ZESZYTY NAUKOWE POLITECHNIKI ŚLĄSKIEJ 2003
Seria: AUTOMATYKA z. 138 Nr kol. 1581

Marek BINKOWSKI
Politechnika Śląska

MODUŁ WYKRYWANIA SYGNAŁU GŁOSOWEGO O ZRÓŻNICOWANEJ AMPLITUDZIE W SYSTEMIE AUTOMATYCZNEGO ROZPOZNAWANIA MOWY

Streszczenie. W pracy opisano przykładowy moduł wykrywania sygnału głosowego, wspomagający system rozpoznawania mowy. Jednym z bloków funkcjonalnych opisywanego modułu jest funkcja wyznaczająca obwiednię sygnału. Zaproponowano działanie modułu przy zastosowaniu różnych funkcji obwiedni: wartości maksymalnej sygnału w oknie, średniej wartości bezwzględnej sygnału w oknie, odchylenia standardowego sygnału w oknie, energii sygnału w oknie oraz entropii sygnału w oknie. W końcowej części pracy porównano rezultaty otrzymane dla przykładowej wypowiedzi, zarejestrowanej wraz z zakłóceniami i szumem otoczenia. Wskazano również możliwości dalszych badań i sposoby poprawy skuteczności wykrywania głosu.

VOICE DETECTION MODULE IN AUTOMATIC SPEECH RECOGNITION SYSTEM

Summary. This work describes an example of voice detection module, supporting speech recognition system. One of functional blocks of described module is an envelope function. Voice detection efficiency has been examined using various envelope functions: maximum value of signal in a window, mean of absolute values of signal in a window, standard deviation of signal in a window, energy of signal in a window and entropy of signal in a window. In last part of the work the results of analysis of an exemplary signal have been presented. The signal has been recorded with accompanying ambient noise and disturbances. Also means of improvements and areas of further exploration has been proposed.

1. Wprowadzenie

Systemy automatycznego rozpoznawania mowy borykają się z wieloma problemami, które wpływają na obniżenie skuteczności i jednoznaczności rozpoznanego sygnału. Proces rozpoznawania mowy jest zwykle wieloetapowy. Typowymi etapami przetwarzania sygnału

mowy w dzisiejszych systemach rozpoznawania są: rejestracja analogowego sygnału dźwiękowego i jego konwersja na postać cyfrową, wstępne przetwarzanie sygnału w celu minimalizacji zakłóceń, ekstrakcja cech związanych z niesioną informacją, klasyfikacja tych cech i wyznaczenie wyniku rozpoznania. W zależności od specyfiki systemu i skali przetwarzania, wynikiem rozpoznania może być pojedynczy fonem, głoska, wyraz, zdanie lub nawet wypowiedź.

Typowy system rozpoznawania mowy zwykle dysponuje bazą cech-wzorców, na podstawie których klasyfikuje zarejestrowaną i wstępnie przetworzoną wypowiedź (lub jej fragmenty, takie jak fonem, głoska czy wyraz). Jest zatem przygotowany (lepiej lub gorzej) do analizy wypowiedzi, złożonej z ustalonych wcześniej elementów konkretnego języka. Działanie systemu polega na dopasowaniu zarejestrowanego sygnału do zapamiętanego wcześniej wzorca i dobraniu najbardziej zbliżonego rezultatu.

Trzeba zauważyć, że typowe systemy rozpoznawania mowy zwykle nie dysponują wzorcami dźwięków nie odpowiadających elementom języka. Baza cech-wzorców nie zawiera zatem danych na temat pozornej ciszy[1], różnych rodzajów szumu, czy też dźwięków otoczenia. Utworzenie tak obszernej bazy byłoby nieopłacalne i wręcz niemożliwe ze względu na nieskończoną różnorodność otaczających nas dźwięków.

Powstaje zatem uzasadniona obawa, że hipotetyczny system rozpoznawania mowy, który skutecznie radzi sobie z rozpoznaniem rzeczywistej wypowiedzi, w przypadku „obcych" sygnałów może dawać niepoprawne rezultaty. Pożądanym rezultatem w przypadku takich obcych sygnałów byłaby, oczywiście, odpowiedź typu „brak wypowiedzi" lub „obcy sygnał". Niestety, istnieje groźba, że sygnał nie będący wypowiedzią zostanie zinterpretowany jako wypowiedź, co pociągnie za sobą powstanie błędnego rezultatu i wpłynie na niewiarygodne działanie systemu.

W związku z tym istnieje potrzeba zbadania możliwości eliminacji wspomnianych „obcych" sygnałów w fazie wstępnego przetwarzania. Eliminacja może polegać na przykład na wykrywaniu sygnału głosowego (mowy) i przesyłaniu do systemu tylko tego sygnału; w czasie, gdy nie jest wykrywana mowa, rejestrowany sygnał nie jest przesyłany do dalszych modułów systemu. W literaturze nie znaleziono wyczerpującego omówienia zastosowań typowych metod detekcji sygnału w systemach automatycznego rozpoznawania mowy. Problem

[1] Próg słyszalności ucha ludzkiego dla częstotliwości 1000 Hz to 20 μPa. Głośność takiego dźwięku jest oznaczana w skali decybelowej przez 0 dB; dźwięk o takiej głośności jest odbierany jako cisza [1]. Oczywiście, istnieją również dźwięki o mniejszej głośności, dlatego powyżej użyto sformułowania „pozorna" cisza.

