Rafał Korycki

Wielokanałowe techniki korekcji nagrań fonicznych w kryminalistyce

Wstęp i cel pracy

Nagrania monofoniczne stanowią większość wśród materiałów nadsyłanych przez zleceniodawców do badań. Jednokanałowe techniki redukcji szumu wydają się być zatem najbardziej upowszechnione wśród ekspertów z zakresu inżynierii dźwięku i fonoskopii. Rozwiązania te zależą od statystycznych modeli sygnałów zakłócających, które mogą być estymowane w czasie braku aktywności mówcy lub tworzone i modyfikowane przez użytkownika. Pomimo wielu lat badań nad jednokanałowymi systemami korekcji nagrań wciąż zdarza się, że efekt ich działania pozostaje niezadowalający. Wystarczy jednak uzyskać trochę dodatkowych informacji na temat sygnału zakłócającego, by możliwe było zastosowanie całego arsenału środków technicznych. Warto mieć to na uwadze, ponieważ coraz większa dostępność i duża popularność przenośnych rejestratorów cyfrowych wyposażonych w dwa mikrofony może przyczynić się do zmniejszenia liczby monofonicznych nagrań przekazywanych do badań w ciągu kilku najbliższych lat.

Celem niniejszej pracy jest krótkie omówienie klasycznych technik redukcji szumu i zakłóceń oraz przedstawienie rozwiązań, które mogą okazać się przydatne podczas analizy nagrań wielokanałowych. Zaprezentowane zostaną techniki filtracji adaptacyjnej, które pozwalają na redukcję sygnałów zakłócających generowanych w celu zapewnienia poufności prowadzonych rozmów. Przedstawione zostaną także algorytmy umożliwiające rozdzielanie jednoczesnych wypowiedzi mówców, co ma niebagatelne znaczenie podczas analizy nagrań rejestrowanych w dużych skupiskach ludzkich. Wszystkie opisane algorytmy zostały zaimplementowane i przetestowane, a wyniki symulacji zaprezentowane w postaci przebiegów czasowych i spektrogramów.

Jednokanałowe techniki poprawy jakości nagrań

Systemy redukcji szumu znacznie różnią się od siebie zarówno pod względem złożoności, jak i efektywności, która przekłada się bezpośrednio na stopień poprawy zrozumiałości przetwarzanej mowy. W większości z nich można wyróżnić wspólne rozwiązania [1, 2]. Pierwszym etapem przetwarzania jest segmentacja, czyli podział sygnału na fragmenty o długości ok. 20–30 milisekund, a następnie mnożenie ich przez tzw. funkcję okna (np. Blackmana, Hanna itp.). Tak przygotowane segmenty poddaje się dyskretnej transformacji Fouriera w celu otrzymania próbek widma. Kolejnym etapem jest tworzenie modeli szumu. Istotne jest, aby algorytm „wiedział" co należy usunąć w celu zapewnienia skutecznego odtworzenia sygnału użytecznego. W większości komercyjnych aplikacji stosowane są dwa rozwiązania: detekcja automatyczna oraz ręczne pobieranie próbek widma szumu. Pierwsze z nich polega na zastosowaniu systemów rozpoznawania mowy, ponieważ wyznaczanie modelu sygnału zakłócającego ma miejsce w przerwach między wypowiedziami. Drugie rozwiązanie wymaga od użytkownika samodzielnego zaznaczenia fragmentu nagrania, w którym występuje wyłącznie sygnał zakłócający. Wyznaczenie estymaty widma mowy niezakłóconej to kolejny etap realizowany przez typowy system redukcji szumu. Konieczna jest w tym przypadku modyfikacja widma sygnału wejściowego stosownie do estymowanego stosunku sygnału do szumu dla każdej dyskretnej wartości częstotliwości. Ostatnim etapem jest wyznaczenie odwrotnej transformacji Fouriera i odtworzenie fazy sygnału.

Przyjmując założenie, że sygnał mowy $s(n)$ oraz sygnał szumu $w(n)$ są addytywne[1], „zaszumione" nagranie można opisać w następujący sposób:

$$y(n) = s(n) + w(n) \quad (1),$$

gdzie n oznacza indeks czasu dyskretnego, czyli numer próbki sygnału. W większości przypadków przyjmuje się, że sygnał mowy nie jest skorelowany z szumem. Ma to jednak uzasadnienie wówczas, gdy sygnały użyteczny i zakłócający generowane są przez niezależne źródła. Jedna z najprostszych metod pozwalająca na poprawę jakości nagrań mowy została szeroko opisana w literaturze już trzy dekady temu [3, 4, 5, 6]. To rozwiązanie sprowadza się do odjęcia estymaty widma szumu z widma sygnału mowy, przyrównania ujemnych różnic do zera, odtworzenia nowego widma z oryginalną wartością przesunięcia fazowego oraz rekon-

strukcji przebiegu czasowego sygnału. Proces można opisać w następujący sposób:

$$D(\omega) = P_S(\omega) - P_N(\omega), \tag{2}$$

$$P'_S(\omega) = \begin{cases} D(\omega), & jeśli\ D(\omega) > 0 \\ 0, & w\ innym\ przypadku \end{cases}$$

$P_S(\omega)$ jest tu widmem zakłóconego sygnału wejściowego, $P_N(\omega)$ stanowi wygładzoną estymatę widma szumu, natomiast $P'_S(\omega)$ jest zmodyfikowanym widmem sygnału. $P_N(\omega)$ otrzymuje się w dwóch etapach. Najpierw uśredniane jest widmo szumu z kilku kolejnych segmentów sygnału, w których nie występuje sygnał mowy, a następnie tak otrzymane widmo jest wygładzane [5].

W przypadku większości metod poprawy jakości sygnału mowy przyjmuje się założenie, że widmo sygnału zakłóconego jest równe sumie widm sygnału oryginalnego oraz szumu. Założenie to jest prawdziwe jedynie w sensie statystycznym i może być spełnione przy zastosowaniu widma wyznaczonego z wykorzystaniem krótkiego okna[2]. Ze względu na fakt, że sygnał mowy oraz sygnał zakłócający nie zawsze spełniają warunek braku wzajemnej korelacji, niektóre składowe $P_S(\omega)$ mogą być ujemne. W związku z powyższym przyrównywane są do zera [3, 4, 6].

Poważnym mankamentem opisanej metody jest powstawanie tzw. szumu muzycznego, który może być subiektywnie odbierany jako dzwonienie lub świergotanie. Wynika to z występowania szczytów i dolin w krótkookresowym widmie szumu białego[3]. Ich amplituda i położenie w dziedzinie częstotliwości mają charakter losowy i zmieniają się z ramki na ramkę. Po odjęciu wygładzonej estymaty widma szumu z bieżącego widma szumu, wszystkie maksima widmowe są redukowane, podczas gdy minima są zerowane (2). W efekcie pojawiają się fluktuacje obwiedni widma szumu. Szersze maksima odbierane są przez słuchacza jako wąskopasmowy sygnał szumu. Węższe natomiast brzmieniem przypominają sygnały tonalne o częstotliwościach zmiennych w czasie, które można określić mianem szumu muzycznego. Jest to jeden z najczęściej pojawiających się efektów podczas korzystania z popularnych aplikacji służących do redukcji szumu i zakłóceń. Jedną z metod minimalizacji tego zjawiska jest rozwiązanie zaproponowane w [6]:

$$P'_S(\omega) = \begin{cases} D^{\gamma}(\omega), & jeśli\ D^{\gamma}(\omega) > \beta P_N(\omega) \\ \beta P_N(\omega), & w\ innym\ przypadku \end{cases} \tag{3}$$

$$D(\omega) + G[P^{\gamma}_S(\omega) - aP^{\gamma}_N(\omega)]$$

Zaleca się tu zachowanie następujących warunków: $\alpha \geq 1$, $0 < \beta << 1$; gdzie α jest współczynnikiem redukcji, natomiast współczynnik β jest związany z poziomem progowym widma. Dzięki tym parametrom możliwa jest zarówno lepsza redukcja szerszych maksimów, jak i zmniejszenie głębokości minimów obwiedni. Ponadto γ jest współczynnikiem redukcji widma mocy, natomiast G współczynnikiem normalizacji [3, 4].

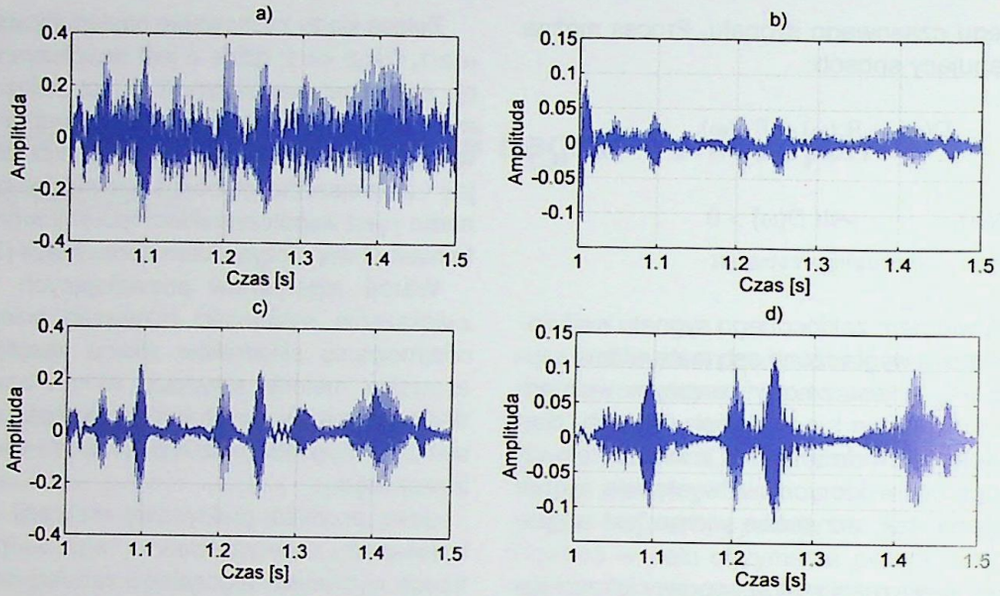
Wśród algorytmów pozwalających na znaczne zwiększenie wydajności rozwiązań polegających na odejmowaniu składników widma znajdują się Bayesowskie metody estymacji widma amplitudowego. Wykorzystuje się w nich funkcje gęstości prawdopodobieństwa sygnału i szumu oraz minimalizuje koszt funkcji błędu.

Jako przykład praktycznej realizacji opisywanych rozwiązań zaprezentowano wyniki implementacji trzech wybranych algorytmów redukcji szumu, w których zastosowano automatyczną detekcję sygnału mowy. Wypowiadane zdanie zostało zakłócone szumem różowym o wartości skutecznej równej –6 dB. W pierwszej z prezentowanych metod zastosowano estymację minimalnego średniego błędu kwadratowego krótkoczasowego widma amplitudowego [7, 8] (ryc. 1b oraz 2b). Pozostałe dwa rozwiązania wykorzystują estymację a priori stosunku sygnału do szumu przedstawionego przez Scalarta i Vieira-Filho [9] (ryc. 1c i 2c) oraz opisanego przez Cohena [11] (ryc. 1d i 2d). W przypadku pierwszego rozwiązania można było zauważyć niewielką poprawę jakości nagrania przy minimalnych zniekształceniach wprowadzanych przez program. Najlepsze efekty uzyskano, stosując dwie ostatnie metody, jednakże w przypadku zilustrowanym na rycinie 1c i 2c wyraźnie słyszalny jest szum muzyczny (ryc. 2).

