1. Ewelina MAJDA-ZDANCEWICZ, 2. Emilia GABRIELCZYK

Wojskowa Akademia Techniczna, Wydział Elektroniki, ul. Gen. Sylwestra Kaliskiego 2
ORCID: 1. 0000-0001-5435-6246, 2. 0000-0003-3953-9057

doi:10.15199/48.2021.10.25

Zastosowanie technik przetwarzania sygnału mowy w celu obiektywnej oceny wysiłku głosowego

Streszczenie. *Analiza cech akustycznych ludzkiego głosu ma wiele celów i może być rozpatrywana z różnych perspektyw. Z punktu widzenia nadmiernej eksploatacji ludzkiego głosu spotykamy się z pojęciem wysiłek głosowy. Celem przeprowadzonych badań jest analiza przydatności wybranych parametrów sygnału mowy w celu obiektywnej oceny wysiłku głosowego. W tym celu należy skupić się przede wszystkim na ocenie wydolności krtani. To ten narząd dostarcza nam największej ilości informacji. W ramach badań wyekstrahowano szereg parametrów w dziedzinie czasu oraz częstotliwości. Uzyskane wyniki pozwalają na stwierdzenie, iż istnieje szereg paramentów, które pozwalają na detekcję zmęczenia mówcy.*

Abstract*. The analysis of the acoustic features of the human voice has many has many objectives and can be considered from different perspectives. The aim of the research is to analyze the usefulness of selected parameters of the speech signal in order to objectively assess the voice effort. For this purpose the research should focus primarily on the assessment of the laryngeal capacity. It is this organ that provides us with the greatest amount of information. As part of the research, a number of parameters in the time and frequency domains were extracted. The results have shown that there are a number of parameters which allow the detection of speaker fatigue.. (**The application of voice signal processing to objective vocal fatigue assessment**).*

Słowa kluczowe: zmęczenie głosu, analiza akustyczna głosu, ekstrakcja cech.
Keywords: vocal fatigue, acoustic analysis of speech signal, feature extraction.

Wstęp

Analiza cech akustycznych ludzkiego głosu ma wiele celów i może być rozpatrywana z różnych perspektyw. Z punktu widzenia nadmiernej eksploatacji ludzkiego głosu spotykamy się z pojęciem wysiłek głosowy. Eksploatacja taka może być rozpatrywana na wielu płaszczyznach. Najłatwiej jest ją zaobserwować u ludzi wykorzystujących własny głos jako podstawowe narzędzie pracy. Przykładami takich zawodów są lektorzy, radiowcy, piosenkarze, a także wykładowcy akademiccy. Stan ten określany jest zwyczajowo jako *zmęczenie głosowe* (ang. *vocal fatigue*). Mogą mu towarzyszyć pewne objawy fizjologiczne tj. *odczuwanie suchości w gardle*, *kaszel*, *zmiana barwy głosu* czy też *chrypka*. Stan zmęczenia głosu jest chwilowy, tak też objawy mu towarzyszące powinny zaniknąć. Przy ciągłym przemęczaniu głosu stan wcześniej nazwany jako chwilowy może zmienić się w stan patologiczny, który zaburza prawidłową prace fałd głosowych. Wszelkiego rodzaju anomalie w barwie głosu, które utrzymują się przez dłuższy czas nazywane są *dysfonią* i są wynikiem utrzymującego się stanu patologicznego lub częstego obciążania głosu.

W celu dokonania obiektywnej oceny wysiłku głosowego należy skupić się przede wszystkim na ocenie wydolności krtani, ponieważ to ten narząd w głównej mierze dostarcza nam informacji niezbędnej do diagnostyki zmęczenia głosu. Obiektywna ocena identyfikacji tego rodzaju stanu możliwa jest przy wykorzystaniu metod cyfrowego przetwarzania sygnału mowy.

