Inżynieria Bezpieczeństwa
Obiektów Antropogenicznych

SELEKCJA CECH OSOBNICZYCH SYGNAŁU MOWY Z WYKORZYSTANIEM ALGORYTMÓW GENETYCZNYCH

Kamil KAMIŃSKI, Andrzej P. DOBROWOLSKI, Ewelina MAJDA
Wojskowa Akademia Techniczna, Warszawa

Streszczenie

W referacie przedstawiono system automatycznego rozpoznawania mówcy zaimplementowany w środowisku *Matlab* oraz pokazano sposoby realizacji i optymalizacji poszczególnych elementów tego systemu. Główny nacisk położono na wyselekcjonowanie cech dystynktywnych głosu mówcy z wykorzystaniem algorytmu genetycznego, który pozwala na uwzględnienie synergii cech podczas selekcji. Pokazano również wyniki optymalizacji wybranych elementów klasyfikatora, m.in. liczby rozkładów Gaussa użytych do zamodelowania każdego z głosów. Ponadto, podczas tworzenia modeli głosów zastosowano model głosu uniwersalnego.
Słowa kluczowe: mowa, głos, system automatycznego rozpoznawania mowy, przetwarzanie sygnałów.

Abstract

The paper presents automatic speaker recognition system, implemented in the Matlab environment, and demonstrates how to achieve and optimize various elements of the system. The main emphasis was put on features selection of speech signal using a genetic algorithm, which takes into account synergy of features. The results of the selected elements of optimizing classifier have been also shown, including the number of Gaussian distributions used to model each of the voices. In addition during creating voice models, the universal voice model have been used.
Key words: speech, voice, automatic speech recognition system, signal processing.

1. WSTĘP

Ze względu na dużą redundancję przebiegu czasowego sygnału mowy, w systemach automatycznego rozpoznawania mówcy oraz mowy (*ASR* – ang. *Automatic Speaker/Speech Recognition*) nie wykorzystuje się bezpośrednio sygnału w dziedzinie czasu. Projektanci systemów ASR poprzez złożone operacje na sygnałach mowy wydobywają z nich zestaw najistotniejszych – pod kątem określonego kryterium – cech dystynktywnych (deskryptorów). Otrzymane cechy tworzą niejednokrotnie liczny zbiór, a jednocześnie nie każda cecha skutecznie reprezentuje analizowany proces [1]. Stąd potrzeba wyselekcjonowania deskryptorów o największej zdolności dyskryminacyjnej. Istnieje wiele metod selekcji cech diagnostycznych, jednakże nie wszystkie z nich uwzględniają synergię cech, która pozwala uzyskać większą zdolność dyskryminacji dla cech występujących w grupie aniżeli dla poszczególnych cech z osobna. Niestety synergia cech w istotny sposób utrudnia ich selekcję, gdyż dobór najistotniejszych deskryptorów na podstawie indywidualnej oceny każdego z nich, może prowadzić do uzyskania nieoptymalnego zbioru.

Prezentowana w artykule metoda selekcji cech z wykorzystaniem algorytmu genetycznego, nie jest obarczona tą niedogodnością, gdyż w wyniku działania operacji genetycznych na pełnym zbiorze cech dystynktywnych możliwe jest otrzymanie optymalnego zbioru deskryptorów. W celu dokonania selekcji cech z wykorzystaniem algorytmu genetycznego autorzy posłużyli się dwiema niezależnymi komercyjnymi bazami głosów, dzięki czemu stało się możliwe sprawdzenie poprawności działania zastosowanego algorytmu selekcji cech oraz wyodrębnienie zbioru takich cech, które pozwalają na poprawną identyfikację mówcy niezależnie od zastosowanej bazy głosów. W kolejnych rozdziałach artykułu przedstawiono architekturę oraz sposób działania zaimplementowanego przez autorów systemu automatycznego rozpoznawania mówcy oraz wyniki badań przeprowadzonej selekcji cech i jej wpływu na wypadkową skuteczność identyfikacji mówców. W artykule znajdują się również wyniki optymalizacji zastosowanego w systemie klasyfikatora wykorzystującego tzw. modele mieszanin Gaussowskich *GMMs* (ang. *Gaussian Mixture Models*) oraz uniwersalny model głosów *UBM* (ang. *Universal Background Model*).

