Akademia Górniczo-Hutnicza im. Stanisława Staszica w Krakowie

Wydział Informatyki, Elektroniki i Telekomunikacji

Katedra Elektroniki

Praca dyplomowa
inżynierska

Implementacja i optymalizacja obliczeniowa algorytmu jednoczesnej separacji sygnałów i usuwania pogłosu

Imię i nazwisko: Aleksander Strzeboński

Kierunek studiów: Elektronika i Telekomunikacja

Typ studiów: Stacjonarne

Opiekun pracy: dr hab. inż. Konrad Kowalczyk, prof. AGH

Kraków 2022

Spis treści

Rozdział 1

Wstęp

1.1 Wprowadzenie

W poniższej pracy poruszona została tematyka separacji źródeł dźwięku z sygnałów mikrofonowych nagranych w warunkach pogłosowych. Separacja sygnałów jest wyzwaniem powszechnym we współczesnych systemach przetwarzania sygnałów dźwiękowych. Pozwala ona wyekstrahować sygnały źródeł dźwięku z sygnału składającego się z wielu wzajemnie zakłócających się źródeł. Dodatkowo, wielokrotnie sygnały posiadają też pogłos. Dzieje się tak, gdy źródła dźwięku znajdują się w niewygłuszonym pomieszczeniu. W celu uzyskania jak najwierniejszej wersji sygnałów źródłowych należy, oprócz separacji, usunąć również pogłos. Algorytmy służące do jednoczesnej separacji sygnałów i eliminacji pogłosu przydają się między innymi jako wstępne przetwarzanie sygnałów mikrofonowych zanim zostaną one poddane automatycznemu rozpoznawaniu mowy (ang. Automatic Speech Recognition - ASR). Przygotowanie odseparowanych i pozbawionych pogłosu sygnałów jest kluczowe do poprawnego rozpoznawania mowy.

Istotną kwestią jest rodzaj czujnika akustycznego. Separacji sygnałów można dokonać, gdy dysponuje się jednym mikrofonem [1], ale lepsze rezultaty można uzyskać przy użyciu macierzy mikrofonowej. Pozwala ona w bardziej efektywny sposób wykorzystać przestrzenne zależności sygnałów pochodzących z różnych źródeł. Przy użyciu macierzy mikrofonowej można wykorzystać metodę kształtowania wiązki (ang. beamforming) w celu wyeliminowania sygnałów niepochodzących z interesującego nas kierunku.

Kolejną kwestią jest podejście do sposobu eliminacji pogłosu pamiętając o równoczesnej potrzebie separacji sygnałów. Jedną z możliwości jest podejście sekwencyjne. Polega ono na wyeliminowaniu najpierw pogłosu, a następnie przeprowadzenie separacji sygnałów. Zakładając, że do odbierania sygnałów zastosowano macierz mikrofonową, do eliminacji pogłosu można użyć metody ważonej predykcji błędu (ang. Weighted Prediction Error - WPE) [2]. Następnie, do separacji sygnałów, można użyć metody kształtowania wiązki polegającej na minimalizacji wariancji sygnałów bez wprowadzania w nich zniekształceń (ang. Minimum Variance Distortionless Response - MVDR) [3] lub liniowo ograniczonej metody kształtowania wiązki, która minimalizuje wariancję (ang. Linearly Constrained Minimum Variance - LCMV) [4].

Można też dokonać jednoczesnej eliminacji pogłosu i separacji źródeł. W tym celu możliwe jest połączenie filtra WPE i metody kształtowania wiązki MPDR (ang. Minimum-Power Distortionless Response) [5] - wariant metody kształtowania wiązki MVDR, w jedną metodę splotowego kształtowania wiązki nazywaną skrótowo WPD (ang. Weighted Power minimization Distortionless response) [6].

W tej pracy skupiono się na podejściu, w którym separacja źródeł i eliminacja pogłosu wykonywana jest w sposób jednoczesny, przy użyciu WPD. W tym celu, do odbierania sygnałów, wykorzystywane są sygnały pochodzące z macierzy mikrofonowej.

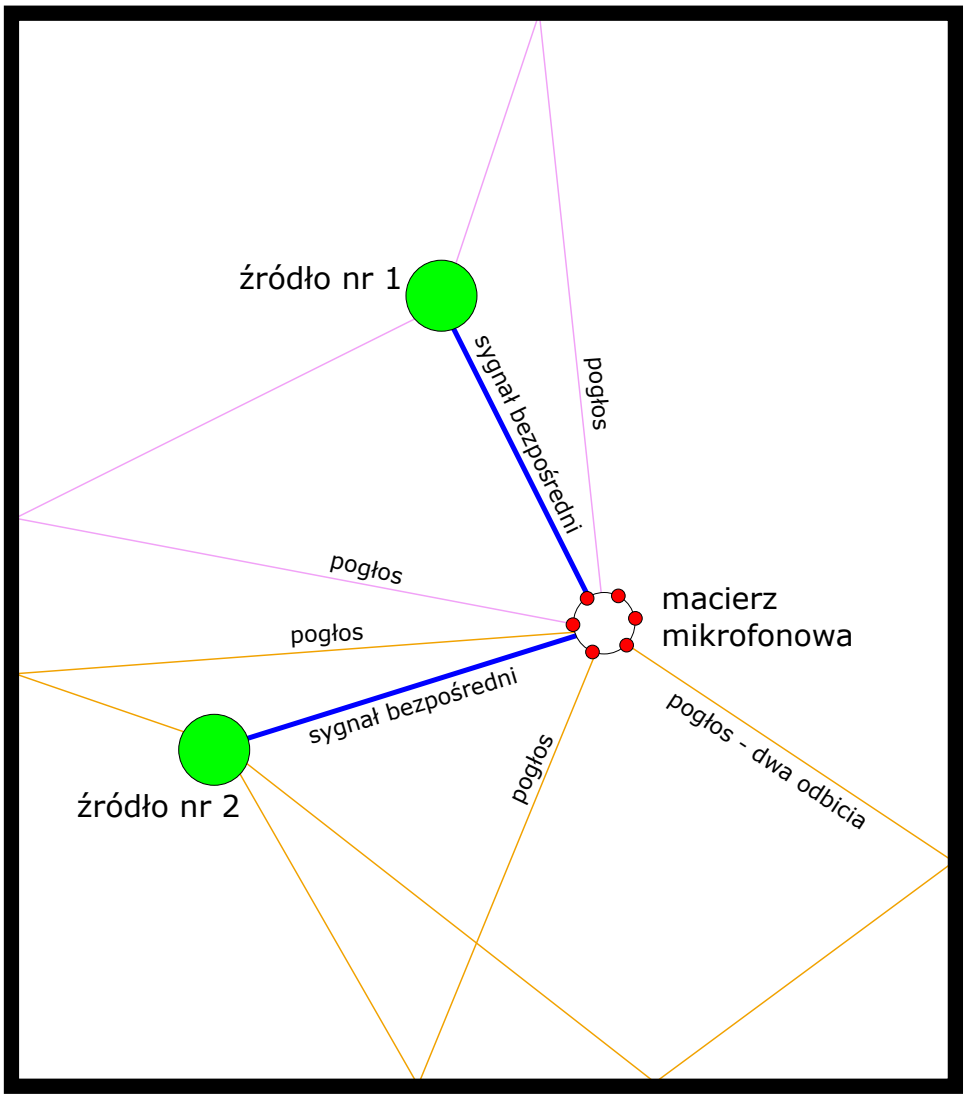
1.2 Cel

Celem pracy inżynierskiej była praktyczna implementacja i optymalizacja obliczeniowa algorytmu WPD służącego do jednoczesnej separacji sygnałów i usuwania pogłosu. Ponadto, celem pracy było również zweryfikowanie działania algorytmu przy różnych warunkach zewnętrznych zarówno pod względem jakości jak i czasu działania.

Na rysunku 4.1 symbolicznie przedstawiono problem, który należy rozwiązać. W niewytłumionym pokoju znajdują się macierz mikrofonowa oraz dwa źródła sygnału - załóżmy, że są to dwie osoby. Osoby mówią jednocześnie w związku z czym ich mowa nakłada się na siebie. Ponadto, ich mowa odbija się od ścian i, po jednym lub większej liczbie odbić, również trafia do mikrofonu tworząc w ten sposób pogłos. Zadaniem implementowanego algorytmu jest taka filtracja sygnałów mikrofonowych by pozostały jedynie sygnały bezpośrednie (eliminacja pogłosu), oddzielone od siebie (separacja sygnałów).

1.3 Zakres wykonanej pracy

Pierwszym etapem wykonywanej pracy było zapoznanie się z literaturą dotyczącą tematyki technik kształtowania wiązki oraz ich użyteczności w kontekście separacji sygnałów [7, 8]. Do implementacji i ewaluacji algorytmu wybrany został język programowania Python. Następnie, przy użyciu generatora odpowiedzi impulsowej pomieszczenia (ang. Room Impulse Response - RIR) [9, 10], stworzony został model pozwalający w łatwy sposób generować sygnały odebrane w poszczególnych mikrofonach macierzy mikrofonowej przy różnych warunkach zewnętrznych takich jak rozmiar pokoju, położenie źródeł etc. Kolejno, zaimplementowany został algorytm WPD służącego do jednoczesnej separacji sygnałów i usuwania pogłosu. Z początku, parametry wejściowe algorytmu WPD takie jak kierunki, z których nadchodzi sygnał z poszczególnych źródeł (ang. Direction of Arrival - DOA) oraz zmienna w czasie moc poszczególnych źródeł założone zostały za znane. Następnie, zaimplementowano również algorytmy służące do estymacji tych parametrów wejściowych. Do estymacji DOA zaimplementowany został algorytm MUSIC w wersji niekoherentnej [11], a do estymacji mocy źródeł zastosowano filtr WPE i metodę kształtowania wiązki LCMV. Kolejnym etapem była optymalizacja obliczeniowa algorytmu. W tym celu



Rysunek 1.1: Symboliczne przedstawienie problemu

wykorzystano między innymi możliwości biblioteki Numpy, która pozwala na szybkie przemnażanie macierzowe tablic wielowymiarowych. Dodatkowo, cześć obliczeń można było skrócić w sposób, który odbijał się na jakości działania algorytmu. Ta zależność między czasem a jakością działania również został zbadana. Następnie, przetestowana została jakość działania algorytmu. W tym celu wykonano szereg testów w różnych warunkach zewnętrznych i przy różnych sygnałach źródłowych. Do ewaluacji wyników testów, posłużyły najpopularniejsze miary pokazujące jakość separacji sygnałów i eliminacji pogłosu [12, 13, 14].

Rozdział 2

Analiza teoretyczna problemu

2.1 Model sygnału

Model sygnału jest zaczerpnięty z [6]. Zakłada się, że sygnał nadawany przez K źródeł odbierany jest przez M mikrofonów w przestrzeni, która ze względu na swoją charakterystykę dokłada do sygnałów pogłos. Kolejnym założeniem jest to, że mikrofony mają charakterystykę dookólną i umieszczone są na płaszczyźnie. Do analizy sygnałów wykorzystuje się krótko-czasową transformatę Fouriera (ang. Short-Time Fourier Transform - STFT). Przy tych założeniach, sygnał odebrany w mikrofonie, przedstawiony w dziedzinie STFT, w częstotliwości f i w ramce czasowej t można zapisać jako:

$$\mathbf{x}_t = \sum_{k=1}^{K} \mathbf{x}_t^{(k)} + \mathbf{n}_t, \tag{2.1}$$

gdzie $\mathbf{x}_t = [x_{1,t}, x_{2,t}, ..., x_{M,t}]^T \in \mathbb{C}^M$ zawiera współczynniki STFT odebranego sygnału, T oznacza transpozycję, $\mathbf{x}_t^{(k)} \in \mathbb{C}^M$ oznacza sygnał odebrany, pochodzący od źródła k, a $\mathbf{n}_t \in \mathbb{C}^M$ oznacza addytywny szum. Implementowany algorytm działa dla każdej częstotliwości niezależnie. Indeksy częstotliwościowe są tutaj i w dalszej części równań pomijane. Sygnał $\mathbf{x}_t^{(k)}$ można rozpisać w sposób następujący:

$$\mathbf{x}_t^{(k)} = \mathbf{d}_t^{(k)} + \mathbf{r}_t^{(k)}, \tag{2.2}$$

$$\mathbf{d}_t^{(k)} = \mathbf{v}^{(k)} s_t^{(k)}, \tag{2.3}$$

$$\mathbf{r}_t^{(k)} = \sum_{\tau=D}^{L_a+D-1} \mathbf{a}_\tau^{(k)} s_{t-\tau}^{(k)}, \tag{2.4}$$

gdzie $\mathbf{d}_t^{(k)}$ oznacza sygnał bezpośredni wraz z wczesnym, pożądanym pogłosem, $\mathbf{r}_t^{(k)}$ oznacza późny, niepożądany pogłos, $\mathbf{v}^{(k)} \in \mathbb{C}^M$ oznacza wektor sterujący reprezentujący zmianę sygnału źródłowego na drodze bezpośredniej źródło - mikrofon, $s_t^{(k)}$ oznacza sygnał nadawany u samego źródła, $\mathbf{a}_\tau^{(k)} \in \mathbb{C}^M$ oznacza wektor reprezentujący zmianę sygnału źródłowego, który, opóźniony o τ ramek czasowych, pojawia się na mikrofonie dokładając się w ten sposób do późnego pogłosu, D reprezentuje najkrótszy odstęp czasowy, mierzony w liczbie ramek czasowych STFT, przy którym

możemy mówić o późnym pogłosie, a L_a oznacza, mierzony również w liczbie ramek, czas trwania późnego pogłosu.

