Rozpoznawanie mówcy – model akustyczny – Support Vector Machines

Paweł Buglewicz, Krzysztof Cieśla, Justyna Olczak

9 lutego 2017

Spis treści

1 Wstęp

Zabezpieczenia biometryczne stają się coraz bardziej popularne. Przyczyną tego jest głównie wygoda, czasem również względy bezpieczeństwa. Dzięki rozpoznawaniu idyndywidualnych i niepowtarzalnych cech danej osoby uzyskuje się dosyć bezpieczny mechanizm zabezpieczający.

Postanowiliśmy się przyjrzeć jak może działać jedno z bardziej popularnych metod identyfikacji – rozpoznawanie osoby **za pomocą próbki głosu**.

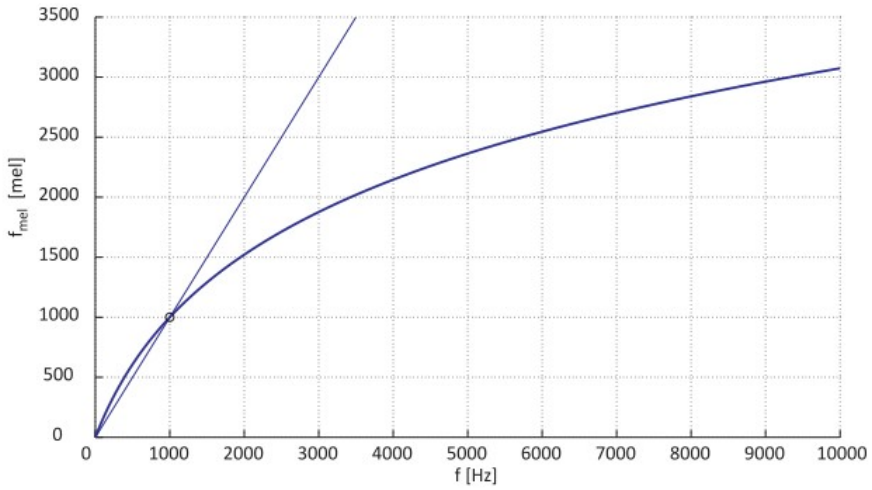
Taki typ identyfikacji może być przeprowadzany na wiele różnych sposobów. Zazwyczaj podział ten przebiega następująco: rozróżniamy **identyfikację słownikową** oraz metody stosujące **model akustyczny**. W metodzie słownikowej bada się charakterystyczne cechy głosu, które towarzyszą wypowiadaniu konkretnych słów lub głosek. Zazwyczaj identyfikacja odbywa się po przeczytaniu podanego przez program słowa lub ciągu słów.

Stosując model akustyczny próbuje się przyporządkować charakterystyczne cechy głosu osoby, **bez wiedzy o tym, jakie słowa dana osoba wypowiada**. Tutaj bada się tembr głosu, jego tempo itp. Stworzenie modelu akustycznego jest prostsze niż implementacja modelu opartego na słowniku, lecz wymaga większej ilości danych wejściowych do prawidłowego działania.

2 MFC (Mel Frequency Cepstrum)

W najprostszych modelach akustycznych, takich jak nasz, jako charakterystycznej cechy głosu używa się współczynników MFC (Mel Frequency Cepstrum).

- Mel - skala wysokości dźwięku mierzona metodą akustyki psychologicznej określającej subiektywny odbiór poziomu dźwięku przez ucho ludzkie względem obiektywnej skali pomiaru częstotliwości dźwięku w hercach.



Rysunek 1: Przykładowa zależność między skalą liniową a mel-skalą.

Uważa się, iż mel-skala lepiej niż skala liniowa odzwierciedla charakterystykę słuchu ludzkiego.

- Cepstrum - odwrotna transformata Fouriera widma sygnału wyrażonego w skali logarytmicznej (decybelowego). Słowo cepstrum jest anagramem słowa spectrum.

