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

Wojskowa Akademia Techniczna, Wydział Elektroniki (1)

Ocena funkcjonalności systemu rozpoznawania mówcy dla zdegradowanej jakości sygnału głosowego

Streszczenie. *W artykule przedstawiono wyniki badań automatycznego systemu rozpoznawania mówcy (ASR – ang. Automatic Speaker Recognition), przeprowadzonych na podstawie komercyjnej bazy głosów TIMIT. Badania prowadzone były pod kątem zastosowania ASR jako systemu automatycznego rozpoznawania rozmówcy telefonicznego. Przedstawiono również wpływ liczebności bazy głosów oraz stopień oddziaływania kompresji stratnej MP3 na skuteczność rozpoznawania mówcy.*

Abstract*. The article presents the results of tests of an automatic speaker recognition system (ASR) conducted on the basis of the TIMIT commercial voice database. The research was conducted with the aim of using ASR as a system for automatic recognition of telephone callers. The impact of the number of voices in the database and the effect of lossy MP3 compression on the effectiveness of speaker recognition has also been shown. (**Evaluation of functionality speaker recognition system for downgraded voice signal quality**).*

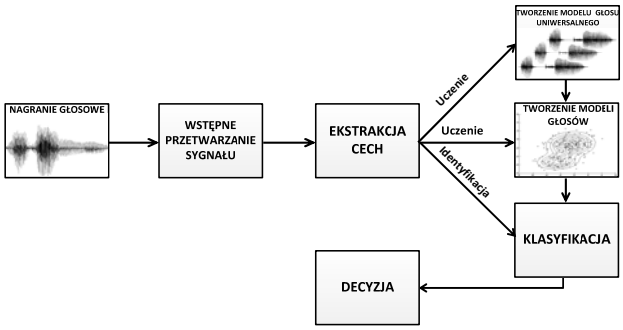
Słowa kluczowe: rozpoznawanie mówców, sygnał mowy, modele mieszanin gaussowskich, model uniwersalny
Keywords: speaker recognition, speech signal, Gaussian Mixtures Models, Universal Background Model

doi:10.12915/pe.2014.08.38

Wstęp

W dobie bardzo dobrze rozwiniętej infrastruktury telekomunikacyjnej istotną rolą dla systemów automatycznego rozpoznawania mówcy ASR może stać się rozpoznawanie rozmówcy podczas wykonywania rozmów telefonicznych. Zastosowań tego typu rozwiązania można znaleźć wiele, poczynając od weryfikacji klientów banku i realizacji ich uwierzytelnionych komend dotyczących zarządzania własnym kontem, a kończąc na identyfikacji głosu osób poszukiwanych lub dzwoniących z pogróżkami [1]. Wiąże się to jednak z analizowaniem sygnałów niskiej jakości. Prowadzone w różnych ośrodkach badawczych badania pokazują, jak duży wpływ na skuteczność rozpoznawania mówcy mają wszelkie procesy obniżające jakość sygnału, z których nie wszystkie są dla ucha ludzkiego wyczuwalne, jednakże dla ASR są bardzo istotne [2]. Ważnym elementem wpływającym na jakość przesyłanych sygnałów głosowych jest szybkość próbkowania, która w przypadku telefonii wynosi 8 kS/s. Zastosowanie ASR jako systemu rozpoznania rozmówcy telefonicznego wiąże się ze stosowaniem różnej – w zależności od potrzeb – wielkości baz głosów. W niniejszym artykule przedstawiono badania wpływu wielkości bazy głosów na skuteczność rozpoznawania mówcy. System przetestowano także w zróżnicowanych wariantach procesu klasyfikacji oraz sprawdzono stopień oddziaływania kompresji stratnej MP3 na skuteczność rozpoznawania mówcy.

W artykule pokazano zaimplementowany w środowisku *Matlab* system automatycznego rozpoznawania mówcy, który w procesie ekstrakcji cech wykorzystuje, tzw. „odcisk głosu" (*VP* – ang. *Voice Print*) [3]. Jako klasyfikatora system używa tzw. *modele mieszanin Gaussowskich* (*GMMs* – ang. *Gaussian Mixture Models*). Dzięki nim możliwe jest wykonanie stosunkowo niewielkich pamięciowo modeli głosów zawierających dużą liczbę cennych informacji o głosie mówcy [4]. Klasyfikator wykorzystuje jako dane startowe w procesie tworzenia modeli głosów tzw. model uniwersalny głosów UBM (ang. *Universal Background Model*), co pozwala na rozpoczęcie tworzenia poszczególnych modeli od wartości zbliżonych do większości ludzkich głosów. Ostatnim ogniwem systemu jest próba podjęcia decyzji o tym, który z utworzonych modeli głosów można z największym prawdopodobieństwem dopasować do zbioru wielowymiarowych punktów stanowiących wektory dystynktywnych cech głosu rozpoznawanego mówcy. Na rys. 1 przedstawiona została architektura prezentowanego systemu ASR. Szerszy opis poszczególnych etapów działania systemu znajduje się w dalszej części artykułu.



