WARSAW UNIVERSITY OF TECHNOLOGY

FACULTY OF MATHEMATICS AND INFORMATION SCIENCE

Ph.D. Thesis

Mohammad K. Nammous, *MBA, M.Sc.*

New Approaches in Speech Recognition
from Isolated Words to Practical Solutions

Supervisor
Professor Khalid Saeed, D.Sc., Ph.D., *M.Sc.*, B.Sc.

WARSAW, 2020

Abstract

New Approaches in Speech Recognition from Isolated Words to Practical Solutions

Voice-print is an economical and natural way to solve the problems of the unauthorized use of multilevel access control, communication devices and computer systems. Speaker recognition methods are achievable via ubiquitous telephone networks and microphones bundled with computers; the software might be the only cost of establishing such a system. Large enterprises are dominating the market with advanced applications available via mobiles for everyone's usage, but this should not discourage researchers from attempting further steps in this domain.

Within his research, the author worked on the different aspects of voice recognition, in particular using machine learning to realise various speaker identification tasks. Security application has been suggested which combines the recognition of a password of three uttered digits and their speaker. The research also touches on speaker verification, as well as language and gender identification cases. The main contribution represents the text-independent speaker identification results achieved for a large set of more than 4k speakers with about 343 hours of speech signals. Various proposed metrics provided up to a 76.9% average accuracy rate for individual voice segments, and 99.5% when considering the testing speech segments of each speaker as one unit. Doubling the amount of the training data yielded a perfect accuracy rate of 100%.

The simplicity of the proposed system was a goal achieved by decreasing the reliance on the preparation phase of noise reduction and data curation. Working on improving the speaker identification algorithms, the author intended mainly to train the models using as little information as possible. This includes exploring diverse percentages of the training and testing datasets, various lengths of the voice samples, as well as the impact of enhancing the feature vectors and using more complex NNs architectures; moreover, how to improve the speaker identification using predicted information about speech (ex. the spoken language).

Keywords: Speaker Recognition, Text-Independent Speaker Identification, Little Training Data, MFCC, Deep Learning, Bidirectional LSTM.

Streszczenie

Nowe podejścia w rozpoznawaniu mowy od pojedynczych słów do praktycznych rozwiązań
'Voice-print' to ekonomiczny i naturalny sposób do rozwiązywania problemów z nieuprawnionym użyciem wielopoziomowej kontroli dostępu urządzeń komunikacyjnych oraz systemów komputerowych. Metody rozpoznawania są osiągalne przez wszechobecnie używany aparat telefoniczny podłączony do sieci i mikrofon komputerowy. Oprogramowanie może być jedynym wydatkiem takiego systemu. Duże przedsiębiorstwa dominują na rynku z dobrze rozwiniętymi aplikacjami dostępnymi za pośrednictwem telefonów komórkowych do dyspozycji każdego człowieka. Dlatego też nie powinno to zniechęcać naukowców do dalszych kroków w tej dziedzinie. W swoim badaniu, autor pracował nad różnymi aspektami rozpoznawania głosu w szczególności maszynowemu nauczaniu realizowania różnych identyfikacji zadań głosowych. Bezpieczeństwem aplikacji są sugestie kombinacji rozpoznających hasło z trzech wypowiadanych cyfr jego mówcy. Podkreślono nieznaczne informacje na temat weryfikacji mówcy, a także przypadków identyfikacji języka i płci. Główny wkład przedstawia identyfikację mówcy niezależnego od tekstu oraz osiągnięte wyniki dla dużego zestawu ponad 4 tysięcy mówców z około 343 godzinami sygnałów mowy. Różne proponowane wskaźniki zapewniły średni wskaźnik dokładności do 76.9% dla poszczególnych segmentów głosu i 99.5% przy rozważaniu testowania segmentów mowy każdego mówcy jako jednej jednostki.Podwojenie ilości testowanych danych dały doskonały wskaźnik dokładności 100%. Prostota proponowanego systemu była celem osiągnięcia zmniejszenia zależności fazy przygotowawczej redukcji szumów i selekcji danych. Pracując nad ulepszeniem algorytmów identyfikacji mówców, autor zamierzał głównie wykształcić modele przy użyciu jak najmniejszej ilości informacji. Praca zawiera uwzględnienie badania zróżnicowanych procentów w tym wykształconych i testowanych zestawów danych, różnych długości próbek głosu, a także wpływ ulepszenia wektorów cech i użyciu bardziej złożonych architektur sieci neuronowe. Ponadto, jak poprawić identyfikację mówcy wykorzystując w tym przewidywanie informacji o mowie (np. język mówiony).

Słowa Kluczowe: rozpoznawanie mówcy, identyfikacja mówcy niezależnego od tekstu, niski poziom treningu w stosunku do testowania, dwukierunkowy LSTM, MFCC, głębokie uczenie.