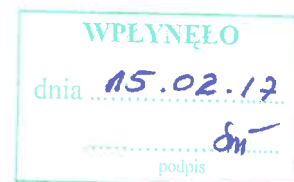


Prof. dr hab. inż. **Krzysztof Ślot**
Instytut Informatyki Stosowanej,
Politechnika Łódzka



Recenzja rozprawy doktorskiej
Pani mgr. inż. **Magdaleny Igras-Cybulskiej**
p.t.:

Analysis of non-linguistic content of speech signals

1. Charakterystyka obszaru tematycznego pracy

Tematyka rozprawy doktorskiej Pani mgr inż. Magdaleny Igras-Cybulskiej zawiera się w dziedzinie przetwarzania sygnału mowy i jest ukierunkowana na poszukiwanie nowych metod analizy, pozwalających na wydobywanie z tego sygnału szerokiego spektrum informacji o charakterze niewerbalnym. Szczegółowe obszary prac podjętych przez Doktorantkę obejmują kilka powiązanych ze sobą, niezwykle interesujących aspektów analizy sygnału mowy:

- poszukiwanie deskryptorów sygnału akustycznego pozwalających na prawidłową reprezentację wybranych kategorii poza-lingwistycznych (takich jak ładunek emocjonalny, intencja) i para-lingwistycznych (takich jak wtrącenia, pauzy, intonacja) przekazu niewerbalnego
- poszukiwanie metod detekcji i identyfikacji treści niewerbalnych
- poszukiwanie związków i relacji istniejących między ilościowymi parametrami opisującymi sygnał mowy a subiektywnymi kategoriami oceny jakości procesu wypowiedzania się

Uzyskanie pomyślnych efektów prac w wymienionych obszarach może umożliwić lub istotnie poprawić jakość realizacji kluczowych funkcjonalności stawianych współcześnie przed systemami analizy mowy. Prawidłowa identyfikacja informacji para-lingwistycznej jest niezbędna dla jakościowej poprawy funkcjonowania systemów translacji mowy na tekst, pozwalając na rekonstrukcję właściwej struktury wypowiedzi, wyznaczonej przez odtworzoną interpunkcję. Identyfikacja procesów psychofizycznych wpływających na formowanie mowy może wpłynąć na zwiększenie poprawności procedur automatycznego rozpoznawania mowy (ang. Automatic Speech Recognition – ASR). Właściwa ocena intonacji zdania może stanowić ważny element ułatwiający interpretację treści zawartej w zdaniu przez algorytmy przetwarzania języka naturalnego (ang. Natural Language Processing - NLP). Poprawna identyfikacja przekazu niewerbalnego i odpowiednie kojarzenie jego komponentów może być punktem wyjścia dla budowy profilu mówcy, obejmującego hipotezy odnośnie jego tożsamości (na bazie indywidualnego charakteru przekazu niewerbalnego), cech osobowościowych (ujawnianych w danym kontekście stanem emocjonalnym lub intencjami), czy poziomu wykształcenia (dedukowanego w drodze analizy błędów i nawyków językowych oraz sposobu intonacji).

Przedstawiona charakterystyka obszaru tematycznego rozprawy uprawnia do sformułowania jednoznacznego wniosku o istotności i aktualności podjętego przedmiotu prac, wpisującego się w dominujący obecnie w naukach technicznych trend badań nad budową systemów inteligentnych. Należy przy tym podkreślić, że zagadnienia szczegółowe poruszone w rozprawie dotyczą problemów trudnych i otwartych, nie posiadających jak dotąd ogólnych rozwiązań. Trudność podjętego zadania jest pochodną faktu, że sygnał mowy należy do jednej z najbardziej złożonych form przekazu informacji z jakim ma do czynienia współczesna technika, a prace nad jego analizą wymagają posiadania odpowiednio wysokich i rozległych kompetencji, obejmujących szereg obszarów wiedzy, takich jak przetwarzanie sygnałów, analiza języka naturalnego i lingwistyka matematyczna, statystyka czy inteligencja obliczeniowa. Brak gotowych rozwiązań wynika w istotnym stopniu z regionalnego kontekstu podjętego problemu – skala badań nad przetwarzaniem i analizą sygnału mowy w odniesieniu do języka polskiego jest bardzo skromna w konfrontacji z dorobkiem naukowym uzyskanym dla języków dominujących w świecie. Jednocześnie automatyczne przenoszenie wiedzy i rozwiązań opracowanych w skali światowej dla innych języków nie jest uzasadnione, więc prace badawcze często muszą mieć pionierski charakter.

Podsumowując przedstawione rozważania chciałbym jednoznacznie stwierdzić, że zarówno stopień trudności podejmowanego zadania, jak i jego istotność oraz aktualność zdecydowanie spełniają wymagania stawiane rozprawom doktorskim, a jego specyfika i obszar tematyczny są adekwatne dla dyscypliny naukowej, której dotyczy przewód (Biocybernetyka i Inżynieria Biomedyczna).

