ASR, ekstrakcja cech,
analiza cepstralna, PCA

DOBROWOLSKI Andrzej P.
MAJDA Ewelina[1]

ANALIZA CEPSTRALNA W SYSTEMACH ROZPOZNAWANIA MÓWCÓW

W prezentowanym referacie poruszono problematykę systemu rozpoznawania mówcy (ASR – ang. Automatic Speakers Recognition). Sygnał mowy w postaci pierwotnej charakteryzuje się dużą nadmiarowością, dlatego konieczna jest ekstrakcja specyficznych cech sygnału, za pomocą których możliwy będzie efektywny opis właściwości sygnał, ważnych z punktu widzenia rozpoznawania mówcy. Z tego względu parametryzacja sygnału w procesie rozpoznawania jest niezwykle istotna. Autorzy podjęli się próby wyboru optymalnego (najbardziej dyskryminującego) zestawu parametrów opisujących sygnału w oparciu o metody przetwarzania homomorficznego. Badania koncentrowały się przede wszystkim na ocenie użyteczności analizy cepstralnej sygnału mowy w systemach rozpoznawania na podstawie pozyskanych w postaci cyfrowej próbek głosu.

CEPSTRAL ANALYSIS IN SPEAKER RECOGNITION SYSTEMS

The present paper addresses issues related to the speaker recognition system (ASR – Automatic Speakers Recognition). In its primary form, a speech signal is characterized by a high redundancy, so it is necessary to extract the specific features of the signal that would allow to efficiently describing the properties thereof that are important from the viewpoint of speaker recognition. Therefore, parameterization of the signal in the process of recognition is extremely important. The authors have attempted to select the optimal (most discriminating) set of parameters describing the signal by using a homomorphic processing method. The study has primarily focused on assessing applicability of the cepstral analysis in speakers recognition systems based on the acquired digitized voice samples.

1. WPROWADZENIE

Automatyczne rozpoznawanie osób znajduje zastosowanie we wszystkich systemach, które dostarczają usług lub informacji zastrzeżonych, szczególnie wtedy, gdy niezbędny jest wysoki stopień bezpieczeństwa tych systemów. Do wzrostu zapotrzebowania na tego typu systemy przyczynia się nie tylko rozwój szeroko rozumianych technik biometrycznych, ale także telekomunikacji i Internetu. Niewątpliwą zaletą systemów rozpoznawania osób na podstawie indywidualnych charakterystyk głosu, jest fakt, że

[1]Wojskowa Akademia Techniczna, Wydział Elektroniki; 00-908 Warszawa; ul. Gen. Kaliskiego 2
E-mail: Andrzej.Dobrowolski@wat.edu.pl, Ewelina.Majda@wat.edu.pl

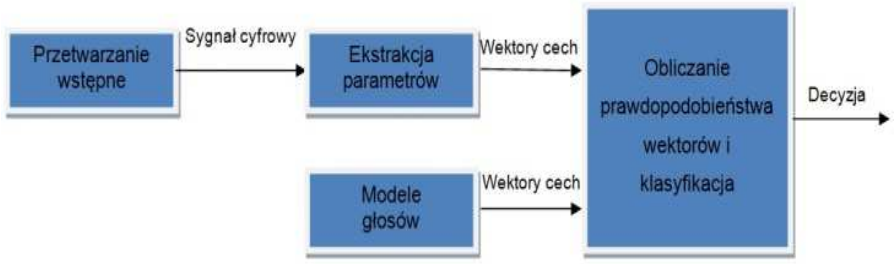
atrybuty te nie mogą zostać zagubione, bądź też zapomniane. W codziennych kontaktach identyfikacja osób na podstawie głosu jest czynnością z łatwością wykonywaną. Powszechność i naturalność tego zjawiska powoduje, że na ogół nie zdajemy sobie sprawy, jakie cechy wypowiedzi są w tym naturalnym procesie uwzględniane i dopiero próba przeniesienia tej czynności na grunt urządzeń technicznych uświadamia nam pełny zakres trudnych do rozwiązania problemów. Wynika to z faktu, że zmysł słuchu i układ nerwowy człowieka są wysoce wyspecjalizowane i wyuczone w odbiorze i analizie sygnału mowy, ale niestety zachodzące przy tym procesy nie są do końca poznane.

