Maciej Bartkowiak
Adam Wereszczyński
Politechnika Poznańska
Wydział Elektroniki i Telekomunikacji
Katedra Telekomunikacji Multimedialnej i Mikroelektroniki
60-965 Poznań, ul. Polanka 3
mbartkow@multimedia.edu.pl

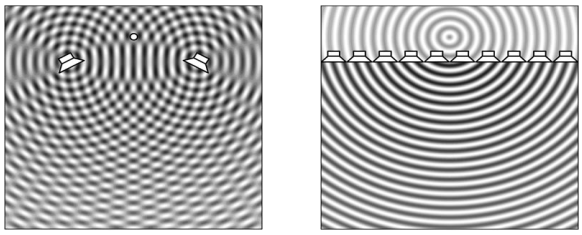
KKRRiT 2015
Łódź, 8-10 kwietnia 2015

Kodowanie wielokanałowego sygnału fonicznego na potrzeby systemów syntezy pola akustycznego

Streszczenie: **Artykuł dotyczy metody kompresji sygnałów fonicznych o wielkiej liczbie kanałów, reprezentujących pole akustyczne w systemach WFS. Idea kompresji polega na wydzieleniu sygnałów reprezentujących indywidualne źródła przestrzenne. Sygnały te, wraz z towarzyszącymi im składowymi tła (rezyduum) kodowane są techniką perceptualną z większą efektywnością niż indywidualne sygnały kanałów.**

1. WSTĘP

Systemy syntezy pola akustycznego, WFS (ang. *wave field synthesis*) pozwalają na realistyczną reprodukcję dźwięku przestrzennego z punktową lokalizacją źródeł dźwięku w pomieszczeniu, co nie jest możliwe w systemach dyskretnej stereofonii [1,2]. Osiągane jest to poprzez wykorzystanie reguły Huygensa oraz dużej liczby (kilkadziesiąt do kilkuset) indywidualnych przetworników głośnikowych odtwarzających falę akustyczną gęsto spróbkowaną w przestrzeni (rys. 1).



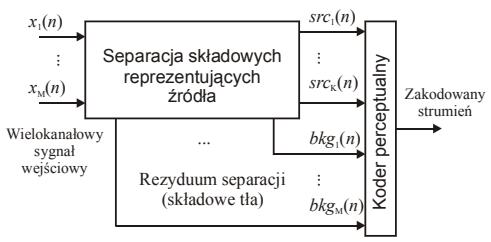
Rys. 1 Pole ciśnienia akustycznego wytwarzane przez parę głośników oraz system WFS.

Przyszłościowe systemy multimedialne związane z telewizją swobodnego punktu widzenia oraz kinem 3D będą dążyły do wykorzystania metody WFS w celu uzyskania pełnego realizmu reprodukcji [3], dlatego dziś prowadzi się badania nad techniczną realizowalnością takich rozwiązań (rys. 2). Koszt eksploatacyjny systemu syntezy pola wynikający z dużej ilości danych (ze względu na liczbę kanałów) może być zredukowany przez stosowanie odpowiednich technik kompresji. Warto zauważyć, że gęste próbkowanie przestrzenne (np. przy pomocy macierzy mikrofonów) powoduje, że sąsiednie kanały reprezentujące zazwyczaj punktowe źródła dźwięku przekazują sygnały o bardzo podobnej treści. Te podobieństwa mogą zostać wykorzystane lokalnie poprzez kodowanie różnicowe (M/S) sąsiadujących par kanałów, ale w przypadku kanałów odległych w przestrzeni taki zabieg jest bezskuteczny ze względu na przesunięcie fazowe i efekty tzw. filtracji grzebieniowej. Z tego powodu obecnie stosowane metody kodowania dźwięku przestrzennego (MPEG-2 NBC [4], MPEG-D SAOC [5]) nie mogą być bezpośrednio wykorzystane do kompresji danych WFS. Również najnowsza norma MPEG-H 3D Audio [6], pomimo tego, że dotyczy systemów wielokanałowych, nie oferuje narzędzia kompresji nadającego się do systemów WFS, ze względu na parametryczną metodę kodowania, która wprowadza artefakty uniemożliwiające prawidłowe odtwarzanie pola akustycznego.

