AKADEMIA GÓRNICZO-HUTNICZA

IM. STANISŁAWA STASZICA W KRAKOWIE

Wydział Informatyki, Elektroniki i Telekomunikacji

Katedra Elektroniki

PRACA DYPLOMOWA

Inżynierska

Temat: Półautomatyczne rozpoznawanie mówców w kryminologii

Semiautomatic speaker recognition in forensic sciences

Imię i nazwisko: Piotr Żelasko
Kierunek studiów: Inżynieria Akustyczna

Opiekun pracy: dr inż. Bartosz Ziółko

Kraków, rok akademicki 2012/2013

OŚWIADCZENIE AUTORA PRACY

Oświadczam, świadomy odpowiedzialności karnej za poświadczenie nieprawdy, że niniejszą pracę dyplomową wykonałem osobiście i samodzielnie i że nie korzystałem ze źródeł innych niż wymienione w pracy.

…….

Podpis

Dziękuję promotorowi, dr. Bartoszowi Ziółce,
za pomoc podczas prowadzenia badań
oraz pani Agacie Trawińskiej
za propozycję tematu pracy i cenne uwagi.

Streszczenie

W pracy opisane są problemy związane z rozpoznawaniem mówców, zarówno w sposób automatyczny jak i nieautomatyczny. Szczególnie uwzględnione zostało zagadnienie właściwej parametryzacji sygnału mowy poprzez ton krtaniowy oraz formanty, wraz z wyjaśnieniem ich pochodzenia. Scharakteryzowano metodę językowo – pomiarową, używaną przez biegłych sądowych na świecie do identyfikacji mówców na podstawie nagrań, oraz zaproponowano sposób na sprawdzenie, czy ilość danych potrzebnych do analizy może zostać zredukowana bez negatywnego wpływu na jakość rozpoznania. Opisany został algorytm DTW (ang. Dynamic Time Warping), który stanowił podstawę do sformułowania procedury testowej. Sprawdzono, które samogłoski spośród grupy sześciu stanowią najlepszy materiał do prowadzenia identyfikacji, a także zbadano, które formanty są najbardziej pomocne przy ustalaniu tożsamości. W pracy zawarto analizę mocnych i słabych stron zaproponowanej procedury oraz ich wpływu na otrzymane wyniki.

Spis treści

1. Wprowadzenie

1.1. O Identyfikacji mówców

Rozpoznawanie mówców jest dziedziną wiedzy, która zaczęła prężnie rozwijać się w drugiej połowie XX wieku, wraz z rosnącym zapotrzebowaniem na precyzyjne rozpoznanie mówców zarówno w sektorze publicznym (np. w celu stwierdzenia tożsamości mówcy wypowiadającego się na nagraniu służącym jako materiał dowodowy w rozprawie sądowej), jak i w sektorze prywatnym (np. by umożliwić komunikację człowiek-maszyna za pomocą głosu) [1, 2]. Samo rozpoznawanie mówców jest tylko częścią nauki zajmującej się przetwarzaniem sygnału mowy, w skład której wchodzą również zagadnienia takie jak: rozpoznawanie mowy, synteza mowy, transmisja i zapis sygnału mowy, poprawa jakości sygnału mowy czy systemy wspomagania niewidomych [2].

W zagadnieniu rozpoznawania mówcy należy dokonać zasadniczego podziału na dwa problemy: identyfikację mówcy, gdy określamy, która spośród znanych nam osób jest mówcą, oraz weryfikację mówcy, gdy sprawdzamy, czy mówca jest konkretną osobą. Zagadnienia te, choć pokrewne, różnią się w podejściu do problemu: podczas identyfikacji mówcy, określa się miarę podobieństwa pomiędzy mówcą nieznanym a mówcami występującymi w bazie danych i określa się tożsamość mówcy na podstawie najlepszego wyniku, natomiast w trakcie weryfikacji, dla konkretnej osoby obliczony zostaje próg odniesienia, do którego odnosi się wynik uzyskany przez mówcę testowanego – jeżeli zmieści się poniżej progu, następuje weryfikacja, natomiast jeżeli próg zostanie przekroczony, mówcę odrzuca się [2].

W niniejszej pracy rozpatrywana jest identyfikacja mówcy na potrzeby kryminologii. Początki rozpoznawania mówcy mają korzenie w lingwistyce. W celu stwierdzenia tożsamości osoby zarejestrowanej na nagraniu, biegli o wykształceniu lingwistycznym analizowali wypowiedzi zarówno pod kątem treści, w celu stwierdzenia charakterystycznej składni zdań i używanego słownictwa, jak i pod kątem sposobu wymawiania, w celu określenia cech artykulacyjnych, nawyków w wymowie i innych [1, 3]. Z upływem czasu, badania czysto lingwistyczne zostały wzbogacone o badania cech akustycznych, mierzalnych ilościowo. Przykłady takich cech opisane są w

dalszych rozdziałach. Uwzględnienie cech takich jak częstotliwość tonu podstawowego czy przebiegów formantowych spowodowało powstanie metody językowo-pomiarowej, łączącej jakościowy opis mówcy wraz z ilościowym, która jest używana do dziś w różnych ośrodkach badawczych na świecie, w tym przez Laboratorium Analizy Mowy i Nagrań Instytutu Ekspertyz Sądowych im. Prof. dr. Jana Sehna w Krakowie (IES) [3].

Wraz z rozwojem nauk informatycznych i technik przetwarzania sygnałów, pojawił się nurt alternatywny, dążący do automatyzacji procesu rozpoznawania mówców. Automatyczne systemy rozpoznawania mówców najpierw przetwarzają sygnał tak, by łatwiej było wydobyć z niego indywidualne cechy mówców (np. preemfaza), a później stosują przekształcenia, uzyskując opis tego, jak ułożony był w danej chwili trakt głosowy, np. za pomocą Mel Frequency Cepstral Coefficients [4] lub Linear Prediction Coding [5]. W następnym kroku dokonywana jest klasyfikacja, mająca na celu znaleźć w bazie danych najbardziej podobnego mówcę. Jednym z klasyfikatorów używanych w takich systemach są ukryte (niejawne) modele Markowa (ang. Hidden Markov Model, HMM) [2, 6], tworzące probabilistyczny opis zjawiska generacji dźwięku w kanale głosowym i znajdujące w bazie danych mówcę maksymalizującego prawdopodobieństwo wypowiedzenia testowanej kwestii.

Metody lingwistyczne oraz metody automatyczne rozpoznawania mówcy bazują na tym samym założeniu – podobieństwo tych samych wypowiedzi pośród tego samego mówcy będzie większe, niż podobieństwo tych samych wypowiedzi pośród dwóch różnych mówców [1], jednak wyniki badań każdej z tych metod ciężko do siebie odnieść, wskutek czego różne – choć nie wszystkie – ośrodki badawcze na świecie preferują tylko jedną z metod [3].