Filtracja adaptacyjna

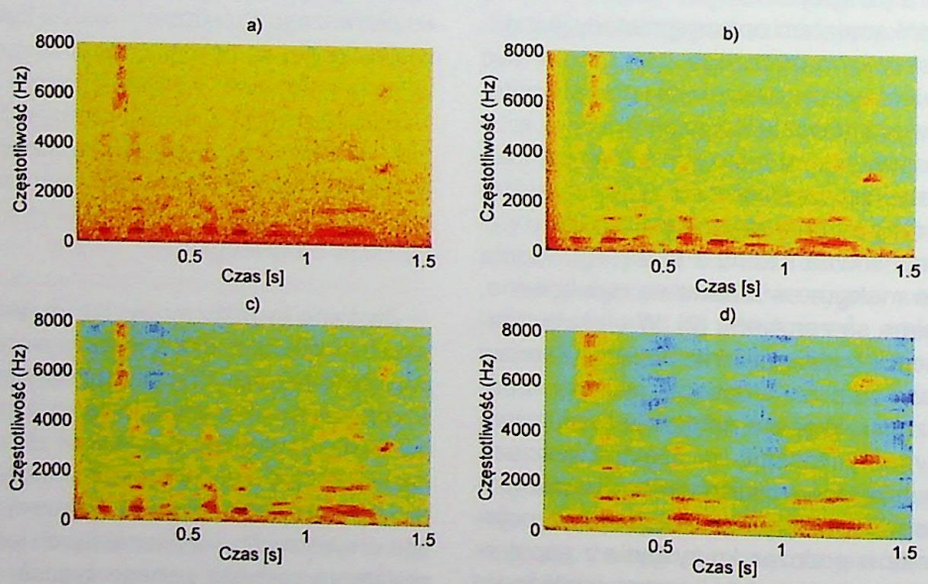
Zarówno sygnały mowy, jak i zakłócenia mogą być rejestrowane za pomocą wielu sensorów, po czym poddawane są filtracji. Wykorzystywany jest fakt, że nawet w przypadku niewielkich odległości między poszczególnymi mikrofonami docierające do nich sygnały różnią się od siebie pod względem natężenia oraz charakterystyki częstotliwościowej i fazowej. Redukcja zakłóceń w systemach wielokanałowych jest realizowana na podstawie zarówno samego sygnału użytecznego, jak i zakłócającego. Istotna jest również szybkość adaptacji (przystosowania) współczynników stosowanych filtrów do zmieniających się warunków akustycznych.

Projektowanie filtrów cyfrowych polega na doborze struktury[4], rzędu[5] oraz wartości współczynników. Tak stworzony układ cechuje się odpowiednią charakterystyką częstotliwościową, dzięki czemu umożliwia wzmacnianie lub tłumienie określonych zakresów czę-



Ryc. 1. Porównanie algorytmów poprawy jakości sygnału mowy. Przebiegi czasowe przedstawiają:
a) sygnał mowy zakłóconej (częstotliwość próbkowania 16 kHz, poziom addytywnego szumu różowego –6 dB),
b) sygnał po korekcji z wykorzystaniem algorytmu opisanego w [7, 8],
c) sygnał po zastosowaniu metody opisanej w [9],
d) sygnał po korekcji metodą przedstawioną w [11]

Fig. 1. *Comparison of single channel audio enhancement algorithms. Time plots as follows:*
a) speech signal corrupted by noise (16 kHz sampling frequency, additive pink noise level –6 dB),
b) speech signal after enhancement by algorithm described in [7, 8],
c) speech signal after enhancement by method described in [9],
d) speech signal after enhancement by algorithm described in [11]
źródło (ryc. 1–19): autor



Ryc. 2. Porównanie algorytmów poprawy jakości sygnału mowy. Spektrogramy przedstawiają:
a) sygnał mowy zakłóconej (częstotliwość próbkowania 16 kHz, poziom addytywnego szumu różowego –6 dB), b) sygnał po korekcji z wykorzystaniem algorytmu opisanego w [7, 8],
c) sygnał po zastosowaniu metody opisanej w [9],
d) sygnał po korekcji metodą przedstawioną w [11]

Fig. 2. *Comparison of single channel audio enhancement algorithms. Spectrograms as follows:*
a) speech signal corrupted by noise (sampling frequency 16 kHz, additive pink noise level –6 dB),
b) speech signal after enhancement by algorithm described in [7, 8],
c) speech signal after enhancement by method described in [9],
d) speech signal after enhancement by algorithm described in [11]

stotliwości. W przypadku dokonywania rejestracji w zmiennych warunkach akustycznych istnieje konieczność dopasowania charakterystyk filtrów do sygnałów zakłócających w taki sposób, by możliwa była skuteczna redukcja tych ostatnich. Filtry adaptacyjne pozwalają na modyfikację charakterystyk przez automatyczną aktualizację wartości współczynników. Proces adaptacji polega na minimalizacji błędu między sygnałem wyjściowym a sygnałem docelowym (lub pożądanym) [12].

Sygnał wejściowy $d(n)$ jest sumą sygnałów użytecznego $s(n)$ oraz zakłócającego $w(n)$. Efektem działania algorytmu jest estymata sygnału zakłócającego. Po odjęciu od siebie tych dwóch sygnałów uzyskuje się sygnał błędu $e(n)$. W zależności od własności sygnału zakłócającego przybliżeniem $s(n)$ może być sygnał błędu lub sygnał wyjściowy. Proces adaptacji uzależniony jest od zastosowanego algorytmu aktualizacji współczynników filtru.

Algorytmy filtracji adaptacyjnej stanowią dość liczną grupę. Dwa najprostsze i jedne z najczęściej stosowanych w praktyce to LMS (*Least Mean Squares*) i RLS (*Recursive Least Squares*) [13, 14].

$$w(n+1) = w(n) + \mu[y(n)e(n)] \quad (4),$$

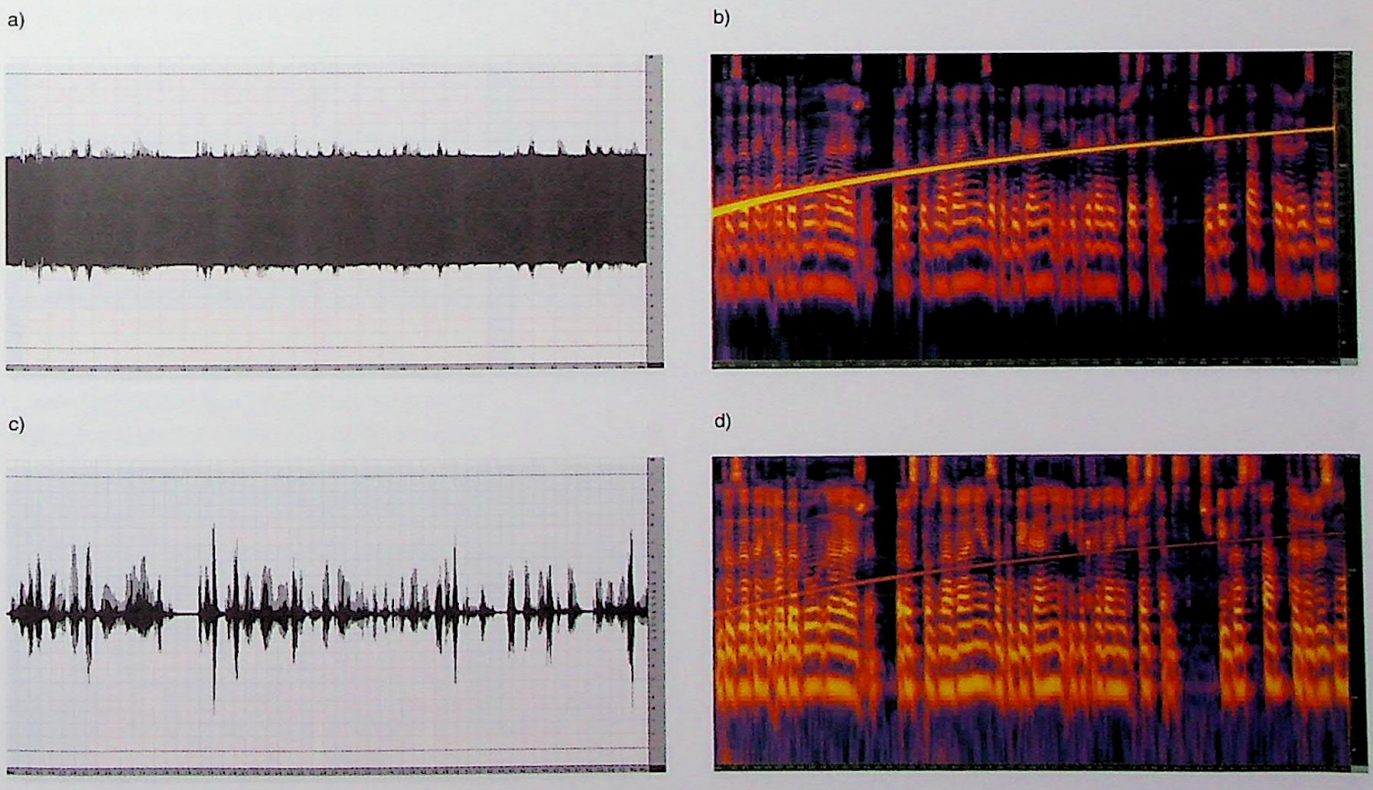
$$w(n) = w(n-1) + k(n)e(n) \quad (5),$$

gdzie:

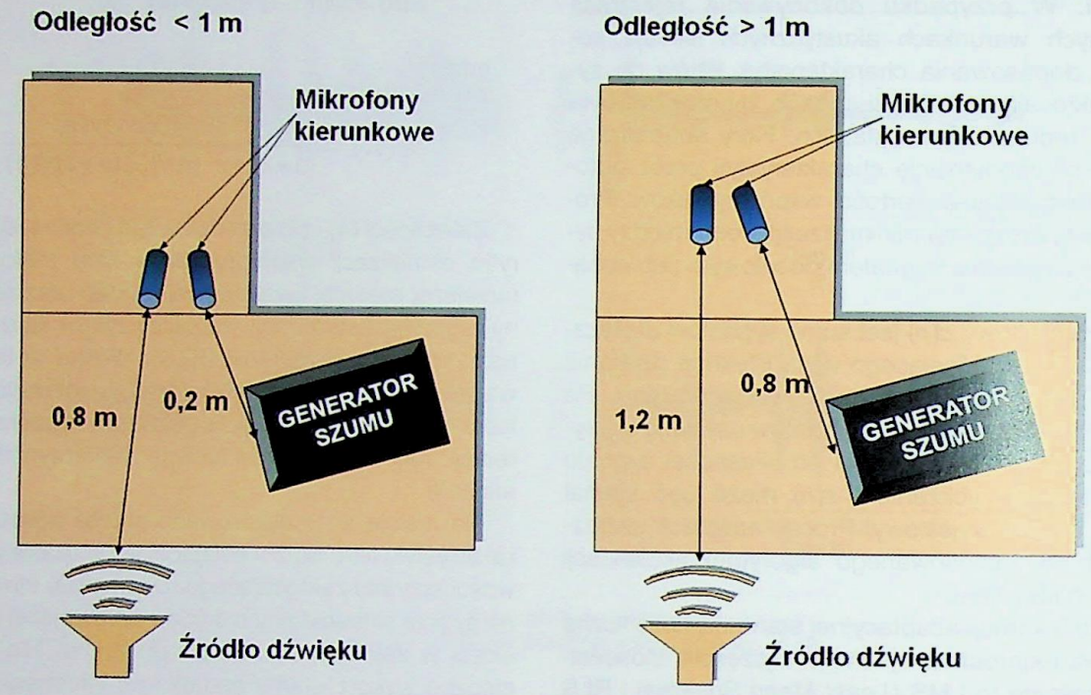
$$k(n) = \frac{\lambda^{-1} R_{yy}^{-1}(n-1)y(n)}{1+\lambda^{-1} y^T(n) R_{yy}^{-1}(n-1)y(n)}$$

Zależności (4) i (5) przedstawiają odpowiednio algorytm aktualizacji współczynników filtru metodą „najmniejszej średniej kwadratowej" (LMS) oraz rekursywny algorytm aktualizacji współczynników filtru metodą najmniejszych kwadratów (RLS). Wektor **w** to wektor wartości wag współczynników filtru, μ jest współczynnikiem szybkości adaptacji, $\mathbf{R}_{yy}$ oznacza macierz autokorelacji, natomiast λ pełni funkcję współczynnika zapominania.