2.2 Postawiony problem

Głównym celem tej pracy inżynierskiej jest jak najlepsza estymacja pożądanego sygnału bezpośredniego wraz z wczesnymi odbiciami, odebranego w referencyjnym mikrofonie o indeksie q, czyli $d_{q,t}^{(k)}$, dla każdego z K źródeł. Parametry które są z założenia znane i posłużą do estymacji to:

- K, czyli liczba źródeł sygnału.
- geometria macierzy mikrofonowej, która jest jednym z czynników służących do estymacji wektorów sterujących $\mathbf{v}^{(k)}$.

Na wejściu, algorytm otrzymuje wieloźródłowy, pogłosowy, zaszumiony sygnał odebrany w M mikrofonach tworzących macierz mikrofonową. W dziedzinie STFT sygnał ten oznaczany jest jako $\mathbf{x}_t$.

Rozdział 3

Opracowane rozwiązanie problemu

3.1 Model rozwiązania

W celu rozwiązania problemu w głównej mierze użyty został algorytm opisany w publikacji [6]. Algorytm ten polega na wyliczeniu filtra WPD, służącego do splotowego kształtowania wiązki, którym następnie filtrować można sygnał odebrany w celu uzyskania pożądanych sygnałów. Do poprawnego wyliczenia WPD, potrzebne są następujące informacje:

- estymacja zmiennej w czasie mocy sygnałów źródłowych:

$$\sigma_t^{(k)} = |u_{q,t}^{(k)}|^2, \tag{3.1}$$

  gdzie $\sigma_t^{(k)}$ oznacza moc, a $u_{q,t}^{(k)} \approx d_{q,t}^{(k)}$ to wstępna aproksymacja sygnału pożądanego otrzymanego w referencyjnym mikrofonie o indeksie q. Sygnał $u_{q,t}^{(k)}$ uzyskany zostanie dzięki zastosowaniu kolejno filtra WPE i metody kształtowania wiązki LCMV.

- wektory sterujące znormalizowane względem wektora sterującego dla referencyjnego mikrofonu, zdefiniowane jako:

$$\tilde{\mathbf{v}}^{(k)} = \mathbf{v}^{(k)} / v_q^{(k)}. \tag{3.2}$$

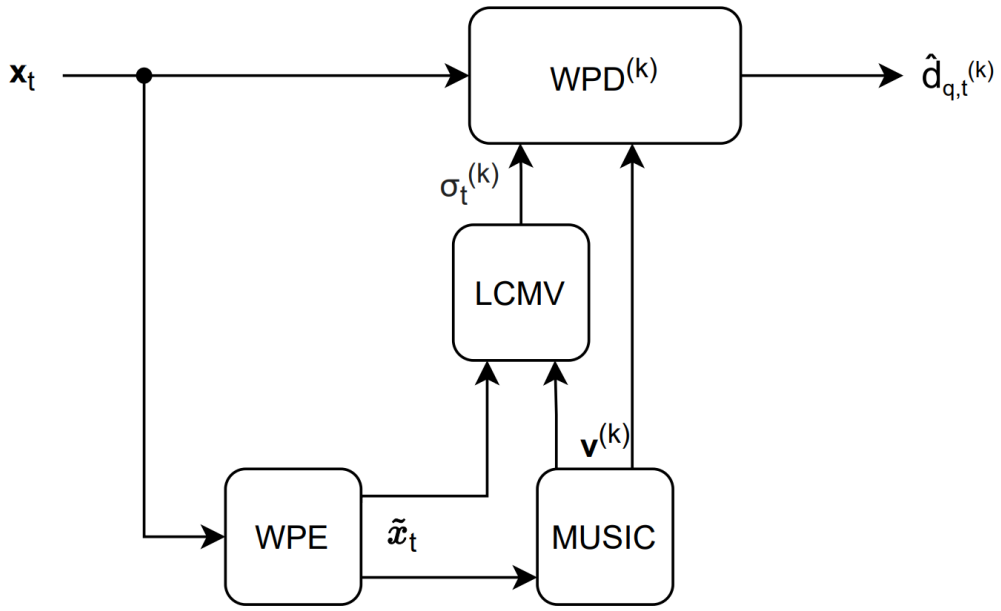
  Miary kątów, z których nadawany jest sygnał obliczone są przy użyciu algorytmu MUSIC. Dzięki tym kątom oraz znajomości geometrii macierzy mikrofonowej, można wyznaczyć wektory sterujące.

Można zatem wyróżnić następujące kroki potrzebne do rozwiązania problemu postawionego w sekcji 2.2:

- obliczenie STFT z sygnału wejściowego,
- aplikacja filtru WPE w celu wstępnej eliminacji pogłosu,

- zastosowanie algorytmu MUSIC w celu obliczenia DOA,
- aplikacja metody kształtowania wiązki LCMV w celu wstępnej estymacji mocy sygnału,
- zastosowanie metody splotowego kształtowania wiązki WPD w celu obliczenia szukanych, odseparowanych i pozbawionych pogłosu sygnałów,
- obliczenie odwrotnego STFT (ang. inverse STFT - iSTFT) z przefiltrowanego sygnału.

Schemat blokowy przetwarzania sygnału przedstawiony został na rysunku 3.1.



Rysunek 3.1: Schemat blokowy przetwarzania sygnału

3.2 Estymacja parametrów filtra przestrzennego WPD

3.2.1 Filtr WPE

Filtr WPE służy do eliminacji pogłosu z sygnału. Stosuje się go w torze przetwarzania, który odpowiedzialny jest za estymację parametrów wejściowych dla filtra WPD. W pracy wykorzystano funkcję biblioteczną, która implementuje filtr WPE. Więcej o filtrze WPE można przeczytać tutaj [2].

3.2.2 Algorytm lokalizacji MUSIC

Algorytm MUSIC służy do estymacji kątów nadchodzenia dźwięków (DOA), dzięki którym można obliczyć wektory sterujące. Jest to najbardziej popularny i najbardziej podstawowy algorytm z rodziny algorytmów estymujących DOA. Więcej o

MUSIC-u oraz ogólnie o estymacji DOA można znaleźć w [11]. W potoku przetwarzania danych, algorytm MUSIC operuje na danych po filtrze WPE. Celem filtra WPE jest usunięcie pogłosu, który znacznie utrudniałby lokalizację źródeł. Pierwszym krokiem w algorytmie MUSIC jest policzenie macierzy korelacji z sygnału wejściowego w dziedzinie STFT oznaczonego na rysunku 3.1 jako $\tilde{\mathbf{x}}_t \in \mathbb{C}^M$. Macierz korelacji można policzyć jako wartość oczekiwaną z sygnałów macierzy mikrofonowej:

$$\mathbf{R} = \sum_{t=1}^{T} \tilde{\mathbf{x}}_t \tilde{\mathbf{x}}_t^H, \tag{3.3}$$

gdzie H oznacza sprzężenie Hermitowskie. Następnie należy dokonać rozkłady macierzy $\mathbf{R}$ na wartości własne i wektory własne. Macierz $\mathbf{W}$ stanowi podmacierz macierzy znormalizowanych wektorów własnych uzyskaną poprzez pominięcie K pierwszych wektorów własnych odpowiadających podprzestrzeni sygnału w celu utworzenia macierzy podprzestrzeni szumu:

$$\mathbf{W} = [\mathbf{y}_{K+1}, \mathbf{y}_{K+2}, ..., \mathbf{y}_M], \tag{3.4}$$

gdzie $\mathbf{W} \in \mathbb{C}^{Mx(M-K)}$ to macierz podprzestrzeni szumu, a $\mathbf{y}_1, ..., \mathbf{y}_M$ to wektory własne macierzy $\mathbf{R}$ korespondujące z wartościami własnymi tej macierzy ułożonymi w kolejności malejącej. Należy pamiętać, że powyższe macierze obliczane są osobno dla każdej częstotliwości f. Miary kątów DOA obliczane są w sposób następujący:

$$\hat{\theta} = \arg\min_{\theta} \sum_{f} \mathbf{v}_f^H(\theta) \mathbf{W}_f \mathbf{W}_f^H \mathbf{v}_f(\theta), \tag{3.5}$$

gdzie $\hat{\theta}$ to wyestymowane miary kątów DOA (należy znaleźć K najmniejszych minimów lokalnych), a $\mathbf{v}_f(\theta) \in \mathbb{C}^M$ to względny wektor sterujący dla sygnału padającego z kierunku θ o częstotliwości f. Znając czasy ($\boldsymbol{\Delta}\mathbf{t}$) o jakie opóźniony jest sygnał na poszczególnych mikrofonach względem dowolnego punktu w przestrzeni, dla danego kierunku i danej częstotliwości, względny wektor sterujący można obliczyć jako:

$$\mathbf{v}(\theta) = \exp(-j2\pi f \boldsymbol{\Delta}\mathbf{t}). \tag{3.6}$$

Wektor sterujący $\mathbf{v}^{(k)}$ właściwy dla źródła k można uzyskać poprzez podstawienie do równania (3.6) kąta $\hat{\theta}_k$, czyli jednego z kątów DOA uzyskanych z równania (3.5). Następnie należy dokonać normalizacji wektorów sterujących względem dowolnego mikrofonu zgodnie z równaniem (3.2). Tak uzyskane wektory sterujące $\tilde{\mathbf{v}}^{(k)}$ będą przydatne w dalszych etapach przetwarzania sygnałów.

3.2.3 Metoda kształtowania wiązki LCMV

Metoda kształtowania wiązki LCMV umożliwia separację sygnałów. W filtrze LCMV można podać wybraną liczbę kierunków i wymagania na wzmocnienie jakie filtr ma mieć dla sygnału przychodzącego z danego kierunku. Filtr LCMV dany jest wzorem:

$$\mathbf{w} = \mathbf{I}^{-1}\tilde{\mathbf{V}}[\tilde{\mathbf{V}}^H\mathbf{I}^{-1}\tilde{\mathbf{V}}]^{-1}\mathbf{g}, \tag{3.7}$$

gdzie $\mathbf{w} \in \mathbb{C}^M$ to szukane współczynniki filtru LCMV, $\mathbf{I}$ to macierz jednostkowa (czasami [4] używa się w jej miejsce macierzy korelacji z równania (3.3)), $\tilde{\mathbf{V}} = [\tilde{\mathbf{v}}^{(1)}, \tilde{\mathbf{v}}^{(2)}, ..., \tilde{\mathbf{v}}^{(K)}] \in \mathbb{C}^{MxK}$ to macierz wektorów sterujących uzyskanych dzięki algorytmowi MUSIC, a $\mathbf{g} \in \mathbb{C}^M$ to wektor pożądanych wzmocnień. Gdy chcemy pozostawić sygnał z jednego źródła o indeksie k, a resztę źródeł stłumić, to należy ustawić k-tą wartość wektora $\mathbf{g}$ na 1, a pozostałe wartości na 0. W ten sposób uzyskać można współczynniki filtru LCMV $\mathbf{w}^{(k)}$. Tak jak i MUSIC, filtr LCMV operuje na sygnałach $\tilde{\mathbf{x}}_t$ pozbawionych pogłosu przez filtr WPE. Przy użyciu uzyskanego filtra, wstępną aproksymację pożądanego sygnału ($u_{q,t}^{(k)}$) uzyskać można w następujący sposób:

$$u_{q,t}^{(k)} = (\mathbf{w}^{(k)})^H \tilde{\mathbf{x}}_t. \tag{3.8}$$

Następnie, stosując równanie (3.1) można uzyskać szukaną estymację zmiennej w czasie mocy sygnałów źródłowych ($\sigma_t^{(k)}$).