Rys. 1. Architektura systemu

Baza głosów

Prezentowane badania zostały wykonane na podstawie bazy nagrań *TIMIT* stworzonej przez *MIT* (ang. *Massachusetts Institute of Technology*), *SRI* (ang. *Stanford Research Institute*) oraz *TI* (ang. *Texas Instruments*). W bazie zgromadzone zostały nagrania 630 mówców obojga płci, nagrywanych z szybkością próbkowania 16 kS/s, przy zapisie jednokanałowym z 16-to bitową rozdzielczością amplitudową. Każdy z głosów reprezentowany jest przez 10 niezależnych nagrań mowy o długości około 3 s. Dla potrzeb prezentowanych badań wykorzystano 200 głosów kolejnych mówców. Wykorzystane głosy nie były w żaden sposób dobierane, wykorzystano 100 kolejnych męskich i 100 kolejnych żeńskich głosów z listy głosów dostępnych w bazie TIMIT.

W celu skorzystania z bazy TIMIT w prezentowanym systemie automatycznego rozpoznawania mówcy użyte nagrania zostały przepróbkowane do szybkości próbkowania 8 kS/s. Pozwala to na sprawdzenie skuteczności systemu w warunkach zbliżonych do transmisji telefonicznej. W celu wyodrębnienia segmentów uczących i niezależnych segmentów testowych zdecydowano, aby zawarte w bazie TIMIT nagrania głosowe każdego z mówców zostały scalone. Następnie z otrzymanego dla każdego mówcy nagrania o długości 30 s, wydzielano 20 s na segment uczący oraz 5 s lub 10 s na segment testowy. Nagrania uczące i testowe pochodziły z odrębnych fragmentów nagranych wypowiedzi.

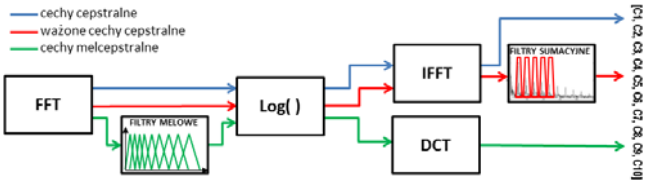
Wstępne przetwarzanie sygnałów mowy

W celu uniezależnienia zgormadzonych nagrań od ustawień sprzętu nagrywającego realizowane jest wstępne przetwarzanie sygnału. Jest to bardzo ważny etap, ponieważ poprzedza on proces generacji cech i ma istotny wpływ na jakość procesu identyfikacji mówcy. W ramach tego etapu realizowana jest filtracja sygnału, normalizacja, segmentacja oraz usuwanie ciszy. Filtracji dokonano wykorzystując cyfrowy filtr pasmowo-przepustowy o skończonej odpowiedzi impulsowej, którego rząd oraz częstotliwość odcięcia poddano wcześniejszemu procesowi optymalizacji. Po filtracji, ze względu na różne poziomy mocy sygnału mowy u różnych mówców, przeprowadzono normalizację zgromadzonych sygnałów w stosunku do maksymalnej wartości dla każdego mówcy. Ten rodzaj skalowania pozwala na zachowanie relacji energetycznych pomiędzy poszczególnymi fragmentami zapisu.

Segmentacja sygnału mowy jest tożsama z operacją okienkowania, czyli mnożenia sygnału przez okno czasowe. W prezentowanym ASR zastosowano okno Hamminga, ze względu na niski poziom listków bocznych jego widma, co minimalizuje zjawisko przecieków częstotliwości. Długość okna oraz przesunięcie poddane były optymalizacji wielokryterialnej we wcześniejszych badaniach [3]. Algorytm eliminacji ciszy polega na wycinaniu ramek niespełniających wyznaczonego empirycznie kryterium mocy w ramce, stanowiącego iloczyn mocy najcichszej ramki nagrania głosowego oraz stałej wyznaczonej w procesie optymalizacji. Dzięki operacji wycinania ciszy możliwe jest analizowanie wyłącznie ramek istotnych z punktu widzenia rozpoznawania mówcy, co wpływa korzystnie na skuteczność rozpoznawania głosów oraz szybkość działania systemu [5].

