prof. dr hab. inż. Krzysztof Marasek
Katedra Multimediów
PJATK
ul. Koszykowa 86
02-008 Warszawa

Warszawa, 2.03.2017

Recenzja

rozprawy doktorskiej mgr. inż. Michała Lenarczyka pt. „Akustyczne i fonetyczne metody przemiany głosu"
Promotor: prof. dr hab. inż. Ryszard Tadeusiewicz

Niniejszą recenzję opracowano na zlecenie Rady Naukowej Instytutu Podstaw Informatyki PAN z dn. 16.01.2017.

1. Charakterystyka tematu, celu i tez badawczych rozprawy

Mowa jako medium komunikacji międzyludzkiej fascynuje badaczy z wielu dziedzin, także informatyków, którzy starają się stworzyć narzędzia pozwalające rozpoznawać i zrozumieć ludzkie wypowiedzi, ale także syntetyzować i zmieniać ludzki głos. Praca p. mgr M. Lenarczyka dotyczy tej ostatniej dziedziny, a doktorant postawił sobie ambitny cel stworzenia metody transformacji głosu mówcy na głos docelowy, odpowiadający cechom innego mówcy. W pracy skupiono się na przemianie cech krótkookresowych – wysokości i barwy głosu, z zamiarem stworzenia systemu działającego w czasie rzeczywistym. Temat ten jest aktualny i potrzebny, choć większość prac dotyczy raczej rozwoju technik na potrzeby wielojęzycznej syntezy mowy (dokonywana transformacja głosu ma stworzyć poliglotę – ten sam głos mówiący w wielu językach). Miarą zainteresowania tym tematem jest zorganizowanie w 2016 konkursu na temat przemiany głosu w ramach światowej konferencji Interspeech, w którym wzięło udział 17 instytutów badawczych.
Teza pracy jest następująca: „Możliwe jest uzyskanie efektu przemiany głosu przez zastosowanie wyselekcjonowanej w pracy parametryzacji sygnału mowy i opracowanych metod uczenia maszynowego. Przy określonych założeniach możliwa jest realizacja tego zadania w czasie rzeczywistym".
Aby obronić tą tezę Autor wykorzystał cztery reprezentacje obwiedni widma (logarytmiczne stosunki przekroju liniowego modelu predykcyjnego (LPC), częstotliwości widma prążkowego LPC, cepstralne współczynniki modelu homomorficznego, współczynniki MFCC modelu linowego) i dwie metody transformacji (sztuczną sieć neuronową o topologii perceptronu oraz klasyfikator wykorzystujący metodę wektorów wspierających SVM). Wyniki zostały przetestowane poprzez badania odsłuchowe, które wskazały na dokonaną zmianę głosu przy stosunkowo niskiej ocenie jakościowej. Moduł oprogramowania dokonujący transformacji głosu jest rozwiązaniem autorskim, natomiast wyznaczenie funkcji transformacji metodami uczenia maszynowego dokonano wykorzystując zewnętrzne pakiety obliczeniowe.

Zagadnienie badawcze uważam zatem za dobrze postawione i aktualne, choć narzucone przez Autora ograniczenia wyraźnie wpłynęły na osiągnięte wyniki.

