Rafał SAMBORSKI

WYKORZYSTANIE FILTRÓW ADAPTACYJNYCH DO EKSTRAKCJI MOWY Z SYGNAŁU ZASZUMIONEGO W SYSTEMACH WIELOMIKROFONOWYCH

STRESZCZENIE *W niniejszym artykule autor przedstawia wyniki badań nad wykorzystaniem filtrów adaptacyjnych w systemach wielomikrofonowych. Zostały przedstawione teoretyczne podstawy działania filtru adaptacyjnego (filtru Wienera) dla układu dwumikrofonowego, a następnie opisana została aplikacja tego filtru w układzie wspierającym komputerowy system rozpoznawania mowy przy obecności zakłóceń. Eliminowane zakłócenia miały nietypowy charakter – ich źródło zbliżone było do punktowego. Takie zakłócenia generowane mogą być np. przez włączony w pomieszczeniu odbiornik RTV. Autor prezentuje również szczegóły techniczne praktycznej implementacji algorytmu oraz prezentuje uzyskane wyniki porównując je z wynikiem działania prostego algorytmu.*

Słowa kluczowe: *filtr Wienera, system wielomikrofonowy, algorytm adaptacyjny*

1. WSTĘP

Jednym z podstawowych problemów, z jakimi borykają się badacze zajmujący się przetwarzaniem mowy jest szum. Jego obecność jest szczególnie

mgr inż. Rafał SAMBORSKI
e-mail: samborski@agh.edu.pl

Zespół Przetwarzania Sygnałów,
Katedra Elektroniki,
Akademia Górniczo-Hutnicza

PRACE INSTYTUTU ELEKTROTECHNIKI, zeszyt 246, 2010

dokuczliwa w komputerowych systemach rozpoznawania mowy ASR (ang. – *Automatic Speech Recognition*). Wynika to z faktu, iż komputer dysponując ograniczoną wiedzą semantyczną i brakiem doświadczenia lingwistycznego, nie jest w stanie uzupełnić rozpoznawanej wypowiedzi w momentach, w których pojawiają się zakłócenia. Przeprowadzone dotychczas badania [5] wskazują na znaczne pogorszenie skuteczności rozpoznawania mowy wraz ze spadkiem stosunku sygnału do szumu SNR (ang.– *Signal To Noise Ratio*).

Należy zastrzec, że jako szum rozumiemy zarówno zakłócenia akustyczne (hałas powodowany przez podmuchy wiatru, włączony odbiornik radiowy), jak i szum mający podłoże elektryczne. Ten drugi rodzaj szumu nie leży w obszarze omawianym przez niniejszy artykuł i w dalszych rozważaniach jego obecność zostanie pominięta.

Jedną z metod walki z zakłóceniami akustycznymi jest zastosowanie, zamiast jednego mikrofonu, matrycy wielu mikrofonów. Klasyczne algorytmy mikrofonowe bazują zwykle na założeniu, iż źródło sygnału użytecznego (mówca) nie przemieszcza się. W takiej sytuacji buduje się algorytm, który obrazowo mówiąc, wzmacnia sygnał pochodzący od mówcy, tłumiąc wszystkie pozostałe. Jest to o tyle łatwe, że położenie mówcy jest zwykle w przybliżeniu znane, jak ma to miejsce w przypadku samochodowego zestawu głośnomówiącego, matryc mikrofonowych stosowanych w telefonii internetowej.

Charakter sytuacji opisanej w artykule jest inny. Mamy do czynienia z mówcą swobodnie poruszającym się w pomieszczeniu, w którym znajduje się punktowe źródło zakłóceń o stałym położeniu jak np. odbiornik RTV.

W rozdziale 2 został zawarty szczegółowy opis badanej sytuacji. Kolejny rozdział przedstawia architekturę opracowanego algorytmu. Przykładowe wyniki eksperymentów zostały zebrane i zestawione z wynikami wcześniejszych prac w rozdziale 4. W podsumowaniu przedstawione zostały propozycje dalszego rozwoju systemu.