3.3 Metoda splotowego kształtowania wiązki WPD

Głównym i ostatnim etapem algorytmu jest zastosowanie metody splotowego kształtowania wiązki WPD. WPD unifikuje metodę kształtowania wiązki MPDR oraz filtr WPE. Dokładne wyprowadzenie przedstawione jest w publikacji [6]. Aby przedstawić sposób działania algorytmu należy zdefiniować wydłużony wektor sygnału - $\bar{\mathbf{x}}_t \in \mathbb{C}^{ML}$, gdzie L oznacza długość filtra splotowego. Założona wartość L jest kompromisem między jakością działania algorytmu, a czasem wyliczania współczynników filtru. Wydłużony wektor sygnału dany jest wzorem:

$$\bar{\mathbf{x}}_t = [\mathbf{x}_t^T, \mathbf{x}_{t-D}^T, \mathbf{x}_{t-D-1}^T, ..., \mathbf{x}_{t-L-D+2}^T]^T \tag{3.9}$$

Do uzyskania pożądanych sygnałów w chwili t użytych zostanie L poprzesuwanych sygnałów mikrofonowych, z czego $L-1$ sygnałów z chwil wcześniejszych począwszy od chwili $t-D$. Sygnały wcześniejsze potrzebne są do usunięcia pogłosu. Zgodnie z modelem sygnału, sygnały z wymienionych chwil wcześniejszych są obecne w mikrofonach w chwili t w postaci pogłosu. Pierwszym krokiem w algorytmie WPD jest wyliczenie macierzy ważonej przestrzennej kowariancji $\mathbf{R}^{(k)} \in \mathbb{C}^{ML \times ML}$:

$$\mathbf{R}^{(k)} = \sum_{t=1}^{T} \frac{\bar{\mathbf{x}}_t \bar{\mathbf{x}}_t^H}{\sigma_t^{(k)} + \epsilon}, \tag{3.10}$$

gdzie $\sigma_t^{(k)}$ to moc k-tego sygnału źródłowego wyestymowana w poprzednim kroku przy pomocy filtru LCMV, ϵ to bardzo mała wartość, którą dodaje się w celu uniknięcia błędów numerycznych przy dzieleniu przez sygnały bliskie zeru. Następnie, współczynniki filtra WPD - $\bar{\mathbf{w}}^{(k)} \in \mathbb{C}^{ML}$ wylicza się w sposób następujący:

$$\bar{\mathbf{w}}^{(k)} = \frac{(\mathbf{R}^{(k)})^{-1} \bar{\mathbf{v}}^{(k)}}{(\bar{\mathbf{v}}^{(k)})^H (\mathbf{R}^{(k)})^{-1} \bar{\mathbf{v}}^{(k)}}, \tag{3.11}$$

gdzie $\bar{\mathbf{v}}^{(k)} = [(\tilde{\mathbf{v}}^{(k)})^T, 0, 0, ..., 0]^T \in \mathbb{C}^{ML}$ to wydłużony wektor sterujący. Mianownik równania (3.11) odpowiedzialny jest za normalizację współczynników filtra WPD.

Dzięki temu sygnał bezpośredni w mikrofonie referencyjnym q nie będzie przeskalowany. Analogicznie jak w równaniu (3.8), filtracji przy użyciu uzyskanego filtra dokonuje się w sposób następujący:

$$\hat{d}_{q,t}^{(k)} = (\bar{\mathbf{w}}^{(k)})^H \bar{\mathbf{x}}_t, \tag{3.12}$$

gdzie $\hat{d}_{q,t}^{(k)} \approx d_{q,t}^{(k)}$ to szukana aproksymacja sygnału bezpośredniego w mikrofonie referencyjnym q.

Rozdział 4

Szczegóły implementacyjne oraz optymalizacja obliczeniowa

4.1 Generator sygnału

Do wygenerowania pogłosowych, wieloźródłowych sygnałów, potrzebne są nieprzetworzone sygnały źródłowe oraz filtry RIR, które są szczególnymi przypadkami filtrów o skończonej odpowiedzi impulsowej (ang. Finite Impulse Response - FIR).

Nieprzetworzone sygnały źródłowe zostały zaczerpnięte z bazy sygnałów mowy EBU SQAM CD [15].

Filtry RIR uzyskane zostały przy pomocy generatora [9] dostępnego z biblioteki dostępnej dla języka programowania Python. Należy wygenerować po jednym filtrze RIR dla każdej pary: źródło sygnału - mikrofon. Informacje potrzebne do wygenerowania wspomnianych filtrów RIR to:

- wymiary pomieszczenia, w którym rozchodzi się dźwięk,
- czas trwania pogłosu związany ze współczynnikami odbicia ścian pomieszczenia,
- maksymalny rząd odbić - maksymalna liczba symulowanych odbić sygnału od ścian pomieszczenia, po ilu sygnał dochodzi do odbiornika,
- prędkość rozchodzenia się sygnału,
- współrzędne położenia źródeł sygnału,
- współrzędne położenia mikrofonów macierzy mikrofonowej,
- częstotliwość próbkowania. Żeby nie doszło do aliasingu, koniecznym jest, aby częstotliwość próbkowania spełniała warunek Nyquista:

$$f_s \geqslant 2B, \tag{4.1}$$

gdzie f_s to częstotliwość próbkowania, a B to szerokość pasma sygnału próbkowanego.

Sygnały źródłowe należy przefiltrować odpowiednimi filtrami RIR odpowiadającymi propagacji od każdego źródła do każdego z mikrofonów. Następnie, dla każdego mikrofonu należy dodać do siebie odpowiednie, przefiltrowane sygnały. Dodatkowo można dodać szum własny mikrofonowy jako addytywny biały szum Gaussowski (ang. Additive White Gaussian Noise - AWGN) [16]. Jest on nieskorelowany pomiędzy mikrofonami.

4.2 Implementacja algorytmu

Mając gotowy generator sygnału można zająć się implementacją algorytmu jednoczesnej separacji sygnałów i usuwania pogłosu opisanego w rozdziale 3.

W pierwszej fazie implementacji algorytmu, zaimplementowano jedynie algorytm WPD bez estymacji parametrów wejściowych czyli wektorów sterujących ($\mathbf{v}^{(k)}$) i zmiennej w czasie mocy poszczególnych sygnałów źródłowych otrzymanej w referencyjnym mikrofonie ($\sigma_t^{(k)}$).

W następnej fazie implementacji, po zweryfikowaniu poprawności działania filtru WPD w oparciu o idealne wartości parametrów, zaimplementowano algorytmy WPE, MUSIC oraz LCMV w celu estymacji wspomnianych powyżej parametrów wejściowych.

4.2.1 Krótko-czasowa transformata Fouriera STFT

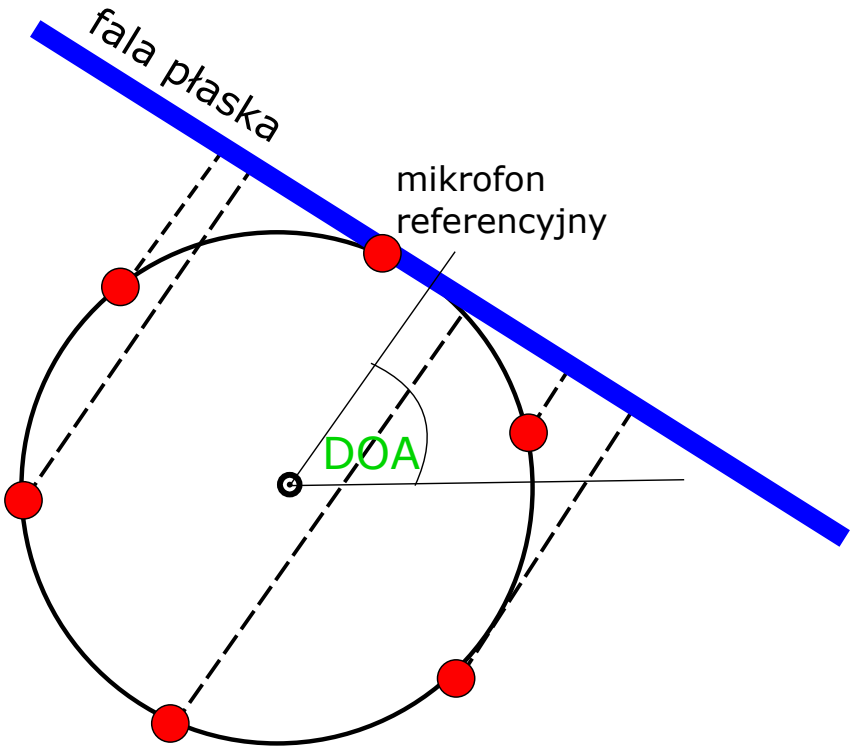
STFT polega na dzieleniu sygnału na krótkie ramki czasowe, a następnie transformacji każdej ramki przy użyciu transformaty Fouriera. Parametry, które otrzymuje na wejściu funkcja wykonująca STFT to:

- rodzaj okna, które służy do ekstrahowania danej ramki czasowej próbek z całego sygnału,
- długość ramki czasowej,
- odległość między początkami sąsiednich ramek czasowych próbek próbek.

4.2.2 Wektory sterujące

Znając DOA źródeł sygnału, geometrię macierzy mikrofonowej i prędkość rozchodzenia się sygnału można wyliczyć wektory sterujące. W tym celu należy policzyć odległości pomiędzy mikrofonem referencyjnym a pozostałymi mikrofonami, które pokonuje fala płaska nadchodząca z danego kierunku DOA. Następnie dzieląc uzyskane odległości przez prędkość sygnału uzyskuje się opóźnienia czasowe sygnału w poszczególnych mikrofonach względem mikrofonu referencyjnego dla danych DOA ($\mathbf{\Delta t}^{(k)}$). W następnym kroku, w celu uzyskania wektorów sterujących, uzyskane opóźnienia należy podstawić do równania (3.6).

Rysunek 4.1 pokazuje skąd biorą się różnice w odległości i jak zależą one od DOA. Dla DOA zaznaczonego na rysunku, sygnał reprezentowany jako fala płaska dociera najpierw do mikrofonu referencyjnego. Następnie, fala opóźniona o odpowiednią ilość czasu dociera do pozostałych mikrofonów. Linią przerywaną zaznaczono odległości



Dystans między mikrofonem a mikrofonem referencyjnym dla danego DOA

Rysunek 4.1: Różnice w sygnale na poszczególnych mikrofonach macierzy mikrofonowej

świadczące o tym o ile opóźniony będzie sygnał na poszczególnych mikrofonach w stosunku do mikrofonu referencyjnego. Odległości te służą do policzenia wektorów sterujących.

4.2.3 Zmienna w czasie moc sygnałów źródłowych w referencyjnym mikrofonie

W pierwszej fazie implementacji algorytmu zastosowano idealną wartość parametru $\sigma_t^{(k)}$ - czyli taką, którą można otrzymać znając wszystkie parametry pokoju oraz czyste sygnały źródłowe. Aby ją wyliczyć, należy spleść sygnał źródłowy z początkową częścią odpowiedzi impulsowej pomieszczenia odpowiadającą dźwiękowi bezpośredniemu. Następnie, po przefiltrowaniu otrzymanym filtrem k-tego sygnału źródłowego, otrzymuje się sygnał, który odebranoby w referencyjnym mikrofonie, gdyby nie było pogłosu oraz innych źródeł. Obliczając moc zgodnie z równaniem (3.1) otrzymać można szukany, idealny (nie jest to aproksymacja) parametr wejściowy $\sigma_t^{(k)}$.

4.2.4 WPD

Kolejnym etapem była implementacja filtru WPD. Pierwszym krokiem była implementacja macierzy ważonej przestrzennej kowariancji $\mathbf{R}^{(k)}$ zgodnie z równaniem (3.10). Z początku, w przytoczonym równaniu, nie było parametru ϵ. Dopiero w trakcie implementacji okazało się, że jest on kluczowy do poprawnego działania algorytmu. Przy braku tego parametru zdarzało się, że dzielono przez bardzo małą liczbę co niekorzystnie wpływało potem na całą sumę i fałszowało macierz kowariancji. Następnie, należało zaimplementować filtr przestrzenny $\bar{\mathbf{w}}^{(k)}$ zgodnie z równaniem (3.11). Kolejno, w celu otrzymania odseparowanego i pozbawionego pogłosu sygnału ($\hat{d}_{q,t}^{(k)} \approx d_{q,t}^{(k)}$) należało zaimplementować równanie (3.12). Po policzeniu odwrotnej krótko-czasowej transformaty Fouriera zweryfikowano czy otrzymane sygnały, a zatem filtr WPD, został zaimplementowany poprawnie.

4.2.5 Estymacja parametrów wejściowych

Po zweryfikowaniu poprawności działania dotychczas zaimplementowanego algorytmu separacji i usuwania pogłosu przy znanych idealnych parametrach wejściowych nadszedł czas na estymację tych parametrów to znaczy kątów DOA oraz mocy sygnałów źródłowych $\sigma_t^{(k)}$.

Pierwszym elementem jest zastosowanie filtru WPE w celu usunięcia późnego pogłosu pomieszczenia z nagrań mikrofonowych. Jako implementację filtru WPE wykorzystano funkcję z biblioteki dostępnej dla języka programowania Python.

Do estymacji DOA zaimplementowano algorytm MUSIC opisany w podrozdziale 3.2.2.

