwościowej ucha ludzkiego, wytrenowanego w rozpoznawaniu mowy i mówców przez tysiące lat ewolucji. W następnej kolejności uzyskane cechy poddawane są logarytmowaniu oraz są dekorelowane za pomocą transformacji kosinusowej.

Opisane powyżej dwa zbiory cech podlegały fuzji, a następnie selekcji. Do tego celu zastosowano algorytm genetyczny, który umożliwia wytypowanie optymalnego wektora cech uwzględniając ich synergię. Algorytm genetyczny ze względu na swoją specyfikę pozwala uniknąć lokalnego „*nasycenia*" funkcji przystosowania i jej przedwczesnego zatrzymania, pozwalając tym samym na uzyskanie optymalnego zbioru cech w znaczeniu globalnym. Ponadto funkcja przystosowania zastosowana w algorytmie została tak zdefiniowana, aby wyselekcjonowane cechy były jak najbardziej zależne od głosu mówcy, a jak najmniej od urządzenia nagrywającego oraz od siebie wzajemnie. Ze względu na fakt, że proces selekcji wykonywany był wyłącznie podczas konstruowania i optymalizacji systemu *ASR*, jego szerszy opis został pominięty w niniejszym artykule. Dokładną analizę tego procesu opisują wcześniejsze publikacje autorów, m.in. [7].

W niniejszym artykule dokonano również próby zastosowania różnicy częstotliwości podstawowej (Δf_0) wyznaczanej metodą cepstralną i metodą autokorelacyjną, jako dodatkowej cech dystynktywnej. Wartość ta wykorzystywana była dotychczas wyłącznie do odrzucania nadmiernie zaszumionych ramek sygnału.

Klasyfikator GMM-UBM

Proces klasyfikacji w prezentowanym systemie ASR realizowany jest w oparciu o liniową kombinację rozkładów Gaussa, która wykorzystując dane uczące pozwala na utworzenie oszczędnych pamięciowo, a zarazem zasobnych w informację osobniczą modeli głosów, zwaną modelami mieszanin Gaussowskich (*GMM* – ang. *Gaussian Mixtures Models*) [8].

Wartości początkowe rozkładów, tj. wartości oczekiwane, macierze kowariancji oraz wagi rozkładów mogą być dobierane w sposób pseudolosowy lub zdeterminowany prze algorytm GMM-UBM [4, 6]. Działanie

tego algorytmu polega na utworzeniu tzw. uniwersalnego modelu głosu *UBM*, który tworzony jest na podstawie głosów wielu osób. Dzięki takiemu podejściu nie ma obaw, że powstałe dane startowe będą silnie odstające, co przekłada się na zwiększenie szybkości tworzenia modeli poszczególnych głosów oraz na ostateczną poprawę skuteczności rozpoznawania mówców [6]. Ponadto uniwersalny model głosu może być wykorzystywany w układzie progowym do normalizacji otrzymanych wyników klasyfikacji [12].

Wszystkie wyuczone modele głosów zgromadzone w bazie wykorzystywane są podczas rozpoznawania tożsamości mówcy poprzez porównywanie ich z wyekstrahowanymi cechami pochodzącymi z nagrania testowego. Model, który najlepiej przybliża dane testowe, uznawany jest za głos najbardziej podobny, co pozwala systemowi powiązać go z rozpoznawanym fragmentem nagrania głosowego.

Układ progowy (decyzyjny)

Dotychczasowe badania autorów nie zawierały układu progowego, co powodowało utożsamienie rozpoznawanej wypowiedzi z głosem najbardziej prawdopodobnego mówcy z bazy, niezależnie od stopnia podobieństwa. W wyniku takiego działania nawet osoba nieznajdująca się w bazie głosów zawsze była utożsamiona z jednym z głosów z bazy (tym, który był najbardziej podobny). W celu zaradzenia takiemu niepożądanemu zjawisku wzbogacono system o układ decyzyjny zawierający ustaloną przez użytkownika wartość progową θ pozwalającą określić stopień akceptacji sygnałów głosowych podlegających uwierzytelnianiu przez system *ASR*.

