BIULETYN INSTYTUTU AUTOMATYKI I ROBOTYKI
NR 31, 2011

Rekonstrukcja zdegradowanych fragmentów nagrań dźwiękowych

Leszek GRAD

Instytut Teleinformatyki i Automatyki, WAT,
ul. Gen. S. Kaliskiego 2, 00-908 Warszawa

STRESZCZENIE: W artykule zostało przedstawione zagadnienie rekonstrukcji zdegradowanych przedziałów w nagraniach dźwiękowych. Przedstawiono krótki przegląd stosowanych metod oraz wyniki badań zastosowania modeli autoregresji AR oraz nieliniowego modelu predykcji (z wykorzystaniem sieci neuronowej). Zbadano także zastosowanie analizy pasmowej w rekonstrukcji utraconych próbek.

SŁOWA KLUCZOWE: rekonstrukcja nagrań, sieci neuronowe, restauracja nagrań dźwiękowych.

Wstęp

W artykule przedstawione zostało zagadnienie rekonstrukcji zdegradowanych przedziałów w nagraniach dźwiękowych[1]. Jest to część szerszego zagadnienia usuwania zakłóceń impulsowych [8], których usuwanie na drodze filtracji nie przynosi zadowalających rezultatów. Badacze starają się różnymi metodami dokonywać rekonstrukcji możliwie najdłuższych zdegradowanych przedziałów.

Często stosowanym podejściem do zagadnienia rekonstrukcji sygnału jest wykorzystanie modelu autoregresji[2] (AR) [1], [2], [5], [8], [9] oraz interpolacja wielomianami[3] [5]. Rekonstrukcja jest realizowana na drodze dwóch ekstrapolacji:

[1] Ten sam problem występuje w przypadku utraty pakietów przy przesyłaniu dźwięku przez sieć komputerową.
[2] Model AR pozwala on na rekonstrukcję przedziałów do 20 ms (ok. 1000 próbek przy częstotliwości próbkowania 44,1 kHz) [9].
[3] Interpolacja wielomianami (niskich rzędów) może być stosowana do rekonstrukcji jedynie krótkich przedziałów.

„w przód" oraz „w tył". Ostateczny wynik rekonstrukcji jest wypadkową prognoz z obydwu końców przedziału. W celu wyznaczenia parametrów modelu AR należy dysponować liczbą próbek co najmniej kilkakrotnie większą od długości prognozy (długości rekonstruowanego przedziału) [2].

W literaturze można znaleźć algorytmy, które bazując na modelu AR pozwalają na rekonstrukcję dość długich przedziałów. W pracy [2] autorzy zastosowali model AR w podpasmach, wykorzystując bank filtrów. Wykazali, że analiza w tym przypadku wymaga zastosowania modeli AR niższych rzędów. Z kolei w artykule [9] zastosowano wielokanałowy model AR, uwzględniający sygnały z innych kanałów przesunięte w czasie. Otrzymane wyniki były zadowalające z uwagi na silną korelację pomiędzy kanałami. Uzyskano dobrej jakości rekonstrukcję odcinka nagrania o długości 240 ms. W artykule [13] przedstawiono rozwiązanie oparte na modelu regresji wykorzystującym analizę czasowo-częstotliwościową (filtracja Gabora).