Generacja cech

Etap generacji cech jest kluczowym elementem procesu automatycznego rozpoznawania mówcy. Błędy popełnione podczas generacji cech są już nie do nadrobienia na dalszych etapach procesu. Prezentowany system wykorzystuje trzy zasadnicze rodzaje dystynktywnych cech głosu. Są nimi: cechy cepstralne, ważone cechy cepstralne oraz cechy melcepstralne [1]. Sposób ich generacji przedstawiony został na rys. 2.



Rys. 2. Schemat procesu generacji cech

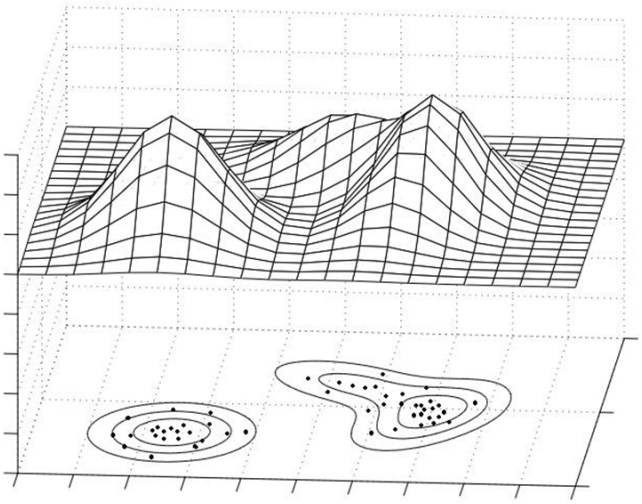
W każdej z prezentowanych technik generacji cech w pierwszej fazie działania obliczane jest widmo amplitudowe, celem analizy sygnału w dziedzinie częstotliwości. Podejście to jest inspirowane obserwacją ludzkich organów komunikacji głosowej, których działanie najwygodniej przedstawia się właśnie w tej dziedzinie. Z drugiej strony sygnał mowy przedstawiany w funkcji czasu charakteryzuje się, z punktu widzenia systemów rozpoznawania mówcy, bardzo dużą redundancją. Obserwując widma amplitudowe sygnałów mowy można łatwo zauważyć, że znacznie bardziej uwydatniają one różnice treści nagranych wypowiedzi niż osobnicze atrybuty związane m.in. z tonem krtaniowym. Dlatego w kolejnych modułach generatora cech sygnał mowy poddawany jest logarytmowaniu, dzięki któremu składowa wolnozmienna nie wymnaża się z amplitudami poszczególnych impulsów pochodzących od pobudzenia, tylko się do nich dodaje. Poddanie takiego sygnału odwrotnej transformacji Fouriera powoduje, że wolnozmienne przebiegi związane z transmitancją traktu głosowego są położone blisko zera na osi czasu cepstralnego zwanego pseudoczasem, a impulsy związane z dźwiękiem krtaniowym zaczynają się mniej więcej w okolicach okresu sygnału krtaniowego i powtarzają się co ten okres. Analizując wysokości kolejnych maksimów znajdujących się na osi czasu cepstralnego uzyskano pierwszą grupę cech zwanych cechami cepstralnymi.

Uzyskanie ważonych cech cepstralnych wiąże się dodatkowo z wykorzystaniem pasmowych filtrów sumacyjnych, które uwzględniają nie tylko maksymalne amplitudy prążków w cepstrum, ale również otaczające je wartości, które jak pokazują dotychczasowe badania, są także istotne dla skuteczności systemu rozpoznawania mówcy.

Ostatnią wykorzystaną techniką generacji cech są cechy melcepstralne. Zasadnicza różnica w procesie ich generacji polega na tym, że uzyskane widmo amplitudowe wymnażane jest przez tzw. filtry melowe, które naśladując ludzki organ słuchu i jego nieliniową wrażliwość na pobudzenia z różnych zakresów częstotliwości powodują poprawę percepcji. W następnej kolejności sygnał poddawany jest logarytmowaniu, podobnie jak w przypadku wcześniej omówionych technik generacji cech. Ostatnim etapem przetwarzania w omawianej ścieżce sygnałowej jest transformacja cosinusowa zapewniająca dekorelację cech.

