PODPIS ZAUFANY
SZYMON
DRGAS
15.07.2024 13:21:03 [GMT+2]
Dokument podpisany elektronicznie
podpisem zaufanym

Załącznik nr 3 – Autoreferat

Szymon Drgas

Wydział Automatyki, Robotyki i Elektrotechniki
Politechnika Poznańska

15 lipca 2024

Spis treści

1 Imię i nazwisko

Szymon Drgas
ORCID: 0000-0002-4603-8894

2 Posiadane dyplomy, stopnie naukowe lub artystyczne – z podaniem podmiotu nadającego stopień, roku ich uzyskania oraz tytułu rozprawy doktorskiej.

1. Doktor nauk technicznych, dyscyplina: automatyka i robotyka, specjalność: cyfrowe przetwarzanie sygnałów, 2013
   - tytuł rozprawy doktorskiej: "Automatic speaker recognition based on multilevel analysis of speech signal"
   - promotor: prof. dr hab. inż. Adam Dąbrowski
2. Magister informatyki stosowanej, Wydział Fizyki Uniwersytetu im. Adama Mickiewicza w Poznaniu, 2006
   - tytuł pracy magisterskiej: "Automatyczne rozpoznawanie mowy przy użyciu niejawnych modeli Markowa"
   - promotor: prof. dr hab. inż. Adam Dąbrowski
3. Magister fizyki, specjalizacja: akustyka, Wydział Fizyki Uniwersytetu im. Adama Mickiewicza w Poznaniu, 2005
   - tytuł pracy magisterskiej: "Dyskryminacja częstotliwości modulacji amplitudowej"
   - promotor: prof. dr hab. Aleksander Sęk

3 Informacja o dotychczasowym zatrudnieniu w jednostkach naukowych lub artystycznych.

1. Politechnika Poznańska, adiunkt, od 2013
2. Politechnika Poznańska, asystent 2006-2013

4 Omówienie osiągnięć, o których mowa w art. 219 ust. 1 pkt. 2 ustawy z dnia 20 lipca 2018 r. Prawo o szkolnictwie wyższym i nauce (Dz. U. z 2021 r. poz. 478 z późn. zm.).

Osiągnięcia naukowe mieszczą się w tematyce "interfejsy człowiek-maszyna" w dyscyplinie naukowej automatyka, elektronika, elektrotechnika i technologie kosmiczne. Wyniki tych osiągnięć dotyczą modelowania i przetwarzania sygnału mowy, w celu jego separacji, identyfikacji, poprawy jakości i zrozumiałości.

4.1 Cykl publikacji

Tytuł głównego osiągnięcia naukowego: **„Interpretowalne i niskokosztowe metody poprawy jakości mowy"**. Obejmuje ono następujące aspekty:

1. Opracowanie generatywnych metod uczenia maszynowego do wyznaczania wzorców charakteryzujących mówcę

2. Niskokosztowe, różniczkowalne metody uczenia maszynowego do poprawy jakości mowy.

Pozostałe osiągnięcia naukowe, które są objęte wnioskiem o nadanie stopnia doktora habilitowanego, a nie zostały włączone do cyklu publikacji omówionego w niniejszym punkcie, są omówione w punkcie 4.2.

W związku ze zaktualizowaniem wartości Journal Impact Factor (JIF) po sporządzeniu wykazu osiągnięć habilitanta przez Bibliotekę Politechniki Poznańskiej w odpowiednich pozycjach dodano aktualną wartość tego wskaźnika (JIF_{2023}).

Artykuły naukowe wchodzące w skład cyklu publikacji

[A1] **Szymon Drgas**. "Speech intelligibility prediction using generalized ESTOI with fine-tuned parameters". W: *Speech Communication* 159 (2024), s. 103068. (**MNiSzW: 140 pkt.; JIF=3,2; JIF_{2023}=2,4**).

[A2] **Szymon Drgas** (70%), Lars Bramsløw, Archontis Politis, Gaurav Naithani i Tuomas Virtanen. "Dynamic Processing Neural Network Architecture for Hearing Loss Compensation". W: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (2023). (**MNiSzW: 140 pkt.; JIF=5,4; JIF_{2023}=4,1**).

[A3] **Szymon Drgas**. "A Survey on Low-Latency DNN-Based Speech Enhancement". W: *Sensors* 23.3 (2023), s. 1380. (**MNiSzW: 100 pkt.; JIF=3,9; JIF_{2023}=3,4**).

[A4] **Szymon Drgas** (85%) i Tuomas Virtanen. "Joint speaker separation and recognition using non-negative matrix deconvolution with adaptive dictionary". W: *Computer Speech & Language* 70 (2021), s. 101223. (**MNiSzW: 100 pkt.; JIF=3,252**).

[A5] **Szymon Drgas** (70%), Magdalena Blaszak i Anna Przekoracka-Krawczyk. "The Combination of Neural Tracking and Alpha Power Lateralization for Auditory Attention Detection". W: *Journal of Speech, Language, and Hearing Research* 64.9 (2021), s. 3603–3616. (**MNiSzW: 100 pkt.; JIF=2,674**).

[A6] **Szymon Drgas** (70%), Tuomas Virtanen, Jörg Lücke i Antti Hurmalainen. "Binary non-negative matrix deconvolution for audio dictionary learning". W: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25.8 (2017), s. 1644–1656. (**MNiSzW: 25 pkt.; JIF=2,95**).

[A7] **Szymon Drgas** (80%) i Adam Dabrowski. "Speaker recognition based on multilevel speech signal analysis on Polish corpus". W: *Multimedia Tools and Applications* 74.12 (2015), s. 4195–4211. (**MNiSzW: 30 pkt.; JIF=1,331**).

[A8] Tomasz Grzywalski i **Szymon Drgas** (50%). "Speech enhancement using U-nets with wide-context units". W: *Multimedia Tools and Applications* 81.13 (2022), s. 18617–18639. (**70 pkt; JIF=3.6**).

[A9] Tomasz Grzywalski i **Szymon Drgas** (45%). "Speech Enhancement by Multiple Propagation through the Same Neural Network". W: *Sensors* 22.7 (2022), s. 2440. (**MNiSzW: 100 pkt.; JIF=3,9**).

[A10] **Szymon Drgas** (85%) i Tuomas Virtanen. "Speaker verification using adaptive dictionaries in non-negative spectrogram deconvolution". W: *Latent Variable Analysis and Signal Separation: 12th International Conference, LVA/ICA 2015, Liberec, Czech Republic, August 25-28, 2015, Proceedings 12*. Springer. 2015, s. 462–469.

[A11] Tomasz Grzywalski i **Szymon Drgas**. "Speech enhancement by iterating forward pass through U-net". W: *2020 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA)*. IEEE. 2020, s. 157–162.

[A12] Tomasz Grzywalski i **Szymon Drgas** (50%). "Using recurrences in time and frequency within U-net architecture for speech enhancement". W: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2019, s. 6970–6974.

[A13] Tomasz Grzywalski i **Szymon Drgas**. "Application of recurrent U-net architecture to speech enhancement". W: *2018 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA)*. IEEE. 2018, s. 82–87.

4.1.1 Charakterystyka osiągnięć naukowych wchodzących w skład cyklu publikacji

4.1.2 Motywacja podjętych badań

W rzeczywistych sytuacjach komunikacji głosowej, oprócz mówcy występują zakłócające źródła dźwięku oraz pogłos. Przykładowo można przywołać scenariusz rozmowy na lotnisku, gdy oprócz osoby, z którą prowadzimy konwersację, można usłyszeć w tle głosy innych osób, komunikaty, dźwięk przylatujących i odlatujących samolotów, itd. W przypadku komunikacji zdalnej, dodatkowo mogą występować zniekształcenia sygnału wynikające na przykład z jego kompresji.

Obecność dodatkowych dźwięków i zniekształceń w kanale komunikacyjnym może znacząco utrudniać zrozumienie treści wypowiedzi oraz zwiększać poziom błędów rozpoznawania mowy i mówcy przez systemy automatyczne. Zakłócenia i zniekształcenia sygnału stanowią bardzo duży problem u osób z uszkodzonym słuchem.

Ocena przetworzonego sygnału może odbywać się pod kątem jego zrozumiałości i jakości. Zrozumiałość mowy jest to odsetek prawidłowo zrozumianych słów. Pomiar zrozumiałości mowy jest trudnym zadaniem, ponieważ wspomniany odsetek poprawnie zrozumianych słów zależy m. in. od kontekstu i używanego słownictwa. Jakość jest jednym z wielu atrybutów sygnału mowy; jest ona bardzo subiektywna i trudna do wiarygodnej oceny. Jest tak chociażby z tego powodu, że różni słuchacze mają różne wewnętrzne standardy takie jak „dobra", „słaba", przekładające się na dużą zmienność wyników. Jakość może być określona przez takie atrybuty jak, „naturalność", „chropowatość", „szorstkość", itd. Z praktycznych względów, typowo korzysta się z do kilku aspektów jakości mowy w zależności od zastosowania.

Systemy do poprawy jakości i zrozumiałości mowy znajdują zastosowanie w aparatach słuchowych, a także na etapie wstępnego przetwarzania w systemach automatycznego rozpoznawania mowy lub mówcy. **Jedną z głównych motywacji moich prac jest potrzeba przeprowadzenia badań i opracowania nowych metod i architektur**

przetwarzania sygnałów, których wyniki mogą być zastosowane w aparatach słuchowych. Główną cechą tych zastosowań jest wymaganie małego poboru mocy, które dyskwalifikuje większość (prawie wszystkie) znanych rozwiązań zawierających obliczeniochłonne współczesne sieci neuronowe. Autor pracował nad nowymi strukturami obliczeniowymi, których celem jest uzyskanie alternatywnych rozwiązań o dużej wydajności. Struktury te zostały przedstawione w kolejnych punktach tego opracowania.

4.1.3 Przebieg prac badawczych składających się na osiągnięcia habilitacyjne

Po zakończeniu doktoratu dotyczącego automatycznego rozpoznawania mówcy, moje prace badawcze przesunęły się w kierunku modelowania mówcy w celu separacji mowy. Interesującym modelem umożliwiającym jednoczesne rozpoznawanie i separację mówcy okazała się nieujemna faktoryzacja macierzy (NMF – non-negative matrix factorization) i jej rozszerzenie – nieujemny rozplot macierzy (NMD – non-negative matrix deconvolution). W artykule [A10] przedstawiłem eksperymenty wskazujące na zasadność zastosowania NMD do modelowania mówcy. Tę ideę rozwinąłem do koncepcji adaptacyjnego słownika i opublikowałem w pracy [A4]. We wspomnianych publikacjach do budowy słownika wykorzystywana była informacja zawarta w etykietach. W celu uzyskania słownika o pożądanych właściwościach w sposób nienadzorowany, opracowałem metodę nieujemnego rozplotu macierzy z binarnymi macierzami aktywacji [A6]. Pokazałem także użyteczność takiego słownika w zadaniu rozpoznawania mówców i separacji sygnałów. Z tym zadaniem wiąże się także problem wyboru źródła dźwięku, na którym słuchacz koncentruje swoją uwagę. W tym celu można wykorzystać informację, zarejestrowaną przez elektrody elektroencefalografu (EEG). Przeprowadzenie badań przy użyciu EEG, było możliwe po podjęciu współpracy z badaczami z Wydziału Fizyki UAM i Centrum NanoBioMedycznego w Poznaniu. W pracy [A5] przedstawiłem możliwość połączenia znanych wcześniej metod NT (neural tracking) i APL (alpha power lateralization) i wyniki wskazujące na poprawioną trafność określania mówcy, na którym swoją uwagę skupia słuchacz.

Wraz z rozwojem sieci neuronowych, podjąłem próby zastosowania architektury U-net do zadania poprawy jakości i zrozumiałości mowy. W przeciwieństwie do separacji sygnałów, we wspomnianym zadaniu z mieszaniny wyodrębniany jest tylko jeden sygnał mowy. Wynikiem tych prac jest publikacja [A13], w której pokazałem (wraz ze współautorem), że zmniejszanie rozdzielczości czasowo-częstotliwościowej map cech na poszczególnych poziomach wpływa negatywnie na wskaźnik signal-to-distortion ratio (SDR) [1]. Pokazałem, że można tego efektu uniknąć stosując warstwy rekurencyjne. Metoda została rozwinięta i opisana w pracy [A12], którą zaprezentowałem na konferencji ICASSP. Zaproponowałem w niej bloki poszerzające kontekst, działające w wymiarze czasu lub częstotliwości. Następnie tę koncepcję rozwinąłem w pracy [A8]. Zaproponowane w tych artykułach sieci neuronowe do poprawy jakości i zrozumiałości mowy zostały także przetestowane w sytuacji wielokrotnej propagacji, w której przekształcony przez daną sieć neuronową sygnał jest ponownie przetwarzany przez tę samą sieć (jedno- lub wielokrotnie) [A11]. W pracy [A9] została przedstawiona metoda trenowania sieci neuronowych w taki sposób, żeby wielokrotnie przetwarzanie przez nie sygnału dawało pozytywny efekt. Tego typu rozwiązanie umożliwia łatwy dobór kompromisu pomiędzy kosztem obliczeniowym a jakością i zrozumiałością mowy. Wskazane powyżej prace (opisane w [A13, A12, A11, A8, A9]) podsumowałem w kontekście stanu wiedzy w [A3].

Opisane powyżej metody uczenia maszynowego opierają się na optymalizacji funk-

cji strat, które określają podobieństwo pomiędzy estymowanymi sygnałami i sygnałami wzorcowymi. Moje zainteresowania skoncentrowały się na funkcjach strat odzwierciedlających zrozumiałość mowy. W pracy [A2] zaproponowałem sieć neuronową do poprawy zrozumiałości mowy dla osób z uszkodzonym słuchem. Kolejnym krokiem było trenowanie parametrów sieci neuronowej, której architektura odpowiada, strukturze obliczeń w modelu ESTOI [2]. Na podstawie tej koncepcji opracowałem metodę FT-GESTOI [A1].

Podsumowanie stanu wiedzy w zakresie identyfikacji separowanych sygnałów Poniżej opisuję ogólną charakterystykę stanu wiedzy, do której odnoszą się moje prace. W rozdziałach dotyczących metod, które składają się na moje główne osiągnięcia naukowe, zawarłem odniesienia do literatury, charakteryzujące stan wiedzy dla poszczególnych zagadnień, które poruszyłem w swoich pracach.

1. Znane metody rozpoznawania mówcy ograniczają się głównie do przypadku, w którym w nagraniu obecny jest głos tylko jednej osoby.

2. Zagadnienia rozpoznawania mówców i separacji sygnałów były traktowane oddzielnie.

3. Znane metody umożliwiające identyfikację źródła dźwięku, na którym słuchacz chce skoncentrować swoją uwagę badane były oddzielnie. Utrudniało to ich bezpośrednie porównanie oraz określenie potencjalnych korzyści płynących z ich połączenia.

W rozdziale 4.1.4 zawarłem wprowadzenie do metod NMF i NMD, na których bazują zaproponowane przeze mnie metody DANMD i BNMD, które opisałem odpowiednio w rozdziałach 4.1.5 i 4.1.6. Natomiast opis metod dotyczących identyfikacji sygnałów bazujących na charakterystyce mówcy i sygnałach EEG zawarłem w rozdziałach 4.1.7 i 4.1.8.

4.1.4 Wprowadzenie do modeli NMF i NMD oraz separacji mówców

Metody separacji sygnałów mowy oparte na nieujemnej faktoryzacji macierzy, znajdują zastosowanie w przypadku małej albo średniej ilości danych. Ich ważną cechą jest możliwość adaptacji/dopasowania modelu podczas testu. W trybie treningu, w przypadku NMD, na podstawie treningowego zbioru danych wyznaczany jest słownik, składający się z atomów - czasowo-częstotliwościowych wzorców, dzięki którym można dokonać rozplotu spektrogramu.

Spektrogram magnitudowy sygnału mowy można reprezentować za pomocą macierzy $\mathbf{X} \in \mathbb{R}^{B \times N}$, gdzie B to liczba pasm częstotliwościowych a N to liczba ramek spektrogramu. Wszystkie elementy macierzy $\mathbf{X}$ są nieujemne. Nieujemna faktoryzacja macierzy (NMF) umożliwia aproksymację macierzy $\mathbf{X}$

$$\mathbf{X} \approx \mathbf{D}\mathbf{A} \,, \tag{1}$$

gdzie $\mathbf{D} \in \mathbb{R}^{B \times H}$ jest macierzą słownika, a macierz $\mathbf{A} \in \mathbb{R}^{H \times N}$ jest macierzą aktywacji. Kolumny macierzy $\mathbf{D}$ to atomy - czyli widma, z których utworzona jest każda kolumna macierzy $\mathbf{X}$. Każda kolumna macierzy $\mathbf{A}$ zawiera nieujemne współczynniki kombinacji liniowej atomów $\mathbf{D}$ będącej aproksymacją odpowiedniej kolumny macierzy $\mathbf{X}$.

Dla danej macierzy $\mathbf{X}$ macierze $\mathbf{D}$ i $\mathbf{A}$ można wyznaczyć za pomocą algorytmów, zaproponowanych przez Lee i Seunga [3]. Rozwiązują one następująco postawione zadania optymalizacji:

$$\begin{aligned} \min \quad & \|\mathbf{X} - \mathbf{DA}\|^2 \\ \text{p.o.} \quad & \mathbf{D} \geq 0 \,, \\ & \mathbf{A} \geq 0 \end{aligned} \tag{2}$$

gdzie $\|\cdot\|$ oznacza normę Frobeniusa a znak $\geq$ odnosi się do wszystkich elementów macierzy, lub

$$\begin{aligned} \min \quad & KL(\mathbf{X} \| \mathbf{DA}) \\ \text{p.o.} \quad & \mathbf{D} \geq 0 \,, \\ & \mathbf{A} \geq 0 \end{aligned} \tag{3}$$

gdzie $KL(\cdot)$ oznacza dywergencję Kulbacka-Leiblera. W przypadku sygnału mowy częściej wykorzystywany jest wariant ze wzoru (3).

NMF można też rozpatrywać jako model generatywny model rozkładu prawdopodobieństwa macierzy $\mathbf{X}$ z parametrami $\mathbf{D}$ i $\mathbf{A}$, w którym obserwacje (czyli elementy macierzy $\mathbf{X}$) są generowane w następujący sposób [4]:

1. losowane są zmienne ukryte z rozkładu Poissona

$$p(C_{bhn}) = \mathcal{P}(C_{bhn}; D_{bh} A_{hn}) \,, \tag{4}$$

2. obserwacje wyznaczane są ze wzoru

$$X_{bn} = \sum_{h=1}^{H} C_{bhn} \qquad p(\mathbf{X}; \mathbf{D}, \mathbf{A}) = \mathcal{P}(\mathbf{X} | \mathbf{DA}) \,. \tag{5}$$

Macierze $\mathbf{D}$ i $\mathbf{A}$ uzyskane przez minimalizację dywergencji KL (wzór (3)) są takie same jak te uzyskane przez maksymalizację wiarygodności $p(\mathbf{X}; \mathbf{D}, \mathbf{A})$.

Model opisany przez wzory (4) i (5) można zmodyfikować tak, żeby macierz $\mathbf{A}$ była zmienną losową o określonym rozkładzie prawdopodobieństwa a'priori. Wówczas optymalizacja macierzy $\mathbf{D}$ odbywa się poprzez zastosowanie algorytmu expectation-maximization (EM) [5].

Interesujący jest model, w którym macierz $\mathbf{A}$ jest macierzą z elementami binarnymi. W tym przypadku rozkładem prawdopodobieństwa a'priori macierzy $\mathbf{A}$ jest rozkład Bernoulliego i obserwacje generowane są w następujący sposób:

1. Każdy element macierzy $\mathbf{A}$ losowany jest z rozkładu Bernoulliego

$$p(A_{hn}) = \mathcal{B}(A_{hn}; \pi_{hn}) \,. \tag{6}$$

2. Obserwacje generowane są z rozkładu Poissona o parametrze zależnym od $\mathbf{A}$:

$$p(\mathbf{X}) = \mathcal{P}(\mathbf{X}; \mathbf{DA}) \,. \tag{7}$$

Uczenie tego modelu polega na wyznaczeniu deterministycznych parametrów (elementów słownika $\mathbf{D}$) dla danego zbioru obserwacji $\mathbf{X}$ poprzez maksymalizację zlogarytmowanej funkcji wiarygodności

$$\mathcal{L} = \sum_{n=1}^{N} \log p(\mathbf{x}_n; \mathbf{D}) \,, \tag{8}$$

gdzie $\mathbf{x}_n$ jest n-tą kolumną macierzy $\mathbf{X}$. Problemem przy zastosowaniu algorytmu EM do optymalizacji słownika dla modelu opisanego przez wzory (6) i (7) jest, to że w kroku E (expectation) algorytmu EM potrzeba wyznaczyć dyskretny rozkład prawdopodobieństwa a posteriori zawierający 2^H prawdopodobieństw. Jednym z możliwych rozwiązań jest zastosowanie przybliżonego algorytmu expectation-truncation (ET) wprowadzonego w [6]. Idea ET polega na znacznym zredukowaniu liczby aktywnych atomów dla każdej kolumny macierzy $\mathbf{X}$ poprzez zastosowanie tzw. funkcji selekcji.