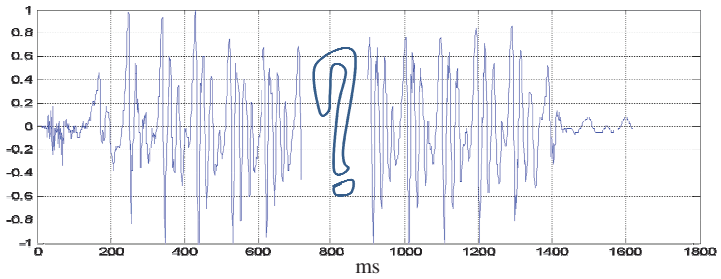
Oprócz wyżej wymienionych technik do rekonstrukcji utraconych próbek stosowane są sztuczne sieci neuronowe [11], [14]. Za ich pomocą można dokonywać ekstrapolacji na drodze predykcji nieliniowej [4], [5], [13]. W pracach [4] i [11] pokazano zastosowanie neuronowej sieci nieliniowej do realizacji predykcji sygnałów w podpasmach (wykorzystano bank filtrów liniowych). W pierwszej z nich wykorzystano sieć z sigmoidalną funkcją aktywacji (z wyjściową warstwą liniową). Otrzymano zadowalające wyniki interpolacji przedziałów do 113 ms (5000 próbek). W drugiej zastosowano sieć z neuronami radialnymi. Przedstawiono zadowalające wyniki dla rekonstruowanych przedziałów o długości 1000 próbek. Ciekawy algorytm rekonstrukcji długich przedziałów zaprezentowano w [10]. Do określania samopodobieństwa w sygnale audio wykorzystano podejście znane z analizy tekstur w obrazach. Idea metody sprowadza się do zastępowania brakującego fragmentu najlepiej dopasowanym niezdegradowanym fragmentem pochodzącym z tego samego nagrania. Uzyskano dobre wyniki dla rekonstrukcji przedziałów o długości rzędu 1s.

Celem przestawionych w niniejszym artykule badań jest weryfikacja przydatności oraz porównanie metod: opartej na modelu AR oraz nieliniowej sieci neuronowej. Zbadano także wpływ podziału sygnału na sygnały podpasmowe na jakość rekonstrukcji.

1. Sformułowanie problemu

Na rys. 1 obrazowo przedstawiono stojące do rozwiązania zadanie. Należy, w sposób automatyczny, uzupełnić brakujący fragment nagrania próbkami sygnału, tak aby nagranie brzmiało poprawnie (w subiektywnym

odbiorze przez człowieka nie było odczuwalne zniekształcenie sygnału). W przypadku degradacji długich przedziałów zadanie jest możliwe do zrealizowania, jeżeli podobny do odtwarzanego fragment nagrania znajduje się nadal w innym miejscu (refren piosenki, akord). Większość z przedstawionych we wstępie metod rozwiązuje ten problem z wykorzystaniem ekstrapolacji dla obu stron brakującego przedziału. Takie podejście zastosowane zostanie także w tym przypadku.



Rys. 1. Nagranie z usuniętym zdegradowanym przedziałem próbek

Do oceny metod wykorzystywane są: stosunek sygnału do szumu SNR [9], błąd średniokwadratowy MSE [4] oraz miara subiektywna PAQM[4] [2][3]. Obliczenie SNR i MSE jest możliwe w przypadku symulowania degradacji (dysponujemy niezniekształconym nagraniem w tym przedziale).

2. Rozwiązanie zadania rekonstrukcji z wykorzystaniem liniowej sieci neuronowej

Niech dany będzie ciąg próbek:

$$u = \left(u(1), u(2), (3), \ldots, u(N)\right) \tag{1}$$

Model AR[5] wykorzystuje korelację próbek sygnału i jest definiowany jako:

$$u(k) = \sum_{i=1}^{p} w(i)u(k-i) + e(k) \tag{2}$$

[4] Ang. *Perceptual Audio Quality Measure*.
[5] Model ten znany jest także pod nazwą LPC (ang. *Linear Predictive Coding*).

gdzie: p – rząd modelu (predykcji), $w(i)$ – współczynniki modelu, $e(k)$ – szum (błąd predykcji).

Wyznaczenie współczynników modelu $w(i)$ dokonuje się na drodze minimalizacji błędu predykcji $e(k)$. Przy założeniu, że sygnał (1) stanowi bezpośrednie lewostronne sąsiedztwo rekonstruowanego przedziału, model (2) może być wykorzystany do ekstrapolowania (odtwarzania brakujących próbek), jednakże kolejne punkty obarczone będą coraz większym błędem, gdyż na wejście układu podawane będą już ekstrapolowane wartości. Rozwiązaniem tego problemu jest zastosowanie wielu modeli, z których każdy przygotowywany jest do wyznaczenia prognozy na jedną, konkretną chwilę czasową. Prowadzi to do zastosowania tylu układów, ile próbek liczy zdegradowany, odtwarzany przedział. Oznaczając przez L liczbę próbek w odtwarzanym przedziale, zadanie to można zdefiniować następująco:

$$\sum_{k=1}^{N-L-p+1} \left\| \mathbf{u}_k' \mathbf{W} - \mathbf{z}_k \right\|^2 \underset{\mathbf{W}}{\rightarrow} min \tag{3}$$

$$\mathbf{U} = \begin{bmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \dots & \mathbf{u}_{N-L-p+1} \end{bmatrix} = \begin{bmatrix} u(1) & u(2) & \dots & u(N-L-p+1) \\ u(2) & u(3) & \dots & u(N-L-p+2) \\ \dots & \dots & \dots & \dots \\ u(p) & u(p+1) & \dots & u(N-L) \end{bmatrix} \tag{4}$$

$$\mathbf{Z} = \begin{bmatrix} \mathbf{z}_1 & \mathbf{z}_2 & \dots & \mathbf{z}_{N-L-p+1} \end{bmatrix} = \begin{bmatrix} u(p+1) & u(p+2) & \dots & u(N-L+1) \\ u(p+2) & u(p+3) & \dots & u(N-L+2) \\ \dots & \dots & \dots & \dots \\ u(p+L) & u(p+L+1) & \dots & u(N) \end{bmatrix} \tag{5}$$

gdzie: $\mathbf{u}_k$ – kolumna macierzy $\mathbf{U}$, $\mathbf{z}_k$ – kolumna macierzy $\mathbf{Z}$, $\mathbf{W}$ – macierz współczynników modeli, współczynniki jednego modelu stanowią kolumnę macierzy $\mathbf{W}$, wymiar macierzy: $p \times L$.

Proces rekonstrukcji jest realizowany na drodze rozwiązania zadania ekstrapolacji w przód, gdzie prognoza jest wyznaczana na podstawie próbek znajdujących się przed rekonstruowanym przedziałem oraz ekstrapolacji w tył, gdzie wykorzystuje się próbki leżące za zdegradowanym przedziałem (analiza wstecz). Wynik rekonstrukcji osiągany jest jako kombinacja wypukła wyników

obu ekstrapolacji:

$$y(i) = a(i)y_p(i) + (1 - a(i))y_t(i),\ i = 1,2,\ldots,L \tag{6}$$

gdzie:

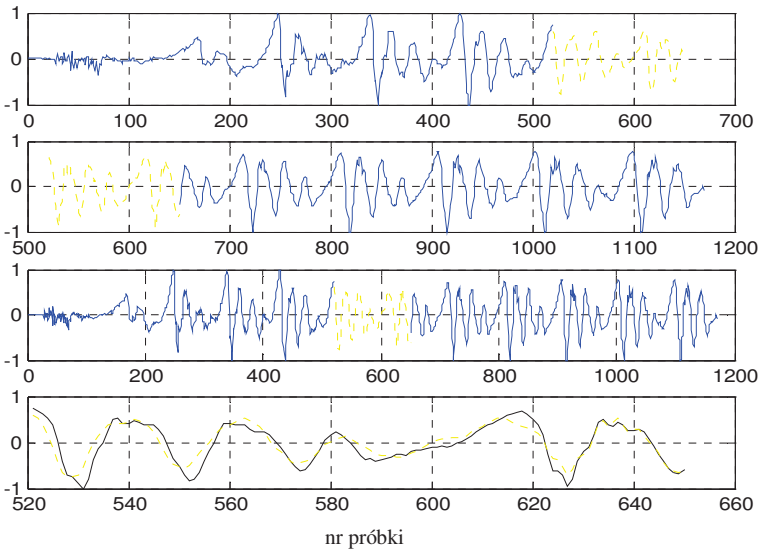
L – szerokość rekonstruowanego przedziału;

$y_p(i)$ – wynik ekstrapolacji w przód;