2. Zawartość rozprawy

Recenzowana praca jest wyjątkowo obszerna - składa się z 6 rozdziałów, polskiego i angielskiego streszczenia, bibliografii, skorowidza, cztero-częściowego suplementu oraz płyty CD z próbkami dźwiękowymi. Dokument liczy w sumie 188 stron (plus 16 stron opisu wykresów). Zamieszczona bibliografia pracy liczy 128 pozycji, z których większość to artykuły z czasopism naukowych i konferencji. W literaturze znalazłem też odwołania do dwóch pozycji Autora, z których jedna to jego praca magisterska, a druga to krótka praca konferencyjna. W pracy nie zamieszczono spisu ilustracji ani wykazu skrótów i oznaczeń, jest za to (skrótowy) indeks pojęć.
Struktura pracy jest logiczna. Praca rozpoczyna się wstępem (rozdział pierwszy) przedstawiającym analizowany problem badawczy, postawione cele i tezę pracy oraz strukturę rozdziałów pracy. Dwa następne rozdziały mają charakter analityczno-teoretyczny. Rozdział drugi stanowi wprowadzenie do fonetyki akustycznej, metod analizy i reprezentacji mowy oraz metod transformacji głosu i oceny jej jakości. Rozdział trzeci opisuje metody ekstrakcji cech z sygnału i rekonstrukcji sygnału na ich podstawie przy spełnionych założeniach modelu powstawania mowy źródło-filtr. Rozdział czwarty opisuje zastosowane metody transformacji cech osobniczych w konwersji mowy oraz wykorzystywany zbiór danych. Ocena jakości i skuteczności zaproponowanej transformacji głosu opisano w zaskakująco zwięzłym rozdziale piątym (10 stron w porównaniu do 30-40-stronnicowych poprzednich rozdziałów). Pracę wieńczy krótkie podsumowanie, w którym zawarto także opis ewentualnych dalszych kierunków badań.
W podsumowaniu nie odniesiono się do postawionej tezy pracy, choć podkreślono wypełnienie celu pracy, to jest stworzenie systemu konwersji głosu działającego z krótkim opóźnieniem.
Rozbudowany suplement składa się z czterech części. Dodatek A opisuje w sposób encyklopedyczny proces artykulacji mowy. Dodatek B definiuje używany w pracy zapis fonetyczny (bazujący na fonetycznym alfabecie słowiańskim). Część C przedstawia ogólny opis właściwości układu akustycznego wykorzystywanego do modelowania toru głosowego i jego analogu elektrycznego (związanego z teorią obwodów). Wreszcie część D suplementu opisuje autorską implementację systemu konwersji głosu, jego architekturę oraz analizę opóźnienia wprowadzanego przez system (maksymalnie do 70 ms). Dodatkowo, na płycie CD, zamieszczono próbki dźwiękowe, do których dostęp zapewniono poprzez dokument pdf.

3. Uwagi krytyczne i wskazówki dotyczące rozprawy

Przede wszystkim chciałbym się tu odnieść do zidentyfikowanych luk badawczych.

I. Praca p. mgr Lenarczyka dotyczyć miała konwersji głosu w czasie rzeczywistym (przy minimalnym opóźnieniu). De facto jednak konieczny jest etap wydobycia parametrów głosów źródłowego i docelowego i cechy te wydobywane są off-line, w procesie uczenia systemu. No ale przecież tak działają dowolne inne systemy konwersji głosu, które raz nauczone mogą działać na danych strumieniowych (np. vokodery fazowe). Nie rozumiem tej różnicy, proszę o wyjaśnienie i porównanie z innymi systemami.

II. W pracy nie ma odniesień do innych, możliwych do zastosowania systemów konwersji głosu. W szczególności, tak jak we wspomnianym w pracy The Voice Conversion Challenge 2016, takim benchmarkiem mógłby być darmowy system konwersji głosu zawarty w pakiecie Festvox.

III. Przeprowadzone eksperymenty ewaluacyjne uważam za bardzo ograniczone – w testach odsłuchowych brało udział 5 lub 9 osób odsłuchujących, którym prezentowano po 60 testów ABX i MOS, przy czym w pracy nie podano jak długie były odtwarzane nagrania. Warto porównać te liczby ze wspomnianym VC challenge lub choćby odnieść się do dowolnego podręcznika statystyki -Trudno tak ograniczone badania uznać za reprezentatywne

IV. Opis implementacji (dodatek D) uważam za niewystarczający – jeśli celem pracy jest minimalizacja opóźnienia konwersji, to spodziewałbym się w rozprawie pogłębionej analizy złożoności obliczeniowej zadania.