Rys. 2 Eksperymentalny system syntezy pola akustycznego skonstruowany na Politechnice Poznańskiej

2. PROPONOWANA TECHNIKA KOMPRESJI

Ogólna idea proponowanej techniki (rys. 3) polega na wyodrębnieniu (separacji) K<<M składowych *src* reprezentujących indywidualne źródła dźwięku zarejestrowane przez zbiór mikrofonów z wejściowego sygnału o M kanałach. Składowe te, jak również zbiór M rezydualnych składowych *bkg* „tła" akustycznego są następnie kodowane tradycyjnym wielokanałowym koderem dźwięku z uwzględnieniem wspólnego modelu perceptualnego. Po stronie dekodera, odtworzone składowe źródeł są sumowane ze składowymi tła na podstawie zbioru współczynników (wag) obliczonych w koderze.



Rys. 3 Ogólna zasada proponowanego rozwiązania

Wyodrębniając wspólne składowe źródeł przestrzennych unika się wielokrotnej reprezentacji tych składowych w sygnale wejściowym. Całkowity strumień bitowy potrzebny do ich zakodowania jest niewielki, gdy

liczba źródeł jest znacznie mniejsza niż liczba kanałów, M. Z drugiej strony, zredukowana energia składowych tła powoduje znaczące zmniejszenie strumienia danych przypadającego na te składowe. Łączny strumień dla K źródeł i M kanałów jest zredukowany pod warunkiem skutecznej separacji. Idealna separacja w ogólności nie jest możliwa, jednak dzięki wykorzystaniu gęstego próbkowania przestrzennego możliwa jest dobra jej aproksymacja.

Większość klasycznych metod separacji sygnałów [7] dotyczy tzw. sumowania natychmiastowego, nie uwzględniającego opóźnień, które mogą być zmienne w czasie (poruszające się źródła dźwięku). Również jakość sygnałów odtworzonych (poziom zniekształceń) nie pozwala na zastosowanie ich w dziedzinie wymagającej wysokiej wierności odtwarzania. Kluczowym elementem techniki proponowanej w tym artykule jest autorska metoda separacji wykorzystująca estymację i kompensację opóźnień międzykanałowych dla poszczególnych składowych źródeł.

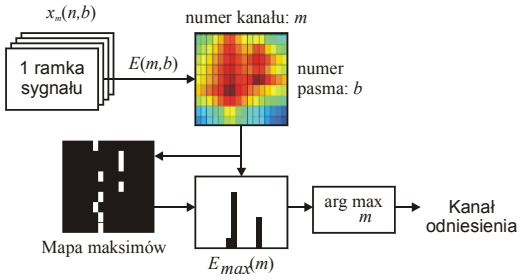
Składowe sygnału wejściowego są analizowane w trójwymiarowej dziedzinie przestrzenno-czasowo-częstotliwościowej, która pozwala traktować obecność składowych widmowych źródeł jako elementy rzadkie. W tym celu sygnał wejściowy jest poddawany dekompozycji na B podpasm za pomocą wejściowego zespołu filtrów o strukturze kaskadowej, przy czym szerokości podpasm zwiększają się z częstotliwością i nie jest zastosowane podpróbkowanie. W każdym kanale i podpaśmie następuje też podział na ramki czasowe z 50% zakładką, oraz okienkowanie, w celu uniknięcia artefaktów wynikających z nieciągłości. Długość ramek (2048 próbek, co odpowiada 23ms przy częstotliwości próbkowania 44.1kHz) jest kompromisem pomiędzy nadmiarowością i granularnością reprezentacji ruchu źródeł.