Baza danych

W pracy wykorzystano nagrania samogłoski „a” o przedłużonej fonacji. Rejestracji dokonano w dwóch trybach: przed oraz po wykonaniu przez badanego tzw. *próby obciążeniowej głosu*. Ciągłe samogłoski, pozbawione artefaktów językowych są wygodnym materiałem do analizy akustycznej głosu. Najczęściej wybiera się głoskę „a”, ponieważ według wielu badaczy najbardziej nadaje się ona do analizy tonu krtaniowego [1]. W założeniu pierwsze zmiany jakie można zaobserwować i nazwać je zmęczeniem głosowym występują po co najmniej godzinnym śpiewie czy też głośnym czytaniu. W ramach przeprowadzonych badań czas ten wydłużono do 2 godzin, aby mieć pewność, że wystąpią zauważalne zmiany.

Materiał badawczy stanowiła pojedyncza próba nagrania zarejestrowana z częstotliwością próbkowania 44,1 kHz i rozdzielczością 16 bitów. Odległość między dyktafonem, a ustami badanego wynosiła ok. 5 cm.

Parametryzacja sygnału mowy

Ekstrakcja parametrów sygnału mowy wymaga przeglądu metod cyfrowego przetwarzania sygnału w odniesieniu do realizowanego zadania. Metody analizy sygnału mowy są silnie uzależnione od struktury sygnału mowy, która jest zdeterminowana przez proces jego wytwarzania. W związku z tym, inaczej bada się fragment tekstu czytanego, inaczej zaś wyizolowanej głoski dźwięcznej. W realizowanych eksperymentach analizie podlega krótki odcinek czasowy, w którym wyizolowano głoskę dźwięczną „a” o przedłużonej fonacji. W celu wyekstrahowania parametrów umożliwiających obiektywną ocenę wysiłku głosowego wykorzystano oprogramowanie *Matlab*. W analizie krótkookresowej wyróżnia się trzy fundamentalne parametry: *częstotliwość podstawową* F_0 (okres podstawowy T_0) oraz służące do opisu jej zmienności parametry z grupy *Jitter* i *Shimmer*, a także parametry będące ich pochodnymi m. in. *RAP*, *APQ3* czy *APQ5*. Obserwacja zmienności częstotliwości podstawowej tonu krtaniowego oraz zmian występujących w głośności sygnału mowy pozwalają na wychwycenie nieprawidłowości w pracy krtani, a tym samym możliwe jest wykrywanie zmian w głosie spowodowanych zmęczeniem głosowym. Znając wartości *częstotliwości podstawowej* F_0 można w łatwy sposób obliczyć *długość okresu podstawowego* T_0, a także wyznaczyć parametry statystyczne tj. *odchylenie standardowe*, *wartość maksymalną* oraz *minimalną*, *wartość częstotliwości pomiędzy poszczególnymi sekundami trwania sygnału*. Dostarcza nam to dodatkowej informacji pozwalającej na bardziej szczegółową analizę zmian zachodzących w głosie [2, 3].

Parametry z grupy *Jitter* definiowane są jako zmiana częstotliwości podstawowej tonu krtaniowego w kolejnych okresach pracy fałdów głosowych. *Jitter* (absolute) przedstawia średnią bezwzględną różnicę między dwoma okresami i wyrażany jest w μs [9]. *Jitter* (relative) przedstawia średnią bezwzględną różnicę pomiędzy dwoma kolejnymi okresami podzieloną przez średni okres.

Kolejnym parametrem z tej grupy jest *RAP*, który reprezentuje średnią bezwzględną różnicę danego okresu i średnią okresu z jej dwoma sąsiadami podzieloną przez okres średni. Zdefiniowano również parametry *PPQn* (ang. *pitch period perturbation quotient*) definiujące względną ocenę zmian krótko lub długo okresowych częstotliwości podstawowej w obrębie analizowanej próbki głosu, przy współczynniku wygładzania zdefiniowanym przez użytkownika. Szersze definicje wyżej wymienionych parametrów można szerzej odnaleźć w [2 - 5].

