Biuletyn WAT
Vol. LXI, Nr 4, 2012

Modelowanie i optymalizacja generatora cech dla systemu rozpoznawania mówcy

EWELINA MAJDA, ANDRZEJ P. DOBROWOLSKI, BOGUSŁAW L. SMÓLSKI

Wojskowa Akademia Techniczna, Wydział Elektroniki,
Instytut Systemów Elektronicznych, 00-908 Warszawa, ul. S. Kaliskiego 2,
ewelina.majda@wat.edu.pl

Streszczenie. W pracy przedstawiono zagadnienia związane z modelowaniem i optymalizacją generatora cech dla systemu automatycznego rozpoznawania mówcy (ang. *Automatic Speaker Recognition* — ASR). Etap generacji cech (parametryzacji sygnału mowy) jest fundamentalny w tego typu systemach, z uwagi na fakt, że unikatowy wektor cech ma decydujące znaczenie w procesie rozpoznawania. Zadaniem generatora cech jest opisanie sygnału mowy za pomocą możliwie mało licznego zbioru deskryptorów, bez utraty informacji istotnych z punktu widzenia rozpoznawania mówcy. Ponadto parametryzacja powinna wykazywać odporność na warunki akustyczne i techniczne rejestracji oraz na zawartość lingwistyczną rejestrowanego materiału. Badania przedstawione w referacie koncentrowały się przede wszystkim na wielokryterialnej optymalizacji wybranych parametrów generatora cech opartego na analizie cepstralnej, uwzględniającej dodatkowo selekcję cech. Oceny otrzymanych wyników dokonano w oparciu o analizę składników głównych (ang. *Principal Component Analysis* — PCA) zbioru deskryptorów wyznaczonych dla próbek głosu pochodzących od 24 mówców.
Słowa kluczowe: automatyczne rozpoznawanie mówcy, analiza cepstralna, ekstrakcja cech, selekcja cech, analiza składników głównych

1. Wprowadzenie

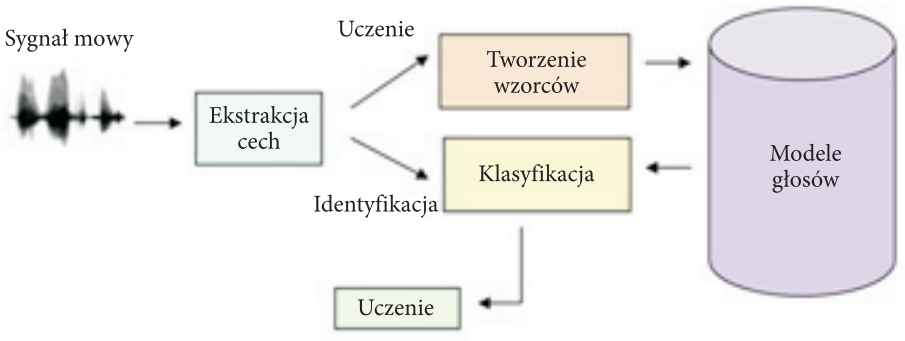
Mowa jest naturalnym i jednym z najbardziej efektywnych sposobów komunikowania się ludzi z otoczeniem. Pod nazwą automatyczne rozpoznawanie w odniesieniu do mowy mieści się wiele różnych rozwiązań technicznych. Ich wspólną cechą jest przetwarzanie sygnału mowy za pomocą urządzenia cyfrowego w celu wydobycia informacji wymaganych dla konkretnych zastosowań. W niniejszym artykule przedstawiono procedurę przetwarzania sygnału mowy w celu identyfikacji mówcy.

Automatyczne rozpoznawanie głosów obejmuje dwie zasadniczo różniące się procedury: identyfikację i weryfikację. *Identyfikacja* mówcy jest procesem decyzyjnym, polegającym na określeniu tożsamości mówcy i wykorzystującym do tego wyłącznie właściwości sygnału mowy (bez deklarowania przez mówcę tożsamości). Z kolei *weryfikacja* to proces decyzyjny, wykorzystujący cechy sygnału mowy do określenia, czy mówca danej wypowiedzi jest faktycznie osobą, której tożsamość deklaruje. Wynikiem weryfikacji jest potwierdzenie lub odmowa potwierdzenia deklarowanej tożsamości.

Bardzo istotną charakterystyką systemów rozpoznawania mówców jest ich zależność od wypowiadanego przez osobę rozpoznawaną tekstu, czyli od ograniczeń narzuconych na materiał lingwistyczny wypowiedzi. Rozróżnia się systemy rozpoznawania mówców zależne od tekstu (ang. *text-dependent*), w których zawartość lingwistyczna materiału treningowego i testowego jest ogólnie taka sama, oraz niezależne od tekstu (ang. *text-independent*), gdy zdania testowe różnią się od zdań uczących przynajmniej pod względem kolejności słów. W szczególności w tym przypadku dopuszcza się możliwość rozpoznawania mówcy nawet niezależnie od języka wypowiedzi [1].

System automatycznego rozpoznawania mówcy identyfikuje/weryfikuje osobę dzięki porównaniu tzw. *wektora cech* z bazą zarejestrowanych modeli głosów. Na rysunku 1 przedstawiono przykładowy schemat takiego systemu. Analiza sygnału mowy, w wyniku której otrzymuje się wektor cech niosący informację o indywidualnych właściwościach głosu mówcy (model głosu), może odbywać się w dwóch trybach: uczenia bądź identyfikacji.

W trybie uczenia nowi mówcy ze znanymi tożsamościami zapisywani są w bazie danych systemu. Tryb identyfikacji polega na porównaniu wyekstrahowanych unikatowych cech głosu nieznanego mówcy z próbkami zawartymi w bazie systemu (klasyfikacja). Zarówno w fazie uczenia jak i identyfikacji używane są te same algorytmy parametryzacji sygnału mowy wyznaczające unikatowy wektor cech, tzw. „odcisk głosu" (ang. *Voice Print*).