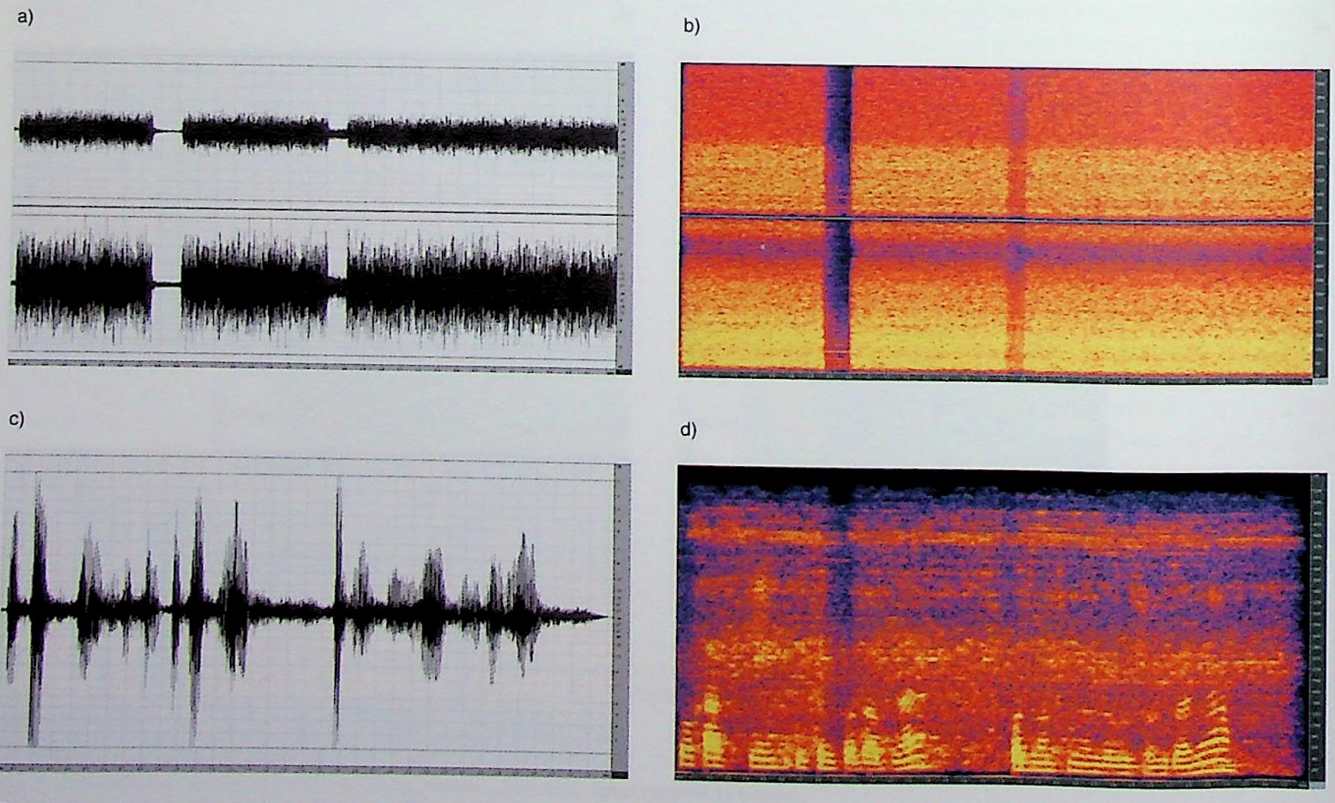
Na rycinie nr 3 zilustrowano proces adaptacji charakterystyki filtru do zmieniającej się w czasie częstotliwości sygnału zakłócającego. Do sygnału mowy dodano sygnał sinusoidalny o liniowo narastającej częstotliwości w zakresie od 500 Hz do 2 kHz. Nie było konieczne wykorzystanie dodatkowej informacji o zakłóceniach. Wystarczyło podanie jako sygnału odniesienia opóźnionej o jedną próbkę kopii sygnału wejściowego. Mimo niewielkiego przesunięcia w czasie sygnał



Ryc. 3. Przebiegi czasowe (a, c) oraz spektrogramy (b, d): sygnału mowy zakłóconego sygnałem sinusoidalnym o liniowo narastającej częstotliwości (a, b) oraz sygnału błędu na wyjściu filtru adaptacyjnego po operacji mającej na celu redukcję sygnału zakłócającego (c, d)
Fig. 3. *Time plots (a, c) and spectrograms (b, d) of: speech signal corrupted by sine wave with linear frequency modulation (a, b) and error signal on the output of adaptive filter after noise reduction operation (c, d)*



Ryc. 4. Rysunek poglądowy ilustrujący rozmieszczenie mikrofonów podczas testu z wykorzystaniem generatora szumu
Fig. 4. *Demonstrative figure which depicts microphone spacing during the test with noise generator*



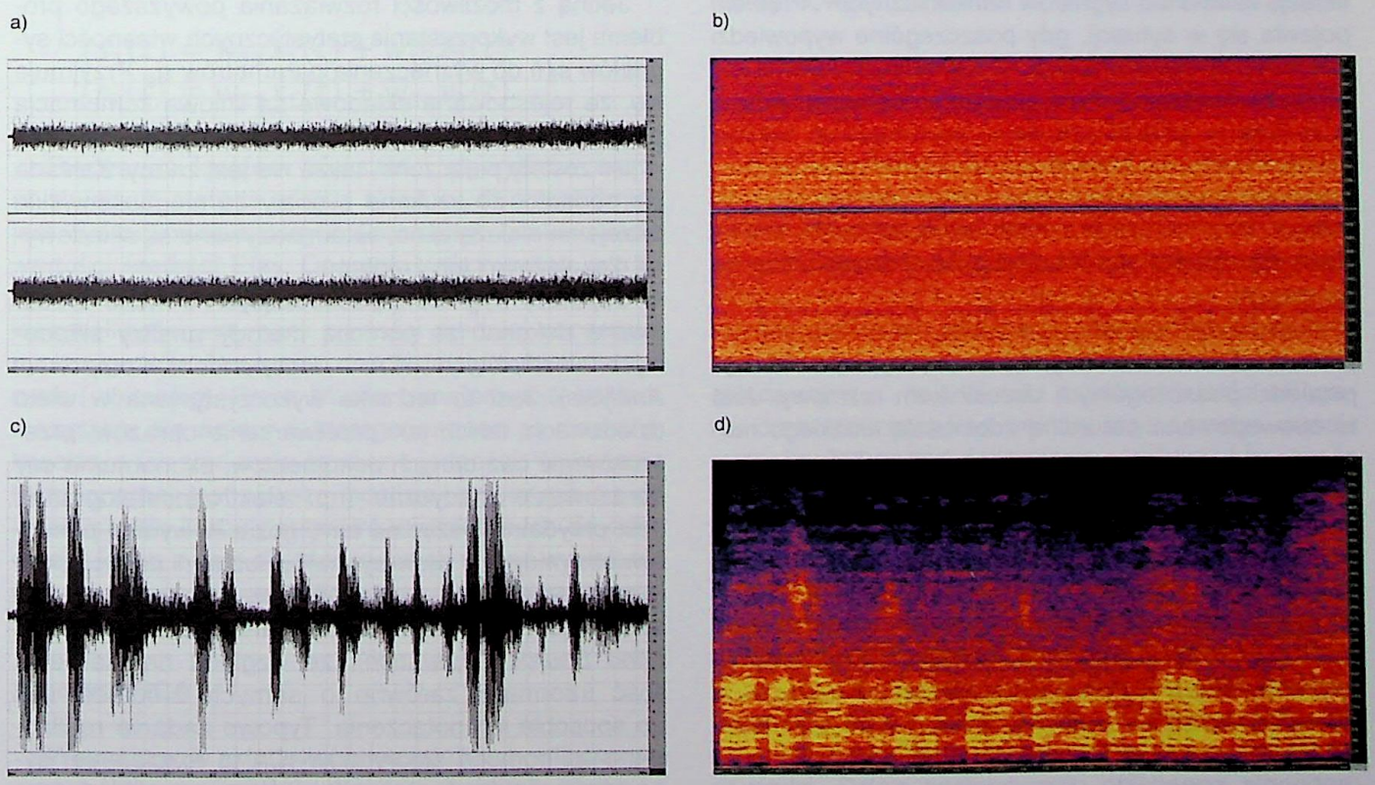
Ryc. 5. Przebiegi czasowe (a, c) oraz spektrogramy (b, d): sygnału mowy zakłóconego generowanym szumem (a, b) oraz sygnału na wyjściu filtru adaptacyjnego (c, d). Odległość mówcy od mikrofonu wynosi 0,8 m
Fig. 5. *Time plots (a, c) and spectrograms (b, d) of: speech signal corrupted by generated noise (a, b) and enhanced signal on the output of adaptive filter (c, d). Distance between microphone and speaker: 0,8 m*

sinusoidalny jest bowiem nadal skorelowany ze swoją kopią. Do obliczeń wykorzystano algorytm LMS.

Przykładem zastosowania filtracji adaptacyjnej do poprawy jakości nagrań rejestrowanych za pomocą więcej niż jednego mikrofonu może być próba odtworzenia treści rozmowy prowadzonej w pomieszczeniu, w którym znajduje się inne źródło dźwięku. Taka sytuacja może mieć miejsce np.: w restauracji, w której włączone jest radio, czy też w samochodzie lub w kabinie pilotów w samolocie, gdzie generatorem sygnałów zakłócających są pracujące silniki. Niekiedy interlokutorzy, chcąc zachować poufność prowadzonej rozmowy, celowo prowadzą ją przy włączonym telewizorze lub stosują rozwiązania w postaci systemów ochrony akustycznej. Takie systemy mogą być wyposażone w generator szumu lub innych sygnałów zakłócających o intensywności zależnej od głośności prowadzonej rozmowy. Aby rozmówcy mogli się wzajemnie słyszeć, urządzenia tego typu wyposażane są w słuchawki z mikrofonami.

Na rycinie 4 zamieszczono rysunek poglądowy ilustrujący rozmieszczenie mikrofonów podczas testu z wykorzystaniem generatora szumu. Zastosowano tu dwa mikrofony o charakterystyce kierunkowej (hiperkardioidalnej). Nagrania zrealizowano dla dwóch różnych odległości od mówcy: 0,8 m oraz 1,2 m oraz od sygnału zakłócającego: 0,2 m oraz 0,8 m. Mikrofony były oddalone od siebie o 10 cm. Zapisu dokonano za pomocą przenośnego cyfrowego rejestratora DAT z częstotliwością próbkowania 44,1 kHz. Po wstępnej analizie uzyskanych nagrań stwierdzono brak występowania w nich jakiejkolwiek informacji lingwistycznej. Wygenerowany szum całkowicie zamaskował sygnał mowy. W celu wykonania korekcji nagrania poszczególne ścieżki poddano synchronizacji, aby zniwelować wpływ opóźnienia sygnału zakłócającego docierającego do poszczególnych mikrofonów. Następnie nagrania poddano filtracji adaptacyjnej z wykorzystaniem filtrów kratowych. Na rycinach 5 oraz 6 przedstawiono przebiegi czasowe i spektrogramy nagrań źródłowych oraz przetworzonych. Dzięki zastosowanej korekcji możliwe było pełne odtworzenie zapisu prowadzonej rozmowy.

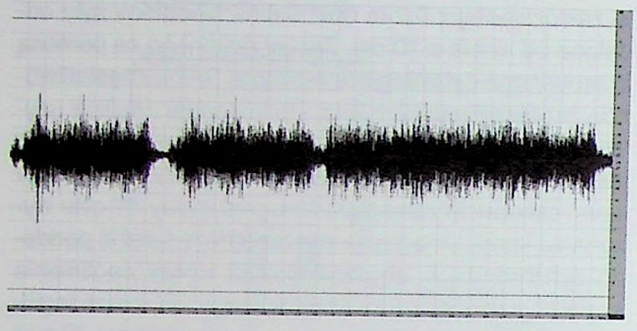
Podczas wykonywania filtracji adaptacyjnej niezwykle istotna jest prawidłowa synchronizacja poszczególnych ścieżek audio. Na rycinie 7 przedstawiono efekt filtracji nagrania rejestrowanego w odległości 0,8 m od mówcy z wykorzystaniem tych samych filtrów, jednakże bez wstępnej synchronizacji. Tym razem tylko w nieznacznym stopniu udało się zredukować zakłócenia, a zapis mowy pozostał nieczytelny.



