POZNAN UNIVERSITY OF TECHNOLOGY ACADEMIC JOURNALS
No 88 Electrical Engineering 2016

Artur ŚLIWIŃSKI*
Krzysztof TOMCZEWSKI*

BADANIA MOŻLIWOŚCI ROZPOZNAWANIA MOWY W AUTONOMICZNYCH SYSTEMACH STEROWANIA

W artykule omówiono zagadnienia dotyczące procesu rozpoznawania mowy w odniesieniu do układów sterowania. Projektowany system przewidziany jest dla prostych platform sprzętowych. W ramach pracy zastosowano do analizy szybką transformatę Fouriera FFT w celu utworzenia identyfikatorów słów. Określono czasy analizy sygnałów. Przeprowadzono wstępne testy opracowanego oprogramowania dla kilku różnych słów wypowiadanych przez osoby różniące się płcią oraz wiekiem. Uzyskano rozpoznawalność około 80% przy czasie obliczeń o połowę krótszym niż czas wymawiania komend. Krótki czas obliczeń pozwala na stosowanie opracowanego oprogramowania w systemach działających w czasie rzeczywistym np. na platformie Raspberry PI z procesorem 700 MHz.

SŁOWA KLUCZOWE: sterowanie, rozpoznawanie mowy, FFT, szybka transformata Fouriera

1. WSTĘP

Rozpoznawaniem komend głosowych nauka zajmuj się od wielu lat. Opracowano szereg różnych algorytmów pozwalających z wysokim prawdopodobieństwem rozpoznawać pojedyncze słowa lub ciągi słów. Nie powstał jednak dotychczas żaden algorytm, który byłby zdolny rozpoznawać tekst mówiony bezbłędnie. Na fakt ten wpływa wiele czynników takich, jak barwa tonu, intonacja, akcent, prędkość wypowiadania słów i cechy charakterystyczne języka. Ta różnorodność zmusza do tworzenia skomplikowanych algorytmów, co wpływa niekorzystnie na czas rozpoznawania słów. Skutkiem tego jest konieczność stosowania do obliczeń wydajnych systemów mikroprocesorowych. Obecnie istnieją rozwiązania pozwalające skutecznie rozpoznać tekst mówiony, niestety wymagają one zazwyczaj bezpośredniego połączenia układu sterowania z Internetem, gdzie znajdują się bazy słów oraz zasoby sprzętowe pozwalające przetworzyć wypowiedzianą komendę w krótkim czasie. Dotychczas najlepsze efekty

* Politechnika Opolska.

uzyskiwane są przy rozpoznawaniu słów w języku angielskim. Dla innych języków efekty są znacznie gorsze. Programy te w większości nie sprawdzają się przy rozpoznawaniu słów wymawianych po polsku.

Sterowanie głosowe ma potencjalnie szerokie zastosowanie, np. w miejscach, gdzie nie ma możliwości korzystania z interfejsów standardowych, w przypadku osób niepełnosprawnych, ze względu na komfort sterowania w systemach takich, jak „inteligentne domy.

Zagadnienie analizy i rozpoznawania tekstu mówionego komplikuje się dodatkowo w sytuacjach, gdy program ma działać na prostych, tanich platformach sprzętowych, w układach sterowania czasu rzeczywistego. W takim przypadku konieczne jest znalezienie kompromisu pomiędzy skutecznością rozpoznawania komend głosowych i czasem analizy sygnału.

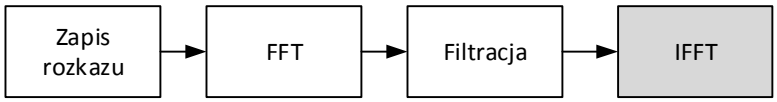
W typowych układach sterowania, np. w systemach inteligentnego zarządzania budynkiem, czas analizy słów powinien być na tyle krótki, żeby sterowanie odbywało się w sposób komfortowy, czyli zbliżony do normalnej konwersacji.