2. OPIS PROBLEMU

Specyfika założonych wymagań projektowych ograniczyła liczbę użytych mikrofonów do dwóch. W dalszej części będziemy zatem rozważać układ złożony z dwóch mikrofonów rejestrujących. Załóżmy, że szum obecny w nagraniu ma wyłącznie charakter losowy i dochodzi z różnych kierunków, a zatem nie występuje korelacja między szumem docierającym do pierwszego i do drugiego mikrofonu. Taką sytuację opisuje model analityczny:

$$\begin{aligned} s_{m1}(t) &= s_{voice}(t) + n_1(t), \\ s_{m2}(t) &= s_{voice}(t-\tau_1) + n_2(t), \end{aligned} \tag{1}$$

gdzie: $s_{voice}(t)$ jest użytecznym sygnałem mowy, który ma być poddawany dalszej analizie, a τ_1 – przesunięciem spowodowanym różnicą w odległości między mikrofonami a mówcą. Sygnały $n_1(t)$ i $n_2(t)$ odpowiadają odpowiednio szumowi docierającemu do pierwszego i drugiego mikrofonu.

W takiej sytuacji, aby wzmocnić sygnał użyteczny należy zsumować sygnały $s_{m1}(t)$ i $s_{m2}(t)$ przesunięte o opóźnienie τ_1. W tym celu obliczamy korelację między sygnałami $s_{m1}(t)$ i $s_{m2}(t)$, a następnie szukamy jej maksimum.

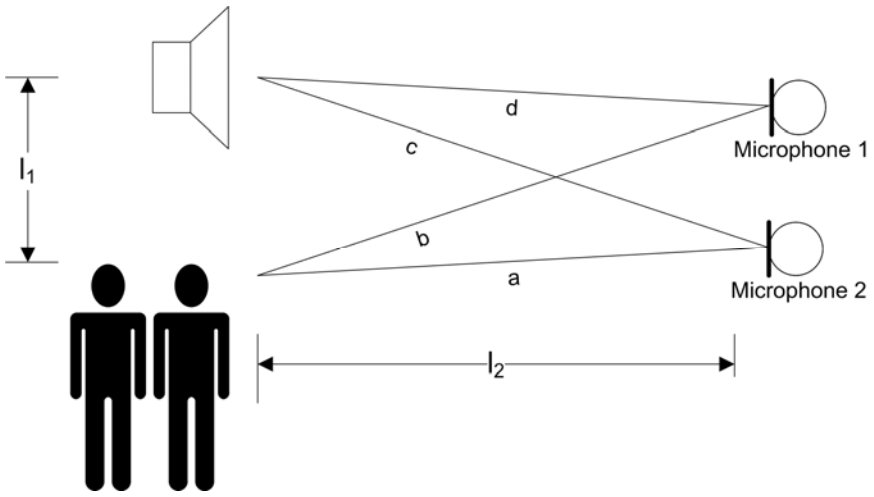
$$\tau_1 = \arg\max_{\tau}\left[\sum_n s_{m1}(n-\tau)s_{m2}(n)\right] \tag{2}$$

W literaturze szeroko omawiane są algorytmy bazujące właśnie na zasadzie sumowania odpowiednio przesuniętych sygnałów. Zaliczyć można do nich beamforming [2], superkierunkowy beamforming [1], i przetwarzanie oparte na właściwościach fazowych [3].