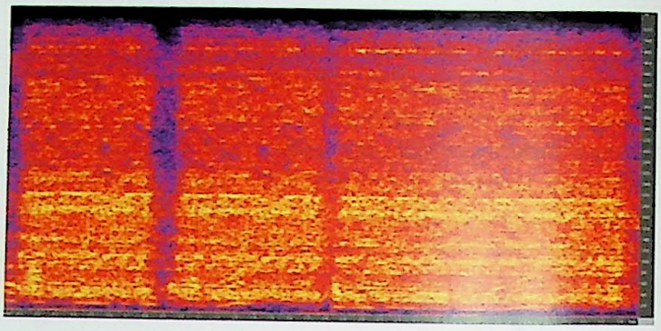
Ryc. 6. Przebiegi czasowe oraz spektrogramy: sygnału mowy zakłóconego generowanym szumem (a, b) oraz sygnału na wyjściu filtru adaptacyjnego (c, d). Odległość mówcy od mikrofonu wynosi 1,2 m

Fig. 6. *Time plots (a, c) and spectrograms (b, d) of: speech signal corrupted by generated noise (a, b) and enhanced signal on the output of adaptive filter (c, d). Distance between microphone and speaker: 1,2 m*

a)



b)



Ryc. 7. Przebieg czasowy (a) oraz spektrogram (b) sygnału mowy zakłóconego generowanym szumem. Sygnał z wyjścia filtru adaptacyjnego przy niewłaściwie dobranych parametrach synchronizacji (przesunięcie o 10 ms w stosunku do optymalnego punktu synchronizacji). Odległość mówcy od mikrofonu wynosi 0,8 m

Fig. 7. *Time plot (a) and spectrogram (b) of speech signal corrupted by generated noise. Enhanced signal on the output of adaptive filter without proper synchronization (10 ms shift in relation to optimum synchronization point). Distance between microphone and speaker: 0,8 m*

Separacja źródeł

Nagrania przekazywane do badań ekspertom z zakresu inżynierii dźwięku i fonoskopii zawierają wiele komponentów, np. głosy kilku mówców, muzykę lub inne sygnały zakłócające. Część z nich uznaje się za niepożądane i próbuje się je usuwać lub przynajmniej redukować za pomocą specjalistycznego oprogramowania. Dostępne narzędzia pozwalają przeważnie na filtrację szumu lub sygnałów harmonicznych[6]. Problem pojawia się w sytuacji, gdy poszczególne wypowiedzi mówców zarejestrowanych w nagraniu zaczynają się na siebie nakładać. Przy zbliżonych barwach głosów uczestników rozmowy może to prowadzić do błędów przy prowadzeniu identyfikacji w obrębie materiału dowodowego. Może także uniemożliwić ekstrakcję osobniczych cech mówców, co jest niezwykle istotne w procesie identyfikacji z wykorzystaniem materiału porównawczego.

Rejestracja nagrania za pomocą dwóch mikrofonów znacznie ułatwia transkrypcję oraz przypisywanie wypowiedzi poszczególnym uczestnikom rozmowy. Jest to spowodowane naturalną zdolnością ludzkiego mózgu do wykonywania przestrzennej filtracji i koncentrowania się tylko na wybranym źródle dźwięku (tzw. efekt *cocktail party*) [15]. Rejestracja stereofoniczna prowadzona jest najczęściej za pomocą dwóch mikrofonów ustawionych blisko siebie (np. dyktafon z wbudowanymi mikrofonami). Dzięki temu uzyskuje się możliwość subiektywnej lokalizacji źródeł dźwięku. Każdy z mikrofonów rejestruje sygnały pochodzące od wszystkich mówców, jednakże proporcje między nimi są różne. Można to zapisać w postaci układu równań:

$$\begin{cases} x_1(t) = a_{11}s_1(t) + a_{12}s_2(t) \\ x_2(t) = a_{21}s_1(t) + a_{22}s_2(t) \end{cases} \quad (6),$$

gdzie a_{11}, a_{12}, a_{21} oraz a_{22} są parametrami zależnymi od odległości źródeł dźwięku od mikrofonów, s_1 i s_2 są źródłami dźwięku (sygnału mowy), natomiast x_1 i x_2 są mieszaninami sygnałów zarejestrowanymi przez poszczególne mikrofony. Przy znanych wartościach parametrów a_{ij} rozwiązanie układu równań nie stanowiłoby problemu. Niestety, powyższy zestaw zmiennych można jedynie estymować, co czyni opisane zagadnienie dużo bardziej złożonym.

Jedną z możliwości rozwiązania powyższego problemu jest wykorzystanie statystycznych własności sygnałów $s_i(t)$ do wyznaczenia parametrów a_{ij}. Przyjmuje się, że rejestrowane składowe są liniową kombinacją pewnych nieznanych zmiennych, przy czym sposób, w jaki zostały połączone, także nie jest znany. Zakłada się ponadto, że szukane sygnały są niegaussowskie i wzajemnie niezależne, zatem nazywane są składowymi niezależnymi lub źródłami.

Separacji sygnałów pochodzących z wielu źródeł można dokonać za pomocą metody analizy składowych niezależnych (ICA – *Independent Component Analysis*). Jest to technika wykorzystywana w wielu dziedzinach, takich jak: przetwarzanie obrazów, przeszukiwanie baz danych dokumentów, ekonometria czy obrazowanie medyczne (np. elektroencefalografia). Jest przydatna wszędzie tam, gdzie w wyniku pomiarów otrzymuje się wiele sygnałów lub serii danych, które następnie należy rozdzielić. Powyższy problem opisywany jest terminem ślepa separacja źródeł (BSS – *Blind Source Separation*), ze względu na niewielką ilość informacji zarówno o samych źródłach, jak i o sposobie ich połączenia. Typowe zadania realizowane za pomocą algorytmów ICA to rozplatanie: sygnałów mowy pochodzących od wielu mówców (i rejestrowanych za pomocą wielu mikrofonów), zapisów fal mózgowych zarejestrowanych za pomocą wielu sensorów, nakładających się sygnałów radiowych docierają-

cych do telefonów bezprzewodowych lub – szczególnie w przemyśle – analiza równoległych serii danych otrzymywanych z wielu czujników [16].

Metoda analizy składowych niezależnych jest zatem statystyczną techniką dekompozycji złożonych grup danych na niezależne podgrupy. W przypadku gdy dwa zarejestrowane sygnały są od siebie niezależne, tzn. obserwacja jednego z nich nie pozwala na znalezienie informacji na temat drugiego, dzięki ślepej separacji źródeł możliwe jest rozdzielenie sygnałów tworzących superpozycję. Problem opisany równaniem (6) można przedstawić, stosując zapis macierzowy:

$$x = As \tag{7},$$

gdzie **s** jest wektorem zawierającym niezależne sygnały źródłowe, **A** jest macierzą mieszającą (kompozycji), natomiast **x** jest wektorem zawierającym zmiksowane sygnały [15]. Istotne jest, aby liczba obserwowanych składników mieszaniny sygnałów (np. liczba mówców) była mniejsza lub równa liczbie zastosowanych sensorów (mikrofonów) [17]. Można także przyjąć założenie, że współczynniki kompozycji a_{ij} są na tyle różne, aby pozwolić na wyznaczenie macierzy **W** odwrotnej do macierzy **A**. Wówczas rozwiązanie problemu miałoby postać:

$$\begin{cases} s_1(t) = w_{11}x_1(t) + w_{12}x_2(t) \\ s_2(t) = w_{21}x_1(t) + w_{22}x_2(t) \end{cases} \tag{8}.$$

Pierwszym krokiem większości algorytmów wykorzystujących metodę analizy składowych niezależnych jest wyśrodkowanie danych (*centering*), tzn. usunięcie wartości średniej E{**x**}. Operacja wykonywana jest jedynie w celu uproszczenia obliczeń i nie oznacza braku możliwości estymacji wartości średniej separowanych sygnałów. Następnie dokonuje się wybielenia danych (*whitening*). Przez liniową transformację wektora *x* uzyskuje się wektor $\tilde{x}$, którego wartości są nieskorelowane, a ich wariancje są równe jedności. Wykonanie takiej operacji jest zawsze możliwe i przyczynia się do zmniejszenia liczby parametrów koniecznych do estymacji. Jedną z częściej stosowanych metod wybielania jest dekompozycja wartości własnych macierzy kowariancji [18].

Algorytm ICA

Wyznaczanie składowych niezależnych odbywa się w sposób iteracyjny. W każdej iteracji aktualizowane są wartości wektora wag **w**. Algorytm ICA polega na maksymalizacji niegaussowości (*nongaussianity*) $\mathbf{w}^T\mathbf{x}$ mierzonej w postaci negentropii $J(\mathbf{w}^T\mathbf{x})$, którą definiuje się następująco:

$$J(y) = H(y_{gauss}) - H(y) \tag{9},$$

gdzie *H* oznacza entropię, natomiast $\mathbf{y}_{gauss}$ jest gaussowską zmienną losową o takiej samej macierzy kowariancji jak **y**. Negentropia jest zawsze nieujemna i osiąga zero tylko wtedy, gdy ma rozkład Gaussa. Ponadto wariancja $\mathbf{w}^T\mathbf{x}$ musi być ograniczona do jedności, co w przypadku wybielonych danych sprowadza się do warunku: $||\mathbf{w}||^2 = 1$. Jako przykład praktycznej realizacji może posłużyć algorytm *FastICA*, szeroko opisany w literaturze [18]:

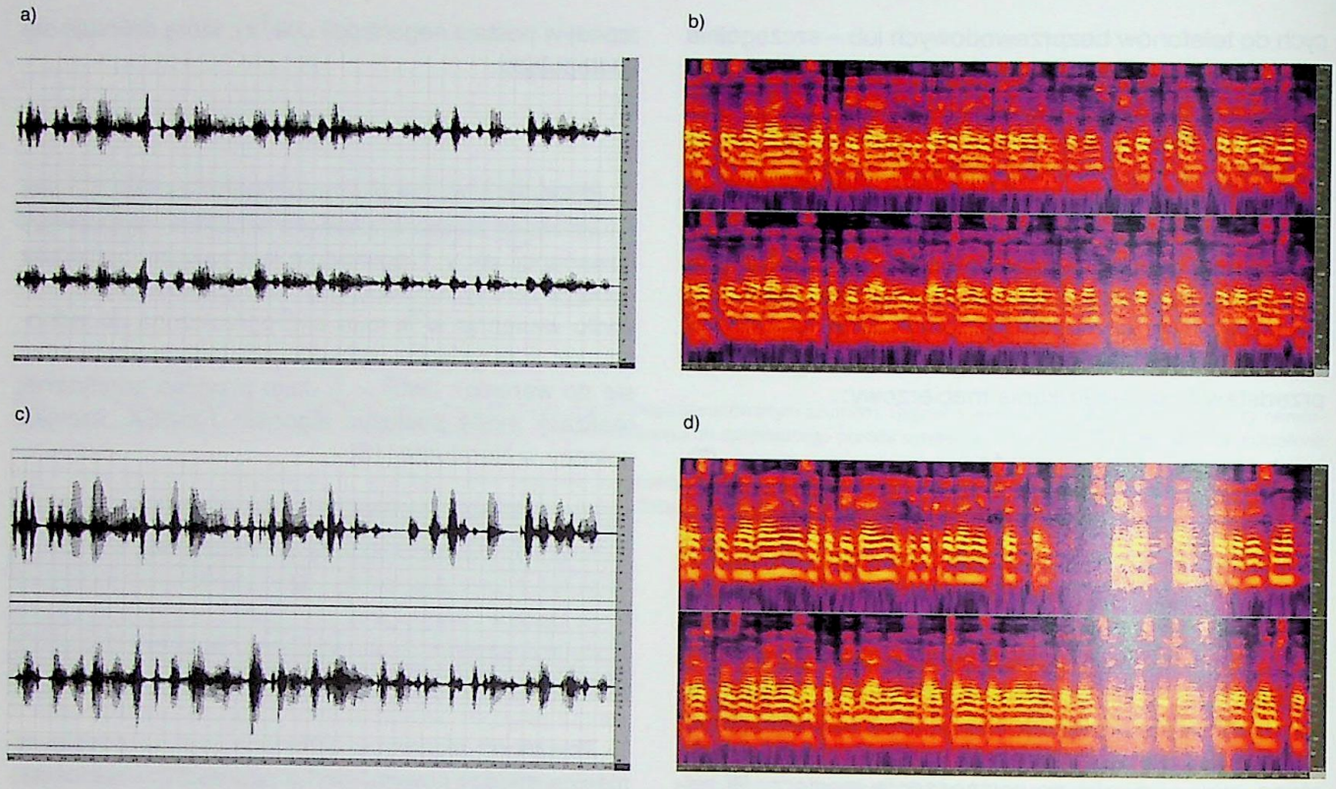
1. Wybierz wartości początkowe wektora wag **w**
2. Niech $\mathbf{w}^+ = E\{\mathbf{x}g(\mathbf{w}^T\mathbf{x})\} - E\{g'(\mathbf{w}^T\mathbf{x})\}\mathbf{w}$
3. Niech $\mathbf{w}^+ = \mathbf{w}^+/||\mathbf{w}^+||$
4. Jeśli algorytm nie jest zbieżny, przejdź do **2**