2. Merytoryczna ocena pracy

Merytoryczną ocenę przedłożonej rozprawy doktorskiej chciałbym poprzedzić zwięzłą charakterystyką układu i zawartości tematycznej poszczególnych części pracy. Punktem wyjścia dla realizacji prac badawczych musi być doskonała znajomość bieżącego stanu dziedziny, co w odniesieniu do rozważanej tematyki, wielowątkowej i intensywnie rozwijanej zarówno na polu teorii jak i zastosowań, oznacza konfrontację z ogromnym wolumenem koncepcji teoretycznych i ich praktycznych implementacji. Wymaganie to jest spełnione przez Doktorantkę w stopniu celującym, a z lektury pracy nieustannie przebija wyśmienita orientacja w szerokim zakresie obszarów tematycznych powiązanych z realizowanymi badaniami. Przedstawiony w rozdziałach drugim i trzecim pracy przegląd stanu wiedzy dziedziny jest wyczerpujący i kompleksowy, obejmując zarówno kompendium informacji na temat nielingwistycznych komponentów wypowiedzi, prezentację dziedziny przetwarzania i analizy sygnału akustycznego (przetwarzanie wstępne, wyznaczanie reprezentacji, redukcja wymiarowości, klasyfikacja), problematykę oceny wyników przeprowadzanych analiz oraz inwentaryzację istniejących technik detekcji występowania w sygnale mowy komponentów nielingwistycznych.

Niezwykle cennym wkładem Doktorantki na rzecz środowiska naukowego zajmującego się się prowadzeniem badań nad analizą sygnału mowy, szczegółowo opisanym w Rozdziale 4 rozprawy, jest przygotowanie przez nią baz nagrań mowy polskiej (samodzielnie lub przy jej istotnym udziale). Bazy dedykowane są różnym celom badawczym: dwie zawierają próbki mowy emocjonalnej (z odgrywanymi oraz spontanicznymi emocjami), dwie zorientowane są na prace nad oceną jakości wypowiedzi i brzmienia głosu, zaś ostatnia jest przeznaczona dla celów analizy pauz występujących w wypowiedziach. Metodyka projektowania baz, obejmująca całokształt środków podjętych w celu uwypuklenia istotnej dla założonego charakteru bazy treści, jest bardzo dojrzała i w pełni odpowiada standardom dziedziny. Przygotowane bazy danych zawierają pełną informację o wszelkich istotnych aspektach procesu rejestracji nagrań, a ich główną zaletą jest skojarzenie z nagraniami informacji opisowych, przygotowanych w trakcie manualnej procedury etykietowania materiału.

W stosunkowo krótkim, piątym Rozdziale rozprawy Doktorantka przedstawia trzy kategorie parametrów sygnału mowy, które nie były dotychczas powszechnie używane w procedurach analizy mowy. Niestety, z lektury rozdziału nie do końca jest jasne, na czym polega nowość lub autorski wkład w modyfikacje procedur wyznaczania deskryptorów dla kolejno przedstawianych podpunktów. Wydaje się, że w pierwszym przypadku, gdzie rozważanymi parametrami są średnie energie komponentów transformaty falkowej zawarte w odpowiednio zdefiniowanych pasmach, Autorka nie modyfikuje, a korzysta z zaproponowanego gdzie indziej deskryptora. Podobnie rzecz się ma w odniesieniu do ostatniego podpunktu, gdzie nie znajduję ani żadnej nowej koncepcji ilościowej oceny brzmienia głosu ani żadnej modyfikacji przytaczanych w podpunkcie deskryptorów ('ostrości', 'chropowatości' czy 'jasności'). Jedynie druga część podrozdziału jest spójna z zawartym w tytule rozdziału twierdzeniem o prezentacji nowych treści – Doktorantka przedstawia w nim algorytm parametryzacji kolejnych okien zawierających sygnał mowy, za pomocą współczynników biliniowej aproksymacji obwiedni (aproksymacji Pade) wyniku złożenia transformacji falkowej i transformacji Fouriera próbek sygnału. Efektem proponowanej procedury są wektory cech reprezentujące fragmenty mowy, które mają stanowić argumenty etapu klasyfikacji realizowanej albo na podstawie oceny odległości Euklidesa do wyznaczonych w analogiczny sposób wektorów referencyjnych (modeli klas), albo na podstawie oceny wartości iloczynów skalarnych z takimi modelami. Słabością proponowanej metody klasyfikacji jest nadmierna wymiarowość problemu: Doktorantka nie podejmuje istotnej dla metodologii klasyfikacji danych próby redukcji przestrzeni cech (choć opisuje odpowiednie metody w rozdziale 2.4.2 oraz stosuje takie metody w pracach opisanych w dalszej części).