Klasyfikator GMM-UBM

Zastosowane w procesie klasyfikacji modele mieszanin gaussowskich GMMs są liniową kombinacją rozkładów normalnych umożliwiającą generację modeli głosów zawierających dużą liczbę cennych informacji w sposób oszczędny z punktu widzenia wymaganej pamięci [6]. Model GMM tworzony jest dla każdego mówcy, modelując wielowymiarowy rozkład gęstości prawdopodobieństwa wyekstrahowany z danych uczących na etapie generacji cech. Parametry startowe modelu mówcy, tj. wartości oczekiwane, macierze kowariancji oraz wagi rozkładów mogą być dobierane w sposób pseudolosowy lub zdeterminowany. W niniejszym artykule przedstawiono badania wpływu zastosowania zróżnicowanych wartości startowych modeli na skuteczność rozpoznawania mówcy.



Rys. 3. Przykład zamodelowania danych uczących przez 3 rozkłady Gaussa

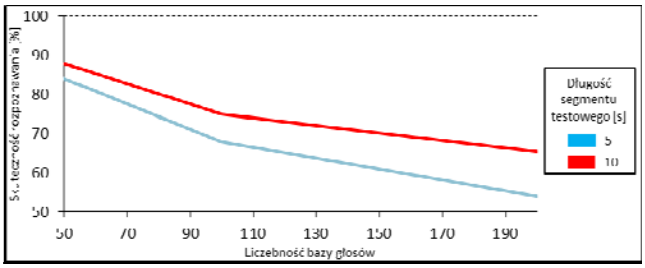
Deterministyczna metoda doboru początkowych wartości modelu głosu mówcy przedstawiona w artykule realizowana jest na podstawie algorytmu GMM-UBM [6].

Metoda ta polega na utworzeniu uniwersalnego modelu głosów UBM wykorzystując jako dane uczące głosy wielu osób [2]. Dzięki temu nie ma obaw, że model wystartuje z silnie odstających wartości początkowych. Ponadto dzięki zastosowaniu punktu startowego w postaci uniwersalnego modelu głosu możliwe jest dopasowanie się określonego modelu mówcy do danych uczących w mniejszej liczbie iteracji, dając w rezultacie wyższą skuteczność identyfikacji mówcy oraz przyśpieszony proces tworzenia poszczególnych modeli głosów (rys. 3). Uniwersalny model UBM wykorzystywany jest również w procesie decyzyjnym do normalizacji otrzymanych wyników wartości prawdopodobieństwa, a ściślej logarytmu prawdopodobieństwa (ang. *log-likelihood*) tego, że sygnał testujący pochodzi od danego mówcy [7].

Wyniki eksperymentów

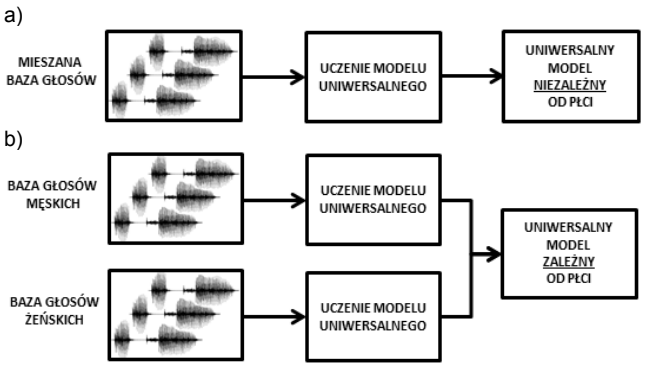
Przedstawione eksperymenty zostały przeprowadzone z wykorzystaniem komercyjnej bazy TIMIT. Baza ta charakteryzuje się niewielką reprezentacją poszczególnych głosów (około 30 s) i jest przeznaczona głównie do testowania procedur automatycznego rozpoznawania mowy. Prezentowany system optymalizowany był dla segmentu uczącego o długości 60 s i segmentu testowego o długości 3 s. Skutkiem testowania systemu dla znacząco odmiennych warunków są stosunkowo niskie wartości otrzymanej skuteczności rozpoznawania. Jednakże dzięki zwiększonemu udziałowi błędnych rozpoznawań, wpływ testowanych wariantów systemu na skuteczność rozpoznawania jest jeszcze bardziej uwydatniony, co ułatwia analizę wyników (system nie nasyca się w okolicy 100 %).

Pierwszy z przeprowadzonych testów (rys. 4) ukazuje spadek skuteczności rozpoznawania wraz ze wzrostem liczebności bazy głosów. Ponadto im długość segmentu testowego jest krótsza, ta zależność jest jeszcze silniejsza.



