AKADEMIA GÓRNICZO-HUTNICZA W KRAKOWIE
IM. STANISŁAWA STASZIC
WYDZIAŁ ELEKTROTECHNIKI, AUTOMATYKI,
INFORMATYKI I ELEKTRONIKI

ROZPOZNAWANIE MÓWCY NA PODSTAWIE DŁOGOOKRESOWEGO HISTOGRMU AMPLITUD SYGNALU MOWY

(ROZPRAWA DOKTORSKA)

ADEL SHOMALI

PROMOTOR:
prof. zw. dr hab. inż. RYSZARD TADEUSIEWICZ

KRAKÓW 1999

Promotorowi prof. zw. dr hab. inż. Ryszardowi Tadeusiewiczowi
Rodzinie – ojcu Afif, matce Afaf, żonie Agnieszce
oraz braciom: Ala'a, Isam i Adi

Podziękowania

Autor pragnie podziękować, przede wszystkim, **promotorowi prof. zw. dr hab. inż. Ryszardowi Tadeusiewiczowi** za wieloletnią opiekę naukową, między innymi za wprowadzenie w świat nauki i podbudowanie wiary we własne możliwości oraz za cenną pomoc przy pisaniu pierwszych opracowań naukowych. Szczególne podziękowania, które bardzo trudno wyrazić słowami, należą się Profesorowi również za Jego czas poświęcony w okresie bezpośredniej pracy nad rozprawą doktorską, jak również za okazywaną wyrozumiałość, cierpliwość oraz za wiele inspirujących i wyczerpujących uwag zgłaszanych w czasie licznych konsultacji. Autor bardzo ceni i szczególnie mile wspomina Jego wnikliwe poprawki do tekstu rozprawy w tym także te o charakterze czysto językowym. Pozwoliło to na uniknięcie wielu usterek merytorycznych i niedociągnięć językowych przyczyniając się do wyraźnego podniesienia wartości naukowej rozprawy przy jednoczesnym gruntownym polepszeniu jej jakości stylistycznej oraz językowej.

Serdeczne podziękowania należą również prof. dr hab. inż. Czesławowi Baszturze za bezinteresowne udostępnienie baz danych głosów i pomoc w ogólnym ukierunkowaniu badań oraz za cenne uwagi dotyczące merytorycznej organizacji rozprawy i jej poprawności językowej.

Autor poczuwa się do miłego obowiązku podziękowania również wszystkim, nie wymienionym z powodu braku miejsca, swoim wykładowcom zarówno ze studiów magisterskich jak i doktoranckich.

Ciepłe wyrazy wdzięczności składam także mojej rodzinie za cierpliwość, pomoc oraz wsparcie podczas mojego pobytu w Polsce. Kolejne wyrazy wdzięczności należą się Mojej Żonie Agnieszce za wsparcie podczas studiów oraz za wstępną poprawę językową rozprawy.

SPIS TREŚCI

ROZDZIAŁ I

Wprowadzenie

1.1. Tematyka, teza oraz cele rozprawy

1.1.1. Tematyka

Rozpoznawanie głosów należy do grupy biometrycznych metod identyfikacji osób. Inne techniki z tej samej grupy polegają na identyfikacji osób na podstawie odcisków palców, kształtu dłoni, obrazu siatkówki oka, wyglądu twarzy oraz badań nad strukturą genetyczną DNA wydobytego z próbek tkanek. Prezentowana rozprawa zajmuje się problematyką automatycznego rozpoznawania mówcy na podstawie histogramów amplitudowych wydobytych z postaci czasowej sygnału mowy.

1.1.2. Teza

Opis sygnału mowy za pomocą długookresowych histogramów amplitudowych zachowuje cechy indywidualne głosu w stopniu umożliwiającym rozpoznawanie mówców.

1.1.3. Cele rozprawy

W pracy podjęto ponadto szczegółowe zagadnienia badawcze. Stanowiły one próby wykazania, że:

1. Zwiększenie długości wypowiedzi, na podstawie których buduje się model głosu danego mówcy, przy zachowaniu przypadkowości ich składu, zwiększa jakość rozpoznawania.
2. Przeprowadzenie rozpoznawania mówcy na podstawie dostatecznie długiej wypowiedzi, składającej się z przypadkowych słów, pozwala na uzyskanie dobrej jakości rozpoznawania niezależnie od następujących czynników:
   - treści wypowiedzi oraz poprawności wymowy,
   - języka wypowiedzi i jej aspektu semantycznego,
   - przeciętnego poziomu szumu lub hałasu towarzyszącego rejestracji sygnału,
   - małej rozdzielczości amplitudowej użytej aparatury.
3. Segmentacja wypowiedzi w sposób ‘ślepy’ nie powinna prowadzić do pogarszania się wyników.
4. Rozpoznawanie mówcy jest również możliwe w przypadku gdy pasmo sygnału mowy jest ograniczone nawet do 1 kHz.
5. Możliwe jest przeprowadzenie rozpoznawania na podstawie wypowiedzi nagranych w jednym języku, nawet gdyby model głosu rozpoznawanej osoby został zbudowany z użyciem wypowiedzi rejestrowanych w innym języku.
6. Rozpoznawanie jest częściowo możliwe również w przypadku otwartego zbioru mówców.

W celu wykonania powyższych badań zostały stworzone narzędzia w postaci bazy danych wejściowych (nagrania głosów) oraz odpowiedniego oprogramowania w języku MATLAB.

1.2. Rozmiar problemu automatycznego rozpoznawania mówcy

Rozpoznawanie mówcy dokonuje się na podstawie pewnych wybranych parametrów wydobytych (obliczanych) z sygnału mowy. W fazie trenowania systemu buduje się dla każdej osoby bazę danych zawierającą wzorcowe wartości tych parametrów. W fazie rozpoznawania wartości te są porównywane z aktualnie wyliczonymi wartościami w celu określenia stopnia ich wzajemnego podobieństwa i w celu podjęcia decyzji o rozpoznawaniu określonej osoby lub o braku podstaw do rozpoznawania. Zatem podstawowym wymaganiem, jakie musi spełnić odpowiedni zestaw parametrów aby rozpoznawanie było możliwe jest zapewnienie dyskryminacji głosów różnych osób na podstawie wartości tych parametrów oraz powtarzalność wartości parametrów dla różnych wypowiedzi tej samej osoby. Za lepszy parametr uważa się taki, którego wartości są dokładnie powtarzalne (lub bardzo zbliżone) dla wypowiedzi tego samego mówcy i stosunkowo znacznie różniące się (dyskryminujące) dla wypowiedzi innych mówców.

Nie ulega wątpliwości, że wartości tych parametrów, nazywanych ogólnie wektorami cech, zależą – w mniejszym lub większym stopniu – od treści wypowiedzi. Zależność parametrów od treści wypowiedzi jest na tyle istotna, że powstały dwie klasy metod rozpoznawania mówców: zależnego od tekstu (*text dependent speaker recognition*) oraz niezależnego od tekstu *(text independent speaker recognition)*. Związane jest z tym kilka problemów, takich jak np. określenie optymalnej długości wypowiedzi wykorzystanych do trenowania metody lub dokonania poprawnego rozpoznawania. Problem ten istnieje dla obu klas metod rozpoznawania.

Cechy indywidualne głosu mogą się zmienić z upływem czasu. Zatem wartości wektorów cech ekstrahowanych z wypowiedzi rejestrowanych w jednej sesji nagrań (np. w fazie trenowania) mogą się różnić od tych zdobytych w innym okresie (np. w fazie rozpoznawania). Na ogół im dłuższy jest odstęp czasu między tymi sesjami tym bardziej ta różnica staje się widoczna. Cechy indywidualne głosu zależą w dodatku od aktualnego stanu zdrowia oraz stanu emocjonalnego mówcy. Jeśli ponadto warunki rejestracji sygnału są różne dla różnych sesji, to problem bardzo się komplikuje. Warunki te dotyczą całego kanału transmisyjnego sygnału od źródła do aparatury rejestrującej i obejmują obecność szumu lub zakłócenia oraz charakterystyki

przenoszenia zarówno kanału transmisyjnego jak i aparatury rejestrującej (w tym cechy mikrofonu).

Na podstawie powyższych argumentów można stwierdzić, że niektóre egzemplarze wektorów cech mogą być niepowtarzalne. Zarówno w fazie trenowania jak i rozpoznawania zalecane jest aby takie wektory nie były brane pod uwagę. Potrzebne jest zatem kryterium oceny na ile dany wektor jest reprezentatywny dla danego mówcy. Na podstawie takiego kryterium można opracować mechanizm selekcji wektorów cech służący do odrzucania niepowtarzalnych wektorów lub do redukcji liczby wektorów (długości ciągu uczącego) opisujących dany głos. Niezależnie od tego istnieje potrzeba określenia siły dyskryminacyjnej danego parametru w celu prowadzenia rankingu parametrów i wyboru tych najbardziej dyskryminujących. Im bardziej dany parametr przyjmuje odległe wartości dla różnych głosów, tym jego siła dyskryminacji jest większa. Ocena tej siły jest jednak ściśle związana z wyborem metody klasyfikacji. Ponadto siła dyskryminacji danego parametru jest zazwyczaj zróżnicowana dla różnych głosów. Problemy selekcji wartości parametrów oraz wyboru najbardziej dyskryminujących z nich są wspólne dla większości systemów rozpoznawania.

Mimo nieustannego postępu techniki komputerowej należy, jak dotąd, brać pod uwagę złożoność obliczeniową metody rozpoznawania oraz rozmiar pamięci potrzebnej do przechowywania opisu modeli poszczególnych głosów. Ponadto im większa jest liczba przykładów (wartości wektorów cech) reprezentujących dany głos, tym bardziej czasochłonny staje się proces rozpoznawania. Po przekroczeniu pewnego poziomu złożoności staje się on niemożliwy do realizacji w czasie rzeczywistym. Oprócz tego zdobycie dużej liczby przykładów wymaga rejestrowania dłuższych nagrań w fazie trenowania, co może być niewygodne zarówno dla użytkowników systemu jak i dla projektantów obrabiających te nagrania.

Niemal wszystkie wymienione wyżej problemy potęgują się w miarę wzrostu liczby mówców. Wraz z tym wzrostem zdolność parametrów do dyskryminacji ulega zazwyczaj osłabieniu. Rośnie bowiem łączna liczba przykładów (długość ciągu uczącego) a ich rozdzielenie na klasy odpowiadające poszczególnym mówcom staje się coraz trudniejsze. Ponadto może zachodzić konieczność redukcji liczby przykładów ze względu na rozmiar pamięci systemu lub niemożność przeprowadzenia procesu rozpoznawania w czasie rzeczywistym. Problem ten bardziej dotyka systemy identyfikacji mówcy niż weryfikacji. W przypadku, gdy dostęp do systemu jest nieograniczony oprócz prawidłowego rozpoznawania głosów, w przypadku mówców dla których

system został trenowany, należy zadbać oto by odrzucał on zdecydowanie obcych mówców.

Odrębna klasa problemów w przypadku systemów rozpoznawania głosu jest związana z współpracą użytkownika z systemem. Chociaż należy zapewnić użytkownikowi pewien stopień wygody korzystania z systemu, to jednak zachodzi konieczność trzymania się pewnych reguł mających na celu ochronę systemu przed intruzami (naśladowcami). Problem tkwi w tym, że ci ostatni są bardziej skłonni do współpracy z systemem niż upoważnieni użytkownicy.

Choć dokonano wielkiego postępu w dziedzinie badań nad automatycznym rozpoznawaniem mówcy, pozostało jeszcze wiele problemów czekających na lepsze rozwiązania. Większość z nich wynika ze zmienności warunków rozpoznawania spowodowanej zarówno przez zmiany w warunkach rejestracji oraz zmienne charakterystyki kanału jak i jest następstwem zmian spowodowanych przez samych mówców. Bardzo ważnym zagadnieniem jest zatem znalezienie parametrów stabilnych w dłuższych interwałach czasowych i w miarę niezależnych od zmian w sposobie mówienia (np. tempo i głośność mowy). Potrzebne są także nowe metody rozpoznawania odporne na zakłócenia towarzyszące kanałowi transmisyjnemu sygnału mowy przez łącza telefoniczne oraz niepodatne na przeciętny poziom szumu środowiska rejestracji sygnału.

Zagadnienie rozpoznawania mówców było przedmiotem wieloletnich badań. Jednak pomimo ogromnego postępu badań, problem pozostaje nie do końca rozwiązany. Występują ponadto trudności w porównywaniu wyników różnych metod rozpoznawania, gdyż większość została zweryfikowana z użyciem innej bazy głosów. Tym niemniej istnieje domniemana zgodność co do klasyfikacji przydatności dotychczas badanych parametrów opisu sygnału mowy do zadania rozpoznawania mówcy.

1.3. Rozprawa

Rozprawa przedstawia metodę automatycznego rozpoznawania mówcy oraz przykładowe wyniki przetwarzania sygnału mowy i rezultaty rozpoznawania. Praca zawiera siedem rozdziałów, bibliografię oraz aneks. W rozdziale pierwszym przedstawiono tematykę, tezę oraz cele rozprawy. Omówiono również pokrótce rozmiar problemu automatycznego rozpoznawania mówcy.

Drugi rozdział poświęcono strukturze systemu automatycznego rozpoznawania mówcy. Zawarto w nim, między innymi, zarys metod opisu i analizy sygnału mowy ze szczególnym uwzględnieniem parametrów opisu sygnału mowy istotnych z punktu

widzenia automatycznego rozpoznawania mówcy. Ponadto przedyskutowano kryteria oraz metody oceny skuteczności parametrów do dyskryminacji różnych głosów. Rozdział ten zawiera szczegółowy opis poszczególnych bloków systemu automatycznego rozpoznawania mówcy począwszy od akwizycji sygnału mowy a skończywszy na klasyfikacji. W rozdziale tym zamieszczono również opis sposobów rozdzielania przestrzeni cech i ich wpływ na błędy rozpoznawania.

Rozdział trzeci zajmuje się problematyką klasyfikacji i porównywania rożnych podejść do zagadnienia automatycznego rozpoznawania mówcy. Zawiera on opis aktualnego stanu wiedzy dotyczącej badań nad systemami automatycznego rozpoznawania mówcy oraz wyniki wcześniejszych eksperymentów z tego zakresu.

Rozdział czwarty poświęcono zagadnieniu reprezentacji sygnału mowy za pomocą histogramów amplitudowych będącej sercem, zaproponowanej w tej rozprawie, metody automatycznego rozpoznawania mówcy. Przedstawiono w nim różne warianty obliczenia wektorów cech (histogramów amplitud) oraz omówiono ich zależność od charakterystyk toru akwizycji sygnału mowy. Zamieszczono również schemat sprzętowej realizacji procesora histogramowego.

W piątym rozdziale opisano metodę automatycznego rozpoznawania mówcy na podstawie długookresowego histogramu amplitud sygnału mowy. Oprócz szczegółowego opisu metody i jej wariantów dokonano także w tym rozdziale wstępnej oceny przydatności zaproponowanego opisu histogramowego do dyskryminacji głosów różnych mówców.

Rozdział szósty przedstawia wyniki eksperymentów polegających na identyfikacji lub weryfikacji mówców. Eksperymenty te przeprowadzono z użyciem różnych baz danych głosów zarówno dla zamkniętego jak i dla otwartego zbioru rozpoznawanych mówców. Dokonano w nim oceny jakości zaproponowanej metody w zależności od różnych warunków przeprowadzenia rozpoznawania.

Rozdział siódmy zawiera podsumowanie rozprawy, opisuje zalety i wady metody oraz przedstawia kierunki dalszych badań. W aneksie zawarto spis funkcji napisanych w języku MATLAB dla celów testowania zaproponowanej metody. Kod źródłowy tych funkcji można znaleźć na dołączonej do rozprawy dyskietce.

ROZDZIAŁ II

Struktura systemu automatycznego rozpoznawania mówcy

2.1. Wprowadzenie

Procedurę rozpoznawania mówców można ogólnie podzielić na trzy etapy (rys. 2.1). Pierwszy etap obejmuje próbkowanie i kodowanie sygnału mowy. W drugim etapie wydzielane są parametry osobnicze głosu stanowiące podstawę do procesu klasyfikacji w trzecim etapie. Klasyfikacji dokonuje się na podstawie podobieństwa uzyskanych wartości parametrów próbek sygnału do ich odpowiedników określonych wcześniej (w tzw. procesie nauczania) dla poszczególnych osób. Proces ten polega na budowaniu modeli dla poszczególnych osób na podstawie wartości ich parametrów osobniczych. Wykorzystuje się przy tym tzw. ciągi uczące. Są to pewne grupy wypowiedzi przypisane znanym mówcom, dla których buduje się modele. Dla budowanych modeli i dla badanych próbek należy określić tzw. funkcje przynależności [Tad92] rozstrzygające stopień podobieństwa parametrów próbki w postaci fragmentu wypowiedzi i określonego wzorca parametrycznego (modelu głosu konkretnego mówcy).

Rys. 2.1: Schemat uproszczony procedury rozpoznawania.

2.2. Metody opisu i analizy sygnału mowy

Sygnał mowy może być analizowany i opisywany różnymi metodami, przy czym każda metoda ma swoje specyficzne zalety i wady. Metody analizy sygnału mowy można podzielić na: czasowe, częstotliwościowe, czasowo-częstotliwościowe, parametryczne oraz oparte na predykcji liniowej [Del93, Fur89, Mak75, Sch81, Tad88]. W obecnym czasie, niektóre z nich odgrywają rolę dominującą i są powszechnie stosowane, inne zaś – w tym czasowe – utraciły wiele ze swego znaczenia. Poniżej przedstawione są charakterystyki metod opisu sygnału mowy w bardzo zwartym skrócie. Obszerniejszy opis tych metod można znaleźć w [Del93, Rab74, Tad88].

Sygnał mowy w postaci czasowej zawiera w istocie wszystkie niezbędne do analizy i rozpoznawania elementy, ale raczej w niedogodnej formie. Na ogół czasowa postać sygnału jest bardzo skomplikowana i może zawierać wiele zbędnych szczegółów. Ponadto, w tej postaci, aspekt osobniczy i semantyczny są silnie związane co

powoduje niechętne stosowanie tej formy sygnału w badaniach nad automatycznym rozpoznawaniem głosu (czy też mowy).

Opis sygnału w dziedzinie częstotliwości jest tą rutynowo stosowaną i najbardziej przydatną formą jego opisu. W odniesieniu do wielu sygnałów, w tym też sygnału mowy, prawdziwe jest stwierdzenie, że widmo amplitudowe – w porównaniu z widmem fazowym – niesie o wiele więcej istotnych informacji o sygnale. Przyjmuje się zwykle, że struktura widma fazowego determinowana jest przede wszystkim przez czynniki losowe. Wskazane jest zatem rozdzielenie tych dwóch składowych co, po dokonaniu analizy widmowej, staje się banalnie proste. Nie brak również przesłanek biologicznych na przydatność analizy częstotliwościowej, gdyż stwierdzono [Tad88], że biologiczny nadajnik formuje, a biologiczny odbiornik analizuje – głównie widmo sygnału i to właśnie w jego amplitudowo-częstotliwościowej formie. Ponadto analiza częstotliwościowa jest punktem wyjścia do innych metod analizy jak np. analiza homomorficzna[1] pozwalająca na w miarę łatwe rozdzielenie wpływów generatora sygnału (źródła krtaniowego) i własności układu modulującego sygnał (kanału głosowego). Czysta analiza częstotliwościowa ma zastosowanie w odniesieniu do sygnału mowy jedynie w kontekście tzw. długoterminowego widma sygnału, uwzględniającego obraz widmowy sygnału w czasie wielokrotnie większym od okresów quasi-stacjonarności widma, lub w odniesieniu do tych fragmentów sygnału, dla których można przyjąć, że w czasie ich generacji widmo sygnału nie podlega istotnym zmianom. Badania wykazały jednak, że zarówno przy transmisji sygnału jak i przy jego automatycznym rozpoznawaniu istotniejsze są zmiany czasowe kształtu widma aniżeli sam kształt w ustalonej chwili. Przykładowo wykazano, że możliwe jest rozpoznawanie niektórych głosek przy słyszeniu jedynie poprzedzających i następujących po nich głosek bez tzw. stanu ustalonego rozważanej głoski!. Sam zaś stan ustalony głoski słuchany ‘w izolacji’ nie jest poprawnie rozpoznawany [Tad88]. Stwierdzono również, że słuchacze nie dostrzegają braku stanu ustalonego percypowanych fonemów. Powyższe argumenty przemawiają za zastosowaniem analizy czasowo-częstotliwościowej sygnału mowy uznawanej za najwłaściwszą i najpełniejszą formę jego opisu.

Analiza czasowo-częstotliwościowa sygnału mowy polega na badaniu zmiany jego widma w kolejnych chwilach czasowych. W celu dogodnego przedstawienia tych zmian wskazane jest stosowanie trójwymiarowej metody reprezentacji tworząc tzw.

[1] Głównie chodzi tu o wyznaczenie tzw. cepstrum sygnału.

obrazy dźwiękowe albo wideogramy. Pojedynczy obraz ma jeden wymiar czasowy (kolejne segmenty czasowe), drugi częstotliwościowy (wybrane pasma), trzeci zaś przedstawia amplitudy składowe widma dla konkretnego segmentu czasowego i konkretnego pasma. Niestety, wspomniane obrazy często nadal zawierają wiele szczegółów utrudniających i ograniczających możliwości ich dalszego wykorzystania. Dotyczy to między innymi ich analizy, porównywania ze wzorcami przy rozpoznawaniu, lub badania skutków zakłóceń i zniekształceń przy transmisji.

W związku z powyższym, uzupełnieniem – a niekiedy alternatywą – owych metod opisu sygnału mowy może być opis parametryczny [Del93, Tad88]. Wśród parametrów opisujących sygnał mowy wymienia się parametry czasowe, częstotliwościowe, korelacyjne, cepstralne oraz związane z techniką predykcji liniowej LPC. Wśród parametrów czasowych warto wymienić: obwiednię amplitud, moc, częstotliwość podstawową tonu krtaniowego, częstotliwość przejść przez zero [Bas93, Tad88], oraz dyskutowane w tej pracy histogramy amplitudowe [Sho98]. Warto podkreślić, iż istnieją różne koncepcje obliczenia oraz aproksymacji wartości tych parametrów. Wśród parametrów częstotliwościowych wymienia się najczęściej momenty widmowe, obwiednię widma amplitudowego (*spectral envelop*) oraz częstotliwości formantowe [Tad88]. Również w przypadku tych parametrów znane są różne metody ich wyznaczenia.

Często stosowaną metodą opisu sygnału mowy jest tzw. technika liniowej predykcji [Del93, Jua93, Mak75]. W kontekście analizy sygnału mowy, technika ta służy między innymi do jego opisu dla celów rozpoznawania oraz spakowania (redukcji ilości informacji). Biorąc pod uwagę, że sygnał głosowy powstaje w wyniku przekształcenia sygnału źródła krtaniowego w trakcie głosowym, metoda ta postuluje przewidywanie jego wartości w danej chwili na podstawie kombinacji liniowej jego poprzednich wartości oraz poprzednich i bieżącej wartości sygnału źródłowego. Ponieważ brak jest jakichkolwiek informacji o sygnale źródłowym, zakłada się jego przypadkowość. Zakłada się także, że wartości sygnału będą przewidywane tylko na podstawie kombinacji liniowej jego poprzednich wartości, przy czym dobór współczynników tej kombinacji (współczynników predykcji liniowej) dokonuje się tak, aby zminimalizować błąd aproksymacji (sumy kwadratów różnic pomiędzy wartościami prawdziwymi a przewidywanymi). Warto podkreślić, że parametry kodowania predykcyjnego określające charakterystykę kanału głosowego mają w związku z tym swoją interpretację fizyczną. Opisują one bowiem kształt kanału głosowego modelowanego jako połączenie kilku

cylindrów o takiej samej długości[2] ale o różnych przekrojach. Model taki jest opisany za pomocą transmitancji zawierającej same bieguny, przy czym ich liczba jest równa liczbie współczynników kodowania predykcyjnego. Parametry kodowania predykcyjnego mają również swoją interpretację w dziedzinie częstotliwości, bowiem opisują wygładzoną obwiednię amplitudową tego widma.

Równie ważna metoda opisu parametrycznego sygnału mowy wywodzi się z analizy cepstralnej. Główną zaletą tej metody jest możliwość rozdzielania składowej sygnału związanej z jego źródłem od tej wynikającej z charakterystyki kanału głosowego. Cepstrum sygnału jest definiowane jako odwrotna transformacja Fouriera logarytmu widma amplitudowego sygnału. Choć widmo sygnału jest obliczane bezpośrednio za pomocą transformaty Fouriera, to jednak często wykorzystuje się do wyznaczania cepstrumu aproksymację widma wyliczoną np. za pomocą parametrów LPC. Badania wykazały, że opis sygnału mowy za pomocą tak wyznaczonego cepstrum daje lepsze wyniki w systemach automatycznego rozpoznawania mowy [Jua93]. Cepstrum takie nosi nazwę LPC-cepstrum dla odróżnienia go od zwykłego cepstrum oznaczonego czasem jako FFT-cepstrum.

2.3. Parametry opisu sygnału mowy istotne z punktu widzenia automatycznego rozpoznawania mówcy

Biorąc pod uwagę problemy automatycznego rozpoznawania mówcy, dyskutowane w rozdziale pierwszym, można określić własności idealnych parametrów opisu sygnału mowy dla celów automatycznego rozpoznawania mówcy. Według [Wol72] najważniejsze własności idealnych parametrów są następujące:

1. Ich wartości są bardzo zbliżone (powtarzalne) dla tego samego mówcy i różniące się dla różnych mówców,
2. Pojawiają się często i naturalnie w mowie,
3. Są łatwe do mierzenia (obliczania),
4. Są mało zmienne w czasie oraz niezależne od stanu zdrowia mówcy,
5. Są niezależne od przeciętnego poziomu hałasu środowiska oraz charakterystyk ewentualnego systemu transmisyjnego,
6. Są niepodatne na próby zmiany głosu przez mówcę lub przynajmniej na próby naśladowania danego głosu.

[2] Liczba tych cylindrów jest równa liczbie współczynników liniowej predykcji.

W przypadku systemów automatycznego rozpoznawania mówcy działających niezależne od treści wypowiedzi (*text independent*), wymaga się również niezależność tych parametrów od tekstu.

Z punktu widzenia automatycznego rozpoznawania mówcy badania wykazały, że znaczenie mają przede wszystkim następujące parametry opisu sygnału mowy [Ata72, Ata74, Bas78, Bol70, Fur81, Mar79, Wol72, Zil98]:

1. Częstotliwość podstawowa tonu krtaniowego,
2. Częstotliwości formantowe,
3. Widmo mocy (długo bądź krótkoterminowe),
4. Parametry analizy przejść przez zero postaci czasowej sygnału mowy,
5. Parametry liniowego kodowania predykcyjnego LPC,
6. Parametry cepstralne.

2.4. Wybór optymalnego zestawu parametrów (wektorów cech); kryteria i metody oceny skuteczności parametrów

Mając do czynienia z dużą liczbą różnorakich parametrów zaczęto poszukiwać pewnych metod wyboru optymalnego (najbardziej dyskryminującego) zestawu parametrów opisujących sygnał oraz opracowania skutecznych metod klasyfikacji. Należy zaznaczyć, że nie istnieją ogólne metody optymalnego wyboru parametrów. Najczęściej o wyborze określonego zestawu parametrów służących do opisu elementów sygnału mowy decyduje nagromadzona wiedza badacza oraz wyniki wcześniejszych doświadczeń. Te właśnie czynniki podkreślają wyższość parametrów LPC oraz cepstralnych nad pozostałymi [Fur79, Fur81].

Zawsze możliwe jest wybranie większej liczby cech, jednak wzrost wymiaru przestrzeni cech i co zatem idzie złożoności obliczeniowej algorytmu rozpoznawania praktycznie uniemożliwia realizację pomysłu, by zaproponować dużo cech i liczyć, że program sam wybierze te dobre. Im więcej cech jest wykorzystywanych tym bardziej rośnie wymiar przestrzeni cech, co powoduje konieczność wykonywania bardziej złożonych i pamięciochłonnych obliczeń. Ponadto wraz ze wzrostem ilości cech wykładniczo rośnie postulowana długość danych uczących.

W celu dokonania redukcji rozmiaru przestrzeni cech wykorzystuje się statystyczne metody polegające na analizie wariancji składowych wektora cech. Przekładem takiego podejścia jest metoda Karthunena-Loeve'go [Cam97]. Ponieważ optymalny wektor cech powinien umożliwiać dokonanie podziału przestrzeni na rozłączne klasy, w

związku z tym problem wyboru takiego wektora jest równoważny problemowi dokonania podziału przestrzeni. W przypadku gdy obserwowane dane są liniowo separowalne, to wystarczające są transformacje liniowe, które pozwalają na podział przestrzeni za pomocą hiperpłaszczyzny.

Podstawowym kryterium wyboru zespołu parametrów z przestrzeni obserwacji jest minimalizacja średniego prawdopodobieństwa błędu rozpoznawania lub też minimalizacja średniego ryzyka błędnej decyzji. Ponieważ stosowanie tego ogólnego kryterium oceny skuteczności parametrów uwzględnia także wpływ algorytmu klasyfikacji, zatem ocena parametrów jest również, w pewnym stopniu, funkcją jakości klasyfikatora. W tej sytuacji często stosuje się kryteria autonomiczne, które pozwalają na oszacowanie zdolności dyskryminacyjnej obrazów lub ich poszczególnych elementów, to jest składowych wektora parametrów. Założenie, że parametry powinny powodować skupienie obrazów w przestrzeni parametrów należących do tej samej klasy, a rozsunięcie obszarów należących do różnych klas, jest podstawową przesłanką do konstrukcji algorytmów oceny skuteczności wybranych parametrów. W zapisie analitycznym można tę przesłankę przedstawić jako kryterium oceny zdolności parametrów do grupowania się w swoich klasach i rozdziału między klasami.

Techniką pozwalającą na wybór lepszych cech jest analiza wariancji *(ANOVA – Analysis of Variance)* wymagająca obliczenia tzw. miary Fishera F pomiędzy funkcjami gęstości prawdopodobieństwa różnych cech próbek:

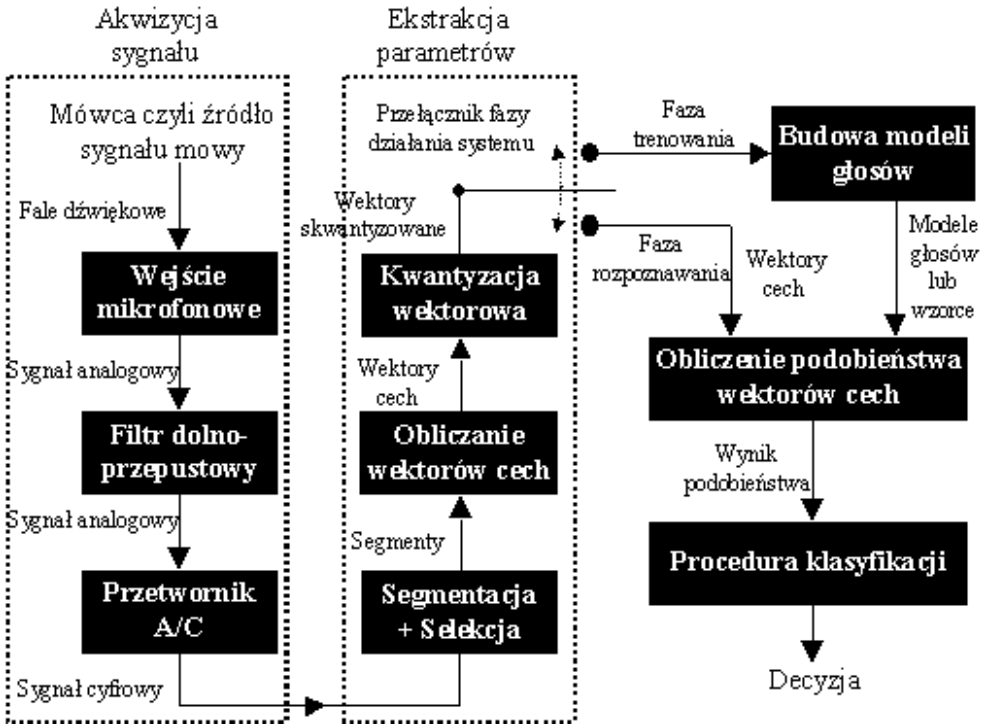
$$F = \frac{\textit{wariancja wartości średniej poszczególnych klas}}{\textit{średnia wariancji wewnątrzklasowych}} \qquad (2.1)$$

Oczywistym jest, że im większa jest wartość miary Fishera, tym łatwiej można dokonać rozdzielania przestrzeni cech na poszczególne klasy. Niestety metoda ta, aby była użyteczna, wymaga obliczenia miary F dla wielu różnych kombinacji cech. Przykładowo dwie cechy mocno skorelowane o dużych wartościach F mogą się okazać mniej efektywne niż inne dwie cechy o mniejszych indywidualnych wartościach F. Przydatność miary F jest dodatkowo osłabiona w przypadku wielomodalnych klas albo gdy klasy mają tą samą wartość średnią (wtedy nawet bardzo dobra wartość F niczego w istocie nie gwarantuje). Inną miarą dyskryminacji może być zaproponowany w [Sho98] stosunek średniej odległości obiektów wewnątrz klas do średniej odległości obiektów różnych klas. Taka miara jest podobna do odwrotności miary Fishera lecz pozbawiona najpoważniejszej wady jaką jest w mierze F brak zależności od wartości średniej. Należy

podkreślić, iż dla konkretnego klasyfikatora rozsądnym kryterium wyboru cech byłby – z całą pewnością – procent błędów systemu rozpoznawania, jednak taka weryfikacja „ex post" jest zbyt kosztowna obliczeniowo.

2.5. Akwizycja sygnału mowy

Szczegółowy schemat rozważanego tu systemu rozpoznawania mówców jest przedstawiony na rysunku 2.2.



Rys. 2.2: Schemat typowego systemu automatycznego rozpoznawania głosu.

W zależności od warunków rejestracji sygnałowi odbieranemu przez mikrofon może, w różnym stopniu, towarzyszyć szum i zakłócenia. Zazwyczaj mówi się o warunkach rejestracji pokojowych (biurowych) oraz laboratoryjnych. Inny interesujący z punktu widzenia zastosowania automatycznego rozpoznawania mówcy jest sygnał o jakości telefonicznej. Tor sygnału od źródła do wejścia przetwornika analogowo cyfrowego jest określany mianem kanału transmisyjnego sygnału. Wpływ charakterystyki kanału transmisyjnego na właściwości otrzymanego sygnału cyfrowego jest na tyle istotny, że metody rozpoznawania bywają klasyfikowane na tej podstawie. W zależności od rodzaju metody buduje się dedykowane bazy głosów dla celów ich weryfikacji.

Kluczowe znaczenie dla procesu rozpoznawania stanowi fakt, czy nagrania trenujące oraz testujące system 'przeszły' przez taki sam (czy nawet ten sam) kanał transmisyjny czy też nie. W tym pierwszym przypadku rozpoznawanie jest łatwiejsze i co najważniejsze przypadek ten jest bardziej interesujący z punktu widzenia zastosowania systemu rozpoznawania mówcy. Dlatego też powstają odrębne bazy głosów przeznaczone do weryfikacji systemów rozpoznawania w których zakłada się, że dane trenujące oraz testujące są dostarczane do systemu taką samą drogą.

Klasa aparatury rejestrującej, w tym rodzaj mikrofonu, decyduje o jakości otrzymanego sygnału cyfrowego. Oprócz części analogowej aparatury decydujący wpływ ma w szczególności zastosowany przetwornik A/C (częstotliwość próbkowania, rozdzielczość amplitudowa oraz błędy przetwarzania). Sygnał analogowy dostarczany przez mikrofon jest filtrowany przed jego przetwarzaniem na postać cyfrową. Szerokość pasma przepustowego filtru jest mniejsza od połowy częstotliwości próbkowania przetwornika analogowo cyfrowego, co wynika z teorii próbkowania Kotielnikowa-Shannona [Sza90]. Twierdzenie to odnosi się do sygnałów o ograniczonym widmie i wynika z analizy struktury widma sygnału po próbkowaniu. Dotychczasowe wyniki różnych badań wskazują na słuszność twierdzenia, że filtrowany sygnał mowy o górnej częstotliwości granicznej przyjętej jako 20kHz zachowuje niemal wszystkie przenoszone w nim informacje. Wskazuje na to między innymi fakt, że graniczna częstotliwość tonów słyszalnych przez ucho ludzkie nie przekracza (na ogół) 20kHz. W istocie możliwe jest dalsze ograniczenie górnej częstotliwości granicznej F_g nawet do 3kHz. Redukcja szerokości pasma sygnału jest bez wątpienia odkupiona pogarszaniem jego jakości oraz utratą niektórych być może istotnych szczegółów. W zależności od ograniczenia pasma sygnału, na mocy twierdzenia Kotielnikowa-Shannona można stwierdzić, że próbki sygnału pobrane z częstotliwością F_p nieco większą od podwójnej wartości F_g mogą reprezentować z pełną dokładnością oryginalny sygnał analogowy. Mimo, że wartość F_p można dobrać dowolnie – głównie w zależności od wymaganej dokładności i stopnia redukcji ilości informacji – to w praktyce jednak najczęściej stosowane są następujące wartości standardowe: 44, 22, 11, 8, 6 oraz 5,5kHz. Natomiast jeśli chodzi o rozdzielczość amplitudową, to stosowane są zazwyczaj następujące wartości standardowe: 8, 12 lub 16 bitów/próbkę.