Zasadnicza część osiągnięć Doktorantki została przedstawiona w szóstym rozdziale rozprawy, podzielonym na dwie części, odpowiadające różnym kategoriom analizowanych, niewerbalnych treści zawartych w mowie. Pierwszy wątek – analiza komponentów paralingwistycznych dotyczy w zasadniczej części detekcji pauz, wtrąceń oraz miejsc akcentowania, ukierunkowanych na opracowanie kompleksowej metody 'interpunkcyjnej' segmentacji wypowiedzi dla potrzeb automatycznej konwersji mowy na tekst. Drugi wątek to analiza przekazu nielingwistycznego, zmierzająca do identyfikacji ładunku emocjonalnego wypowiedzi oraz intencjonalnego wskazywania odpowiedniej interpretacji treści werbalnych. Z uwagi na znaczenie i obszerność materiału zawartego w Rozdziale 6, szczegółowe odniesienie się do zawartych tam treści zostanie przedstawione w następnym podpunkcie recenzji.

Ostatnia część rozprawy zmienia perspektywę prezentacji osiągnięć Doktorantki i dotyczy zastosowania opracowanych przez nią, lub przy jej wydatnym udziale, metod i algorytmów w rzeczywistych aplikacjach. W pierwszej z nich, budzącym moje autentyczne uznanie systemie profilowania w czasie rzeczywistym osób dzwoniących pod numer alarmowy, wykorzystany został opracowany przez Doktorantkę moduł rozpoznawania i śledzenia zmian emocji. Drugi obszar praktycznej weryfikacji opracowanych przez Doktorantkę metod analizy to aplikacja służąca do oceny postępów procesu treningu wokalnego. Aplikacja ta, dzięki określeniu ilościowych miar skorelowanych z jakością sposobu wypowiadania się, pozwala na obiektywizację oceny efektów treningu, umożliwiając indywidualizację jego tempa i stosowanych środków. Obydwie aplikacje zostały pomyślnie przetestowane w praktyce, stanowiąc niezwykle mocną weryfikację osiągnięć Doktorantki.

2.1 Ocena wkładu Doktorantki do dziedziny analizy sygnału mowy

Doktorantka formułuje w pracy dwie tezy badawcze oraz określa zbiór celów szczegółowych, wyznaczających kierunki prac prowadzonych dla uzyskania pozytywnej weryfikacji tych tez. Pierwsza teza pracy stanowi, że akustyczne deskryptory sygnału mowy pozwalają na identyfikację

paralingwistycznych i pozalingwistycznych komponentów przekazu zawartego w wypowiedzi. Druga, że klasyfikacja tych komponentów może być z powodzeniem dokonana z wykorzystaniem standardowych metodologii automatycznego rozpoznawania. Przedstawione brzmienie też nie jest dosłowną transkrypcją oryginału, ale zawiera w sobie najistotniejsze ich przesłanie (i pomija elementy o niezbyt klarownym sensie, np. 'universal cues', 'improve its performance' - względem czego?).

Ocenę naukowej wartości prac Doktorantki oraz potencjalnych korzyści płynących z uzyskanych wyników, a także merytorycznego poziomu rozważań i stosowanej metodologii badawczej, chciałbym przedstawić w formie odstępującej od chronologii prezentacji materiału rozprawy, subiektywnie szeregując jej osiągnięcia. Z tej perspektywy, jako dwa najważniejsze dokonania Doktorantki upatruję: przedstawienie kompleksowej metodyki rekonstrukcji struktury wypowiedzi, bazującej na analizie paralingwistycznych parametrów mowy oraz pionierskie prace nad identyfikacją pozalingwistycznego przekazu determinującego sposób interpretacji treści wypowiedzi.

2.1.1 Rekonstrukcja syntaktycznej struktury zdania i detekcja wtrąceń

Segmentacja wypowiedzi na frazy wyznaczone znakami interpunkcji wraz z eliminacją wtrąceń niepowiązanych z treścią jest jednym z podstawowych wymagań niezbędnych dla prawidłowego rozumienia zdań mowy naturalnej, dlatego też konsekwencje poradzenia sobie z tym niezwykle trudnym problemem są trudne do przecenienia. Doktorantka podchodzi do zagadnienia z pozycji analizy informacji zawartej w akustycznym sygnale mowy, a więc próbuje uzupełnić algorytmy rozpoznawania mowy w sposób pozwalający na generację tekstu semantycznie i syntaktycznie odpowiadającego wypowiadanemu treści. Trudność zadania podjętego przez Doktorantkę jest pochodną złożoności sygnału mowy, w którym przemieszane jest szerokie spektrum treści werbalnych i niewerbalnych, zawartych w wypowiedzi oraz dodatkowych, rozmaitych czynników zakłócających.

Nasuującą się podstawą rekonstrukcji struktury interpunkcyjnej zdania jest detekcja chwil występowania pauz. Doktorantka w wysmienity sposób rozprawia się z naiwnością takiego oczekiwania, przedstawiając niezwykle ciekawe wyniki analiz świetnie zaprojektowanych eksperymentów, mających na celu ocenę korelacji istniejących między znakami interpunkcji tekstu, wypowiadanego przez różne osoby w różnych kontekstach, a wykrywaniem w sygnale mowy pauzami. Zaprezentowane statystyki, dotyczące trzech rodzajów wyróżnianych pauz: międzywyrazowych, oddechowych i wtrąceń ('filled pauses'), wskazują generalnie niewielkie powiązanie rozważanych zdarzeń akustycznych z interpunkcją, wyraziście zarysowując skalę trudności problemu. W obliczu bezskuteczności prostych rozwiązań, Autorka proponuje szereg doskonałych pomysłów częściowych, których agregacja prowadzi do kompleksowej metodologii postępowania, dającej w efekcie rezultaty o praktycznym wymiarze.

