ZESZYTY NAUKOWE POLITECHNIKI ŚLĄSKIEJ 2001
Seria: AUTOMATYKA z. 133 Nr kol. 1526

Ewa BIELIŃSKA

METODY EKSTRAKCJI CECH OSOBOWYCH MÓWCY

Streszczenie. Przedstawiono metody ekstrakcji cech osobowych mówcy, bazujące na analizie cepstralnej i metodzie liniowej predykcji. Zaproponowano metodę ekstrakcji cech wykorzystującą częstotliwość występowania powtarzalnych biegunów modelu wielomianowego w poszczególnych ramkach wypowiedzi. Porównano właściwości ośmiu metod ekstrakcji cech ze względu na miarę rozproszenia wewnętrznego i miarę rozproszenia zewnętrznego. Podważono zasadność wymiarowości wektora cech w metodach cepstralnych.

METHODS OF SPEAKER INFORMATION EXTRACTING

Summary. The article is concerned with methods of speaker information extracting, that are based on cepstral analysis and linear predictive coding. A method using poles location in frames was proposed and compared with eight other methods of speaker information extracting. The comparison was made due to a measure of internal and external dispersion. Dimension of the vector of features, applied in methods based on cepstral analysis was discussed.

1. Wprowadzenie

Od pewnego czasu obserwowany jest wzrost zainteresowania zagadnieniem rozpoznawania mowy i równolegle - zagadnieniem rozpoznawania mówcy. Rozpoznawanie mówcy stosuje się w wielu systemach zabezpieczeń rozpoznających uprawnienia osoby wydającej polecenie dotyczące uruchomienia sprzętu takiego, jak: samochód, komputer, otwieranie drzwi do pomieszczeń dostępnych tylko osobom uprawnionym itp. Rozpoznanie mówcy na podstawie próbki jego głosu może znaleźć zastosowanie przy biometrycznej identyfikacji osób, ale dotychczas uzyskiwana dokładność jest jeszcze za mała, by stanowić konkurencję do rozpoznawania osób na podstawie ich linii papilarnych. Rozpoznanie osoby wypowiadającej tekst może mieć duże znaczenie w systemach rozpoznających mowę, współpracujących z dużą liczbą użytkowników. Każdy z takich systemów umożliwia wywołanie pewnej, ograniczonej liczby komend głosowych. Z reguły są to komendy do edycji i dyktowania tekstu. Najpierw mikrofon przetwarza głos mówiącego, odbierany jako drgania powietrza, na postać analogową - zmienny prąd elektryczny. Karta dźwiękowa uzupełniona odpowiednim oprogramowaniem przetwarza sygnał analogowy na sygnał cyfrowy i od tego momentu zadanie

rozpoznania mowy czy zadanie rozpoznania mówcy pozostaje domeną odpowiednio wykonanego oprogramowania. Żaden z komercyjnych systemów nie obsługuje języka polskiego, a dyktowanie poleceń w jednej z dostępnych wersji językowych: angielskiej, niemieckiej lub francuskiej wymaga od użytkownika perfekcyjnej wymowy, gdyż w przeciwnym przypadku pojawiają się trudności z porozumieniem się z komputerem. Rozpoznawany tekst, oprócz warstwy znaczeniowej, zawiera dane charakteryzujące osobę wypowiadającą ten tekst. Wczytanie profilu odpowiedniego użytkownika daje szansę lepszej analizy poleceń i sygnału mowy ciągłej. Z punktu widzenia automatycznej identyfikacji mówca generuje sygnał scharakteryzowany zbiorem cech. Cechy te charakteryzują zarówno mówcę, jak i wypowiadany przez niego tekst. Z kilkusekundowego fragmentu wypowiedzi należy wyodrębnić te cechy, które charakteryzują samego mówcę a nie specyficzny fragment tekstu. Rozpoznanie osoby na podstawie fragmentu jej wypowiedzi dokonane automatycznie, to znaczy bez udziału człowieka, nazywa się automatyczną identyfikacją mówcy. Przez analogię do człowieka, który może rozpoznać tylko tę osobę, którą zna, automatyczna identyfikacja mówcy polega na wybraniu ze zbioru cech charakteryzujących różnych mówców zestawu cech najbardziej zbliżonych do cech wypowiedzi badanej osoby.

2. Założenia

Mowa jest ciągiem złożonych dźwięków powstających w wyniku pobudzania kanału głosowego, zwanego inaczej torem akustycznym. Źródłem energii do wytwarzania tych dźwięków jest powietrze wydychane z płuc. W przypadku głosek dźwięcznych (a, e, i, y...) wiązadła głosowe przetwarzają strumień powietrza na quasi-okresowy ciąg impulsów. W przypadku głosek szczelinowych (trących, szumowych – sz., cz., c, s, f...) pobudzenie kanału głosowego ma charakter szumu powstającego w wyniku turbulencji strumienia powietrza przez przewężenie w kanale głosowym. Głoski zwarte (wybuchowe - p, b, k, t,...) powstają w wyniku całkowitego zamknięcia toru akustycznego, wytworzenia nadciśnienia powietrza i gwałtownego uwolnienia tego powietrza. W celu zastosowania sformalizowanych metod rozpoznawania mówcy należy przyjąć pewne założenia i określić model generowania sygnału mowy.

1. Zakłada się, że wymienione sygnały pobudzające można łącznie potraktować jako szerokopasmowe pobudzenie kanału głosowego.
2. Kanał głosowy można opisać modelem filtru wolnozmiennego w czasie, modyfikującego widmo pobudzenia przez swoją charakterystykę częstotliwościową.
3. Zakłada się wzajemną niezależność źródła pobudzającego i kształtu kanału głosowego. Na mocy tego założenia mechanizm tworzenia mowy można przedstawić modelem pokazanym na rys.1.
4. Ponieważ w czasie generacji ciągłego sygnału mowy kształt kanału głosowego ulega względnie powolnym zmianom, zakłada się, że w przedziale czasowym 10-20 msec. własności filtru pozostają stałe. Wewnątrz każdego z takich przedziałów filtr może być opisany odpowiedzią impulsową, $h(t)$, charakterystyką częstotliwościową, $H(\omega)$ lub zbiorem współczynników filtru.
5. Zakłada się, że model kanału głosowego, jak i charakter pobudzenia zawierają informacje o cechach osobowych mówcy. Cechy pobudzenia znajdują głównie zastosowanie w analizie foniatrycznej. Charakterystyczne cechy kanału głosowego mówcy zawarte są w pierwszych współczynnikach cepstralnych.

6. W dalszym ciągu zakłada się, że nagrany fragment wypowiedzi jest wystarczającą reprezentacją cech charakteryzujących mówcę.