Parametry z grupy *Shimmer* definiowane są jako zmiany amplitudy fali dźwiękowej (sygnału mowy) w kolejnych cyklach pracy fałd głosowych. Parametr nazywany jako *Shimmer* przedstawia średnią bezwzględną różnicę między amplitudami dwóch kolejnych okresów podzielonych przez średnią amplitudę. Parametry z grupy *APQn* przedstawiają iloraz zakłóceń amplitudy w ciągu *n* okresów, czyli średnią bezwzględną różnicę między amplitudą okresu, a średnią amplitudą jego (*n*-1) sąsiadów podzieloną przez średnią amplitudę.

Parametry *Jitter*, *Shimmer* oraz F_0 są bezpośrednio związane z drganiami fałdów głosowych. Są to najczęściej używane parametry w analizie aparatu głosu [2, 3, 4].

Kolejnym parametrem jest *PVI* (ang. *pathology vibrato index*). Vibrato oznacza szybką oraz regularną fluktuację *częstotliwości podstawowej tonu krtaniowego* F_0, która pojawia się podczas przedłużonej fonacji samogłoski.

Ponadto w ramach procesu ekstrakcji cech wyekstrahowano również szereg widmowych deskryptorów:

- *spectralCentroid,*
- *spectralCrest,*
- *spectralDecrease,*
- *spectralFlatness,*
- *spectralKurtosis,*
- *spectralSkewness,*

Współczynnik szczytu definiuje się poprzez stosunek wartości maksymalnych w paśmie do średniej arytmetycznej widma energii. Jest to jeden z parametrów opisujących miarę płaskości widma sygnału. Przedstawiony przebieg jest ciągiem wartości współczynników szczytu wyznaczonych dla każdego okresu tonu podstawowego określonego na podstawie analizy widmowej sygnału.

Kolejnym analizowanym deskryptorem jest tzw. *nachylenie widma (ang. Spectral Decrease),* które definiuje spadek widma amplitudowego. Wartość tego parametru zdefiniowana jest w oparciu o poniższy wzór:

$$decrease = \frac{\sum_{k=b_1+1}^{b_2} \frac{s_k - s_{b_1}}{k-1}}{\sum_{k=b_1+1}^{b_2} s_k} \quad (1)$$

gdzie:s_k -jest wartością widmową próbki *k; b_1* i b_2 – są krawędziami pasma w przedziałach, dla których należy obliczyć spadek widma

W ramach artykułu przedstawiono wybrane wyniki w zakresie oceny wyżej zdefiniowanych deskryptorów. Szerszy opis wszystkich można znaleźć w [4, 5].

Analiza wyników

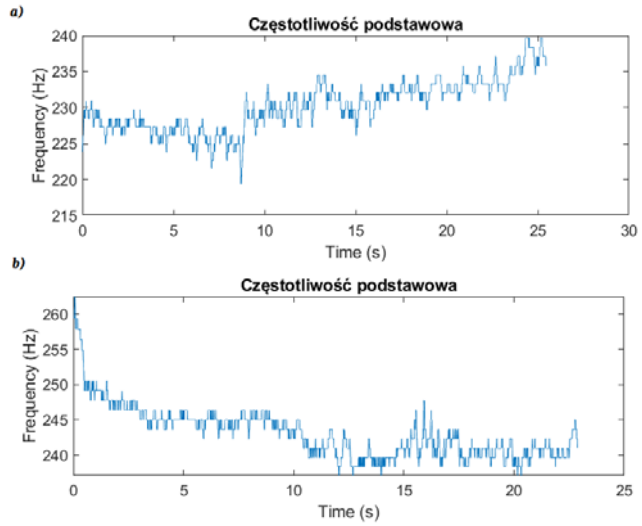
Wyekstrahowane deskryptory wraz z obliczonymi wartościami liczbowymi dla sygnału mowy przed i po próbie obciążeniowej głosu stanowią bazę do oceny przydatności ich zastosowania w celu detekcji zmęczenia głosu. Badania przeprowadzono w dwóch etapach. Pierwszym z nich było porównanie liczbowych wartości wybranych deskryptorów sygnału mowy przed i po dokonaniu próby obciążeniowej głosu zestawionych zbiorczo w tab.1. Drugą metodą zastosowaną przez autorów była analiza zmian wartości wybranych parametrów w czasie.

