Barbara WILK

Politechnika Rzeszowska, Katedra Metrologii i Systemów Diagnostycznych

doi:10.15199/48.2015.11.70

Wyznaczanie wartości chwilowej częstotliwości podstawowej tonu krtaniowego za pomocą analizy falkowej sygnału mowy

Streszczenie. *Jednym z ważniejszych parametrów charakteryzujących źródło mowy dźwięcznej jest częstotliwość podstawowa tonu krtaniowego (F_0), która odpowiada częstotliwości podstawowej drgań fałd głosowych. W artykule zaprezentowano metodę wyznaczania wartości chwilowych tej częstotliwości za pomocą analizy falkowej sygnału mowy. Metodę zastosowano do badania wyizolowanych głosek w celu oceny zmienności częstotliwości podstawowej tonu krtaniowego.*

Abstract. *One of the basic parameters characterizing voiced speech is the fundamental vocal frequency (pitch, F_0), which corresponds to the rate of the vocal folds vibration. In this paper, a method based on wavelet analysis of speech signal for determining instantaneous fundamental frequency is presented. This method was applied to some isolated voiced vowels for assessment of pitch frequency variability. (**Determination of instantaneous vocal fundamental frequency using wavelet analysis of speech signal**).*

Słowa kluczowe: mowa dźwięczna, częstotliwość podstawowa tonu krtaniowego, transformata falkowa, kontur intonacyjny.
Keywords: voiced speech, fundamental vocal frequency, wavelet transform, pitch contour.

Wstęp

W powszechnie stosowanym modelu generowania mowy dźwięcznej przyjmuje się, że quasi-stacjonarny segment sygnału mowy jest wynikiem splotu sygnału okresowego (tzw. pobudzenia krtaniowego) oraz odpowiedzi impulsowej filtru liniowego, który reprezentuje charakterystykę amplitudowo-częstotliwościową traktu głosowego. W przypadku generowania mowy bezdźwięcznej pobudzenie ma charakter losowy.

W procesie wytwarzania mowy dźwięcznej biorą udział fałdy głosowe (zwane potocznie strunami głosowymi), których drgania modulują strumień powietrza wypływający z płuc do kanału głosowego. Fałdy głosowe stanowią wolny, ostry brzeg błony śluzowej warg głosowych, których sprężyste podłoże tworzą więzadła głosowe. Fałdy głosowe pełnią funkcję zaworu, który zamyka lub otwiera głośnię, gdy fałdy te zbliżają się do siebie lub oddalają pod wpływem zmian ciśnienia słupa powietrza w okolicy podgłośniowej. Jednocześnie zakończenia nerwowe zlokalizowane w tej okolicy reagują na zmiany ciśnienia i przenoszą bodźce do mózgu. Ośrodkowy układ nerwowy wysyła impulsy do nerwów krtaniowych, które koordynują pracę krtani.

Jednym z ważniejszych parametrów charakteryzujących źródło mowy dźwięcznej jest częstotliwość podstawowa tonu krtaniowego (oznaczana jako F_0), tj. częstotliwość podstawowa drgań fałd głosowych, która jest przede wszystkim funkcją masy, sprężystości (sztywności) oraz współczynnika napięcia warg głosowych [1]. Wartość częstotliwości F_0 zależy m. in. od płci, wieku oraz stanu emocjonalnego mówcy. Dla mężczyzn częstotliwość F_0 przyjmuje wartości z przedziału (80–480) Hz, zaś dla kobiet z przedziału (160–960) Hz (uwzględniono mowę i śpiew).

Krótkoczasowa zmienność częstotliwości podstawowej tonu krtaniowego w czasie wypowiedzi jest związana przede wszystkim z intonacją i stanem emocjonalnym mówcy. Postawę obiektywną mówca wyraża zazwyczaj poprzez intonację rosnąco-opadającą, zaś wyrazem postawy subiektywnej jest np. intonacja rosnąca w czasie zadawania pytania. Do oceny zmienności częstotliwości F_0 stosuje się różne parametry (np. Jitter, RAP, PPQ), których definicje podano np. w pracy [2].