2.6. Segmentacja sygnału

Obliczenie parametrów dokonywane jest po segmentacji sygnału. Zazwyczaj wydzielone segmenty zachodzą na siebie (*overlapping*) w celu częściowego

zapobiegania niekorzystnym skutkom dzielenia fragmentów o jednorodnej (jednolitej) strukturze. Sposób dokonania segmentacji ma istotny wpływ na pracę systemu. W zależności od długości czasowej segmentów mówi się o parametrach długo oraz krótkookresowych. W tym pierwszym przypadku systemy mogą mieć długość nawet całych wyrazów i wydobywane są z nich pewne wartości średnie wybranej wielkości fizycznej (np. moc, częstotliwość podstawowa tonu krtaniowego, widmo). W tym drugim zaś segmenty są bardzo krótkie ponieważ przedmiotem zainteresowania jest zazwyczaj rozkład czasowy danej wielkości fizycznej. Zrozumiały jest fakt, że w tym drugim przypadku trudniej jest włamać się do systemu rozpoznawania. Pierwsza koncepcja segmentacji bywa częściej wykorzystywana w systemach rozpoznawania mówcy na podstawie wypowiedzi swobodnej czyli niezależnej od treści (*text independent*), zaś druga w systemach rozpoznawania zależnego od tekstu (*text dependent*). W tym ostatnim wypadku długość segmentu jest podobna jak w systemach rozpoznawania mowy i jest rzędu kilkudziesięciu milisekund. W niektórych przypadkach parametry są wydobyte tylko z niektórych wybranych fragmentów sygnału (selekcja segmentów). Przykładowo określenie częstotliwości tonu krtaniowego dokonuje się tylko na podstawie mowy dźwięcznej, odrzucając tym samym fragmenty mowy bezdźwięcznej. Również eliminowane mogą być fragmenty ciszy występujące pomiędzy słowami a nawet sylabami.

Wydzielone segmenty sygnału są mnożone przez funkcję tzw. okna. Kształt zastosowanego okna determinuje rozdzielczość częstotliwościową oraz tłumienie bocznych prążków widma amplitudowego sygnału [Del93, Jua93]. W przypadku wydobycia parametrów LPC zastosowanie okna ogranicza błąd aproksymacji sygnału w pobliżu granic segmentu [Jua93].

2.7. Obliczanie wektorów cech

Jednym z ważniejszych zagadnień związanych z ekstrakcją parametrów jest złożoność obliczeniowa tego zadania. Sygnał mowy w postaci pierwotniej jest przebiegiem czasowym, a postać ta jest punktem wyjścia do dalszych transformacji sygnału. Zazwyczaj sygnał jest transformowany do dziedziny częstotliwości w celu obliczenia parametrów częstotliwościowych. Obliczenia parametrów cepstralnych wymaga zaś przeprowadzenia dwóch transformacji FFT przy czym otrzymane widmo po pierwszej transformacji FFT jest logarytmowane przed przeprowadzeniem tej drugiej. Obliczenie parametrów kodowania predykcyjnego wiąże się z koniecznością rozwiązania układu równań liniowych. Liczba tych równań jest równa liczbie współczynników

aproksymacji sygnału [Fur89, Mak75]. Postęp techniki cyfrowej, a w szczególności dostępność procesorów sygnałowych umożliwia wykonania tych operacji w czasie rzeczywistym. Jak dotąd zachodzi jednak potrzeba ograniczenia złożoności obliczeniowej algorytmu ekstrakcji parametrów. W miarę postępu techniki kryterium oceny metody rozpoznawania pod względem sprawności algorytmu obliczeniowego stanie się raczej drugorzędne.

2.8. Kwantyzacja wektorowa

Złożoność obliczeniowa algorytmów klasyfikacji zależy przede wszystkim od liczby wektorów należących do przestrzeni cech. Dlatego też istnieje konieczność redukcji ilości tych wektorów w przestrzeni. Stosunkowo złożoną, ale skuteczną metodą dokonania tego zadania jest kwantyzacja wektorowa VQ (*Vector Quantization*). Technika ta opiera się na fakcie, iż zakres zmienności współrzędnych wektora cech razem wziętych jest mniejszy od zakresu ich zmienności indywidualnej. Kwantyzerem wektorowym *P*-punktowym nazywa się odwzorowanie, które każdemu wektorowi cech przypisuje jeden i tylko jeden numer najbliższego wektora, ze z góry ustalonego słownika zawierającego *P* wektorów. Zakres zmian wektorów cech wyznacza pewną przestrzeń *S*, którą dla potrzeb kwantowania dzielimy na *P* podprzestrzeni S_i i w każdej z tych podzielonych podprzestrzeni określamy jeden wektor skwantowany $\boldsymbol{y}_i$. Zbiór wektorów $\boldsymbol{y}$ nazywa się również (wymiennie) książką kodową lub alfabetem bazowym. Kwantyzacja wektorowa polega na tym, że każdemu wektorowi $\boldsymbol{x}_j$ należącemu do przestrzeni, przypisuje się najbliższy numer (indeks) wektora skwantowanego $\boldsymbol{y}$, który spełnia warunek:

$$d\left(\boldsymbol{x}_j, \boldsymbol{y}\right) = \min_i d\left(\boldsymbol{x}_j, \boldsymbol{y}_i\right) \qquad (2.2)$$

Spośród zalet kwantyzacji wektorowej warto wymienić: wspomniane już wcześniej zmniejszenie rozmiaru przetwarzanej informacji oraz redukcję obliczeń koniecznych do przeprowadzenia klasyfikacji. Natomiast wadami tej metody są: konieczność przechowywania książki kodowej oraz błąd kwantyzacji wektorowej. Ponieważ książka kodowa zawiera skończoną liczbę elementów, to wybór najlepszego reprezentanta jest zawsze obarczony pewnym błędem, zwanym błędem kwantyzacji wektorowej, który maleje wraz ze wzrostem rozmiaru książki kodowej. Jednakże bez względu na rozmiar książki błąd ten zawsze występuje. Zatem wybór rozmiaru książki kodowej jest

pewnego rodzaju kompromisem pomiędzy dokładnością kwantyzacji, a rozmiarem potrzebnej pamięci i szybkością działania procesu identyfikacji wektorów.

Aby stworzyć książkę kodową należy dysponować odpowiednio dużym zbiorem ciągów uczących. Ponadto należy zdefiniować miarę podobieństwa opartą zwykle na odległości pomiędzy wektorami cech, tak aby móc dokonać klasteryzacji ciągu uczącego, określić procedury wyliczania tzw. centroid i zaproponować algorytm dopasowania rozpoznawanych wektorów. Jednym z najczęściej stosowanych algorytmów klasteryzacji jest algorytm *K*-średnich (*K-means*), który można przedstawić w następującej postaci:

- Inicjalizacja: dokonuje się arbitralnego wyboru P wektorów jako początkowej zawartości książki kodowej.
- Poszukiwanie najbliższego sąsiada: dla każdego wektora treningowego znajduje się w książce kodowej wektor najbliższy przypisując do niego dany wektor treningowy.
- Uaktualnienie centroid: oblicza się nowe wartości centroid z uwzględnieniem dodanych wektorów.
- Iteracja: powyższe dwa kroki są powtarzane, aż do momentu uzyskania zakładanej dokładności.

Mimo, iż powyższa procedura funkcjonuje poprawnie, praktyka pokazuje, że tworzenie książki kodowej na zasadzie kolejnego podwajania liczby centroid jest bardziej skuteczne. Ponadto zaletą takiego podejścia jest fakt, iż redukuje się w ten sposób ilość operacji porównania potrzebnych do klasyfikacji badanego wektora. Najbardziej naturalny algorytm wielopoziomowej kwantyzacji wektorowej można przedstawić w następującej postaci:

- Inicjalizacja: obliczenie centroidy wszystkich wektorów $\boldsymbol{Y}$.
- Podział centroid: dokonuje się rozbicia każdej grupy wektorów na dwie części, dzieląc jej elementy w zależności od tego, która z następujących dwóch centroid $\boldsymbol{Y}^+$ oraz $\boldsymbol{Y}^-$ jest bliższa danemu wektorowi:

$$\boldsymbol{Y}^+ = \boldsymbol{Y} \cdot (1+\varepsilon) \quad , \quad \boldsymbol{Y}^- = \boldsymbol{Y} \cdot (1-\varepsilon) \tag{2.3}$$

Typowe wartości ε wynoszą $0.01 \le \varepsilon \le 0.05$.

- Obliczenie nowych centroid: dla każdej rozbitej podprzestrzeni obliczane są dwie nowe centroidy.

- Iteracja: powyższe dwa kroki są powtarzane, aż do momentu uzyskania wymaganej liczby centroid.

Należy zaznaczyć, iż dla danej liczby bitów kodowania sygnału, kwantyzacja wektorowa w porównaniu z kwantyzacją skalarną znacznie poprawia jakość kodowania (zmniejsza rozproszenie wektorów względem centroid). Oznacza to również, iż dla ustalonego poziomu rozproszenia liczba bitów kodowania dla kwantyzacji wektorowej jest mniejsza od liczby bitów potrzebnej do uzyskania tego samego poziomu rozproszenia w przypadku zastosowania kwantyzacji skalarnej. Doświadczenia pokazują, iż 10 bitowy kwantyzator wektorowy posiada porównywalną dokładność jak 24 bitowy kwantyzator skalarny [Jua93]. Wynika to w zasadzie z istoty idei kwantyzacji wektorowej, opierającej się na wykluczeniu wystąpienia pewnych kombinacji kodowych (ciągów wartości skalarnych) nie mających sensu z fizycznego punktu widzenia kodowanego sygnału lub nie używanych w danym sygnale ze względów semantycznych. W przypadku kodowania mowy oznacza to (w ograniczonym stopniu) wykluczenie wszystkich sygnałów dźwiękowych nie będących „mową", bowiem mowa stanowi małą częścią wszystkich możliwych sygnałów dźwiękowych.

Budowa kontekstowych książek kodowych polega na tworzeniu wielu książek kodowych, z których każda byłaby stworzona i odpowiednio przystosowana do kodowania konkretnej klasy sygnałów. W przypadku rozpoznawania mówcy każdej książce może odpowiadać jedna osoba. Ogólnie centroidy książki kodowej są posegregowane w hierarchicznych strukturach drzewiastych, pozwalających na szybkie wyszukiwania najbliższej centroidy. Na ogół dzięki temu udaje się redukować liczbę operacji obliczania odległości z P do $log(P)$, gdzie P to liczba centriod w przestrzeni. Gdy występują korelacje pomiędzy kolejnymi wartościami, czy ogólnie fragmentami kodowanego sygnału można budować hierarchiczne książki kodowe co pozwala na użycie tzw. kwantyzacji macierzowej *matrix Quantization*. Przykład zastosowania tej kwantyzacji w dla celów automatycznego rozpoznawania głosu można znaleźć w [Che93]. W praktyce hierarchiczna kwantyzacja wektorowa polega na kolejnym wektorowym kodowaniu, otrzymanych z poprzedniego poziomu kodowania, numerów wektorów będących de facto numerami centroid książki kodowej z poprzedniego etapu kodowania. Pewną ciekawą odmianą kwantyzacji wektorowej jest tzw. kwantyzacja zależna, w której ogranicza się zbiór możliwych do kodowania centroid w zależności od wyniku kwantyzacji

poprzedzających ich wektorów. Podejście to pozwala na szybką realizację kwantyzacji oraz powiązania informacji zawartej w kolejnych wektorach.

Ciekawą i skuteczną koncepcją kwantyzacji wektorowej dla celu rozpoznawania mówcy jest nowo proponowany algorytm kwantyzacji [Yua99] nazywanej kwantyzacją binarną. Algorytm wyznaczenia centroid dla kwantyzacji binarnej przedstawia się następująco:

1. Oblicza się wartość centroidy dla całego ciągu uczącego,
2. Oblicza się różnicy między poszczególnymi wektorami ciągu uczącego i centroidy,
3. Rozbija się ciąg uczący na 2^n podciągów wektorów w zależności od znaków poszczególnych współrzędnych wektorów ich różnic od centroidy. Znak odpowiedniej współrzędnej w wektorze różnic jest kodowane jako zero bądź jeden (stąd też nazwa metody) zaś wynikowa kombinacja stanowi numer hipersześcianu do którego zalicza się wektor ciągu uczącego odpowiadający tej kombinacji.
4. Dla każdego takiego podciągu zawierającego wektory należące do jednego hipersześcianu oblicza się centroidę reprezentującą tego hipersześcianu.

W przypadku kwantyzacji binarnej istnieje także możliwość wielopoziomowego wyznaczania centroid poprzez klusteryzację wektorów należących do jednego hipersześcianu zgodnie z zaproponowanym powyżej algorytmem. Alfabetem bazowym kwantyzacji binarnej są właśnie numery kolejnych hipersześcianów, na które dzieli się przestrzeń wektorów.

Optymalny wybór reprezentatywnej centroidy $\boldsymbol{y}^*$ dla grupy Z wektorów dokonywany jest w ogólnym przypadku na podstawie następującego równania:

$$\boldsymbol{y}^* = \arg\min_{\boldsymbol{y}} \frac{1}{Z} \sum_{i=1}^{Z} d(\boldsymbol{x}_i, \boldsymbol{y}) \qquad (2.4)$$

gdzie $\boldsymbol{x}_i$ oznacza kolejny wektor z Z-elementowego otoczenia centroidy.

Z powyższego równania wynika, że obliczenie centroidy bezpośrednio zależy od wybranej metryki. Przykładowo dla przestrzeni z metryką Euklidesa[3] centroidami są po prostu wartości średnie:

$$\boldsymbol{y}^* = \frac{1}{Z}\sum_{i=1}^{Z}\boldsymbol{x}_i \qquad (2.5)$$

W przypadku metryki ulicznej centroidą jest wektor którego kolejne współrzędne są obliczane jako mediany wartości dla poszczególnych współrzędnych wektorów grupy. Natomiast w ogólnym przypadku mogą to być jakieś inne miary tendencji centralnej w odpowiednich podzbiorach danych.

2.9. Określenie stopnia podobieństwa wektorów cech

W metodach minimalno-odległościowych klasyfikacja oparta jest na odległości (metryce) wektorów badanych i wzorcowych określanych wcześnie w procesie uczenia się [Tad92]. Najczęściej stosowana jest metryka Euklidesowa, definiowana wzorem:

$$\rho(\boldsymbol{x},\boldsymbol{y}) = \sqrt{\sum_{i=1}^{N}(x_i - y_i)^2} \qquad (2.6)$$

Niestety metryka ta ma pewne wady. Pierwsza z nich wynika z ewentualnych różnic przedziałów wartości poszczególnych składowych wektora $\boldsymbol{x}$. Jeśli jedna składowa jest duża, a wszystkie pozostałe są niewielkie, to powyższy wzór uzależnia odległość ρ wyłącznie od wartości tej dużej składowej. Sytuację tą można poprawić poprzez zastosowanie uogólnionej metryki Euklidesowej o postaci:

$$\rho(\boldsymbol{x},\boldsymbol{y}) = \sqrt{\sum_{i=1}^{N}\lambda_i(x_i - y_i)^2} \qquad (2.7)$$

Mnożniki normalizujące λ_i mogą być w różny sposób związane z wartościami składowych wektora $\boldsymbol{x}$, na przykład można je wyliczyć z następujących wzorów:

$$\lambda_i = \frac{1}{\max\limits_{x\in U} x_i - \min\limits_{x\in U} x_i} \qquad (2.8)$$

[3] Ogólnie z metryką Mahalanobisa.

$$\lambda_i = \sqrt{\frac{L-1}{\sum_{x \in U}\left(x_i - \frac{1}{L}\sum_{x \in U} x_i\right)^2}} \qquad (2.9)$$

gdzie L jest długością ciągu uczącego U.

Kolejny zarzut związany z metryką Euklidesową polega na małej efektywności obliczeniowej. Prostsza pod tym względem i dająca dobre rezultaty w przypadku zagadnień praktycznych jest metryka Hamminga, zwana też metryką uliczną:

$$\rho(\boldsymbol{x},\boldsymbol{y}) = \sum_{i=1}^{N} |x_i - y_i| \qquad (2.10)$$

lub uogólniona metryka uliczna:

$$\rho(\boldsymbol{x},\boldsymbol{y}) = \sum_{i=1}^{N} \lambda_i |x_i - y_i| \qquad (2.11)$$

Prostsza jest również metryka Czebyszewa:

$$\rho(\boldsymbol{x},\boldsymbol{y}) = \max_{1 \le i \le N} |x_i - y_i| \qquad (2.12)$$

Jednak aby móc z niej sensownie skorzystać należy wcześniej dokonać normalizacji cech. Metryki Euklidesa, Hamminga i Czebyszewa są szczególnymi przypadkami metryki Minkowskiego:

$$\rho(\boldsymbol{x},\boldsymbol{y}) = \left[\sum_{i=1}^{N} |x_i - y_i|^t\right]^{\frac{1}{t}} \qquad (2.13)$$

przy czym dla $t=2$ metryka ta staje się metryką Euklidesa, dla $t=1$ jest to metryka Hamminga, natomiast gdy $t \to \infty$ to metryka ta dąży do metryki Czebyszewa. Ciekawe własności samonormalizujące ma metryka Camberra podana wzorem:

$$\rho(\boldsymbol{x},\boldsymbol{y}) = \sum_{i=1}^{N} \left|\frac{x_i - y_i}{x_i + y_i}\right| \qquad (2.14)$$

Wspólną wadą wszystkich powyższych metryk jest fakt oparcia ich koncepcji na założeniu o ortogonalności bazy przestrzeni cech. Założenie to w ogólnym przypadku jest niczym nie uzasadnione [Tad92], ponieważ wybór cech x_i jest zwykle podyktowany możliwościami pomiarowymi i nie ma żadnych przesłanek, że tak uzyskane wartości

wyznaczają bazę ortogonalną. Zatem celowe jest stosowanie także innych metryk, na przykład metryki Mahalanobisa:

$$\rho(\boldsymbol{x},\boldsymbol{y}) = \sqrt{(\boldsymbol{x}-\boldsymbol{y})^T \boldsymbol{C}^{-1}(\boldsymbol{x}-\boldsymbol{y})} \qquad (2.15)$$

gdzie $\boldsymbol{C}$ jest macierzą kowariancji cech. Zastosowanie macierzy $\boldsymbol{C}$ powoduje „wyprostowanie" nieortogonalnego układu współrzędnych na drodze apriorycznej transformacji liniowej. Metryka Mahalanobisa jest przykładem metryki kwadratowej o ogólnym wzorze:

$$\rho(\boldsymbol{x},\boldsymbol{y}) = \sqrt{(\boldsymbol{x}-\boldsymbol{y})^T \boldsymbol{T}(\boldsymbol{x}-\boldsymbol{y})} \qquad (2.16)$$

gdzie $\boldsymbol{T}$ jest macierzą transformacji. Przy odpowiednim doborze macierzy T metryka ta może uwzględniać różne własności przestrzeni cech. Na przykład dla T diagonalnej otrzymuje się uogólnioną metrykę Euklidesa, natomiast przyjęcie T zawierającej elementy niezerowe wyłącznie poza główną przekątną prowadzi do metryki Bhattacharyya. Metryka ta jest stosunkowo często wskazywana jako optymalna w praktycznych zadaniach rozpoznawania i grupowania cech.

2.10. Minimalno-odległościowe metody klasyfikacji

Rozmaite metody klasyfikacji mogą być zastosowane do różnych zadań rozpoznawania, takich jak rozpoznawanie mowy, głosów lub obrazów, identyfikacja stanu technicznego maszyn, diagnostyka chorób oraz poszukiwania geologiczne. Choć dla konkretnego zadania natura rozpoznawanych obiektów może być dowolna, to proces klasyfikacji jest podobny. Klasyfikacja ma być prowadzona w sytuacji braku apriorycznej informacji na temat reguł przynależności obiektów do poszczególnych klas, a jedyna możliwa do wykorzystania informacja jest zawarta w tzw. ciągu uczącym, w skład którego wchodzą obiekty o znanej przynależności. Ponieważ wstępny etap procesu rozpoznawania obejmuje zwykle pomiar cech opisujących rozpoznawane obiekty, zaś dalsze postępowania polega wyłącznie na analizowaniu tych cech, bez powracania do oryginalnego obiektu, to możliwe jest zunifikowane podejście do metod rozpoznawania. Cechy rozpoznawanych obiektów są kodowane za pomocą tzw. wektorów cech. Siła dyskryminacyjna rozpatrywanej metody klasyfikacji jest zdeterminowana zarówno wyborem tych cech jak i kryterium rozdzielającym ich przestrzeń na obszary odpowiadające poszczególnym klasom.

Zarówno wybór parametrów (cech obiektów) jak i kryterium rozdzielającego ich przestrzeń na klasy odpowiadające poszczególnym kategoriom, determinuje siłę dyskryminacji rozważanej metody rozpoznawania. Jednakże im lepszy jest wybór parametrów, tym łatwiej dobiera się kryterium rozdzielenia klas odpowiadających poszczególnym kategoriom w przestrzeni parametrów. Z powyższych rozważań wynika z jednej strony konieczność koncentrowania wysiłków na badaniach nad ekstrakcją dobrych parametrów i z drugiej strony możliwość korzystania z standardowych metod klasyfikacyjnych. Wśród tych metod wyróżnia się metody: minimalno-odległościowe, aproksymacyjne, wzorców, z wykorzystaniem sieci neuronowych oraz probabilistyczne [Tad92]. Wspomniane metody można adaptować i modyfikować tak, aby najlepiej pasowały do wybranych parametrów. Wydaje się pewny fakt, że większość tych metod klasyfikacji jest w stanie produkować zadawalające wyniki pod warunkiem, że wybrane parametry są dostatecznie dobre.

Nie ulega wątpliwości, iż umiejscowienie granic danej klasy w przestrzeni cech zależy bezpośrednio od sposobu jej reprezentacji w tej przestrzeni. Innymi słowy podział przestrzeni sygnałów wejściowych, zależy od przyjętej metody klasyfikacji określającej tzw. funkcje przynależności, rozstrzygające stopień podobieństwa parametrów badanego obiektu (jeden lub więcej wektorów cech) i określonego wzorca (np. modelu głosu konkretnego mówcy). We wszystkich metodach minimalno-odległościowych funkcję przynależności wiąże się z pojęciem odległości punktów w przestrzeni cech, co z jednej strony pozwala łatwo wyobrazić sobie istotę proponowanych metod, a z drugiej strony prowadzi do stosunkowo dobrej jakości rozpoznawania.

Dla celów dalszego opisu metod klasyfikacji wprowadzono następujące oznaczenia:

- liczba klas: L_k, zbiór klas: $\boldsymbol{K}=\{ k^i \}$, $i=1,2,..., L_k$
- liczba wektorów: L_h, zbiór wektorów ciągu uczącego: $\boldsymbol{H}= \{ \boldsymbol{h}^j \}; j=1,2,....,L_h$
- $\boldsymbol{H}^i$: podzbiory ciągu $\boldsymbol{H}$ reprezentujące poszczególne klasy, $i=1,2,..., L_k$,
- $C^i(\boldsymbol{h}^j)$ funkcja przynależności obiektu $\boldsymbol{h}^j$ do klasy i.

Niezależnie od zastosowanej metody klasyfikacji umiejscowienie granic klas jest też ściśle związane z przyjętą miarą odległości (metryką) punktów w przestrzeni cech. Obszar danej klasy jest określany bowiem na bazie jednego bądź kilku punktów przestrzeni (wektorów cech odpowiadających rzeczywistym obiektom) oraz przyjętej metryki. Zazwyczaj obszar klasy jest sumą kilku obszarów, z których każdy jest otoczeniem pojedynczego punktu, przy czym sumowanie dotyczy wszystkich punktów

reprezentujących daną klasę. Przez otoczenie danego punktu rozumie się zbiór wszystkich punktów przestrzeni, których odległości od niego jest mniejsza od pewnej zadanej wartości granicznej określającej rozmiar tego otoczenia. Widać zatem, że kształt otoczenia jest bezpośrednio zależny od przyjętej metryki.

2.10.1. Metoda *k* najbliższych sąsiadów *kNN*

Metoda ta w podstawowym wariancie (dla $k=1$), nazwanym algorytmem najbliższego sąsiada *NN* (*Nearest Neighbour*) zakłada, iż badany obiekt należy do tej klasy, do której należy najbliższy mu (w sensie metryki) obiekt z ciągu uczącego. W związku z tym najpoważniejszą wadą metody *NN* jest duża wrażliwość na ewentualne błędy ciągu uczącego. Przykładowo, jeżeli w ciągu uczącym występuje chociażby jeden element nie reprezentatywny (lub błędnie zakwalifikowany), to wówczas całe jego otoczenie będzie błędnie klasyfikowane. Aby efekt ten ograniczyć stosuje się pewną odmianę rozważanej metody polegającą na uwzględnieniu nie jednego lecz k sąsiadów klasyfikowanego obiektu. W metodzie tej arbitralnie wybierany parametr k w dużej mierze determinuje własności metody.

Dla metody *kNN* klasyfikacja nieznanego obiektu $\boldsymbol{h}$ przebiega w sposób następujący. Po pierwsze obliczane są odległości tego obiektu do wszystkich obiektów ciągu uczącego:

$$\boldsymbol{O} = \left\{ O^l \right\} = \left\{ \rho\left(\boldsymbol{h}, \boldsymbol{h}^l\right) \right\} \quad ; \quad l = 1,2,...,L_h \tag{2.17}$$

Następnie zapamiętuje się k najmniejszych odległości a także odpowiadające im numery obiektów otrzymując w ten sposób dwa zbiory $\boldsymbol{O}^s$, $\boldsymbol{N}^s$:

$$\boldsymbol{O}^s = \left\{ O^s = \rho(\boldsymbol{h}, \boldsymbol{h}^{N^s}) \right\} \quad ; \quad s = 1,2,...,k \tag{2.18}$$

przy czym:

$$k << L_h$$
$$\forall s,\ O^s \le O^{s+1}$$
$$\forall O^n \in \boldsymbol{O} - \boldsymbol{O}^s,\ O^n \ge O^k \in \boldsymbol{O}^s$$

$$\boldsymbol{N}^s = \left\{ N^s : \rho(\boldsymbol{h}, \boldsymbol{h}^{N^s}) \in \boldsymbol{O}^s \right\} \quad ; \quad s = 1,2,...,k \tag{2.19}$$

przy czym: $N^s \in \langle 1, L_h \rangle$

Biorąc pod uwagę, że poszczególnym numerom obiektów ciągu uczącego odpowiadają konkretne klasy w przestrzeni cech, rozbija się otrzymany zbiór $\boldsymbol{N}^s$ na podzbiory (ewentualnie puste) związane z poszczególnymi klasami $\boldsymbol{N}^{s,i}$. Funkcje przynależności dla metody *kNN* można teraz wyznaczyć na podstawie liczebności podzbiorów $\boldsymbol{N}^{s,i}$:

$$C^i(\boldsymbol{h}) = \#\boldsymbol{N}^{s,i} \quad ; \quad i = 1,2,...,L_k \tag{2.20}$$

2.10.2. Metoda potencjału najbliższych sąsiadów *pkNN*

Metoda ta zaproponowana przez autora pracy jest hybrydowym połączeniem metody *kNN* oraz metody funkcji potencjalnych. Koncepcja tej ostatniej, mało znanej dziś metody wprowadzonej w latach siedemdziesiątych przez Ajzermana, Brawermana i Rozenoera [Ajz76], polega na tym, aby funkcje przynależności budować jako superpozycje kilku funkcji składowych przypominających swoim kształtem rozkład potencjału elektrycznego wokół ładunku punktowego[4]. Funkcje składowe (potencjalne) są silnie malejącymi funkcjami odległości od punktu generującego potencjał (obiektu należącego do ciągu uczącego $\boldsymbol{H}$). Stosowana jest zazwyczaj następująca postać funkcji składowej :

$$F(\boldsymbol{h},\boldsymbol{h}^j) = \frac{1}{1+\mu.\rho^2(\boldsymbol{h},\boldsymbol{h}^j)} \tag{2.21}$$

przy czym: $\boldsymbol{h}^j \in \boldsymbol{H}$, $\mu > 0$, ρ : metryka.

Wtedy funkcje przynależności mogą mieć następującą postać:

$$C^i(\boldsymbol{h}) = \sum_{j=1}^{j=k} F(\boldsymbol{h},\boldsymbol{h}^j) \tag{2.22}$$

przy czym: $\boldsymbol{h}^j \in \boldsymbol{H}^i$.

Klasyfikację metodą *pkNN* przeprowadza się w sposób następujący. Po pierwsze, dla klasyfikowanego obiektu i dla każdej z klas z osobna, znajduje się k najbliższych sąsiadów tego obiektu należących do zbioru $\boldsymbol{H}^i$ reprezentującego klasę i. Zbiór tych sąsiadów w zależności od klasy jest oznaczony symbolem $\boldsymbol{H}^{s,i}$. Należy zwrócić uwagę, że w przypadku tej metody parametr k reprezentuje liczbę sąsiadów pochodzących z każdej pojedynczej klasy, a nie łączną liczbą sąsiadów ze wszystkich klas[5].

[4] Stąd wzięto nazwę metody.
[5] Jak to ma miejsce w przypadku zwykłej metody *kNN*.

Zatem liczebność każdego pojedynczego zbioru $\boldsymbol{H}^{s,i}$ wynosi k. Zakładamy przy tym, że każda klasa jest reprezentowana w ciągu uczącym przez liczbę wzorców większa od ustalonej liczby k. Dla każdej klasy elementy odpowiadającego jej zbioru $\boldsymbol{H}^{s,i}$ uznaje się za obiekty generujące potencjał. Zakłada się, że funkcje potencjalne mają następującą postać:

$$F(\boldsymbol{h},\boldsymbol{h}^{s,i}) = \frac{1}{1+\mu.\rho^2(\boldsymbol{h},\boldsymbol{h}^{s,i})} \tag{2.23}$$

przy czym: $\boldsymbol{h}^{s,i} \in \boldsymbol{H}$, $\mu > 0$, ρ : metryka.

Klasyfikowany obiekt uznaje się za należący do tej klasy, która generuje największy potencjał w odpowiadającym mu punkcie przestrzeni. Zatem funkcje przynależności w tym przypadku także mają postać następującą:

$$C^i(\boldsymbol{h}) = \sum_{s=1}^{k} F(\boldsymbol{h},\boldsymbol{h}^{s,i}) \tag{2.24}$$

2.10.3. Metoda wzorców[6]

Metoda ta stanowi pewne uproszczenie metody *kNN*, polegające na reprezentacji wszystkich obiektów każdego ciągu $\boldsymbol{H}^i$ za pomocą pewnej liczby wzorców na ogół za pomocą jednego obiektu będącego pewną średnią (centroidą) dla danej klasy. Jednakże w zależności od rozproszenia obiektów ciągu $\boldsymbol{H}^i$ można podjąć próbę reprezentacji klasy za pomocą więcej niż jednego obiektu, gdy rozkłady gęstości w przestrzeni mają charakter wielomodalny. W przypadku tej metody o przynależności obiektu do konkretnej klasy i decyduje jego odległość od ustalonego obiektu będącego elementem wzorca $\boldsymbol{W}^i$ klasy i. Kluczowym problemem staje się przy tym sposób definicji wzorca. W najprostszym przypadku wzorzec klasy i można utożsamiać z pewnym podzbiorem zbioru $\boldsymbol{H}^i$ ciągu uczącego. Aby określić funkcje przynależności dla tej metody należy jeszcze przyjąć maksymalną odległość ε określającą granice otoczenia poszczególnych elementów wzorca. Obszar wzorca $\boldsymbol{OW}^i$ jest zazwyczaj zdefiniowany jako suma wszystkich obszarów otaczających poszczególne jego elementy. Funkcję przynależności można formalnie zapisać w następującej postaci:

[6] Metoda wzorców jest uznawana za odmianę metod minimalno-odległościowych [Tad92].

$$C^i(\boldsymbol{h}) = \begin{cases} 1 & \text{gdy} \quad \boldsymbol{h} \in \boldsymbol{OW}^i \\ 0 & \text{w przeciwnym przypadku} \end{cases} \tag{2.25}$$

przy czym obszar wzorca $\boldsymbol{OW}^i$ jest zdefiniowany w następujący sposób:

$$\boldsymbol{OW}^i = \left\{ \boldsymbol{h} : \exists\, \boldsymbol{h}^j \in \boldsymbol{W}^i, \rho\left(\boldsymbol{h}, \boldsymbol{h}^j\right) < \varepsilon \right\} \tag{2.26}$$

W szczególności wartość stałej ε może być różna dla różnych wzorców klas jak i różnych obiektów tego samego wzorca, co może być wykorzystane w celu polepszenia jakości klasyfikacji. Przyjęcie odpowiednio dużej wartości ε dla wybranego obiektu $\boldsymbol{h}^j \in \boldsymbol{W}^i$ może doprowadzić do 'znikania' reszty obiektów wzorca $\boldsymbol{W}^i$ wraz z ich otoczeniem w obszarze tego obiektu. Wtedy cała klasa może być reprezentowana przez ten jeden obiekt zwany zwykle obiektem modalnym lub krótko modą.

2.11. Selekcja wektorów cech

W każdym ciągu uczącym istnieje pewne prawdopodobieństwo wystąpienia elementów niereprezentatywnych. Element danej klasy jest uznawany za niereprezentatywny lub niepowtarzalny jeśli odbiega od innych w sposób wyraźny. Biorąc pod uwagę, że zaproponowana metoda rozpoznawania mówcy bazująca na dyskryminacji amplitudowej z założenia jest metodą niezależną zarówno od tekstu jak i języka nagranych wypowiedzi, należy spodziewać się wystąpienia wśród próbek sygnału fragmentów o charakterze niepowtarzalnym. Jest to całkowicie prawdopodobne chociażby dlatego, że reprezentowanie całego słownika wyrazów danego języka za pomocą fragmentu tekstu o długości rzędu kilkuset słów (ok. 2 minuty) jest teoretycznie niemożliwe. Z tego względu, wiarygodne stwierdzenie niepowtarzalności danego fragmentu wiąże się z koniecznością badania odpowiednio długiego nagranego tekstu. Dodatkowym powodem wystąpienia niepowtarzalnych fragmentów jest zachowanie pełnej swobody w nagraniu wypowiedzi jak i rozdzielanie sygnału w sposób ślepy. Wówczas mogą występować fragmenty nie reprezentatywne, na przykład zawierające dłuższy odcinek ciszy[7].

Eliminację obiektów niepowtarzalnych można przeprowadzić osobno dla poszczególnych zbiorów $\boldsymbol{H}^i$ ciągu uczącego. Algorytm eliminacji, zaproponowany przez autora pracy, przedstawia się następująco:

[7] Jak początki nagrania lub fragmenty obejmujący przerwę w czytaniu.

1. Dla każdego obiektu klasy *i* oblicza się jego odległość do wszystkich innych obiektów tej samej klasy:

$$\forall \boldsymbol{h}^x, \boldsymbol{h}^y \in \boldsymbol{H}^i \quad \text{oblicza się} \quad O^{x,y} = \rho(\boldsymbol{h}^x, \boldsymbol{h}^y) \ \text{ przy czym x} \neq \text{y} \qquad (2.27)$$

2. Dla każdego obiektu zapamiętuje się odległość do jego najbliższego sąsiada tworząc ciąg $\boldsymbol{O}^s$:

$$\forall \boldsymbol{h}^x \in \boldsymbol{H}^i \ \text{wybiera się} \ \boldsymbol{h}^s : \forall \boldsymbol{h}^y \neq \boldsymbol{h}^s, O^{x,s} \leq O^{x,y}, \qquad \boldsymbol{O}^s = \left\{ O^{x,s} : \boldsymbol{h}^x \in \boldsymbol{H}^i \right\} \qquad (2.28)$$

3. Otrzymany zbiór najbliższych odległości $\boldsymbol{O}^s$ jest sortowany według rosnących wartości otrzymując ciąg $\boldsymbol{O}_u$. Podczas procesu sortowania zapamiętywane są także numery obiektów odpowiadających posortowanemu ciągowi $\boldsymbol{O}_u$. Sortowany ciąg numerów oznaczony jest symbolem $\boldsymbol{N}^s$.
4. W zależności od procentu eliminowanych obiektów oznaczonego jako o_p odrzuca się te, których numery występują na końcu ciągu numerów $\boldsymbol{N}^s$ czyli tych, których odległości do najbliższego sąsiada są większe od pozostałych (tzn. punkty odosobnienia).

2.12. Sposoby rozdzielania przestrzeni cech, błędy rozpoznawania

W niniejszym rozdziale przedyskutowane zostaną w zwartym skrócie rodzaje błędów rozpoznawania oraz ich zależność od założenia dotyczącego zbioru rozpoznawanych klas. Głębszą analizę tego zagadnienia wraz z pełną metodologią i założeniami matematycznymi można znaleźć w pracy profesora Basztury [Bas87] poświęconej automatycznemu rozpoznawaniu głosów w zbiorach otwartych[8].

Błędy rozpoznawania mogą być różne w zależności od tego czy klasa badanego obiektu została uwzględniona w przestrzeni cech[9] (*target pattern*) czy też nie (*non-target pattern*). W pierwszym przypadku można mówić o następujących przypadkach błędów:

- **błędne odrzucanie** (*false rejection*) przez własną klasę (błąd klasyfikacji pierwszego rodzaju) polegające na nie zaliczeniu obiektu do jego własnej klasy.

[8] Można to również znaleźć w książce profesora Basztury [Bas92, Bas93].
[9] Zbudowano dla niej model w procesie uczenia.

- błąd **niejednoznaczności** wynikający z niemożności jednoznacznego wskazania klasy, do której należy badany obiekt (obiekt wydaje się być przynależny do więcej niż jednej klasy).
- błąd **nieokreślenia** występujący w przypadku nie stwierdzenia przynależności obiektu do żadnej z klas.

W przypadku gdy, badany obiekt nie należy do żadnej z klas uwzględnionych w przestrzeni[10], można mówić o **błędnej akceptacji** (*false acceptance*) tego obiektu (błąd klasyfikacji drugiego rodzaju) oraz o błędzie **niejednoznaczności.** Nie występuje tutaj błąd **nieokreślenia** gdyż nie stwierdzenie przynależności obiektu do żadnej z klas jest w tym wypadku decyzją prawidłową.

