prof. dr hab. inż. Tomasz P. Zieliński
Katedra Telekomunikacji
Wydział Informatyki, Elektroniki i Telekomunikacji
Akademia Górniczo-Hutnicza
Kraków

Kraków 12.04.2016

R E C E N Z J A

rozprawy doktorskiej pt.

Wykorzystanie szerokopasmowej matrycy wielomikrofonowej w rozpoznawaniu mówcy

Pana mgr inż. Rafała Samborskiego

Streszczenie

1. Krótka charakterystyka rozprawy

Recenzowana rozprawa doktorska dotyczy zagadnienia automatycznego rozpoznawania mówcy podczas operacji automatycznej analizy nagrań rozmów wielu osób, np. podczas spotkań konferencyjnych (próba odpowiedzi na pytanie: kto kiedy mówi). W terminologii angielskiej problem ten nosi nazwę diaryzacji (anotacji, etykietowania) nagrań, tzn. ich podziału na oddzielne wypowiedzi poszczególnych mówców. Zwykle rozwiązanie tego zadania sprowadza się do segmentacji nagrania na części, czyli oddzielenia od siebie wypowiedzi różnych mówców, a następnie przyporządkowania poszczególnych fragmentów poszczególnym osobom.

Na stronie 15 Autor pracy stawia następującą tezę:

„Wykorzystanie kilku strumieni cech w znaczący sposób polepsza skuteczność systemu diaryzacji nagrań. Poprzez dynamiczny dobór proporcji pomiędzy informacją pochodzącą z klasycznego systemu identyfikacji mówcy opartego o cechy częstotliwościowe (MFCC) a informacją związaną z lokalizacją mówcy (TDOA) następuje znacząca poprawa wyników algorytmu w stosunku do istniejących rozwiązań."

W literaturze światowej istnieją podobne rozwiązania do tego, które jest zaprezentowane w pracy, łączące algorytm rozpoznawania mówcy i estymację położenia mówcy względem mikrofonów. W szczególności w pracy [5] z 2007 roku, cytowanej w rozprawie. Dokonuje się w niej fuzji metody częstotliwościowej MFCC i przestrzennej TDOA (dokładnie GCC-PHAT), **automatycznie** zmieniając wartości współczynników wagowych na podstawie BIC (*Bayesian Information Criterion*). Ważenie stałe i dynamiczne stosuje się także w pracy D. Vijayasenana i F. Valente [79] z 2011 roku. Dynamicznie zmienne wartości wag są w tym przypadku proporcjonalne do odwrotności entropii. To samo rozwiązanie, tylko bez automatycznego ważenia, jest przedstawione w pracach [57, 58] (2007, 2012), autorów pracy [5]. W metodach tych także wykorzystuje się detektor występowania mowy w rejestrowanym sygnale (VAD), tak jak w recenzowanej rozprawie.

Podstawową różnicą rozwiązania zaproponowanego w doktoracie (rys. 6.4) w stosunku istniejących rozwiązań, podanych powyżej, jest **estymacja na bieżąco stosunku sygnału do szumu SNR oraz**

dynamiczna zmiana współczynników wagowych obu strumieni z wykorzystaniem krzywej z rysunku 7.5 (linia przerywana). Osiągnięte dzięki temu zmniejszenie błędu diaryzacji nagrań (w funkcji SNR) jest pokazane na rysunkach 7.6 i 7.7. W tabelach 7.1 i 7.2 porównano zaproponowaną metodę z alternatywnymi podejściami, znanym z literatury, uzyskując podobne wyniki. Moim zdaniem **na korzyść zaproponowanej metody w stosunku do rozwiązań konkurencyjnych przemawia jej skuteczność i prostota.**

2. Zawartość merytoryczna rozprawy

Praca składa się z ośmiu rozdziałów, w tym wstępu i podsumowania, bibliografii (nieuwzględnionej w spisie treści) oraz spisów: rysunków (3 strony), tablic (1 strona), skrótów i oznaczeń (2 strony) – łącznie liczy ona 94 strony. W pracy brak jest wymaganego streszczenia w języku angielskim. Układ rozprawy jest następujący.

W rozdziale pierwszym, wstępie (6 stron), na początku podkreślono rosnącą rolę systemów wielomikrofonowych oraz podano typowe przykłady ich zastosowań (m.in. lokalizacja źródeł dźwięku, redukcja zakłóceń i kształtowanie kierunkowości odbioru). Następnie skrótowo na tym tle zaprezentowano cechy sygnału mowy, które są wykorzystywane do rozpoznawania mowy lub mówcy (rys. 1.1 [43]: spektralne, prozodyczne, wysokopoziomowe) oraz przedstawiono wyniki badań światowych, pokazujących jak bardzo skuteczność rozpoznawania pojedynczych fonemów zależy od poziomu szumu (tabela 1.1 [46]). Na tym tle zaprezentowano cel pracy: poprawę jakości sygnału wejściowego do klasycznego algorytmu rozpoznawania mowy/mówcy metodą cyfrowego przetwarzania sygnałów, pochodzących z macierzy mikrofonowej, a nie z pojedynczego mikrofonu. Zaproponowano także zrobienie fuzji metod częstotliwościowych rozpoznawania mówców (cechy MFCC) z estymacją położenia mówców (wykorzystanie uogólnionej korelacji wzajemnej (GCC) i mikstur gaussowskich). Oryginalnością tego podejścia w stosunku do podobnych znanych z literatury [5] miał być dynamiczny dobór wartości współczynników wagowych dla dwóch strumieni informacyjnych **w zależności od warunków akustycznych**. Postawiono tezę, że takie postępowanie zwiększy skuteczność rozpoznawania w stosunku do układów o stałych wartościach wag. Pod koniec rozdziału przedstawiono strukturę pracy.