Następnie, należało zaimplementować algorytm umożliwiający oszacowanie mocy źródła dla każdego punktu czasowo-częstotliwościowego ($\sigma_t^{(k)}$). Jako pierwsze podejście, w pracy posłużono się metodą kształtowania wiązki MPDR z iteracyjną estymacją mocy [17]. Równania pozwalające obliczyć filtr MPDR przypominają analogiczne

równania (3.10), (3.11) oraz (3.12) z tym, że używa się niewydłużonych wektorów sterujących oraz sygnału:

$$\mathbf{R}^{(k)} = \sum_{t=1}^{T} \frac{\mathbf{x}_t \mathbf{x}_t^H}{\hat{\sigma}_t^{(k)} + \epsilon}, \tag{4.2}$$

$$\mathbf{w}^{(k)} = \frac{(\mathbf{R}^{(k)})^{-1}\mathbf{v}^{(k)}}{(\mathbf{v}^{(k)})^H (\mathbf{R}^{(k)})^{-1}\mathbf{v}^{(k)}}, \tag{4.3}$$

$$\hat{u}_{q,t}^{(k)} = (\mathbf{w}^{(k)})^H \mathbf{x}_t, \tag{4.4}$$

$$\hat{\sigma}_t^{(k)} = |\hat{u}_{q,t}^{(k)}|^2. \tag{4.5}$$

Wartość $\hat{\sigma}_t^{(k)}$ to estymacja szukanego parametru wejściowego $\sigma_t^{(k)}$. Jako że występuje ona na wejściu algorytmu w równaniu (4.2) oraz na wyjściu algorytmu, można obliczać ją iteracyjnie. Jako $\hat{\sigma}_t^{(k)}$ w kroku pierwszym iteracji bierze się moc pogłosowego, wieloźródłowego sygnału otrzymanego w mikrofonie referencyjnym ($x_{q,t}$). Ta metoda jednak, nie okazała się skuteczna. Cały proces liczenia kolejnych iteracji był czasochłonny, a poziom separacji nie był satysfakcjonujący i w rezultacie obliczona moc $\sigma_t^{(k)}$ w znacznym stopniu pogarszała późniejsze działanie algorytmu WPD.

Dlatego na późniejszym etapie pracy użyto metody kształtowania wiązki, w której oprócz kierunku, który ma być pozostawiony w wyniku separacji, podaje się też kierunek, który ma zostać stłumiony. Finalnie zdecydowano się na implementację metody kształtowania wiązki LCMV, która posiada pożądaną własność. W celu implementacji tej metody skorzystano z algorytmu opisanego w podrozdziale 3.2.3.

4.3 Optymalizacja obliczeniowa

4.3.1 Unikanie pętli

Głównym problemem dotyczącym czasu wykonywania algorytmu była potrzebna liczba obliczeń. Należy pamiętać, że wiele części algorytmu wykonywanych jest dla każdej częstotliwości, dla każdej ramki czasowej, dla każdego źródła sygnału i dla każdego mikrofonu. Naturalnym wydawałoby się więc zrobienie bardzo dużej liczby pętli, często pętli zagnieżdżonych. Nie jest to jednak dobre podejście. Biblioteka Numpy zoptymalizowana jest w kierunku wykonywania operacji na tablicach wielowymiarowych. Bardzo wiele funkcji takich jak np. mnożenie macierzowe lub odwracanie macierzy zaimplementowanych jest w taki sposób, że jako ich argument można podać tablicę o dowolnie dużej liczbie wymiarów, czyli tak naprawdę zbiór macierzy. Wystarczy jedynie określić, które dwa wymiary wielowymiarowej tablicy reprezentują faktyczne wymiary macierzy, na których chcemy przeprowadzać daną operację. Jakiekolwiek pętle z kolei, spowalniają obliczenia. Z tego powodu, liczbę pętli ograniczono do minimum.

Dla przykładu: w równaniu (3.11) obliczany jest filtr przestrzenny $\bar{\mathbf{w}}^{(k)}$. Równanie to trzeba policzyć dla każdej częstotliwości i dla każdego źródła sygnału (k), co sugerowałoby użycie dwóch pętli. W implementacji tego równania nie pojawia

się jednak żadna pętla. Zamiast pętli, tablice reprezentujące macierze mają poszerzone wymiary. Tablica reprezentująca wektory sterujące ($\bar{\mathbf{v}}^{(k)}$) jest tablicą 3-wymiarową, gdzie pierwszy wymiar to częstotliwość, drugi wymiar to numer źródła, a trzeci wymiar to elementy poszczególnego wektora. Jest to zatem tak naprawdę zbiór wszystkich wektorów sterujących, które użyte są w algorytmie. Analogicznie, tablica reprezentująca macierz kowariancji ($\mathbf{R}^{(k)}$) jest reprezentowana przez tablicę 4-wymiarową, gdzie wymiar pierwszy to częstotliwość, wymiar drugi to numer źródła a pozostałe dwa wymiary to wymiary pojedynczej macierzy kowariancji.

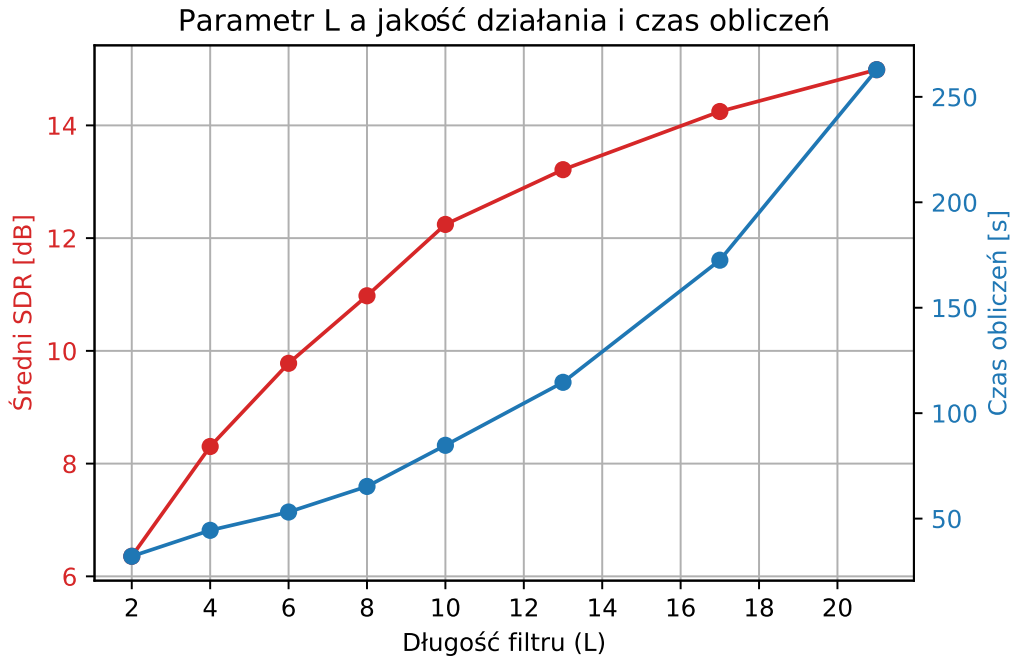
Alternatywnie, gdyby implementować to samo równanie przy pomocy dwóch zagnieżdżonych pętli, czas wykonywania wydłużyłby się. Dla 257 częstotliwości i dwóch źródeł implementacja z pętlami jest wolniejsza o około 20%. Z drugiej strony, unikanie używania pętli wiąże się z większym użyciem pamięci RAM, gdyż potrzeba wtedy przechowywać w pamięci RAM większe tablice.

Podobnie zachowuje się również czas liczenia macierzy kowariancji $\mathbf{R}^{(k)}$ zgodnie z równaniem (3.10). Jest to część algorytmu, której obliczanie zajmuje zdecydowanie największą ilość czasu (około 88% całego czasu). W tym przypadku implementacja, z pętlami używałaby aż trzech zagnieżdżonych pętli: pętla po częstotliwościach, pętla po źródłach sygnału, pętla po czasie. Niemożliwym jest w realnym scenariuszu, zaimplementowanie tego równania bez użycia jakiejkolwiek pętli, gdyż w tym celu potrzebne by były dziesiątki GB pamięci RAM. Da się jednak zaimplementować to równanie używając tylko jednej pętli po częstotliwościach. W stosunku do takiej implementacji, implementacja używająca trzech zagnieżdżonych pętli okazuje się być o około 42% wolniejsza.

4.3.2 Długość filtra eliminującego pogłos

Czas trwania obliczeń zależy w bardzo dużym stopniu od parametru L, czyli długości filtra eliminującego pogłos. Na filtr składa się jedna grupa M współczynników odpowiedzialnych za zostawianie sygnału bezpośredniego i $L-1$ grup po M współczynników odpowiedzialnych za eliminację pogłosu z kolejnych $L-1$ ramek czasowych. Jednocześnie, gdy filtr będzie zbyt krótki, późny pogłos nie będzie eliminowany co w znacznym stopniu wpłynie na jakość działania algorytmu. Warto zauważyć, że pogłos przychodzi z bardzo różnych kierunków, więc gdy pogłos nie będzie poprawnie eliminowany, to również separacja sygnałów nie będzie odbywała się w sposób poprawny.

Kluczowym, z punktu widzenia optymalizacji obliczeniowej, jest dobranie takiej wartości parametru L, która będzie kompromisem pomiędzy długim czasem wykonywania a jakością działania algorytmu. W tym celu zmierzono jak zmienia się czas wykonywania algorytmu w zależności od różnych wartości parametru L. Jakość działania algorytmu w każdym z przypadków oceniono przy pomocy stosunku sygnału do zniekształceń (ang. Signal to Distortion Ratio - SDR) na wyjściu algorytmu, o którym więcej informacji można znaleźć w [13]. Obliczono SDR dla każdego źródła sygnału a następnie uśredniono te wartości otrzymując średni SDR. Wyniki zaprezentowano na rysunku 4.2. Przedstawia on porównanie tego jak zmienia się wykres czasu obliczeń oraz średniego SDR w zależności od wartości parametru L.



Rysunek 4.2: Porównanie jakości i czasu wykonywania algorytmu w zależności od L

Biorąc pod uwagę oba wykresy, wybrano wartość $L = 13$ jako kompromis między SDR, który chcemy zmaksymalizować a czasem obliczeń, który chcemy zminimalizować.

W powyższym eksperymencie separowano dwa źródła o czasie trwania sygnału równym 7 sekund, częstotliwości próbkowania równej 16 kHz i przy pomocy macierzy mikrofonowej posiadającej 10 mikrofonów. Żeby algorytm działał w czasie rzeczywistym, należałoby zastosować małą wartość parametru L oraz zmniejszyć częstotliwość próbkowania lub liczbę mikrofonów. Zmniejszyłoby to czas obliczeń kosztem jakości separacji i usuwania pogłosu.

Rozdział 5

Weryfikacja eksperymentalna oraz jej rezultaty

5.1 Parametry używane w eksperymentach oraz miary ewaluacyjne

5.1.1 Informacje ogólne

W celu zweryfikowania działania zaimplementowanego algorytmu przeprowadzono szereg doświadczeń. Jakość działania implementacji oceniona została na podstawie następujących miar oceniających jakość przetworzenia sygnału na podstawie sygnału idealnego, jaki chcielibyśmy otrzymać oraz sygnału otrzymanego, czyli aproksymacji sygnału idealnego. Pierwsze trzy miary dotyczą ewaluacji separacji sygnałów, a kolejne dwie wskazują na skuteczność usuwania pogłosu pomieszczenia:

- Stosunek sygnału do zniekształceń (ang. Signal to Distortion Ratio - SDR) [13] - miara, która mówi w jakim stopniu udało się odtworzyć sygnały źródłowe,
- Stosunek sygnału do zakłóceń (ang. Signal to Interference Ratio - SIR) [13] - miara, która mówi w jakim stopniu udało się odseparować źródła oraz usunąć pogłos,
- Stosunek sygnału do artefaktów (ang. Signal to Artifact Ratio - SAR) [13] - miara, która mówi w jakim stopniu udało się uniknąć wprowadzenia do sygnału artefaktów w trakcie jego przetwarzania,
- Percepcyjna ewaluacja jakości mowy (ang. Perceptual Evaluation of Speech Quality - PESQ) [12] - miara, która bazuje na ludzkiej percepcji; mówi ona w jakim stopniu przetworzony sygnał mowy jest zrozumiały,
- Odległość Kepstralna (ang. Cepstral Distance - CD) [14] - miara, która mówi jak bardzo pogłos „oddalił" sygnał otrzymany od sygnału oryginalnego (im mniejsza wartość CD, tym lepiej).

To, jak będzie działał algorytm zależy od wielu parametrów, którymi można manipulować i badać jaki mają one wpływ na separację źródeł i eliminację pogłosu.