W omawianej powyżej sytuacji zakłócenia docierały do mikrofonów z różnych, losowych kierunków. Rozważany przypadek ma nieco inny charakter. Mamy do czynienia z sytuacją, w której w pomieszczeniu znajduje się swobodnie poruszający się mówca oraz źródło zakłóceń, o charakterze punktowym. Może to być np. włączony odbiornik radiowy. Zakłócenia docierają zatem do mikrofonów docierają z jednego kierunku. Taką sytuację prezentuje rysunek 1, a opisują poniższe równania:

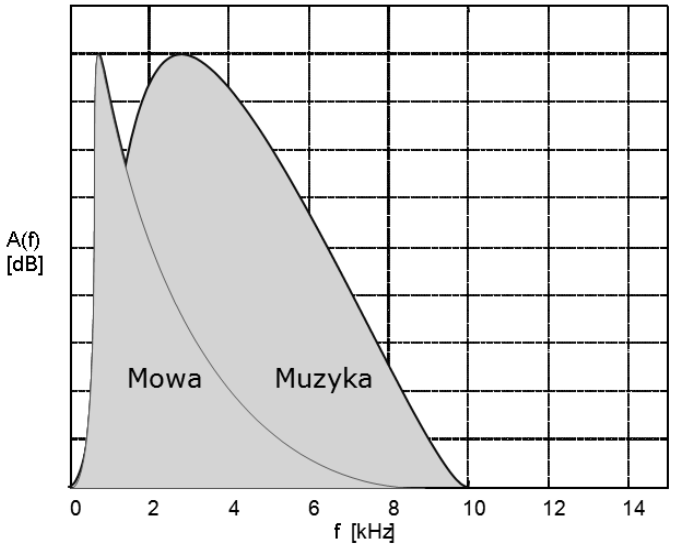
$$\begin{aligned} s_{m1}(t) &= s_{voice}(t) + s_{dist}(t) + n_1(t), \\ s_{m2}(t) &= s_{voice}(t-\tau_1) + s_{dist}(t-\tau_2) + n_2(t), \end{aligned} \tag{3}$$

gdzie $s_{dist}(t)$ jest zakłóceniem drugiego typu (źródło punktowe). Przesunięcie τ_2 nie jest równe τ_1 gdyż różnice odległości od źródła zakłócającego i mówcy są zwykle różne (nie zachodzi $|a-b| = |c-d|$). Obecne są również zakłócenia omówione wcześniej: $n_1(t)$ oraz $n_2(t)$.



Rys. 1. Przykładowa sytuacja, w której działa omawiany algorytm.
W pomieszczeniu znajdują się rozmówcy (mówca) oraz punktowe źródło zakłóceń (np. odbiornik RTV). W pewnej odległości od nich umieszczone zostały mikrofony

Ponieważ źródłem zakłóceń są zwykle inni mówcy lub urządzenie RTV, to większość ich mocy jest zgromadzona w tym samym paśmie, co w przypadku sygnału użytecznego (rys. 2). Zawodzi więc stosowana często w takich sytuacjach filtracja pasmowoprzepustowa lub pasmowozaporowa, która jest skuteczna np. w przypadku zakłóceń powodowanych przez sieć elektryczną. Można jednak wykorzystać fakt, iż źródło zakłóceń ma odmienne położenie w przestrzeni. Poprzez znalezienie przesunięcia τ_2 jesteśmy w stanie utworzyć taką różnicę sygnałów, w której obecność sygnału zakłócającego będzie znacznie zmniejszona.



Rys. 2. Porównanie charakteru widmowego mowy i muzyki.
Pasmo muzyki jest znacznie szersze niż pasmo mowy i pokrywa się z nim w szerokim zakresie częstotliwości

3. ALGORYTM ADAPTACYJNY

Przyjrzyjmy się jeszcze raz sytuacji przedstawionej na rysunku 1. Działanie omawianego dalej algorytmu opiera się na istnieniu różnicy pomiędzy odległościami $d_1 = |a-b|$ i $d_2 = |c-d|$. Gdyby odległości te były równe, a co za tym idzie równe byłyby opóźnienia τ_1 i τ_2, zmniejszanie amplitudy sygnału zakłócającego pociągałoby za sobą zmniejszenie amplitudy sygnału użytecznego. Powstaje zatem pytanie: jaka różnica między d_1 i d_2 zapewni poprawne działanie algorytmu?