Wypowiedź słowna oprócz informacji o treści wypowiedzi, niesie również informacje związane z wewnętrzną strukturą jej źródła. Sygnał mowy jest nośnikiem zarówno cech fizjologicznych, jak i behawioralnych, dzięki czemu należy do biometryk zapewniających wysoki stopniu zróżnicowania. Te osobnicze informacje odzwierciedlające indywidualne cechy głosu mówcy są wynikiem różnic w budowie organów artykulacyjnych (traktu głosowego) u różnych osób, nawyków nabytych w trakcie nauki mówienia i stopnia opanowania danego języka. W praktyce istnieją mniejsze lub większe związki pomiędzy charakterystykami biometrycznymi i cechami mówcy takimi jak jego płeć, wiek, stan zdrowia, nastrój, pochodzenie, język narodowy. Z tych właśnie względów analiza głosu jest przedmiotem badań specjalistów z wielu dziedzin, ale pomimo trwających już dziesiątki lat badań, sygnał mowy należy uznać za bardzo złożony i trudny do pełnej (czyli analogicznej do analizy wykonywanej przez zmysł słuchu) interpretacji.

2. PROCEDURA ASR

Automatyczne rozpoznawanie mówcy zwane także automatycznym rozpoznawaniem głosów, jest procesem realizującym szereg reguł decyzyjnych na mierzalnych cechach sygnału mowy, mających na celu określenie czy dana wypowiedź należy do określonego mówcy lub zbioru mówców. W ogólności procedurę rozpoznawania osób można podzielić na 3 etapy (rys. 1). Blok przetwarzania wstępnego odpowiada za odbiór sygnału z mikrofonu oraz jego wstępne przetworzenie, uwzględniające poprawę jakości sygnału. W drugim etapie następuje analiza sygnału mowy, w wyniku, której otrzymuje się wartości parametrów niosących informację o indywidualnych cechach głosu mówcy niezależne od treści wypowiedzi. Ostatni etap klasyfikacji dokonuje się na podstawie podobieństwa uzyskanych parametrów próbek sygnału do ich odpowiedników określonych wcześniej (w tzw. procesie nauczania) dla poszczególnych osób. Wynikiem działania systemu jest binarna decyzja o rozpoznaniu mówcy, bądź też jego odrzuceniu. Z punktu widzenia systemu rozpoznawania mówcy najważniejszym etapem jest zapewnienie odpowiedniego zestawu parametrów, aby rozpoznanie było możliwe. Podstawowym wymaganiem dla takiego zestawu jest zapewnienie dyskryminacji głosów różnych osób na podstawie wartości parametrów oraz powtarzalność wartości tych parametrów dla różnych wypowiedzi tej samej osoby. Za lepszy parametr uważa się taki, którego wartości są dokładnie powtarzalne (lub bardzo zbliżone) dla wypowiedzi tego samego mówcy i stosunkowo znacznie różniące się dla wypowiedzi różnych mówców. W celu ekstrakcji odpowiednich parametrów z sygnału mowy należy pozyskany sygnał poddać procesowi parametryzacji, gdyż od niej w dużej mierze zależy skuteczność oraz szybkość działania całego systemu rozpoznawania mówcy. Mając do czynienia z dużą ilością różnorodnych

parametrów poszukiwane są pewne metody wyboru optymalnego (najbardziej dyskryminującego) zestawu parametrów opisujących sygnał.



Rys. 1. Schemat procedury rozpoznawania mówców

3. METODY OPISU SYGNALU MOWY

Pierwotną i podstawową formą, w której występuje sygnał mowy jest postać czasowa. W tej postaci w istocie zawarte są wszystkie niezbędne do analizy i rozpoznania elementy, ale w raczej niedogodnej formie. Jest to spowodowane dużą nadmiarowością informacji zawartej w takiej postaci. Wzorując się na „ludzkim" sposobie analizy sygnału mowy, znaczna część metod komputerowych bazuje na analizie częstotliwościowej, która jest podstawową i rutynowo stosowaną metodą opisu tego typu sygnałów, a także punktem wyjścia do innych zaawansowanych metod parametryzacji.