Dąży się do otrzymania opisu czy modelu wzorców mówcy w przestrzeni cech, który może być wykorzystany do identyfikacji mówcy na podstawie testowej próbki wypowiedzi. Niewątpliwie ważnym krokiem w procesie identyfikacji jest wydobycie z wypowiedzi informacji wystarczającej do rozpoznania mówcy, ale z drugiej strony forma i rozmiar uzyskanych informacji muszą umożliwiać efektywne modelowanie mówcy. Ilość danych generowanych nawet przy krótkiej wypowiedzi jest bardzo duża.

Rys. 1. Schemat generacji sygnału mowy
Fig.1. Speech generation scheme

Zazwyczaj sygnały mowy próbkowane są z częstotliwością 8 *kHz* lub wyższą. Przy wykorzystaniu 8 bitów na próbkę otrzymuje się dziesiątki tysięcy bajtów na kilkusekundową wypowiedź. O ile tak ogromna ilość informacji potrzebna jest do scharakteryzowania fali głosowej, to zasadnicze cechy charakteryzujące proces mówienia zmieniają się względnie wolno. Sygnał mowy może być sparametryzowany w obrębie względnie długich, bo trwających 10-20 msec. fragmentów mowy, zwanych ramkami. Jeżeli wypowiedź z 20ms ramki może być reprezentowana przez 14-wymiarowy wektor cech, mówi się, że osiągnięto poziom redukcji danych :

$$r = \frac{20 \cdot 10^{-3} \cdot 8 \cdot 10^{3}}{14} = 11.4, \tag{1}$$

przy częstotliwości próbkowania 8kHz. Proces redukcji danych przy jednoczesnym zachowaniu klasyfikacji informacji nazywa się ekstrakcją cech. Uzyskana w wyniku ekstrakcji cech *n*-wymiarowa przestrzeń cech nazywa się przestrzenią mówcy. W procesie rozpoznawania mówcy można więc wyodrębnić trzy etapy:

1. ekstrakcję cech,
2. określenie modelu,
3. przyrównanie wzorców i wybór właściwego mówcy według założonego kryterium.

Metody ekstrakcji cech osobowych mówcy można podzielić w zależności od wykorzystywanego aparatu matematycznego na:

- metody cepstralne, dla których wymagane przekształcenia sygnału mowy wykonywane są w dziedzinie częstotliwości;
- metody liniowej predykcji, w których przekształcenia potrzebne do analizy sygnału mowy dokonywane są w dziedzinie czasu.

3. Metody cepstralne

Analiza cepstralna jest metodą przetwarzania sygnału mowy wykorzystującą uogólnioną zasadę superpozycji zdefiniowaną dla systemów liniowych. Podstawą metody jest założenie, że widmo sygnału mowy można traktować jako iloczyn składowej źródła sygnału oraz składowej opisującej kanał głosowy. Sygnał źródłowy ze swej natury zmienia się szybciej niż sygnał charakteryzujący kanał głosowy. Na mocy założenia, że informacja o cechach osobowych mówcy zawarta jest w sygnale opisującym kanał głosowy, dla dalszej analizy celowe

jest oddzielenie sygnału źródłowego od sygnału kanału głosowego. Ponieważ rozdzielenie czynników jest trudniejsze niż rozdzielenie składników sumy, zamiast bezpośrednich sygnałów rozpatruje się ich logarytmy. Ponieważ logarytm iloczynu jest równy sumie logarytmów poszczególnych czynników, zamiast rozdzielać czynniki iloczynu dokonuje się rozdziału obu składników logarytmu widma sygnału mowy, różniących się charakterystykami częstotliwościowymi. W procesie rozpoznawania mówcy wykorzystuje się krótkoterminowe widma, wyznaczane dla 10-20 msec. ramek, na które została podzielona wypowiedź. Pierwszym krokiem algorytmu ekstrakcji cech na podstawie fragmentu wypowiedzi jest sprowadzenie zarejestrowanego sygnału mowy, *s(t)*, do dziedziny częstotliwości, *S(ω)*, np. za pomocą szybkiej transformaty Fouriera. Taka transformacja pozwala przejść z opisu *s(t)* jako funkcji splotu:

$$s(t) = v(t) * h(t) \tag{2}$$

na wygodniejszy opis w dziedzinie częstotliwości:

$$S(\omega) = V(\omega)H(\omega), \tag{3}$$

gdzie: *S(ω), V(ω), H(ω)* są transformatami Fouriera poszczególnych sygnałów *s(t), v(t), h(t)*. Odwrotna transformata Fouriera z logarytmu widma sygnału nazywana jest cepstrum zespolonym (odwrócenie pierwszych 4 liter w słowie spectrum),

$$\bar{s} = F^{-1}(\ln F(s(t))) \tag{4}$$

Dziedzinę, w której bada się amplitudy cepstrum, nazywa się queferency (przez analogię do angielskiego słowa, frequency). Niekiedy dziedzinę tę utożsamia się ze specyficznie pojmowanym czasem. Dla klasy szeregów minimalnofazowych można zastąpić widmo S(ω) modułem widma, |S(ω)|, [13]. Pozwala to uprościć obliczenia, które wykonywane są teraz na zbiorze liczb rzeczywistych. Wykazano, że dla szeregów minimalnofazowych nie prowadzi to do utraty dokładności. Problem jednak leży w tym, że sygnał mowy jest ogólnie minimalnofazowy. Wprowadzone poprzednio uproszczenie ma mały wpływ na przebieg analizy sygnału mowy, gdyż dla sygnałów nieminimalnofazowych wartości cepstrum zachowują informacje o module widma a nie o jego fazie. Dla sygnału mowy, zawierającego się w pojedynczej ramce, cepstrum rzeczywiste obliczane jest jako:

$$\hat{s}(ramka) = FFT^{-1}(\ln|FFT(ramka)|) \tag{5}$$

Po zlogarytmowaniu obu stron (4) uzyskuje się log(|FFT(ramka)|:

$$\log|FFT(ramka)| = \ln S(\omega) = \ln V(\omega) + \ln H(\omega) \tag{6}$$

W wyniku operacji logarytmowania nastąpiło rozdzielenie części okresowej, charakteryzującej sygnał pobudzający od części charakteryzującej kanał głosowy. Jeżeli ω_0 jest częstotliwością pobudzającego sygnału okresowego, to w logV(ω) występują piki dla częstotliwości będących wielokrotnościami częstotliwości podstawowej, $\omega = n\omega_0$. Dla cepstrum sygnału mowy wyliczanego w każdej ramce zachodzi zależność:

$$\hat{s}(t) = \hat{v}(t) + \hat{h}(t) \tag{7}$$

Zwraca uwagę fakt, że cepstrum rozdzieliło sygnał mowy na dwie składowe:

- niskoczęstotliwościową, przedstawiającą własności kanału głosowego,
- wysokoczęstotliwościową, przedstawiającą własności pobudzenia krtaniowego.