1.1. Cele pracy

Celem niniejszej pracy jest zaproponowanie sposobu na przetestowanie jakości rozpoznania mówcy w zależności od wyboru badanych samogłosek spośród zbioru {a, e, y, i, o, u} oraz od wyboru formantów spośród zbioru {F1, F2, F3, F4} oraz wykonanie odpowiednich testów. Badania takie umotywowane są chęcią poprawy jakości identyfikacji mówcy przez biegłego metodą językowo – pomiarową, przy jednoczesnym zmniejszeniu liczby czynników potrzebnych do przeprowadzenia identyfikacji. O priorytecie doboru samogłosek do badań decyduje również częstość występowania konkretnej samogłoski w języku polskim, która wpływa na ilość dostępnego materiału, na podstawie którego ma zostać dokonana identyfikacja [7], oraz fakt, że niektóre formanty podczas nagrywania w telefonie komórkowym (lub innym urządzeniu elektronicznym) mogą ulec zniekształceniu ze względu na filtry dolnoprzepustowe, filtry antyaliasingowe oraz proces kodowania mowy np. w popularnym formacie mp3 [8, 9].

Teza, która jest sprawdzana w pracy, brzmi następująco: niektóre samogłoski silniej indywidualizują danego mówcę niż pozostałe, tj. ich przebiegi czasowe niosą więcej informacji o charakterystycznych cechach mówcy. Sprawdzono również, czy jakość identyfikacji nie ucierpi poprzez zmniejszenie liczby parametrów, czyli oparcie analizy tylko na części formantów, zamiast na wszystkich czterech.

2. Metodyka oraz sposób parametryzacji

2.1. Częstotliwość tonu krtaniowego F0

Jednym z często używanych parametrów przy opisie sygnału mowy jest częstotliwość tonu krtaniowego F0 [10, 11]. Parametr taki jest możliwy do wyekstrahowania tylko w wypadku głosek dźwięcznych, ponieważ w trakcie ich wymawiania pracują mięśnie wewnętrzne krtani, pobudzając ją do drgania. Widmo tonu krtaniowego składa się z wielu harmonicznych o amplitudzie zanikającej wraz z częstotliwością [11].

Częstotliwość tonu krtaniowego zależy od płci oraz indywidualnych cech mówcy. W ramach wypowiedzi tego samego mówcy występuje ciągła zmienność F0, zależna od intonacji, stanu emocjonalnego, tempa mowy oraz intencji wypowiadającego. Orientacyjne zakresy F0 w zależności od wysokości głosu podane są w tab. 2.1.

Tabela 2.1. Częstotliwości tonu krtaniowego dla różnych głosów [11].

Bas	80 – 320 Hz
Baryton	100 – 400 Hz
Tenor	120 – 480 Hz
Alt	160 – 640 Hz
Mezzosopran	200 – 800 Hz
Sopran	240 – 960 Hz

Jedną z metod określenia częstotliwości tonu krtaniowego jest analiza przejść przez zero. W celu wykonania jej, dokonuje się filtracji dolnoprzepustowej sygnału mowy oraz usuwa się składową stałą, a następnie zlicza się odległości czasowe pomiędzy kolejnymi przejściami przez zero, które wykrywane są w tych miejscach, gdzie sygnał zmienia wartości z ujemnych na dodatnie i vice versa [10]. Metoda ta pozwala nie tylko na określenie tonu podstawowego, ale także na stwierdzenie jego zmienności – jest dzięki temu chętnie stosowana podczas badania schorzeń kanału głosowego i analizy mowy patologicznej [11].

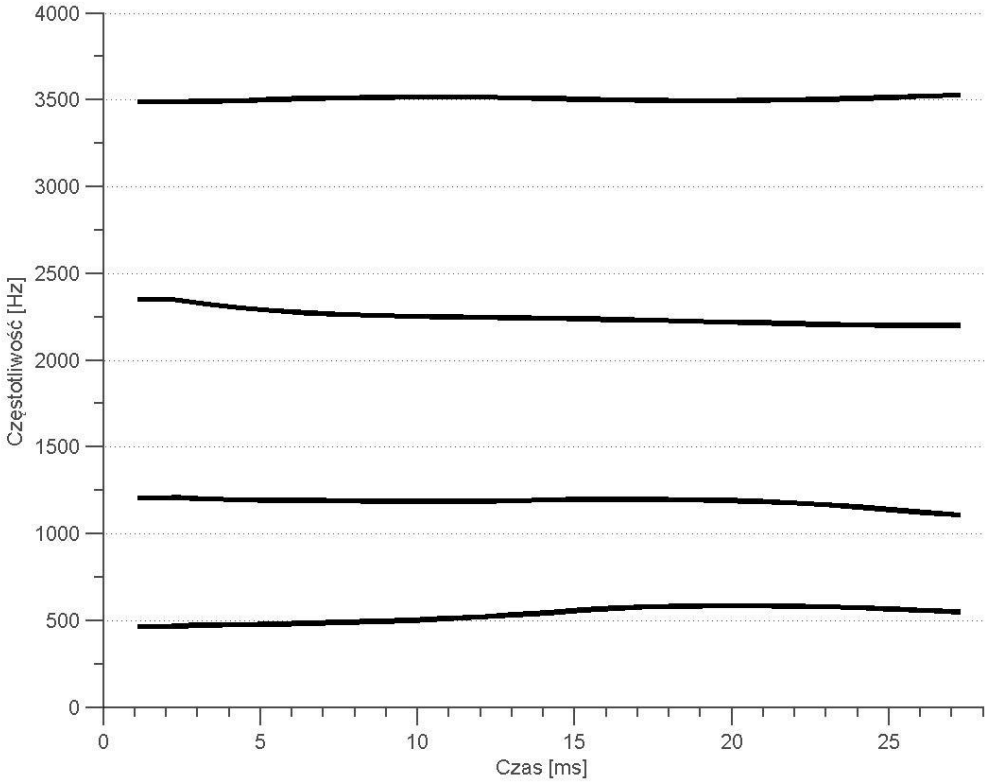
F0 jest chętnie stosowane w systemach weryfikacji i identyfikacji mówców oraz rozpoznawania mowy [2], nie może być jednak uznane za cechę jednoznacznie określającą tożsamość w taki sposób, jak np. odcisk palca. Przykładowo, większość mężczyzn cechuje się częstotliwością tonu podstawowego 90 – 140 Hz, a więc dopiero wartości F0 wykraczające poza ten zakres będą silnie wskazywać na konkretnego mówcę podczas identyfikacji [12]. Z tego powodu częstotliwość tonu krtaniowego jest brana pod uwagę przez biegłych sądowych podczas ustalania tożsamości nieznanej osoby, ale nie jest parametrem wystarczającym do opisu cech akustycznych zawartych w dźwięcznych fragmentach wypowiedzi [3].

2.2. Formanty

Fala dźwiękowa generowana poprzez drganie krtani ulega modyfikacjom, gdy przechodzi przez kanał głosowy człowieka. Formalnie proces ten opisany jest jako filtracja widma tonu krtaniowego za pomocą transmitancji modelu kanału głosowego. Wskutek przejścia tonu krtaniowego przez filtr traktu głosowego, w widmie tonu krtaniowego pojawiają się lokalne maksima, zwane formantami, a częstotliwości, w których występują, zwane są częstotliwościami formantowymi [10, 11]. Wynikają one z rezonansów zachodzących wewnątrz traktu głosowego, który jest modelowany przez falowód o zmiennej średnicy – taki zabieg pozwala na próby przewidywania, jak może układać się trakt głosowy podczas wypowiadania konkretnej samogłoski [13].