W rozdziale drugim „Matryce wielomikrofonowe" (11 stron) przedstawiono podstawowe pojęcia dotyczące matryc wielomikrofonowych: aperturę i jej rodzaje, funkcję wrażliwości i charakterystykę kierunkową apertury oraz aliasing przestrzenny. W szczególności podano wzory dla apertury liniowej ciągłej i dyskretnej dla źródła w polu dalekim.

W rozdziale trzecim „Lokalizacja źródeł akustycznych" (11 stron) przedstawiono narzędzia obliczeniowe, które były wykorzystywane w pracy do estymacji położenia mówcy na podstawie sygnałów z pary mikrofonów. W pierwszej części, bazując na [44], podano definicję zwykłej i uogólnionej (ważonej) funkcji korelacji wzajemnej GCC (3.11) oraz poszczególne rodzaje GCC dla różnych funkcji ważących: normalizacji Rotha (3.13), wygładzonej transformacji koherentnej (3.16) i transformacji fazowej (GCC-PHAT) (3.19). Natomiast w drugiej części rozdziału wytłumaczono poglądowo na konkretnym przykładzie rzeczywistym jak filtr adaptacyjny, dopasowujący dwa sygnały do siebie (w układzie ACC - *adaptive correlation canceller*), może być wykorzystany do wyznaczenia różnicy odległości pomiędzy dwoma mikrofonami a lokalizowanym źródłem dźwięku (poprzez znalezienie maksimum jego odpowiedzi impulsowej). Pod koniec przedstawiono rozwiązanie problemu dla przypadku wielu par mikrofonów, czyli algorytm SRP [89].

W rozdziale czwartym „Filtracja adaptacyjna" (11 stron) przedstawiono wybrane zagadnienia dotyczące ww. filtracji. Wyprowadzono równania adaptacyjnego filtru Wienera [83], zaprezentowano ich zapis zaproponowany przez Levinsona [48] oraz podano iteracyjny, rekurencyjny algorytm Shermana-Morrisona (SM) [70] do implementacji równań Wienera w praktyce. Następnie przedstawiono najważniejsze równania, dotyczące filtracji LMS i NLMS oraz porównano złożoność obliczeniową algorytmów filtracji adaptacyjnej SM, LMS i NLMS (tablica 4.1, str. 46). Wspomniano o filtracji w dziedzinie częstotliwości (FDAF) oraz podpasmach (SAF). Pod koniec rozdziału zaprezentowano typowy, dwumikrofonowy układ adaptacyjny do poprawy jakości sygnału mowy (rys. 4.3, str. 48) z układem VAD (*Voice Activity Detector*), który włącza adaptowanie się filtru w przypadku kiedy w

sygnale nie ma mowy tylko same zakłócenia. W pracy zastosowano układ VAD pracujący w dziedzinie częstotliwości [21] oraz filtr adaptacyjny o długości N=256 próbek z algorytmem SM. **Dla kilkugodzinnych nagrań w różnych warunkach akustycznych (zakłóceniach) filtr adaptacyjny, pracujący jako klasyczny układ ACC, poprawiał SNR sygnału mowy średnio o 2.9 dB, podczas gdy zastosowanie korelacji wzajemnej i metody *delay-sum* – o około 2 dB. Na tej podstawie stwierdzono, że tak mały zysk okupiony dużo wyższą złożonością obliczeniową algorytmu filtra adaptacyjnego nie uzasadnia jego zastosowania w dalszej części badań.** Wyciągnięty wniosek jest poprawny, ale tylko w odniesieniu do zmniejszenia zakłóceń adaptacyjnym układem ACC, a nie kształtowania kierunkowości macierzy mikrofonowych.

W rozdziale piątym „Kształtowanie wiązki" (10 stron) przedstawiono klasyczne algorytmy kształtowania kierunku wiązki metodą sumowania opóźnionych (*delay-sum*) albo filtrowanych (*filter-sum*) sygnałów z kilku mikrofonów. Badany układ składał się z 4 mikrofonów oddalonych od siebie o 7 cm, filtry miały 1 (K=0, *delay-sum*) lub 10 (K=9, *filter-sum*) współczynników. Zadano charakterystyki kierunkowe, a następnie wyznaczono zestawy współczynników filtrów metodą minimalizacji funkcji kosztu (5.19), wykorzystując iteracyjny algorytm Neldera-Meada [55] do nieliniowej optymalizacji bez ograniczeń (bez konieczności obliczania pochodnych). Wykazano symulacyjnie i pomiarowo, że w przypadku dostrojenia układu *filter-sum* (po 5506 iteracjach) listki boczne charakterystyki kierunkowo-częstotliwościowej badanego zestawu mikrofonów obniżają się o 5dB. Nie wykorzystano powszechnie stosowanych, adaptacyjnych układów: LCMV (*Linearly Constrained Minimum Variance*), zaproponowanego przez Frosta [28,7], oraz GSC (*Generalized Sidelobe Canceller*).

