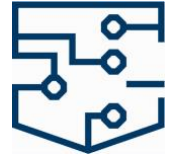




**POLITECHNIKA
GDAŃSKA**

Wydział Elektroniki,
Telekomunikacji i
Informatyki



Mateusz Michalski (kierownik) - research, Keele pitch database,
Michał Pałka - Keele pitch database, GitLab - zapoznanie się z kodem
Łukasz Wróblewski - GitLab - zapoznanie się z kodem
Patrik Kosior - zapoznanie się z IFE & obróbka bazy danych
Radosław Wierzbicki - zapoznanie się z IFE & obróbka bazy danych

Baza sygnałów mowy do treningu i walidacji neuronowego estymatora częstotliwości tonu krtaniowego

Opiekun: dr hab. inż. Marek Blok

Raport z przeglądu metody IFE oraz wykorzystywanych tam baz danych i metod oceny F0

1. Wprowadzenie do metody IFE

Metoda IFE (Instantaneous Frequency Estimation) to nowoczesne podejście do estymacji tonu podstawowego (F0) w mowie, które wykorzystuje głębokie sieci neuronowe do oceny tonu podstawowego na podstawie ekstrakcji natychmiastowej częstotliwości i mocy składowych sygnału (Dokument "IFE: NN-aided Instantaneous Pitch Estimation" autorstwa Marka Bloka, Jana Banasa i Mariusza Pietrolaja, strona 0).

IFE jest szczególnie użyteczne w interakcjach człowiek-komputer opartych na mowie, takich jak tłumienie szumów, separacja źródeł i identyfikacja mówcy. Metoda ta oferuje natychmiastową estymację tonu, co oznacza, że estymacja tonu jest obliczana próbkę po próbce, co jest znaczącym ulepszeniem w porównaniu do tradycyjnych metod, które polegają na okienkowaniu czasowym (Dokument "IFE: NN-aided Instantaneous Pitch Estimation" autorstwa Marka Bloka, Jana Banasa i Mariusza Pietrolaja, strona 0).

2. Architektura sieci neuronowej w IFE

IFE wykorzystuje architekturę głębokiej sieci neuronowej opartej na perceptronie wielowarstwowym. Warstwa wejściowa została zaprojektowana tak, aby dokładnie pasować do wyników Audio Front End. Przeprowadzono początkowe wyszukiwanie siatki w celu ustalenia optymalnych parametrów sieci neuronowej, w tym liczby warstw ukrytych, funkcji aktywacji, rozmiaru partii i liczby epok (Dokument "IFE: NN-aided Instantaneous Pitch Estimation" autorstwa Marka Bloka, Jana Banasa i Mariusza Pietrolaja, strona 2).

3. Eksperymenty i wyniki

W trakcie badań przeprowadzono eksperymenty z różnymi wariantami dystrybucji klas - liniowymi i logarytmicznymi. Wyniki pokazały, że dystrybucja logarytmiczna daje lepsze wyniki, zwłaszcza przy mniejszej liczbie klas (Dokument "IFE: NN-aided Instantaneous Pitch Estimation" autorstwa Marka Bloka, Jana Banasa i Mariusza Pietrolaja, strona 2).

Wyniki testów IFE na syntetycznych danych mowy, jak również na rzeczywistych danych mowy, pokazały, że metoda ta jest porównywalna z innymi zaawansowanymi estymatorami tonu, takimi jak YIN i CREPE. Co więcej, IFE wykazało nawet do 6% poprawy dokładności w porównaniu do tych metod, zachowując jednocześnie przewagę wyższej rozdzielczości czasowej (Dokument "IFE: NN-aided Instantaneous Pitch Est

W kontekście baz danych wykorzystywanych w metodzie IFE, warto zauważyć, że nie użyto żadnych rzeczywistych danych ani publicznych zestawów danych do celów szkolenia. Zamiast tego, dane do szkolenia zostały wygenerowane syntetycznie. Dane te składały się z segmentów syntetycznych samogłosek o długości 500 ms, przerywanych okresami ciszy o długości od 0 do 200 ms. Ton podstawowy (F_0) wahał się między 50 Hz a 400 Hz, a badano dwa rodzaje dystrybucji - logarytmiczną i liniową (Dokument "IFE: NN-aided Instantaneous Pitch Estimation" autorstwa Marka Bloka, Jana Banasa i Mariusza Pietrolaja, strona 1).

Dodatkowo, dane były testowane na syntetycznych danych mowy, jak również na rzeczywistych danych mowy. W trakcie testów rozważano liniową i logarytmiczną dystrybucję klas wyjściowych, a także liczbę klas wyjściowych. Syntetyczny zestaw danych walidacyjnych został zmieszany z białym szumem gaussowskim na zmiennych poziomach, a wyniki dla różnych stosunków sygnał/szum (SNR) zostały uzyskane (Dokument "IFE: NN-aided Instantaneous Pitch Estimation" autorstwa Marka Bloka, Jana Banasa i Mariusza Pietrolaja, strona 1).

