WYBRANE ASPEKTY DEKOMPOZYCJI ZŁOŻONYCH SYGNAŁÓW AKUSTYCZNYCH STEREOFONICZNYCH i WIELOKANAŁOWYCH

dr inż. Jacek Wierzbicki
Katedra Mechaniki i Wibroakustyki
AGH
Al. Mickiewicza 30
30 – 059 Kraków

mgr inż. Grzegorz Łukasz Augustyn
Katedra Mechaniki i Wibroakustyki
AGH
Al. Mickiewicza 30
30 – 059 Kraków

1. WSTĘP

Problem analizy i dekompozycji sygnałów akustycznych jest niezwykle trudny i złożony. Jest on obecny zarówno w obcowaniu z otaczającym nas środowiskiem, jak i w wirtualnej rzeczywistości. Często przypuszcza się, że jest to brakujące ogniwo do pełnego poznania i opisu zasad funkcjonowania ludzkiego organizmu, procesów słyszenia i percepcji bodźców. Przedstawienie wszystkich aspektów dekompozycji w ich pełnej formie jest rzeczą niemożliwą i zbyt obszerną, dlatego zakres publikacji został ograniczony do jednej konkretnej klasy sygnałów – sygnałów akustycznych oraz przedstawia jedynie wybrane jej metody (dla sygnałów stereofonicznych i wielokanałowych), jednak wyniki i metody opisane w tej pracy mogą być rozszerzane i stosowane w innych dziedzinach nauki, takich jak np. medycyna, astronomia, informatyka, przetwarzanie danych i sygnałów.

Dotychczasowe metody dekompozycji opracowane na podstawie narzędzi z teorii sygnałów, teorii zbiorów, optymalizacji oraz sztucznej inteligencji z różnym powodzeniem są stosowane do separacji cech danego sygnału. Nad problemami dekompozycji pracują największe ośrodki na świecie takie jak MIT Media Laboratory, Bell Laboratories i IBM Inc. w USA, Fraunhoffer Institut w Niemczech, firmy – Creative Labs, Philips. W Polsce zagadnieniami tymi zajmują się naukowcy z Politechniki Wrocławskiej, Uniwersytetu im. Adama Mickiewicza w Poznaniu, Politechniki Gdańskiej i Akademii Górniczo – Hutniczej w Krakowie. Mimo to liczba publikacji na ten temat jest bardzo mała, co jest związane ze stopniem skomplikowania problemu, zarówno pod względem teoretycznym jak i w postaci konkretnej aplikacji. Zwykle publikacje te dotyczą jednego z etapów dekompozycji sygnału lub zawierają opis metody skutecznej w przypadku bardzo wysublimowanej klasy sygnałów. Jednak metody te są ciągle rozwijane – niektóre z nich bazują na już znanych sposobach analizy sygnału, a niektóre wykorzystują nowe osiągnięcia nauki. Szczególnie dotyczy to dziedziny przetwarzania sygnałów, gdzie zaczyna się stosować metody sztucznej inteligencji, co – oprócz podniesienia stopnia skomplikowania metody – zwiększa skuteczność narzędzi używanych do dekompozycji.

Znamiennym dla czasów, w których żyjemy, jest postęp techniki informatycznej, a co za tym idzie, cyfrowego przetwarzania sygnałów, toteż w pracy skupiono się głównie na cyfrowej reprezentacji sygnału, gdyż takiej postaci dotyczą - w większości przypadków – wszelkie prace aplikacyjne, natomiast postać analogowa sygnału akustycznego stanowi często postać wyjściową do rozwiązania danego problemu i jest wygodna w opisie wielu zjawisk na gruncie teoretycznym. Mimo to należy zdawać sobie sprawę, że postać cyfrowa sygnału

stanowi jedynie niedoskonałe przybliżenie rzeczywistości, która, rozważana w naszej skali, jest jednak analogowa.