W rozdziale szóstym „Matryce wielomikrofononowe w rozpoznawaniu mówcy" (11 stron) przedstawiono główne wyniki rozprawy, czyli autorskie propozycje układowe, zaproponowane przez Doktoranta. Na początku krótko scharakteryzowano literaturę z zakresu współczesnych systemów telekonferencyjnych oraz diaryzacji nagrań. Potem opisano klasyczny algorytm Reynoldsa [62] rozpoznawania mówcy, wykorzystujący modelowanie współczynników melcepstralnych (MFCC) za pomocą mikstur gaussowskich (GMM).
Następnie zaproponowano „unikatowe rozwiązanie nieopisane dotąd w literaturze" (rys. 6.4, str. 69), czyli dodatkowe użycie uogólnionej korelacji wzajemnej z normalizacją fazową (GCC-PHAT) do określenia położenia mówcy oraz **oryginalne połączenie obu algorytmów**: (MFCC+GMM) i (GCC-PHAT+GMM). Sygnał z mikrofonu o największej energii podawano na algorytm MFCC+GMM i dokonywano klasyfikacji mówcy. Równocześnie obliczano wartości TDOA za pomocą GCC-PHAT dla trzech par mikrofonów {1,2}, {2,3} i {3.4}. Trzyelementowe wektory poddawano także modelowaniu GMM i na tej podstawie podejmowano decyzję o lokalizacji mówcy, czyli ponownie go rozpoznawano. Następnie dokonywano **dynamicznej fuzji** obu wyników klasyfikacji (6.15) (średnia ważona) **z odpowiednimi wagami, dobieranymi zgodnie z (7.3) w zależności od zmierzonego poziomu zakłóceń SNR** (pomiaru dokonywano podczas braku wypowiedzi mówców – korzystano z układu VAD). Dla każdego mówcy obliczano logarytmiczny wskaźnik wiarygodności LLR, że wypowiada się właśnie ten mówca. Osobno dla cech częstotliwościowych i położenia. Wybierano największą ważoną sumę. Wraz z poprawą jakości sygnału większe znaczenie odgrywał strumień danych, związany z lokalizacją mówcy.

W rozdziale siódmym „Wyniki eksperymentów" (14 stron) przedstawiono rezultaty z przeprowadzonych badań eksperymentalnych. Pomieszczenie miało wymiary 6,4 na 3,3 metra. Nad stołem na wysokości 0,5 metra były umieszczone w linii 4 mikrofony oddalone od siebie o 10 centymetrów. Było 6 mówców, z których w każdej sesji losowano pięciu uczestników rozmowy. Mówcy byli oddaleni od mikrofonów od 1 do 2,5 metra. Całkowity czas nagrań wynosił 28 minut. Wypowiedzi mówców trwały od 1 do 15 sekund. Wypowiedź treningowa dla każdego z mówców wynosiła 60 sekund. Podczas rozpoznawania nagrania były dzielone na fragmenty jednosekundowe z przesunięciem 250 ms (dla GCC: ramka 500ms co 250ms, dla MFCC: ramka 25ms co 10ms). Sygnały były próbkowane 16-bitowo z częstotliwością 16 kHz. Poziom SNR był równy 30 dB. Stworzona i wykorzystywana baza nagrań pod względem technicznym spełniała wymagania *National Institute of Standards and Technology* (NIST), ale poziom trudności był mniejszy, gdyż zawsze wypowiadał się tylko jeden mówca. Do oceny wyników wykorzystano miarę diaryzacji DER (7.1), zdefiniowaną przez NIST, dla zastosowanej

prostszej bazy redukujący się do (7.2). Przeprowadzono testy z zakłóceniem typu przyjęcie (*cocktail party*) oraz biuro (*office*):

- tylko dla rozpoznawania MFCC+GMM (32 komponenty); wyniki na rys. 7.2: DER=30% dla SNR > 20 dB;
- tylko dla rozpoznawania GCC+GMM (1 komponent); wyniki na rys. 7.3: DER=25% dla SNR > 17,5 dB, ale DER jest o wiele większy dla SNR<17,5dB niż dla metody MFCC+GMM;
- dla rozpoznawania łącznego; wówczas podczas treningu jest dobierany najlepszy zestaw wag dla obu strumieni (rys. 7.4), z którego wynika optymalna krzywa przełączania wag w zależności od zmierzonej wartości SNR – równanie (7.3) i rys. 7.5.

Rysunki 7.6 i 7.7 są głównym dowodem korzyści, wynikających z zastosowania rozwiązania zaproponowanego w recenzowanej pracy: zmniejszenie DER z 30% do wartości około 17,5% dla metody hybrydowej, zawsze jest lepiej niż dla stałych wartości wag.