Na podstawie analizy krótkoczasowej zmienności częstotliwości F_0 (np. w przypadku badania wyizolowanych głosek dźwięcznych o przedłużonej fonacji) można m. in.:
- wykryć anomalie struktury anatomicznej oraz określić przyczyny ograniczenia czynności ruchowej fałd głosowych, które występują w różnych stanach patologicznych (np. [2, 3, 4]),
- obiektywnie ocenić poprawę stanu głosu w czasie leczenia i rehabilitacji,
- ocenić stopień depresji pacjenta (np. [5])
- rozpoznawać emocje mówcy (np. [6]).

W niniejszym artykule krótko omówiono wybrane metody stosowane do wyznaczania częstotliwości podstawowej tonu krtaniowego oraz zaproponowano metodę do określania wartości chwilowej częstotliwości F_0, która bazuje na analizie falkowej sygnału mowy i nie wymaga segmentacji tego sygnału.

Metody stosowane do wyznaczania częstotliwości podstawowej tonu krtaniowego

Do wyznaczania wartości częstotliwości podstawowej tonu krtaniowego opracowano różne metody (np. [4, 7, 8, 9, 10]). Najczęściej stosowane są metody bazujące na:
- funkcji AMDF,
- funkcji autokorelacji,
- Cepstrum.

Wspomniane metody wyznaczają wartość F_0 na podstawie przetwarzania segmentu sygnału mowy, który zazwyczaj reprezentuje co najmniej kilka cykli pracy fałd głosowych. W praktyce segment sygnału mowy jest wydzielany za pomocą ruchomego okna czasowego (zwanego ramką), przy czym zakłada się, że obrębie tego okna częstotliwość F_0 jest stała. Należy podkreślić, że wymienione metody pozwalają jedynie na wyznaczenie wartości F_0 znamiennej dla danego segmentu sygnału mowy.

W celu zwiększenia dokładności wyznaczania wartości częstotliwości F_0 wprowadzono różne modyfikacje tych metod. Na przykład w pracach [4, 5, 9] zaproponowano zastosowanie funkcji autokorelacji do sygnałów reprezentujących różne rozwinięcia falkowe sygnału mowy.

Wartość chwilową częstotliwości F_0 można obliczyć na podstawie czasu trwania cyklu pracy fałd głosowych (tj. okresu tonu krtaniowego), w którym następuje otwarcie i zamknięcie głośni. Detekcja poszczególnych cykli w sygnale mowy stanowi istotny problem z uwagi na złożoną strukturę czasowo-częstotliwościową tego sygnału, tzn. trudno jest odróżnić lokalne maksima sygnału związane z drganiem fałd głosowych od maksimów fal wywołanych rezonansowym zachowaniem narządów toru głosowego. Opracowano różne rozwiązania tego problemu, np. w pracy

[8] zaproponowano zastosowanie filtru, który tłumi składowe sygnału mowy o częstotliwościach formatowych (tj. częstotliwościach rezonansowych znamiennych dla traktu głosowego).

Do wyznaczania wartości chwilowej częstotliwości F_0 wykorzystuje się także odpowiednie rozwinięcia falkowe uzyskane za pomocą dyskretnej (DWT) bądź ciągłej transformaty falkowej (CWT) - np. [3, 9]). Przekształcenie falkowe dzięki dekompozycji badanego sygnału na składowe dobrze zlokalizowane w dziedzinie czasu i częstotliwości umożliwia opisanie lokalnej regularności sygnału [11]. Wykazano związek transformaty falkowej z wykładnikiem Lipschitza stosowanym w matematyce do badania i opisu lokalnej regularności funkcji. Jeżeli falka o zwartym nośniku jest pochodną funkcji skalującej i ma n zerowych momentów, to transformata falkowa może być interpretowana jako wieloskalowy operator różniczkowy n - tego stopnia. Detekcja punktów osobliwych i nieregularnych struktur sygnału sprowadza się zatem do badania lokalnych maksimów modułu rozwinięcia falkowego. Gdy falka posiada jeden moment zerowy, to lokalne maksima modułu transformaty falkowej występują przy dużej zmienności sygnału, zaś brak maksimów modułu rozwinięcia falkowego przy małych skalach oznacza, że badany sygnał jest lokalnie regularny. W przypadku mowy dźwięcznej duża zmienność i nieciągłość sygnału mowy występuje w chwili zamknięcia głośni przez struny głosowe.