Na bazie NMF powstało rozszerzenie, tzw. nieujemny rozplot macierzy (non-negative matrix deconvolution, NMD [7]). Macierz $\mathbf{X}$ ze spektrogramem jest reprezentowana jako kombinacja liniowa przesuwanych w czasie wzorców czasowo-częstotliwościowych (atomów)

$$\mathbf{X} \approx \sum_{t=0}^{T-1} \mathbf{D}(t) \overset{t\rightarrow}{\mathbf{A}} , \tag{9}$$

gdzie $\overset{t\rightarrow}{\cdot}$ oznacza przesunięcie wszystkich elementów macierzy o t pozycji na prawo, np:

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \end{bmatrix} \qquad \overset{2\rightarrow}{\mathbf{A}} = \begin{bmatrix} 0 & 0 & 1 & 2 \\ 0 & 0 & 5 & 6 \end{bmatrix} , \tag{10}$$

a macierze $\mathbf{D}(0), \ldots, \mathbf{D}(T-1)$ zawierają słownik.

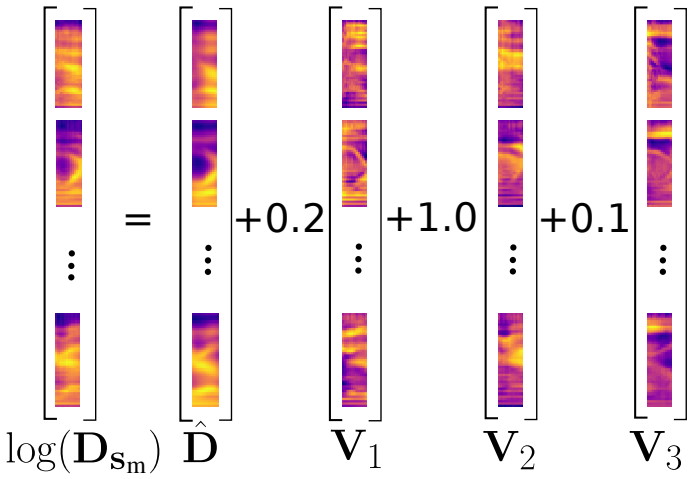
4.1.5 Metoda DANMD - słownik o niewielkiej liczbie parametrów

Motywacja: Metody separacji sygnałów mowy oparte na nieujemnej faktoryzacji macierzy znajdują zastosowanie w przypadku małej albo średniej ilości danych. Ich ważną cechą jest możliwość adaptacji/dopasowania parametrów modelu w fazie testu, np. kiedy dla macierzy $\mathbf{D}$ uzyskanej w fazie treningu, w fazie testu optymalizowana jest tylko macierz $\mathbf{A}$, tak żeby macierz wygenerowana przez model była jak najbliższa obserwacji $\mathbf{X}$.

Właściwości modelu NMD dają możliwość opracowania metody, która jednocześnie realizuje dwa zadania – separację sygnałów i rozpoznawanie mówców w nagraniu zawierającym mieszaninę głosów wielu osób. Zadanie jednoczesnej separacji i identyfikacji może mieć różne warianty: tzn. działać w zbiorze zamkniętym lub w zbiorze otwartym. W zadaniach dla zbioru zamkniętego nieznane są tożsamości mówców w mieszaninach ze zbioru testowego, ale wiadomo, że nagrania ich głosów (w izolacji) są w zbiorze treningowym. W wariancie dla zbioru otwartego nagrania izolowanych głosów nie występują w zbiorze treningowym.

Wcześniejsze znane z literatury metody do separacji i rozpoznawania sygnałów korzystające z NMF pokazały, że rozwiązanie oparte na adaptacji słownika może przynieść dobre rezultaty w przypadku niewielkich zbiorów treningowych. Zarówno dla zadania dla zbioru otwartego jak i zamkniętego ważna jest adaptacja słownika. We wcześniejszych znanych metodach z adaptacją słownika [8, 9], każdy atom jest adaptowany niezależnie.

W pracy [A4] zaproponowałem ograniczenie słowników do liniowej rozmaitości, co skutkuje redukcją liczby parametrów do zaadaptowania, co może także wpłynąć na lepszą separację dla krótkich nagrań. **Zaproponowałem metodę DANMD (dictionary adaptive NMD), w której parametryczny słownik może adaptować się do mówców wypowiadających się jednocześnie w nagraniu dźwiękowym.** W mojej metodzie słownik jest sumą tzw. *słownika niezależnego od mówcy* i liniowej kombinacji *komponentów reprezentujących zmienność mówców*. Współczynniki tej kombinacji liniowej stanowią parametry słownika.



Rysunek 1: Ilustracja przykładowego słownika adaptacyjnego

W rozważanym słowniku parametrycznym każdy atom jest funkcją parametrów słownika. Optymalne wartości adaptowanych parametrów słownika zależą od wszystkich ramek spektrogramu. **Jest to nowość w stosunku do innych metod, w których wartości optymalne parametrów adaptowanego słownika zależą jedynie od ramek spektrogramu, w których są aktywne.** To może prowadzić do mniejszej ilości danych wymaganych do adaptacji.

Metoda:
W zaproponowanej metodzie spektrogram mówcy m jest modelowany jako

$$\bar{\mathbf{X}}_m = \sum_{t=0}^{T-1} \mathbf{D}_{\mathbf{s}_m}(t) \overset{t\rightarrow}{\mathbf{A}}_m \ , \tag{11}$$

gdzie

$$\mathbf{D}_{\mathbf{s}_m}(t) = \exp\left(\hat{\mathbf{D}}(t) + \sum_{k=1}^{K} \mathbf{V}_k(t) s_{mk} \right) \ , \tag{12}$$

gdzie $\mathbf{s}_m = [s_{m1} \ \ldots \ s_{mK}]^{\mathrm{T}}$ jest wektorem parametrów charakteryzujących mówcę m a $\mathbf{D}_{\mathbf{s}_m}(t)$ jest $B \times P$ macierzą słownika z parametrami $\mathbf{s}_m$. Macierz $\hat{\mathbf{D}}(t)$ o wymiarach $B \times P$ nazywana jest słownikiem niezależnym od mówcy, a $\mathbf{V}_1(t), \ldots, \mathbf{V}_K(t)$ są komponentami reprezentującymi zmienność mówców. Słownik niezależny od mówcy i komponenty reprezentujące zmienność mówców są hiperparametrami słownika. Ilustracja słownika adaptacyjnego jest przedstawiona na rys. 1, gdzie macierz $\mathbf{V}_k$ zawiera macierze $\mathbf{V}_k(1), \ldots, \mathbf{V}_k(T)$ które zostały przeorganizowane, tak żeby zobrazować atomy o wymiarach $B \times T$. Analogicznie, macierz $\hat{\mathbf{D}}$ jest otrzymana z elementów $\hat{\mathbf{D}}(0), \ldots, \hat{\mathbf{D}}(T-1)$.

Macierz spektrogramu obserwacji mieszaniny M_A mówców jest modelowana jako

$$\mathbf{X} \approx \bar{\mathbf{X}} = \sum_{m=1}^{M_A} \bar{\mathbf{X}}_m \ . \tag{13}$$

Obserwowana macierz spektrogramu jest w tym przypadku aproksymowana przez sumę M_A macierzy, gdzie M_A jest liczbą mówców występujących w nagraniu. Dla mieszanin dwóch mówców, podczas separacji dla danych hiperparametrów słownika i macierzy obserwowanych spektrogramów, rozwiązywany jest następujący problem optymalizacji

$$
\begin{aligned}
\min_{\mathbf{s}_1,\mathbf{s}_2,\mathbf{A}} \quad & D_{KL}\left(\mathbf{X}\|\bar{\mathbf{X}}\right) && +\lambda_{\mathbf{s}}\left(\|\mathbf{s}_1\|_1+\|\mathbf{s}_2\|_1\right) \\
& && +\lambda_{\mathbf{A}}(\|\mathbf{A}_1\|_1+\|\mathbf{A}_2\|_1) \\
& && +\lambda_{\mathbf{s}_1\mathbf{s}_2}\mathbf{s}_1^{\mathrm{T}}\mathbf{s}_2 \\
& && +\lambda_{\mathbf{A}_1\mathbf{A}_2}\|\mathbf{A}_1\odot\mathbf{A}_2\|_1 \\
\text{p.o.:} \quad & \mathbf{A}\geq 0 && \\
& \mathbf{s}_1\geq 0, \|\mathbf{s}_1\|_2=1 && \\
& \mathbf{s}_2\geq 0, \|\mathbf{s}_2\|_2=1 && ,
\end{aligned}
\tag{14}
$$

gdzie parametry słownika $\mathbf{s}_1$ i $\mathbf{s}_2$ oraz macierz aktywacji

$$
\mathbf{A}=\left[\begin{array}{c}\mathbf{A}_1\\ \mathbf{A}_2\end{array}\right] \tag{15}
$$

są optymalizowane tak, żeby zminimalizować dywergencję Kullbacka-Leiblera pomiędzy spektrogramem mieszaniny i jej modelem. Optymalizowana funkcja zawiera także składniki które wprowadzają kary dla niechcianych wartości parametrów modelu, które są opisane poniżej. Zmienna $\lambda_{\mathbf{A}}$ jest wagą dla składnika wymuszającego rzadkość macierzy $\mathbf{A}$. Zmienna $\lambda_{\mathbf{s}_1\mathbf{s}_2}$ jest wagą składnika, który nakłada karę za podobieństwo wektorów $\mathbf{s}_1$ i $\mathbf{s}_2$, a $\lambda_{\mathbf{A}_1\mathbf{A}_2}$ jest podobną wagą,ale dla $\mathbf{A}_1$ i $\mathbf{A}_2$. Symbol $\odot$ oznacza iloczyn Hadamarda (indywidualne mnożenie odpowiednich elementów macierzy). Zadanie optymalizacji (wzór (14)) może być rozwiązane przy użyciu algorytmu 1. Macierz $\mathbf{A}$ jest poprawiana przy użyciu reguły mnożeniowej

$$
\mathbf{A}\leftarrow\sum_{t=0}^{T-1}\mathbf{A}\odot\frac{\mathbf{D}^{\mathrm{T}}(t)\overset{\leftarrow t}{\left(\frac{\mathbf{X}}{\bar{\mathbf{X}}}\right)}}{\mathbf{D}^{\mathrm{T}}(t)\overset{\leftarrow t}{(\mathbf{1}_{B\times N})}+\lambda_{\mathbf{A}}\mathbf{1}_{P\times N}+\lambda_{\mathbf{A}_1\mathbf{A}_2}\mathbf{A}^{\mathrm{R}}}\,, \tag{16}
$$

gdzie

$$
\mathbf{A}^{\mathrm{R}}=\left[\begin{array}{c}\mathbf{A}_2\\ \mathbf{A}_1\end{array}\right] \tag{17}
$$

i $\mathbf{1}_{A\times B}$ jest macierzą o rozmiarach A na B wypełnioną jedynkami.

Istnieją różne sposoby na uzyskanie hiperparametrów słownika ($\hat{\mathbf{D}}(t)$ i $\mathbf{V}_1(t),\ldots,\mathbf{V}_K(t)$ dla $t=0,\ldots,T-1$), które rozważałem dla zadań separacji ze zbioru otwartego i zbioru zamkniętego, które opisałem w [A4].

Wyniki eksperymentów

Eksperymenty przeprowadziłem dla nagrań ze zbioru TIDIGITS. Wyniki przedstawione w [A4] pokazują, że metoda DANMD daje lepsze wyniki (w sensie SDR) niż rozwiązanie bazowe - NMD z łączonym słownikiem. Metoda z łączonym słownikiem działa w taki sposób, że w fazie testu stosowane jest NMD ze słownikiem będącym połączeniem słowników trenowanych osobno dla wszystkich mówców ze zbioru treningowego. Separacja i rozpoznawanie realizowane są w oparciu o atomy, dla których odpowiednie wartości w macierzy aktywacji są największe. Metoda DANMD dała też lepszy wynik niż w przypadku sieci neuronowej BLSTM (bi-directional long short-term memory) trenowanej przy użyciu metody uPIT (utterance-level permutation invariant training) [10]. Uzyskałem też pozytywny wynik, dla przypadku ze zbiorem otwartym – wskaźnik SDR był wyraźnie dodatni.

Algorithm 1 Optymalizacja parametrów modelu DANMD

$\mathbf{s}_1 \leftarrow$ ZEROS()
$\mathbf{s}_2 \leftarrow$ ZEROS()
$\mathbf{A}_1 \leftarrow$ RAND()
$\mathbf{A}_2 \leftarrow$ RAND()
repeat
    Obliczyć macierz $\mathbf{A}^{\text{new}}$ przy użyciu wzoru (16), zaktualizować jedyni te wiersze macierzy $\mathbf{A}$ odnoszące się do $\mathbf{A}_1$.
    Obliczyć $\bar{\mathbf{X}}$.
    Obliczyć ponownie macierz $\mathbf{A}^{\text{new}}$ przy użyciu Wzoru (16), poprawiać jedynie wiersze $\mathbf{A}$ odnoszące się do $\mathbf{A}_2$.
    Obliczyć $\bar{\mathbf{X}}$.
    Zaktualizować $\mathbf{s}_1$ przy pomocy metody gradientowej.
    Obciąć ujemne wartości $\mathbf{s}_1$.
    Znormalizować $\mathbf{s}_1$ do $f(\text{iter})$.
    Zaktualizować $\mathbf{s}_2$ przy pomocy metody gradientowej.
    Obciąć ujemne wartości $\mathbf{s}_2$.
    Znormalizować $\mathbf{s}_2$ do $f(\text{iter})$.
    Zaktualizować słownik korzystając ze zaktualizowanych wektorów $\mathbf{s} = [\mathbf{s}_1^{\mathrm{T}}\ \mathbf{s}_2^{\mathrm{T}}]^{\mathrm{T}}$.
until zbieżność

4.1.6 Metoda BNMD - nieujemny rozplot macierzy z binarnymi macierzami aktywacji

Motywacja: W przypadku zastosowania standardowej procedury NMD do wyznaczenia słownika, uzyskane atomy często dążą do mało charakterystycznych wzorców - pojedynczych maksimów. Może to wynikać z rozkładów a'priori nakładanych na aktywacje. W literaturze istnieją metody bazujące na metodyce expectation-truncation przeznaczone do nieujemnej faktoryzacji macierzy [6]. **Nie ma jednak doniesień o metodach umożliwiających nieujemny rozplot macierzy z binarną macierzą aktywacji.**
Cel: Opracowanie metody uczenia słownika, w której zastosowany rozkład a'priori wymusi uczenie charakterystycznych atomów czasowo-częstotliwościowych
Metoda: W pracy [A6] zaproponowałem metodę BNMD (binary non-negative matrix deconvolution), w której dzięki metodyce expectation-truncation było możliwe dopasowanie modelu NMD, w którym rozkład a'priori macierzy aktywacji jest rozkładem Bernoulliego. Model BNMD można zapisać w następujący sposób

$$p(\mathbf{X}|\mathbf{A}) = \prod_{b=1}^{B}\prod_{n=1}^{N} \mathcal{P}\left(X_{bn}; \sum_{h=1}^{H}\sum_{t=0}^{T-1} D_{bh}(t) A_{h,n-t}\right) , \tag{18}$$

gdzie prawdopodobieństwo a'priori elementów macierzy $\mathbf{A}$ jest określone poprzez wzór:

$$p(\mathbf{A};\pi) = \prod_{h=1}^{H}\prod_{n=1}^{N} \pi^{A_{hn}}(1-\pi)^{1-A_{hn}} . \tag{19}$$

Energia swobodna dla tego modelu jest określona za pomocą zależności

$$\mathcal{F} = \sum_{\mathbf{A}\in\mathcal{K}} q(\mathbf{A}) \log p(\mathbf{X}, \mathbf{A}; \mathbf{D}(0), \ldots, \mathbf{D}(T-1)) + \mathcal{H}[q] , \tag{20}$$

gdzie $\mathcal{H}[\cdot]$ oznacza entropię Shannona, a zbiór $\mathcal{K}$ jest zdefiniowany jako

$$\mathcal{K} = \left\{ \mathbf{A} \in \mathcal{E} : \sum_{h=1}^{H} \sum_{n=1}^{N} A_{hn} \leq \gamma \wedge \left(\overset{\forall}{(h,n)} \notin \mathcal{D} : A_{hn} = 0 \right) \right\} , \tag{21}$$

przy czym $\mathcal{D}$ jest zbiorem indeksów h, n dla których funkcja selekcji przyjmuje najmniejsze wartości (ich liczba jest ograniczona przez odpowiedni parametr) a $\mathcal{E} = \{0,1\}^{H \times N}$.

Liczba kombinacji ukrytych zmiennych (elementów macierzy $\mathbf{A}$) może być bardzo duża w typowych zastosowaniach przetwarzania mowy. Nawet po ograniczeniu wszystkich możliwych wartości macierzy $\mathbf{A}$ tj. zbioru $\mathcal{E}$ do zbioru $\mathcal{K}$, liczba ta może być niemożliwa do uwzględnienia w obliczeniach, tak żeby zakończyły się one w akceptowalnym czasie. W pracy [A6] zaproponowałem rozwiązanie polegające na podziale spektrogramu na segmenty. Ze spektrogramu wyodrębniane są segmenty o długości L kroczące z długością kroku S. Segment o indeksie k: $Y_{b,(k-1)S+l}$ o długości L jest modelowany jako

$$\bar{X}_{b,(k-1)S+l} = \sum_{t=0}^{T-1} \sum_{h=1}^{H} D_{bh}(t) A_{h,(k-1)S+l-t}, \qquad \text{dla } l = 1 \ldots, L \tag{22}$$

co można zapisać w postaci macierzowej jako

$$\mathbf{X}_k \approx \sum_{t=0}^{T-1} \mathbf{D}(t) \overset{t\rightarrow}{\mathbf{A}}_k , \tag{23}$$

gdzie $\mathbf{X}_k$ jest k-tym segmentem obserwacji, a $\mathbf{A}_k \in \{0,1\}^{H \times L+T-1}$ jest k-tym segmentem macierzy aktywacji. Warto podkreślić, że macierz aktywacji dla segmentu $\mathbf{X}_k$ zawiera wszystkie elementy macierzy $\mathbf{A}$, które mają wpływ na elementy w segmencie $\mathbf{X}_k$. Dla indeksów $k < T$, dla niedodatnich indeksów, elementy macierzy $\mathbf{A}$ mają wartość zero.

Dla wersji z segmentowaniem funkcja energii swobodnej jest wyrażona w następujący sposób

$$\mathcal{F} = \sum_{k \in \mathcal{O}} \sum_{\mathbf{A}_k \in \mathcal{K}} q_k(\mathbf{A}_k) \log p(\mathbf{X}_k, \mathbf{A}_k | \mathbf{D}(0), \ldots, \mathbf{D}(T-1)) + \mathcal{H}[q_k] , \tag{24}$$

gdzie zbiór $\mathcal{K}_k$ jest otrzymywany w podobny sposób jak $\mathcal{K}$ z tą różnicą, że odnosi się jedynie do k'tego segmentu a $\mathcal{O}$ zawiera indeksy K_{cut} segmentów, dla których wartości $\sum_{\mathbf{A}_k \in \mathcal{K}_k} p(\mathbf{A}_k, \mathbf{X}_k | \mathbf{D}(0), \ldots, \mathbf{D}(T-1))$ są największe.

Ze względu na aproksymacje w kroku E, zbieżność zaproponowanego algorytmu uczącego BNMD nie jest gwarantowana. Jednakże eksperymenty wykazały, że w praktyce funkcja wiarygodności nie zmniejsza się znacząco podczas uczenia.

Zbiór $\mathcal{K}_k$ wyznaczany dla k-tego segmentu jest przy użyciu funkcji selekcji, która na podstawie bieżącego słownika i segmentu, wskazuje, które elementy macierzy aktywacji mogą być aktywne. Można to zrobić np. przy użyciu wzoru:

$$f(\underline{\mathbf{X}}, \mathbf{D}_h) = \frac{\text{trace}(\underline{\mathbf{X}}^{\mathrm{T}} \mathbf{D}_h)}{\|\underline{\mathbf{X}}\|_F \|\mathbf{D}_h\|_F} , \tag{25}$$

gdzie $\underline{\mathbf{X}}$ oznacza T-ramkowy fragment spektrogramu a

$$\mathbf{D}_h = \begin{bmatrix} D_{1h}(0) & \ldots & D_{1h}(T-1) \\ \vdots & \ddots & \vdots \\ D_{Bh}(0) & \ldots & D_{Bh}(T-1) \end{bmatrix} . \tag{26}$$

Zaproponowałem następujące funkcje selekcji [A6]:

1. podobieństwo kosinusowe,
2. podobieństwo kosinusowe w skali logarytmicznej,
3. funkcja $f_{log,regul}$.

Procedura uczenia słownika przy użyciu BNMD jest przedstawiona w pseudokodzie 2.