Tabela 1. Lista wyekstrahowanych cech sygnału mowy wraz z otrzymanymi wynikami liczbowymi

Grupa parametrów	Parametr	Wartość przed próbą	Wartość po próbie
Cechy czasowe	Maksymalny czas fonacji [s]	25,45	22,89
	Średnia wartość T_0 [ms]	4,37	4,33
	Odchylenie standardowe T_0 [ms]	0,29	0,28
Cechy często-tliwościowe	Średnia wartość F_0 [Hz]	229,0	241,7
	Odchylenie standardowe F_0 [Hz]	3,43	3,57
	Wartość minimalna F_0 [Hz]	218,71	231,02
	Wartość maksymalna F_0 [Hz]	239,31	252,41
	Wartość średnia F_0 po 2s [Hz]	228,17	250,87
	Wartość średnia F_0 między 2, a 4s [Hz]	227,64	246,07
	Wartość średnia F_0 między 4, a 6s [Hz]	226,60	244,94
	Wartość średnia F_0 między 6, a 8s [Hz]	225,48	244,80
	Wartość średnia F_0 między 8, a 10s [Hz]	226,94	244,64
	Wartość średnia F_0 między 10, a 12s [Hz]	229,70	241,68
	Wartość średnia F_0 między 12, a 14s [Hz]	231,66	239,49
	Wartość średnia F_0 między 14, a 16s [Hz]	229,81	240,98
	Wartość średnia F_0 między 16, a 18s [Hz]	232,06	241,10
	Wartość średnia F_0 między 18, a 20s [Hz]	232,57	240,48
	Jitter (relative) [%]	0,13	0,27
	Jitter (absolute) [μs]	24	49
	RAP [%]	0,08	0,18
	PPQ5 [%]	0,11	0,18
	PPQ11 [%]	0,14	0,21
	PVI	2,02	3,36
Cechy amplitudowe	Wartość maksymalna sygnału	0,63	0,73
	Wartość skuteczna sygnału	0,23	0,21
	Średnia wartość U_{pp}	0,86	0,83
	Odchylenie standardowe U_{pp}	0,07	0,13
	Shimmer [%]	1,90	2.48
	APQ3 [%]	1,10	1,45
	APQ5 [%]	1,16	1,42

Maksymalny czas fonacji jest to parametr wydolnościowy informujący o tym, jak długo badany może wypowiadać samogłoskę „a” o przedłużonej fonacji. Przed śpiewaniem jego wartość wynosiła 25,45 sekund, natomiast po 22,89 sekundy. Oznacza to, że maksymalny czas fonacji skrócił się o ok. 10% po przeprowadzeniu próby obciążeniowej w postaci 2-godzinnego śpiewania. Świadczy to o zmniejszonej wydolności głosu badanego. U osoby badanej wysiłek głosowy zmniejszył okres trwania cyklu pracy fałd głosowych T_0 tzn. zwiększyła się częstotliwość podstawowa tonu krtaniowego, jej wartość maksymalna i minimalna oraz jej odchylenie standardowe. Przed śpiewaniem rozkład utrzymuje się mniej więcej na stałym poziomie, po śpiewaniu możemy zaobserwować tendencję spadkową. Oznacza to, że po wysiłku głosowym nasza zdolność do utrzymania stałego poziomu wartości częstotliwości tonu krtaniowego maleje. W ramach realizowanych badań dokonano również oceny zmian

częstotliwości podstawowej w ramach zdefiniowanych odstępach czasowych. Obliczono zatem wartości średnie F_0 po każdych 2 sekundach trwania sygnału. Przebieg zmian wartości częstotliwości podstawowej podczas wypowiadania samogłoski „a” o przedłużonej fonacji zilustrowano na rys. 1.