2. WYBRANE METODY DEKOMPOZYCJI DLA SYGNAŁÓW STEREOFONICZNYCH I WIELOKANAŁOWYCH

Rozwój nowoczesnych technologii doprowadził do rozwinięcia technik pomiarowych i audio – wizualnych bazujących na metodach rejestracji i odtwarzania sygnałów wielokanałowych (np. kino domowe). Popularność tego typu narzędzi spowodowała zapotrzebowanie na opracowanie metod skutecznie dekomponujących uprzednio uzyskane sygnały np. w celu ich powtórnego miksowania lub rozdzielenia obecnych w nich informacji.
Autorzy artykułu przeprowadzili eksperymenty związane głównie z sygnałami stereofonicznymi, ale umożliwiło to rozwinięcie teoretycznych podstaw zastosowanych algorytmów na sygnały wielokanałowe.
Pierwszą z metod jest metoda separacji sygnałów stereofonicznych. Podstawy tej metody wywodzą się z zasad stereofonii amplitudowo – fazowej. W tym przypadku sygnały docierające do uszu słuchacza są zróżnicowane zarówno w fazie, jak i amplitudzie. Dzięki temu, tworzy się tzw. panorama sygnału, która tworzy pozorną scenę muzyczną pozwalającą określić położenie poszczególnych źródeł dźwięku, a więc poniekąd rozpoznanie ich ilości i położenia w przestrzeni. Co więcej, dla niskich częstotliwości (duże wartości długości fali) efekty fazowe można zaniedbać bez znaczącego pogorszenia jakości separacji. Stosując więc odpowiedni algorytm dostrajania wag kombinacji liniowej naszych dwóch sygnałów w dziedzinie amplitud i faz (aczkolwiek przy wykorzystaniu wspomnianego uproszczenia, ten fragment można pominąć), możliwa jest separacja dwóch sygnałów nagranych w technice dwumikrofonowej, stereofonicznej.

Matematycznie możemy zapisać to w postaci układu równań, który dla czasowej reprezentacji układu będzie miał postać

$$\begin{aligned} y_1(t) &= a_{11}x_1(t) + a_{12}x_2(t - c_1\tau) \\ y_2(t) &= a_{21}x_1(t - c_2\tau) + a_{22}x_2(t) \end{aligned} \tag{2.1}$$

gdzie:
x_1, x_2 – sygnały zapisane na nośniku
a_{ij} – współczynniki kombinacji liniowej amplitud sygnałów (współczynniki układu równań)
τ - elementarna wartość opóźnienia czasowego
c_1, c_2 – współczynniki opóźnienia czasowego (jeżeli źródła i mikrofony tworzą układ symetryczny to $c_1=c_2=c$)

Rys. 2.1. Sposób rejestracji nagrania stereo duetu gitarowego Jim Hall i Pat Metheny w Rigth Track Studios w Nowym Jorku.

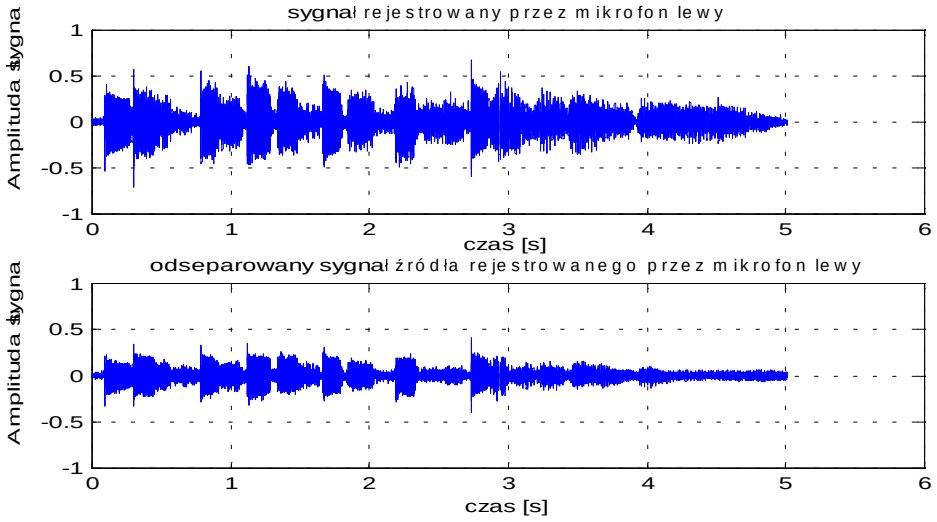
Tenże układ równań przypomina bardzo układ stosowany w metodzie BSSD (z ang. Blind Source Separation and Deconvolution) separacji źródeł [8], [9]. System i algorytm jest niemal identyczny, jednak stopień skomplikowania jest znacznie mniejszy. Cała trudność polega na prawidłowym określeniu opóźnienia elementarnego i jego współczynników (w praktyce pominięcie tego członu nieznacznie pogarsza osiągnięte rezultaty) oraz usunięcie pogłosu pomieszczenia, którego udział w wypadkowym sygnale staje się znaczący dopiero wtedy, gdy dane źródło zostało wyeliminowane, a jedynym śladem po jego istnieniu jest właśnie pogłos.

