AKADEMIA GÓRNICZO-HUTNICZA IM. STANISŁAWA STASZICA W KRAKOWIE

DZIEDZINA NAUK INŻYNIERYJNO-TECHNICZNYCH

DYSCYPLINA AUTOMATYKA, ELEKTRONIKA, ELEKTROTECHNIKA I TECHNOLOGIE KOSMICZNE

AUTOREFERAT ROZPRAWY DOKTORSKIEJ

Dyskryminatywne reprezentacje mówców dla zadań rozpoznawania mówców i diaryzacji

Autor: Magdalena Marta Rybicka

Promotorzy rozprawy:
dr hab. inż. Konrad Kowalczyk
Profesor uczelni Akademii Górniczo-Hutniczej im. Stanisława Staszica
dr Jesús Villalba
Assistant Research Professor na Johns Hopkins University

Praca wykonana: Akademia Górniczo-Hutnicza im. Stanisława Staszica w Krakowie, Wydział Informatyki, Elektroniki i Telekomunikacji

Kraków, 2026

AGH University of Krakow

FIELD OF SCIENCE ENGINEERING AND TECHNOLOGY

SCIENTIFIC DISCIPLINE AUTOMATION, ELECTRONICS, ELECTRICAL ENGINEERING AND SPACE TECHNOLOGIES

SUMMARY OF DOCTORAL THESIS

Towards discriminative speaker representations for speaker recognition and diarization

Author: Magdalena Marta Rybicka

Supervisors:
dr hab. inż. Konrad Kowalczyk
Associate Professor at AGH University of Krakow
dr Jesús Villalba
Assistant Research Professor at Johns Hopkins University

Completed in: AGH University of Krakow,
Faculty of Computer Science, Electronics and Telecommunications

Krakow, 2026

Wprowadzenie

W ostatnich latach możemy zaobserwować coraz więcej postępów technologicznych, które mają na celu ułatwienie i usprawnienie naszego życia, takich jak wirtualna i rozszerzona rzeczywistość (ang. virtual and augmented reality), Internet Rzeczy (ang. Internet of Things), „inteligentne" domy (ang. 'smart' homes) czy asystenci/agenty konwersacyjne wykorzystujący sztuczną inteligencję, np. ChatGPT, Google Gemini, Grok i wiele innych. Rozwiązania te są zaprojektowane w taki sposób, aby były łatwe i intuicyjne w użyciu. Jednym ze sposobów realizacji tego celu jest integracja technologii mowy, która może służyć jako naturalny środek do nawigacji i sterowania, a także stanowić istotne źródło informacji o użytkowniku lub prowadzonej konwersacji. W istocie, technologia mowy już teraz znajduje zastosowanie w wielu codziennych zadaniach, takich jak głosowe wybieranie numeru, dyktowanie i wypowiadanie komend dla inteligentnych asystentów, tłumaczeniu wypowiedzi na wybrany język w czasie rzeczywistym (ang. on-the-fly), częściowej automatyzacji obsługi klienta, czy też poprzez zapytania do asystentów głosowych (np. Siri lub Alexa). Zasadne jest również odniesienie się do niedawnej sytuacji pandemicznej, która wymusiła przejście znacznej części użytkowników na tryb pracy zdalnej i wykorzystywanie narzędzi audiowizualnych do prowadzenia spotkań. Narzędzia te implementują algorytmy detekcji aktywności głosowej oraz poprawy jakości sygnału mowy, co umożliwia uzyskanie lepszej jakości komunikacji. Aktualne kierunki badawcze koncentrują się na obszarze tzw. inteligencji konwersacyjnej (ang. conversational intelligence), w ramach której interakcje pojedynczego użytkownika z systemem przetwarzającym informacje w sposób pasywny ewoluują w stronę systemów aktywnie zaangażowanych zdolnych do realizacji złożonych zadań, takich jak automatyczne generowanie dokumentacji ze spotkań, weryfikacja faktów, ekstrakcja informacji o uczestnikach spotkania czy też wspieranie procesów tzw. uczenia się we współpracy (ang. collaborative learning) np. w grupach rówieśniczych. Tego typu systemy składają się z komponentów wykorzystujące różnorodne modalności, między innymi systemy dialogowe (ang. dialogue systems) odpowiedzialne za podążanie za kontekstem wypowiedzi, przetwarzanie obrazu wideo (ang. video processing) i ekstrakcja informacji, mechanizmy rozumowania oparte na wiedzy zdroworozsądkowej (ang. common-sense reasoning), a także moduły umożliwiające analizę konwersacji wieloosobowych.

