Prof. dr hab. inż. Andrzej Dobrucki
Politechnika Wrocławska
Wydział Elektroniki
Katedra Akustyki i Multimediów
Wybrzeże Wyspiańskiego 27
50-370 Wrocław
Tel. 71 320 30 68
E-mail: andrzej.dobrucki@pwr.edu.pl

Wrocław 28.02.2017

Recenzja rozprawy doktorskiej

Mgr. inż. Radosława Weychana

zatytułowanej:

Speaker recognition based on transcoded speech for human-machine interfaces

1. Problem badawczy i jego znaczenie

Jaki jest najważniejszy problem rozważany w rozprawie?

Rozprawa dotyczy problemu rozpoznawania mówców na podstawie krótkich zdań lub komend w zastosowaniu do komunikacji człowiek-maszyna. Zdania, na podstawie których następuje rozpoznawanie mówcy mogą wcześniej podlegać kodowaniu stratnemu i być transmitowane przez kanały telekomunikacyjne, na skutek czego następuje degradacja ich jakości.

Czy ma on charakter naukowy?

Rozpoznawanie mówcy należy do klasycznych problemów na styku akustyki i informatyki. Zagadnieniu temu poświęcona jest ogromna liczba publikacji. Również problem rozpoznawania mówców na podstawie sygnałów o obniżonej jakości jest współcześnie szeroko rozwijane. Zadanie ma więc niewątpliwie charakter naukowy, zwłaszcza, że Autor przedstawia w rozprawie doktorskiej szereg oryginalnych pomysłów przyczyniających się do poprawy efektywności rozpoznawania.

Czy ma on znaczenie praktyczne?

Poprawa efektywności rozpoznawania mówcy w zastosowaniu do interakcji człowiek-maszyna ma bardzo duże znaczenie praktyczne. Mogą wystąpić dwa rodzaje błędów: fałszywe rozpoznanie mówcy, który nie wypowiada zadanej (lub swobodnej) sentencji lub fałszywe nierozpoznanie mówcy wypowiadającego rzeczoną sentencję. W zależności od zastosowania dąży się do uniknięcia jednego lub drugiego rodzaju błędu, wystąpienie błędów jest jednak nieuniknione. Autor do oceny efektywności algorytmów rozpoznawania przyjął miarę równego prawdopodobieństwa wystąpienia obu rodzaju błędów (punkt EER na krzywej DET – Detection Error Tradeoff – kompromis detekcji błędów).

2. Wkład autora

Jaki jest najważniejszy wkład autora opisywane w rozprawie?

Aby ocenić wkład Autora, należy dokonać krótkiego przeglądu treści pracy. Rozprawa doktorska składa się z pięciu rozdziałów, z których pierwszy stanowi wprowadzenie, zaś ostatni – podsumowanie. We wprowadzeniu Autor przedstawia cele pracy oraz formułuje jej tezę. Teza ta ma postać:

Efektywność rozpoznawania mówców na podstawie sygnałów niskiej jakości (szybkość próbkowania 8 kS/s, rozdzielczość 10 bitów przy przetwarzaniu analogowo cyfrowym) może być poprawiona poprzez:

- Zastosowanie algorytmów aktywności mowy (VAD)
- Rozpoznanie techniki kodowania
- Wybór właściwego modelu mówcy z bazy danych mówców.

Teza ta jest sformułowana zasadniczo poprawnie, chociaż w rozprawie doktorskiej należałoby unikać wchodzenia w szczegóły, takie jak określenie szybkości i rozdzielczości przetwarzania analogowo-cyfrowego. Brak jest mi również określenia w tezie istotnego praktycznego rezultatu badań jakim jest porównanie realizacji algorytmów we wbudowanych systemach działających w oparciu o procesory pracujące w arytmetyce zmiennoprzecinkowej i stałoprzecinkowej. W rozdziale 1 Autor przedstawia znaczenie problematyki rozpoznawania mówców dla dyscypliny Automatyka i Robotyka, a także układ pracy.