W kontekście dyskusji możliwych błędów w systemie rozpoznającym istotny dylemat stanowi sposób rozdzielania przestrzeni cech, można bowiem dokonać tego dwojako, co istotnie wpływa na proporcje prawdopodobieństwa wystąpienia poszczególnych błędów. Pierwszy 'ścisły sposób' rozdzielania jest związany z tzw. otwartym zbiorem klas (*open set*) i polega na takim ograniczeniu obszarów klas należących do przestrzeni, że w tej ostatniej świadomie pozostawia się puste obszary, określane zazwyczaj mianem 'dziur'. Jest to jednoznaczne z postulowaniem, iż wynikiem rozpoznawania może być decyzja typu 'obiekt zupełnie obcy'[11], co oznacza możliwość wystąpienia błędu nieokreślenia jeśli w rzeczywistości klasa badanego obiektu została uwzględniona w przestrzeni cech. Drugi 'luźny sposób' rozdzielania przestrzeni jest zazwyczaj stosowany w przypadku domkniętego zbioru klas (*closed set*) tzn. w sytuacji gdy liczba rozpatrywanych klas jest ograniczona a wynik klasyfikacji musi być pozytywny przynajmniej dla jednej z nich. Typowym przykładem takiego zbioru jest np. zbiór rozpoznawanych jednostek mowy w systemach automatycznego rozpoznawania mowy przy założeniu prawidłowego korzystania z systemu (wypowiedzi nie należące do zbioru są wykluczone z zasady). Innym przykładem może być zbiór głosów w systemie rozpoznawania mówcy do którego dostęp ma tylko ograniczona liczba mówców i brak możliwości pojawienia się osoby z zewnątrz. W tych wypadkach można zakładać, że w obszarach znajdujących się pomiędzy rozpatrywanymi klasami nie znajdują się żadne inne klasy niż te uwzględnione w procesie uczenia. Wobec tego można dokonać

[10] Nie zbudowano modelu dla jego klasy w procesie uczenia.
[11] Obiekt nie należy do żadnej z klas uwzględnionych w przestrzeni.

odpowiedniego rozszerzenia granic obszarów rozpatrywanych klas tak aby ‘pochłonęły’ one – częściowo lub całkowicie – istniejące puste obszary. Należy podkreślić, że w systemach ze zbiorem otwartym stosowanie tego sposobu rozdzielania wiąże się z niebezpieczeństwem wynikającym z założenia o wyłączności istnienia skończonej liczby klas. Założenie takie choć eliminuje możliwość wystąpienia błędu nieokreślenia, maksymalizuje jednak prawdopodobieństwo występowania błędnej akceptacji.

W systemach identyfikacji ze zbiorem domkniętym każdy klasyfikowany obiekt musi mieć z definicji swoją klasę w przestrzeni co oznacza aprioryczny brak możliwości wstąpienia błędu nieokreślenia. Zatem rozszerzenie granic klas jest w tej klasie zadań w pełni uzasadnione, gdyż może znacznie poprawiać jakość rozpoznawania. Zalecane jest również, aby podział przestrzeni między klasami dokonać biorąc pod uwagę miary skupienia obiektów poszczególnych klas. Przykładowo kosztem ograniczenia rozmiaru jednych klas o stosunkowo skupionych obiektach można rozszerzyć obszary innych o bardziej rozproszonych obiektach przyczyniając się tym samym do zmniejszenia błędów rozpoznawania dla tych drugich. Dla ‘luźnego’ sposobu rozdzielania przestrzeni zapobieganie błędowi akceptacji może być osiągnięte poprzez ograniczone rozszerzanie obszarów klas tak aby pochłonęły one tylko częściowo obszary dziur. Innym sposobem ograniczenia tego błędu byłoby zastosowanie tzw. rozpoznawania grupowego polegającego na jednoczesnym rozpoznawaniu kilku obiektów tego samego pochodzenia (czyli należących do tej samej klasy). Wówczas gdy klasa do której w rzeczywistości należy obiekt rozpoznawanej grupy nie jest uwzględniona w przestrzeni cech, istnieje możliwość, iż poszczególne obiekty grupy zostają zaklasyfikowane do różnych klas, co pozwala na odmowę rozpoznawania gdy te cząstkowe rozpoznawania są niezgodne. Jednoczesne rozpoznawanie kilku obiektów (próbek mowy tego samego mówcy) pozwala także na przyjęcie skutecznego kryterium rozpoznawania dla domkniętego zbioru klas polegającego np. na tzw. głosowaniu czyli wskazaniu klasy, do której zaliczono najwięcej obiektów z grupy. Idea rozpoznawania grupowego może być jeszcze połączona z koncepcją ograniczenia rozmiaru klasy poprzez przyjęcie pewnego progu określającego czy odległość obiektu grupy od danego modelu klasy daje mu prawo do głosowania czy też nie. Zagadnienie rozpoznawania grupowego opisano dokładnie w punkcie 2.14. W istocie, granice rozszerzenia należy określić racjonalnie przede wszystkim na podstawie liczby rozpoznawanych klas. Oczywistym jest, że w miarę wzrostu liczby klas można pozwolić sobie na szerszą eskalację ich obszarów, jednak staje się to też coraz trudniejsze. W szczególności w przypadku zastosowania

rozpoznawania grupowego, korzystne jest zwiększanie liczby klas uwzględnionych w przestrzeni, bowiem na ogół zwiększa to możliwość 'rozproszenia' obiektów obcej grupy pomiędzy klasami zmniejszając tym samym prawdopodobieństwo wystąpienia błędnej akceptacji. Osobny problem obu wspomnianych sposobów rozdzielania przestrzeni stanowi zachodzenie na siebie więcej niż jednej klasy, co może stanowić źródło błędu niejednoznaczności. Naturalna jest komplikacja tego problemu w miarę zwiększania liczby rozpatrywanych klas.

2.13. Porównanie minimalno-odległościowych metod klasyfikacji

Najistotniejszą zaletą korzystnie odróżniającą metod *kNN* oraz *pkNN* od metody wzorców[12] jest większa swoboda reprezentacji w tych metodach tzw. klas wielomodalnych[13], przy czym odnosi się to do przypadku gdy wzorce są jedno-obiektowe. Niemniej jednak metoda *k*NN ma poważne wady. Po pierwsze wymaga ona obliczenia odległości obiektu do wszystkich elementów ciągu uczącego, a następnie wybierania *k* najmniejszych z nich. Oznacza to, że złożoność obliczeniowa metody *k*NN jest na ogół duża i szybko rośnie ze wzrostem liczby klas oraz liczby obiektów reprezentujących pojedynczą klasę. W związku z tym metoda ta jest bardziej wymagająca zarówno z punktu widzenia mocy obliczeniowej realizującego rozpoznawanie komputera, jak i pojemności jego pamięci. Po drugie mimo, iż w najbliższym sąsiedztwie danego obiektu leżą obiekty nawet z jednej i tej samej klasy nie oznacza to, że badany obiekt należy do tej klasy, bowiem nie wiadomo na ile jest on od nich odległy[14]. Zatem metoda *kNN* biorąca pod uwagę tylko liczebność sąsiednich obiektów bez względu na ich odległość (tzw. klasyczne kryterium klasyfikacji) rozdziela przestrzeń cech w sposób całkiem "luźny". Jak już wspomniano w systemach z otwartym zbiorem klas konsekwencją tego będzie całkowita eliminacja błędu nieokreślenia jednak kosztem maksymalizacji prawdopodobieństwa wystąpienia błędnej akceptacji. Błędna akceptacja w rozważanym przypadku wiąże się z sytuacją, w której klasyfikowany obiekt leży na tyle daleko od wykrytych najbliższych sąsiadów, że błędne byłoby jego zaliczanie do którejkolwiek klasy. Ograniczenie tego błędu można osiągnąć wprowadzając pewien kres górny (próg) dopuszczalnych odległości do rozpatrywanych sąsiadów (ograniczenie tzw.

[12] Ogólnie od innych metod klasyfikacji.
[13] Posiadających więcej niż jedno skupisko obiektów w przestrzeni cech.
[14] Obiekty mogą być najbliższe, ale nie wystarczająco bliskie.

promienia sąsiedztwa). Dzieje się to jednak kosztem wzrostu prawdopodobieństwa wystąpienia błędu nieokreślenia, który może mieć miejsce w przypadku decyzji o przynależności typu „obiekt zupełnie obcy". Kolejne błędy i niepewności towarzyszące klasycznemu kryterium klasyfikacji *kNN* mogą dać o sobie znać w przypadku, gdy w sąsiedztwie danego obiektu znajduje się kilka bardzo bliskich obiektów z jego klasy ale i znacznie więcej dalszych obiektów z innej klasy. Polepszenie jakości klasyfikacji można wtedy osiągnąć poprzez optymalny dobór wartości parametru *k*. W tym celu w systemach weryfikacji wartość ta powinna być różnie dobierana dla różnych klas w zależności od ich struktury.

Nie ulega wątpliwości, że w systemach otwartych należy przyjąć inne, zmodyfikowane kryterium, bogatsze niż liczebność najbliższych sąsiadów. W szczególności można zaproponować różne inne hybrydowe kryteria opierające się zarówno na liczebności sąsiadów jak i ich odległości od klasyfikowanego obiektu. Przykładem takiego kryterium jest np. opracowana przez autora metoda potencjału najbliższych sąsiadów *pkNN*. Innym ciekawym rozwiązaniem byłoby zastosowanie sieci neuronowej w celu ustalania miary ważności kolejnych sąsiadów w zależności od ich odległości. Zagadnienie to nie będzie jednak podejmowane w tej pracy.

Należy zaznaczyć, że podobnie jak dla metody *kNN* złożoność obliczeniowa oraz rozmiar wymaganej pamięci metody *pkNN* są stosunkowo duże. Zakładając, że każda klasa jest reprezentowana przez L_{hk} obiektów, w przypadku metody *kNN* liczba operacji porównywania wymaganych przy wyznaczeniu *k* najbliższych sąsiadów – np. metodą prostego wybierania – wynosi:

$$LO_{kNN} = \frac{(L_k L_{hk})!}{(L_k L_{hk} - k_1)!} \quad ; \quad k_1 < L_k L_{hk} = L_h \tag{2.29}$$

Zakładając, że dla metody *pkNN* klasy są reprezentowane przez taką samą liczbę obiektów, to liczba operacji porównywania dla tej metody jest dana wzorem:

$$LO_{pkNN} = L_k \frac{L_{hk}!}{(L_{hk} - k_2)!} \quad ; \quad k_2 < L_{hk} \tag{2.30}$$

Zakładając ponadto, że:

$$k_1 = L_k . k_2 \tag{2.31}$$

można stwierdzić, że dla przeciętnych wartości L_k, L_{hk}, k_2 lub k_1 spełniona jest następująca zależność:

$$LO_{pkNN} << LO_{kNN} \tag{2.32}$$

Dla pełnego porównywania złożoności obliczeniowej owych metod należy jeszcze uwzględnić operacje obliczenia wartości funkcji potencjalnych dla metody *pkNN*.

W obu wspomnianych metodach istnieje możliwość doświadczalnego wyboru optymalnej wartości k tak, aby zminimalizować prawdopodobieństwo wystąpienia błędów. W przypadku gdy wartość parametru k jest porównywalna z liczbą obiektów reprezentujących pojedynczą klasę, do klasyfikacji można stosować metodę wzorców. Gdy liczba obiektów wzorca jest stosunkowo mała, metoda ta jest jedną z najprostszych metod klasyfikacji. Jednakże gdy liczba ta jest taka sama jak dla metod *kNN* oraz *pkNN*, to złożoność obliczeniowa i rozmiar wymaganej pamięci dla tej metody stają się zbliżone do tych dla metod *kNN* oraz *pkNN*. Tym niemniej obiekty wzorca reprezentujące daną klasę w wyniku rożnego rodzaju operacji uśredniania mogą się różnić od ich odpowiedników dla metod *kNN* oraz *pkNN*. Pomijając tę różnicę można stwierdzić, że dla $k=1$ wynik klasyfikacji dla wszystkich trzech rozpatrywanych metod (*kNN*, *pkNN*, wzorców) może być identyczny, przy czym wymaga to odpowiedniego doboru rozmiaru sąsiedztwa obiektów dla metody *kNN* oraz rozmiaru otoczenia obiektów wzorców jak i ograniczenia pól potencjału dla metody *pkNN* przy dodatkowym pominięciu kwadratu w wzorze (2.23).

W celu wyznaczenia wielo-obiektowego wzorca można wykorzystać technikę kwantyzacji wektorowej VQ (*Vector Quantization*). Zastosowanie tej techniki nazywanej również klasteryzacją, ma na celu znalezienie optymalnego wzorca minimalizującego rozproszenie obiektów $\boldsymbol{H}^i$ wokół niego. Za kryterium optymalności wzorca można przyjąć minimalizację sumy odległości obiektów $\boldsymbol{H}^i$ do najbliższych im obiektów szukanego wzorca. Postawienie takiego kryterium optymalności pozwala na znalezienie wzorca również za pomocą algorytmu lub strategii ewolucyjnej [Mic96]. Kluczowe zagadnienie stanowi wybór liczby obiektów wzorca reprezentujących pojedynczą klasę. Przy wyborze tym można kierować się znajomością optymalnej wartości parametru k[15] dla metody *kNN* lub *pkNN* oraz adekwatnej liczby obiektów ciągu uczącego.

[15] Wyznaczonej na podstawie analizy wyników metody *kNN* lub *p-kNN* dla różnych wartości k.

Chociaż metoda wzorców, będąca pewnym uproszczeniem metody *kNN*, ma lepsze właściwości z punktu widzenia sprawności obliczeniowej[16], to jednak zakres jej zastosowania jest wyraźnie mniejszy. Tym niemniej w niektórych przypadkach metoda ta mimo swej prostoty bywa skuteczniejsza od innych. Dlatego też ma ona pierwszeństwo zastosowania i dopiero gdy zawiedzie podejmowane są próby aplikacji innych metod.

Zdaniem autora spośród wspomnianych metod, metoda *pkNN* pozwala na najbardziej wnikliwą analizę struktur rozpoznawanych klas. Zapewnia ona największą elastyczność w podejmowaniu decyzji, eliminując jednocześnie wady klasycznego kryterium klasyfikacji metody *kNN*. Ponadto może stanowić ona podstawę do wstępnego wyznaczenia optymalnej liczby obiektów wzorca klasy.

2.14. Rozpoznawanie grupowe na podstawie głosowania

Idea rozpoznawania grupowego polega na jednoczesnej klasyfikacji kilku obiektów tego samego pochodzenia. Wówczas, gdy klasa rozpoznawanej grupy nie jest uwzględniona w przestrzeni cech, istnieje możliwość, iż poszczególne obiekty tej obcej grupy zostają zaklasyfikowane do różnych klas, co może pozwalać na poprawne rozpoznawanie na podstawie „głosowania" czyli wskazania klasy, do której zaliczono najwięcej obiektów z grupy. W tym celu po zakończeniu procedury klasyfikacji pojedynczych obiektów przeprowadza się głosowanie. W celu ograniczenia błędnej akceptacji można zakładać tylko częściowe ‘pochłonięcie’ dziur przestrzeni cech przez rozpatrywane klasy. W przypadku rozpoznawania grupowego można to osiągnąć wprowadzając dla pojedynczego obiektu grupy pewnego progu[17] określającego, czy posiada on prawo do głosowania za daną klasę czy też nie. Dla danej klasy próg ten można optymalnie wyznaczyć badając jej strukturę. W najprostszym przypadku za próg można przyjąć średnią lub medianę odległości obiektów do własnej klasy. Ogólnie próg ten najlepiej określić, badając posortowany ciąg odległości obiektów do własnej klasy. Wtedy ustala się go tak, aby zapewnić prawo głosowania dla pewnego wybranego procentu obiektów własnych. Przykładowo próg 75 procentowy to taki, który zapewnia prawo głosowania dla 75% obiektów danej klasy. Badania wykazały, że wybór wartości progu ma zasadniczy wpływ na jakość identyfikacji. W szczególności przyjęcie małego progu może

[16] Pod warunkiem, że liczba obiektów wzorca jest znacznie mniejsza od L_{hk}.
[17] Pewna odległość od obiektów klasy lub wartość funkcji przynależności dla pojedynczego obiektu.

pozwolić na podjęcie decyzji tylko na podstawie krótkich fragmentów wypowiedzi lecz spełniających względnie ostrzejsze wymogi podobieństwa do fragmentów wzorcowych (względnie duża wartość funkcji przynależności). Oznacza to również możliwość ignorowania innych, mniej reprezentatywnych fragmentów wypowiedzi. W systemach z otwartym zbiorem rozpoznawanych klas zastosowanie idei rozpoznawania grupowego wiąże się z koniecznością określenia minimalnej liczby głosów wymaganej do uznania klasyfikowanego obiektu za należący do danej klasy.

2.15. Analiza stopnia rozłączności klas w przestrzeni cech

Niezależnie od metody klasyfikacji warunkiem koniecznym i wystarczającym do prawidłowego rozpoznawania jest rozłączność klas w przestrzeni cech[18]. Dla konkretnej klasy i oznacza to, że minimalna wartość jej funkcji przynależności dla własnych obiektów musi być większa od maksymalnej jej wartości dla obcych obiektów. Biorąc pod uwagę obiekty ciągu uczącego, można zatem zdefiniować miarę dyskryminacji dla danej klasy D^i jako stosunek tych dwóch wartości:

$$D^i = 100.\frac{W_{min}}{O_{max}} = 100.\frac{\min\limits_{arg\,W} C^i(\boldsymbol{h}^W)}{\max\limits_{arg\,O} C^i(\boldsymbol{h}^O)} \qquad (2.33)$$

przy czym: $\boldsymbol{h}^W \in \boldsymbol{H}^i$: obiekt własnej klasy, $\boldsymbol{h}^O \in (\boldsymbol{H} - \boldsymbol{H}^i)$: obiekt obcy.

W przypadku gdy wartość miary D^i jest większa od stu, klasę można uznać za wystarczająco dyskryminowaną. Oznacza to, że w myśl przyjętej metody klasyfikacji obiekty reprezentujące klasę i w ciągu uczącym (obiekty zbioru $\boldsymbol{H}^i$) są w pełni rozdzielone w przestrzeni cech od wszystkich innych obiektów ciągu uczącego $\boldsymbol{H}$. Wtedy o jakości klasyfikacji nowych nieznanych obiektów wyrażonej np. jako prawdopodobieństwo występowania błędu, decyduje głównie stopień reprezentatywności ciągu uczącego. Jeżeli wartość miary jest jednak mniejsza lub równa sto oznacza to, że przyjęta przestrzeń cech jest wadliwa i że żadna metoda klasyfikacji nie 'potrafi' w pełni rozdzielić obiektów zbioru $\boldsymbol{H}^i$ od innych. Zatem przy takiej metodzie istnieje niezerowe prawdopodobieństwo występowania błędów rozpoznawania. Naturalnie im większa jest miara D^i tym mniejsze jest prawdopodobieństwo występowania błędów i odwrotnie. Należy dodać, że prawdopodobieństwo występowania błędnej akceptacji przez klasę i

[18] Odnosi się to tylko do zamkniętego zbioru rozpoznawanych klas.

zależy jeszcze – a może przede wszystkim – od prawdopodobieństwa istnienia obiektów obcych, dla których wartość funkcji przynależności C^i należy do przedziału (W_{min}, O_{max}). Analogicznie prawdopodobieństwo występowania błędnego odrzucania zależy od prawdopodobieństwa istnienia obiektów własnych w tym samym przedziale. Warto zauważyć, że manipulując wartościami W_{min} lub O_{max}, poprzez eliminowanie niektórych obiektów z ciągu uczącego, można odpowiednio zmniejszyć prawdopodobieństwo występowania jednego błędu kosztem drugiego. Wartość graniczna, określająca czy obiekt może być klasyfikowany jako należyty do danej klasy, bywa określona mianem progu detekcji.

ROZDZIAŁ III

Klasyfikacja i porównywanie metod automatycznego rozpoznawania mówcy

3.1. Wprowadzenie

Rozpoznawanie głosów należy do grupy biometrycznych metod identyfikacji osób [Abs97]. Inne techniki z tej samej grupy polegają na identyfikacji osób na podstawie odcisków palców, kształtu dłoni, obrazu siatkówki oka, wyglądu twarzy, grupy krwi oraz badań nad strukturą genetyczną DNA wydobytego z próbek tkanek.

Sygnał mowy powstaje w wyniku sekwencji kompleksowych transformacji informacji i dźwięku na kilku poziomach: semantycznym, lingwistycznym, artykulacyjnym oraz akustycznym. Różnice w charakterze tych transformacji dla różnych głosów mogą się objawiać w formie różnic w akustycznych właściwościach wynikowego sygnału mowy. Charakterystyki osobnicze głosu występujące w sygnale akustycznym wynikają z różnic anatomicznych w wymiarach i kształcie aparatu głosowego mówcy oraz są związane z właściwościami oddziaływań fizjologicznych i psychologicznych powstających w części mózgu i systemu nerwowego, sterującego procesem mówienia. Zatem oprócz różnic o charakterze czysto biologicznym, mówcy różnią się między sobą wyuczonymi zwyczajami i manierami w sposobie mowy. Niezależnie od tego sposób wymowy zależy również od aktualnego stanu psychicznego i fizycznego mówcy. Metody automatycznego rozpoznawania mówcy usiłują wydobyć z sygnału mowy i wykorzystać do dyskryminacji mowy różnych osób cechy indywidualne głosu w celu ich rozpoznawania.

Istniejący trend w rozwoju biometrycznych technik identyfikacji jest ukierunkowany na zwiększenie ich skuteczności, aż do uzyskania pełnej niezawodności. Podstawową motywacją do badań w tej dziedzinie jest potrzeba stworzenia skuteczniejszych i zarazem wygodniejszych w użyciu systemów rozpoznawania osób. Technika automatycznego rozpoznawania mówcy pozwala na zweryfikowanie jego tożsamości w celu kontroli dostępu do różnego rodzaju usług, takich jak: głosowe karty telefoniczne, obsługa konta bankowego przez telefon, zakupy telefoniczne, dostęp do różnego rodzaju systemów informacyjnych, poczta głosowa, ochrona dostępu do poufnych baz danych, zdalny dostęp do komputerów oraz sterowanie urządzeń głosem przez osoby do tego uprawnione. Innym odmiennym obszarem zastosowania systemów automatycznego rozpoznawania mówcy jest sądownictwo i kryminalistyka [Bas95].

Zaletą automatycznego rozpoznawania osób na podstawie indywidualnych charakterystyk głosu jest to, że atrybuty osobnicze nie mogą zostać zgubione (np. jak klucz) lub zapomniane (jak skomplikowany kod lub hasło). Niestety mogą one jednak

tymczasowo się zmienić w wyniku choroby (np. katar) i to w stopniu uniemożliwiającym ich wykorzystanie w systemie automatycznego rozpoznawania mówcy. Choć głos nie może zostać 'skradziony', to istnieje realna możliwość jego naśladowania lub nagrania. Zasadnicza część zastosowań automatycznego rozpoznawania mówcy jest związana z transmisją sygnału mowy w liniach telefonicznych. Biorąc pod uwagę rosnącą powszechność dostępu do telefonii oraz komputerów, koszt budowy automatycznych systemów rozpoznawania mówcy redukuje się do kosztu oprogramowania. Natomiast korzyści ze zastosowania tych systemów mogą być ogromne. Przykładowo firma księgowa Ernst & Young oszacowała, że hakerzy w Stanach Zjednoczonych kradną rocznie na nielegalnych transakcjach ok. 3 do 5 bilionów dolarów. Zastosowanie automatycznego systemu rozpoznawania mówcy mogłoby znacznie ograniczyć te straty. Jednak w takich rozwiązaniach gdy wymagany jest wysoki poziom zabezpieczenia systemu, to musi on być zintegrowany z innymi systemami zabezpieczenia, takimi jak inteligentna karta (*smart card*). Tym niemniej jakość dzisiejszych systemów rozpoznawania mówcy pozwala na indywidualne ich zastosowanie w wielu praktycznych aplikacjach.

Istnieje wiele komercyjnych systemów automatycznego rozpoznawania mówcy. Wśród popularnych producentów takich systemów można wymienić firmy: ITT, Lernout & Hauspie, T-NETIX, Veritel, Voice Control Systems i Sprint produkująca popularną karty telefoniczną *Sprint's Voice FONCARD®* używaną w Stanach Zjednoczonych.

Szerszy przegląd metod i problematyki związanej z automatycznym rozpoznawaniem mówcy można znaleźć np. w pracach: [Ata76, Bas93, Cam97, Dod85, Fur89, Fur96, Ros76, Ros91].

3.2. Klasyfikacja systemów automatycznego rozpoznawania mówcy

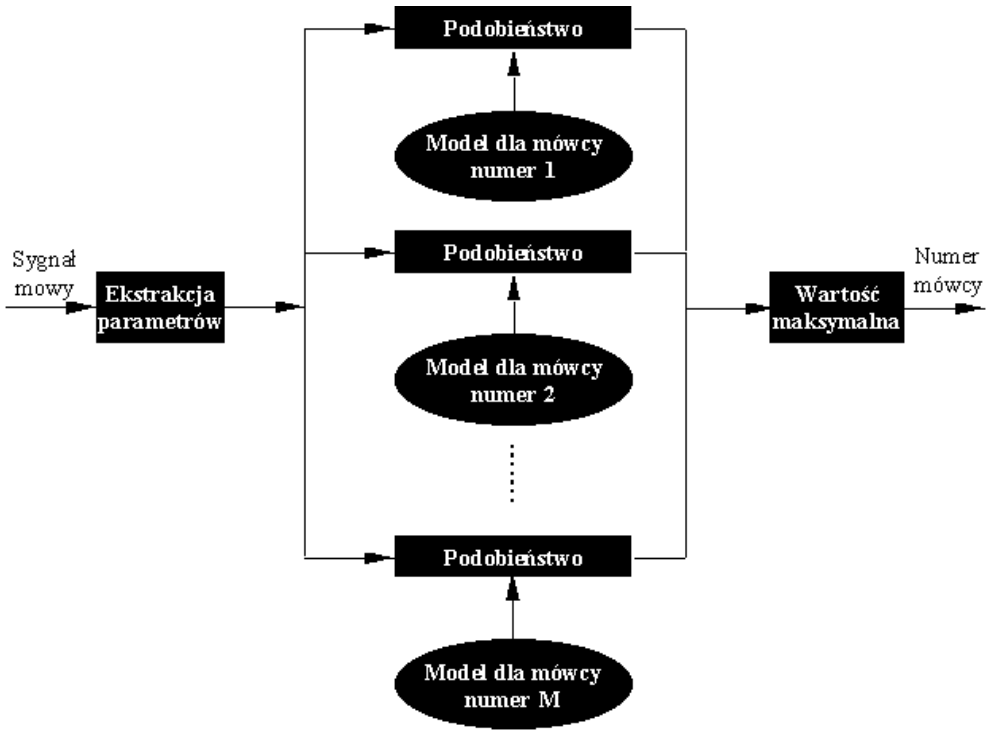
Systemy automatycznego rozpoznawania mówcy można klasyfikować według różnych kryteriów w zależności od następujących aspektów [Cam97, Dod85, Fur96, Ros76, Ros91]:

- rodzaju podejmowanej decyzji o rozpoznawaniu. System automatycznego rozpoznawania może pracować w trybie weryfikacji lub identyfikacji mówcy.
- rozmiaru populacji oraz założenia o otwartości (bądź nie) zbioru rozpoznawanych mówców (*open/closed speakers' set*).

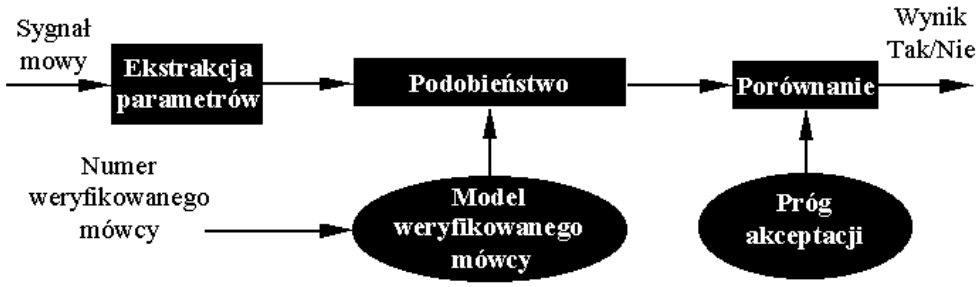
- zależności rozpoznawania od tekstu wypowiedzi. Tekst wypowiedzi – na podstawie której przeprowadza się rozpoznawanie mówcy – może być dowolny, z góry ustalony lub proponowany przez system w sposób losowy.

3.2.1. Identyfikacja a weryfikacja mówcy – procedura decyzyjna

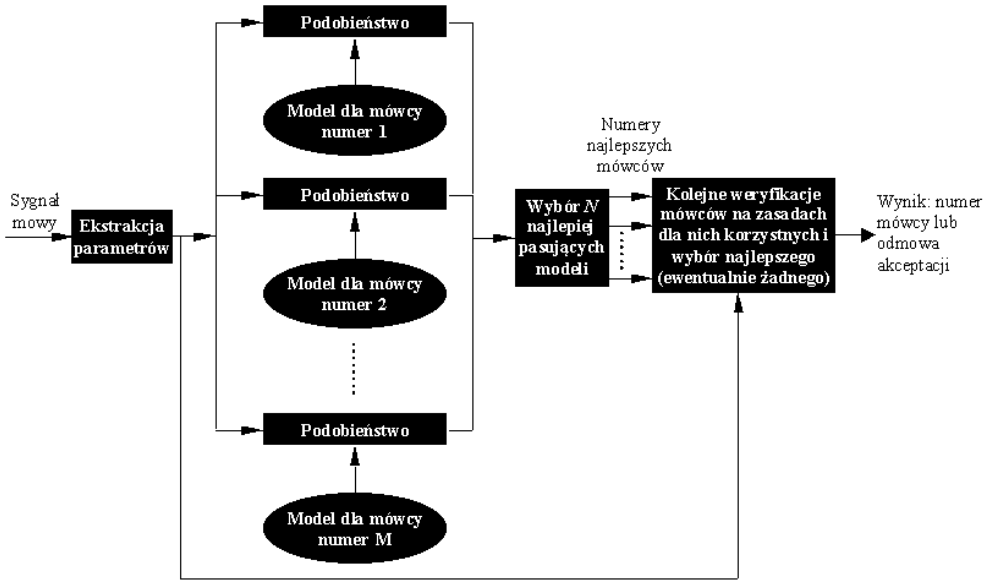
Automatyczne rozpoznawanie głosów obejmuje dwie zasadniczo różniące się procedury: identyfikację i weryfikację [Cam97, Fur89, Fur96]. Przez weryfikację rozumie się sprawdzenie na ile prawdopodobne jest przypisanie danej wypowiedzi pewnej 'podejrzanej' osobie. Natomiast identyfikacja polega na przypisaniu wypowiedzi do którejś z osób należących do ustalonej populacji. Należy zauważyć, że o ile weryfikacja jest jednoznaczna (jest pozytywna tylko dla jednej osoby) to identyfikacja osoby jest równoważna kolejnej weryfikacji wszystkich osób należących do populacji. W istocie jest jednak różnica między stopniem szczegółowości testów stosowanych przy weryfikacji i przy identyfikacji. Weryfikacja wymaga znacznie wyższego stopnia zgodności dla wydania pozytywnej decyzji. Rysunek 3.1a przedstawia schemat działania procedury identyfikacji w przypadku zamkniętego zbioru rozpoznawanych mówców (liczebność zbioru wynosi *M*). Schemat działania procedury weryfikacji jest przedstawiony na rysunku 3.1b.



Rys. 3.1a: Procedura identyfikacji mówcy w zamkniętym zbiorze o liczebności *M*.



Rys. 3.1b: Procedura weryfikacji mówcy.



Rys. 3.2: Algorytm identyfikacji mówcy w zbiorze otwartym.

W systemach automatycznej weryfikacji istnieje możliwość wyboru parametrów metody klasyfikacji tak, aby minimalizować prawdopodobieństwo wystąpienia błędów dla konkretnego mówcy [Sho99]. Z kolei w systemach identyfikacji w których zakłada się, że zbiór identyfikowanych mówców jest zbiorem otwartym można przeprowadzić wstępną redukcję przestrzeni poszukiwań (liczby podejrzanych klas) za pomocą algorytmu identyfikacji[1] przy warunkach jednakowych dla wszystkich klas. Następnie można dokonać oddzielnych weryfikacji kilku klas, które uzyskały najlepsze wyniki we wstępnym procesie identyfikacji. Weryfikację każdej z tych klas przeprowadza się wtedy na warunkach dla niej preferencyjnych tzn. zapewniających małe błędy

[1] Przy założeniu, że badany głos należy do pewnego zamkniętego zbioru.

weryfikacji. W zależności od wyniku wykonanych procedur weryfikacji można podjąć ostateczną decyzję o identyfikacji mówcy bądź jego odrzucaniu (dla rozpatrywanej populacji badany mówca jest obcy). Schemat procedury identyfikacji w otwartym zbiorze rozpoznawanych klas przedstawiony jest na rysunku 3.2.

W rozdziale drugim przedyskutowano zależność błędów rozpoznawania od sposobu rozdzielania przestrzeni cech pomiędzy klasami. Stwierdzono wówczas, że proporcje pomiędzy poszczególnymi błędami zależą przede wszystkim od stopnia pochłaniania przestrzeni przez rozpoznawane klasy. Na ogół rozszerzenie obszarów klas zmniejsza prawdopodobieństwo błędnego odrzucania kosztem zwiększenia prawdopodobieństwa błędnej akceptacji i odwrotnie.

Ogólnie rzecz biorąc, zadaniem procedury klasyfikacji jest określenie stopnia podobieństwa badanego obrazu obiektu do określonego wzorca (modelu). W przypadku weryfikacji podobieństwo to jest określane jedynie dla konkretnego modelu 'podejrzanego' mówcy. Natomiast w przypadku identyfikacji określa się poszczególne stopnie podobieństwa próbki do każdego z modeli należących do przestrzeni cech (dla każdego mówcy z rozpatrywanej populacji). Zadaniem procedury decyzyjnej jest podjęcie najbardziej właściwej decyzji w oparciu o wynik (dla weryfikacji) bądź wyniki (dla identyfikacji) procedury rozpoznawania. Tabela 3.1 przedstawia możliwe warianty decyzyjne dla procedury identyfikacji oraz weryfikacji. Ponadto w niektórych systemach podjęte mogą być decyzje przejściowe np. o przedłużeniu czasu badania lub powtórzeniu próby rozpoznawania.

Procedura →	**Identyfikacja**	**Weryfikacja**
Zamknięty zbiór mówców	Identyfikacja badanego mówcy jako mówcy numer #	Akceptacja bądź odrzucanie niezależnie od typu zbioru mówców
	Brak możliwości jednoznacznego określenia numeru mówcy	
Otwarty zbiór mówców	Identyfikacja badanego mówcy # jako mówcy o numerze #	
	Brak możliwości jednoznacznego określenia numeru mówcy	
	Odrzucanie -badany mówca # nie należy do populacji	

Tabela 3.1: Warianty decyzyjne w systemach automatycznego rozpoznawania mówcy.

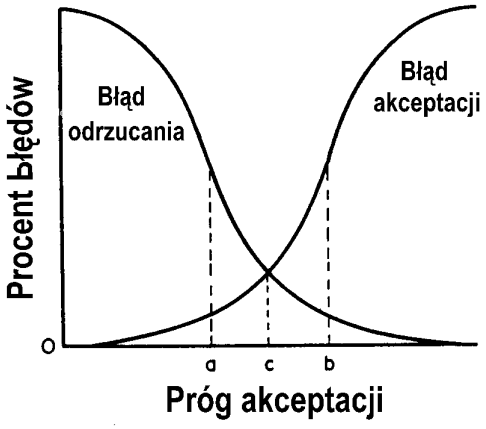
Stopień rozszerzenia obszaru danej klasy w przestrzeni można jednoznacznie powiązać z wartością pewnego progu (*threshold*) jej funkcji przynależności, określającą czy dany obiekt należy do klasy czy też nie. Biorąc pod uwagę, że procedura identyfikacji jest teoretycznie równoważna procedurom weryfikacji kolejnych mówców, to możliwe jest zunifikowane podejście do zagadnienia doboru wartości progu. Zazwyczaj

wybór tej wartości dokonuje się tak, aby spełniony był jeden z następujących warunków:

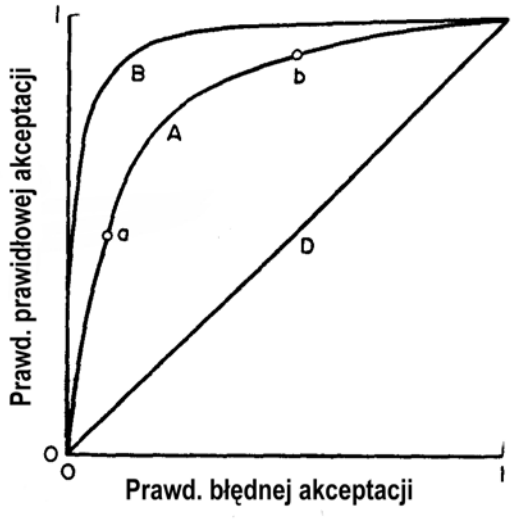
1) równe błędy odrzucania i akceptacji (Neyman–Pearson) – przypadek c na rysunku 3.3,
2) stały błąd akceptacji lub odrzucenia – przypadki a, b na rysunku 3.3,
3) stały stosunek błędu akceptacji do błędu odrzucania,
4) minimalny sumaryczny błąd (suma błędu akceptacji i odrzucania).

O wyborze jednego z powyższych warunków decydują skutki wystąpienia poszczególnych błędów. Zarówno w systemach weryfikacji jak i identyfikacji wartość progu może być różna dla różnych mówców. Ponadto może być ona zmieniana w zależności od ryzyka włamania się do systemu (np. wyższy próg w nocy). Rysunek 3.3 przedstawia typowe zależności błędów akceptacji oraz odrzucania od wartości przyjętego progu.

Ponieważ błędy akceptacji oraz odrzucania są skorelowane, to oszacowanie jakości metody rozpoznawania wiąże się z koniecznością określenia wartości obu błędów. Dlatego też często podawana jest zależność jednego z nich od drugiego. Inny sposób przedstawienia jakości systemu weryfikacji mówcy stanowią tzw. charakterystyki operacyjne odbiorcy (*Receiver operating characteristic ROC*). Przykłady takich charakterystyk dla trzech różnych typów systemów przedstawiono na rysunku 3.4 gdzie oś x-ów reprezentuje prawdopodobieństwo błędnej akceptacji (akceptacja obcego mówcy) zaś oś y-ów prawdopodobieństwo prawidłowej akceptacji (akceptacja właściciela modelu). Łatwo wywnioskować, że charakterystyka B odpowiada najlepszemu przypadkowi z trzech przedstawionych.



Rys 3.3: Zależność błędów odrzucania oraz akceptacji od wartości progu.