Rys. 1. Przebieg częstotliwości podstawowej tonu krtaniowego w dziedzinie czasu dla samogłoski „a” o przedłużonej fonacji: a) przed próbą obciążeniową b) po próbie obciążeniowej

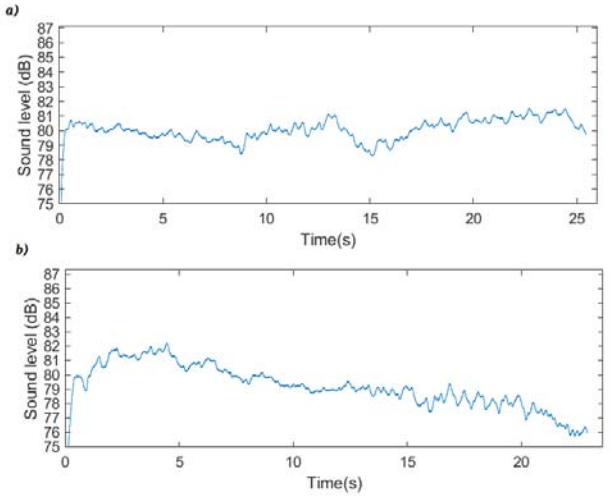
Przechodząc do analizy zmian amplitudy sygnału mowy w pierwszej kolejności należy wspomnieć o wartości maksymalnej oraz skutecznej sygnału głosowego. Wartość skuteczna sygnału pozostaje praktycznie na tym samym poziomie. Jednocześnie wartość maksymalna sygnału wzrosła. Najprawdopodobniej jest to spowodowane tzw. „nadwyrężeniem głosu”. W tej sytuacji badany czuje, że efektywność jego głosu zmalała, niemniej jednak próbuje nadrobić te różnice, czego konsekwencją może być chwilowy wzrost amplitudy głosu. Wnioski te częściowo potwierdzają przebiegi zmiany głośności dźwięku mowy w czasie – rys. 2. Zwróćmy uwagę, że w pierwszej próbie, czyli nagraniu przed śpiewaniem poziom natężenia głosu przez cały czas utrzymuje się mniej więcej na równym poziomie. W przypadku drugiej próby obserwuje się spadek wartości wraz ze zwiększaniem długości nagrania.

Wspominając o zmianach w zakresie amplitudy sygnału mowy warto przyjrzeć się parametrom z grupy *Shimmer* oraz *APQ*, które charakteryzują niestałość amplitudy sygnału mowy [6]. Im większe wartości reprezentują, tym większa niestabilność amplitudy głosu w analizowanych sygnałach. Różnica między wartościami wyznaczonymi dla próbki głosu zarejestrowanej przed i po obciążeniu głosowym wynosi ok. 30% dla parametrów tj. *Shimmer*, *APQ3*, *APQ11*. Dla *APQ5* wynosi ok. 22%, natomiast dla *APQ55* aż 66%. Zmiany te można uznać za znaczące.

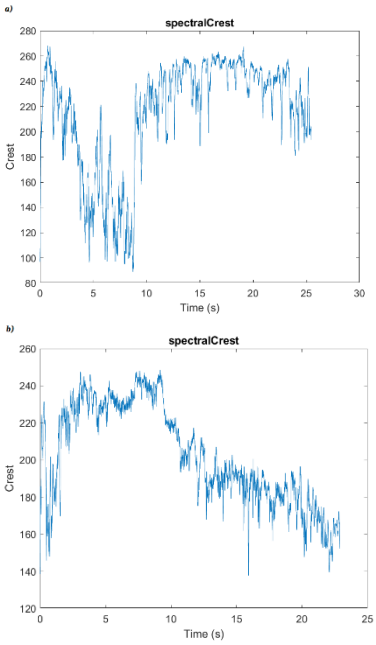
Warto zwrócić uwagę również na parametr *PVI* - wartości tego współczynnika mogą także informować np. o zmianach ciśnienia śródpiersiowego [7]. Zwiększona wartość świadczy o zaistniałych przemianach. Przypatrując się wyznaczonym wielkościom liczbowym, widać pojawienie się zmęczenia głosowego, jednak w stopniu umiarkowanym, nie wpływającym szkodliwie na głos.

