Metody wykrywania oddechu za pomocą sztucznej inteligencji

Tomasz Sankowski
Piotr Sulewski
Aleksandra Bruska
Jan Walczak
Politechnika Gdańska, 2024

1. Opis problemu

Projekt ma na celu badanie oraz implementację różnych metod klasyfikacji dźwięku w celu wykrywania oddechu u człowieka. Wykrywanie oddechu ma odbywać się w czasie rzeczywistym z wykorzystaniem komputerowego mikrofonu. Wykrywane będą trzy klasy: wdech, wydech oraz przerwa pomiędzy oddechami (cisza). Badaniu zostaną poddane następujące podejścia:

- Wykorzystanie modelu VGGish w celu otrzymania wektora cech dźwięku oraz klasyfikacja na podstawie tych cech za pomocą losowego lasu decyzyjnego.
- Klasyfikacja spektrogramu dla pół/ćwierćsekundowych nagrań za pomocą sieci neuronowej przystosowanej do klasyfikacji obrazów.
- Wykorzystanie współczynników cepstralnych częstotliwości w skali melowej, jako zbioru cech dźwięku pozwalających na klasyfikację za pomocą sieci neuronowej.
- Wykrywanie ekstremów głośności oraz klasyfikowanie go naprzemiennie jako wdech oraz wydech.

Wszystkie implementacje metod wraz z danymi znajdują się na publicznym repozytorium na GitHubie pod podanym linkiem:

https://github.com/tomaszsankowski/Breathing-Classification

2. Dane testowe oraz treningowe

Dane testowe oraz treningowe składają się z łącznie 1000 około trzysekundowych nagrań, a dokładniej z 200 nagrań ciszy, 400 nagrań wdechu oraz 400 nagrań wydechu. Pobierane dane zawierają zarówno oddychanie przez nos oraz usta, jak i poprzez przyłożenie mikrofonu do krtani.

3. Wykorzystanie modelu VGGish

3.1. Model VGGish

VGGish to gotowy model, dostępny na GitHubie pod podanym linkiem:
https://github.com/tensorflow/models/tree/master/research/audioset/vggish.
Model VGGish pobiera sekundowe (a właściwie to 0.975s) nagrania, z których następnie zwracany jest wektor 128 cech dźwięku. Cechy te w teorii mają być uniwersalne dla każdego dźwięku. Model VGGish wykorzystuje mel-spektrogramy oraz specjalnie wyuczony model oparty na popularnej rodzinie konwolucyjnych sieci neuronowych VGG. Spektrogramy to wykresy widma amplitudowego w kolejnych chwilach czasu, które tworzą macierz wartości, pozwalając na reprezentację graficzną za pomocą obrazu. Mel-spektrogramy to rodzaj spektrogramów, w których pasma częstotliwościowe nie są równomiernie rozmieszczone, jak ma to miejsce w tradycyjnych spektrogramach. Zamiast tego, pasma są zagęszczone zgodnie z percepcją ludzkiego słuchu. Człowiek lepiej radzi sobie z rozróżnianiem niskich częstotliwości, dlatego mel-spektrogram może lepiej odzwierciedlać ludzkie postrzeganie dźwięków

Mel-spektrogramy (lub spektrogramy) są następnie używane do trenowania sieci neuronowych. Takie podejście jest jednym z najbardziej powszechnych sposobów klasyfikacji dźwięku.

3.2. Sposób badania

Dane podzielone zostały na testowe i treningowe. Model VGGish automatycznie dzieli nagrania na jednosekundowe elementy. Oznakowane wektory cech z danych treningowych użyte zostały w celu wytrenowania klasyfikatora losowego lasu decyzyjnego RandomForest z biblioteki sklearn.ensemble. Używane były domyślne parametry (zastosowanie Grid Search w celu zbadania możliwych kombinacji parametrów wykazało, że domyślne parametry osiągają w tym problemie najlepsze rezultaty).

3.3. Wyniki

3.3.1 Dokładność (accuracy)

Dla danych testowych, wygenerowany klasyfikator wykazał dokładność na poziomie około 87%.

3.3.2 Macierz pomyłek

Poniższa tabela prezentuje macierz pomyłek, w której wiersze odpowiadają poprawnym decyzjom klasyfikacyjnym, a kolumny decyzjom przewidzianym przez klasyfikator.

		Wartość przewidziana		
		Wdech	Wydech	Cisza
Wartość rzeczywista	Wdech	113	26	5
	Wydech	4	146	4
	Cisza	6	5	85

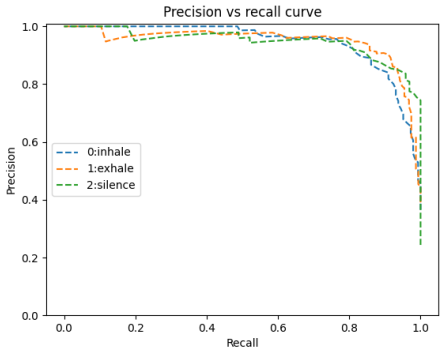
Macierz pomyłek prezentuje niewielką liczbę błędnych decyzji, przy czym pomyłki rozkładają się w większości równomiernie pomiędzy klasami. Wyjątkiem jest klasyfikacja wdechu jako wydechu - pomyłka ta występuje kilkukrotnie częściej niż pozostałe błędy. Może wynikać to z faktu, że niektóre nagrania wdechu w używanym zbiorze testowym są ciężkie do sklasyfikowania nawet dla człowieka, ponieważ brzmią bardzo podobnie do wdechu. Niestety takim przypadkom ciężko przeciwdziałać.