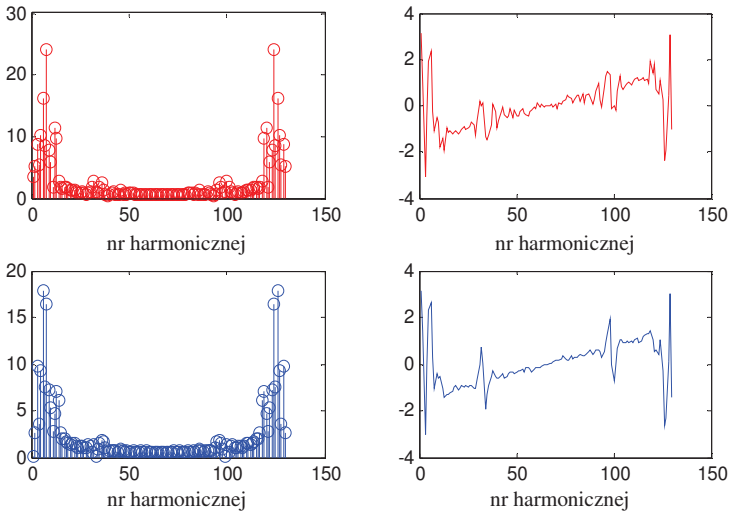
$y_t(i)$ – wynik ekstrapolacji w tył;

$a(i)$ – współczynniki kombinacji, $0 \le a(i) \le 1$.

Na rys. 2 przedstawiony został wynik rekonstrukcji sygnału. Wynik ekstrapolacji w przód przedstawiono na wykresie pierwszym od góry (prognozę oznaczono linią przerywaną), wynik ekstrapolacji w tył przestawiony został na wykresie drugim od góry (prognozę oznaczono linią przerywaną). Złożenie obu prognoz i ostateczny wynik odtworzenia sygnału przedstawiono na wykresie trzecim od góry (ekstrapolowany przedział oznaczono linią przerywaną). Na wykresie pierwszym od dołu przedstawiony został, na tle sygnału niezdegradowanego, wynik interpolacji (linia przerywana). Przedział, w którym dokonywana była interpolacja, liczył $L = 130$ próbek, rząd predykcji $p = 120$. Błąd średniokwadratowy wyniósł $MSE = 0{,}0303$, co wydaje się dobrym rezultatem.



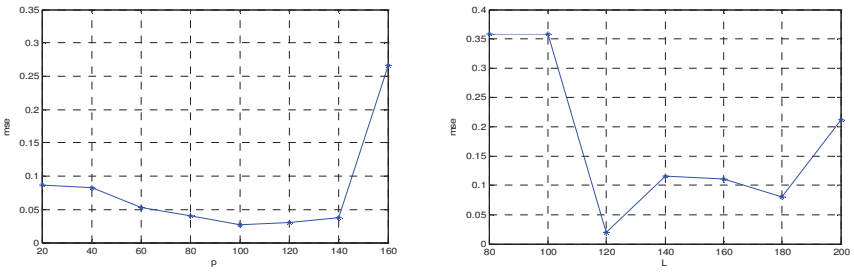
Rys. 2. Rekonstrukcja z wykorzystaniem modeli AR, opis w tekście



Rys. 3. Widmo amplitudowe i fazowe przebiegu oryginalnego (wykresy dolne) i zrekonstruowanego (wykresy górne). Analizowany zakres częstotliwości 0-5,5 kHz

W celu pełniejszej oceny jakości interpolacji przedstawione zostały widma sygnału oryginalnego i wyniku interpolacji (rys. 3). Widma nie różnią się istotnie, pozytywnym jest niewprowadzanie do wyniku interpolacji wyższych harmonicznych (niewystępujących w oryginale).

Z wykorzystaniem modeli AR przeprowadzone zostały badania pozwalające dobrać rząd predykcji w interpolacji sygnału dźwiękowego (rys. 4, wykres lewy). Wynika z nich, że dobre rezultaty osiąga się dla rzędów od $p = 60$ do $p = 140$. Badanie mające określić maksymalną długość interpolowanego przedziału wykazało, że dla przedziałów powyżej 180 próbek błąd średniokwadratowy dość szybko rośnie (rys. 4, wykres prawy).