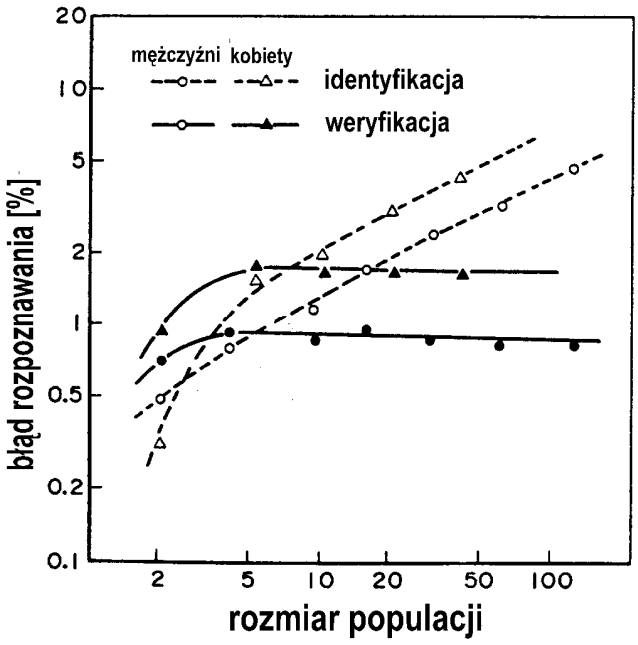
Rys 3.4: Kształt typowych charakterystyk operacyjnych odbiorcy ROC.

3.2.2. Rozmiar populacji, otwartość zbioru rozpoznawanych mówców

Kolejnym czynnikiem podziału systemów automatycznego rozpoznawania mówców – zdaniem autora najważniejszym z punktu widzenia oceny skuteczności metody rozpoznawania – jest rozmiar populacji[2] oraz założenie o otwartości (bądź jej brak) zbioru rozpoznawanych mówców. Podział ten jest szczególnie istotny, ponieważ wraz ze wzrostem liczby mówców, rośnie wykładniczo prawdopodobieństwo błędnego rozpoznawania [Fur89]. Sytuacja ta dotyczy w szczególności procedury identyfikacji mówcy ale występuje również w przypadku weryfikacji mówcy, gdy siła dyskryminacji wektora cech jest stosunkowo mała [Fur96]. Rysunek 3.5 przedstawia typowe zależności błędów procedury weryfikacji i identyfikacji od rozmiaru populacji.

Należy zaznaczyć, że procedura weryfikacji jest jednakowa zarówno dla otwartego jak i dla zamkniętego zbioru rozpoznawanych mówców. Procedura identyfikacji natomiast różni się w sposób zasadniczy dla obu tych przypadków. Zdaniem autora w dotychczasowych badaniach nie poświęcono należytej uwagi temu szczególnie ważnemu zagadnieniu. Tym niemniej wyczerpujący opis założeń metodologicznych i algorytmu rozpoznawania mówców w zbiorach otwartych można znaleźć w [Bas87] lub [Bas93].

[2] Przyjmuje się [Bas93], że liczba mówców mniejsza od 30 zaliczana jest do małych populacji.



Rys 3.5: Błąd rozpoznawania w zależności od rozmiaru populacji dla procedury identyfikacji oraz weryfikacji [Fur89].

3.2.3. Rozpoznawanie zależne oraz niezależne od tekstu

Rozważa się dwa podejścia do rozpoznawania głosu [Cam97, Fur96]: rozpoznawanie wypowiedzi ustalonej z ograniczonego zbioru wypowiedzi (*Text Dependent Speaker Recognition*) bądź swobodnej czyli dowolnej (*Text Independent Speaker Recognition*). W szczególności w tym drugim przypadku dopuszcza się rozpoznawania głosu niezależnie nawet od języka wypowiedzi[3]. W tym pierwszy zaś wymagane jest, aby mówca używał tej samej wypowiedzi w fazie treningu systemu jak i w procesie rozpoznawania. Pewną odmianę tego podejścia stanowi metoda rozpoznawania wypowiedzi wybieranej przez system w sposób losowy (z ustalonego zbioru wypowiedzi) i 'podawanej' użytkownikowi wizualnie bądź drogą dźwiękową (*Text–Prompted Speaker Recognition*). Ma to zapobiec ewentualnym próbom oszukania systemu rozpoznawania poprzez odtworzenie zdobytego wcześniej nagrania dla danego użytkownika. Przykład takiego rozwiązania opisany jest np. w pracach [Mat93, Hig91].

[3] Oznacza to między innymi możliwość rozpoznawania głosu mówcy w przypadku gdy model jego głosu został stworzony dla innego języka niż ten użyty w procesie rozpoznawania [Sho98].

W przypadku rozpoznawania zależnego od tekstu stosuje się zazwyczaj jedno z dwóch podejść: technikę dopasowania opartą na programowaniu dynamicznym *DTW* (*Dynamic Time Warping*) [Fur81, Fur96] lub ukryte modele Markowa HMM (*Hidden Markov Models*) [Che95, Mat95, Rey95, Tis91]. Natomiast metody rozpoznawania niezależnego od tekstu bazują na jednym z następujących podejść [Fur96]:

- długookresowych modelach statystycznych,
- technice kwantyzacji wektorowej,
- ergodycznych modelach Markowa,
- selekcji zdarzeń specyficznych,
- selekcji tych samych segmentów quasi-stacjonarnych np. głoska "a",
- sieciach neuronowych.

3.3. Metody rozpoznawania mówców niezależne od tekstu bazujące na długookresowych modelach statystycznych

W zależności od długości czasowej fragmentu wypowiedzi, na podstawie którego oblicza się pojedynczy wektor cech mówi się o metodach rozpoznawania bazujących na parametrach długo- lub krótkookresowych. W tym pierwszym przypadku wydobywane są pewne wartości średnie wybranej wielkości fizycznej (np. moc, częstotliwość podstawowa tonu krtaniowego, widmo krótkoterminowe). W tym drugim zaś przedmiotem zainteresowania jest zazwyczaj rozkład czasowy danej wielkości fizycznej obrazujący czasową zmienność sygnału. Zrozumiały jest fakt, że w tym drugim przypadku trudniej jest włamać się do systemu rozpoznawania. Pierwsza koncepcja bywa częściej wykorzystywana w systemach rozpoznawania na podstawie wypowiedzi swobodnej, zaś druga w systemach rozpoznawania zależnego od tekstu. W niektórych przypadkach parametry są wydobywane tylko z wybranych fragmentów sygnału. Przykładowo określenie częstotliwości tonu krtaniowego dokonuje się tylko na podstawie mowy dźwięcznej, odrzucając tym samym fragmenty mowy bezdźwięcznej. Również eliminowane mogą być fragmenty ciszy występujące między słowami a nawet sylabami.

Metody rozpoznawania mówców niezależne od tekstu są często realizowane z wykorzystaniem długookresowych modeli statystycznych. Opierają się one na założeniu, że choć niektóre parametry opisu sygnału mowy mogą być zmienne w czasie, to ich wartość średnia na dostatecznie długim odcinku czasowym dąży do stałej wartości zależnej od globalnych, niezależnych od tekstu cech osobniczych głosu. W związku z tym można podjąć próbę dyskryminacji głosów na podstawie długoterminowych

średnich wartości wektorów cech oraz ich wariancji. Ze względu na fizjologiczne zmiany w mowie parametry opisujące sygnał mowy nie posiadają w istocie stałych wartości średnich lub stałych wariancji. Jednak w miarę wzrostu czasu uśredniania ich wartości stają się bardziej skupione [Mar77, Mar79]. Wśród parametrów opisujących sygnał mowy wykorzystywanych w tych metodach można wymienić: częstotliwość podstawową tonu krtaniowego, znormalizowaną wariancję energii oraz widmo długookresowe.

Badania wykazują [Mar77], że względnie stałe wartości średnie różnych parametrów można osiągnąć dla wypowiedzi o długości kilkudziesięciu sekund. W szczególności wykazano, że w obrębie jednej klasy odchylenie standardowe dla trzech różnych parametrów maleje w przedziale uśredniania $t \in (0.7s - 70s)$ proporcjonalnie do $t^{-1/3}$. Ponadto między klasowa wariancja wartości średnich parametrów rośnie ze wzrostem czasu uśredniania, co ułatwia dyskryminację.

Zaletą długookresowych pomiarów cech jest brak konieczności czasowego dopasowania obiektu do wzorca przed obliczeniem ich wzajemnego podobieństwa. Ponadto liczba wektorów cech ekstrahowanych z wypowiedzi jest w tym przypadku znacznie mniejsza niż w przypadku parametrów krótkookresowych, co w zasadniczy sposób skraca proces klasyfikacji. Fakt ten ma pierwszorzędne znaczenie w systemach identyfikacji ponieważ pozwala na jej przeprowadzenie w czasie rzeczywistym oraz na redukcję rozmiaru zarówno pamięci operacyjnej jak i pamięci stałej potrzebnej do przechowywania informacji o modelach poszczególnych głosów.

3.4. Wyniki wcześniejszych badań

Badania nad automatycznym rozpoznawaniem mówcy są prowadzone szeroko i intensywnie w różnych laboratoriach, instytucjach oraz na licznych uczelniach. Wśród liderów w tej dziedzinie badań, którym udało się skonstruować kilka generacji systemów automatycznego rozpoznawania mówców wymienić można: koncern AT&T, firma Bolt, Beranek and Newman BNN, Instytut Sztucznej Inteligencji Dolle Molle (Szwajcaria), Instytut Technologii Massachusetts–Laboratorium Lincolon MIT-LL, Instytut Politechniki Rensselaer RPI, Narodowy Uniwersytet Tsing Hua (Tajwan) NTHU, firmę Nippon Telegraph and Telephone NTT oraz Uniwersytet Nagoya (Japonia), Uniwersytet Rutgers RU oraz koncern Texas Instruments TI.

Pierwsze systematyczne badania w dziedzinie rozpoznawania głosów przeprowadził Kersta. Jego metoda polegała na wizualnym porównaniu spektrogramów wypowiedzi. Niedoskonałości tej metody spowodowały skierowania uwagi badaczy na inne zobiektywizowane metody rozpoznawania głosów. Zasadniczy wkład w badaniach nad automatycznym rozpoznawaniem mówcy mają m. in. Japończycy: Furui i Itakura, Amerykanie: Atal, Doddington, Pruzanski, Sambur, Rosenberg, Wolf oraz wielu innych.

W Polsce badania nad automatycznym rozpoznawaniem głosów rozpoczęto z kilkuletnim opóźnieniem. Pierwsze publikacje pojawiły się na przełomie lat sześćdziesiątych i siedemdziesiątych i pochodziły z trzech ośrodków: Instytutu Podstawowych Problemów Techniki PAN w Warszawie, Pracowni Fonetyki Akustycznej IPPT PAN w Poznaniu i Instytutu Telekomunikacji i Akustyki Politechniki Wrocławskiej. Aktualnie, szerzej zakrojone prace w zakresie komputerowego rozpoznawania głosów, prowadzone są w ITA Politechniki Wrocławskiej. W pracach autorów polskich koncentrowano się najczęściej na podstawowej problematyce, jaką jest dobór parametrów dyskryminujących głosy. Do takich prac należą m. in. wcześniejsze prace Gubrynowicza, Jassema, późniejsze publikacje m.in. Majewskiego, Basztury i innych. Ostatnie prace ujmowały szerzej problematykę automatycznego rozpoznawania mówcy, włączając w to badania wpływu algorytmu rozpoznawania [Bas87], funkcji podobieństwa [Bas91] oraz rzeczywistych warunków transmisji sygnału mowy [Bas81]. Większość tych prac powstała przy doniosłym udziale profesora Basztury któremu zawdzięcza się główny postęp w badaniach przeprowadzonych w Polsce. Niektóre wyniki badań przeprowadzonych przez profesora Baszturę zostały zaprezentowane w publikacjach zagranicznych aktualnie wymienianych wśród czołowych prac w tym zakresie[4].

Tabela 3.2 przedstawia przegląd opisywanych w literaturze metod automatycznego rozpoznawania mówcy. W kolumnie „metoda" zaznaczono jądro algorytmu klasyfikacji. Jakość metod można oceniać na podstawie procentu błędów identyfikacji (I) lub weryfikacji (V) dla zadanej długości wypowiedzi (w sekundach). Materiał fonetyczny (baza głosów) klasyfikowano jako: laboratoryjny 'Lab', pokojowy (biurowy) 'Pok' lub telefoniczny 'Tel'.

W scenariuszu rozwoju badań nad automatycznym rozpoznawaniem mówcy, wynikającym z analizy literatury, można zauważyć, że istnieje generalny trend do zwiększania rozmiaru populacji oraz koncentracji badań nad rozpoznawaniem mówcy

[4] http://www.lia.univ-avignon.fr/personel/BESACIER/biblio/tout-en.html

w warunkach normalnych (biurowych oraz przez łącza telefoniczne). Ogólnie w miarę upływu czasu uzyskiwano coraz lepsze wyniki. Na ogół jednak nie można wyciągnąć takiego wniosku bezpośrednio z przedstawionych w tabeli wyników, ponieważ brak jest możliwości łatwego porównania wyników różnych metod rozpoznawania. Problem ten jest omówiony w następnym podrozdziale. Rozważając dalsze wnioski wypływające z tabeli można stwierdzić, że najczęściej używane są parametry cepstralne oraz parametry kodowania predykcyjnego. Najnowsze badania są częściej przeprowadzane z wykorzystaniem niejawnych modeli Markowa HMM stosowanych jako metoda dopasowania obiektów do wzorców. Niejawne modele Markowa, podobnie jak w systemach rozpoznawania mowy, zastępują obecnie stosunkowo bardziej ograniczone algorytmy dopasowania obiektów do wzorców oparte na programowaniu dynamicznym DTW.

Autor/ autorzy	Organizacja	Wektor cech	Metoda	Materiał fonetyczny	Zależność od tekstu[5]	Rozmiar populacji	Procent błędów
Wolf 72	BNN, MIT–LL	Widmowe, F_0	Wzorców	Lab.	Zależne	21	I:1.5%@2.5s V:4%@2.5s
Atal 72	AT&T	Obwiednia F_0	Wzorców	Lab.	Zależne	10	I:3%@2s
Atal 74	AT&T	LPC	Wzorców	Lab.	Zależne	10	I:2%@0.5s V:2%@1s
					Niezależne		I:7%@7s
Markel et al. 77	SCRL	F_0, LPC, wzmocnienie	Statystyczna długookresowa	Lab.	Niezależne	4	
Basztura, Majewski 78	ITiA–PW	APPZ	Wzorców	Lab	długoterm. niezależne i zależne	10	I:3%@30s
Markel, Davis 79		LPC	Statystyczna długookresowa	Lab.	Niezależne	17	I:2%@39s
Furui 81	AT&T	Cepstrum	DTW	Tel.	Zależne	10	V:0.2@3s
Doddington 85	TI	Bank filtrów	DTW	Lab.	Zależne	200	V:0.8@6s
Soong et al. 85	AT&T	LPC	Kwantyzacja wektorowa	Tel.	10 cyfr	100	I:5%@1.5s I:1.5%@3.5s
Higgins, Wohlford 86	ITT	Cepstrum	DTW	Lab.	Niezależne	11	V:10%@2.5s V:4.5%@10s

[5] Należy podkreślić, że użycie ograniczonego słownika wyrazów nie koniecznie oznacza, że rozpoznawanie musi być zależne od tekstu. Przykładem tego są prace Che i Lin 1995 oraz Yuan et al.1999.

Autor/ autorzy	Organizacja	Wektor cech	Metoda	Materiał fonetyczny	Zależność od tekstu	Rozmiar populacji	Procent błędów
Basztura, Majewski, Jurkiewicz 87	ITiA–PW	APPZ	Wzorców	Lab.	Niezależne	20	V:1-9%@2s
Attill, et al. 88	RPI	Cepstrum, LPC	Statystyczna długookresowa	Lab.	Zależne	90	V:1%@3s
Basztura, Zuk 91	ITiA–PW	APPZ	Metody minimalno-odległościowe	Lab.	Niezależne i zależne	20	I:3%@2s
Higgins et al. 91	ITT	LPC Cepstrum	DTW	Pok.	Zależne	186	V:1.7%@10s
Tishby 91	AT&T	LPC	HMM	Tel.	10 cyfr	100	V:3%@1.5s V:1%@3.5s
Chen, Lin, Wang 93	NTHU	LPC cepstrum	Kwantyzacja macierzowa	Lab.	Niezależne (10 cyfry)	50	I:0.8%@1.5s
Reynols, Carlson 95	MIT-LL	Mel-Cepstrum	HMM	Pok.	Zależne	138	I:0.8%@10s V:0.2%@10s
Che, Lin 95	RU	Cepstrum	HMM	Pok.	Zależne	138	I:0.6%@2.5s I:0.1%@10s V:0.6%@2.5
Reynolds 96	MIT–LL	Mel-Cepstrum	HMM	Tel.	Niezależne	416	V:11%@3s V:6%@10s V:3%@30s
Zilovic et al. 98	BCR, RU	LPC Cepstrum	Kwantyzacja wektorowa	Pok.	Niezależne	20	I:4%@2.5s
				SNR 30dB			I:9%@2.5s
Yuan et al. 99	NU	LPC	Kwantyzacja binarna	Pok.	Niezależne (10 cyfr)	42	I:1%@2s
SCRL – Speech Communication Research Laboratory, Inc., Santa Barbara, California BCR – Bell Communication Research California, NU – Uniwersytet Nanjing, Chiny							

Tabela 3.2: przegląd wcześniejszych prac z zakresu automatycznego rozpoznawania mówcy (tabela adaptowana z [Cam97] oraz [Bas93] i uzupełniona).

W literaturze można też zaobserwować, że metody rozpoznawania mówcy zależne od tekstu są częściej przedmiotem badań, niż inne z omawianych tu metod. Wynika to z faktu, że są one na ogół prostsze w realizacji i uzyskują lepsze wyniki. Ponadto są one bardziej związane z zagadnieniem rozpoznawania treści mowy, będącym przedmiotem bardziej wnikliwych i szerszych badań. Warto przypomnieć, że jednym z ważniejszych celów badań nad zależnością sygnału mowy od aspektu osobniczego jest próba eliminowania cech osobniczych występujących w sygnale tak, aby umożliwić jego sprawniejszego wykorzystania w systemach automatycznego

rozpoznawania treści mowy. Podsumowując przegląd badań w zakresie automatycznego rozpoznawania mówcy należy zauważyć, że często doniesienia na ten temat mają charakter fragmentaryczny, a duża część prac była i jest prowadzona niejako w cieniu badań nad automatycznym rozpoznawaniem treści mowy.

3.5. Porównywanie i weryfikacja skuteczności metod rozpoznawania

Trudności w porównywaniu metod automatycznego rozpoznawania mówcy opisywanych przez różnych autorów wynikają w zasadzie z następujących przyczyn:

1. Wyniki badań uzyskano w różnych pracach na podstawie różnych materiałów badawczych (różnych baz danych głosów) zarejestrowanych w różnych warunkach.
2. W różnych pracach występują różnice w liczbie sesji nagrań oraz odstępach czasowych między nimi.
3. Obserwuje się zależność wyników rozpoznawania od długości danych trenujących oraz testujących, a te nie są ujednolicone.
4. Różne rozmiary badanej populacji mówców.
5. Wyniki przedstawione przez różnych autorów uzyskane są dla różnych procedur rozpoznawania: weryfikacji lub identyfikacji.
6. Wyniki dla procedury identyfikacji zależą zasadniczo od tego, czy zbiór rozpoznawanych mówców jest zamknięty czy otwarty. W tym drugim przypadku oprócz zależności od rozmiaru populacji, wyniki zależą też od liczby uwzględnionych w badaniach intruzów nie należących do podstawowej populacji, dla których system został wytrenowany. Zależność ta również zachodzi w przypadku weryfikacji mówcy.
7. Istnieją różnice w trybach pracy systemów rozpoznawania, głównie w założeniach procedury decyzyjnej. Systemy mogą bowiem posiadać różne kryteria podjęcia decyzji o akceptacji bądź odrzuceniu mówców, o powtórzeniu próby rozpoznawania lub kontynuacji procedury rozpoznawania.
8. Niemożliwe jest porównywanie wyników metod rozpoznawania niezależnego od tekstu z ich odpowiednikami dla metod rozpoznawania zależnego od tekstu.

Likwidując powyższe różnice (całkowicie lub częściowo) można by było dokonać porównania jakości niektórych metod w sposób „globalny” tzn. traktując metodę jako połączenie wyboru wektorów cech (parametrów) oraz algorytmu klasyfikacji. Jednak takich badań do tej pory wyczerpująco nie prowadzono.

Problem porównywania różnych podejść do zagadnienia rozpoznawania mówcy jest na tyle istotny, że w połowie lat dziewięćdziesiątych zdecydowano się na budowanie baz danych głosów przystosowanych do konkretnych założeń poszczególnych podejść do zadania rozpoznawania. Celem budowy tych baz jest przede wszystkim dostarczenie badaczom jednakowych materiałów doświadczalnych przystosowanych do obiektywnej oceny konkretnego podejścia i w pewnym stopniu narzucających warunki przeprowadzenia eksperymentów. Przykładowo, poszczególne głosy w bazie mogą być podzielone a priori na te należące do zakładanej populacji oraz obce (intruzi). Często w literaturze anglojęzycznej te pierwsze są określane mianem owiec (*sheep*) zaś te drugie mianem wilków (*wolfs*). Dane zazwyczaj bywają również dzielone na trenujące oraz testujące. Jednakże niektóre bazy nie mają jednoznacznie określonych procedur przedstawiania wyników. Jest to ich dość istotny i stale aktualny mankament. Choć proponowane bazy danych są rozmaite to mogą być przypadki w których do konstruowanej metody żadna baza w pełni nie pasuje. Przypadki takie należą jednak do rzadkości.

Próbą rozwiązania powyższych problemów jest program oceny systemów rozpoznawania mówcy *(Speaker Recognition Evaluation*) organizowany przez Narodowy Instytut Standardów i Technologii (*National Institute of Standards and Technology NIST*) od 1995 roku i sponsorowany przez rząd amerykański. Ściśle określone procedury przedstawiania wyników jak i standardowa baza danych z zarejestrowanymi wypowiedziami i podziałem na dane do nauczania i testowania pozwalają na organizowanie corocznych konkursów na najefektywniejszy system rozpoznawania mowy [Nis98]. Brak jest jednakże analogicznych działań dla innych języków narodowych jak i szerokiego bezpłatnego dostępu do danych przygotowywanych przez NIST dla naukowców z całego świata.

Do czołowych producentów baz danych głosów należy koncern danych lingwistycznych Uniwersytetu Pensylwanii (*Linguistic Data Consortium* LDC – University of Pennsylvania[6]).
Wśród najważniejszych oferowanych przez niego baz danych można wymienić:

- *Speaker Recognition Evaluation Test-Set*: wspomniana wcześnie baza dla testowania metod rozpoznawania mówców niezależnego od tekstu przez łącza telefoniczne.

[6] http://www.ldc.upenn.edu

- *YOHO Speaker Verification Corpus*[7]: baza dla potrzeb automatycznej weryfikacji mówców. Jakość nagrań w tej bazie jest laboratoryjna zaś jej rozmiar wynosi 1.5 gigabajt (przy częstotliwości próbkowania 8kHz).
- *KING Corpus for Speaker Verification Research*[8]: baza stworzona dla potrzeb automatycznego rozpoznawania mówców w zbiorach zamkniętych i testowania wpływu telefonicznego kanału transmisyjnego na jakość rozpoznawania. Dlatego też dane są podzielone według ich jakości na laboratoryjne oraz telefoniczne.
- *TIMIT*: baza zawierająca szereg wypowiedzi w różnych dialektach amerykańskich języka angielskiego.

Niestety ze względu na kosztowność tych baz nie są one dostatecznie powszechnie stosowane. Warto podkreślić, iż trend tworzenia tych baz idzie śladem badań nad automatycznym rozpoznawaniem mówcy. Z punktu widzenia zastosowania coraz ważniejszą rolę pełnią systemy jednokanałowe w których dane trenujące oraz testujące zarejestrowane są za pośrednictwem takiego samego a nawet tego samego kanału, co stanowi zasadnicze ułatwienie procesu rozpoznawania. Z tych względów najnowsza baza *Speaker Recognition Evaluation Test-Set 98* została stworzona właśnie dla potrzeb weryfikacji jakości jednokanałowych systemów automatycznego rozpoznawania mówcy.

3.6. Subiektywne metody rozpoznawania mówcy

Słuchowe rozpoznawanie mówcy przez człowieka należy do subiektywnych metod rozpoznawania [Bas93]. Badania w tej dziedzinie były przeprowadzone już w 1954 r [Pol54]. Podobnie jak w innych dziedzinach sztucznej inteligencji próbuje się konstruować automatyczne systemy rozpoznawania mówców naśladujące, doskonalsze jak dotąd, systemy biologiczne. Dopiero próba budowy takiego automatycznego systemu może uświadomić badaczom jakie wspaniałe są możliwości realizacji tego zadania przez człowieka [Ros76]. Poza unikatową, ludzką zdolnością do odbioru i dekodowania mowy, system słuchowy zapewnia inne działania funkcjonalne jak np.: lokalizację źródła dźwięku, odczuwanie przyjemności ze słuchania muzyki oraz omawianą w tej pracy zdolność do rozpoznawania osób na podstawie ich głosów. Wymienianie zalet biologicznych systemów rozpoznawania wychodzi jednak poza zakres niniejszej pracy.

[7] Baza stworzona w 1989r przez ITT pod licencją rządu Stanów Zjednoczonych.
[8] Baza stworzona w 1987r przez ITT pod licencją rządu Stanów Zjednoczonych.

Badania nad subiektywnymi metodami rozpoznawania mówcy potrzebne są również w celu określenia poziomu rozsądnych oczekiwań odnośnie automatycznych systemów realizacji tego zadania [Pru66]. Przykładem takich badań jest praca [Pru66] mająca na celu określenie zależności jakości rozpoznawania mówców przez ludzi od długości oraz składu badanej wypowiedzi. Wywnioskowano tam między innymi, że istotniejszy wpływ na jakość rozpoznawania ma liczba różnych fonemów występujących w wypowiedzi aniżeli jej długość. Stwierdzono również, że jakość rozpoznawania opartego na badaniu izolowanej samogłoski jest zróżnicowana dla różnych samogłosek. Innym ciekawym zaobserwowanym zjawiskiem było pogorszenie jakości wyników w przypadku przeprowadzenia rozpoznawania na podstawie wstecznie odtwarzanych wypowiedzi (*reverse playback*).

Należy podkreślić, że na zdolność człowieka do rozpoznawania mówców ogromny wpływ ma nieprzerwany, długoletni trening zarówno jego systemu słuchowego jak i jego złożonego ośrodka klasyfikacyjno–decyzyjnego (mózg). Wspomniany proces treningowy jest w dużej mierze związany z kompleksowym charakterem danych trenujących. Wskazuje na to chociażby przeprowadzony w [Pru66] eksperyment z rozpoznawaniem mówcy na podstawie izolowanych samogłosek. Stwierdzono bowiem, że zdolność człowieka do rozpoznawania znanej mu 10 elementowej – w dodatku zamkniętej – populacji mówców na podstawie krótkich (117 ms) izolowanych samogłosek jest rzędu 50 do 60%. Należy dodać, że wynik rozpoznawania w tych samych warunkach ale dla 2,5 sekundowego zadania wyniósł aż 98%.

W niektórych badaniach ponoć stwierdzono, że niektóre automatyczne systemy rozpoznawania mówcy przewyższały ludzkie zdolności do realizacji tego zadania [Ata74, Ata76, Ros73]. Przykładowo badania prowadzone przez Rosenberga [Ros73] dowiodły, że w zadaniu automatycznej weryfikacji 40 mówców jakość weryfikacji osiągniętej przez ludzi wyniosła 96% zaś przez system automatycznego rozpoznawania wykorzystujący trzech zestawów parametrów aż 98%. Wyniki te wymagają jednak dalszych potwierdzeń i badań.

Oprócz słuchowej metody rozpoznawania mówcy przez człowieka do subiektywnych metod należy również rozpoznawanie mówców na podstawie wizualnej inspekcji spektrogramów ich wypowiedzi [Bol70, Bol73, Kre62]. Temat ten wykracza jednak zdecydowanie poza obszar tej pracy.

3.7. Podsumowanie

Choć dokonano wielkiego postępu w dziedzinie badań nad automatycznym rozpoznawaniem mówcy, pozostało jeszcze wiele problemów czekających na lepsze rozwiązania. Większość z nich wynika ze zmienności warunków rozpoznawania spowodowanej zarówno przez zmiany w warunkach rejestracji oraz zmienne charakterystyki kanału jak i jest następstwem zmian spowodowanych przez samych mówców. Bardzo ważnym zagadnieniem jest zatem znalezienie parametrów stabilnych w dłuższych interwałach czasowych i w miarę niezależnych od zmian w sposobie mówienia (np. tempo i głośność mowy). Potrzebne są także nowe metody rozpoznawania odporne na zakłócenia towarzyszące kanałowi transmisyjnemu sygnału mowy przez łącza telefoniczne oraz niepodatne na przeciętny poziom szumu środowiska rejestracji sygnału.

Systemy automatycznego rozpoznawania mówcy lub mowy są bardzo wymagające pod względem szybkości oraz rozmiaru pamięci komputera realizującego tego zadania. Dlatego też znaczący postęp w badaniach nad tym zagadnieniem zawdzięczamy głównie burzliwemu postępowi techniki komputerowej. Tym niemniej trzeba obiektywnie stwierdzić, że nawet mimo stosowania najnowszej technologii w ostatnich latach nie dokonano znacznego postępu w określeniu indywidualnych charakterystyk głosu mówcy [Fur96]. Tymczasem można sądzić, że właśnie bliższe zapoznanie się z tymi charakterystykami zaowocuje stworzeniem bardziej skutecznych metod rozpoznawania mówcy. Oprócz tego w celu umożliwienia rzetelnego porównywania różnorakich metod rozpoznawania mówcy zachodzi konieczność rozpowszechniania dostępu do różnych baz danych, co nie jest łatwe.

ROZDZIAŁ IV

Reprezentacja sygnału mowy za pomocą histogramów amplitudowych

Niniejszy rozdział poświęcono zagadnieniu reprezentacji sygnału za pomocą histogramu amplitud sygnału mowy będącej sercem zaproponowanej metody rozpoznawania mówcy.

4.1. Reprezentacja sygnału mowy za pomocą histogramów amplitudowych

Zgodnie z twierdzeniem Kotielnikowa-Shannona [Sza90] sygnał analogowy o ograniczonej do F_g górnej częstotliwości można reprezentować z pełną dokładnością za pomocą przeliczalnej liczby próbek uzyskanych w wyniku równomiernego próbkowania z częstotliwością $F_p \geq 2.F_g$. Dowód tego twierdzenia opiera się na analizie widma próbkowanej postaci sygnału. Teoretycznie rzecz biorąc, postać analogowa sygnału można zrekonstruować na podstawie ciągu próbek sygnału zgodnie z następującym wzorem[Sza90]:

$$x(t) = \sum_{n=-\infty}^{\infty} x(nT_p) \frac{\sin \omega_g (t - nT_p)}{\omega_g (t - nT_p)} \qquad (4.1)$$

gdzie: $\omega_g = 2\pi F_g$: górna pulsacja graniczna sygnału,

$T_p = 1/F_p$: kwant próbkowania.

$x(nT_p)$: wartość sygnału w chwili $t = nT_p$ (wartość próbki o numerze n).

Zaproponowany opis sygnału mowy za pomocą histogramów amplitudowych polega na uproszczeniu reprezentacji ciągu jego próbek $x(nT_p)$. Uproszczenia tego dokonujemy posługując się opisem histogramowym. Należy zaznaczyć, iż histogramowy opis sygnału mowy jest częścią obszeniejszych badań podjętych przez autora i zmierzających do szerszego wykorzystania reprezentacji tego sygnału w dziedzinie czasu. W szczególności zaproponowano uproszczony sposób kodowania sygnału mowy za pomocą jego ekstremów czasowych [Sho99c] dla celów kompresji oraz automatycznego rozpoznawania mowy. Ponadto podjęto zakończoną powodzeniem próbę rozpoznawania ograniczonego słownika wyrazów na podstawie opisu histogramowego oraz algorytmu *DTW*.

4.1.1. Histogram amplitudowy sygnału

Niech *x(n)* będzie *N* elementowym sygnałem cyfrowym[1] o amplitudach z przedziału $\pm X_{max}$. Niech wektor poziomów amplitud *v(i)* przyjętych przy sporządzaniu histogramu o długości *2p+1* spełnia następujące założenia:

1. $i \in (-p,p)$
2. $v(p) \leq X_{max}$
3. $v(0)=0$
4. $v(i+1) \geq v(i) \ \forall i$
5. $v(-i)=-v(i) \ \forall i$

Niech ponadto wektor amplitud granicznych *U(i)* o długości *2(p+1)* będzie zdefiniowany na podstawie wektora poziomów amplitud *v(i)* w następujący sposób:

1. $U(1)=- X_{max}$
2. $U(i) = \dfrac{v(-p+i-2)+v(-p+i-1)}{2} \qquad dla \quad i = 2,3,...,p+1$
3. $U(i)=-U(2p+3-i) \quad dla \quad i=p+2,p+3,...,2p+1$
4. $U(2p+2)= X_{max}$

Zbiór wartości próbek sygnału *x(n)* będzie podzielony na następujące *2P+1* podzbiory Z_i w następujący sposób:

$$Z_i=\{x(n): U(i+1) > x(n) \geq U(i)\} \ ; \ i=1,2,3,...,2p+1 \qquad (4.2)$$

Jeżeli liczebność zbioru będzie oznaczona symbolem #, to histogram amplitudowy h_x^v sygnału *x(n)* sporządzony na podstawie wektora amplitud *v* jest dany wzorem:

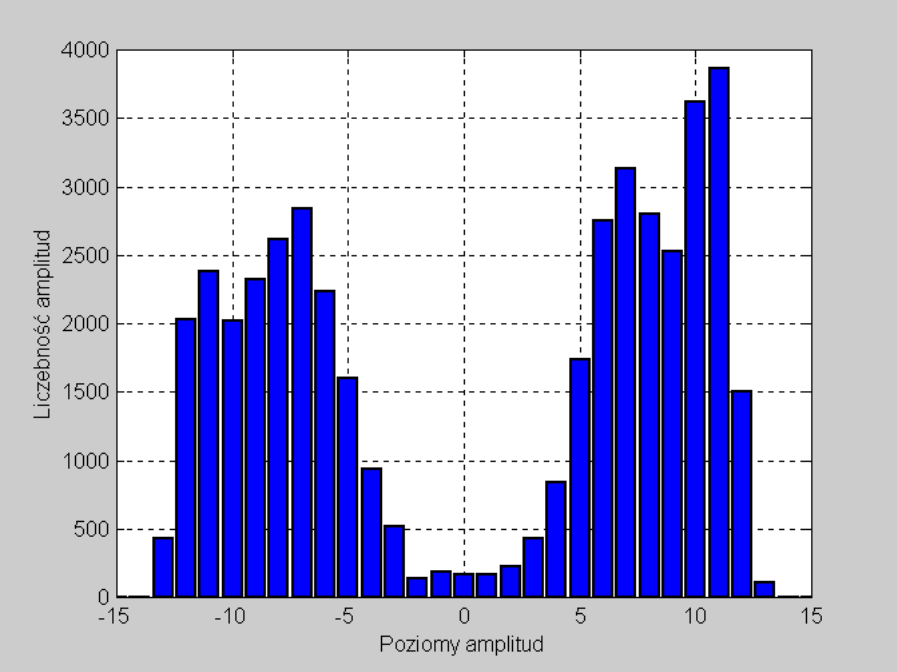
$$h_x^v = (\#Z_1, \#Z_2, \#Z_3, ..., \#Z_{2p+1}) \qquad (4.3)$$

Ponieważ sygnał *x(n)* składa się z *N* próbek to zachodzi następujący związek:

$$\sum_{i=1}^{2p+1} \#Z_i = N \qquad (4.4)$$

[1] Dyskretnym w dziedzinie czasu oraz amplitudy.

Rysunek 4.1 przedstawia przykładowy histogram amplitudowy fragmentu wypowiedzi o długości 2 sekund (głos damski), nagranej w komorze bezechowej za pomocą wysokiej jakości aparatury rejestrującej[2]. Częstotliwość próbkowania wyniosła 22kHz zaś rozdzielczość amplitudowa 16 bitów.



Rys. 4.1: Histogram amplitudowy dla wypowiedzi o długości 2 sekundy.

Histogram wykonano dla zakresu amplitud rozdzielonego symetrycznie na 31 poziomów ($p=15$) w sposób logarytmiczny. Tabela 4.1 przedstawia użyte poziomy dyskryminacji. Ponieważ wektor amplitud v jest symetryczny względem zera przedstawiono wartości tylko dla dodatnich i.

i	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$v(i)$	0	1	2	4	8	16	32	64	128	256	512	1024	2048	4096	8192	1638

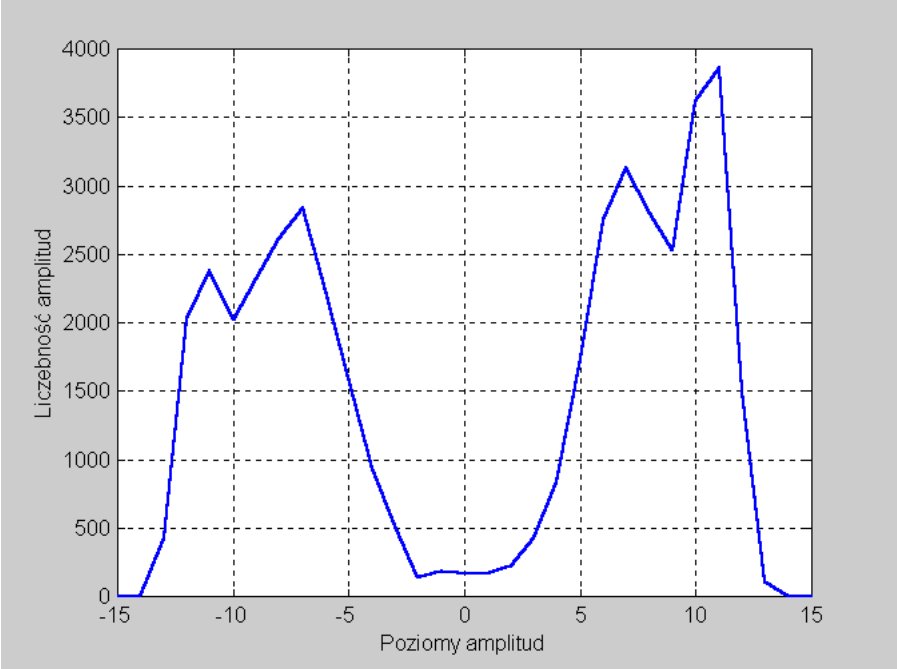
Tabela 4.1: Poziomy amplitud histogramu z rysunku 4.1.