Częstotliwości formantowe zależą od czynników niezmiennych, takich, jak wymiary kanału głosowego indywidualnej osoby, oraz zmiennych, do których zaliczyć można: rodzaj wypowiadanego fonemu, ułożenia kanału głosowego i aparatu mowy w danej chwili czasu, indywidualne nawyki językowe i sposób wymowy oraz inne. W trakcie wypowiadania jednego fonemu, częstotliwości formantowe ulegają zmianom, co związane jest ze zmianami ułożenia toru głosowego. Formanty pozwalają więc na określenie, co jest mówione, przez kogo, oraz w jaki sposób, ponieważ ich konfiguracja jest charakterystyczna dla danej samogłoski [10] i choć wykorzystywane są z powodzeniem do identyfikacji oraz weryfikacji mówców [2], to nie zostało wciąż wykazane, że stanowią cechę unikatową dla danej osoby, tzn. że nie znajdzie się druga

osoba, która może zostać scharakteryzowana przez ten sam zestaw przebiegów formantowych [1].



Rycina 2.1. Przykładowe przebiegi czasowe pierwszych czterech formantów podczas wypowiadania samogłoski „a”.

W tej pracy, formanty są podstawowymi i jedynymi cechami akustycznymi opisującymi samogłoski, które stanowią podstawę do identyfikacji mówcy. Ze względu na ograniczenia techniczne, zazwyczaj występujące podczas rejestracji mowy za pomocą szeroko dostępnego sprzętu elektronicznego, np. telefonów komórkowych, następuje zawężenie pasma częstotliwościowego nagrania do granic ok. 200 – 4000 Hz, co pozwala na ekstrakcję i analizę pierwszych czterech formantów, oznaczanych jako F1, F2, F3 oraz F4.

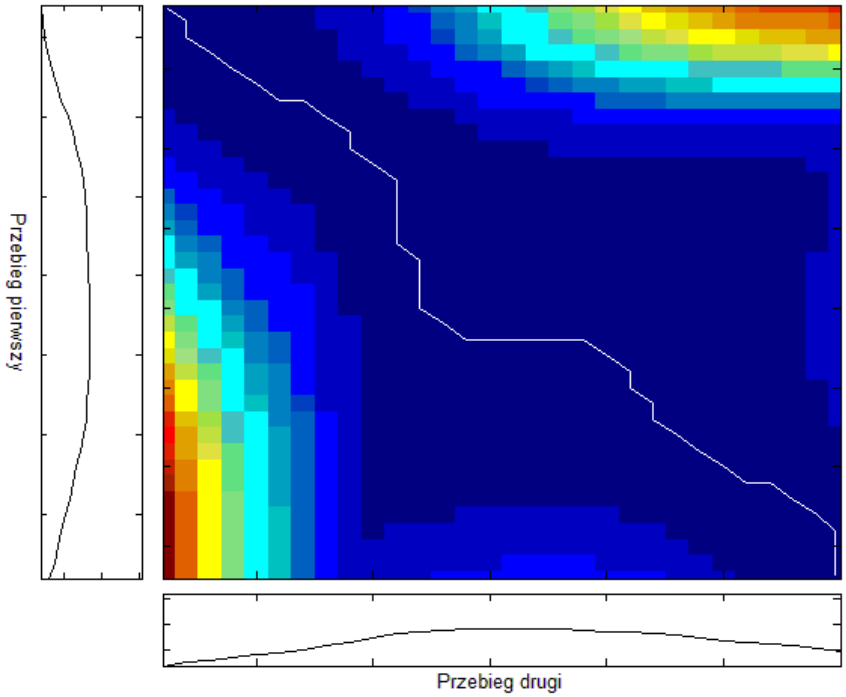
Ekstrakcja przebiegów formantowych z nagrania dokonywana jest za pomocą programu STx Austriackiej Akademii Nauk [14]. Obróbka wygląda następująco – sygnał mowy jest okienkowany z długością ramki 22 ms, oraz zachodzeniem się ramek (overlapping) 95%, a następnie w każdej ramce liczone jest FFT (ang. Fast Fourier Transform) sygnału w celu otrzymania widma. Po tych operacjach stosowana jest

predykcja liniowa w celu otrzymania współczynników LPC, dzięki którym można oszacować obwiednię widma i znaleźć jej lokalne maksima oraz ich położenie na osi częstotliwości w danej ramce. Pomiędzy wartościami otrzymanymi z każdej ramki występuje odstęp czasowy 1.1 ms.

2.3. Algorytm DTW

DTW (ang. Dynamic Time Warping) jest algorytmem, którego zastosowania obejmują często dopasowanie czasowe dwóch sygnałów w różnych dziedzinach czasu, tzn. gdy są one podobne kształtem, ale charakterystyczne cechy ich przebiegów nie znajdują się w tych samych momentach czasu. Znane są praktyczne zastosowania algorytmu np. w systemach rozpoznawania mówców [15] oraz podczas dopasowania wektorów cech sygnału testowanego do wektorów cech sygnału referencyjnego [2].

Zadaniem algorytmu jest znalezienie optymalnej ścieżki dopasowania pomiędzy dwoma sygnałami. Ścieżka taka charakteryzuje się najmniejszym możliwym kosztem przejścia z początku macierzy kosztów – w punkcie (1,1) – do jej końca w punkcie (N,M), gdzie N – długość pierwszego sygnału, M – długość drugiego sygnału.



Rycina 2.2. Zilustrowana macierz kosztów wraz z optymalną ścieżką dopasowania DTW. Kolory cieplejsze oznaczają większą odległość pomiędzy próbkami sygnału, a kolory chłodniejsze mniejszą.

Macierz kosztów formułowana jest w oparciu o obliczoną według przyjętej metryki odległość pomiędzy próbkami sygnałów w różnych chwilach czasu. W dalszej kolejności znajdywana jest optymalna ścieżka dopasowania, poprzez poszukiwanie w każdym kolejnym punkcie – począwszy od punktu (1,1) – następnego punktu o najmniejszej wartości, aż do końca ścieżki, czyli punktu (*M*,*N*). Ograniczenia, jakie zostały nałożone na algorytm, to monotoniczność – w żadnym kroku indeksy *i* oraz *j* (gdzie i = *1* … *N*, *j* = 1 … *M*) wskazujące na obecny punkt ścieżki nie mogą zmaleć i przynajmniej jeden z nich musi wzrosnąć, oraz ciągłość – w żadnym kroku indeksy *i* oraz *j* nie mogą powiększyć się o więcej niż 1. Dodatkowe ograniczeniem jest wymuszenie, by algorytm nie dopuścił do zbyt dużego odchylenia ścieżki od przekątnej macierzy – spowodowałoby to pominięcie ważnych cech dopasowywanych sygnałów [15] – co osiągnięte jest poprzez ustalenie progu *r*, który definiuje maksymalną możliwą różnicę pomiędzy indeksami *i* oraz *j*. Jeżeli próg *r* zostanie osiągnięty, zawężana jest liczba możliwości wyboru dalszego kierunku ścieżki dopasowania.

