Biometryczne
wspomaganie interakcji człowiek-komputer
Biometria głosu

Bartłomiej Stasiak

bartlomiej.stasiak@p.lodz.pl
basta@ics.p.lodz.pl

Instytut Informatyki
Politechnika Łódzka

2017

Plan wykładu

Wstęp

- Głos jest cechą biometryczną najłatwiejszą do pozyskania
  - Możliwość akwizycji zdalnej (linie telefoniczne, telefony komórkowe, VoIP, etc.)
  - Możliwość akwizycji bez wiedzy użytkownika
- Nie jest cechą tak niepowtarzalną jak linie papilarne, czy tęczówka
- Może podlegać znaczącym zmianom (infekcje górnych dróg oddechowych, stan emocjonalny, proces starzenia)
- Stosunkowo podatny na próby fałszerstwa (np. *replay attack*)
- Uważany zwykle za cechę behawioralną (pomimo pewnej zależności od anatomii traktu głosowego)

Podział metod rozpoznawania mówcy

- Rozpoznawanie zależne od tekstu (ang. *Text-dependent speaker recognition*)
  - Zastosowanie: jako metoda uwierzytelniania (zwykle dostęp zdalny)
  - Wykorzystanie w połączeniu z hasłem
  - Częsty wariant – losowy wybór słów do powiedzenia (np. ciąg cyfr), w celu przeciwdziałania próbom podszywania się (*replay attacks*)
- Rozpoznawanie niezależne od tekstu (ang. *Text-independent speaker recognition*)
  - Zastosowanie: pasywne rozpoznawanie mówcy
  - Identyfikacja niepożądanych klientów (*blacklisting*) np. przez centra telefoniczne *call centers*
  - Analiza danych z monitoringu, podsłuchów, etc. (analiza fonoskopijna)

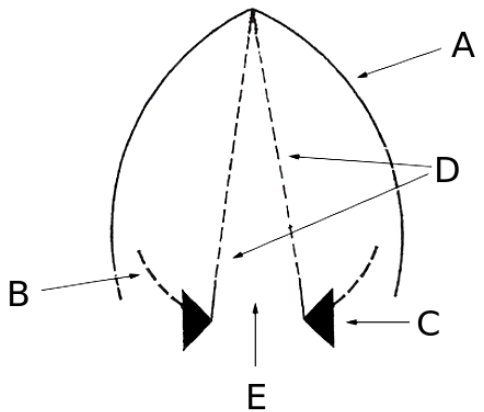
Podział metod rozpoznawania mówcy

- Powstawanie sygnału mowy jest skomplikowanym procesem zależnym od:
  - Czynników *socjolingwistycznych*
    - Poziom i charakter wykształcenia, wykonywany zawód
    - Dobór słownictwa
    - Różnice dialektu/odmiany języka
  - Czynników *anatomiczno-fizjologicznych*
    - Kształt poszczególnych odcinków traktu głosowego
    - Długość traktu głosowego
    - Własności i sposób wykorzystania narządów artykulacyjnych
- W związku z tym metody biometryczne obejmują:
  - analizę wysokiego poziomu, tj. lingwistyczną (*language generation*)
  - analizę niskiego poziomu, tj. akustyczną (*speech production*)

Aparat mowy

- Aparat mowy obejmuje część układu oddechowego człowieka
  - Aparat oddechowy (płuca, przepona, oskrzela, tchawica)
  - Aparat fonacyjny (krtań, więzadła głosowe)
  - Aparat artykulacyjny (jama gardła, jama nosowa, jama ustna)

Krtań i więzadła głosowe



- A – chrząstka tarczowa
- B – chrząstka pierścieniowa
- C – chrząstki nalewkowate
- D – więzadła głosowe
- E – szpara głośni

- Wymawianie głosek dźwięcznych powoduje napięcie więzadeł i przysunięcie ich do siebie
- Napięte więzadła głosowe na skutek przepływu powietrza okresowo zwierają się i rozwierają
- Częstotliwość zależy od ich długości, grubości i stopnia napięcia mięśniowego – zależnych z kolei m.in. od płci i wieku:
  - mężczyźni ok. 125 Hz
  - kobiety ok. 210 Hz
  - dzieci ok. 300 Hz

Aparat artykulacyjny

- Aparat artykulacyjny obejmuje (ruchome) *narządy artykulacyjne* (dolna warga, część przednia, środkowa, tylna i korzeń języka), które podczas artykulacji zbliżają się, bądź zwierają z (nieruchomym) *miejscem artykulacji* (górna warga, górne zęby, dziąsła, podniebienie twarde, podniebienie miękkie)
- To zbliżenie bądź zwarcie jest typowe dla spółgłosek i umożliwia ich klasyfikację, na zasadzie narząd-miejsce artykulacji, np:
  - dwuwargowe
  - wargowo-zębowe
  - międzyzębowe
  - przedniojęzykowo-zębowe
  - przedniojęzykowo-dziąsłowe
  - przedniojęzykowo-zadziąsłowe
  - spółgłoski tylnojęzykowe-twardopodniebienne
  - ...