2.1 Współczynniki MFC

Aby uzyskać współczynniki MFC należy:

1. Zastosować transformatę Fouriera do próbki dźwiękowej (do odpowiednio krótkiej jej części).

    - Transformata Fouriera: jest operatorem liniowym określanym na pewnych przestrzeniach funkcyjnych.
      Rozkłada funkcję okresową na szereg funkcji okresowych tak, że uzyskana transformata podaje, w jaki sposób poszczególne częstotliwości składają się na pierwotną funkcję.
    - Literatura sugeruje, że odpowiednia długość próbki to 3s.

2. Przenieść uzyskane wartości na skalę mel za pomocą odpowiednich okien czasowych

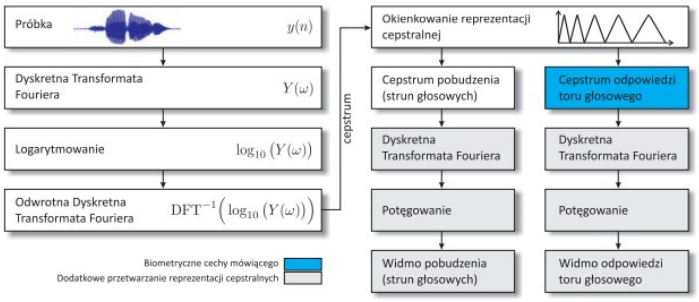
    - Mel-cepstralne = odpowiedzi toru głosowego. Skupiona jest w początkowych elementach reprezentacji.
    - Okno czasowe to funkcja opisująca sposób zbierania próbek z sygnału, musi ona być ograniczona

3. Zlogarytmować uzyskane wartości

4. Zastosować dyskretną transformatę kosinusową do uzyskanego wyniku.

    - Dyskretna transformata kosinusowa = rodzaj blokowej transformacji danych. Jest szczególnie popularny w stratnej kompresji danych.

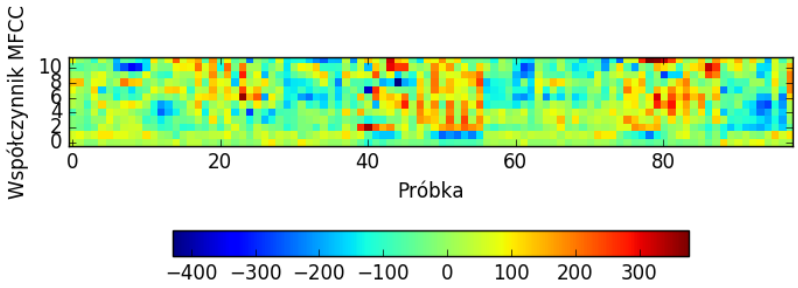
5. MFC to amplitudy uzyskanego widma.



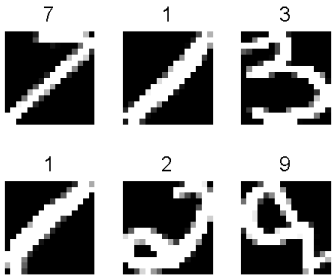
Rysunek 2: Estymacja w dziedzinie cepstralnej - schemat działania.

3 Metoda SVM

Współczynniki MFC dla każdego okna czasowego to wektor liczb rzeczywistych, a dla całej 3-sekundowej próbki - odpowiednia macierz, w której kolumnach znajdują się wektory współczynników MFC okien czasowych składających się na trzysekundową próbkę.



