Detekcja mowy w filmach

Magdalena Cisowska 160527

29 października 2019

1 Wprowadzenie

Detekcja mowy w sygnale znajduje szerokie zastosowanie. Bardzo często jest to pierwsza część przetwarzania sygnału zawierającego mowę. Przykładem jest wykrywanie mowy w rozmowach telefonicznych, pozwalających następnie na rozpoczęcie procesu rozpoznawania jej w wykrytych fragmentach sygnału. Spotyka się także rozwiązania, w których jakość rozmowy telefonicznej jest obniżana, gdy w sygnale nie zostaje w danej chwili wykryta mowa. Z punktu widzenia filmów, detekcja mowy również może znaleźć zastosowanie, np. w automatycznej translacji mowy na napisy.

2 Przegląd dostępnych metod detekcji

Według źródeł [2][4] istnieje wiele metod detekcji mowy w sygnale. Używa się do tego GMM (*Gaussian Mixture Model*), sieci neuronowych czy różnego rodzaju klasyfikatorów (np. SVM - *Support Vector Machine*). Praca ta skupi się na dwóch podejściach - głębokiego uczenia DNN (*Deeep Neural Network*) oraz wspomnianego klasyfikatora SVM. Przedstawione zostaną również dwie metody ekstrakcji cech z sygnału dźwiękowego zawierającego mowę.

2.1 Podejście pierwsze

W modelu stosowanym w [2] struktura głębokiej sieci neuronowej przedstawia się następująco: 1053 neuronów wejściowych, 3 ukryte, w pełni połączone warstwy po 512 neuronów, oraz warstwa wyjściowa składająca się z dwóch neuronów. Funkcje aktywacji warstw ukrytych to *ReLU* (*Rectified Linear Unit*).

Jako dane wejściowe wykorzystane zostało 13 ustandaryzowanych współczynników melcepstralnych danej ramki wejściowej sygnału. Podziału sygnału dokonano co 10 ms, na ramki o długości 25 ms. Aby nadać sekwencji wejściowej więcej informacji o kontekście, oprócz współczynników MFCC (*Mel Frequency Cepstral Coefficients*) z bieżąco rozpatrywanej ramki, dołączono także MFCC z 40 poprzedzających i następujących ramek - łącznie 81 ramek.

Współczynniki MFCC powstają z cepstrum sygnału przedstawionego w skali melowej. Aby uzyskać skalę melową, należy filtrować sygnał bankiem filtrów o charakterystyce trójkątnej, z tymże szerokość filtra zwiększa się wraz z częstotliwością, co odpowiada charakterystyce wrażliwości ludzkiego ucha na wysokości dźwięku. Cepstrum uzyskuje się za pomocą odwrotnej transformaty Fouriera z logarytmu widma badanego sygnału.

Według autorów skuteczność tego systemu wynosi od 80 % do 90 %. Warto jednak wspomnieć, że zbiór uczący podzielony został na nagrania zawierające m.in. mowę i dźwięk, mowę

i szum oraz mowę i śpiew. Użyty zbiór nagrań to korpus 65 godzin nagrań filmów z serwisu *YouTube* - zbiór *HAVIC* [3].

2.2 Podejście drugie

Model użyty w [4] znacząco różni się od przedstawionego powyżej. Do klasyfikacji ramek (takich samych, jak w podejściu pierwszym), jako zawierające mowę lub nie, posłużyło 5 cech. Były to:

- logarytm energii ramki:
$$E = log(\sum_{n=1}^{L} x(n)^2), \tag{1}$$
- ilość przejść przez zero sygnału ramki,
- znormalizowany współczynnik autokorelacji (względem opóźnienia jednej próbki) dany wzorem:
$$C = \frac{\sum_{n=1}^{L-1} x(n)x(n-1)}{\sqrt{\sum_{n=1}^{L-1} x(n)^2 \sum_{n=1}^{L-1} x(n-1)^2}}, \tag{2}$$
- pierwszy współczynnik zidentyfikowanego modelu funkcji sygnału ramki,
- logarytm błędu zidentyfikowanego modelu względem rzeczywistego sygnału ramki.