Ponieważ będziemy operować na sygnałach dyskretnych zapiszmy (3) w następującej postaci

$$\begin{aligned} s_{m1}(nT_s) &= s_{voice}(nT_s) + s_{dist}(nT_s) + n_1(nT_s), \\ s_{m2}(nT_s) &= s_{voice}(nT_s - k_1T_s) + s_{dist}(nT_s - k_2T_s) + n_2(nT_s), \end{aligned} \tag{4}$$

gdzie T_s jest czasem próbkowania, a $n = 0,1,2\ldots$ oznacza numery kolejnych próbek sygnałów. Niech prędkość dźwięku w powietrzu równa będzie v. Wtedy

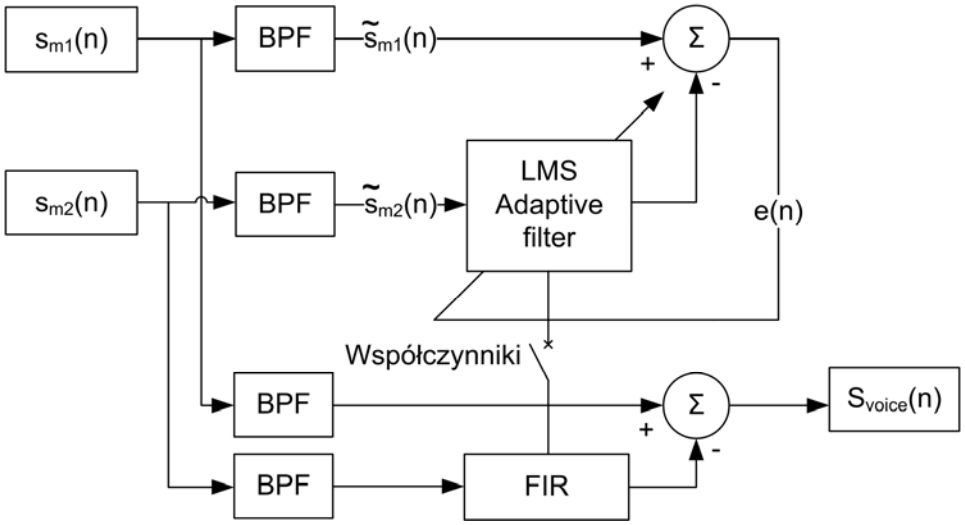
$$d_1 - d_2 = v(k_1 - k_2)T_s. \tag{5}$$

Przeprowadzone badania wskazują, że dla typowej częstotliwości próbkowania f_s = 44 100 kHz (Ts = 23 μs) wystarczająca jest różnica rzędu dziesięciu próbek. Zakładając, że v = 330 ms^{-1} otrzymujemy d_1 – d_2 = 7,5 cm. Oznacza to, że dla poprawnego działania opisywanego algorytmu w zupełności wystarczy różnica między d_1 – d_2 na poziomie kilku centymetrów.

Rysunek 3 przedstawia architekturę omawianego układu. Na wejście podawane są sygnały pochodzące z dwóch mikrofonów. W trakcie badań przetestowane zostały mikrofony pojemnościowe dwóch rodzajów: mikrofony wielkomembranowe, mikrofony małomembranowe. Oba zastosowane rodzaje mikrofonów posiadały kardioidalną charakterystykę kierunkową. Przeprowadzone pomiary wskazują, iż lepsze rezultaty udało się osiągnąć stosując mikrofony małomembranowe. Odległość między mikrofonami mieściła się w przedziale d_m = 10 ÷ 100 cm w zależności od geometrii badanej sceny.