2. BAZA GŁOSÓW

Podczas prowadzonych badań autorzy wykorzystywali dwie powszechnie znane bazy głosów dedykowane do badań związanych z rozpoznawaniem mowy i mówcy. Pierwszą z nich jest baza nagrań *TIMIT* stworzona przez *MIT* (ang. *Massachusetts Institute of Technology*), *SRI* (ang. *Stanford Research Institute*) oraz *TI* (ang. *Texas Instruments*) [2]. Baza zawiera nagrania głosowe 192 kobiet i 438 mężczyzn, zarejestrowane z szybkością próbkowania 16 kS/s, przy zapisie jednokanałowym z 16-to bitową rozdzielczością amplitudową. Każdy z mówców reprezentowany jest przez 30 sekundowe nagranie utworzone przy użyciu 10 niezależnych 3 sekundowych nagrań głosowych. W celu równomiernego wykorzystania głosów męskich i żeńskich dla potrzeb prezentowanych badań wykorzystano 200 głosów kolejnych mówców. Użyte głosy nie były w żaden sposób dobierane, wykorzystano 100 kolejnych męskich i 100 kolejnych żeńskich głosów z listy głosów dostępnych w bazie *TIMIT*. Nagrania wykorzystane w prezentowanych badaniach zostały przepróbkowane do szybkości próbkowania 8 kS/S. Pozwala to na sprawdzenie skuteczności systemu w warunkach zbliżonych do transmisji telefonicznej, co poszerza spektrum zastosowań prezentowanego systemu.

Kolejną bazą głosów użytą podczas badań była baza *2002 NIST Speaker Recognition Evaluation*. Jest to baza dedykowana dla systemów rozpoznawania mówcy niezależnych od tekstu. Baza obejmuje ponad 156 godzin nagrań, rejestrowanych z szybkością próbkowania 8 kS/s oraz 16 kS/s [3]. Na nagrania składają się zarówno wypowiedzi pojedynczych mówców, jak również dialogi nagrywane w różnych warunkach. Dla potrzeb prowadzonych badań wykorzystano nagrania pojedynczych mówców rejestrowane z szybkością próbkowania 8 kS/s. W bazie znajduje się 330 nagrań spełniających powyższe kryterium (191 głosów żeński i 139 głosów męskich). Podobnie jak w przypadku bazy *TIMIT* dla zrównoważenia liczby głosów męskich i żeńskich wykorzystano 200 głosów kolejnych mówców (100 głosów żeńskich i 100 męskich). Nagrania występujące w bazie *2002 NIST Speaker Recognition Evaluation* gromadzone są w formacie **.sph* (*sphere format*), co powoduje konieczność właściwego wczytania plików lub ich konwersji. Autorzy do tego celu wykorzystali dostępny w środowisku *Matlab* pakiet narzędzi *VOICEBOX* przeznaczonych do przetwarzania sygnału mowy [4].

W celu wyodrębnienia części uczącej i części testowej nagrań zawartych w bazach *TIMIT* oraz *2002 NIST Speaker Recognition Evaluation* zdecydowano, aby przeznaczyć na część uczącą 25 s każdego z nagrań, a na część testową pozostałe niezależne 5 s

nagrań. W przypadku bazy *TIMIT* 10 nagrań głosowych reprezentatywnych dla każdego z mówców zostało scalone, a następnie podzielone na dwie części (25 s i 5 s). Natomiast w bazie *2002 NIST Speaker Recognition Evaluation* zgromadzone nagrania przypadające na jednego mówcę były znacznie dłuższe, co pozwoliło na wycięcie z każdego z nich 25 s segmentu uczącego i 5 s niezależnego segmentu testowego.

3. WSTĘPNE PRZETWARZANIE SYGNAŁÓW MOWY

Pierwszym i bardzo ważnym etapem działania systemu automatycznego rozpoznawania mówcy jest proces wstępnego przetwarzania sygnału, pozwalający na uniezależnienie zgromadzonych nagań od ustawień sprzętu nagrywającego. W ramach tego etapu realizowana jest filtracja, normalizacja oraz selekcja ramek. Filtracja przeprowadzona została poprzez zastosowanie filtru pasmowo-przepustowego o skończonej odpowiedzi impulsowej, którego częstotliwość odcięcia oraz rząd poddano wcześniejszej optymalizacji [5].