Parametry te można podzielić na dwie kategorie: parametry zewnętrze oraz parametry wewnętrzne - zależne od implementacji algorytmu.

5.1.2 Parametry zewnętrzne

Parametry zewnętrzne to takie, które zależą od środowiska, w którym dokonywane są doświadczenia oraz od sprzętu, którym odbierany jest sygnał. Algorytm służący do jednoczesnej separacji i eliminacji pogłosu nie ma wpływu na parametry zewnętrzne. Do parametrów zewnętrznych należą:

- Wymiary pomieszczenia, w którym rozchodzi się dźwięk. W tej pracy jest to zawsze pokój o wymiarach 10 m x 10 m x 3 m wysokości.
- Czas pogłosu.
- Maksymalny rząd odbić. W tej pracy rząd ten jest ustalony na stałe na nieskończoność. Oczywiście, nie będzie występować nieskończenie wiele odbić, gdyż liczba odbić ograniczona jest również przez czas trwania pogłosu.
- Prędkość rozchodzenia się sygnału. W tej pracy parametr ten został ustawiony na 343 [m/s], czyli przybliżoną wartość prędkości dźwięku.
- Liczba źródeł sygnału (K)
- Położenie źródeł sygnału. W tej pracy źródła sygnału zawsze położone są na wysokości 1 m, i w odległości 2 m od macierzy mikrofonowej - ważne, aby źródła nie były zbyt blisko odbiornika, żeby można było założyć, że sygnał, który dochodzi do odbiornika jest falą płaską. Ważne też, aby miary kątów DOA nie były zbyt blisko siebie, gdyż wtedy algorytm MUSIC będzie miał problem z rozróżnieniem źródeł. Zbyt bliskie położenie źródeł wymusza też bardzo „ostrą” (wąskie listki główne) charakterystykę filtra przestrzennego, co wpływa negatywnie na jakość działania algorytmu separacji.
- Sygnał odbierany, a w szczególności jego charakter. Czy jest to sygnał szerokopasmowy, czy jest to mowa, muzyka etc. W tej pracy badano sygnały o pasmie częstotliwości z przedziału [0, 8] [kHz].
- Długość odbieranego sygnału. Odebrany sygnał można przetwarzać w całości lub dzielić na kilkusekundowe bloki i każdy blok przetwarzać osobno. W tej pracy użyto sygnałów o długości 7 s, których nie dzielono na bloki.
- Częstotliwość próbkowania. W tej pracy częstotliwość próbkowania wynosiła 16 kHz, a szerokość pasmowa sygnału wynosiła 8 kHz, czyli następowało próbkowanie krytyczne.
- Położenie macierzy mikrofonowej. W tej pracy macierz mikrofonowa jest zawsze umieszczona w odległości 5.5 m od dwóch ścian, na wysokości takiej samej jak źródła sygnału czyli 1 m - macierz nie leży idealnie w środku pokoju.

- Liczba mikrofonów w macierzy mikrofonowej (M). Liczba ta musi być większa niż liczba źródeł sygnału. W tej pracy ta liczba zawsze wynosić będzie 10.
- Kształt macierzy mikrofonowej. W tej pracy jest to zawsze okrąg o promieniu 4 cm, na którym w równych odstępach ułożone są mikrofony. Żeby nie doszło do zjawiska aliasingu przestrzennego [7], maksymalna odległość między mikrofonami w macierzy mikrofonowej powinna być mniejsza niż połowa długości najkrótszej fali w sygnale:

$$d_{\max} < \frac{\lambda_{\min}}{2}, \tag{5.1}$$

$$d_{\max} < \frac{v}{2f_{max}}, \tag{5.2}$$

  gdzie $d_{\max}$ to największa odległość między mikrofonami macierzy mikrofonowej, $\lambda_{\min}$ to minimalna długość fali w sygnale, v to prędkość rozchodzenia się sygnału, a f_{max} to największa częstotliwość składowa w sygnale. W praktyce jednak, lepiej sprawdzają się macierze mikrofonowe o nieco większych wymiarach z uwagi na to, że wtedy można zaobserwować większe zmiany w wartościach sygnału w poszczególnych mikrofonach w danej chwili czasowej.
- Stosunek sygnału do szumu (ang. Signal to Noise Ratio - SNR). W tej pracy ustawiony jest minimalny poziom szumu - SNR wynosi 50 dB.

5.1.3 Parametry wewnętrzne

Parametry wewnętrzne to takie, które zależą od algorytmu. Implementując algorytm można mieć wpływ na jakość jego działania poprzez odpowiednie dobranie parametrów wewnętrznych. Do parametrów wewnętrznych należą:

- Parametry STFT:
  - Rodzaj okna. W tej pracy zastosowano okno Hanna [16].
  - Długość ramki czasowej. W tej pracy wartość tego parametru wynosi 32 ms.
  - Odległość między początkami sąsiednich ramek czasowych. W tej pracy ustalono tę wartość na 8 ms, co oznacza, że sąsiednie ramki czasowe nachodzą na siebie w obszarze o czasie trwania równym 24ms.
- Parametry WPE oraz WPD:
  - Ilość ramek czasowych, w których znajduje się pożądany, krótki pogłos (D). W tej pracy założono, że liczba ta wynosi 1, to znaczy, że każdy pogłos obecny poza ramką czasową, w której pobrano próbkę sygnału, uznawany jest za pogłos niepożądany i należy go wyeliminować.
  - Długość filtra eliminującego pogłos (L). W tej pracy wartość ta ustawiona została na 13.
  - Współczynnik ϵ, czyli liczba odpowiedzialna ze eliminację potencjalnych błędów numerycznych w równaniu (3.10). W tej pracy wartość ta została ustawiona na 10^{-5}.

5.1.4 Parametry które będą zmieniane

W następnych sekcjach tego rozdziału przedstawione zostaną rezultaty uzyskane w wyniku kilku eksperymentów. Wszystkie parametry zostały podane w podsekcjach powyżej. Jedyne parametry, które będą zmieniane w zależności od rodzaju eksperymentu to:

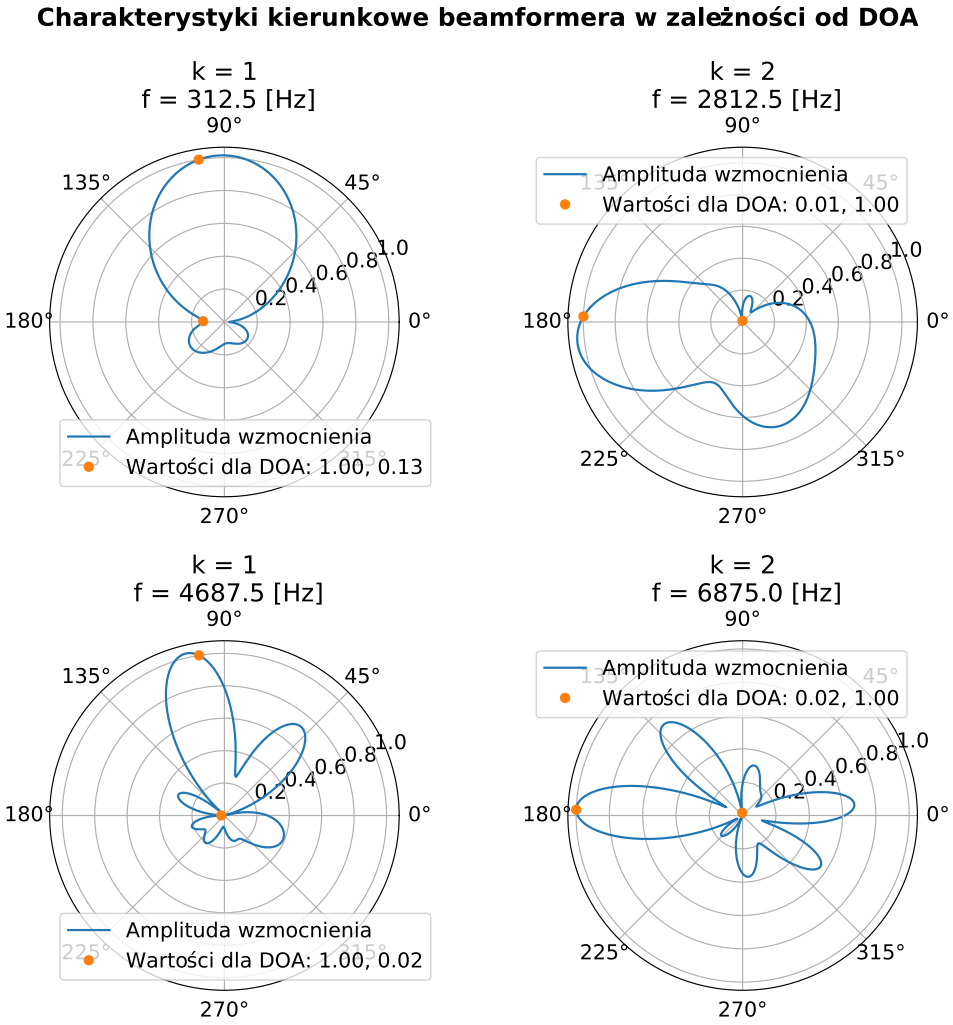
- czas pogłosu,
- liczba źródeł sygnału (K),
- rodzaje sygnałów źródłowych.

5.2 Skuteczność działania WPD dla idealnych parametrów wejściowych

Najlepsze wyniki można otrzymać, gdy weźmie się idealne wartości parametrów wejściowych filtra, pomijając etap estymacji tych parametrów. Tymi parametrami jest moc sygnału $\sigma_t^{(k)}$ oraz miary kątów DOA. Rezultaty przedstawione w tej sekcji uzyskano przy czasie pogłosu równym 0.4 s oraz liczbie źródeł sygnału równej 2. Kąty DOA dla tych źródeł wynosiły 99° oraz 178°. Sygnałami źródłowymi były odpowiednio mowa kobieca oraz mowa męska.

5.2.1 Beamforming

Biorąc pierwsze M współczynników rozszerzonego filtra przestrzennego $\bar{\mathbf{w}}^{(k)}$ omawianego w równaniu (3.11), można otrzymać zwykły filtr przestrzenny $\mathbf{w}_0^{(k)}$ operujący na pojedynczej ramce czasowej sygnału w dziedzinie STFT. Dla takiego filtru przestrzennego wyrysować można jego charakterystykę kierunkową. Rysunek 5.1 przedstawia charakterystyki kierunkowe uzyskanego filtru przestrzennego dla kilku źródeł sygnału (k) oraz dla kilku przykładowych częstotliwości (f). Oś z kątami na wykresach przedstawia kierunki przychodzenia sygnału a na osi wartości można odczytać amplitudę wzmocnienia dla danego filtru przestrzennego w każdym kierunku. Wyrysowane charakterystyki obliczono jako amplituda iloczynu filtru przestrzennego ze znormalizowanymi wektorami sterującymi dla kierunków (θ), czyli wyrysowano $|\mathbf{w}_0^{(k)}\tilde{\mathbf{v}}(\theta)|$ w funkcji kąta θ. Wykres ten mówi, jaka część sygnału bezpośredniego z danego kierunku θ zostaje zachowana w sygnale przefiltrowanym po dokonaniu filtracji przy użyciu filtru przestrzennego. Kolorem pomarańczowym zaznaczono wartości amplitudy dla kierunków, w których umieszczone są źródła sygnału. Można zauważyć, że gdy główna wiązka filtru przestrzennego skierowana jest w danym kierunku k, to jego wartość w tym kierunku wynosi 1 - całość sygnału jest zostawiana (podstawowe wymaganie dla separacji źródeł), natomiast drugie źródło jest tłumione o około 10 - 20 dB. Niskie częstotliwości nie są tak dobrze tłumione jak wyższe częstotliwości.