Pierwszym elementem wzmiankowanego pakietu pomysłów jest koncepcja wykorzystania dynamiki zmian długości trwania głosek na przestrzeni wypowiadanych fraz. Hipoteza o tendencji wydłużania czasu artykulacji ostatnich głosek zdania zostaje przez Doktorantkę starannie zweryfikowana (wyniki eksperymentów podlegają rzetelnej weryfikacji statystycznej), dostarczając pierwszego wyróżnika granic wypowiedzi. Skupienie uwagi na korelacji dynamiki zmian parametrów czasowych głosek ze strukturą zdania, skutkuje kolejnym doskonałym pomysłem wykorzystania do tego samego celu tendencji występujących w ewolucji energetycznych parametrów akustycznych – zmniejszania mocy/energii dla kończących frazy głosek. Trzecim istotnym komponentem zbioru wyróżników granic fraz wypowiedzi są pauzy oddechowe. Detekcja pauz oddechowych, w odróżnieniu od zwykłych przerw międzywyrazowych, wymaga zupełnie innego podejścia, bowiem pauzom tym w sytuacji rejestracji dźwięku za pomocą nieodległego

mikrofonu, odpowiada specyficzna struktura sygnału akustycznego. Autorka proponuje heurystyczny algorytm detekcji pauz oddechowych, łączący analizę czasowo-energetyczną oraz spektralną, zbudowany na bazie wiedzy pozyskanej z przeprowadzonych przez nią wcześniej eksperymentów. Doktorantka uzupełnia zbiór tych deskryptorów o typowo stosowane w analizie mowy parametry, takie jak statystyki ewolucji częstotliwości tonu krtaniowego, współczynniki MFCC i inne, tworząc obszerną reprezentację sygnału mowy. Po poddaniu zbioru procedurze liniowej redukcji wymiarowości i przeprowadzeniu na tej podstawie eksperymentów detekcji granic fraz, uzyskuje rezultaty istotne z praktycznego punktu widzenia.

Na tle zasługujących na najwyższe uznanie wnikliwych i pomysłowych prac nad zbudowaniem omówionego zbioru deskryptorów, lekkim zaskoczeniem jest pominięcie przez Doktorantkę na etapie klasyfikacji danych tych metod, które uznawane są obecnie za najskuteczniejsze i użycie zamiast nich szeregu stosunkowo 'słabych' klasyfikatorów. Dlaczego zamiast kNN (a w istocie – NN), 'naive Bayes' czy liniowej analizy dyskryminacyjnej nie zostały uwzględnione SVM, GMM, lasy decyzyjne czy klasyfikatory wykorzystujące 'wzmacnianie' (np. Ada-Boost)? Użycie LDA w przestrzeni liniowej jest równoznaczne założeniu, że problem jest prosty, użycie NN jest równoważne założeniu, że istnieje 100% pewność, że zarówno procedura etykietowania jak i ekstrakcji cech jest bezbłędna, użycie podstawowej wersji klasyfikatora Bayesa jest równoważne założeniu 'unimodalności' i Gaussowskości klas. Żaden z trzech wymienionych warunków moim zdaniem nie występuje, a sposobem na poradzenie sobie z konsekwencjami niespełnienia powyższych wymagań byłoby użycie innych metod klasyfikacji.

W wątku prac skierowanym na doskonalenie systemów analizy języka naturalnego (NLP) i automatycznego rozpoznawania mowy (ASR), ciekawe i wartościowe są również badania Doktorantki nad pozostałymi kategoriami przekazu paralingwistycznego: detekcją występowania wtrąceń, analizą intonacji (akcentowania) oraz brzmienia głosu. W odniesieniu do pierwszego zagadnienia, Autorka dla dominującego w mowie polskiej wtrącenia 'yyy' weryfikuje przydatność specyficznych dla jego detekcji deskryptorów (stabilność formantów i wydłużony czas trwania głoski). Ponieważ analizie poddawane są głoski wstępnie zidentyfikowane (na podstawie formantów F1 i F2) jako 'y', nasuwa się pytanie, czy ten zabieg jest konieczny. Rezygnacja z tej procedury być może przyniosłaby możliwość detekcji również innych typów pauz, redukując jednocześnie skutki odrzucenia potencjalnej pauzy z uwagi na błędną identyfikację głoski. Oceniając znaczenie dokonań w obszarze detekcji wtrąceń, chcę stwierdzić, że chociaż ładunek oryginalności naukowej jest w tym przypadku niewielki, przeprowadzone prace są pionierskie z punktu widzenia analizy mowy polskiej, mając bardzo duże znaczenie praktyczne (pozwalając na przykład na eliminację w rozpoznawanej wypowiedzi zbędnych elementów). Również istotne z punktu widzenia praktycznego są wyniki prac Doktorantki w zakresie ilościowego modelowania rozmaitych (nieprecyzyjnych) kategorii określających brzmienie głosu, zastosowane w opracowanym i opisanym na końcu rozprawy programie do wspomagania treningu wokalnego.