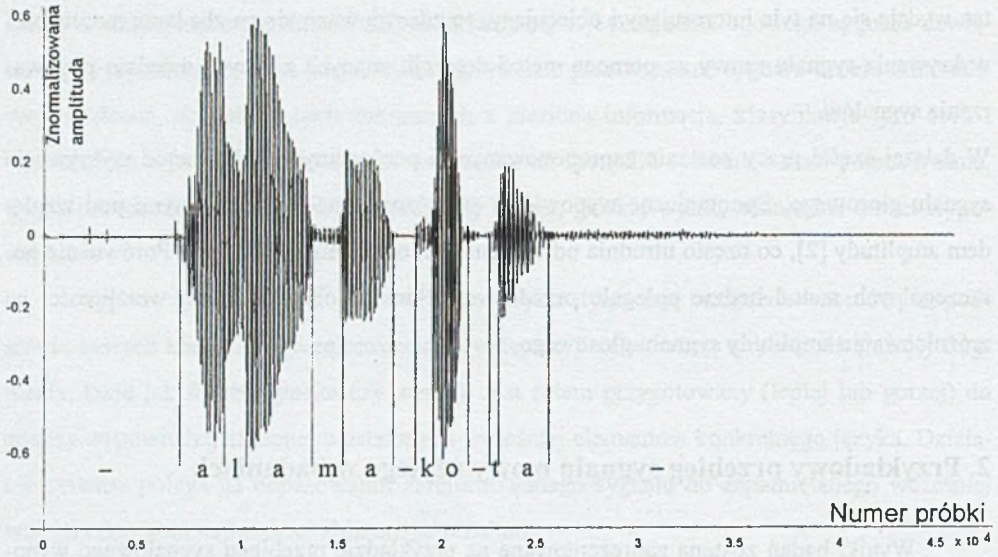
ten wydaje się na tyle interesujący i obiecujący, że zdecydowano się na zbadanie możliwości wykrywania sygnału mowy za pomocą metod detekcji, znanych z innych dziedzin przetwarzania sygnałów.

W dalszej części pracy zostanie zaproponowanych i porównanych kilka metod wykrywania sygnału głosowego. Spontaniczne wypowiedzi są zróżnicowane między innymi pod względem amplitudy [2], co często utrudnia odróżnienie ich od szumu i hałasu tła. Porównanie poszczególnych metod będzie polegało przede wszystkim na obserwacji ich wrażliwości na zróżnicowanie amplitudy sygnału głosowego.

2. Przykładowy przebieg sygnału mowy użytego w badaniach

Wyniki badań zostaną zaprezentowane na przykładzie przebiegu sygnałowego wypowiedzi „Ala ma kota", przedstawionego na rysunku 1. Widoczny na rysunku przebieg można podzielić na trzy fragmenty. Pierwszy fragment reprezentuje czas przed rozpoczęciem wypowiedzi. W tym czasie zarejestrowano pewne obce sygnały, takie jak szum, trzaski mikrofonu oraz odległe szczekanie psa. Drugi fragment to sama wypowiedź. Można zaobserwować, że występują w niej krótkie przerwy, podczas których nie jest wypowiadana żadna głoska. Takie przerwy są typowe dla spontanicznych, swobodnych wypowiedzi i stanową pewne dodatkowe utrudnienie, z którym należy się liczyć. Trzeci fragment reprezentuje czas po zakończeniu wypowiedzi. W tym czasie zarejestrowano wzmożony szum, którego źródłem było powietrze nagromadzone w płucach mówcy i wydychane na siatkę mikrofonu.

Wypowiedziom często towarzyszą różne dodatkowe dźwięki wydawane przez samego mówcę, które nie niosą żadnej informacji i w przypadku naturalnej komunikacji są przez nas po prostu ignorowane. Dopiero po zobrazowaniu sygnału w postaci wykresu okazuje się, jak duży wpływ mogą mieć na działanie automatycznych systemów rozpoznawania mowy. W tym przypadku widać, że sygnał szumu wydychanego powietrza ma amplitudę zbliżoną do amplitudy niektórych głosek w samej wypowiedzi, takich jak „m", „k" czy „t". Jak się okaże, takie podobieństwo może być poważnym utrudnieniem w skutecznym wykrywaniu sygnału głosowego.