Rys. 1. Schemat procedury rozpoznawania mówców

2. Charaktertstyka zjawisk związanych z generacją mowy

Z fizycznego punktu widzenia proces komunikacji przy pomocy mowy polega na generowaniu i odbiorze bodźców akustycznych. Narząd mowy jest wyspecjalizowanym układem umożliwiającym generowanie szerokiej gamy dźwięków. Steruje on strumieniem powietrza wypływającym z płuc, umożliwiając kodowanie użytecznej informacji w postaci zmian chwilowego ciśnienia. Sygnał mowy oprócz informacji o treści wypowiedzi niesie również informacje związane z wewnętrzną strukturą jej źródła. Te osobnicze informacje odzwierciedlające indywidualne cechy głosu są wynikiem różnic w budowie traktu głosowego, nawyków nabytych w trakcie nauki mówienia oraz stopnia opanowania języka.

Dźwięki mowy wytwarzane są w tzw. organie mowy, którego zasadniczymi elementami są płuca, tchawica, krtań, gardło, nos, jama nosowa oraz usta. Część drogi głosowej leżącą powyżej krtani nazywa się kanałem głosowym. Kształt jego przekroju poprzecznego może się znacznie zmieniać pod wpływem ruchów języka, warg i szczęki (tzw. narządów artykulacyjnych), umożliwiając wymawianie (artykulację) różnych głosek. Z punktu widzenia wytwarzania mowy zasadniczym elementem krtani są tzw. fałdy (struny) głosowe. Przestrzeń pomiędzy fałdami głosowymi nazywa się głośnią. Fałdy głosowe mogą się otwierać i zamykać, wpływając na przepływ powietrza z płuc. Dźwięk wytwarzany w trakcie wydostawania się powietrza z płuc przez fałdy głosowe, które wykonują szybkie ruchy (periodyczne lub quasi-periodyczne) zamykające i otwierające głośnię, nazywa się dźwiękiem krtaniowym [2]. Głoski wytwarzane przy udziale drgań fałdów głosowych nazywają się dźwięcznymi. Wysokość głosu, a ściślej jego częstotliwość podstawowa, zmienia się w trakcie mowy w związku z naturalną intonacją i w przypadku głosu męskiego wynosi średnio 100-130 Hz, a dla głosu żeńskiego osiąga średnią wartość równą 200-260 Hz. Częstotliwość podstawowa w mowie zmienia się od 60 do 200 Hz u mężczyzn i od 180 do 400 Hz u kobiet [3].

Dźwięk krtaniowy stanowi sygnał wejściowy dla kanału głosowego, w którym jego widmo podlega znacznym modyfikacjom. Kanał głosowy zachowuje się jak układ filtrów (rezonatorów) o określonych częstotliwościach rezonansowych tak, że widmo tonu krtaniowego po przejściu przez układ tych filtrów charakteryzuje się pewnymi maksimami lokalnymi nazywanymi *formantami*.

Pierwotną i podstawową formą, w której rejestruje się sygnał mowy, jest przebieg czasowy. Przyjmując, że dla quasi-stacjonarnych fragmentów mowy trakt głosowy jest układem liniowym niezmiennym w czasie, sygnał mowy $s(t)$ można przedstawić jako splot impulsowego pobudzenia generowanego w głośni $e(t)$ i odpowiedzi impulsowej traktu głosowego $h(t)$:

$$s(t) = e(t) * h(t). \tag{1}$$

Dziedzina czasu nie jest jednak najwłaściwsza do przeprowadzania dalszych operacji, ponieważ sygnał mowy charakteryzuje się w niej bardzo dużą redundancją.

Znacznie efektywniejsze z punktu widzenia dalszej analizy jest przetransformowanie sygnału do dziedziny częstotliwości. Jednym z głównych powodów takiego podejścia jest próba naśladowania natury, która w toku milionów lat ewolucji wykształciła organ mowy człowieka, w którym sygnał mowy jest generowany — a następnie odbierany i analizowany przez organ słuchu — w dziedzinie częstotliwości. Znaczna część komputerowych metod przetwarzania mowy opiera się więc na analizie częstotliwościowej, która umożliwia zastąpienie splotu realizowanego w dziedzinie czasu iloczynem widma pobudzenia (krtaniowego) i transmitancji toru głosowego (zmiennej w takt artykulacji) [4, 5].

Ponieważ widmo dźwięku krtaniowego jest zmodulowane w amplitudzie przez funkcję przenoszenia traktu głosowego, korzystne jest wyznaczenie w pierwszej fazie logarytmu widma, gdyż w ten sposób multiplikatywny związek pobudzenia i traktu głosowego zostaje zastąpiony związkiem addytywnym, co znacznie upraszcza późniejszą separację obu składników. Przedstawione rozumowanie prowadzi wprost do metod *przetwarzania homomorficznego*, a w szczególności do koncepcji *cepstrum* [6, 7].

Ponieważ obliczanie logarytmu zespolonego wiąże się z komplikacjami wynikającymi z konieczności zapewnienia ciągłości fazy, a w przypadku sygnału mowy zasadnicza informacja zawarta jest w amplitudzie widma, w praktyce wyznacza się najczęściej tzw. *cepstrum rzeczywiste*, formalnie zdefiniowane następująco

$$c(t) = F^{-1}\left\{ \ln\left(\left| F\{s(t)\} \right| \right) \right\}, \tag{2}$$

co dla sygnałów dyskretnych sprowadza się do postaci

$$c(n) = IDFT\left(\ln\left(\left| DFT\left(s(n) \cdot w(n) \right) \right| \right) \right) \tag{3}$$

i ostatecznie

$$c(n) = \frac{1}{N}\sum_{m=0}^{N-1} C(m) \mathrm{e}^{\mathrm{j}2\pi\frac{mn}{N}} = \frac{1}{N}\sum_{m=0}^{N-1} \ln\left(\left| \sum_{n=0}^{N-1} s(n)\, w(n) \mathrm{e}^{-\mathrm{j}2\pi\frac{mn}{N}} \right| \right) \mathrm{e}^{\mathrm{j}2\pi\frac{mn}{N}}. \tag{4}$$

Ze względu na okresowość jądra transformaty Fouriera, logarytm z modułu widma amplitudowego $C(m)$ jest okresowy i jednocześnie spełnia zależność

$$C(m) = C(-m) = C(N-m). \tag{5}$$

Jest więc funkcją parzystą (symetria względem osi rzędnych), a zatem w jego rozwinięciu występują tylko funkcje kosinusoidalne (parzyste). Nie ma więc znaczenia, czy w ostatnim etapie zastosuje się prostą, czy odwrotną transformację Fouriera, czy po prostu tylko transformację kosinusową. Pozwala to na prostą

interpretację *cepstrum rzeczywistego* jako widma zlogarytmowanego widma amplitudowego [6, 7].