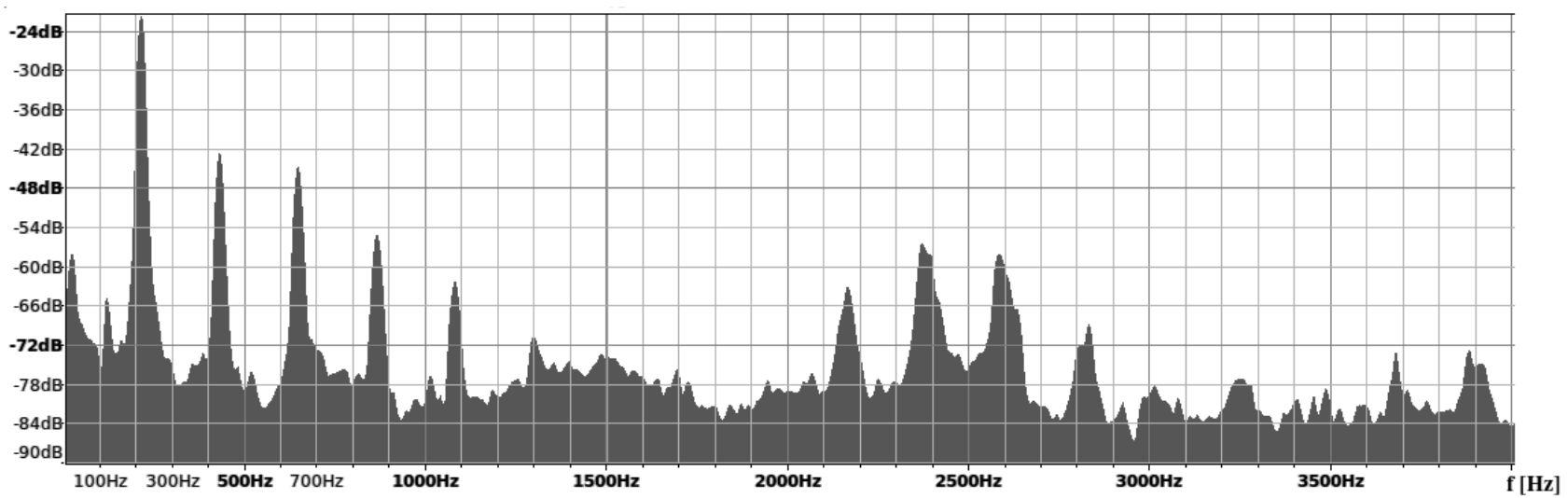
Aparat artykulacyjny

- Innym (równoległym) sposobem klasyfikacji spółgłosek jest ich dźwięczność/bezdźwięczność
- W przeciwieństwie do spółgłosek, samogłoski są typowo dźwięczne, a narządy artykulacyjne nie ograniczają swobodnego przepływu powietrza przez trakt głosowy
- Klasyfikacji samogłosek dokonujemy ze względu na:
  - położenie języka
    - przednie, środkowe, tylne
    - wysokie (przymknięte), średnie, niskie (otwarte)
  - kształt warg
    - zaokrąglone
    - niezaokrąglone
  - wykorzystanie jamy nosowej jako dodatkowego rezonatora

Metody analizy sygnału mowy

Analiza widmowa

- Podstawową metodą analizy sygnału mowy jest analiza częstotliwościowa (p. $\rightarrow$ instrukcja laboratoryjna)
- Najczęściej stosuje się transformację Fouriera (DFT) do przekształcenia oryginalnego sygnału mowy z dziedziny czasu do dziedziny częstotliwości:

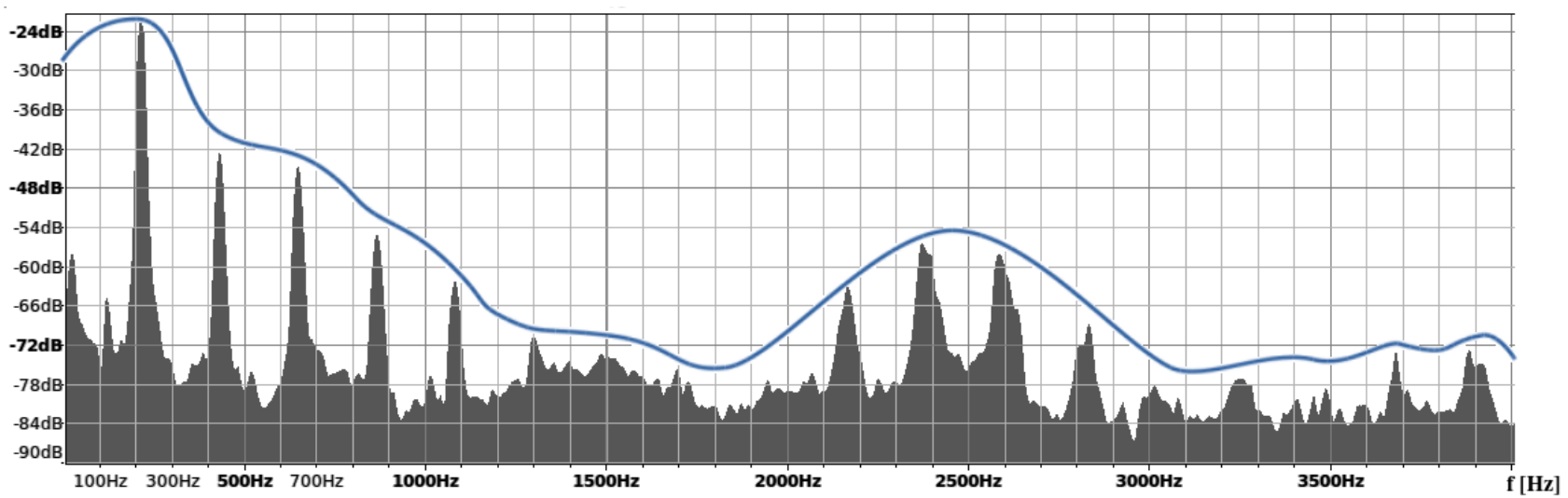


Przykładowe widmo sygnału mowy (głoska 'i')

Metody analizy sygnału mowy

Analiza widmowa

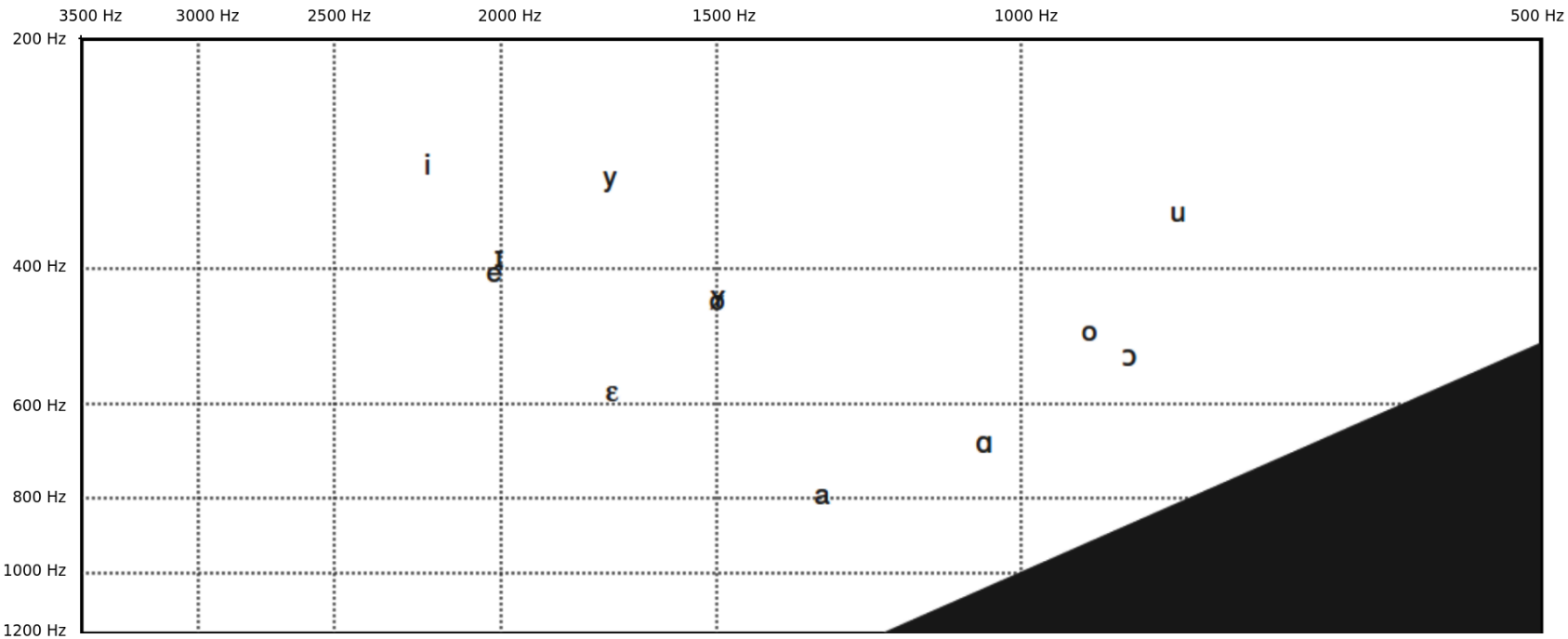
- Podstawową metodą analizy sygnału mowy jest analiza częstotliwościowa (p. $\rightarrow$ instrukcja laboratoryjna)
- Najczęściej stosuje się transformację Fouriera (DFT) do przekształcenia oryginalnego sygnału mowy z dziedziny czasu do dziedziny częstotliwości:



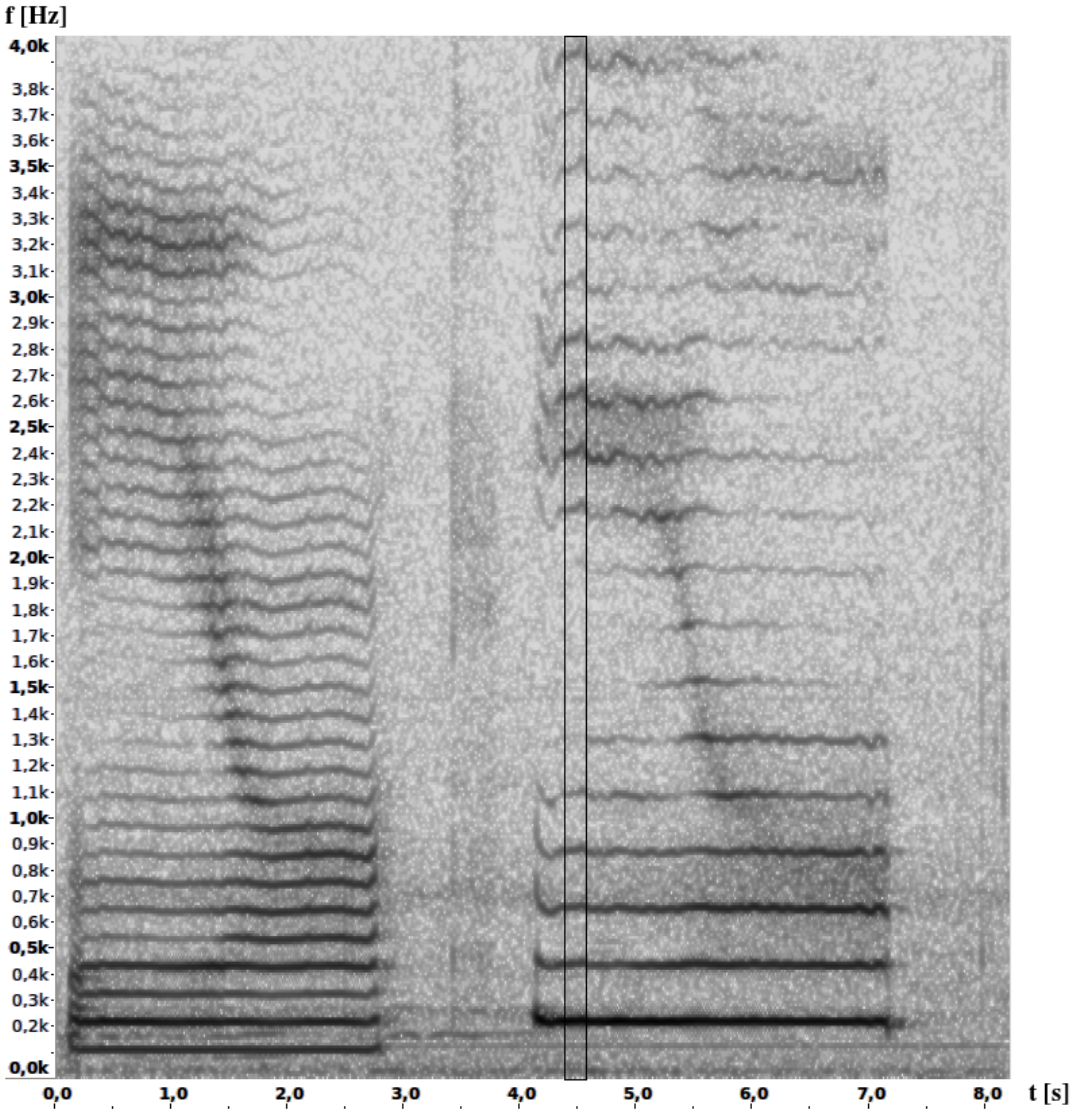
Przykładowe widmo sygnału mowy (głoska 'i') + obwiednia widmowa

Analiza widmowa

- Analiza obwiedni widmowej (położenie i szerokość pasm formantowych - przynajmniej dwóch) dostarcza informacji m.in. o rodzaju samogłoski