Jako doświadczenie wykorzystano realizację stereofoniczną, dwumikrofonową nagrania sesji studyjnej

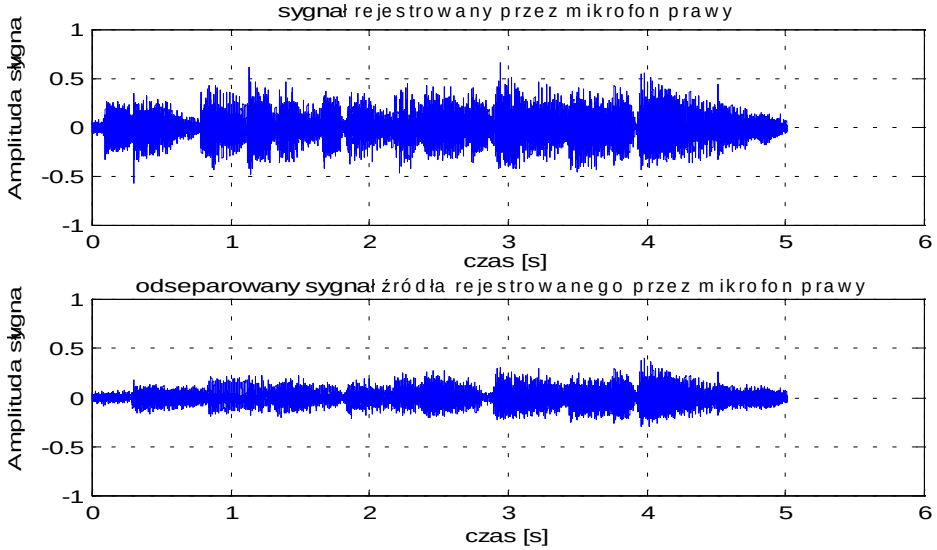
duetu Jim Hall i Pat Metheny, która odbyła się w Right Track Studios w Nowym Jorku. Ustawienie mikrofonów do rejestracji gry obu gitarzystów jest pokazany na rys. 2.1.

Nagranie zostało zrealizowane przez firmę Telarc w technologii 24-bitowej, dzięki czemu jakość nagrania jest najwyższa z możliwych. Do analizy wykorzystano utwór „Lookin' up" Jima Halla. Utwór poddano operacji resamplingu na częstotliwość 48 kHz, w celu łatwiejszego przetwarzania. Zależności fazowe obu sygnałów w początkowej fazie przetwarzania pominięto. Odpowiedni kod programu do separacji duetów nagranych techniką wielomikrofonową napisano w pakiecie inżynierskim MATLAB.

W efekcie uzyskano dość dobrą separację obu sygnałów, bazując jedynie na nagraniu z płyty kompaktowej bez słyszalnego pogorszenia jakości brzmieniowej i artystycznej badanego materiału muzycznego. Również wszelkie niuanse brzmieniowe, a także interpretacyjne zostały poprawnie przeniesione w procesie analizy. Efekty separacji są przedstawione na rys. 2.2 oraz 2.3 gdzie porównano ze sobą sygnał oryginalny każdego z kanałów z sygnałem dedykowanego instrumentu uzyskanym w wyniku dekompozycji.

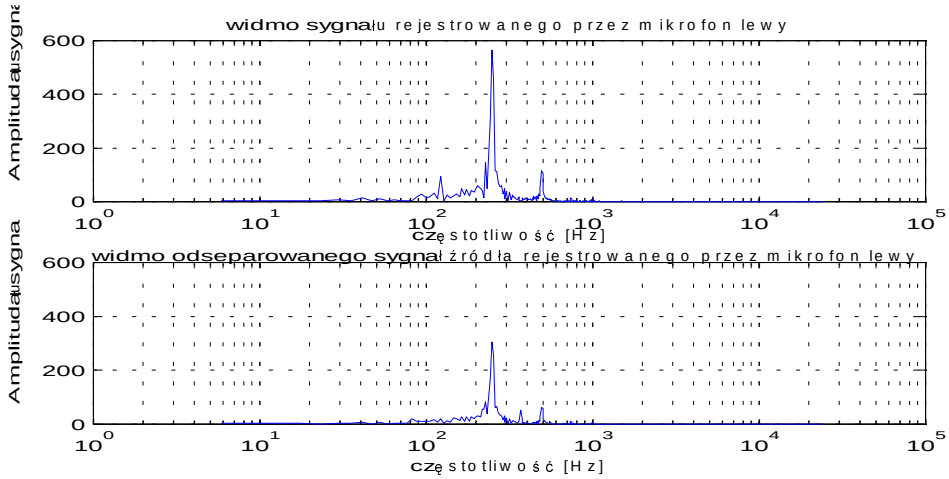


Rys. 2.2. Porównanie sygnału zarejestrowanego przez mikrofon lewy z sygnałem uzyskanym w wyniku operacji separacji.