Algorithm 2 Uczenie słownika przy użyciu BNMD

Inicjalizacja słownika $\{\mathbf{D}(t)\}_{t=0}^{T-1}$ przy użyciu losowych liczb dodatnich lub przy użyciu atomów uzyskanych za pomocą kwantyzacji wektorowej.
repeat
  for każdy segment $k \in \{1 \dots K\}$ **do**
    Obliczyć funkcję selekcji dla całej wypowiedzi przy użyciu jednej z funkcji selekcji.
    Obliczyć I' kandydatów (macierzy ze zbioru $\mathcal{K}_k$, dla których funkcja selekcji przyjmuje najmniejsze wartości).
    Obliczyć przybliżone wartości prawdopodobieństw a'posteriori dla każdego stanu ze zbioru $\mathcal{K}_k$ przy użyciu wzoru

$$q_k(\mathbf{A}_k) = \frac{p(\mathbf{X}_k|\mathbf{A}_k)p(\mathbf{A}_k)}{\sum_{\mathbf{A}'_k \in \mathcal{K}_k} p(\mathbf{X}_k|\mathbf{A}'_k)p(\mathbf{A}'_k)} \,. \tag{27}$$

    Obliczyć oczekiwaną macierz aktywacji

$$\langle \mathbf{A}_k \rangle_{\mathrm{ET}} = \sum_{\mathbf{A}_k \in \mathcal{K}_k} q_k(\mathbf{A}_k)\mathbf{A}_k \,, \tag{28}$$

    Obliczyć rekonstrukcję

$$\hat{\mathbf{X}}_k = \sum_{t=0}^{T-1} \mathbf{D}(t)\langle \overset{t\rightarrow}{\mathbf{A}_k} \rangle_{\mathrm{ET}} \tag{29}$$

    Zachować w pamięci licznik i mianownik ze wzoru (30) dla bieżącego k.
  end for
  Dla każdego segmentu obliczyć $\sum_{\mathbf{A}_k \in \mathcal{K}_k} p(\mathbf{A}_k, \mathbf{X}_k|\mathbf{D}(0), \dots, \mathbf{D}(T-1), \sigma^2)$ i przypisać do $\mathcal{O}$ zbiór K_{cut} indeksów segmentów, które mają największe wartości.
  Zaktualizować słownik za pomocą wzoru

$$\mathbf{D}(t) \leftarrow \mathbf{D}(t) \odot \sum_{k \in \mathcal{O}} \frac{\begin{pmatrix} \mathbf{X}_k \\ \hat{\mathbf{X}}_k \end{pmatrix} \left[(\langle \overset{t\rightarrow}{\mathbf{A}_k} \rangle_{\mathrm{ET}}) \right]^{\mathrm{T}}}{\mathbf{1} \left[\langle \overset{t\rightarrow}{\mathbf{A}_k} \rangle_{\mathrm{ET}}) \right]^{\mathrm{T}}} \,. \tag{30}$$

until zbieżność

Wyniki:
Wyniki eksperymentów, które przedstawiłem w [A6] pokazują, że użycie słowników wyuczonych za pomocą BNMD daje lepszą trafność rozpoznawania mówców niż systemy bazujące na GMM i I-vector. Jest to najbardziej zauważalne dla małych SNR, gdzie

trafność dla GMM wyniosła 68%, podczas gdy dla rozplotu ze słownikiem BNMD zawierającym 50 atomów 88%. Poprawa została także w stosunku do NMD i VQ (vector quantization). Okazało się, że słowniki uzyskane przy użyciu kwantyzacji wektorowej są lepsze niż NMD (trafności uśrednione po wszystkich testowanych SNR wyniosły odpowiednio 87.8% i 75.4%) Słownik BNMD dał w tym przypadku 91.3%. Ta poprawa w stosunku do innych metod jest najbardziej zauważalna dla słowników o małych liczbach atomów. Wraz ze wzrostem liczby atomów poprawa maleje. Słownik BNMD daje lepsze wyniki w stosunku do słownika zbudowanego ze wzorców uzyskanych na postawie etykiet fonetycznych. Dla 250-atomowego słownika BNMD dało wynik 92,5% a wzorców uzyskanych na podstawie etykiet otrzymano wartość 90,9%.
Dla zadania separacji mowy słowniki wyuczone dają nieznacznie lepsze SDR (8,5 w porównaniu do 8,3 dB). Jednakże większe wartości SDR mogą być zaobserwowane dla największych testowanych SNR (-3–9 dB). Na przykład dla 50-atomowego słownika BNMD poprawa była o 0,5 dB większa dla 9 dB SNR ale 0,2 dB.

Podsumowując, w pracy [A6] zaproponowałem metodę umożliwiającą nieujemny rozplot macierzy z binarną macierzą aktywacji. Słowniki uzyskane za pomocą tej metody okazały się skuteczne w zadaniach rozpoznawania mówcy i separacji sygnałów.

4.1.7 Łączenie cech charakteryzujących mówców za pomocą AdaGrad

Motywacja:
Wysokopoziomowe cechy sygnału mowy (np. cechy prozodyczne) dostarczają uzupełniającej informacji do klasycznych cech spektralnych i sprawiają, że system do rozpoznawania mówcy jest bardziej odporny na zakłócenia [11, 12]. Rozpoznawanie mówcy w kontekście separacji sygnałów ma na celu identyfikację odseparowanych źródeł dźwięku występujących w analizowanej mieszaninie. W pracy [A7] zastosowałem połączenie cech spektralnych, prozodycznych, artykulacyjnych i leksykalnych przy użyciu metody AdaGrad, która łączy macierze jąder dla poszczególnych cech. Podobieństwo ze względu na mówcę dla dwóch różnych nagrań, w przeciwieństwie do bardziej skomplikowanych klasyfikatorów, określano za pomocą podobieństwa kosinusowego z zastosowaną normalizacją typu *z-norm* i *z-norm2*. **Celem pracy było przygotowanie i przeprowadzenie eksperymentów testujących połączenie metod normalizujących i algorytmu Ada-Grad dla niewielkiego zbioru danych jakim jest baza nagrań PUEPS opisana w [13].**
Metoda:
W opracowanym przeze mnie systemie, dla każdego nagrania wyznaczane są wektory $\mathbf{x}_1, \ldots, \mathbf{x}_K$ reprezentujące ze względu na różne cechy: np. widmowe, prozodyczne, fonetyczne i leksykalne, tak jak w [A10]. Do porównywania cech zastosowałem kosinusową funkcję jądra

$$(\mathbf{K}_k)_{ij} = \frac{\mathbf{x}_{ki}^{\mathrm{T}} \mathbf{x}_{kj}}{\|\mathbf{x}_{ki}\| \|\mathbf{x}_{kj}\|} ,$$

gdzie $\mathbf{K}_k$ oznacza macierz jądra zestawu cech o indeksie k, której elementy zawierają podobieństwo kosinusowe pomiędzy parami nagrań ze zbioru danych. Normalizacja z-norm polega porównywaniu cech i-tego mówcy do nagrań innych mówców, zawartych w zbiorze B. Następnie statystyki (średnia $\mu_z^{(i)}$ i wariancja $\sigma_z^{(i)}$) są obliczane z wynikającego zbioru punktacji $\{(\mathbf{K}_k)_{ij}\}_{j \in B}$. Kolejnym krokiem jest wyznaczenie nowej macierzy jądra,

tak żeby uwzględniała normalizację

$$(\bar{\mathbf{K}})_{ij} = \frac{(K_k)_{ij} - \mu_z^{(i)}}{\sigma_z^{(i)}} \ .$$

Dalej macierze z punktacją są obliczane według wzoru

$$\mathbf{S}_k = \bar{\mathbf{K}}_k - \theta_k^{\mathrm{EER}} \mathbf{1}\mathbf{1}^{\mathrm{T}} \ ,$$

gdzie $\mathbf{1}$ jest wektorem wypełnionym jedynkami, a θ_k^{EER} jest progiem, dla którego błąd pominięcia jest taki sam jak fałszywy alarm. System przetestowałem także dla metody z-norm2, gdzie punktacje były przeskalowane w taki sposób, że wszystkie elementy na przekątnej macierzy jądra były jednostkowe. Zastosowany algorytm AdaGrad można zapisać w następujący sposób:

1. **Wejście**: Macierze punktacji i etykiet $\{(\mathbf{S}_k, \mathbf{D})\}_{k=1}^{K}$, gdzie

$$(\mathbf{D})_{ij} = \begin{cases} 1 & \text{jeśli } \mathbf{x}_i \text{ i } \mathbf{x}_j \text{ reprezentują nagranie tego samego mówcy} \\ 0 & \text{w przeciwnym przypadku} \end{cases}$$

2. **Inicjalizacja**: $\mathbf{W} = 1/m$, $\hat{\mathbf{S}} = \mathbf{0}$

3. **Dla każdej pary**$(\mathbf{S}_k, \mathbf{D})$

   (a) $S^+ = \{(i,j) : (\mathbf{S}_k)_{ij}(\mathbf{D})_{ij} > 0\}$,$S^- = \{(i,j) : (\mathbf{S}_k)_{ij}(\mathbf{D}_k)_{ij} < 0\}$

   (b) $W^+ = \sum_{(i,j)\in S^+}(\mathbf{W}_{ij}|(\mathbf{S}_k)_{ij}|)$
   $W^- = \sum_{(i,j)\in S^-}(\mathbf{W}_{ij}|(\mathbf{S}_k)_{ij}|)$

   (c) $\lambda_k = \frac{1}{2}\log\left(\frac{W^+}{W^-}\right)$

   (d) $\mathbf{W}_{ij} = \mathbf{W}_{ij}\exp(-\lambda_k(\mathbf{D}_k)_{ij}(\mathbf{S}_k)_{ij})$

   (e) $\mathbf{W} = \frac{\mathbf{W}}{\mathbf{1}^{\mathrm{T}}\mathbf{W}\mathbf{1}}$, gdzie $\mathbf{1}$ jest wektorem kolumnowym, którego wszystkie składowe są równe 1, $\mathbf{1}^{\mathrm{T}}\mathbf{W}\mathbf{1}$ jest sumą wszystkich elementów macierzy $\mathbf{W}$.

   (f) $\hat{\mathbf{S}} = \hat{\mathbf{S}} + \lambda_k \mathbf{S}_k$

4. **Wyjście**: Macierz jądra $\hat{\mathbf{S}}$.

Wyniki:
Przeprowadzone przeze mnie eksperymenty na autorskiej bazie PUEPS pokazały, że proste metody punktowania bazujące na podobieństwie kosinusowym i normalizacji (w tym zaproponowana z-norm2) dają lepsze rezultaty niż klasyfikator SVM (support vector machine). Zastosowanie algorytmu AdaGrad w efektywny sposób połączyło informacje pochodzącą z różnych rodzajów cech, tj. spektralnych, prozodycznych, leksykalnych i artykulacyjnych.

4.1.8 Metoda do wyboru odseparowanego sygnału na podstawie EEG łącząca neural tracking (NT) i alpha power lateralization (APL)

Motywacja: System separacji sygnałów mowy może być zastosowany do wyodrębnienia wybranego mówcy z mieszaniny wielu głosów. W przypadku mieszaniny dwóch mówców dwa odseparowane można zaprezentować dychotycznie (jeden z sygnałów do lewego ucha,

a drugi do prawego). Następnie przy użyciu sygnałów EEG zarejestrowanych u słuchacza można podjąć próbę wskazania, na którym z głosów koncentruje on swoją uwagę. Tego typu technika nosi nazwę wykrywania uwagi słuchowej (AAD – auditory attention detection) [14]. Do określenia mówcy można zastosować, z pewną dokładnością, takie metody jak śledzenie neuronowe (NT - neural tracking) [15, 16, 17, 18] lub lateralizacja mocy fal alfa (APL) [19, 20, 21, 22, 23]. Metoda NT polega na tym, że rytm bodźca wyrażony poprzez obwiednię amplitudową sygnału mowy jest odzwierciedlony w aktywności neuronowej podczas odsłuchów. W technice APL wyznaczane są różnice mocy sygnałów pomiędzy parami elektrod umieszczonymi symetrycznie z lewej i prawej strony głowy. Te różnice zależą od kierunku, z którego dociera dźwięk pochodzący od źródła dźwięku, na którym skoncentrowany jest słuchacz. **Wspomniane metody (NT i APL) są zazwyczaj testowane osobno, co sprawia że niemożliwe jest bezpośrednie porównanie ich dokładności dla zadania AAD.**

Cel: Opracowanie metody łączącej informacje z neural tracking i lateralizację fal alfa w celu poprawy dokładności detekcji uwagi słuchowej.

Metoda:

W pracy [A5] zaproponowałem łączenie metod NT i APL. W technice NT rekonstrukcja obwiedni sygnału mowy (na którym skoncentrowana jest uwaga słuchacza) z sygnałów EEG jest określona wzorem

$$r(n) = \sum_{f=1}^{F} \sum_{t=1}^{\tau} c_f(t) e_f(n-t) \;, \tag{31}$$

gdzie $c_f(t)$ jest t-tym współczynnikiem filtru f-tego dekodera, F jest liczbą filtrów i τ jest rzędem tych filtrów. Współczynniki $c_f(t)$ są uzyskane przez minimalizację sumy kwadratów różnicy pomiędzy obwiednią sygnału $s(n)$ i jej rekonstrukcją $r(n)$. Dodatkowo, w optymalizowanej funkcji jest uwzględniony składnik regularyzujący bazujący na normie L2. Rozwiązywane zadanie optymalizacji można zapisać postaci

$$\min_{c_f(t)} \;\; \sum_{n=1}^{N} (r(n; c_f(t)) - s(n))^2 + \mu \sum_{t=1}^{\tau} \sum_{f=1}^{F} c_f^2(t) \;, \tag{32}$$

gdzie μ jest parametrem regularyzacji. Współczynniki filtru mogą być obliczone według następującego wzoru

$$\mathbf{c} = (\mathbf{E}^{\mathrm{T}}\mathbf{E} + \mu\mathbf{I})^{-1}\mathbf{E}^{\mathrm{T}}\mathbf{s} \;, \tag{33}$$

gdzie $\mathbf{E} = [\mathbf{e}(\tau+1) \;\ldots\; \mathbf{e}(N)]^{\mathrm{T}}$ i

$$\mathbf{e}(n) = [e_1(n-1) \;\ldots\; e_F(n-1)\; e_1(n-2) \;\ldots\; e_F(n-2) \;\ldots\; e_1(n-\tau) \;\ldots\; e_F(n-\tau)]^{\mathrm{T}} \;, \tag{34}$$

natomiast

$$\hat{\mathbf{c}} = [c_1(1) \;\ldots\; c_F(1) \;\; \ldots\; c_1(\tau) \;\ldots\; c_F(\tau)]^{\mathrm{T}} \tag{35}$$

i $\mathbf{s}$ zawierają próbki obwiedni czasowej sygnału mowy $s_1, \ldots, s_N$.

Współczynniki filtrów wyznaczałem osobno dla każdego z uczestników eksperymentu na podstawie przykładów, które zawierały sygnały EEG zarejestrowane podczas słuchania 50-sekundowych nagrań mowy (jednocześnie dwóch głosów w sposób dychotyczny). W fazie testu dla każdej próby, na podstawie sygnałów EEG rekonstruowałem obwiednie sygnałów przy użyciu filtru uzyskanego w fazie treningu. Wynikowa rekonstrukcja

była porównywana do obwiedni z obu sygnałów (z lewego i z prawego ucha) prezentowanych słuchaczowi, poprzez obliczenie podobieństwa kosinusowego. Słuchany sygnał był wskazywany na podstawie zmiennej

$$d_{\mathrm{NT}} = d_m - d_f \ . \tag{36}$$

Moduł APL bazuje na klasyfikacji cech wyrażających lateralizację mocy fal alfa w sygnałach EEG przy użyciu SVM. W celu wyznaczenia mocy fal alfa, wszystkie sygnały EEG są przetwarzane przez filtr pasmowo-przepustowy o częstotliwościach odcięcia 8 i 12 Hz (filtr rzędu 255 o skończonej odpowiedzi impulsowej, zaprojektowany metodą okien przy użyciu okna Hamminga). Następnie obliczono moc sygnału dla wszystkich kanałów EEG (okno wyznaczania mocy pokrywało cały 50 s. sygnał). Kolejnym krokiem było obliczenie różnic logarytmów mocy pomiędzy parami kontralateralnymi. W ten sposób określone zostały wskaźniki APL dla następujących par elektrod EEG, ((Fp1, Fp2), (F3, F4), (F7, F8), (FC5, FC6), (FC1, FC2), (T7, T8), (C3, C4), (CP5, CP6), (CP1, CP2), (P7, P8), (P3, P4), (PO9, PO10), i (O1, O2)). W efekcie dla każdego nagrania uzyskałem 13-wymiarowy wektor cech. SVM klasyfikujący cechy APL trenowałem na wydzielonym zbiorze treningowym. W czasie testów, wynikiem klasyfikacji wektora cech APL ($\mathbf{x}_{\mathrm{APL}}$) była tzw. punktacja

$$d_{\mathrm{APL}} = \mathbf{w}_{\mathrm{SVM}}^{\mathrm{T}} \mathbf{x}_{\mathrm{APL}} + b_{\mathrm{SVM}} \tag{37}$$

gdzie $\mathbf{w}_{\mathrm{SVM}}$ i b_{SVM} są parametrami wytrenowanego SVM.

Punktacje decyzyjne z systemów NT i APL (d_{NT} i d_{APL}) były łączone przy użyciu regresji logistycznej zaimplementowanej w pakiecie SKLEARN. Trening tego klasyfikatora był przeprowadzony przez rozwiązanie następującego zadania optymalizacji

$$\min_{\mathbf{w}} \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} + C\sum_{i=1}^{n} \log(\exp(-y_i(\mathbf{x}_i^{\mathrm{T}}\mathbf{w})) + 1) \ , \tag{38}$$

gdzie dane treningowe dla danego uczestnika eksperymentu są zawarte w parach $(\mathbf{x}_i, y_i)$ dla $i = 1, \ldots, n$, gdzie $\mathbf{x}_i$ oznacza wektor zawierający punktacje decyzyjne z podsystemów APL i NT dla i-tego nagrania treningowego, podczas y_i równe jest 1, gdy uwaga była skupiona na głosie męskim lub zero w przeciwnym przypadku, $\mathbf{w}$ zawiera parametry klasyfikatora, natomiast C jest parametrem regularyzacji. W opisanych eksperymentach C było ustawione na domyślną wartość 1.

Przetestowałem także przypadek, w którym waga kombinacji była stała dla wszystkich uczestników eksperymentu odsłuchowego. Punktacje decyzyjne z modułów NT i APL były zdefiniowane przez następującą kombinację liniową:

$$s = \alpha d_{NT} + (1 - \alpha) d_{APL} \ , \tag{39}$$

gdzie d_{NT} i d_{APL} są punktacjami decyzyjnymi odpowiednio z modułów NT i APL, gdzie α jest wagą kombinacji.

Wyniki:

Eksperymenty przeprowadziłem na zbiorze danych, który zarejestrowałem wspólnie ze współautorkami pracy [A5] w Centrum NanoBioMedycznym w Poznaniu. Dane pochodziły od 13 słuchaczy. Uzyskane wyniki przedstawione w [A5] pokazują, że czas trwania nagrania zmniejsza trafność we wszystkich testowanych systemach. W systemie bazującym na NT, zaobserwowałem statystycznie istotną różnicę pomiędzy 50 i 25 sekundowymi nagraniami ($p = 0.006$) a także pomiędzy 25 i 12.5 s ($p = 0.00002$). Podobnie, dla systemu łączącego NT i APL, wartość p wyniosła odpowiednio 0.00004 i 0.007. Dla systemu

bazującego na APL, zmniejszenie czasu trwania spowodowało istotną różnicę jedynie dla zmiany z 50 na 25 s. (p=0.002). W rezultacie dla każdego czasu trwania, kombinację NT i APL porównywałem do lepiej działającego modułu (NT lub APL). Kombinacja dała istotną poprawę średniej dokładności dla najdłuższych (50 s) i średnich (25 s) czasów trwania (odpowiednio $p = 0.005$ and $p = 0.011$). Dla najkrótszych próbek, poprawa nie była statystycznie istotna ($p = 0.51$). **Podsumowując, opracowałem system AAD łączący informacje z dwóch modułów: NT i APL. Uzyskane wyniki pokazują, że dla wystarczająco długich sygnałów kombinacja modułów NT i APL daje statystycznie istotną poprawę dokładności w porównaniu do lepiej działającego pojedynczego modułu.**

4.1.9 Sieci neuronowe do poprawy jakości i zrozumiałości mowy

W ostatnich latach można zaobserwować znaczny postęp w zakresie poprawy jakości mowy przy użyciu sieci neuronowych [24, 25, 26]. W porównaniu do zadania separacji sygnałów, które rozważałem we wcześniejszych punktach, wymienione publikacje opisują sieci neuronowe, które wyodrębniają z mieszaniny mowy i tła akustycznego jedynie sygnał mowy. Takie systemy na podstawie sygnału zaszumionego, lub jego reprezentacji (np. spektrogramu lub cech takich jak MFCC (mel-frequency cepstral coefficients)), wyznaczają tzw. maski, lub samą reprezentację oszacowanego czystego sygnału mowy (np. STFT).

W pierwszych metodach poprawy jakości mowy, bazujących na sieciach neuronowych, zakłócona mowa była reprezentowana jako spektrogram magnitudowy (krótko-terminowa transformata Fouriera, STFT – short-term Fourier transform), a sieci neuronowe były trenowane tak, żeby na wyjściach generowały maski, np. (IBM - ideal binary mask, IRM - ideal ratio mask) lub bezpośrednio spektrogram magnitudowy czystego sygnału mowy [27]. W ostatnim czasie opisano wiele sieci neuronowych, które bezpośrednio przekształcają sygnał wejściowy w dziedzinie czasu na czysty sygnał [28, 29, 30, 31].

W metodach bazujących na STFT, każdy czasowo-częstotliwościowy element maski jest funkcją obszaru czasowo-częstotliwościowego spektrogramu sygnału wejściowego. Ten obszar nazywany jest polem recepcyjnym. Rozmiar pola recepcyjnego zależy od architektury sieci neuronowej. Im większe pole recepcyjne, tym większy kontekst czasowo-częstotliwościowy może być użyty do wyznaczenia maski. W wielu przypadkach proste zwiększanie kontekstu wiąże się ze zwiększeniem liczby parametrów i przetrenowaniem sieci neuronowej. Wiąże się to także ze zwiększeniem liczby operacji arytmetycznych koniecznych do wyliczenia wyjścia sieci oraz wymagań dotyczących pamięci.