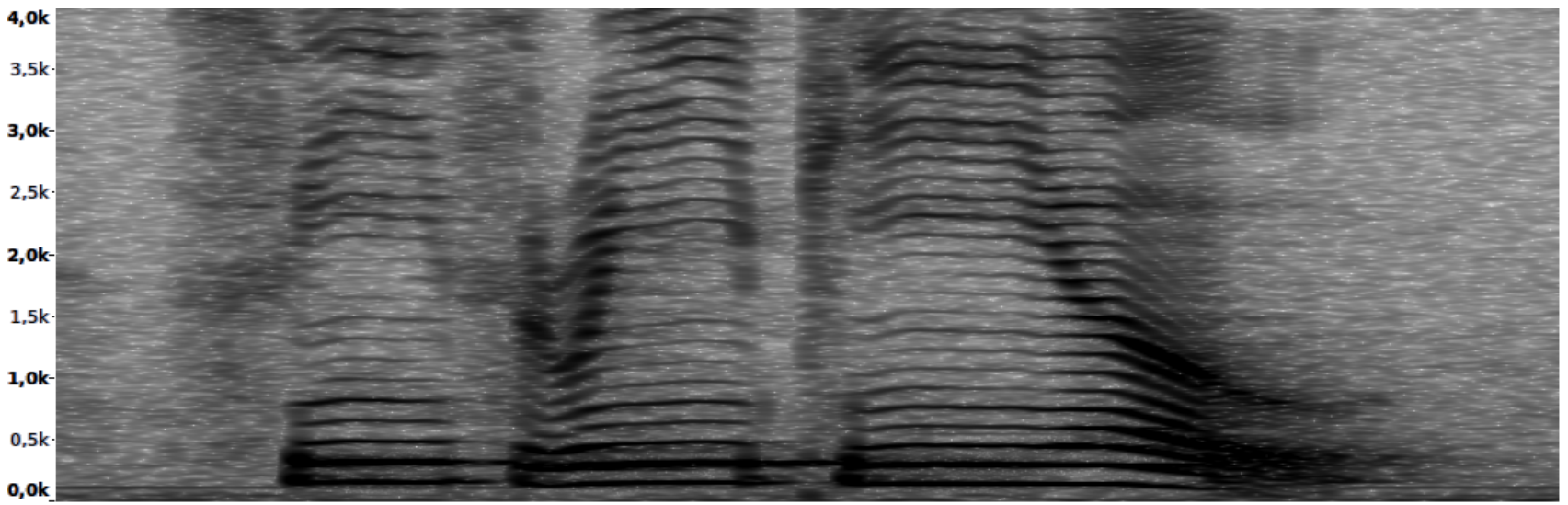
Analiza widmowa



- W celu analizy zmienności sygnału mowy w czasie stosujemy ciąg kolejnych widm amplitudowych, czyli spektrogram

Analiza widmowa

- Spektrogram pozwala nie tylko śledzić położenie pasm formantowych, ale również m.in. analizować dźwięczność/bezdźwięczność głosek



Systemy rozpoznawania mówcy zależne od tekstu

- Systemy zależne od tekstu (ang. *text-dependent*) dzielą się na:
  - Systemy ze stałym tekstem
    - Tekst wypowiedzi podczas weryfikacji jest identyczny jak w fazie zapisu użytkownika (*enrollment)*
    - Każdy użytkownik może mieć inny tekst (np. hasło)
  - Systemy ze zmiennym tekstem
    - Przy każdym użyciu systemu użytkownik proszony jest o wypowiedzenie innego tekstu (np. ciąg liczb w zmienionej kolejności)

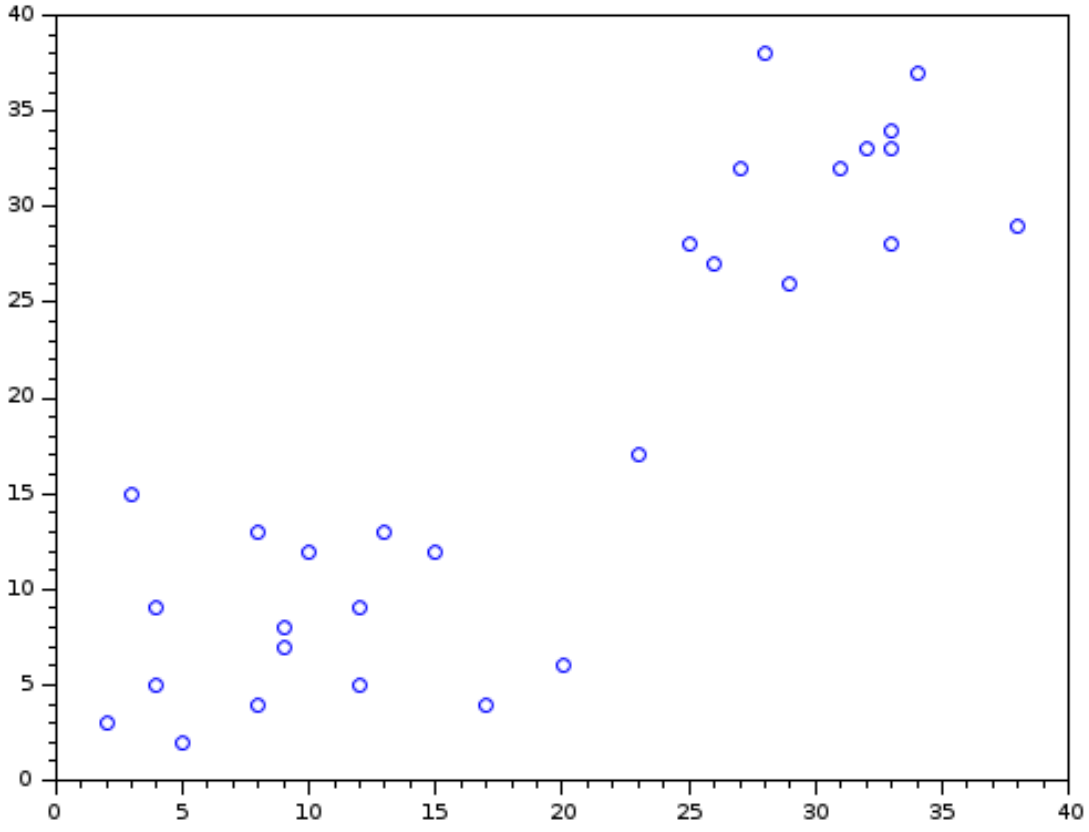
Systemy rozpoznawania mówcy zależne od tekstu