2.1.2 Identyfikacja stosunku mówcy do przekazywanej treści

Niezwykłe interesującym wątkiem prac naukowych Doktorantki są pionierskie w odniesieniu do języka polskiego, ale również mało dotychczas obecne w badaniach światowych, zagadnienie identyfikacji niewerbalnego przekazu, istotnie modyfikującego jego treść, jakim może być ironia lub ugrzecznienie. Umiejętność dostrzegania tego subtelного poziomu przekazu jest zdecydowanym atrybutem wyższych czynności poznawczych, niezwykle pożądanym w opracowywaniu inteligentnych systemów komunikacji maszyna-człowiek. Doktorantka w zasługujący po raz kolejny na uznanie, poprawny metodologicznie sposób rozpoczyna badania od zaprojektowania serii eksperymentów, z których pierwszy ma na celu określenie punktu odniesienia dla prac - diagnozy skuteczności człowieka w rozpoznawaniu rozważanych kategorii przekazu niewerbalnego. Przy okazji podsumowania uzyskanych wyników identyfikuje zbiór stanów

emocjonalnych najczęściej mylonych z ironią, co może być wykorzystane jako dodatkowa informacja do wykorzystania na dalszych etapach analizy. Tak jak to miało miejsce w większości wcześniejszych eksperymentów, do reprezentacji sygnału mowy Doktorantka wykorzystała obszerną pulę deskryptorów akustycznych sygnału mowy, tym razem dokonując ich wstępnej selekcji na podstawie wyników przeprowadzonej analizy dyskryminacyjnej. Również tak jak poprzednio, z nieznanых powodów Doktorantka postanowiła okroić arsenał metod klasyfikacji danych i nie uwzględnić w nim algorytmów najskuteczniejszych.

Uzyskane rezultaty, mimo deklarowanego przez Doktorantkę dystansu, uważam za istotne, bowiem sformułowanie hipotezy o występowaniu w wypowiedzi ironii, nawet jeżeli jest to robione na umiarkowanym poziomie pewności, może być elementem kluczowym dla właściwego zrozumienia treści, które może być dokonywana na późniejszym etapie analizy. Podobnie wnikliwe i skrupulatne są badania związane z detekcją ugrzecznienia, przy czym część elementów wdrożonej procedury jest podobna jak w poprzednim przypadku.

2.1.3 Biometryczne wykorzystanie cech para-lingwistycznych

W toku prowadzonych przez Doktorantkę prac nad paralingwistycznymi komponentami przekazu, kilkakrotnie wskazuje ona biometryczną przydatność informacji zawartych w statystykach czasu trwania głosek, pauz, wtrąceń czy sposobie intonacji. Całkowicie podzielam ten pogląd, ale sugerowany przez Doktorantkę sposób wykorzystania wspomnianych parametrów w analizie biometrycznej uważam za niewłaściwy. Doktorantka stwierdza istnienie indywidualnego charakteru relacji między wspomnianymi deskryptorami, co prowadzi ją do sformułowania ryzykownej hipotezy o możliwości poprawy skuteczności rozpoznawania mówcy poprzez uwzględnienie wymienionych parametrów jako dodatkowych elementów wektora cech biometrycznych. Rezultaty przeprowadzonych eksperymentów są zimnym prysznicem, jednoznacznie podważającym rozbudzone oczekiwania (należy w tym miejscu wyrazić uznanie Doktorantki dla rzetelności przytaczania wyników 'niewygodnych'). Takie wyniki nie powinny dziwić, bowiem przyjęcie założenia o wystarczającej z punktu widzenia biometrii specyficzności czasu trwania głosek lub ich energii oznaczałoby, że problemy biometrii głosu już dawno powinny zostać pomyślnie rozwiązane. Powodów, dla których do biometrycznego wykorzystania wskazanych przez Doktorantkę parametrów należy podchodzić ostrożnie, jest wiele. Pierwszym z brzegu jest problem braku możliwości dokładnego wyznaczenia czasu trwania głoski (pauzy), skutkujące wprowadzeniem błędu reprezentacji. Kolejnym jest silne uzależnienie proponowanych deskryptorów od stanu emocjonalnego, modyfikujące ich wartości w stopniu nieporównywalnie większym niż w przypadku parametrów toru głosowego, odzwierciedlanych przez parametry standardowo stosowane w biometrii głosu. W konsekwencji, dodawanie problematycznych biometrycznie komponentów do modelu reprezentującego mówcę (do tzw. 'voiceprint') w najlepszym razie nie poprawia wyników analizy. Nie oznacza to jednak, jak wcześniej wspomniałem, że Doktorantka bezzasadnie doszukuje się w rozważanych parametrach biometrycznego potencjału. Aby go wykorzystać, należy zmienić perspektywę analizy biometrycznej na schemat weryfikacji, gdzie komponenty paralingwistyczne mogłyby służyć potwierdzaniu lub podważaniu hipotezy o tożsamości mówcy, sformułowanej na podstawie analizy 'klasycznej' (np. wspomnianą przez Doktorantkę metodą 'Identity vectors'). Należy podkreślić, że w swoich pracach Doktorantka w istocie stworzyła podstawy dla tworzenia biometrycznego modelu stylu wypowiedzania się, obejmującego tempo, melodię, styl akcentowania, wtrącenia, statystyki pauz, który mógłby okazać się niezwykle przydatny w sugerowanej, hipotetycznej analizie walidacyjnej, co można uznać za istotny wkład autorski.