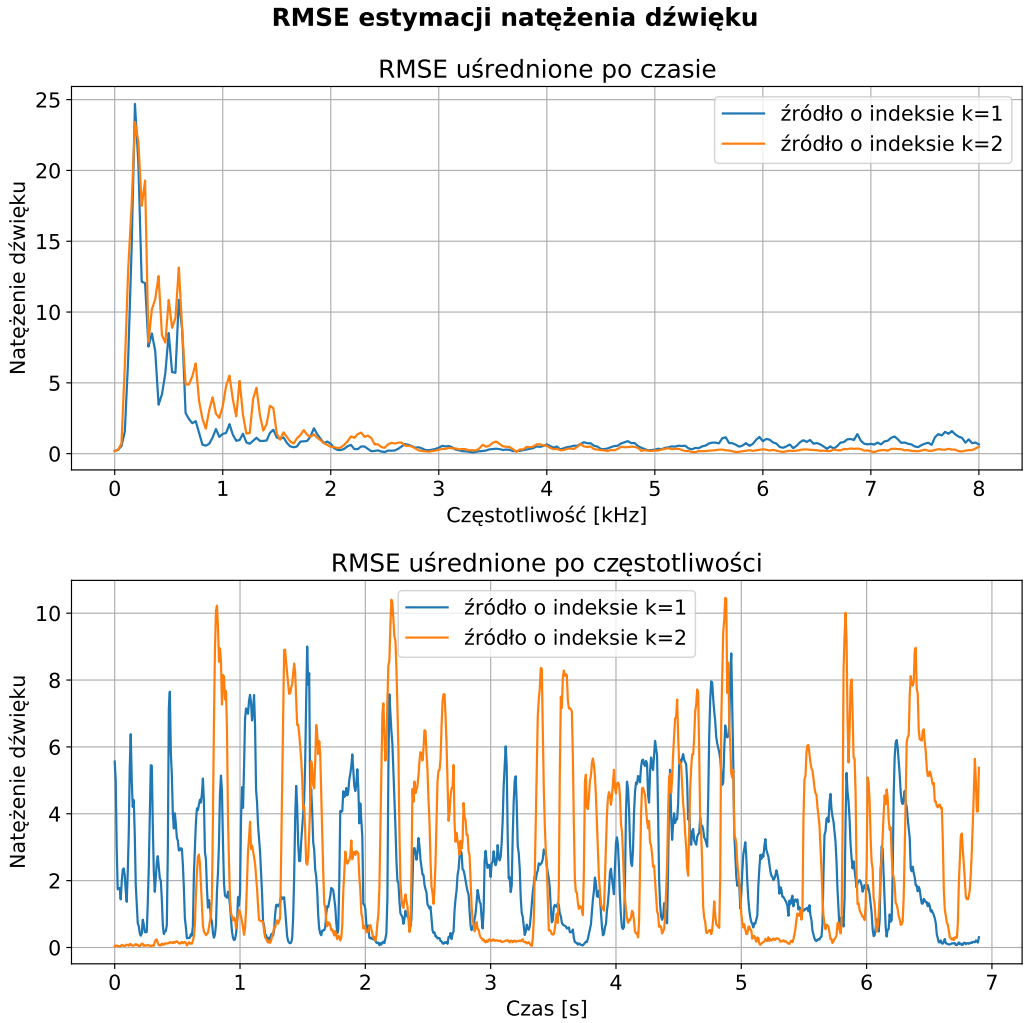


Rysunek 5.1: Charakterystyki kierunkowe filtru przestrzennego

5.2.2 Błędy estymacji

Porównanie sygnału wyestymowanego z sygnałem oczekiwanym można przeprowadzić przy pomocy błędu średniokwadratowego (ang. Root Mean Square Error - RMSE) estymacji natężenia dźwięku, czyli $\sqrt{E(|\hat{d}_q^{(k)} - d_q^{(k)}|^2)}$, gdzie uśrednianie może odbywać się po częstotliwości lub po czasie. Rysunek 5.2 przedstawia RMSE estymacji natężenia dźwięku dla obu źródeł uśrednione kolejno po czasie i częstotliwości.

Na wykresach widać, że największe problemy występują z estymacją sygnału w niskich częstotliwościach. Efektu tego można było się spodziewać patrząc na charakterystykę kierunkową filtra przestrzennego dla częstotliwości 312.5 Hz z rysunku 5.1. Dla niskich częstotliwości fala przychodząca z danego kierunku jest długa i co za tym idzie wartości sygnału odebranego w poszczególnych mikrofonach różnią się



Rysunek 5.2: Błędy estymacji

bardzo nieznacznie, co utrudnia cały proces ekstrahowania z sygnału odebranego jego przestrzennych właściwości.

5.2.3 Miary jakości separacji sygnałów i eliminacji pogłosu

W poniższej tabeli zaprezentowano obliczone miary oceniające jakość działania algorytmu. Użyte miary zostały opisane w sekcji 5.1.1.

Tabela 5.1: Miary jakości

Miary jakości działania algorytmu WPD bez estymacji parametrów wejściowych					
Numer źródła (k)	SDR [dB]	SIR [dB]	SAR [dB]	PESQ	CD
	Na wejściu				
$k = 1$	-4.31	-3.23	7.19	1.05	6.74
$k = 2$	1.30	3.35	7.19	1.13	6.75
	Na wyjściu				
$k = 1$	18.33	37.46	18.38	2.54	3.06
$k = 2$	15.60	37.78	15.65	2.66	2.92
	Poprawa				
$k = 1$	22.64	40.70	11.19	1.49	3.68
$k = 2$	14.30	32.42	8.46	1.53	3.83

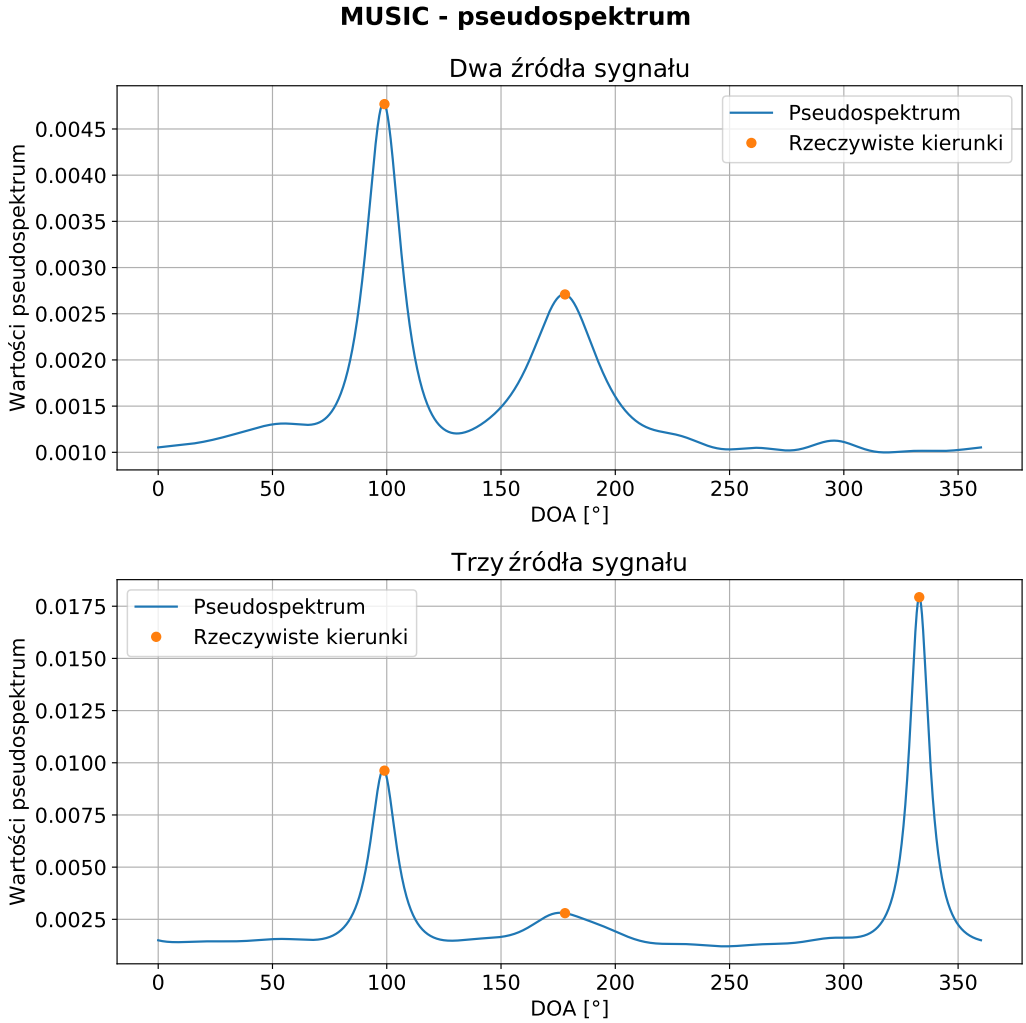
Statystyki dla wejścia to statystyki dla sygnału odbieranego w mikrofonie referencyjnym ($x_{q,t}$). Statystyki dla wyjścia zostały obliczone dla wyestymowanego sygnału pożądanego źródła $\hat{d}_{q,t}^{(k)}$. W dolnej części tabeli przedstawiono poprawę wyznaczoną jako różnica miary na wyjściu oraz na wejściu filtra. W ogólnym ujęciu sygnały na wyjściu są wysokiej jakości. Szczególnie dobrze zadziałało usuwanie interferencji, o czym świadczy wysoki SIR. Biorąc sygnał dla jednego źródła, zupełnie nie słychać źródła drugiego. Słychać natomiast, że sygnał jest lekko zniekształcony o czym świadczy nie tak wysoki SAR.

5.3 Estymacja parametrów wejściowych

W tej sekcji zbadana zostanie jakość estymacji parametrów wejściowych. Rezultaty przedstawione w tej sekcji, podobnie jak w poprzedniej sekcji, uzyskano przy czasie pogłosu równym 0.4 s.

5.3.1 MUSIC

Do estymacji kątów DOA użyto algorytmu MUSIC. Na rysunku 5.3 zaprezentowano pseudospektra przestrzenne w funkcji DOA uzyskane w wyniku działania algorytmu MUSIC. Pseudospektrum to odwrotność parametru minimalizowanego w równaniu (3.5). Estymowane kierunki to współrzędne „x" największych K maksimów lokalnych pseudospektrum. Na pomarańczowo zaznaczono wartości pseudospektrum dla prawdziwych DOA. Poniższe wykresy uzyskano przeprowadzając eksperyment kolejno dla dwóch źródeł sygnału o DOA równych 99° i 178° oraz dla trzech źródeł sygnału o DOA równych 99°, 178° i 333°. Wszystkie sygnały źródłowe były mową.



Rysunek 5.3: Działanie algorytmu MUSIC

Można zauważyć, że gdy źródła sygnału są oddalone od siebie, to algorytm działa bardzo dobrze. Wyestymowane DOA to w pierwszym przypadku 98.5° i 177.5°, a w drugim przypadku 98.5°, 175.9° i 332.9°. Problem może pojawić się gdy źródła sygnału są zbyt blisko siebie. Wtedy może zdarzyć się sytuacja, w której dwa bliskie maksima lokalne zlewają się w jedno maksimum lokalne o DOA będącym wypadkową dwóch rzeczywistych DOA. Taki błąd uniemożliwiłby dalsze działanie algorytmu.

5.3.2 Estymacja zmiennej w czasie mocy sygnałów źródłowych

Drugim z parametrów wejściowych, który należy wyestymować są zmienne w czasie moce sygnałów źródłowych $\sigma_t^{(k)}$. Zgodnie z równaniem (3.1), szukaną moc $\sigma_t^{(k)}$ oblicza się ze wstępnej aproksymacji sygnału pożądanego $(u_{q,t}^{(k)})$. Rzeczoną aproksymację można ewaluować przy pomocy takich samych narzędzi jak końcową estymację sy-

gnałów źródłowych ($\hat{d}_{q,t}^{(k)}$). W tabeli poniżej prezentuję obliczone miary oceniające jakość aproksymacji sygnałów źródłowych ($d_{q,t}^{(k)}$) obliczonym sygnałem $u_{q,t}^{(k)}$. Są to te same miary co w podsekcji 5.2.3. Eksperyment przeprowadzony został dla tych samych warunków co eksperymenty w sekcji 5.2.

Tabela 5.2: Miary jakości

Miary jakości estymacji mocy sygnałów źródłowych					
Numer źródła (k)	SDR [dB]	SIR [dB]	SAR [dB]	PESQ	CD
	Na wejściu				
$k=1$	-4.31	-3.23	7.19	1.05	6.74
$k=2$	1.30	3.35	7.19	1.13	6.75
	Na wyjściu				
$k=1$	9.04	12.91	11.54	1.30	4.75
$k=2$	10.17	17.28	11.19	1.94	4.65
	Poprawa				
$k=1$	13.35	16.15	4.36	0.25	1.99
$k=2$	8.87	13.92	4.00	0.81	2.00

Można zauważyć, że estymacja mocy jest przyzwoita, ale w porównaniu do wyników estymacji ze znanymi parametrami wejściowymi otrzymane wartości są zdecydowanie niższe. Negatywnie wpływa to później na działanie algorytmu WPD, który korzysta z wyestymowanych w tej sekcji parametrów wejściowych.