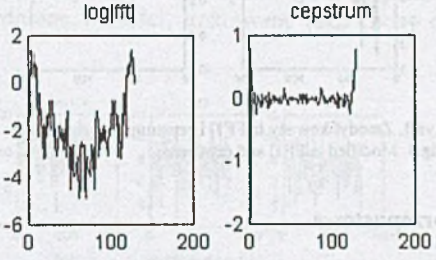
Analiza cepstrum pobudzenia krtaniowego wykorzystywana jest do analizy foniatrycznej, natomiast analiza cepstrum kanału głosowego wykorzystywana jest do analizy treści wypowiedzi. Składniki odpowiadające własnościom kanału głosowego mieszczą się w cepstrum w pobliżu $t=0$.

3.1. Algorytm podstawowy

W literaturze opisane są różne sposoby ekstrakcji cech z wykorzystaniem metod cepstralnych. Podstawowy algorytm ekstrakcji cech badanego mówcy z fragmentu jego wypowiedzi, wykorzystujący cepstrum uproszczone, można przedstawić następującym schematem, [5]:

1. Pomiar fali głosowej;
2. Podział sygnału na 10-20ms ramki, zachodzące wzajemnie na siebie;
3. Okienkowanie sygnału w każdej ramce w celu zmniejszenia zniekształceń;
4. Ewentualne uzupełnienie wartości sygnału wewnątrz ramki zerami tak, aby liczba próbek w ramce była wielokrotnością dwu;
5. Wyliczenie logarytmu widma modułu sygnału dla każdej ramki;
6. Obliczenie odwrotnej transformaty Fouriera, F^{-1};
7. Przyjęcie pierwszych kilku (np. 14) wartości cepstrum za cechy charakterystyczne;
8. Uśrednienie cech po wszystkich ramkach.

Z reguły wprowadzenie każdej czystej idei do praktyki wiąże się z wprowadzeniem pewnych modyfikacji, wynikających z przesłanek heurystycznych. Podobnie rzecz się ma z podstawowym algorytmem ekstrakcji cech mówcy. Analizując ln|FFT| i cepstrum wyznaczone podstawowym algorytmem dla sygnału mowy w pojedynczej ramce, pokazane na rys.2 zauważamy na wykresie ln|FFT| występowanie składowej szybkozmiennej, związanej z sygnałem pobudzenia i składowej wolnozmiennej, związanej z właściwościami kanału głosowego. Przebieg ln|FFT| przedstawiony jest na wykresie w funkcji przesunięcia, k, a nie jak zazwyczaj w funkcji częstotliwości. Zależność między częstotliwością, f, a przesunięciem, k, i częstotliwością próbkowania, f_p, jest następująca:



Rys. 2. Wykres ln|FFT| i cepstrum dla sygnału mowy w pojedynczej ramce
Fig. 2. Diagram of ln|FFT| and cepstrum for a single frame

$$f = \frac{k}{N} f_p, \tag{8}$$

gdzie N jest liczbą próbek przetwarzanego sygnału mowy. Cepstrum, przedstawione jest na rys.2 również w funkcji przesunięcia, k, a nie w funkcji czasu, τ. Zależność między zmiennymi k, τ, f_p jest następująca:

$$\tau = \frac{k}{f_p} \tag{9}$$

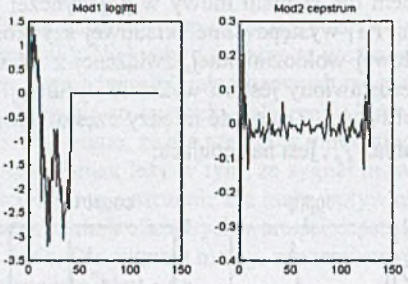
W literaturze dotyczącej przetwarzania sygnału mowy próbki występujące na wykresie cepstrum w zakresie 0 - 5ms przypisuje się składowym związanym z kanałem głosowym. Dla częstotliwości próbkowania 8kHz zakres ten odpowiada zakresowi 40 próbek. Jeżeli chcemy rozdzielić składowe zawarte w sygnale mowy, to należy ln|FFT| poddać filtracji dolnoprzepustowej w celu otrzymania składowej związanej z kanałem głosowym i filtracji górno przepustowej w celu otrzymania składowej związanej z pobudzeniem. Ponieważ w zagadnieniu identyfikacji mówcy interesujące są niskie częstotliwości, przed wyliczeniem cepstrum dokonuje się dodatkowych, pośrednich transformacji ln|FFT|. Stąd biorą się kolejne modyfikacje podstawowego algorytmu wyznaczania cech.

3.2. Filtracja dolnoprzepustowa

Najprostsza modyfikacja ,[10], polega na wymnożeniu ln|FFT| przez ciąg:

$$l(n) = \begin{cases} 1, & \text{dla } |n| \le 40 \\ 0, & \text{dla } |n| > 40 \end{cases} \tag{10}$$

Zmodyfikowane w opisany sposób ln|FFT| i wynikające z przyjętej modyfikacji cepstrum, dla pojedynczej przykładowej ramki pokazuje rys.3.



Rys.3. Zmodyfikowany ln|FFT| i cepstrum
Fig.3. Modified ln|FFT| and cepstrum

3.3. Filtracja pasmowo przepustowa

Kolejna modyfikacja wynika ze spostrzeżenia, że sygnał mowy, zawierający się na ogół w przedziale częstotliwości z zakresu 200 - 8000 Hz, ze względu na dolnoprzepustowe właściwości ucha zewnętrznego i środkowego może być, bez zauważalnego obniżenia jego zrozumiałości, rozpatrywany w zakresie 300-3500 Hz, [13], a nawet jeszcze bardziej zawężonym. Dlatego ln|FFT|, wyliczany i przedstawiany na wykresach w funkcji k, można od dołu ograniczyć wartością:

$$k_1 = \frac{300\ N}{f_p}, \tag{11}$$

a od góry:

$$k_2 = \frac{3500\ N}{f_p}, \tag{12}$$

gdzie: f_p jest częstotliwością próbkowania sygnału mowy, a N jest liczbą próbek w ramce. Takie obcięcie ln|FFT| spowoduje zmniejszenie całkowitej liczby próbek w ramce i wiążący się z tym spadek dokładności. Jako antidotum stosuje się rozciągnięcie obciętego ln|FFT| do poprzedniego zakresu N próbek z zastosowaniem liniowej interpolacji.