Dodatkowym utrudnieniem w rozpoznawaniu głosu są szumy i zakłócenia wnoszone przez układy rejestrujące dźwięk i tło emitowane przez otoczenie. Pierwszym etapem obróbki danych jest usunięcie z zarejestrowanego przebiegu sygnałów stanowiących szumy i sygnałów o częstotliwościach wychodzących poza pasmo, w którym zawiera się mowa ludzka. W celu eliminacji części zakłóceń można zastosować wstępne filtrowanie sygnału. Stosuje się do tego celu najczęściej filtry dolnoprzepustowe lub środkowoprzepustowe [1]. Mogą to być analogowe filtry sprzętowe lub filtry implementowane w programie sterującym.

Przedstawione w niniejszej pracy badania mają na celu określenie wpływu filtrowania i częstotliwości próbkowania sygnału dźwiękowego na skuteczność rozpoznawania komend oraz czas analizy sygnału przy zastosowaniu Szybkiej Transformaty Fouriera (FFT), w celu określenia możliwości wykorzystania do analizy dźwięku w układach sterownia prostych platform sprzętowych takich, jak np. Raspberry Pi.

2. FILTRACJA SYGNAŁU

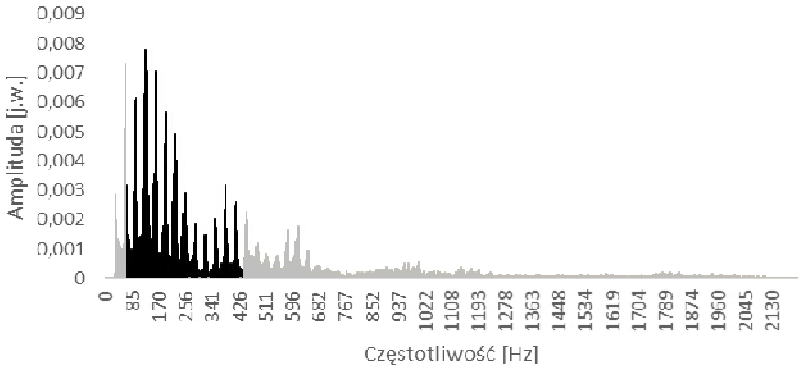
Sygnał do układu sterowania pobierany jest poprzez przetwornik analogowo-cyfrowy i zapisywany w postaci cyfrowej w pamięci modułu sterującego. Zarejestrowane próbki sygnałów dźwiękowych zostają poddane filtracji ze składowych o częstotliwościach znajdujących się poza pasmem emitowanym przez aparat mowy człowieka, czyli z informacji nieistotnych dla rozpoznawania komend. W ramach wstępnych badań filtracja została wykonana zgodnie ze schematem blokowym pokazanym na rys. 1.



Rys. 1. Schemat blokowy algorytmu filtracji sygnału rozkazu dźwiękowego

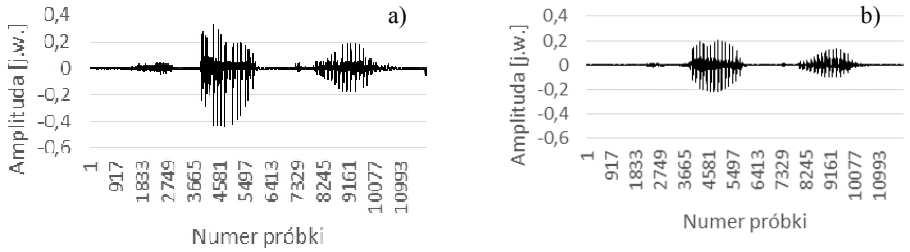
Do filtracji użyto filtr numeryczny środkowoprzepustowy oparty o transformatę Fouriera. Na spektrum widmowym w dziedzinie częstotliwości uzyskanym poprzez Szybką Transformatę Fouriera nałożono okno eliminujące sygnały wychodzące poza zakres częstotliwości istotnych do analizy oraz sygnały o bardzo małej amplitudzie.

Na rysunku 2 przedstawiono widmo sygnału dźwiękowego z zaznaczonym zakresem częstotliwości filtracji.