2.1. Struktura kodera

Koder sygnału (rys. 4) dokonuje iteracyjnej analizy przestrzenno-częstotliwościowej dla każdej ramki sygnału, identyfikując w każdej iteracji najsilniejszą składową wspólną, której obecność w poszczególnych kanałach może być obserwowana z różnym opóźnieniem. Opóźnienia te są następnie estymowane, przy czym jako odniesienie przyjmuje się kanał, w którym energia danej składowej jest największa. W kolejnym kroku, opóźnienia są kompensowane (ramki w poszczególnych kanałach są odpowiednio przesuwane w czasie), dzięki czemu następuje wyrównanie czasowe dla bieżącej składowej. Procedura analizy składowych głównych pozwala na wyodrębnienie wspólnej dominującej składowej ze skompensowanych kanałów, przy czym ta składowa jest odejmowana od kanałów przed przywróceniem im pierwotnej pozycji w czasie. Wyodrębniona składowa, po złożeniu ramek i podpasm jest przekazywana do kodera perceptualnego, a cykl analizy powtarza się tak długo, jak długo całkowita energia pozostałych po wyodrębnieniu kanałów rezyduum ulega istotnej redukcji.

2.2. Wyznaczanie kanału odniesienie

W każdej iteracji algorytmu separacji, kanał odniesienia powinien reprezentować mikrofon, który znajduje się najbliżej aktualnie analizowanego źródła, oferując dla tego źródła najlepszą wartość SNR (przy czym za szum uważa się tutaj składowe pozostałych źródeł, które mogą zakłócać proces estymacji opóźnień). W celu jego identyfikacji, wyznacza się mapę energii w dziedzinie przestrzeń-częstotliwość (rys. 5).



Rys. 5 Wyznaczanie kanału odniesienia

Binarna mapa maksimów zawiera tylko jedną wartość niezerową dla każdego kanału. Funkcja $E_{max}(m)$ zawiera sumaryczną energię w każdym kanale wyznaczoną tylko dla składowych określonych mapą maksimów.

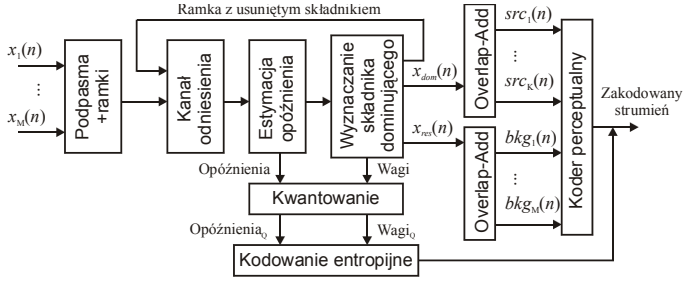
2.3. Estymacja opóźnień

Wyznaczenie przesunięcia w czasie między parą sygnałów zakłóconych obecnością innych sygnałów (o innym opóźnieniu) często wymaga zastosowania metod widmowych. W ogólności, przesunięcie w czasie skutkuje zmianą argumentu transformaty Fouriera,

$$F\{x_s(n-\Delta n)\} = X_s(k)\exp(-j2\pi k\,\Delta n). \qquad (1)$$

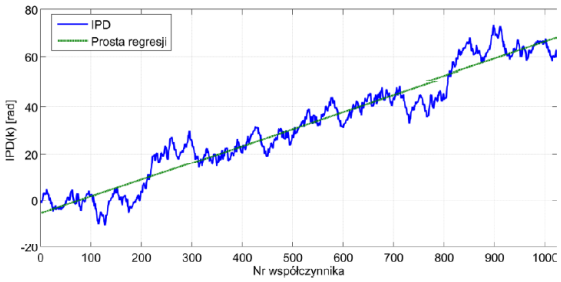
Analiza przesunięcia fazy między zakłóconymi sygnałami obarczona jest jednak dużym błędem losowym (rys. 6). W zaproponowanym autorskim podejściu przesunięcie identyfikuje się w dziedzinie czasu na podstawie najsilniejszego maksimum odwrotnej transformaty Fouriera z funkcji IPD (międzykanałowa różnica fazy). Obliczenie to można również zinterpretować jako wyznaczanie maksimum funkcji korelacji wzajemnej dla sygnałów wybielonych widmowo,

$$x_{imp}(m,n) = F^{-1}\{\exp(-j\,IPD_s(m,k))\}, \qquad (2)$$



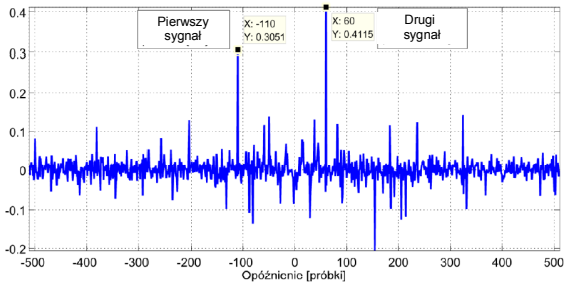
Rys. 4 Schemat blokowy kodera

gdzie $$IPD_s(m,k) = \arg\left\{X_s(k)\ X_m^*(k)\right\}. \quad (3)$$



Rys. 6 Wyznaczanie opóźnienia na podstawie trendu liniowego z funkcji różnicy fazy, IPD