Rys. 4. Wpływ liczebności bazy głosów na skuteczność rozpoznawania mówcy

Przetestowano również wpływ zróżnicowanych wartości początkowych modeli głosów na skuteczność rozpoznawania mówcy. Pierwszym testowanym wariantem było pseudolosowe dobranie parametrów początkowych rozkładów, tj. wartości początkowych, macierzy kowariancji i wag rozkładów. Drugi wariant (rys. 5.a) wykorzystuje parametry uniwersalnego modelu głosu UBM jako początkowe wartości, które są adaptowane przez modele poszczególnych mówców. W tym wariancie do utworzenia modelu uniwersalnego wykorzystano mieszaną bazę głosów, bez podziału na płeć. Natomiast trzeci wariant (rys. 5.b) również wykorzystuje uniwersalny model UBM, jednakże w tym przypadku jest on utworzony na podstawie bazy głosów z uwzględnieniem podziału na płeć. Dla tego przypadku w chwili tworzenia modeli poszczególnych mówców użytkownik systemu deklaruje płeć mówcy tworzonego modelu. Na tej podstawie wprowadzane są parametry początkowe uniwersalnego modelu żeńskiego bądź uniwersalnego modelu męskiego.



Rys. 5. Schematy przetestowanych wariantów uniwersalnego modelu UBM

W związku z licznymi nagraniami głosowymi dostępnymi w Internecie w formacie *.mp3 mogącymi posłużyć jako baza dla systemów ASR, dokonano sprawdzenia wpływu zastosowanej kompresji MP3 (ang. MPEG-1/MPEG-2 Audio Layer 3) na skuteczność rozpoznawania mówcy. Kompresja MP3, jest jedną z najpopularniejszych stratnych kompresji dźwięku. Charakteryzuje się następującymi cechami:

- wykorzystuje model psychoakustyczny, uwzględniający percepcję dźwięku przez człowieka. Model psychoakustyczny uwzględnia zakres słyszalnych częstotliwości, a także maskowanie jednych dźwięków przez inne, powodując, że ciche dźwięki zbliżone (zarówno w dziedzinie czasu jak i częstotliwości) do głośnego dźwięku nie są słyszalne;
- wykorzystuje zapis danych w postaci zmiennoprzecinkowej, co zmniejsza ich rozmiar;
- umożliwia wybór określonej przepływność bitowej: stałej *CBR* (ang. *Constant Bit Rate*), zmiennej *VBR* (ang. *Variable Bit Rate*) lub dostępnej *ABR* (ang. *Available Bit Rate*). W prezentowanym eksperymencie wykorzystano stałą przepływność o wartości 128 kb/s. Są to warunki uznawane powszechnie za wystarczające, aby dla większości słuchaczy różnica dźwięku po kompresji w stosunku do dźwięku oryginalnego była praktycznie niesłyszalna;
- wykorzystuje kodowanie niezależnie każdego z kanałów oraz możliwość doboru najlepszego algorytmu kompresji dla każdej ramki (ang. *Joint Stereo*).

Zbiorcze wyniki wpływu kompresji MP3 oraz doboru parametrów początkowych modeli głosów zostały przedstawione w tab. 1.

Tab. 1. Wyniki wpływu stosowanych parametrów początkowych oraz zastosowania kompresji MP3 na skuteczność rozpoznawania mówcy

Format plików	Skuteczność rozpoznawania [%]		
	Losowe parametry startowe	UBM niezależny od płci	UBM zależny od płci
wav	84	86	88
mp3	72	74	78

Wyniki przeprowadzonych badań wskazują, że testowany system automatycznego rozpoznawania mówcy może z powodzeniem służyć jako system rozpoznawania rozmówcy telefonicznego, jednakże charakteryzuje się dużą wrażliwością na liczebność bazy głosów. Skutecznie zaimplementowano w systemie algorytm GMM-UBM, wykazując najwyższą skuteczność dla wariantu adaptującego parametry UBM zależnego od płci, jako dane startowe dla poszczególnych modeli głosów. Stosowanie UBM pozwala zwiększyć skuteczność rozpoznawania mówcy średnio o 5%, co jest istotnym zyskiem z punktu

widzenia ASR. W artykule przedstawiony został również test stosowania stratnej kompresji audio (MP3). Wyniki pokazały, że prawie niewyczuwalne dla ludzkiego słuchu różnice w jakości dźwięku przed i po kompresji MP3 stanowią duże utrudnienie dla systemu automatycznego rozpoznawania mówcy, czego rezultatem jest spadek skuteczności prawidłowej identyfikacji o ponad 10%. Dalsze badania będą wiązać się z przetestowaniem wpływu medium transmisyjnego i związanego z nim zróżnicowanego kodowania stosowanego w powszechnej komunikacji telefonicznej na skuteczność prawidłowej identyfikacji mówców.

LITERATURA

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

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

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

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

[5] 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

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

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

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