MFCC próbki głosu Krzyśka, nr 16



Przykładowe zastosowanie SVN –
rozpoznawanie pisma odręcznego

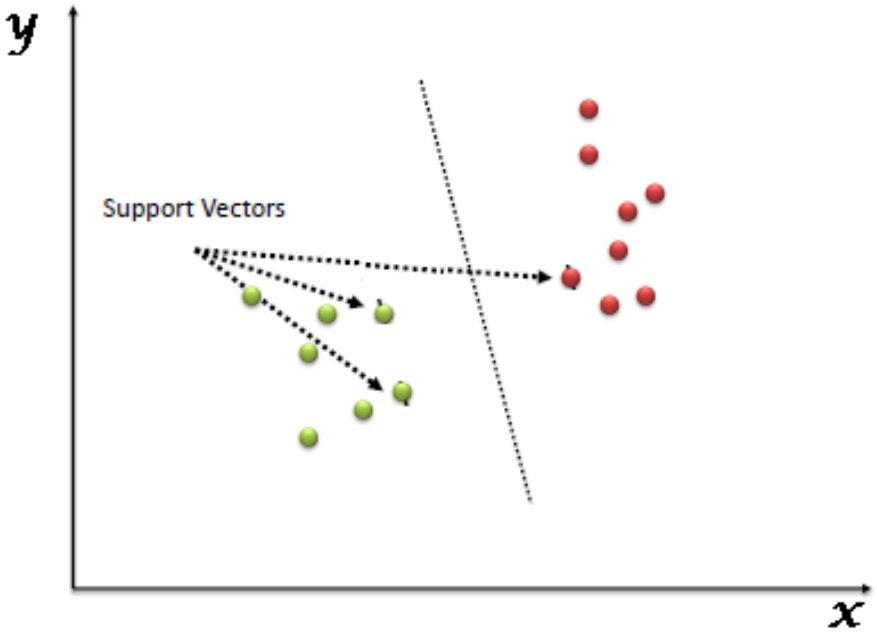
Rysunek 3: Dla mechanizmu SVN nie ma znaczenia, czy dane są współczynnikami MFC, czy wartościami pikseli obrazów przedstawiających cyfry – oba mają jednoznaczną reprezentację macierzową i wektorową.

Taka tablica niczym nie różni się od macierzowej reprezentacji obrazu w skali szarości. Dlatego można tu zastosować analogiczne metody jak przy rozpoznawaniu obrazów (rys.3).

Jedną z metod, pozwalających na identyfikację przynależności próbki do pewnego zbioru, jest SVM - **Support Vector Machine**. Jest to jeden z podstawowych algorytmów uczenia maszynowego. Można go podzielić na dwie fazy: uczenie i predykcję. W pierwszej fazie dane są rozmieszczone w wieloelementowej przestrzeni i każdemu przypisana jest etykieta - zbiór musi być opisany a priori. Ilość wymiarów zależy od wymiarowości próbki. Dla obrazu

w skali szarości - jest to ilość pikseli.

Następnie algorytm szuka wielowymiarowych płaszczyzn, które jak najlepiej oddzielają punkty z daną etykietą od innych. Aby zaoszczędzić pamięć, algorytm zachowuje tylko punkty (wektory), które są położone blisko tej granicy (pokazane na rysunku 4).



Rysunek 4: Wektory położone najbliżej płaszczyzny dzielącej są jedynymi, które definiują tę płaszczyznę. Nazywamy je **wektorami nośnymi** (ang. *support vectors*).

Gdy płaszczyzna jest wyznaczona można przejść do drugiej fazy: przewidywania. Odbywa się to poprzez wstawianie nowych punktów do tej samej przestrzeni i sprawdzanie, po której stronie płaszczyzny się znajdują. Tym sposobem można przyporządkować im odpowiednią etykietę. Te punkty nie modyfikują już położenia płaszczyzny.