3.4. Dekompozycja sygnału mowy na pasma częstotliwości

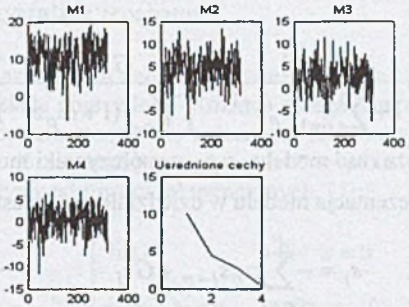
W [4] A.Czyżewski przytacza wyniki zastosowania skali melowej (skali wysokości dźwięku) do rozpoznawania mowy. Tak zwane współczynniki melcepstralne, M_i, wyznacza w F podpasmach widma wyliczonego z wykorzystaniem transformaty Fouriera, obliczonej przy zastosowaniu okna Hamminga, na podstawie następującej zależności:

$$M_i = \sum_{k=1}^{F} \ln X_k^2 \cos\{i(k-0.5)\frac{\pi}{F}\}, \qquad (13)$$

gdzie $\log X_k^2$ jest logarytmem energii w paśmie o numerze k. Algorytm, wyliczający współczynniki melcepstralne, działa według poniższego schematu:

1. Pomiar fali głosowej;
2. Filtracja pasmowo przepustowa zarejestrowanego sygnału mowy, z wykorzystaniem F filtrów;
3. Podział sygnału w każdym paśmie na 10-20ms ramki, zachodzące wzajemnie na siebie;
4. Okienkowanie sygnału w każdej ramce w celu zmniejszenia zniekształceń;
5. Ewentualne uzupełnienie wartości sygnału wewnątrz ramki zerami tak, aby liczba próbek w ramce była wielokrotnością dwu;
6. Wyliczenie widma sygnału dla każdej ramki i każdego pasma;
7. Wyliczenie logarytmu energii sygnału dla każdej ramki i każdego pasma;
8. Wyliczenie współczynników melcepstralnych dla każdej ramki;
9. Uśrednienie wartości współczynników melcepstralnych dla ramek.

Wartości kolejnych współczynników melcepstralnych, M1, M2, M3, M4, w poszczególnych ramkach i ich uśrednione wartości, traktowane jako cechy osobowe, pokazane są na rys.4.



Rys.4.Współczynniki melcepstralne
Fig. 4. Melcepstral coefficients

3.5. Wygładzanie cepstrum

Dla rzeczywistego sygnału cyfrowego A. Czyżewski, [4], opisuje metodę wykorzystującą cepstrum wygładzone, którą zastosował z powodzeniem do analizy sygnału wadliwej wymowy. Współczynniki cepstrum wyznacza według zależności:

$$C_r = \sum_{i=1}^{m} \ln s_i \cos\left(r\frac{i\pi}{m}\right), \tag{14}$$

gdzie: $r=1,2,..,R$ - indeks współczynników cepstrum,

s_i - wartość próbki sygnału mowy w chwili i,

$m = \frac{N}{f_p} f_c$,

N - liczba próbek w ramce,

f_p - częstotliwość próbkowania,

f_c - maksymalna częstotliwość uwzględniana w analizie cepstralnej.

Wygładzanie cepstrum wykonywane jest według zależności:

$$W_n = \sum_{r=1}^{R} C_r \cos(r\frac{n\pi}{m}) \quad \text{dla } n = 1,2,...m \tag{15}$$

Ostatecznie wygładzone współczynniki cepstralne wyliczane są jako:

$$\tilde{C}_r = \sum_{i=1}^{m} W_i \ln s_i \cos(r\frac{i\pi}{m}). \tag{16}$$

4. Metody liniowej predykcji

Liniowa predykcja jest jedną z najczęściej stosowanych technik w analizie sygnału mowy. Wykorzystuje ona, taki sam jak analiza cepstralna, liniowy filtracyjny model generacji sygnału mowy. Zakłada, że w analizowanym, krótkim przedziale czasu (ramce) sygnał mowy może być traktowany jako stacjonarny i stąd opisany jest liniowym modelem autoregresywnym o postaci wielomianowej lub zerobiegunowej i stałych, w rozpatrywanym przedziale czasu, współczynnikach:

$$S(z) = \frac{GV(z)}{1 + \sum_{p=1}^{P} a_p z^{-1}} = \frac{GV(z)}{\prod_{p=1}^{P} (1 + \alpha_p z^{-1})} \tag{17}$$

W przyjętym modelu P oznacza rząd modelu, a_p - współczynniki modelu wielomianowego, a α_p - bieguny modelu. Reprezentacja modelu w dziedzinie czasu jest następująca:

$$s_i = -\sum_{p=1}^{P} a_p s_{i-p} + Gv_i \tag{18}$$

Równanie (18) ma charakter predyktora i na jego podstawie analizowany, spróbkowany sygnał mowy, $s(i)$, może być prognozowany jako liniowa kombinacja ważonych poprzednich próbek tego sygnału, zsumowana z pobudzeniem, Gv_i, gdzie współczynnik G jest wzmocnieniem. Współczynnik wzmocnienia G często bywa pomijany w zagadnieniach typu rozpoznawanie mowy czy rozpoznawanie mówcy, by uodpornić algorytmy na zmiany energii sygnału mowy. Przedstawiony model, (18), nazywany jest często modelem liniowej predykcji (LP),

a współczynniki a_p nazywane są współczynnikami predykcji. Błąd predykcji definiowany jest jako różnica między wartością aktualną sygnału a jego oceną wyliczoną na podstawie poprzednich próbek,

$$e_i = s_i - \hat{s}_i = s_i + \sum_{p=1}^{P} a_p s_{i-p} \tag{19}$$

Minimalizacja sumy kwadratów błędów predykcji prowadzi do następującego układu równań:

$$\sum_{r=1}^{P} a_{pr} \sum_{i=P}^{N-1} s_{i-p} s_{i-r} = -\sum_{i=P}^{N-1} s_i s_{i-r}, \tag{20}$$

którego rozwiązaniem jest zbiór parametrów modelu, a_{pr}. Sumy iloczynów $s_i s_{i-k}$ występujące w równaniu (20) są elementami funkcji autokorelacji. Minimalny błąd predykcji określony jest zależnością:

$$e_{\min} = \sum_{i=P}^{N-1} s_i^2 + \sum_{r=1}^{P} \sum_{i=P}^{N-1} s_i s_{i-r} a_{pr} \tag{21}$$

Autokorelacja sygnału mowy pełni kluczową rolę w procesie wyznaczania współczynników modelu LP metodą liniowej predykcji. Funkcję autokorelacji, R_k, ciągu, s_i, stacjonarnego w przedziale $0<i<N-1$, można aproksymować następująco:

$$R_k = \frac{1}{N-k} \sum_{i=k}^{N-1} h_i s_i h_{i-k} s_{i-k}, \tag{22}$$

gdzie h_i jest funkcją wagową, czyli oknem czasowym, stosowanym w celu wygładzenia efektów brzegowych wynikających ze skończonego ciągu danych. Najczęściej jako funkcję wagową stosuje się okno Hamminga. Współczynniki a_{pr} można wyznaczać wykorzystując np. rekurencyjną procedurę Durbina.