Rys. 1. Wypowiedź „Ala ma kota" zarejestrowana i skonwertowana na postać cyfrową. Myślniki reprezentują przerwy w wypowiedzi

Fig. 1. The recorded and digitalized phrase „Ala ma kota". The dashes represent gaps in spoken phrase

3. Rola i budowa modułu wykrywania sygnału głosowego

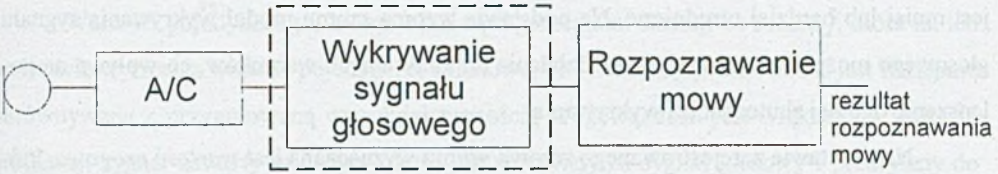
W tym rozdziale zostanie przedstawiony moduł wykrywania sygnału głosowego, będący przedmiotem badań. Omówiona zostanie jego rola w systemie rozpoznawania mowy, jego budowa oraz poczynione założenia. W następnym rozdziale zostaną wymienione i opisane poszczególne funkcje obwiedni, wykorzystane w tym module podczas badań.

3.1. Rola prezentowanego modułu

Ostatecznym celem badań jest skonstruowanie modułu, który będzie wspomagał pracę systemu rozpoznawania mowy. Rolą modułu będzie wstępne przetwarzanie sygnału, polegające na wykryciu sygnału głosowego.

Na wejście modułu jest podawany sygnał zarejestrowany przez mikrofon i spróbkowany przez przetwornik analogowo-cyfrowy. Na wyjściu modułu pojawiają się te fragmenty sygnału, które zostaną przez moduł uznane za sygnał głosowy. Pozostałe fragmenty, będące szumem, zakłóceniami i innymi obcymi dźwiękami występującymi w czasie, gdy nie jest rejestrowana wypowiedź, zostaną odrzucone. Moduł pełni zatem rolę bramki, która przepuszcza

jedynie fragmenty sygnału odpowiadające wypowiedzi. Rysunek 2 przedstawia umiejscowienie modułu wykrywania sygnału głosowego w przykładowym systemie rozpoznawania mowy.

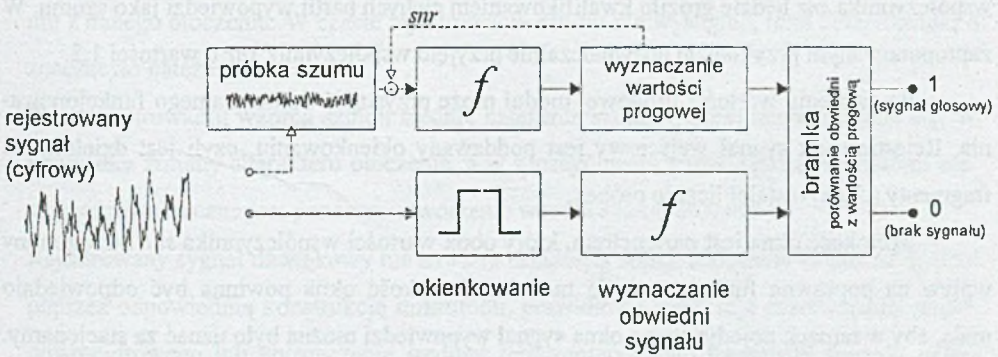


Rys. 2. Umiejscowienie modułu wykrywania sygnału głosowego w systemie rozpoznawania mowy
Fig. 2. Location of voice detection module in speech recognition system

3.2. Ogólna budowa modułu

Ogólne działanie modułu polega na wyznaczaniu obwiedni rejestrowanego sygnału. Obwiednia jest wektorem E, którego składowe są wyznaczane za pomocą funkcji obwiedni *f* na podstawie wartości rejestrowanego sygnału. Pojedyncza składowa obwiedni, E(k), jest wyznaczana na podstawie wydzielonego k-tego fragmentu sygnału, zwanego oknem. Po wyznaczeniu wartości obwiedni E(k) jest ona porównywana z ustaloną wartością progową. W rezultacie porównania podejmowana jest decyzja, czy dany fragment sygnału należy zakwalifikować jako sygnał głosowy, czy też jako sygnał obcy.

Moduł wykrywania sygnału głosowego zaprojektowano w taki sposób, by umożliwiał porównanie skuteczności wykrywania przy zastosowaniu różnych funkcji obwiedni *f*. Rysunek 3 przedstawia ogólną budowę modułu.



Rys. 3. Ogólna budowa modułu wykrywania sygnału głosowego
Fig. 3. General structure of voice detection module

Wstępnym etapem działania modułu jest zarejestrowanie i zapamiętanie wzorca szumu otoczenia. Dysponowanie takim wzorcem jest istotne z tego względu, że w zależności od natężenia i charakterystyki częstotliwościowej szumu otoczenia, wykrywanie sygnału głosowego jest mniej lub bardziej utrudnione. Na podstawie wzorca szumu moduł wykrywania sygnału głosowego może zaadaptować swoje działanie do konkretnych warunków, co wpływa na polepszenie ogólnej skuteczności wykrywania.

Na podstawie zarejestrowanego wzorca szumu wyznaczana jest *wartość progowa*, która następnie pełni kluczową rolę w funkcjonowaniu modułu. Wartość progowa jest wyznaczana za pomocą tej samej funkcji obwiedni *f*, która w dalszej, normalnej pracy modułu służy do wyznaczania obwiedni rejestrowanego sygnału.