W celu podjęcia decyzji o akceptacji bądź odrzuceniu rozpoznawanego głosu, w pierwszej kolejności, realizowana jest normalizacja wyniku podlegającego porównaniu z progiem θ. Podczas normalizacji następuje zestawienie wiarygodności $p(X \mid \lambda_{hyp})$ świadczącej o tym, że sygnał testowy pochodzi od danego mówcy z alternatywną wiarygodnością $p(X \mid \lambda_{\overline{hyp}})$ mówiącą o tym, że sygnał pochodzi od innego nieznanego mówcy z tej samej populacji (3) [12].

$$\Lambda(X) = \log\left(\frac{p(X \mid \lambda_{hyp})}{p(X \mid \lambda_{\overline{hyp}})}\right) = \log p(X \mid \lambda_{hyp}) - \log p(X \mid \lambda_{\overline{hyp}}) \tag{3}$$

W przeciwieństwie do podejścia tradycyjnego [11, 12], ze względu na wykorzystywane środowisko obliczeniowe (*Matlab*), zastosowany przez autorów klasyfikator poszukuje minimalnej wartości sumy ujemnego logarytmu gęstości prawdopodobieństwa (ang. *negative log-likelihood*).

$$\log p(X \mid \lambda_{hyp}) = \max_{1 \le k \le N} \sum_{t=1}^{T} \log p\left(x_t \mid \lambda_k\right) = \min_{1 \le k \le N} \sum_{t=1}^{T} -\log p\left(x_t \mid \lambda_k\right) \tag{4}$$

gdzie: N – liczba wszystkich głosów w bazie, T – liczba wektorów cech wyekstrahowanych z sygnału testowego. Dlatego ostatecznie, wyciągając znaki przed logarytmy wiarygodności, zależność (3) można zapisać w postaci:

$$\Lambda(X) = -\log p(X \mid \lambda_{hyp}) - (-\log p(X \mid \lambda_{\overline{hyp}})) = \log p(X \mid \lambda_{\overline{hyp}}) - \log p(X \mid \lambda_{hyp}) \tag{5}$$

Określenie wartości $\log p(X \mid \lambda_{\overline{hyp}})$ realizowane jest w systemie na dwa sposoby. Pierwszy z nich wykorzystuje bezpośrednio uniwersalny model głosu *UBM*:

$$\log p(X \mid \lambda_{\overline{hyp}}) = \log p(X \mid \lambda_{UBM}) \tag{6}$$

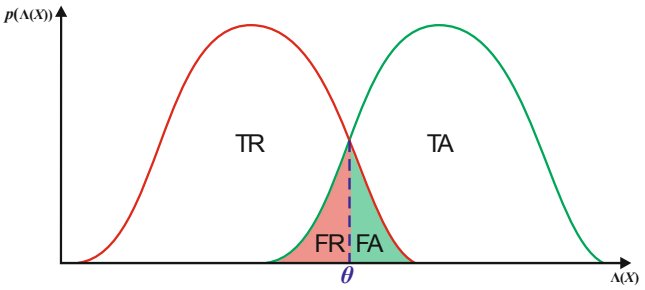
Natomiast drugi sposób uwzględnia średnie wartości logarytmów wiarogodności najbardziej podobnych do sygnału testowego modeli głosów w bazie, jednakże nie uwzględniając w tym zbiorze modelu najbardziej dopasowanego:

$$\log p(X \mid \lambda_{\overline{hyp}}) = \mu_M(X) = \frac{\sum_{k=2}^{M+1} \underset{1 \le k \le N}{\text{sort}} \log p\left(X \mid \lambda_k\right)}{M} \tag{7}$$

Liczba uwzględnionych modeli M podlegała optymalizacji.