3.3.3 Precyzja i czułość

Dla klasyfikacji poszczególnych klas otrzymano przedstawione poniżej wartości precyzji (rozumianej jako iloraz klasyfikacji prawdziwie pozytywnych przez sumę prawdziwie oraz fałszywie pozytywnych) oraz czułości (jako iloczynu klasyfikacji prawdziwie pozytywnych przez sumę prawdziwie pozytywnych i fałszywie negatywnych). Dodatkowo obliczono wartość F1, stanowiącą średnią harmoniczną precyzji oraz czułości.

	Wdech	Wydech	Cisza
Precyzja	0.89	0.84	0.91
Czułość	0.81	0.93	0.89
F1	0.85	0.88	0.90

Zależność między wartościami precyzji i czułości dla różnych wartości progu klasyfikacji (w tym wypadku prawdopodobieństwa wystarczającego do przewidzenia danej klasy) prezentuje się następująco:



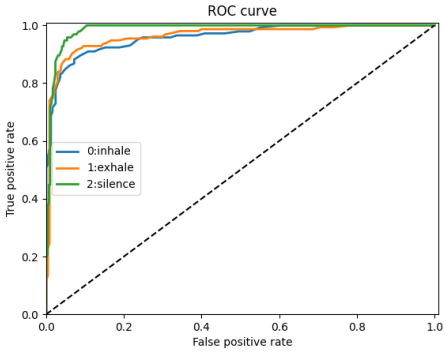
Wyniki uzyskane dla poszczególnych klas uśredniono stosując metodę macro-averaging, przyznającą każdej klasie taką samą wagę. Ostatecznie otrzymano wyniki:

Precyzja: 0.88
Czułość: 0.87
F1: 0.88

Uzyskane wartości są zbliżone, co świadczy o odpowiednim kompromisie pomiędzy precyzją a czułością, a zatem właściwie dobranym progu decyzyjnym.

3.3.4 Krzywa ROC

Wykres krzywej ROC (receiver operating characteristic) dla badanej metody przedstawiony został poniżej.



Oś pionowa (true positive rate) reprezentuje *czułość*, czyli odsetek obserwacji należących do danej klasy, które zostały przez badany klasyfikator poprawnie zaklasyfikowane do tej klasy. Oś pozioma (false positive rate) reprezentuje wartość 1-*specyficzność*, czyli odsetek obserwacji nienależących do danej kasy, które zostały niepoprawnie zaklasyfikowane jako należące do tej klasy.
Dla każdej z analizowanych klas obliczono wartość AUC, czyli pole pod krzywą ROC. Wyniosły one odpowiednio:

- Wdech: 0.96
- Wydech: 0.96
- Cisza: 0.98

Na tej podstawie można wywnioskować, że klasyfikator najlepiej radzi sobie z wykrywaniem ciszy - osiąga wysoką wartość TPR, utrzymując jednocześnie FPR na akceptowalnie niskim poziomie.

3.3.5 Klasyfikacja w czasie rzeczywistym

Klasyfikator osiągnął wysoką skuteczność zarówno w wykrywaniu oddechu w warunkach cichych, jak i w otoczeniu z lekkim szumem.

3.4. Wnioski

To podejście okazało się skuteczne. Zaobserwowaną wadą jest konieczność analizy jednosekundowych próbek, podczas których oczywiście istnieje szansa na nagranie fragmentów kilku klas (np. wdech i wydech na jednym jednosekundowym nagraniu). Jednym z przebadanych sposobów na radzenie sobie z tego typu problemami było powielanie nagrań poprzez kopiowanie próbek dźwięku. Możemy zastosować takie podejście ze względu na fakt, że wdech, wydech oraz cisza brzmią jednorodnie na całej swojej długości, co oznacza, że każdy fragment nagrania jest rozpoznawalny tak samo dobrze, jak całe nagranie. W ten sposób do modelu VGGish przekazywane było na przykład sklonowane półsekundowe nagranie, które model traktował jako sekundowe. Model dla takich sklonowanych nagrań wykazywał praktycznie tak samo dobre wyniki, a odświeżanie było dzięki temu dwa razy częstsze. Równie dobrze wyniki wykazało powielanie ćwierćsekundowych nagrań.

4. Analiza spektralna dźwięku

4.1. Użycie spektrogramu

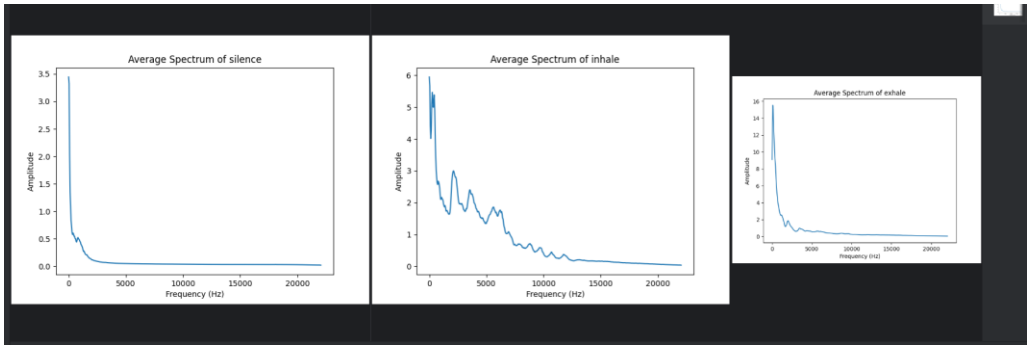
Spektrogram to obraz powstały z próbki dźwięku. Na osi X znajdują się kolejne widma obliczone za pomocą szybkiej transformacji Fouriera, natomiast na osi Y znajdują się kolejne badane pasma. Kolor każdego piksela odpowiada amplitudzie na danym paśmie w danym momencie czasu nagrania. Dzięki zastosowaniu spektrogramu, rozpoznawanie oraz klasyfikację dźwięku sprowadza się do klasyfikacji obrazów, co pozwala na zastosowanie sieci neuronowych właśnie do tego przystosowanych. Takie podejście jest jednym z podstawowych sposobów rozpoznawania i klasyfikacji dźwięku.