Rys. 2. Widmo sygnału uzyskane po transformacie Fouriera z zaznaczonym pasmem wydzielonym po filtracji

Usunięcie z przebiegów składowych niezawierających informacji istotnych z punktu widzenia rozpoznawania mowy upraszcza analizę sygnału, poprawia jej jakość oraz przyspiesza obliczenia [2]. Następnie obliczono współczynniki korelacji pomiędzy analizowanymi słowami wypowiedzianymi przez różne osoby. W przypadku braku filtracji wartości korelacji dla słów różnych były wyższe, a dla jednakowych niższe, niż dla sygnałów po filtracji. Dodatkowo usunięto z widma prążki o bardzo małej amplitudzie. Poziom odcięcia szumów ustalono porównując uzyskiwane wartości korelacji między jednakowymi i różnymi słowami wypowiedzianymi przez różne osoby. Na rysunku 3 przedstawiono przykładowy przebieg zarejestrowany przy wypowiedzeniu słowa „światło" (rys. 3a) i tego samego sygnału po filtracji wykonanej zgodnie z algorytmem pokazanym na rys. 1 (rys. 3b). Wadą tego rozwiązania jest stosunkowo długi czas trwania analizy wynikający z wielkości analizowanej próbki.



Rys. 3. Przebieg zarejestrowany przy podaniu komendy „światło” (a) i ten sam przebieg po filtracji (b)

Zaznaczone na szaro dwa końcowe etapy algorytmu pokazanego na rys. 1 są opcjonalne, gdyż służyły do ustalenia zakresów filtracji, a do dalszej analizy częstotliwościowej nie jest konieczne odtwarzanie sygnału w dziedzinie czasu.

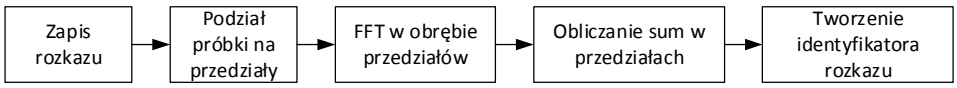
3. ALGORYTM ROZPOZNAWANIA KOMEND

Jako kryterium oceny zbieżności komend ze wzorcem przyjęto uzyskiwaną wartość współczynnika korelacji, obliczanego zgodnie z (1).

$$r_{xy} = \frac{\sum (X_i - \overline{X}) \cdot (Y_i - \overline{Y})}{\sqrt{\sum (X_i - \overline{X})^2 \cdot (Y_i - \overline{Y})^2}} \tag{1}$$

gdzie: X_i, Y_i – wartości z wzorca i analizowanego sygnału, $\overline{X}, \overline{Y}$ – wartości średnie z wzorca i analizowanego sygnału.

W celu przyspieszenia identyfikacji rozkazu dla każdego słowa utworzono identyfikator, a następnie przeprowadzono obliczenia współczynnika korelacji. W tym celu wykonano podział wzorca i analizowanego rozkazu na stałą liczbę przedziałów czasowych. Dla każdego z ustalonych przedziałów oddzielnie wykonano szybką transformatę Fouriera i filtrowanie. Na rys. 4 przedstawiono algorytm zastosowany do tworzenia identyfikatorów wzorców i analizowanych rozkazów. Jest on modyfikacją i rozwinięciem algorytmu zastosowanego do wstępnej analizy przy ustalaniu zakresów filtracji.



Rys. 4. Schemat blokowy algorytmu generacji identyfikatora rozkazu

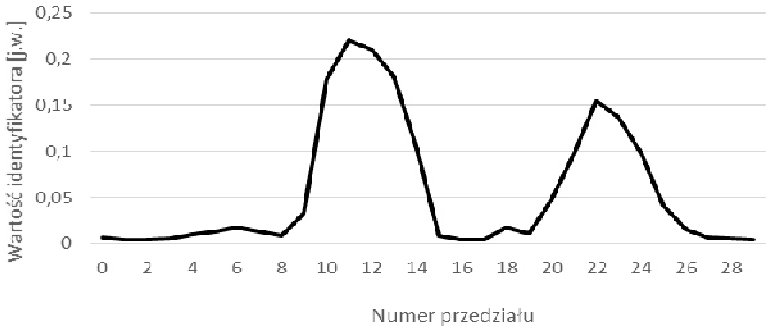
Pierwszym etapem po zarejestrowaniu rozkazu jest odcięcie fragmentów nagrania ciszy przed słowem oraz po słowie w celu pozostawienia tylko sygnału

do analizy. Następnym krokiem jest podział uzyskanego sygnału na N przedziałów. Dzięki temu uzyskane identyfikatory, charakteryzujące analizowane rozkazy, są zawsze tej samej wielkości. Umożliwia to porównywanie komend wypowiadanych z różną szybkością.