Dzięki wybieleniu widmowemu, sygnał x_{imp} ma przebieg impulsowy, a lokalizacja jego maksimow jest bardzo precyzyjna (rys. 7). Zaletą tej metody jest fakt, że składowe zakłócające nie zmieniają położenia najsilniejszego impulsu w czasie, lecz generują inne mniejsze impulsy, dzięki czemu przy prawidłowej identyfikacji głównego maksimum, błąd estymacji jest niewielki.



Rys. 7 Wyznaczanie opóźnienia na podstawie maksimum sygnału impulsowego w dziedzinie czasu

2.4. Wyznaczanie składowych źródeł oraz tła

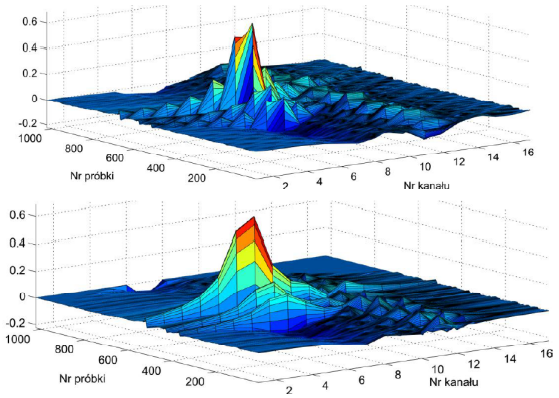
Po dokonaniu estymacji opóźnień dla wszystkich kanałów sygnału w bieżącej ramce, następuje przesunięcie kanałów w czasie, które powoduje wyrównanie położenia składowej dominującej (rys. 8). W tak skorygowa nych sygnałach identyfikowane są następnie składowe wspólne metodą PCA [8]. W tym celu, dla każdego podpasma tworzony jest wektor próbek $V(n)=[x_1(n), x_2(n), \ldots x_M(n)]^T$ oraz obliczana jest macierz autokowariancji $\mathbf{R}_{VV}=\underline{V}\ \underline{V}^T/M$. Macierz optymalnego przekształcenia $\mathbf{C}$ konstruowana jest w taki sposób, że każdy wiersz $\mathbf{C}$ jest rozwiązaniem równania

$$\left[\underline{C}_k - \lambda_k \mathbf{I}\right]\underline{X} = 0, \quad (4)$$

gdzie λ_k jest k-tą wartością własną $\mathbf{R}$, $\mathbf{I}$ jest macierzą jednostkową, a $\underline{X} = [1, 1, \ldots 1]^T$. Pierwszy wektor własny $\underline{C}_1$ skojarzony z największą wartością własną λ_{max} pozwala na rzutowanie danych wszystkich kanałów na w przestrzeni M-wymiarowej na oś o maksymalnej wariancji,

$$x_{dom}(n) = \underline{C}_1\ \underline{V}(n), \quad (5)$$

co skutkuje wyznaczeniem jednej ramki składowej dominującej. Złożenie wszystkich ramek x_{dom} daje w wyniku składowe indywidualnych źródeł src_k, poddawane następnie kompresji perceptualnej.