Widmo amplitudowe sygnału mowy wyznaczane najczęściej za pomocą szybkiej transformaty Fouriera (ang. *Fast Fourier Transform* — FFT) złożone jest z czynnika szybkozmiennego (wynikającego z pobudzenia) oraz czynnika wolnozmiennego (wynikającego z bieżącej konfiguracji narządów artykulacyjnych) modulującego amplitudę impulsowego sygnału pobudzenia. Podobnie wygląda interpretacja logarytmu widma amplitudowego, przy czym tu składowa wolnozmienna nie wymnaża się z amplitudami poszczególnych impulsów pochodzących od pobudzenia, tylko się do nich dodaje. Obliczenie widma takiego sygnału powoduje, że wolnozmienne przebiegi związane z transmitancją traktu głosowego są położone blisko zera na osi pseudoczasu, a impulsy związane z dźwiękiem krtaniowym zaczynają się mniej więcej w okolicach okresu sygnału krtaniowego i powtarzają się co ten okres. Informacja związana z transmitancją traktu głosowego jest skupiona w okolicy pseudoczasu zerowego, a zatem w tym obszarze należy poszukiwać zwięzłej informacji na temat tego, *co jest mówione*. Natomiast dla pseudoczasów powyżej okresu dźwięku krtaniowego informacja o tym, co jest mówione, jest zminimalizowana, pozostaje jedynie czytelna informacja dotycząca dźwięku krtaniowego, a ponieważ jest on ściśle związany z budową anatomiczną krtani i głośni, należy poszukiwać tam informacji osobniczej.

3. Parametryzacja sygnału mowy

Z punktu widzenia systemu rozpoznawania mówcy najważniejszym etapem jest generacja odpowiedniego zestawu deskryptorów numerycznych jak najlepiej charakteryzujących rozpoznawanych mówców. Celem parametryzacji sygnału mowy na potrzeby ASR jest takie przekształcenie czasowego przebiegu wejściowego, by uzyskać możliwie małą liczbę deskryptorów zawierających informacje istotne dla danego mówcy, przy jednoczesnej minimalizacji ich wrażliwości na zmienność sygnału nieistotną z punktu widzenia ASR. Wyboru tych deskryptorów dokonano, kierując się analizą przedstawionego wyżej procesu generacji mowy i poszukując elementów związanych z cechami osobniczymi.

3.1. Metodyka badań

Rejestracji czasowych przebiegów sygnału akustycznego mowy dokonano w Instytucie Systemów Elektronicznych Wydziału Elektroniki WAT z zastosowaniem mikrofonu dynamicznego Monacor DM-500, karty dźwiękowej komputera oraz oprogramowania Matlab. Przestrzenne zmiany ciśnienia akustycznego generowane przez mówcę rejestrowane są w pewnym punkcie przestrzeni za pomocą mikrofonu, którego zadaniem jest zamiana ciśnienia akustycznego na napięcie. O warunkach

rejestracji decydują charakterystyki mikrofonu i przetwornika A/C. Pożądane jest, aby ich jakość była wystarczająco dobra i aby elementy te nie miały znaczącego wpływu na strukturę zarejestrowanego sygnału. Podczas badania, odległość mikrofonu od ust osoby mówiącej wynosiła ok. 10 cm. Dodatkowo mikrofon został wyposażony w osłonę, która zapobiegała zniekształceniom towarzyszącym *sybilantom* (tzw. głoski świszczące: s, sz, cz, ć) i *głoskom wybuchowym* (p, b, t). Materiał fonetyczny obejmował różnorodne teksty będące fragmentami typowego dialogu oraz wypowiedzi o charakterze podniosłym i zabawnym. Grupa biorąca udział w doświadczeniu składała się z szesnastu mężczyzn i ośmiu kobiet.

W opisywanych w literaturze opracowaniach stosowane są różne strategie dotyczące wyboru częstotliwości próbkowania. Mniejsza częstotliwość próbkowania oznacza mniejszą liczbę danych do przetworzenia, ale utratę części informacji. Większa częstotliwość próbkowania z kolei oznacza więcej danych i niekoniecznie lepszą jakość rozpoznawania. Projektując system rozpoznawania mówcy, należy znaleźć kompromis między wiernością zapisu sygnału, w kontekście zachowania cech osobniczych, a ilością danych zajmujących pamięć komputera i wpływających na szybkość obliczeń. Badania pilotażowe przeprowadzono z sygnałami próbkowanymi z częstotliwościami 44100 Hz, 22050 Hz i 11025 Hz i w rezultacie przyjęto wartość 22050 Hz, przy 16-bitowej rozdzielczości amplitudowej oraz rejestracji jednokanałowej (monofonicznej). Z zarejestrowanego materiału badawczego została utworzona baza danych zawierająca identyfikator mówcy oraz odpowiadające mu próbki sygnału akustycznego.

3.2. Przetwarzanie wstępne

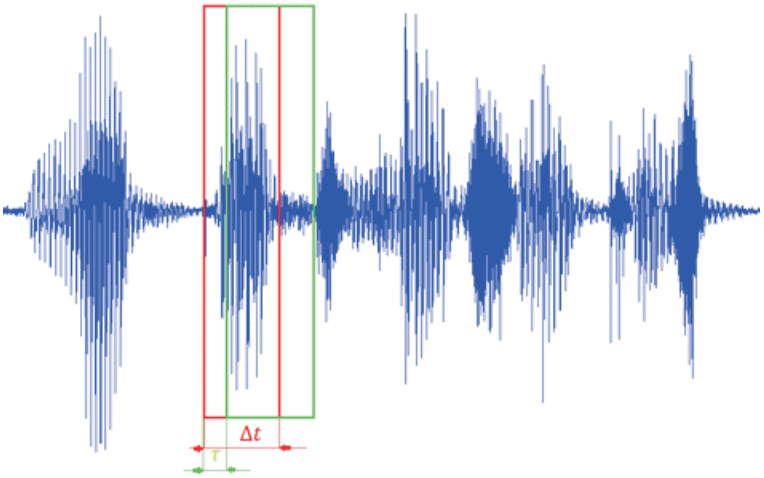
Wstępne przetwarzanie sygnału mowy jest bardzo ważnym etapem obróbki danych, ponieważ poprzedza wprowadzenie sygnału do generatora cech i ma fundamentalne znaczenie dla jakości procesu identyfikacji mówcy. Głównym celem wstępnej obróbki sygnału mowy jest jak największe uniezależnienie zarejestrowanych sygnałów akustycznych od ustawień sprzętu nagrywającego. Na tym etapie przeprowadzana jest filtracja, a także normalizacja, gdyż te dwie procedury w dużym stopniu niwelują różnice wynikające z różnych charakterystyk częstotliwościowych torów pomiarowych oraz z różnych poziomów głośności. W aplikacji zastosowano cyfrowy filtr pasmowo-przepustowy o skończonej odpowiedzi impulsowej. Zakładając brak zniekształceń i zakłóceń sygnału, pominięto kwestie dotyczące tłumienia odbić, zakłóceń i szumów. Zagadnienia te będą jednak uwzględniane w dalszych badaniach.