Taki podział zakresu amplitud będzie wykorzystany do obliczenia większości przedstawionych w tym rozdziale histogramów. Należy podkreślić, że oczywisty związek

[2] Sygnał ten został wykorzystany także w innych badaniach opisanych w dalszej części niniejszego rozdziału.

powyższych liczb z kodowaniem binarnym stanowi ułatwienie dla ewentualnej realizacji sprzętowej (bądź programowej, w języku maszynowym) wykonania procesora histogramów. Zagadnieniu temu poświęcono podrozdział 4.5.

Na rysunku 4.2 przedstawiono histogram z rysunku 4.1 w postaci wykresu. Taka forma przedstawiania histogramu będzie częściej używana w niniejszej pracy. Wbrew pozorom histogram amplitudowy nie jest symetryczny względem poziomu zerowego mimo, że z sygnału wejściowego usunięto składową stałą (jeśli ewentualnie była) przed obliczeniem histogramu. Problem ten będzie dyskutowany w punkcie 4.1.6. We wszystkich przedstawionych w tej pracy histogramach sygnały wejściowe zostały pozbawione składowej stałej. Ponadto w niektórych przypadkach dla celów ilustracyjnych energia sygnału wejściowego została zmieniona (sygnał został wzmocniony bądź stłumiony).



Rys. 4.2: Histogram amplitudowy przedstawiony w formie wykresu.

4.1.2. Unormowanie czasowe histogramu

Ze względu na możliwość wykonania histogramu dla wypowiedzi o różnych czasach trwania zachodzi potrzeby unormowania czasowego histogramu. W tym celu liczebność odpowiednich amplitud w sygnale może być reprezentowana w procentach

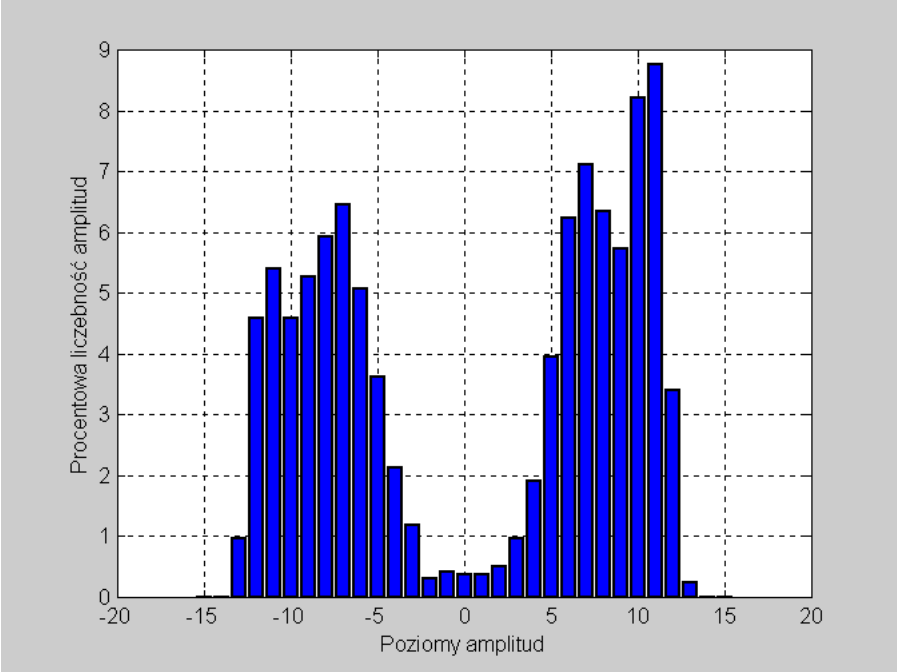
łącznej liczby próbek sygnału. Zatem jeśli pierwotny sygnał ma długość N próbek to jego unormowany czasowo histogram h_{unorm} jest dany wzorem:

$$h_{unorm}(i) = 100\frac{h(i)}{N}\% \qquad \forall i \tag{4.5a}$$

Ogólniej wzór 4.5a ma następującą postać:

$$h_{unorm}(i) = C\frac{h(i)}{N} \qquad \forall i \tag{4.5b}$$

gdzie C to długość histogramu unormowanego wyrażona jako liczba próbek sygnału. Rysunek 4.3 przedstawia unormowany czasowo histogram z rysunku 4.1.



Rys. 4.3: Unormowany czasowo histogram amplitudowy.

Chociaż kształt histogramu nie zmienia się w wyniku unormowania czasowego, jednak operacja ta ma zasadnicze znaczenie, bowiem uniezależnia metrykę (miarę odległości między histogramami) od długości czasowej sygnałów. Unormowanie czasowe histogramu jest również potrzebne w przypadku operowania różnymi częstotliwościami próbkowania, bowiem liczba próbek sygnału o ustalonej długości czasowej jest wprost proporcjonalna do częstotliwości jego próbkowania.

4.1.3. Addytywność histogramów

Niech sygnał *x(n)* będzie podzielony na dwa fragmenty o różnych długościach. Wówczas histogram wypadkowy sygnału jest sumą histogramów obliczanych dla obu jego fragmentów. Związek ten jest spełniony wyłącznie dla nie unormowanych histogramów. W przypadku histogramów unormowanych należy dokonać odpowiedniego skalowania histogramów w zależności od stosunku długości czasowej reprezentowanych przez nie fragmentów sygnału. Gdy długość tych fragmentów jest równa, unormowany histogram wypadkowy jest równy średniej wartości obu histogramów. Związek ten jest również spełniony w przypadku większej liczby fragmentów sygnału pod warunkiem, że ich długość jest jednakowa.

4.1.4. Rozdzielczość histogramu – histogramy logarytmiczne oraz liniowe

Zakres amplitud sygnału można podzielić ogólnie na różne przedziały, przy czym mogą one być rozłożone w sposób logarytmiczny lub liniowy. W obu przypadkach istotny wpływ na kształt histogramu ma liczba przedziałów. Ze względu na dużą dynamikę sygnału mowy, logarytmiczny sposób podzielenia zakresu amplitud pozwala stworzyć wyraźniejszy obraz sygnału. Dla histogramu liniowego wektor poziomów amplitud *v(i)* histogramu ma następującą postać:

$$v(i) = \frac{X_{max}}{p} i \quad ; \quad i \in (-p,p) \tag{4.6}$$

Natomiast wektor amplitud dla histogramu logarytmicznego jest dany wzorem:

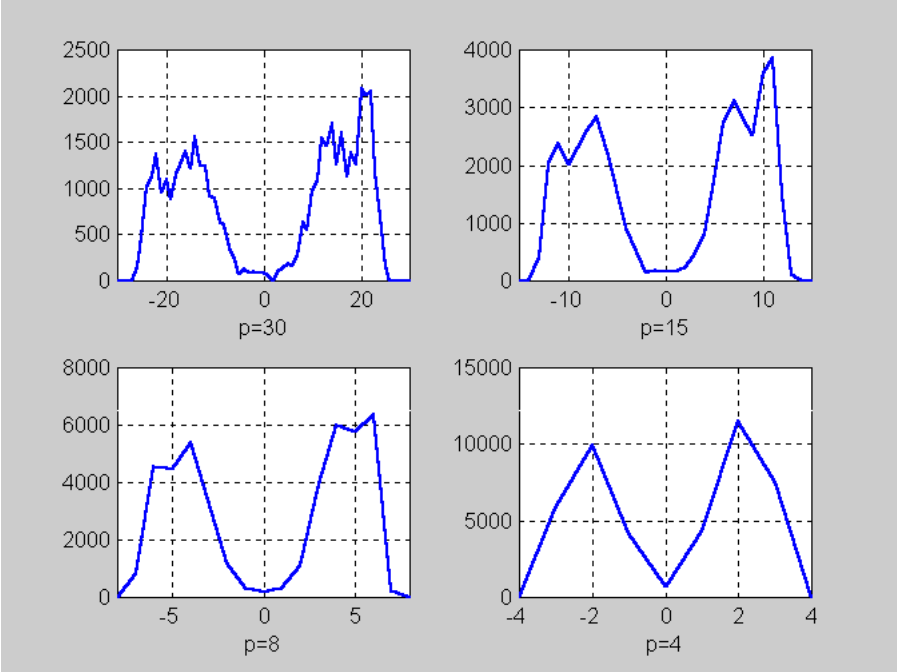
$$v(i) = exp\left[\frac{ln(X_{max})}{p} i \right] \qquad dla \quad i > 1 \tag{4.7}$$

Wartości wektora *v* dla $i \leq 0$ można obliczyć korzystając z założenia 5 (patrz 4.1.1).

W celu przedstawienia wpływu rozdzielczości amplitudowej histogramu na jego kształt na rysunku 4.4 przedstawiono histogram z rysunku 4.1 dla czterech rozkładów poziomów amplitud (różne rozdzielczości). We wszystkich przedstawionych przypadkach zakres został podzielony w sposób logarytmiczny. Użyto następujących wektorów poziomów amplitud[3]:

[3] Ponieważ wektory amplitud są symetryczne względem zera przedstawiono tylko ich dodatnie współrzędne.

- V_1: 0, 1, 2, 3, 4, 6, 8, 12, 16, 28, 32, 56, 64, 112, 128, 224, 256, 447, 512, 891, 1024, 1778, 2048, 3548, 4096, 7079, 8192, 14125, 16384, 28184, 32768
- V_2: 0, 1, 2, 4, 8, 16, 32, 64, 128, 256, 512, 1024, 2048, 4096, 8192, 16384, 32768
- V_3: 0, 2, 8, 32, 128, 512, 2048, 8192, 32768
- V_4: 0, 8, 128, 2048, 32768

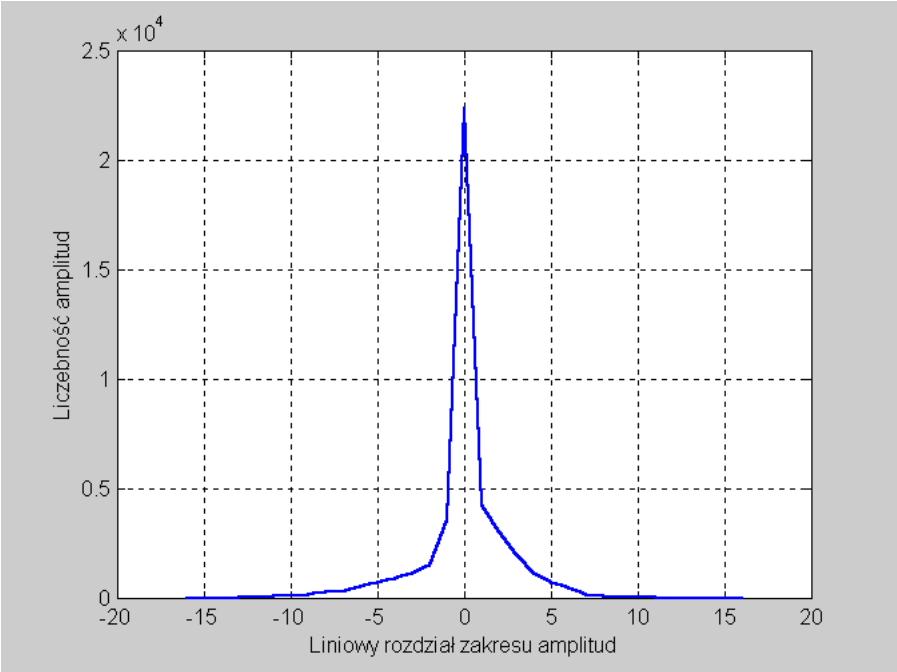


Rys. 4.4: Wpływ rozdzielczości histogramu na jego kształt.

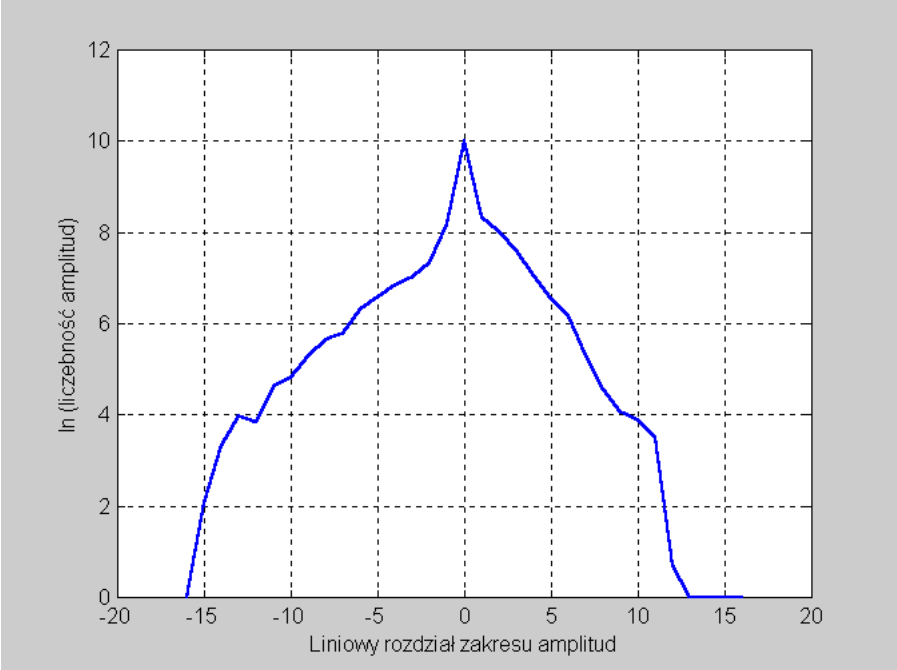
Do celów porównawczych na rysunku 4.5 przedstawiono histogram z rysunku 4.1 dla przypadku liniowego podziału zakresu amplitud. Widać na nim, że przy liniowym podziale nastąpiło pogorszenie obrazu (mniej szczegółów) ze względu na „scalenie się" informacji o małych amplitudach, które z naturalnych względów stanowią większość amplitud występujących w sygnale. Dlatego też bardziej interesujący jest kształt histogramu zlogarytmowanego (patrz rysunek 4.6) zdefiniowanego według następującego wzoru :

$$h_{log}(i)=log\ [h(i)+1] \quad lub \quad h_{ln}(i)=ln\ [h(i)+1] \tag{4.8}$$

Jedynkę we wzorze 4.7 dodano aby zapobiec operacji logarytmowania wartości zerowych.



Rys. 4.5: Kształt histogramu z liniowym podziałem zakresu amplitud.

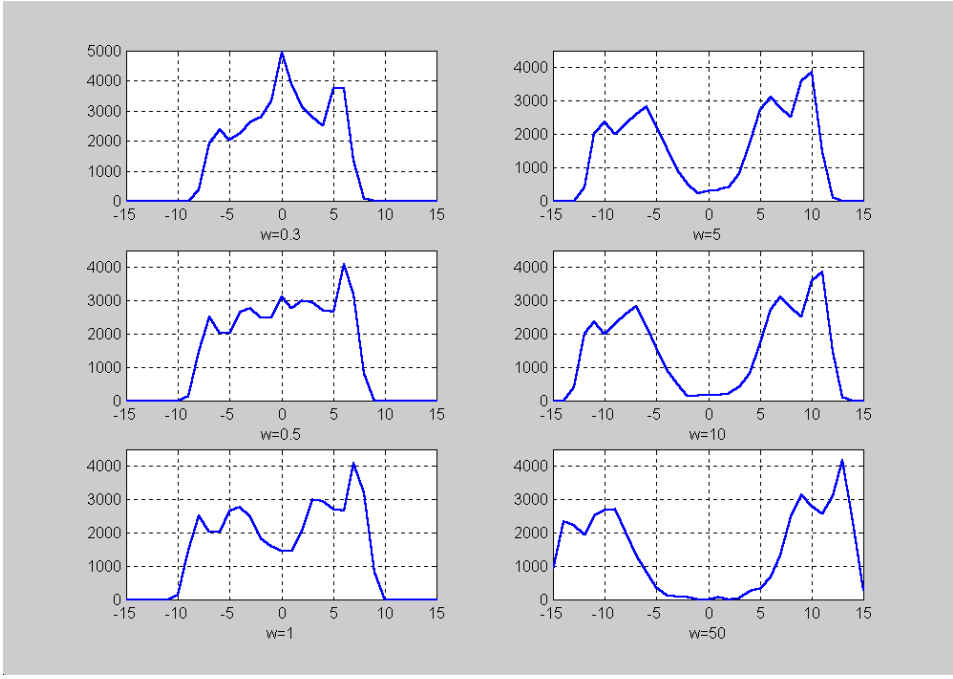


Rys. 4.6: Kształt histogramu z liniowym zakresem amplitud i skalą logarytmiczną.

Oczywistym jest, że w praktyce wybierając rozdzielczość amplitudową histogramu (liczbę poziomów *2p+1* oraz wektor poziomów amplitud *v*) bierze się pod uwagę rozdzielczość próbkowania samego sygnału.

4.1.5. Zależność histogramu od energii sygnału

W odróżnieniu od niektórych używanych w literaturze parametrów opisu sygnału mowy, zaproponowany w tej pracy opis histogramowy zależy bezpośrednio od poziomu mocy sygnału. W celu zbadania zależności kształtu histogramu od energii sygnału na rysunku 4.7 przedstawiono kilka wersji histogramu z rysunku 4.1 odpowiadających różnym poziomom wzmocnienia (energii) sygnału. Zmienna *w* na rysunku określa względną wartość wzmocnienia sygnału.



Rys. 4.7: Zależność kształtu histogramu od energii sygnału.

W celu uniezależnienie kształtu histogramu od poziomu energii sygnału można unormować jego amplitudy np. względem średniej wartości ich modułu. W praktyce oznacza to wzmocnienie bądź stłumienie sygnału tak, aby jego energia była równa z góry zakładanej wartości. Sygnał *x(n)* o długości *N* próbek, unormowany względem średniej wartości jego energii jest dany następującym wzorem:

$$x_{u1}(n) = c_1 \frac{N.x(n)}{\sum_{n=1}^{N}\left[x(n)\right]^2} \tag{4.9}$$

gdzie stała c_1 określa energię sygnału $x_{u1}(n)$. W przypadku unormowania względem średniej wartości modułu sygnału jego unormowana postać jest dana następującym wzorem:

$$x_{u2}(n) = c_2 . \frac{N.x(n)}{\sum_{n=1}^{N}\left|x(n)\right|} \tag{4.10}$$

Innym rozwiązaniem tego problemu może być unormowanie wartości wektora poziomów amplitud v np. względem średniej wartości modułu amplitud sygnału lub jego średniej energii.

4.1.6. Zależność histogramu od fazy sygnału

Jak już wspomniano w punkcie 4.1.1. histogramy amplitudowe nie są symetryczne względem poziomu zerowego. W niektórych przypadkach zachodzi konieczność uniezależnienia histogramu od fazy sygnału (np. w celu porównywania dwóch sygnałów z których jeden ma odwróconą fazę względem drugiego skutkiem 'przejścia' przez inny kanał posiadający wzmacniacz odwracający fazę o 180 stopni). Ponieważ badania wykazują, że dla zdecydowanej większości histogramów istnieje globalne maksimum liczebności amplitud zależne od fazy sygnału, dlatego zaproponowano odzwierciedlenie histogramów względem poziomu zerowego tak, aby ich maksima znalazły się zawsze po tej samej strony (po stronie dodatnich bądź ujemnych amplitud). Ze względu na prostotę operacji: wykrywania maksimum oraz zwierciadlanego odbicia histogramu względem zera nie przedstawiono tutaj ich matematycznego opisu.

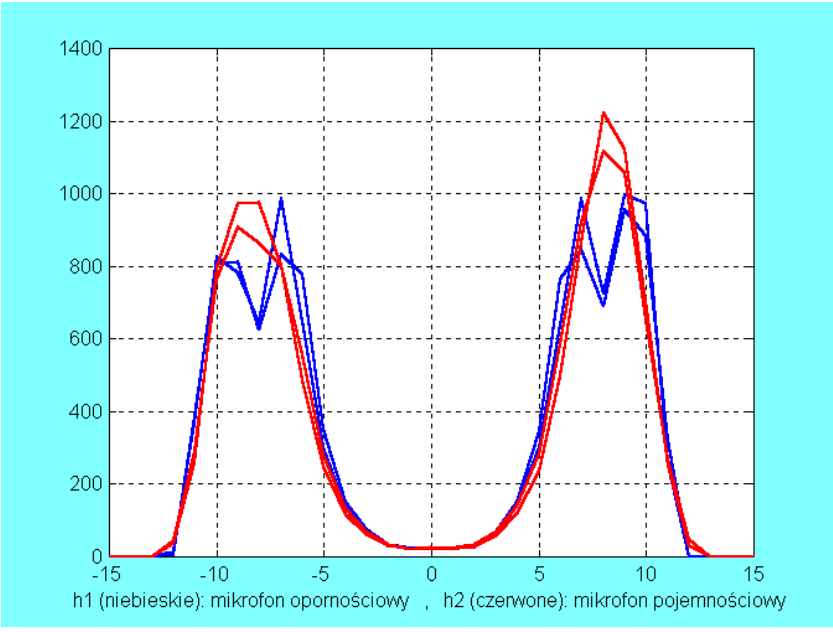
4.1.7. Wpływ dynamiki mikrofonu na kształt histogramu

Badania wykazały, że dynamika użytego mikrofonu ma znaczący wpływ na kształt histogramu. W szczególności zauważono, że histogramy sygnału mowy zarejestrowanego w warunkach laboratoryjnych za pomocą wysokiej klasy aparatury i czułego mikrofonu o dużej dynamice posiadają na ogół cztery wierzchołki (dwa dla dodatnich amplitud oraz dwa dla ujemnych). Natomiast histogramy sygnału zarejestrowanego w warunkach pokojowych za pomocą karty multimedialnej komputera PC oraz mikrofonu pojemnościowego ogólnego przeznaczenia posiadają z reguły tylko dwa

stosunkowo wyraźne wierzchołki. W celu ilustracji wpływu mikrofonu na kształt histogramu przeprowadzono następujące eksperyment. Nagrano dwie wypowiedzi o długości 5 sekund w warunkach pokojowych za pomocą karty multimedialnej komputera PC oraz mikrofonu pojemnościowego ogólnego przeznaczenia. Nagrania powtórzono w tej samej sesji po zmianie mikrofonu na opornościowy. W tabelce 4.2 podano dane techniczne użytych mikrofonów. Obliczono cztery histogramy odpowiadające zarejestrowanym wypowiedziom. Rysunek 4.8 przedstawia otrzymane wyniki. Widać, że histogramy niebieskie odpowiadające nagraniom z mikrofonem opornościowym o większej dynamice posiadają cztery wierzchołki, zaś histogramy czerwone odpowiadające nagraniom z mikrofonem pojemnościowym o mniejszej dynamice posiadają w zasadzie jeden wierzchołek. Świadczy to o mniejszej precyzji przenoszenia szczegółów sygnału w przypadku użycia gorszego mikrofonu.

Producent	Nazwa komercyjna	Rodzaj	Czułość	Impedancja	Zakres częstotliwości
TARGET Sound line	Internet microphone	Pojemnościowy	42 dB 0 dB =1 Pa, 1 kHz	2000 Ω	20 Hz – 20 kHz
ALGA Co. Ltd. Japan	Dynamic Microphone	Opornościowy	67 dB ± 3 dB dla 1 kHz	600 Ω	80 Hz – 12 kHz

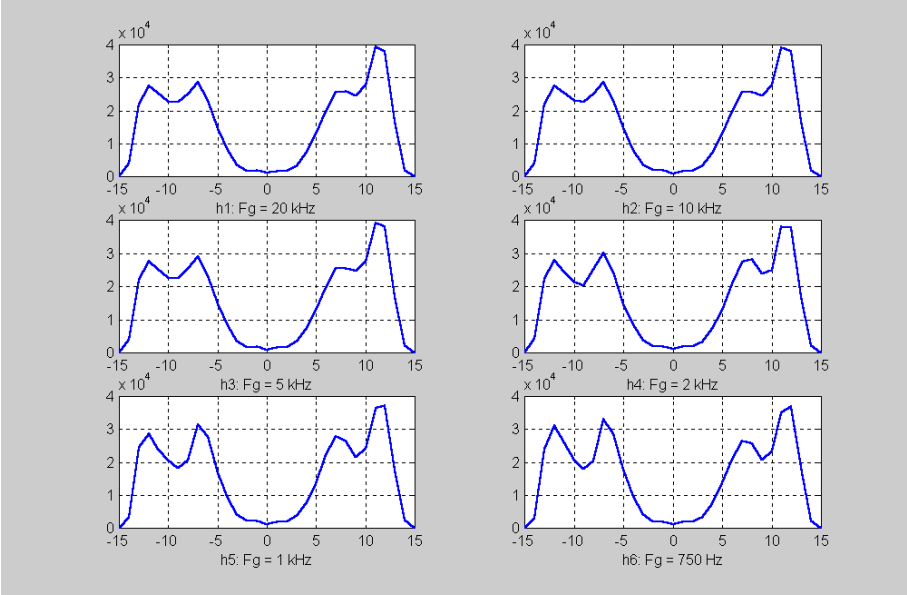
Tabela 4.2: Dane techniczne użytych w doświadczeniu mikrofonów.



Rys. 4.8: Wpływ rodzaju użytego do nagrania mikrofonu na kształt histogramu sygnału.

4.1.8. Związek między pasmem przenoszenia sygnału a jego histogramem

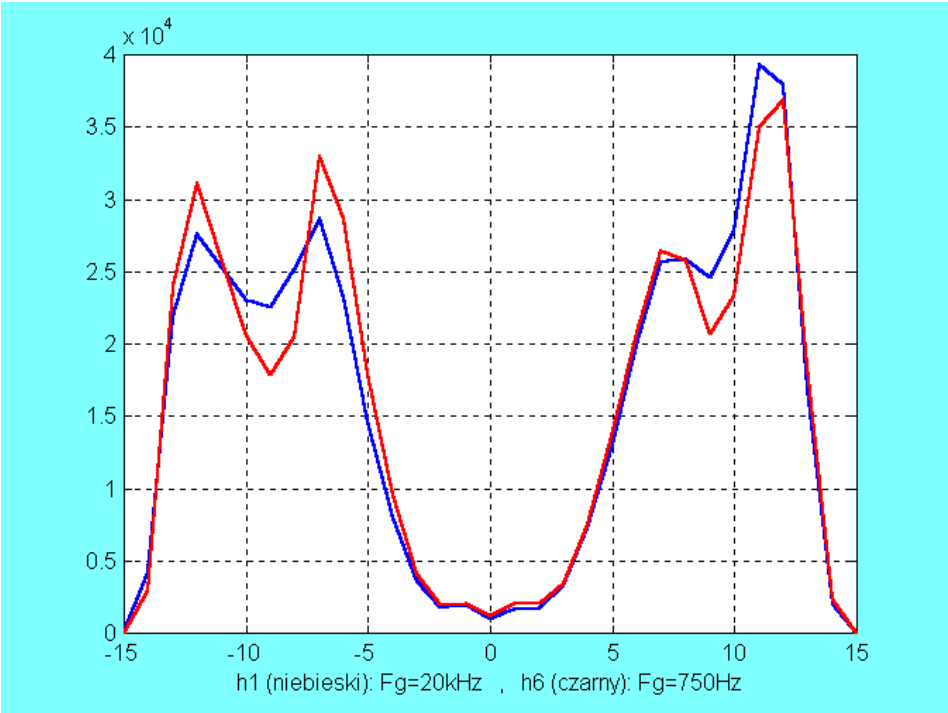
Jak już podkreślono w rozdziale drugim opis sygnału mowy w dziedzinie częstotliwości jest jedną z najważniejszych form jego reprezentacji. Warto zatem zbadać związek pomiędzy kształtem histogramu sygnału a jego pasmem przenoszenia. Niestety nie istnieje teoretyczny, prosty sposób powiązania składu częstotliwościowego sygnału z kształtem jego histogramu. Jednak w celu doświadczalnego zbadania tego związku wykonano kilka histogramów dla jednego sygnału[4] po ograniczeniu jego widma kolejno do następujących częstotliwości górnych: 20kHz, 10kHz, 5kHz, 2kHz, 1kHz oraz 750Hz, przy czym częstotliwość próbkowania sygnału była wspólna dla wszystkich wersji sygnału i wynosiła 44kHz. Rysunek 4.9 przedstawia otrzymane histogramy.



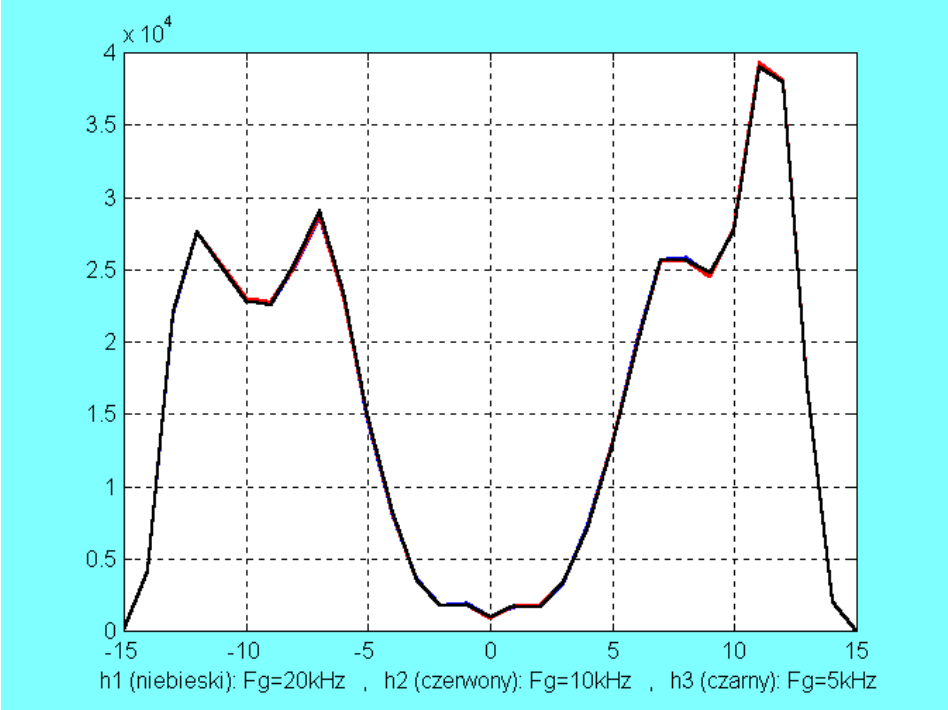
Rys. 4.9: Zależność kształtu histogramu od górnej częstotliwości granicznej widma sygnału.

Bardziej zauważalne zmiany w kształcie histogramu można dostrzec między histogramami h1 oraz h6 odpowiadającymi dwóm wersjom sygnału o znacznie różniących się częstotliwościach granicznych widma. W celu lepszej wizualizacji na rysunku 4.10a przedstawiono te dwa histogramy razem. Histogramy h1, h2 oraz h3 odpowiadające kolejnym częstotliwościom granicznym 20kHz, 10kHz oraz 5kHz nie wykazują w zasadzie istotnych różnic, co widać na rysunku 4.10b przedstawiającym je razem.

[4] Jest to ten sam sygnał wykorzystany w punkcie 4.1.1.

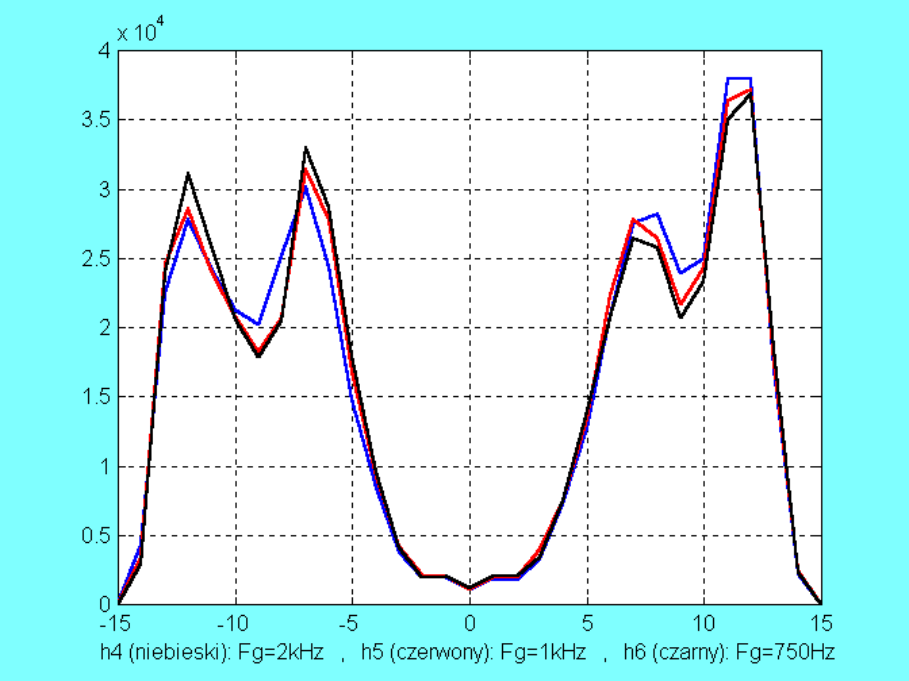


Rys. 4.10a: Porównanie histogramów h1 i h6 z rysunku 4.8.



Rys. 4.10b: Porównanie histogramów h1, h2 oraz h3 odpowiadających częstotliwościom granicznym 20, 10 oraz 5kHz.

Z drugiej strony histogramy h4, h5 oraz h6 odpowiadające częstotliwościom granicznym 2kHz, 1kHz oraz 750Hz wykazują względnie więcej różnic w kształcie zarówno między sobą jak i w stosunku do histogramów h1, h2 i h3. Do celów porównawczych przedstawiono te pierwsze razem na rysunku 4.10c. Z przedstawionych badań wynika, że histogramy amplitudowe generalnie lepiej zachowują informacje o niższych częstotliwościach. Filtracja tych wyższych nie spowodowała bowiem zauważalnych zmian w kształcie histogramów.

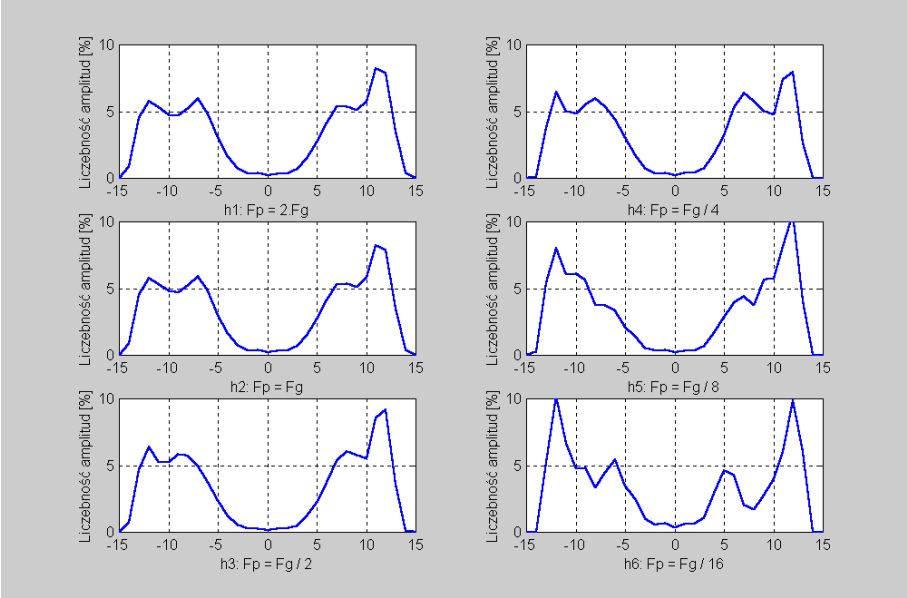


Rys. 4.10c: Porównanie histogramów h4, h5 oraz h6 odpowiadających częstotliwościom granicznym 2kHz, 1kHz oraz 750Hz.

Ponieważ ograniczeniu pasma przenoszenia sygnału towarzyszy często obniżanie jego częstotliwości próbkowania[5], podjęto również próbę zbadania wpływu obniżenia częstotliwości próbkowania na kształt histogramu. Na podstawie wykonanych badań można stwierdzić, że o ile spełniony jest warunek $F_p \geq 2.F_g$ wynikający z twierdzenia Kotielnikowa-Shannona, to kształt histogramu nie zależy w sposób istotny od częstotliwości próbkowania F_p. W przypadku nie spełnienia warunku $F_p \geq 2.F_g$ badania wykazały, że histogramy amplitudowy w pewnych granicach nadal zachowują duży

[5] zgodnie z twierdzeniem o próbkowaniu Kotielnikowa-Shannona.

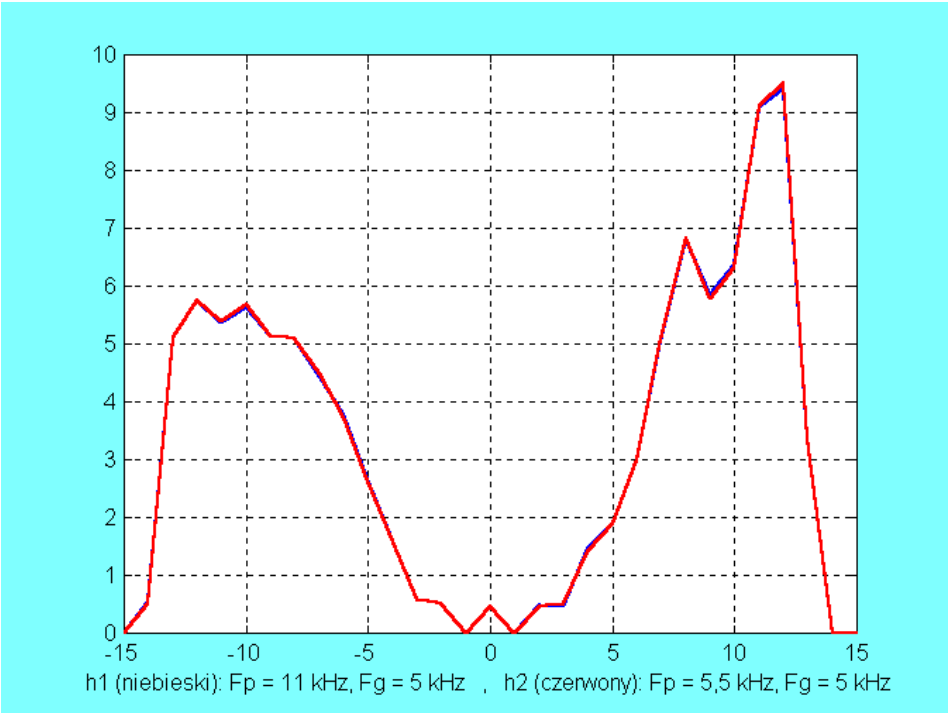
stopień niezależności od częstotliwości próbkowania F_p. Rysunek 4.11 przedstawia histogramy czterech wersji sygnału o F_g=*20kHz* (ten sam sygnał o histogramie *h1*) po jego próbkowaniu – bez wcześniejszego ograniczenia F_g – z sześcioma różnymi częstotliwościami: 44kHz, 22kHz, 11kHz, 5,5kHz, 2,75kHz oraz 1375Hz odpowiadające w przybliżeniu kolejno: $2F_g$, F_g, $F_g/2$, $F_g/4$, $F_g/8$ oraz $F_g/16$. Ponieważ liczba próbek w sygnale jest proporcjonalna do F_p dokonano unormowania czasowego powyższych histogramów w celu umożliwienia ich porównania.