Do tej pory w algorytmach opisanych powyżej macierz mikrofonowa była tylko wykorzystywana do estymacji TDOA i lokalizacji mówcy, a nie do poprawy jakości sygnałów przed analizą MFCC. W dalszej części rozdziału wykorzystano informację TDOA do przesunięcia względem siebie sygnałów z poszczególnych mikrofonów i do ich zsumowania. Otrzymano w ten sposób uzdatnianie sygnału metodą *delay-sum*. Potem pracując na takim sygnale przeprowadzano rozpoznawanie mówcy metodą MFCC+GMM. Otrzymane wyniki pokazano na rys. 7.8: zmniejszenie DER o 10% dla SNR<0dB, ale równocześnie zwiększenie DER o kilka procent dla SNR>15dB. **Szkoda, że w tym miejscu pracy nie zastosowano adaptacyjnych algorytmów LCMV i GSC kształtowania wiązki typu *filter-sum*, bardziej odpornych na zmieniające się warunki akustyczne!** Otrzymany wynik mógłby być lepszy.

W tablicy 7.1 porównano wyniki z pracą [78], osobno dla MFCC i GCC. Pomimo tego, że w bazie NIST wartość SNR jest o 15 dB mniejsza, metoda Vijayasenana et al. [78] tylko dla cech MFCC ma dwukrotnie mniejszy DER niż podobna implementacja autora, a dla cech GCC zbliżony, co dziwi. Prosiłbym Doktoranta o komentarz podczas obrony.

W ostatnim, ósmym rozdziale, podsumowaniu (2 strony), podkreślono zalety zaproponowanego rozwiązania oraz, co szczególnie cenne, wymieniono kierunki dalszych badań, które mają wyeliminować niedoskonałości lub ograniczenia obecnego rozwiązania (udoskonalenie przeprowadzania fuzji decyzji, np. zastosowanie głębokich sieci neuronowych, optymalizacja liczby mikrofonów i ich rozmieszczenia, zastosowanie innych niż GMM metod modelowania mówców, zastosowanie innych algorytmów rozpoznania mówców, np. iVectors, zastosowanie zaawansowanych algorytmów kształtowania wiązki, np. LCMV, GSC).

Pracę zamyka **wykaz cytowanej literatury** (8 stron), liczący 89 pozycji. Ogólnie jest on reprezentatywny, aktualny i dobrze dobrany, chociaż można mieć do niego kilka zastrzeżeń (o czym poniżej).

3. Poprawność metod badawczych

Badania wykonano poprawnie pod względem metodologicznym, chociaż moim zdaniem nie w sposób kompletny, ponieważ nie wykorzystano wszystkich istniejących możliwości. W rozdziale 3 pokazano, jak można estymować czas opóźnienia sygnałów docierających do pary mikrofonów (szkoda, że tylko dla jednego mówcy). W rozdziale 4 zaimplementowano układ dwumikrofonowy (szkoda, że tylko dwu!) i po przeprowadzeniu wielogodzinnych testów w różnych warunkach akustycznych stwierdzono, że poprawa jakości sygnału metodą *delay-sum* z użyciem GCC-PHAT jest tylko nieznacznie gorsza (o 0.9 dB) od adaptacyjnego układu ACC z rekurencyjnym algorytmem Shermana-Morrisona. Szkoda, że dokładniej nie opisano przebiegu tych testów. Następnie w rozdziałach 6 i 7 porównano kilka coraz bardziej złożonych konfiguracji algorytmicznych testowanego rozwiązania do diaryzacji nagrań (ale bez konfiguracji *filter-sum* macierzy mikrofonowej):

- tylko GCC-PHAT+GMM na macierzy mikrofonów,
- tylko MFCC+GMM na losowo wybranym sygnale z macierzy mikrofonów,

- połączenie GCC-PHAT+GMM z MFCC+GMM na najsilniejszym sygnale z macierzy mikrofonów (ze stałymi wagami oraz z adaptacją wag podczas fuzji klasyfikatorów),
- połączenie GCC-PHAT+GMM z MFCC+GMM, pracującym na sygnale sumarycznym, uzdatnionym metodą *delay-sum* (ze stałymi wagami oraz z adaptacją wag podczas fuzji).

Otrzymano wyraźne zwiększenie skuteczności proponowanej metody. Rezygnując z zastosowania metody *filter-sum* świadomie wybrano rozwiązanie gorsze pod względem skuteczności i elastyczności, ale mniej złożone obliczeniowo.

Rozdział piąty nie jest połączony z resztą rozprawy i dotyczy badań, które są dopiero planowane w przyszłości (wielka szkoda, że nie zrealizowano ich w ramach recenzowanej pracy, co niestety obniża jej ocenę). Pokazano w nim możliwości łącznej optymalizacji „statycznej" współczynników filtrów mikrofonowych algorytmem Neldera-Meada, w celu podniesienia selektywności kierunkowej macierzy mikrofonów. Uzyskano obniżenie listków bocznych charakterystyki o 5 dB. Ale tego wyniku nie wykorzystano w dalszej części rozprawy, gdyż w ogóle nie wykorzystywano w niej konfiguracji *filter-sum* macierzy mikrofonowej.