Poziomy mocy sygnału mowy u różnych mówców są bardzo zróżnicowane, dlatego po filtracji przeprowadzono normalizację zgromadzonych sygnałów w stosunku do wartości maksymalnej dla każdego mówcy. Dzięki temu skalowaniu zachowane zostały relacje energetyczne pomiędzy poszczególnymi fragmentami zapisu.

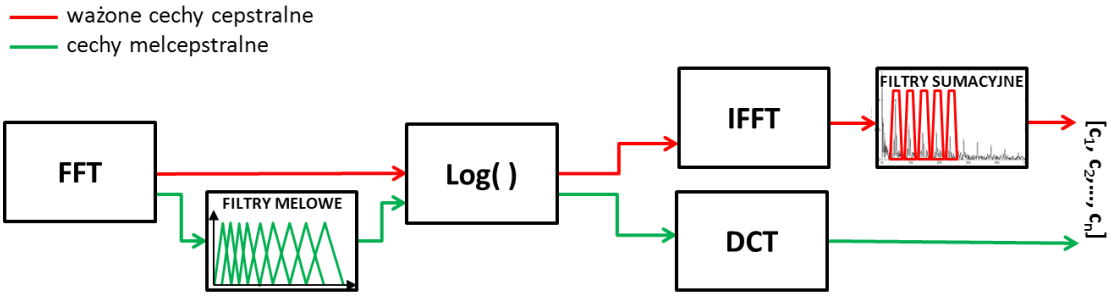
W następnej kolejności realizowana była segmentacja sygnału mowy, która jest tożsama z operacją okienkowania. W prezentowanym systemie zastosowano okno Hamminga, którego widmo charakteryzuje się niskim poziomem listków bocznych, a tym samym znacznie redukuje zjawisko przecieku widma. Długość okna oraz przesunięcie również poddawane były procesowi optymalizacji w trakcie wcześniejszych badań.

Algorytm selekcji ramek realizowany jest trzyetapowo. W pierwszej kolejności eliminowane są ramki niespełniające wyznaczonego empirycznie kryterium mocy w ramce, stanowiącego iloczyn mocy najcichszej ramki oraz stałej wyznaczonej w procesie optymalizacji. Pozwala to na wyeliminowanie ciszy z analizowanych fragmentów mowy. W drugim etapie selekcjonowane są wyłącznie ramki zawierające głoski dźwięczne, ze względu na fakt, iż w nich zawarta jest większość informacji o głosie mówcy istotna z punktu widzenia jego identyfikacji. Klasyfikacji na ramki zawierające głoski dźwięczne i bezdźwięczne dokonuje się na podstawie funkcji autokorelacji. Pierwsze maksimum funkcji autokorelacji występujące dla zerowego przesunięcia niesie informację o energii ramki, natomiast drugie pozwala określić dźwięczność ramki. Wartość drugiego maksimum funkcji autokorelacji w ramce porównywana jest z wyznaczonym w procesie optymalizacji progiem dźwięczności.

Ostatni etap eliminacji ramek polega na porównywaniu wartości częstotliwości podstawowej wyznaczonej dwiema metodami. Operacja wyznaczania częstotliwości podstawowej metodą cepstralną jest mniej dokładna, ale bardziej niezawodna niż metoda autokorelacji, szczególnie przy silnym zaszumieniu. Dlatego porównanie tych wartości pozwala na określenie stopnia zaszumienia ramki sygnału mowy. Jeżeli różnica częstotliwości podstawowej wyznaczonej dwiema niezależnymi metodami będzie większa od wartości progowej wyznaczonej w procesie optymalizacji to ramka zostanie odrzucona, ze względu na zbytnie zaszumienie sygnału, niekorzystnie wpływające na ostateczny wynik działania systemu ASR.

