

Opinia nt. rozprawy doktorskiej mgra inż. Stanisława Gmyrka
pt.: **“Propozycje algorytmów parametryzacji, odpornych na zmienność tonu krtaniowego, dla automatycznego rozpoznawania mowy”**, wykonanej pod kierunkiem prof. dra hab. Ryszarda Makowskiego oraz promotora pomocniczego dra inż. dr inż. Roberta Hossy

1. Tytuł rozprawy doktorskiej

Przedmiotem recenzji jest rozprawa doktorska mgra inż. Stanisława Gmyrka pt.: **“Propozycje algorytmów parametryzacji, odpornych na zmienność tonu krtaniowego, dla automatycznego rozpoznawania mowy”**.

2. Ocena układu rozprawy doktorskiej, w tym informacje o jej poszczególnych częściach składowych

Rozprawa składa się z pięciu rozdziałów oraz literatury, uzupełnionych trzema dodatkami (A, B, C), wykazem skrótów, spisem rysunków i tabel oraz bibliografią liczącą 114 pozycji. Układ pracy jest logiczny i przejrzysty i spełnia wymagania stawiane pracom doktorskim.

Rozdział pierwszy pełni funkcję wprowadzenia: przedstawia krótką klasyfikację systemów ARM (Automatycznego Rozpoznawania Mowy), precyzuje tezę rozprawy, która stanowi jednocześnie cel badań (sformułowany w Podsumowaniu: „Celem badawczym pracy było zaprojektowanie algorytmów parametryzacji odpornych na niepożądany wpływ zmian częstotliwości tonu krtaniowego na współczynniki cepstralne”) oraz podaje założenia badawcze i zadania, mające na celu udowodnienie tezy rozprawy. W końcowej części tego rozdziału znajduje się odniesienie do układu pracy doktorskiej.

Rozdział drugi stanowi wprowadzenie teoretyczne do tematyki prowadzonych przez doktoranta badań; omówiono w nim schemat konwencjonalnego systemu ARM, model sygnału mowy (model źródło-filtr Fanta), metody parametryzacji sygnału z obejmujące różne podejścia, tj. wykorzystujące filtrację sygnału, metody liniowe LPC (Linear Predictive Coding), reprezentacje cepstralne (MFCC (ang. Mel Frequency Cepstral Coefficients, współczynniki mel-cepstralne), HFCC (ang. Human Factor Cepstral Coefficient, współczynniki cepstralne uwzględniające właściwości percepcji dźwięku przez człowieka), BFCC (ang. Bark Frequency Cepstral Coefficients/(ang. Basiliar membranę Frequency-band Cepstral Coefficient)) i inne), techniki kompensacyjne (odporna parametryzacja, grupowanie mówców, normalizacja, adaptacja) oraz przegląd nowoczesnych architektur systemów ARM (hybrydowe

HMM-DNN, architektury łączące kryte Modele Markowa (HMM) z głębokimi sieciami neuronowymi (DNN, deep neural models, End-to-End, Transformer). Dokładniej omówiono metodę parametryzacji HFCC oraz uwarunkowania dotyczące wykorzystania metod/parametrów odwołujących się do ludzkiej percepcji.

Rozdział trzeci odnosi się do autorskich dokonań i wkładu autora: analizy wpływu zmienności częstotliwości tonu krtaniowego F_0 na współczynniki cepstralne, zmiany długości ramki sygnału w kontekście szczegółowego opisu dwóch autorskich algorytmów: PS-HFCC (ang. Pitch Synchronized Human Factor Cepstral Coefficient) oraz IF-HFCC (ang. Inverse Filtering Human Factor Cepstral Coefficient). Metody te stanowią autorskie rozwiązania problemu korekcji widma ramki i zostały wprowadzone ze względu na potrzebę kompensacji oraz poszukiwanie metod odpornej parametryzacji sygnału.

Rozdział czwarty przedstawia metodologię badań, charakterystykę danych (modelowych i rzeczywistych), statystyczny model fonemów, opis modelu GMM, miary skuteczności rozpoznawania oraz wyniki eksperymentów z analizą trzech miar skuteczności dla sygnałów modelowych i rzeczywistych, w tym globalną analizą błędów.