4.2. Sposób badania

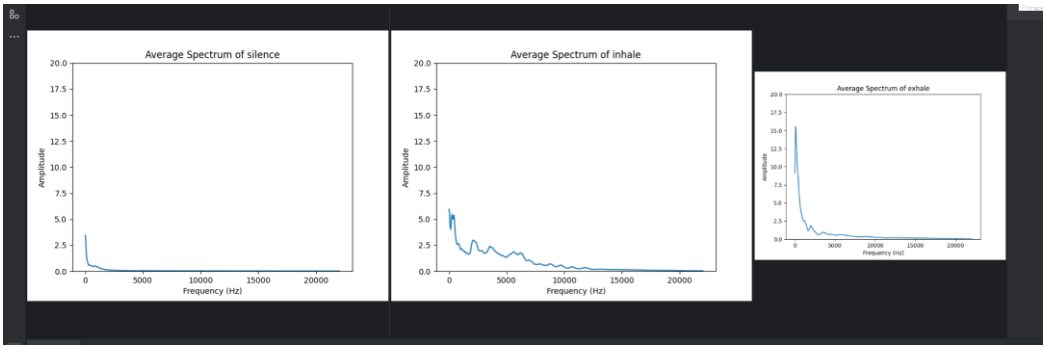
Zbiór danych testowych i treningowych został podzielony na ćwierć bądź półsekundowe, oznaczone nagrania. Dla każdego takiego nagrania tworzony jest spektrogram. Następnie obrazy używane są do uczenia sieci neuronowej. Zastosowany został gotowy model sieci neuronowej do klasyfikacji obrazów MobileNet z biblioteki Tensorflow.

Proste zastosowanie gotowej funkcji do tworzenia spektrogramów z biblioteki librosa nie dawało pożądanych efektów. Przygotowany model posiadał bardzo niską dokładność i nie radził sobie w czasie rzeczywistym. Wynikało to najprawdopodobniej z zapisania obrazu jako png oraz zmianie jego rozdzielczości z powodu wymaganej przez używaną przez nas sieć rozdzielczości 224 x 224 pikseli. Mogło to powodować niepotrzebne straty cennych cech, niezbędnych przy badaniu podobnych dźwięków jakimi są wdech i wydech. Nie pomagało również stosowanie bardziej skomplikowanych sieci, typu EfficientNet czy VGG. W modelach opartych na tych sieciach dochodziło do przetrenowania.

Zadziałanie analizy spektralnej wymagało wyekstrahowania maksymalnej ilości cech z dźwięku. Jednym z problemów mogły być słabe jakościowo dane treningowe. Poprzednio używany zbiór 1000 nagrań zawierał również takie niemożliwe do klasyfikacji nawet przez człowieka. W tym celu stworzono testowy zbiór 90 nagrań (30 wdechów, 30 wydechów oraz 30 nagrań ciszy), zawierający sprawdzone i przefiltrowane ręcznie przez człowieka dane wybrane w taki sposób, aby były możliwe do bezproblemowego rozpoznania przez słuchającą je osobę. Poprzednio używane dane mogły przez niedbałość w ich zbieraniu jedynie działać na niekorzyść modelu, a analizowany przez nas problem niekoniecznie wymaga dużej ilości danych do poprawnego wyuczenia. Ponadto zbadane zostały widma sygnału dla każdej z klas. Poniżej przedstawiono średnią arytmetyczną widm (wartości amplitud dla każdej częstotliwości zostały po prostu uśrednione). Dało to następujące efekty:



Przeskalowana oś Y dla wydechu oraz ciszy powoduje, że same w sobie wykresy stają się niemal identyczne. Można to bardzo łatwo wyjaśnić: podczas wydechu człowiek właściwie dmucha w mikrofon co powoduje powstanie dźwięku podobnego do bardzo głośnego szumu, tego samego szumu, który zawiera się w nagraniach ciszy. Nadanie uśrednionym widmom równej skali pozwala już na zauważenie różnic pomiędzy każdą z klas:



Zauważyć można, że największe różnice widoczne są na niższych częstotliwościach. Na częstotliwościach powyżej 10 kHz amplitudy pasm są dla wszystkich trzech klas podobne: bliskie zeru. Oznacza to, że najlepiej pod uwagę brać głównie niższe częstotliwości. Już na tym etapie można spekulować, że utworzony model będzie najprawdopodobniej wydech od ciszy rozróżniał głównie po głośności.