Pierwszym stopniem omawianego układu jest filtracja pasmowo- przepustowa, która ma za zadanie pozbawić badane sygnały zakłóceń niskoczęstotliwościowych pochodzących od podmuchów powietrza, Zabieg ten tłumi rów-

nież pasma o wysokiej częstotliwości, które nie wpływają na czytelność mowy. W praktycznej implementacji dolna częstotliwość filtru wynosiła f_{c1} = 200 ÷ 300 Hz, a górna częstotliwość filtru pasmowo przepustowego – f_{c1} = 5000 ÷ 7000 Hz.



Rys. 3. Architektura omawianego układu opartego o filtr adaptacyjny

Przefiltrowany sygnał podawany jest na wejście algorytmu adaptacyjnego, tak jak zostało to przedstawione na rysunku 3. Filtr adaptacyjny w tej konfiguracji, ma na celu zmniejszenie ilości szumu obecnego w sygnale wejściowym. W celu wyjaśnienia zasady działania filtru adaptacyjnego załóżmy, że sygnał $\widetilde{s}_{m1}(n)$ jest tzw. sygnałem obserwowanym, a sygnał $\widetilde{s}_{m2}(n)$ posłuży do wyznaczenia sygnału estymowanego $s_{est}(n)$.

Algorytm wyznaczania optymalnego filtru został opisany przez Wienera w latach czterdziestych XX wieku. Kryterium działania tego rozwiązania jest minimalizacja wartości oczekiwanej błędu średniokwadratowego [4]

$$Q = E\left\{e^2(n)\right\}, \tag{6}$$

gdzie $e(n)$ jest różnicą między sygnałem obserwowanym a sygnałem referencyjnym. Powyższy wzór można, zatem zapisać jako:

$$Q = E\left\{\left(\widetilde{s}_{m1}(n) - s_{est}(n)\right)^2\right\}. \tag{7}$$

Sygnał estymowany $s_{est}(n)$ otrzymywany jest w wyniku przejścia sygnału $\widetilde{s}_{m2}(n)$ przez filtr o transmitancji $\mathbf{h}^T$. Można go zatem zapisać, jako:

$$s_{est}(n) = \mathbf{h}^T \widetilde{\mathbf{s}}_{m2,n}\,, \tag{8}$$

gdzie:

$\mathbf{h} \in \Re^{N+1}$ jest wektorem zawierającym współczynniki filtru, a

$$\widetilde{\mathbf{s}}_{m2,n} = \left[s_{obs}(n), s_{obs}(n-1), \ldots, s_{obs}(n\_N)\right]^T \in \Re^{N+1} \tag{9}$$

jest wektorem zawierającym N+1 próbek sygnału $\widetilde{s}_{m2}(n)$.

Łącząc ze sobą (7) i (8) otrzymujemy:

$$Q = E\left\{\left(\widetilde{s}_{m1}(n) - \mathbf{h}^T \widetilde{\mathbf{s}}_{m2,n}\right)^2\right\}, \tag{10}$$

co równoważnie można zapisać jako:

$$Q = E\left\{\widetilde{s}_{m1}^2(n)\right\} - 2\mathbf{h}^T \mathbf{\Phi}_{corr} + \mathbf{h}^T \mathbf{\Phi}_{obs} \mathbf{h}\,, \tag{11}$$

gdzie: $\mathbf{\Phi}_{corr} = E\left\{\widetilde{s}_{m2}(n)\widetilde{s}_{m1}(n)\right\}$ jest wektorem korelacji, a $\mathbf{\Phi}_{obs} = E\left\{\widetilde{s}_{m2}(n)\widetilde{s}_{m2}^T(n)\right\}$ jest macierzą autokorelacji.