Celem systemów przetwarzających mowę jest zdolność do analizy swobodnych wypowiedzi w różnorodnych warunkach, np. zacisza domowego, biura, restauracji, czy też przyjęcia (lub sytuacjach typu 'cocktail party'), co jednocześnie stwarza szerokie spektrum

wyzwań i uwarunkowań akustycznych. Aby opracować uniwersalny system przetwarzania mowy, który jest odporny na zróżnicowane warunki oraz zdolny do rozwiązywania złożonych zadań, konieczne jest połączenie różnych systemów zaprojektowanych do pojedynczych problemów. Na przykład, systemy do transkrypcji (ang. transcription systems) mogą składać się z połączenia systemów: poprawy jakości mowy (ang. speech enhancement) lub separacji mowy (ang. speech separation), automatycznego rozpoznawania mowy (ang. automatic speech recognition), rozpoznawania mówców (ang. speaker recognition) oraz diaryzacji mówców (ang. speaker diarization). Celem pierwszego z tych systemów, tj. poprawy jakości mowy, jest redukcja lub całkowite usunięcie niepożądanych inferencji, takich jak szum czy pogłos, które utrudniają zrozumienie i pogarszają jakość sygnału mowy. Zadaniem separacji mowy jest wydzielenie wypowiedzi jednej lub wielu osób do odrębnych strumieni audio. Automatyczne rozpoznawanie mowy służy do rozpoznawania wypowiedzianych słów (czyli ich transkrypcji). Rozpoznawanie mówców służy rozpoznaniu tożsamości danego mówcy. W ramach tego zadania wyróżnia się dwa podtypy: identyfikację oraz weryfikację. Pierwszy z nich polega na ustaleniu tożsamości danej osoby, z kolei drugi ma na celu potwierdzenie lub odrzucenie, czy dana osoba jest tym, za kogo się podaje. Diaryzacja mówców jest zadaniem polegającym na segmentacji nagrania wypowiedzi wieloosobowych na fragmenty, ze wskazaniem kiedy mówi ta sama osoba. Diaryzacja jest często wykorzystywana jako ogniwo łączące różne zadania, np. transkrypcja wywiadu otrzymana z systemu rozpoznawania mowy może dostarczać wypowiedzi mówców w postaci jednego ciągłego tekstu, bez rozróżnienia wypowiedzi poszczególnych mówców. W połączeniu z systemem diaryzacji tekst może zostać przyporządkowany do poszczególnych rozmówców – np. do osoby prowadzącej wywiad i respondenta – co pozwala na uzyskanie pełniejszej informacji i lepszego zrozumienia nagrania. W związku z tym, że niniejsza praca koncentruje się na reprezentacjach mówców w kontekście wybranych zadań systemów przetwarzania mowy, istotne jest podkreślenie znaczenia zadania rozpoznawania mówców, którego rozwiązania są często implementowane jako nowe metody dla innych obszarów, takich jak diaryzacja mówców, rozpoznawanie języka, rozpoznawanie emocji, czy też detekcja chorób na podstawie mowy.

Problemy oraz cele badawcze

Rozwiązania oparte na głębokich sieciach neuronowych (ang. deep neural networks) przyniosły znaczący przełom w możliwościach współczesnych systemów technologicz-

nych. Niemniej jednak społeczność naukowa nadal podejmuje intensywne wysiłki w celu poprawy wydajności, zrozumienia oraz interpretacji decyzji podejmowanych przez głębokie sieci neuronowe. Jedną z popularnych praktyk jest przenoszenie rozwiązań z jednej dziedziny do innej (np. stosowanie metod przetwarzania obrazów w systemach przetwarzania mowy), co przyczynia się do postępu w nowym obszarze. Jednakże zdarza się, że proponowane metody nie są dostosowane do specyfiki domeny mowy. Kolejnym istotnym trendem w tej dziedzinie jest dążenie do budowy uniwersalnych systemów, które integrują wiele zadań, umożliwiając opracowanie generycznych modeli zdolnych do ekstrakcji różnorodnych informacji.