Głębokie sieci neuronowe oparte na warstwach w pełni połączonych (inaczej nazywanych: dense) mogą przekształcać wektor otrzymany poprzez połączenie kilku następujących po sobie ramek spektrogramu. W takim przypadku jednak, liczba parametrów rośnie stosunkowo szybko. Z tego względu sieci neuronowe zbudowane z warstw w pełni połączonych mają zazwyczaj niezbyt duże pole recepcyjne [32].

W celu zwiększenia efektywności powiększania pola recepcyjnego w literaturze można znaleźć propozycje następujących rozwiązań. Wśród nich są rekurencyjne sieci neuronowe (RNN – recurrent neural network) [32], sieci splotowe (CNN – convolutive neural networks) [33, 34], sploty z dylatacjami [35]. Można także znaleźć kombinacje warstw rekurencyjnych i splotowych [36].

Pole recepcyjne CNN może być powiększone poprzez zwiększanie głębokości sieci, co wiąże się z problemem zanikającego gradientu (vanishing gradient [37, 38]). Dodatkowo

można zastosować operacje takie jak max-pooling lub zwiększenie kroku splotu (stride). Te techniki nie zwiększają liczby parametrów sieci neuronowej, ale zmniejszają rozdzielczość czasowo-częstotliwościową map cech w warstwach ukrytych. W przypadku architektury U-net, informacja przestrzenna może być zachowana dzięki połączeniom skrótowym. Architektura U-net do poprawy jakości sygnału została zaproponowana przez habilitanta w [A13]. Zmniejszanie rozdzielczości czasowo-częstotliwościowej map cech w warstwach ukrytych może jednak powodować aliasing. Jedną z możliwości rozwoju sieci neuronowych do poprawy jakości i zrozumiałości mowy jest projektowanie architektur w taki sposób, żeby zwiększyć pole recepcyjne bez zmniejszania rozdzielczości czasowo-częstotliwościowej map cech.

Funkcja strat, w oparciu o którą jest trenowana sieć neuronowa, może być ukierunkowana na poprawę jakości lub zrozumiałości mowy. Proste metryki, takie jak np. odległość euklidesowa pomiędzy sygnałem zaszumionym i czystym, nie odzwierciedlają zrozumiałości mowy.

Efektywne algorytmy do poprawy zrozumiałości mowy mają duże znaczenie w kontekście aparatów słuchowych. Do negatywnych aspektów uszkodzonego słuchu należy w dużym stopniu zmniejszona zrozumiałość mowy, zwłaszcza w hałaśliwych sytuacjach. Wynika ona ze zniekształceń wzorców aktywności neuronowej [39]. Zniekształcone wzorce aktywności neuronowej wywołanej przez różne dźwięki są mniej jednoznaczna i bardziej wrażliwe na dźwięki tła akustycznego.

Poprawa zrozumiałości mowy może także uwzględniać ubytki słuchu. U osób z uszkodzonym słuchem część informacji jest utracona. U osób z sensorycznym uszkodzeniem słuchu można zaobserwować zmniejszoną rozdzielczość częstotliwościową. Oprócz tego, problemem jest wyrównywanie głośności (loudness recruitment) polegające na tym, że dla małych poziomów sygnału dźwięk jest niesłyszalny, a wraz ze wzrostem poziomu sygnału można zaobserwować szybki przyrost głośności. Po przekroczeniu pewnego poziomu ciśnienia akustycznego, wrażenie głośności jest takie jak u słuchaczy ze słuchem normalnym.

W celu predykcji zrozumiałości mowy u osób z uszkodzonym słuchem można połączyć model uszkodzonego słuchu wraz z metodami do predykcji zrozumiałości mowy. Dodatkowo, jeśli połączenie tych modeli jest różniczkowalne, to może być ono zastosowane do trenowania sieci neuronowych do poprawy zrozumiałości mowy.

Zaproponowane przeze mnie metody do poprawy jakości i zrozumiałości mowy są opisane w punktach 4.1.10–4.1.13.

Podsumowanie stanu wiedzy dotyczącego metod poprawy jakości i zrozumiałości mowy

1. Większość systemów bazuje na spektrogramie magnitudowym.

2. Wiele architektur sieci neuronowych ma ograniczone pole recepcyjne i nie wykorzystuje efektywnie parametrów. Znaczna część architektur jest nieprzyczynowa lub charakteryzuje się znaczną latencją.

3. Wielokrotne przetwarzanie sygnału przez tę samą sieć nie prowadzi do poprawy jakości odszumiania. Metoda *progressive learning* jest zaprojektowana do efektywniejszego trenowania sieci neuronowej.

4. Sieci neuronowe do poprawy jakości i zrozumiałości mowy charakteryzują się dużym kosztem obliczeniowym (w porównaniu do klasycznych metod, duże liczby warstw sieci neuronowych i filtrów warstw splotowych).

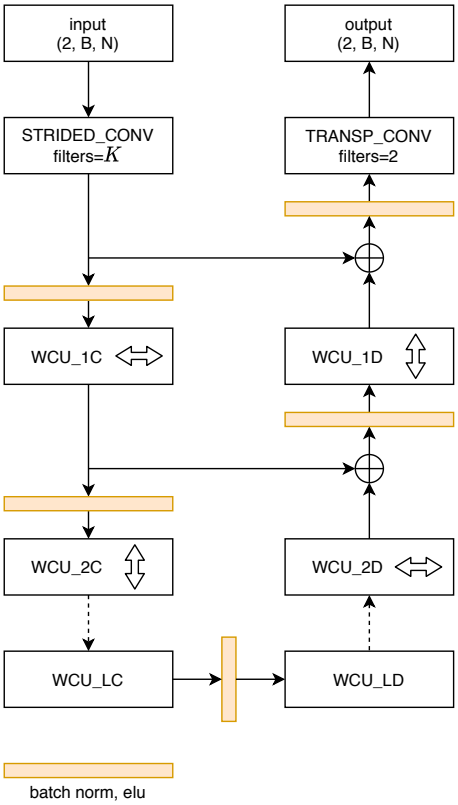
5. Algorytmy do poprawy jakości mowy nie uwzględniają możliwych ubytków słuchu.

6. Funkcje strat służące do trenowania systemów poprawy jakości mowy nie odzwierciedlają dobrze zrozumiałości mowy we wszystkich warunkach.

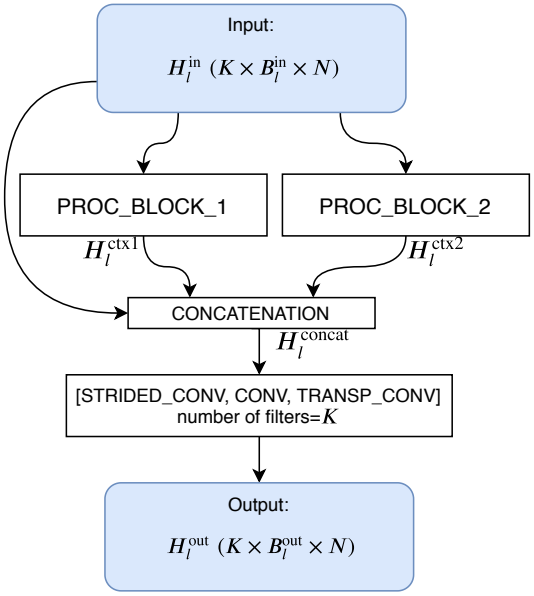
7. Znane sieci neuronowe do predykcji zrozumiałości mowy bazują na reprezentacjach wyznaczonych przez duże sieci neuronowe do automatycznego rozpoznawania mowy lub przez połączenie wielu sieci neuronowych trenowanych do różnych zadań.

4.1.10 Metoda do poprawy jakości mowy bazująca na sieci neuronowej z wide-context units

Motywacja: Architektura U-net sieci neuronowej zaproponowana oryginalnie do analizy obrazów medycznych [40] może być zastosowana do poprawy jakości i zrozumiałości mowy. Oryginalna architektura U-net bazuje na warstwach splotowych i jej pole recepcyjne zależy w dużym stopniu od głębokości sieci. W pracy [A13] pokazałem, że U-net można zastosować do poprawy jakości sygnału mowy. Dodatkowo pokazałem, że po usunięciu operacji max-pooling zmniejsza się pole recepcyjne sieci, a co za tym idzie zmniejsza się średnia wartość wskaźnika SDR określającego jakość wyniku działania sieci neuronowej. Eksperymenty w opisywanej pracy pokazują także, że pogorszenie wyniku związane z usunięciem operacji max-pooling można skompensować poprzez dodanie warstwy rekurencyjnej GRU (gated recurrent unit) pomiędzy enkoderem i dekoderem. **W pracach [A12, A8] zaproponowałem użycie bloku zastępującego warstwy splotowe w U-net, posiadającego następujące właściwości: (1) możliwie mała liczba warstw pomiędzy wejściem a wyjściem, (2) duże pole recepcyjne, (3) niewielka liczba parametrów, (4) małe wymagania pamięciowe.**
Cel: Opracowanie komponentu sieci neuronowej w architekturze U-net w taki sposób, żeby sieć uwzględniała większy kontekst przy niewielkiej liczbie parametrów.
Metoda: Ogólna struktura zaproponowanej sieci neuronowej do poprawy jakości i zrozumiałości jest przedstawiona na rys. 2. Jest to architektura oparta na U-net [40], która zawiera enkoder i dekoder (odpowiednio bloki po lewej i po prawej stronie sieci neuronowej) oraz połączenia skrótowe pomiędzy odpowiednimi blokami enkodera i dekodera na każdym poziomie sieci. W porównaniu do oryginalnej sieci U-net opracowane rozwiązanie posiada bloki WCU (wide-context units) zamiast zwykłych warstw splotowych. WCU sprawiają, że elementy maski wyjściowej zależą od znacznie większego kontekstu w czasie i częstotliwości, niż w przypadku warstw splotowych. Bloki WCU są indeksowane od 1 do L: WCU_1C, ..., WCU_LC dla bloków WCU w enkoderze i WCU_1D, ..., WCU_LD dla odpowiednich WCU w dekoderze, gdzie L jest liczbą poziomów zaproponowanej sieci neuronowej. Zaproponowana strategia projektowania sieci daje możliwość uniknięcia zmniejszania rozdzielczości czasowo-częstotliwościowej dla map cech w warstwach ukrytych. Zależność od szerokiego kontekstu jest uzyskiwana przez zastosowanie warstw rekurencyjnych lub splotów z dylatacjami. Jak pokazano na rys. 2 WCU poszerzające kontekst w czasie i częstotliwości kształtują wzór szachownicy. Dzięki temu, na każdej ścieżce od wejścia do wyjścia sieci neuronowej są na przemian WCU poszerzające kontekst w wymiarze czasu i częstotliwości.



Rysunek 2: Ogólna architektura sieci neuronowej do poprawy jakości mowy. We wszystkich obrazkach rozmiary tensorów są opisane jako: (liczba kanałów/cech, liczba pasm częstotliwościowych, liczba ramek).



Rysunek 3: Ogólna struktura bloku WCU

WCU Ogólna budowa bloku WCU jest przedstawiona na rys. 3. Niech mapa cech podawana na wejście l-tego bloku WCU to $H_l^{\text{in}} \in \mathbb{R}^{K \times B_l^{\text{in}} \times N}$ a wyjście tego bloku to $H_l^{\text{out}} \in \mathbb{R}^{K \times B_l^{\text{out}} \times N}$, gdzie K jest liczbą cech, która jest taka sama na wszystkich poziomach sieci neuronowej, B_l^{in} and B_l^{out} oznaczają liczbę pasm częstotliwościowych, odpowiednio na wejściu i na wyjściu bloku. WCU składa się z dwóch równolegle połączonych bloków przetwarzających (processing blocks), które razem dają mapę cech: $H_l^{\text{ctx1}} \in \mathbb{R}^{J \times B_l^{\text{in}} \times N}$, $H_l^{\text{ctx2}} \in \mathbb{R}^{J \times B_l^{\text{in}} \times N}$, gdzie J jest liczbą cech na wyjściach bloków przetwarzających (PROC_BLOCK_1 i PROC_BLOCK_2), które są sklejone razem z mapą wejściową H_l^{in} wzdłuż pierwszego wymiaru (wymiaru cech), czego wynikiem jest $H_l^{\text{concat}} \in \mathbb{R}^{(K+2J) \times B_l^{\text{in}} \times N}$. Bloki PROC_BLOCK_1 i PROC_BLOCK_2 są odpowiedzialne za kodowanie różnych aspektów szerokiego kontekstu (np. wcześniejszych i późniejszych elementów kontekstu z różnymi współczynnikami dylatacji). Połączenie skrótowe może zapobiec utracie informacji i łagodzić efekt zanikającego gradientu. Mapa H_l^{concat} jest dalej przetwarzana przez warstwę splotową w celu zredukowania liczby cech z $K + 2J$ do K. W eksperymentach testowane były także warianty, w których ta warstwa splotowa ma krok 2 w wymiarze częstotliwości. W tym przypadku w dekoderze stosowana jest transponowana warstwa splotowa, dzięki której możliwe jest odtworzenie wymiarowości. Te warianty nazywane są odpowiednio CONV, STRIDED_CONV i TRANSP_CONV.

Wyniki:

W celu przetestowania zaproponowanej metody, zaplanowałem eksperymenty z wykorzystaniem zbiorów danych WSJ0 i TIMIT zmieszanymi z sygnałami zakłócającymi ze zbiorów Noisex, DCASE i freesound. Ich wyniki są przedstawione w pracy [A8]. Dla eksperymentów z sieciami neuronowymi niezależnymi od SNR- i rodzaju szumu, U-net z WCU opartymi na rekurencjach dała lepsze wyniki niż porównywane sieci: GRN [41], ResBLSTM [42] i U-net WCU ze splotami z dylatacjami. Poprawa została uzyskana dla wskaźników SI-SDR, STOI [43] i PESQ [44]. Podobnie, w przypadku sieci zależnej od SNR i rodzaju szumu, zaproponowana architektura U-net z WCU dała lepsze wyniki niż porównywana architektura GRN. Dodatkowo, WCU bazujące na rekurencjach dała lepszą jakość niż sieć ze splotami z dylatacjami. Ogólnie, uzyskane wyniki sugerują, że

zaproponowana budowa sieci neuronowej jest wysoce efektywna dla zadań poprawy jakości i zrozumiałości mowy. Zebrane dane potwierdzają przypuszczenie, że sieci neuronowe do poprawy jakości mowy powinny efektywnie agregować kontekst (zarówno w czasie jak i częstotliwości) jednocześnie utrzymując lokalizacje tej informacji zarówno w czasie jak i częstotliwości. **Podsumowując, w pracach [A12, A8] zaproponowałem bloki o szerokim kontekście jako rozwiązanie problemu niewystarczającego pola recepcyjnego w sieciach neuronowych typu U-net. Wyniki przeprowadzonych eksperymentów pokazują poprawę wskaźników SI-SDR, STOI i PESQ w porównaniu do znanych z literatury sieci neuronowych do poprawy jakości mowy.**

4.1.11 Metoda Multi-pass

Motywacja:
Sieci neuronowe do poprawy jakości sygnału mowy, oprócz tego, że zmniejszają poziom zakłóceń mogą wprowadzać artefakty. Objawia się to tym, że oprócz braku pełnego odszumienia, mowa w sygnale wyjściowym sieci neuronowej jest zniekształcona. Artefakty wydają się być przyczyną, w wyniku której wielokrotne użycie sieci neuronowej, tzn. dalsze odszumianie częściowo odszumionych nagrań nie daje dodatkowej poprawy jakości mowy. Wytrenowanie sieci neuronowej, w taki sposób, że jej wielokrotne użycie da poprawę wyników, umożliwia dopasowywanie zasobów (operacji arytmetycznych i pamięci) potrzebnych do przetwarzania przez sieć neuronową w trybie testu. Jedną z technik umożliwiających trenowanie sieci neuronowych w taki sposób, żeby dobrze działały dla różnych stopni zaszumienia jest *progressive learning* [45], w której połączonych jest szeregowo wiele sieci neuronowych ze współdzielonymi wagami. Każda z tych podsieci ma swoje wyjście, dla którego wzorcowe wartości opisuje nagranie czyste z szumem o różnych wartościach SNR - od najmniejszego do największego. Metoda multi-pass, którą zaproponowaliśmy ze współautorem w [A9] różni się od *progressive learning* [45, 46, 47] tym, że w każdej iteracji na wejście odpowiedniego bloku jest też dostarczany sygnał wejściowy (zaszumiony). Dodatkowo, w każdej iteracji sieć próbuje usunąć cały szum, a z każdą kolejną iteracją proces ten daje coraz lepsze wyniki.
Metoda:
Wiele znanych sieci neuronowych do poprawy jakości mowy można podzielić na trzy części: warstwy wejściowe, warstwy bazowe i warstwy wyjściowe. W artykułach [A11, A9] zaproponowaliśmy układ, w którym sygnał jest przetwarzany przez warstwy wejściowe. Po tym następuje wielokrotne przetwarzanie przez takie same warstwy bazowe, przy czym przed każdym ponownym przetworzeniem przez warstwy bazowe, dodana jest reprezentacja uzyskana z warstw wejściowych. Po każdej iteracji (ponownym przetwarzaniu) przez warstwy wejściowe, sygnał przetwarzany jest przez warstwy bazowe i wyjściowe. Sieć jest trenowana w taki sposób, żeby średni błąd (wartość funkcji strat) ze wszystkich iteracji był jak najmniejszy.

Ogólny schemat przetwarzania multi-pass do poprawy jakości i zrozumiałości mowy jest przedstawiony na rys. 4. Tensor wejściowy zawierający spektrogram jest przetwarzany przez warstwy wejściowe w celu wyznaczenia jego reprezentacji. Następnie jest ona przetwarzania przez tzw. bazową podsieć L razy. W architekturze występują połączenia skrótowe które dodają reprezentacje cech sygnału wejściowego do wyjściowej mapy cech z każdej iteracji. Wyjście bazowej podsieci z każdej iteracji jest przetwarzane przez warstwy wyjściowe, które rzutują reprezentację na maskę cIRM [48], O_l dla $l = 1, \ldots, L$, która jest

później wykorzystywana do wyznaczenia odszumionego sygnału. Jako warstwy bazowe w pracy [A9] zastosowano następujące podsieci:

1. U-net z warstwami splotowymi z dylatacjami, o architekturze podobnej do WCU [A8],
2. ResBLSTM [42],
3. TranformerNet - sieć wykorzystująca bloki MHA (multi-head attention) [49].

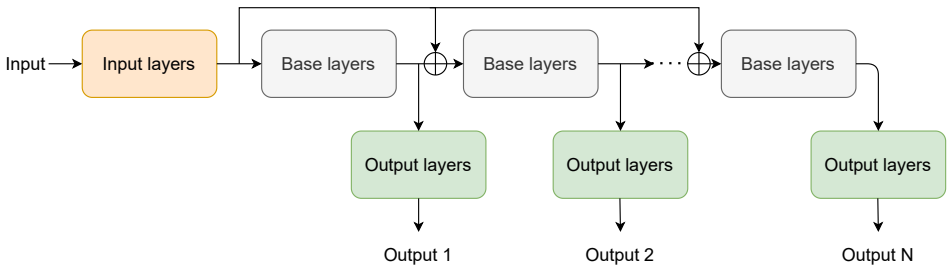
Podczas treningu wyznaczana jest funkcja strat dla każdego wyjścia

$$\mathcal{L}_l = \frac{1}{2NBT} \sum_{n=1}^{N} \sum_{b=1}^{B} \sum_{t=1}^{T} \sum_{p=1}^{2} ((O_l)_{nbtp} - M_{nbtp})^2 \,, \tag{40}$$

gdzie M_{nbtp} jest docelową maską. Indeksy n,b,t,p w tych tensorach oznaczają odpowiednio: p'ty przykład w partii danych (batch), pasmo częstotliwościowe, numer ramki spektrogramu i wybór część rzeczywistą/urojoną.

Całkowita funkcja strat jest wykorzystywana do wyliczenia gradientu, który jest sumą strat ze wszystkich wyjść podzieloną przez liczbę podsieci

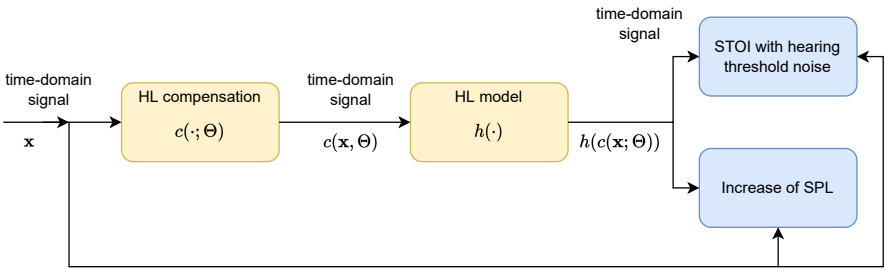
$$\mathcal{L} = \frac{1}{L} \sum_{l=1}^{L} \mathcal{L}_l \,. \tag{41}$$



Rysunek 4: Ogólna architektura multi-pass speech enhancement

Wypadkowa funkcja strat umożliwia wytrenowanie sieci w taki sposób, żeby ta sama podsieć mogła przetwarzać sygnał wielokrotnie. W ciągu drugiej, trzeciej i każdej następnej iteracji sieć może wykorzystać wcześniej uzyskaną reprezentację sygnału oraz reprezentację sygnału wejściowego. Połączenia skrótowe dodawane do reprezentacji sygnału wejściowego do wyjścia z warstw bazowych przed każdą następną iteracją są krytyczną częścią zaproponowanej architektury. Zapobiegają one akumulacji zniekształceń wprowadzanych przez warstwy bazowe.

