States of Art

Skład osobowy: Iga Gaj 155082, Adam Ledwoń 155290

W kontekście rozpoznawania płci na podstawie częstotliwości głosu można wyróżnić kilka podejść, opartych na różnych metodach sztucznej inteligencji. Poniżej opisaliśmy trzy znane koncepcje rozwiązywania tego problemu:

1. **Metoda oparta na analizie cech akustycznych i klasyfikacji SVM (Support Vector Machine)**

Opis metody: SVM jest algorytmem klasyfikacyjnym, który stara się znaleźć najlepszą granicę (hiperpłaszczyznę) dzielącą dane w przestrzeni cech, tak aby jak najlepiej oddzielić jedną klasę od drugiej. W przypadku rozpoznawania płci na podstawie głosu, cechy te mogą obejmować:

- **Częstotliwość podstawową (F0)** – zależną od wysokości głosu, która zazwyczaj jest niższa u mężczyzn, a wyższa u kobiet.
- **Spektrogram** – przedstawienie widma częstotliwości w funkcji czasu, umożliwiające analizę zmian akustycznych w czasie.
- **Formanty** – charakterystyczne częstotliwości rezonansowe, które różnią się w zależności od płci.
- **Długość i intensywność dźwięku** – różnice w tym zakresie także mogą pomóc w klasyfikacji głosu.

2. **Sieci neuronowe (Deep Learning)**
    a) **Sieci konwolucyjne (CNN - Convolutional Neural Networks)**

Opis metody: CNN jest zaprojektowany do analizy danych o strukturze przestrzennej, takich jak obrazy czy dane dźwiękowe reprezentowane w postaci widm (np. mel-spektrogramów). Kluczowe elementy CNN to:

- **Warstwa konwolucyjna (Convolutional Layer)** - Stosuje tzw. filtry przesuwające się po danych wejściowych, aby wykrywać lokalne wzorce, takie jak krawędzie, tekstury czy kształty. Wyjściem jest tzw. mapa cech, która reprezentuje wykryte wzorce.
- **Warstwa aktywacji (Activation Layer)** - Często stosuje się funkcję ReLU (Rectified Linear Unit), która wprowadza nieliniowość, eliminując ujemne wartości (zeruje je), co pozwala modelowi uczyć się bardziej złożonych wzorców.
- **Warstwa poolingowa (Pooling Layer)** - Zmniejsza wymiarowość danych, redukując liczbę parametrów i poprawiając wydajność obliczeniową. Najczęściej stosuje się max pooling, który wybiera najwyższą wartość w określonym obszarze, zachowując najważniejsze cechy.
- **Warstwy w pełni połączone (Fully Connected Layers)** - Po wyodrębnieniu cech z danych wejściowych za pomocą warstw konwolucyjnych i poolingowych, dane są przekształcane do postaci wektora i przekazywane do klasyfikatora.

    b) **Sieci rekurencyjne (np. LSTM - Long Short-Term Memory)**

Opis metody: Podstawową jednostką LSTM jest komórka pamięci (memory cell), która kontroluje przepływ informacji za pomocą trzech głównych bramek:

- **Brama zapominania (forget gate)**: Decyduje, które informacje z poprzedniego stanu powinny zostać zapomniane.
    - Działa na zasadzie mnożenia przez wagę w zakresie [0,1], gdzie 0 oznacza całkowite zapomnienie, a 1 pełne zachowanie.
- **Brama wejściowa (input gate)**: Kontroluje, które nowe informacje z bieżącego wejścia zostaną dodane do stanu komórki.
    - Analizuje bieżące dane i aktualizuje stan pamięci.
- **Brama wyjściowa (output gate)**: Ustala, które informacje z bieżącego stanu komórki mają być przekazane do następnej warstwy lub kroku czasowego.

3. **Modelowanie probabilistyczne (np. Gaussian Mixture Models – GMM)**

Opis metody: GMM to probabilistyczne modele statystyczne, które służą do modelowania rozkładu danych jako mieszaniny wielu rozkładów normalnych (gaussowskich). GMM zakłada, że dane pochodzą z kilku różnych grup (klastrów), z których każda jest opisana przez własny rozkład normalny.

Model GMM można interpretować jako połączenie rozkładów normalnych z określoną średnią i wariancją oraz wag klastrów która odzwierciedla proporcję danych należących do tych klastrów (wagi sumują się do 1).

Cały model GMM to suma ważona K rozkładów normalnych.

GMM wykorzystuje algorytm EM (Expectation-Maximization), aby dopasować model do danych dzielący się na etap oczekiwania (oblicza prawdopodobieństwo przynależności każdego punktu danych do każdego klastra, bazując na bieżących parametrach $\mu k, \Sigma k, \pi k$) i etap maksymalizacji (aktualizuje parametry modelu $\mu k, \Sigma k, \pi k$ w celu maksymalizacji prawdopodobieństwa obserwowanych danych).

Metoda		**Mocne strony**	**Słabe strony**
SVM		● Wysoka dokładność przy odpowiednio dobranych cechach. ● Efektywność w przestrzeni o wysokim wymiarze (gdy bierzemy pod uwagę wiele cech akustycznych). ● Dobre radzenie sobie z dużymi zbiorami danych.	● Bardzo wrażliwe na jakość cech. ● Trudności z uzyskaniem dobrych wyników, gdy dane głosowe są zróżnicowane (np. różne akcenty, zakłócenia tła, zmiany w jakości nagrania). ● Wrażliwość na szum.
Sieci neuronowe	CNN	● Automatyczne wykrywanie hierarchicznych wzorców, eliminując konieczność ręcznej analizy danych.	● Duże wymagania obliczeniowe.

		• Redukcja liczby parametrów (max pooling). • Odporność na lokalne przesunięcia.	• Brak interpretowalności (działa jak "czarna skrzynka"). • Do analizy plików audio konieczne jest wcześniejsze przekształcenie ich na postać widmową.
	LSTM	• Skuteczność w analizie długich sekwencji. • Uniwersalność (sygnały audio, tekst i wideo). • Odporność na zanikający gradient (zapamiętuje zarówno krótkoterminowe, jak i długoterminowe zależności).	• Duże wymagania obliczeniowe. • Czasochłonny trening. • Złożoność modelu.
GMM		• Może modelować różnorodne kształty klastrów, w tym eliptyczne. • Wyniki klasteryzacji są probabilistyczne, co pozwala oszacować pewność przynależności punktu do klastra. • Działa równie dobrze na danych jednowymiarowych, jak i wielowymiarowych.	• Zakłada, że dane w każdym klastrze mają rozkład normalny. • Wyniki mogą być zależne od początkowych parametrów, co może prowadzić do lokalnych minimów. • Algorytm EM może być kosztowny dla dużych zbiorów danych.