Niniejsza praca doktorska bada reprezentacje mówców w kontekście zadań rozpoznawania oraz diaryzacji mówców. Prezentuje ona metody zaprojektowane do poprawy systemów rozpoznawania mówców, ze szczególnym uwzględnieniem optymalnego procesu uczenia oraz zachowania cech i informacji charakterystycznych dla poszczególnych mówców. Praca dodatkowo wprowadza wyjaśnialne (ang. explainable) i dyskryminatywne reprezentacje mówców w kontekście zadania diaryzacji. Na zakończenie, w oparciu o poprzednie wyniki i wnioski, zaprezentowane są badania, których celem jest zapełnienie luki w domenie jednoczesnej diaryzacji oraz separacji mówców (ang. joint speaker diarization and separation) poprzez zaproponowanie generycznych metod, które działają skutecznie w warunkach naturalnej konwersacji, kiedy wypowiedzi nakładają się w sposób nieznaczny (ang. low-overlap speech) oraz w warunkach, kiedy wiele osób mówi niemalże jednocześnie (ang. high-overlap speech).

Opisane metody są ujęte w postaci następujących celów badawczych i hipotez:

1. Odpowiednie wartości hiperparametrów funkcji celu opartej na mierze kątowej (ang. angular-based loss function) poprawiają wydajność oraz zbieżność (ang. convergance) procesu uczenia systemu rozpoznawania mówców;

2. Zastosowanie cech wieloskalowych (ang. multi-scale features), zwiększenie rozdzielczości czasowej oraz uwzględnienie zależności częstotliwościowych przyczyniają się do poprawy skuteczności modeli rozpoznawania mówców;

3. Informacje zakodowane w wektorach osadzeń kodera na poziomie ramek (ang. frame-level encoder embeddings) modelu diaryzacji niosą względne informacje o mówcach i umożliwiają ich rozróżnianie w obrębie jednego nagrania;

4. Informacje o mówcach, wyekstrahowane przy użyciu metod zaproponowanych dla diaryzacji, mogą być wykorzystane w modelu jednoczesnej diaryzacji i separacji mówców, co pozwala na poprawę działania dla obu zadań.

Osiągnięcia badawcze pracy

Osiągnięcia i propozycje ukierunkowane na rozwiązanie przedstawionych problemów zostały zaprezentowane w formie serii pięciu publikacji obejmujących zagadnienia związane z rozpoznawaniem oraz diaryzacją mówców:

I **M. Rybicka** and K. Kowalczyk, "*On Parameter Adaptation in Softmax-Based Cross-Entropy Loss for Improved Convergence Speed and Accuracy in DNN-Based Speaker Recognition*", Interspeech, Shanghai, China, 2020.

II **M. Rybicka**, J. Villalba, P. Żelasko, N. Dehak, K. Kowalczyk, "*Spine2Net: SpineNet with Res2Net and Time-Squeeze-and-Excitation Blocks for Speaker Recognition*", Interspeech, Brno, Czech Republic, 2021.

III **M. Rybicka**, J. Villalba, N. Dehak and K. Kowalczyk, "*End-to-End Neural Speaker Diarization with an Iterative Refinement of Non-Autoregressive Attention-based Attractors*", Interspeech, Incheon, South Korea, 2022.

IV **M. Rybicka**, J. Villalba, T. Thebaud, N. Dehak and K. Kowalczyk, "*End-to-End Neural Speaker Diarization with Non-Autoregressive Attractors*", IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2024.

V **M. Rybicka**, K. Kowalczyk, T. Thebaud, N. Dehak, J. Villalba, "*Joint Diarization and Separation Using SepFormer with Non-Autoregressive Attractors*", IEEE Signal Processing Letters, 2025.

Zaproponowane rozwiązania potwierdzają postawione hipotezy i wnoszą istotny wkład w rozwój metod rozpoznawania oraz diaryzacji mówców poprzez optymalizację podejść opartych na głębokich sieciach neuronowych. W dalszej części niniejszego autoreferatu publikacje te będą przywoływane za pomocą przypisanych im powyżej cyfr rzymskich.