2.1.4 Identyfikacja emocji na podstawie przekazu niewerbalnego

Prace nad oceną ładunku emocjonalnego wypowiedzi zrealizowane przez Doktorantkę zawierają się w kontekście niezwykle intensywnych badań, które zaowocowały bogatym bagażem koncepcji, rozwiązań i doświadczeń, chociaż nie doprowadziły jak dotąd do wypracowania żadnych standardowych metod o powszechnie zaakceptowanej skuteczności. Istotną motywacją dla podjęcia prac nad rozpoznawaniem emocji jest silne zapotrzebowanie praktyczne, co w konfrontacji z trudnością dokonania autentycznego wkładu merytorycznego, najczęściej skłania do powielania gotowych, umiarkowanie skutecznych przepisów.

Mimo obiektywnej trudności dokonania oryginalnego wkładu do dziedziny, Doktorantce udało się znaleźć miejsce na sformułowanie własnej, ciekawej propozycji, odnoszącej się nie tyle do samej procedury rozpoznawania, co do towarzyszącego jej aspektu - budowy nowego sposobu reprezentacji ładunku emocjonalnego, ukierunkowanego na specyficzny kontekst (rozmów kierowanych na numer telefonu alarmowego). Zaproponowany przez Doktorantkę dwuwymiarowy model reprezentacji emocji, wyznaczony w efekcie przeprowadzenia analizy asocjacyjnej zebranych przez Nią danych eksperymentalnych, zakłada parametryzację emocji względem dwóch kategorii pojęciowych: napięcia emocjonalnego i poczucia zagrożenia. W tak określonej przestrzeni cech lokuje następnie szereg kategorii emocji, adekwatnych dla rozważanego kontekstu aplikacyjnego (niepokój, stres, strach, panika, złość, frustracja itd.), a następnie próbuje skorelować je z ilościowymi deskryptorami sygnału mowy emocjonalnej. Zaproponowany model mógłby stanowić niezwykle interesującą podstawę dla procesu klasyfikacji emocji. Niestety, z dalszego ciągu przedstawionych rozważań wynika, że Doktorantka z niego nie korzysta, dokonując klasyfikacji emocji w szablonowy sposób: budując dla poszczególnych emocji modele statystyczne, a następnie wykorzystując je w procesie oceny podobieństwa do nieznanej próbki. Takie podejście nie wyróżnia się koncepcyjną oryginalnością i nie wykracza poza schemat postępowania typowy dla prac w rozważanej dziedzinie.

Jako podstawę klasyfikacji emocji Doktorantka testuje trzy sposoby reprezentacji sygnału mowy: współczynniki dyskretnej transformaty falkowej (DWT), kombinację współczynników DWT i DFT i wreszcie współczynniki MFCC. W pierwszym przypadku, klasyfikacja jest dokonywana na podstawie oceny odległości Euklidesa próbki do klasy, wykorzystującej trywialny model klasy, która jest zastępowana wartością średnią (co implikuje założenie o Gaussowskim charakterze rozkładu emocji w klasie, z jednostkową macierzą kowariancji). Nasuwa się więc tu pytanie, czy Doktorantka próbowała zweryfikować hipotezę o normalności rozkładu? Czy próbowała zweryfikować założenie o braku korelacji między cechami stosowanymi do reprezentacji mowy? W drugim eksperymencie, gdzie mamy do czynienia z zasadniczo dłuższymi wektorami cech, Doktorantka zdecydowała się użyć jako miary podobieństwa próbki do klasy wartości iloczynu skalarnego odpowiadających sobie wektorów. Wyniki są raczej słabe, przy czym wydaje się, że główną przyczyną może być tu zdecydowanie nadmierna wymiarowość danych w stosunku do objętości posiadanego zbioru przykładów, a także, jak w poprzednim przypadku, skrajnie uproszczony sposób modelowania klas. Biorąc pod uwagę słabe wyniki klasyfikacji, nie wiem dlaczego Doktorantka podsumowuje je twierdząc, że reprezentacja sygnału za pomocą kombinacji współczynników DWT-DFT stanowi obiecującą metodę identyfikacji emocji. W ostatnim z eksperymentów Doktorantka potwierdza prezentowane w wielu innych doniesieniach, przyzwoite wyniki klasyfikacji emocji z wykorzystaniem współczynników MFCC. Przy okazji opisu stosowanej przez Nią procedury wspomina o utworzeniu na podstawie danych z różnych baz, niezwiązanych z reprezentacją emocji, modelu UBM, co w kontekście celu jej analiz wymaga dodatkowego komentarza. Rozumiem, że Autorka zakłada neutralność emocjonalną wypowiedzi w tych bazach.