4. GENERACJA DYSTYNKTYWNYCH CECH GŁOSU

Generacja cech jest kluczowym elementem systemu automatycznego rozpoznawania mówcy. Błędy popełnione na tym etapie są już nie do nadrobienia w dalszych etapach działania systemu. Prezentowany w artykule system ASR wykorzystuje dwa rodzaje dystynktywnych cech głosu. Są nimi ważone cechy cepstralne i cechy melcepstralne [6]. Sposób ich generacji przedstawiono na rys. 1.



Rys. 1. Schemat procesu generacji cech

Obie zaprezentowane techniki generacji cech w pierwszej fazie działania realizują transformację sygnału do dziedziny częstotliwości, gdyż analiza sygnału czasowego obarczona jest zbyt dużą redundancją (proces ten jest inspirowany działaniem ludzkich organów komunikacji głosowej). Otrzymane widmo amplitudowe znacznie bardziej uwydatnia różnice treści nagrywanych wypowiedzi niż osobnicze atrybuty związane m.in. z tonem krtaniowym. Dlatego konieczne są dalsze operacje na sygnale pozwalające wydobyć osobnicze cechy głosu mówcy.

W przypadku generacji ważonych cech cepstralnych kolejnym procesem jest logarytmowanie widma amplitudowego, dzięki któremu multiplikatywny związek między składową wolnozmienną i amplitudami poszczególnych impulsów pochodzących od pobudzenia zamienia się na związek addytywny. Poddanie takiego sygnału odwrotnej transformacji Fouriera powoduje, że wolnozmienne przebiegi związane z transmitancją traktu głosowego zostają usytuowane blisko zera na osi czasu cepstralnego zwanego pseudoczasem, natomiast impulsy związane z dźwiękiem krtaniowym zaczynają się mniej więcej w obrębie okresu sygnału krtaniowego i powtarzają się co ten okres. Ostatnim etapem generacji ważonych cech cepstralnych jest wymnożenie otrzymanego sygnału w dziedzinie pseudoczasu przez bank filtrów sumacyjnych, które uwzględniają nie tylko maksymalne amplitudy prążków w cepstrum, ale także wartości je otaczające, które również niosą informację osobniczą o głosie mówcy.

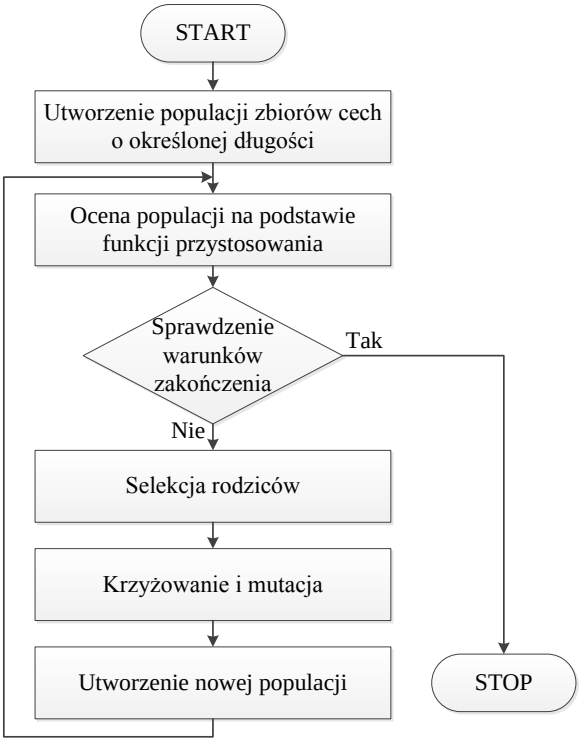
Podczas generacji cech melcesptralnych otrzymane po transformacji amplitudowe widmo Fouriera wymnażane jest przez bank filtrów melowych, które naśladując ludzki organ słuchu i jego nieliniową wrażliwość na pobudzenia z różnych zakresów częstotliwości, powodują poprawę percepcji. Następnie przetworzony sygnał jest logarytmowany, podobnie jak w przypadku cech cepstralnych. Ostatnim etapem generacji cech melcepstralnych jest poddanie sygnału transformacji cosinusowej w celu dekorelacji cech.