Wymagana przez sieć rozdzielczość obrazu wymaga dostosowania parametrów do tworzenia spektrogramów. W tym celu zastosować można szybką transformację Fouriera z biblioteki scipy w Pythonie, która pozwoli na zwrócenie macierzy, składającej się z wektora obliczonych za pomocą okna przesuwnego widm w kolejnych momentach trwania dźwięku. Ważne jest dobranie parametrów transformacji tak, aby transformacja zwróciła macierz jak najbardziej zbliżoną do wymaganych 224 x 224 pikseli. Oczywiście nadmiarowe widma można z macierzy obciąć bez większej straty jakości cech. Dzięki zastosowaniu odpowiedniej liczby próbek używanych do transformacji manipulować można przedziałami pasm, które spektrogram brać ma za uwagę. Przykładowo, jeżeli transformacja zostanie wykonana dla 1024 punktów, czyli otrzymano 513 pasm częstotliwościowych w równych odległościach, a częstotliwość próbkowania wynosi 44,1 kHz, czyli badamy częstotliwości do około 22 kHz (zakres słyszalności człowieka), to użycie do spektrogramu 224 pierwszych pasm spowoduje, że pod uwagę brane będą jedynie pasma częstotliwości od 0 Hz do około 10 kHz.

Ostatnim problemem jest wymagany przez MobileNet format RGB obrazu. Spektrogram nie jest do końca obrazem, tylko macierzą składającą się z wektorów widm, gdzie każde widmo to

wektor amplitud na kolejnych pasmach częstotliwości. Spektrogram może być prezentowany za pomocą kolorowego obrazu, jednak zapisywanie spektrogramu w formacie używającym kolorów RGB (na przykład w formacie png) mogłoby powodować niepotrzebną utratę cech. Dlatego spektrogramy najlepiej traktować jako macierze (format npy obsługiwany przez bibliotekę numpy w Pythonie pozwala na zapisywanie macierzy liczb bez kompresji na komputerze) i tak przekazywać do modelu w celu uczenia. Wymagany przez MobileNet trzywymiarowy format koloru rozwiązano poprzez nałożenie na siebie trzy razy tej samej macierzy (rozmiar 224x224x1 zamieniono w ten sposób na rozmiar 224x224x3), co nie powinno utrudnić sieci odpowiedniej klasyfikacji. Ponadto wyłączono sieć bazową MobileNet wraz z pretrenowanymi wagami z uczenia, ponieważ powodowało to przetrenowanie modelu, najpewniej z powodu zbyt dużej liczby parametrów. Finalnie uczeniu poddane były jedynie warstwa wejściowa zastosowana przed warstwami MobileNet w celu zmiany rozmiaru danych wejściowych z jedno na trójwymiarowy kolor, a po bazowej sieci MobileNet: warstwa poolingu, warstwa spłaszczająca, warstwa Dropout z ze współczynnikiem 50% oraz wyjściowa warstwa gęsta z funkcją aktywacji softmax służąca do klasyfikacji na 3 klasy.

Klasyfikacja w czasie rzeczywistym polega w tym momencie na pobieraniu ćwierć lub półsekundowych nagrań, tworzeniu z nich macierzy 224x224 reprezentujących spektrogramy, a następnie klasyfikacji tychże spektrogramów przez wcześniej wytrenowaną sieć neuronową.

4.3. Wyniki

Zbadano efektywność klasyfikacji modelu dla nagrań o długości 0.5s oraz 0.25s oraz dla punktów użytych dla transformacji Fouriera w ilości: 512, 1024, 2048 oraz 4096. Najlepsze wyniki zaobserwowano dla modeli uczonych na spektrogramach, które stworzono dla 2048 punktów szybkiej transformacji Fouriera (0.5s oraz 0.25s). Tych właśnie modeli wyniki będą porównywane.

4.3.1 Dokładność (accuracy)

Dokładność dla kolejnych modeli wyniosła:

- Mobile Net, 2048, 0.5 - 83%
- Mobile Net, 2048, 0.25 - 83%

4.3.2 Macierz pomyłek

Uzyskano następujące macierze pomyłek:

	Wartość przewidziana			
		Wdech	Wydech	Cisza
	Mobile Net, 2048, 0.5s			
Wartość rzeczywista	Wdech	75	31	0
	Wydech	15	92	1
	Cisza	15	2	140
	Mobile Net, 2048, 0.25s			
	Wdech	184	41	1
	Wydech	43	181	5
	Cisza	25	13	281

4.3.3 Precyzja i czułość

Dla kolejnych modeli uzyskano przedstawione poniżej wartości precyzji i czułości dla każdej z analizowanych klas.

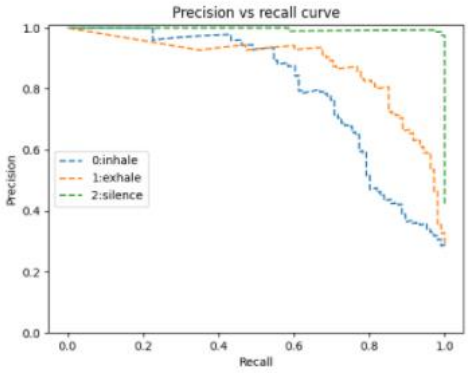
	Mobile Net, 2048, 0.5s			Mobile Net, 2048, 0.25s		
	Wdech	Wydech	Cisza	Wdech	Wydech	Cisza
Precyzja	0.71	0.74	0.99	0.73	0.77	0.98
Czułość	0.71	0.85	0.89	0.81	0.79	0.88
F1	0.71	0.79	0.94	0.77	0.78	0.93

Uzyskane wyniki uśredniono za pomocą metody macro-averaging, nadającej każdej klasie równą wagę.