Następnym krokiem była klasyfikacja ramek przy użyciu SVM. Klasyfikator, na podstawie wyodrębnionych cech, umieszcza je w 5 wymiarowej przestrzeni, a następnie stara się utworzyć linię separującą zbiór elementów zawierających mowę od tych nie zawierających. Autorzy nie wspominają o dokładności implementacji.

3 Implementacja i wyniki

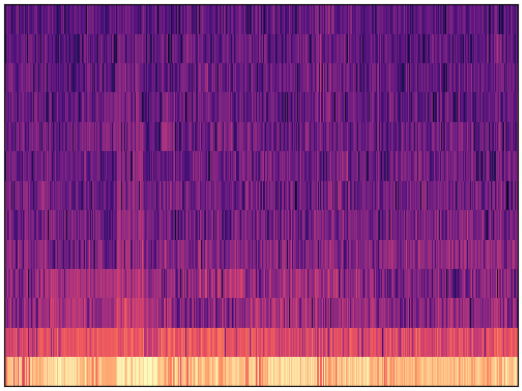
Niniejsza praca bazuje na implementacjach dwóch opisanych wyżej podejść. Dane wejściowe przygotowane zostały w ten sam sposób, dokonano jedynie niewielkich modyfikacji modeli detekcji.

3.1 Zbiór uczący i jego przygotowanie

Zestaw filmów użytych do detekcji mowy w sygnale to zbiór *AVA-Speech* [1]. Zawiera on kilkadziesiąt filmów w różnych językach, z których w każdym oznaczone są 15 minutowe fragmenty. Oznaczenia jako mowa czysta, mowa z zakłóceniami oraz brak mowy dokonane zostały ręcznie przez twórców. Zbiór ten jest darmowy i dostępny do pobrania. W implementacji wspomniane fragmenty zostały wycięte, a następnie wyodrębniono z nich sygnał audio. Ten został następnie pocięty na 25 ms ramki, co okres 10 ms. Uzyskano w ten sposób dane wejściowe do obydwu poniższych modeli. Liczba ramek dla każdego filmu wynosiła 90000. Podczas testów algorytmów wybrano 3 pliki, 2 uznane wg autora za łatwe w ocenie, oraz jeden trudniejszy.

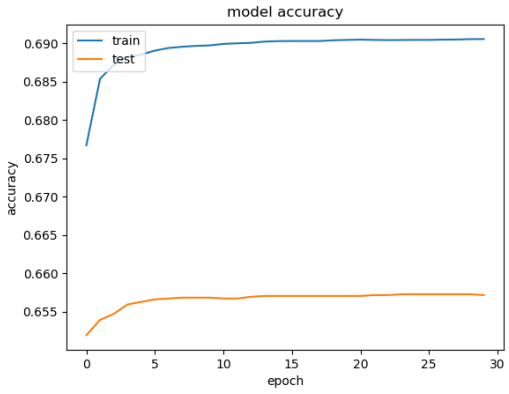
3.2 NN

Z uwagi na złożoność modelu i żmudność obliczeń, postanowiono, że jako dane wejściowe wykorzystane zostaną współczynniki MFCC z łącznie 41 ramek. Zgodnie z [2] dane te ustandaryzowano tak, aby miały zerową wartość średnią oraz jednostkową wariancję.



Rysunek 1: Mel spektrogram przykładowego pliku audio ze zbioru *AVA-Speech*.

Struktura sieci neuronowej wygląda następująco: 533 neuronów wejściowych, dwie warstwy ukryte po 128 neuronów oraz 2 neurony wyjściowe. Funkcja aktywacji to *ReLU*, współczynnik uczenia wynosił 0.0001. Uczenia dokonano najpierw na pojedynczych filmach. Próbki podzielono na zbiór testowy oraz 10 % zbiór walidacyjny, wykorzystano *mini batch* wielkości 50 próbek. Proces uczenia dla każdego pliku charakteryzował się niepoprawnymi wartościami skuteczności sieci. Rozbieżność pomiędzy skutecznością zbioru treningowego oraz walidacyjnego była bardzo duża. Po wytrenowaniu sieć błędnie klasyfikowała ramki testowe, a sekwencja ocen po 30 epokach przypominała ciąg losowych wartości 0 lub 1. Z kolei po 50 epokach sieć jednakowo klasyfikowała cały zbiór testowy jako brak mowy.



Rysunek 2: Przebieg uczenia sieci dla 30 epok.