Pierwsze dwie publikacje (I oraz II) skupiają się na zadaniu **rozpoznawania mówców**. System rozpoznawania mówców zawiera kilka etapów, gdzie kluczowym jest ekstrakcja reprezentacji mówcy. W przypadku zadania weryfikacji, reprezentacje te są w kolejnych krokach przetwarzane i porównywane między sobą, zwracając wynik ich podobieństwa,

który pozwala stwierdzić, czy nagrania pochodzą od tego samego mówcy, czy też nie. Obecnie ekstraktory reprezentacji mówców są oparte na głębokich sieciach neuronowych. W literaturze dużo uwagi jest poświęcone modyfikacjom i poprawie architektury sieci neuronowej, jej komponentów, uwzględniając również funkcję straty zastosowaną do treningu modelu. W niniejszej pracy zaproponowano dwie funkcje celu oparte na mierze kątowej. Rodzina tych funkcji jest szeroko stosowana i wykazuje poprawę wydajności systemów rozpoznawania mówców poprzez wprowadzenie modyfikacji, które prowadzą do odpowiedniej separacji reprezentacji pochodzących od różnych mówców (ang. between-class distance) przy jednoczesnym dążeniu do jak najbardziej zbliżonych reprezentacji dla wypowiedzi pochodzących od tego samego mówcy (ang. within-class distance). Istotnym wkładem zaproponowanych funkcji kosztu jest automatyczna adaptacja hiperparameterów, których wartości dostosowują się do poprawności odpowiedzi sieci i jej zbieżności w bieżącym kroku treningu. Przedstawiona propozycja wpłynęła na poprawę skuteczności oraz przyspieszenie zbieżności treningu sieci dla zadania rozpoznawania mówców. Następnie zaproponowano wiele rozwiązań dla samej architektury sieci. W pierwszej publikacji zmodyfikowano powszechnie stosowaną strukturę, tzw. model rezydualny (ang. residual model), czyli ResNet [3], poprzez zwiększenie rozdzielczości czasowej oraz dostosowanie architektury do zadania rozpoznawania mówców, co skutkuje poprawą skuteczności. Druga z publikacji także wprowadza dalsze ulepszenia dla modelu rezydualnego. Ta powszechnie stosowana struktura może zostać scharakteryzowana jako model ze zmniejszającą się skalą (ang. scale-decreased design). Oznacza to, że informacja wraz z przetwarzaniem przez sieć jest jednostajnie podpróbkowywana. Tego typu przetwarzanie może się przyczynić do utraty części informacji. Aby temu zapobiec dla zadania mówców zaadoptowany został model ze skalą permutowaną (ang. scale-permuted design), tzw. SpineNet [1]. Struktura ze skalą permutowaną oznacza, że rozmiar przetwarzanej informacji, tzw. mapy cech (ang. feature map) może się dowolnie zwiększać lub zmniejszać w trakcie przetwarzania. Ponadto, zaimplementowana stru-ktura pozwala na łączenie map cech wieloskalowych, na podstawie których generowana jest reprezentacja mówcy, podczas gdy w standardowym podejściu wykorzystywana jest ostatnia mapa cech, która jest najbardziej podpróbkowana. Przeprowadzono również badania, które pozwalają na integrację dodatkowych modułów w celu dalszej poprawy wydajności informacji uzyskanej z nagrania, a mianowicie Res2Net [2] oraz Time-Squeeze-and-Excitation (T-SE) [5, 6]. Oba moduły zostały pierwotnie zaproponowane dla zadania przetwarzania obrazów dla modeli o architekturze rezydualnej. Blok Res2Net modyfikuje bazowe bloki struktury rezydualnej w celu poszerzenia pola recepcyjnego (ang. receptive field)

oraz umożliwia uchwycenie cech wieloskalowych o drobniejszej rozdzielczości w ramach pojedynczego bloku rezydualnego. Z kolei celem bloku T-SE jest rekalibracja zależności pomiędzy mapami cech wzdłuż wymiaru kanałów oraz częstotliwości.