- Parametryzacja sygnału mowy
  - Analiza spektrogramu, zazwyczaj z wykorzystaniem współczynników MFCC lub LPC (p. → instrukcja laboratoryjna)
- Identyfikacja/weryfikacja mówcy
  - Porównywanie wzorców (np. ciągów wektorów MFCC) za pomocą DTW, ang. *Dynamic Time Warping* (p. → instrukcja laboratoryjna)
  - Metody statystyczne (najczęściej HMM, ang. *Hidden Markov Models*)
    - Trening HMM (algorytm Bauma–Welcha)
    - Wyszukiwanie najlepiej dopasowanego wzorca (algorytm Viterbi'ego)
  - Wykorzystanie HMM jest bardziej kosztowne obliczeniowo niż DTW, ale daje większą elastyczność (zwłaszcza w systemach ze zmiennym tekstem)

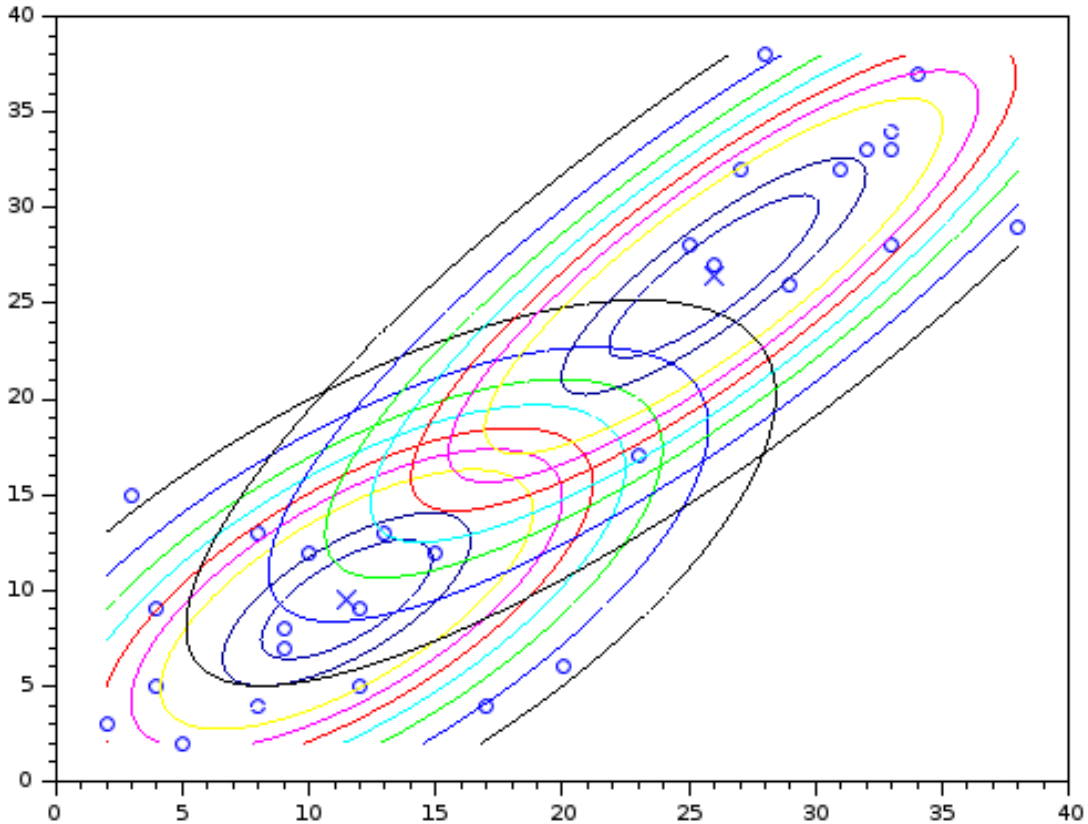
Systemy rozpoznawania mówcy niezależne od tekstu