W architekturze multi-pass, w przeciwieństwie do metody progressive learning, każdy krok poprawy jakości jest wykonywany przy tym samym zbiorze warstw i wag, które są trenowane w celu dostarczenia najlepszego sygnału wyjściowego po każdym przejściu. W trakcie przetwarzania, sieć multi-pass może być wykorzystywana do przeprowadzenia dowolnej liczby iteracji odszumiania pomiędzy 1 and L. W tym przypadku, jedynie ostatnie wyjście powinno być wykorzystane, ponieważ zawiera sygnał maski cIRM poprawiony najwięcej razy. W technice *progressive learning*, każdy etap przetwarzania jest realizowany przez oddzielną sieć która usuwa pewną, predefiniowaną ilość szumu.



Rysunek 5: Ogólny schemat trenowania sieci neuronowej do kompensacji ubytków słuchu

Wyniki:
Zaproponowaną metodę przetestowano dla dobranych przeze mnie zbiorów nagrań mowy WSJ0 i TIMIT oraz szumów ze zbiorów Noisex, DCASE i FreeField. Architektury bazowe, z których wyodrębniono warstwy wejściowe, bazowe i wyjściowe to: Dilated U-net, ResBLSTM i TransformerNet. **Uzyskane w [A9] pokazują, że zastosowanie wielokrotnego przetwarzania przez warstwy bazowe poprawia wskaźniki określające jakość sygnału wyjściowego. Przyrost SI-SDR, STOI i PESQ zmniejsza się jednak z każdą kolejną iteracją. Możliwe, jest zatem dostosowywanie kompromisu pomiędzy jakością a potrzebnymi zasobami, dla wytrenowanej już sieci.**

4.1.12 Metoda DPN (Dynamic processing network)

Motywacja:
W pracy [A2] podjąłem problem projektowania uczących się modeli, które przetwarzają sygnał mowy w celu skompensowania ubytków słuchu. Wśród opisywanych wcześniej w literaturze rozwiązań można znaleźć liniowe moduły wzmacniające poziom sygnału w podpasmach lub stosunkowo złożone sieci neuronowe takie jak np. ConvTASNet [50]. Prace [51, 52, 53] są nastawione na ogólne wykazanie korzyści wynikających z zastosowania systemów uczących się w porównaniu do klasycznego przetwarzania sygnałów w aparatach słuchowych. W pracy [A2] podjąłem się zagadnienia doboru architektury sieci neuronowej na podstawie budowy modelu ubytku słuchu. **Zaproponowałem dwa systemy uczące się: a) różniczkowalny procesor dynamiki, który można poddać trenowaniu i b) sieć splotową ogólnego przeznaczenia. Ponadto systematycznie określiłem wpływ decyzji podczas projektowania sieci neuronowej na obiektywne metryki zrozumiałości mowy.**
Metoda:
Ogólna struktura zaproponowanego przeze mnie systemu do kompensacji uszkodzenia słuchu jest zaprezentowana na rys. 5. Odnosi się ona do scenariusza słuchania jednousznego, w którym sygnał mowy jest przetwarzany przez aparat słuchowy, a następnie zostaje odtworzony do ucha użytkownika, który ma ubytek słuchu. Tak więc, sygnał wejściowy w dziedzinie czasu jest na początku przetworzony przez sieć neuronową do kompensacji ubytku słuchu. Następnie, przetworzony sygnał jest zdegradowany przez model uszkodzonego słuchu, który jest różniczkowalną wersją modelu przedstawionego w pracy [54].

Przekształcenie sygnału wejściowego przez sieć do kompensacji, a następnie model uszkodzonego słuchu może być zapisane w następujący sposób

$$\mathbf{y} = h(c(\mathbf{x};\Theta)) = (h \circ c)(\mathbf{x};\Theta)\ , \tag{42}$$

gdzie $\mathbf{x}$ jest wektorem zawierającym próbki sygnału wejściowego w dziedzinie czasu, h jest

funkcją symulującą uszkodzony słuch, a c jest funkcją (siecią neuronową z parametrami Θ) kompensującą ubytki słuchu.

Funkcja strat STOI jest to obiektywna metryka zaprojektowana tak, żeby odzwierciedlała zrozumiałość mowy. Bazuje ona na korelacji obwiedni porównywanych sygnałów (przetworzonego i referencyjnego) w pasmach tercjowych, i w związku z tym, nie zależy od poziomu sygnałów. Bez żadnych dodatkowych kryteriów, maksymalizacja STOI doprowadzi do zwiększenia poziomu sygnału mowy. W związku z tym w pracy [A2] zastosowałem funkcję strat, która zawiera STOI i dodatkowy składnik, który zapobiega zwiększeniu poziomu sygnału powyżej poziom określany przez czysty (wzorcowy) sygnał mowy. Dlatego parametry sieci kompensującej c są optymalizowane tak, żeby zminimalizować funkcję strat

$$\mathcal{L} = -\alpha \mathrm{STOI}(\mathbf{x}, \mathbf{y}) + (1-\alpha)\max(0, L(\mathbf{y}) - L(\mathbf{x})) \tag{43}$$

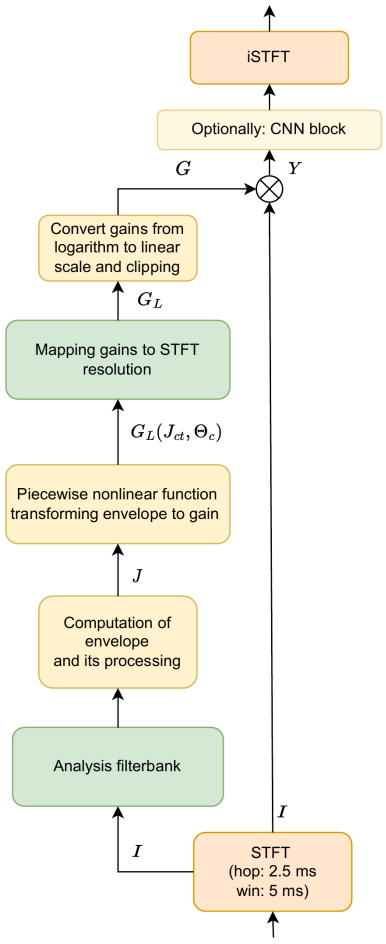
względem parametrów Θ, gdzie $\mathrm{STOI}(\cdot,\cdot)$ jest użytą implementacją STOI, $L(\cdot)$ jest poziomem danego sygnału (np. dla N-wymiarowego wektora $\mathbf{x}$ poziom jest obliczany jako $L(\mathbf{x}) = 10\log_{10}(\|\mathbf{x}\|^2)/N$), a $\alpha \in [0,1]$ jest wagą umożliwiającą ustawienie względnego wpływu tych dwóch składników. Warto dodać, że STOI nie uwzględnia tego, czy sygnał jest powyżej progu słyszenia czy nie. Dlatego zaproponowałem dodanie szumu progowego (zaraz po modelu uszkodzonego słuchu) przed obliczeniem STOI. Podobne rozwiązanie zostało użyte w znanej metryce HASPI [55].

Różniczkowalny model uszkodzonego słuchu Model uszkodzonego słuchu wykorzystany do trenowania bazuje na modelu z pracy [54][1]. Składa się on z dwóch głównych części tj. rozmycia widmowego i wyrównywania głośności, które wykonuje ekspansję dynamiki [56].

Architektury sieci neuronowych do kompensacji ubytków słuchu W pracy [A2] zaproponowałem architektury do kompensacji ubytków słuchu (zob. $c(\cdot;\Theta)$ na rys. 5). Istnieje kilka pożądanych cech, które powinna spełniać sieć kompensująca. Po pierwsze, w praktycznych zastosowaniach tego typu sieć powinna być przyczynowa, a latencja wprowadzana przez tę sieć powinna być mniejsza niż 10 ms [57]. W pracy [A2] te wymagania zostały spełnione poprzez zastosowanie przyczynowych warstw splotowych oraz krótkoterminowej transformaty Fouriera (STFT) z oknem 5 ms. Dodatkowo, obliczeniochłonność sieci powinna być ograniczona, co jest związane z możliwościami obliczeniowymi procesora oraz poborem mocy w potencjalnym aparacie słuchowym. W zaproponowanej sieci DPN, liczba operacji arytmetycznych jest stosunkowo mała w porównaniu z generycznymi sieciami neuronowymi. Na przykład, zaproponowana przeze mnie sieć DPN, w swojej podstawowej postaci wykonuje 12874 operacji arytmetycznych na jedną ramkę spektrogramu, podczas gdy jeden poziom warstwy splotowej 2D (w eksperymentach stosowałem - 48 filtrów z jądrem 7×3 , przyjmującej na wejściu 48 cech dla każdej jednostki czasowo-częstotliwościowej), wykonuje ponad 12 milionów operacji na ramkę spektrogramu.

Dynamic processing network W pracy [A2] zaproponowałem sieć DPN (dynamic processing network), która jest różniczkowalnym modelem obliczeniowym, który może

[1] https://github.com/claritychallenge



Rysunek 6: Struktura zaproponowanej sieci DPN

wykonywać przetwarzanie dynamiki sygnału w pasmach częstotliwościowych. Przy przetwarzaniu dynamiki sygnał jest wzmacniany albo tłumiony przy użyciu zmiennego w czasie wzmocnienia, które zależy od wartości skutecznej wyznaczonej z sygnału lokalnie w czasie. Parametry sieci DPN mogą być ustawione w oparciu o tablice wzmocnień, które są wyznaczone np. za pomocą procedury Camfit [58], w której konfiguracja wielopasmowego kompresora jest otrzymana w oparciu o model percepcji głośności. Następnie, parametry kompresora mogą być dodatkowo optymalizowane, tak żeby zmaksymalizować przewidywaną zrozumiałość mowy na podstawie (43). W podstawowej postaci DPN, jej parametry są inicjalizowane w oparciu o Camfit przy użyciu procedury opisanej w następnym paragrafie, po czym parametry są dostrajane.

Ogólna architektura DPN jest przedstawiona na rys. 6. Na początku jest wyliczane STFT sygnału wejściowego do DPN. Następnie, przetwarzanie jest rozdzielone na dwie gałęzie. Lewa gałąź na rys. 6 oblicza wzmocnienia na podstawie obwiedni sygnałów w pasmach częstotliwościowych. Wzmocnienia są później użyte do przemnożenia sygnałów z prawej gałęzi. Wzmocnienia i obwiednie są obliczane w pasmach częstotliwościowych uzyskanych przez bank filtrów analizujących, a następnie wzmocnienia są z powrotem przeliczone na pasma częstotliwości STFT przy użyciu zestawu filtrów syntezujących. Na końcu zastosowana jest odwrotna STFT do przekształcenia reprezentacji sygnału do dziedziny czasu. Filtry analizy i syntezy są wykorzystywane do zredukowania liczby pasm częstotliwościowych, a co za tym idzie, parametrów do trenowania, ze względu na to, że w każdym paśmie częstotliwościowym wyuczana jest osobna funkcja przekształcająca obwiednie na wzmocnienie.

Wzmocnienie jest funkcją łamaną wartości skutecznej sygnału, zależy od parametrów, które mogą być dostrojone przy pomocy procedury uczenia. W tej podstawowej postaci jedynie parametry funkcji łamanych są dostrajane. Rozważałem także rozszerzenia, w których parametry filtrów przetwarzających obwiednie i filtrów analizy oraz syntezy (rys. 6) były także optymalizowane.

Blok STFT na rys. 6 oblicza tensor $I \in \mathbb{R}^{B \times T_{\text{comp}} \times 2}$ z sygnału wejściowego $\mathbf{x}$, gdzie $B = N_{\text{FFT}}/2 + 1$, T_{comp} jest liczbą ramek, ostatni wymiar odnosi się do części rzeczywistych i urojonych spektrogramu. Przed obliczeniem wzmocnienia, tensor STFT jest decymowany do C pasm częstotliwościowych. Wartość skuteczna w paśmie częstotliwościowym c dla ramki czasowej t jest obliczana według wzoru

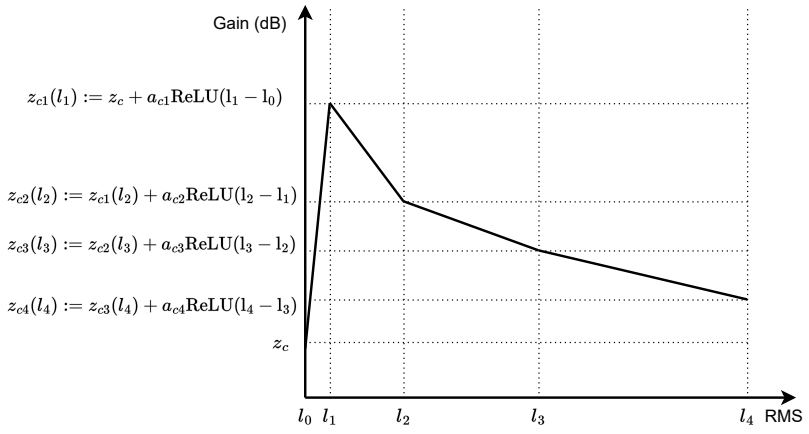
$$J_{ct} = \sqrt{\frac{2}{N_{\text{FFT}}} \sum_{b=1}^{B} F_{cb}(I_{bt1}^2 + I_{bt2}^2)} \,, \tag{44}$$

gdzie F_{cb} zawiera odpowiedź filtru analizy c dla każdego pasma częstotliwościowego b. Charakterystyka F_{cb} jest równa jedności dla częstotliwości STFT b pomiędzy częstotliwościami odcięcia c-tego filtru i zero w przeciwnym przypadku. Częstotliwości odcięcia dla pasm $c = 1, \ldots, C$ są rozmieszczone logarytmicznie. Obwiednie mogą być opcjonalnie przetworzone przez filtry wygładzające opisane w [A2].

Wartości wzmocnienia w każdym kanale częstotliwościowym $c = 1, \ldots, C$ są wyznaczane w oparciu o funkcję łamaną

$$G_L(J_{ct}, \Theta_c) = \sum_{i=1}^{H} a_{ci} \text{ReLU}(J_{ct} - l_{i-1}) + z_c \,, \tag{45}$$

gdzie $\text{ReLU}(x) = \max(x, 0)$. Wzór (45) reprezentuje zależność pomiędzy wzmocnieniem a wartością skuteczną (w skali liniowej) sygnału paśmie częstotliwościowym i jego wzmoc-



Rysunek 7: Schematyczna ilustracja funkcji łamanej wykorzystanej w sieci do przetwarzania dynamiki (DPN)

nieniem (w skali decybelowej). Nachylenia segmentów funkcji łamanej zależą od trenowanych parametrów $\Theta_c = (a_{c1}, \ldots, a_{cH}, z_c)$, a $l_0, \ldots, l_H$ są ustalonymi wartościami skutecznymi które są granicami pomiędzy następującymi po sobie segmentami funkcji (45). W eksperymentach w [A2] $l_0, \ldots, l_H$ są wartościami skutecznymi, które odnoszą się do wartości poziomów -10 to 111 dB SPL z krokiem 1 dB. Funkcja łamana z równania (45) jest pokazana na rys. 7.

Wzmocnienie (w dB) dla pasma częstotliwościowego STFT b i ramki czasowej t jest obliczane przez $G_{LJ_{ct}}$ dla pasma częstotliwościowego c, które zawiera b w swoim paśmie przepustowym:

$$G_{Lbt} = \sum_{c=1}^{C} F_{cb} G_L(J_{ct}; \Theta_c) \ , \tag{46}$$

i przekształcone do skali liniowej przy użyciu formuły

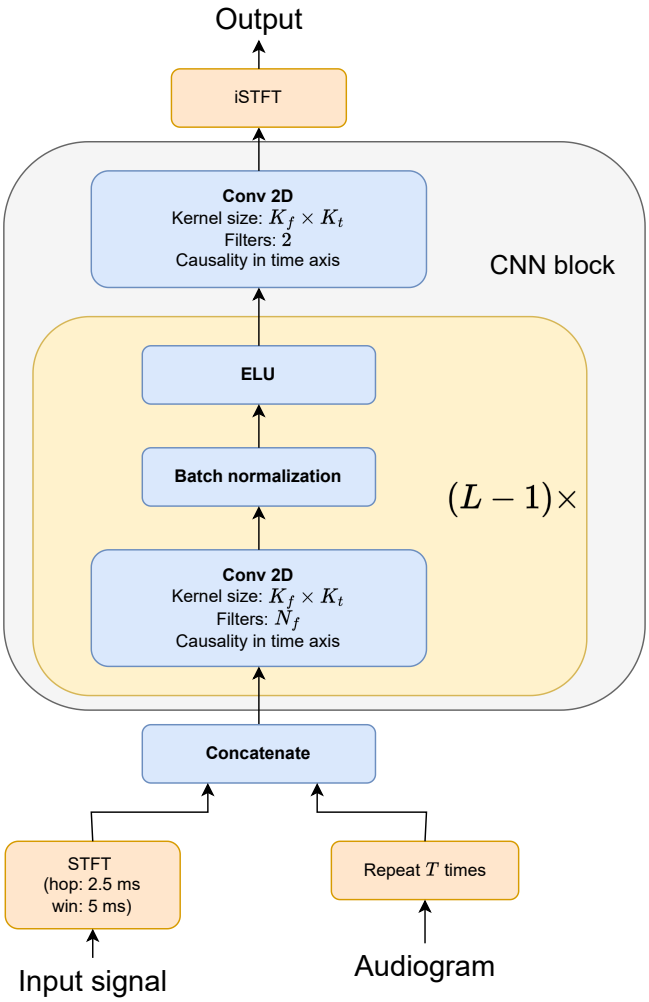
$$G_{bt} = 10^{G_{Lbt}/20} \ . \tag{47}$$

Następnie wzmocnienia G_{bt} są obcinane, w celu ograniczenia maksymalnego wzmocnienia do 10000 (80 dB). Ostatecznie sygnał wejściowy w dziedzinie STFT jest przemnożony przez odpowiadające wzmocnienie G_{bt} jako

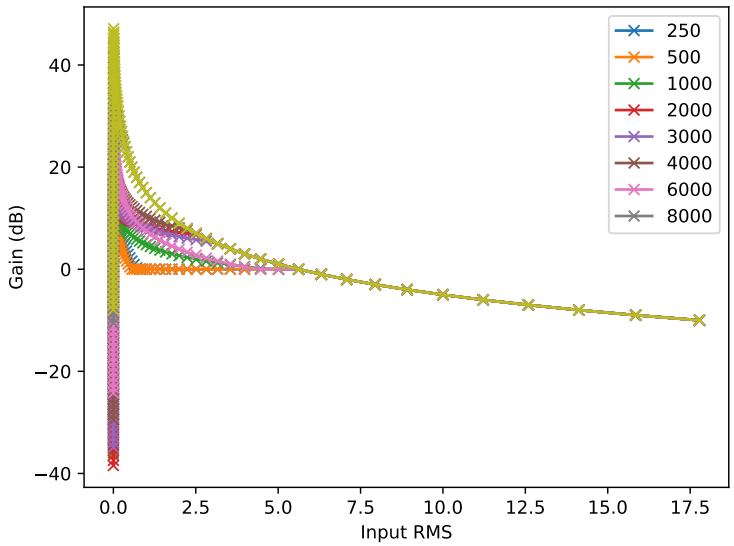
$$Y_{btk} = I_{btk} G_{bt} \ , \tag{48}$$

dla $b = 1, \ldots, B$, $t = 1, \ldots, T_{\text{comp}}$ i $k = 1, 2$, gdzie $k = 1$ odnosi się do części rzeczywistej a $k = 2$ do części urojonej spektrogramów Y i I. Tensor Y może być opcjonalnie przetwarzany przez blok CNN (oznaczony na rys. 8).

Inicjalizacja DPN na podstawie tabeli wzmocnień Camfit Parametry DPN przedstawione w poprzedniej sekcji mogą być wyznaczone na podstawie procedury Camfit [58]. Ze względu na to, że DPN jest modelem różniczkowalnym, zainicjalizowane parametry mogą być dodatkowo dostrojone ze względu na funkcję strat (43). Procedura Camfit umożliwia określenie charakterystyki kompresji dynamiki w aparatach słuchowych w oparciu jedynie o audiogram. Parametry kompresora są wyprowadzone tak, żeby odzyskać charakterystykę percepcji głośności dla słuchaczy z uszkodzonym słuchem. Wynik



Rysunek 8: Architektura sieci DNN do kompensacji ubytków słuchu



Rysunek 9: Tabele wzmocnień uzyskane przy użyciu procedury Camfit (1 RMS = 85 dB SPL)

procedury Camfit jest często w postaci tabeli wzmocnień L_{cj}, która wskazuje wzmocnienie dla każdego pasma częstotliwościowego c i wartości skutecznej sygnału wejściowego z indeksem j. Przykładowe krzywe narysowane na podstawie tabeli wzmocnień są pokazane na rys. 9. Bieżąca sekcja pokazuje jak przekształcić tabele wzmocnień na parametry $(\Theta_1, \ldots, \Theta_C)$ zaproponowanego modelu DPN. W celu wyrażenia tabeli wzmocnień jako kombinacji przesuwanych funkcji ReLU (zob. równanie (45)) $z_c = L_{c0}$ i parametry a_{ci} mogą być wyznaczone za pomocą wzoru

$$[a_{c1} \ \ldots \ a_{cH}]^{\mathrm{T}} = \mathbf{A}^{-1} \begin{bmatrix} L_{c1} - z_c \\ \vdots \\ L_{cH} - z_c \end{bmatrix} , \tag{49}$$

gdzie

$$A_{ij} = \mathrm{ReLU}(l_i - l_{j-1}) , \tag{50}$$

dla $i = 1, \ldots, H$ i $j = 1, \ldots, H$. Wartości wejściowe w J_{ct} są skalowane tak, że reprezentują wartość skuteczną odnoszącą się do poziomu ciśnienia akustycznego.