Wstępnie próbki dzielono na różne liczby przedziałów i obliczano współczynniki korelacji. Na tej podstawie ustalono najmniejszą liczbę przedziałów, dla jakiej nie następowało wyraźne pogorszenie wartości korelacji.

Do badań w ramach tej pracy przyjęto podział rozkazów na 30 jednakowych przedziałów czasowych.

Następnie, dla widma w każdym utworzonym w ten sposób przedziale, wykonano analizę widmową FFT oraz filtrację w celu poprawy jakości sygnału. Ostatnim krokiem realizacji algorytmu jest obliczenie sum amplitud prążków dla każdego przedziału i utworzenie identyfikatora rozkazu złożonego z 30 wartości. Przykład uzyskanego w ten sposób identyfikatora słowa „światło" przedstawiono w formie wykresu na rys. 5.



Rys. 5. Identyfikator słowa „światło"

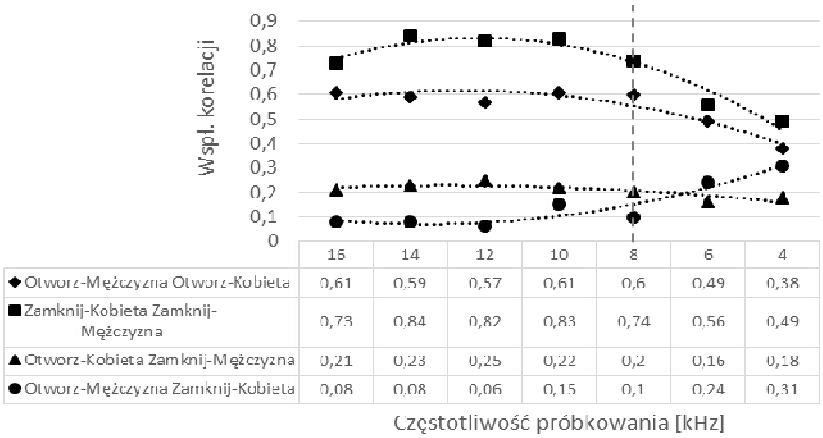
Tak wygenerowane identyfikatory wykorzystano następnie do porównania rejestrowanych rozkazów z wzorcami poprzez wyznaczanie wartości współczynników korelacji.

4. WPŁYW CZĘSTOTLIWOŚCI PRÓBKOWANIA NA WARTOŚCI WSPÓŁCZYNNIKA KORELACJI

Jednym z czynników wpływających w istotny sposób na czas analizy rozkazu jest częstotliwość próbkowania zarejestrowanego sygnału. Im jest ona wyższa, tym lepsza jest jakość nagrania. Z drugiej strony, im większa liczba zarejestrowanych próbek, tym bardziej czasochłonna jest analiza rozkazu. Dla systemów czasu rzeczywistego istotne jest określenie możliwie niskiej częstotliwości próbkowania, przy jakiej nie następuje jeszcze istotne

pogorszenie wyników. Wpływa to na minimalne wymagania dotyczące stosowanego sprzętu, a co za tym idzie koszt budowy systemu sterowania.