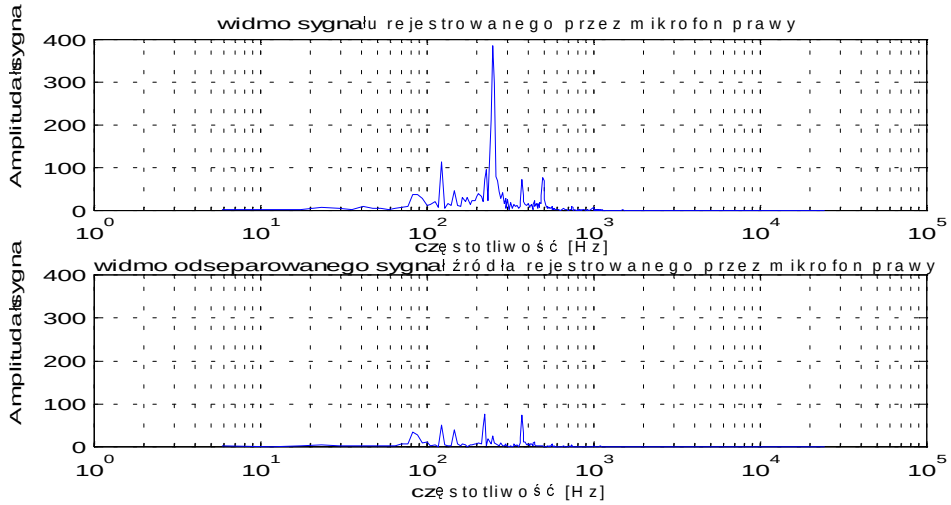


Rys. 2.3. Porównanie sygnału zarejestrowanego przez mikrofon prawy z sygnałem uzyskanym w wyniku operacji separacji źródeł.

Na rys. 2.4 oraz 2.5 można zaobserwować efekt działania algorytmu w dziedzinie częstotliwości, gdzie szczególnie jego skutki są widoczne w postaci wyraźnego uwidocznienia się składowych sygnału separowanego, w porównaniu do sygnału rejestrowanego.



Rys. 2.4. Efekt separacji źródeł dla sygnału rejestrowanego przez mikrofon lewy w dziedzinie częstotliwości.



Rys. 2.5. Efekt separacji źródeł dla sygnału rejestrowanego przez mikrofon prawy w dziedzinie częstotliwości.

Wyniki uzyskane tą metodą skłaniają do uczynienia kroku dalej i podania układu równań, dotyczącego układów wielomikrofonowych budowanych w postaci tablic lub sfer składających się z równomiernie rozmieszczonych nań mikrofonów. Układ równań znacznie się wtedy komplikuje, gdyż zależności pomiędzy sygnałami są już bardziej złożone, aczkolwiek nadal jest to kombinacja liniowa

$$\begin{aligned} y_1(t) &= a_{11}x_1(t-b_{11}\tau) + a_{12}x_2(t-b_{12}\tau) + \ldots + a_{1M}x_M(t-b_{1\,M-1}\tau) \\ &\ldots \\ y_N(t) &= a_{N1}x_1(t-b_{N1}\tau) + a_{N2}x_2(t-b_{N2}\tau) + \ldots + a_{NM}x_M(t-b_{NM}\tau) \end{aligned} \tag{2.2}$$

gdzie:
M – liczba źródeł M=N
N – ilość kanałów rejestracji

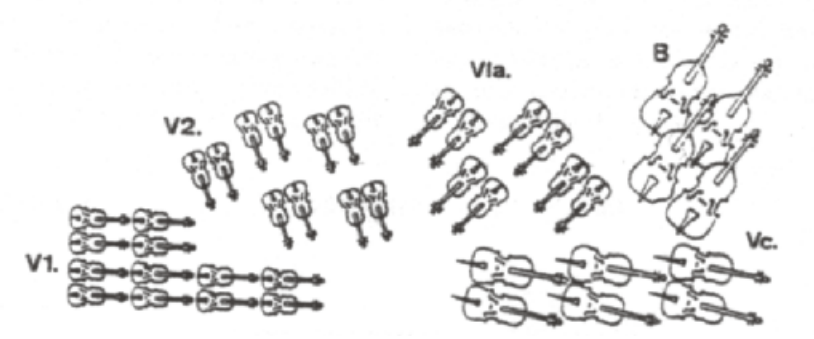
Układ powyższy można zapisać w postaci macierzowej i wtedy