4.1. Metody LP wykorzystujące cepstrum

Idea metod LP rozpoznających cechy osobowe mówcy na podstawie cepstrum jest taka sama jak opisana w rozdziale poprzednim. Różnica polega na tym, że do wyliczenia cepstrum zamiast transformaty Fouriera stosuje się zależności rekurencyjne, (24), (25), co powoduje przyspieszenie obliczeń. Z parametrycznego modelu sygnału mowy można wyliczyć cepstrum według następujących zależności rekurencyjnych, [7]:

$$c_n = \begin{cases} \ln G & \text{dla } n = 0 \\ a_n + \frac{1}{n} \sum_{i=1}^{n-1} c_i a_{n-i} & \text{dla } n > 0 \end{cases} \tag{23}$$

Zerowy współczynnik cepstrum c_0 jest dalej pomijany ze względu na dużą wrażliwość na współczynnik wzmocnienia modelu, G. Dla modelu zadanego poprzez zera, α_i, zależności rekurencyjne pozwalające wyznaczyć cepstrum są następujące:

$$c_n = \begin{cases} \ln G & \text{dla } n = 0 \\ \frac{1}{n}\sum_{i=1}^{p} \alpha_i^n & \text{dla } n > 0 \end{cases} \quad (24)$$

Działanie algorytmu wykorzystującego parametryczny model AR i sposób wyliczania cepstrum wg zależności rekurencyjnej (23) przedstawia poniższy schemat:

1. Pomiar fali głosowej;
2. Podział sygnału na 10--20ms ramki, zachodzące wzajemnie na siebie;
3. Okienkowanie sygnału w każdej ramce w celu zmniejszenia zniekształceń;
4. Wyliczenie współczynników modelu AR(n) sygnału dla każdej ramki;
5. Wyliczenie cepstrum dla każdej ramki wg wzoru rekurencyjnego, (23);
6. Przyjęcie pierwszych kilku (np. 14) wartości cepstrum za cechy charakterystyczne;
7. Uśrednienie cech po wszystkich ramkach.

4.2. Metoda LP i lokalizacja biegunów

Idea metody polega na założeniu, że filtr modelujący kanał głosowy

$$H(z^{-1}) = \frac{G}{\prod_{i=1}^{P}(1-\alpha_i z^{-1})} \quad (25)$$

można zdekomponować na część modelującą cechy osobowe mówcy i część modelującą cechy wypowiedzi związane z treścią,

$$H(z^{-1}) = H_1(z^{-1})H_2(z^{-1}) = \frac{G_1}{\prod_{i=1}^{p_1}(1-\alpha_i z^{-1})}\frac{G_2}{\prod_{i=1}^{p_2}(1-\alpha_i z^{-1})} \quad (26)$$

W tym przypadku $G = G_1 G_2$ i $P = p_1 + p_2$. Analizując modele zidentyfikowane dla poszczególnych ramek, *j=1,...,lram*, badanego fragmentu wypowiedzi pod kątem powtarzalności biegunów w poszczególnych ramkach można wyodrębnić bieguny powtarzalne, należące do modelu $H_{1_j}(z^{-1})$ i bieguny różne, przynależne do modeli $H_{2_j}(z^{-1})$. Zbiór biegunów powtarzalnych tworzy zbiór cech osobowych mówcy. Podstawowy algorytm lokalizacji biegunów działa w następujący sposób:

1. Podział sygnału na 10-20ms ramki, zachodzące wzajemnie na siebie;
2. Okienkowanie sygnału w każdej ramce w celu zmniejszenia zniekształceń;
3. Wyliczenie współczynników modelu *AR(n)* sygnału dla każdej ramki;
4. Wyliczenie biegunów modelu dla każdej ramki;
5. Wybór i zliczenie odpowiadających sobie biegunów w przebiegu wypowiedzi;
6. Posortowanie biegunów według częstości występowania;
7. Wybór biegunów o największej częstości występowania;
8. Uśrednienie po wszystkich ramkach biegunów o największej częstotliwości występowania .

5. Porównanie metod ekstrakcji cech

Aby porównać opisane metody ekstrakcji cech należy wybrać wartości kryterialne, umożliwiające dokonanie takiego porównania. Dobrą metodę powinno cechować możliwie duże skupienie cech tego samego mówcy, wyznaczanych dla różnych wypowiedzi, przy jednoczesnym możliwie dużym rozproszeniu cech dla różnych mówców.

5.1. Kryteria oceny metod

Przyjęto, że metodę ekstrakcji cech charakteryzują dwie wielkości:

- miara rozproszenia cech tego samego mówcy, uzyskiwanych badaną metodą dla różnych wypowiedzi, nazywana miarą rozproszenia wewnętrznego. W charakterze miary rozproszenia cech osobowych danego mówcy można wykorzystać macierz kowariancji, C, której elementy wyznaczane są w następujący sposób:

$$c(p,r)=\frac{1}{L_r}\sum_{i=1}^{L_r}(x(i,p)-\bar{x}(i,p))(x(i,r)-\bar{x}(i,r)), \tag{27}$$

gdzie: $p,r=1,2,\ldots,14$ jest wymiarem przestrzeni cech, L_r jest liczbą różnych wypowiedzi tego samego mówcy, na podstawie których wyznaczano dla niego wektory cech. Pierwiastek elementów leżących na przekątnej głównej macierzy kowariancji jest miarą dyspersji poszczególnych cech danego mówcy;

- miara rozproszenia cech uzyskiwanych badaną metodą dla różnych mówców, nazywana miarą rozproszenia zewnętrznego. Za miarę rozproszeń zewnętrznych można przyjąć macierz B, której elementy wylicza się według następującej reguły:

$$b(p,r)=\frac{1}{M}\sum_{m=1}^{M}\left\{\bar{X}_{L_r}(m,p)-\overline{\bar{X}_{L_r}}(p))(\bar{X}_{L_r}(m,p)-\overline{\bar{X}_{L_r}}(p))\right\}, \tag{28}$$

gdzie: $\bar{X}_{L_r}(m,p)$ oznacza wartość cechy p dla mówcy m, uśrednioną po wszystkich L_r wypowiedziach mówcy m,