V. Przegląd literatury i istniejących rozwiązań jest stosunkowo ograniczony. Brakuje mi odniesień do prac prof. Hirose, szczególnie tych wykorzystujących rachunek tensorowy czy też prac wykorzystujących sieci neuronowe z mechanizmem uwagi np. neural vocoder [4]

VI. Konkluzja „Patrząc ogólnie na wyniki uzyskane we wszystkich badaniach, należy jasno stwierdzić, że jakość głosu po przemianie jest niezadowalająca" oraz „... zaś od tego czasu dokonano licznych poprawek i ulepszeń. Wskazane wydaje się poddanie systemu przemiany głosu ponownej ocenie odsłuchowej" (str.128) wskazuje na świadomość Autora, iż złożona rozprawa jest niekompletna.

Poniżej zamieściłem uwagi i pytania szczegółowe, dotyczące poszczególnych aspektów pracy, podając je według kolejności stron (ew. rozdziałów):

- str.4 w sformułowaniu tematu pracy doktorant pisze „Innymi słowy: czy głos można uznać za cechę niepodrabialną, biometryczną? Odpowiedź brzmi nie". To bardzo odważna teza, idąca wbrew nie tylko obiegowym opiniom, ale i wynikom pracy wielu naukowców i instytucji. Niestety, w całej pracy nie znalazłem przeprowadzonego dowodu prawdziwości tego stwierdzenia, więc uważam je za gołosłowne. Z kolei zdanie na temat adaptacji systemów rozpoznawania mowy jest nieścisłe, bowiem istnieją systemy adaptacji dostosowujące modele akustyczne, a nie tylko cechy akustyczne.
- str. 8 „Znacznie trudniej zmienić wysokość czy barwę głosu, które wynikają z budowy narządów głosowych i w niewielkim stopniu podlegają świadomej kontroli" - Obie wspominane cechy jak najbardziej podlegają świadomej kontroli, można przecież zmieniać i sposób artykulacji i tryb fonacji i to na dowolnie długo.
- str. 25 i wcześniej - Model źródło-filtr jest od dawna uważany za dyskusyjny. Odsyłam choćby do prac Titzego [6], przy czym sprzężenie obu elementów zależy od stopnia otwarcia głośni i dotyczy głównie wyższych głosów. W wielu trybach fonacji głośnia pozostaje (częściowo) otwarta, indukując istotne zmiany w charakterystyce generowanego dźwięku (np. pogorszenie dobroci pierwszego rezonansu) czy zmiany nachylenia obwiedni widma. Efekt ten oczywiście nie może być prawidłowo modelowany przez liniowe modele predykcyjne. Nieco tu spekuluję, ale być może wybór głosów z bazy CORPORA był poniekąd nieszczęśliwy – wybrano głosy w których głośnia nie była w pełni zamykana, co z kolei mogło prowadzić do błędów modelowania, a co za tym idzie obniżenia jakości wynikowej konwersji głosu. Wybór taki warto było oprzeć o analizę GCI (glottal closure instant). Sprzężenie źródło-filtr wpływa też na barwę głosu.
- W pracy wspominano o fazach artykulacji głosek, nie przeanalizowano jednak wpływu prozodii na artykulację, a w szczególności pozycji akcentowanej/nieakcentowanej