Przed obliczeniem wartości progowej składowe wzorca szumu są mnożone przez współczynnik *snr* (*signal to noise ratio* — stosunek sygnału do szumu). Współczynnik ma wartość większą od 1 i określa, o ile większa powinna być wartość obwiedni sygnału od wartości obwiedni szumu, by sygnał został uznany za wypowiedź. Na przykład, gdy współczynnik *snr* ma wartość 1,2, wartość obwiedni sygnału musi być co najmniej 1,2 raza większa od wartości obwiedni szumu, by została zakwalifikowana jako wypowiedź. Dzięki temu obliczana wartość progowa zapewnia pewną tolerancję, zmniejszając tym samym prawdopodobieństwo zakwalifikowania szumu jako sygnału.

Dobór wartości współczynnika *snr* ma bardzo duży wpływ na skuteczność wykrywania sygnału mowy. Przyjęcie zbyt małej wartości współczynnika *snr* spowoduje, że szum otoczenia o chwilowej głośności nieznacznie większej od głośności zarejestrowanego wzorca szumu zostanie zakwalifikowany jako wypowiedź. Z kolei przyjęcie zbyt dużej wartości współczynnika *snr* będzie groziło kwalifikowaniem cichych partii wypowiedzi jako szumu. W zaproponowanym przykładzie doświadczalnie przyjęto współczynnik *snr* o wartości 1,2.

Po ustaleniu wartości progowej moduł może przystąpić do normalnego funkcjonowania. Rejestrowany sygnał wejściowy jest poddawany okienkowaniu, czyli jest dzielony na fragmenty (okna) o stałej liczbie próbek.

Szerokość okna jest parametrem, który obok wartości współczynnika *snr* ma ogromny wpływ na poprawne funkcjonowanie modułu. Szerokość okna powinna być odpowiednio mała, aby w ramach pojedynczego okna sygnał wypowiedzi można było uznać za stacjonarny. Z drugiej strony okno nie może być zbyt wąskie, aby wyznaczana wartość obwiedni nie była zbytnio uzależniona od chwilowych wartości szumu. Dobór szerokości okna jest uzależniony przede wszystkim od zastosowanej częstotliwości próbkowania dźwięku i jest przeprowadza-

ny doświadczalnie. W przykładzie przedstawionym w tej pracy zastosowano częstotliwość próbkowania 16 kHz i dla takiej częstotliwości przyjęto okno o szerokości 100 próbek (czyli obejmujące sygnał o czasie trwania 6,25 ms).

Wartości pojedynczego, k-tego okna są argumentami funkcji obwiedni *f*, która na ich podstawie wyznacza wartość pojedynczej składowej obwiedni, E(k). Ta wartość jest następnie porównywana z przygotowaną wcześniej wartością progową. Jeśli jest większa od wartości progowej, sygnał zawarty w danym oknie jest uznawany za sygnał głosowy i przesyłany do dalszych modułów systemu rozpoznawania mowy. W przeciwnym razie sygnał w danym oknie jest uznawany za sygnał obcy, który nie wymaga rozpoznawania, ponieważ nie zawiera informacji językowych.

3.3. Założenia

Przygotowując się do budowy modułu wykrywania sygnału głosowego, poczyniono następujące założenia:

- Dźwięk rejestrowany przez mikrofon i przetwarzany przez przetwornik analogowo-cyfrowy zawiera jedynie sygnał mowy oraz szum otoczenia. Nie zawiera natomiast innych dźwięków o charakterystyce podobnej do sygnału mowy, a w szczególności dźwięków o natężeniu zbliżonym do natężenia sygnału głosowego.
- Sygnał odpowiadający wypowiedzi jest wyraźnie głośniejszy od sygnału otoczenia. Założenie to można spełnić między innymi poprzez umieszczenie mikrofonu w bezpośredniej bliskości ust mówcy (jeśli hałas otoczenia nie dominuje nad głosem mówcy).
- Praca modułu wykrywania sygnału mowy rozpoczyna się od zarejestrowania wzorca szumu z danego otoczenia. W czasie rejestrowania szumu nie występują inne obce dźwięki o znaczącym natężeniu.
- Po zarejestrowaniu wzorca szumu średnie natężenie szumu otoczenia nie zmienia się. W przypadku zmiany charakteru otoczenia, a w szczególności zmiany głośności szumu otaczającego konieczne jest ponowne utworzenie wzorca szumu otoczenia.
- Rejestrowany sygnał dźwiękowy nie zawiera składowej stałej. Założenie to można spełnić poprzez odpowiednią konstrukcję mikrofonu, prawidłową kalibrację przetwornika analogowo-cyfrowego lub wyznaczenie średniej reprezentatywnego fragmentu sygnału cyfrowego i odjęcie jej od tego sygnału.

- Moduł wykrywania sygnału głosowego pełni rolę pomocniczą w systemie rozpoznawania mowy. Wszelkie opóźnienia wprowadzone przez ten moduł dodają się do sumarycznego czasu upływającego od zarejestrowania wypowiedzi do jej rozpoznawania. Dlatego należy dążyć do minimalizacji opóźnień wprowadzanych przez ten moduł.