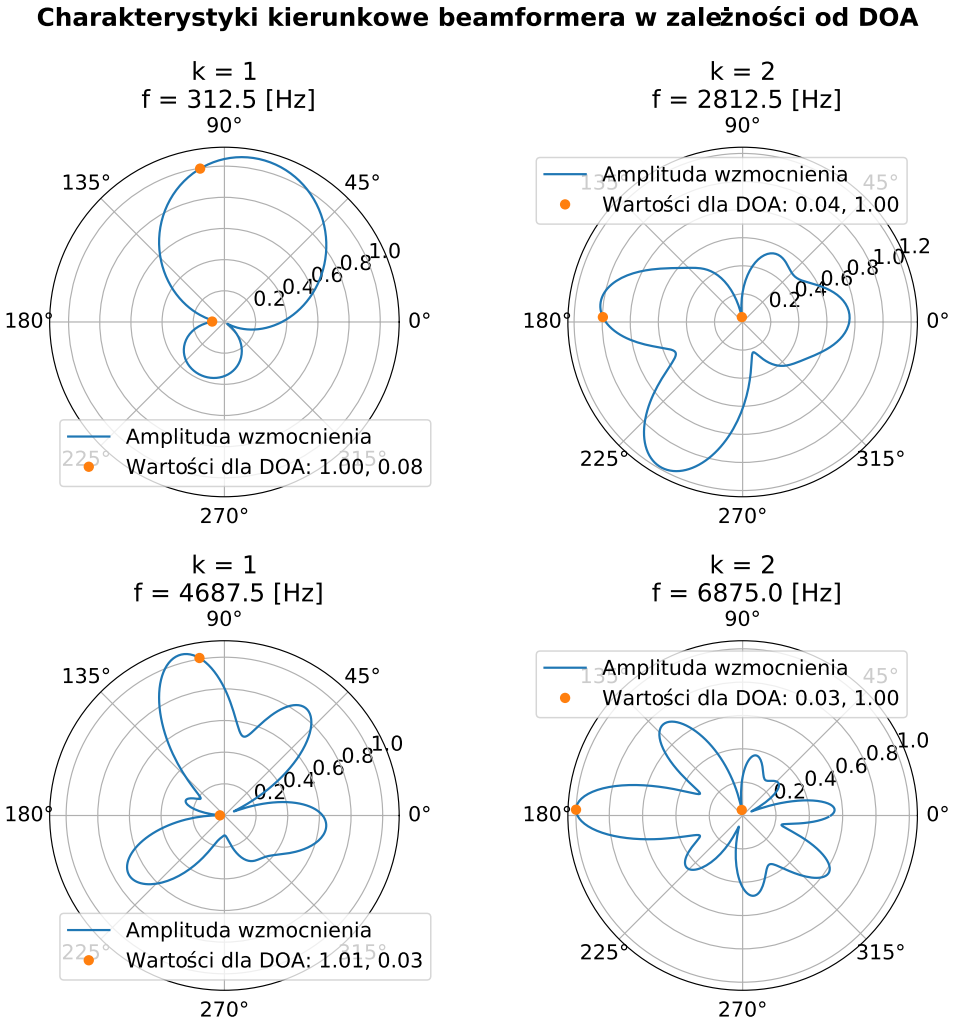
5.4 Skuteczność działania WPD z estymacją parametrów wejściowych

Finalnym algorytmem zaimplementowanym w tej pracy jest algorytm jednoczesnej separacji sygnałów i usuwania pogłosu przy pomocy WPD z estymacją parametrów wejściowych. Rezultaty przedstawione w tej sekcji uzyskano dla tych samych warunków co w sekcji 5.2, to znaczy: czas pogłosu równy 0.4 s oraz 2 źródła sygnału z kierunków 99° i 178°. Sygnały źródłowe to odpowiednio mowa kobieca oraz mowa męska.

5.4.1 Beamforming

Analogicznie jak w podsekcji 5.2.1, na rysunku 5.4 zaprezentowano charakterystyki kierunkowe filtru przestrzennego WPD, przy obliczaniu którego bazowano na estymacji parametrów wejściowych.

Można zauważyć sporą różnicę w kształtach wykresów z rysunku 5.4 w porównaniu do wykresów z rysunku 5.1. Najważniejsze jednak wartości jakimi są amplitudy dla faktycznych kierunków DOA pozostały niemal niezmienione. Można zauważyć, że dla częstotliwości $f = 4687.5$[Hz] amplituda dla kierunku, w którym skierowana

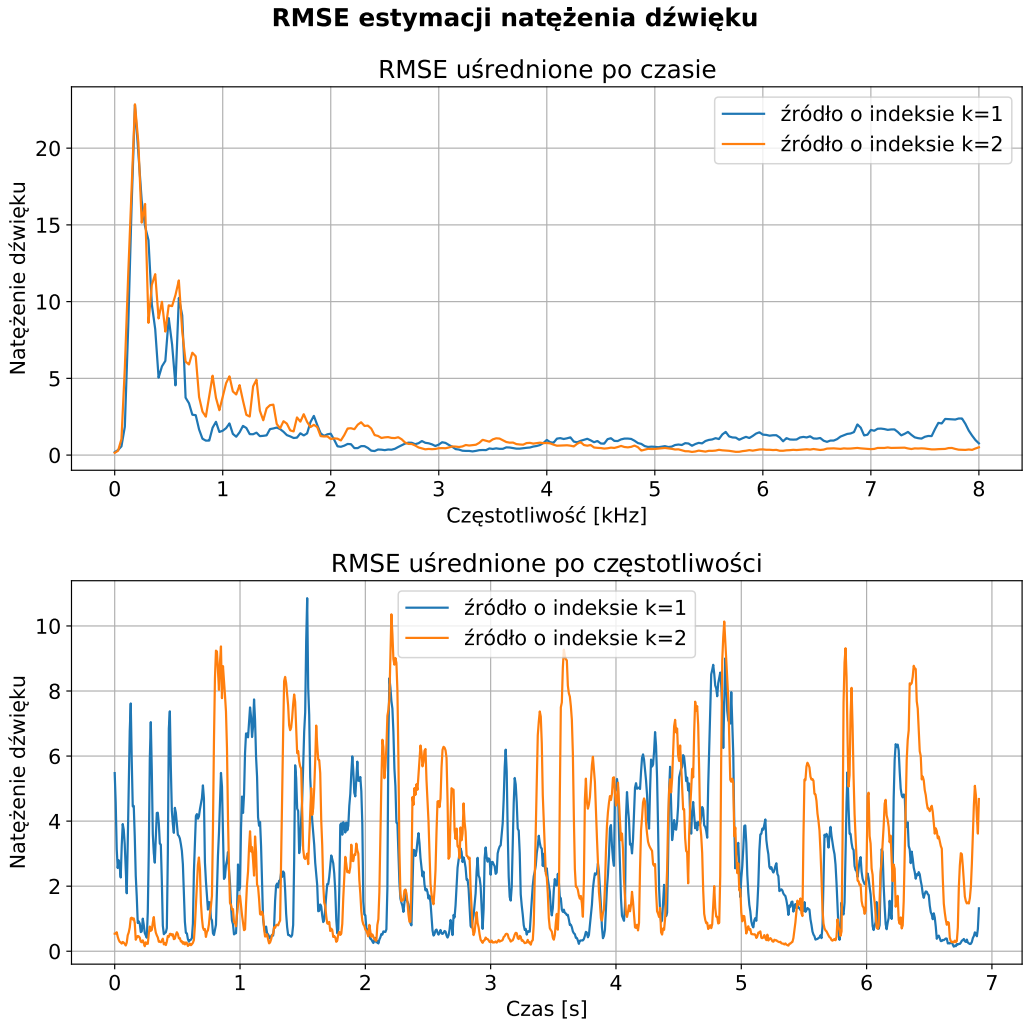


Rysunek 5.4: Charakterystyki kierunkowe filtru przestrzennego z estymacją parametrów wejściowych

jest główna wiązka filtru przestrzennego jest nieco większa niż 1. Wynika to z błędu estymacji miary kąta DOA. Na wykresie zaznaczono rzeczywistą miarę kąta DOA czyli 99°, a filtr przestrzenny kierowany był w stronę wyestymowanego kierunku, jakim było 98.5°. Nie są to jednak różnice, które miałyby istotny wpływ na działanie filtru. Widać też, że maksima lokalne filtru przestrzennego nie zawsze są tam gdzie faktyczne DOA oraz występuje sporo stosunkowo dużych wartości amplitud dla kierunków zupełnie niezwiązanych z DOA. Ta własność wpływa potem na jakość działania całego algorytmu.

5.4.2 Błędy estymacji

Podobnie jak w analogicznej podsekcji 5.2.2, na rysunku 5.5, możemy zobaczyć błąd estymacji sygnałów źródłowych. Błąd przedstawiony jest w postaci RMSE z estymacji sygnałów $d_q^{(k)}$ sygnałami $\hat{d}_q^{(k)}$. Do estymacji, tak jak w całej tej sekcji, użyto algorytmu WPD z estymacją parametrów wejściowych. RMSE uśredniano kolejno po czasie i po częstotliwości. Uzyskane wykresy bardzo mocno przypominają analogiczne wykresy z rysunku 5.2



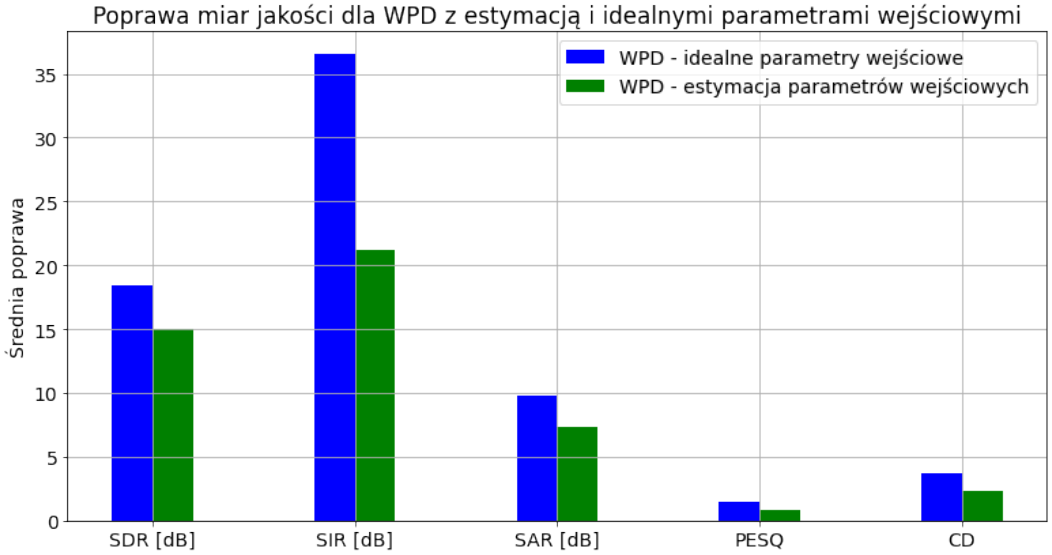
Rysunek 5.5: Błędy estymacji z użyciem algorytmu z estymacją parametrów wejściowych

5.4.3 Miary jakości separacji i eliminacji pogłosu

Analogicznie jak w podsekcji 5.2.3, w poniższej tabeli zamieszczono miary oceniające jakość działania algorytmu.

Tabela 5.3: Miary jakości

Miary jakości działania algorytmu WPD z estymacją parametrów wejściowych					
Numer źródła (k)	SDR [dB]	SIR [dB]	SAR [dB]	PESQ	CD
	Na wejściu				
$k = 1$	-4.31	-3.23	7.19	1.05	6.74
$k = 2$	1.30	3.35	7.19	1.13	6.75
	Na wyjściu przy estymacji parametrów wejściowych				
$k = 1$	13.90	19.37	15.41	1.60	4.49
$k = 2$	13.09	23.20	13.56	2.27	4.37
	Poprawa				
$k = 1$	18.22	22.60	8.22	0.55	2.25
$k = 2$	11.79	19.85	6.37	1.14	2.38



Rysunek 5.6: Porównanie WPD z idealnymi i estymowanymi parametrami wejściowymi

Na rysunku 5.6 porównano przy użyciu diagramów słupkowych uzyskaną poprawę z poprawą, jaką uzyskano w podsekcji 5.2.3, czyli kiedy algorytm WPD korzystał z idealnych wartości parametrów wejściowych. Można dzięki temu zauważyć, jak bardzo niedokładności estymacji parametrów wejściowych wpłynęły na końcowy rezultat. Parametrem, który najbardziej się pogorszył jest SIR mówiący o tym, jak dobrze eliminowane są zakłócenia. W tym przypadku głównym zakłóceniem dla da-

nego źródła jest drugie źródło. W istocie, gdy słucha się sygnałów wynikowych, słychać w tle nie do końca odseparowane drugie źródło.

5.5 Skuteczność działania WPD z estymacją parametrów wejściowych przy trzech źródłach

W tej sekcji zaprezentowane zostanie działanie algorytmu WPD z estymacją parametrów wejściowych przy obecności trzech źródeł sygnału. Warunki w tym eksperymencie są takie same jak w sekcji 5.4 z tą różnicą, że w tej sekcji sygnał wejściowy składa się z trzech źródeł sygnału o DOA równych 99°, 178° oraz 333°. Źródła sygnału to kolejno: mowa kobieca i mowa męska - te same co w poprzednich eksperymentach, oraz mowa kobieca - dodatkowa. Uzyskane miary ewaluacyjne (te same co w podsekcji 5.2.3) zaprezentowane są w poniższej tabeli.