$$\begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_N \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1M} \\ a_{21} & a_{22} & \dots & a_{2M} \\ \dots & \dots & \dots & \dots \\ a_{N1} & a_{N2} & \dots & a_{NM} \end{bmatrix} \cdot \begin{bmatrix} x_1(t-b_{11}\tau) & x_2(t-b_{12}\tau) & \dots & x_M(t-b_{1M}\tau) \\ x_1(t-b_{21}\tau) & x_2(t-b_{22}\tau) & \dots & x_M(t-b_{2M}\tau) \\ \dots & \dots & \dots & \dots \\ x_1(t-b_{N1}\tau) & x_2(t-b_{N2}\tau) & \dots & x_M(t-b_{NM}\tau) \end{bmatrix} \quad (2.3)$$

co stanowi układ równań, którego proponowana przez autora nazwa powinna brzmieć „układ równań dla tablicy wielomikrofonowej" (w jęz. ang. *multichannel microphone array source separation equations* - MMASSE). W ten sposób dobierając starannie współczynniki a_{ij} oraz b_{ij} tak, aby zawsze $b_{ii}=0$, powinniśmy skutecznie przeprowadzać separację wielu źródeł wykorzystując technikę wielomikrofonową. Szczególnie przydatne jest to w technice dźwięku dookólnego, który zdobywa sobie coraz większą popularność wśród odbiorców. Umożliwiłoby to precyzyjniejsze ustalanie opóźnień czasowych i stosunku amplitud poszczególnych sygnałów, a także ponowne ich złożenie w ten sposób, aby ich lokalizacja przestrzenna była jeszcze bliższa rzeczywistości.

Poważnym problemem, a w zasadzie ograniczeniem tego sposobu separacji jest brak metod automatycznego wyznaczania optymalnych wag algorytmu i układu równań oraz – od strony czysto akustycznej – niezbyt dobry efekt działania dla nagrań w pomieszczeniach o znaczących czasach pogłosu (większych niż 0,5 sekundy). Słyszalne jest wtedy pogorszenie efektu separacji. Dekompozycja złożonego sygnału akustycznego tą metodą powinna jednak doskonale sprawdzać się w studiach radiowych, muzycznych, oraz w diagnostyce w warunkach pola swobodnego (np. komorze bezechowej).

Zagadnienie znacznie się komplikuje, jeżeli mamy do czynienia z mniejszą ilością mikrofonów niż źródeł. Zazwyczaj taka sytuacja występuje w nagraniach koncertowych, a już zupełnie skomplikowany i złożony problem powstaje w przypadku nagrań muzyki klasycznej w dużych składach, gdzie jest to realizowane za zwyczaj za pomocą dwóch lub czterech tylko par mikrofonowych (rys. 2.6). Niemożliwe wtedy staje się określenie współczynników układu równań, gdyż istnieje jego nieskończenie wiele jego rozwiązań. Wtedy, jedynym możliwym algorytmem jest algorytm separacji poszczególnych grup instrumentów odpowiadających rozmieszczeniu na podium orkiestry symfonicznej (i rzeczywiście czasami spotkać można nagrania, gdzie każdej z grup instrumentów przypisana jest para mikrofonowa). Jednak w obrębie tejże grupy problem dekompozycji dotyczy już stricte problemu dekompozycji złożonych sygnałów akustycznych monofonicznych.



Rys. 2.6. Przykład współczesnego rozmieszczenia grup instrumentów w orkiestrze smyczkowej. Oznaczenia: V1 – skrzypce I, V2 – skrzypce II, Vla – altówki, B – kontrabasy, Vc - wiolonczele

W opisanej grupie nagrań istnieje jednak jeszcze jeden przypadek szczególny. Dotyczy on mianowicie nagrań stereofonicznych, gdzie orkiestra stanowi tło i podkład muzyczny dla solisty, który jest zazwyczaj nagrywany z wykorzystaniem osobnego mikrofonu i w efekcie jego sygnał jest obecny zarówno w jednym jak i w drugim kanale w