Kolejne trzy publikacje (III, IV oraz V) dotyczą zadania **diaryzacji mówców**. Zaproponowane metody są oparte na podejściu diaryzacyjnym typu end-to-end, co oznacza, że model bezpośrednio określa wynik diaryzacji (aktywności poszczególnych mówców), na podstawie otrzymanych danych wejściowych. Kluczową cechą tego podejścia jest transformacja sekwencji wejściowej sieci na sekwencję wektorów osadzeń (ang. embedding sequence), gdzie na każdy zestaw cech przypada jeden wektor. W literaturze wektory osadzeń na poziomie ramek (ang. frame-level embeddings) były wykorzystywane na wiele sposobów w celu ekstrakcji informacji diaryzacyjnej, informacji na temat mówców, czy też do określania liczby mówców w nagraniu. W tym zakresie ważną procedurą jest estymacja atraktorów (ang. attractors), które określają reprezentacje mówców występujących w danym nagraniu. Najbardziej popularne jest podejście autoregresywne (ang. autoregressive), znane jako Encoder-Decoder-based Attractors [4], gdzie osadzenia na poziomie ramek są wykorzystane do oszacowania atraktorów. W niniejszej pracy zaproponowana została metoda nieautoregresywnej estymacji atraktorów (ang. Non-Autoregressive Attractor estimation - NAA). Główna różnica między podejściem autoregresywnym a nieautoregresywnym polega na tym, że w podejściu autoregresywnym kolejne atraktory estymowane są sekwencyjnie na podstawie poprzednich wyników, natomiast w podejściu nieautoregresywnym wszystkie atraktory wyznaczane są równocześnie. Zaprezentowane podejście wyznacza reprezentacje mówców dla zadania diaryzacji poprzez wykorzystanie właściwości osadzeń na poziomie ramek występujących w strukturze modelu diaryzacji, zapewniając bardziej wyjaśnialny (ang. explainable) proces estymacji atraktorów, w przeciwieństwie do standardowego podejścia autoregresywnego. Ważne, aby wspomnieć, że w literaturze zostały zaproponowane inne nieautoregresywne rozwiązania, jednakże po raz pierwszy w kontekście diaryzacji zostało ono zaproponowane przez autorkę tej pracy. Początkowo, metoda NAA była zaprezentowana jedynie na nagraniach zawierających dwóch mówców oraz z założeniem, że liczba mówców w nagraniu jest znana. W toku dalszych badań, procedura estymacji atraktorów została rozwinięta i rozszerzona poprzez zaproponowanie wariacji NAA, które pozwoliły na jej zastosowanie do warunków bardziej generycznych, obejmujących zmienną, większą niż dwa i nieznaną liczbę mówców. Ewaluacje systemów uwzględniały wiele baz danych, zarówno nagrań symulowanych, jak i rzeczywistych. Kolejnym krokiem w rozwoju metody NAA było jej zintegrowanie z modelem dla zadania **jednoczesnej diaryzacji i separacji mówców**. Badania

miały na celu zniwelowanie luki pomiędzy zadaniami diaryzacji i separacji, które są w zasadzie bardzo podobne. Celem diaryzacji jest wskazanie aktywności każdego z mówców, co może być zaprezentowane poprzez wskazanie prawdopodobieństwa występowania danego mówcy w danej ramce czasowej. Z kolei separacja mówców, która polega na rozdzieleniu sygnałów audio odpowiadających poszczególnym mówcom, typowo jest wykonywana poprzez estymację masek, które mogą być interpretowane jako aktywności mówców w domenie czasowo-częstotliwościowej. W związku z tym zaproponowano rozwiązanie w formie pojedynczej struktury trenowanej jednocześnie dla obu zadań, które pozwala na ekstrakcję zarówno wyniku diaryzacji, jak i separacji.

Podsumowanie

Wkład badawczy i wyniki przedstawione w rozprawie doktorskiej potwierdzają postawione w niej hipotezy i cele. Ponadto, wnosi ona istotny wkład do dziedzin rozpoznawania mówców i diaryzacji poprzez optymalizację podejść opartych o głębokie sieci neuronowe. W szczególności badania skupiają się na różnych aspektach reprezentacji mówców w kontekście wspomnianych zadań i systemów.

Publikacje I oraz II wprowadzają szereg optymalizacji struktury sieci neuronowej, których celem jest uzyskanie bardziej dyskryminatywnych reprezentacji mówców dla zadania rozpoznawania mówców. Prace te proponują funkcje kosztu oparte na mierze kątowej, które adaptują swoje parametry w zależności od aktualnego etapu zbieżności oraz dokładności, aby zapewnić optymalny proces uczenia. W rezultacie uzyskano poprawę zarówno skuteczności, jak i szybkości zbieżności treningu. W obu publikacjach zaproponowano modyfikacje architektury mające na celu zwiększenie zależności czasowych i częstotliwościowych, co prowadzi do lepszej ekstrakcji informacji o mówcach.