W rozdziale 2 Autor przedstawia stan wiedzy związany z przedmiotem rozprawy, w tym metody parametryzacji sygnału mowy oraz metody klasyfikacji, metody określania aktywności głosowej (VAD), standardy kompresji stratnej sygnału mowy, a także metodykę określania dokładności algorytmów. Rozdział ten jest oparty zasadniczo o literaturę, chociaż Autor sam zrealizował i przetestował oprogramowanie, w którym zastosował opisane algorytmy. Jako metodę parametryzacji wybrano metodę wyznaczania współczynników melowo-cepstralnych, z zastosowaniem dyskretnej transformaty kosinusowej. Jest to popularna metoda parametryzacji sygnału mowy. Wśród algorytmów klasyfikacji rozważano ich kilka, ale w dalszej części skupiono się na metodzie kwantyzacji wektorowej (VQ) oraz mieszanin gaussowskich (GMM). Stosowano też rozwinięcie modelu GMM-UBM (mieszaniny gaussowskie – uniwersalny model tła). Metody te są również często stosowane w technice rozpoznawania mówców. Autor stwierdził, że efektywność algorytmów rozpoznawania mówców zmniejsza się na skutek występowania w próbkach – zarówno referencyjnych, jak i badanych fragmentów ciszy. Redukcja efektywności jest tym większa, im jakość próbek mowy jest gorsza. Z tego względu Autor zdecydował się na stosowanie algorytmów VAD – rozpoznawania aktywności głosowej, a następnie usuwania z próbek fragmentów, w której aktywności tej nie było, czyli fragmentów ciszy. W rozdziale 2 wdrożono i przebadano 4 algorytmy VAD: algorytm oparty na obliczaniu energii w ramkach sygnału, oraz trzech algorytmów opracowanych przez Janga. W dalszej części rozprawy stosowano głównie algorytm „energetyczny" oraz algorytm HOD (różnic wyższego rzędu) Janga. W dalszej części rozdziału 2 Autor prezentuje standardy kodowania z zastosowaniem kompresji stratnej sygnału mowy. Dzieli je na dwie grupy: standardy stosowane w telefonii GSM oraz standardy stosowane w systemach multimedialnych. Te drugie stosowane są nie tylko do kodowania mowy, ale też, a może przede wszystkim, do kodowania sygnałów fonicznych, czyli również muzyki. Nic dziwnego, że jakość kodowanych za pomocą tych algorytmów mowy, jest na ogół wyższa, niż standardów GSM. Generalnie, Autor stwierdza, że włączenie do modelu mówcy stosowanych koderów wpływa na skuteczność rozpoznawania. W kolejnej części tego rozdziału omawiane są miary dokładności (skuteczności) algorytmów rozpoznawania. Jak wspomniano wyżej, Autor mierzył w dalszym ciągu jakość rozpoznawania za pomocą wyznaczania punktu EER na wykresie DET. Jest to również podejście standardowe. Rozdział 2 jest rozdziałem opracowanym na podstawie literatury, czyli wykazującym orientację Autora w temacie rozprawy i znajomości współcześnie stosowanych metod.

Zasadniczymi rozdziałami rozprawy, będącymi całkowicie dorobkiem Autora są rozdziały 3 i 4. Rozdział 3 dotyczy zaproponowanych przez Autora udoskonaleń algorytmów rozpoznawania mówców biorących pod uwagę niską jakość rozpoznawanych sygnałów. Jako przedmioty rozpoznawania, a także do uczenia modeli Autor wykorzystał wypowiedzi znajdujące się w dostępnej światowej bazy TIMIT, oraz opracowanej przez siebie bazy wypowiedzi polskich SPU. Należy