3. Procedura oparta o algorytm DTW i jej testy

3.1. Opis danych

Przed przedstawieniem samej metody, należałoby przybliżyć, w jaki sposób wyglądają dane, które zostały wykorzystane. W projekcie użyto dwóch zestawów danych, jeden składający się z ośmiu mówców, a drugi z dwudziestu jeden mówców, których wypowiedzi zostały nagrane. Z każdej wypowiedzi zostały wydobyte wszystkie samogłoski, a następnie umieszczone w grupach {a, e, y, i, o, u}. W każdej z grup, każda z samogłosek zawiera adnotację o tym, w jakim kontekście wystąpiła w zdaniu i w jaki sposób została zaakcentowana – przykładowo, zapisana jest samogłoska *a*, która wystąpiła po spółgłosce *p* i przed spółgłoską *s*, mocno zaakcentowana. Ostatecznie, każda z tak zapisanych spółgłosek opisana jest poprzez przebiegi czasowe formantów {F1, F2, F3, F4}.

Niestety, w praktyce liczba powtórzeń takiej samej samogłoski w tym samym kontekście wypowiedzianej przez tego samego mówcę nie przekracza trzech prób, a w niektórych przypadkach w zbiorze danych występuje tylko jedna taka próba. Ponadto należy wspomnieć, że w wielu próbach niektóre z formantów mają brakujące fragmenty lub w ogóle nie występują, co zazwyczaj jest spowodowane charakterystyczną wymową. W wypadku nie występowania któregoś z formantów, nie jest on brany pod uwagę przy klasyfikacji. Problem niekompletności danych nie jest zagadnieniem poruszanym w tej pracy.

Dane pozyskane są z nagrań przeprowadzonych przez Laboratorium Analizy Mowy i Nagrań Instytutu Ekspertyz Sądowych im. Prof. dr. Jana Sehna w Krakowie. Rejestracja została dokonana w wypadku pierwszego zestawu w przystosowanym pomieszczeniu, a w wypadku drugiego zestawu w pomieszczeniach mieszkalnych. Zastosowany sprzęt to mikrofony kierunkowe Schoeps CCM-40 ustawione w odległości 20-30 cm od mówcy, rejestratory cyfrowe Marantz PMD 671 i HHB DRM 85-C, a format zapisu to wave PCM z częstotliwością próbkowania 44.1 kHz oraz rozdzielczością 16 bit.

3.2. Opis procedury

Przedstawiona w dalszych akapitach procedura została opracowana w celu przeprowadzenia badań nad przydatnością poszczególnych samogłosek podczas różnicowania mówców. Procedura ta opiera się na założeniu, że te same samogłoski wypowiadane w tych samych kontekstach przez jednego mówcę będą do siebie bardziej podobne, niż gdy będą je wypowiadać dwaj różni mówcy. Znalezienie samogłosek, które lepiej spełniają to założenie od pozostałych, pozwoli na uproszczenie procesu identyfikacji mówcy, poprzez pominięcie części danych, niosących mniej informacji o indywidualnych cechach akustycznych danego mówcy. Zagwarantowanie, że analizowane dane niosą informacje o dystynktywnych cechach mówcy, pozwoli również na dokładniejsze określenie jego tożsamości.

Wybór algorytmu DTW w celu zweryfikowania przytoczonego założenia został umotywowany wiedzą o przydatności algorytmu podczas rozpoznawania mówców, łatwością implementacji oraz przejrzystością sposobu jego działania. Na początkowych etapach projektu rozważane były inne metody parametryzacji danych, zostały jednak odrzucone z kilku powodów. Po pierwsze, nie mogły zostać użyte narzędzia tak wyrafinowane, jak np. ukryte modele Markowa, ze względu na niewystarczającą liczbę prób poszczególnych samogłosek w ramach pojedynczego mówcy – model zbudowany na niedostatecznej liczbie danych mógłby nie zadziałać poprawnie [8]. Ograniczenie takie wymusiło użycie prostszej metody testowania. Po drugie, rozważono możliwość opisu czasowych przebiegów formantów za pomocą wielomianów poprzez aproksymację wielomianową – nie wybrano tego sposobu z obawy, że tak znaleziony wielomian nie będzie opisywał wyłącznie odcinka czasu, w którym znajduje się formant, ale całą dziedzinę czasu w której jest zdefiniowany.

Parametryzacja danych z pomocą DTW wygląda następująco – na wejście algorytmu podawane są dwa sygnały, a na jego wyjściu otrzymywana jest odległość pomiędzy nimi. Odległość – liczona w metryce euklidesowej – jest sumą odległości zapisanych w każdym punkcie ścieżki dopasowania, znormalizowaną względem długości ścieżki. Dla każdego mówcy w bazie danych, określane są progi rozpoznania odpowiednie dla każdej samogłoski. Próg taki, to maksymalna odległość DTW uzyskana pomiędzy próbami tej samej samogłoski w tym samym kontekście w obrębie

jednego mówcy – gwarantuje to, że wszystkie przebiegi tego samego mówcy zostaną zakwalifikowane pozytywnie w dalszym etapie.

Procedura testowania jakości rozpoznania zaczyna się od zebrania takich samych samogłosek (wypowiedzianych w tych samych kontekstach) spośród wszystkich mówców i umieszczenia ich w jednej puli, na chwilę zapominając, który mówca je wypowiedział. Następnie formułowana jest symetryczna macierz odległości, gdzie oblicza się odległość DTW pomiędzy każdą parą samogłosek w puli – każda komórka macierzy zawiera na tym etapie tyle odległości, ile formantów jest branych pod uwagę. Każdy wiersz tej macierzy opisuje odległości jednej wybranej samogłoski od wszystkich pozostałych w puli, gdzie liczba wierszy równa jest liczbie samogłosek w puli. Przekątna macierzy to odległości zerowe, ponieważ są tam obliczone odległości pomiędzy tymi samymi samogłoskami. W następnym etapie, każdy z wierszy jest odnoszony do odpowiednich dla danego mówcy progów rozpoznania. W trakcie klasyfikacji zliczane są samogłoski poprawnie zaakceptowane *TA* (ang. True Acceptance), poprawnie odrzucone *TR* (ang. True Rejection) oraz fałszywie zaakceptowane *FA* (ang. False Acceptance). Fałszywe odrzucenie samogłoski nie jest możliwe ze względu na metodę przyjęcia progu.

Miarą jakości systemu są parametry *Precision*, *Recall* i *F*, zdefiniowane jak poniżej:

$$Precision = \frac{TA}{TA+FA} \quad \text{(wzór 3.1)}$$

$$Recall = \frac{TA}{TA+FR} \quad \text{(wzór 3.2)}$$

$$F = 2\,\frac{Precision*Recall}{Precision+Recall} \quad \text{(wzór 3.3)}$$

Parametr *Precision* określa jak wiele spośród zaakceptowanych samogłosek zostało zaakceptowanych poprawnie, *Recall* stwierdza, jak wiele spośród samogłosek, które powinny zostać zaakceptowane, zostało zaakceptowane, a *F* jest jednoliczbowym wskaźnikiem dającym pojęcie o jakości systemu.

3.3. Testowanie z użyciem pierwszego zestawu danych

Pierwszy zestaw danych składa się z samogłosek {a, e, y, i, o, u} wypowiadanych przez ośmiu różnych mówców. Wyniki testowania przedstawione są w tabelach.