3.3. Generator cech osobniczych

Sygnał mowy jest sygnałem o zmiennej w czasie strukturze częstotliwościowej, dlatego parametryzacji poddawane są kolejne fragmenty sygnału, a nie sygnał jako

całość. Fragmenty, na jakie dzielony jest sygnał, nazywane są ramkami (rys. 2). Najczęściej długość ramki Δt powiązana jest z jej przesunięciem (skokiem) τ zależnością [4]

$$\tau = \frac{1}{3}\Delta t. \tag{6}$$



Rys. 2. Ilustracja przesunięcia ramki — przedstawiono dwa sąsiednie położenia ramki

Jednym z pierwszych zadań autorów było ustalenie podstawowego parametru generatora cech, jakim jest długości ramki. Czasy trwania poszczególnych jednostek fonetycznych są różne i zależne od określonego mówcy. Jednostki składające się z głosek dźwięcznych charakteryzują się czasem trwania z przedziału od 10 ms do nawet 200 ms [4]. Zakres zmienności jest więc znaczny i decyzja dotycząca wyboru długości ramki jest niezmiernie ważna w projektowanym systemie ASR. Badania dotyczące optymalizacji poszczególnych parametrów generatora cech przedstawione są w kolejnym rozdziale.

Podział sygnału na ramki powoduje powstawanie nieciągłości w przetwarzanym sygnale, co wiąże się ze zjawiskiem przecieku częstotliwości. Aby zminimalizować to zjawisko sygnał z każdej ramki należy poddać procesowi okienkowania, czyli wymnożenia przez odpowiednią funkcję okna. Dzięki temu następuje wygładzenie nieciągłości i usunięcie z widma fałszywych składowych. Autorzy zastosowali charakteryzujące się dobrymi właściwościami *okno Hamminga*

$$w(n) = 0{,}54 - 0{,}46\cos\left(\frac{2\pi n}{N}\right);\ 0 \le n \le N. \tag{7}$$

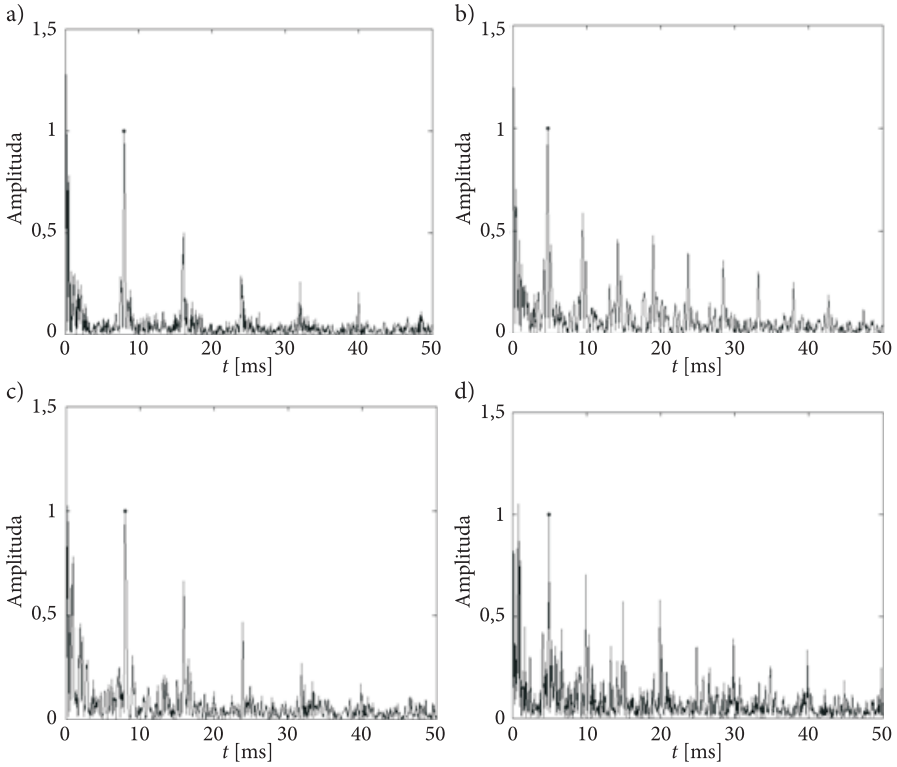
Ze względu na to, że istotna informacja związana z mówcą i niesiona przez dźwięk krtaniowy zawarta jest w dźwięcznych fragmentach mowy, podczas analizy

należy brać pod uwagę jedynie „ramki dźwięczne". Fragmenty dźwięczne charakteryzują się regularnym występowaniem maksimów (co okres tonu podstawowego), w przeciwieństwie do fragmentów bezdźwięcznych, które przypominają sygnał aperiodyczny. Klasyfikacja fragmentów sygnału mowy na dźwięczne i bezdźwięczne dokonywana jest w systemie za pomocą funkcji autokorelacji. Aby sprawdzić, czy analizowana głoska jest dźwięczna, należy wyznaczyć drugie globalne maksimum i sprawdzić jego poziom (pierwsze maksimum występuje oczywiście dla przesunięcia zerowego). Jeżeli jest ono większe od pewnej wartości odniesienia p_v, to dany fragment należy uznać za dźwięczny, w przeciwnym przypadku za bezdźwięczny. Ustalenie optymalnego progu p_v to kolejny fragment procedury optymalizacyjnej opisanej w rozdziale 4.