Rys. 4.11: Wpływ obniżenia F_p próbkowania sygnału do wartości niższej niż $2F_g$ na kształt histogramu.

Generalnie rzecz biorąc można stwierdzić, że obniżenie częstotliwości próbkowania sygnału F_p do wartości równej nawet $F_g/4$ nie niesie ze sobą względnie dużych zmian w kształcie jego histogramu. Dalsze szczegółowe badania wykazały, że kształty histogramów odpowiadających częstotliwościom $F_p \cong 2F_g$ oraz $F_p \cong F_g$ nie wykazują w zasadzie żadnych różnic. Wniosek ten nieco zaskakujący w świetle znanych twierdzeń o efektach zwierciadlanych występujących w zbyt wolno próbkowanych sygnałach został sprawdzony dla kilku mówców różnej płci, dla różnej długości wypowiedzi (rzędu kilku sekund) oraz dla częstotliwości próbkowania F_p równej 44kHz, 22kHz, 11kHz oraz 5,5kHz. Przykład dwóch takich histogramów dla wypowiedzi o długości 5 sekund (głos

damski nagrany w warunkach laboratoryjnych) i F_p=*11kHz* przedstawiono na rysunku 4.12. Widać na nim, że oba histogramy nieomal się pokrywają.

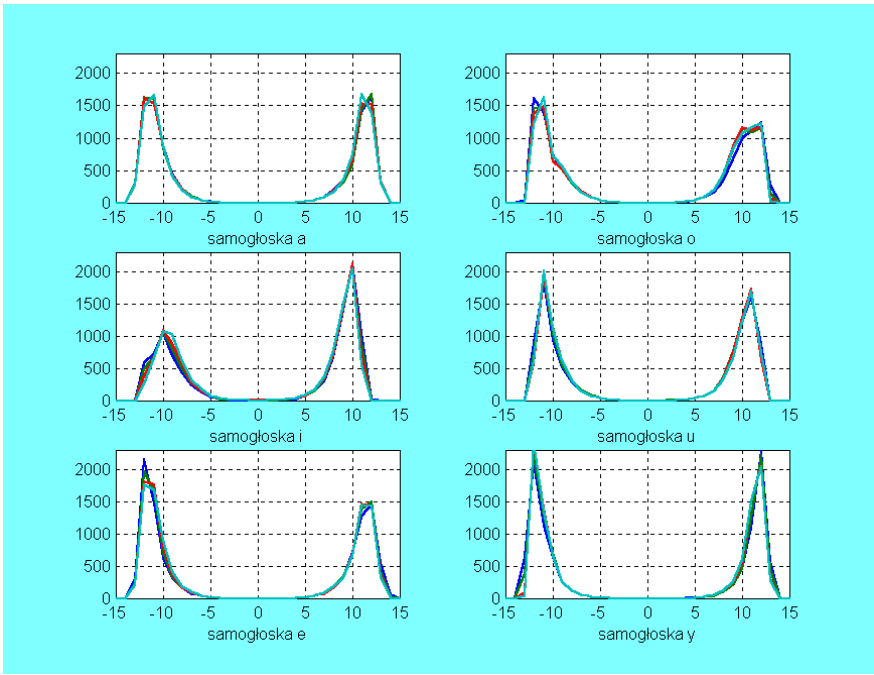


Rys. 4.12: Pokrywanie się histogramów tego samego sygnału próbkowanego z dwoma różnymi częstotliwościami: $F_p \cong 2F_g$ oraz $F_p \cong F_g$.

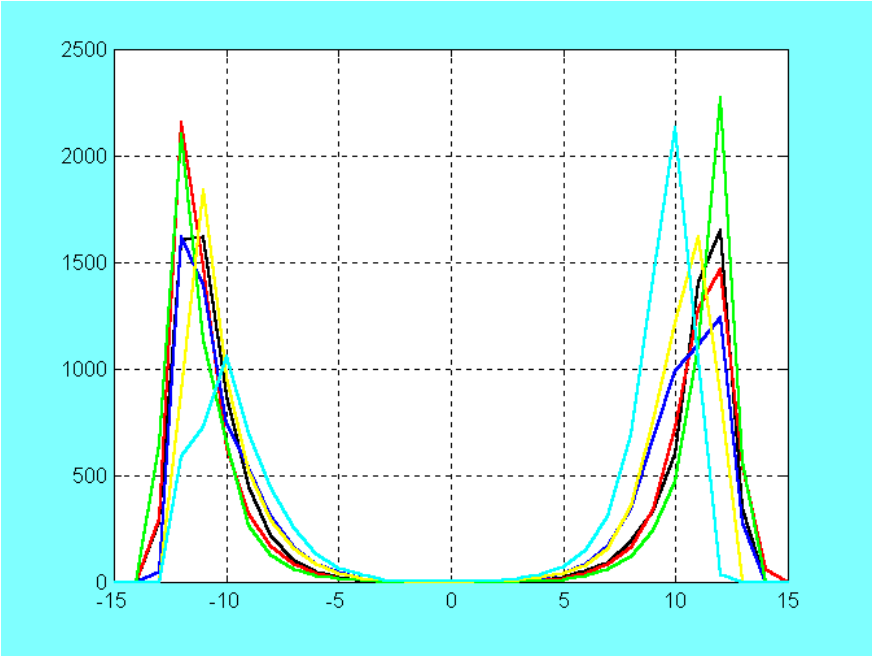
4.2. Przykładowe histogramy wybranych głosek języka polskiego

W nawiązaniu do przedstawionych powyżej badań w podrozdziale tym podjęto próbę zbadania kształtu histogramu dla kilku głosek języka polskiego w stanie ustalonym. Do badania wybrano następujące samogłoski: ‘a’, ‘y’, ‘e’, ‘i’, ‘o’ i ‘u’, głoski szumowe ‘s’ i ‘sz’ , nosowe ‘n’ i ‘m’ oraz boczną ‘l’. Głoski te nagrano głosem damskim i próbkowano z częstotliwością 44kHz używając 16 bitowej rozdzielczości amplitudowej. Nagrania zarejestrowano w warunkach pokojowych z użyciem uniwersalnego mikrofonu pojemnościowego i karty dźwiękowej komputera PC. Rysunek 4.13a przedstawia histogramy wybranych samogłosek (4 histogramy dla każdej głoski) zaś na rysunku 4.13b porównywano kształt tych histogramów (po jednym dla każdej samogłoski). Należy podkreślić, iż występujące ‘scalenie się’ wierzchołków histogramów w stosunku do tych wcześniej przedstawionych jest przede wszystkim wynikiem małej dynamiki użytego mikrofonu. Wykazują na to także inne badania podjęte w ramach tej pracy. Ponieważ warunki nagrania były wspólne dla wszystkich głosek to nie

przeprowadzono tutaj operacji unormowania energii (sprowadzenia energii głosek do tego samego poziomu) ani też ewentualnego zwierciadlanego odbicia histogramów względem poziomu zerowego.



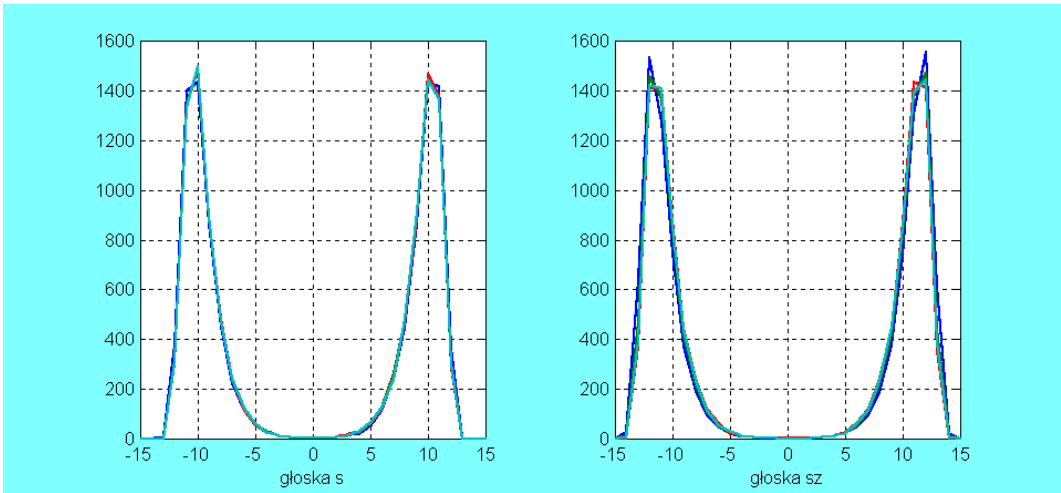
Rys. 4.13a: Histogramy różnych samogłosek języka polskiego.



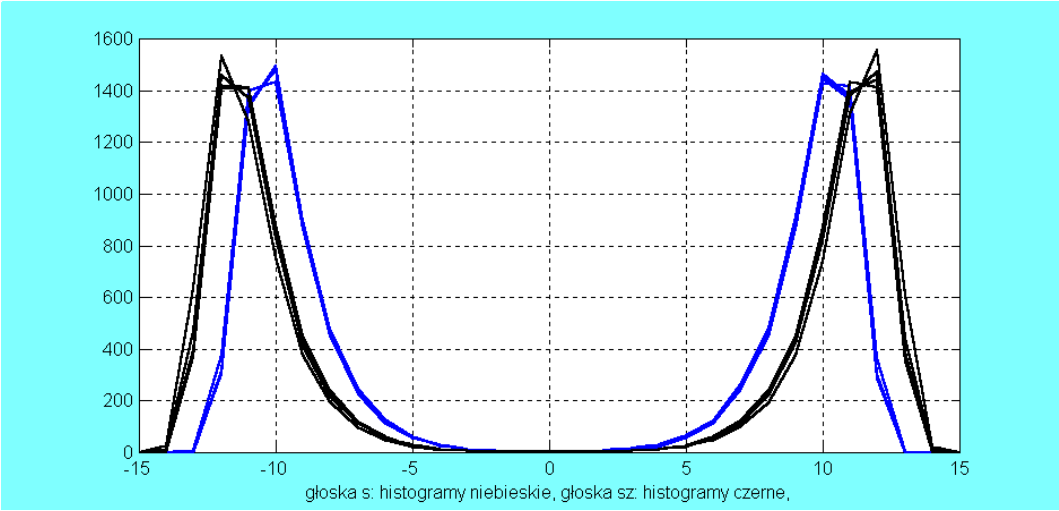
Rys. 4.13b: Porównanie histogramów samogłosek języka polskiego.

a: czarny, e: czerwony, o: niebieski, y: zielony, u: żółty, i: błękitny.

Rysunek 4.14a przedstawia histogramy głosek szumowych ‘s’ oraz ‘sz’ (cztery na każdym wykresie). Biorąc pod uwagę, że widmo tych głosek jest skupione wokół wyższych częstotliwości oraz fakt, że histogramy amplitudowy są stosunkowo mało czułe na zmiany składowych widma dla tych częstotliwości można było przewidzieć, że różnice w kształcie histogramów głosek ‘s’ i ‘sz’ będą raczej nieznaczne. Jednak ze względu na różne wartości energii tych głosek występuje różnica w położeniu wierzchołków ich histogramów, co przedstawiono na rysunku 4.14b. Należy wspomnieć, iż histogramy te mogłby się pokrywać gdyby przeprowadzono operację wyrównania poziomu energii sygnałów.

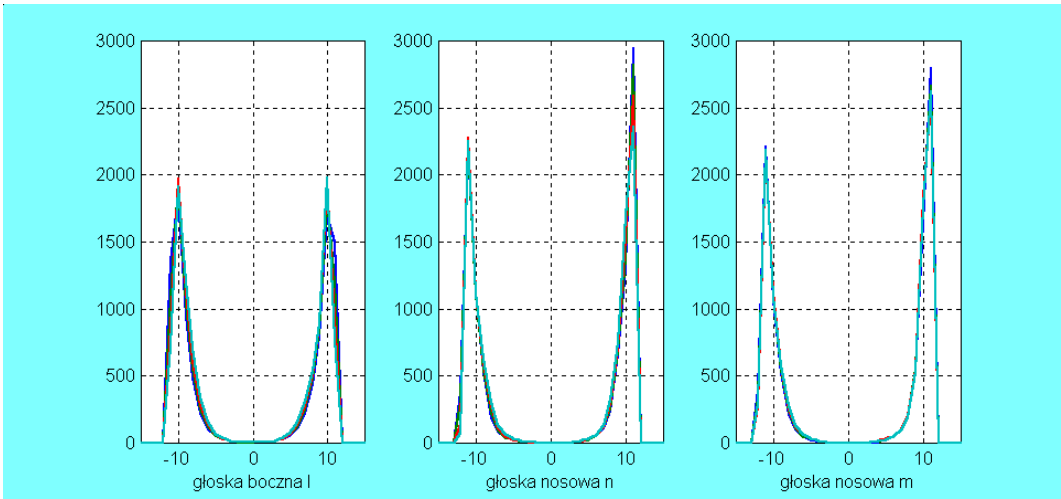


Rys. 4.14a: Histogramy głosek szumowych ‘s’ oraz ‘sz’ języka polskiego.



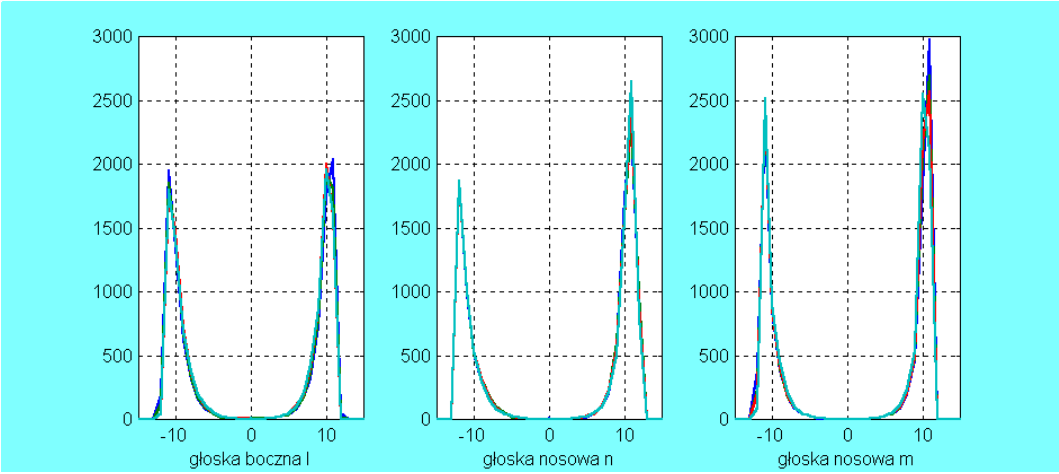
Rys. 4.14b: Porównywanie histogramów głosek szumowych: ‘s’, ‘sz’.

Rysunek 4.15 przedstawia histogramy głosek nosowych ’n’ i ‘m’ oraz głoski bocznej ‘l’. O ile histogram głoski ‘l’ można odróżnić od histogramów głosek ‘n’ oraz ‘m’, o tyle histogramy tych dwóch ostatnich nie są łatwo rozróżnialne. Różnią się one nieznacznie wysokością wierzchołka występującego dla ujemnego zakresu amplitud.

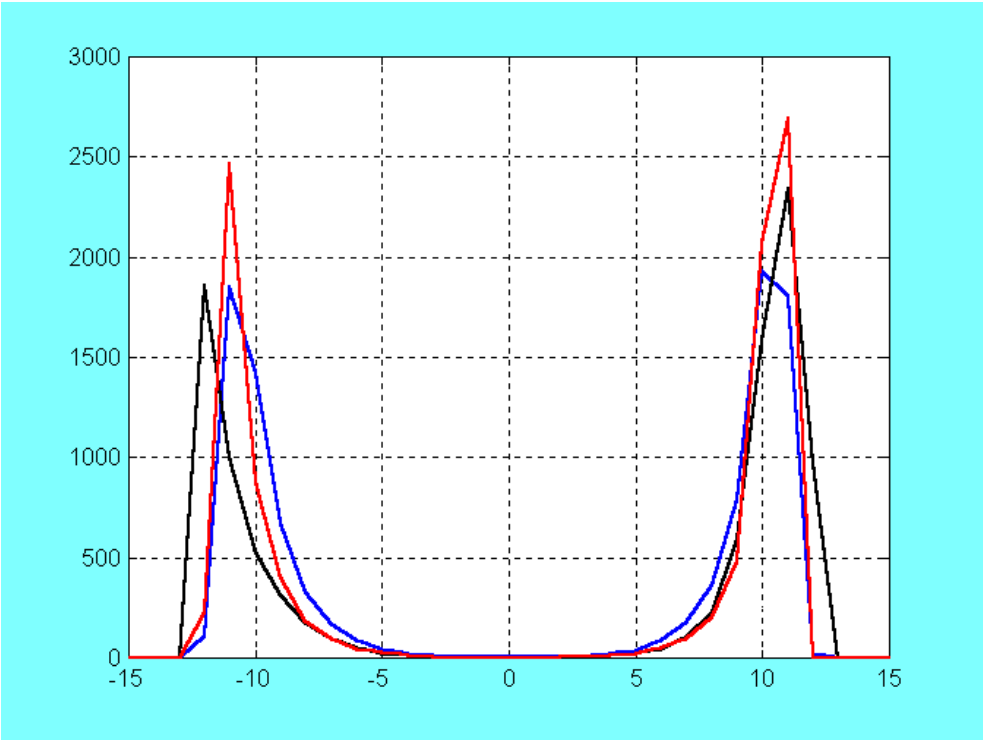


Rys. 4.15: Histogramy głosek nosowych ’n’ i ‘m’ oraz głoski bocznej ‘l’ – głos damski.

Rysunek 4.16 (a,b) przedstawia histogramy tych samych głosek nagranych głosem męskim w takich samych warunkach. Widać na nim, iż w tym przypadku wszystkie głoski są rozróżnialne.



Rys. 4.16a: Histogramy głosek nosowych ’n’ i ‘m’ oraz głoski bocznej ‘l’ – głos męski.



Rys. 4.16b: Porównanie histogramów głosek ‘l’, ‘n’ oraz ‘m’ – głos męski.

4.3. Inne warianty obliczenia histogramów

4.3.1. Histogramy pochodnych sygnału – wpływ preemfazy

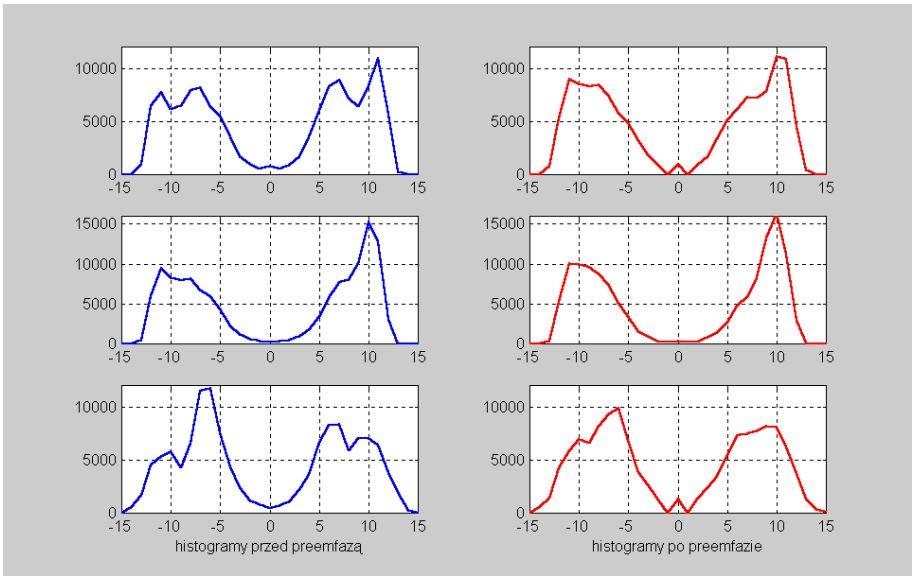
W systemach automatycznego rozpoznawania mowy oraz w systemach identyfikacji mówcy sygnał mowy często jest poddawany operacji preemfazy przed ekstrakcją parametrów. Operacja ta ma na celu wyrównanie energii poszczególnych składowych sygnału poprzez wzmocnienie wyższych składowych częstotliwości posiadających z reguły względnie mniejszą energię. W wyniku zastosowania operacji preemfazy zwiększany jest stosunek sygnału do szumu (*Signal to Noise Ratio SNR*). W celu maksymalnej poprawy stosunku *SNR* zalecane jest aby operację preemfazy przeprowadzono przed przetwarzaniem analogowo-cyfrowym sygnału [Jua93]. Z matematycznego punktu widzenia operacja preemfazy może być interpretowana jako różniczkowanie sygnału. W przypadku wykonania operacji preemfazy po przetwarzaniu sygnału na postać cyfrową jest ona w praktyce realizowana w postaci filtru o następującej transmitancji *z*:

$$H(z)=1-a.z^{-1} \qquad , \qquad 0.9 \le a \le 1.0 \qquad (4.11)$$

Postać sygnału po operacji preemfazy jest dana wzorem:

$$x_p(n)=x(n) - a.x(n-1) \qquad (4.12)$$

Na rysunku 4.17 przedstawiono trzy pary histogramów dla trzech sygnałów (przed i po operacji preemfazy). Sygnały odpowiadają trzem wypowiedziom (każda o długości trzech sekundy) nagranych głosem damskim w warunkach laboratoryjnych. Widać na nim, że kształt histogramu może ulegać znaczącej zmianie w wyniku przeprowadzenia operacji preemfazy.



Rys 4.17: Wpływ operacji preemfazy na kształt histogramu.

4.3.2. Histogramy selektywne mniej zależne od tempa mowy

Badania dowodzą [Fur89], iż zmiany w tempie mowie są w 65% wynikiem zmian odstępów między słowami oraz sylabami. Inaczej mówiąc zmiana szybkości mowy w dużej mierze wynika ze skrócenia lub wydłużenia odstępu czasowego między słowami (fragmentów ciszy) a niekiedy sylabami. W celu ograniczenia wpływu tempa mowy na kształt histogramu można je obliczać po wyeliminowaniu z wypowiedzi występujących w niej fragmentów ciszy. Ogólnie sprowadza się to do tego, że sygnał jest przepuszczany przez tzw. bramkę szumu przed obliczeniem jego histogramu. Bramka szumu jest zazwyczaj stosowana w celu ‘wyciszania’ fragmentów występujących między słowami, które zazwyczaj zawierają sam szum. Posiada ona dwa parametry: długość czasową t_b oraz próg amplitudowy p_b. W wyniku działania bramki wycisza się z sygnału fragmenty o bezwzględnej amplitudzie mniejszej od progu p_b i czasie

trwania dłuższym od t_b[6]. Dopóki chwilowa wartość amplitudy sygnału przechodzącego przez bramkę utrzymuje się powyżej progu p_b, bramka pozostaje otwarta, przepuszczając sygnał w postaci niezmienionej. Gdy chwilowa amplituda sygnału spadnie poniżej tego progu bramka zamyka się i rozpoczyna się proces zliczania czasu zamknięcia bramki. W tym czasie sygnał wejściowy jest zapamiętywany przez bramkę. Bramka pozostaje zamknięta do momentu gdy chwilowa bezwzględna amplituda sygnału wejściowego znowu przekroczy próg p_b. Jeśli od momentu zamknięcia do momentu ponownego otwarcia bramki nie minął czas t_b, wówczas w momencie otwarcia zapamiętywany wcześnie fragment sygnału wejściowego jest dołączany z powrotem do wyjściowego sygnału (przepuszczony bez zmian). W przeciwnym wypadku fragment ten zostaje wyzerowany (ewentualnie wyeliminowany) a licznik czasu zamknięcia zostanie wyzerowany. W ten sposób stan bramki powraca do stanu pierwotnego i cykl się powtarza.

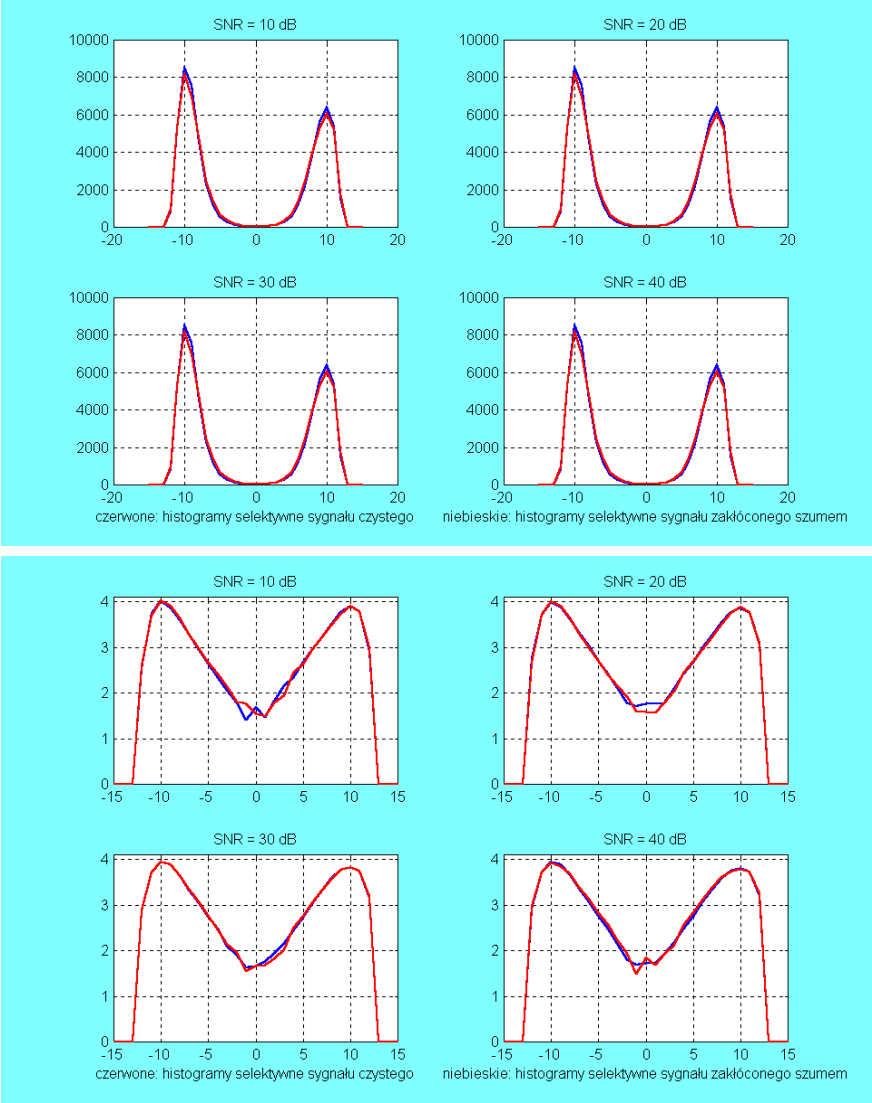
Należy wspomnieć, że wybór parametrów bramki dokonywany jest w sposób arbitralny. W przypadku sygnału mowy optymalny wybór wartości parametru t_b jest w dużej mierze zależny od tempa mowy. W praktyce parametr ten dobiera się z przedziału 50–150ms. Wybór progu natomiast jest ściśle związany z poziomem szumu występującym w sygnale. Bramka szumów, jak sama nazwa wskazuje, jest stosowana w celu filtracji ewentualnego hałasu występującego w przerwach między fragmentami sygnału niosącymi informację (mowa). W tym przypadku odcinki są zerowane (wyciszane). Natomiast w przypadku eliminacji przerw między słowami owe fragmenty są odcinane z sygnału. W tym wypadku dopuszcza się użycia mniejszych od 50ms wartości parametru t_b. Na ogół w wyniku przepuszczania sygnału przez bramkę szumu mogą zostać obcięte fragmenty sygnału zawierające także mowę. Sytuacja ta ma miejsce w przypadku głosek o stosunkowo niższej energii, porównywalnej do energii samego szumu w sygnale. Wtedy istotny wpływ ma – przede wszystkim – odpowiedni dobór wartości progu bramki w zależności od jakości sygnału. Jak już wspomniano w rozdziale drugim oraz trzecim niektóre systemy automatycznego rozpoznawania mówcy wykorzystują tylko część informacji zawartej w sygnale mowy mającej względnie większą energię. Technika ta polegająca na selekcji tylko pewnych fragmentów sygnału mowy przed ekstrakcją parametrów będzie również przedmiotem badania niniejszej pracy. W celu odróżnienia histogramów obliczanych po przepuszczeniu sygnału przez bramkę szumu od tych zwykłych będą one dalej nazywany histogramami selektywnymi.

[6] W szczególności fragmenty te mogą być ‘wycinane’ z sygnału, sam sygnał zaś zostanie ‘ściśnięty’.

4.4. Odporność histogramów na biały szum

W punktach 4.1.8 przedyskutowano wpływ pasma przenoszenia sygnału mowy na kształt jego histogramu amplitudowego. Pasmo przenoszenia sygnału jest uwarunkowane między innymi charakterystyką kanału transmisyjnego oraz aparatury rejestrującej. Ogólnie rzecz biorąc jakość sygnału jest również zależna od rodzaju mikrofonu oraz dokładności przetwornika analogowo-cyfrowego. Oprócz tego w warunkach naturalnych[7] istotny wpływ ma również poziom szumu towarzyszącego sygnałowi. Dlatego też podjęto tu próbę oszacowania odporności histogramów amplitudowych na biały szum. Badania wykazały, że histogramy selektywne są bardziej odporne na obecność szumu w sygnale, co było łatwe do przewidzenia. Dlatego też będą one przedmiotem badań w tym podrozdziale. Celem oszacowania odporności histogramów selektywnych na obecność białego szumu w sygnale przeprowadzono następujący eksperyment. Do sygnału nagranego w warunkach laboratoryjnych (wypowiedź o długości 1,5 sekundy) dodano numerycznie biały szum i porównywano histogramy selektywne kilku wersji sygnału o różnych stosunkach sygnału do szumu *SNR*. Badania przeprowadzono dla następujących wartości *SNR*: 10, 20, 30 oraz 40dB. Histogram obliczono dla sygnału wyjściowego bramki szumu o stałej długości czasowej $t_b=20ms$, próg bramki jednak był dobierany w zależności od *SNR*. Badania wykazały, że próg bramki zapewniający zadawalające podobieństwo histogramów selektywnych sygnału zakłóconego oraz oryginalnego zależy zarówno od parametru *SNR* jak i od składu wypowiedzi (dokładnie od energii występujących w niej głosek mającej decydujący wpływ co do podatność sygnału na zakłócenia). Im stosunek *SNR* jest większy tym próg bramki powinien być większy i co za tym idzie sygnał wyjściowy bramki jest krótszy. Niech stosunek łącznej długości odrzucanych z sygnału fragmentów w wyniku jego przepuszczenia przez bramkę szumu do długości sygnału wejściowego bramki będzie oznaczony symbolem δ. Rysunek 4.18 przedstawia cztery pary histogramów selektywnych odpowiadających poszczególnym wymienionym powyżej wartościom stosunku *SNR*. Tabela 4.3 podaje wartości δ odpowiadające histogramom z rysunku 4.18 w zależności od stosunku *SNR*. Wartości progu bramki zostały tak dobrane, aby zapewnić odpowiednie wartości δ gwarantujące pewne wizualne podobieństwo histogramów sygnału oryginalnego oraz zakłóconego białym szumem.

[7] np. gdy sygnał mowy jest zarejestrowany w warunkach pokojowych (biurowych) lub przesyłany w łączach telefonicznych.



Rys. 4.18: Odporność histogramów selektywnych na biały szum.
(a): skala liniowa na osi 'y', (b): skala logarytmiczna na osi 'y'.

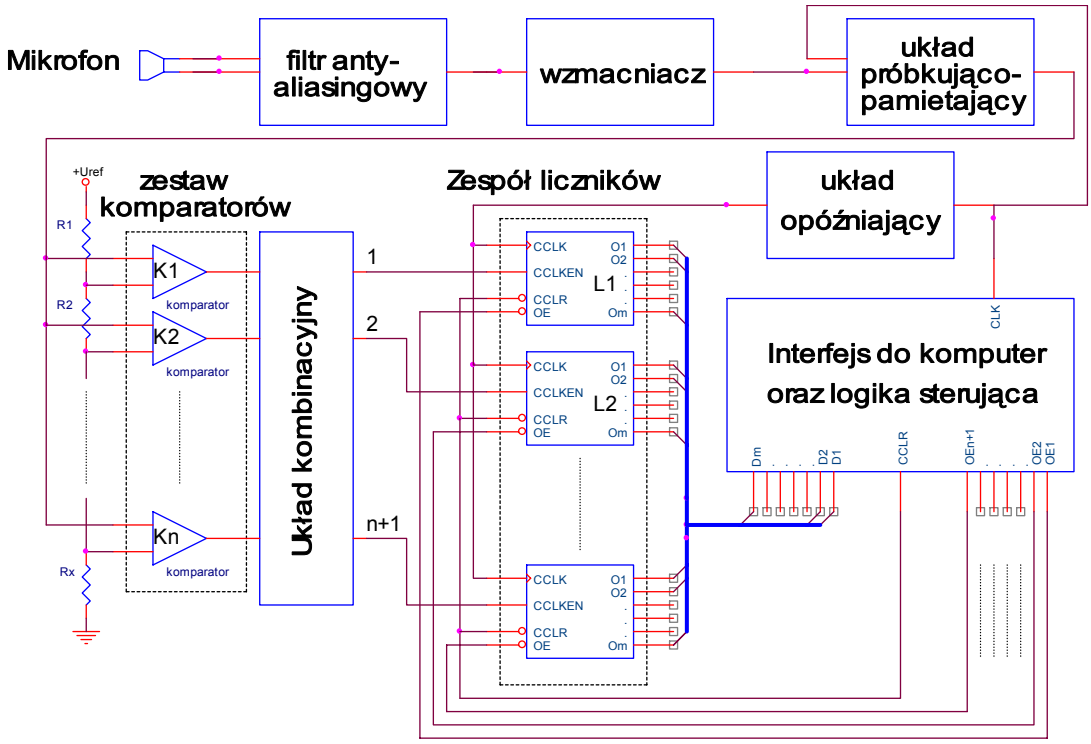
SNR [dB]	10	20	30	40
δ [%]	35	30	20	15

Tabela 4.3: Wymagane wartości δ zapewniające podobieństwo histogramów selektywnych sygnału czystego oraz zakłóconego.

Należy zaznaczyć, że możliwa jest również filtracja sygnału zakłóconego przed jego przepuszczeniem przez bramkę szumu. Pozwala to na wstępną redukcję stosunku *SNR* i w efekcie pozwala na zmniejszanie wartości δ wymaganych w celu zapewnienia zakładanego stopnia podobieństwa histogramów sygnału czystego oraz zakłóconego.

4.5. Realizacja sprzętowa procesora histogramowego

Jedną z najważniejszych zalet histogramów amplitudowych jako wektorów cech jest bardzo prosta procedura ich ekstrakcji. W przypadku kwantyzacji sygnału zgodnie z wektorem amplitud histogramu wymaga ona jedynie zliczania krotności wystąpienia poszczególnych amplitud w sygnale. Co więcej kwantyzacja sygnału np. na 31 poziomów oznacza zasadnicze uproszczenie struktury przetwornika analogowo-cyfrowego. Przykładowo w przypadku równoległych przetworników oznacza to znaczną redukcję liczby komparatorów oraz eliminację dekodera [Bara94] co umożliwia łatwą realizację przetwornika[8] przy dodatkowym ograniczeniu jego kosztu i poprawy parametrów (np. poboru mocy, impedancji wejściowej). Natomiast w przetwornikach nadążnych zmniejszanie liczby poziomów kwantyzacji oznacza między innymi uproszczenie struktury zawartego w nim przetwornika cyfrowo analogowego oraz skrócenie czasu przetwarzania [Bara94].



Rys. 4.19: Realizacja sprzętowa procesora histogramowego.

[8] Ze względu na pobór mocy oraz ograniczenie powierzchni układu scalonego nie produkuje się przetworników równoległych o rozdzielczości lepszej niż 8 bitów (255 komparatorów).

W przetwornikach równoległych sygnały z wyjść komparatorów mogą być bezpośrednio wykorzystane do odblokowania odpowiedniego licznika amplitud. Schemat takiego procesora histogramowego jest przedstawiony na rysunku 4.19. Układ kombinacyjne o prostej strukturze odpowiednio mnoży sygnały z komparatorów w celu obliczenia sygnałów sterujących procesem zliczania. Taki układ może się składać z samych bramek NAND oraz inwerterów. W danej chwili aktywne może być tylko jedno wyjście tego układu, co powoduje zwiększenie licznika odpowiedniej wartości amplitudy w momencie dotarcia impulsu do wejścia zegarowego licznika. Impuls ten jest odpowiednio opóźniony w stosunku do impulsu próbkującego o czas potrzebny na zadziałanie komparatorów oraz układu kombinacyjnego. Układ sterujący po czytaniu wyników dla jednego histogramu zeruje liczniki i proces zliczania się powtarza. Stany wyjść liczników są odczytywane po kolej. W tym celu układ sterujący wykorzystuje sygnały otwarcia buforów wyjściowych liczników *OE*. Jeśli zachodzi konieczność blokowania procesu zliczania na czas odczytu liczników to można to zrealizować za pomocą odpowiedniego połączenia między układem sterującym i licznikami lub układem kombinacyjnym.