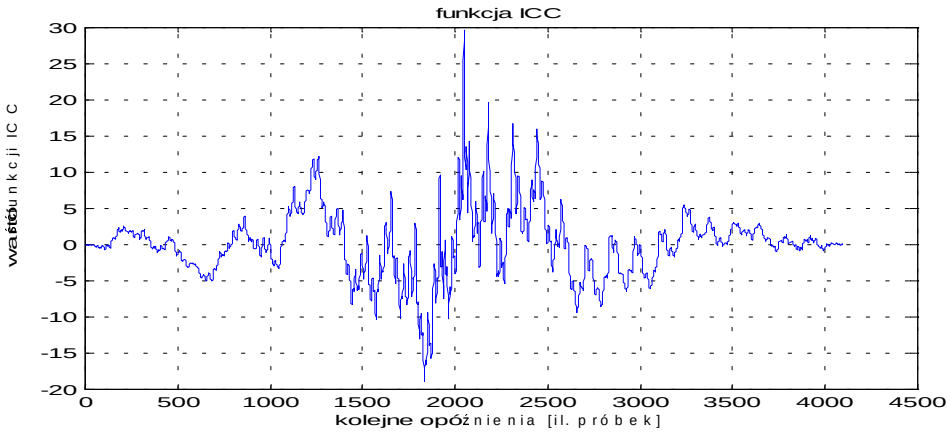
tych samych proporcjach, bez opóźnień fazowych dociera do naszych uszu, dzięki czemu jest on zazwyczaj lokalizowany pomiędzy głośnikami, czyli w środku pozornej sceny muzycznej (panoramy). Przyjmując takie założenie dotyczące sygnału solisty, zauważmy, że badając jednocześnie sygnały zarówno z jednego, jak i z drugiego kanału zarejestrowanej i odbieranej muzyki, powinny być one wysoce skorelowane dla tegoż sygnału. A jeśli tak, to możemy dokładnie to zbadać, wykorzystując do tego znaną nam już funkcję korelacji wzajemnej między dwoma kanałami wchodzącymi w skład zbioru muzycznego (tzw. korelację międzykanałową – *interchannel cross-correlation* – ICC).

Oznaczmy ciągi liczbowe kanału prawego i lewego (reprezentacja cyfrowa) jako x_n oraz y_n i wtedy nieobciążony estymator funkcji korelacji międzykanałowej ma postać

$$ICC = \frac{1}{N-m}\sum_{k=1}^{N-m} x_k\, y_{k+m} \text{ dla m=0,1,2,...,K} \tag{2.4}$$

gdzie:
K – maksymalne opóźnienie równe długości wektora danych

Obliczając w pierwszym fragmencie funkcję ICC dla segmentu danych o długości K=2048 otrzymujemy postać funkcji korelacji międzykanałowej pokazaną na rys. 2.7.



Rys. 2.7. Postać funkcji ICC dla nagrania stereofonicznego utworu z wokalistką osadzoną centralnie w panoramie muzycznej.

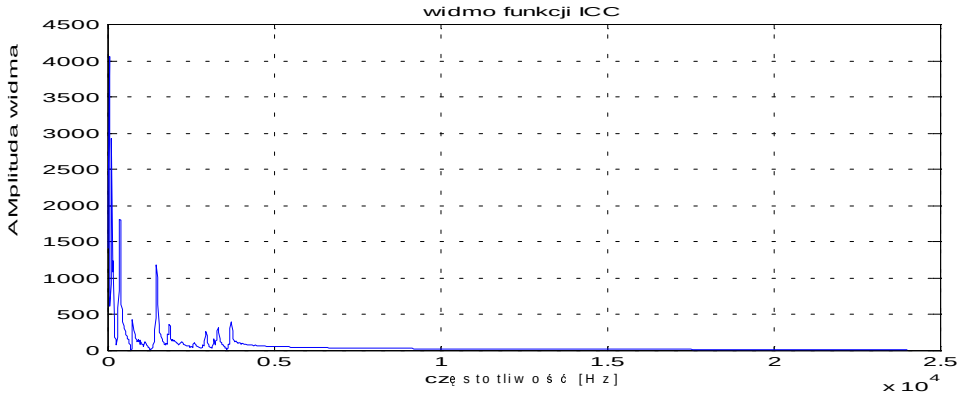
Funkcja ICC niesie ze sobą informację o korelacji pomiędzy kanałem lewym i prawym, stąd można przyjąć, że jeżeli wyznaczymy transformację Fouriera funkcji ICC, to otrzymamy maksima widma w miejscach odpowiadających widmu głosu wokalistki. Stąd

$$ICC(k) = \frac{1}{L}\sum_{n=1+L}^{2\cdot L} ICC(n)e^{-j2\pi kn/L} \tag{2.5}$$

gdzie:
L – długość analizowanego segmentu czasowego i funkcji ICC (tu: L=2048)

W wyniku tej operacji otrzymujemy widmo pokazane na rys. 2.8. Otóż okazuje się, że rzeczywiście dominującymi składowymi widma są składowe odpowiadające partii solisty, ale oprócz tego w widmie znalazły się inne niepożądane składowe nisko i wysokoczęstotliwościowe. Dlatego też zachodzi potrzeba ograniczenia pasma do zakresu widma wynikającego z rozpiętości skali głosu żeńskiego, tak, aby nie usunąć z sygnału pozostałych składowych, nie związanych z usuwanym sygnałem.