- Podejście podstawowe – analiza widmowa
  - Parametryzacja sygnału mowy (analiza spektrogramu, zazwyczaj z wykorzystaniem współczynników MFCC lub LPC)
  - Konstrukcja modelu mówcy (*Speaker Model*, SM / *Universal Background Model*, UBM)
    - Kwantyzacja wektorowa (VM, ang. *Vector Quantization*)
    - Modele mieszane (GMM, ang. *Gaussian Mixture Models*)
- Wykorzystanie informacji wyższego poziomu
  - Systemy fonotaktyczne
    - Blok dekodera fonetycznego (zwykle w oparciu o HMM)
    - Blok modelowania statystycznego (n-gramy) – modelowanie częstości użycia głosek i sekwencji głosek dla danego mówcy
  - Systemy prozodyczne
    - Analiza sekwencji cech prozodycznych (wysokość głosu, energia)
    - Blok modelowania statystycznego (n-gramy) dla wzorców prozodycznych

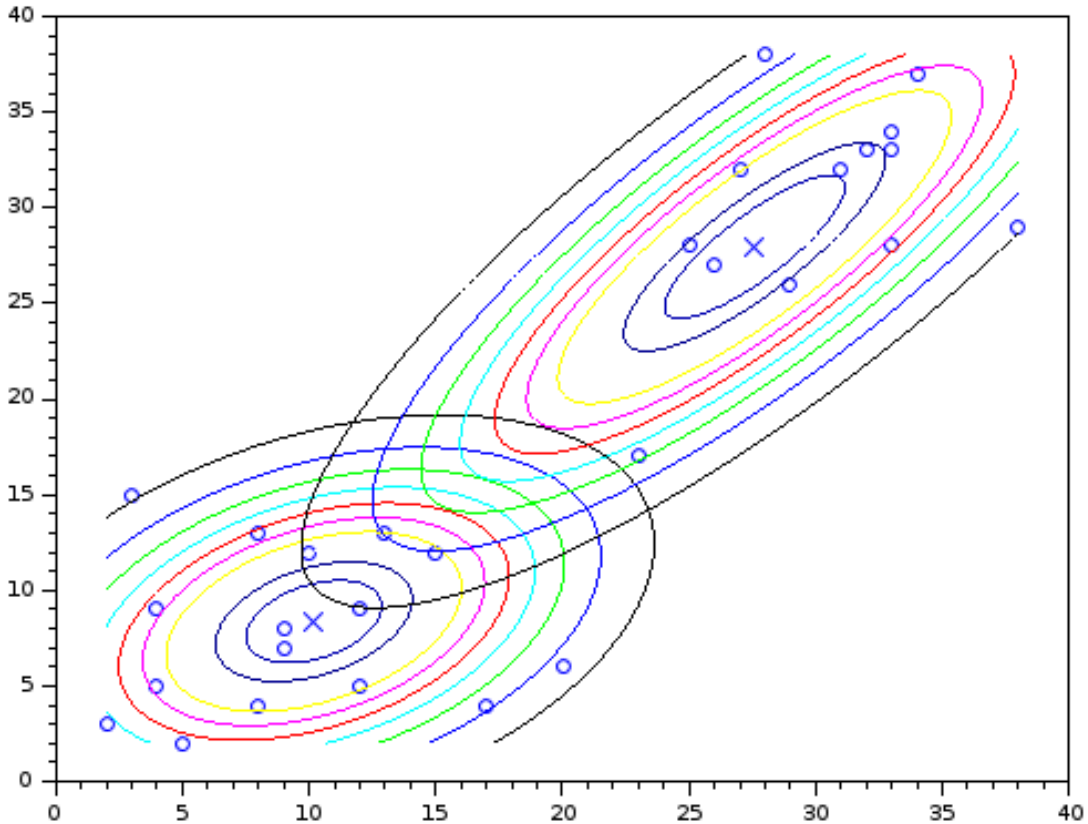
Gaussian Mixture Models (GMM) – przykład