Zbieżność algorytmu określana jest na podstawie różnicy między nowymi ($\mathbf{w}^+$) a poprzednimi wartościami wektora **w**. Wymienione powyżej funkcje *g* oraz *g'* są pierwszą oraz drugą pochodną niekwadratowych funkcji *G*, które dobierane są eksperymentalnie, np.:

$$G_1(u) = 1/a_1 \log\cosh a_1 u, \quad G_2(u) = -exp(-u^2/2) \tag{10},$$
$$g_1(u) = \tanh(a_1 u), \quad g_2(u) = u\, exp(-u^2/2), \tag{11},$$

gdzie *a* jest stałą dobieraną doświadczalnie i mieści się w granicach $1 \le a \le 2$ (najczęściej $a = 1$). W praktyce także wartość oczekiwaną należy zastąpić jej estymatą wyznaczoną na podstawie odpowiednio dobranej krótkiej serii danych. Algorytm pozwala na wyznaczenie tylko jednej składowej $\mathbf{w}^T\mathbf{x}$. Aby zwiększyć ich liczbę, obliczenia należy zrealizować osobno dla każdej składowej $\mathbf{w}_1,\ldots,\mathbf{w}_n$. W celu przeciwdziałania osiąganiu przez poszczególne wektory tego samego maksimum, konieczna jest dekorelacja wyników $\mathbf{w}_1^T\mathbf{x},\ldots,\mathbf{w}_n^T\mathbf{x}$ po każdej iteracji. Można to osiągnąć, stosując, np. ortogonalizację Grama-Schmidta [17].

Opisane rozwiązanie postanowiono przetestować na nagraniu stereofonicznym powstałym przez zmiksowanie dwóch nagrań monofonicznych przesuniętych w panoramie odpowiednio 60% w lewo oraz 20% w prawo. Tak utworzony zapis zawierał głosy obydwu mówców zarówno w jednym, jak i w drugim kanale, co znacznie utrudniało zrozumienie poszczególnych wypowiedzi. Dzięki zastosowaniu opisanego algorytmu możliwe było całkowite rozdzielenie poszczególnych wypowiedzi (ryc. 8).



Ryc. 8. Przebiegi czasowe (a, c) oraz spektrogramy (b, d): nagrania będącego mieszaniną dwóch zapisów monofonicznych (a, b) oraz nagrania powstałego po zastosowaniu algorytmu opisanego w [18]
Fig. 8. *Time plots (a, c) and spectrograms (b, d) of: convoluted mixture of two single channel recordings (a, b) and output recording after using the algorithm described in [18]*

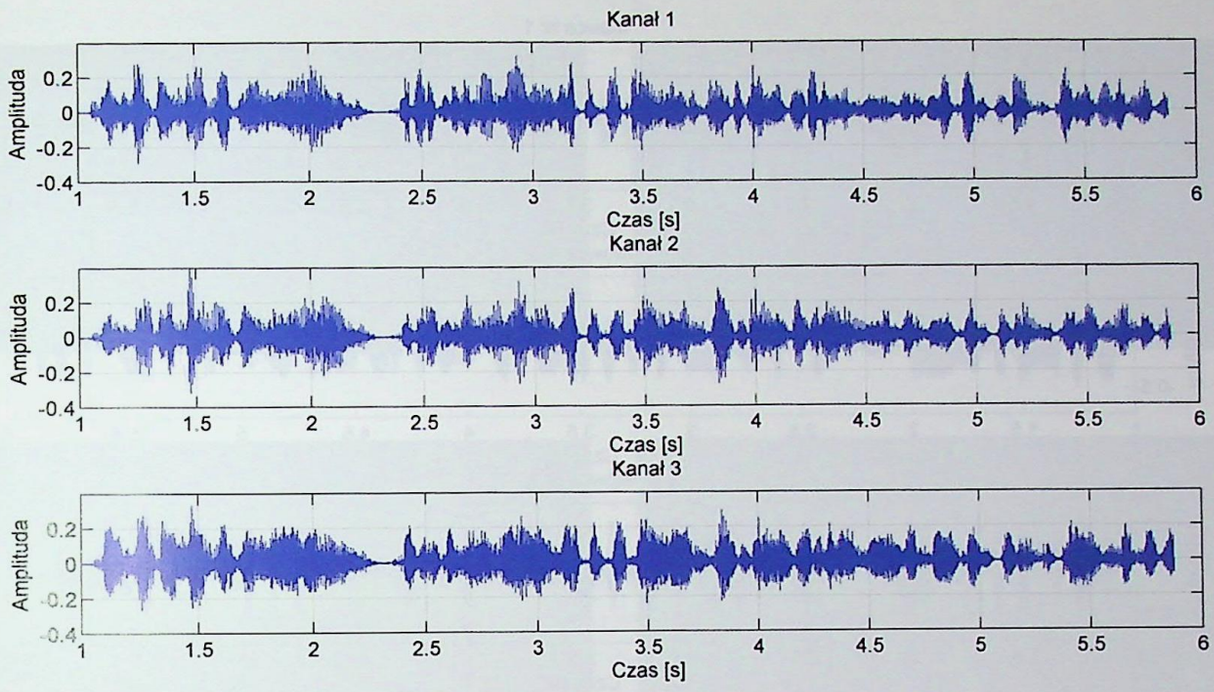
Naturalnie liczbę mówców w mieszaninie można zwiększać, o ile odpowiednio rosnąć będzie także liczba sensorów. Powyższe założenie sprawdzono dla nagrania trójkanałowego i zapisów wypowiedzi trzech osób, które zostały połączone według poniższej zależności (9):

$$\begin{cases} Kanał\ 1 = 0{,}6s_1 + 0{,}3s_2 + 0{,}3s_3 \\ Kanał\ 2 = 0{,}3s_1 + 0{,}7s_2 + 0{,}3s_3 \\ Kanał\ 3 = 0{,}3s_1 + 0{,}4s_2 + 0{,}6s_3 \end{cases} \quad (12),$$

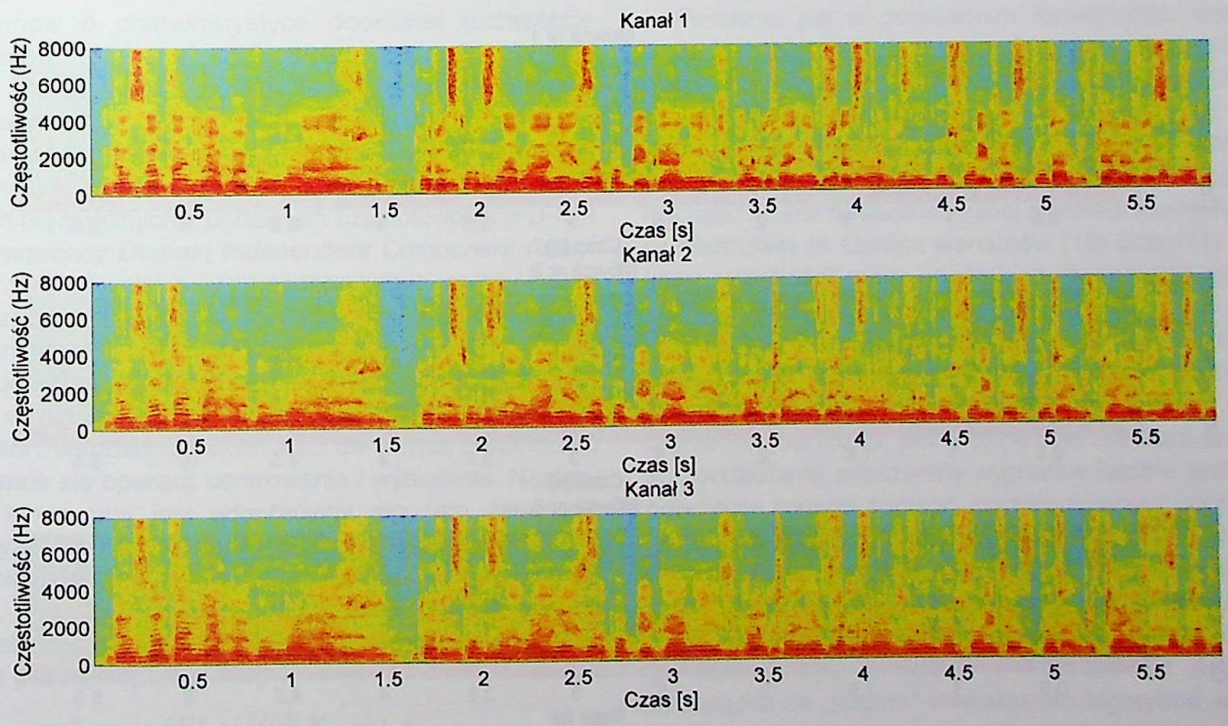
gdzie jako s_i oznaczono zapisy pochodzące od poszczególnych mówców. Przebiegi czasowe oraz spektrogramy dla każdego z kanałów przedstawiono na rycinach 9 oraz 10. Także i tym razem możliwa była pełna separacja poszczególnych źródeł. Wyniki separacji przedstawiono na rycinach 11 oraz 12.