Rys. 8 Kompensacja opóźnień między kanałami

Wektor składowych rezydualnych tworzony jest przez różnicę

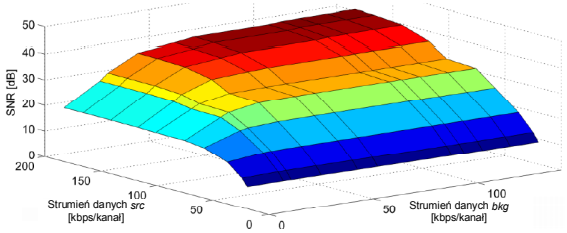
$$\underline{V}_{res}(n) = \underline{V}(n) - \underline{D}_1 x_{dom}(n), \quad (6)$$

gdzie $\underline{D}_1$ oznacza pierwszą kolumnę macierzy odwrotnej, $\mathbf{D} = \mathbf{C}^{-1}$. Składowe rezydualne, po odwróceniu kompensacji opóźnienia, stanowią dane wejściowe kolejnej iteracji kodera. W końcowej iteracji, składowe te, po złożeniu ramek, tworzą M sygnałów bkg_m, również kierowanych do kompresji perceptualnej.

2.5. Kodowanie perceptualne

Do łącznego zakodowania zestawu K+M składowych sygnału wielokanałowego może być wykorzystany dowolny koder perceptualny, pod warunkiem odpowiednio małej stratności kompresji, gwarantującej transparentną rekonstrukcję. Koder taki wymaga jednak zastosowania odpowiednio zmodyfikowanego modelu psychoakustycznego, który uwzględnia różne sytuacje sumowania składowych po stronie dekodera. Aby uniknąć zjawiska *demaskowania* zniekształceń kompresji, progi maskowania muszą być obliczone wg. tzw. najgorszego scenariusza maskowania dla sum i różnic kanałów.

Dodatkowy problem konfiguracyjny stanowi właściwe ustalenie proporcji pomiędzy docelową prędkością bitową dla składowych *src* oraz *bkg*. Przeprowadzone eksperymenty wskazują, że jakość zrekonstruowanego sygnału zależy przede wszystkim od jakości składowych źródeł (rys. 9).

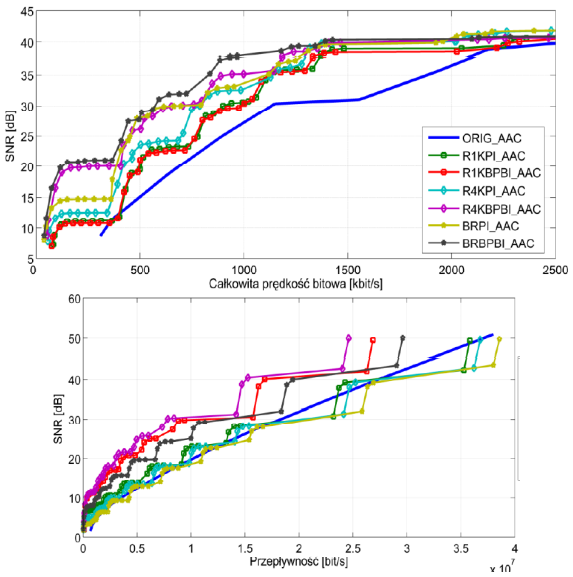


Rys. 9. Jakość zdekodowanego sygnału w funkcji prędkości bitowej dla składowych src i bkg

Aby umożliwić prawidłową rekonstrukcję sygnałów po stronie dekodera potrzebne jest przesłanie współczynników $\underline{D}_1$, które stanowią wagi sumowania w poszczególnych ramkach, pasmach i kanałach. Dane, te po zakodowaniu różnicowym są źródłem strumienia o wielkości nie przekraczającej 1kb/s.

3. WYNIKI EKSPERYMENTALNE

Przedstawiona technika kodowania została zaimplementowana w systemie Matlab i poddana badaniom eksperymentalnym, uwzględniającym różne scenariusze danych (statyczne i ruchome pojedyncze i wielokrotne źródła dźwięku) oraz szeroki zakres prędkości transmisji. Ze względu na dużą liczbę indywidualnych eksperymentów, tylko niektóre wyniki mogły zostać poddane ocenie odsłuchowej. We wszystkich przypadkach wyznaczono jakość rekonstrukcji sygnału w postaci błędu średniokwadratowego, porównując ją do jakości uzyskanej przy tradycyjnym niezależnym kodowaniu kanałów techniką MPEG AAC (rys. 10).