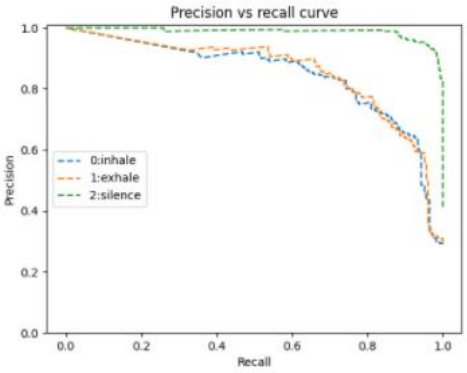
	Mobile Net, 2048, 0.5s	Mobile Net, 2048, 0.25s
Precyzja	0.81	0.83
Czułość	0.82	0.83
F1	0.81	0.83

Uzyskano kompromis między precyzją a czułością.

Wykresy zależności między precyzją a czułością dla kolejnych modeli zaprezentowane są poniżej.



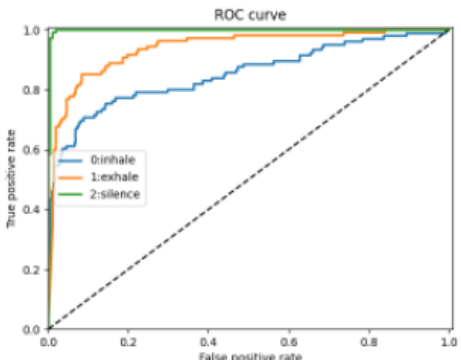
Mobile Net, 2048, 0.5s



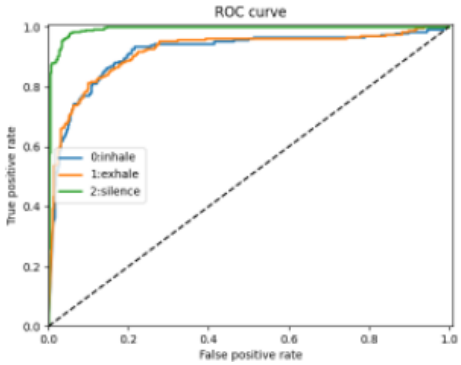
Mobile Net, 2048, 0.25s

4.3.4 Krzywa ROC

Dla każdego modelu wyrysowano krzywą ROC oraz obliczono wartość AUC (area under curve).



Mobile Net, 2048, 0.5s



Mobile Net, 2048, 0.25s

Krzywe obrazują zależność między odsetkiem klasyfikacji prawdziwie pozytywnych (TPR) a odsetkiem fałszywie pozytywnych (FPR).

Wyznaczono także wartości AUC – pola pod wykresem krzywych:

	Mobile Net, 2048, 0.5s	Mobile Net, 2048, 0.25s
Wdech	0.85	0.91
Wydech	0.94	0.92
Cisza	0.99	0.99

Na podstawie powyższych danych można zauważyć, że wszystkie przetestowane modele najlepiej radzą sobie z wykrywaniem ciszy, osiągając wysoki poziom TPR dla niskiego FPR. Jest to szczególnie widoczne dla pierwszego modelu (Mobile Net, 2048, 0.5), dla którego krzywa ROC osiąga wartość bliską 1 niemal na całym przedziale. W przypadku wydechu pierwszy model (Mobile Net, 2048, 0.5) także radzi sobie lepiej od pozostałych, choć zauważalnie gorzej niż dla ciszy. Dla wdechu jego skuteczność spada i efektywniejszy okazuje się wówczas drugi model (Mobile Net, 2048, 0.25), który osiąga zbliżone wyniki dla klas wdech i wydech.

4.4. Wnioski

Skutecznie rozwiązano problem dokładności analizy spektralnej, poprzez opisane powyżej dokładne ekstrahowanie cech. Model wykazuje dobrą skuteczność przy zarówno analizie gotowych nagrań jak i przy detekcji w czasie rzeczywistym, jednakże zdarzają się również złe klasyfikacje podczas jego używania. Wyższą skuteczność można by uzyskać najprawdopodobniej po dokładniejszym zbadaniu widm sygnału oraz hiperparametrów sieci, bądź łącząc jego działanie z jakąś inną metodą w celu wspólnego podejmowania decyzji o klasyfikacji. Nie jest to zaskoczeniem, że najlepsze wyniki uzyskano dla 2048 punktów do szybkiej transformacji Fouriera – zgodnie z obliczeniami omówionymi w punkcje 2. tego rozdziału, oznacza to, że model brał pod uwagę częstotliwości do około 5kHz, czyli właśnie te, w których najbardziej zauważalne były różnice pomiędzy widmami klas.

5. Wykorzystanie współczynników cepstralnych

5.1. Współczynniki cepstralne (MFCC)

Współczynniki cepstralne (Mel Frequency Cepstral Coefficients - MFCC) to opis częstotliwościowy dźwięku, który wykorzystuje skalę melową, aby lepiej odzwierciedlać ludzką percepcję dźwięku. W porównaniu ze spektrogramami, MFCC skupiają się na informacjach w paśmie słyszalnym dla człowieka, co może prowadzić do lepszej klasyfikacji dźwięku.