W związku z pierwszym i drugim założeniem, proponowany moduł w obecnej postaci nie jest przystosowany do dowolnych warunków akustycznych. Jego funkcjonowanie można sprawdzić w specyficznych warunkach, na przykład w studiu nagraniowym, gdzie wszelkie obce dźwięki o znaczącym poziomie głośności są wyciszone.

4. Badane funkcje obwiedni

W niniejszej pracy zostanie zbadane i porównane działanie modułu wykrywania sygnału głosowego korzystającego z pięciu różnych funkcji obwiedni. Dobór funkcji podyktowany był obserwacją własności specyficznych dla sygnału mowy, a także ostatnim z wymienionych powyżej założeń. Skuteczność wykrywania sygnału mowy zostanie zbadana na przykładzie z zastosowaniem następujących funkcji obwiedni:

- wartości maksymalnej sygnału w oknie,
- średniej wartości bezwzględnej sygnału w oknie,
- odchylenia standardowego sygnału w oknie,
- energii sygnału w oknie,
- entropii sygnału w oknie.

W każdym przypadku zostanie podany matematyczny opis funkcji oraz uzasadnienie jej wyboru. W dalszej części pracy zostaną porównane rezultaty uzyskane za pomocą poszczególnych funkcji.

4.1. Wartość maksymalna sygnału w oknie

Najprostszą i najbardziej intuicyjną z omawianych funkcji obwiedni jest funkcja wyznaczająca wartość maksymalną sygnału w oknie[2]. Obserwując przebieg sygnału dźwiękowego łatwo zauważyć, że w miejscach, gdzie zarejestrowano wypowiedź, amplituda sygnału jest

[2] Ze względu na fakt, że większość przebiegów dźwiękowych jest w przybliżeniu symetryczna względem osi zerowej, równie dobrze można zastosować w tym miejscu funkcję wyznaczającą minimalną wartość sygnału w oknie.

zwykle wyraźnie większa od miejsc, w których nie ma wypowiedzi. Dlatego najprostszą metodą wykrycia sygnału głosowego jest porównanie lokalnego maksimum zarejestrowanego sygnału z wartością progową, będącą maksymalną wartością próbki szumu (skorygowaną przez współczynnik *snr*).

Oznaczając numer okna przez *k* oraz indeks próbki w danym oknie przez *i*, można wyrazić funkcję wyznaczającą obwiednię dźwięku następującym wzorem:

$$E(k) = \max_i(s_k(i))$$

E(k) jest skalarem reprezentującym wartość obwiedni dla *k*-tego okna sygnału, zaś $s_k(i)$ jest *i*-tą próbką w *k*-tym oknie sygnału. Wektor *E* zawiera wartości opisujące obwiednię sygnału dźwiękowego.

Jeśli wektor reprezentujący wzorzec szumu zostanie oznaczony przez *N*, zaś jego kolejne składowe ponumerowane przez indeks *i*, wówczas funkcję wyznaczającą wartość progową można wyrazić takim wzorem:

$$T = \max_i(|N(i) * snr|)$$

T jest skalarem wyznaczanym jeden raz, po pobraniu wzorca szumu danego otoczenia. Współczynnik *snr* (zdefiniowany wcześniej) reprezentuje zakładany stosunek sygnału do szumu i przyjmuje wartość większą od 1. Wartość progowa jest wartością maksymalną z wszystkich próbek wzorca szumu.

Dla każdego okna sygnału jest wyznaczana wartość maksymalna — jest ona wartością obwiedni w danym oknie. Następnie jest ona porównywana w wartością progową, będącą maksimum modułu wzorca szumu (z korektą *snr*). Gdy maksymalna wartość sygnału w oknie jest większa od wartości progowej, sygnał zawarty w oknie jest uznawany za sygnał głosowy i przesyłany do dalszego przetwarzania. W przeciwnym razie wartości sygnału w danym oknie są zerowane.

4.2. Średnia wartość bezwzględna sygnału w oknie

Zaletą poprzedniej metody jest jej prostota i wynikająca stąd szybkość działania. Ma ona jednak pewną wadę, polegającą na silnej zależności rezultatu od wartości chwilowych sygnału w oknie. Wszelkie przypadkowe duże wartości, nie należące do wypowiedzi, mogą wpłynąć na błędne rozpoznanie. W założeniach przyjęto, co prawda, że takie wartości nie wystąpią w sygnale, lecz w rzeczywistych sygnałach akustycznych zdarzają się one nader czę-

sto, co wpływa na błędną detekcję. Wady tej nie ma funkcja, która wyznacza wartość obwiedni jako średnią wartość bezwzględną sygnału w oknie.

Oznaczając numer okna przez *k*, indeks próbki w danym oknie przez *i* oraz liczbę próbek w oknie przez *I*, można wyrazić funkcję wyznaczającą obwiednię dźwięku następującym wzorem:

$$E(k) = \frac{\sum_i |s_k(i)|}{I}$$

E(k) jest skalarem reprezentującym wartość obwiedni dla *k*-tego okna sygnału. Oznaczając wektor reprezentujący wzorzec szumu przez *N*, zaś jego kolejne wartości przez *N(i)*, funkcję wyznaczającą wartość progową można wyrazić takim wzorem:

$$T = \frac{\sum_i |N(i) * snr|}{I}$$

Wartość progowa jest średnią arytmetyczną z modułu wszystkich próbek wzorca szumu (z korektą *snr*).