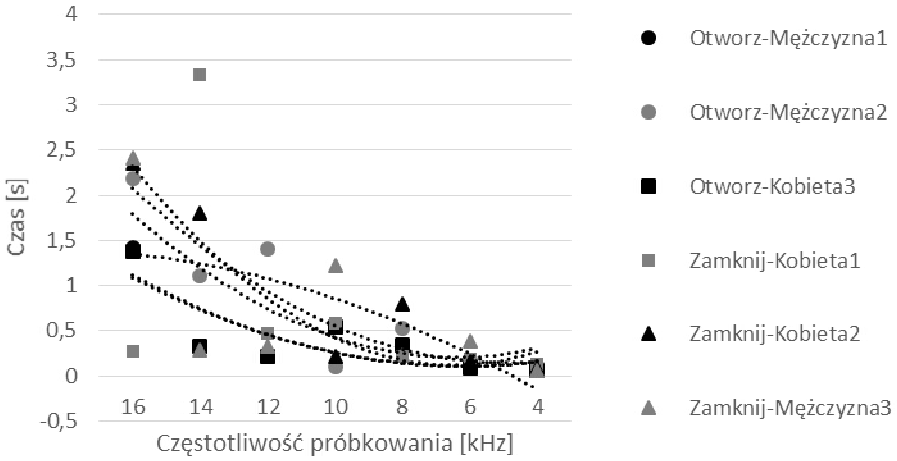
W celu określenia najniższej częstotliwości próbkowania wykonano serię obliczeń współczynnika korelacji dla testowego zestawu słów wymawianych przez różne osoby, zapisanych z różnymi częstotliwościami próbkowania. Przykładowe wyniki tych obliczeń zestawiono na rys. 6. Na uzyskanych zależnościach widoczne jest pogorszenie wyników, polegające na zmniejszaniu się odstępu pomiędzy wartościami współczynników korelacji dla jednakowych i różnych słów, przy częstotliwości poniżej 8 kHz. Natomiast na rys. 7 pokazano zależności czasu trwania obliczeń na platformie Raspberry PI wersja B z procesorem ARM1176JZF-S 700 MHz od częstotliwości próbkowania sygnałów. Czasy akceptowalne (poniżej 1 s) uzyskano przy częstotliwościach poniżej 10 kHz. Biorąc pod uwagę czas analizy i jakość uzyskiwanych wyników do dalszych badań przyjęto częstotliwość próbkowania 8 kH.



	16	14	12	10	8	6	4
Otwórz-Mężczyzna Otwórz-Kobieta	0,61	0,59	0,57	0,61	0,6	0,49	0,38
Zamknij-Kobieta Zamknij-Mężczyzna	0,73	0,84	0,82	0,83	0,74	0,56	0,49
Otwórz-Kobieta Zamknij-Mężczyzna	0,21	0,23	0,25	0,22	0,2	0,16	0,18
Otwórz-Mężczyzna Zamknij-Kobieta	0,08	0,08	0,06	0,15	0,1	0,24	0,31

Rys. 6. Zestawienie wyników obliczeń współczynników korelacji dla sygnałów zapisanych przy różnych częstotliwościach próbkowania

W kolejnym etapie badań wykonano próbę identyfikacji 2 słów wypowiedzianych przez 4 różne osoby. Wyniki tych badań pokazano na rys. 8 – 10. Współczynniki korelacji obliczono dla wszystkich konfiguracji tych słów. Jako kryterium uznania słów za jednakowe przyjęto wartość korelacji większą niż 0,5. Pola jasne w tabelach oznaczają słowa zidentyfikowane jako jednakowe, natomiast ciemne jako różne. Na tej podstawie określono prawdopodobieństwo prawidłowego rozpoznania słów, jako jednakowe lub różne.

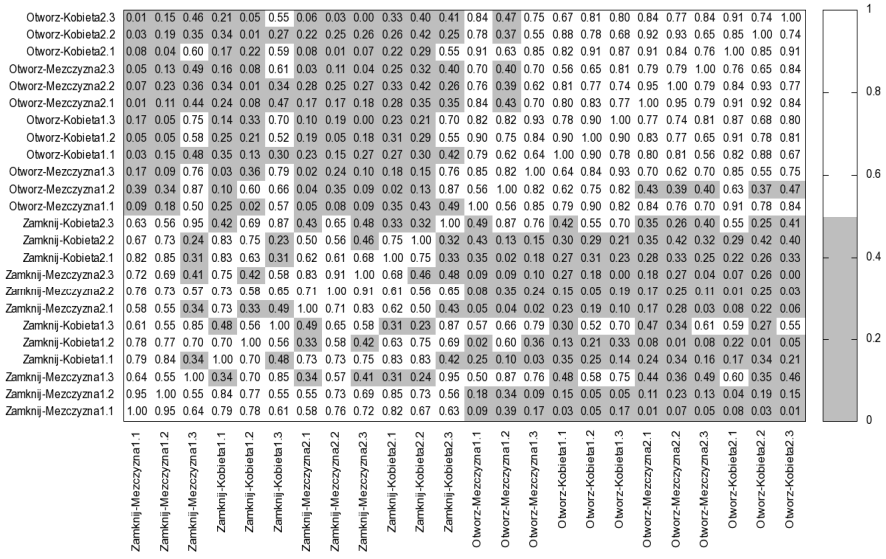


Rys. 7. Zestawienie czasów obliczeń współczynników korelacji dla sygnałów zapisanych z różnymi częstotliwościami próbkowania