Przyczyn takiego zachowania się algorytmu może być dużo. Być może dane wejściowe nie pozwalały na dostatecznie dokładną klasyfikację poszczególnych ramek. Możliwe jest także, że zbiór uczący *HAVIC* wykorzystany w pracy [2] był dużo wyższej jakości, a wybrany model został dostosowany specjalnie pod ten konkretny zestaw danych.

3.3 Klasyfikator SVM

Podejście to charakteryzuje się szybszym działaniem. Zestaw cech został tak dobrany, że mimo, iż każda z osobna nie byłaby wystarczająca do poprawnej klasyfikacji, o tyle razem tworzą bardzo dobry zestaw cech danej ramki audio. Testy wykonano na kilku plikach - najpierw model dopasowywano i testowano na danych pochodzących z tego samego filmu. Następnie klasyfikacji dokonano na jednym zbiorze danych, a testu na innym, pochodzącym z innego filmu. Wyniki implementacji przedstawia tabela 1 .

Identyfikacji modelu dokonano za pomocą układu równań Yule-Walkera - identyfikacja parametrów modelu autoregresyjnego na podstawie oszacowanych współczynników autokorelacji sygnału. Rząd modelu wynosił 12, podobnie jak w [4].

Tabela 1: Skuteczność klasyfikatora SVM w zależności od rodzaju danych wejściowych oraz testowych.

Nr pliku uczącego	**Nr pliku testowego**	**Skuteczność**
1	1	85.62%
2	2	68.75 %
3	3	42.45%
1	2	42.3%
3	1	50.23%
1	3	69.92%

3.4 Wnioski

W celu dokonania oceny poprawności działania algorytmów przez człowieka, wyjście klasyfikatorów zamieniane było na plik z napisami do poszczególnych filmów. Pozwoliło to ocenić jakość nie tylko na podstawie porównywania danych procentowych, lecz także obserwując, czy predykcje zgadzają się z sytuacją na ekranie. Dzięki temu zabiegowi widać było wyraźną różnicę pomiędzy obydwoma algorytmami dla "najlepszego" pliku, jakim wybrano plik nr 1. DNN, mimo, że charakteryzował się pewną skutecznością procentową, nie potrafił wygenerować napisów, w których pokrycie się z rzeczywistością byłoby choć zbliżone do akceptowalnego. Z kolei dla metody SVM, udało się to zrobić tylko dla jednego pliku. Dla pliku nr 1 oceny w dużej części przypadków pokrywały się z rzeczywistością, co skutkowało bardzo dobrym odbiorem przez słuchacza. Podczas mowy dźwięcznej, nie zdarzało się algorytmowi popełniać błędów, tak samo było w chwilach ciszy oraz większości hałasów. Zdarzało się natomiast, że niektóre hałasy błędnie kwalifikowane były jako mowa.

Rysunek 3: Przykłady napisów prawidłowo zakwalifikowanych klatek metodą SVM dla pliku nr 1.

To, co odróżniało plik 1 od reszty, to przester, jaki słyszalny był za każdym razem, gdy w sygnale pojawiała się mowa. Aktorzy bardzo często mówili głośno oraz mieli podobne do siebie głosy. Mogło to w znaczącym stopniu ułatwić detekcję mowy. Wydaje się, że metoda SVM lepiej poradziła sobie z zadaniem, choć świadczy o tym tylko jeden udany przypadek. Metoda ta jest jednak dużo szybsza, a więc i wygodniejsza w użyciu.

Źródła

[1] Sourish Chaudhuri et al. "AVA-Speech: A Densely Labeled Dataset of Speech Activity in Movies". In: *CoRR* abs/1808.00606 (2018). arXiv: 1808.00606. URL: http://arxiv.org/abs/1808.00606.

[2] N. Ryant, Mark Liberman, and Jiahong Yuan. "Speech activity detection on youtube using deep neural networks". In: *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH* (Jan. 2013), pp. 728–731.

[3] Stephanie Strassel et al. "Creating HAVIC: Heterogeneous Audio Visual Internet Collection". In: ().

[4] *Voice Activity Detection (VAD) Tutorial*. URL: http://practicalcryptography.com/miscellaneous/machine-learning/voice-activity-detection-vad-tutorial/.