Miara jakości układu progowego

W sytuacji idealnej, układ decyzyjny akceptuje mówców znajdujących się w bazie głosów i uzyskujących wyniki $\Lambda(X)$ większe od wartości progowej θ (*TA* – ang. *True Acceptance*), natomiast odrzuca mówców z niższymi wynikami, którzy są spoza bazy (*TR* – ang. *True Rejection*). W rzeczywistości może jednak dojść do jeszcze dwóch zdarzeń. Pierwszym z nich jest akceptacja osoby niepowołanej (*FA* – ang. *False Acceptance*), której głos jest dla systemu zbliżony do jednego z zarejestrowanych w bazie. Natomiast druga sytuacja powoduje odrzucenie upoważnionego użytkownika systemu pomimo obecności jego głosu w bazie (*FR* – ang. *False Rejection*). Na rys. 2 przedstawiono dwa rozkłady gęstości prawdopodobieństwa uzyskiwanych przez mówców wyników $\Lambda(X)$ podlegających weryfikacji w układzie progowym. Zielony rozkład reprezentuje sytuację zgodności rozpoznawanego fragmentu nagrania głosowego z modelem w bazie, natomiast czerwony zdarzenie, w którym głos wypowiadającej się osoby nie znajduje się w bazie głosów. Niebieską przerywaną linią zaznaczony jest próg θ.



Rys. 2. Rozkłady gęstości prawdopodobieństwa uzyskiwanych przez mówców wyników $\Lambda(X)$ podlegających weryfikacji w układzie progowym.

Istnieje wiele sposobów określania jakości detekcji realizowanej przez układ decyzyjny. Do podstawowych należy określenie relacji pomiędzy czułością klasyfikatora (*TAR* – ang. *True Acceptance Rate*) (8), a wskaźnikiem fałszywych akceptacji (*FAR* – ang. *False Acceptance Rate*) (9). Graficzną reprezentację tej zależności stanowi tzw. krzywa cech eksploatacyjnych narzędzia (*ROC* – ang. *Receiver Operating Characteristics*), przedstawiona na rys. 3.a) [1, 10] Dzięki tej charakterystyce możliwe jest porównanie klasyfikatorów (np. K_1 i K_2) między sobą oraz z klasyfikatorem losowym (K_r). Jakość klasyfikatora jest tym wyższa im krzywa repetująca dany klasyfikator znajduje się bliżej lewego górnego rogu wykresu, a pole pod nią jest większe. Ponadto pole powierzchni pod krzywą (*AUC* – ang. *Area Under Curve*) jest często wykorzystywane do liczbowej oceny klasyfikatora.

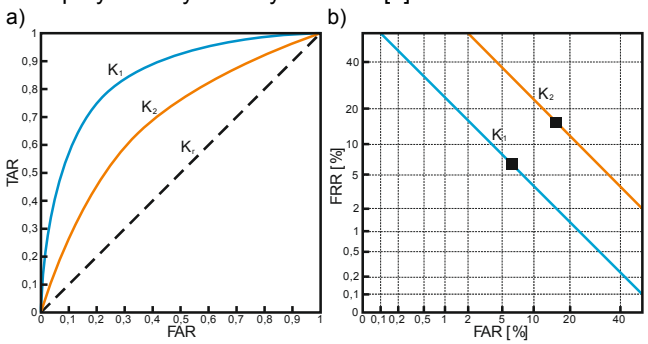
(8) $$TAR = \frac{TA}{TA + FR}$$

(9) $$FAR = \frac{FA}{FA + TR}$$

Kolejnym istotnym kryterium porównawczym klasyfikatorów jest zależność błędu II rodzaju, jakim jest wskaźnik fałszywych odrzuceń (*FRR* – ang. *False Rejection Rate*) (10), względem wskaźnika fałszywych akceptacji FAR, który stanowi błąd I rodzaju.

(10) $$FRR = \frac{FR}{FR + TA}$$

Zobrazowaniem tej funkcji jest tzw. krzywa kompromisu błędów wykrywania (*DET* – ang. *Detection Error Trade-off*) przedstawiona na rys. 3.b). Z punktu widzenia użytkownika system *ASR* można ją odczytywać jako zależność niewygody użytkowania systemu związanej z odrzucaniem głosów zarejestrowanych wcześniej w bazie względem niebezpieczeństwa związanego z nieupoważnioną akceptacją intruza. Punkt, w którym dla danego klasyfikatora wartości *FRR* i *FAR* są sobie równe nazywany jest stopą błędu zrównoważonego (*EER* – ang. *Equal Error Rate*). Pozwala on na określenie jakości klasyfikatora za pomocą jednej wartości. Jakość systemu jest tym wyższa im niższa jest wartość *EER*. Na rysunku 3 b) za pomocą kwadratów zaznaczono punkty reprezentujące wartości *EER* przykładowych klasyfikatorów [3].