Dodatkowym ograniczeniem zastosowanym przez autorów przy wyborze tzw. reprezentatywnych dla danego mówcy ramek jest detekcja aktywności mówcy. W trakcie rejestracji pojawiają się często fragmenty sygnału, podczas których mówca nie jest aktywny. Zastosowanie kolejnego parametru odpowiadającego za odrzucenie ramek tego typu ma na celu przede wszystkim eliminację ciszy z nagrania oraz odrzucenie ramek będących potencjalnie szumem, a więc takich, które mogą powodować błędną ekstrakcję cech. W takim podejściu w pierwszej kolejności należy określić statystykę sygnału $P(n)$, na podstawie której będzie dokonywana selekcja, a następnie zastosować kryterium decyzyjne. Zwykle dokonuje się odniesienia wartości $P(n)$ do pewnego ustalonego progu. W zależności od wielkości, na jakiej bazuje selekcja, algorytmu jej wyznaczania oraz wartości progu, wyniki selekcji będą różne. Autorzy zdecydowali się oprzeć na wartości mocy składowej zmiennej, czyli na wariancji sygnału. Ustalenie dodatkowego parametru, jakim jest próg mocy p_p, było więc kolejnym zadaniem optymalizacji wielokryterialnej, który opisany został w następnym rozdziale.

Klasyczna metoda rozplotu cepstralnego, w przypadku analizy pod kątem rozpoznawania mówcy, polega na usunięciu niepożądanego składnika poprzez wyzerowanie próbek cepstrum dla pseudoczasu w okolicach zera i poszukiwaniu unikatowych cech każdego mówcy dla czasów powyżej okresu dźwięku krtaniowego. Przydatność cepstrum rzeczywistego do celów rozpoznawania mówcy można łatwo zauważyć, analizując przebiegi przedstawione na rysunku 3 — informacje o wypowiadanej głosce zacierają się, natomiast zarysowuje się wyraźne zróżnicowanie w zależności od mówcy.

Na etapie generacji cech zdefiniowano 9 deskryptorów numerycznych różnicujących mówców. Należą do nich: częstotliwość podstawowa F_{av} (deskryptor nr 1), będąca odwrotnością położenia drugiego maksimum cepstrum, oraz wartości 7 kolejnych maksimów cepstrum unormowanego c_1-c_7 (deskryptory nr 3-9). Dla każdego mówcy dokonywano uśredniania zbioru cech cepstralnych w zbiorze reprezentatywnych ramek i dodatkowo uzupełniano zbiór deskryptorów o odchylenie standardowe częstotliwości podstawowej σ (deskryptor nr 2).



Rys. 3. Moduły cepstrum rzeczywistego głosek *a* i *e*, a) głosem męskim, b) głosem żeńskim

4. Wielokryterialna optymalizacja systemu

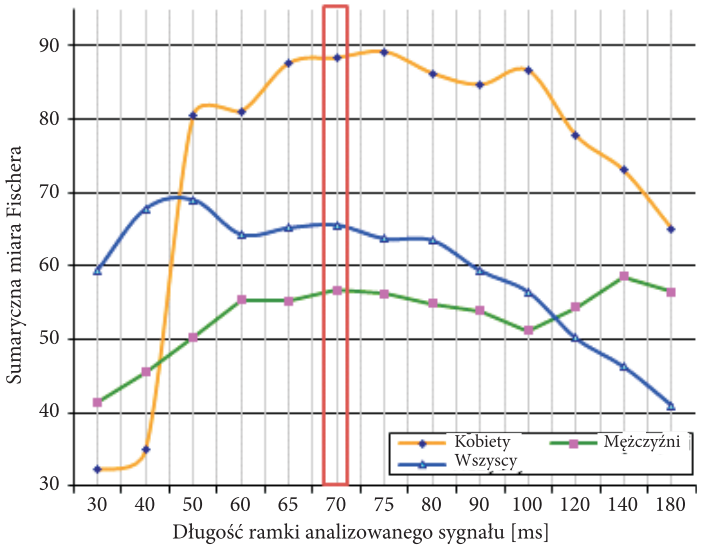
W poprzednim rozdziale przedstawiono ogólny schemat projektowanego systemu ASR. W zależności od tego, jaką funkcję ma spełniać dany system (rozpoznawanie treści wypowiedzi bądź tożsamości mówcy), należy dobrać optymalne parametry układu, uwzględniając przyjętą procedurę ekstrakcji wektora cech oraz tryb rejestracji. Autorzy mieli za zadanie optymalizację systemu w oparciu o cztery zasadnicze parametry: długość ramki (Δt) oraz jej przesunięcie (τ), próg dźwięczności ramki (p_v) i próg minimalnej mocy (p_p).

Ze względu na szerokie zakresy zmian wszystkich optymalizowanych parametrów autorzy postanowili w pierwszej kolejności dokonać wstępnego wyboru wartości parametrów w oparciu o współczynniki istotności Fischera definiowane zależnością

$$F_{ij}(f) = \frac{\left|c_i - c_j\right|}{\sigma_i + \sigma_j}, \tag{8}$$

gdzie c_i i c_j są wartościami średnimi i-tej i j-tej cechy, natomiast σ_i oraz σ_j ich odchyleniami standardowymi.

Obliczeń współczynników istotności Fischera dokonano dla 9 scharakteryzowanych wyżej deskryptorów bazując na 8 klasach, wśród których wyróżniono cztery kobiety i czterech mężczyzn. Równomierny podział na kobiety i mężczyzn nie był przypadkowy. Należy zwrócić uwagę na fakt, iż wartość danego deskryptora może mieć dużą siłę dyskryminacyjną pomiędzy poszczególnymi kobietami, lecz znacznie mniejszą wśród mężczyzn. Z tego powodu obliczeń współczynników istotności Fischera dokonano w trzech podklasach: *Kobiet*, *Mężczyzn* oraz w podklasie *Wszyscy*. Ponieważ klas jest więcej niż dwie, współczynniki istotności Fischera obliczono dla wszystkich par oraz wyznaczono ich sumę (sumaryczny współczynnik istotności Fischera). W pierwszym etapie parametrem optymalizowanym była długość ramki (Δt). Uzyskane wyniki zobrazowano na rysunku 4.