4. Ważność i aktualność tematyki

Tematyka pracy jest niezwykle trudna, ważna i aktualna. **Trudna**, gdyż dotyczy bardzo skomplikowanego zagadnienia rozpoznawania mowy/mówcy, które badane już jest od ponad 30 lat (wczesne prace L. Rabinera), ale do tej pory nie przyniosły one jednoznacznie pozytywnego wyniku. Jak widać z przedstawionych w pracy rezultatów, dodanie metod lokalizacji mówcy jest korzystne tylko w przypadku większego współczynnika SNR (>20 dB), czyli ciężar rozwiązania problemu nadal spoczywa na metodach częstotliwościowych rozpoznawania głosu. **Ważna**, gdyż chcemy obecnie wszystko „automatyzować", w tym przypadku automatycznie, dokładnie protokołować rozmowy wielu osób (kto, kiedy, co powiedział). Osoby te mogą przemieszczać się, a zakłócenia (tło) akustyczne może być zmienne i duże. **Aktualna**, gdyż patrząc na obecną skuteczność DER istniejących systemów, widzimy, iż jest ona daleka od oczekiwań użytkowników.

5. Oryginalność rozprawy - wkład własny Autora

Rozwiązanie zaprezentowane w pracy wykorzystuje ogólną koncepcję, która jest stosowana na świecie od około 10-ciu lat przez kilka dużych zespołów badawczych. Zespoły te stosują te same elementy składowe, tzn. GCC-PHAT do określenia położenia mówcy oraz MFCC+GMM do identyfikacji mówcy na podstawie cech widmowych. Oryginalną propozycją Doktoranta jest estymowanie poziomu zakłóceń SNR i na tej podstawie zmienianie współczynników wagowych podczas fuzji dwóch klasyfikatorów: położenia i mówcy. Inni stosują do tego celu BIC [5] lub odwrotność entropii [79]. Jak wykazały przeprowadzone w pracy badania, zaproponowane rozwiązanie zmniejsza błąd diaryzacji DER dla wyższych wartości SNR. Ale trudno powiedzieć czy jest lepsze od innych metod, ponieważ Doktorant pracował na swojej bazie nagrań. Baza ta była stworzona według rekomendacji NIST, ale jednak: 1) jej nagrania charakteryzują się większą wartością SNR niż nagrania baz NIST o około 15 dB (30 zamiast 15 dB - stwierdzenie samego Doktoranta na str. 77) oraz 2) zawsze mówi tylko jedna osoba, co moim zdaniem stanowi dużą różnicę.

Mgr inż. Samborski jest badaczem posiadającym jedną bardzo wartościową publikację ([30]: *Speech Communication* **2015**), ale niezwiązaną z doktoratem. Jego dorobek bezpośrednio związany z tematyką rozprawy doktorskiej jest starszy i skromniejszy (IASTED Int. Conf. on Signal Processing, Pattern Recognition and Applications **2010**: *Speech extraction in dual-microphone system*, IEEE Int. Symp. on Industrial Electronics **2010**: *Wiener filtration for speech extraction*, konferencja krajowa **2011**: *Wavelet-Fourier speaker recognition,* IEEE Int. Symp. on Signal Processing and Information Technology **2012**: *Speaker localization employing phase features and wavelet transform*,). Na tej podstawie należy stwierdzić, że metoda Doktoranta i postawiona w pracy teza rozprawy nie zostały jeszcze poddane szerszej ocenie.

Pomimo powyższych uwag, przedstawione w pracy wyniki są przekonujące, gdyż pokazują systematyczny wzrost skuteczności rozwiązań proponowanych przez Doktoranta w kolejnych etapach badań.

6. Znajomość światowego stanu wiedzy i literatury

Z lektury rozprawy wynika, że Doktorant zna stan wiedzy światowej z zakresu zagadnień poruszanych w rozprawie, **bardzo dobrze** z zakresu rozpoznawania mowy i mówców oraz filtracji adaptacyjnej, **dobrze** z zakresu macierzy mikrofonowych oraz **dostatecznie** z zakresu diaryzacji nagrań. Moim zdaniem najsłabszą stroną bibliograficzną/metodologiczną pracy jest nie omówienie i nie zastosowanie w niej adaptacyjnych metod kształtowania kierunkowości macierzy mikrofonowych (metody LCMV i GSC, szeroko omówione w [7]**)** oraz brak rozbudowanego, czytelnego przedstawienia i porównania w pracy aktualnie stosowanych metod z zakresu diaryzacji nagrań, czyli głównego, najambitniejszego obszaru badań, tzn. jakie metody były do tej pory stosowane (schematy blokowe), jak pojawiały się one historycznie, jak są ze sobą powiązane oraz jaką skuteczność oferują. Są cytowania artykułów przeglądowych, np. [76], ale artykuły te nie są dokładniej omówione. Brak jest cytowań najnowszych prac, np.

- X. Anguera, S. Bozonnet, N. Evans, C Fredouille, G. Friedland, O. Vinyals: "Speaker Diarization: A Review of Recent Research", *IEEE Trans. Audio, Speech, Language Processing*, vol. 20, no. 2, pp. 356–370, 2012.