W ramach eksperymentów przeanalizowano również przebieg wybranych parametrów widmowych w funkcji czasu dla zarejestrowanych sygnałów celem oceny zmian wartości danej cechy w funkcji czasu.

Pierwszym z analizowanych parametrów jest przebieg zmian wartości widmowego współczynnika szczytu (ang. *spectral crest*) – rys. 3.



Rys. 2. Przebiegi głośności dźwięku sygnału mowy w dziedzinie czasu dla samogłoski „a” o przedłużonej fonacji: a) przed próbą obciążeniową b) po próbie obciążeniowej

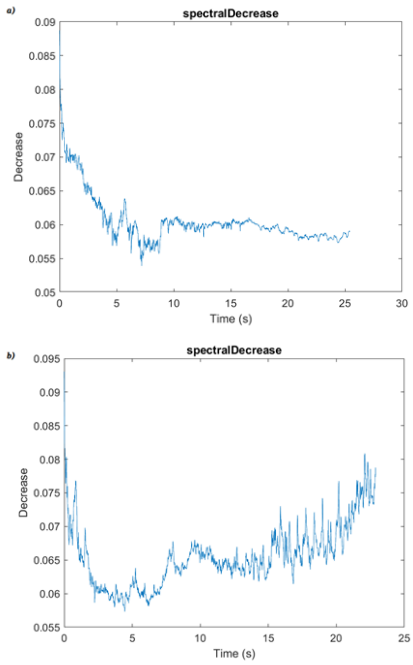


Rys. 3. Przebieg widmowego współczynnika szczytu w dziedzinie czasu dla samogłoski „a” o przedłużonej fonacji: a) przed próbą obciążeniową b) nagranie po próbie obciążeniowej

Przedstawiony przebieg jest ciągiem wartości współczynników szczytu wyznaczonych dla każdego okresu tonu podstawowego określonego na podstawie analizy widmowej sygnału. Fale dźwiękowe zwykle charakteryzują się wysokimi wartościami współczynników szczytu ze względu na swój zmienny w czasie charakter. Przyglądając się obu przebiegom widać znaczące różnice. Wartości otrzymane w przypadku sygnału mowy przed próbą obciążeniową są zdecydowanie większe. Ponadto pomiędzy 5, a 10 sekundą przebiegu możemy zaobserwować znaczny ich spadek, po czym następuje kolejny ich wzrost, i finalnie wartości te oscylują w granicach 240. W przypadku zmian wartości tego współczynnika dla głosu obarczonego próbą obciążeniową zmiany te mają nieco innych charakter. W pierwszej kolejności należy podkreślić fakt, że średnia wartość tego współczynnika zmalała. Maksymalne wartości dochodzą do 240. Ponadto spadek wartości grzbietów widmowych w tym

przypadku następuje przeciwnie niż w poprzednim przypadku w pierwszych sekundach jego trwania oraz jest znacznie krótszy. Po nim następuję skok wartości, a od ok. 10 sekundy następuje systematyczny spadek wartości, który odbywa się nieprzerwanie do końca czasu nagrania.

Kolejnym analizowanym deskryptorem jest tzw. nachylenie widma *(ang. Spectral Decrease)*. Przebiegi zmian wartości tego współczynnika przedstawiono na rys. 4. Przed śpiewaniem wartość *SpetralDecrease* utrzymuje się na mniej więcej równym poziomie. Dla drugiego wariantu widzimy znaczny spadek widma od ok. 5 sekundy trwania sygnału. Świadczy to o słabnącej jakości sygnału mowy, ponieważ krawędzie (zbocza) sygnału liczone są w miejscach, gdzie częstotliwość podstawowa tonu krtaniowego jest niższa od jej uśrednionej wartości. Jeżeli wartość częstotliwości podstawowej spada w znaczącym stopniu (tzn. możliwe są do zaobserwowania zmiany na rys.4.) możemy mówić o zmęczeniu głosowym.