Istnieje również możliwość sprzętowej realizacji bramki szumu. W tym celu między układem kombinacyjnym a licznikami należy wprowadzić rejestry przesuwne typu FIFO. Wówczas układ sterujący powinien mieć dostęp do wyjściowych sygnałów układu kombinacyjnego lub do wyjściowych sygnałów z komparatorów. Wartości progu p_b oraz czasu t_b bramki mogą być np. ustalane przez komputer nadrzędny i zapamiętywane w układzie sterującym. Długość rejestrów FIFO powinna być dłuższa od iloczynu F_p i t_b. Zaimplementowany w układzie sterującym licznik zlicza czas w którym amplituda sygnału (dostarczana do niego np. z wyjść komparatorów) nie przekracza zadanego progu p_b. W momencie gdy stan wyjściowy tego licznika będzie równy iloczynowi F_p*t_b to rejestry przesuwne FIFO są zerowane i podtrzymywane w tym stanie do momentu w którym wartość sygnału znowu przekroczy próg p_b. Wtedy licznik sterownika jest zerowany i cykl może się powtarzać. Odpowiednia logika w układzie sterującym powinna być przeznaczona do podjęcia decyzji o czytaniu stanu poszczególnych liczników amplitud i ich zerowaniu.

W przypadku programowej realizacji procesu zliczania wyjścia komparatorów mogą być podłączane bezpośrednio do układu sterującego bądź do portu komputera nadrzędnego. W przypadku gdy zakres amplitud jest rozdzielony symetrycznie na 31 poziomów ($p=15$) w sposób logarytmiczny zgodnie z wartościami podanymi w tabelce

4.1 to realizacja programowa procesora histogramowego w języku maszynowym (zarówno segregacji jak i zliczanie amplitud) staje się bardzo prosta ponieważ odpowiednie poziomy amplitud są związane ze stanem poszczególnych 16 bitowego słowa reprezentującego pojedynczą próbkę. Zatem segregacja amplitud można wykonać za pomocą operacji manipulacji bitami.

ROZDZIAŁ V

Rozpoznawanie mówcy na podstawie długookresowego histogramu amplitud sygnału mowy

Niniejszy rozdział poświęcono opisowi metody automatycznego rozpoznawania mówcy na podstawie długookresowego histogramu amplitud sygnału mowy. Zawarto w nim szczegółowy opis metody i jej wariantów oraz dokonano wstępnej oceny przydatności zaproponowanego opisu histogramowego do dyskryminacji głosów różnych mówców.

5.1. Założenia metody

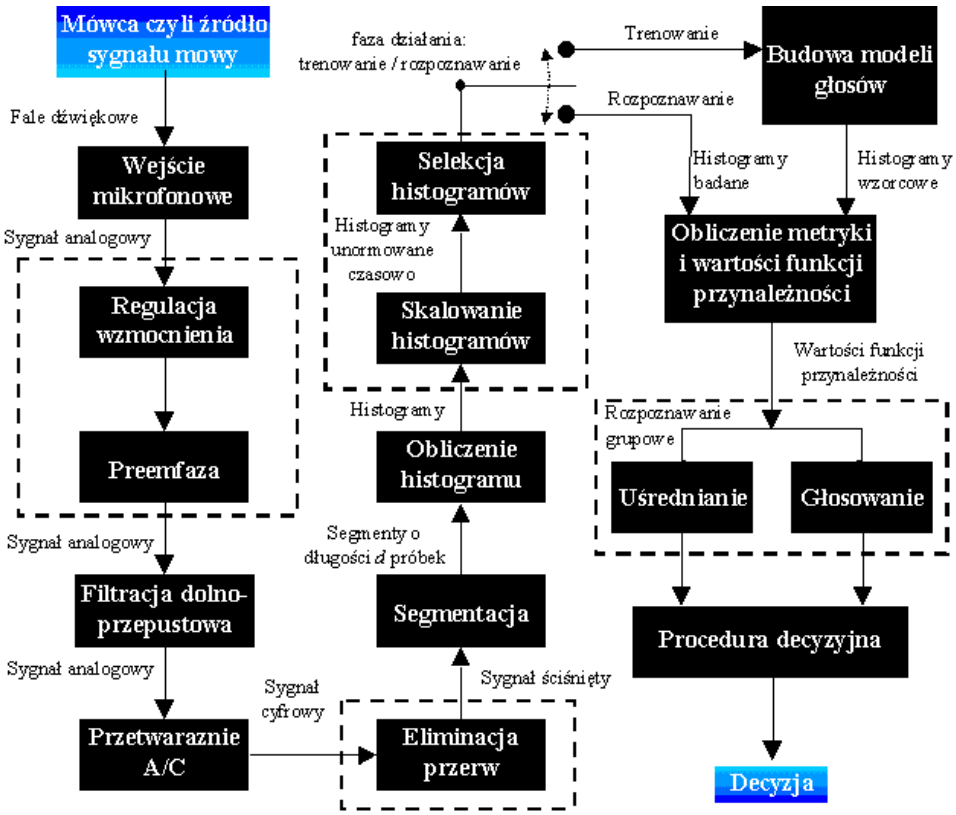
Budując opisaną poniżej metodę oparto się na następujących założeniach:

1. Informacje w sygnale mowy (w tym cechy osobnicze mówcy) są zapisane w jego wartościach bądź w ich zmianach (w pochodnych).
2. Zwiększanie czasu trwania sygnałów (długość ciągów uczących lub danych testujących), na podstawie których będzie budowany model głosu lub zostanie podjęta decyzja o rozpoznawaniu powinno zwiększyć moc dyskryminacyjną metody. Podobna relacja jest słuszna w stosunku do umysłu ludzkiego, łatwiej bowiem jest rozpoznać głos, którego częściej (dłużej) się słucha.
3. Treść oraz język wypowiedzi nie powinny mieć wpływu na rozpoznawanie głosu jeśli wypowiedź jest dostatecznie długa.
4. Wypowiedzi będzie można segmentować w sposób 'ślepy' chociaż nie jest to jakoś szczególnie zalecane. Założenie to opiera się na fakcie, iż człowiek potrafi identyfikować znany mu głos nawet jeśli np. końcówka wypowiedzi jest odcięta.
5. Każda cecha osobnicza (nawet niekiedy wada wymowy) powinna zwiększać prawdopodobieństwo identyfikacji głosu jeśli jest ona powtarzalną cechą (właściwością) mowy danej osoby i występuje często.

5.2. Opis metody rozpoznawania mówcy

W swoim podstawowym wariancie, dyskutowana w niniejszej rozprawie metoda rozpoznawania mówcy opiera się na opisie sygnału mowy za pomocą histogramów amplitudowych oraz zaproponowanej w rozdziale drugim metodzie klasyfikacji ‘potencjału najbliższych sąsiadów’. Sposób działania metody jest zgodny z ogólnym schematem prezentowanym na rysunku 2.1 oraz 2.2 rozdziału drugiego. Rysunek 5.1 przedstawia kompleksowy schemat działania metody, przy czym bloki opcjonalne zaznaczono liniami przerywanymi.

W dalszej części niniejszego rozdziału zamieszczono szczegółowy opis metody rozpoznawania oraz jej wariantów z pominięciem zagadnień podjętych wcześnie w rozdziale drugim oraz w czwartym.



Rys. 5.1: Szczegółowy schemat działania metody automatycznego rozpoznawania mówcy na podstawie długookresowego histogramu amplitud sygnału mowy.

5.2.1. Warianty wstępnego przetwarzania sygnału

Cyfrowy sygnał wejściowy może być poddany operacji preemfazy przed obliczeniem histogramów amplitudowych. W niektórych przypadkach można na tym etapie dokonać normalizacji amplitudowej sygnału względem jego energii (wzmocnić bądź stłumić sygnał w celu zapewnienia z góry zakładanego poziomu energii sygnału). Do tego celu może służyć wzmacniacz z automatyczną regulacją wzmocnienia na analogowym torze sygnału. Zarówno operacja preemfazy jak i unormowanie poziomu energii sygnału mogą być wykonywane programowo po przekształceniu sygnału na postać cyfrową. Możliwe jest także przepuszczanie sygnału przez bramkę szumu (realizowana sprzętowo bądź programowo) w celu późniejszego obliczania histogramów selektywnych. Histogramy są obliczane po segmentacji sygnału (w sposób "ślepy") na fragmenty

o długości *d*. W przypadku rozpatrywanej metody segmenty nie są mnożone przez funkcję okna (tzn. używane jest okno prostokątne).

5.2.2. Obliczenie i selekcja histogramów

Z każdego fragmentu o długości *d* oblicza się pojedynczy histogram. Zakres amplitud tych histogramów może być podzielony w sposób liniowy bądź logarytmiczny. Ponadto liczebność amplitud może być wyrażona w sposób bezpośredni lub w stosunku do łącznej liczby próbek w rozpatrywanym fragmencie (unormowanie czasowe). Istnieje również możliwość logarytmowania wartości histogramu. Zasady obliczenia histogramów zostały zaprezentowane w rozdziale czwartym (punkty 4.1 – 4.4).

Biorąc pod uwagę, że rozpatrywana metoda rozpoznawania mówcy jest z założenia metodą niezależną od tekstu jak i języka nagranych wypowiedzi, należy spodziewać się wystąpienia fragmentów sygnału – i co zatem idzie histogramów – o charakterze niepowtarzalnym. Jest to całkowicie prawdopodobne zwłaszcza w przypadku, gdy łączna długość danych trenujących nie przekracza kilkuset słów (ok. 2 minuty), co uniemożliwia reprezentowanie całego słownika wyrazów danego języka na ich podstawie zwłaszcza, gdy pojedynczy obiekt odpowiada dłuższej wypowiedzi (dłuższej jednostce mowy). Z tego względu, wiarygodne stwierdzenie niepowtarzalności danego fragmentu wiąże się z koniecznością badania odpowiednio długiego nagranego tekstu, co może się okazać niepraktyczne. Dodatkowym powodem wystąpienia niepowtarzalnych fragmentów jest również zachowanie pełnej swobody w nagraniu wypowiedzi jak i ich rozdzielanie w sposób ślepy. Wówczas mogą występować fragmenty nie reprezentatywne zawierające np. dłuższy odcinek ciszy[1]. Dlatego też może zachodzić konieczność eliminacji niepowtarzalnych (niereprezentatywnych) histogramów. Do tego celu w zaproponowanej metodzie dopuszcza się możliwość użycia mechanizmu selekcji zaproponowanego w punkcie 2.9. Mechanizm ten został zaprojektowany z myślą o zastosowanej metodzie klasyfikacji ‘potencjału najbliższych sąsiadów’. W fazie uczenia się systemu selekcja histogramów może być użyta wraz z algorytmem kwantyzacji wektorowej w celu optymalnego wyboru histogramów reprezentujących poszczególne głosy w budowanej przestrzeni cech. Nic nie stoi na przeszkodzie by mechanizm selekcji został również zastosowane w fazie rozpoznawania, co pozwala na podjęcie decyzji o

[1] jak początki nagrania lub fragmenty obejmujące przerwę w czytaniu.

rozpoznawaniu tylko na podstawie wybranych fragmentów. Jednak może on być zbędne w przypadku realizacji idei rozpoznawania grupowego.

5.2.3. Miary podobieństwa histogramów

W fazie rozpoznawania oblicza się metrykę między badanym histogramem a histogramami wzorcowymi i określa się wartość jego funkcji przynależności do poszczególnych klas (modeli) zgodnie z założeniami metody *pkNN* (wariant podstawowy metody).

Zaproponowana metoda rozpoznawania mówcy była testowana z użyciem dwóch różnych metryk: Euklidesowej oraz Camberra. Operacja logarytmowania histogramu może być traktowana jako część operacji obliczenia metryki między histogramami. Włączając operację logarytmowania w zapisie metryki można mówić o przebadaniu następujących czterech przypadków:

1. $$\rho 1(x,y) = \sqrt{\sum_{i=1}^{N} (x_i - y_i)^2} \tag{5.1}$$

2. $$\rho 2(x,y) = \sqrt{\sum_{i=1}^{N} [log(x_i + 1) - log(y_i + 1)]^2} \tag{5.2}$$

3. $$\rho 3(x,y) = \sum_{i=1}^{N} \left| \frac{x_i - y_i}{x_i + y_i} \right| \tag{5.3}$$

4. $$\rho 4(x,y) = \sum_{i=1}^{N} \left| \frac{log(x_i + 1) - log(y_i + 1)}{log(x_i + 1) + log(y_i + 1)} \right| \tag{5.4}$$

5.2.4. Warianty klasyfikacji

Należy podkreślić, iż serce zaproponowanej metody rozpoznawania stanowi użycie histogramów amplitudowych jako wektorów cech. Dlatego też metoda ta została testowana z użyciem różnych metod klasyfikacji. Przykładowo w publikacji [Sho98] oceniono skuteczność rozpoznawania na podstawie histogramów amplitudowych z użyciem tradycyjnych metod klasyfikacji: metoda wzorców oraz metoda *kNN*. Wyniki różnych prowadzonych badań wykazały jednak, że lepsze wyniki można uzyskać stosując

inne ulepszone metody klasyfikacji. W szczególności zaproponowano metodę potencjału najbliższych sąsiadów *pkNN* jako połączenie idei metody funkcji potencjalnych oraz metody *kNN*. Nic nie stoi na przeszkodzie aby stosować także inne rozbudowane algorytmy dopasowania próbek do wzorców. Możliwe jest zastosowanie sieci neuronowych, techniki *DTW* opartej na programowaniu dynamicznym, niejawnych modeli Markowa *HMM* oraz klasyfikatory oparte na technice kwantyzacji wektorowej, binarnej lub macierzowej. W szczególności w przypadku krótkookresowych histogramów można podjąć próbę rozpoznawania mówców zależnego od tekstu.

Oprócz metody klasyfikacji *pkNN* zaproponowano również w punkcie 2.14 ideę rozpoznawania grupowego polegającego na podjęciu decyzji o rozpoznawaniu w oparciu o kilka kolejnych histogramów. Przykładowo wartość funkcji przynależności dla grupy histogramów można obliczyć jako sumę poszczególnych wartości tej funkcji dla każdego z nich z osobna. Możliwa jest także klasyfikacja na podstawie głosowania jak to zostało opisane w punkcie 2.14.

5.3. Podsumowanie warunków przeprowadzenia rozpoznawania

W tabelce 5.1 zamieszczono najważniejsze warianty warunków przeprowadzenia rozpoznawania. Ogólnie są one związane z jednym z następujących zagadnień:

1. jakością sygnału wejściowego oraz warunkami jego rejestracji,
2. wstępnym przetwarzaniem sygnału,
3. obliczeniem oraz selekcją histogramów,
4. wyborem metryki,
5. użytą procedurą klasyfikacji,
6. procedurą decyzyjną (identyfikacja lub weryfikacja),
7. rozmiarem rozpatrywanej populacji,
8. otwartością zbioru rozpoznawanych mówców,
9. zależnością rozpoznawania od tekstu.

Nr	Kategoria	Warianty (*skrót oznaczenia*)
1	Jakość sygnału wejściowego	Laboratoryjna *lab*, pokojowa *pok* lub telefoniczna *tel*
2	Przetwarzanie A/C	Pasmo przenoszenia sygnału F_g Częstotliwość próbkowania F_p Rozdzielczość *B*

Nr	Kategoria	Warianty (*skrót oznaczenia*)
3	Przetwarzanie wstępne	Filtracja szumu *filtr* Preemfaza *preem, a* Normalizacja energii N_E Bramka szumu: t_b, p_b, δ Segmentacja: *d*
4	Obliczenia histogramów	Podział zakresu amplitud: liniowe *hlin* lub logarytmiczne *hlog* Liczba poziomów amplitud $L_b=2P+1$ Unormowanie czasowe U_t , długość *C*
5	Selekcja histogramów	Procent odrzucanych: o_p
6	Metryka w przestrzeni	$\rho 1$, $\rho 2$, $\rho 3$ lub $\rho 4$
7	Metoda klasyfikacji	Wzorców *wz*, *kNN*, *pkNN*,
8	Klasyfikacja grupowa	Głosowanie *Gg* Uśrednianie dla *Gq* kolejnych histogramów
9	Procedura decyzyjna	Identyfikacja *I*, weryfikacja *W*
10	Rozmiar populacji	Liczba mówców *M*
11	Zbiór mówców	Otwarty *O*, zamknięty *Z*
12	Zależność od tekstu	Zależne *dep*, niezależne *indep*

Tabela 5.1: Warunki przeprowadzenia rozpoznawania; skróty oznaczeń.

5.4. Wstępna ocena siły dyskryminującej histogramów

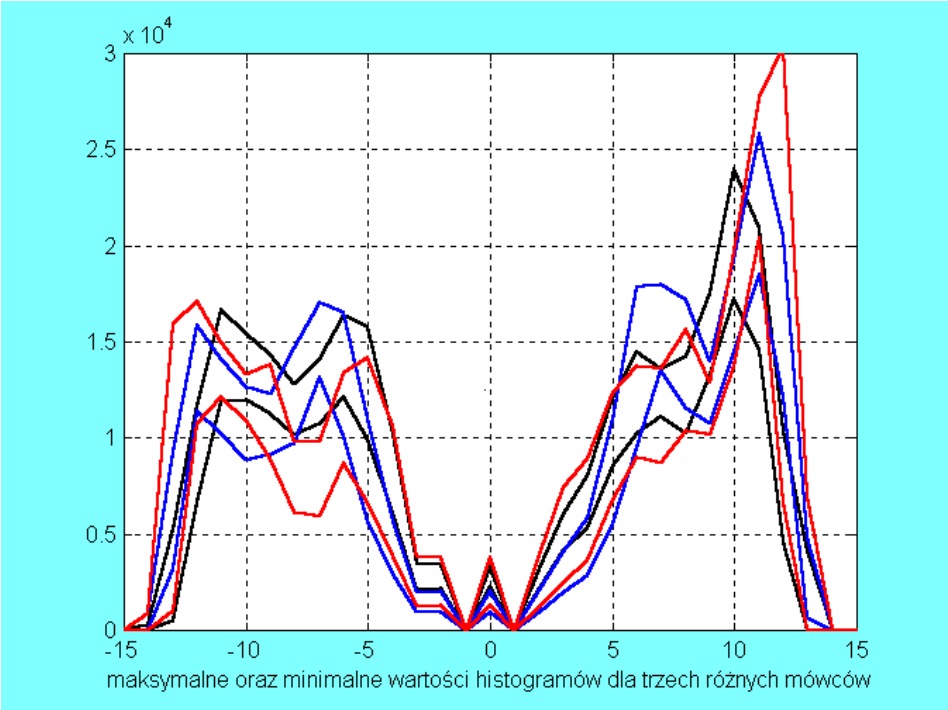
W celu wstępnej ilustracji możliwości dyskryminacji głosów na podstawie długookresowych histogramów amplitudowych skorzystano z nagrania tylko trzech głosów damskich o łącznej długości ok. 6 minut. Nagrania zarejestrowano w komorze bezechowej za pomocą wysokiej klasy aparatury rejestrującej. Częstotliwość próbkowania wynosiła $F_p=44{,}1kHz$ zaś rozdzielczość amplitudowa $B=16$ bitów/próbkę. Osoby biorące udział w nagraniach czytały w sposób ciągły i swobodne tekst typu gazetowego.

Rysunek 5.2a przedstawia minimalne oraz maksymalne wartości histogramów dla poszczególnych mówców. Wartości te zostały obliczone dla 32 histogramów[2] uzyskanych przy następujących warunkach:

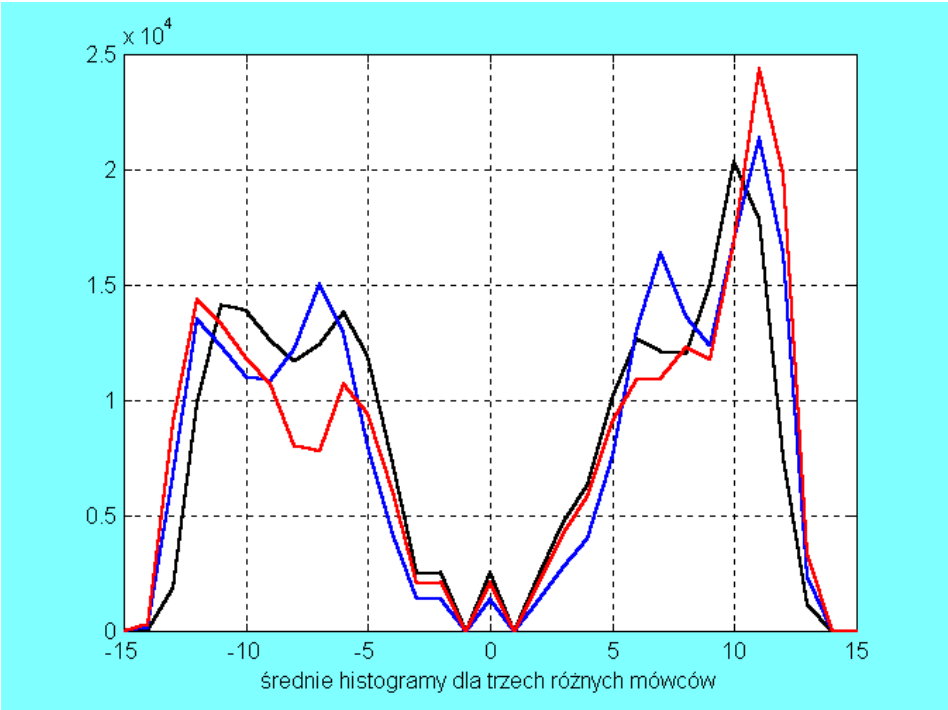
$$lab,\ F_g=20kHz,\ F_p=44{,}1kHz,\ B=16,\ U_t,\ hlog,\ L_p=31,\ d=10.8\ sek.,\ M=3.$$

[2] Łączna długość nagrań wynosi 6 minut a więc dla $d=10.8$ uzyskuje się ok. 360 /10.8 ≅ 32 histogramy.

Ponadto na rysunku 5.2b przedstawiono średnie histogramy dla poszczególnych mówców.



Rys 5.2a: Wizualna ocena możliwości dyskryminacji głosów na podstawie histogramów.



Rys 5.2b: Wizualna ocena możliwości dyskryminacji głosów na podstawie histogramów.

Rysunek 5.2 pozwala jedynie na wizualną ocenę możliwości dyskryminacji głosów na podstawie histogramów amplitudowych sygnału mowy. Inną obiektywną ocenę siły dyskryminacyjnej histogramów można uzyskać na podstawie miary dyskryminacji *S* zaproponowanej w [Sho98]. Stanowi ona stosunek średniej odległości obiektów wewnątrz klas (między obiektami pojedynczej klasy) do średniej odległości obiektów różnych klas. Taka miara jest podobna do odwrotności miary Fishera lecz pozbawiona najpoważniejszej wady jaką jest w tej ostatniej – brak zależności od wartości średniej.

5.4.1. Wpływ długości czasowej wypowiedzi

Badania opisane w tym punkcie uwzględniają wpływu długości wypowiedzi (parametru *d*) na stosunek *S*. Były one prowadzone przy następujących warunkach:

lab, F_g*=20kHz,* F_p*=44,1kHz, B=16,* U_t*, hlog,* L_p*=31,* $\rho 3$*, M=3.*

Tabela 5.2 przedstawia otrzymane wyniki. Zapewnienie warunku U_t jest bardzo istotne z punktu widzenia porównywania wyników ponieważ gwarantuje ono możliwość porównywania metryk między histogramami o różnych *d* a więc i wartości stosunku *S*.

d [sek]→	1,35	2,7	5,4	10,8	21,6	43,2
Mówca 1	0,82	0,57	0,41	0,27	0,20	0,03
Mówca 2	0,81	0,61	0,40	0,28	0,26	0,25
Mówca 3	0,92	0,91	0,67	0,61	0,47	0,47

Tabela 5.2: Miara skupienia *S* w zależności od długości czasowej histogramu *d* – odległości między histogramami obliczone metryką $\rho 3$.

Należy zauważyć, że stosunek *S* systematycznie maleje wraz ze wzrostem długości *d*. Jest on ponadto zróżnicowany dla różnych mówców (najlepszy dla pierwszego mówcy i najgorszy dla trzeciego mówcy niezależnie od wartości *d*).

5.4.2. Wpływ metryki

Badania opisane w poprzednim punkcie powtórzono dla pozostałych trzech metryk. Tabela 5.3 przedstawia otrzymane wyniki.

Metryka	*d [sek]*→	1,35	2,7	5,4	10,8	21,6	43,2
$\rho 1$	Mówca 1	0,86	0,72	0,59	0,46	0,36	0,15
$\rho 1$	Mówca 2	0,90	0,79	0,63	0,51	0,48	0,44
$\rho 1$	Mówca 3	0,93	0,92	0,79	0,71	0,61	0,61
$\rho 2$	Mówca 1	0,94	0,85	0,73	0,76	0,85	0,53
$\rho 2$	Mówca 2	0,79	0,76	0,75	0,74	0,73	0,90
$\rho 2$	Mówca 3	0,98	0,98	0,94	0,96	0,90	1,10
$\rho 4$	Mówca 1	0,93	0,78	0,65	0,81	1,03	0,75
$\rho 4$	Mówca 2	0,64	0,68	0,76	0,72	0,60	0,92
$\rho 4$	Mówca 3	0,97	0,98	1,01	1,01	1,03	1,68

Tabela 5.3: Miara skupienia *S* w zależności od długości czasowej histogramu *d*. Odległości między histogramami obliczone kolejno metrykami: $\rho 1$, $\rho 2$, $\rho 4$.

Na podstawie wyników przedstawionych w tabelach 5.2 oraz 5.3 można wyciągnąć następujące wnioski:

1. Najlepsza okazała się metryka $\rho 3$ (Camberra bez logarytmowania). Szersze badania przedstawione w rozdziale szóstym także wskazują na słuszność tego twierdzenia lecz tylko dla względnie szerszego pasma przenoszenia sygnałów.
2. Wyniki dla metryk $\rho 1$ oraz $\rho 3$ są zdecydowanie lepsze niż dla metryk $\rho 2$ lub $\rho 4$ i co najważniejsze dla tych pierwszych spełniona jest zawsze następująca zależność:

$$\frac{\Delta S(d)}{\Delta d} \leq 0 \qquad (5.5)$$

Zatem logarytmowanie wartości histogramu prowadzi do pogarszania się stosunku *S*. Oznacza to, że histogramy bez logarytmowania lepiej dyskryminują głosy różnych mówców. W związku z powyższym badania opisane w rozdziale szóstym były prowadzone jedynie z użyciem metryki $\rho 1$ bądź $\rho 3$.

5.4.3. Ocena dla histogramów z liniowym podziałem zakresu amplitud

Badania z punktu 5.4.1 oraz 5.4.2 powtórzono dla przypadku histogramów liniowych *hlin*, $L_p=31$. Wyniki przedstawiono w tabelce 5.4. Ogólnie rzecz biorąc można stwierdzić, że wyniki dla histogramów typu *hlog* są lepsze niż dla typu *hlin*. Ponadto dla histogramów *hlin* zostały zachowane wnioski z punktu 5.4.2 słuszne dla histogramów typu *hlog*[3]. Jako wyjątek należy zanotować fakt, że w niektórych

[3] W przypadku mówcy trzeciego tylko częściowo.

przypadkach, dla mniejszych wartości *d*, wyniki dla metryk z logarytmem ($\rho 2$, $\rho 4$) są niekiedy lepsze niż te dla zwykłych metryk ($\rho 1$, $\rho 3$).

Metryka	*d [sek]*→	1,35	2,7	5,4	10,8	21,6	43,2
$\rho 1$	Mówca 1	0,81	0,78	0,68	0,56	0,38	0,14
$\rho 1$	Mówca 2	1,01	0,90	0,72	0,54	0,46	0,39
$\rho 1$	Mówca 3	0,95	1,06	0,96	0,94	0,82	0,93
$\rho 2$	Mówca 1	0,90	0,91	0,88	0,78	0,83	0,38
$\rho 2$	Mówca 2	0,82	0,76	0,74	0,83	0,83	0,96
$\rho 2$	Mówca 3	0,94	0,94	1,00	1,05	1,03	1,24
$\rho 3$	Mówca 1	0,72	0,61	0,48	0,38	0,31	0,03
$\rho 3$	Mówca 2	0,93	0,68	0,44	0,31	0,29	0,32
$\rho 3$	Mówca 3	0,93	1,05	0,93	0,91	0,77	0,92
$\rho 4$	Mówca 1	0,85	0,87	0,80	0,61	0,61	0,15
$\rho 4$	Mówca 2	0,75	0,68	0,74	0,94	0,91	0,96
$\rho 4$	Mówca 3	0,84	0,83	0,88	0,95	0,83	0,93

Tabela 5.4: Miara skupienia *S* w zależności od długości czasowej histogramu typu *lin* – badania prowadzono kolejno dla metryk: $\rho 1$, $\rho 2$, $\rho 3$ oraz $\rho 4$.

5.4.4. Ocena dla selektywnych histogramów

W tym punkcie podjęto próbę oceny skuteczności dyskryminacji głosów na podstawie histogramów selektywnych. W tym celu badania zależności *S* od *d* prowadzono dla wszystkich metryk przy następujących stałych warunkach:

lab, F_g=*20kHz*, F_p=*44.1kHz*, *B=16*, N_g, t_b=*50ms*, p_b=*5%*[4], U_t, *hlog*, L_p=*31*, $\rho 3$, *M=3*.

Ponadto oszacowano wartość δ (stosunek łącznej długości odrzucanych z sygnału fragmentów we wyniku jego przepuszczenia przez bramkę szumu do długości wejściowego sygnału bramki) dla różnych mówców. Badania wykazały, że wartość δ jest bardzo zbliżona dla różnych mówców i wynosi ok. 20%. Aby moc porównywać otrzymane wyniki z wynikami dla zwykłych histogramów, wykonano histogramy selektywne dla sygnałów wyjściowych bramki o długości równej 80% kolejnych wartości *d* użytych w punktach 5.4.1-5.4.3. Na uwagę zasługuje fakt, iż w odróżnieniu od wszystkich powyższych badań dokonano tutaj normalizacji energii sygnałów

[4] Wartość progu bramki dobrano jako 5% wartości maksymalnej bezwzględnej amplitudy sygnału.

(sprowadzono energię sygnałów wszystkich trzech głosów do tego samego poziomu) utrudniając tym samym możliwość ich dyskryminacji. Tabela 5.5 przedstawia otrzymane wyniki.

Metryka	*d [sek]*→	1,35	2,7	5,4	10,8	21,6	43,2
$\rho 1$	Mówca 1	0,90	0,87	0,77	0,62	0,43	0,18
$\rho 1$	Mówca 2	0,90	0,82	0,73	0,59	0,54	0,41
$\rho 1$	Mówca 3	0,91	0,87	0,80	0,62	0,52	0,42
$\rho 2$	Mówca 1	0,89	1,01	1,18	0,82	0,37	0,09
$\rho 2$	Mówca 2	1,02	0,95	0,68	0,74	1,12	1,62
$\rho 2$	Mówca 3	0,94	0,80	0,94	1,45	1,57	0,38
$\rho 3$	Mówca 1	0,84	0,81	0,73	0,48	0,30	0,06
$\rho 3$	Mówca 2	0,91	0,80	0,71	0,49	0,47	0,47
$\rho 3$	Mówca 3	0,82	0,78	0,73	0,60	0,49	0,32
$\rho 4$	Mówca 1	0,85	1,03	1,34	0,76	0,04	0,00
$\rho 4$	Mówca 2	1,03	0,92	0,43	0,56	1,17	1,95
$\rho 4$	Mówca 3	0,89	0,67	0,88	1,69	1,86	0,04

Tabela 5.5: Miara skupienia *S* w zależności od długości czasowej histogramu selektywnego typu *log* – badania prowadzono kolejno dla metryk: $\rho 1$, $\rho 2$, $\rho 3$ oraz $\rho 4$.

Pomimo sprowadzania sygnałów do tego samego poziomu energii nadal uzyskano dobre wyniki, co więcej w przypadku mówcy trzeciego nastąpiła zauważalna poprawa. Ma to silny związek z faktem, iż histogramy selektywne obliczane są na podstawie 'ściśniętego' sygnału mowy po eliminacji występujących w nim przerw (ogólnie fragmentów o małej energii). W związku z tym większy jest 'repertuar' mowy, na podstawie którego powstaje pojedynczy histogram, co umożliwia lepszą reprezentację danego głosu. Ponadto bierze się pod uwagę głoski o większej energii a więc bardziej odporne na szum towarzyszący mowie. Eliminacja przerw natomiast w dużej mierze prowadzi do uniezależnienie histogramu od tempa mowy co w przypadku małych wartości *d* jest tym bardziej uzasadnione bowiem trudno jest wówczas mierzyć tempo mowy. Poza nielicznymi wyjątkami również w tych badaniach spełnione są wnioski przytaczane w punkcie 5.4.2 choć można mieć wiele wątpliwości co do absolutnej przewagi metryki $\rho 1$ nad $\rho 2$ oraz $\rho 3$ nad $\rho 4$. Jakość wyników uzyskanych różnymi metrykami stała się wyraźnie zależna od głosu mówcy. Należy również zauważyć, iż została zachwiana absolutna przewaga wyników dla pierwszego mówcy w stosunku do trzeciego.

5.5. Opis metody na tle innych badań

Zaproponowana metoda jest tylko elementem szerokiego zakresu badań nad automatycznym rozpoznawaniem mówcy prowadzonych na całym świecie (przegląd można znaleźć w pracach [Ata76, Bas92, Cam97, Dod85, Fur89, Fur96, Jun98, Ros76, Ros91]). Aktualny stan tych badań przedstawiony jest w rozdziale trzecim rozprawy.

Spośród szerokiej gamy różnorodnych metod rozpoznawania mówcy zaproponowana metoda opiera się na wykorzystaniu nowych parametrów opisu sygnału mowy w postaci histogramu amplitudowego tego sygnału w dziedzinie czasu. Nie ulega wątpliwości, że jest to główny element odróżniający zaproponowaną metodę od innych, omawianych w literaturze. Opis ten jest jedną z form reprezentacji sygnału mowy w dziedzinie czasu odgrywającą w obecnym czasie zdecydowanie mniejszą rolę zarówno w systemach automatycznego rozpoznawania mowy jak i automatycznego rozpoznawania mówcy. Większość najnowszych metod opiera się bowiem na parametrach cepstralnych sygnału oraz na parametrach kodowania predykcyjnego LPC. Niektóre rozwiązania wykorzystują technikę predykcji liniowej LPC w celu wyznaczenia cepstrum sygnału. Jeśli uznać reprezentację sygnału mowy za pomocą współczynników predykcji liniowej (lub innych pochodnych) za opis w dziedzinie czasu[5] można stwierdzić, że reprezentacja tego sygnału w dziedzinie czasu jest nadal bardzo przydatna. Na uwagę zasługuje również fakt, iż cepstrum sygnału jest wyznaczane w dziedzinie będącej analogią dziedziny czasu[6].

Histogramy amplitudowe sygnału mowy można uznać za przybliżoną formę, przedstawionej w [Bie81] funkcji gęstości prawdopodobieństwa amplitud chwilowych sygnału mowy. Funkcja ta została wykorzystana w wymienionej pracy do celów statystycznego opisu rozkładu parametrów sygnału mowy.

Wśród parametrów opisujących sygnał mowy bezpośrednio w dziedzinie czasu i wykorzystywanych w systemach automatycznego rozpoznawania mówcy można wymienić: obwiednię amplitudową, rozkład czasowy energii sygnału oraz parametry przejść przez zero. Znaczna część badań wykorzystujących te ostatnie parametry została przeprowadzona w Polsce głównie przez profesora Baszturę [Bas78]. Rozwiązania

[5] Współczynniki predykcji liniowej mają swoją interpretację zarówno w dziedzinie czasu jak i częstotliwości. Mają one ponadto swoją interpretację fizyczną.

[6] Cepstrum sygnału wyznaczane jest w wyniku odwrotnej transformacji Fouriera widma sygnału po jego logarytmowaniu.

wykorzystujące między innymi rozkład energii sygnału można znaleźć w [Mar77, Mar79].

Należy podkreślić, że zaproponowany wektor cech w stosunku do innych, opisanych w literaturze, charakteryzuje się wyjątkową łatwością pomiaru. Ponadto istnieje możliwość bardzo prostej sprzętowej realizacji procesu jego ekstrakcji (roz. 4.5).

Algorytm klasyfikacji na podstawie potencjału *k* najbliższych sąsiadów *pkNN,* wykorzystany w zaproponowanej metodzie, stanowi oryginalną, zaproponowaną przez autora pracy metodę klasyfikacji. Stanowi ona hybrydowe połączenie metody *kNN* oraz mało znanej dziś metody funkcji potencjalnych, wprowadzonej w latach siedemdziesiątych przez Ajzermana, Brawermana i Rozenoera [Ajz76]. Problem klasyfikacji jest wspólny dla większości systemów rozpoznawania. W szczególności stosunkowo bogata literatura poświęcona jest zagadnieniu rozpoznawania obrazów [Ajz76, Dud73, Jaj93, Tad92].

Niektóre systemy automatycznego rozpoznawania mówcy wykorzystują tylko część informacji zawartej w sygnale mowy, mającej względnie większą energię. Koncepcja rozpoznawania polegająca na analizie tylko wybranych fragmentów sygnału mowy była już wcześnie zastosowana. Parametry opisu sygnału (wektory cech) są wtedy wydobywane tylko z wybranych fragmentów sygnału. Przykładowo określenie częstotliwości tonu krtaniowego dokonuje się na podstawie mowy dźwięcznej zaś obliczenie parametrów kodowanie predykcyjnego przeprowadza się często na podstawie samogłosek. Również mogą być eliminowane z sygnału fragmenty ciszy występujące między słowami a nawet sylabami. Badania wykazują [Fur94, Wal98], że przeprowadzenie rozpoznawania mówcy tylko na podstawie wybranych fragmentów sygnału mowy, charakteryzujących się względnie wyższym poziomem energii, prowadzi do znacznej poprawy jakości uzyskanych wyników. Tym niemniej w większości dotychczasowych prac analizowany jest cały przebieg sygnału mowy.