zaznaczyć, że stosowana jest identyfikacja niezależna od wypowiadanego tekstu. Autor zastosował i rozwinął to idee zasygnalizowane w rozdziale poprzednim idee. I tak, Autor wdrożył i gruntownie przebadał wpływ zastosowania wykrywania aktywności głosowej na skuteczność rozpoznawania. Badane było usuwanie ciszy na początku i na końcu wypowiedzi, a także, jako osobne zadanie usuwanie ciszy również w środku wypowiedzi (przerwy między wyrazami). Przebadane były różne (wymienione w rozdziale 2) algorytmy VAD, oraz algorytmy klasyfikacji VQ (kwantyzacja wektorowa) oraz GMM (mieszaniny gaussowskie). Bardzo istotny okazał się wpływ szybkości próbkowania na efektywność rozpoznawania. Wnioski wynikające z tych badań nie zawsze są oczywiste, np. dla szybkości próbkowania 8 kS/s rozpoznawanie było lepsze dla próbek niepoddanych algorytmowi VAD (usuwanie ciszy na krańcach próbki dźwiękowej), dla klasyfikatora VQ. Klasyfikacja GMM daje znacznie większe różnice rozpoznawania w zależności od szybkości próbkowania. Stosując algorytm GMM rozpoznawalność jest lepsza dla lepszej jakości próbek niż dla klasyfikatora VQ, ale znacznie gorsza dla próbek gorszej jakości (mała szybkość próbkowania). Kolejnym udoskonaleniem wprowadzonym przez Autora rozprawy do algorytmu rozpoznawania jest włączenie do modelu mówcy informacji o rodzaju kodera i sposobie kodowania. Badania przeprowadzono osobno dla kodowania GSM z jego standardami FR (Full Rate), HR (Half Rate), EFR (Enhanced Full Rate) i AMR (Adaptive Multi-Rate) oraz różnych algorytmów kodowania multimedialnego (MP3, Ogg Vorbis, WMA i AAC). Dla kodowania GSM zbadano wpływ długości ramki na efektywność wykrywania rodzaju kodera. Następnie, wprowadzono model kodera do algorytmu rozpoznawania mówcy. Założono, że wielokrotne kodowanie wprowadza coraz mniejszą degradację sygnału. Stwierdzono, że najbardziej degraduje sygnał (mierzony jako stosunek sygnał/szum) koder AMR, najmniej – kodery FR i EFR. Na podstawie eksperymentów opracowano hiperboliczny model zależności między błędem średniokwadratowym a numerem współczynnika MFCC. Model sprawdza się bardzo dobrze przy wielokrotnym kodowaniu. Przeprowadzono eksperymenty rozpoznawania mówców z uwzględnieniem typu kodowania. Ogólnie, błędy rozpoznawania (mierzone wartością EER) są duże, ale mogą być zmniejszone przez uwzględnienie w algorytmie modelu kodera. Określono czas obliczeń z użyciem różnych algorytmów VAD. Okazuje się, że algorytm „energetyczny" usuwania ciszy jest bardzo efektywny i dla bazy SPU pracuje nawet szybciej, niż algorytm bez usuwania ciszy z wnętrza próbki. Nota bene, Autor używa tu określenia „middle" na usuwanie ciszy z wnętrza. Wydaje mi się, że lepszym określeniem byłoby słowo „internal" lub „content", ponieważ „middle" znaczy tyle co „środkowy", a tu chodzi bardziej o „wewnętrzny". W końcowej części rozdziału 3 oceniana jest skuteczność rozpoznawania przy użyciu różnych kodeków multimedialnych w zależności od szybkości transmisji (przepływności bitowej). Jak należało się spodziewać, jakość sygnału mowy kodowana za pomocą kodeków multimedialnych jest wyższa niż z użyciem kodeków GSM, co w istotny sposób zmniejsza wartość EER. Podobnie, jak w przypadku koderów GSM badano skuteczność rozpoznawania mówców przy użyciu innego kodeka na etapie nauki (treningu), a innego na etapie testowania. Stwierdzono zależność rozkładu współczynników MFCC zarówno od kodeka, jak i przepływności. Skuteczność rozpoznawania, nie zawsze jest największa na etapie treningu i testowania, z wyjątkiem kodera Ogg, dla którego zdecydowanie najmniejsza wartość EER wystąpiła, jeśli OGG był użyty zarówno na etapie treningu, jak i testów (rys. 3.65). W rozdziale 3 opisano bardzo wiele eksperymentów, których wyniki mają praktyczne znaczenie. Omówiono je powyżej dość pobieżnie, ponieważ szczegółowe omawianie tych wyników spowodowałoby znaczne wydłużenie recenzji. Na podkreślenie zasługuje fakt samodzielnego zaprojektowania oraz przeprowadzenia wszystkich eksperymentów przez Doktoranta, co wymagało ogromnego nakładu pracy.

Kolejny, 4 rozdział pracy dotyczy sprzętowego wdrożenia zaproponowanych powyżej, udoskonalonych algorytmów jako systemów wbudowanych, pracujących w czasie rzeczywistym. Jak wspomniałem wyżej, było to jednym z celów pracy, ale nie znalazło odzwierciedlenia w tezie