Brak jest także cytowania interesującego omówienia istniejących, powszechnie dostępnych bibliotek programowych do diaryzacji nagrań:

- E. Kiktova, J. Juhar: "Comparison of Diarization Tools for Building Speaker Database", *Information and Communication Technologies and Services*, vol. 15, no. 4, pp. 314-319, 2015 (http://advances.utc.sk/index.php/AEEE/issue/view/56, http://advances.utc.sk/index.php/AEEE/article/view/1468/1085).

oraz informacji, w jakim zakresie korzystano w nich w doktoracie.

Brak jest cytowania doktoratów z zakresu diaryzacji nagrań, których teksty są dostępne w sieci Internet. Przykładowo:

- Sree Harsha Yella: *Speaker diarization of spontaneous meeting room conversations*, EPFL Lozanna 2015,
- Mary Tai Knox: *Speaker Diarization: Current Limitations and New Directions*, UC Berkeley 2013,
- Jordi Luque Serrano: *Speaker Diarization and Tracking in Multiple-Sensor Environments*, UPC Barcelona 2012.

Podsumowując, należy stwierdzić, że wykaz literatury jest bardzo **różnorodny i bogaty, ale niekompletny i nie wystarczająco omówiony** w zakresie zagadnień bezpośrednio związanych z tematem rozprawy.

7. Uwagi pozytywne – zalety i mocne strony rozprawy

Zadaniem recenzenta na podkreślenie zasługują następujące, pozytywne aspekty rozprawy:

1) Praca dotyczy bardzo trudnego, wielowątkowego problemu, nad którym pracują na świecie duże zespoły badawcze, dlatego sam fakt zbudowania przez Doktoranta praktycznie działającego systemu diaryzacji o podobnej skuteczności jak rozwiązana światowe, zdecydowanie przemawia na korzyść rozprawy.
2) Rozwiązanie zaproponowane przez Doktoranta jest co prawda podobne do rozwiązań światowych, ale jednak jest inne (dokonywanie fuzji dwóch strumieni decyzyjnych na podstawie zmierzonego poziomu zakłóceń, a nie z wykorzystaniem kryterium BIC lub entropii). Dodatkowo

rozwiązanie to jest proste i skuteczne, co wykazano w pracy, a to świadczy o jego dużej wartości aplikacyjnej.

3) Badania dotyczące bezpośrednio tematu pracy wymagały od Doktoranta zbudowania swojej bazy nagrań (tzw. korpusu). Zrobiono to **zgodnie z „zasadami sztuki"**, określonymi przez NIST (http://www.nist.gov/itl/), jednak poziom zakłóceń w nagraniach był mniejszy (ok. 30dB) niż w bazach danych NIST (ok. 15 dB). Jak mniemam, postąpiono tak, ponieważ planowano przeprowadzić badania dotyczące wpływu poziomu zakłóceń na skuteczność różnych rodzajów fuzji dwóch strumieni informacyjnych. Jednak Doktorant osobiście nie ustosunkowuje się w pracy do tej kwestii.
4) Wyniki dotyczące samej diaryzacji, przedstawione w rozdziale 7, nie zajmują zbyt wiele stron, ale wymagały od Doktoranta dużego nakładu pracy. **Jednoznacznie przekonują mnie one o wartości końcowego rozwiązania**.

8. Uwagi krytyczne – wady i słabe strony rozprawy

Pod adresem recenzowanej rozprawy można sformułować wymienione poniżej uwagi krytyczne.

1) Autor w rozprawie za mało miejsca poświęca dokładnemu omówieniu istniejących rozwiązań światowych bezpośrednio związanych z zagadnieniem diaryzacji, centralnym problemem pracy (tylko rozdziały 6 i 7, razem 25 stron na 94). Bardziej koncentruje się na powszechnie znanych podstawach, dotyczących macierzy mikrofonowych (np. na metodach estymacji parametru TDOA, na poprawie jakości sygnału techniką *delay-sum* i *filter-sum*, na samej możliwości kształtowania kierunkowości macierzy mikrofonowej w układzie *filter-sum* – co ciekawe sam nie korzysta z tej możliwości w końcowym rozwiązaniu).
2) Pierwsze rozdziały pracy (od 2 do 5) są zbyt podręcznikowe, a ich wyniki eksperymentalne – oczywiste i bardziej ilustracyjne niż naukowe. Natomiast w ostatnich rozdziałach pracy, o wiele bogatszych i ciekawszych naukowo, nie przedstawiono w przejrzysty sposób szczegółów algorytmicznych proponowanych rozwiązań. Brak jest czytelnych, ale jednak precyzyjnych schematów blokowych oraz wyeksponowanego, prostego opisu kolejnych kroków badań (w stylu: 1), 2) 3), …).
3) Uważam, że rozwiązanie zaprezentowane w pracy jest oryginalne i ciekawe, ale nie jest wyjątkowo nowatorskie. Połączenie cech klasycznego rozpoznawania mówcy, bazującego na metodzie (MFCC+GMM), z jego lokalizacją, estymowaną metodą GCC-PHAT, było już wcześniej opisywane w literaturze [5,57,58,78-80]. Co więcej, dostępny jest bezpłatny pakiet programowy DiarTK dla tej metody [78]. Doktorant twierdzi (m.in. w tezie rozprawy), że jego oryginalnym osiągnięciem jest dodanie algorytmu dynamicznego ważenia (6.15) obu wskaźników rozpoznawania mówcy (indywidualne cechy głosu oraz położenie mówcy) w zależności od poziomu zakłóceń (SNR). Tak, to prawda. Ale inni też stosowali wzór (6.15) tylko inaczej adaptowali wagi (używając BIC lub entropię). W pracy [5] (2007): „*Automatic weighting for the combination of TDOA and acoustic features in speaker diarization of meetings*" także porównywano „sztywną" i automatyczną fuzję cech kierunkowych i osobowych, jednak nie wykorzystywano SNR tylko BIC (*Bayesian Information Criterion*). Natomiast w pracy [79] (2011) „*An information theoretic combination of MFCC and TDOA features for speaker diarization*" zrobiono to samo, tylko wykorzystując jako wagi odwrotność entropii. **Oryginalność rozwiązania Doktoranta polega na zastosowaniu nowego kryterium podczas fuzji, poziomu SNR a nie kryterium BIC lub entropii. Należało to precyzyjnie sformułować.**
4) Jak pokazuje tabela 7.2 na str. 84 osiągnięte wyniki są dobre, ale nie najlepsze, w porównaniu z innymi metodami. Może należało jeszcze wykazać, że zaproponowana metoda ma inne zalety w porównaniu z rozwiązaniami alternatywnymi, np. być może jest o wiele mniej złożona obliczeniowo.
5) Dobrze, że Doktorant przedstawił porównanie swojej metody z innymi. Szkoda jednak, że nie wykorzystał tego samego korpusu NIST (http://nist.gov/speech/tests/rt/: RT06, RT06s, RT07,