Podsumowując, ocena naukowa przedstawionego wątku tematycznego skłania do mieszanych wniosków. Z jednej strony mamy do czynienia z propozycją stworzenia interesującego punktu wyjścia dla zagadnienia rozpoznawania ładunku emocjonalnego w sygnale mowy, który zostaje jednak porzucony na rzecz szablonowej techniki postępowania.

2.2 Uwagi końcowe

Zaprezentowane w rozprawie badania Doktorantki nad analizą nielingwistycznego przekazu zawartego w mowie rozwinęły się w kilka wątków, w ramach których Doktorantka przedstawiła szereg interesujących i doskonale zweryfikowanych koncepcji. Wskazane w recenzji w odniesieniu do niektórych z nich mankamenty w niczym nie umniejszają całościowej, jednoznacznie pozytywnej opinii o niezaprzeczalnych osiągnięciach merytorycznych Autorki i istotnym wkładzie do dziedziny analizy sygnału mowy. Lista potencjalnych korzyści płynących z opracowanych przez Doktorantkę metod detekcji para-lingwistycznych i pozalingwistycznych komponentów przekazu słownego jest obszerna i z pewnością może być inspiracją lub pomocą dla dalszych prac badawczych prowadzonych w tym obszarze. Ogromnym atutem rozprawy jest metodologiczna dojrzałość: precyzyjne definiowanie celów, staranne przygotowanie materiału eksperymentalnego, analiza istniejącego stanu wiedzy i zwykle kompetentny wybór metod postępowania, a wreszcie, wzorowy sposób analizy uzyskiwanych wyników, uprawniający do formułowania istotnych statystycznie wniosków lub nowych, sprawdzanych dalej hipotez. Doktorantka niezwykle rzetelnie konfrontuje wyniki uzyskiwane w drodze stosowania jej koncepcji, z efektami osiąganymi znanymi metodami, nie ukrywając ponoszonych czasami porażek. Spora część prowadzonych przez nią badań ma charakter podstawowy i jest opatrzona ciekawą i wzbogacającą dyskusją, potwierdzającą lub negującą panujące poglądy oraz interesującymi interpretacjami o trudnych do przecenienia konsekwencjach.

Zaproponowane przez Doktorantkę koncepcje, ich implementacja i weryfikacja całkowicie wykazują słuszność sformułowanych w rozprawie tez. Przedstawione w pracy rozważania potwierdzają zarówno hipotezę o możliwości identyfikacji komponentów nielingwistycznych na bazie opisu sygnału mowy za pomocą znanych parametrów akustycznych, jak i hipotezę o możliwości prawidłowego rozpoznawania tych komponentów z wykorzystaniem znanych metodologii klasyfikacji danych.

3. Uwagi szczegółowe

Mimo, że praca napisana została starannie, Autorka nie ustrzegła się w niej od usterek technicznych rozmaitej natury. Dodatkowo, pewna grupa sformułowań i wywodów zawartych w pracy jest niedostatecznie jasna i wymaga doprecyzowania, lub trudno ją uznać za poprawną. Celem niniejszej części jest przedstawienie tych, szczegółowych zastrzeżeń. W zamieszczonej niżej, ułożonej chronologicznie liście, miejsce wystąpienia usterki (lub kontrowersyjnego ustępu) jest zwykle identyfikowane poprzez podanie numeru strony (a czasami, również numeru wiersza, w konwencji: 'g' - liczone od góry, lub 'd' - od dołu:

- Rys. 4.2, 4.3 – Skala osi pionowej jest błędna (jednostka to 0.2 procent lub 0.5 procent)
- str 60: rys.5.6 jest trudny do zrozumienia bez dodatkowych wyjaśnień – czemu odpowiadają krzywe kropkowane a czemu – ciągle z punktami pomiarowymi.
- Str. 61, w. 13g: nieuzasadnione logicznie wprowadzenie zdania („Thereby ...”)
- str. 63, w 3d: Autorka miała zapewne na myśli „role of pauses”, a nie ‘meaning of pauses’
- str 64, w 5d: Autorka miała zapewne na myśli Tab. 6.4 a nie Fig. 6.4.