Rys. 10. Porównanie jakości rekonstrukcji sygnału dla różnych wariantów kodowania z jakością oferowaną przez koder MPEG AAC (krzywa ORIG_AAC). Górny wykres: dwa nieruchome źródła, dolny wykres: 2 ruchome źródła, liczba kanałów: 16

Charakterystyczny „schodkowy" kształt krzywych jakości obiektywnej (SNR w funkcji prędkości bitowej) na prezentowanych wykresach wynika ze stopniowej zmiany części strumienia przypadającej na składowe *bkg*. W przypadku niewielkiej liczby równocześnie aktywnych nieruchomych i ruchomych źródeł dźwięku otrzymane wyniki wskazują na znaczącą przewagę efektywności kompresji zaproponowanej techniki kodowania. W przypadku, gdy liczba źródeł dźwięku rośnie – spada dokładność separacji, a co za tym idzie, również zysk efektywności staje się marginalny. W każdym przypadku zysk efektywności w znaczącym stopniu zależy od relacji liczby źródeł do liczby kanałów sygnału.

W tabeli 1 pokazano podsumowanie wyniku testów odsłuchowych dla 16 kanałów sygnału oryginalnego (zarejestrowanego macierzą 16 mikrofonów). Wyniki te zasadniczo potwierdzają rezultaty oceny obiektywnej.

Tab. 1. Prędkości transmisji [kb/s] wymagane dla zapewnienia dobrej jakości zrekonstruowanego sygnału (powyżej 80 punktów w skali MUSHRA)

Liczba źródeł	*Nowa technika*	*AAC*
1 nieruchome	130-202	1536-1920
1 ruchome	1319-3790	2480-3592
2 nieruchome	673-2637	1929-2986
6 nieruchomych	4223-8700	4722-11740

Zaprezentowane podejście umożliwiające łączne kodowanie wielokanałowego sygnału fonicznego dla systemu WFS wykorzystuje separację wspólnych składowych sygnałów. Z pewnością uzyskany zysk efektywności kompresji (w porównaniu do bezpośredniego kodowania kanałów) zależy od dokładności metody separacji. Istnieje potencjalna możliwość zwiększenia zysku kompresji przez zastosowanie bardziej złożonego modelu formowania sygnału, który oprócz opóźnienia uwzględnia międzykanałowe odpowiedzi impulsowe. Również nieliniowe metody separacji mogą zaoferować tutaj dokładniejszą reprezentację sygnałów źródeł.

4. PODZIĘKOWANIE

Praca finansowana ze środków przyznanych przez Ministerstwo Nauki i Szkolnictwa Wyższego na działalność statutową w roku 2015 w ramach projektu 08/84/DSPB/0160.

BIBLIOGRAFIA

[1] E. Verheijen, *Sound Reproduction by Wave Field Synthesis*, rozprawa doktorska, Technische Universitaet Delft, 2010

[2] S. Spors, H. Teutsch, R. Rabensteing, *High-Quality Acoustic Rendering with Wave Field Synthesis*, Vision, Modeling, and Visualization Conf. (VMV 2002), str. 101-108, Erlangen, 2002

[3] M. Rébillat, B. Katz, E. Corteel, *SMART-I 2: Spatial multi-user audio-visual real-time interactive interface*, *A broadcast application context*, 3DTV Conference: The True Vision-Capture, Transmission and Display of 3D Video, str. 1-4, IEEE, 2009

[4] ISO/IEC JTC1/SC29/WG11 MPEG, *International Standard ISO/IEC 13818-3, Generic Coding of Moving Pictures and Associated Audio: Audio*, ISO, 1997

[5] ISO/IEC MPEG, *International Standard ISO/IEC 23003-2: Information Technology – Part 2: MPEG Spatial Audio Object Coding*. Int. Org. Stand., 2010

[6] ISO/IEC MPEG, *International Standard ISO/IEC 23008-3: Information technology — High efficiency coding and media delivery in heterogeneous environments — Part 3: 3D audio*. ISO, 2014

[7] P. Comon , Ch. Jutten (red.), *Handbook of Blind Source Separation: Independent Component Analysis and Applications*, Academic Press, 2010

[8] I.T. Jolliffe, *Principal Component Analysis (Springer Series in Statistics)*. Springer, 2002