Algorytm ICA z wykrywaniem kierunku (DOA)

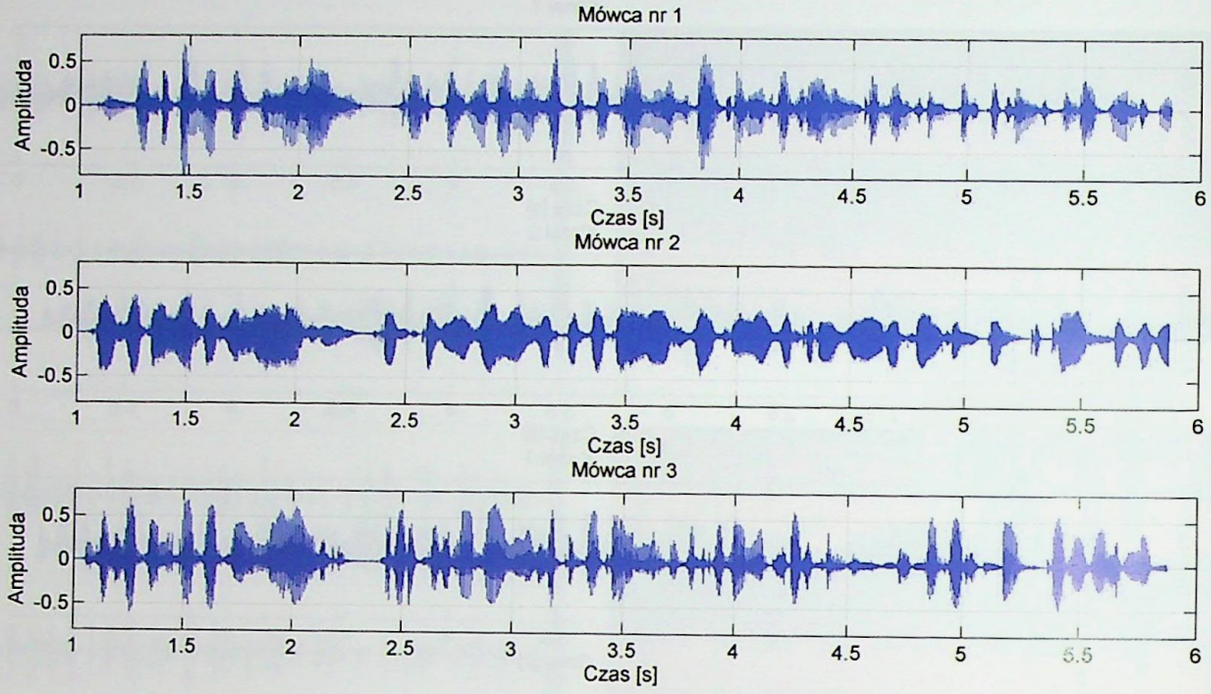
Opisany powyżej problem analizy składowych niezależnych został zdefiniowany przy założeniu, że każdy z sensorów (mikrofonów) rejestruje sygnał mieszaniny źródeł z zachowaniem różnych proporcji między tymi źródłami. Odpowiada to występowaniu macierzy miksującej **A** (por. równanie (3) oraz (4)). W przypadku ograniczenia liczby źródeł do dwóch modelowany system rejestracji można porównać do stereofonii natężeniowej, w której lokalizacja źródeł dźwięku następuje na podstawie różnicy głośności między poszczególnymi kanałami. Niestety, większość stereofonicznych nagrań analizowanych przez ekspertów z zakresu inżynierii dźwięku i fonoskopii tworzona jest z wykorzystaniem przenośnych rejestratorów, w których odległości między mikrofonami nie są duże (rzędu 2–3 cm). Skutkuje to niewielkimi różnicami natężeń między poszczególnymi kanałami. Taki system rejestracji należy zatem modelować jako stereofonię fazową, w której lokalizacja źródeł dźwięku następuje na podstawie różnicy czasów dotarcia poszczególnych sygnałów do mikrofonów. Różnice między opisanymi systemami zostały zilustrowane na rycinach 13a i 13b, gdzie przedstawiono wykresy panoramy dla nagrania dwukanałowego będącego mieszaniną dwóch zapisów monofonicznych (przesuniętych w panoramie odpowiednio 60% w lewo oraz 20% w prawo – por. ryc. 8) oraz nagrania stereofonicznego będącego rejestracją jednoczesnych wypowiedzi dwóch osób utrwalonych za pomocą dwóch mi-



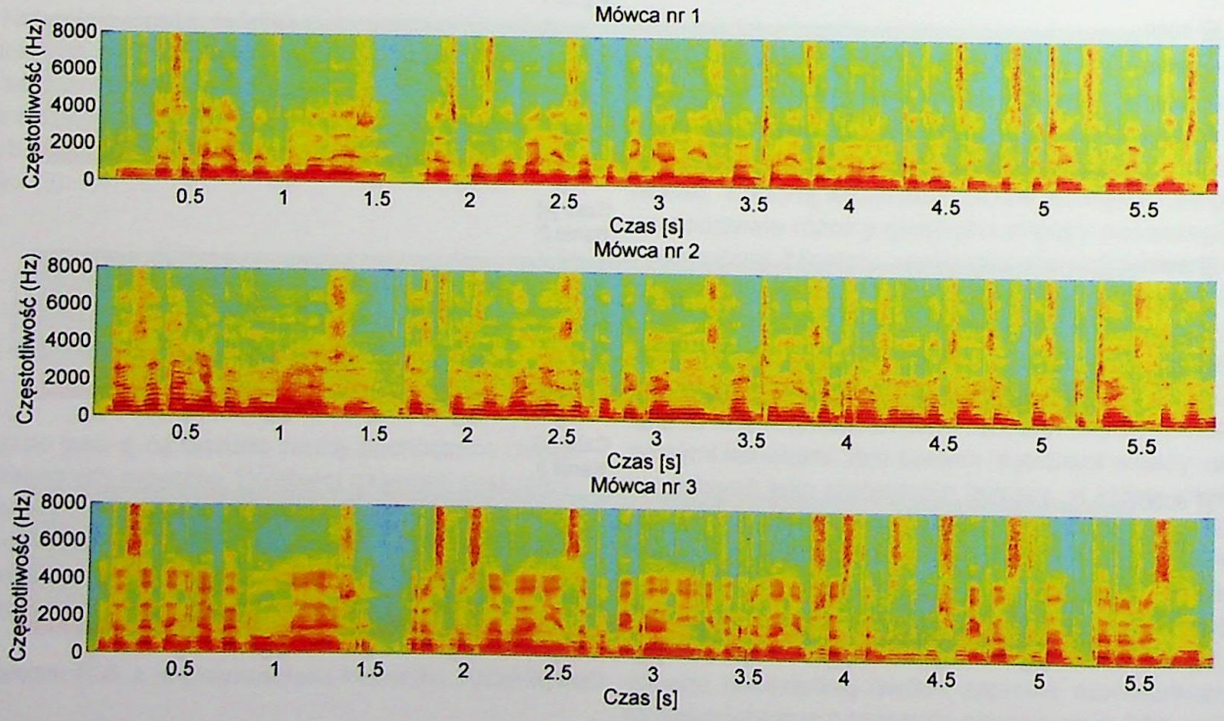
Ryc. 9. Przebiegi czasowe sygnałów mowy trzech mówców zmiksowane według zależności (9)
Fig. 9. *Time plots of speech signals coming from three speakers, mixed as denoted in equation (9)*



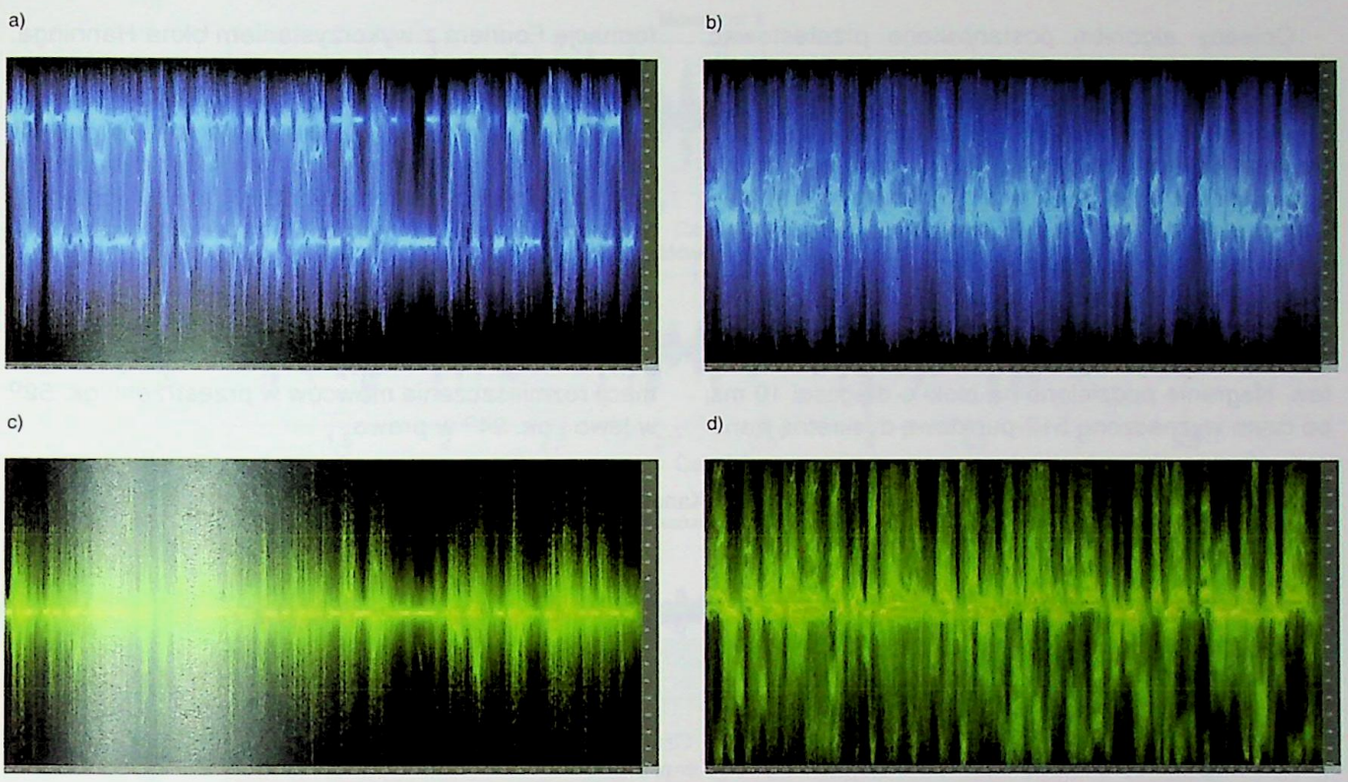
Ryc.10. Spektrogramy sygnałów mowy trzech mówców zmiksowane według zależności (9)
Fig. 10. *Spectrograms of speech signals coming from three speakers, mixed as denoted in equation (9)*



Ryc.11. Przebiegi czasowe rozdzielonych sygnałów mowy trzech mówców
Ryc.11. *Time plots of deconvoluted speech signals of three speakers*



Ryc.12. Spektrogramy rozdzielonych sygnałów mowy trzech mówców
Ryc.12. *Spectrograms of deconvoluted speech signals of three speakers*



Ryc.13. Porównanie wykresów panoramy (a, b) oraz charakterystyk fazowych (c, d) dla dwóch nagrań stereofonicznych: nagrania będącego mieszaniną dwóch zapisów monofonicznych (a, c) oraz zarejestrowanego za pomocą pary mikrofonów (b, d)
Fig. 13. *Comparison of panorama plots (a, b) and phase plots (c, d) of two stereo recordings: made by using convoluted mixture of two single channel recordings (a, c) and made by using two microphones (b, d)*

krofonów o charakterystyce dookólnej (ustawienie mówców: –60° oraz +20° od osi mikrofonów). W obydwu przypadkach subiektywne wrażenia przestrzenne i lokalizacja mówców były zbliżone.