Wyniki: Zarówno dla modeli niezależnych i zależnych od słuchacza wyniki przeprowadzonych przeze mnie eksperymentów (zawarte w pracy [A2]) pokazują, że dostrajanie sieci DPN z wagami zainicjalizowanymi za pomocą procedury Camfit prowadzi do poprawy przewidzianej zrozumiałości mowy w porównaniu do ustalonych parametrów Camfit. Najlepsze wyniki dla modelu niezależnego od słuchacza uzyskałem dla kombinacji DPN i CNN. Eksperymenty, w których techniki kompensacji były zastosowane do głównych komponentów modelu uszkodzonego słuchu pokazały, że sieć splotowa może częściowo skompensować rozmycie widmowe. Pokazałem także, że zarówno sieć splotowa i wyuczona DPN mogą częściowo skompensować wyrównanie głośności. Wydajna obliczeniowo DPN daje lepsze wyniki niż CNN z 6 warstwami splotowymi.

Należy podkreślić, że model DPN w swojej podstawowej postaci 12874 operacji arytmetycznych na jedną ramkę spektrogramu, podczas gdy jedna warstwa splotowa sieci

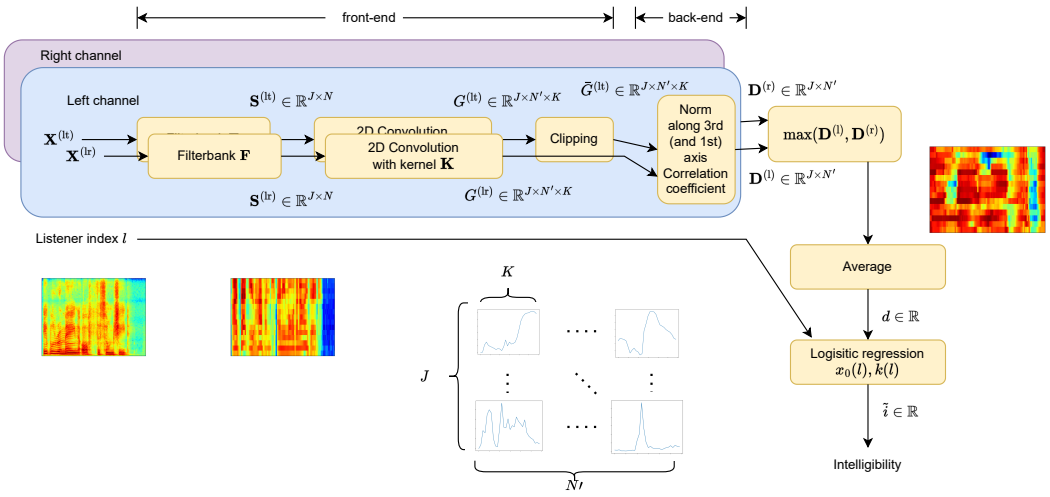
neuronowej zastosowanej w przeprowadzonych eksperymentach — 48 filtrów z maską o wymiarach 7×3, przyjmującą na wejściu mapę 48 cech dla każdej jednostki czasowo-częstotliwościowej), wykonuje ponad 12 milionów operacji arytmetycznych na ramkę spektrogramu. **Podsumowując, w pracy [A2] opracowałem architekturę DPN do kompensacji ubytków słuchu, która daje lepsze wartości wskaźnika HASPI niż porównywane sieci splotowe. Zaproponowana architektura charakteryzuje się niskim kosztem obliczeniowym**

4.1.13 Metoda FT-GESTOI

Motywacja:
Uczenie maszynowe systemów do poprawy zrozumiałości mowy musi odbywać się z użyciem odpowiednio dobranej funkcji strat. W przypadku sieci neuronowych taka funkcja strat, jeśli jest funkcją różniczkowalną, daje możliwość zastosowania algorytmu propagacji wstecznej do wyznaczenia pochodnych cząstkowych parametrów modelu. **W pierwszym konkursie Clarity dotyczącym predykcji zrozumiałości mowy [59], najlepiej działające metody korzystały z wielu osobno trenowanych sieci neuronowych (np. [60]) lub obliczeniochłonnych głębokich sieci neuronowych, które są trenowane w oparciu o dodatkowy zbiór danych [61, 62]. Nie jest do końca jasne jak takie złożone sieci będą działały na nowym zbiorze nagrań. Nieprzewidywalne zachowanie sieci neuronowych może być zademonstrowane przy użyciu *adversarial examples* (np. [63]).**

W pracy [A1] opisałem opracowaną przeze mnie metodę FT-GESTOI (fine-tuned generalized ESTOI), która bazuje na ESTOI [2], przy czym filtry zarówno we front- i back-endzie są dostrajane przy użyciu zbioru danych. Metoda odnosi się do ograniczenia, w którym wszystkie lokalne podobieństwa są traktowane z jednakową wagą. Zróżnicowanie wag osiągnąłem dzięki wprowadzeniu gałęzi uwagi (TA). Ograniczenie związane z niezależnością metryki podobieństwa od charakterystyki mowy rozwiązałem przez wyznaczenie odporności sygnału referencyjnego na zniekształcenia. Do tego celu użyłem sieci neuronowych do automatycznego rozpoznawania mowy. W zaproponowanej metodzie zastosowałem model uszkodzonego słuchu, który nie występuje w ESTOI. Metoda FT-GESTOI jest różniczkowalna; dzięki czemu może być potencjalnie zastosowana jako funkcja strat do trenowania algorytmów do poprawy zrozumiałości mowy.

Podstawowy wariant FT-GESTOI W podstawowym wariancie FT-GESTOI, ESTOI jest rozszerzone o model uszkodzonego słuchu MSBG [54] i uwzględnia kombinację lewego i prawego kanału w oparciu o zasadę lepszego ucha. Filtry do analizy spektralnej i do przetwarzania wyodrębnionych obwiedni mogą być zoptymalizowane. Dzięki tym modyfikacjom, wynikowy model predykcji zrozumiałości mowy ma stosunkowo nieduże wymagania obliczeniowe i jest interpretowalny jak ESTOI, przy czym wykorzystuje informacje zawarte w zbiorze treningowym. ESTOI jest intruzywną metryką zrozumiałości mowy, która przyjmuje na wejście spektrogramy magnitudowe sygnałów referencyjnych i testowych, $\mathbf{X}^{(\mathrm{t})} \in \mathbb{R}^{B \times N}$ i $\mathbf{X}^{(\mathrm{r})} \in \mathbb{R}^{B \times N}$, gdzie B oznacza liczbę częstotliwości STFT a N jest liczbą ramek spektrogramu. Ta metryka wyznacza korelację 300-ms segmentów czasowo-częstotliwościowych wyodrębnionych z pasm tercjowych. Chociaż oryginalne ESTOI jest sformułowane dla jednego kanału, może być ono w prosty sposób rozszerzone do dwóch kanałów poprzez użycie metody lepszego ucha, tj. przed wyznaczeniem ostatecznej predykcji zrozumiałości, pośrednia zrozumiałość dla regionów czasowo-częstotliwościowych



Rysunek 10: Architektura zaproponowanej metody do predykcji zrozumiałości mowy. Pierwsza litera w etykiecie zmiennej ((lt), (lr), (rt), (rr)) oznacza kanał (lewy (l) lub prawy (r)), druga litera oznacza sygnał referencyjny lub testowy

jest wyznaczona odpowiednio dla każdego z kanałów. Te pośrednie zrozumiałości z obu kanałów mogą być połączone przez użycie lokalnej wartości zrozumiałości z tego ucha, dla którego była ona większa. Ta operacja jest wykonywana dla każdego regionu czasowo-częstotliwościowego.

Schemat przetwarzania w metodzie FT-GESTOI pokazany jest na rys. 10. Na wejście systemu podane są spektrogramy referencyjny i testowy dla obu kanałów (lewego i prawego). Dla lewego kanału spektrogram testowy i referencyjny są oznaczone przez: $\mathbf{X}^{(\mathrm{lt})}$, $\mathbf{X}^{(\mathrm{lr})}$, a dla prawego kanału: $\mathbf{X}^{(\mathrm{rt})}$ i $\mathbf{X}^{(\mathrm{rr})}$. Sygnały wejściowe dla kanałów lewego i prawego są przetwarzane równolegle przez te same bloki obliczeniowe. To przetwarzanie daje w wyniku macierze $\mathbf{D}^{(\mathrm{l})}$ i $\mathbf{D}^{(\mathrm{r})}$, które zawierają podobieństwa pomiędzy testowym i referencyjnym sygnałem, odpowiednio dla lewego i prawego ucha. Są to tzw. zrozumiałości pośrednie. Te macierze są łączone metodą lepszego ucha, uśrednione i przekształcone przez funkcję logistyczną, której parametry mogą być zależne od indeksu l. Wynikowa liczba jest predykcją zrozumiałości mowy.

Kroki przetwarzania w modelu FT-GESTOI są następujące. Po pierwsze, wykonywana jest transformacja STFT sygnałów wejściowych do spektrogramów tercjowych przez mnożenie przez macierz $\mathrm{softplus}(\mathbf{F}) \in \mathbb{R}^{J\times B}$, gdzie J jest liczbą filtrów, a funkcja

$$\mathrm{softplus}(x) = \log(\exp(x) + 1) \tag{51}$$

jest zastosowana indywidualnie do każdego elementu w celu zapewnienia nieujemności macierzy wynikowej. Elementy F_{jb} macierzy $\mathbf{F}$ są inicjalizowane w taki sposób, że po przekształceniu przez softmax każdy wiersz jest charakterystyką filtru tercjowego:

$$F_{jb} = \begin{cases} 0.55 & \text{for } (j,b) : f_l(j) \le b \le f_u(j) \\ -6 & \text{w innym przypadku} \end{cases}, \tag{52}$$

gdzie f_l i f_u są odpowiednio dolnymi i górnymi częstotliwościami odcięcia filtrów tercjowych zaczynających się od wybranej częstotliwości (w przypadku pracy [A1] 150 Hz). Wartość 0.55 w równaniu (52) po przekształceniu przez softmax() jest równa jeden, a -6 jest odwzorowane w pewną wartość bliską zeru. W przypadku FT-GESTOI, macierz

$\mathbf{F}$ może być zoptymalizowana przy użyciu zbioru treningowego. Przekształcenie z magnitudowego STFT do wartości skutecznych wyjść filtrów zdefiniowanych przez $\mathbf{F}$ jest przeprowadzane jako

$$\mathbf{S}^{(\cdot \mathrm{t})} = \sqrt{\mathrm{softplus}(\mathbf{F})(\mathbf{X}^{(\cdot \mathrm{t})})^2}, \tag{53}$$

$$\mathbf{S}^{(\cdot \mathrm{r})} = \sqrt{\mathrm{softplus}(\mathbf{F})(\mathbf{X}^{(\cdot \mathrm{r})})^2}, \tag{54}$$

gdzie $\cdot^2$ i $\sqrt{\cdot}$ są odpowiednio funkcjami potęgowania i pierwiastkowania stosowanymi indywidualnie do elementów macierzy. Ta operacja jest stosowana oddzielnie do obu kanałów, tj., l i r musi być podstawiona za kropkę w górnych indeksach $\mathbf{S}$ i $\mathbf{X}$. Taka notacja jest używana w pozostałej części opisu metody.

W STOI/ESTOI następnym krokiem po transformacji spektrogramu wejściowego w spektrogram tercjowy jest zastosowanie okna kroczącego o długości K do każdego pasma częstotliwościowego j. Wynikiem tej operacji jest $G^{(\mathrm{t})} \in R^{J \times N' \times K}$ i $G^{(\mathrm{r})} \in R^{J \times N' \times K}$, gdzie N' jest liczbą pozycji okna kroczącego w spektrogramie. W standardowych ustawieniach ESTOI jeden segment (długość okna kroczącego) odnosi się do około 300 ms. Jest to długość wybrana także w metryce STOI. Taki czas trwania obejmuje większość sylab. W języku angielskim średnia długość jednej sylaby to ok. 200 ms, a zakres czasu trwania sylab jest pomiędzy 125 i 400 ms [64]. Elementy tensorów $G^{(\cdot \mathrm{t})}$ i $G^{(\cdot \mathrm{r})}$ są wyznaczone przez

$$G^{(\cdot \mathrm{t})}_{jni} = S^{(\cdot \mathrm{t})}_{j,n-K/2+i} \qquad \text{for } i = 1, \ldots, K \tag{55}$$

i

$$G^{(\cdot \mathrm{r})}_{jni} = S^{(\cdot \mathrm{r})}_{j,n-K/2+i} \qquad \text{for } i = 1, \ldots, K\ . \tag{56}$$

Ten krok może być określony jako przypadek szczególny dwuwymiarowej warstwy splotowej i może być zaimplementowany jako

$$G^{(\cdot \mathrm{t})}_{jni} = \sum_{l=0}^{K-1} S^{(\cdot \mathrm{t})}_{j,n-K/2+l} K_{li}\ , \tag{57}$$

$$G^{(\cdot \mathrm{r})}_{jni} = \sum_{l=0}^{K-1} S^{(\cdot \mathrm{r})}_{j,n-K/2+l} K_{li}\ , \tag{58}$$

gdzie K_{li} są elementami macierzy $\mathbf{K} = \mathbf{I} \in \mathbb{R}^{K \times K}$, która reprezentuje K jąder o rozmiarze $1 \times K$. Kiedy macierz $\mathbf{K}$ jest zainicjalizowana jako macierz jednostkowa, dla każdej pozycji n jądra, wartości w tensorach $G^{(\cdot \mathrm{t})}$ i $G^{(\cdot \mathrm{r})}$, wzdłuż wymiaru kanału (numeru filtru), występuje K kolejnych próbek obwiedni sygnału. Kiedy macierz $\mathbf{K}$ jest dostrojona, to umożliwia bardziej ogólną operację, tj. daje w wyniku K różnych sum ważonych próbek objętych przez okno kroczące (w zakresie jądra). Na przykład, możliwe jest, że po dostrojeniu nie tylko k-ta próbka z okna kroczącego jest wprowadzona do G, ale również różnica pomiędzy tą próbką i jej otoczeniem w czasie.

Następnie przeprowadzane jest obcinanie $G^{(\cdot \mathrm{t})}$ w celu zredukowania krótko-terminowych zniekształceń (większych niż 15 dB). Wyliczane są normy dla wektorów $\mathbf{g}^{(\cdot \mathrm{r})}_{jn} = [G^{(\cdot \mathrm{r})}_{jn1}, \ldots, G^{(\cdot \mathrm{r})}_{jnK}]^{\mathrm{T}}$ i $\mathbf{g}^{(\cdot \mathrm{t})}_{jn} = [G^{(\cdot \mathrm{t})}_{jn1}, \ldots, G^{(\cdot \mathrm{t})}_{jnK}]^{\mathrm{T}}$, gdzie operacja obcinania jest wyrażona jako

$$\bar{G}^{(\cdot \mathrm{r})}_{jnk} = \min \left\{ \frac{\|\mathbf{g}^{(\cdot \mathrm{r})}_{jn}\|}{\|\mathbf{g}^{(\cdot \mathrm{t})}_{jn}\|} G^{(\cdot \mathrm{t})}_{jnk}, (1+c) G^{(\cdot \mathrm{r})}_{jnk} \right\}\ , \tag{59}$$

przy czym $c = 10^{15/20}$. Dotychczas opisane kroki FT-GESTOI opisują tzw. front-end.

Na tym etapie, sygnały testowe i referencyjne są reprezentowane odpowiednio przez tensory $\bar{G}^{(\cdot \mathrm{t})}$ i $G^{(\cdot \mathrm{r})}$. Następnym krokiem jest porównanie tych reprezentacji indywidualnie dla lewego i prawego kanału, czego wynikiem są dwie macierze o wymiarach $J \times N'$: $\mathbf{D}^{(\mathrm{l})}$ i $\mathbf{D}^{(\mathrm{r})}$ zawierające pośrednie składowe zrozumiałości. To porównanie $\bar{G}^{(\cdot \mathrm{t})}$ i $G^{(\cdot \mathrm{r})}$ nazywane jest back-endem i jest przeprowadzane w następujących krokach. Po pierwsze, tensory są standaryzowane wzdłuż trzeciego wymiaru, tj. dla każdej pary j, n średni wektor z K elementów wzdłuż trzeciego wymiaru jest odjęty, a wynik jest wyskalowany tak, żeby jego norma była równa jedności. Następnie podobna standaryzacja jest przeprowadzana wzdłuż pierwszego wymiaru (j). Po standaryzacji, dla każdego pasma częstotliwościowego j i dla każdego kroku czasowego n, obliczany jest iloczyn skalarny wektorów $[\bar{G}^{(\mathrm{t})}_{jn1}, \ldots, \bar{G}^{(\mathrm{t})}_{jnK}]$ i $[G^{(\mathrm{r})}_{jn1}, \ldots, G^{(\mathrm{r})}_{jnK}]$. To tworzy macierze $\mathbf{D}^{(\mathrm{l})} \in \mathbb{R}^{J \times N'}$ i $\mathbf{D}^{(\mathrm{r})} \in \mathbb{R}^{J \times N'}$ odpowiednio dla lewego i prawego ucha. Te macierze zawierają pośrednie składowe zrozumiałości mowy, zlokalizowane w czasie i częstotliwości. Macierze $\mathbf{D}^{(\mathrm{l})}$ i $\mathbf{D}^{(\mathrm{r})}$ są łączone na podstawie zasady lepszego ucha indywidualnie dla każdego obszaru czasowo-częstotliwościowego

$$D_{jn} = \max\{D^{(\mathrm{l})}_{jn}, D^{(\mathrm{r})}_{jn}\} \; . \tag{60}$$

W celu połączenia składowych pośredniej zrozumiałości dla każdego kroku czasowego d_n jest wyliczane jako

$$d_n = \sum_{j=1}^{J} D_{jn} \tag{61}$$

i ostatecznie te wartości są uśredniane po czasie:

$$d = \frac{1}{N'} \sum_{n=1}^{N'} d_n \; . \tag{62}$$

Wartość pośredniej zrozumiałości mowy d jest przekształcana przy użyciu funkcji logistycznej w celu obliczenia predykcji zrozumiałości mowy

$$\tilde{i} = \frac{1}{1 + \exp(-k(d - x_0))} \; , \tag{63}$$

gdzie x_0 i k są niezależne od słuchacza. W przypadku zależnym od słuchacza, $x_0(l)$ i $k(l)$ są optymalizowane indywidualnie dla każdego słuchacza.

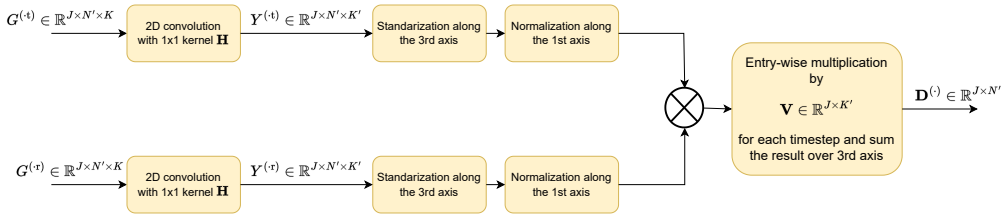
Ogólnie FT-GESTOI jest funkcją g spektrogramów testowego i referencyjnego oraz indeksu słuchacza, parametryzowaną przez $\Theta = (\mathbf{F}, \mathbf{K}, x_0(1), \ldots, x_0(L), k(1), \ldots, k(L))$, gdzie L jest liczbą słuchaczy, dla których estymowana jest funkcja g

$$g : (\mathbf{X}^{(\mathrm{t})}, \mathbf{X}^{(\mathrm{r})}, l; \Theta) \to \mathbb{R} \; . \tag{64}$$

Dla zbioru danych $\{(\mathbf{X}^{(\mathrm{lt})}_m, \mathbf{X}^{(\mathrm{lr})}_m, \mathbf{X}^{(\mathrm{rt})}_m, \mathbf{X}^{(\mathrm{rr})}_m, l_m, i_m)\}_{m=1}^{M}$ z przykładami zawierającymi wejścia i wyjścia FT-GESTOI, parametry Θ mogą być adaptowane, żeby zminimalizować funkcją strat określoną jako

$$\min_{\Theta} \qquad \mathcal{L} = \frac{1}{M} \sum_{m=1}^{M} (g(\mathbf{X}^{(\mathrm{lt})}_m, \mathbf{X}^{(\mathrm{lr})}_m, \mathbf{X}^{(\mathrm{rt})}_m, \mathbf{X}^{(\mathrm{rr})}_m, l_m; \Theta) - i^m)^2 \; . \tag{65}$$

Model FT-GESTOI jest trenowany w następujący sposób:



Rysunek 11: Uogólniony back-end z uczącymi się filtrami modulacyjnymi

1. Ustawić parametry FT-GESTOI tak, żeby model wykonywał takie obliczenia jak ESTOI.

2. Obliczyć FT-GESTOI dla każdego przykładu treningowego.

3. Dopasować krzywą logistyczną niezależnie dla każdego słuchacza (w przypadku zależnym od słuchacza) lub korzystając z danych dla wszystkich słuchaczy (w przypadku niezależnym od mówcy), przy użyciu metody Levenberg-Marquardt [65],

4. Dostroić parametry FT-GESTOI, przy użyciu całego zbioru danych dla ustalonych parametrów krzywych logistycznych.

Back-end z uczącymi się filtrami modulacyjnymi (LMFB) W swoim podstawowym wariancie, model FT-GESTOI opisany w poprzedniej sekcji, porównuje segmenty obwiedni czasowych w pasmach częstotliwościowych przy użyciu back-endu. Te segmenty obwiedni czasowej mogą być przekształcone liniowo przed pozostałymi standardowymi krokami. Wynikowy schemat przetwarzania może być interpretowany jako zastosowanie filtrów modulacyjnych, tj, dla każdego pasma częstotliwościowego j i filtru k, uzyskany jest sygnał o długości N', który jest obwiednią częstotliwościową z pasma$j$ przefiltrowaną przez filtr k. Zaproponowane przeze mnie rozszerzenie FT-GESTOI polegające na zastosowaniu wyuczonych filtrów modulacyjnych (LMFB) jest przedstawione na rys. 11, przed użyciem standardowego back-endu, wektory $\bar{\mathbf{g}}_{jn}^{(\cdot \mathrm{t})} = [\bar{G}_{jn1}^{(\cdot \mathrm{t})} \ \ldots \ \bar{G}_{jnK}^{(\cdot \mathrm{t})}]$ and $\mathbf{g}_{jn}^{(\cdot \mathrm{r})} = [G_{jn1}^{(\cdot \mathrm{r})} \ \ldots \ G_{jnK}^{(\cdot \mathrm{r})}]$ są liniowo przekształcone do potencjalnie niższej wymiarowości przy użyciu macierzy $\mathbf{H} \in \mathbb{R}^{K'\times K}$ jako

$$\mathbf{y}_{jn}^{(\cdot \mathrm{t})} = \mathbf{H}\bar{\mathbf{g}}_{jn}^{(\cdot \mathrm{t})} \ , \tag{66}$$

$$\mathbf{y}_{jn}^{(\cdot \mathrm{r})} = \mathbf{H}\mathbf{g}_{jn}^{(\cdot \mathrm{r})} \ . \tag{67}$$

Uczone filtry modulacyjne są zaimplementowane przy użyciu warstwy splotowej 2D z filtrami o wymiarach 1×1. Po przekształceniu, tworzone są tensory Y_T i Y_R o wymiarach $J \times N' \times K'$, po czym następną operacje z back-endu ESTOI (standaryzacja wzdłuż trzeciego wymiaru a potem wzdłuż pierwszego wymiaru), co daje w wyniku $\tilde{Y}_T$ i $\tilde{Y}_R$. Po tym następuje obliczenie macierzy $\mathbf{D}^{(\mathrm{l})}$ i $\mathbf{D}^{(\mathrm{r})}$ za składowymi pośredniej zrozumiałości jako

$$D_{jn}^{(\cdot)} = \sum_{k=1}^{K'} V_{jk} \tilde{Y}_{jnk}^{(\cdot \mathrm{t})} \tilde{Y}_{jnk}^{(\cdot \mathrm{r})} \ , \tag{68}$$

gdzie dla każdego pasma j i indeksu cechy k istnieje trenowalna waga V_{jk}. Macierze $\mathbf{D}^{(\mathrm{l})}$ i $\mathbf{D}^{(\mathrm{r})}$ są łączone jak pokazano na rys. 10.