$\bar{X}_{L_r}(m,r)$ oznacza wartość cechy r dla mówcy m, uśrednioną po wszystkich L_r wypowiedziach mówcy m,

$$\bar{X}_{L_r}(m,p)=\frac{1}{L_r}\sum_{i=1}^{L_r}x(m,i,p) \tag{29}$$

$$\bar{X}_{L_r}(m,r)=\frac{1}{L_r}\sum_{i=1}^{L_r}x(m,i,r) \tag{30}$$

$\overline{\bar{X}_{L_r}}(p)$ oznacza uśrednioną po wszystkich M mówcach średnią z wypowiedzi, $\bar{X}_{L_r}(m,p)$,

$\overline{\overline{X}_{L_r}}(r)$ oznacza uśrednioną po wszystkich M mówcach średnią z wypowiedzi, $\overline{X}_{L_r}(m,r)$,

$$\overline{\overline{X}_{L_r}}(p) = \frac{1}{M}\sum_{m=1}^{M}\overline{X}_{L_r}(m,p) \tag{31}$$

$$\overline{\overline{X}_{L_r}}(r) = \frac{1}{M}\sum_{m=1}^{M}\overline{X}_{L_r}(m,r) \tag{32}$$

5.2. Przebieg badań

Przeprowadzenie badań wymagało:

- utworzenia eksperymentalnej bazy danych;
- wstępnego przetworzenia zarejestrowanych sygnałów mowy;
- ekstrakcji cech z wykorzystaniem wybranego algorytmu;
- wyliczenia wielkości kryterialnych;
- porównania metod według przyjętych kryteriów.

5.2.1. Tworzenie bazy danych

Badany zbiór obejmował dziewięć osób: pięć kobiet i czterech mężczyzn. Każda z badanych osób generowała przynajmniej 6 różnych wypowiedzi. Wypowiadane sekwencje trwały od 3 do 15 sec. Trzy pierwsze sekwencje wypowiadane przez każdą badaną osobę były identyczne, trzy kolejne różniły się i obejmowały:

- dowolny fragment czytanego tekstu, ok. 12sec.
- dowolna recytacja, ok. 15sec.
- dowolny tekst improwizowany, ok. 14 sec.

5.2.2. Wstępne przetwarzanie sygnału

Ciągły sygnał mowy próbkowany był z częstotliwością *8 kHz*. Dyskretny sygnał mowy podlegał wstępnej obróbce obejmującej wycięcie ciszy i normalizację sygnału.

5.2.3. Ekstrakcja cech

Uzyskany sygnał był przetwarzany w celu ekstrakcji cech osobowych mówcy z wykorzystaniem następujących algorytmów:

- podstawowy *(podst)*,
- filtracja pasmowo przepustowa *(filtr)*,
- nieliniowa transformacja skali częstotliwości *(melwar)*,
- wygładzanie cepstrum *(cepwyg)*,
- LP- model wielomianowy *(LPw)*,
- LP- model biegunowy *(LPb)*,
- Lokalizacja biegunów - algorytm podstawowy *(biegp)*,
- Lokalizacja biegunów - algorytm uproszczony *(biegu)*.

Algorytmy: *podst, filtr, cepwyg, LPw, LPb* w charakterze cech osobowych mówcy przyjmowały 14 pierwszych wartości cepstrum, algorytm *melwar* - 4 współczynniki melcepstralne. Algorytm *biegp* przyjmował za cechy mówcy bieguny, których powtarzalność w ramkach była większa niż 80%. Algorytm *biegu* przyjmował za cechy charakteryzujące mówcę uśrednione wartości biegunów, których część rzeczywista była większa od 0.8, a moduł części urojonej mniejszy od 0.3.

5.3. Podsumowanie wyników

Uzyskane rezultaty będą przedstawione oddzielnie dla metod wykorzystujących cepstrum i metod wykorzystujących lokalizację biegunów.

5.3.1. Metody wykorzystujące cepstrum

W tablicy1 przedstawiono miarę rozproszenia zewnętrznego metod wykorzystujących w charakterze cech osobowych wartości współczynników cepstrum.

Tablica 1

Miary rozproszenia zewnętrznego

metoda b	podst	filtr	melwar	cepwyg	LPw	LPb
$b_{1,1}$	2.643	1.41	1.46	8.57	2.779	1.38
$b_{2,2}$	.201	.002	.330	1.52	.018	.100
$b_{3,3}$	.029	.000	.191	6.69	.090	.012
$b_{4,4}$	.022	.000	.022	.91	.015	.007
$b_{5,5}$	.002	.000	--	5.30	.005	.006
$b_{6,6}$	.001	.000	--	.82	.004	.008
$b_{7,7}$	.002	.000	--	6.01	.001	.006
$b_{8,8}$	.003	.000	--	.45	.001	.006
$b_{9,9}$	.003	.001	--	4.90	.005	.007
$b_{10,10}$	.002	.001	--	.16	.001	.006
$b_{11,11}$	.003	.001	--	4.20	.000	.005
$b_{12,12}$	.002	.001	--	.11	.001	.004
$b_{13,13}$	.003	.001	--	4.19	.000	.003
$b_{14,14}$	.002	.001	--	.09	.001	.002

Miary rozproszenia wewnętrznego dla każdej z przedstawionych metod są kilka rzędów niższe niż odpowiednie miary rozproszenia zewnętrznego i przykładowo dla algorytmu podstawowego, dla dwóch badanych osób wynoszą:

Tablica 2

Miary rozproszenia zewnętrznego

Ewa	.54	.88	.14	.06	.08	.02	.02	.03	.02	.04	.01	.02	.04	.02	$*10^{-4}$
Pat	10.24	.28	.08	.08	.06	.03	.07	.04	.04	.03	.02	.08	.06	.12	$*10^{-4}$

5.3.2. Metody wykorzystujące lokalizację biegunów

Ekstrakcję cech przeprowadzono dwoma metodami:

- metodą podstawową,
- metodą zmodyfikowaną.

Metoda podstawowa

Zastosowanie analizy rozkładu biegunów modelu AR identyfikowanego dla każdej z ramek, na które podzielona została wypowiedź, pod kątem powtarzalności biegunów w ramkach pozwoliło stwierdzić, że:

1. Dla każdej z badanych osób w każdej z ramek powtarza się para biegunów zespolonych o części rzeczywistej dodatniej, zawierającej się w przedziale (0.8 - 0.99) i częściach urojonych z zakresu (0.1j - 0.3j).
2. Powtarzalność takich biegunów wahała się w granicach 80-99%.