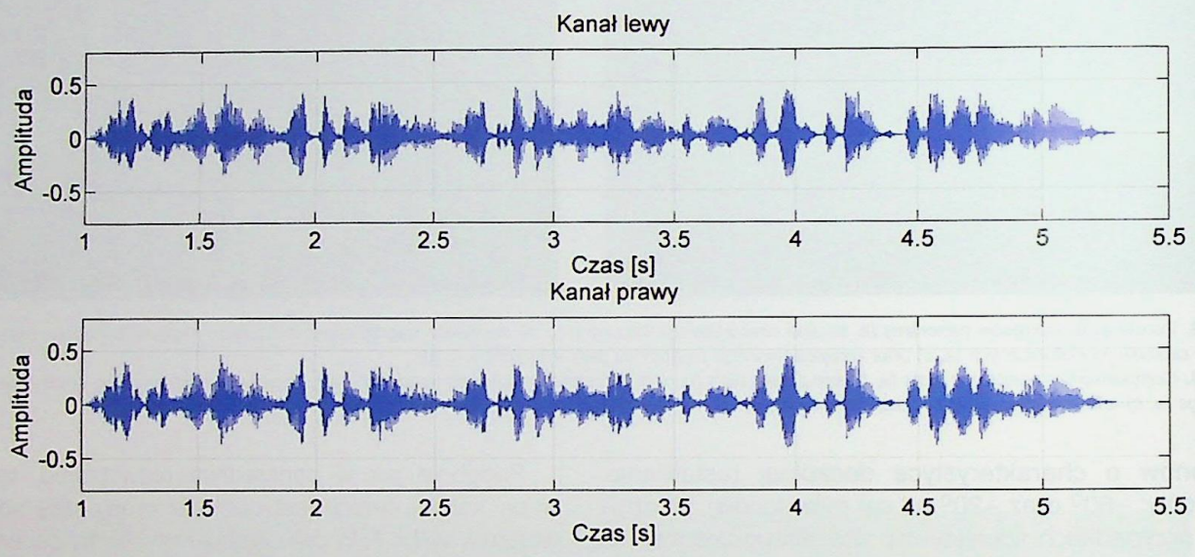
Mając na uwadze inny model tworzenia mieszaniny sygnałów, można zastosować algorytm analizy składowych niezależnych z podziałem częstotliwości (FD-ICA – *Frequency Domain Independent Component Analysis*). Sygnał wejściowy z każdego kanału należy zatem podzielić na ramki z wykorzystaniem funkcji okna (np. Hanninga), a następnie wyznaczyć dyskretną transformację Fouriera dla każdego zestawu próbek. Podobnie jak w przypadku opisanego wcześniej rozwiązania, wektory wartości częstotliwości dla każdej chwili czasu poddaje się operacji centrowania i wybielania. Następnie stosowany jest adaptacyjny algorytm iteracyjny, który wyznacza wartości macierzy separacji $\mathbf{W}(f)$ dla każdej grupy danych „czas-częstotliwość" w taki sposób, aby maksymalizować negentropię. Wzór opisujący sposób modyfikacji wartości wektorów macierzy $\mathbf{W}(f)$ jest następujący [19]:

$$\begin{aligned} \boldsymbol{w}^{+} &= \boldsymbol{w}(E\{g(|\boldsymbol{w}^{H}\boldsymbol{X}_{w}|^{2}) + \\ &+ (|\boldsymbol{w}^{H}\boldsymbol{X}_{w}|^{2})g'(|\boldsymbol{w}^{H}\boldsymbol{X}_{w}|^{2})\}) \\ &-E\{g(|\boldsymbol{w}^{H}\boldsymbol{X}_{w}|^{2})(\boldsymbol{X}_{w}^{H}\boldsymbol{w})\boldsymbol{X}_{w}\}. \end{aligned} \tag{13}$$

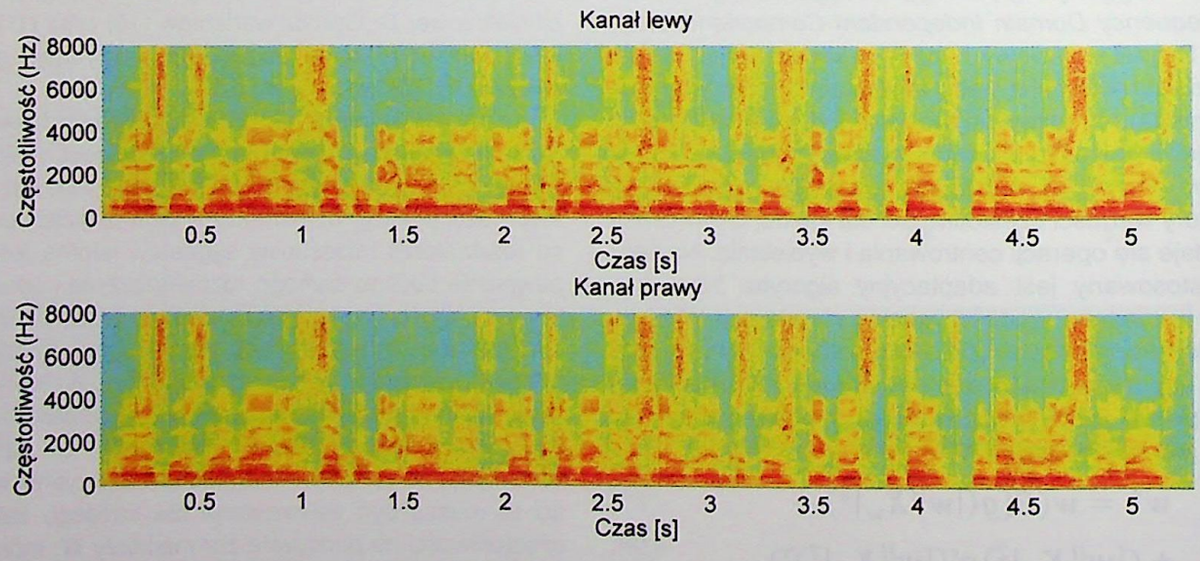
Podobnie jak w poprzednim rozwiązaniu, wektor $\mathbf{w}$ po każdej iteracji jest normalizowany, aby spełnić warunek $|\mathbf{w}^2| = 1$. W celu zwiększenia liczby separowanych źródeł obliczenia należy realizować osobno dla każdej składowej $\mathbf{w}_1,\ldots,\mathbf{w}_n$. Konieczna jest także dekorelacja wyników po każdej iteracji. Dokładność algorytmu zależna jest od zastosowanej niekwadratowej funkcji nieliniowej *G*. Oprócz wariantów (10) oraz (11) ciekawe własności mają uogólnione dystrybucje gaussowskie (GGD) opisane w [19].

Każdy wiersz macierzy separacji odpowiada wektorowi separacji dla różnych źródeł. Kolejność występowania wierszy macierzy **W** dla każdego zakresu częstotliwości jest różna, jednakże w celu realizacji procesu rozdzielania mieszaniny sygnałów istotne jest zapewnienie takiego samego rozmieszczenia i uporządkowania wektorów separacji dla każdego źródła w każdym zakresie częstotliwości. W tym celu stosuje się metodę polegająca na wyznaczaniu modelu kierunkowego (DP – *Directivity Pattern*) oraz kierunku, z którego dociera sygnał (DOA – *Direction Of Arrival*). W związku ze „ślepym" charakterem algorytmu wartości te muszą być estymowane dla każdego zakresu częstotliwości na podstawie zer macierzy **W**, które występują na pozycjach odpowiadających kierunkowi, z jakiego odbierany jest sygnał [20].

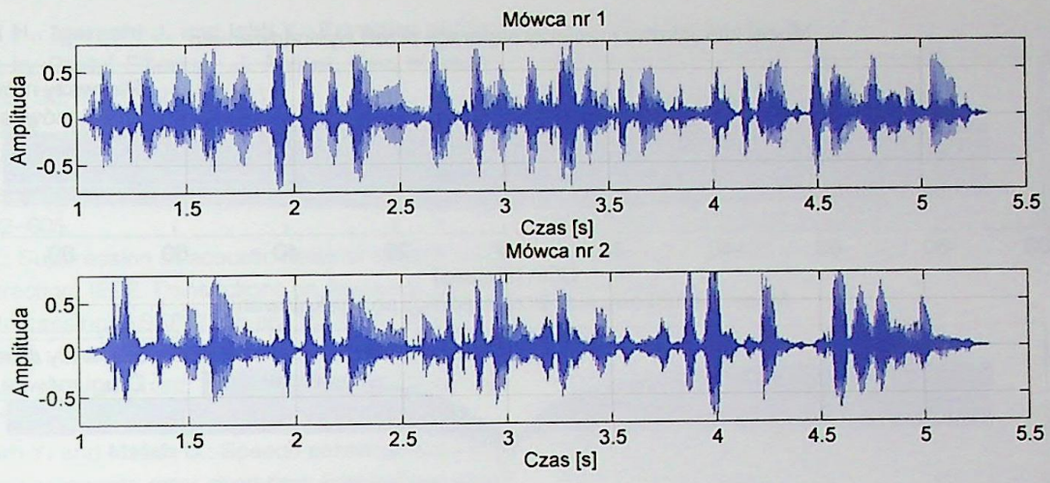
Opisany algorytm postanowiono przetestować, wykorzystując w tym celu stereofoniczne nagranie jednoczesnych wypowiedzi dwóch mówców zarejestrowane w pomieszczeniu biurowym za pomocą cyfrowego rejestratora DAT oraz dwóch mikrofonów o charakterystyce dookólnej (ustawienie mówców: -60^{o} oraz $+20^{o}$ od osi mikrofonów). Zapisu dokonano z częstotliwością próbkowania równą 44,1 kHz. Na rycinie 14 przedstawiono przebiegi czasowe stworzonego nagrania, natomiast na rycinie 15 zamieszczono spektrogramy dla poszczególnych kanałów. Nagranie podzielono na bloki o długości 10 ms, po czym wyznaczono 512-punktową dyskretną transformację Fouriera z wykorzystaniem okna Hanninga. Jako funkcję nieliniową przyjęto wybraną w [20] uogólnioną dystrybucję gaussowską (GGD). Na rycinie 16 przedstawiono przebiegi czasowe sygnałów mowy poszczególnych mówców po przeprowadzonej operacji separacji, natomiast na rycinie 17 ich spektrogramy. Z kolei na rycinie 18 zaprezentowano wykresy przedstawiające efekt wyznaczania modelu kierunkowego macierzy separacji (każda linia odpowiada jednemu prążkowi widma). Czerwonymi gwiazdkami na osi odciętych zaznaczono wynik estymacji rozmieszczenia mówców w przestrzeni: ok. 52^{o} w lewo i ok. 24^{o} w prawo.



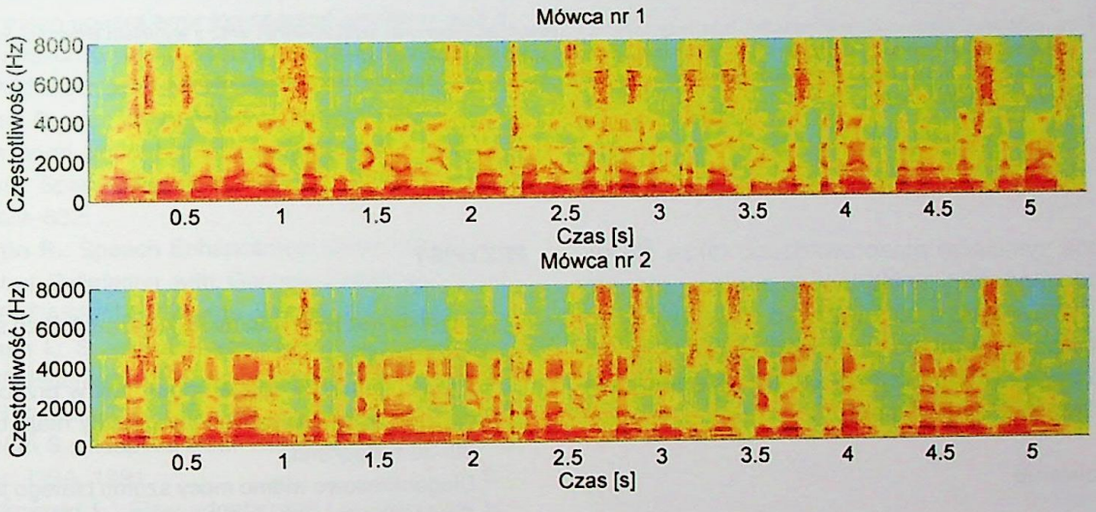
Ryc.14. Przebiegi czasowe zarejestrowanych za pomocą dwóch mikrofonów sygnałów mowy dwóch mówców
Fig. 14. *Time plots of speech signals of two speakers recorded by using two microphones*



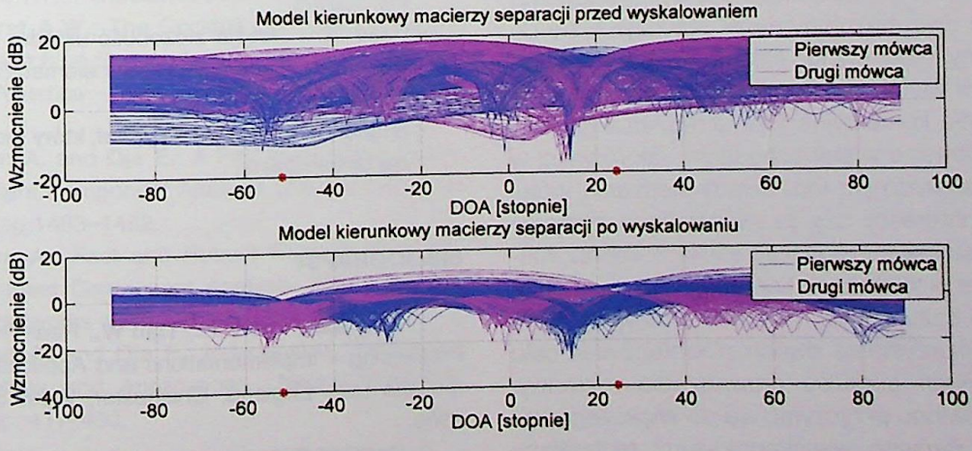
Ryc.15. Spektrogramy zarejestrowanych za pomocą dwóch mikrofonów sygnałów mowy dwóch mówców
Fig. 15. *Spectrograms of speech signals of two speakers recorded by using two microphones*