Gałąź uwagi czasowej (temporal attention – TA) W przypadku ESTOI, sygnały testowy i referencyjny są indywidualnie porównywane dla każdej pozycji okna kroczącego (jądra $\mathbf{K}$) o długości 300 ms. Wynikiem tej operacji są komponenty pośredniej zrozumiałości d_n zlokalizowane w czasie (zobacz równanie (61)). W ESTOI te wartości są uśredniane z jednakowymi wagami w celu uzyskanie pośredniej zrozumiałości mowy (zob. równanie (62)).

W pracy [A1] przedstawiłem także rozszerzenie FT-GESTOI. Bazuje ono na gałęzi uwagi czasowej, która oblicza wagi na podstawie sygnału referencyjnego dla wszystkich pozycji n okna kroczącego. Innymi słowy, oblicza sygnał ważący. Co więcej, może istnieć wiele sygnałów (głowic), które zależą od różnych aspektów krótkoterminowego widma sygnału referencyjnego.

Odporność mowy na zakłócenia Nagrania dźwiękowe zdań różnią się odpornością na zakłócenia przez takie czynniki jak przewidywalność słów w zdaniu, szybkość mówienia, precyzja artykulacji, charakterystyka mówcy. Jeśli słowa są łatwe do przewidzenia i wyraźnie wyartykułowane, wtedy jest większa szansa na to, że zdegradowane nagranie będzie trafnie rozpoznane. W kontekście metody FT-GESTOI mniejsze wartości pośredniej zrozumiałości d mogą korespondować do większej zrozumiałości mowy i. W pracy [A1] zaproponowałem korektę pośredniej zrozumiałości w oparciu przewidywaną odporność mowy na zakłócenia.

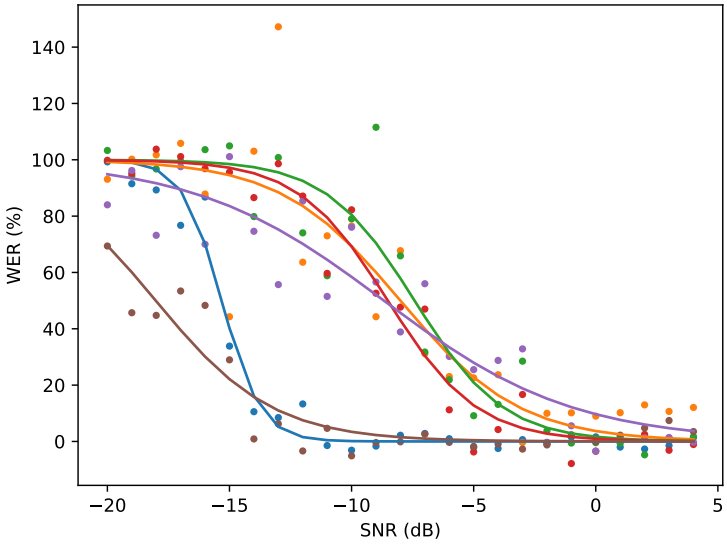
Odporność na zakłócenia każdego sygnału jest oceniana przy użyciu systemu automatycznego rozpoznawania mowy (ASR - automatic speech recognition). Należy podkreślić, że chociaż ASR może być bezpośrednio używane do predykcji zrozumiałości mowy na podstawie sygnału testowego, w moim rozwiązaniu ocena odporności może być przeprowadzona offline. Jej wynik może być użyty jako dodatkowe wejście sieci FT-GESTOI. Nie zwiększa to znacząco kosztu obliczeniowego predykcji zrozumiałości mowy i może być wykorzystane w trenowaniu sieci neuronowych do poprawy zrozumiałości mowy. Chociaż podczas procedury treningowej zazwyczaj dysponuje się większymi zasobami obliczeniowymi, funkcja kosztu bazująca na ASR opartym na głębokiej sieci neuronowej, może zużyć znacznie więcej zasobów, np. jeśli planowane jest trenowanie systemu do poprawy jakości sygnału mowy przy użyciu pojedynczej karty GPU. Najlepszy system bazujący na ASR w Clarity prediction challenge potrzebuje 4 GB na każde 10 s sygnału.

Odporność nagrań oceniałem przy użyciu szumu speech-shaped noise (SSN) poprzez mieszanie go z sygnałem mowy dla SNR od -20 do 5 dB z krokiem 1 dB. Dla każdego SNR wyznaczano błąd rozpoznawania mowy (WER – word error rate). W następnym kroku krzywa logistyczna była dopasowywana, żeby modelować zależność pomiędzy SNR i WER jako

$$(a_m, s_{0m}) = \underset{(a,s_{0m})}{\arg\min} \left\| \mathbf{e}_m - \frac{1}{1 - \exp[a(\mathbf{s}_m - s_{0m}\mathbf{1})]} \right\|^2 , \tag{69}$$

gdzie składowe $\mathbf{e}_m$ są wartościami WER dla odpowiadających im SNR w $\mathbf{s}_m$ dla m-tego przykładu. Dla każdego nagrania, s_{0m} jest przechowywane jako wartość odzwierciedlająca odporność dla danego nagrania, tj. wartość SNR dla której WER jest równy 0.5 jest uzyskiwana dla systemu ASR.

Na rys. 12 pokazano WER i dopasowane krzywe logistyczne dla sześciu losowo wybranych nagrań referencyjnych. Można zauważyć, że dla czterech nagrań, ich s_0 są w okolicach -6 dB, podczas gdy dla dwóch nagrań wartość ta jest poniżej -20 dB. To pokazuje, że istnieją nagrania w mniejszym stopniu odporne na konkretny rodzaj zniekształceń (SSN



Rysunek 12: WER uzyskane dla różnych SNR (w zakresie od -20 do 5 dB) i krzywych logistycznych dopasowanych do nich dla sześciu losowo wybranych nagrań (różne kolory)

noise). Możliwe jest, że to wynika z mniej starannej artykulacji, która wpływa na niejednoznaczność podczas dekodowania tekstu w ASR. Podobna niejednoznaczność może też wystąpić u ludzi.

Wyniki:

Metodę FT-GESTOI przetestowałem przy użyciu zbioru danych z Clarity prediction challenge 1, a wyniki zawarte są w pracy [A1]. W przypadku zadania *closed-track*, najlepszy rezultat uzyskała sieć FT-GESTOI- LMFB-TA-R i okazała się lepsza od modelu bazowego CPC-1 o 7,17%, a w porównaniu do zwycięskiego systemu z CPC-1 dała lepszy rezultat o 1,17% RMSE. Ta poprawa jest spójna dla innego indeksu tj. współczynnika korelacji (pomiędzy przewidzianymi, a prawdziwymi wartościami zrozumiałości mowy). Dla FT-GESTOI-LMFB- TA-R współczynnik korelacji wyniósł 0,81, podczas gdy dla najlepszego systemu biorącego udział w CPC-1 współczynnik korelacji miał wartość 0,79. Po wyłączeniu LMFB i TA, ale nadal korzystając z odporności, wynik (RMSE = 22,01) był lepszy od FT-GESTOI-LMFB-TA choć nieznacznie gorszy od FT-GESTOI-LMFB-TA-R. Dla zadania *open-track*, najlepszy wynik uzyskano dla FT-GESTOI-LMFB-TA-R, która dała lepszy wynik niż ASR [61].

W metodzie FT-GESTOI jest wykonywanych 79256 operacji na ramkę. Kiedy zastosowane jest rozszerzenie LMFB, koszt obliczeniowy zaoszczędzony przez redukcję wymiarowości jest większy niż koszt samej transformacji. Dla porównania, liczba operacji potrzebnych w samym enkoderze do przetworzenia 1s sygnału to 734 375 424. Podane wartości liczbowe zostały uzyskane przy użyciu narzędzia THOP do enkodera z pakietu SpeechBrain.

Do określenia pamięci alokowanej podczas używania modelu FT-GESTOI wykorzystałem profiler PyTorch. W najprostszym wariancie FT-GESTOI-INIT potrzebne było 111 MB, dla rozszerzeń to było dodatkowe 17MB. TransformerASR zaalokowało ponad 2 GB pamięci.

4.1.14 Podsumowanie cyklu publikacji

Podsumowując powyższy opis cyklu powiązanych tematycznie artykułów pod tytułem: „Interpretowalne i niskokosztowe metody poprawy jakości i zrozumiałości mowy", pokazuje etapy mojego rozwoju naukowego. Do najważniejszych osiągnięć stanowiących wkład w rozwój dyscypliny automatyka, elektronika, elektrotechnika i technologie kosmiczne zaliczam:

1. Opracowanie generatywnych metod uczenia maszynowego do wyznaczania wzorców charakteryzujących mówcę
   (a) Opracowanie metody DANMD umożliwiającej separację sygnałów z wykorzystaniem słownika/słowników o małej liczbie parametrów [A10, A4],
   (b) Opracowanie metody BNMD umożliwiającej nieujemny rozplot macierzy z binarną macierzą aktywacji [A6],
   (c) Opracowanie systemu do automatycznego rozpoznawania mówcy, łączącego różne cech, bazującego na AdaGrad i z-norm [A7].
   (d) Opracowanie metody łączenia dwóch różnych podejść do detekcji uwagi słuchowej: NT i APL [A5].
2. Niskokosztowe, różniczkowalne metody uczenia maszynowego do poprawy jakości i zrozumiałości mowy.
   (a) Opracowanie modułu sieci neuronowej WCU potencjalnie wykorzystującej szeroki kontekst w cechach wejściowych [A13, A12, A8, A3],
   (b) Opracowanie procedury (multi-pass) do trenowania sieci neuronowej, w której sygnał jest wielokrotnie przetwarzany przez ten sam moduł [A11, A9, A3],
   (c) Opracowanie metody (DPN) uczenia się procesora dynamiki [A2],
   (d) Opracowanie metody (FT-GESTOI) uczenia niskokosztowej funkcji strat do predykcji zrozumiałości mowy [A1].

Bibliografia dotycząca cyklu publikacji

[1] Jonathan Le Roux, Scott Wisdom, Hakan Erdogan i John R Hershey. "SDR–half-baked or well done?" W: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2019, s. 626–630.

[2] Jesper Jensen i Cees H Taal. "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers". W: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 24.11 (2016), s. 2009–2022.

[3] Daniel Lee i H Sebastian Seung. "Algorithms for non-negative matrix factorization". W: *Advances in neural information processing systems* 13 (2000).

[4] Ali Taylan Cemgil. "Bayesian inference for nonnegative matrix factorisation models". W: *Computational Intelligence and Neuroscience* 2009.1 (2009).

[5] Arthur P Dempster, Nan M Laird i Donald B Rubin. "Maximum likelihood from incomplete data via the EM algorithm". W: *Journal of the royal statistical society: series B (methodological)* 39.1 (1977), s. 1–22.

[6] Jörg Lücke i Julian Eggert. "Expectation truncation and the benefits of preselection in training generative models". W: *The Journal of Machine Learning Research* 11 (2010), s. 2855–2900.

[7] Paris Smaragdis. "Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs". W: *Independent Component Analysis and Blind Signal Separation: Fifth International Conference, ICA 2004, Granada, Spain, September 22-24, 2004. Proceedings 5*. Springer. 2004, s. 494–499.

[8] Tuomas Virtanen i Ali Taylan Cemgil. "Mixtures of gamma priors for non-negative matrix factorization based speech separation". W: *International Conference on Independent Component Analysis and Signal Separation*. Springer. 2009, s. 646–653.

[9] Xabier Jaureguiberry, Pierre Leveau, Simon Maller i Juan Jose Burred. "Adaptation of source-specific dictionaries in non-negative matrix factorization for source separation". W: *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2011, s. 5–8.

[10] Morten Kolbæk, Dong Yu, Zheng-Hua Tan i Jesper Jensen. "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks". W: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25.10 (2017), s. 1901–1913.

[11] William M Campbell, Joseph P Campbell, Terry P Gleason, Douglas A Reynolds i Wade Shen. "Speaker verification using support vector machines and high-level features". W: *IEEE Transactions on Audio, Speech, and Language Processing* 15.7 (2007), s. 2085–2094.

[12] Douglas Reynolds, Walter Andrews, Joseph Campbell, Jiri Navratil, Barbara Peskin, Andre Adami, Qin Jin, David Klusacek, Joy Abramson, Radu Mihaescu i in. "The SuperSID project: Exploiting high-level information for high-accuracy speaker recognition". W: *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03)*. T. 4. IEEE. 2003, s. IV–784.

[13] Adam Dabrowski, Szymon Drgas, Pawel Pawlowski i Julian Balcerek. "Development of PUEPS corpus of emergency telephone conversations". W: *Language Resources for Public Security Applications* (2012), s. 8.

[14] James A O'Sullivan, Alan J Power, Nima Mesgarani, Siddharth Rajaram, John J Foxe, Barbara G Shinn-Cunningham, Malcolm Slaney, Shihab A Shamma i Edmund C Lalor. "Attentional selection in a cocktail party environment can be decoded from single-trial EEG". W: *Cerebral cortex* 25.7 (2015), s. 1697–1706.

[15] N. Ding i J. Z. Simon. "Emergence of neural encoding of auditory objects while listening to competing speakers". W: *Proceedings of the National Academy of Sciences* 109.29 (lip. 2012), s. 11854–11859. DOI: 10.1073/pnas.1205381109. URL: https://doi.org/10.1073%2Fpnas.1205381109.

[16] Elana M. Zion Golumbic, Nai Ding, Stephan Bickel, Peter Lakatos, Catherine A. Schevon, Guy M. McKhann, Robert R. Goodman, Ronald Emerson, Ashesh D. Mehta, Jonathan Z. Simon, David Poeppel i Charles E. Schroeder. "Mechanisms Underlying Selective Neuronal Tracking of Attended Speech at a "Cocktail Party"". W: *Neuron* 77.5 (mar. 2013), s. 980–991. DOI: 10.1016/j.neuron.2012.12.037. URL: https://doi.org/10.1016%2Fj.neuron.2012.12.037.

[17] Nima Mesgarani i Edward F. Chang. "Selective cortical representation of attended speaker in multi-talker speech perception". W: *Nature* 485.7397 (kw. 2012), s. 233–236. DOI: 10.1038/nature11020. URL: https://doi.org/10.1038%2Fnature11020.

[18] Cort Horton, Michael D'Zmura i Ramesh Srinivasan. "Suppression of competing speech through entrainment of cortical oscillations". W: *Journal of Neurophysiology* 109.12 (czer. 2013), s. 3082–3093. DOI: 10.1152/jn.01026.2012. URL: https://doi.org/10.1152%2Fjn.01026.2012.

[19] Samuel Thorpe, Michael D'Zmura i Ramesh Srinivasan. "Lateralization of Frequency-Specific Networks for Covert Spatial Attention to Auditory Stimuli". W: *Brain Topography* 25.1 (czer. 2011), s. 39–54. DOI: 10.1007/s10548-011-0186-x. URL: https://doi.org/10.1007%2Fs10548-011-0186-x.

[20] Malte Wöstmann, Björn Herrmann, Burkhard Maess i Jonas Obleser. "Spatiotemporal dynamics of auditory attention synchronize with speech". W: *Proceedings of the National Academy of Sciences* 113.14 (mar. 2016), s. 3873–3878. DOI: 10.1073/pnas.1523357113. URL: https://doi.org/10.1073%2Fpnas.1523357113.

[21] Sarah Tune, Malte Wöstmann i Jonas Obleser. "Probing the limits of alpha power lateralization as a neural marker of selective attention in middle-aged and older listeners". W: *European Journal of Neuroscience* 48.7 (lut. 2018), s. 2537–2550. DOI: 10.1101/267989. URL: https://doi.org/10.1101%2F267989.

[22] J. R. Kerlin, A. J. Shahin i L. M. Miller. "Attentional Gain Control of Ongoing Cortical Speech Representations in a Cocktail Party". W: *Journal of Neuroscience* 30.2 (sty. 2010), s. 620–628. DOI: 10.1523/jneurosci.3631-09.2010. URL: https://doi.org/10.1523%2Fjneurosci.3631-09.2010.

[23] Malte Wöstmann, Lorenz Fiedler i Jonas Obleser. "Tracking the signal cracking the code: speech and speech comprehension in non-invasive human electrophysiology". W: *Language, Cognition and Neuroscience* 32.7 (grud. 2016), s. 855–869. DOI: 10.1080/23273798.2016.1262051. URL: https://doi.org/10.1080%2F23273798.2016.1262051.

[24] Po-Sen Huang, Minje Kim, Mark Hasegawa-Johnson i Paris Smaragdis. "Deep learning for monaural speech separation". W: *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2014, s. 1562–1566.

[25] Po-Sen Huang, Minje Kim, Mark Hasegawa-Johnson i Paris Smaragdis. "Joint optimization of masks and deep recurrent neural networks for monaural source separation". W: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 23.12 (2015), s. 2136–2147.

[26] DeLiang Wang i Jitong Chen. "Supervised speech separation based on deep learning: An overview". W: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 26.10 (2018), s. 1702–1726.

[27] Yuxuan Wang, Arun Narayanan i DeLiang Wang. "On training targets for supervised speech separation". W: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 22.12 (2014), s. 1849–1858.

[28] Jonathan Le Roux, Gordon Wichern, Shinji Watanabe, Andy Sarroff i John R Hershey. "The phasebook: Building complex masks via discrete representations for source separation". W: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2019, s. 66–70.

[29] Ashutosh Pandey i DeLiang Wang. "A new framework for CNN-based speech enhancement in the time domain". W: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 27.7 (2019), s. 1179–1188.

[30] Zhong-Qiu Wang, Jonathan Le Roux, DeLiang Wang i John R Hershey. "End-to-end speech separation with unfolded iterative phase reconstruction". W: *Proc. Interspeech 2018*. 2018, s. 2708–2712.

[31] Zhong-Qiu Wang, Ke Tan i DeLiang Wang. "Deep learning based phase reconstruction for speaker separation: A trigonometric perspective". W: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2019, s. 71–75.

[32] Jitong Chen i DeLiang Wang. "Long short-term memory for speaker generalization in supervised speech separation". W: *The Journal of the Acoustical Society of America* 141.6 (2017), s. 4705–4714.

[33] L. Hui, M. Cai, C. Guo, L. He, W. Q. Zhang i J. Liu. "Convolutional maxout neural networks for speech separation". W: *2015 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*. Grud. 2015, s. 24–27. DOI: 10.1109/ISSPIT.2015.7394335.

[34] Emad M Grais i Mark D Plumbley. "Single channel audio source separation using convolutional denoising autoencoders". W: *2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. 2017, s. 1265–1269.

[35] Fisher Yu i Vladlen Koltun. "Multi-scale context aggregation by dilated convolutions". W: *arXiv preprint arXiv:1511.07122* (2015).

[36] Wenhao Yuan. "A time–frequency smoothing neural network for speech enhancement". W: *Speech Communication* 124 (2020), s. 75–84. ISSN: 0167-6393. DOI: https://doi.org/10.1016/j.specom.2020.09.002. URL: https://www.sciencedirect.com/science/article/pii/S0167639320302703.

