

## **Abstract**

The goal of speech processing systems is to handle natural, free-flowing speech across diverse conditions such as homes, offices, restaurants, and noisy cocktail parties. Each of these present unique challenges for system design. Building a universal and robust system capable of managing such a complex goal is achieved by combining different speech processing approaches, where each contributes with complementary types of information. A key trend towards the development of generic models is to integrate multiple tasks and leverage methods from other fields to extract sufficiently diverse information. Deep neural networks have significantly advanced speech technology systems, but research still continues to focus on improving performance, interpretability, and adapting cross-domain solutions.

This thesis investigates speaker representations for speaker recognition and diarization tasks, proposing optimization strategies to improve systems while retaining speaker-specific information. It further introduces explainable, discriminative representations for diarization and develops generic methods for joint diarization and separation, effective in both low- and high-overlap speech. The research contributions are presented through a series of five publications focused on speaker recognition and diarization tasks.

The first two of these publications focus on the speaker recognition task. The speaker recognition system involves several steps, the core of which is the extraction of speaker representations. In this thesis, the extraction method is based on a deep neural network approach. Two angular-based speaker objective functions are introduced, which adapt their hyperparameter values based on network performance and convergence in the current training step. Next, multiple proposals are presented for the architecture itself that increase processed temporal resolution, preserve and enhance processed information.

The next three publications concern the speaker diarization task. The development is based on an end-to-end approach to diarization, where the model directly estimates speaker activity from the input. An important procedure in this context is the estimation of the so-called attractors, which are representations of the speakers present in a given recording. In this thesis, the method of Non-Autoregressive Attractor (NAA) estimator is

---

introduced. The approach estimates speaker representations for diarization by leveraging the properties of the embeddings present in the structure of the diarization model, providing a more explainable process of the attractor generation for the speaker diarization task, in contrast to the more obscure standard autoregressive method. The proposed NAA approach has been developed further and applied for joint speaker diarization and separation, at the same time aiming to bridge the gap between diarization and speech separation tasks.

## Streszczenie

Celem systemów przetwarzania mowy jest możliwość przetwarzania naturalnej wypowiedzi w różnorodnych warunkach takich jak zacisze domowe, biura, restauracje, czy głośne imprezy. Każdy z zaprezentowanych scenariuszy stawia odmienne wyzwania podczas projektowania systemu. Opracowanie uniwersalnego i niezawodnego systemu zdolnego do realizacji tego złożonego zadania może zostać osiągnięte poprzez połączenie różnych systemów przetwarzania mowy, z których każdy wnosi komplementarne rodzaje informacji. Istotnym kierunkiem badań jest dążenie do systemów generycznych, poprzez łączenie wielu zadań i wykorzystywanie metod z innych dziedzin w celu uzyskania różnorodnych informacji. Głębokie sieci neuronowe (ang. deep neural networks) znacząco rozwinęły możliwości technologii mowy, niemniej jednak nadal trwają badania koncentrujące się na poprawie wydajności, interpretowalności oraz integracji rozwiązań z różnych dziedzin.

Niniejsza praca bada reprezentacje mówców dla zadań rozpoznawania mówców i diaryzacji, proponując strategie optymalizacji służące poprawie systemów przy jednoczesnym zachowaniu informacji charakterystycznych dla mówcy. W dalszej części wprowadza wyjaśnialne, dyskryminatywne reprezentacje dla diaryzacji i rozwija generyczne metody dla zadania jednoczesnej diaryzacji i separacji, skutecznych zarówno dla nagrań zawierających w niewielkim, jak i znaczącym stopniu mowę wielu osób wypowiadających się jednocześnie (ang. overlap speech). Wkład naukowy pracy jest zaprezentowany w postaci serii pięciu publikacji skupiających się na zadaniach rozpoznawania i diaryzacji mówców.

Pierwsze dwie publikacje dotyczą zadania rozpoznawania mówców. System rozpoznawania mówców zawiera kilka etapów, gdzie kluczowym jest ekstrakcja reprezentacji mówcy. W tej pracy metoda ekstrakcji opiera się na podejściu wykorzystującym głębokie sieci neuronowe. Zaproponowano dwie funkcje kosztu oparte na mierze kątowej, których wartości hiperparametrów adaptują się w zależności od poprawności odpowiedzi sieci oraz jej zbieżności w bieżącym kroku treningu. Następnie przedstawiono szereg propozycji dla samej architektury modelu, które pozwalają na zwiększenie przetwarzanej rozdzielczości czasowej oraz zachowują i wzmacniają przetwarzaną informację.

Kolejne trzy publikacje dotyczą zadania diaryzacji mówców. Rozwój metody jest oparty na podejściu typu end-to-end dla diaryzacji, gdzie model bezpośrednio estymuje aktywność mówców na podstawie informacji wejściowej. Ważną procedurą w tym zakresie jest estymacja tzw. atraktorów, czyli reprezentacji mówców występujących w danym nagraniu. W niniejszej pracy zaproponowano nieautoregresywną estymację atraktorów (ang. Non-Autoregressive Attractor (NAA) estimation). Podejście to wyznacza reprezentacje mówców

---

poprzez wykorzystanie właściwości wektorów osadzeń (ang. embeddings) obecnych w architekturze modelu diaryzacji, zapewniając bardziej wyjaśnialny (ang. explainable) proces estymacji atraktorów dla zadania diaryzacji mówców, w przeciwieństwie do standardowego podejścia autoregresywnego. Opracowaną metodę rozszerzono i wykorzystano w zadaniu jednoczesnej diaryzacji i separacji mówców, aby zniwelować istniejącą lukę pomiędzy tymi procesami.