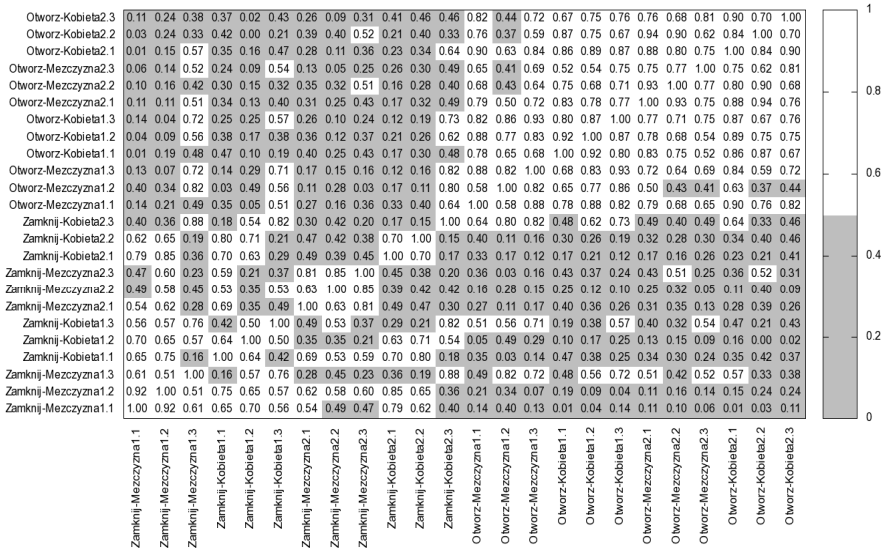
Tabela 1. Zestawienie czasów obliczeń współczynników korelacji dla sygnałów zapisanych z różnymi częstotliwościami próbkowania

	16	**14**	**12**	**10**	**8**	**6**	**4**
Otworz-Mężczyzna1	1,425364	0,331386	0,216284	0,539068	0,338798	0,083846	0,056185
Otworz-Mężczyzna2	2,181367	1,114148	1,404466	0,106379	0,529953	0,170077	0,121588
Otworz-Kobieta3	1,382763	0,330928	0,215729	0,538717	0,339484	0,084006	0,056168
Zamknij-Kobieta1	0,274501	3,329846	0,475465	0,588344	0,212229	0,182635	0,12497
Zamknij-Kobieta2	2,355828	1,800065	0,267737	0,212254	0,79172	0,156954	0,084506
Zamknij-Mężczyzna3	2,406093	0,282972	0,333432	1,218523	0,234863	0,385085	0,059303

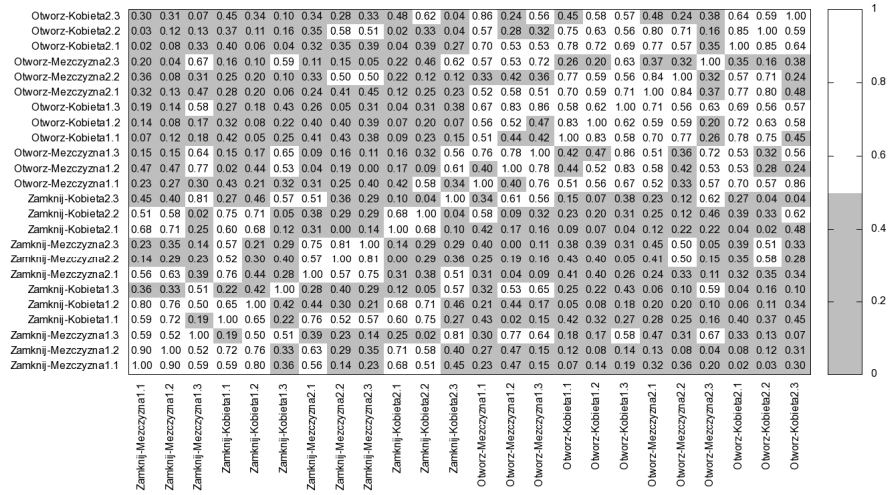
Rezultaty uzyskane dla trzech częstotliwości próbkowania zestawiono na rys. 11. Widoczna jest na nim wyraźna zależność prawdopodobieństwa rozpoznania rozkazów od częstotliwości próbkowania (dla jednakowych słów). W przypadku porównywania słów różnych zależność ta jest niewielka.