Kolejną operacją jest znormalizowanie widma względem maksymalnej wartości składowej zawartej w analizowanym widmie. W ten sposób rozpiętość amplitud widma ICC zawierać się będzie w przedziale [0,1]. Taki wektor po bezpośrednim nałożeniu na widmo sygnału spowodowałby, że uzyskalibyśmy odpowiednik splotu w dziedzinie częstotliwości sygnału z korelacją ICC. Efekt tej operacji nie jest przekonujący, dlatego wykonano nieco inną operację – nieliniowy splot z odejmowaniem widmowym.



Rys. 2.7. Widmo funkcji ICC.

Zapisano to jako

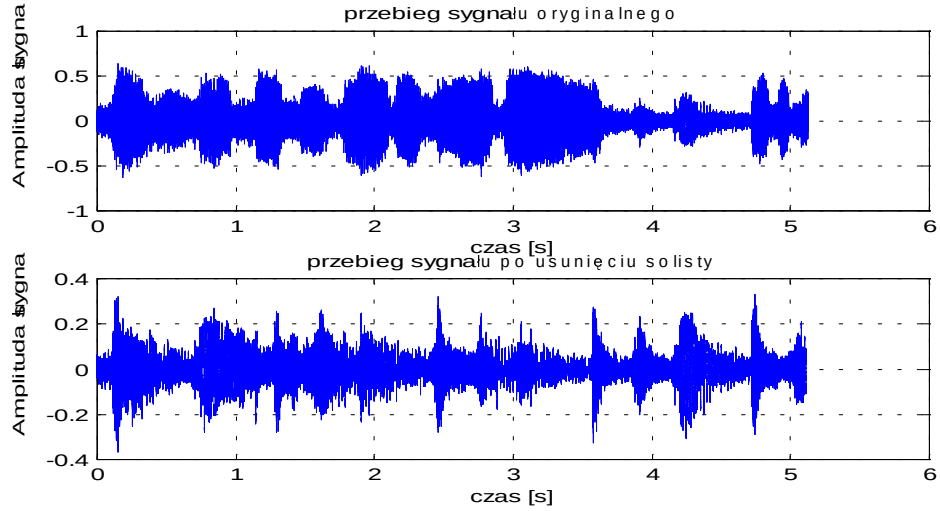
$$W(f) = F\{x(t)\} - [\log(ICC(f)_{norm})]^{\alpha(f)} \tag{2.6}$$

gdzie:
W(f) – wynikowe widmo sygnału po usunięciu składnika odpowiadającego śpiewowi wokalistki
$ICC(f)_{norm}$ – unormowane do jedynki i ograniczone widmo funkcji ICC
α (f) - wykładnik potęgi, której podstawą jest wartość widma funkcji ICC, zależny od częstotliwości

W wyniku tego operacja usuwania niepożądanego składnika z widma przebiega łagodniej, powodując mniejsze zniekształcenia oraz uwzględnia poziom składowych słabiej zaznaczonych w widmie oryginalnego przebiegu i w widmie śpiewu wokalistki.

Kolejnym krokiem jest wykonanie odwrotnej transformacji Fouriera uzyskanego widma, z uwzględnieniem fazy przebiegu oryginalnego. Jeżeli wcześniej założyliśmy, że obliczenia będziemy przeprowadzać w pętli, pobierając kolejne segmenty danych o długości L, oraz że stosujemy np. 25% nakładkowanie, to musimy także dokonać operacji miksowania uzyskanych fragmentów, tak, aby uniknąć nieciągłości na krańcach badanych przedziałów. Doświadczenie oraz nieformalne testy odsłuchowe wskazują na miksowanie próbek według wskazówek zawartych w [1].

Wynikiem tak wykonanych operacji jest sygnał bez partii solisty pokazany na rys. 2.8 w odniesieniu do sygnału oryginalnego.



Rys. 2.8. Efekt separacji solisty z nagrania stereofonicznego w odniesieniu do sygnału oryginalnego.