Rys. 3. Zobrazowanie jakości przykładowych klasyfikatorów za pomocą a) krzywej *ROC* oraz b) krzywej *DET*

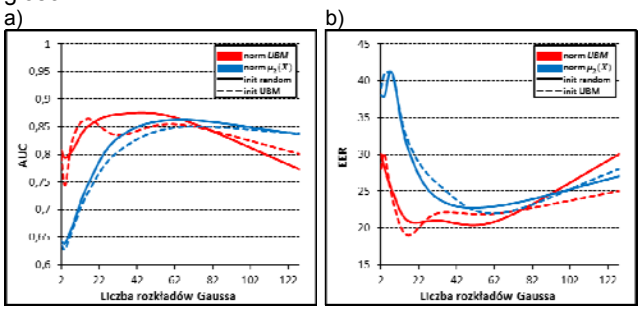
Wyniki eksperymentów

Przedstawione badania w głównej mierze skupione były na wyborze optymalnego wariantu systemu *ASR*, zapewniającego satysfakcjonujące wyniki działania układu decyzyjnego. W celu przetestowania tego układu wykorzystano 200 głosów (100 żeńskich i 100 męskich) z bazy *TIMIT*, a następnie dokonano rozdziału ich na dwie części zawierające każda po 100 głosów (50 żeńskich i 50 męskich). Pierwsza część głosów posłużyła do utworzenia modeli głosów i nagrań testowych dla wszystkich 100 mówców. Natomiast z drugiej części głosów utworzono wyłącznie nagrania testowe, który miały sprawdzić zdolność systemu do odrzucania mówców nieznanych względem zamkniętej bazy modeli głosów z pierwszej setki nagrań. Podczas badań porównywano dwa sposoby normalizacji wyników logarytmów wiarygodności (5) uzyskiwanych przez poszczególnych mówców.

Pierwszy z nich (czerwone linie na rys. 4.) wykorzystywał do tego celu uniwersalny model głosu wg wzoru (6). Podczas badań dokonano próby tworzenia modelu *UBM* z różnej liczby głosów, tj. 430 pozostałych głosów niebiorących udziału w badaniu, 100 głosów (50 żeńskich oraz 50 męskich) oraz kilku wariantów mniejszej grupy głosów. Ostatecznie ze względu na najwyższe uzyskiwane rezultaty do utworzenia modelu głosu uniwersalnego wybrano 100 kolejnych głosów niebiorących udziału w badaniu (50 żeńskich i 50 męskich). Taki wybór umożliwia utworzenie modelu *UBM* „*nieodstrojonego*" w kierunku którejkolwiek płci, ze względu na obecność równolicznych podgrup męskich i żeńskich. Możliwe jest również utworzenie dwóch modeli *UBM* zależnie od płci. Wątek ten nie został opisany w niniejszym artykule, jednak był testowany przez autorów we wcześniejszych badaniach [6].

Drugi sposób (niebieskie linie na rys. 4.) wyznaczania składowej normującej wyniki, realizowany był wg wzoru (7). Uwzględnia on średnią wartość logarytmów wiarygodności $\mu_M(X)$ uzyskiwanych w grupie najbardziej prawdopodobnych modeli, bez uwzględnienia modelu najbardziej dopasowanego. Liczba modeli M, których wyniki podlegały uśrednieniu została zoptymalizowana. W rezultacie zdecydowano, aby do uśredniania wykorzystywać 2 najbardziej prawdopodobne modele głosów, bez uwzględniania modelu najbardziej dopasowanego.

Porównywanie składowych normujących uzyskiwane rezultaty realizowane był z uwzględnieniem różnej liczby rozkładów Gaussa. Ze względu na mnogość otrzymywanych wyników nie przedstawiono krzywych *ROC* i *DET* dla tego zestawienia. Posłużono się natomiast wartościami *AUC* i *EER* bezpośrednio wynikającymi z tych krzywych (rys. 4.). Ponadto porównano wpływ doboru wartości początkowych (wartości losowe – linie ciągłe i wartości zdeterminowane prze model *UBM* – linie przerywane) podczas generowania modeli poszczególnych głosów.