W pierwszej kolejności dokonano testów, jakie kryterium kwalifikacji samogłoski powinno zostać wybrane podczas przeprowadzania dalszych testów. Problem polega na tym, że samogłoska opisana przez maksymalnie cztery formanty może uzyskać maksymalnie cztery wyniki klasyfikacji, tzn. każdy formant klasyfikowany jest osobno. Aby dokonać klasyfikacji samogłoski, należy sformułować jeden wynik klasyfikacji na podstawie uzyskanych czterech. Na tym etapie rozważane były dwie koncepcje: jedna z nich, nazwana lżejszym kryterium, polega na wyciągnięciu średniej arytmetycznej ze wszystkich wyników klasyfikacji, natomiast druga – ostrzejsze kryterium – to wybór tego wyniku klasyfikacji, który wypadł najgorzej.

Testowanie obydwu kryteriów odbyło się na całym zbiorze danych, tj. wszystkich dostępnych samogłoskach, czyli zbiorze {a, e, y, i, o, u}. Porównanie wyników przedstawiono w tab. 3.1. Po wyłonieniu samogłosek dających najlepsze wyniki rozpoznania w dalszym etapie testowania tego zbioru danych, przeprowadzono jeszcze raz testy kryteriów, w celu zweryfikowania ich słuszności. Testy te przedstawiono w tab. 3.2 i uwzględniają one zbiór samogłosek {a, e, y}.

Wyniki przedstawione w tab. 3.1 pozwoliły jednoznacznie stwierdzić, że lepszym kryterium będzie kryterium ostrzejsze – parametr *Precision* wzrasta z wartości 0.38 do wartości 0.55, a *F* wzrasta z wartości 0.55 do wartości 0.71, co oznacza, że identyfikacja przebiegła sprawniej.

Tabela 3.1. Porównanie skuteczności identyfikacji przy zastosowaniu kryterium średniej z wyników poszczególnych formantów (lżejsze) i kryterium najgorszego wyniku ze wszystkich formantów (ostrzejsze). Oznaczenia: TA – liczba poprawnie zaklasyfikowanych samogłosek, TR – liczba poprawnie odrzuconych samogłosek, FA – liczba fałszywie zaklasyfikowanych samogłosek, FR – liczba fałszywie odrzuconych samogłosek. Precision, Recall i F opisane są wzorami (3.1), (3.2), (3.3).

	TA	TR	FA	FR	Precision	Recall	F
Lżejsze kryterium	5547	91966	8954	0	0.38253	1	0.55337
Ostrzejsze kryterium	5547	96315	4605	0	0.54639	1	0.70667

Tabela 3.2 Ponowne porównanie skuteczności identyfikacji przy kryterium lżejszym oraz ostrzejszym, tym razem na zbiorze danych ograniczonym do samogłosek {a, e, y}.

	TA	TR	FA	FR	Precision	Recall	F
Lżejsze kryterium	2325	47670	1686	0	0.57966	1	0.7339
Ostrzejsze kryterium	2325	48685	671	0	0.77603	1	0.8739

W tab. 3.1 oraz 3.2 można zauważyć, że zgodnie z założeniem związanym z doborem progu klasyfikacji, liczba poprawnie zaklasyfikowanych (*TA*) oraz fałszywie odrzuconych (*FR*) samogłosek nie ulega zmianie.

Wybór kryterium klasyfikacji umożliwił przejście do kolejnego etapu: sprawdzenia, jak dobrze przebiega rozpoznanie w zależności od wybranej samogłoski. Porównanie wyników z klasyfikacji na podstawie pojedynczych samogłosek zostało przedstawione w tab. 3.3 – należy mieć na uwadze, że testów dokonano przy użyciu takiej liczby formantów, jaką dysponowano. Oznacza to, że testowane są tutaj samogłoski opisano zarówno przez cztery formanty, jak i przez jeden.

Trzy samogłoski cechujące się najlepszymi wynikami klasyfikacji to samogłoski *y*, *a* i *e*. Osiągnęły one parametr *Precision* na poziomach odpowiednio 0.82, 0.81 i 0.72, oraz parametr *F* na poziomach 0.9, 0.89 i 0.83. Pozostałe trzy samogłoski – *u, i* oraz *o* – osiągnęły gorsze wyniki, z których najlepszy to *Precision* na poziomie 0.63 oraz *F* wynoszące 0.77 dla samogłoski *u*, a najgorszy to *Precision* równe 0.41 i *F* wynoszące 0.59 dla samogłoski *o*.

Na podstawie otrzymanych wyników można zatem wyróżnić trzy samogłoski – *y, a, e*, które uzyskały parametr *Precision* powyżej poziomu 0.7 i parametr *F* powyżej poziomu 0.8, tym samym uzyskując znacznie lepsze rozpoznanie od pozostałych trzech – *u, i,* o. Jest to podstawa, aby wnioskować o ich większej przydatności podczas rozpoznawanie mówców. Należy zwrócić jednak uwagę, że różni się liczba dostępnych prób dla każdej z samogłosek – np. samogłoska *o* testowana była 32223 razy, a samogłoska *i* tylko 6969 razy, co daje około 4.5 razy mniej danych w jej przypadku. Stwierdzenie zatem, czy na pewno możliwe jest porównywanie uzyskanych w ten sposób wyników wymaga przeprowadzenia dalszych badań.

Tabela 3.3. Porównanie wyników rozpoznania w grupie 8 mówców z pomocą procedury w zależności od rozpoznawanej samogłoski. Zbiór testowanych samogłosek to {a, e, y, i, o, u}. Podobnie jak w tabelach 3.1 i 3.2, kolumny od drugiej do piątej opisują liczbę przypadków przyporządkowanych do odpowiedniej kategorii klasyfikacji, a kolumny od szóstej do ósmej przedstawiają parametry opisujące jakość rozpoznania.

Samogłoska	TA	TR	FA	FR	Precision	Recall	F
a	835	18497	195	0	0.81068	1	0.89544
o	2327	26509	3387	0	0.40725	1	0.57878
u	648	14566	380	0	0.63035	1	0.77327
e	892	16427	345	0	0.7211	1	0.83795
i	247	6555	167	0	0.59662	1	0.74735
y	598	13760	132	0	0.81918	1	0.9006

3.4. Testowanie z użyciem drugiego zestawu danych

Drugi zestaw danych składa się z samogłosek {a, e, y, i, o, u} wypowiadanych przez dwudziestu jeden mówców. Wyniki testowania, podobnie jak w poprzednim punkcie, przedstawione są w tabelach.

Ostrzejsze kryterium klasyfikacji zostało przejęte z poprzedniego zbioru danych, tak więc samogłoska zostaje uznana za zaklasyfikowaną, jeżeli najgorszy spośród czterech wyników klasyfikacji opisujących ją formantów zmieścił się poniżej progu ustalonego dla mówcy, względem którego samogłoska jest testowana.