Rozdział piąty stanowi zwięzłe podsumowanie wraz z wnioskami końcowymi.

Warto zauważyć, że każdy z rozdziałów kończy się syntetycznym podsumowaniem, co ułatwia odniesienie się do treści przedstawionych w danym rozdziale.

Załącznikami do rozprawy doktorskiej są Dodatki, które zasługują na pozytywną ocenę: Dodatek A zawiera przegląd algorytmów estymacji F_0 (w tym opis algorytmu YIN zastosowanego w pracy), Dodatek B opisuje algorytm detekcji dźwięczności oraz procedurę IAIF, natomiast Dodatek C prezentuje pełne macierze błędów FER dla obu metod odpornych.

W ogólności, recenzowana rozprawa doktorska ma charakter eksperymentalny, dlatego rozdział drugi pełni funkcję uzasadnienia wyboru wykorzystywanych narzędzi, a nie typowego przedstawienia podstaw teoretycznych. Jednak podział materiału między rozdziałami można by ewentualnie w niewielkim stopniu zmodyfikować; na przykład opis modelu GMM i odniesienie do algorytmu Bauma-Welcha (rozdział 4.1.3, przy okazji: ten algorytm z nazwy pojawia się jedynie w streszczeniach) mogłyby zostać włączone do rozdziału 2 jako część podstaw teoretycznych.

Z kolei rozdział piąty jest stosunkowo krótki i w znacznej mierze powtarza wnioski z rozdziału czwartego, co osłabia jego rolę jako syntetycznego podsumowania. **Ponadto w Podsumowaniu można byłoby wyraźniej odnieść się do ograniczeń metody, do pozycji uzyskanych wyników w odniesieniu do aktualnego stanu wiedzy oraz do planów dalszych badań.**

3. Ocena celu rozprawy doktorskiej

Teza pracy i cel zostały jasno sformułowane i dotyczą zaprojektowania algorytmów parametryzacji odpornych na zmienność częstotliwości tonu krtaniowego, które poprawią jakość rozpoznawania fonemów dźwięcznych (cel rozprawy podany w streszczeniach oraz Podsumowaniu).

Zanim jednak skomentuję tezę pracy, chciałabym odnieść się do tytułu rozprawy. A mianowicie, tytuł rozprawy doktorskiej trafnie wskazuje na problematykę odpornej parametryzacji sygnału mowy, **jednak rozważenia – w mojej opinii – wymagałoby ograniczenie bezpośredniego odniesienia do automatycznego rozpoznawania mowy**. Z przedstawionej pracy wynika bowiem, że jej głównym osiągnięciem jest analiza wpływu zmienności częstotliwości tonu krtaniowego na proces parametryzacji oraz opracowanie autorskich metod ograniczających ten wpływ. I niewątpliwie taki jest rzeczywisty wkład naukowy rozprawy. Automatyczne rozpoznawanie mowy stanowi przede wszystkim obszar eksperymentalnej weryfikacji proponowanych rozwiązań.

Ponadto, wobec bardzo dynamicznego rozwoju technologii rozpoznawania mowy oraz szybkiej ewolucji stosowanych obecnie modeli i architektur, wskazane byłoby rozważenie tytułu podkreślającego bardziej uniwersalny, analityczny charakter prowadzonych badań, co zresztą jest ich ogromną zaletą. Takie sformułowanie lepiej eksponowałoby zasadniczy wkład naukowy rozprawy, niezależny od konkretnego obszaru zastosowania, zwłaszcza, że modele głębokie najczęściej wykorzystują na wejściu reprezentację 2D w postaci spektrogramów w skali melowej.

Za najbliższy rzeczywistemu wkładowi naukowemu pracy uznałabym:

„Algorytmy parametryzacji sygnału mowy odporne na zmienność częstotliwości tonu krtaniowego” lub

„Analiza i kompensacja wpływu zmienności tonu krtaniowego w parametryzacji sygnału mowy”

ponieważ eksponują one główny problem badawczy i opracowane rozwiązania, nie wiążąc ich z konkretną generacją technologii rozpoznawania mowy, która zmienia się niemal z dnia na dzień.