Poszukiwany filtr Winera jest w istocie wektorem $\mathbf{h}_{Wiener}$, który minimalizuje kryterium (10) i może zostać zapisany jako:

$$\mathbf{\Phi}_{obs} \mathbf{h}_{Wiener} = \mathbf{\Phi}_{corr}\,. \tag{12}$$

Warto wspomnieć, że powyższa operacja zachodzi w momencie, gdy żaden z sygnałów $s_{m1}(n)$ i $s_{m2}(n)$ nie zawiera sygnału użytecznego (mowy). Oznacza to, że współczynniki filtru poszukiwane są jedynie w momentach, gdy do mikrofonów dociera jedynie sygnał zakłócający. Takie działanie sprawia, że adaptacja filtru zachodzi w ten sposób, by minimalizować obecność sygnału zakłócającego w sygnale wyjściowym. W praktycznej realizacji sygnałem do rozpoczęcia adaptacji może być decyzja operatora przesłuchującego nagrania (jeśli mamy do czynienia z postprocessingiem danych już nagranych) lub wykrycie odpowiedniego momentu przez algorytm voice activity detection (VAD).

Przeprowadzone eksperymenty wykazały, że do poprawnej adaptacji zastosowanego filtru Wienera o 100 współczynnikach wystarczy 0,8 ÷ 1,2 s nagrania. Takie przerwy dosyć często pojawiają się między poszczególnymi wypowiedziami mówcy, więc nie jest trudno odnaleźć wystarczająco długi

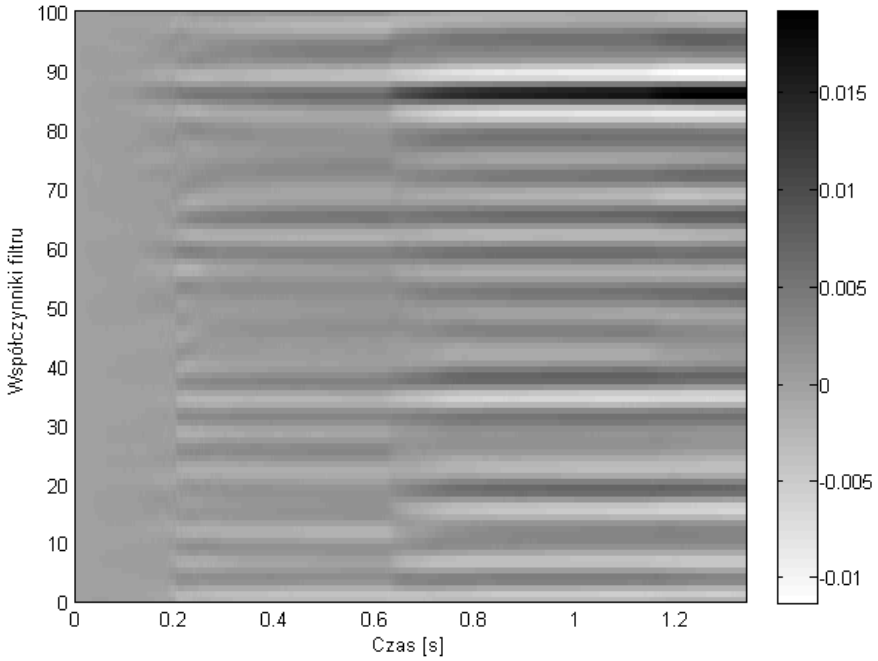
przedział czasu. Gdy współczynniki filtru są już ustabilizowane zostają one przeniesione do bloku zwykłej filtracji FIR, gdzie następuje właściwa filtracja sygnału. Sygnał wynikowy otrzymywany jest jako

$$s_{voice}(n) = \widetilde{s}_{m1}(n) - \mathbf{h}_{Wiener}^{T} \widetilde{\mathbf{s}}_{m2,n} . \tag{13}$$

W chwili, gdy ulegną zmianie warunki akustyczne (w szczególności, gdy zmieni sie położenie mikrofonów lub źródła zakłócającego), operacja adaptacji musi zostać przeprowadzona ponownie.

4. WYNIKI BADAŃ

Przebieg przykładowej adaptacji filtru prezentuje rysunek 4. Widoczny jest wyraźnie proces adaptacji – dopasowywania się współczynników filtru, który w tym wypadku osiąga zadowalający poziom po ok. 0,8 s.