Rys. 4. Wykres sumarycznej miary Fischera dla poszczególnych podklas w zależności od długości ramki analizowanego sygnału

Z wykresu wyraźnie widać, że dla małej długości ramki (30-40 ms) współczynniki Fischera są niewielkie. Zdecydowany przyrost następuje w okolicach 50 ms, a dla długości ramki przekraczającej 90 ms wartości współczynników w podklasach *Kobiety* i *Wszyscy* znacząco spadają. Należało więc dokonać wyboru czasu trwania ramki z przedziału od 60 do 80 ms. Warto podkreślić fakt, że nie istnieje taka długość ramki, dla której współczynniki istotności Fischera osiągają maksimum we wszystkich trzech podklasach, dlatego należało dokonać pewnego wyboru kompromisowego. Ostatecznie autorzy zdecydowali się na długość ramki wynoszącą 70 ms. Dla sprawdzenia poprawności wyboru wykonano kilka serii dokładniejszych badań, potwierdzających, że optymalna długość ramki $\Delta t = 70$ ms.

Kolejnym parametrem, który należało poddać optymalizacji, był krok (τ), z jakim realizowane będzie przesuwanie ramki wzdłuż analizowanego sygnału mowy. Podczas rozwiązania tego problemu należy uwzględnić fakt, że mniejsza wartość przesunięcia daje większą liczbę ramek, co przekłada się na wydłużenie czasu obliczeń. Próba poszukiwania wartości przesunięcia ramki odbywała się jednocześnie z optymalizacją dwóch pozostałych parametrów (p_v oraz p_p). Ze względu na dużą ilość informacji zawartej w danych wejściowych, jakimi w rozważanym przypadku są 9-wymiarowe wektory cech, autorzy zdecydowali się na optymalizację w oparciu o analizę składników głównych (ang. *Principal Component Analysis* — PCA). Istotą tej metody jest zamiana dużej ilości informacji zawartej we wzajemnie skorelowanych danych wejściowych w zbiór statystycznie niezależnych składników uszeregowanych według ich ważności. Był to jeden z najbardziej pracochłonnych etapów badań. Prace polegały na obserwacji zmian położenia wektorów cech poszczególnych mówców na płaszczyźnie PCA_1/PCA_2. Badań dokonano w oparciu o trzy 8-osobowe zbiory mówców. I tutaj również powtórzył się problem wyboru optymalnych wartości parametrów, bowiem pewien zestaw parametrów zapewniający idealne rozróżnienie w jednym zbiorze mówców nie najlepiej sprawdzał się w przypadku innego zbioru. Należało więc dokonać wyboru kompromisowego, rozważając wszystkie 24 osoby biorące udział w eksperymencie.

Zgodnie z literaturą [4] wyznaczanie częstotliwości podstawowej metodą cepstralną jest mniej dokładne, lecz bardziej niezawodne niż metodą autokorelacyjną, w szczególności przy silnie zaszumionym sygnale mowy. W poszukiwaniu możliwość uzyskania większej stabilności deskryptorów zastosowano dodatkowe ograniczenie przy wyborze poprawnych ramek, polegające na porównaniu wartości częstotliwości podstawowej otrzymanej w oparciu o funkcję autokorelacji oraz w oparciu o cepstrum. Ostatecznie ustalono, że jeżeli różnice pomiędzy wartościami częstotliwości podstawowej ramki, wyznaczonymi za pomocą tych dwóch metod, różnią się o więcej niż 15%, ramka taka zostaje automatycznie odrzucona i nie bierze udziału w generacji deskryptorów.

Zbiór zoptymalizowanych wartości parametrów generatora cech określony dla 30-sekundowych wycinków głosu przedstawiono w tabeli 1.

Tabela 1

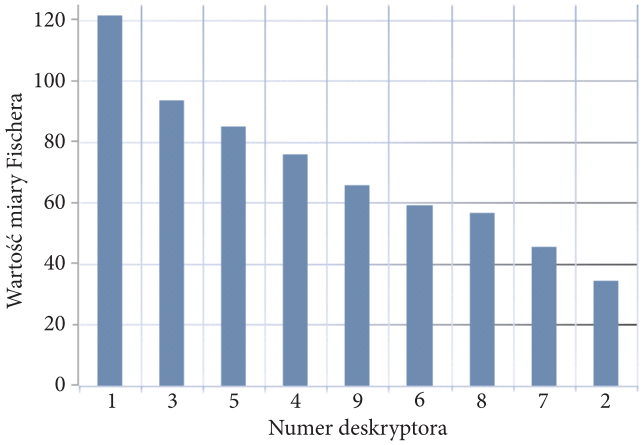
Zoptymalizowane wartości parametrów generatora cech

Parametr		Wartość
Długość ramki	Δt	70 ms
Przesunięcie ramki	τ	18 ms
Próg dźwięczności	p_v	10%
Próg mocy	p_p	20%
Próg różnic częstotliwości podstawowej	p_f	15%

5. Selekcja cech cepstralnych

Zdefiniowane na etapie generacji cech deskryptory stanowią maksymalny zbiór potencjalnych cech dystynktywnych, które mogą być wykorzystane w systemie automatycznego rozpoznawania wzorca reprezentującego badany obiekt. Badania prowadzone na świecie pokazują, że nie zawsze użycie maksymalnego zestawu cech prowadzi do najlepszych wyników, gdyż nie są one jednakowo ważne w procesie rozpoznania wzorców. Pewne cechy mogą mieć postać szumu pomiarowego, pogarszając możliwość rozpoznania danego mówcy, natomiast cechy silnie skorelowane mają zwykle niekorzystny wpływ na jakość klasyfikacji, dominując nad innymi i tłumiąc w ten sposób ich korzystny wpływ [8]. Ważnym elementem procesu staje się zatem ocena jakości deskryptorów i zastosowanie metod selekcji przy tworzeniu optymalnego wektora cech, na podstawie którego będzie dokonywana klasyfikacja (identyfikacja, weryfikacja).