5.2. Sposób badania

Podobnie jak w przypadku analizy spektrogramów, zbiór danych testowych i treningowych podzielono na półsekundowe, oznaczone nagrania. Dla każdego nagrania obliczono współczynniki MFCC. Następnie dane te zostały użyte do uczenia sieci neuronowej. W tym badaniu testowano własne modele sieci, oparte na:

- kilku warstwach konwolucyjnych oraz kilku warstwach gęstych. Zastosowano także technikę regularyzacji “Dropout”, która pomogła zapobiec przeuczaniu się modelu.
- trzech warstwach LSTM oraz trzech warstwach gęstych.

Liczbę warstw konwolucyjnych jak i gęstych dostosowywano metodą prób i błędów. Mimo wielu prób treningu sieci opisanej w podpunkcie 1., opisany model dla danych walidacyjnych wykazywał niższą dokładność, dlatego skupiono się na zastosowaniu sieci rekurencyjnej.

5.3. Wyniki

5.3.1 Dokładność (accuracy)

Dla danych testowych osiągnięto dokładność na poziomie 87%.

5.3.2 Macierz pomyłek

Poniższa tabela prezentuje macierz pomyłek, w której wiersze odpowiadają poprawnym decyzjom klasyfikacyjnym, a kolumny decyzjom przewidzianym przez klasyfikator.

		Wartość przewidziana		
		Wdech	Wydech	Cisza
Wartość rzeczywista	Wdech	27	2	1
	Wydech	3	27	0
	Cisza	5	1	24

5.3.3 Precyzja i czułość

Dla analizowanych klas uzyskano następujące wartości precyzji i czułości. Obliczono także wartość F1.

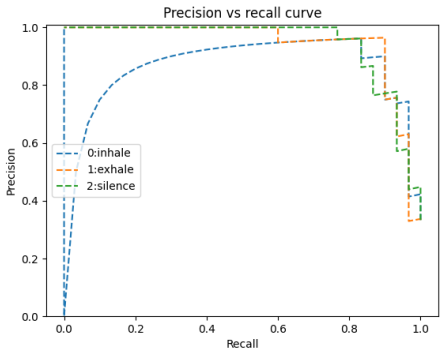
	Wdech	Wydech	Cisza
Precyzja	0.90	0.77	0.96
Czułość	0.90	0.90	0.80
F1	0.90	0.83	0.87

Do uśrednienia wyników otrzymanych dla poszczególnych klas wykorzystano metodę macro-averaging, przyznającą każdej klasie taką samą wagę. Ostatecznie otrzymano wyniki:

Precyzja: 0.88
Czułość: 0.87
F1: 0.87

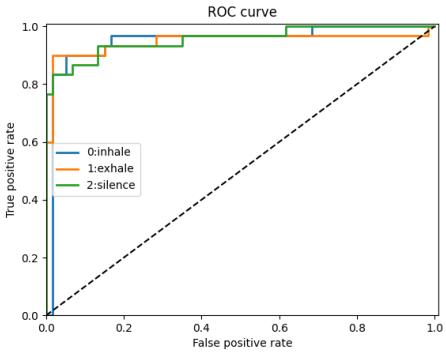
Uzyskane wartości są zbliżone, zatem osiągnięty został kompromis między precyzją a czułością.

Analizowana zależność przedstawiona została na poniższym wykresie.



5.3.4 Krzywa ROC

Wykres krzywej ROC (receiver operating characteristic) dla badanej metody przedstawiony został poniżej.



Wartości AUC (area under curve) dla poszczególnych klas wyniosły odpowiednio:

- Wdech: 0.95
- Wydech: 0.95
- Cisza: 0.96

Powyższe wyniki prowadzą do wniosku, że model osiąga podobną skuteczność dla każdej z analizowanych klas, z lekką przewagą wykrywania ciszy, jeśli wziąć pod uwagę pole pod wykresem. Dla niskiego odsetka fałszywie pozytywnych (FPR poniżej 0.05) najwyższą wartość odsetka prawdziwie pozytywnych (TPR) udało się osiągnąć dla klasy wydechu.

5.4. Wnioski

To podejście okazało się skuteczne. Model dokonywał poprawnej klasyfikacji w większości przypadków. Występowały natomiast problemu w przypadku zwiększonego tempa oddychania. Wynika to z faktu, że model pobiera półsekundowe próbki i w przypadku, wystąpienia dwóch różnych oddechów, to model sklasyfikuje je jako jeden.

6. Wykrywanie ekstremów głośności

6.1. Podejście

Badanie ekstremów głośności opierało się jedynie na wczytywaniu wejścia z mikrofonu i wykorzystaniu informacji o zmianach głośności w prostym algorytmie rejestrującym te zmiany. Nie wykorzystano w nim uczenia maszynowego, a jedynie informacje o schematyczności ludzkiego oddechu.

6.2. Sposób badania

6.2.1 Przygotowanie algorytmu

Aby algorytm działał poprawnie, konieczne było określenie jaki próg głośności jest potrzebny, aby algorytm mógł zarejestrować zmianę i dokonać stosownej operacji.