W procesie generacji sygnału mowy bierze udział głośnia oraz trakt głosowy obejmujący w szczególności jamę ustną i nosową oraz język i usta. Zasadniczą rolę w procesie mówienia (i oddychania) odgrywają fałdy głosowe, często zwane strunami głosowymi, a właściwie ich krawędzie czyli więzadła głosowe. Szparę pomiędzy więzadłami nazywa się szparą głosową (szparą głośni), a wraz z przyległymi fałdami głośnią. Podczas spokojnego oddychania oraz w czasie artykulacji bezdźwięcznych elementów mowy więzadła są rozsunięte i powietrze swobodnie przepływa przez szparę głośni. W czasie wymawiania głosek dźwięcznych więzadła, na skutek dochodzących do nich impulsów nerwowych, na przemian zwierają się i rozwierają pod naporem sprężanego powietrza. Obserwowana w tym czasie gołym okiem szpara pomiędzy fałdami głosowymi jest złudzeniem optycznym spowodowanym bezwładnością wzroku ludzkiego, który nie jest w stanie zarejestrować szybko następujących po sobie faz zamykania i otwierania głośni. Obserwacja w zwolnionym tempie pokazuje, że więzadła zwierają się rytmicznie aż do pełnego zamykania głośni. Proces generacji dźwięku krtaniowego nazywany bywa fonacją (udźwięcznianiem).

Określająca wysokość głosu liczba cykli zwarć i rozwarć więzadeł na sekundę, zależy od ich długości, grubości i napięcia (a te od płci i wieku). 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 [2]. Częstotliwość podstawowa w mowie zmienia się od 60 do 200 Hz u mężczyzn i od 180 do 400 Hz u kobiet. W przypadku śpiewu zakres zmian

częstotliwości podstawowej, zwany skalą głosu, jest znacznie szerszy. Zgodnie z nomenklaturą muzyczną w klasyfikacji podstawowej wyróżnia się trzy zasadnicze głosy męskie: bas (73-294 Hz), baryton (98-392 Hz) i tenor (123-494 Hz) oraz trzy głosy żeńskie: alt (165-652 Hz), mezzosopran (195-784 Hz) i sopran (247-900 Hz).

Strumień powietrza tłoczony przez głośnię jest modyfikowany w trakcie przejścia przez trakt głosowy, którego charakterystyka amplitudowo-częstotliwościowa charakteryzuje się kilkoma maksimami nazywanymi formantami. Częstotliwości tych maksimów są chwilowymi częstotliwościami rezonansowymi traktu głosowego wynikającymi z bieżącego stanu procesu artykulacji. Przyjmując, że dla quasi-stacjonarnych fragmentów mowy trakt głosowy jest układem liniowym niezmiennym w czasie, sygnał mowy można przedstawić jako splot impulsowego pobudzenia generowanego w głośni i odpowiedzi impulsowej traktu głosowego.

Ponieważ transformata Fouriera równomiernie poprzesuwanych impulsów Diraca

$$\mathrm{III}(\mathrm{t}) = \sum_{n=-\infty}^{\infty} \delta(t - nT_0) \tag{1}$$

jest także sumą impulsów Diraca

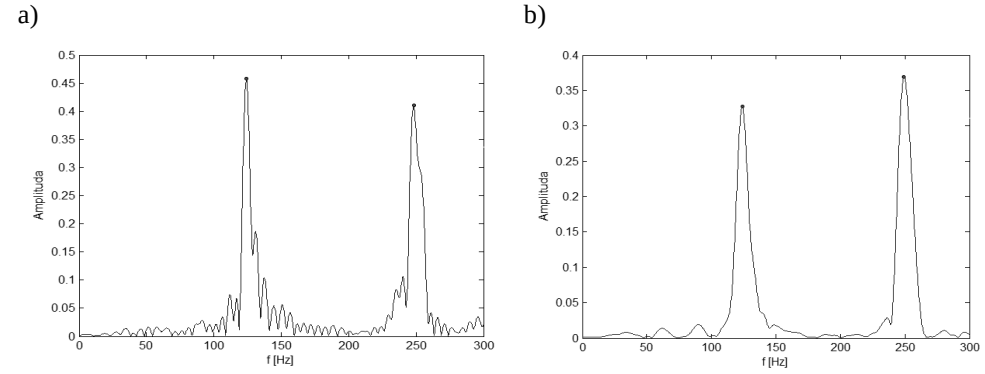
$$\omega_0 \cdot \sum_{m=-\infty}^{\infty} \delta(\omega - m\omega_0); \qquad \omega_0 = \frac{2\pi}{T_0} \tag{2}$$