RT09, …) co inni badacze, tylko swoje dane, gdyż uniemożliwia to obiektywną ocenę osiągniętego wyniku. Dlaczego tak postąpiono? Prosiłbym na odpowiedź na to pytanie podczas obrony.

6) Doktorant zastosował stosunkowo prostą, typową metodę (GCC-PHAT), także stosowaną przez innych badaczy zajmujących się diaryzacją nagrań, a nie metodę LCMV lub GSC, dlatego nie należało się spodziewać, że szala zwycięstwa przechyli się na jego korzyść.
7) W pracy nie podaje się w sposób wyeksponowany szczegółowej informacji na temat dostępnych narzędzi programowych dla diaryzacji nagrań (typu *open source*), a takowe istnieją (w pkt. 7 recenzji podano publikację, w której je omówiono). Doktorant nie informuje czytelnika, z których bibliotek korzystał i w jakim stopniu. Na stronie 84 pisze tylko o istnieniu przybornika DiarTK [78], ale nie mówi, że jego autorzy także wykorzystywali w [79] ważoną fuzję klasyfikatorów MFCC+GMM i TDOA+GMM, dynamicznie przestrajaną odwrotnością entropii. Wynika stąd, że poziom trudności **programowej**, wymagany do realizacji doktoratu, nie był bardzo duży: należało „tylko" dodać lub zmienić regułę adaptacyjną. W tabeli 7.2 Autor porównuje skuteczność swojego rozwiązania z innymi algorytmami [5, 58, 75, 78]. Czy dla nich także istnieją gotowe programy? Należało o tym napisać.
8) Stworzona baza nagrań jest prostsza do analizy niż baza NIST: mówi tylko jeden mówca, czyli poziom trudności jest mniejszy (algorytm zaproponowany zwracał tylko jedną hipotezę). W tym aspekcie wyniki z tabeli 7.1 nie są optymistyczne: błąd DER dla metody MFCC z [78] jest aż dwukrotnie mniejszy niż dla metody MFCC, wykorzystanej w recenzowanej pracy, a błędy dla TDOA są porównywalne.
9) Badania eksperymentalne przedstawione w rozdziale 3 i 4, dotyczące zastosowania filtracji adaptacyjnej do poprawy jakości sygnału mowy, nie są przekonujące z powodu ich ograniczonego zakresu. Bardziej przypominają one „demonstrator" idei, niż pracę badawczą. W rozdziale 3 jeden eksperyment pokazuje, że filtr się dostraja do opóźnienia, a w rozdziale 4 przeprowadza się analizę sygnałów, co prawda kilkugodzinnych, ale zarejestrowanych tylko w jednym pomieszczeniu. Nie przebadano różnych konfiguracji filtrów adaptacyjnych i doboru ich parametrów. Dodatkowo, moim zdaniem, cała dyskusja dotycząca filtracji adaptacyjnej powinna znaleźć się w rozdziale 4.
10) W pracy brak jest wymaganego streszczenia w języku angielskim.