W tym celu pobrane zostały próbki z danych treningowych tj. wdechy, wydechy i cisza. Długość próbek dla trzech rodzajów próbek została znormalizowana. Głośność dźwięku, dla każdego rodzaju próbek została uśredniona. Aby uzyskać interesujący nas próg głośności wystarczyło dodać odpowiednie średnie tj.:

1. Średni próg głośności dla wdechu = średnia głośność wdechu + średnia głośność ciszy
2. Średni próg głośności dla wydechu = średnia głośność wydechu + średnia głośność ciszy

Algorytm opiera się o schematyczność oddechu - zakłada, że po wdechu zawsze jest cisza, po której następuje wydech i po wydechu zawsze następuje cisza, po której jest wdech. Algorytm otrzymuje na wejściu aktualnie pobraną próbkę z mikrofonu i porównuje jej głośność z wcześniej wyliczonymi średnimi. Zapisuje poprzedni stan, kiedy wykrył wyższą głośność (wdech/wydech) i aktualny stan, żeby sprawdzić, czy pomiędzy próbkami nastąpiła cisza. Jeśli tak to zmienia aktualnie przechowywany stan. Algorytm zwraca informacje o tym co aktualnie zarejestrował.

6.2.2 Przygotowanie programu działającego w czasie rzeczywistym.

Program używa wyżej opisanego algorytmu do wyświetlania wyniku - jaki stan oddechu jest aktualnie rejestrowany. W pierwszych kilku sekundach działania, program wczytuje szum z mikrofonu. Potem uśrednia otrzymaną wartość i przy każdym kolejnym pobraniu próbek normalizuje bufor wejściowy, aby jak najdokładniej określać głośność. Przygotowaną próbkę wysyła do algorytmu i według zwróconej wartości koloruje wykres na odpowiedni kolor (niebieski: cisza, czerwony: wdech, zielony: wydech).
Program może działać w dwóch trybach: oryginalnym i eksperymentalnym. Tryb oryginalny opisany jest wyżej, tryb eksperymentalny polega na porównaniu sąsiednich wartości jednego, analizowanego fragmentu próbki. Jeśli fragment próbki jest wyjątkowo krótki i nie pasuje do sąsiadujących z nim fragmentów to taki wynik zostanie potraktowany jako błąd wejścia i program kontynuuje swoją pracę poprawnie.

6.3. Wyniki

Opisana metoda nie wykorzystywała żadnego z modeli uczących więc przedstawienie wyników liczbowych jest trudne. W warunkach idealnej ciszy i systematycznego, równomiernego oddechu, program działa z dużą skutecznością. Zdarzają się jednak pomyłki powodowane przez dźwięki zewnętrzne i nagłe zmiany w głośności wdechu/wydechu.

6.4. Wnioski

Opisana metoda jest bardzo podatna na zakłócenia. Jeśli dookoła mikrofonu znajdują się inne źródła dźwięku, niż oddychający człowiek, to program zgubi rytm oddechu i będzie niepoprawnie analizował zachodzące zmiany. Zakłóceniem może okazać się też nierównomierny oddech – chwilowe zawieszenie wdechu/wydechu i kontynuowanie go po krótkiej chwili może przyczynić się do niepoprawnego zarejestrowania głośności i tym samym do przestawienia algorytmu. Częściowe rozwiązanie problemu stanowi tryb eksperymentalny

7. Podsumowanie

7.1. Porównanie wyników każdej z metod

7.1.1. Porównanie dokładności oraz F1

Do porównania rezultatów uzyskanych przy pomocy zastosowanych metod
wybrane zostały następujące metryki wydajności:

- Dokładność, rozumiana jako iloraz prawidłowych predykcji przez całkowitą ich liczbę.
- Wartość F1, obliczona przy pomocy uzyskanych wartości precyzji i czułości oraz uśredniona dla analizowanych klas przy zastosowaniu metody macro-averaging, nadającej taką samą wagę każdej klasie.

Porównanie w formie tabeli obejmuje następujące podejścia:

- Klasyfikacja za pomocą lasu losowego na podstawie cech uzyskanych przy wykorzystaniu modelu VGGish.
- Klasyfikacja spektrogramu dla ćwierćsekundowych nagrań wykorzystująca sieć Mobile Net z przyjętą liczbą 2048 punktów Fouriera.
- Klasyfikacja przy wykorzystaniu współczynników cepstralnych częstotliwości w skali melowej (MFCC).

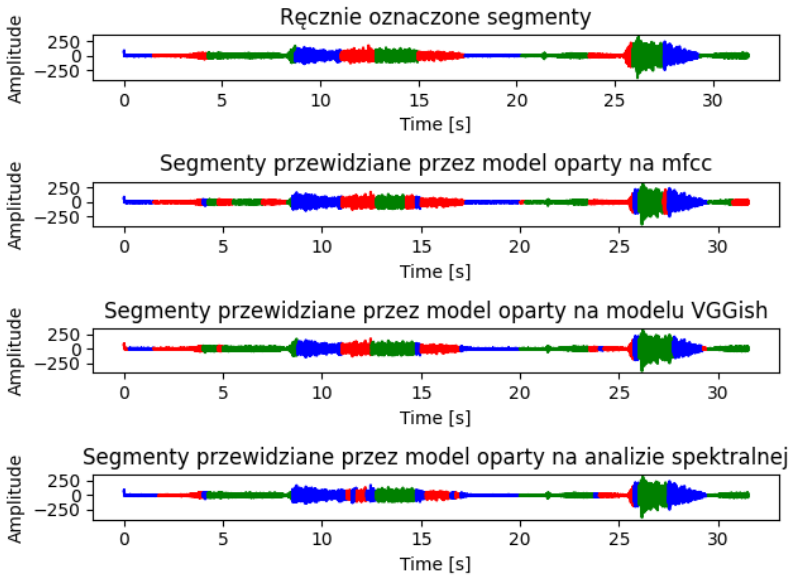
Ze względu na odmienną specyfikę i trudności z uzyskaniem wiarygodnych wartości analizowanych metryk, podejście związane z wykrywaniem ekstremów głośności nie zostało uwzględnione w poniższym zestawieniu.