rozprawy. W rozdziale 4 Autor rozważa zarówno problemy dotyczące wyboru typu procesora (zmienno- lub stałoprzecinkowy) oraz jego szczegółowego modelu, a także wyboru odpowiedniego języka programowania. Ostatecznie testowo system zmiennoprzecinkowy zbudowany w oparciu o platformę Raspberry Pi2, wyposażoną w czterordzeniowy procesor ARM Cortex-A7 i 1 GB pamięci RAM. Jako języka programowania użyto języka Python, ze względu na biblioteki tego języka, w istotny sposób wspierające obliczenia. Zrealizowano i przetestowano opisane wyżej algorytmy, oraz oceniono ich efektywność. Na podstawie przeprowadzonych eksperymentów opracowano zależność wiążącą definiowaną funkcję kosztu odzwierciedlającą czasochłonność obliczeń z błędem identyfikacji charakteryzowanym przez wartość EER. System rozpoznawania mówców zrealizowano również na procesorze sygnałowym TMS320C5515 pracującym ze stałym przecinkiem. Do programowania użyto w tym przypadku środowisko MATLAB. System o arytmetyce stałoprzecinkowej zmniejsza koszt eksploatacji w czasie rzeczywistym, ale z kolei, ze względu na ograniczony zakres dynamiki procesora, błędy obliczeń są większe. Zbadany został wpływ długości słowa na efektywność rozpoznawania mówcy. Długość ta (16 lub 32 bity) bardzo istotnie wpływa na EER przy zastosowaniu klasyfikatora VQ, natomiast prawie nie wpływa przy stosowaniu mieszanin gaussowskich. Należy jednak stwierdzić, że błąd rozpoznawania dla klasyfikatora VQ jest zdecydowanie mniejszy. Zbadano działanie różnych algorytmów aktywności głosowej.

Rozdział 5 stanowi podsumowanie pracy. Autor skrótowo przedstawił swoje osiągnięcia, a także oszacował wzrost efektywności rozpoznawania mówców uzyskany w wyniku wdrożenia modyfikacji zaproponowanych algorytmów. W szczególności, zmniejszenie dokładności obliczeń uzyskane poprzez zastosowanie 16 bitowego stałoprzecinkowego procesora przy 10-bitowej rozdzielczości danych wejściowych nie wpływa zasadniczo na dokładność rozpoznawania, natomiast zwiększenie dokładności uzyskane w wyniku modyfikacji algorytmów, głównie o zastosowanie VAD i rozpoznawanie koderów, jest znaczne.

Stwierdzam, że własny wkład Autora rozprawy doktorskiej do problematyki rozpoznawania mówców w oparciu o wypowiedzi o zdegradowanej jakości jest znaczny. Autor wykonał olbrzymią pracę, aby wdrożyć i przebadać swoje idee.

3. Poprawność

Rozprawa przygotowana jest bardzo starannie. Wszystkie stwierdzenia i wnioski oparte są o wyniki szczegółowo opisanych eksperymentów. Z tego powodu uważam je za wiarygodne. Mam pewne uwagi do prezentacji wyników. Wykresy przedstawiające zależności FAR/FRR, zawierają często dużą liczbę krzywych, przez co są trudne do wizualnej analizy, mimo opisu i stosowania różnych kolorów i typów linii do krzywych. Dlatego, w takich przypadkach wskazany byłby komentarz w jaki sposób przedstawione na wykresach zmienne wpływają na jakość rozpoznawania. Takie komentarze czasem występują, a czasem nie i trzeba samemu analizować wykresy, co bywa dość żmudne.

4. Wiedza kandydata

Jak wskazałem wyżej, rozdziały 1 i 2 powstały głównie w oparciu o literaturę. Autor rozprawy nie ograniczył się do zacytowania źródeł, ale też wdrożył i przebadał znane algorytmy, które w następnych rozdziałach były modyfikowane. Autor wykazał się bardzo dużą wiedzą w zakresie własnego przedmiotu rozprawy, czyli rozpoznawania mówców, ale także w tematyce zbliżonej, jak parametryzacja sygnałów mowy i algorytmy klasyfikacji. Ma też dużą wiedzę dotyczącą kodowania mowy i technicznych rozwiązań w tym zakresie. Wykazuje wiedzę z zakresu realizacji sprzętowej i programowej algorytmów, a także umiejętności praktycznego jej zastosowania. Rozdziały 3 i 4 są działami, w których na podstawie wiedzy o charakterze erudycyjnym przedstawionej w rozdziałach

1 i 2 realizuje w sposób twórczy i zarazem konsekwentny własne pomysły. Rozdział 5 jest rozdziałem podsumowującym rozprawę. Znajduje się tam deklaracja o udowodnieniu tez rozprawy, z czym się zgadzam. W rozprawie Autor cytuje 142 pozycje bibliograficzne. Kompletna bibliografia dotycząca problemów poruszanych w rozprawie liczyłaby wiele tysięcy pozycji i oczywiście nie jest możliwe jej opanowanie przez jednego badacza. Wiele pozycji literatury cytowanej w rozprawie ma charakter encyklopedyczny, czyli jest podręcznikami, w których podsumowano istniejącą wiedzę. Takie podejście też uważam za uzasadnione. Inne z kolei cytowane publikacje są materiałami firmowymi dotyczącymi sprzętu i oprogramowania lub standardami dotyczącymi np. kodowania sygnałów mowy. W cytowanej bibliografii znajduje się też 12 prac Autora rozprawy, opublikowanych na ogół z Promotorem, prof. Adamem Dąbrowskim i innymi badaczami. Prace te opublikowano w języku angielskim w czasopismach i na konferencjach.