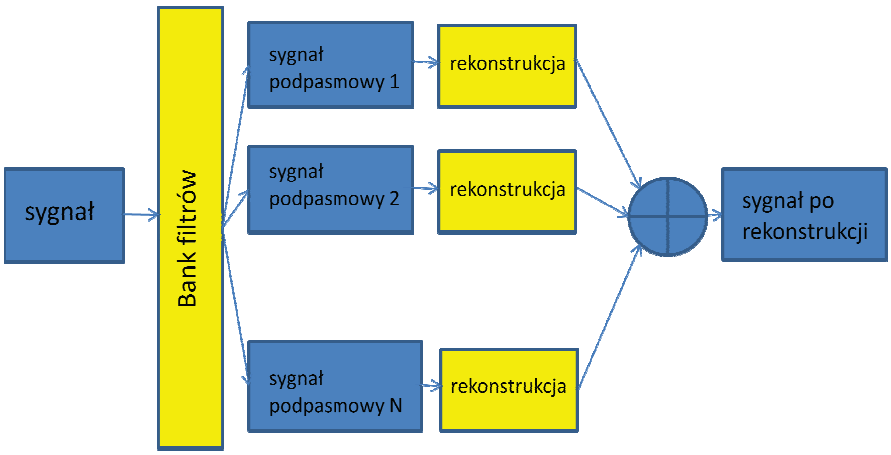


Rys. 4. Wykres lewy – wpływ rzędu modelu AR na jakość rekonstrukcji. Wykres prawy – wpływ długości rekonstruowanego przedziału na jakość rekonstrukcji. Wskaźnik jakości – błąd średniokwadratowy

3. Rekonstrukcja sygnału z zastosowaniem analizy pasmowej

W tej części eksperymentu zastosowano modele AR do rekonstrukcji sygnału uprzednio poddanego analizie pasmowej. Zastosowano banki filtrów o różnej liczbie pasm. Na rys. 5 przedstawiono schemat eksperymentu, a na rys. 6 przebiegi czasowe sygnałów podpasmowych przy podziale na 5 pasm częstotliwości.

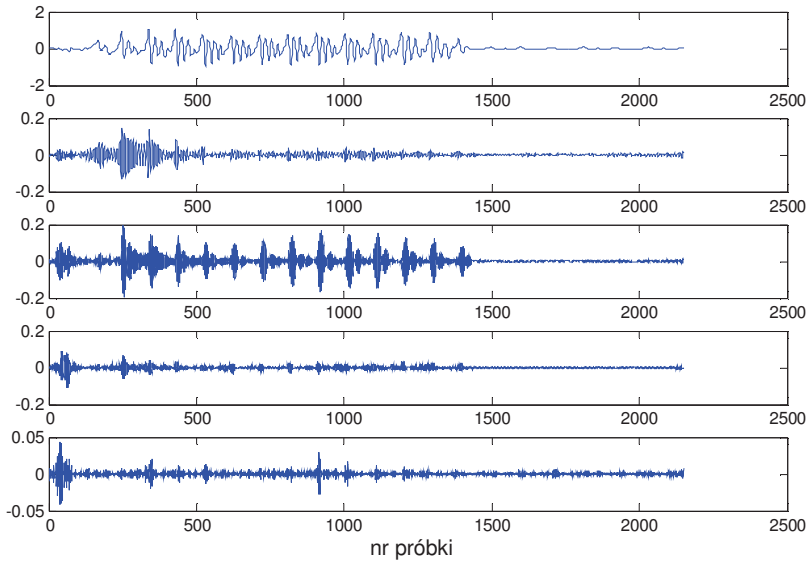
Wyniki rekonstrukcji w pasmach częstotliwości przedstawiono na rys. 7 oraz w tab. 1. Na rys. 7 przedstawiono przebiegi wyniku rekonstrukcji na tle sygnału oryginalnego, przy podziale na 5 pasm częstotliwości. Można zauważyć, że model AR słabo radzi sobie z rekonstrukcją w wyższych pasmach częstotliwości. W tab. 1 zamieszczono ocenę rekonstrukcji (wartości błędu średniokwadratowego) przy podziale na różną liczbę pasm. Wyniki nie wskazują na to, aby przeprowadzenie rekonstrukcji sygnału oddzielnie dla wydzielonych pasm częstotliwości poprawiło wynik.