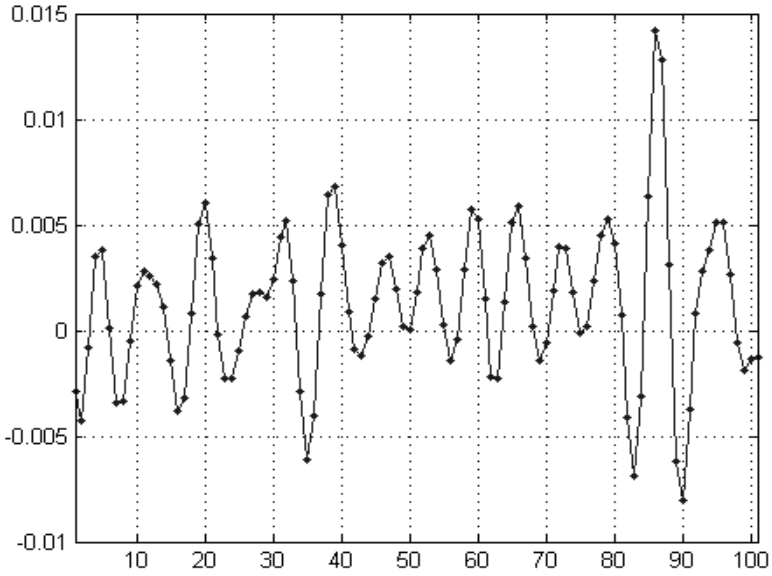


Rys. 4. Przebieg adaptacji filtru Wienera podczas jednego z eksperymentów. Widoczna jest stabilizacja współczynników po ok. 0,8 s

Na rysunku 5 przedstawione zostały wartości współczynników filtru z rysunku 4 po 0,8 s od momentu rozpoczęcia adaptacji. Te wartości współczynników zostały przekazane do bloku FIR. Widoczne jest wyraźne maksimum

dla współczynnika $\mathbf{h}_{\text{Wiener}}$ [86]. Obecność tego maksimum jest bezpośrednio związana z geometrią układu przedstawionego na rysunku 1. Najwyższą wartość ma, bowiem współczynnik odpowiadający takiej ilości próbek, o jaką przesunięte są względem siebie sygnały $s_{m1}(n)$ i $s_{m2}(n)$. Korzystając z (5) możemy wyliczyć odpowiednią odległość podstawiając $k_1 - k_2 = 86$, Ts = 23 μs i v = 330 ms^{-1}.

$$d_1 - d_2 \approx 65 \text{ [cm]}. \tag{14}$$



Rys. 5. Przykładowe współczynniki filtru Wienera.
Wyraźne maksimum odpowiada ilości próbek, o jakie są przesunięte względem siebie sygnały wejściowe

Tabela 1 porównuje poprawę SNR uzyskaną przy użyciu opisanego algorytmu w porównaniu z badanym wcześniej algorytmem Sum-And-Delay opisanym w [6]. Widać poprawę o ok. 1 dB w stosunku do wcześniejszego rozwiązania. Warto nadmienić, że jest to poprawa uśredniona. W niektórych przypadkach (w zależności od warunków akustycznych i rodzaju zakłóceń), poprawa była znacznie większa.

TABELA 1
Porównanie wyniku działania opisanego algorytmu opartego o filtr Wienera oraz wcześniej badanego algorytmu Sum-And-Delay

Algorytm	Poprawa SNR [dB]
Sum-And-Delay	2.0
Algorytm adaptacyjny	2.9

5. PODSUMOWANIE

Przedstawiony algorytm średnio o 3 dB poprawia stosunek sygnału do szumu w rejestrowanym sygnale. Taka poprawa subiektywnie nie wydaje się znacząca, ale jest już słyszalna. Dla komputerowego systemu rozpoznawania mowy taka poprawa oznacza jednak (w zależności od początkowego SNR) poprawę skuteczności rozpoznania od 5% do 10% [5]. Taki preprocessing jest szczególnie istotny w środowiskach, gdzie zakłócenia akustyczne są szczególnie duże.