W tab. 3.4, podobnie jak w tab. 3.3, zostało przedstawione porównanie wyników klasyfikacji samogłosek w zależności od wybranej samogłoski z grup {a, e, y, i, o, u} przy użyciu tylu formantów, iloma dysponowano w danym przypadku. Do opisu jakości klasyfikacji dodano jeden nowy parametr, opisany w tabeli jako *Ranga*. Jest on obliczony jako iloraz średniej z miejsc, które zajmują poprawnie zaklasyfikowane samogłoski (*TA*) i średniej liczby wszystkich samogłosek poddawanych klasyfikacji w ramach jednego kontekstu samogłoski – przykładowo, jeżeli podczas rozpoznawania samogłoski *a* mocno akcentowanej, w otoczeniu spółgłosek *p* i *s*, należącej do mówcy *X*, zostały poprawnie rozpoznane 3 samogłoski należące do *X* i zajęły miejsca 1, 3 i 5 (wg najlepszych wyników), oraz zostało zaklasyfikowanych błędnie 5 samogłosek należących do innych mówców, parametr *Ranga* zostanie obliczony jako iloraz liczb 3 (średnio zajęte miejsce) i 8 (liczba wszystkich zaklasyfikowanych samogłosek) oraz podany w skali procentowej jako 37.5%. Jego proponowana interpretacja jest taka, że dana samogłoska średnio uzyskuje gorszy wynik niż podany procent wszystkich rozpoznawanych samogłosek.

Należy również zwrócić uwagę, że w tab. 3.4 wartości *FR* nie są zerowe. Wynika to z błędów procedury występujących w przypadku, gdy dwie samogłoski tego samego mówcy, porównywane do siebie, nie zawierają wspólnych formantów – przykładowo, jedna z prób jest opisana przez F1 i F2, a druga tylko przez F3. Wskutek braku materiału do porównania, samogłoska taka jest odrzucona przez system. Skala tego błędu nie jest jednak duża – tam, gdzie występuje najwięcej *FR*, czyli w testach samogłoski *i*, są to 142 przebiegi na 161317, co daje ok. 0.09% wszystkich testów.

Tabela 3.4. Porównanie wyników rozpoznania w grupie 21 mówców z pomocą procedury w zależności od rozpoznawanej samogłoski. Zbiór testowanych samogłosek to {a, e, y, i, o, u}. Podobnie jak w tab. 3.3, kolumny od drugiej do piątej opisują liczbę przypadków przyporządkowanych do odpowiedniej kategorii klasyfikacji, a kolumny od szóstej do ósmej przedstawiają parametry opisujące jakość rozpoznania. W ostatniej rubryce znajduje się nowy parametr, opisujący średnią pozycję, jaką samogłoski poprawnie zaklasyfikowane zajęły pośród wszystkich zaklasyfikowanych samogłosek w stosunku do średniej liczby sprawdzanych samogłosek.

Samogłoska	TA	TR	FA	FR	Precision	Recall	F	Ranga
a	6511	159943	23983	52	0.21352	0.99207	0.3519	17.8 %
o	4272	131138	24407	36	0.14896	0.99164	0.25929	26.9 %
u	2252	62876	12080	34	0.15713	0.98512	0.27159	24.6 %
e	8076	222176	55282	56	0.12747	0.99311	0.22611	27.4 %
i	3071	53233	23037	42	0.11763	0.98650	0.21049	33.4 %
y	5175	123641	32501	142	0.13736	0.97329	0.24153	29.3 %

Wyniki przedstawione w tab. 3.4 są różne od tych przedstawionych w tab. 3.3. Przede wszystkim, jakość rozpoznania drastycznie zmalała – jest to rezultatem znacznego zwiększenia zbioru danych. Najlepsze rozpoznanie wyszło w przypadku samogłoski *a*, która osiągnęła *F* równe 0.35 oraz *Ranga* na poziomie 17.8%, natomiast na drugim i trzecim miejscu znajdują się samogłoski *u* oraz *o*, z wynikami *F* odpowiednio 0.27 i 0.26 oraz *Rangą* na poziomach 24.6% i 26.9%. Najgorsze rozpoznanie w tym zestawie przypada samogłosce *i* – *F* znajduje się na poziomie 0.21, a *Ranga* wynosi 33.4%.

W tab. 3.5, 3.6, 3.7, 3.8, 3.9 oraz 3.10 przedstawiono wyniki rozpoznania dla różnych samogłosek w zależności od wyboru formantu lub formantów. Sprawdzono następujące konfiguracje: {F1}, {F2}, {F3}, {F4}, {F1, F2, F3, F4}, {F2, F3, F4}, {F2, F3} oraz {F3, F4}.

W przypadku każdej z samogłosek {a, e, y, i, o, u} następuje tym lepsze rozpoznanie mówców, im wyższy formant zostaje wybrany – *F* wzrasta nawet ok. 2.5 razy w przypadku samogłoski *i*. Wyjątkiem jest jedynie samogłoska *o*, gdzie najlepsze rozpoznanie uzyskane jest poprzez wybranie formantu F3. Wśród samogłosek *a, o, e, i, y* zauważalny jest wzrost parametru *Ranga* na formancie F2, co wskazywałoby na większe podobieństwo fałszywie zakwalifikowanych (*FA*) do poprawnie zakwalifikowanych (*TA*) przebiegów formantowych niż w wypadku innych formantów. Podobnie jak w wypadku testowania zależności rozpoznania od wybranej samogłoski, należy też zauważyć, że liczba dostępnych do testowania formantów różni się znacznie w zależności od wybranego formantu – przykładowo, dla samogłoski *a* formant F1 przetestowano 165713 razy, a formant F4 tylko 21267 razy, co stanowi ok. 12.8% liczby testów formantu F1. Porównanie wyników testowania wszystkich czterech formantów do testowania z każdego osobna jest tym bardziej problematyczne, ze względu na stosunkowo niewielką ilość samogłosek opisanych przez wszystkie cztery formanty.

Tabela 3.5. Porównanie wyników rozpoznania w grupie 21 mówców z pomocą procedury w zależności od wybranych formantów. Wyniki w tabeli przedstawione są dla samogłoski *a*.

A	TA	TR	FA	FR	Precision	Recall	F	Ranga
F1	5949	119520	40244	0	0.12879	1	0.22818	23.3%
F2	3887	69201	20687	0	0.15818	1	0.27315	26.2%
F3	3161	49326	10865	0	0.22537	1	0.36784	22.4%
F4	1588	15709	3970	0	0.28571	1	0.44444	22.7%
F1, F2, F3, F4	711	4345	198	0	0.78218	1	0.87778	10.6%
F2, F3, F4	819	5230	326	0	0.71528	1	0.83401	12.8%
F3, F4	1043	7564	689	0	0.60219	1	0.75171	14.9%
F2, F3	2205	30317	2630	0	0.45605	1	0.62642	16.7%

Tabela 3.6. Porównanie wyników rozpoznania w grupie 21 mówców z pomocą procedury w zależności od wybranych formantów. Wyniki w tabeli przedstawione są dla samogłoski *o*.

O	TA	TR	FA	FR	Precision	Recall	F	Ranga
F1	4122	113313	33218	0	0.11039	1	0.19883	30.3%
F2	1545	29395	6488	0	0.19233	1	0.32261	32.1%
F3	1559	29679	5371	0	0.22496	1	0.3673	29.5%
F4	823	11299	2903	0	0.22088	1	0.36184	33.2%
F1, F2, F3, F4	261	1853	45	0	0.85294	1	0.92063	18.3%
F2, F3, F4	266	1889	92	0	0.74302	1	0.85256	21.7%
F3, F4	459	4944	428	0	0.51747	1	0.68202	21.7%
F2, F3	760	9712	648	0	0.53977	1	0.70111	25.2%

Tabela 3.7. Porównanie wyników rozpoznania w grupie 21 mówców z pomocą procedury w zależności od wybranych formantów. Wyniki w tabeli przedstawione są dla samogłoski *u*.