W kontekście przyszłych kierunków badań, sugeruje się, że syntetyczny zestaw danych powinien być rozszerzony o dźwięki podobne do spółgłoskowych. Możliwe jest również dostrojenie Audio Front End (AFE) do lepszego dopasowania do docelowych sygnałów mowy, na przykład poprzez zmianę liczby filtrów, typu filtrów i jądra. Możliwe jest również przeprowadzenie dalszego wyszukiwania siatki w celu optymalizacji architektury sieci neuronowej (Dokument "IFE: NN-aided Instantaneous Pitch Estimation" autorstwa Marka Bloka, Jana Banasa i Mariusza Pietrolaja, strona 0).

Zakres projektu

Niniejszy projekt ma na celu gromadzenie i adnotację bazy sygnałów mowy w kontekście referencyjnych przebiegów tonu krztaniowego (F_0). Priorytetem jest walidacja oraz trening specjalizowanego neuronowego estymatora częstotliwości tonu krztaniowego (Instantaneous Frequency Estimation Neural Network, IFE NN).

Cel Projektu

Cel główny to konstrukcja i walidacja bazy sygnałów mowy, która będzie zawierała odpowiadające im referencyjne wartości tonu krztaniowego (F_0). Baza ta posłuży do treningu i walidacji modelu neuronowego estymującego F_0 (IFE NN) oraz do analizy wpływu różnych modyfikacji architektury i parametrów modelu na jakość estymatów F_0 .

Opis Zebranych Nagrań

Zebrane nagrania mowy charakteryzują się zróżnicowanymi parametrami, takimi jak długość czasowa (w sekundach) oraz częstotliwość próbkowania (w Hz). Planujemy adnotację tych nagrań przy użyciu etykiet zawierających referencyjne wartości F_0 , ekstrahowane przy pomocy algorytmu Praat

Uzasadnienie Metody

Wybrano metodę Praat dla anotacji ze względu na jej potwierdzoną skuteczność i popularność w dziedzinie analizy parametrów mowy. Jest ona w stanie dostarczyć precyzyjnych etykiet F_0 , co jest kluczowe dla efektywnego treningu modelu IFE NN.

Postępy i Eksperymenty

W fazie wstępnej projektu dokonaliśmy analizy literaturowej i eksploracyjnej odnośnie metody IFE NN. Planujemy w najbliższym czasie przystąpić do fazy eksperymentalnej, która obejmie trening i walidację modelu IFE NN w kontekście zebranych nagrań.

Plan Prac i Eksperymentów

W planach mamy porównanie efektywności estymacji F0 przy użyciu metody Praat oraz IFE NN na wybranych próbkach. Proces ten zostanie zautomatyzowany dzięki opracowaniu dedykowanego oprogramowania, potencjalnie bazującego na architekturach sieci neuronowych. Proces walidacji obejmie różne metryki, w tym RMSE (Root Mean Square Error).

Plan wykonania testów IFE NN:

1. Przygotowanie Danych: Pierwszym krokiem jest przygotowanie danych do treningu i testowania. W tym przypadku zostaną przygotowane nagrania mowy, które są już opatrzone etykietami z referencyjnymi wartościami F0.
2. Preprocessing: Następnym krokiem jest przetworzenie tych danych, aby były one odpowiednie do treningu sieci. Zostaną wykorzystane techniki normalizacji, ekstrakcji cech itd.
3. Trening Modelu: przygotowanie odpowiedniej architektury sieci neuronowej i zestawu danych do treningu.
4. Walidacja i Testy: Po wytrenowaniu modelu zostaną wykonane serie testów, aby ocenić jego skuteczność. W tym celu użyte zostaną różne metryki, takie jak RMSE (Root Mean Square Error) dla oceny dokładności estymacji.
5. Analiza Wyników: Po przeprowadzeniu testów, wyniki zostaną przeanalizowane, aby zrozumieć, w jakim zakresie model jest efektywny i czy wymaga dalszych modyfikacji.
6. Ocena Generalna: Na koniec, zostaną zestawione wyniki z innymi technikami estymacji F0 (na przykład, metodą Praat) dla porównania jakości estymacji

Podsumowanie

Metoda IFE to nowoczesne podejście do estymacji tonu podstawowego (F0) w mowie, które wykorzystuje głębokie sieci neuronowe i syntetyczne dane do szkolenia. Metoda ta oferuje natychmiastową estymację tonu, co jest znaczącym ulepszeniem w porównaniu do tradycyjnych metod. Testy przeprowadzone na syntetycznych i rzeczywistych danych mowy pokazały, że IFE jest porównywalne z innymi zaawansowanymi estymatorami tonu, takimi jak YIN i CREPE, a nawet wykazuje do 6% poprawę dokładności. W kontekście przyszłych badań, istnieje wiele możliwości dalszego rozwoju i optymalizacji metody IFE, w tym rozszerzenie syntetycznego zestawu danych, dostrojenie Audio Front End (AFE) i optymalizacja architektury sieci neuronowej.

Projekt znajduje się w fazie gromadzenia danych i planowania eksperymentów. Na podstawie zebranych nagrań i literatury planujemy trening i walidację neuronowego estymatora częstotliwości tonu krtaniowego.