Przykładowo, części rzeczywiste biegunów modeli w ramkach dla jednej z badanych osób, uszeregowane według malejącej powtarzalności dla 8 różnych przykładowych wypowiedzi wynosiły:

Tablica 3

Bieguny rzeczywiste Ewy

wypowiedź / re(biegun)	1	2	3	4	5	6	7	8
1/2	.934	.963	.926	.928	.927	.880	.881	.906
3/4	-.620	-.642	-.682	.866	-.346	-.691	-.614	.831
5/6	-.206	-.058	-.232	.766	.838	-.144	.899	.327
7/8	.348	-.674	.398	-.499	-.567	.365	-.350	-.093

Odpowiednie części urojone wynosiły:

Tablica 4

Bieguny urojone Ewy

wypowiedź / im(biegun)	1	2	3	4	5	6	7	8
1/2	.197	.185	.189	.202	.210	.199	.257	.200
3/4	.426	.368	.318	.296	.650	.346	.419	0
5/6	.708	.728	.704	0	0	.693	0	.715
7/8	.681	0	.640	.650	.377	.610	.719	.704

Miara rozproszenia wewnętrznego dla poszczególnych biegunów wynosi:

- dla części rzeczywistych:
  cr=[0.0005, 0.0005, 0.4032, 0.2965, 0.2157, 0.1417, 0.1780, 0.2121]

- dla części urojonych:
  ci=[0.0004, 0.0004, 0.0284, 0.1474, 0.2447, 0.3644, 0.3397, 0.3311]

Dla większości badanych osób miara rozproszenia wewnętrznego jest dla dwóch pierwszych biegunów, najbardziej powtarzalnych, znacznie mniejsza niż dla pozostałych biegunów. Dla dwóch badanych osób uzyskano wyniki odbiegające od tak sformułowanego wniosku:

- Dla Bogny rozkład pierwszych trzech biegunów pokazuje tabl. 5 i tabl. 6.

Tablica 5

Bieguny rzeczywiste Bogny

wypowiedź / re(biegun)	1	2	3	4	5	6
1	.924	.939	.923	-.725	.896	.924
2	.924	.939	.923	.932	.896	.924
3	-.668	-.0079	-.524	.932	-.682	.723
...	...	...	...	...	...	...

Tablica 6

Bieguny urojone Bogny

wypowiedź / im(biegun)	1	2	3	4	5	6
1	.245	.236	.243	0	.332	.240
2	-.245	-.236	-.243	.263	-.332	-.240
3	0	.782	.613	-.263	0	0
...	...	...	...	...	...	...

- Dla Andrzeja rozkład pierwszych trzech biegunów pokazuje tabl. 7 i tabl. 8.

Tablica 7

Bieguny rzeczywiste Andrzeja

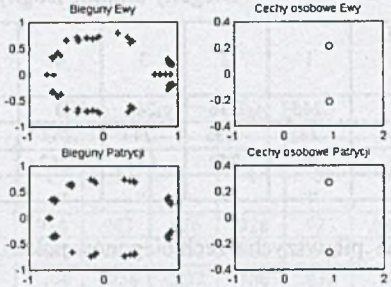
wypowiedź / re(biegun)	1	2	3	4	5	6
1	.880	.907	.905	-.682	.885	.854
2	.894	.915	.911	.906	.892	.871
3	.361	.910	.906	.906	-.681	.871
...	...	...	...	...	...	...

Tablica 8

Bieguny urojone Andrzeja

wypowiedź im(biegun)	1	2	3	4	5	6
1	-.149	-.170	-.156	0	-.114	-.128
2	.179	0	.167	.235	.182	.278
3	.653	.179	0	-.235	0	-.278
...	...	...	...	...	...	...

Przyglądając się rozkładowi biegunów dla tych, odbiegających od zaobserwowanego wzoru, mówców możemy zaobserwować, że w wypowiedzi czwartej najczęściej występował biegun ujemny, rzeczywisty, a dopiero w następnej kolejności biegun urojony zespolony, o dużej dodatniej części rzeczywistej. Powtarzalność tych trzech biegunów była podobna, z niewielką przewagą ujemnego bieguna rzeczywistego, co mogło być spowodowane występowaniem zakłócenia podczas nagrywania wypowiedzi. Przesunięcie kolejności występowania biegunów znalazło odzwierciedlenie w mierze rozproszenia wewnętrznego. Przedstawione wyniki skłaniają do przyjęcia założenia, że za cechy kanału głosowego mówcy, niezależne od wypowiadanych sekwencji zdaniowych, odpowiedzialne są dwa bieguny zespolone, o dodatnich częściach rzeczywistych, zawierających się w przedziale (0.8 - 0.99) i częściach urojonych z zakresu (0.1j - 0.3j). Ten zakres zmienności biegunów modeluje wolnozmienne właściwości kanału głosowego. Przykładowy rozkład biegunów dla dwóch z badanych osób i najczęściej występujące w ramkach bieguny, przyjęte za cechy osobowe mówców, pokazano na rys.5.

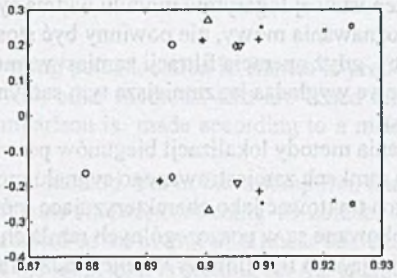


Rys.5. Rozkład biegunów dla Ewy i Patrycji
Fig. 5. Poles of Eva and Patrycja

Rysunek 6 przedstawia rozkład biegunów odpowiedzialnych za cechy osobowe badanych dziewięciu mówców. Miara rozproszenia zewnętrzneego wynosi dla tego przypadku:

$$b(p)=[0.7324 ;0.0465;0.7270;0.0391]$$

Jest ona dość duża dla części rzeczywistych i mała dla części urojonych biegunów, niemniej jednak ze względu na to, że porównywane będą między sobą pary biegunów zespolonych, rokowania możliwości identyfikacji mówcy na podstawie kryterium biegunów są pomyślne.