9. Błędy edycyjne

W pracy występują drobne błędy edycyjne:

- najczęściej interpunkcyjne (brak przecinków);
- często zamiast „liczba" używane jest słowo „ilość" do rzeczy przeliczalnych (np. ilość mikrofonów na str. 23, 37, str. 45, ilość wymiarów na str. 25, ilość próbek (str. 67), ilości komponentów (str. 74, 75);
- nazwy tabel są pod a nie nad tabelami, np. tabl. 4.1 (str. 46), tabl. 4.2 (str. 49), itd.;
- często rysunki mają angielskie opisy osi, np. rys. 3.4 (str. 35), 3.5 (str. 36), 6.3 (str. 68),
- w spisie treści nie ma wykazu literatury;
- w wykazie skrótów i oznaczeń nie wszystkie terminy zostały przetłumaczone na język polski (np. MAP nie);
- oś rysunku 2.5 (str. 21) nie jest opisana,
- w równaniu (3.3) (str. 29) powinno być też wprowadzone oznaczenie $s_2(n)$, wykorzystywane potem w równaniu (3.18) (str. 33);
- brak „2" w argumencie funkcji $\exp(\pi j f \tau_0)$ w równaniu (3.10) (str. 32)
- brak jednostek na osi poziomej na rys. 3.1 (str. 30);
- brak pełnej nazwy przy pierwszym użyciu akronimu GCC (str. 32);
- brak „f" w argumencie funkcji $\exp(2\pi j \tau)$ w równaniach (3.14), (3.15), (3.17), (3.19), (3.20), (3.25), (3.26), (3.27) … i następnych (patrz równanie (6c) w [44]);
- albo złe cytowanie albo niepoprawnie opisana publikacja [25]: brak Hoanga jako autora;

- w rozdziale 4 na rysunku 4.1 w pierwszych równaniach (4.1)-(4.7) stosuje się oznaczenie $z(n)$, a potem jako indeks macierzy korelacji wzajemnej **R** stosuje się błędne oznaczenie $d(n)$, czyli pisze $\mathbf{R}_{d,x}$ zamiast $\mathbf{R}_{z,x}$ (równania od (4.8) do (4.14), także w tekście na str. 44);
- w równaniach (4.13) i (4.14), w celu zwiększenia ich zrozumiałości, lepiej było oznaczyć wektor małą, pogrubioną literą, a nie dużą, pogrubioną, czyli napisać $\mathbf{r}_{z,x}$ zamiast $\mathbf{R}_{z,x}$.
- na str. 42 i 43 nie uzupełniono cytowania dotyczącego algorytmu Shermana-Morrisona, pozostało „[?]";
- zmienne w tekście powinny być pisane tak jak w równaniach, czyli kursywą; na stronie 42 mamy „N+1";
- str. 43: w równaniu (4.22) w dwóch kolejnych wierszach napisano to samo;
- str. 44: równanie (4.24) jest niepoprawne - $x(n-i)$ jest zbyteczne;
- str. 47, rys. 4.2, także str. 65: określenie „bank filtrów" jest stosowane w polskiej literaturze, ale nie oddaje one aspektu „współpracy" poszczególnych filtrów; osobiście preferowałbym termin „zestaw, zespół, grzebień filtrów";
- str. 51: duża, biała „plama" pod rysunkiem 5.1; nie powiedziano dlaczego równanie (5.2) można zapisać w takiej postaci (przy założeniu fali płaskiej generowanej przez źródła w dalekim polu);
- str. 54: dlaczego w równaniu (5.15) odbierany sygnał czasowy jest zespolony?
- str. 55: zaproponowane rozwiązanie optymalizacyjne doboru wag filtrów z rys. 5.2 za pomocą algorytmu Neldera-Meada dotyczy jednej, określonej („sztywnej") sytuacji pomiarowej, tzn. układ nastraja wagi na określone pozycje mówców;
- po str. 58 występują dwie puste strony;
- str. 62: moim zdaniem stwierdzenie, że zastosowanie TDOA pojawia się dopiero „ostatnio" w diaryzacji nie jest prawdziwe; jest stosowane od około 10 lat [5,57];
- str. 65: w równaniu (6.9) brakuje modułu, powinno być $|Y(m)|$;
- str. 66: zbyt wysoki poziom ogólności, nie podano szczegółów dotyczących GMM;
- str. 91, niepełne dane pozycji [62], dwa razy 2000;

9. Podsumowanie

Podsumowując, chciałbym jednoznacznie stwierdzić, że recenzowana rozprawa doktorska Pana mgr inż. Rafała Samborskiego posiada szereg istotnych zalet i pomimo licznych niedociągnięć spełnia wymagania stawiane przez „Ustawę o stopniach naukowych i tytule naukowym" (artykuł 13, ustęp 1) z 2003 roku, z kolejnymi nowelizacjami, w tym w 2011 i 2014 roku, gdyż „stanowi **oryginalne rozwiązanie** problemu naukowego oraz **wykazuje ogólną wiedzę teoretyczną** kandydata w danej dyscyplinie naukowej oraz **umiejętność samodzielnego prowadzenia pracy naukowej**". W rozprawie zaproponowano bowiem oryginalne i ciekawe rozwiązanie, które jedynie nie najlepiej opisano. Wykazano jednak ponad wszelką wątpliwość jego wartościowość oraz zdolności naukowe Doktoranta. Ostatnia współautorska publikacja Doktoranta w czasopiśmie *Speech Communication* wydawnictwa Elsevier, trochę z innej tematyki, jednoznacznie potwierdza dojrzałość badawczą autora recenzowanej rozprawy. Dlatego z pełnym przekonaniem wnioskuję o dopuszczenie Pana mgr inż. Rafała Samborskiego do dalszych etapów przewodu doktorskiego, w tym do oficjalnej obrony.

Tomasz Zieliński

Kraków,12.04.2016