Podejście	Metryka	
	Dokładność	F1
VGGish, las losowy	87%	0.88
Analiza spektogramów	83%	0.83
MFCC	87%	0.87

Z zestawienia wynika, że zastosowane podejścia cechują się podobnym poziomem wydajności dla analizowanego problemu, przy czym zastosowanie współczynników cepstralnych częstotliwości oraz modelu VGGish w połączeniu z lasem losowym pozwoliło uzyskać najwyższe wartości analizowanych metryk.

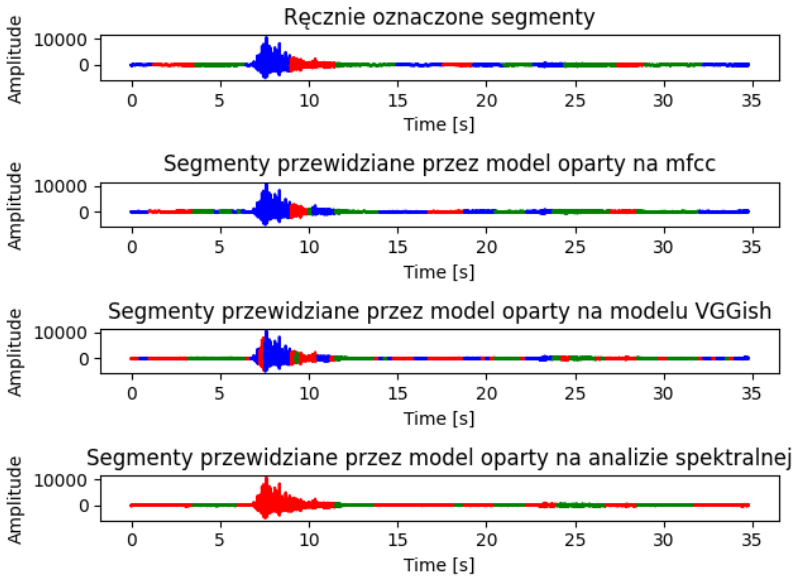
7.1.2. Porównanie działania modeli na identycznym, dłuższym nagraniu oraz na różnej jakości mikrofonów.

Kolejną metodą porównania modeli jest sprawdzenie ich sprawowania na około trzydziestosekundowych nagraniach. Nagrania te utworzono przy pomocy mikrofonów różnej jakości oraz ręcznie oznaczono występujące w nim klasy.
Dla dobrego mikrofonu, na którym tworzona była duża część danych treningowych wyniki są następujące:



Modele bardzo dobrze obrazują rytm oddychania z nagrania, natomiast zauważyć można błędne klasyfikacje pojedynczych, ćwierćsekundowych odcinków na całej długości wykresów co mimo wszystko nie wpływa na wartość informacyjną takich wykresów. Najgorzej wypadł model oparty na analizie spektralnej, którego najczęstszym błędem widocznym na wykresie była klasyfikacja wdechu jako ciszy. Wszystkie modele w około 26 sekundzie nagrania klasyfikują parę segmentów jako ciszę, wbrew temu jak oznaczone zostało nagranie przez człowieka, natomiast może wynikać to z krótkiej przerwy podczas zamiany wdechu na wydech co nie zostało oznaczone przez oznaczającego nagranie.

Dla mikrofonu niskiej jakości znajdującego się w tanich słuchawkach dousznych zaobserwowano następujące wyniki:



Zauważyć można więcej błędów. Najlepiej poradził sobie model oparty na współczynnikach cepstralnych. Pomimo większej ilości błędów dobrze wypadł również model oparty na modelu VGGish. Najgorzej poradził sobie model oparty na analizie spektralnej. Dla słabej jakości mikrofonu nie wykrywał on w ogóle ciszy. Cisza była w większości wykrywana jako wdech. Pomijając to, model ten dobrze zmieniał klasy podczas zmiany fazy oddychania, jednak problemy z wykrywaniem ciszy powodują, że model nie nadaje się do użycia przy stosowaniu złej jakości mikrofonów.

Podsumowując, rodzaj mikrofonu ma decydujący wpływ na jakość predykcji analizowanych modeli. Najbardziej uniwersalny okazał się model oparty na współczynnikach cepstralnych. Daje on satysfakcjonujące efekty zarówno dla dobrych jak i złych jakościowo mikrofonów. Najgorzej ze słabej jakości mikrofonami radzi sobie model oparty na analizie spektralnej. Dla niskojakościowych mikrofonów predykcje modelu są na tyle złe, że nie nadają się do jakiegokolwiek użycia.

7.2. Wnioski i możliwe kierunki dalszych badań

Finalnie problem nie został w pełni rozwiązany. Przedstawione techniki wykrywania oddechu wykazują wystarczającą skuteczność, w celu stosowania ich na przykład do określania przybliżonej częstotliwości oddychania, zwłaszcza przy zastosowaniu dodatkowych technik w celu ewentualnej korekcji pomyłek modeli. Mimo to, zaimplementowane techniki detekcji oddechu są podatne na pomyłki, co może się okazać niedopuszczalne przy ich zastosowaniach na przykład w medycynie, która może wymagać praktycznie perfekcyjnej klasyfikacji.