- str. 16 –„ informacja o fazie jest pomijana, ponieważ ucho ludzkie jest zasadniczo nieczułe na różnice fazy między składowymi ..." nie jest ścisłe. Odsyłam do literatury, gdzie wielokrotnie wykazano, że faza silnie wpływa na percepcję, a w szczególności zrozumiałość mowy. Dlatego też wiele systemów TTS wykorzystuje także informację o fazie sygnału.
- str. 18 – nie wchodząc w szczegóły, stwierdzenie, że słyszalna wysokość tonu wiąże się z odstępem pomiędzy kolejnymi harmonicznymi jest nieścisłe. Jest kilka teorii percepcji wysokości tonu (pitch), dla rekreacji polecam zapoznanie się z „Huggins pitch"
- str. 19 – głoska /t/ w fazie zwarcia jest „niemal" zupełną ciszą - dlaczego niemal? Co generuje dźwięk, jeśli nie ma przepływu powietrza?
- str. 23 – EGG daje tylko informacje o powierzchni styku fałd głosowych, a nie o przepływie powietrza.
- str. 25 – prozodyczne właściwości sygnału mowy nie były brane pod uwagę. To jest oczywiście silne ograniczenie, natomiast jest jeszcze jedna cecha krótkookresowa, która umknęła prowadzonym analizom, mianowicie nazalizacja. W ogóle uważam za słabość pracy brak pojęcia „ustawienia" (*setting* za Laver'em [7]), które silnie wpływa na artykulację, także w reżimie krótookresowym.
- str. 27 i kolejne – brakuje we wzorach wyjaśnienia czynnika N (rząd predykcji), dyskusyjne jest także pominięcie współczynnika wzmocnienia toru. Zastanawia mnie brak w tym rozdziale (2.3) wzmianki o PLP, która to parametryzacja sygnału mowy wydaje się dobrze powiązana z odbiorem dźwięku i stanowi pomost pomiędzy LPC a metodami MFCC.
- str. 32 – szkoda, ze pisząc o falkach Autor nie wspomniał o constant-Q, która to metoda znacznie ułatwia manipulacje głosem i mogłaby pomóc w zwalczaniu problemów efektu fazowego (str. 81) [3]
- rozdział 2.5 jest generalnie dobrze napisany, chciałbym jednak zwrócić uwagę na metodologię testów MUSHRA (ITU-R BS.1534-3) używaną, o ile wiem, także do oceny konwersji głosu, a pozwalająca na efektywniejsze testy.
- str. 49 – dobór rzędu modelu LPC na potrzeby analizy mowy nie jest zadaniem trywialnym i akurat ten temat warto byłoby przeanalizować dokładniej. Zwracam uwagę, że dobór ten powinien uwzględniać cechy mówcy [1,2].
- str. 59 – Rysunek 3.10 jest nieco mylący, bo zestawia ze sobą amplitudy filtrów pasmowych i współczynniki MFCC, które nie mają bezpośredniego związku interpretacyjnego z filtrami. Pierwsze współczynniki MFCC można jeszcze interpretować jako stosunki mocy w pasmach filtrów melowych, ale trudno zauważyć jakieś odpowiedniości dla współczynników wyższego rzędu.
- str. 67 i dalsze – wyznaczanie współczynnika dźwięczności przyjmującego wartości z przedziału <0,1> jest pomysłem ciekawym i czytelnie uzasadnionym. W pracy nie znalazłem jednak wyjaśnienia, jak taki współczynnik zachowuje się dla głosek zwartych (dźwięcznych i bezdźwięcznych).
- str. 73 i dalsze opisują wokoder fazowy i jego zastosowanie. Ta część jest dobrze napisana, wnioski są prawidłowe, jeszcze raz wyrażę w tym miejscu żal, że Autor nie zastosował metody [3] do ograniczenia efektu fazowego.
- str. 84 w pracy założono, że anotacje korpusu CORPORA są prawidłowe. Przytoczony na tej stronie fragment anotacji ma regularne przesunięcie początku segmentu względem końca poprzedniego o 5 ms. Z czego to wynika? Czy to mogło mieć jakiś wpływ na transformacje? Czy,