Dla każdego okna sygnału jest wyznaczana średnia arytmetyczna wartości bezwzględnych wszystkich próbek tego okna — jest ona wartością obwiedni w danym oknie. Następnie jest ona porównywana z wartością progową. Gdy wartość obwiedni jest większa od wartości progowej, sygnał zawarty w oknie jest uznawany za sygnał głosowy i przesyłany do dalszego przetwarzania. W przeciwnym razie zawartość okna jest zerowana.

4.3. Odchylenie standardowe sygnału w oknie

Następną funkcję dobrano opierając się na następującej obserwacji: gdy przebieg dźwięku nie zawiera sygnału głosowego, jego wartości są skupione w pobliżu zera (jeśli zgodnie z założeniem, sygnał nie zawiera składowej stałej); z kolei w przypadku zarejestrowanej wypowiedzi wartości przebiegu są bardziej rozproszone po obu stronach osi. Cechę tę dobrze reprezentuje odchylenie standardowe sygnału.

Oznaczając numer okna przez *k*, indeks próbki w danym oknie przez *i* oraz liczbę próbek w oknie przez *I* (w prezentowanym przykładzie *I=100*), odchylenie standardowe w *k*-tym oknie sygnału można wyznaczyć, korzystając z następującego wzoru:

$$E(k) = \sqrt{\frac{\sum_i (s_k(i) - \bar{s}_k)^2}{I-1}}, \qquad \bar{s}_k = \frac{\sum_i s_k(i)}{I}$$

Wielkość $\overline{s}_k$ jest średnią arytmetyczną wartości sygnału w *k*-tym oknie. *E(k)* jest skalarem reprezentującym wartość obwiedni dla *k*-tego okna sygnału. Oznaczając wektor reprezentujący wzorzec szumu przez *N*, zaś jego kolejne wartości przez *N(i)*, funkcję wyznaczającą wartość progową można wyrazić takim wzorem:

$$T = \sqrt{\frac{\sum_i \left[(N(i) - \overline{N}) * snr\right]^2}{I - 1}}, \qquad \overline{N} = \frac{\sum_i N(i)}{I}$$

Wielkość $\overline{N}$ jest średnią arytmetyczną wartości we wzorcu szumu. Wartość progowa jest odchyleniem standardowym składowych zarejestrowanego wzroca szumu (z korektą *snr*).

Dla każdego okna sygnału jest wyznaczane odchylenie standardowe próbek w tym oknie — jest ono wartością obwiedni w danym oknie. Następnie wartość obwiedni jest porównywana z wartością progową. Gdy wartość obwiedni jest większa od wartości progowej, sygnał zawarty w oknie jest uznawany za sygnał głosowy i przesyłany do dalszego przetwarzania. W przeciwnym razie zawartość okna jest zerowana.

4.4. Energia sygnału w oknie

Fala dźwiękowa jest nośnikiem energii akustycznej. Gdy mikrofon umieszczony w bezpośredniej bliskości mówcy rejestruje wypowiedź, energia zarejestrowanego sygnału jest stosunkowo duża w porównaniu z sytuacją, gdy w pobliżu mikrofonu nie ma żadnego źródła dźwięku. Badając energię sygnału akustycznego można wykryć bliskie źródło dźwięku. Jeśli zgodnie z założeniem źródłem tym jest mówca, będzie można stwierdzić, czy w danej chwili jest rejestrowana wypowiedź.

Oznaczając numer okna przez *k*, indeks próbki w danym oknie przez *i* oraz liczbę próbek w oknie przez *I*, energię sygnału w *k*-tym oknie można wyznaczyć, korzystając z następującego wzoru:

$$E(k) = \frac{\sum_i \left(s_k(i)\right)^2}{I}$$

W podobny sposób jest wyznaczana wartość progowa na podstawie wzorca szumu *N*:

$$T = \frac{snr * \sum_i \left(N(i)\right)^2}{I}$$

Dla każdego okna sygnału jest wyznaczana energia — jest ona wartością obwiedni w danym oknie. Następnie wartość ta jest porównywana z wartością progową. Gdy wartość obwiedni

jest większa od wartości progowej, sygnał zawarty w oknie jest uznawany za sygnał głosowy i przesyłany do dalszego przetwarzania. W przeciwnym razie zawartość okna jest zerowana.

4.5. Entropia sygnału w oknie

Jedną z charakterystycznych cech sygnału mowy jest jego uporządkowanie. W czasie wypowiadania poszczególnych fonemów narządy mowy są ułożone w określony sposób, dzięki czemu mówca wydaje dźwięk o uporządkowanym przebiegu; to pozwala słuchaczowi rozpoznać każdy fonem i zrozumieć wypowiedź. Z drugiej strony szum otoczenia i inne nakładające się na siebie zakłócenia zwykle mają charakter przypadkowy. Badając uporządkowanie sygnału, można wykryć, czy zawiera on sygnał głosowy.