U	TA	TR	FA	FR	Precision	Recall	F	Ranga
F1	1969	48405	15336	0	0.11378	1	0.20432	31.2%
F2	733	10583	2962	0	0.19838	1	0.33107	29.8%
F3	686	11563	2756	0	0.1993	1	0.33236	34.4%
F4	347	3741	1273	0	0.2142	1	0.35282	33.9%
F1, F2, F3, F4	39	255	36	0	0.52	1	0.68421	26.9%
F2, F3, F4	48	327	62	0	0.43636	1	0.60759	27.9%
F3, F4	125	1316	198	0	0.387	1	0.55804	32.4%
F2, F3	244	2564	370	0	0.39739	1	0.56876	28.2%

Tabela 3.8. Porównanie wyników rozpoznania w grupie 21 mówców z pomocą procedury w zależności od wybranych formantów. Wyniki w tabeli przedstawione są dla samogłoski *e*.

E	TA	TR	FA	FR	Precision	Recall	F	Ranga
F1	7592	172448	84618	0	0.082334	1	0.15214	33.7%
F2	4847	94295	45635	0	0.096014	1	0.17521	37.4%
F3	3857	78581	23430	0	0.14135	1	0.24769	35.7%
F4	1692	23924	6388	0	0.20941	1	0.3463	34.6%
F1, F2, F3, F4	673	6849	342	0	0.66305	1	0.79739	22.8%
F2, F3, F4	696	7003	469	0	0.59742	1	0.74798	26.6%
F3, F4	1071	13476	1367	0	0.43929	1	0.61043	30.1%
F2, F3	2583	47026	8952	0	0.22393	1	0.36592	29.6%

Tabela 3.9. Porównanie wyników rozpoznania w grupie 21 mówców z pomocą procedury w zależności od wybranych formantów. Wyniki w tabeli przedstawione są dla samogłoski *i*.

I	TA	TR	FA	FR	Precision	Recall	F	Ranga
F1	2991	43229	29885	0	0.090978	1	0.16678	37.1%
F2	914	15062	5835	0	0.13543	1	0.23855	45.3%
F3	963	15577	4404	0	0.17943	1	0.30427	44%
F4	486	5048	1221	0	0.28471	1	0.44323	43.1%
F1, F2, F3, F4	105	1088	37	0	0.73944	1	0.8502	39.3%
F2, F3, F4	110	1114	48	0	0.6962	1	0.8209	41.6%
F3, F4	290	2654	287	0	0.5026	1	0.66897	44.9%
F2, F3	361	5384	1032	0	0.25915	1	0.41163	38.2%

Tabela 3.10. Porównanie wyników rozpoznania w grupie 21 mówców z pomocą procedury w zależności od wybranych formantów. Wyniki w tabeli przedstawione są dla samogłoski *y*.

Y	TA	TR	FA	FR	Precision	Recall	F	Ranga
F1	4783	90370	50265	0	0.086888	1	0.15988	35.3%
F2	2770	45526	24247	0	0.10253	1	0.18599	39.5%
F3	2715	44143	15980	0	0.14523	1	0.25362	34.6%
F4	1191	12085	6541	0	0.15404	1	0.26695	37.7%
F1, F2, F3, F4	597	5874	1165	0	0.33882	1	0.50615	24.5%
F2, F3, F4	610	5924	1331	0	0.31427	1	0.47824	28.4%
F3, F4	888	8681	2622	0	0.25299	1	0.40382	31.7%
F2, F3	1677	26384	6763	0	0.1987	1	0.33152	30%

Można natomiast porównać między sobą wyniki testowania z uwzględnieniem wszystkich czterech formantów do wyników testowania z uwzględnieniem tylko trzech ostatnich – F1 zostaje wtedy pominięty ze względu na najmniejszy wkład w rozpoznanie mówcy. Różnice w rozpoznaniu nie są duże – największy odnotowany spadek *F* wynosi niecałe 0.08 w przypadku samogłoski *u*, a *Ranga* wzrasta najwyżej o 3.8% w przypadku samogłoski *e*. Przy interpretacji wyników rozpoznania dla samogłosek *u* (tab. 3.7) oraz *i* (tab. 3.9) należy zachować ostrożność – danych w zbiorze testowym jest bardzo mało.

Idąc dalej w tym kierunku, przetestowano również konfigurację tylko dwóch formantów – {F2, F3} oraz {F3, F4}, w celu stwierdzenia, czy różnica w rozpoznaniu będzie duża. Punktem odniesienia jest rozpoznanie przy zastosowaniu wszystkich czterech formantów. W wypadku konfiguracji {F2, F3}, zauważalny jest znaczny spadek w poprawnym rozpoznaniu dla każdej samogłoski – zakres tego spadku wynosi od ok. 0.12 (samogłoska *u*) do ok. 0.44 (samogłoska *i*) dla parametru *F*. Podobnie parametr *Ranga* ulega pogorszeniu w każdym wypadku, z wyjątkiem jedynie samogłoski *i*, gdzie następuje niewielka poprawa o ok. 1.1%. Konfiguracja {F3, F4} wypada lepiej od konfiguracji {F2, F3} w wypadku samogłosek *a, e, i, y* uzyskując nawet o 0.25 wyższy parametr *F*, natomiast niewiele gorzej w wypadku samogłosek *o* i *u*, gdzie *F* spada w granicach 0.1 – 0.2.

4. Dyskusja i wnioski

Zastosowana w pracy metoda jest metodą prostą, nie wymagającą trenowania specyficznego modelu, ze względu na naturę danych. Niewielka liczba prób – maksymalnie trzy – opisująca samogłoskę z uwzględnieniem jej akcentacji i otoczenia fonemów wymagałaby stworzenia dobrze dopracowanego modelu w pełni automatycznego, aby rozpoznanie mierzalne ilościowo mogło dać dobre i w pełni wiarygodne wyniki. Jednakże, w świetle otrzymanych wyników, prosta procedura okazuje się być wystarczająco dobra do rozpoczęcia sprawdzania zależności pomiędzy jakością rozpoznania mówcy a zastosowanymi w tym celu parametrami. Należy jednak mieć na uwadze kilka czynników, które stwarzają prawdopodobieństwo, że otrzymane wyniki nie odwzorowują w pełni rzeczywistości.

Przede wszystkim, podstawowym problemem jest wrażliwość procedury na ilość dostępnych danych. Zastosowana metoda testowania ma zarówno cechy metody weryfikacyjnej – stosowanie progów rozpoznania – oraz identyfikacyjnej – określenie miary podobieństwa próby do wzorca w odniesieniu do pozostałych prób. W swojej książce L. R. Rabiner przestrzega, że metody weryfikacji mają pewną charakterystyczną przypadłość – nawet przy systemach bardzo wyrafinowanych, wraz ze wzrostem liczby mówców, rośnie prawdopodobieństwo, że znajdzie się osoba o takich cechach charakterystycznych, że zostanie zaakceptowana przez system mimo, że nie jest mówcą wzorcowym [2]. Tym bardziej problem ten tyczy się prostej procedury proponowanej w pracy – jest to zauważalne szczególnie podczas porównywania wyników rozpoznania zaprezentowanych w tab. 3.3 i 3.4. Duże rozbieżności w liczbie danych dostępnych dla różnych samogłosek (przykładowo, w tab. 3.4. różnica między liczbą testowań samogłoski *e* i samogłoski *i* wynosi 207 193) powodują, że wyniki zaprezentowane w tab. 3.4 nie mogą zostać uznane za wiarygodne. Problem ten tyczy się w dużo mniejszym stopniu wyników zaprezentowanych w tab. 3.3 – tam różnica między liczbą testowań tych samych samogłosek wynosi 10 695. Pomimo to, należy pamiętać, że dokładny wpływ przyrostu liczby danych nie został zbadany w tej pracy.