- str. 66 Autorka zakłada Gaussowskość rozkładów długości trwania pauzy dla każdego z użytkowników (jest to warunek stosowności analizy ANOVA), która być może zachodzi, ale co do czego nie ma pewności.
- str. 67, w 3d: Autorka sugeruje wykorzystanie indywidualnych rozkładów prawdopodobieństw występowania pauz o różnej długości jako elementów modelu mieszanin Gaussowskich, które chciałaby wykorzystać w klasyfikacji. Nie bardzo rozumiem czego miałaby taka klasyfikacja dotyczyć (każdy komponent mieszaniny modeluje oczekiwane długości pauz dla mówcy, więc mieszanina reprezentuje zbiór mówców)
- rys. 6.10b jest kompletnie nieczytelny i wbrew sugestii, niczego na nim się w łatwy sposób nie da zauważyć.
- str. 72 – nie bardzo rozumiem jak ma się przedstawiony przez Autorkę schemat weryfikacji: jeden zbiór treningowy i cztery testowe do powszechnie przyjętego schematu walidacji krzyżowej
- str. 78 – tekst sugeruje, że rys. 6.22 powinien zawierać dwie części: odpowiadającą pauzie i odpowiadającą normalnej mowie, tymczasem zawiera tylko pierwszą z nich
- str. 80 ‘pattern parameters’ to chyba ‘reference parameters’?
- str. 81: ‘... wavelet ... parameters are counted’ - chyba nie o to chodzi.
- 81...: ‘sygnalization’, ‘sygnalize’, ‘sygnalization’, ‘Cryterias’, ‘summarization’ - to jest wzbogacanie angielskiego przez Doktorantkę, które chyba jednak się nie przyjmie
- 82 – tabela 6.15: jednostki energii to nie sekundy ...
- 84 ‘determinate’ zamiast ‘determine’; co Autorka ma na myśli pisząc: ‘executing the prototype’?
- Nie rozumiem wniosków podsumowujących parametry oceny procesu detekcji pauz. Powszechnie przyjęty sposób definiowania parametrów „Precision” (P) i „Recall” (R) to: $P = TP/(TP+FP)$, $R = TP/(TP+FN)$. Stosując te definicje, wyliczenia Doktorantki są niezrozumiałe ($TP=340$, $FP=513$). Wydaje się, że Autorka pomyliła obydwa terminy, chociaż sprawa jest bardziej zawiłana, bo najpierw twierdzi Ona, że całkowita liczba pauz wynosiła 340 oraz że TP, czyli liczba poprawnie wykrytych pauz to też 340. Skąd więc bierze się FN (czyli, liczba niewykrytych pauz) różne od zera? Natomiast bardzo pomysłowy i ciekawy jest eksperyment sprawdzający subiektywną istotność pomyłek typu FP, popełnianych przez zaproponowany algorytm.
- str. 92 Co to jest ‘phoneme level’?
- str. 96 6.1.3 Detekcja granic Z analiz przedstawionych przez Doktorantkę wynika istnienie tendencji wzrostowych lub spadkowych różnych parametrów. Nasuwa się więc tu również inna droga klasyfikacji danych – potraktowanie ich sekwencji wektorów (a nie wektorów) i zastosowanie metod klasyfikacji sekwencji (a więc np. metod HMM lub DTW). Czy Doktorantka rozważała takie podejście?
- str. 110: ‘automatize’
- str. 111: Histogramy rozkładu stosunku czasu trwania samogłosek względem wartości średniej są ewidentnie niegaussowskie, dlatego więc warto by poinformować, jakiego rodzaju test statystyczny został przeprowadzony.
- str. 112. Podpis pod rysunkiem nie jest prawidłowy: po pierwsze, zawiera on cztery elementy (a nie dwa, jak sugeruje podpis), po drugie, modele probabilistyczne przynależności samogłoski do klasy ‘akcentowana’ znajdują się u dołu (a nie po lewej). Swoją drogą, ciekawe czego dotyczą wykresy w lewej części rysunku.
- Str. 119: zamiast ‘tow’ powinno być ‘two’
- str. 123: podpisy pod osiami x rysunków powinny brzmieć „Parameter index” (bo stosowany podpis znaczy coś innego ...)
- str. 128. Jakie są różnice między uwarunkowaniami eksperymentów przedstawionych na rys. 6.64
- str. 134 Nie do końca jasne jest, jak wyglądała procedura redukcji wymiarowości wektora cech: czy analizie PCA poddane zostały wszystkie początkowe parametry, czy tylko ich

podzbiór, uznany w wyniku wcześniejszych testów statystycznych za istotny w odróżnianiu klas.

- str. 143 – kilka potknięć technicznych w tekście strony

4. Wniosek końcowy

Podsumowując przedstawioną recenzję, chciałbym jednoznacznie stwierdzić, że opiniowana praca doktorska Pani magister inżynier Magdaleny Igras-Cybulskiej pt. „Analysis of non-linguistic content of speech signals” spełnia z wyrażnym nadmiarem wymagania określone w odnośnej ustawie o stopniach i tytule naukowym i tym samym wnioskuję o **dopuszczenie Doktorantki do jej publicznej obrony**. Dodatkowo, biorąc pod uwagę wysoką wartość merytoryczną rozprawy, wnoszę jednocześnie o rozważanie możliwości **wyróżnienia rozprawy**.

Kinga Słot