Autorzy postanowili wstępnie zastosować selekcję opartą na metodzie Fischera. Zgodnie z jej założeniami duża wartość sumarycznego współczynnika istotności Fischera oznacza dobrą zdolność dyskryminacyjną cechy pomiędzy klasami, a mała oznacza, że wartości cechy należące do obu klas są rozproszone i potencjalnie przemieszane ze sobą, co dyskwalifikuje ją jako cechę diagnostyczną. Sumaryczne współczynniki istotności Fischera poszczególnych deskryptorów przedstawione są na rysunku 5.



Rys. 5. Wykres sumarycznego współczynnika istotności Fischera poszczególnych deskryptorów

Wyniki wskazują, że najkorzystniejsze są deskryptory o numerach 1, 3, 5, 4, 9 i 6. Zdecydowanie najmniejszą wartość otrzymano dla deskryptora nr 2 (wariancja częstotliwości podstawowej). Niezależnie od sumacyjnej wartości dyskryminacyjnej poszczególnych cech, budując każdy automatyczny system klasyfikacji, warto

sprawdzić siłę dyskryminacyjną deskryptorów pracujących w zespole. Często okazuje się bowiem, że włączenie równoległego działania wielu cech na raz może zmienić jakość danej cechy. Pewne cechy (nawet te gorsze), współpracując ze sobą, wzbogacają się nawzajem, podnosząc swoją wartość dyskryminacyjną. Autorzy przeprowadzili taką analizę, śledząc zmiany położeń poszczególnych wektorów określających mówcę na płaszczyźnie PCA_1/PCA_2.

Na podstawie obliczonych miar Fischera oraz obserwacji zmian położenia wektorów cech w oparciu o transformację PCA określono optymalny 5-wymiarowy wektor cech **vp** („odcisk głosu" — ang. *Voice Print* — VP) składający się z częstotliwości podstawowej oraz czterech cech cepstralnych

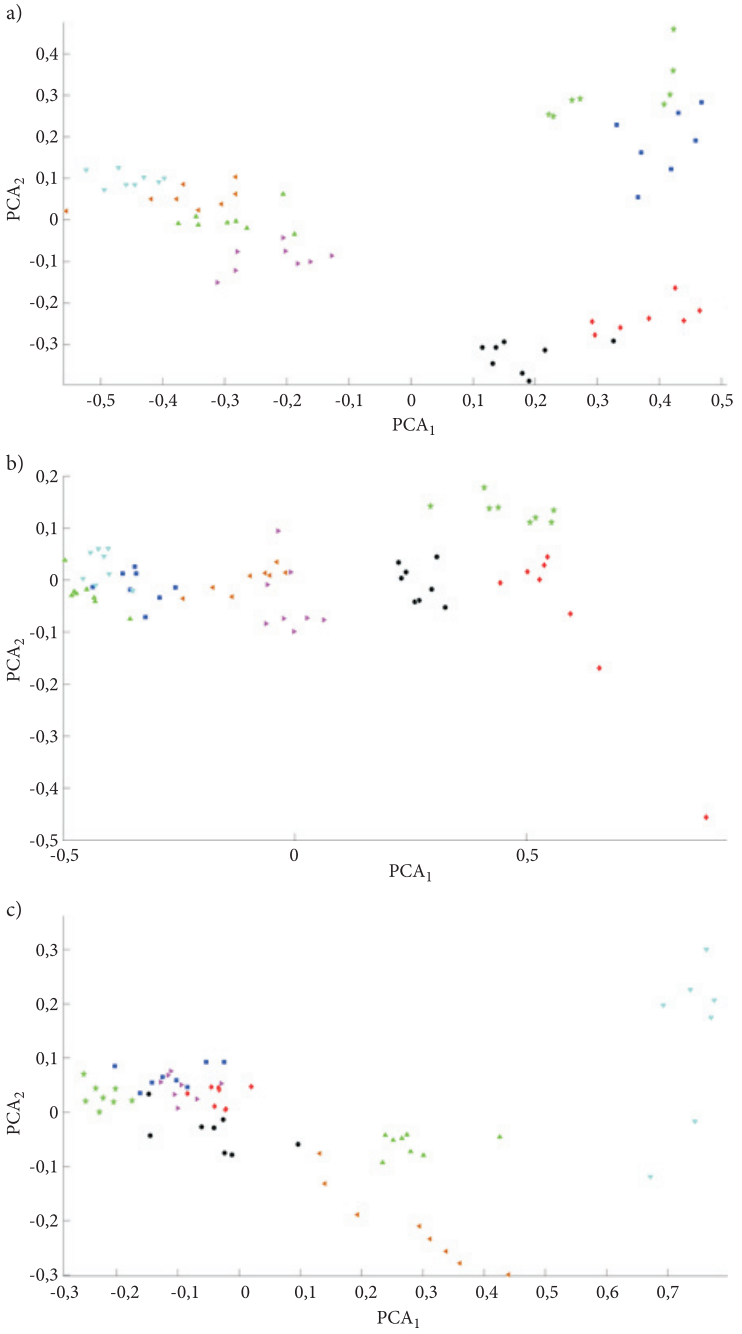
$$\begin{cases} vp_1 = F_{av} = \dfrac{1}{N}\displaystyle\sum_{j=1}^{N} F_j, \\ vp_{i-1} = \dfrac{1}{N}\displaystyle\sum_{j=1}^{N} c_{i,j}, \qquad i = 3,4,5,6, \end{cases} \tag{9}$$

gdzie: N — liczba poprawnych ramek;
F_j — częstotliwość podstawowa j-tej ramki, wyznaczana z cepstrum rzeczywistego;
$c_{i,j}$ — wartość i-tego maksimum cepstrum rzeczywistego j-tej ramki.

6. Wyniki badań

W wyniku optymalizacji wielokryterialnej parametrów projektowanego systemu oraz selekcji deskryptorów otrzymano ostateczny model generatora cech osobniczych projektowanego systemu rozpoznawania mówcy. Przykładowe wyniki przeprowadzonej transformacji PCA przedstawione są na rysunku 6. Poszczególne wyniki odnoszą się do trzech rozłącznych 8-osobowych zbiorów mówców. Każdy mówca reprezentowany jest przez osiem 5-wymiarowych *odcisków głosu*. Warto zwrócić uwagę na wyniki analiz przedstawione na rysunku 6c, który przedstawia wyłącznie samych mężczyzn, w przeciwieństwie do dwóch poprzednich, gdzie przedstawiono zbiory z jednakową liczbą kobiet i mężczyzn (kobiety grupują się po prawej stronie płaszczyzny, natomiast mężczyźni po lewej).