5. Inne uwagi

Rozprawa doktorska napisana została w języku angielskim. Język rozprawy oceniam pozytywnie. Jedynym konsekwentnie stosowanym niepoprawnym określeniem jest „parametrization". Poprawnie po angielsku powinno być „parameterization". Poprawne określenie „parameterize" występuje na str. 77, wiersz 4 od dołu. Czasem użyłbym innego słowa niż Autor, np. wskazanego już wyżej słowa „middle" dotyczącego usuwania fragmentów ciszy z wewnętrznych części wypowiedzi. Na str. 56 użyty jest termin „capacitor microphone". W literaturze przedmiotu używa się raczej „condenser microphone". Na tej samej stronie Autor opisuje eksperyment, którego elementem jest nagrywanie mówcy w komorze bezechowej. Stanowisko do nagrań przedstawione jest na rys. 3.5. Nie ma informacji, jaki był wpływ blatu stołu na nagrany sygnał. Czy nie byłoby celowe przykrycie stołu choćby kocem, w celu zmniejszenia odbić od blatu? Z innych pomyłek wymieniłbym

- Na str. 14 współczynniki cepstralne (wzór 2.5) są oznaczane tak samo, jak próbki sygnału (wzór 2.1), tzn. $s(n)$.
- Na str. 32, wiersz 9 od dołu sklejono dwa wyrazy „partsconsists".
- Na str.112 wiersz 11 od dołu jest niezrozumiałe słowo „aoerating". Chodzi chyba o „operating"?
- Na str. 130 występuje „rage" zamiast „range" (wiersz 9 od dołu).
- Na str. 140, w ostatnim wierszu punktu 7, nazwisko powinno być „Bayes" zamiast „Beyes".
- Na str. 147, pod wzorem (4.1) liczba składników mieszaniny gaussowskiej oznaczona jest zarówno przez N_{GMM}, jak i N_{FFT}. W tym drugim przypadku jest to oczywista pomyłka.
- Na str. 172, w cytowanej pozycji 41, występuje literówka" diagramsa" zamiast „diagrams".

Biorąc pod uwagę objętość pracy pisanej w obcym (jak sądzę) dla Autora języku, liczba tych pomyłek jest zadziwiająco mała. Pomyłki i niezręczności w żadnym przypadku nie umniejszają wartości pracy, którą oceniam bardzo wysoko.

6. Podsumowanie

Biorąc pod uwagę opinie zaprezentowane w poprzednich punktach i wymagania zdefiniowane przez artykuł 13 Ustawy z dnia 14 marca 2003 r. o stopniach naukowych i tytule naukowym (z późniejszymi zmianami)[1] moja ocena rozprawy pod względem trzech podstawowych kryteriów jest następująca:

[1] http://www.nauka.gov.pl/g2/oryginal/2013_05/b26ba540a5785d48bee41aec63403b2c.pdf

A. Czy rozprawa zawiera oryginalne rozwiązanie problem naukowego? (wybierz jedną opcję stawiając znak **X**)

[X]	[ ]	[ ]	[ ]	[ ]
Zdecydowanie TAK	*Raczej TAK*	*Trudno powiedzieć*	*Raczej NIE*	*Zdecydowanie NIE*

B. Czy po przeczytaniu rozprawy zgadzasz się, że kandydat posiada ogólną wiedzę teoretyczną w dyscyplinie Informatyka lub Automatyka i Robotyka?

[X]	[ ]	[ ]	[ ]	[ ]
Zdecydowanie TAK	*Raczej TAK*	*Trudno powiedzieć*	*Raczej NIE*	*Zdecydowanie NIE*

C. Czy kandydat umiejętność samodzielnego prowadzenia pracy naukowej?

[X]	[ ]	[ ]	[ ]	[ ]
Zdecydowanie TAK	*Raczej TAK*	*Trudno powiedzieć*	*Raczej NIE*	*Zdecydowanie NIE*

Ponadto, biorąc pod uwagę wysoki poziom naukowy rozprawy oraz ogromny nakład pracy wykonany przez Autora rozprawy, co wykazałem powyżej w recenzji rekomenduję wyróżnienie rozprawy doktorskiej.

Andrzej Dokurek

Podpis