1859
Politecnico
di Torino
ScuDo
Scuola di Dottorato ~ Doctoral School
WHAT YOU ARE, TAKES YOU FAR

Doctoral Dissertation
Doctoral Program in Computer and Control Engineering (36thcycle)

Speaker Verification and Language Recognition

By

Salvatore Sarni
******

Supervisor(s):
Prof. S. Cumani

Politecnico di Torino
2024

Declaration

I hereby declare that, the contents and organization of this dissertation constitute my own original work and does not compromise in any way the rights of third parties, including those relating to the security of personal data.

Salvatore Sarni
2024

* This dissertation is presented in partial fulfillment of the requirements for **Ph.D. degree** in the Graduate School of Politecnico di Torino (ScuDo).

Abstract

Automatic systems for identity recognition are ubiquitous. A person's identity can be found and matched exploiting some of its biometrical traits. The trivial operation of unlocking a smartphone hides some form of verification, usually face or fingerprint verification. Personal assistants answer our questions only after verifying our identity through our voices. Speaker verification is the main focus of this work. In speaker verification, given an identity and an audio utterance, the system should be able to assess if the audio was spoken by such identity.

The starting point of the verification pipeline is the audio. Due to their nature, voice recordings come with different durations and from various sources. A fixed dimensional utterance representation that contains speaker-discriminant information is then required. The advances in the field of Deep Learning made Deep Neural Networks (DNN) the state-of-the-art technique to process utterances and extract such representations, namely the embeddings. In this work, we explore the latest and most common architecture employed for the speaker verification task.

The nature of audio and the need to model its long-range temporal dependencies resulted in a variety of solutions, from Time Delay Neural Networks (TDNN) to Residual Networks (ResNet) and finally, the latest experiments with Transformers models. Architectures are typically trained on a set of background speakers using a multi-class classification paradigm. The network processes the acoustic features and aggregates them using a pooling layer, obtaining a fixed-length embedding from which speaker posterior probabilities are then computed, typically with a softmax layer.

The need to identify or verify a language arises for a plethora of applications. Language-dependent systems may be required to automatically adapt their model to the language being spoken. Given the small number of languages compared to that of potential speakers, the number of classes is reduced and while training

data is generally abundant, low-resource languages pose a significant challenge. Nevertheless, speaker verification methods can be successfully adapted. In this work, we focus on extending state-of-the-art DNN-based embedding models to the language recognition task.

To assess whether two embeddings convey the same information, being the speaker or the language, a scoring system is needed. Various backend classifiers can be used for this purpose, ranging from distance metrics to probabilistic models. However, no matter how the score is obtained, it needs to be interpreted and each application may have a specific threshold for accepting or rejecting the same speaker (or language) hypothesis. An optimal system should exhibit consistent performance across different applications, regardless of the chosen threshold.

Different techniques can be employed to address the calibration problem. In this study, we introduce a new generative model that can effectively use additional information from the audio, such as utterance duration. Moreover, logistic regression and score normalization are two distinct state-of-the-art approaches used to improve calibration. In this study, we explore a combination of both methods as well as a possible alternative to the score normalization approach.