Dalsza rozbudowa algorytmu przewiduje zwiększenie ilości mikrofonów oraz zastosowanie również innych kryteriów poza wartością oczekiwaną błędu średniokwadratowego. Pozwoli to na lepszą eliminację tych artefaktów, które pochodzą nie tylko bezpośrednio od źródła, a także w wyniku odbić sygnałów od ścian pomieszczenia. Eksperymenty opisywane w literaturze wykazują możliwość poprawy SNR o 5 dB, co przekłada się na poprawę skuteczności rozpoznania nawet o 15%.

LITERATURA

1. Bitzer J., Simmer K. U. i Kammeyer K. D.: Theoretical noise reduction limits of the generalized sidelobe canceller (GSC) for speech enhancement. Proc. IEEE Int. Conference on Acoustics, Speech, Signal Processing, 5:2965–2968, 1999.
2. DeMuth G.: Frequency domain beamforming techniques. Proc. IEEE Int. Conference on Acoustics, Speech, Signal Processing, 2:713–715, 1977.
3. Halupka D., Rabi A. S., Aarabi P., i Sheikholeslami A.: Low-power dual-microphone speech enhancement using field programmable gate arrays. IEEE Transactions on Signal Processing, vol. 55, no. 7, pp. 3526–3535, 2007.
4. S. Haykin. Adaptive Filter Theory. Prentice-Hall, Nowy Jork, 1996.
5. Kollmeier B., Brand T., Meyer B.: Perception of Speech and Sound. Springer Handbook of Speech Processing. Springer-Verlag, Berlin Heidelberg, 2008.
6. Ziółko M., Ziółko B., Samborski R.: Dual-microphone speech extraction from signals with audio background. Proc. IEEE International Conference on Intelligent Information Hiding and Multimedia Signal Processing, 2009.

Rękopis dostarczono dnia 17.08.2010 r.
Opiniował: dr hab. inż. Stefan F. Filipowicz – prof. PW

ADAPTIVE FILTERS IN THE EXTRACTION OF SPEECH FROM DISTURBED SIGNAL IN THE MULTI-MICROPHONES SYSTEMS

Rafał SAMBORSKI

ABSTRACT *This paper presents the results of the research into the utilization of adaptive filters in multi-microphone systems. The theoretical basis of adaptive filtration* (*Wiener filtration*) *in dual-microphone system was explained. The author presents an application of a filter in a speech enhancement system for automatic speech recognition in presence of strong disturbances. Particularly, disturbances generated by point sources (such as radio set or TV) were considered. The author also presents the technical details of the practical implementation of the algorithm. The results of this algorithm were compared with previously obtained results of the Sum-And-Delay algorithm. Opportunities for further development of this solution were also suggested.*

Mgr inż. Rafał SAMBORSKI – absolwent Elektroniki na Wydziale Elektrotechniki, Automatyki, Informatyki i Elektroniki Akademii Górniczo-Hutniczej. W 2009 roku rozpoczął studia doktoranckie na tym samym wydziale. Jednocześnie współpracuje z Zespołem Przetwarzania Sygnałów w Katedrze Elektroniki AGH. Jego zainteresowania naukowe obejmują przede wszystkim systemy wielomikrofonowe w szczególności w kontekście wsparcia komputerowych systemów rozpoznawania mowy.

IEl, Warszawa 2010. Nakład 110 + 10 egz. Ark. wyd. 18,61. Ark. druk. 13,75. Pap. off. Kl.III. 80 g.
Oddano do druku we wrześniu 2010 r. Druk ukończono we wrześniu 2010 r.
Redakcja – Wydawnictwo Instytutu Elektrotechniki
Indeks nr 37656