Teza pracy brzmi:

„Możliwe jest zaprojektowanie metody odpornej parametryzacji sygnału mowy dla systemów ARM, kompensującej wpływ okresowości pobudzenia i jej zmienności, o skuteczności nie mniejszej niż skuteczność metod znanych z literatury.”

Teza została sformułowana naukowo poprawnie (ale ostrożnie), bowiem przyjmuje jako kryterium sukcesu skuteczność nie mniejszą niż w metodzie referencyjnej HFCC, co można weryfikować eksperymentalnie. Zakres tej weryfikacji jest przy tym precyzyjnie ograniczony do fonemów dźwięcznych mowy polskiej, co autor konsekwentnie uzasadnia przyjętą metodą kompensacji.

Moim zdaniem teza mogłaby być sformułowana bardziej ambitnie; **wyniki eksperymentalne wykazują wyraźną poprawę (a nie tylko dorównanie metodzie referencyjnej), co warto podkreślić.**

Z kolei założenie ograniczające bazę do głosów męskich jest uzasadnione praktycznie, lecz zawęża ogólność wniosków i powinno zostać w Podsumowaniu wyraźniej omówione jako ograniczenie pracy.

W pracy podano również motywację prowadzonych prac badawczych. Jak podaje autor rozprawy, motywacją do opracowania algorytmów było jak najwierniejsze odwzorowanie ludzkiej percepcji dźwięku w zastosowanych metodach parametryzacji. Jest to bardzo istotny cel, dlatego zarówno cel, jak i tezę niniejszej rozprawy doktorskiej należy ocenić jako wartościowe i aktualne. **Jednak podanie hipotezy w formie jakościowo-ilościowej wzbogaciłoby wnioski wynikające z prowadzonych eksperymentów.**

4. Ocena zastosowanych metod badawczych

Autor stosuje metodologię właściwą dla badań w obszarze przetwarzania sygnałów mowy oraz w zastosowaniach systemów ARM, w szczególności w specyficznych zastosowaniach, w których istotne jest podejście analityczne.

Autor rozprawy podaje następujące założenia badań:

- baza nagrań wykorzystywana w badaniach składa się z krótkich wypowiedzi (pojedynczych wyrazów) różnych mówców płci męskiej,
- w badaniach skoncentrowano się na wyselekcjonowanych fonemach dźwięcznych mowy polskiej, a to ograniczenie wynika z badanych metod kompensacji,
- badania wykonano, korzystając z bazy nagrań zawierającej wypowiedzi 40 przypadkowych mówców zarejestrowanych w różnych regionach Polski,
- skuteczność poszczególnych metod kompensacji weryfikowano w oparciu o rozpoznawanie fonemów w pojedynczych ramach sygnału.

Weryfikacja zaproponowanych rozwiązań przebiega dwuetapowo: najpierw na sygnałach modelowych (generowanych syntetycznie na podstawie modelu źródło-filtr Fanta), następnie na rzeczywistych nagraniach 42 mówców.

Baza nagrań jest odpowiednio zróżnicowana regionalnie i fonetycznie. Grupowanie mówców (algorytm oparty na modelu UBM, Universal Background Model) zwiększa wiarygodność wyników i pozwala ocenić, w jakim stopniu poprawa wynika z zastosowania metod odpornych, a w jakim z jednorodności grupy akustycznej.

Przyjęta metodologia jest poprawna, a przedstawienie wyników - przekonujące. **Jednak pomimo pozytywnego odbioru przeprowadzonych badań, nasuwają się pewne uwagi:**

- Baza obejmuje głosy męskie, co autor uzasadnia w kontekście przyjętej metodologii (estymacja F_0), jednak ogranicza to możliwość uogólnienia wyników na głosy kobiece i dziecięce.