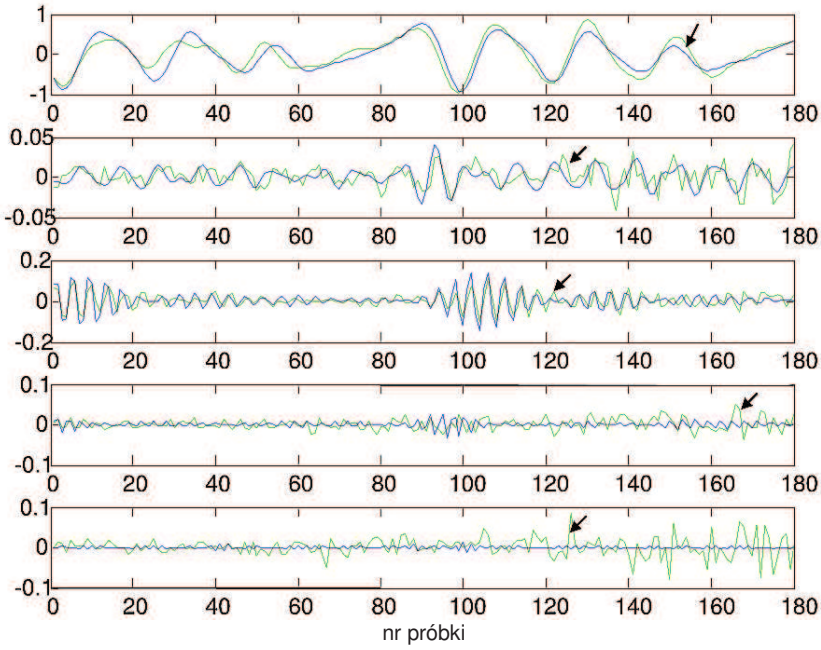
Rys. 5. Schemat eksperymentu rekonstrukcji sygnałów podpasmowych

Tab. 1. Błąd rekonstrukcji w funkcji liczby pasm

Liczba pasm	MSE
1	0,0436
2	0,0554
3	0,0523
5	0,0437
7	0,0518



Rys. 6. Przebiegi czasowe sygnałów podpasmowych przy podziale na 5 pasm częstotliwości



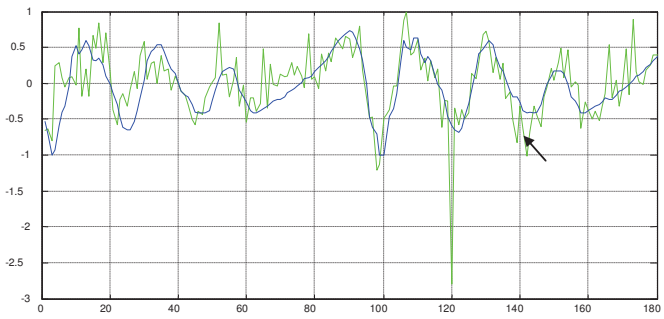
Rys. 7. Wynik rekonstrukcji w pasmach częstotliwości, przy podziale na 5 pasm. Przedstawiono jedynie przebiegi w rekonstruowanym przedziale, strzałkami oznaczono wynik rekonstrukcji. Parametry: *p=100*, *L=180*

4. Rekonstrukcja metodą nieliniowej sieci neuronowej

Ostatnim etapem badań było zastosowanie do rekonstrukcji sygnału neuronowej sieci nieliniowej, co oznacza zastosowanie nieliniowego modelu predykcji. W przypadku zastosowania sieci neuronowej macierze $\mathbf{U}$ i $\mathbf{Z}$ przedstawione wzorami (4) i (5) stanowią dane uczące $\langle \mathbf{U}, \mathbf{Z} \rangle$. Oczekiwania w tym przypadku są większe niż w przypadku modelu AR[6]. W tab. 2 przedstawiono wyniki dla różnych modeli jednokierunkowej sieci nieliniowej. We wszystkich przypadkach można zaobserwować pojawienie się silnych fluktuacji w wyniku rekonstrukcji (rys. 8).