5. SELEKCJA CECH DYSTYNKTYWNYCH Z WYKORZYSTANIEM ALGORYTMU GENETYCZNEGO

Na etapie generacji cech wygenerowano maksymalnie liczny zbiór cech dystynktywnych, które mogą być wykorzystane w systemie automatycznego rozpoznawania mówcy. Prowadzone na świecie badania dowodzą, że nie zawsze użycie maksymalnego zestawu cech pozwala uzyskać najlepsze wyniki. Selekcja cech daje niejednokrotnie możliwość otrzymania większej lub takiej samej dokładności klasyfikacji dla zredukowanego wektora cech, co z kolei przekłada się na skrócenie czasu obliczeń.

Podczas oceny jakości cech niektóre z nich mogą mieć postać szumu pomiarowego pogarszającego możliwość rozpoznawania danego wzorca, natomiast inne mogą być ze sobą silnie skorelowane, co powoduje dominację tych cech nad pozostałymi i zwykle niekorzystnie wpływa na jakości klasyfikacji.

Istotnym elementem jest wybór metody selekcji cech. W literaturze dostępnych jest wiele różnych metod selekcji poczynając od szybkich metod rankingowych, a kończąc na czasochłonnych zawierających złożone klasyfikatory. Do najbardziej znanych miar jakości i metod selekcji cech należą: *współczynnik Fishera*, *t-statystyki*, *korelacja wzajemna*, *sekwencyjna selekcja w przód*, *algorytmy genetyczne*, *liniowa analiza dyskryminacyjna*.



Rys. 2. Schemat blokowy selekcji cech przy użyciu algorytmu genetycznego

Autorzy w prezentowanym systemie zastosowali algorytm genetyczny do selekcji najistotniejszych z punktu widzenia systemu ASR cech osobniczych. Metoda ta uwzględnia synergię cech i umożliwia otrzymanie optymalnego zbioru cech, jednakże jest dość czasochłonna. Zasada działania zaimplementowanego klasyfikatora cech wykorzystującego algorytm genetyczny została przedstawiona na rys. 2.
W pierwszej kolejności w maksymalnym potencjalnym zbiorze cech dystynktywnych obliczana jest informacja wzajemna (5.3) pomiędzy poszczególnymi cechami oraz pomiędzy cechami, a wektorem przynależności klasowej. Pozwala ona na określenie, jak bardzo znajomość zmiennej *Y* ułatwia przewidywanie wartości zmiennej *X*. Innymi słowy jak bardzo poznanie jednej ze zmiennych zmniejsza niepewność drugiej zmiennej. W celu wyznaczenia informacji wzajemnej koniecznie jest określenie stopnia nieuporządkowania poszczególnych zmiennych zwanego entropią. Niska entropia zmiennej zapewnia jej wysoką przewidywalność, co jest efektem pożądanym. Entropię każdej zmiennej oraz entropię łączną wyznacza się wg zależności (5.1) i (5.2) [7]:

$$H(X) = -\sum_{x \in X} p(x) \log_2 p(x) \tag{5.1}$$

$$H(X,Y) = -\sum_{x \in X} \sum_{y \in Y} p(x,y) \log p(x,y) \tag{5.2}$$

$$I(Y;X) = H(X) + H(Y) - H(X,Y) \tag{5.3}$$

W rezultacie utworzona zostaje macierz informacji wzajemnych pomiędzy wszystkimi potencjalnymi cechami $I(yh_i; y)$ oraz wektor informacji wzajemnej pomiędzy każdą z cech a wektorem przynależności klasowych $I(yh_i; yh_j)$.
Niestety obliczanie informacji wzajemnej dla dużego zbioru cech jest bardzo czasochłonne. Dlatego powtarzanie wspomnianych obliczeń dla każdej możliwej kombinacji byłoby niewykonalne w rozsądnym czasie stosując deterministyczne formy rozwiązań. W przypadku algorytmu genetycznego bazowe informacje wzajemnie maksymalnego zbioru cech wykorzystywane są każdorazowo przy ocenie przystosowania populacji osobników. Populacja początkowa tworzona jest w sposób pseudolosowy poprzez generację wektorów (chromosomów) zawierających numery cech. Następnie każdy chromosom w populacji jest oceniany na podstawie funkcji przystosowania (5.5), wykorzystując uśrednione wartości informacji wzajemnej (5.4) dla występujących w populacji zbiorów cech [7]:

$$V = \frac{1}{N_n} \sum_{i=1}^{N_n} I(yh_i; y) \tag{5.4}$$

$$P = \frac{1}{N_n{}^2} \sum_{i=1}^{N_n} \sum_{j=1}^{N_n} I(yh_i; yh_j) \tag{5.5}$$

$$S = V - P \tag{5.6}$$

Jeżeli różnica pomiędzy maksymalną wartością funkcji przystosowania w danej populacji, a wartością średnią funkcji przystosowania w populacji nie przekroczy wyznaczonego empirycznie poziomu, to dalsze obliczenia są przerywane. Przerwanie obliczeń następuje również w przypadku osiągnięcia granicznej liczby pokoleń, która stanowi wartość stałą.
Jeżeli maksymalna wartość funkcji przystosowania nie spełnia warunków zakończenia i wykazuje wzrost przystosowania względem średniej wartości tej funkcji, to kolejnym krokiem algorytmu jest selekcja osobników najlepiej przystosowanych. Selekcja odbywa się w sposób pseudolosowy, jednakże faworyzuje osobniki lepiej przystosowane wg zależności (5.7) [7]

$$round\left(N \frac{e^{ar} - 1}{e^{r} - 1} \right), \tag{5.7}$$

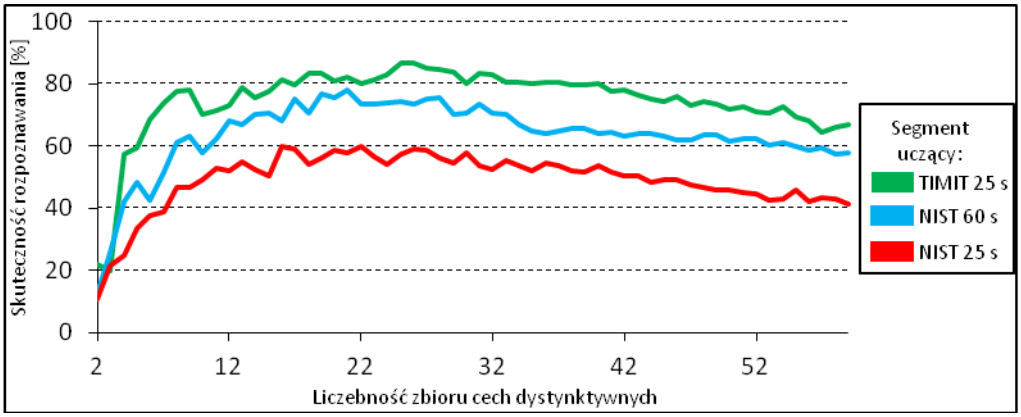
gdzie N stanowi wielkość populacji, r jest wartością pseudolosową z przedziału [0;1], natomiast a jest wartością stałą.
Następnie dwa wylosowane wg powyższego wzoru chromosomy podlegają krzyżowaniu. Krzyżowanie odbywa się wielopunktowo w ten sposób, że każda cecha dla nowego chromosomu dobierana jest pseudolosowo spośród dwóch cech występujących na danym indeksie w wektorach cech tzw. chromosomów rodziców. Procedura ta jest powtarzana, aż do otrzymania nowego wektora o tej samej długości co wektory rodzicielskie. W przypadku wystąpienia duplikacji cech w nowym wektorze, następuje mutacja wektora poprzez podmienienie cechy przez cechę nie występującą w wektorze. W wyniku operacji genetycznych powstaje nowa populacja, która poddawana jest sprawdzeniu funkcją przystosowania. Powyższe operacje są powtarzane, aż do spełnienia warunków stopu. Ostatecznym wynikiem jest wektor cech najlepiej przystosowanego chromosomu z ostatniego pokolenia.