ze względu na wagę tej etykietyzacji w procesie uczenia systemu transformacji głosu (str.90), nie należało przeprowadzić testów choćby jednorodności transkrypcji?
- str. 85 – „materiał dźwiękowy... podzielono na rozdzielne podzbiory przeznaczone do uczenia i testowania": czy nie wybrano zatem oddzielnego zbioru do bieżącej ewaluacji postępów uczenia (learn, validation, test)? Na str. 88 jest mowa o zbiorze walidacyjnym...
- str. 98 – po raz kolejny Autor wspomina o tym, czego nie zrobił – w tym wypadku chodzi o metodę analizy składowych głównych.
- Str. 100 – odnośnik do pracy [38] jest chyba nie na temat, aczkolwiek zgadzam się, że funkcja taka jest używana też w analizie mowy
- rozdział 4.4.2 dotyczący wykorzystania sieci neuronowej w przemianie głosu nie budzi wielu uwag, jest dobrze napisany. Moim zdaniem wybór prostej topologii sieci jest tu akurat rozsądny, choć (może) zastosowanie autoenkodera mogłoby dać sugestię co do sensowności zaproponowanego zestawu cech opisujących obwiednię widma
- str.111 – „... obliczenia prowadzące do wyniku można znaleźć w licznych źródłach ...." Warto byłoby podać odnośnik do tych źródeł.
- rozdział 4.5 – dobrze napisany, ale rysunki 4.14-4.19 są nieczytelne, co ma z nich wynikać?
- str. 123 - Moje zastrzeżenia do części eksperymentalnej wyraziłem już wcześniej, ale proszę o wyjaśnienie, czy w części testów dotyczących SVM brały udział te same osoby co w pierwszej części (SNN). To bowiem wpływa na stosowane metody analiz statystycznych.
- str. 123 „ W obydwu eksperymentach transformację tonu zrealizowano przy pomocy metody parametrycznej" Dlaczego? Mamy w pracy ciekawy rozdział na temat wokodera fazowego, Autor wskazał w nim na dokonane przez niego modyfikacje usprawniające działanie, a teraz tego nie zastosował? Proszę o wyjaśnienie.
- str. 124 – jaki wyniki w skali MOS otrzymały głosy źródłowe, bez przekodowania? Dlaczego, pomimo wiedzy, iż testowanie samego kodowania mogło wpłynąć na ocenę jakości głosu po transformacji nie przeprowadzono dodatkowych testów?
- str. 134 w końcowych wnioskach Autor sugeruje użycie metody DTW do zrównoleglenia nagrań, choć na str. 89 pisze „ W pracy zrezygnowano z ... DTW .. na rzecz prostszej i bardziej niezawodnej metody, wykorzystującej dostępną anotację czasową korpusu"

4. Ocena strony redakcyjnej

Od strony edytorskiej praca jest bardzo dobra. Właściwie nie znalazłem w niej literówek, błędnych oznaczeń, czy odnośników. Także stylistycznie praca jest na dobrym poziomie. Sam układ treści także nie budzi większych zastrzeżeń, choć moim zdaniem można było w niej zrezygnować z opisów powszechnie znanych technik i zastąpić je odnośnikami do literatury. Sugerowałbym szerszy opis innych technik konwersji głosu dających znakomite rezultaty (np. [4], [5]). W podsumowaniu pracy zabrakło odniesienia się do postawionej w niej tezy.

5. Wnioski końcowe recenzji

Trudno o jednoznacznie pozytywną ocenę przedstawionej od recenzji rozprawy mgr. inż. Michała Lenarczyka.

Z jednej strony praca jest dobra warsztatowo, doktorant wykazał się w niej wiedzą z zakresu cyfrowego przetwarzania sygnałów, ale z drugiej strony praca sprawia wrażenie kończonej w znacznym pośpiechu, bez solidnego doświadczalnego udokumentowania wyników zaproponowanych metod konwersji głosu.

Uważam, że rozprawa wykazuje zdolność doktoranta do samodzielnego rozwiązania złożonego problemu badawczego przy użyciu nowatorskich metod i własnych osiągnięć (parametryzacja mowy na potrzeby konwersji głosu, ciągły współczynnik dźwięczności, stosowanie metod maszynowego uczenia).

Z punktu widzenia recenzenta rozsądne wydaje się uzupełnienie części doświadczalnej pracy o nowe testy odsłuchowe z udziałem większej liczby testerów i zaimplementowanych, ale nie przetestowanych w złożonej dysertacji metod obróbki sygnału oraz eksperymenty, które jasno wykażą zalety estymacji współczynnika dźwięczności jako wielkości ciągłej (co Autor przedstawia w konkluzji pracy jako najważniejsze osiągnięcie badawcze).

W konkluzji stwierdzam, że po zaproponowanych uzupełnieniach opiniowana rozprawa spełni wymogi stawiane przez ustawę o stopniach i tytułach naukowych dla prac doktorskich, a jej autor zasługuje na przyznanie stopnia naukowego doktora nauk technicznych. Stawiam więc wniosek o dopuszczenie tej rozprawy do publicznej obrony.