Wartość chwilową częstotliwości F_0 można także wyznaczyć na podstawie analizy tzw. widma Hilberta [12]. W metodzie tej najpierw przeprowadzana jest empiryczna dekompozycja sygnału mowy na składowe (tzw. funkcje *IMF*), a następnie za pomocą transformaty Hilberta wyznaczane są wartości chwilowe częstotliwości i amplitudy tych składowych, co stanowi podstawę do utworzenia wspomnianego widma. Metoda ta nie jest powszechnie stosowana z uwagi na czasochłonne obliczenia.

Wartość okresu tonu krtaniowego można również określić za pomocą tzw. metody filtru odwrotnego [10], tj. na podstawie analizy sygnału resztkowego, który reprezentuje błąd predykcji modelu AR przyjętego dla traktu głosowego. W przypadku mowy dźwięcznej w sygnale resztkowym występują charakterystyczne impulsy, które są związane z cyklicznym oraz krótkotrwałym przepływem powietrza przez drgające fałdy głosowe. Modelowanie parametryczne wymaga spełnienia warunku dotyczącego stacjonarności sygnału, dlatego sygnał mowy musi być poddany segmentacji.

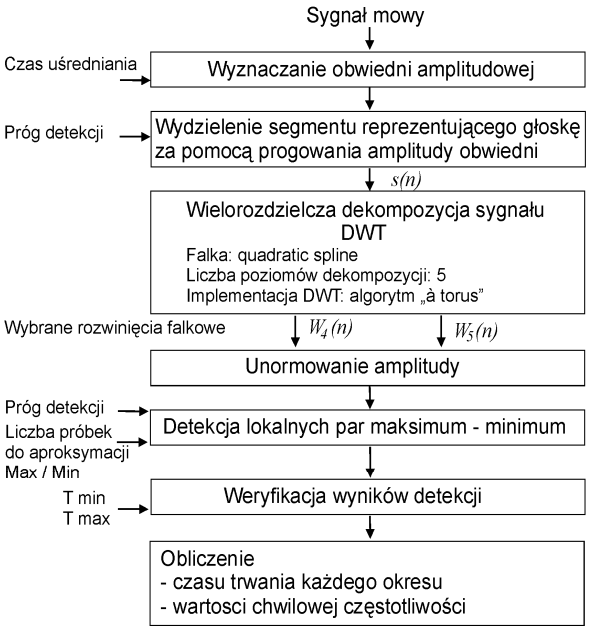
Na podstawie przeglądu literatury przedmiotowej można stwierdzić, że nadal prowadzane są prace w zakresie opracowania nowych oraz modyfikacji klasycznych metod wyznaczania częstotliwości F_0 na podstawie sygnału mowy.

Opis metody

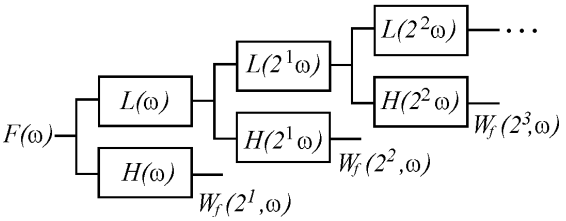
Na rysunku 1 przedstawiono algorytm opracowany do wyznaczania wartości chwilowej częstotliwości podstawowej tonu krtaniowego na podstawie analizy (w trybie off-line) wyizolowanych głosek dźwięcznych o naturalnej lub wydłużonej fonacji.

W ramach wstępnego przetwarzania zarejestrowanego sygnału wyznaczana jest obwiednia amplitudowa, która stanowi podstawę do wydzielenia głosek. Lokalizację czasową (tj. początek i koniec) każdej głoski określa się za pomocą metody progowania obwiedni amplitudowej. Jednocześnie w ten sposób można wyeliminować wpływ zakłóceń występujących pomiędzy zarejestrowanymi głoskami. Do dalszej analizy wybierana jest jedna lub kilka głosek w zależności od celu badania.

Następnie przeprowadzana jest wielorozdzielcza dekompozycja sygnału mowy za pomocą falki „quadratic spline" według algorytmu „à torus" [11], którego schemat blokowy przedstawiono na rysunku 2.