- Liczba mówców (42 x 150 słów języka polskiego) jest stosunkowo duża, ale dla badań o charakterze statystycznym większa baza zwiększyłaby wiarygodność wniosków.
- SNR (Signal-to-Noise Ratio) na poziomie 35 dB (nagrania w samochodzie) wprowadza pewien poziom zakłóceń, który może w sposób niekontrolowany współoddziaływać z analizowanymi efektami parametryzacji. Brak osobnej analizy wpływu poziomu szumu na skuteczność obu metod jest pewnym niedoborem.
- Badania ograniczono do rozpoznawania na poziomie pojedynczych ramek (FER), nie przedstawiając wyników rozpoznawania sekwencji fonetycznych ani słów (WER). Warto rozważyć plany weryfikacji na wyższym poziomie przetwarzania. Czy autor przewiduje takie badania?
- W przypadku metody PS-HFCC skuteczność jest silnie uzależniona od dokładności algorytmu estymacji F_0 (YIN) oraz precyzji wyznaczania granic okresów. Warto zaproponować analizę wrażliwości metody na błędy estymacji F_0 .

Wybór trzech miar skuteczności jest trafny i dobrze uzasadniony:

- **Odchylenia standardowe współrzędnych wektora obserwacji** pozwalają bezpośrednio ocenić redukcję rozrzutu współczynników cepstralnych niezależnie od modelu klasyfikatora.
- **Dywergencja Kullbacka-Leiblera (KLD)** między rozkładami GMM umożliwia ocenę separowalności fonemów w przestrzeni cech; zastosowanie aproksymacji UT (Unscented Transform) w przypadku braku analitycznej formuły dla mieszanin gaussowskich stanowi rozwiązanie metodologicznie poprawne.
- **Frame Error Rate (FER)** stanowi bezpośrednią miarę jakości klasyfikacji ramek sygnału, bliską praktycznej efektywności systemu ARM.

5. Uwagi do dyskusji:

1. Niektóre z uwag do dyskusji znalazły się we wcześniejszych punktach oceny - chciałabym, aby doktorant odniósł się do nich.
2. Zaproponowane metody PS-HFCC i IF-HFCC różnią się zasadniczo mechanizmem kompensacji. Pierwsza z tych metod eliminuje efekt obciążenia poprzez synchronizację ramki z T_0 , a druga dąży do separacji pobudzenia i traktu głosowego poprzez filtrację odwrotną. Wyniki wskazują, że PS-HFCC jest skuteczniejsza - widać poprawę w 80% przypadków wobec 75% dla IF-HFCC, a jej przewaga jest szczególnie wyraźna w przypadku samogłosek. Metodę IF-HFCC cechują natomiast większe oscylacje wyników między grupami mówców. Warto rozważyć, czy jest to związane z wrażliwością algorytmu IAIF na warunki rejestracji.

3. Autor słusznie wskazuje, że IF-HFCC nie eliminuje efektu obciążenia, co stanowi istotne ograniczenie tej metody i może wyjaśniać gorsze wyniki w porównaniu z PS-HFCC. Interesującą perspektywą badawczą byłoby połączenie obu metod (wstępnie sygnalizowane w pracy [31]).

Uwagi formalne i redakcyjne:

1. Praca jest napisana bardzo starannie; język jest poprawny i precyzyjny.
2. Liczba rysunków i tabel jest obszerna; część rysunków (szczególnie wykresy FER dla grup g1–g6) jest podobna do siebie; ich liczba mogłaby zostać zredukowana na rzecz bardziej zwartego zestawienia wyników oraz przeniesienia wykresów szczegółowych do Załącznika.
3. Przydatna byłaby podsumowująca tabela porównawcza wyników FER dla obu metod w układzie fonem $\times$ grupa, co pozwoliłoby czytelnikowi na szybkie porównanie rezultatów.
4. W treści rozprawy nie znalazłam odwołania do Dodatku A.
5. Wyniki zawarte w tabelach są zapisane z separatorem dziesiętnym kropką (separatorem dziesiętnym w tekście pisanym w języku polskim jest przecinek).