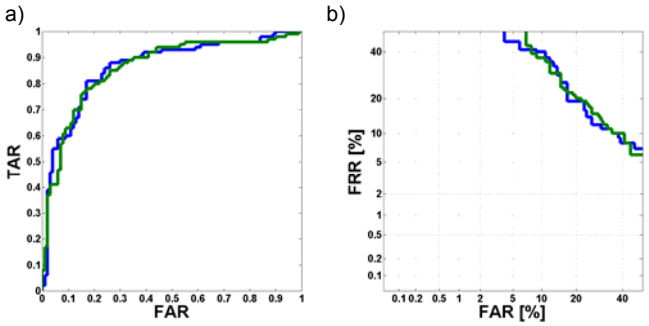


Rys. 4. Porównanie wariantów systemu z wykorzystaniem wartości a) *AUC* i b) *EER* w funkcji liczby rozkładów Gaussa

Z uzyskanych wyników widać, że układ decyzyjny oparty o normowanie wyników średnią rezultatów $\mu_2(X)$ dla dwóch najbardziej prawdopodobnych modeli głosów radzi sobie znacznie lepiej niż układ uwzględniający normowanie wyników za pomocą modelu głosu uniwersalnego. Ponadto stosowanie *UBM* jako elementu inicjalizującego wartości początkowe podczas generowania modeli głosów pozwala tylko, w niektórych przypadkach nieznacznie poprawić uzyskiwane rezultaty. Na podstawie otrzymanych wyników (rys. 4.) wybrano liczbę 16 rozkładów Gaussa jako wartość optymalną pozwalającą uzyska najlepsze wyniki przy jednoczesnej oszczędności pamięciowej względem stosowania większej liczby rozkładów Gaussa

Ostatnim przeprowadzonym przez autorów badaniem było sprawdzenie zasadności stosowania różnicy częstotliwości podstawowej Δf_0 wyznaczanej dwiema metodami tj. autokorelacyjną i cepstralną, jako cechy dystynktywnej. Do tego celu posłużono się krzywymi *ROC* (rys. 5 a) i *DET* (rys. 5 b). Doświadczenie przeprowadzono dla najlepszego wariantu systemu uzyskanego z poprzedniego badania tzn. dla normowania wyników średnią $\mu_2(X)$, z uwzględnieniem *UBM* wyłącznie jako danych inicjalizujących generowanie poszczególnych modeli głosów. Modele głosów utworzono za pomocą 16 rozkładów Gaussa.

Przeprowadzono również badanie wpływu stosowania Δf_0 na skuteczność właściwej identyfikacji poszczególnych mówców. W celu zwiększenia populacji podlegającej badaniu wykonano modele wszystkich 200 głosów biorących udział w badaniu, a nie jak dotychczas 100 głosów. Takie podejście pokazuje wyniki identyfikacji w zamkniętym zbiorze mówców, jednakże ze względu na powiększenie populacji modeli głosów pozwala na zwiększenie wiarygodności otrzymywanych wyników oraz umożliwia porównanie otrzymanych rezultatów z licznymi wcześniejszymi badaniami autorów w zamkniętym zbiorze mówców.



Rys. 5. Porównanie wariantów systemu (wektor cech z Δf_0 – linia zielona i bez Δf_0 – linia niebieska) z wykorzystaniem wartości a) *AUC* i b) *EER* w funkcji liczby rozkładów Gaussa

Tab. 1. Wyniki skuteczności rozpoznawania mówców w zależności od wariantu zastosowanych cech dystynktywnych i liczebności populacji

Wariant cech	Skuteczność rozpoznawania [%]	
	100 głosów (50 żeńskich i 50 męskich)	200 głosów (100 żeńskich i 100 męskich)
Zbiór cech bez Δf_0	**92**	**85**
Zbiór cech z Δf_0	**91**	**87,5**