Rys. 1. Algorytm wyznaczania wartości chwilowej częstotliwości F_0



Rys. 2. Algorytm „à torus" do implementacji DWT

$L(2^j\omega)$ i $H(2^j\omega)$ oznaczają transmitancje widmowe filtrów o skończonej odpowiedzi impulsowej odpowiednio dolnoprzepustowego oraz górnoprzepustowego na poziomie dekompozycji k:

$$L(2^j\omega) = exp(i2^j\omega/2)\cdot\left(cos\frac{2^j\omega}{2}\right)^3 \tag{1}$$

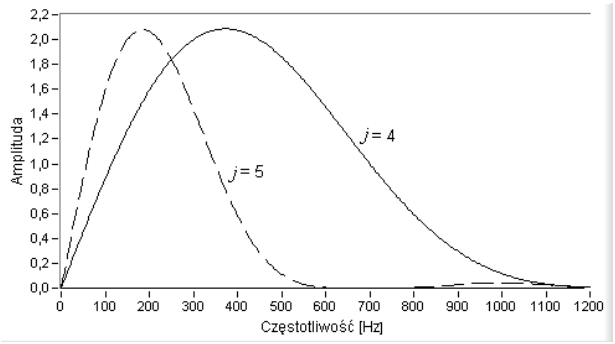
$$H(2^j\omega) = 4i\,exp(i2^j\omega/2)\cdot\left(sin\frac{2^j\omega}{2}\right) \tag{2}$$

Podczas wyznaczania dyskretnej transformaty falkowej (DWT) wg schematu na rysunku 2 nie realizuje się operacji decymacji, która w algorytmie Mallata jest wykonywana na każdym poziomie dekompozycji [11]. Takie rozwiązanie zapewnia wspólną oś czasu dla współczynników reprezentujących różne poziomy dekompozycji sygnału, co ma duże znaczenie przy detekcji charakterystycznych punktów w badanym sygnale. Współczynniki rozwinięcia falkowego, tzw. detale dla każdego poziomu dekompozycji k wyznaczane są wg wzoru:

$$\begin{aligned} W_f(2^{j+1},\omega) &= H(\omega)\cdot F(\omega) \quad j=0, k=1 \\ &= H(2\omega)\cdot L(\omega)\cdot F(\omega) \quad j=1, k=2 \\ &= H(2^{j-1}\omega)\cdot L(2^{j-2}\omega)\cdot\ldots\cdot L(\omega)\cdot F(\omega) \quad j\geq 2, k\geq 3 \end{aligned} \tag{3}$$

gdzie $F(\omega)$, $W_f(2^{j+1},\omega)$ oznaczają odpowiednio widma Fouriera badanego sygnału $f(n)$ i detalu na poziomie k.

W przypadku sygnału mowy spróbkowanego z częstotliwością 11025 Hz wielorozdzielcza dekompozycja została przeprowadzona z uwzględnieniem pięciu poziomów, przy czym do detekcji poszczególnych okresów tonu krtaniowego są wykorzystywane tylko rozwinięcia falkowe dla skal 2^4 oraz 2^5. Charakterystyki amplitudowo-częstotliwościowe filtrów pasmowoprzepustowych, które są znamienne dla tych skal zaprezentowano na rysunku 3. Łatwo zauważyć, że częstotliwość podstawową tonu krtaniowego dla głosu męskiego można określić tylko na podstawie detalu odpowiadającego skali 2^5. Natomiast w przypadku głosu kobiecego o dużej wartości częstotliwości F_0 należy wykorzystać detal związany ze skalą 2^4.



Rys. 3. Charakterystyki amplitudowo-częstotliwościowe filtrów przyjętych do wyznaczania F_0

Przyjęta do dekompozycji falka „quadratic spline" posiada jeden moment zerowy, co jak wcześniej wspomniano, umożliwia wykrycie chwili czasowej, w której występuje duża zmienność badanego sygnału. Na przykład określenie takiej chwili czasu sprowadza się do ustalenia miejsca zerowego pary: ujemne minimum i dodatnie maksimum.