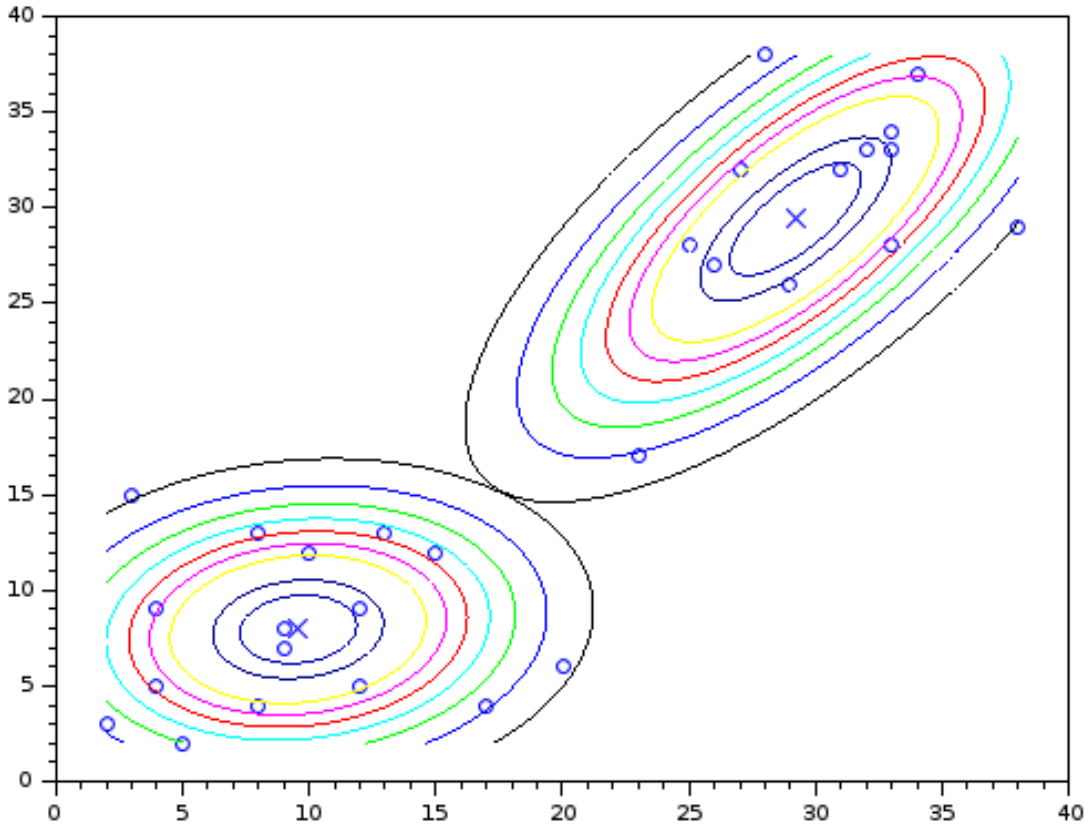
Gaussian Mixture Models (GMM) – przykład



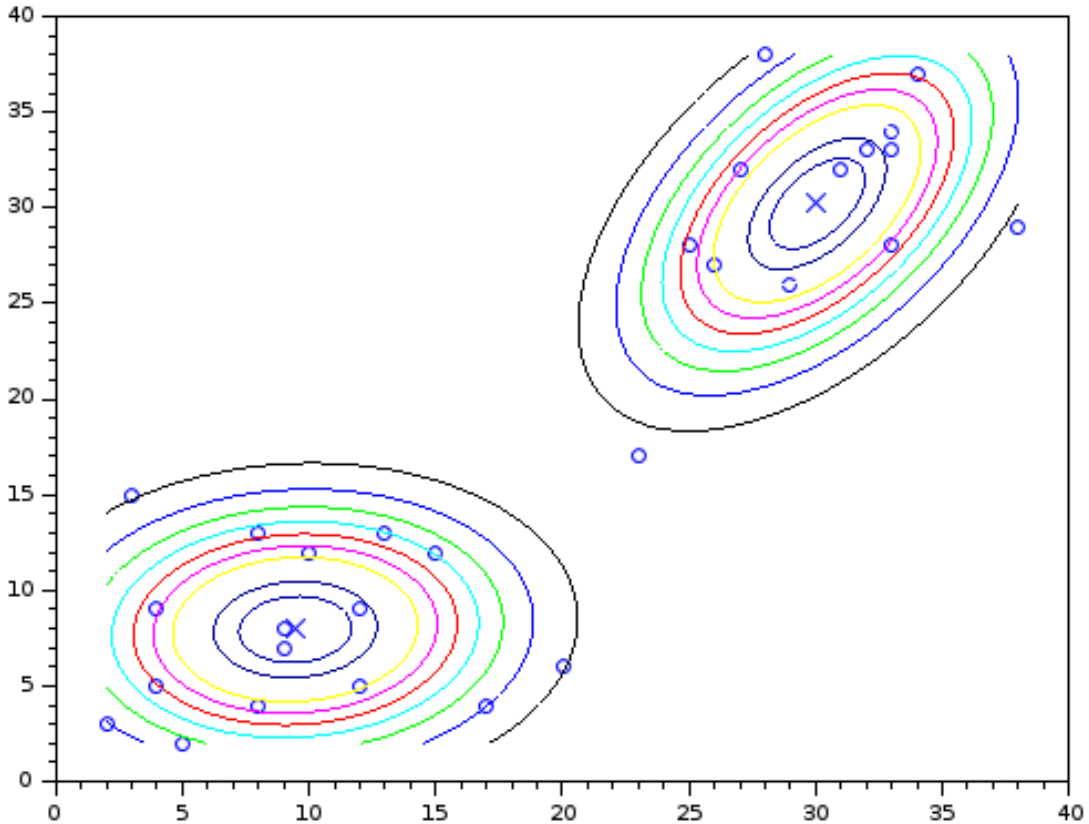
Gaussian Mixture Models (GMM) – przykład



Gaussian Mixture Models (GMM) – przykład



Gaussian Mixture Models (GMM) – przykład



Dziękuję za uwagę