GŁOSOWA ŁĄCZNOŚĆ Z KOMPUTEREM x EKSPLORACJA DANYCH BIOMEDYCZNYCH

Abstrakt

Celem realizowanego projektu było stworzenie systemu zdolnego do rozpoznawania zestawu komend głosowych. Stworzone rozwiązanie osiągnęło 71% skuteczności w rozpoznawaniu zadanych słów.

Słowa kluczowe

Python, mfcc, voice recognition, SVM, speech recognition, LibROSA

Wstęp

System, który został stworzony w ramach przedstawianego projektu docelowo ma służyć jako część większego oprogramowania służącego do kontroli inteligentnego budynku. Takie rozwiązania stają się coraz popularniejsze ze względu na swoją wszechstronność, jak i możliwość użytkowania przez starszych ludzi w celu ułatwienia ich egzystencji.

Metodologia

13 studentów i studentek w wieku od 22 do 25 lat nagrało po 4 zestawy komend głosowych. Wspomniane komendy zostały wcześniej wyłonione na podstawie ich potencjalnej użyteczności w tworzonym systemie. Wybrano między innymi słowa takie jak: otwórz, zamknij, zapal, telewizor, włącz i wyłącz. Pełen zestaw liczył 13 słów. Autorzy nagrań zostali poproszeni o stworzenie ich w różnych momentach dnia, ponieważ ludzki głos zmienia się wraz z cyklem dobowym – inaczej można brzmieć, gdy jest się zaspanym, a inaczej, gdy głodnym (czyli smutnym).

W implementacji wykorzystano język Python oraz biblioteki scikit-learn oraz libROSA, ponieważ udostępniają one możliwości uczenia maszynowego oraz ekstrakcji cech z plików głosowych.

Do budowy klasyfikatora opartego na maszynie wektorów wspierających (SVM) wykorzystano metryki wyliczone z użyciem biblioteki LibROSA. Były to między innymi:

- krótkoczasowa transformata Fouriera,
- Mel-frequency cepstrum,
- chromagramy,
- kontrast spektrogramów,
- inne.

Wyniki i analiza

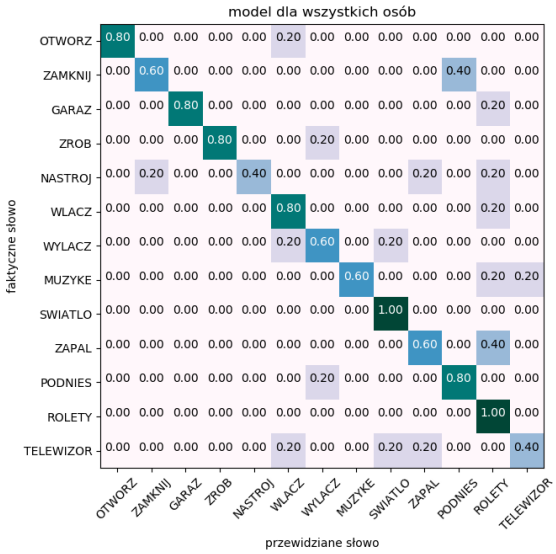
Zbudowano modele dla wszystkich studentów, z podziałem na płeć oraz dla poszczególnych osób. Otrzymane skuteczności klasyfikacji słów przedstawia Tabela 1.

Tabela 1 – Skuteczności wybranych modeli.

model	skuteczność [%]
jedna osoba	85
kobiety	62
mężczyźni	74
wszystkie osoby	71

Wysoka skuteczność modelu dla jednej osoby nie powinna być traktowana wiążąco, ze względu na fakt, że zbiór nagrań liczył 4 zestawy. Zdecydowano się użyć 3 do trenowania modelu oraz 1 do walidacji. Taka ilość danych nie jest wystarczająca do statystycznie ważnego wnioskowania. Macierz konfuzji klasyfikacji na podstawie nagrań wszystkich osób zaprezentowana jest na Rysunku 1. Modele zależne od płci wykazały, że głos kobiecy jest trudniejszy do analizy i klasyfikacji.

Analiza wygenerowanych macierzy konfuzji pozwala stwierdzić, że zaimplementowane modele są wystarczające by w poziomie zadawalającym móc rozpoznawać słowa.



Rysunek 1 – Macierz konfuzji dla modelu ogólnego.

Na przekątnej każdej macierzy widać było wysokie wartości, co oznacza wysoką skuteczność modelu. Co ciekawe, słowa do siebie bardzo podobne takie jak włącz i wyłącz były rozróżniane od siebie częściej niż słowa do siebie zupełnie nie podobne.

Maszyna wektorów wspierających, która nie należy do najbardziej zaawansowanych algorytmów poradziła sobie z problemem klasyfikacji ku zaskoczeniu autora.

Wnioski

Prosty system oparty o SVM pozwolił sklasyfikować nagrania utworzone przez 13 różnych osób. Zestaw słów poddawany klasyfikacji był dobrany tak, aby mógł być potencjalnie wykorzystany do rozwiązań dostępnych w inteligentnych budynkach. Proponowane rozwiązanie offline działa na poziomie 71% skuteczności. Oczywiście, rezultat ten może być polepszon poprzez odpowiednią obróbkę sygnału głosowego, użycie atrybutów o największym powinowactwie do słowa czy też użycie algorytmu klasyfikacyjnego u wyższej złożoności, jak na przykład sieć neuronowa.

Referencje

[1] https://librosa.github.io/librosa/feature.html
[2] http://aqibsaeed.github.io/2016-09-03-urban-sound-classification-part-1/