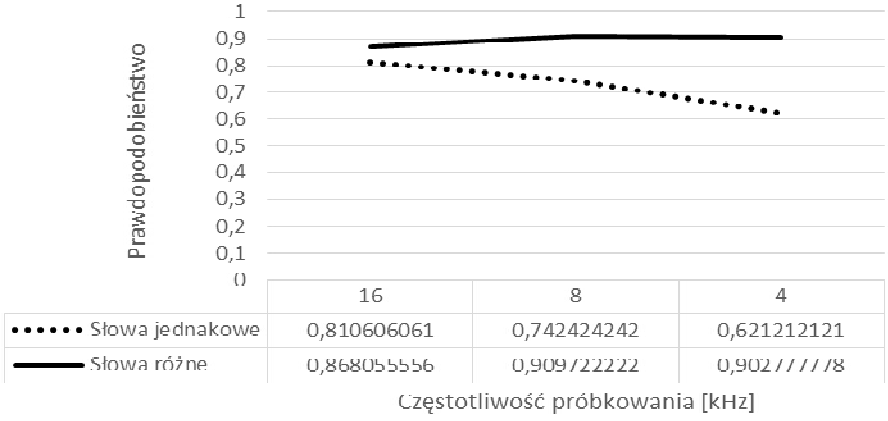
Rys. 8. Wartości korelacja dla częstotliwości próbkowania 16 kHz
(format opisu w tabeli: rozkaz-płeć.osoba.nr próbki)



Rys. 9. Wartości korelacja dla częstotliwości próbkowania 8 kHz
(format opisu w tabeli: rozkaz-płeć.osoba.nr próbki)



Rys. 10. Wartości korelacja dla częstotliwości próbkowania 4 kHz
(format opisu w tabeli: rozkaz-płeć.osoba.nr próbki)



Rys. 11. Zależność prawdopodobieństwa prawidłowej identyfikacji słów
od częstotliwości próbkowania

4. PODSUMOWANIE

Przeprowadzone wstępne analizy możliwości wykorzystania platformy Raspberry PI do analizy rozkazów słownych pokazała, że w przypadku tej platformy konieczne jest ograniczenie częstotliwości próbkowania rejestrowanych sygnałów do ok. 8 kHz. Dla takiej częstotliwości nie następuje jeszcze istotne pogorszenie skuteczności zastosowanego algorytmu, przy

zachowaniu akceptowalnych czasów obliczeń dla systemów pracujących w czasie rzeczywistym. Ze względu na 80% rozpoznawalność słów jednakowych konieczne jest rozszerzenie tej metody o dodatkowe kryteria oceny. Dalsze prace zostaną ukierunkowane na poprawę współczynnika rozpoznawalności słów.

5. LITERATURA

[1] Walendowski P., Zastosowanie sieci neuronowych typu SVM do rozpoznawania mowy, Praca doktorska, Politechnika Wrocławska, 2008.
[2] Igarashi T., Hughes J. F., Voice as Sound: Using Non-verbal Voice Input for Interactive Control, Computer Science Department, 2001.

AUTONOMUS CONTROL SYSTEMS SPEECH RECOGNITION POSSIBILITIES RESEARCH

The paper presents issues related to the process of speech recognition in control systems. The system to be designed is dedicated for simple hardware platforms that do not have high computing power. In order to create word identifiers, Fast Fourier Transformation (FFT) was used. The project specified signal analysis time, after which, preliminary software tests were carried out for several different words pronounced by people of various gender and age. The result was voice recognition at the level of approximately 80%, with calculation time being half of command pronouncing time. Due to short calculation time, the software may be used in systems working in real time, e.g. on 700 MHz processor Raspberry PI platform.

(Received: 14. 02. 2016, revised: 8. 03. 2016)