Wielkością, która dobrze reprezentuje stopień uporządkowania sygnału, jest entropia. Im bardziej nieuporządkowany jest sygnał, tym większa jego entropia. Oznaczając numer okna przez k oraz indeks próbki w danym oknie przez i, entropię sygnału w k-tym oknie można wyznaczyć, korzystając z następującego wzoru[3]:

$$E(k) = -\sum_i s_k(i)^2 \log\left(s_k(i)^2\right)$$

W podobny sposób jest wyznaczana wartość progowa na podstawie wzorca szumu N (z korektą snr):

$$T = -\sum_i \left(N(i) * snr\right)^2 \log\left[\left(N(i) * snr\right)^2\right]$$

Znaki minus w powyższych wzorach pozwalają zastosować bezpośrednie porównanie obwiedni z wartością progową. Gdy wartość obwiedni jest większa od wartości progowej, sygnał zawarty w oknie jest uznawany za sygnał głosowy i przesyłany do dalszego przetwarzania. W przeciwnym razie zawartość okna jest zerowana.

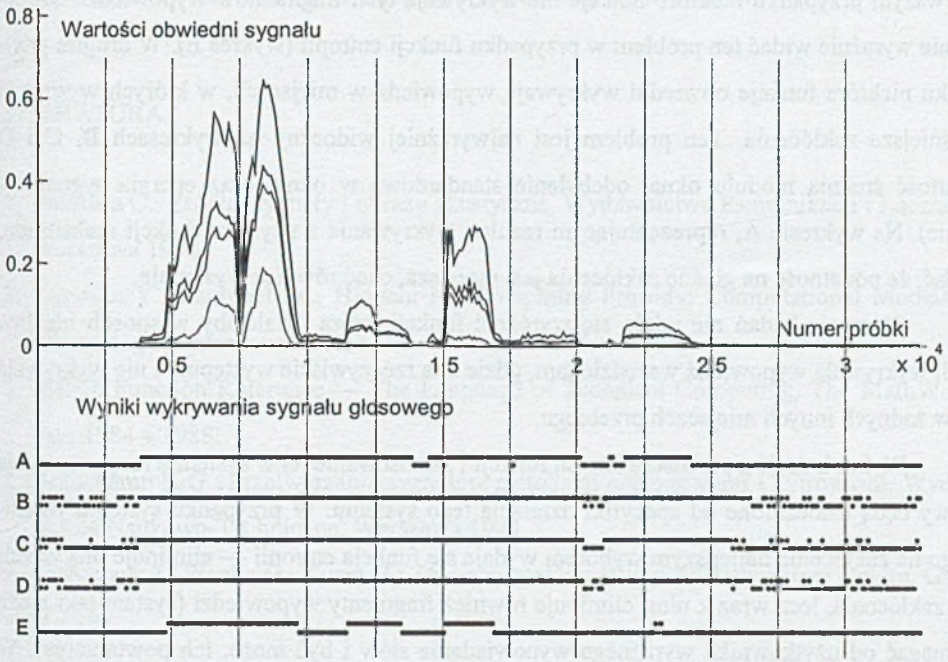
5. Wyniki badań

Rysunek 4 przedstawia przykładowe wyniki wykrywania sygnału głosowego, uzyskane za pomocą poszczególnych funkcji obwiedni. Materiałem wejściowym jest, przedstawiona wcześniej, przykładowa wypowiedź „Ala ma kota".

---

[3] Nieznormalizowana entropia Shannona. Dodatkowe założenie: 0·log(0)=0 [3].

Krytycznymi parametrami funkcjonowania modułu są współczynnik *snr* oraz szerokość okna służącego do obliczania wartości chwilowych obwiedni sygnału. W prezentowanym przykładzie doświadczalnie dobrano współczynnik *snr* (stanowiący mnożnik składowych zarejestrowanego wzorca szumu) o wartości 1,2. Przyjęto szerokość okna 100 próbek sygnału cyfrowego, co przy częstotliwości próbkowania sygnału równej 16 kHz odpowiada 6,25 ms.



Rys. 4. Rezultaty wykrywania sygnału głosowego za pomocą różnych funkcji obwiedni: A — wartość maksymalna okna; B — wartość średnia modułu okna; C — odchylenie standardowe w oknie; D — energia sygnału w oknie; E — entropia sygnału w oknie

Fig. 4. The results of detection of voice signal with various envelope functions: A — maximum value in a window; B — mean of absolute values in a window; C — standard deviation of signal in a window; D — energy of signal in a window; E — entropy in a window

W górnej części rysunku przedstawiono kształty obwiedni uzyskanych za pomocą poszczególnych funkcji (oprócz obwiedni uzyskanej za pomocą funkcji entropii, ponieważ uzyskane wartości wykraczały poza skalę rysunku). Ta część rysunku ma charakter poglądowy.

W dolnej części rysunku mieszczą się rezultaty wykrywania sygnału głosowego za pomocą poszczególnych funkcji obwiedni. Wykres sporządzony dla każdej funkcji reprezentuje dwa możliwe stany: górny — sygnał zakwalifikowany jako wypowiedź; dolny — sygnał zakwalifikowany jako brak wypowiedzi.