Wyniki zamieszczone na rys. 5. i w tab. 1. pokazują, że trudno jest jednoznacznie określić zasadność stosowania Δf_0, jako cechy dystynktywnej. W przypadku identyfikacji w zbiorze zamkniętym cecha ta wpływa na poprawę otrzymywanych rezultatów w zbiorze 200 głosów, jednakże w przypadku ograniczenia bazy głosów powoduje pogorszenie wyników. W przypadku układu decyzyjnego sytuacja jest podobna, gdyż tylko dla niektórych ustawień progu θ możliwe jest uzyskanie lepszych rezultatów stosując Δf_0. Z fizycznego punktu widzenia cecha ta niesie głównie informację o stopniu zaszumienia ramki głosu i jej korelację z konkretnymi głosami mówców można uznać za znikomą, natomiast korelacja ze środowiskiem oraz torem akustycznym konkretnych sesji nagraniowych może być bardzo duża. Reasumując stosowanie Δf_0 jako dodatkowej cechy dystynktywnej nie przyniosło zakładanych rezultatów, dlatego nie będzie ona wykorzystywana do tego celu w dalszych badaniach autorów.

Przedstawione w artykule badania pozwoliły na wzbogacenie dotychczasowej realizacji systemu o skuteczny układ decyzyjny, co znacznie rozszerza skalę jego zastosowań. Po dokonaniu optymalizacji układu progowego dla zastosowanej bazy głosów uzyskano wartości *AUC* = 0,87 oraz *EER* = 17, co jest wynikiem satysfakcjonującym. Dalsze badania autorów będą wymagały sprawdzenia jakości układu progowego na innych komercyjnych bazach głosów.

LITERATURA

[1] Bolle R. M., Connell J. H., Pankanti S., Ratha N. K., Senior A. W., TAO Biometria, *WNT*, (2008)

[2] Dobrowolski A. P., Majda E., Cepstral analysis in the speakers recognition systems, *15th IEEE SPA Conference*, (2011), 85-90

[3] J. P. Cambell., Speaker Recognition: A Tutorial, *IEEE*, 85 (2007), nr 9, pp. 1437-1462

[4] Janicki, A., Staroszczyk, T., Klasyfikacja mówców oparta na modelowaniu GMM-UBM dla mowy o różnej jakości, *Krajowe Sympozjum Telekomunikacji i Teleinformatyki*, (2011)

[5] Kamiński K., Majda E., Dobrowolski A. P., Automatic speaker recognition using Gaussian Mixture Models, *17th IEEE SPA Conference,* (2013), 220-225

[6] Kamiński K., Dobrowolski A. P., E. Majda, Ocena funkcjonalności systemu rozpoznawania mówcy dla zdegradowanej jakości sygnału głosowego, *Przegląd Elektrotechniczny*, 90 (2014) nr 8, 164-167

[7] Kamiński K., Dobrowolski A. P., E. Majda-Zdancewicz, Posiadała D., Optymalizacja systemu automatycznego rozpoznawania mówcy w warunkach zróżnicowanych torów akustycznych, *XIV Krajowa Konferencja Elektroniki (KKE'15)*, 2015

[8] Kamiński K., Wojtuń J., Piotrowski Z., Subscriber authentication using GMM and TMS320C6713DSP, *Przegląd Elektrotechniczny*, 88 (2012) nr 12a, 127-130

[9] Kamiński K., Dobrowolski A. P., System automatycznego rozpoznawania mówcy z wykorzystaniem techniki cepstralnej i modeli mieszanin gaussowskich, *Przegląd Elektrotechniczny*, 89 (2013) nr 9, 87-93

[10] Osowski S., Metody i narzędzia eksploracji danych, *BTC*, (2013)

[11] Reynolds, D. A., Gaussian Mixture Models, *Encyclopedia of Biometric Recognition*, Springer, (2008)

[12] Reynolds, D. A., Quatieri, T. F., Dunn, R. B., Speaker Verification Using Adapted Gaussian Mixture Models, *Digital Signal Processing*, 10 (2000), 19-41

***Autorzy**: mgr inż. Kamil Kamiński, dr hab. inż. Andrzej P. Dobrowolski, dr inż. Ewelina Majda-Zdancewicz, Wojskowa Akademia Techniczna, Wydział Elektroniki, ul. gen. S. Kaliskiego 2, 00-908 Warszawa,*
E-mail: kamil.kaminski@wat.edu.pl, adobrowolski@wat.edu.pl, emajda@wat.edu.pl