Rys. 4. Przebieg zmian parametry *Spectral decrease* w dziedzinie czasu dla samogłoski „a" o przedłużonej fonacji: a) przed próbą obciążeniową b) po próbie obciążeniowej

Podsumowanie

Sygnał mowy niesie ze sobą nie tylko treść słowną, informacje o mówcy czy emocjach, ale również może zawierać informacje o kondycji narządów wewnętrznych człowieka. Obiektywna analiza sygnału mowy przy użyciu cyfrowych algorytmów przetwarzania sygnału mowy może pozwolić na wydobycie tych wszystkich informacji i wykorzystanie ich w celu oceny stanu głosu danego człowieka. Badania przeprowadzone przez autorów pracy wykazały, że istnieje szereg paramentów, które pozwalają na detekcję *zmęczenia mówcy*. Obserwacja zmienności częstotliwości podstawowej tonu krtaniowego oraz zmian występujących w głośności pozwalają wychwycić nieprawidłowości w pracy krtani, a tym samym można zaobserwować, jak zmienia się głos po wysiłku głosowym. Ocena skuteczności wybranych cech z uwzględnieniem zaprojektowanego klasyfikatora i po poszerzeniu bazy sygnałów głosowych będzie przeprowadzona podczas realizacji kolejnego etapu badań. Ponadto warto rozszerzyć badania na mowę spontaniczną i analizę zmian wybranych parametrów sygnału mowy w tym zakresie. Tego typu badania mogą okazać się przydatnym narzędziem pozwalającym określić stopień szkodliwości zawodu (u osób, którym podmiotem pracy jest głos), a przede wszystkim ułatwić wykrycie pojawiających się stanów patologicznych głosu.

Praca została sfinansowana przez Wojskową Akademię Techniczną w ramach projektu nr UGB 22-850.

***Autorzy**: dr inż. Ewelina Majda-Zdancewicz Wojskowa Akademia Techniczna, Wydział Elektroniki, ul. Gen. Sylwestra Kaliskiego 2, 00-908 Warszawa, E-mail: ewelina.majda@wat.edu.pl, inż. Emilia Gabrielczyk, Wojskowa Akademia Techniczna, Wydział, Elektroniki, ul. Gen. Sylwestra Kaliskiego 2, 00-908 Warszawa, E-mail: emilia.gabrielczyki@student.wat.edu.pl;*

LITERATURA

[1] E. Niebudek-Bogusz, Ocena parametrów analizy akustycznej głosu u kobiet zdrowych, *Otolaryngologia*, nr 3(1), s. 33-39, 2004

[2] J. Mekyska i wsp., Robust and complex approach of pathological speech signal analysis, *Neurocomputing*, vol. 167, 1, pp. 94-111, 2015

[3] Teixeira, João Paulo; Oliveira, Carla and Lopes, Carla, "Vocal Acoustic Analysis – Jitter, Shimmer and HNR Parameters", Procedia Technology – Elsevier, Vol. 9, pp 1112-1122, 2013

4] Y. Maryn i wsp., Acoustic measurement of overall voice quality: a meta-analysis, *The Journal of the Acoustical Society of America*, nr 126(5), s. 2619-34, 2009

[5] Peeters, Geoffroy & Giordano, Bruno & Susini, P. & Misdariis, Nicolas & Mcadams, Stephen. (2011). The Timbre Toolbox: Extracting audio descriptors from musical signals. *The Journal of the Acoustical Society of America*. 130. 2902-16. 10.1121/1.3642604.

[6] S. Hadjitodorov, P. Mitev, A computer system for acoustic analysis of pathological voices and laryngeal diseases screening, *Medical Engineering & Physics*, nr 24, 6, s. 419-429, 2002

[7]https://www.hospital.com.pl/index.php?option=com_attachments&task=download&id=4072