Tabela 5.4: Miary jakości

Miary jakości estymacji sygnałów źródłowych					
Numer źródła (k)	SDR [dB]	SIR [dB]	SAR [dB]	PESQ	CD
	Na wejściu (3 źródła)				
$k = 1$	-7.36	-6.43	7.11	1.04	7.31
$k = 2$	-3.16	-1.96	7.11	1.06	7.78
$k = 3$	-2.38	-1.10	7.11	1.05	7.17
	Na wyjściu przy estymacji parametrów wejściowych (3 źródła)				
$k = 1$	11.90	18.37	13.07	1.43	4.81
$k = 2$	11.37	19.02	12.24	1.88	4.77
$k = 3$	10.97	18.08	11.98	1.65	4.76
	Poprawa (3 źródła)				
$k = 1$	19.26	24.81	5.95	0.39	2.50
$k = 2$	14.53	20.98	5.12	0.82	3.01
$k = 3$	13.35	19.17	4.87	0.60	2.41
	Na wyjściu przy estymacji parametrów wejściowych (2 źródła)				
$k = 1$	13.90	19.37	15.41	1.60	4.49
$k = 2$	13.09	23.20	13.56	2.27	4.37
	Poprawa (2 źródła)				
$k = 1$	18.22	22.60	8.22	0.55	2.25
$k = 2$	11.79	19.85	6.37	1.14	2.38

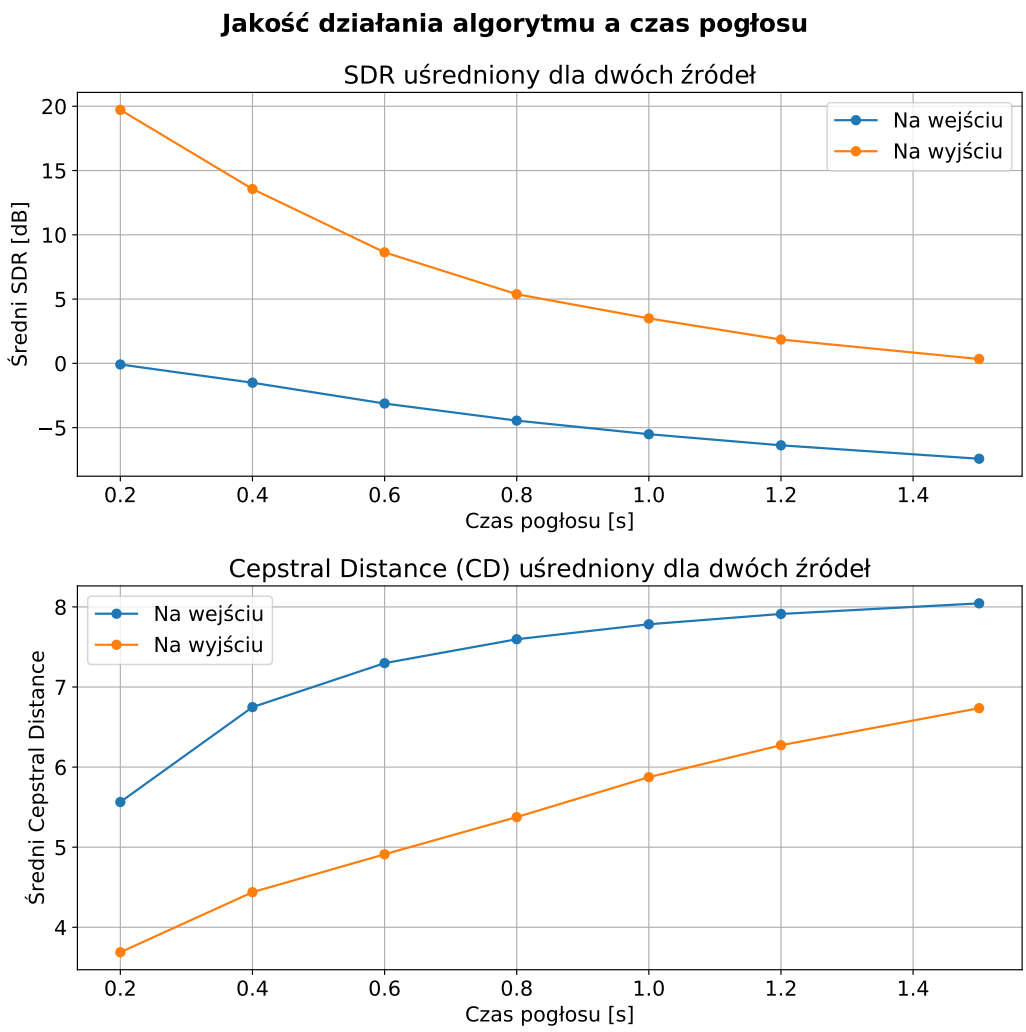
W częściach tabeli nazwanych „Na wyjściu przy estymacji parametrów wejściowych (2 źródła)" oraz „Poprawa (2 źródła)" przytoczono miary jakie uzyskano w podsekcji 5.4.3, czyli kiedy algorytm WPD działał na sygnale złożonym z dwóch źródeł sygnału. Algorytm w dobrym stopniu radzi sobie z trzema źródłami. Porównując miary dla dwóch i trzech źródeł można zauważyć, że dołożenie trzeciego źródła wiąże się z zauważalnym, chociaż nie drastycznym spadkiem jakości sygnałów wynikowych.

Patrząc na to, jaką poprawę udało się uzyskać w wyniku działania algorytmu dla dwóch i trzech źródeł, można dojść do wniosku, że algorytm w obu przypadkach radzi sobie podobnie z delikatną przewagą przypadku z dwoma źródłami. Do podobnych wniosków można dojść słuchając sygnałów wynikowych. Dla trzech źródeł tło składające się z dwóch pozostałych sygnałów jest trochę głośniejsze, ale wciąż nie ma problemu ze zrozumieniem osoby mówiącej.

5.6 Skuteczność działania WPD z estymacją parametrów wejściowych przy różnych czasach pogłosu

W tej sekcji zaprezentowane zostanie zachowanie algorytmu WPD z estymacją parametrów wejściowych przy różnych czasach pogłosu. Warunki w tym eksperymencie są takie same jak w sekcji 5.4 z tą różnicą, że w tej sekcji zmieniany będzie czas pogłosu ustawiany w generatorze odpowiedzi impulsowych pomieszczenia. Do tej pory czas pogłosu zawsze ustawiony był na 0.4 s.

Na rysunku 5.7 zaprezentowano jak zmieniają się miary SDR oraz CD na wyjściu i na wejściu algorytmu wraz ze zmianą czasu pogłosu. Można zauważyć, że algorytm jest bardzo czuły na zmiany czasu pogłosu i dosyć szybko obniża się jego jakość działania. Należy jednak pamiętać, że długość filtra przestrzennego (L) nie jest zmieniana i wraz ze wzrostem czasu pogłosu coraz więcej późnego pogłosu przekracza założone 13 ramek czasowych STFT i nie jest w ogóle eliminowane. Aby temu zapobiec, należałoby sukcesywnie zwiększać wartość L, aczkolwiek to z kolei prowadziłoby do dłuższego czasu działania algorytmu.



Rysunek 5.7: WPD z estymacją parametrów wejściowych a czas pogłosu

Rozdział 6

Podsumowanie

W pracy zaimplementowano i zoptymalizowano pod kątem obliczeniowym algorytm jednoczesnej separacji źródeł i eliminacji pogłosu. Do uzyskania zamierzonego efektu użyto metody splotowego kształtowania wiązki WPD. Do estymacji parametrów wejściowych użyto filtra WPE, algorytmu lokalizacji źródeł dźwięku MUSIC oraz filtru przestrzennego LCMV.

W ramach ewaluacji poprawności działania zaimplementowanego algorytmu przeprowadzono szereg testów, w których badane były zarówno poszczególne etapy algorytmu, jak i algorytm jako całość. Uzyskano rezultaty świadczące o wysokiej jakości odseparowanych i pozbawionych pogłosu sygnałów, które otrzymano jako wynik działania algorytmu.

W przyszłości należałoby rozważyć dalsze przyspieszenie działania algorytmu i próbę uruchomienia go w czasie rzeczywistym. Dodatkowo można spróbować zastosować inny algorytm estymacji parametrów wejściowych, szczególnie mocy, i zbadać, czy polepszy to jakość działania filtru przestrzennego WPD. Jednym z możliwych rozwiązań jest iteracyjne wywołanie algorytmu WPD, gdzie w drugim i kolejnych krokach jako estymowaną moc sygnałów źródłowych można by było zastosować moc obliczoną na bazie sygnałów odseparowanych i pozbawionych pogłosu, otrzymanych w krokach wcześniejszych. Znacząco wydłużyłoby to czas obliczeń, ale mogłoby w dużym stopniu poprawić jakość działania algorytmu.

Bibliografia

[1] Tuomas Virtanen. "Monaural Sound Source Separation by Nonnegative Matrix Factorization With Temporal Continuity and Sparseness Criteria". W: *IEEE Transactions on Audio, Speech, and Language Processing* 15.3 (2007), s. 1066–1074. DOI: 10.1109/TASL.2006.885253.

[2] Takuya Yoshioka i Tomohiro Nakatani. "Generalization of Multi-Channel Linear Prediction Methods for Blind MIMO Impulse Response Shortening". W: *IEEE Transactions on Audio Speech and Language Processing* 20 (grud. 2012), s. 2707–2720. DOI: 10.1109/TASL.2012.2210879.

[3] Hakan Erdogan i in. "Improved MVDR Beamforming Using Single-Channel Mask Prediction Networks". W: wrz. 2016, s. 1981–1985. DOI: 10.21437/Interspeech.2016-552.

[4] Oliver Thiergart i Emanuel A.P. Habets. "An informed LCMV filter based on multiple instantaneous direction-of-arrival estimates". W: *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings* (paź. 2013), s. 659–663. ISSN: 15206149. DOI: 10.1109/ICASSP.2013.6637730.

[5] Tomohiro Nakatani i Keisuke Kinoshita. "A unified convolutional beamformer for simultaneous denoising and dereverberation". W: *IEEE Signal Processing Letters* 26 (grud. 2018), s. 903–907.

[6] Tomohiro Nakatani i in. "Simultaneous Denoising, Dereverberation, and Source Separation Using a Unified Convolutional Beamformer". W: paź. 2019, s. 224–228. DOI: 10.1109/WASPAA.2019.8937285.

[7] I.A. McCowan. *Robust Speech Recognition Using Microphone Arrays*. Queensland University of Technology, Brisbane, 2001.

[8] B.D. Van Veen i K.M. Buckley. "Beamforming: a versatile approach to spatial filtering". W: *IEEE ASSP Magazine* 5.2 (kw. 1988), s. 4–24. ISSN: 1558-1284. DOI: 10.1109/53.665.

[9] Emanuël Habets. *Room Impulse Response Generator*. Wrz. 2010. URL: https://www.researchgate.net/publication/259991276_Room_Impulse_Response_Generator.

[10] Jont Allen i David Berkley. "Image method for efficiently simulating small-room acoustics". W: *The Journal of the Acoustical Society of America* 65 (kw. 1979), s. 943–950. DOI: 10.1121/1.382599.

[11] Yeo-Sun Yoon, L.M. Kaplan i J.H. McClellan. "TOPS: new DOA estimator for wideband signals". W: *IEEE Transactions on Signal Processing* 54.6 (czer. 2006), s. 1977–1989. ISSN: 1941-0476. DOI: 10.1109/TSP.2006.872581.

[12] A.W. Rix i in. "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs". W: *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*. T. 2. Maj 2001, 749–752 vol.2. DOI: 10.1109/ICASSP.2001.941023.

[13] E. Vincent, R. Gribonval i C. Fevotte. "Performance measurement in blind audio source separation". W: *IEEE Transactions on Audio, Speech, and Language Processing* 14.4 (2006), s. 1462–1469. DOI: 10.1109/TSA.2005.858005. URL: https://hal.inria.fr/inria-00544230/document.

[14] Y. Tohkura. "A weighted cepstral distance measure for speech recognition". W: *ICASSP '86. IEEE International Conference on Acoustics, Speech, and Signal Processing*. T. 11. Kw. 1986, s. 761–764. DOI: 10.1109/ICASSP.1986.1169214.

[15] *Sound quality assessment material recordings for subjective tests*. URL: https://tech.ebu.ch/publications/sqamcd.

[16] Tomasz P. Zieliński. *Cyfrowe przetwarzanie sygnałów od teorii do zastosowań*. 2005.

[17] Christoph Boeddeker i in. "Jointly Optimal Dereverberation and Beamforming". W: *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Maj 2020, s. 216–220. DOI: 10.1109/ICASSP40776.2020.9054393. URL: https://arxiv.org/abs/1910.13707.