Dodatkowe pytania do dyskusji:

1. Jak zachowuje się metoda PS-HFCC w przypadku znacznych błędów estymacji F_0 , np. przy niskim SNR lub fonemach o niestabilnej częstotliwości podstawowej?
2. Czy autorskie metody były testowane lub porównywane z innymi podejściami odpornymi opisanymi w literaturze (np. MVDR-MFCC (Minimum Variance Distortionless Response-Mel-Frequency Cepstral Coefficients), RASTA-PLP, Relative Spectral Transformation-Perceptual Linear Prediction)? Obie metody są wykorzystywane w cyfrowym przetwarzaniu sygnałów audio i automatycznym rozpoznawaniu mowy. **RASTA-PLP** dodatkowo filtruje sygnał w celu usunięcia szumów i zakłóceń o charakterze wolnozmiennym (np. pogłosu czy szumu tła), dzięki czemu jest odporniejsza na trudne warunki akustyczne.
3. Jakie są plany weryfikacji skuteczności na wyższym poziomie przetwarzania (rozpoznawanie słów, WER (Word Error Rate))?
4. Czy zaproponowane metody nadają się do zastosowania w systemach czasu rzeczywistego, biorąc pod uwagę konieczność bieżącej estymacji F_0 oraz ekstrakcji pojedynczych okresów?

5. W jaki sposób metoda PS-HFCC (Pitch Synchronized Human Factor Cepstral Coefficients) radzi sobie z fonemami o niestacystycznej wartości F_0 w obrębie ramki (np. podczas prozodycznych zmian tonu)?

6. Ocena, czy rozprawa doktorska stanowi oryginalne rozwiązanie problemu naukowego

Zaproponowana teza badawcza ma wyraźny walor oryginalności. Przedstawione autorskie wyniki pozwoliły udowodnić tę tezę. Najbardziej wartościowe wydaje się włączenie mniej typowych metryk oceny skuteczności proponowanych metod, dostosowanych do specyfiki tych algorytmów.

Tematyka odpornej parametryzacji sygnału mowy pozostaje aktywnym obszarem badań, mimo dynamicznego rozwoju systemów głębokiego uczenia (End-to-End). Aktualność pracy można uzasadnić na kilku poziomach.

- Nawet w nowoczesnych architekturach neuronowych etap ekstrakcji cech akustycznych nie został wyeliminowany. Jak sam autor wykazuje w przeglądzie literatury, systemy oparte na modelach Transformer, CTC czy RNN-T nadal wykorzystują spektrogramy w skali melowej lub współczynniki cepstralne (MFCC, PLP) jako wejście do enkoderów. Problemy związane z quasiokresowością pobudzenia krtaniowego, będące przedmiotem pracy, degradują te reprezentacje niezależnie od tego, czy dalsze przetwarzanie realizuje klasyczny klasyfikator GMM czy głęboka sieć neuronowa. Poprawa jakości cech na etapie parametryzacji ma zatem potencjał przekładający się na skuteczność całego łańcucha przetwarzania.

- Systemy ARM działające w warunkach ograniczonych zasobów obliczeniowych, wbudowane systemy sterowania głosowego, interfejsy człowiek-maszyna w systemach przemysłowych, medycznych czy motoryzacyjnych, nadal w znacznej mierze opierają się na architekturach klasycznych (HMM-GMM) lub hybrydowych, ze względu na niskie wymagania pamięciowe i czasowe. W tych zastosowaniach odporna parametryzacja jest kluczowym narzędziem poprawy skuteczności, a jej znaczenie praktyczne jest niekwestionowane.

- Zagadnienie zmienności tonu krtaniowego jest problemem fundamentalnym w akustyce mowy, dotyczącym bezpośrednio fizycznych właściwości sygnału, a nie jedynie specyfiki konkretnego algorytmu rozpoznawania. Podejście zaproponowane przez autora, tj. analiza zjawiska od strony modelu sygnału (model Fanta, efekt obciążenia, harmoniczne F_0) oraz projektowanie kompensacji na poziomie reprezentacji widmowej, ma charakter ogólny i może być przenoszone na inne metody parametryzacji oraz architektury systemów.

Znaczenie dla akustyki inżynierskiej i przetwarzania sygnałów

Rozprawa doktorska wnosi wartościowy wkład w kilku wymiarach.