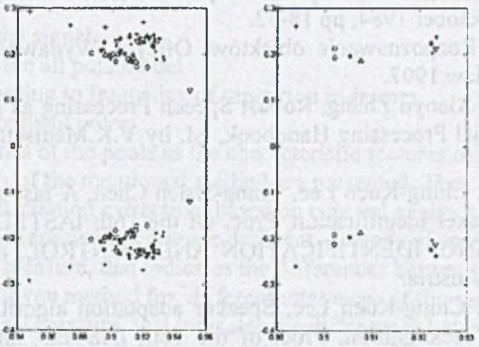


Rys. 6. Cechy osobowe 9 mówców, przedstawione rozkładem najczęściej występujących biegunów
Fig. 6.Personal features for 9 speakers represented by most frequent poles

Metoda uproszczona

Przyjmując założenia dotyczące rozkładu biegunów w modelu kanału głosowego mówcy można uprościć program wyliczający bieguny, ograniczając procedurę do wyszukiwania i zapamiętywania w poszczególnych ramkach biegunów o części rzeczywistej dodatniej, większej od 0.8 i części urojonej, mniejszej od 0.3. Wynik działania programu pokazano na rys.7. Miara rozproszenia zewnętrznego wynosi w tym przypadku:

$$b(p)=[0.7363;0.0416;0.7390\ 0.0379]$$



Rys. 7. Uśrednione bieguny kanału głosowego dla badanych mówców
Fig. 7. Averaged poles for considered speakers

6. Podsumowanie

1. W literaturze dotyczącej sposobów analizy sygnału mowy i identyfikacji mówcy przyjmuje się za cechy charakteryzujące mówcę czternaście pierwszych współczynników cepstrum. Z przeprowadzonych badań wynika, że dla większości metod wystarczy rozpatrywać co najwyżej 4 pierwsze cechy, gdyż dla cech dalszych współczynnik rozproszenia zewnętrznego jest bliski zeru, co oznacza, że na podstawie tych cech nie można rozróżnić mówców.

2. Modyfikacje podstawowego algorytmu metody cepstralnej, polegające na filtracji dolno- lub pasmowo przepustowej logarytmu modułu widma sygnału, stosowane głównie dla algorytmów rozpoznawania mowy, nie powinny być stosowane w algorytmach rozpoznawania mówcy, gdyż operacja filtracji zamiast wzmocnić poszczególne indywidualne cechy osobowe wygładza je; zmniejsza tym samym szanse rozróżnienia mówców między sobą.
3. Przeprowadzone badania metody lokalizacji biegunów potwierdzają przypuszczenia, że w poszczególnych ramkach zarejestrowanego sygnału mowy powtarzają się pary biegunów, które można traktować jako charakteryzujące mówcę.
4. Należy zbadać, jak lokowane są w poszczególnych ramkach bieguny przy zmienianiu rzędu modelu AR opisującego filtr liniowy w pojedynczej ramce.
5. Należy sprawdzić, jakie efekty przyniesie lokalizacja zer zastosowana obok lokalizacji biegunów dla celów rozróżnienia mówców.

LITERATURA

1. Atal B.S., Linear Prediction Analysis of Speech Signals, in Programs for Digital Signal Processing, John Wiley and Sons, 1979.
2. Basztura Cz., Komputerowe Systemy Diagnostyki Akustycznej, PWN, Warszawa 1996.
3. Brągoszewski P., Pogadaj z komputerem, PC World Komputer, Nr 3/2000, pp 115-120.
4. Czyżewski A., Dźwięk cyfrowy. Wybrane zagadnienia teoretyczne, technologia, zastosowania, Akademicka Oficyna Wydawnicza EXIT, Warszawa 1998.
5. Gish H., Schmidt M., Text-independent Speaker Identification, IEEE Signal Processing Magazine, October 1984, pp 18-32.
6. Kurzyński M. Rozpoznawanie obiektów, Oficyna Wydawnicza Politechniki Wrocławskiej, Wrocław 1997.
7. Mammone R., Xiaoyu Zhang, Robust Speech Processing as an Inverse Problem, in The Digital Signal Processing Handbook, ed. by V.K.Madisetti, D.B.Williams, IEEE Press, 1998.
8. Ming-Tzaw-Lin, Ching-Kuen Lee, Ching-Hsien Chen, A fast search method for text-independent speaker identification, Proc. Of the 16th IASTED International Conference MODELLING, IDENTIFICATION AND CONTROL, held February 17-19th, 1997, Insbruck, Austria.
9. Ming-Tzaw-Lin, Ching-Kuen Lee, Speaker adaptation algorithms for speaker independent speech reckognition, Proc. of the 16th IASTED International Conference MODELLING, IDENTIFICATION AND CONTROL, held February 17-19th, 1997, Insbruck, Austria.
10. Oppenheim A, Schaffer R., Cyfrowe Przetwarzanie Sygnałów, WKił, Warszawa 1979.
11. Sherman Ong, Yih-Sheng Lin, Miles Moody, Sridha Sridharan, Text independent speaker reckognition using Fisher's discriminant, Proc. of the 16th IASTED International Conference MODELLING, IDENTIFICATION AND CONTROL, held February 17-19th, 1997, Insbruck, Austria.
12. Szabatin J., Podstawy teorii sygnałów, WKił, Warszawa 1982.
13. Tadeusiewicz R., Sygnał mowy, WKiŁ 1988.

Recenzent: Prof.dr hab. inż. Ryszard Tadeusiewicz
Wpłynęło do Redakcji 1.03.2001 r.

Abstract

In the article a method using poles location in frames is proposed for speaker information extracting and compared with other methods, that are based on cepstral analysis and linear predictive coding. The comparison is made according to a measure of internal and external dispersion.

The idea of the pole based method lies in the assumption, that the registered speech signal may be segmented into separate frames, containing 10-20msec fragments of speech. The signal in frames may by modelled as an output of a linear stationary filter excited in the input by a white noise signal. As the main goal is seeking for the characteristic features which allow to distinguish one speaker from another, the further assumption is, that the speech signal restricted in each of the frames contains information about as well the meaning of the utterance as the speaker itself. Hence, the further assumption is, that the filter model inside every frame may be decomposed into two parts, one which models personal speaker features and the other which represents the meaning of the utterance.

$$H(z^{-1}) = \frac{G_1}{\prod_{i=1}^{p}(1-\alpha_i z^{-1})} \frac{G_2}{\prod_{i=1}^{p}(1-\beta_i z^{-1})}$$

Under this assumptions the problem of speaker information extraction consists in seeking poles or pair of poles repeated in each frame.

The basic algorithm of pole based method may be represented by the following steps:

1. Measurement of a speech signal
2. Segment the signal into 10-20msec frames
3. For each frame
   - Window the signal
   - Calculate the all pole model
4. Sort the poles according to frequency of repetition in frames
5. Choose the most frequent poles
6. Take the mean values of the pools as the characteristic features of the speaker.

In the article the results of the mentioned method are presented. They are compared to the six methods of speaker information extracting, based on cepstral analysis and linear predictive coding. Two indexes are applied to compare different methods of speaker features extracting. The first is an internal measure, that indicates the differences between speaker features calculated with the use of a given method for different utterances of the same speaker.

The second is an external measure, that indicates the differences between features of various speakers calculated with the use of the same method .