W celu ograniczenia zakresu zmian amplitudy do przedziału (-1,1) wybrane detale, tj. $W_4(n)$ oraz $W_5(n)$ zostały poddane operacji normowania amplitudy wg wzoru:

$$unW_k(n) = \frac{W_k(n)}{max\left\{\left|W_k(n)\right|\right\}}, \qquad k = 4,5 \tag{4}$$

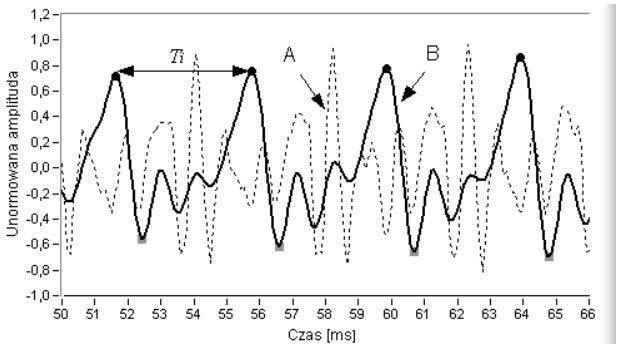
Do wykrywania i weryfikacji lokalnych maksimów oraz minimów występujących w wybranych rozwinięciach falkowych opracowano algorytm, który bazuje na aproksymacji wielomianem drugiego stopnia wartości współczynników danej skali, leżących w okolicach lokalnych maksimów i powyżej założonego progu detekcji. Należy podkreślić, że aproksymacja szczytu za pomocą paraboli i przyjęcie maksimum paraboli jako wartości maksymalnej zapewnia większą dokładność lokalizacji czasowej niż metoda wyszukiwania maksymalnych wartości współczynników falkowych. Wyniki detekcji lokalnych maksimów są weryfikowane w celu odrzucenia „fałszywych maksimów", przy czym jako kryterium weryfikacji przyjęto minimalny oraz maksymalny czas trwania okresu tonu krtaniowego typowy dla płci mówcy.

Na rysunkach 4, 5 przedstawiono segment sygnału mowy (o unormowanej amplitudzie) reprezentujący głoskę /a/ wraz z wybranym rozwinięciem falkowym, odpowiednio $unW_4(n)$ i $unW_5(n)$ (oznaczonych na rysunkach jako B oraz C). Łatwo zauważyć, że dużym zmianom amplitudy sygnału mowy towarzyszy pojawienie się w wybranych detalach lokalnych maksimów i minimów.

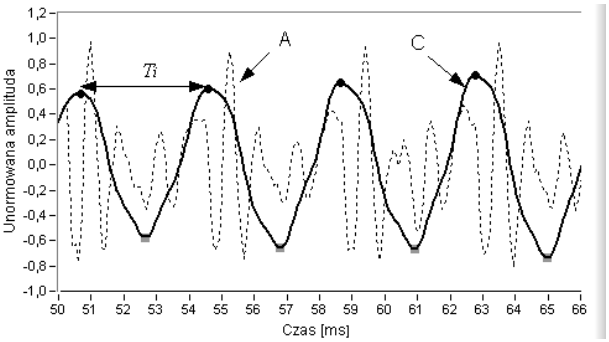
Rozwinięcie falkowe dla skali 2^5 (na rys.5) reprezentuje tylko ton krtaniowy, którego okres można obliczyć na podstawie lokalizacji czasowej sąsiednich maksimów.

Natomiast na rysunku 6 zaprezentowano badany sygnał mowy z wydzielonymi segmentami, które odpowiadają poszczególnym okresom tonu krtaniowego, które określono za pomocą rozwinięcia falkowego dla skali 2^5.

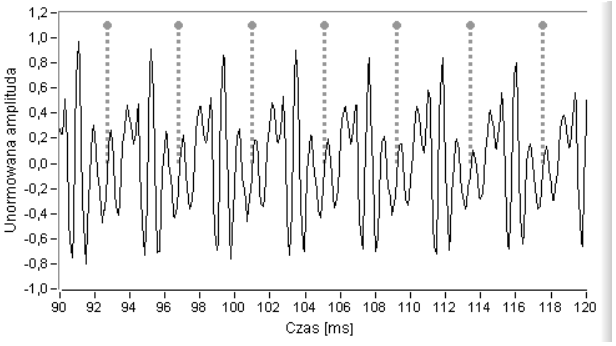
Chwilową wartość częstotliwości podstawowej tonu krtaniowego określa się jako odwrotność czasu trwania okresu T_i.