Na rysunku widać, że wszystkie funkcje poprawnie wykrywają wypowiedź w obszarach, w których jej występowanie nie budzi żadnych wątpliwości. Podobnie wszystkie wykrywają brak wypowiedzi tam, gdzie amplituda sygnału jest znikoma.

Więcej problemów sprawiają obszary, w których występowanie wypowiedzi nie jest ewidentne oraz takie, w których wypowiedzi nie ma, lecz występuje głośniejszy szum. W pierwszym przypadku niektóre funkcje nie wykrywają tych fragmentów wypowiedzi. Szczególnie wyraźnie widać ten problem w przypadku funkcji entropii (wykres E). W drugim przypadku niektóre funkcje obwiedni wykrywają wypowiedź w miejscach, w których występują głośniejsze zakłócenia. Ten problem jest najwyraźniej widoczny na wykresach B, C i D (wartość średnia modułu okna, odchylenie standardowe w oknie oraz energia sygnału w oknie). Na wykresie A, reprezentującym rezultat wykrywania z użyciem funkcji maksimum, widać, że podatność na głośne zakłócenia jest mniejsza, choć również występuje.

W czasie badań nie udało się wyróżnić funkcji, która działałaby w sposób idealny, czyli wykrywała wypowiedź wszędzie tam, gdzie ona rzeczywiście występuje i nie wykrywała jej w żadnych innych miejscach przebiegu.

Wybór którejś z przedstawionych funkcji i zastosowanie jej w systemie rozpoznawania mowy będą uzależnione od specyfiki działania tego systemu. W przypadku systemu wrażliwego na zakłócenia najlepszym wyborem wydaje się funkcja entropii — eliminuje ona wszelkie zakłócenia, lecz wraz z nimi eliminuje również fragmenty wypowiedzi (system taki może wymagać od użytkownika wyraźnego wypowiadania słów i być może, ich powtarzania). W przypadku systemu posiadającego dodatkowy mechanizm eliminacji zakłóceń (na przykład w postaci wzorców prostych zakłóceń, zawartych w bazie wiedzy) dobrym wyborem może się okazać funkcja maksimum. Ze względu na swą prostotę zapewnia ona szybkie działanie modułu wykrywania mowy i jednocześnie eliminuje większość obszarów, w których występuje sam szum.

6. Kierunki dalszych badań

W pracy nie zbadano możliwości zastosowania analizy częstotliwościowej ani falkowej do wykrywania sygnału głosowego. Pominięto ten obszar badań ze względu na budowę samego modułu, która uniemożliwiała proste zaimplementowanie tych metod. Integralną częścią modułu jest bramka, w której pojedyncza wartość obwiedni jest porównywana z warto-

ścią progową, zaś w przypadku metod częstotliwościowych i falkowych liczba uzyskiwanych danych uniemożliwia tego typu proste porównania. Jedynie całkowite przebudowanie struktury modułu umożliwiłoby zastosowanie transformaty Fouriera czy dekompozycji falkowej do wykrywania głosu. Ten kierunek badań wydaje się obiecujący.

Sprawdzenia wymaga również wpływ doboru wartości współczynnika korygującego *snr* na rezultaty wykrywania w różnych typach otoczenia akustycznego.

LITERATURA

1. Basztura C.: Źródła, sygnały i obrazy akustyczne, Wydawnictwa Komunikacji i Łączności, Warszawa 1988.
2. Sagisaka Y., Campbell Y., Higuchi N.: Computing Prosody: Computational Models for Processing Spontaneous Speech.
3. Matlab Function Reference — The Language of Technical Computing, The MathWorks, Inc., 1984 – 1988.
4. Beauchamp K.G.: Przetwarzanie sygnałów metodami analogowymi i cyfrowymi. Wydawnictwa Naukowo-Techniczne, Warszawa 1978.
5. Izydorczyk J., Płonka G., Tyma G.: Teoria sygnałów: Wstęp. Wydawnictwo Helion, Gliwice 1999.

Recenzent: Dr hab.inż. Zdzisław DUDA
Prof. Politechniki Śląskiej

Wpłynęło do Redakcji dnia 04 lipca 2002 r.

Abstract

This work describes an example of voice detection module. The purpose of this module is to support speech recognition system, by preprocessing recorded audio signal and eliminating parts of it, which do not contain voice signal. The voice detection module can make use of several envelope functions. In this work five envelope functions has been examined:

- Maximum value in a window.
- Mean of absolute values in a window.
- Standard deviation of signal in a window.
- Energy of signal in a window.
- Entropy in a window.

In last part of the work the results of analysis of an examplary voice signal have been presented. The signal has been recorded with accompanying ambient noise and disturbances in order to evaluate the functionality of module in conditions approximate to real. The disturbances have affected functionality, giving results inconsistent with expectations — the module has detected voice signal in some points in time, where there have have been no voice recorded, only noise and disturbances. The results of testing of described module in current form are not satisfactory enough for use it as a speech recognition support. A few suggestions have been included on how it's efficiency could be improved.