Tab. 2. Ocena jakości rekonstrukcji dla różnych modeli sieci nieliniowej. Parametry eksperymentu: p=100, L=180

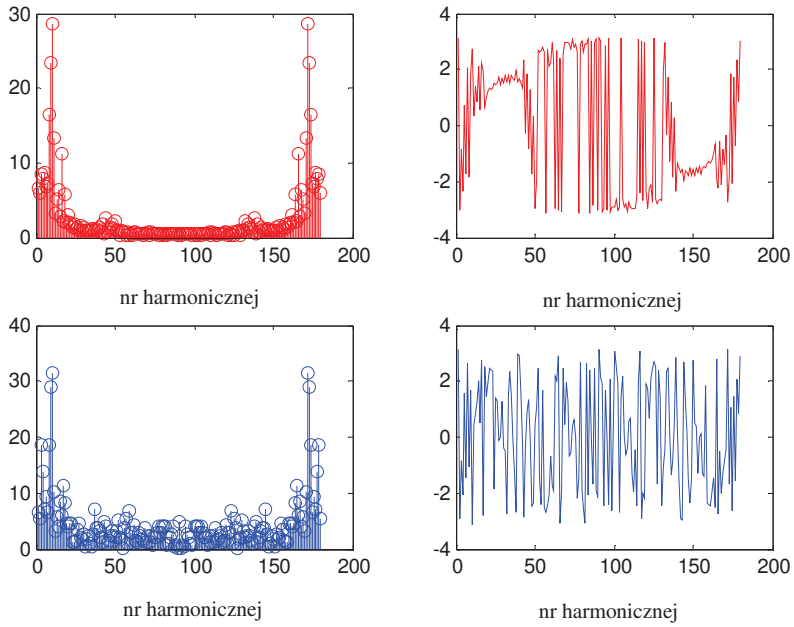
Nr modelu	Opis modelu sieci neuronowej	MSE
1	Sieć nieliniowa o strukturze dwuwarstwowej [2L L]; funkcja aktywacji: pierwsza warstwa – tanh, druga – liniowa; metoda uczenia: BP z momentum i zmiennym współczynnikiem szybkości uczenia.	0,0795
2	Sieć nieliniowa o strukturze dwuwarstwowej [2L L]; funkcja aktywacji: pierwsza warstwa – tanh, druga – tanh; metoda uczenia: BP z momentum i zmiennym współczynnikiem szybkości uczenia.	0,1106
3	L sieci, każda o strukturze [2 1]; funkcja aktywacji: pierwsza warstwa – tanh, druga – liniowa; metoda uczenia: LM (Levenberga-Marquardta).	0,1237



Rys. 8. Wynik rekonstrukcji na tle sygnału oryginalnego dla ostatniego modelu sieci z tab. 2. Wynik oznaczono strzałką

[6] Model AR odpowiada neuronowi liniowemu.

Owe silne fluktuacje bardzo dobrze widoczne są także w widmie sygnału będącego wynikiem rekonstrukcji (rys. 9). Obserwujemy duże rozbieżności widm sygnału oryginalnego i wyniku w zakresie wyższych częstotliwości. W sygnale wynikowym pojawiły się niewystępujące w oryginale harmoniczne o wyższych częstotliwościach. Zmniejszenie błędu daje filtracja dolnoprzepustowa dokonana na wyniku rekonstrukcji. Poprawę jakości rekonstrukcji powinno także przynieść skrócenie procesu uczenia sieci tak, aby zachowała swoje właściwości uogólniania informacji. Potwierdzają to wyniki zamieszczone w tab. 3, gdzie tę samą sieć trenowano wiele razy zmieniając w procesie uczenia liczbę epok i lepsze rezultaty uzyskano dla niezbyt wielkiej liczby epok.



Rys. 9. Widmo amplitudowe i fazowe przebiegu oryginalnego (wykresy dolne) i zrekonstruowanego (wykresy górne). Analizowany zakres częstotliwości 0-5,5 kHz

Tab. 3. Błąd rekonstrukcji w funkcji liczby epok uczenia sieci nieliniowej (model 3 z tab. 2)