to widmo dźwięku krtaniowego jest ciągiem impulsów, przy czym jeśli odstęp impulsów w dziedzinie czasu wynosi T_0, to odstęp w dziedzinie częstotliwości wynosi $F_0 = 1/T_0$.

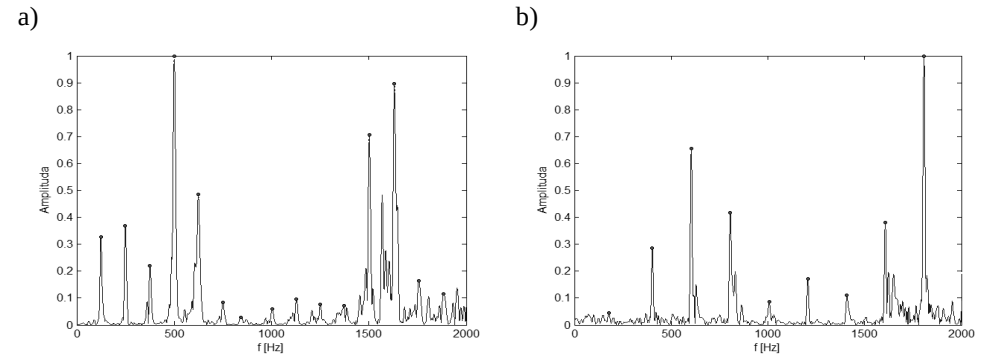
Przyjęcie liniowego modelu traktu głosowego, w którym pobudzenie splata się z odpowiedzią impulsową filtru w dziedzinie czasu pozwala – w świetle (1) i (2) – na stwierdzenie, że widmo fragmentów mowy dźwięcznej jest iloczynem rozłożonych w odstępie F_0 na osi częstotliwości impulsów Diraca (idealizowane widmo impulsów emitowanych z głośni) i transmitancji traktu głosowego. Skończony czas otwarcia głośni w trakcie fonacji uwzględniany jest w rozważaniach teoretycznych w postaci dodatkowego członu w transmitancji traktu głosowego. Podczas praktycznych badań sygnału mowy, fragmenty sygnału wycinane są za pomocą wybranej funkcji okienkującej, której widmo splata się z widmowymi impulsami Diraca i w konsekwencji w miejscu spodziewanych impulsów Diraca pojawia się powielone na każdym z nich widmo okna, co zilustrowano na rys. 2.

Na rys. 3 przedstawiono widma amplitudowe głoski *a* wypowiadanych przez mężczyznę i kobietę. Łatwo zauważyć, a ściśle potwierdziły to badania wstępne, że na podstawie widma amplitudowego łatwiej jest odróżnić wypowiadane głoski niż mówców. Istotną informacją rozróżniającą mówców jest częstotliwość podstawowa dźwięku, która – co oczywiste w przypadku porównywania wypowiedzi mężczyzny i kobiety – może ewentualnie posłużyć jako parametr różnicujący. Jednak w przypadku porównania np. dwóch mężczyzn jest to informacja o niewielkiej użyteczności, tym bardziej, że częstotliwość podstawowa fluktuuje w tak intonacji zdania. Na rys. 2 oraz 3 wyraźnie widoczna jest okresowość widma wynikająca z impulsów dźwięku krtaniowego, można

więc obliczyć odwrotną transformatę Fouriera z modułu widma i na jej podstawie wyznaczyć okres podstawowy pobudzenia krtaniowego. Ponieważ jednak sygnał jest zmodulowany w amplitudzie przez funkcję przenoszenia traktu głosowego, korzystniej jest wyznaczyć najpierw logarytm z modułu widma, a dopiero potem poddać go odwrotnej transformacji Fouriera, gdyż w ten sposób multiplikatywny związek pobudzenia i traktu głosowego zastąpiony zostanie związkiem addytywnym, co znacznie upraszcza późniejszą separację obu składników. Przedstawione rozumowanie prowadzi wprost do metod przetwarzania homomorficznego, a w szczególności do koncepcji cepstrum.