- W wymiarze **metodycznym** autor proponuje dwa oryginalne algorytmy parametryzacji, tj. PS-HFCC i IF-HFCC, będące twórczymi modyfikacjami istniejących metod cepstralnych. Metoda PS-HFCC, oparta na synchronizacji długości ramki z bieżącym okresem podstawowym i uśrednianiu widm pojedynczych okresów, jest eleganckim rozwiązaniem problemu efektu obciążenia, dobrze uzasadnionym matematycznie i skutecznym eksperymentalnie. Metoda IF-HFCC, wykorzystująca filtrację odwrotną IAIF do estymacji transmitancji kanału głosowego, nawiązuje do klasycznych badań nad modelem źródło-filtr i stanowi jej praktyczne zastosowanie w kontekście systemów ARM. Oba algorytmy mogą zostać uogólnione na inne metody parametryzacji cepstralnej, co zwiększa ich wartość naukową.
- W wymiarze **eksperymentalnym** autor opracował własną bazę nagrań polskich głosów męskich, przeprowadził ręczną segmentację i etykietyzację fonetyczną oraz zrealizował kompleksową analizę wyników z zastosowaniem trzech miar skuteczności, w tym rzadziej stosowanej w pracach krajowych miary KLD (Dywergencja Kullbacka-Leiblera) między rozkładami GMM. Stanowi to solidny wkład metodyczny w zakresie ewaluacji systemów ARM dla języka polskiego, który jest stosunkowo słabo reprezentowany w międzynarodowej literaturze przedmiotu.
- W wymiarze **aplikacyjnym** uzyskane wyniki, tj. redukcja błędu FER dla fonemu r z 62% do 53%, dla m z 52% do 44,5%, dla samogłoski a do poziomu 4–5% po grupowaniu mówców, mają charakter praktycznie istotny i potwierdzają użyteczność zaproponowanych rozwiązań w rzeczywistych systemach przetwarzania mowy.

Wobec powyższego, należy uznać, że praca jest aktualna, osadzona w nurcie badań nad przetwarzaniem sygnałów akustycznych i w pełni mieści się w profilu dyscypliny Automatyka, Elektronika, Elektrotechnika i Technologie Kosmiczne. Rozwiązuje konkretny, dobrze zdefiniowany problem inżynierski, stosuje właściwe metody, a uzyskane wyniki są weryfikowalne i porównywalne z wynikami z literatury. Stanowi tym samym wartościowy wkład zarówno w dorobek krajowej akustyki inżynierskiej, jak i w szerszy obszar cyfrowego przetwarzania sygnałów mowy. Wymagało to od autora zarówno głębokiego zrozumienia teorii przetwarzania sygnałów, jak i umiejętności implementacyjnych oraz eksperymentalnych.

W konkluzji stwierdzam, że ocena przedłożonej mi do recenzji rozprawy doktorskiej p. mgra inż. **Stanisława Gmyrka** jest **pozytywna** i spełnia wymagania stawiane rozprawom doktorskim w aktualnie obowiązującej Ustawie o stopniach i tytule naukowym. W związku z tym wnoszę o **dopuszczenie rozprawy doktorskiej p. mgra inż. Stanisława Gmyrka do publicznej obrony.**

Uzasadnienie wyróżnienia

Jednocześnie wnoszę o wyróżnienie rozprawy doktorskiej. W uzasadnieniu chciałabym się odnieść do pracy opublikowanej w czasopiśmie Archives of Acoustics

- jest to bardzo ważne czasopismo, jedno z nielicznych na świecie w dziedzinie akustyki. Ponadto doktorant jest autorem kilku innych artykułów (są to pozycje [27]-[31], cytowane w rozdziale Literatura; doktorant jest pierwszym autorem). Dlatego, zgodnie z przyjętymi zasadami wyróżniania rozpraw doktorskich w Politechnice Wrocławskiej (Uchwała nr 81/3/RDn02/2-24-2028 Rady Naukowej w sprawie zaopiniowania Zasad wyróżniania rozpraw doktorskich w dyscyplinie naukowej Automatyka, Elektronika, Elektrotechnika i Technologie Kosmiczne) **wnoszę o wyróżnienie rozprawy doktorskiej.**

 
prof. dr hab. inż. Bożena Kostek