Wyniki uzyskane w wyniku zastosowania tego algorytmu zdecydowanie skłaniają do optymizmu, jeżeli patrzy się na wykresy. Jednak testy odsłuchowe pokazują różnicę

pomiędzy praktyką, a teoretycznymi przewidywaniami. Otóż, rzeczywiście, partia solisty została niemal całkowicie usunięta, jednak składowe sygnału znajdujące się w obrębie widma głosu partii solisty zostały też poważnie stłumione, aczkolwiek są słyszalne. Pojawiły się też zniekształcenia wynikające z efektu uśredniania widma oraz jego ostrych wahań (jeżeli w widmie pojawia się nieciągłość pierwszej pochodnej, czyli widmo nie jest funkcją gładką, pojawią się efekty dzwonienia, zniekształcenia modulacyjne i podobne do tych z jakimi mamy do czynienia w przypadku operacji odejmowania widmowego). Dlatego też postanowiono delikatnie zniekształcić uzyskane wyniki poprzez uwypuklenie pozostałości po operacji usuwania solisty w zakresie obejmującym jego rejestr na skali muzycznej, poprzez nałożenie okna Hamminga na ten właśnie fragment widma. Efekt uzyskany w wyniku takich działań jest nieco lepszy, jednak nadal słyszalne są zniekształcenia. Drugim spostrzeżeniem jest dominacja niektórych składowych widma (niskich rejestrów i perkusji). Aczkolwiek najbardziej znaczący udział w sygnale mają teraz składowe niskotonowe. Stąd propozycja, aby w tego typu operacjach wprowadzać jednak preemfazę, co uwypukli informację zawartą w pozostałych rejestrach. Ogólnie jednak należy powiedzieć, że metoda daje oczekiwany rezultat, jednak jakość będzie zależna bardzo od konkretnego materiału muzycznego, rozdzielczości (czyli długości) analizowanych fragmentów, oraz od tego jak wielki zakres nakładkowania zastosujemy. Najlepiej byłoby, gdyby istniała możliwość analizy sygnału próbka po próbce – niestety taka operacja wymagałaby niesamowitej mocy obliczeniowej i bardzo sprawnego algorytmu. Obecnie taka metoda jest opracowywana przez jednego z autorów artykułu, jak również opracowywane są metody automatycznego wyznaczania wag potrzebnych do dekompozycji sygnałów z zastosowaniem inteligentnych systemów decyzyjnych.

3. WNIOSKI

Opisane powyżej narzędzia z większym lub mniejszym powodzeniem pozwalają na separację i dekompozycję sygnału muzycznego, jednak nie uwzględniają one charakteru sygnałów składowych (wykorzystujemy tylko proste związki wynikające z metod jakimi te sygnały zostały zarejestrowane). Następnym krokiem będą już naprawdę zaawansowane metody dekompozycji sygnałów akustycznych, wykorzystujące inteligentne systemy decyzyjne w celu wyznaczenia wag i współczynników koniecznych do rozwiązania podanych w artykule równań. Pomimo jeszcze eksperymentalnego charakteru podanych metod wydaje się być pewne, iż w niedługim czasie znajdą one zastosowanie w wielu komercyjnych narzędziach do dekompozycji wielokanałowych sygnałów akustycznych zarówno dla użytkowników profesjonalnych, jak i do zastosowań domowych.

LITERATURA

[1] A. Czyżewski: *Dźwięk cyfrowy*. Akademicka Oficyna Wydawnicza EXIT Warszawa 1998

[2] A. Gołaś: *Sterowanie dźwiękiem w pomieszczeniach zamkniętych*. Wydawnictwa AGH Kraków 2000

[3] A. V. Oppenheim, R. W. Schafer: *Discrete – Time Signal Processing*. Prentice – Hall, Englewood Cliffs, New Jersey 1989

[4] B. Urbański: *Rejestracja sygnałów fonicznych*. WKŁ Warszawa 1990

[5] W. F. Druyvesteyn, J. Garas: *Personal sound*. J. Audio Eng. Soc., Vol. 45, No. 9, November 1997

[6] R. C. Maher: *Evaluationof a method for separating digitized duet signals*. J. Audio Eng. Soc., Vol. 38, No. 12, December 1990

[7] J. Meyer: *The sound of orchestra*. J. Audio Eng. Soc., Vol. 41, No. 4, April 1993

[8] A. Westner, V. M. Bove, Jr.*: Blind separation of real world audio signals using overdeermined structures*. Proc. ICA'99, January 11-15, Aussois, France
[9] A. Westner, V. M. Bove, Jr.*: Applying blind source separation and deconvolution to real-world acoustic environments*. MIT MEdia Lab, Cambridge, Massachusetts USA 1999