Rys. 2. Widmo głoski *e* wypowiadanej głosem męskim: a) okno prostokątne, b) okno Hanninga



Rys. 3. Widma głosek e wypowiadanych: a) głosem męskim, b) głosem żeńskim; zastosowano okno Hanninga

4. PODSTAWY ANALIZY CEPSTRALNEJ

Jedną ze szczególnych metod parametryzacji jest analiza cepstralna opierająca się na tzw. technice homomorficznej. Cepstrum zespolone zdefiniowane jest następująco:

$$c_z(t) = \mathcal{F}^{-1}\left\{ \ln\left(\mathcal{F}\{x(t)\} \right) \right\} \tag{3}$$

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

$$c(t) = \mathcal{F}^{-1}\left\{ \ln\left(\left| \mathcal{F}\{x(t)\} \right| \right) \right\} \tag{4}$$

co dla sygnałów dyskretnych sprowadza się do postaci

$$c(n) = IDFT\left(\ln\left(\left| DFT\left(x(n) \cdot w(n) \right) \right| \right) \right) \tag{5}$$

i ostatecznie

$$c(n) = \frac{1}{N}\sum_{m=0}^{N-1} C(m) \mathrm{e}^{\mathrm{j}2\pi\frac{mn}{N}} = \frac{1}{N}\sum_{m=0}^{N-1} \ln\left(\left| \sum_{n=0}^{N-1} x(n)\, w(n) \mathrm{e}^{-\mathrm{j}2\pi\frac{mn}{N}} \right| \right) \mathrm{e}^{\mathrm{j}2\pi\frac{mn}{N}} \tag{6}$$

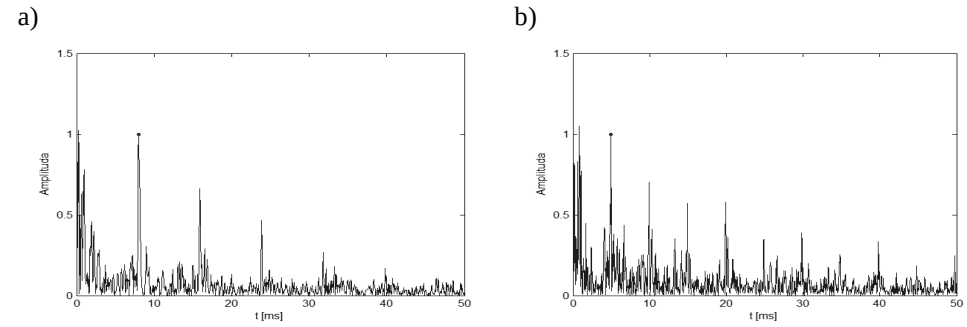
Ze względu na okresowość jądra transformaty Fouriera, logarytm z modułu widma amplitudowego $C(m)$ jest okresowy i jednocześnie spełnia zależność

$$C(-m) = C(N-m) \tag{7}$$

Jest więc funkcją parzystą (symetria względem osi 0y), zatem w jego rozwinięciu występują tylko funkcje kosinusoidalne (parzyste). Jest więc bez 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 zlogarytowanego widma amplitudowego. Obserwując widmo amplitudowe sygnału mowy można łatwo zauważyć, że jest ono złożone z czynnika szybkozmiennego wynikającego z pobudzenia i wolnozmiennego modulującego amplitudę kolejnych impulsów wynikających z 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 transmitancja 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.