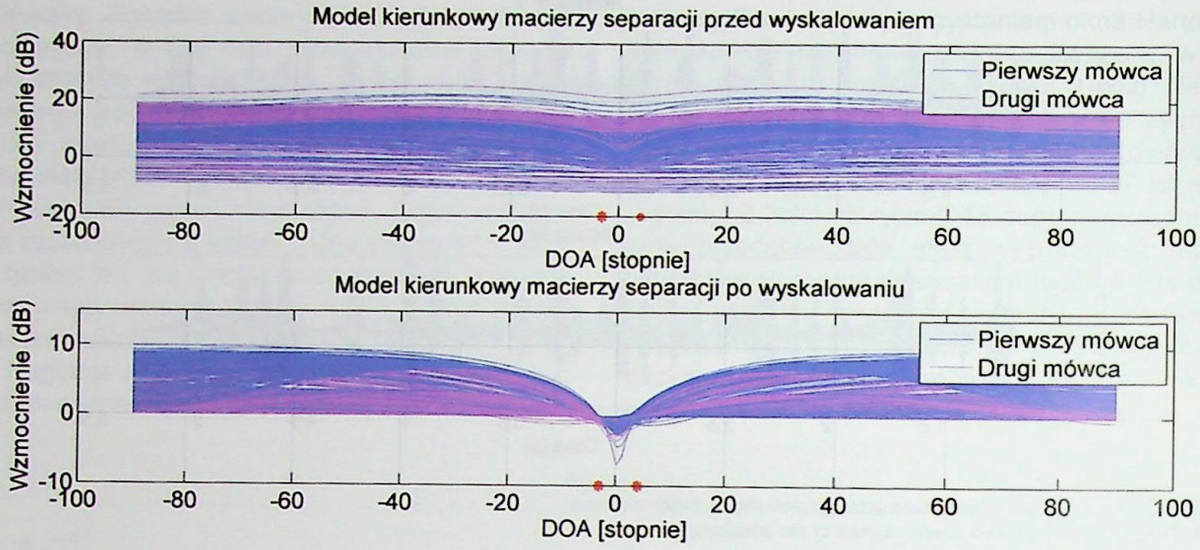
Ryc. 16. Przebiegi czasowe odseparowanych sygnałów mowy dwóch mówców
Fig. 16. *Time plots of separated speech signals of two speakers*



Ryc.17. Spektrogramy odseparowanych sygnałów mowy dwóch mówców
Fig. 17. *Spectrograms of separated speech signals of two speakers*



Ryc.18. Wykresy przedstawiające modele kierunkowe macierzy separacji (niewyskalowany oraz wyskalowany) wraz z wynikiem estymacji kierunku, z którego dociera sygnał (czerwone gwiazdki). Rejestracji dokonano za pomocą dwóch mikrofonów
Fig. 18. *Directivity patterns of separation matrix (unscaled and scaled) with estimation of direction of arrival (red asterix). Recording was made by using two microphones.*



Ryc.19. Wykresy przedstawiające modele kierunkowe macierzy separacji (niewyskalowany oraz wyskalowany) wraz z wynikiem estymacji kierunku, z którego dociera sygnał (czerwone gwiazdki). Nagranie jest mieszaniną dwóch zapisów monofonicznych
Fig. 19. *Directivity patterns of separation matrix (unscaled and scaled) with estimation of direction of arrival (red asterix). Recording is the convoluted mixture of two single channel recordings*

Podobną symulację przeprowadzono także dla nagrania będącego mieszaniną dwóch zapisów monofonicznych (60% w lewo o 20% w prawo – ryc. 19). Zgodnie z oczekiwaniami wyznaczony kierunek DOA był bliski 0^{o}, ponieważ w tym przypadku nie występuje przesunięcie fazy między kanałami.

Podsumowanie

Nagrania stereofoniczne stanowią mniejszość wśród materiałów dowodowych nadsyłanych do badań, jednak coraz większa dostępność dyktafonów wyposażonych w dwa mikrofony może przyczynić się do szybkiej zmiany tego stanu. Rośnie liczba dźwiękowych systemów wielokanałowych znajdujących zarówno zastosowanie komercyjne (np. urządzenia głośnomówiące instalowane w samochodach), jak również w sprzęcie specjalistycznym (np. układy śledzenia mówcy podczas konferencji czy zaawansowane systemy monitoringu instalowane podczas wielkich imprez masowych). Nawet w telefonach komórkowych instalowane są aplikacje bazujące na filtracji adaptacyjnej, które umożliwiają wykorzystanie dwóch mikrofonów w celu tłumienia zakłóceń podczas prowadzenia rozmowy. Wszystko to powinno przyczynić się do większego zainteresowania zarówno wielokanałowymi technikami redukcji szumu, jak i układami pozwalającymi na separację głosów poszczególnych mówców tłumie.

PRZYPISY

1 Sygnały addytywne to takie, które można ze sobą łączyć (sumować).
2 Przyjmuje się, że sygnał mowy jest w przybliżeniu stacjonarny, gdy analizowane segmenty mają długość ok. 20–30 milisekund.
3 Długookresowe widmo mocy szumu białego jest płaskie.
4 Rozróżnia się filtry FIR (*Finite Impulse Response*) o skończonej odpowiedzi impulsowej oraz filtry typu IIR (*Infinite Impulse Response*) o nieskończonej odpowiedzi impulsowej.
5 Rząd filtru określa złożoność układu; im większy rząd, tym więcej współczynników i elementów opóźniających tworzy filtr.
6 Sygnał harmoniczny to sygnał, który można opisać funkcją sinusoidalną.

BIBLIOGRAFIA

1. **Kuo S.M., Lee B.H., Tian W.**: Real-Time Digital Signal Processing – Implementations and Applications, John Wiley & Sons Ltd., England, Chichester, West Sussex, England 2006.

2. **Vaseghi S.V.**: Advanced Digital Signal Processing and Noise Reduction, John Wiley & Sons Ltd., England, Chichester, West Sussex, England 2006.

3. **Suzuki H.**, **Igarashi J.** and **Ishii Y.**: Extraction of Speech in Noise by Digital Filtering, „J. Acoust. Soc. of Japan" Aug. 1977, Vol. 33, No. 8, pp. 105–411.

4. **Curtis R.A.**, **Niederjohn R.J.**: An Investigation of Several Frequency – Domain Methods for Enhancing the Intelligibility of Speech in Wideband Random Noise, ICASSP, April 1978, pp. 602–605.

5. **Boll S.**: Suppression of acoustic noise in speech using spectral subtraction, IEEE Transactions on Acoustics Speech and Signal Processing, ASSP-27(2) pp. 113–120, 1979.

6. **Berouti M.**, **Schwartz R.**, **Makhoul J.**: Enhancement of speech corrupted by acoustic noise, IEEE ICASSP'79, Washington 1979, pp. 208–211.

7. **Ephraim Y.** and **Malah D.**: Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator, IEEE Transactions on Acoustics, „Speech, Signal Processing" Dec. 1984, vol. ASSP-32, no. 6, pp. 1109–1121.

8. **Ephraim Y.** and **Malah D.**: Speech enhancement using a minimum mean square error log-spectral amplitude estimator, IEEE Trans. on Acoust., „Speech, Signal Processing" Apr. 1985, vol. ASSP-33, pp. 443–445.

9. **Scalart P.** and **Vieira-Filho J.**: Speech enhancement based on a priori signal to noise estimation, 21st IEEE Int. Conf. Acoust. Speech Signal Processing, Atlanta, GA, May 1996, pp. 629–632.

10. **Martin R.**: Speech Enhancement Using MMSE Short Time Spectral Estimation with Gamma Distributed Speech Priors, IEEE ICASSP'02, Orlando, Florida, May 2002.

11. **Cohen I.**: Speech Enhancement Using a Noncausal A Priori SNR Estimator, IEEE Signal Processing Letters, Vol. 11, No. 9, Sep. 2004, pp. 725–728.

12. **Haykin S.**: Adaptive Filter Theory, Prentice-Hall International, Inc. USA, 1991.

13. **Rutkowski L.**: Filtry adaptacyjne i adaptacyjne przetwarzanie sygnałów, WNT, Warszawa 1994.

14. **Zieliński P.**: Cyfrowe Przetwarzanie Sygnałów, od teorii do zastosowań, WKŁ, Warszawa 2007, s. 205.

15. **Bronkhorst A.W.**: The Cocktail Party Phenomenon: A Review of Research on Speech Intelligibillity in Multiple-Talker Conditions, Acustica – acta acustica 2000, Vol. 86, pp. 117–128.

16. **Hyvärinen A.** and **Oja E.**: A Fast Fixed-Point Algorithm for Independent Component Analysis, „Neural Computation" 1997, 9(7), pp.1483–1492.

17. **Hyvärinen A.**: Fast and Robust Fixed-Point Algorithms for Independent Component Analysis, IEEE Transactions on Neural Networks 10(3), pp. 626–634, 1999.

18. **Hyvärinen A.** and **Oja E.**: Independent Component Analysis: Algorithms and Applications, „Neural Networks" 2000, 13(4–5), pp. 411–430.

19. **Prasad R.**, **Saruwatari H.**, **Shikano K.**: Blind Separation of Speech by Fixed-Point ICA with Source Adaptive Negentropy Approximation, IEICE Trans. Fundamentals, Vol. E-88A (7), 2005.

20. **Prasad R.**, **Saruwatari H.**, **Lee A.**, **Shikano K.**: A Fixed-Point ICA Algorithm for Convoluted Speech Signal Separation, 4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA 2003), 2003, Nara, Japan.

Streszczenie

W pracy nakreślono problem redukcji addytywnego szumu i zakłóceń w nagraniach jedno- i wielokanałowych oraz wyjaśniono zasadę działania algorytmów detekcji mowy i widmowej redukcji szumu. Opisano metody filtracji adaptacyjnej, które mogą zostać wykorzystane do poprawy zrozumiałości mowy w miejscach, w których stosowane są generatory szumu i zakłóceń. Przedstawiono także techniki ślepej separacji, które stosowane są w celu oddzielania głosów mówców z mieszanin rejestrowanych przez dwa lub więcej mikrofonów. Ponadto opisano techniki analizy składowych niezależnych wraz z metodami wyznaczania modelu kierunkowego i estymacją kierunku, z którego dociera dźwięk. Powyższe rozwiązania zostały omówione w kontekście poprawy jakości nagrań, a efekty ich działania zaprezentowano w postaci wykresów.

Słowa kluczowe: *korekcja nagrań, nagrania wielokanałowe, filtracja adaptacyjna, efekt cocktail party, ślepa separacja źródeł, analiza składowych niezależnych, ICA.*

Summary

The paper addresses the problem of additive noise and disturbance reduction in single and multichannel audio recordings. It explains several algorithms for speech detection and spectral subtraction of noise. It describes adaptive filtering methods, which can be used for speech intelligibility enhancement in noisy environments where noisy generators are used. The paper introduces also blind source separation methods, used in order to extract speakers' voice from convoluted mixtures recorded by two or more microphones. Further, it describes independent component analysis techniques with directivity pattern computation and arrival direction estimation. The paper presents the described tools in the context of audio enhancement. Their effectiveness is presented on sample plots.

Keywords: *audio enhancement, multichannel audio recordings, adaptive filtration, cocktail party phenomenon, blind source separation, independent component analysis, ICA.*