[37] Sunitha Basodi, Chunyan Ji, Haiping Zhang i Yi Pan. "Gradient amplification: An efficient way to train deep neural networks". W: *Big Data Mining and Analytics* 3.3 (2020), s. 196–207.

[38] Razvan Pascanu, Tomas Mikolov i Yoshua Bengio. "On the difficulty of training recurrent neural networks". W: *International Conference on Machine Learning*. PMLR. 2013, s. 1310–1318.

[39] Nicholas A Lesica. "Why do hearing aids fail to restore normal auditory perception?" W: *Trends in neurosciences* 41.4 (2018), s. 174–185.

[40] Olaf Ronneberger, Philipp Fischer i Thomas Brox. “U-net: Convolutional networks for biomedical image segmentation”. W: *International Conference on Medical image computing and computer-assisted intervention*. Springer. 2015, s. 234–241.

[41] Ke Tan, Jitong Chen i DeLiang Wang. “Gated residual networks with dilated convolutions for monaural speech enhancement”. W: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 27.1 (2019), s. 189–198.

[42] Aaron Nicolson i Kuldip K Paliwal. “Deep learning for minimum mean-square error approaches to speech enhancement”. W: *Speech Communication* 111 (2019), s. 44–55.

[43] Cees H Taal, Richard C Hendriks, Richard Heusdens i Jesper Jensen. “An algorithm for intelligibility prediction of time–frequency weighted noisy speech”. W: *IEEE Transactions on Audio, Speech, and Language Processing* 19.7 (2011), s. 2125–2136.

[44] Antony W Rix, Michael P Hollier, Andries P Hekstra i John G Beerends. “Perceptual Evaluation of Speech Quality (PESQ) The New ITU Standard for End-to-End Speech Quality Assessment Part I–Time-Delay Compensation”. W: *Journal of the Audio Engineering Society* 50.10 (2002), s. 755–764.

[45] Andong Li, Minmin Yuan, Chengshi Zheng i Xiaodong Li. “Speech enhancement using progressive learning-based convolutional recurrent neural network”. W: *Applied Acoustics* 166 (2020), s. 107347.

[46] Tian Gao, Jun Du, Li-Rong Dai i Chin-Hui Lee. “SNR-Based Progressive Learning of Deep Neural Network for Speech Enhancement.” W: *Proc. Interspeech 2016*. 2016, s. 3713–3717.

[47] Tian Gao, Jun Du, Li-Rong Dai i Chin-Hui Lee. “Densely connected progressive learning for lstm-based speech enhancement”. W: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2018, s. 5054–5058.

[48] Donald S Williamson i DeLiang Wang. “Time-frequency masking in the complex domain for speech dereverberation and denoising”. W: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25.7 (2017), s. 1492–1501.

[49] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser i Illia Polosukhin. “Attention is all you need”. W: *Advances in neural information processing systems* 30 (2017).

[50] Yi Luo i Nima Mesgarani. “Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation”. W: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 27.8 (2019), s. 1256–1266.

[51] Zehai Tu, Jisi Zhang, Ning Ma, Jon Barker i in. “A two-stage end-to-end system for speech-in-noise hearing aid processing”. W: *Proc. Clarity* (2021), s. 3–5.

[52] Zehai Tu, Ning Ma i Jon Barker. “DHASP: Differentiable hearing aid speech processing”. W: *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2021, s. 296–300.

[53] K Zmolikova i JH Cernock. “BUT system for the first clarity enhancement challenge”. W: *Proc. Clarity* (2021), s. 1–3.

[54] Yoshito Nejime i Brian CJ Moore. "Simulation of the effect of threshold elevation and loudness recruitment combined with reduced frequency selectivity on the intelligibility of speech in noise". W: *The Journal of the Acoustical Society of America* 102.1 (1997), s. 603–615.

[55] James M Kates. "Principles of digital dynamic-range compression". W: *Trends in Amplification* 9.2 (2005), s. 45–76.

[56] Shanqing Cai, Wei-Li D Ma i Eric D Young. "Encoding intensity in ventral cochlear nucleus following acoustic trauma: implications for loudness recruitment". W: *Journal of the Association for Research in Otolaryngology* 10.1 (2009), s. 5–22.

[57] Michael A Stone, Brian CJ Moore, Katrin Meisenbacher i Ralph P Derleth. "Tolerable hearing aid delays. V. Estimation of limits for open canal fittings". W: *Ear and Hearing* 29.4 (2008), s. 601–617.

[58] Brian CJ Moore, Brian R Glasberg i Michael A Stone. "Development of a new method for deriving initial fittings for hearing aids with multi-channel compression: CAMEQ2-HF". W: *International Journal of Audiology* 49.3 (2010), s. 216–227.

[59] Jon Barker, Michael Akeroyd, Trevor J Cox, John F Culling, Jennifer Firth, Simone Graetzer, Holly Griffiths, Lara Harris, Graham Naylor, Zuzanna Podwinska i in. "The 1st Clarity Prediction Challenge: A machine learning challenge for hearing aid intelligibility prediction." W: *Interspeech*. 2022, s. 3508–3512.

[60] Mark Huckvale i Gaston Hilkhuysen. "ELO-SPHERES intelligibility prediction model for the Clarity Prediction Challenge 2022". W: *Proc. Interspeech 2022*. T. 2022. ISCA. 2022, s. 3934–3938.

[61] Zehai Tu, Ning Ma i Jon Barker. "Exploiting hidden representations from a DNN-based speech recogniser for speech intelligibility prediction in hearing-impaired listeners". W: *Proc. Interspeech 2022*. 2022, s. 3488–3492.

[62] Zehai Tu, Ning Ma i Jon Barker. "Unsupervised uncertainty measures of automatic speech recognition for non-intrusive speech intelligibility prediction". W: *Proc. Interspeech 2022*. 2022, s. 3493–3497.

[63] Raphael Olivier i Bhiksha Raj. "There is more than one kind of robustness: Fooling whisper with adversarial examples". W: *Proc. Interspeech 2023*. 2023, s. 4394–4398.

[64] David Poeppel i M Florencia Assaneo. "Speech rhythms and their neural foundations". W: *Nature Reviews Neuroscience* 21.6 (2020), s. 322–334.

[65] Jorge J Moré. "The Levenberg-Marquardt algorithm: implementation and theory". W: *Numerical Analysis*. Springer. 1978, s. 105–116.

4.2 Prace naukowe niewchodzące w skład głównego osiągnięcia naukowego

W niniejszym punkcie jest przedstawiony mój dorobek naukowy, który nie został przeze mnie włączony do omówionego w punkcie 4.1 cyklu publikacji stanowiącego główne osiągnięcia naukowe. Ten dorobek naukowy również mieści się w tematyce „interfejsy człowiek-maszyna". W szczególności dotyczy on zagadnień:

1. Rozpoznawania mowy, w tym mowy szeptanej,

2. Optymalizacji parametrów ougólnionego podobieństwa kosinusowego,

3. Wykrywania patologii głosu,
4. Wykrywania nieprawidłowych dźwięków osłuchowych,
5. Filtry cząsteczkowe.

Poniżej znajduje się lista publikacji, które dotyczą tego dorobku.

Publikacje niewchodzące w skład głównego osiągnięcia naukowego

[E1] **Szymon Drgas** i Adam Dabrowski. "Generalized cosine similarity in I-vector based automatic speaker recognition systems". W: *2013 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA)*. IEEE. 2013, s. 73–77.

[E2] **Szymon Drgas** i Rafal Zdunek. "Automatic tuning of hyperparameters for NMF-based face recognition system". W: *2016 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA)*. IEEE. 2016, s. 40–44.

[E3] Talar Sadalla, Piotr Kozierski, Dariusz Horla, Wojciech Giernacki i **Szymon Drgas**. "Particle Filtering in Servo Drive Velocity Control with Fractional-Order PI Controller". W: *Automation 2018: Advances in Automation, Robotics and Measurement Techniques*. Springer. 2018, s. 265–275.

[E4] Piotr Kozierski, Talar Sadalla, **Szymon Drgas**, Adam Dąbrowski i Wojciech Giernacki. "Polish whispery speech recognition—Minimum sampling frequency". W: *2017 22nd International Conference on Methods and Models in Automation and Robotics (MMAR)*. IEEE. 2017, s. 611–615.

[E5] Piotr Kozierski, Talar Sadalla, **Szymon Drgas**, Adam Dąbrowski i Joanna Zietkiewicz. "The impact of vocabulary size and language model order on the Polish whispery speech recognition". W: *2017 22nd International Conference on Methods and Models in Automation and Robotics (MMAR)*. IEEE. 2017, s. 616–621.

[E6] Piotr Kozierski, Talar Sadalla, **Szymon Drgas** i Adam Dąbrowski. "Allophones in automatic whispery speech recognition". W: *2016 21st International Conference on Methods and Models in Automation and Robotics (MMAR)*. IEEE. 2016, s. 811–815.

[E7] Piotr Kozierski, Talar Sadalla, **Szymon Drgas**, Adam Dąbrowski i Dariusz Horla. "Kaldi toolkit in Polish whispery speech recognition". W: *Przeglad Elektrotechniczny* 92.11 (2016), s. 301–304.

[E8] Piotr Kozierski, Talar Sadalla, Adam Owczarkowski i **Szymon Drgas**. "Particle filter in multidimensional systems". W: *2016 21st international conference on methods and models in automation and robotics (MMAR)*. IEEE. 2016, s. 806–810.

[E9] Adam Dąbrowski, Julian Balcerek, **Szymon Drgas**, Tomasz Marciniak, Andrzej Meyer i Paweł Pawłowski. "Nowoczesne systemy łączności i transmisji danych na rzecz bezpieczeństwa : szanse i zagrożenia". W: red. Andrzej R. Pach, Zbigniew Rau i Michał Wągrowski. Wolters Kluwer Polska, 2013. Rozd. Klasyfikacja i rozpoznawanie osób na podstawie rozmów na telefony alarmowe.

[E10] Adam Dąbrowski, Piotr Kardyś, Marek Portalski, Damian Cetnarowicz, **Szymon Drgas** i Paweł Pawłowski. "Układy elektroniczne jako elementy ludzkiego ciała i człowiek jako element układów elektronicznych". W: *Elektronika: konstrukcje, technologie, zastosowania* 54.9 (2013), s. 53–57.

[E11] Tomasz Grzywalski, Riccardo Belluzzo, **Szymon Drgas**, Agnieszka Cwalińska i Honorata Hafke-Dys. "Interactive Lungs Auscultation with Reinforcement Learning Agent". en. W: *Proceedings of the 11th International Conference on Agents and Artificial Intelligence ICAART 2019. Volume 2*. SciTePress, 2019, s. 824–832. DOI: 10.5220/0007573608240832. URL: http://www.insticc.org/Primoris/Resources/PaperPdf.ashx?idPaper=75736.

[E12] Tomasz Grzywalski, Adam Maciaszek, Adam Biniakowski, Jan Orwat, **Szymon Drgas**, Mateusz Piecuch, Riccardo Belluzzo, Krzysztof Joachimiak, Dawid Niemiec, Jakub Ptaszynski i Krzysztof Szarzynski. "Parameterization of Sequence of MFCCs for DNN-based voice disorder detection". W: *2018 IEEE International conference on big data (big data)*. IEEE. 2018, s. 5247–5251.

[E13] Piotr Kozierski, Talar Sadalla, **Szymon Drgas**, Adam Dabrowski, Joanna Zietkiewicz i Wojciech Giernacki. "Acoustic Model Training, using Kaldi, for Automatic Whispery Speech Recognition." W: *FedCSIS (Position Papers)*. 2018, s. 109–114.

[E14] Piotr Kozierski, Jacek Michalski, Talar Sadalla, Wojciech Giernacki, Joanna Zietkiewicz i **Szymon Drgas**. "New grid for particle filtering of multivariable nonlinear objects". W: *2018 Federated Conference on Computer Science and Information Systems (FedCSIS)*. IEEE. 2018, s. 1073–1077.

Mój wkład w wyżej wymienionych pracach jest następujący:

1. Opracowanie systemu automatycznego rozpoznawania mówcy w oparciu o uogólnioną odległość kosinusową, do optymalizacji zaproponowałem algorytm wykorzystujący grupy Liego ([E1]).

2. Analiza automatycznego strojenia hiperparametrów modeli NMF przy użyciu procesów gaussowskich [E2].

3. Wykrywanie nieprawidłowych dźwięków układu oddechowego zarejestrowanych przez stetoskop. W firmie StethoMe zaproponowałem sposób zbierania danych i ekstrakcji cech. System częściowo bazujący na tych elementach został opisany w pracy [E11].

4. Brałem udział pracach zespołu przygotowującego system wykrywania patologii wytwarzania mowy na konkurs FEMH voice pathology [E12]. Mój wkład polegał na analizie i interpretacji wyników.

5. Brałem udział w przygotowaniu zbioru nagrań do trenowania i testowania automatycznego rozpoznawania mowy szeptanej [E13, E5, E6, E7].

6. Analiza działania filtrów cząsteczkowych [E14, E8].

4.3 Prace naukowe opublikowane przed uzyskaniem stopnia doktora nauk technicznych

Przed uzyskaniem stopnia doktora nauk technicznych m.in. opublikowałem artykuły związane ze zrozumiałością mowy, do których odniosłem się w mojej pracy doktorskiej. Dwie ważne prace wymienione są poniżej.

Artykuły związan z doktoratem

[D1] Szymon Drgas i Magdalena A Blaszak. "Perceptual consequences of changes in vocoded speech parameters in various reverberation conditions". W: *Journal of Speech, Language, and Hearing Research* (2009).

[D2] Szymon Drgas i Magdalena A Blaszak. "Perception of speech in reverberant conditions using AM–FM cochlear implant simulation". W: *Hearing research* 269.1-2 (2010), s. 162–168.

5 Informacja o wykazywaniu się istotną aktywnością naukową albo artystyczną realizowaną w więcej niż jednej uczelni, instytucji naukowej lub instytucji kultury, w szczególności zagranicznej.

Wizyty naukowe:

1. Finlandia, Tampere University of Technology (obecnie: Tampere University), Audio Research Group, prof. Tuomas Virtanen, sierpień 2013, 5.08.2013-25.08.2013, wspólne badania dotyczące przetwarzania sygnału mowy
2. Niemcy, Carl von Ossietzky, Universitat, Machine Learning Group, prof. Jörg Lücke, 23.02.2015-18.03.2015, wspólne badania dotyczące uczenia reprezentacji sygnałów
3. Finlandia, Tampere University of Technology (obecnie: Tampere University), Audio Research Group, prof. Tuomas Virtanen, 3.06.2015-24.06.2015, wspólne dotyczące badania uczenia reprezentacji sygnałów

W wyniku wymienionych wyżej wizyt naukowych powstały wspólne artykuły: [A2, A6, A4, A7].

Współpraca z zagranicą:
Współpraca z dr. Larsem Bramsløwem z Eriksholm Research Center (Dania),
Udział w działalności EEG Research Hub, prof. Rob van der Lubbe, University of Twente (Holandia).

Projekty:
Po uzyskaniu stopnia doktora brałem udział w następujących projektach:

1. Wydział Informatyki Politechniki Poznańskiej, Konkurs badawczy dla młodych pracowników nauki Pro-IDEAS-2013, "Cyfrowe metody analizy sceny dźwiękowej (CMASD)", numer projektu 09/93/DSMK/4106 - **kierownik**
2. Stypendium celowe w ramach projektu Inżynier Przyszłości, Wzmocnienie potencjału dydaktycznego Politechniki Poznańskiej. Projekt nr 012/2/2014/MD: "Ekstrakcja informacji o mówcy z nagrań jednokanałowych" - **stypendysta**

3. Uczestniczyłem w pracach firmy StethoMe realizującej projekt nr POIR.01.01.01-00-0528/16-02: "Unikalna w skali świata technologia identyfikacji i klasyfikacji dźwięków z badania osłuchowego przeprowadzonego w warunkach domowych, z wykorzystaniem inteligentnych algorytmów, jako wsparcie zdalnej diagnostyki i monitorowania chorób układu oddechowego." w ramach programu Szybka Ścieżka.

6 Informacja o osiągnięciach dydaktycznych, organizacyjnych oraz popularyzujących naukę lub sztukę.

W Politechnice Poznańskiej jestem zatrudniony na stanowisku adiunkta w grupie pracowników badawczo-dydaktycznych, co jest powiązane z realizacją zadań dydaktycznych i organizacyjnych.

6.1 Działalność dydaktyczna

Moja działalność dydaktyczna na Politechnice Poznańskiej związana jest z prowadzeniem m.in. akustyki technicznej. Oprócz tego prowadzę zajęcia z algebry liniowej oraz optymalizacji, przedmiotów które obejmują zagadnienia, na których bazują metody uczenia maszynowego. Jestem odpowiedzialny za następujące przedmioty:

1. Akustyka techniczna (Automatyka i robotyka, 2 stopień, specjalność systemy wizyjne). wykład
2. Teoria i metody optymalizacji (Automatyka i robotyka, 2 stopień, specjalność systemy wizyjne), wykład i ćwiczenia
3. Algebra z geometrią, (Automatyka i robotyka, 1 stopień, studia niestacjonarne), wykład i ćwiczenia

Oprócz tego, po uzyskaniu stopnia doktora prowadziłem następujące zajęcia:

1. Algebra z geometrią (automatyka i robotyka, sztuczna inteligencja (w języku ang.), 1 stopień), ćwiczenia
2. Przetwarzanie sygnałów i informacji (automatyka i robotyka, 1 stopień), laboratoria

Jestem promotorem 14 prac dyplomowych inżynierskich i 6 prac magisterskich.

6.2 Działalność organizacyjna i popularyzująca naukę

Moja działalność organizacyjna charakteryzowana jest przez poniższe zestawienie:

1. Członek komitetu organizacyjnego IEEE Signal Processing Algorithms, Arrangements, and Applications (SPA). Konferencja odbywa się co roku na Politechnice Poznańskiej.
2. Opieka nad sekcją akustyki i psychoakustyki koła naukowego Decybel na Politechnice Poznańskiej.
3. Udział w pracach komitetu naukowego konferencji OSKA (Uniwersytet im. Adama Mickiewicza)

4. Redaktor działu “Speech, computational acoustics and signal processing” w czasopiśmie Archives of Acoustics wydawanym przez Polską Akademię Nauk.

5. Recenzje w takich czasopismach jak: IEEE/ACM Transactions on Audio Speech, and Language Processing; Signal Processing Letters, Computer Speech and Language, Archives of Acoustics, Multimedia Tools and Applications, PlosONE.

6. Recenzje artykułów konferencyjnych z ICASSP, Interspeech, SPA.

7. Ciągłe, wieloletnie (od 2021 r.) wsparcie pracy grupy składającej się z psychiatrów, neurologów, psychologów, fizyków przygotowującej projekt “Alzheimer Prediction Project”. Celem projektu ma być opracowanie testów funkcjonalnych wczesnego wykrywania choroby Alzheimera w fazie prodromalnej. Jednym ze sposobów ma być wczesne wykrywanie w oparciu o płynność mowy i języka. Zespół składa się ze specjalistów z Uniwersytetu im. Adama Mickiewicza, Polskiej Akademii Nauk, Uniwersytetu z Oldenburgu. Opiekuję się stroną internetową (`alz.put.poznan.pl`), na której zawarte są informacje o działalności zespołu a także doniesienia naukowe z zakresu wczesnej diagnostyki choroby Alzheimera.

8. Zaangażowanie w budowę i rozwój laboratorium akustyki technicznej, które posiada komorę bezechową (box-in-box) oraz trzy komory odsłuchowe umożliwiające przeprowadzanie eksperymentów psychoakustycznych. Dzięki uzyskanemu przeze mnie finansowaniu możliwe było powiększenie możliwości tego laboratorium.

9. Ciągłe, wieloletnie prezentowanie możliwości laboratorium akustyki technicznej dla innych ośrodków naukowych i szkół odwiedzających Politechnikę Poznańską.

10. Członkostwo w the EEG Research Hub [2] - grupy prowadzonej przez prof. Roba van der Lubbe.

7 Oprócz kwestii wymienionych w pkt. 1-6, wnioskodawca może podać inne informacje, ważne z jego punktu widzenia, dotyczące jego kariery zawodowej.

Od roku 2016 współpracowałem z firmą StethoMe, która projektuje i rozwija stetoskop działający w oparciu o sztuczną inteligencję (StethoMe AI), która analizuje i wykrywa nieprawidłowe dźwięki osłuchowe w układzie oddechowym. StethoMe umożliwia wykonanie profesjonalnego badania osłuchowego w warunkach domowych. Natychmiast informuje o pojawieniu się nieprawidłowości osłuchowych. Przyspiesza niezbędne wizyty lekarskie, a eliminuje te niepotrzebne. Umożliwia wysłanie wyników na odległość, a także wspiera lekarza w procesie diagnostycznym o obiektywizujące badanie osłuchowe.

Będąc w zespole realizującym projekt “Unikalna w skali świata technologia identyfikacji i klasyfikacji dźwięków z badania osłuchowego przeprowadzonego w warunkach domowych, z wykorzystaniem inteligentnych algorytmów, jako wsparcie zdalnej diagnostyki i monitorowania chorób układu oddechowego.” W trakcie wspólnych prac badawczych powstały publikacje [E12, E11], których jestem współautorem opisane w punkcie 4.2 W pracy [A13] opisany jest system do wykrywania patologii wytwarzania mowy, który został wysłany na konkurs FEMH (Far-Eastern Memorial Hospital) voice data challenge, gdzie zajął drugie miejsce.

[2] `https://www.utwente.nl/en/bms/eeg-research-hub/`