Cepstra rzeczywiste odpowiadające widmom z rys. 3 przedstawione są na rys. 4. Informacja związana z transmitancją traktu głosowego jest skupiona w okolicy czasu zerowego, a zatem w tym obszarze należy poszukiwać zwięzłej informacji na temat tego *co jest mówione*. Natomiast dla czasów powyżej okresu dźwięku krtaniowego informacja o tym co jest mówione jest zminimalizowana, pozostaje jedynie czytelna informacja dotycząca dźwięku krtaniowego. Ponieważ dźwięk krtaniowy związany jest ściśle z budową anatomiczną krtani i głośni, jest wiec zarazem dobrym nośnikiem informacji osobniczej. Klasyczna metoda rozplotu cepstralnego, w przypadku analizy pod kątem rozpoznawania mówcy, polega więc na usunięciu niepożądanego składnika poprzez wyzerowanie próbek cepstrum dla pseudoczasu w okolicach zera.

Przydatność cepstrum rzeczywistego do celów rozpoznawania mówcy można łatwo zauważyć analizując wzrokowo przebiegi przedstawione na rys. 4 – informacje o wypowiadanej głosce zacierają się, natomiast zarysowuje się wyraźne zróżnicowanie w zależności od mówcy.



Rys. 4. Moduły cepstrum rzeczywistego głosek e wypowiadanych: a) głosem męskim, b) głosem żeńskim; zastosowano okno Hanninga

5. PRZEPROWADZONE BADANIA

Obiecujące wyniki początkowych eksperymentów wykorzystujących analizę cepstralną sygnału mowy dały możliwość szerszego badania sygnału mowy w oparciu o unormowane cepstrum reczywiste. Podjęto więc próbę wyekstrahowania z każdego wycinka czasowego mowy dźwięcznej zbioru 10 cepstralnych cech dystynktywnych. Ze względu na to, że istotna informacja związana z mówcą zawarta jest jedynie w tzw. dźwięcznych fragmentach mowy zarejestrowany materiał fonetyczny obejmował samogłoski *a*, *e*, *i*, *o*, *u*. powtarzane przez każdego z uczestników 3 razy. Grupa biorąca udział w doświadczeniu składała się z 5 mężczyzn i 5 kobiet. Nagrania zarejestrowano w warunkach pokojowych z użyciem uniwersalnego mikrofonu biurkowego MT383, karty dźwiękowej komputera oraz oprogramowania Matlab. Sygnały próbkowane były z częstotliwością 22050 Hz, a rozdzielczość amplitudowa wynosiła 16 bitów. Dla pozyskanych danych wejściowych, przy użyciu oprogramowania *Matlab* wyznaczono widma amplitudowe z wykorzystaniem *szybkiej transformacji Fouriera* (*FFT*). Algorytm *FFT* o podstawie 2 jest bardzo efektywną

procedurą wyznaczania *dyskretnej transformacji Fouriera* (*DFT*), pod warunkiem, że liczba próbek sygnału wejściowego jest potęgą 2. Aby móc wykorzystać wydajność obliczeniową FFT należało w każdym z analizowanych sygnałów zapewnić liczbę próbek wejściowych będącą całkowitą potęga liczby 2. Przyjęto 65536-punktowe FFT, opierając się na założeniu, że taka ilość próbek zapewni wystarczająco gęstą ziarnistość wynikowego widma. W celu minimalizacji zjawiska przecieku zastosowano okno Hanninga, zdefiniowane wzorem:

$$w(n) = 0{,}54 - 0{,}46\cos\left(\frac{2\pi n}{N}\right) \tag{8}$$

Do analiz wykorzystywane było jedynie widmo amplitudowe sygnału niosące informacje użyteczne z punktu widzenia systemów automatycznego rozpoznawania mówcy. Każde z analizowanych widm było ponadto normalizowane, a wyświetlany przebieg ograniczany był do połowy częstotliwości próbkowania ze względu na symetryczność DFT. Następnym etapem było obliczenie cepstrum rzeczywistego z wykorzystaniem zależności (6).
Jako cechy charakterystyczne zdecydowano się wybrać częstotliwość podstawową mówcy, która jest odwrotnością pierwszego maksimum w cepstrum oraz wartości n-1 kolejnych maksimów. Przyjęto n=9, w wyniku czego uwzględniono 8 kolejnych wartości maksymalnych unormowanego cepstrum rzeczywistego. Dla każdego mówcy na podstawie zarejestrowanego pojedynczej wypowiedzi, zawierającej 5 samogłosek dokonywano uśredniania zbioru cech cepstralnych i dodatkowo uzupełniono go o odchylenie standardowe częstotliwości podstawowej. Pełny zbiór cech dystynktywnych określony jest zależnościami:

$$\begin{cases} F_{av} = \dfrac{1}{N}\displaystyle\sum_{j=1}^{N} F_j, \quad F_j = \dfrac{1}{N_j T_p} \\ \sigma = \sqrt{\dfrac{\displaystyle\sum_{j=1}^{N}(F_j - F_{av})^2}{N-1}} \\ c_i = \dfrac{1}{N}\displaystyle\sum_{j=1}^{N} c_j, \qquad i = 1,2,...,8 \end{cases} \tag{9}$$

gdzie:
F_1 – częstotliwość podstawowa, dla każdego kolejnego segmentu analizy mowy (dla tego samego mówcy)
N_1 – numer próbki odpowiadającej pierwszemu maksimum ceptrum, dla każdego kolejnego segmentu analizy mowy (dla tego samego mówcy)
T_p – okres próbkowania,
N- liczba analizowanych segmentów dla tego samego mówcy
c_i – wartości kolejnych 8 cech każdego mówcy,

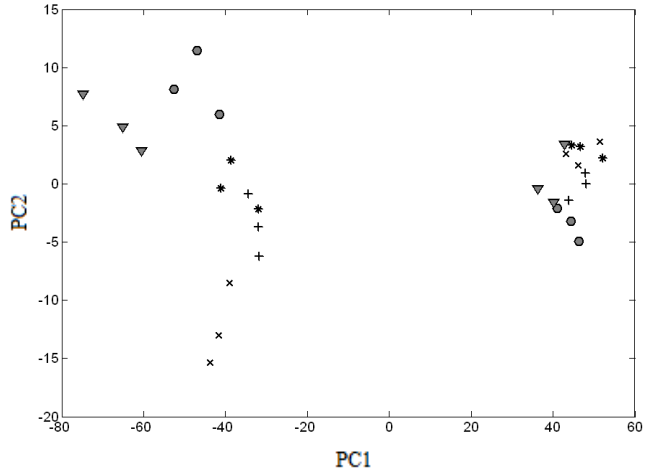
c_j – wartość cechy dla każdego kolejnego analizowanego segmentu mowy (dla tego samego mówcy),
N – ilość analizowanych segmentów dla danego mówcy,
σ – odchylenie standardowe częstotliwości podstawowej,
F_{av} – wartość średnia częstotliwości podstawowej.

W wyniku tej operacji otrzymano 10-wymiarowy wektor cech, będący *VoicePrintem* analizowanego mówcy. Zgodnie z założeniami dla każdego mówcy otrzymano 3 oddzielne wektory cech. Ze względu na dużą ilość informacji zawartej w danych wejściowych, jakimi w naszym przypadku są wektory cech opisujące każdego z uczestników, do redukcji wymiaru zdecydowano się zastosować metodę PCA. Analiza składowych głównych (*Principal Component Analysis – PCA*) jest metodą statystyczną określoną przez przekształcenie liniowe $y=\boldsymbol{W}x$ transformujące opis stacjonarnego procesu stochastycznego w postaci wektora x, w ten sposób, że przestrzeń wyjściowa o zredukowanym wymiarze zachowuje najważniejsze informacje dotyczące procesu. Innymi słowy PCA zamienia dużą ilość informacji zawartej we wzajemnie skorelowanych danych wejściowych w zbiór statystycznie uporządkowanych niezależnych składników według ich ważności. PCA jest przykładem systemu uczącego się bez nadzoru. Zadaniem takiego systemu jest opisanie obserwowanych danych (wyekstrahowanych wektorów cech) na podstawie wyłącznie ich samych. W wyniku PCA otrzymano dwie składowej główne, które są liniowymi funkcjami zmiennych oryginalnych. Wyniki transformacji PCA dla 10 mówców przedstawione są na rys. 5. Na podstawie przeprowadzonej analizy można w pierwszej kolejności zauważyć, że transformacja PCA pozwala na bezproblemowe odróżnienie płci mówiącego. Kobiety zauważalnie grupują się po lewej stronie płaszczyzny, natomiast mężczyźni szeregują się po stronie prawej. Na rys. 6 przedstawiono powiększone fragmenty powierzchni z rys. 5 – osobno dla kobiet i mężczyzn (skala obu rysunków jest różna).