Publikacje III oraz IV proponują uniwersalną nieautoregresywną metodę estymacji atraktorów, która wykazuje lepszą skuteczność w porównaniu z innymi metodami oraz bezpośrednio wykorzystuje informacje dostarczane przez koder sieci diaryzacyjnej. W szczególności metoda ta wykorzystuje informacje o mówcy zakodowane w osadzeniach na poziomie ramek, co umożliwia wyjaśnialną ekstrakcję oraz poprawę reprezentacji mówców. Metoda została również rozszerzona umożliwiając przetwarzanie nagrań, w których jest nieznana i zmienna liczba mówców.

Publikacja V kontynuuje wykorzystanie podejścia nieautoregresywnego w kontekście wspólnego zadania diaryzacji i separacji mówców, potwierdzając uniwersalność zaproponowanej metody. Ponadto, przedstawia ona system, który integrujące zadania separacji

i diaryzacji w jeden, wspólny model. Uzyskane wyniki wskazują na poprawę dokładności zaproponowanych metod w warunkach dużego udziału mówców wypowiadających się jednocześnie, typowych dla zadania separacji, jak również w warunkach małego stopnia nakładania się wielu wypowiedzi, charakterystycznych dla rozmów i zadania diaryzacji.

Istnieje wiele potencjalnych kierunków dalszego rozwoju przedstawionych badań. Po pierwsze, analizy informacji o mówcy zakodowanej w atraktorach, których celem byłaby weryfikacja, czy atraktory posiadają bezwzględną informację o mówcach oraz czy mogą być wykorzystywane w kontekście zadania rozpoznawania mówców. Kolejnym celem jest rozszerzenie właściwości modeli diaryzacji oraz jednoczesnej diaryzacji i separacji do przetwarzania długich nagrań. Obecnie zaproponowane modele osiągają dobre wyniki dla stosunkowo krótkich nagrań (do 60 sekund), gdzie niektóre zastosowania diaryzacji lub separacji mówców obejmują nagrania trwające od kilku minut do nawet kilku godzin. Taki kierunek badań mógłby również ewoluować w stronę diaryzacji i separacji wykonywanej w czasie rzeczywistym. Dalszy rozwój zaproponowanych metod mógłby także obejmować integrację zaproponowanych rozwiązań z systemami automatycznego rozpoznawania mowy, zarówno jako zadania służącemu ocenie modeli od strony ich praktycznego zastosowania, jak i poprzez bezpośrednie włączenie modelu ASR do wspólnego modelu, w pełni odpowiadając na pytanie „kto, co i kiedy powiedział". Jednym z problemów w rozwoju i treningu systemu jednoczesnej diaryzacji i separacji jest brak rzeczywistych danych odniesienia (ground truth) dla zadania separacji, które są niezbędne do uczenia modelu w sposób nadzorowany. Aby rozwiązać ten problem, połączenie technik nienadzorowanych dla separacji z technikami nadzorowanymi dla diaryzacji mogłoby umożliwić przetwarzanie i wykorzystanie nagrań z różnych zbiorów danych, zwiększając tym samym ilość dostępnych danych treningowych.

Bibliografia

[1] Du, X., Lin, T. Y., Jin, P., Ghiasi, G., Tan, M., Cui, Y., Le, Q. V., and Song, X. (2020). SpineNet: Learning Scale-Permuted Backbone for Recognition and Localization. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11589–11598.

[2] Gao, S.-H., Cheng, M.-M., Zhao, K., Zhang, X.-Y., Yang, M.-H., and Torr, P. (2021). Res2Net: A New Multi-Scale Backbone Architecture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(2):652–662.

[3] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778.

[4] Horiguchi, S., Fujita, Y., Watanabe, S., Xue, Y., and Nagamatsu, K. (2020). End-to-End Speaker Diarization for an Unknown Number of Speakers with Encoder-Decoder Based Attractors. In *Proc. Interspeech 2020*, pp. 269–273.

[5] Hu, J., Shen, L., and Sun, G. (2018). Squeeze-and-Excitation Networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7132–7141.

[6] Kataria, S., Nidadavolu, P. S., Villalba, J., Chen, N., Garcia-Perera, P., and Dehak, N. (2020). Feature Enhancement with Deep Feature Losses for Speaker Verification. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7584–7588.