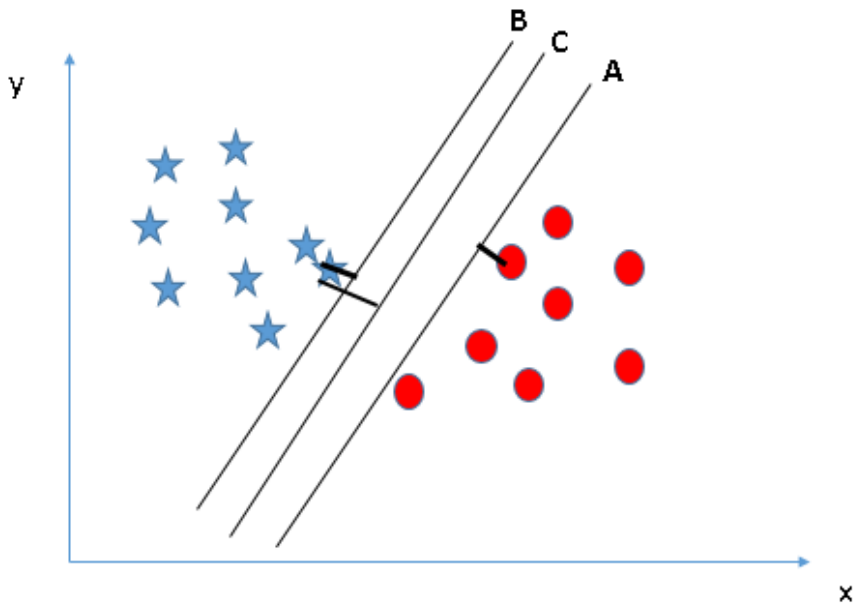
3.1 Wymiarowość próbki i margines separacji

W zależności od geometrii i typu danych, płaszczyzna może być linią, krzywą, płaszczyzną lub hiperpłaszczyzną.

- **Linia:** W przypadku liniowo separowalnym, metoda ta gwarantuje zna-

lezienie takiej płaszczyzny, która ma maksymalny tzw. margines separacji;

- **Płaszczyzna:** W przypadku nieseparowalnym liniowo, metoda SVM pozwala na znalezienie płaszczyzny, która klasyfikuje obiekty i jednocześnie przebiega możliwie daleko od typowych skupień dla każdej z klas;



Rysunek 5: Przykład klasyfikacji - płaszczyzna

- **Hiperpłaszczyzna:** W przypadku nieseparowalnym liniowio, można również za pomocą metody SVM znaleźć krzywoliniową granicę klasyfikacji o dużym marginesie separacji.

3.2 Zalety i Wady SVM

Zalety:

- Skuteczna w przestrzeniach wielowymiarowych;
- Używa tylko części próbki tzw. wektorów nośnych do nauki modelu, więc oszczędzamy pamięć;
- Jest wszechstronna, można stoswać różne modele podziałów przestrzeni.

Wady:

- Jeśli jest więcej wymiarów niż próbek, to wyniki są nieprawidłowe;
- Nie oblicza błędu klasyfikacji - nie wiemy, jak poprawny jest nasz podział.

4 Dane

W naszym projekcie próbowaliśmy rozpoznać za pomocą SVM 4 osoby. Pierwszym krokiem było nagranie próbek głosowych. Zostały one podzielone na ścieżki o długości 3s - każda taka ścieżka to jeden punkt danych.

Dla każdej ścieżki została wyznaczona macierz współczynników MFC. Jest to macierz o wymiarach 12x120. Taki rozmiar macierzy jest arbitralnym wyborem, proponowanym w literaturze, gdzie zostało empirycznie pokazane, że taki podział próbki daje najlepsze efekty. 12 to ilość współczynników MFC, 120 to ilość 25 ms okien czasowych. Problem jest więc $12 \cdot 120 = 1440$ wymiarowy.

5 Podsumowanie

Nie udało nam się uzyskać zamierzonego efektu. Prawdopodobną przyczyną jest to, że nie spełnione zostało jedno z wymagań algorytmu (o którym dowiedzieliśmy się za późno):

Jednym z wymagań, aby SVM dawało wiarygodne i poprawne wyniki, jest to, iż ilość próbek służących do wytrenowania musi być większa niż ilość wymiarów. W naszym przypadku mogłoby się to sprowadzić do godzinnych nagrań, lub do zmniejszenia ilości wymiarów.

Pierwsza z opcji jest mało wygodna z punktu widzenia użytkownika, który chce używać identyfikacji głosowej jako zabezpieczenia biometrycznego.

Druga jest możliwa do wykonania, jednak niezalecana w literaturze, ponieważ zmiana wymiarowości wiąże się ze zmianą długości okna lub pobieranej próbki, a to z kolei wymaga, aby próbka była wyższej jakości (lepsze nagranie, lepsze filtrowanie itp.)