Odnośniki

[1] http://www.rle.mit.edu/soundtosense/conference/pdfs/fulltext/Saturday%20Posters/SB-Vallabha-STS.pdf
[2] http://clas.mq.edu.au/speech/acoustics/speech_spectra/fft_lpc_settings.html http://www1.se.cuhk.edu.hk/~lfsun/
[3] http://www.cs.tut.fi/sgn/arg/klap/schoerkhuber-JAES-2013.pdf
[4] http://www.josesotelo.com/speechsynthesis/
[5] https://www.microsoft.com/en-us/research/project/voice-conversion-with-neural-network/
[6] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2811547/
[7] Laver, J., "The Gift of Speech", University Press,. Edinburgh, 1991.

prof. dr hab. inż. Krzysztof Marasek
Katedra Multimediów
PJATK
ul. Koszykowa 86
02-008 Warszawa

Warszawa, 12.11.2017

Uzupełnienie recenzji

rozprawy doktorskiej mgr. inż. Michała Lenarczyka pt. „Akustyczne i fonetyczne metody przemiany głosu"
Promotor: prof. dr hab. inż. Ryszard Tadeusiewicz

W dniu 8.11.2017 otrzymałem aneks zawierający dwa dodatki do rozprawy. W dodatku E przedstawiono opis syntezy sygnału z użyciem metody *true envelope*, a w dodatku F zawarto wyniki dodatkowych testów odsłuchowych porównujących parametryczne i nieparametryczne metody kodowania w przemianie głosu.

Dodatek E zawiera kompetentny opis estymacji obwiedni widma sygnału rekurencyjną metodą kolejnych przybliżeń kształtu widma sygnału. Ze względów obliczeniowych wykorzystano metodę Roebla i Rodeta, której opis zmodyfikowano w stosunku do oryginalnego artykułu, tak aby był on zgodny z tematyką rozprawy. W drugiej części tego dodatku wskazano na zastosowanie tej metody w przemianie głosu, co zilustrowano na rysunkach przedstawiających parametryczną analizę i syntezę sygnału dla głosu kobiecego, męskiego i chłopięcego.

Dodatek F podsumowuje dodatkowe testy odsłuchowe w których przeanalizowano jakość transformacji głosu przy użyciu wokodera fazowego i sieci neuronowej (zastosowanej do cech obwiedni otrzymanych metodą true envelope). Starannie opisano przeprowadzone eksperymenty, w których tym razem wzięła udział duża grupa testerów (22 osoby). Wyniki przedstawiono w postaci testów preferencji (ABX) oraz MOS. Wynika z nich polepszenie skuteczności przemiany głosu przy wykorzystaniu transformacji obwiedni w wokoderze fazowym, natomiast parametr jakościowy (rys. F.2) pozostał niski, bądź pogorszył się. Testy preferencji przy wykorzystaniu techniki rozszerzania pasma wokodera fazowego wykazały ich lepszą percepcję. Wreszcie rys. F.4 podsumowuje preferencje słuchaczy dla kodowania parametrycznego i wokodera fazowego. W tym wypadku wyniki nie są jednoznaczne. Pewnym brakiem w tej części jest nieobecność testów statystycznych wskazujących na istotność otrzymanych wyników. Cieszy mnie sugestia wykorzystania w przyszłych pracach testu MUSHRA, którego wykorzystanie zaproponowałem w recenzji.

Podsumowując, uważam że przedstawiony aneks w sposób wyczerpujący odpowiada na zastrzeżenia sformułowane w mojej recenzji i rozwiewa wątpliwości. Nie mam zatem żadnych obiekcji co do dopuszczenia rozprawy p. Lenarczyka do dalszych etapów postępowania.

W konkluzji stwierdzam, że po przedstawionych uzupełnieniach opiniowana rozprawa spełnia wymogi stawiane przez ustawę o stopniach i tytułach naukowych dla prac doktorskich, a jej autor zasługuje na przyznanie stopnia naukowego doktora nauk technicznych. Stawiam zatem wniosek o dopuszczenie tej rozprawy do publicznej obrony.