Rys. 4. Sygnał mowy (A) i rozwinięcie falkowe dla skali 2^4 (B)



Rys. 5. Sygnał mowy (A) i rozwinięcie falkowe dla skali 2^5 (C)



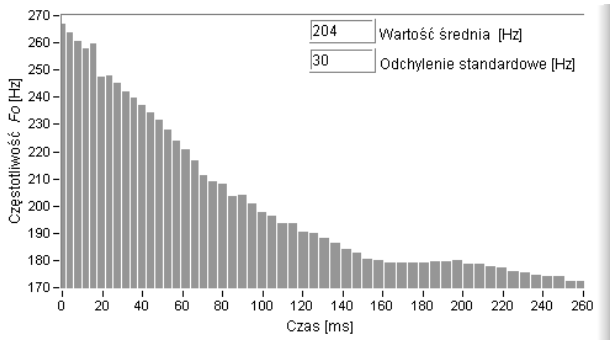
Rys. 6. Wynik podziału sygnału mowy reprezentującego głoskę /a/ na segmenty opowiadające okresowi tonu krtaniowego

Asymetria filtru cyfrowego opisanego wzorem (3) jest przyczyną opóźnienia np. par lokalnych ekstremów o przeciwnych znakach w odniesieniu do oryginalnego sygnału, a wartość tego opóźnienia (wyrażana jako liczba próbek) zależy od skali i wynosi około 2^{k-1}-1 dla detalu na poziomie k. Stałe opóźnienie jednak nie wpływa na dokładność wyznaczania czasu trwania okresów tonu krtaniowego.

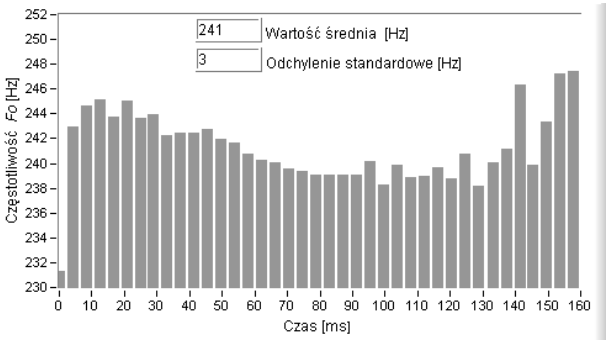
Przykładowe wyniki

Wyizolowane głoski dźwięczne kilku mówców (wypowiadane w pozycji siedzącej) zarejestrowano za pomocą mikrofonu i karty dźwiękowej (z 16-bitowym przetwornikiem A/C), która współpracuje z komputerem klasy PC. Mikrofon był umieszczony w odległości ok. 10 cm od ust mówcy.

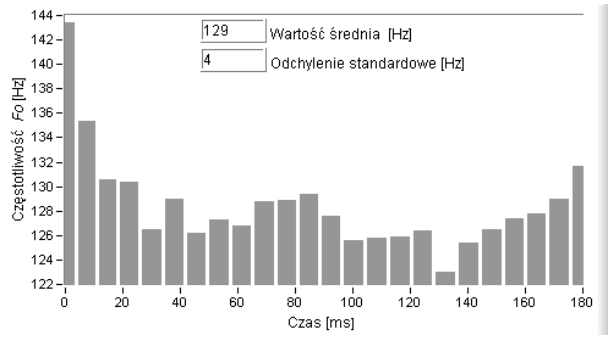
Na rysunkach 7, 8 przedstawiono wyniki końcowe analizy wyizolowanej głoski /a/ wypowiadanej przez dwie kobiety w różnym wieku. Łatwo zauważyć, że zmiany częstotliwości podstawowej tonu krtaniowego w dużym stopniu zależą od intonacji. W przypadku głosów kobiecych wartość częstotliwości F_0 zmniejsza się wraz z wiekiem. Natomiast na rysunku 9 zaprezentowano wykres zmienności częstotliwości podstawowej tonu krtaniowego podczas wypowiedzi /Ola/ przez mężczyznę.