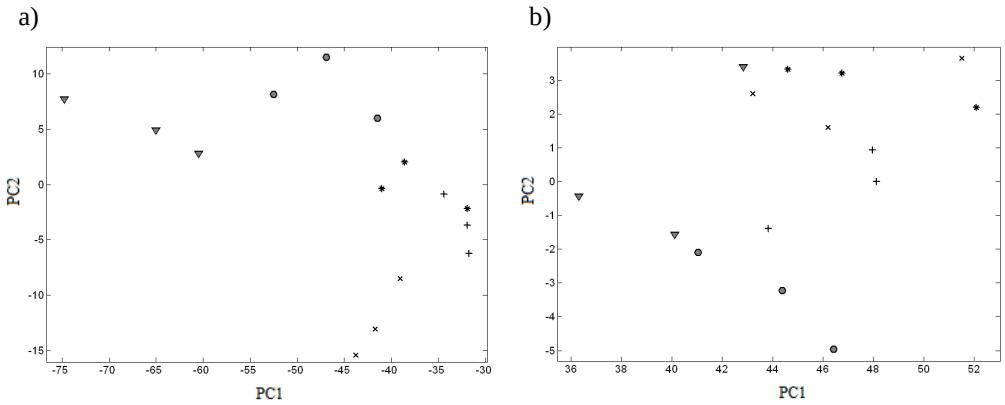
Analizując otrzymane wyniki można stwierdzić, że transformacja PCA umożliwia odseparowanie poszczególnych mówców od siebie, co jest wyraźnie zauważalne na przedstawionych wykresach. Dodatkowo należy podkreślić, że uśrednienie parametrów cepstralnych następowało z niewielkiej liczby segmentów, co miało bezpośredni wpływ na stosunkowo duży rozrzut punktów obrazujących poszczególne osoby.

6. WNIOSKI

Przeprowadzone eksperymenty pozwoliły na pozytywną ocenę użyteczności analizy cepstralnej w odniesieniu do parametryzacji sygnału mowy. Po zastosowaniu transformacji PCA zaobserwowano ewidentne oddzielenie się poszczególnych mówców na płaszczyźnie. Każda z osób praktycznie koncentruję się w rozdzielnych obszarach. Zaobserwować można jednak nieznaczne zachodzenie na siebie 2 mówców zarówno w przypadku mężczyzn jak i kobiet. Jest to spowodowane małą liczbą uśrednień dla każdego uczestnika. Można więc oczekiwać, że ich większa ilość jeszcze bardziej powinna podkreślać cechy wspólne i o ile parametry wyznaczone w ramach jednego segmentu mogą mieć niewielką korelację z numerem mówcy, to po uśrednieniu korelacja ta powinna być znacząco wyższa. W szczególności autorzy zajmują się obecnie ustaleniem stabilnego kryterium oceny segmentu, który będzie zapewniał poprawną generację cech.



Rys. 5. Wyniki transformacji PCA (po lewej stronie głosy żeńskie po prawej męskie)



Rys. 6. Wynik transformacji PCA dla kobiet (a) i mężczyzn (b)

7. BIBLIOGRAFIA

[1] A. Shomali, *Rozpoznawanie mówcy na podstawie długookresowego histogramu amplitud sygnału mowy*, Rozprawa doktorska, AGH, 1999

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

[3] J. Ming, T. Hazen, J. R Glass, D. A. Reynolds, *Robust Speaker Recognition In Noisy Conditions*, IEEE Transactions on Audio, Speech, and Language Processing, vol. 15, no. 5, 2007, pp. 1711-1723

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