6. KLASYFIKATOR GMM-UBM

W procesie klasyfikacji zastosowano liniową kombinację rozkładów normalnych GMMs, które umożliwiają generację modeli głosów zawierających dużą liczbę cennych informacji, a zarazem oszczędnych z punktu widzenia wymaganej pamięci [8]. Model GMM tworzony jest dla każdego mówcy, modelując wielowymiarowy rozkład gęstości prawdopodobieństwa na podstawie wyekstrahowanych cech dystynktywnych wygenerowanych w oparciu o dane uczące. Parametry startowe modelu mówcy, tj. wartości oczekiwane, macierze kowariancji oraz wagi rozkładów mogą być dobierane w sposób pseudolosowy lub zdeterminowany wg algorytmu GMM-UBM [8]. Metoda ta polega na utworzeniu głosu uniwersalnego na podstawie głosów wielu osób. Pozwala to na dopasowanie się modelu określonego mówcy w mniejszej liczbie iteracji, co wpływa korzystnie na skuteczności identyfikacji oraz na przyśpieszenie procesu tworzenia modeli głosów.

7. WYNIKI EKSPERYMENTÓW

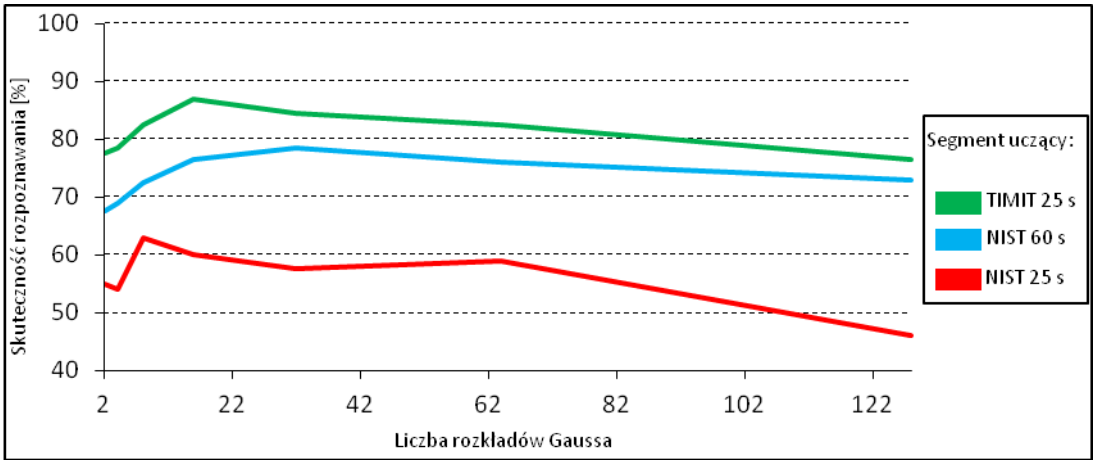
Głównym celem prezentowanych badań było zastosowanie skutecznego algorytmu selekcji dystynktywnych cech głosów mówców. Wykorzystano do tego celu algorytm genetyczny, dla którego danymi wejściowymi był maksymalny potencjalny zbiór cech dystynktywnych, wektor przynależności klasowych oraz określona długość pożądanego wektora cech po jego redukcji. Długość ta podlegała optymalizacji, której wyniki zilustrowano na rys. 3. Do testów wykorzystano dane pochodzące z bazy *TIMIT* i z bazy *NIST*. Długość segmentu uczącego dla każdego mówcy wynosiła 25 s, a długość odcinka testowego 5 s. W związku z tym, że system optymalizowany był we wcześniejszych badaniach [5] dla znacznie dłuższego segmentu uczącego, postanowiono wykonać również test dla 60 s segmentu uczącego z wykorzystaniem nagrań z bazy *NIST*.



Rys. 3. Skuteczność rozpoznawania z zależności od liczebności zbioru cech dystynktywnych

Wyselekcjonowane wektory cech zostały porównane, następnie wykonano kilka powtórzeń selekcji dla wektorów cech charakteryzujących się największą skutecznością. W rezultacie wybrano zbiór 22 cech pozwalający uzyskać najwyższą skuteczność identyfikacji.

W kolejnym etapie badań dokonano sprawdzenia dla jakiej liczby rozkładów Gaussa następuje najlepsza generalizacja danych, a tym samym najwyższa skuteczność identyfikacji. Testy przeprowadzono również dla trzech różnych długości segmentów uczących.