Liczba epok	MSE
5	0,0928
7	0,0689
8	0,0794
10	0,0598
20	0,0692
150	0,1026

Podsumowanie

W artykule przedstawiono wyniki badań zastosowania modelu AR oraz nieliniowej sztucznej sieci neuronowej do rekonstrukcji sygnałów dźwiękowych. Nieliniowy model sieci neuronowej przebadano w kilku odmianach. W przypadku modelu AR wyniki są zadowalające. Z badań wynika, że rekonstrukcja przedziałów do 180 próbek jest obarczona niewielkim błędem, powyżej tej wartości błąd rośnie. Przeprowadzone z wykorzystaniem tego modelu badania rekonstrukcji z podziałem na pasma częstotliwości nie przyniosły poprawy. W przypadku zastosowania nieliniowej sieci neuronowej, co oznaczało przyjęcie nieliniowego modelu predykcji, otrzymane wyniki odbiegały od oczekiwanych. Co prawda, w tym przypadku można rekonstruować dłuższe przedziały, lecz w wyniku rekonstrukcji pojawiają się obce, wysokie harmoniczne. Dopiero zastosowanie filtracji dolnoprzepustowej przynosi poprawę rezultatu.

Literatura

[1] BISCAINHO L. W. P., DINIZ P. S. R, ESQUEL P. A. A., *A Model for an ARMA Process Split in Sub-Bands*, Proc. 2000 IEEE Intern. Symposium on Circuits and Systems, Vol. III, IEEE May 2000, pp. 97 – 100.

[2] BISCAINHO L. W. P., DINIZ P. S. R., ESQUEL P. A. A., *ARMA Process in Sub-Bands with Application to Audio Restoration*, Proceedings of IEEE International Symposium on Circuits and Systems, Sydney, Vol. 2 (2001), pp. 157 – 160.

[3] BEERENDS J. G., STEMERDING J. A., *A perceptual audio quality measure based on a psychoacoustic sound representation*, J.Audio Eng. Soc., Vol. 40, Dec. 1992, pp. 963 – 978.

[4] COCCHI G., UNCINI A., *Subbands Audio Signal Recovering using Neural Nonlinear Prediction*, Proceedings ICASSP 2001, Vol. 2, 7-11 May 2001, pp. 1289 – 1292.

[5] CZYŻEWSKI A., *Dźwięk cyfrowy*, Wybrane zagadnienia teoretyczne, technologia, zastosowania, Akademicka Oficyna Wydawnicza EXIT, Warszawa 2001.

[6] CZYŻEWSKI A., *Learning Algorithms for Audio Signal Enhancement, Part 1: Neural Network Implementation for the Removal of Impulse Distortions*, Journ. of Audio Engineering Society, Vol. 45, No. 10, Oct. 1996, pp. 815 – 831.

[7] GRAD L., *Zastosowania liniowego i nieliniowego modelu predykcji w analizie sygnałów mowy*, „Biuletyn IAiR", nr 10/1999, str. 25 – 40.

[8] GRAD L., *Restauracja nagrań dźwiękowych – usuwanie zakłóceń impulsowych*, „Biuletyn IAiR", nr 27/2009, str. 85 – 95.

[9] LIN H., GODSILL S. J., *The Multi-channel AR Model for Real-time Audio Restoration*, IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, Oct. 2005, pp. 335 – 338.

[10] LU L., MAO Y., WENYIN L., ZHANG H. J., *Audio Restoration by Constrained audio Texture Synthesis*, Proceedings ICASSP 2003, July 2003, Vol. 3, pp. 405 – 408.

[11] OSOWSKI S., *Sieci neuronowe do przetwarzania informacji*, Oficyna Wydawnicza Politechniki Warszawskiej, Warszawa, 2000.

[12] SENG K. P., HUI L. E., MING T. K., *Multimedia Signal Processing Using AI*, Communications, APPC 2003 Asia-Pacific Conference, Vol. 2, pp. 825 – 829.

[13] WOLF P. J., GODSILL S. J., *Interpolation of missing data value for audio signal restoration using a Gabor regression model*, Proceedings ICASSP 2005, IEEE, Vol. 5, pp. 517 – 520.

[14] ŻURADA J., BARSKI M., JĘDRUCH W., *Sztuczne sieci neuronowe*, Wydawnictwo Naukowe PWN, Warszawa, 1996.

Reconstruction of degraded parts of audio signals

ABSTRACT: The paper presents an issue of neural networks use for reconstruction of audio signals. Various models of neural networks were tested. The use of sub-band reconstruction was examined.

KEYWORDS: interpolation, neural networks, audio restoration.

Praca wpłynęła do redakcji: 16.02.2012