Główną zaletą transformacji PCA jest możliwość obserwacji rozkładu poszczególnych wektorów cech wygenerowanych dla każdego mówcy przy pomocy zoptymalizowanego generatora cech na płaszczyźnie, pomimo że oryginalny wektor cech jest 5-wymiarowy. Dzięki temu można dokonać wstępnej klasyfikacji poszczególnych mówców. Warto zauważyć, że dla każdego mówcy otrzymano dość dobrą powtarzalność wygenerowanych wektorów, mimo dużej różnorodności



Rys. 6. Rozkład danych zrzutowanych na dwa najważniejsze składniki główne PCA; odpowiednio 1., 2. i 3. zbiór mówców

treści zarejestrowanych wypowiedzi (dialog, wypowiedź poważna i wypowiedź żartobliwa). Świadczy to o spełnieniu podstawowego warunku projektowanego systemu, jakim jest uniezależnienie generowanego wektora cech od treści i charakteru wypowiedzi.

Należy podkreślić fakt, że obserwacja otrzymanych wektorów cech odbywała się jedynie w funkcji dwóch pierwszych składowych głównych. Dalsze badania wykazują, że kolejne składowe potwierdzają jeszcze lepszą separację poszczególnych mówców. Najbardziej jest to widoczne w przypadku, gdy dwóch mówców nieznacznie „zachodzi na siebie" w funkcji pierwszej oraz drugiej składowej. Analiza tych wektorów w funkcji PCA_3 i PCA_4 pozwala na stwierdzenie, że klasy te jednak wyraźnie się rozdzielają.

Należy również dodać, że wśród uczestników badania były matka oraz córka. Wydawać by się mogło, że ze względu na bliskie pokrewieństwo występujących w eksperymencie osób wystąpi częściowe pokrycie się dwóch mówców. Jednak również te dwie klasy są wyraźnie separowalne. Widoczne jest to na rysunku 6a, gdzie wektory cech oznaczające odpowiednio matkę i córkę to punkty zaznaczone kolorem niebieskim oraz zielonym. Świadczy to niezaprzeczalnie o dobrych własnościach dyskryminacyjnych zaproponowanego generatora cech.

7. Podsumowanie

Etap parametryzacji sygnału jest bardzo ważny ze względu na fakt, że niewłaściwych wyników nie można skorygować w dalszych etapach. Przeprowadzone eksperymenty pozwoliły na zoptymalizowanie modelu generatora cech w projektowanym systemie rozpoznawania mówcy. Dokonano wielokryterialnej optymalizacji wybranych parametrów i zastosowano ostateczną selekcję deskryptorów. Wyniki przedstawione przy pomocy transformacji PCA wyglądają bardzo obiecująco. Każdy z mówców koncentruje się w oddzielnym obszarze. Dodatkowe badania przeprowadzone przy uwzględnieniu większej liczby składowych PCA potwierdzają przedstawione wnioski. Można więc oczekiwać, że wygenerowane wektory cech są przede wszystkim unikatowe dla każdego mówcy, ale również odporne na tekst wypowiadany przez badanego.

Ostatnim etapem procesu rozpoznawania mówcy jest przeprowadzenie klasyfikacji — obecnie autorzy zajmują się właśnie tym zagadnieniem. Z dotychczasowej analizy wynika, że do klasyfikacji zostanie zastosowana nieliniowa sieć SVM, a głównym celem autorów będzie dobranie optymalnych parametrów tej sieci, zapewniających minimalne błędy klasyfikacji.

Praca naukowa finansowana ze środków na naukę w latach 2010-2012 jako projekt rozwojowy.

Artykuł wpłynął do redakcji 9.03.2012 r. Zweryfikowaną wersję po recenzji otrzymano w maju 2012 r.

LITERATURA

[1] S. Furui, *Recent advantages in speaker recognition*, Pattern Recognition Letters, 18, 1997, 859-1872.

[2] T. Kinnunen, H. Li, *An overview of text-independent speaker recognition: From feature to supervectors*, Speech Communication, 2010, 12-40.

[3] Z. Pawłowski, *Foniatryczna diagnostyka wykonawstwa emisji głosu śpiewaczego i mówionego*, Impuls, 2005.

[4] Z. Ciota, *Metody przetwarzania sygnałów akustycznych w komputerowej analizie mowy*, Exit, 2010.

[5] A. Dobrowolski, E. Majda, *Ocena przydatności wybranych cech sygnału mowy w systemach automatycznego rozpoznawania mówcy*, Przegląd Elektrotechniczny, R. 87, 10, 2011, 193-197.

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

[7] A. Dobrowolski, E. Majda, *Application of homomorphic methods of speech signal processing in speakers recognition system*, Przegląd Elektrotechniczny, artykuł w recenzji.

[8] S. Osowski, T. Markiewicz, M. Kruk, W. Kozłowski, *Metody sztucznej inteligencji do wspomagania diagnostyki patologii tkanek*, red. A. Michalski, *Metrologia w medycynie — wybrane zagadnienia*, WAT, Warszawa, 2011, 91-126.

E. MAJDA, A.P. DOBROWOLSKI, B.L. SMÓLSKI

Modeling and optimization of features generator for speaker recognition systems

Abstract. The paper presents issues related to modeling and optimization of the features generator for the speaker recognition system (ASR – *Automatic Speakers Recognition*). Parameterization's stage of the speech signal (features generation) is fundamental in this type of systems, due to the fact that the unique vector of features is crucial in the process of recognition. The task is to describe the speech signal using descriptors as little as possible, without loss of relevant information to the speaker recognition. In addition, parametrization should have robust to acoustic and technical registration conditions and the recorded linguistic material. The research presented in this paper is focused primarily on the multicriteria optimization of selected parameters of the features generator based on cepstral analysis, additionally allowing features selection. Finally, evaluation of the results was based on the analysis of main components, a set of descriptors for the samples voice acquired from 24 speakers.

Keywords: Automatic Speaker Recognition (ASR), cepstral analysis, features extraction, features selection, principal component analysis (PCA)