Rys. 4. Skuteczność rozpoznawania w zależności od liczby rozkładów Gaussa użytych do zamodelowania każdego z głosów mówców

Na podstawie otrzymanych wyników widać, że dla różnych długości segmentów uczących liczba GMMs dającą najwyższą skuteczność identyfikacji jest niejednakowa. Postanowiono wybrać liczbę 16 jako kompromisową wartość liczby rozkładów Gaussa, pozwalającą na najlepszą generalizację.

Ostatnim etapem prowadzonych badań było zastosowanie modelu uniwersalnego do utworzenia modeli głosów powstałych z wyselekcjonowanych cech dystynktywnych w procesie uczenia. Do jego utworzenia wykorzystano głosy pochodzące z bazy *TIMIT*, niebiorące udziału w testowaniu (92 głosów żeńskich i 338 głosów męskich). Dokonano również próby wykorzystania głosów pochodzących z bazy *NIST* do utworzenia modelu głosu uniwersalnego, jednakże otrzymano znacznie słabszą skuteczność identyfikacji, spowodowaną najprawdopodobniej znacznie niższą jakością nagrań w porównaniu do bazy *TIMIT*. Ostateczna skuteczność identyfikacji dla prezentowanego systemu ASR przedstawiona została w tab. 1.

TABELA 1

Wyniki skuteczności prawidłowej identyfikacji mówcy dla zbioru 22 cech dystynktywnych

Segment uczący	50 osób (25 kobiet + 25 mężczyzn)		100 osób (50 kobiet + 50 mężczyzn)		200 osób (100 kobiet + 100 mężczyzn)	
	prawidłowo rozpoznane	procentowa skuteczność	prawidłowo rozpoznane	procentowa skuteczność	prawidłowo rozpoznane	procentowa skuteczność
TIMIT	49	98,0%	91	91,0%	176	88,0%
NIST 60	48	96,0%	82	82,0%	154	77,0%
NIST 25	41	82,0%	65	65,0%	123	61,5%

8. WNIOSKI

Przeprowadzone badania pozwoliły wyselekcjonować zbiór cech dystynktywnych pozwalających na skuteczną identyfikację mówców. Zredukowano wymiar wektora z 60 do 22 cech co stanowi znaczący zysk pamięciowy oraz zwiększa szybkość obliczeń. Dzięki zastosowaniu algorytmu genetycznego podczas selekcji uwzględniono synergię cech, co znacząco przełożyło się na wypadkową skuteczność systemu względem wcześniejszych prób selekcji cech z wykorzystaniem miary Fishera i techniki *PCA* (ang. *Principal Component Analysis*) [5]. Dalsze prace wiążą się z koniecznością wykonania stosownej normalizacji wyników klasyfikatora GMM, a następnie utworzenia układu

decyzyjnego umożliwiającego odrzucenie mówcy nieznajdującego się w zamkniętej bazie głosów.

Literatura

1) Osowski S., *Metody i narzędzia eksploracji danych*, BTC, Legionowo, 2013.
2) Garofolo J. S. et al., *TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1*, Linguistic Data Consortium, Philadelphia, 1993.
3) Martin A., Przybocki M., *2002 NIST Speaker Recognition Evaluation LDC2004S04*, Linguistic Data Consortium, Philadelphia, 2004.
4) Brookes M., *VOICEBOX: Speech Processing Toolbox for MATLAB*, http://www.ee.ic.ac.uk/np/staff/dmb/voicebox/voicebox.html, 2002.
5) Kamiński K., Majda E., Dobrowolski A. P., *Automatic speaker recognition using Gaussian Mixture Models*, 17th IEEE SPA Conference, 2013, s. 220-225.
6) Dobrowolski A. P., Majda E., *Cepstral analysis in the speakers recognition systems*, 15th IEEE SPA Conference, 2011, s. 85-90.
7) Ludwig O., Nunes U., *Novel Maximum-Margin Training Algorithms for Supervised Neural Networks*, IEEE Transactions on Neural Networks, tom 21, nr 6, s. 972-984, 2010.
8) Reynolds, D. A., Quatieri, T. F., Dunn, R. B., *Speaker Verification Using Adapted Gaussian Mixture Models, Digital Signal Processing*, nr 10, s. 19-41, 2000.
9) Goldberg D. E., *Algorytmy genetyczne i ich zastosowanie*, WNT, Warszawa, 2003