Pomimo, że mogłoby to wydawać się dobrym rozwiązaniem, nie dokonano ograniczenia zbioru danych w taki sposób, by ujednolicić liczbę danych dostępnych dla każdej z samogłosek. Po pierwsze skutkiem takiej operacji byłoby odrzucenie bardzo dużej liczby danych (co widać, gdy weźmie się pod uwagę wspomniane różnice w liczbie testowań), a po drugie nie jest jasne, którą część nadmiarowych danych należałoby wybrać do testowania, a którą odrzucić. Nie została również zastosowana kroswalidacja (ang. *crossvalidation*), ponieważ wtedy do ustalenia progu odniesienia dla danego mówcy zostałyby zastosowane w najlepszym wypadku dwie próby, a w pozostałych wypadkach ustalenie takiego progu nie byłoby możliwe – są do tego potrzebne przynajmniej dwie próby ze względu na naturę algorytmu DTW. Dzięki zastosowaniu pełnego zbioru danych do testowania, możliwe było wybranie progu spośród trzech różnych możliwości (odległości pomiędzy próbą pierwszą i drugą, drugą i trzecią oraz pierwszą i trzecią).

Korzystając z wyników pierwszego zbioru danych, gdzie co prawda dostępnych było mniej mówców, ale różnice pomiędzy liczbami testowań nie były tak duże, jak w drugim zbiorze danych, można stwierdzić, że samogłoski *a, e* i *y* mocniej charakteryzują mówców niż pozostałe, tj. *o, u* oraz *i*. Nie zostało to potwierdzone podczas testowania drugiego zbioru danych, jednak występujące w nim duże różnice pomiędzy liczbą dostępnych danych zakłócają wyniki testów, czyniąc je niewiarygodnymi. W związku z brakiem potwierdzenia wyników, nie należy stwierdzenia o wyższości wspomnianych trzech samogłosek podczas identyfikacji traktować jako kategorycznego, a raczej jako wskazówkę przy doborze priorytetowych samogłosek.

Problem zbyt różnej liczby danych w znacznie mniejszym stopniu dotyczy testów przeprowadzonych w celu sprawdzenia zależności jakości rozpoznania od wyboru formantów. W tab. 3.5, 3.6, 3.7, 3.8, 3.9 oraz 3.10 widać, że podczas testowania różnych konfiguracji formantów liczba testowań pozostaje w każdym wypadku stosunkowo mała, szczególnie pomiędzy konfiguracjami {F1, F2, F3, F4} oraz {F2, F3, F4}. Na podstawie tych testów stwierdzono, że im wyższy formant jest brany pod uwagę, tym większy ma on wpływ na wynik klasyfikacji, a tym samym – niesie więcej informacji o charakterystycznych cechach mówcy. To stwierdzenie potwierdzają testy konfiguracji różnych formantów, gdzie, co prawda, najlepsze rezultaty osiągane są, gdy weźmie się pod uwagę każdy z formantów, ale rezultaty osiągane przez testowanie

dwóch najwyższych formantów (F3 i F4) są lepsze lub porównywalne, niż gdy testowane są formanty środkowe (F2 i F3). W wypadku, gdy liczbę parametrów trzeba by z jakiegoś powodu ograniczyć, lepiej zatem zrezygnować z zastosowania najniższego formantu.

Wprowadzony na potrzeby uzyskania dodatkowych wyników z testowania drugiego zbioru danych parametr *Ranga* daje pogląd na to, jaką pozycję zajmują poprawnie zaklasyfikowane przebiegi wśród wszystkich pozostałych. Pozwala on na odróżnienie, czy przebieg należący do referencyjnego mówcy średnio znalazł się np. na piątym miejscu ze stu, czy na piętnastym. W wypadku zastosowania tej lub podobnej procedury do dokonania identyfikacji – lub uproszczenia tego procesu przez zawężenie liczby prób, które trzeba wziąć pod uwagę – pozwala on na określenie, ile prób o najlepszych wynikach należy faktycznie rozpatrzeć. Pomógł on również sformułować część wniosków w pracy, wykazując pewną korelację z parametrem *F*, a także upewnić się, że procedura radzi sobie z obróbką drugiego, większego zestawu danych.

5. Bibliografia

[1] A. Trawińska (2009): *Analiza mowy i nagrań*, (w:) *Postępy w naukach sądowych*, Kała M. (red.), Wydawnictwo Instytutu Ekspertyz Sądowych, Kraków, 117-134.

[2] L. R. Rabiner, R. W. Schafer, *Digital Processing of Speech Signals*, New Jersey 1978.

[3] K. Klus, A. Trawińska, *Forensic Speaker Identification by the Linguistic-Acoustic Method in KEÚ AND IES* (w:) Problems of Forensic Sciences 2009, vol. LXXVIII, 160-174.

[4] S. B. Davis and P. Mermelstein, *Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences*, IEEE Transactions on Acoustics, Speech and Signal Processing, 1980, vol. ASSP-28, pp. 357-366, no. 4.

[5] J. Makhoul, *Spectral linear prediction: properties and applications*, IEEE Transcations, 1975, vol. ASSP-23, pp. 283-296.

[6] B. Ziółko, W. Kozłowski, M. Ziółko, R. Samborski, D. Sierra, J. Gałka, *Hybrid Wavelet-Fourier-HMM Speaker Recognition*, *International Journal of Hybrid Information Technology*, vol. 5, No. 4, October, 2011.

[7] B. Ziółko, J. Gałka, M. Ziółko, *Polish phoneme statistics obtained on large set of written texts*, dane o publikacji

[8] B. Ziółko, M. Ziółko, *Przetwarzanie mowy*, Wydawnictwa AGH, 2011.

[9] T. Zieliński, *Cyfrowe przetwarzanie sygnałów*, Wydawnictwo Komunikacji i Łączności, 2009.

[10] R. Tadeusiewicz, *Sygnał Mowy*, Wydawnictwo Komunikacji I Łączności, Warszawa 1988.

[11] M. Kłaczyński, *Zjawiska wibroakustyczne w kanale głosowym człowieka*, Akademia Górniczo-Hutnicza im. Stanisława Staszica, Kraków, 2007.

[12] Nolan F. *Speaker Recognition and Forensic Phonetics*, (in) Hardcastle W.J., Laver J., *The handbook of Phonetic Sciences*, Blackwell Publishers, Oxford, 1999, pp. 744–767

[13] http://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-345-automatic-speech-recognition-spring-2003/lecture-notes/

[14] http://www.kfs.oeaw.ac.at/index.php?lang=en

[15] H. Sakoe, S. Chiba, *Dynamic Programming Algorithm Optimization for Spoken Word Recognition*, (w:) IEEE Transatctions of Acoustics, Speech and Signal Processing, vol. ASSP-26, No. 1, Luty 1978