Rys. 7. Wykres zmienności częstotliwości podstawowej tonu krtaniowego dla głoski /a/ wypowiedzianej przez kobietę (50 lat) z intonacją w formie rozkazującej



Rys. 8. Wykres zmienności częstotliwości podstawowej tonu krtaniowego dla głoski /a/ wypowiedzianej przez kobietę (24 lata) z intonacją w formie oznajmującej



Rys. 9. Wykres zmienności częstotliwości podstawowej tonu krtaniowego dla wypowiedzi /Ola/ z intonacją w formie oznajmującej - głos mężczyzny (60 lat)

Podsumowanie

Opisany w niniejszej pracy algorytm jest przeznaczony do wyznaczania wartości chwilowej częstotliwości podstawowej tonu krtaniowego na podstawie analizy falkowej sygnału mowy dźwięcznej. Dyskretną transformatę falkową zastosowano w celu uzyskania sygnału przydatnego do detekcji poszczególnych okresów tonu krtaniowego. Metoda ta jest dedykowana do badania wyizolowanych głosek dźwięcznych.

Uzyskane wyniki umożliwiają ocenę krótkoczasowej zmienności częstotliwości F_0 oraz mogą być wykorzystane do aproksymacji konturu intonacyjnego.

Zaprezentowana w artykule metoda nie wymaga segmentacji oryginalnego sygnału mowy na stacjonarne segmenty. Zaproponowany algorytm do detekcji i weryfikacji lokalnych maksimów i minimów występujących w wybranych rozwinięciach falkowych zapewnia dużą dokładność ich lokalizacji, co ma istotne znaczenie podczas wydzielania poszczególnych okresów.

W ramach kontynuacji pracy przewiduje się modyfikację opracowanego algorytmu w celu badania dłuższych wypowiedzi, w tym również rejestrowanych w obecności typowych zakłóceń akustycznych.

LITERATURA

[1] Yao et al, Classification of speech under stress based on modeling of the vocal folds and vocal tract, *EURASIP Journal on Audio, Speech, and Music Processing,* (2013), 1-17

[2] Moran R. J., et al, Telephony-based voice pathology assessment using automated speech analysis, *IEEE Trans. Biomed. Eng.*, 53 (2006), n.3, 468-477

[3] Cnockaert L., et al, Fundamental frequency estimation and vocal tremor analysis by means of Morlet Wavelet Transforms, *IEEE ICASSP* (2005), 383-396

[4] Manfredi C., et al, A comparative analysis of fundamental frequency estimation methods with application to pathological voices, *Medical Engineering & Physics*, 22 (2000), 135-147

[5] Ozdas A., et al, Investigation of vocal jitter and glottal flow spectrum as possible cues for depression and near-term suicidal risk, *IEEE Trans. Biomed. Eng.*, 51 (2004), n.9, 1530-1539

[6] Kotti M., Paternó F., Speaker-independent emotion recognition exploiting a psychologically-inspired binary cascade classification schema, *Int J Speech Technol,* 15 (2012), 131-150

[7] Veprek P., Scordilis M. S., Analysis, enhancement and evaluation of five pitch determination techniques, *Speech Communication*, 37 (2002), 249-270

[8] Yegnanarayana B., K. et al, Event-based instantaneous fundamental frequency estimation from speech signals, *IEEE Trans. on Audio, Speech, and Language Processing.*, 17 (2009), n.4, 614-624

[9] Bernardin S. L., Foo S. Y., Wavelet processing for pitch period estimation, *IEEE Proceedings of the 30th Southeastern Symposium on System Theory* (2006), 426-429

[10] Zieliński T. P., Cyfrowe przetwarzanie sygnałów – od teorii do zastosowań, *WKŁ*, Warszawa, 2005

[11] Mallat S.: A Wavelet Tour of Signal Processing, *Academic Press,* San Diego-London, 1998

[12] Feng Z., et al, Pitch period estimation of voice signal based on EEMD and Hilbert Transform, *IEEE Proceedings of 2nd International Asia Conference on Informatics in Control, Automation and Robotics,* (2010), 365-368

***Autor**: dr inż. Barbara Wilk, Politechnika Rzeszowska, Katedra Metrologii i Systemów Diagnostycznych, ul. W. Pola A, 35-959 Rzeszów E-mail: bmwilk@prz.edu.pl.*