W literaturze opisane zostały rożne podejścia do zagadnienia rozpoznawania mówcy niezależnego od tekstu (patrz roz. 3.4). Zupełną nowość stanowi natomiast możliwość rozpoznawania głosu niezależnie od języka. Oznacza to, że możliwe jest przeprowadzenie rozpoznawania na podstawie wypowiedzi nagranych w jednym języku, nawet gdyby model głosu rozpoznawanej osoby został zbudowany z użyciem wypowiedzi rejestrowanych w innym języku.

ROZDZIAŁ VI

Omówienie eksperymentów i wyników

W niniejszym rozdziale zamieszczono wyniki eksperymentów podjętych w celu zbadania jakości rozpoznawania mówców na podstawie histogramów amplitudowych sygnału mowy. W celu rzetelnej oceny możliwości zaproponowanej metody rozpoznawania mówcy badania prowadzono z użyciem kilku baz danych głosów.

6.1. Badania wykorzystujące klasyczne metody klasyfikacji

W tym podrozdziale przedstawiono wyniki badań opartych na nagraniach głosów siedmiu mówców. Ocenę siły dyskryminacyjnej histogramów amplitudowych przeprowadzono za pomocą miary dyskryminacji *S* oraz dwóch tradycyjnych metod klasyfikacji: metody wzorców oraz metody *kNN*, przy czym we wszystkich trzech przypadkach jako miarę odległości między histogramami przyjęto metrykę Euklidesową. Badania zostały przeprowadzone przy założeniu, że zbiór rozpoznawanych klas jest zbiorem zamkniętym. Poniżej podano warunki przeprowadzenia eksperymentów (patrz tabela 4.4):

1- jakość sygnału wejściowego: *pok,*
2- przetwarzanie A/C: F_g=10kHz, F_p=22kHz, B=16 bitów,
3- bramka szumu: *nie zastosowano,*
4- przetwarzanie wstępne: N_E, *d=10sek,*
5- obliczenia histogramów: *hlog,* $L_b=31$,
6- selekcja histogramów: $o_p=0\%$,
7- metryka: *Euklidesowa*,
8- metoda klasyfikacji: *wzorców* lub *kNN,*
9- klasyfikacja grupowa: *Gq=2 do 8,*
10- procedura decyzyjna: *I,*
11- rozmiar populacji: *M=7,*
12- zbiór mówców: *Z,*
13- zależność od tekstu: *indep*.

6.1.1. Opis materiału fonetycznego, przetwarzanie danych wstępnych

Materiał fonetyczny składa się z nagrania głosów siedmiu osób (trzy kobiety, cztery mężczyźni). Osoby uczestniczące w doświadczeniu czytały teksty typu gazetowego o różnych długościach. Głosy męskie nagrano w dwóch różnych językach: polskim oraz arabskim. Częstotliwość próbkowania wynosiła 22050Hz a rozdzielczość amplitudowa 16 bitów. Głosy były zarejestrowane w warunkach pokojowych za

pomocą karty muzycznej komputera osobistego i mikrofonu oporowego o względnej dużej dynamice. Tabela 6.1 przedstawia dane wejściowe eksperymentu.

Nr. osoby	Płeć	Język
1	Żeńska	Polski
2	Żeńska	Polski
3	Żeńska	Polski
4	Męska	Polski, Arabski
5	Męska	Polski, Arabski
6	Męska	Polski, Arabski
7	Męska	Polski, Arabski

Tabela 6.1 : Dane wejściowe eksperymentu.

Dane wejściowe zostały rozdzielone w sposób ślepy na 8 fragmentów o długości ok. 10 sekund. Dla każdego fragmentu wykonano histogram z logarytmicznie podzielonym zakresem amplitud zgodnie z wektorem amplitud przedstawionym w rozdziale czwartym (tabela 4.1). Histogramy uzyskane na podstawie danych wejściowych zostały podzielone na 11 klas w zależności od mówcy i od języka, przy czym każda klasa zawierała 8 histogramów (obiektów). Tabela 6.2 przedstawia oznaczenia histogramów.

Nr klasy	Oznaczenie	Płeć	Język	Numery	Oznaczenie
1	Ag	Ż	P	1 – 8	Hagp1 – Hagp8
2	Iw	Ż	P	9 – 16	Hiwp1 – Hiwp8
3	Jo	Ż	P	17 – 24	Hjop1 – Hjop8
4	Ad	M	A	25 – 32	Hada1 – Hada8
5	Ad	M	P	33 – 40	Hadp1 – Hadp8
6	Os	M	A	41 – 48	Hosa1 – Hosa8
7	Os	M	P	49 – 56	Hosp1 – Hosp8
8	Fr	M	A	57 – 64	Hfra1 – Hfra8
9	Fr	M	P	65 – 72	Hfrp1 – Hfrp8
10	Tr	M	A	73 – 80	Htra1 – Htra8
11	Tr	M	P	81 – 88	Htrp1 – Htrp8
Uwagi: H: histogram, P: polski, A: arabski, łączna liczba histogramów wynosi 88.					

Tabela 6.2 : Numeracja klas i histogramów.

Na wstępie badano odległości histogramów w układzie ‘każdy z każdym’ i oceniono siłę dyskryminacji za pomocą miary *S*. Badania wykazały, że wartość stosunku *S* wynosi w najgorszym przypadku ok. 0.67 a w najlepszym ok. 0.015. Dalszą ocenę siły dyskryminacyjnej histogramów dokonano na podstawie dwóch tradycyjnych metod klasyfikacji a mianowicie metody wzorców oraz metody *kNN*.

6.1.2. Zastosowanie metody wzorców

Używając jako dane trenujące 6 histogramów z każdej z 11-tu klas obliczono histogram średni i przyjęto go za model (wzorzec) klasy. Celem weryfikacji modeli obliczono odległości dwóch pozostałych histogramów (dane testujące) od modeli poszczególnych klas. Tabela 6.3 przedstawia otrzymane wyniki:

Numer klasy	Oznaczenie histogramu	Model klasy 1	Model klasy 2	Model klasy 3	Model klasy 4	Model klasy 5	Model klasy 6	Model klasy 7	Model klasy 8	Model klasy 9	Model klasy 10	Model klasy 11
1	Hagp7	4	25	27	91	78	124	91	154	77	99	71
1	Hagp8	5	28	38	102	84	131	99	169	86	104	76
2	Hiwp7	16	2	12	77	81	95	67	124	66	70	68
2	Hiwp8	19	3	7	67	74	85	58	113	58	64	63
3	Hjop7	27	14	2	40	54	55	30	81	30	41	36
3	Hjop8	40	24	6	36	52	50	25	76	30	40	36
4	Hada7	85	78	60	4	23	14	12	38	17	22	67
4	Hada8	106	94	68	5	49	7	10	14	11	22	57
5	Hadp7	77	72	48	19	12	50	33	75	38	57	81
5	Hadp8	64	74	62	36	2	76	59	114	60	80	102
6	Hosa7	102	88	64	13	69	2	8	10	8	11	4
6	Hosa8	97	82	59	9	58	1	5	13	8	12	44
7	Hosp7	93	81	54	11	66	6	3	18	8	16	34
7	Hosp8	80	66	43	6	47	6	1	23	7	12	38
8	Hfra7	114	104	78	15	72	5	16	3	11	20	56
8	Hfra8	159	139	114	34	106	14	37	3	30	37	83
9	Hfrp7	62	64	43	13	51	13	11	25	3	14	27
9	Hfrp8	80	79	59	11	54	6	12	13	5	16	39
10	Htra7	113	93	70	22	78	12	17	21	21	8	73
10	Htra8	93	75	54	17	64	13	13	24	15	3	64
11	Htrp7	90	88	58	39	105	29	20	39	21	43	8
11	Htrp8	78	79	51	42	100	37	25	51	25	52	4

Tabela 6.3 : Weryfikacja modeli budowanych na podstawie sześciu histogramów.

Na podstawie otrzymanych wartości można wyciągać następujący wniosek: dla każdego modelu istnieje pewna odległość L rozdzielająca histogramy jego klasy od histogramów innych klas tak, że te pierwsze leżą bliżej własnego modelu. Przykładowe wartości odległości L dla poszczególnych klas są podane poniżej:

Numer klasy	1	2	3	4	5	6	7	8	9	10	11
L	6	4	7	6	13	3	4	4	6	9	3

Tym samym wykazano, że wszystkie badane histogramy leżą bliżej wzorca własnej klasy niż wzorców innych klas.

6.1.3. Zastosowanie metody *k* najbliższych sąsiadów *kNN*

W tym punkcie przedstawiono wyniki badań wykonanych z zastosowaniem klasycznego kryterium metody *kNN*, polegającego na klasyfikacji badanego obiektu w zależności od liczebności jego sąsiadów, przy czym aktualnie klasyfikowany histogram był wykluczony z ciągu uczącego. Tabela 6.4 przedstawia krotności wystąpienia numerów poszczególnych klas (nazywane dalej wskaźnikiem przynależności do klasy) dla tych samych histogramów z poprzedniego punktu (puste pola są zerami):

Numer klasy histogramu	Klasa #1	Klasa #2	Klasa #3	Klasa #4	Klasa #5	Klasa #6	Klasa #7	Klasa #8	Klasa #9	Klasa #10	Klasa #11
1	6		1								
1	6		1								
2		7									
2		6	1								
3		1	6								
3			6				1				
4				3	1	2	1				
4				4		1			2		
5				3	4						
5				1	6						
6						6	1				
6						5	2				
7				1		1	5				
7						2	5				
8						1		5	1		
8						1		6			
9				4*					3		
9						4*			3		
10						2				5	
10							1			6	
11							1				6
11											7

Tabela 6.4 : Krotności wystąpienia numerów poszczególnych klas.

Należy zauważyć, że generalnie rozpoznawanie jest poprawne, chociaż występują błędy w rozpoznawaniu dla klasy nr 9. Błędy te można korygować w różny sposób np. zmieniając wartości parametru *k*. Tabela 6.5 przedstawia krotności wystąpienia numerów poszczególnych klas przy różnych wartościach parametru *k* dla poprzednio rozpatrywanych dwóch histogramów z klasy nr 9:

Wartość parametru *k*	Klasa #1	Klasa #2	Klasa #3	Klasa #4	Klasa #5	Klasa #6	Klasa #7	Klasa #8	Klasa #9	Klasa #10	Klasa #11
1									1		
1									1		
3				1					2		
3						1			2		
4				1					3		
4						2			2		

Tabela 6.5 : Krotności wystąpienia numerów poszczególnych klas dla różnych wartości *k*.

6.1.4. Wpływ długości wypowiedzi na wynik rozpoznawania

Wpływ długości wypowiedzi na wynik rozpoznawania przedstawiono dla histogramów klasy nr 9 ze względu na to, że wyniki dla tej klasy były najgorsze. W tym celu obliczono wskaźniki przynależności poszczególnych 8-miu histogramów tej klasy przy założeniu, że wartość parametru *k* wynosi 7 (jako dane trenujące przyjęto wszystkie obiekty klasy 9 z wyłączeniem aktualnie klasyfikowanego obiektu). Na podstawie otrzymanych wyników przedstawionych w tabelce 6.6 można stwierdzić, że podjęcie decyzji na podstawie sumy wskaźników odpowiadających dłuższej wypowiedzi jest łatwiejsze i co najważniejsze – trafniejsze, niż podjęcie decyzji na podstawie wskaźnika przynależności dla pojedynczego histogramu (krótszej wypowiedzi).

Numer klasy histogramu	Klasa #1	Klasa #2	Klasa #3	Klasa #4	Klasa #5	Klasa #6	Klasa #7	Klasa #8	Klasa #9	Klasa #10	Klasa #11
1						1	1	1	4		
2						1	2		4		
3				2			1		4		
4				2			1		4		
5				3		1		2	1		
6						1	2		4		
7				4					3		
8						4			3		
Suma				11		8	7	3	27		

Tabela 6.6 : Poprawa jakości rozpoznawania dla dłuższych wypowiedzi.

6.1.5. Wpływ języka na jakość rozpoznawania – badania w zamkniętym zbiorze

Na podstawie wyników otrzymanych w punkcie 6.1.3 można stwierdzić, że o ile w przypadku niektórych osób istnieje sprzężenie pomiędzy modelami ich głosu dla różnych języków (np. osoba 4, 5) o tyle w przypadku innych (np. osoba 6, 7) takiego sprzężenia nie stwierdzono. Tak więc z badań wynika, że wskazane jest przyjęcie

dwóch różnych modeli dla tej samej osoby w zależności od języka wypowiedzi. W celu bliższego zbadania zależności wyników rozpoznawania od języka w przypadku istnienia sprzężenia wykonano następujący eksperyment. Dla osoby nr 4 (klasa 4 – język arabski, klasa 5 – język polski) nagrano 16 wypowiedzi każda o długości ok. 10 sekund w trzech różnych językach (polski, arabski oraz angielski) i wykonano odpowiedni histogram dla każdej z nich. Histogramy te oznaczono A1-A16. Podjęto niezależną próbę rozpoznawania każdego z nich za pomocą metody klasyfikacji *kNN* przy *k=8*. Za obiekty wzorcowe mówcy # 4 przyjęto osiem histogramów użytych w poprzednich punktach. Tabela 6.7 przedstawia wskaźniki przynależności dla poszczególnych histogramów A1-A16 (puste pola są zerami).

Histogram	Język	Klasa #1	Klasa #2	Klasa #3	Klasa #4	Klasa #5	Klasa #6	Klasa #7	Klasa #8	Klasa #9	Klasa #10	Klasa #11
A1	Arabski				1	7						
A2	Arabski				1	7						
A3	Arabski				1	7						
A4	Arabski				1	7						
A5	Polski				1	7						
A6	Polski				3	5						
A7	Polski				4	3		1				
A8	Polski				6	2						
A9	Polski				6	2						
A10	Polski				1	7						
A11	Angielski				4	3		1				
A12	Angielski				3	5						
A13	Angielski				4	4						
A14	Angielski				5	3						
A15	Angielski				2	6						
A16	Angielski				5	3						
Suma		0	0	0	48	78	0	2	0	0	0	0

Tabela 6.7 : Wpływ języka na wyniki rozpoznawania.

Z tabeli 6.7 widać, że wszystkie pierwsze 4 histogramy, odpowiadające fragmentom wypowiedzi nagranym w języku arabskim były bliższe klasie 5 odpowiadającej językowi polskiemu. Odwrotną relację można stwierdzić dla histogramów 8, 9. Pomimo braku modelu dla języka angielskiego wszystkie histogramy odpowiadające wypowiedziom nagranym w języku angielskim zostały zdecydowanie dobrze zaliczone do klasy czwartej bądź piątej czyli do klas odpowiadających właściwej osobie mówiącej.

6.2. Rozpoznawanie na podstawie selektywnych histogramów z zastosowaniem metody klasyfikacji potencjału najbliższych sąsiadów *pkNN*

W niniejszym podrozdziale przedstawiono wyniki badań opartych na nagraniach głosów 21 mówców. Ocenę siły dyskryminacyjnej histogramów amplitudowych przeprowadzono za pomocą różnych błędów rozpoznawania. Jako miarę odległości między histogramami przyjęto metrykę Euklidesową lub metrykę Camberra. Badania zostały przeprowadzone dla otwartego oraz zamkniętego zbioru rozpoznawanych mówców. Poniżej podano warunki przeprowadzenia eksperymentów opisanych w tym podrozdziale:

1- jakość sygnału wejściowego: *pok,*
2- przetwarzanie A/C: F_g=10kHz, F_p=22kHz, B=16 bitów,
3- bramka szumu: *różne wartości parametrów bramki (patrz tabela 6.9),*
4- przetwarzanie wstępne: N_E, *d=1sek,*
5- obliczenie histogramów: *hlog,* L_b*=31*,
6- selekcja histogramów: o_p*=5%,*
7- metryka: *Euklidesowa lub Camberra*
8- metoda klasyfikacji: *pkNN, zakres zmienności parametru k: od 1 do 20.*
9- klasyfikacja grupowa: *Gq=2 do 25,*
10- procedura decyzyjna: *I,*
11- rozmiar populacji: *M=21,*
12- zbiór mówców: *O,*
13- zależność od tekstu: *indep*.

6.2.1. Materiał fonetyczny

Opisany poniżej eksperyment jest oparty na sesji nagrań dla 21 osób. Biorąc pod uwagę, że zaproponowana metoda charakteryzuje się dużym stopniem niezależności jakości wyników od języka [Sho98, Sho99], dla niektórych osób zarejestrowano nagrania w dwóch lub trzech różnych językach. Nagrania zostały zarejestrowane za pomocą karty muzycznej komputera osobistego w warunkach pokojowych (poziom szumu tła około 20 dB). Częstotliwość próbkowania wynosiła 22kHz a rozdzielczość amplitudowa (liczba bitów/próbkę) 16 bitów. W celu wyeliminowania wpływu wzmocnienia kanału dokonano unormowania wartości średniej amplitud sygnałów. Zgodnie z wynikami badań zrealizowanymi w pracy [Sho98] histogramy amplitudowe dla sygnałów wykonano dla pełnego zakresu amplitud rozdzielonego w sposób logarytmiczny na 31 poziomów. Tabela 6.8 przedstawia ogólną charakterystykę wyników, podając spis osób

oraz odpowiadające im numery klas a także przybliżone dane o długości nagrań. Na uwagę zasługuje fakt, iż wiek rozważanych mówców jest zbliżony (22 do 28 lat). Ponadto mówcy o numerach: 1, 2 oraz 8 to trzej bracia[1], których głosy są z trudem rozróżniane nawet przez osoby dobrze ich znające.

Klasa	Płeć	Etykieta	Języki	Długość [sek]	Klasa	Płeć	Etykieta	Języki	Długość [sek]
1	M	Adi	Ar	182	12	M	Osm	Ar, Pl, Hb	270
2	M	Adl	Ar, An, Pl	167	13	M	Slm	Ar, Pl	198
3	Ż	Afa	Ar	96	14	Ż	Aga	Pl	196
4	M	Aym	Ar, Pl	179	15	M	Mrn	Pl	101
5	M	Fyz	Ar, Pl	180	16	M	Bgd	Pl	105
6	M	Hdi	Ar, Pl	211	17	M	Drk	Pl	97
7	M	Hsn	Ar	94	18	M	Ptr	Pl	98
8	M	Ism	Ar	197	19	M	Tmk	Pl	95
9	M	Iyd	Ar, Pl	192	20	Ż	Iwo	Pl	131
10	M	Mhd	Ar, Pl	247	21	Ż	Jol	Pl	235
11	M	Nsr	Ar, Pl	164	Ar: arabski, An: angielski, Hb: hebrajski, Pl: polski				

Tabela 6.8: Spis mówców, ich etykiety, języki oraz długości nagrania.

6.2.2. Wstępne przetwarzanie danych wejściowych

Zarejestrowany sygnał wejściowy po wstępnym unormowaniu poziomu energii został poddany operacji mającej na celu wyeliminowanie z niego fragmentów charakteryzujących się względnie niższym poziomem energii (operacja selekcji zdarzeń głosowych). W tym celu przepuszczono go przez bramkę szumu. Otrzymany na wyjściu bramki ‘ściśnięty’ sygnał podlegał kolejnej operacji normalizacji amplitudy polegającej na dzieleniu wartości jego chwilowych amplitud przez ich bezwzględną średnią wartość[2] i mnożeniu ich razy stałą wartość określającą, z góry zakładany, poziom energii sygnału. Unormowana wersja sygnału była dalej dzielona na kolejne fragmenty o długości jednej sekundy i dla każdego takiego fragmentu obliczono histogram amplitudowy, który stanowił pojedynczy obiekt ciągu uczącego.

Eksperymenty z identyfikacją mówców zostały wykonane dla różnych wartości parametrów bramki szumu. W celu adaptacji parametrów bramki do poziomu energii

[1] Autor rozprawy i jego dwaj bracia.
[2] Bezwzględna wartość średnia była obliczana dla wyjściowego sygnału bramki.

wejściowego sygnału dokonano normalizacji wartości jej progu w stosunku do bezwzględnej średniej amplitudy tego sygnału[3] oznaczonej symbolem μ. Oczywistym jest, że im krótsza jest długość czasowa bramki lub większa jest wartość jej progu, tym więcej fragmentów sygnału zostaje 'wyciętych'. Tabela 6.9 przedstawia użyte parametry bramki oraz odpowiadające im średnie wartości stosunku δ (stosunek łącznej długości odrzucanych z sygnału fragmentów, w wyniku jego przepuszczenia przez bramkę szumu, do długości sygnału wejściowego bramki).

Oznaczenia sesji badań	Wartość progu bramki	Długość czasowa bramki	Wartość δ
A	Zero	----	0.00
B	$\mu / 2$	20 ms	0.31
C	μ	100 ms	0.22
D	μ	10 ms	0.38
E	$1{,}3.\ \mu$	10 ms	0.42
F	$2.\mu$	20 ms	0.49

Tabela 6.9: Użyte w badaniach wartości parametrów bramki szumu.

6.2.3. Opis przeprowadzonych eksperymentów

Badania wykonano dla różnych kombinacji następujących warunków przeprowadzenia identyfikacji:

- sześciu różnych sesji badań w zależności od parametrów zastosowanej bramki szumu,
- czasowej długości identyfikowanych próbek (do 25 sekund),
- wartości parametru *k* metody *pkNN* ze zakresu 1 do 20,
- dwóch metryk odległości między histogramami: Euklidesowej lub Camberra,
- otwartego lub zamkniętego zbioru rozpoznawanych klas.

Oceny jakości identyfikacji dokonano na podstawie błędów identyfikacji, przy czym wyodrębniono następujące rodzaje błędów:

- błąd odrzucania własnych obiektów przy zapewnieniu zerowego błędu akceptacji obcych obiektów,

[3] Oznacza to, że w tym wypadku bezwzględna wartość średnia była obliczana dla wejściowego sygnału bramki.

- błąd akceptacji obcych obiektów przy zapewnieniu zerowego błędu odrzucania własnych obiektów,
- minimalna, względem rozmiaru klasy, suma błędów akceptacji i odrzucania,
- błąd identyfikacji w przypadku zamkniętego zbioru rozpoznawanych klas[4].

Pierwsze trzy błędy użyto do oceny jakości identyfikacji w przypadku otwartego zbioru identyfikowanych klas, zaś czwarty w przypadku zamkniętego zbioru.

6.2.4. Wyniki badań dla otwartego zbioru identyfikowanych klas

Opisane w niniejszej pracy badania identyfikacji mówców w zbiorze otwartym zostały wykonane zgodnie z dwuetapowym algorytmem identyfikacji mówcy w zbiorach otwartych, zaprezentowanym w rozdziale trzecim (patrz rys. 3.2). Pierwszy etap identyfikacji przeprowadzono metodą *pkNN* dla wartości *k=1*. W drugim etapie zweryfikowano trzech mówców, którzy uzyskali najlepsze wyniki w pierwszym etapie. Weryfikację każdego z tych mówców przeprowadzono dla wartości parametru *k* preferencyjnej dla niego tzn. zapewniającej małe błędy weryfikacji. W tym celu dokonano ich weryfikacji dla wartości *k=1* do *k=20* i wybrano najlepsze wyniki odpowiadające jednej optymalnej dla konkretnego mówcy wartości parametru *k*. Badania wykazały, że poza nielicznymi przypadkami optymalne wartości parametru *k* to *k=1* lub *k=2*.

6.2.4.1. Błąd odrzucania własnych obiektów przy zapewnieniu zerowego błędu akceptacji obcych obiektów

Dla danej klasy określenie błędu odrzucania własnych obiektów przy zapewnieniu zerowego błędu akceptacji obcych obiektów polega na takim ograniczeniu rozmiaru klasy w przestrzeni, żeby nie obejmował on obcych obiektów i znalezieniu procentu jej własnych obiektów, które wobec takiego ograniczenia znalazły się poza jej obszarem. Badania wykonano dla różnej długości wypowiedzi na podstawie której podejmuje się decyzje o rozpoznawaniu, przy czym długość ta stanowiła wielokrotność podstawowej długości odpowiadającej pojedynczemu obiektowej (jedna sekunda). Identyfikację z użyciem dłuższej wypowiedzi, czyli kilku obiektów, dokonano na podstawie średniej ich indywidualnych wartości funkcji przynależności (identyfikacja grupowa). Innymi

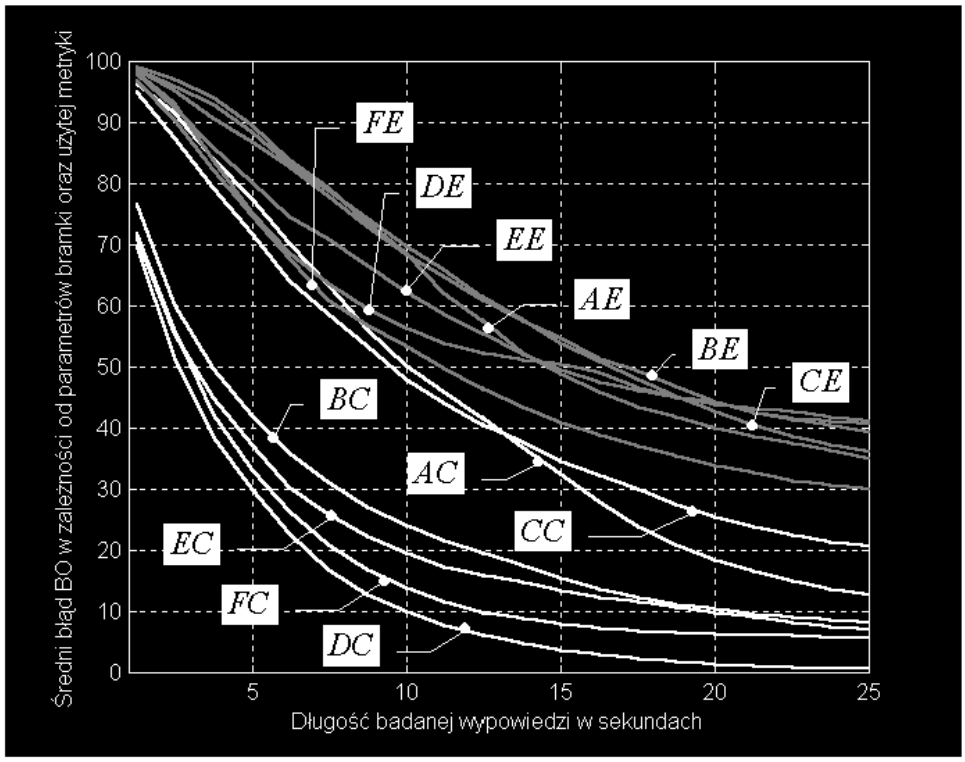
[4] Dla systemu z zamkniętym zbiorem rozpoznawany klas odrzucanie danego obiektu przez własną klasę oznacza automatycznie, że nastąpiła również błędna jego akceptacja przez obcą klasę. A więc te dwa błędy są równoważne.

słowy, tak obliczoną średnią przyjęto jako wartość funkcji przynależności dla całej wypowiedzi (dla grupy obiektów).

W zależności od parametrów bramki szumu oraz użytej metryki odległości między histogramami badania wykonano dla 12 różnych przypadków identyfikacji. Użyto metryki Camberra oraz Euklidesowej oznaczonych kolejno literkami *C* oraz *E*. W celu skrótowego opisu warunków przeprowadzenia rozpoznawania użyto koncepcji ich oznaczenia za pomocą dwóch liter z których pierwsza odnosi się do sesji badań (parametrów bramki – patrz tabela 6.9) zaś druga określa zastosowaną metrykę odległości między histogramami. Przykładowo skrót *BC* oznacza, że identyfikacja została przeprowadzona dla progu bramki szumu wynoszącego $\mu/2$, długości czasowej bramki wynoszącej 20 ms oraz, że zastosowano metrykę Camberra jako miarę odległości między histogramami.

Rysunek 6.1 przedstawia 12 wykresów błędów w funkcji długości badanej wypowiedzi. Są to wartości błędów uśrednianych dla całej populacji (21 mówców) i odpowiadających wyżej wymienionym 12 przypadkom przeprowadzenia identyfikacji (6 różnych sesji x 2 metryki). Wykresy 'białe' odpowiadają metryce Camberra zaś 'szare' metryce Euklidesowej. Można na rysunku zaobserwować wyraźną przewagę tej pierwszej nad drugą. Jeśli chodzi o zależność wyników od parametrów bramki szumu to dalsze szczegółowe badania, prowadzone oddzielnie dla każdego z mówców wykazały, że optymalny dobór tych parametrów zależy od rozpatrywanego mówcy. Najlepsze wyniki uzyskano dla sesji *D*, *E* oraz *F*, jakość wyników dla tych sesji była bardzo zbliżona. Najgorsze okazały się wyniki z sesji *C* (parametry bramki $\mu/2$, 100 ms), wyniki dla sesji *B* były na ogół lepsze niż te dla sesji *A*.

Chcąc poprawić jakość identyfikacji dokonano próby jej przeprowadzenia w następujący sposób. We wstępnym etapie przeprowadzono identyfikację, przy założeniu zamkniętości zbioru rozpoznawanych mówców, dla parametrów bramki z sesji *F* oraz metryki Camberra, przy czym wartość parametru *k* metody była równa jeden na tym etapie. W drugim etapie dokonano weryfikacji trzech mówców, dla których wyniki identyfikacji z pierwszego etapu były najlepsze, przy czym w odróżnieniu od badań przeprowadzonych wcześnie weryfikacji danego mówcy dokonano także na preferencyjnych dla niego wartościach parametrów bramki szumu. Rysunek 6.2 przedstawia poszczególne błędy odrzucania dla 21 mówców przy zapewnieniu zerowego błędu akceptacji obcych obiektów. Można zaobserwować, że najgorsze wyniki uzyskano kolejno dla mówców #4, #12 i #16.



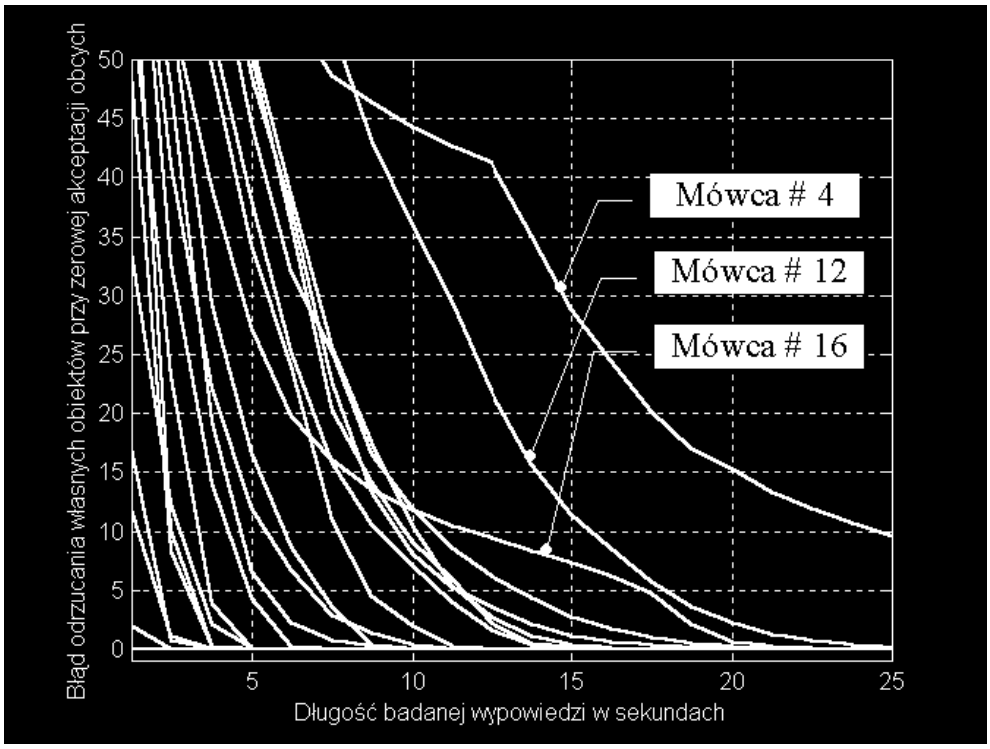
Rys. 6.1: Zależność średniego błędu odrzucania własnych obiektów przy zerowym błędzie akceptacji obcych obiektów od parametrów bramki szumu oraz metryki odległości między histogramami.

Badania dowiodły, że dla zdecydowanej większości mówców optymalne[5] warunki przeprowadzenia identyfikacji są następujące:

1) wartości parametru *k* metody *pkNN*: *k=1* oraz *k=2*,
2) parametry bramki szumu: próg μ = 1-2, czasowa długość rzędu 10-20 ms,
3) miara odległości: metryka Camberra.

Badania wskazują również na bardzo duży stopień niezależności w/w optymalnych warunków od długości badanej wypowiedzi. Ponadto wykresy błędów wskazują na systematyczne ich malenie w miarę wzrostu długości wypowiedzi, na podstawie której podejmowana jest decyzja o identyfikacji danego mówcy. Ponieważ błędy odrzucania obliczono dla zerowego błędu akceptacji oznacza to, że gdy wartość błędu odrzucania równa jest zeru to tym samym dokładność klasyfikacji jest stu procentowa.

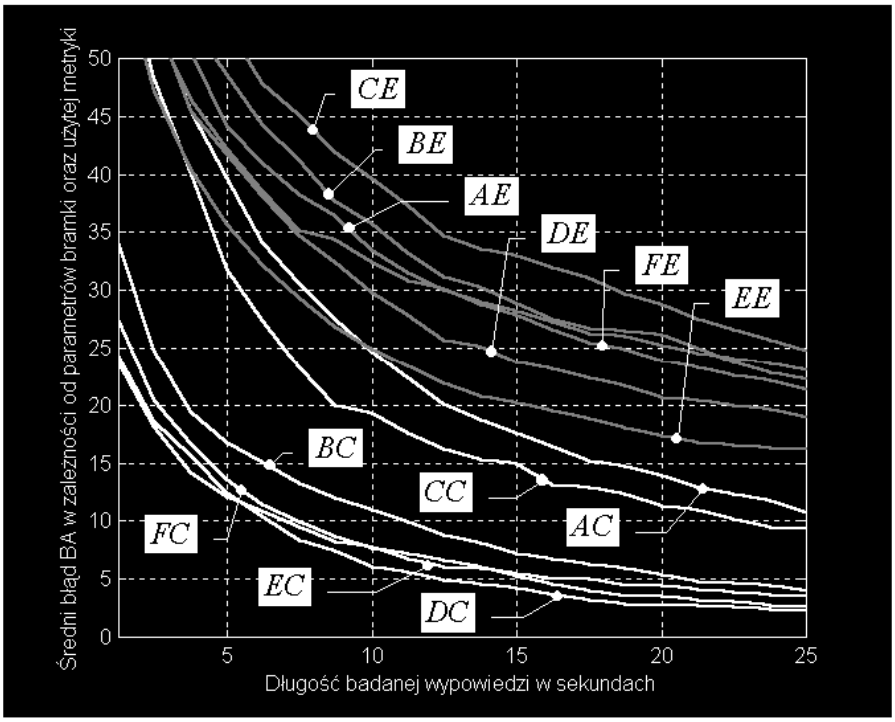
[5] Kryterium optymalności w tym wypadku jest minimalizacja błędnego odrzucania własnych obiektów przy zerowej akceptacji obcych obiektów.



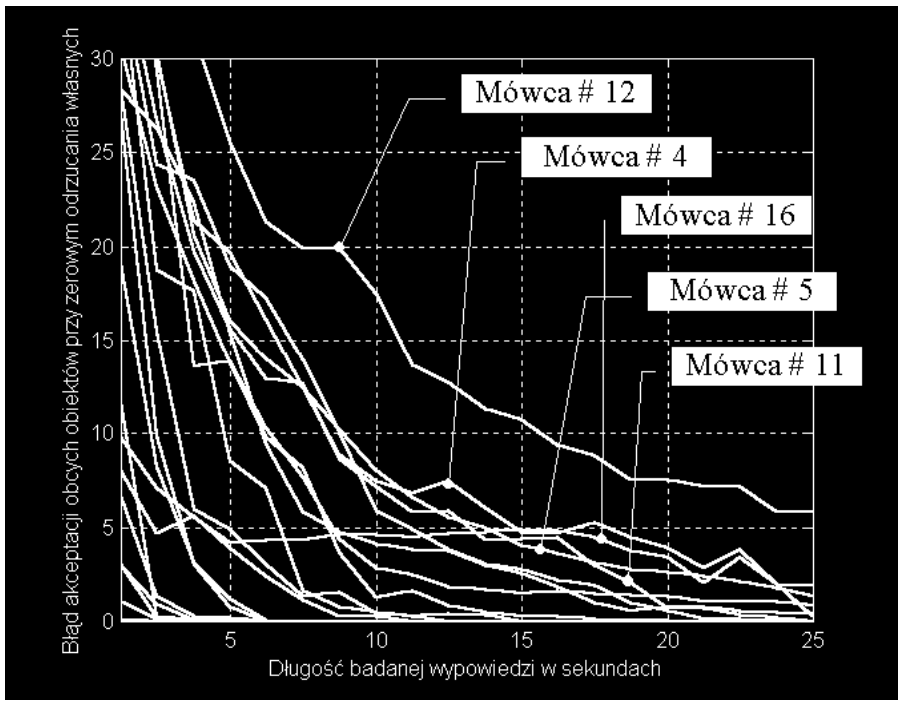
Rys. 6.2: Błędy odrzucania własnych obiektów przy zerowym błędzie akceptacji obcych obiektów dla poszczególnych 21 mówców

6.2.4.2. Procent błędnej akceptacji obcych obiektów przy zerowym błędzie odrzucania własnych obiektów

Analogicznie do badań przeprowadzonych w punkcie 6.2.4.1 dokonano także próby oceny jakości identyfikacji na podstawie błędnej akceptacji obcych obiektów przy zapewnieniu zerowego odrzucania własnych obiektów. Procentowe wartości tych błędów przedstawiono na rysunkach 6.3 oraz 6.4 odpowiadających analogicznym rysunkom 6.1 oraz 6.2. Najlepsze wyniki uzyskano dla sesji *D*, *E* oraz *F*, jakość wyników dla tych sesji była bardzo zbliżona. Biorąc pod uwagę wyniki dla metryki Camberra najgorsze okazały się wyniki z sesji *A* (bez zastosowania bramki szumu), wyniki dla sesji *B* były znacznie lepsze niż te dla sesji *C*. Rysunek 6.4 przedstawiający procenty poszczególnych błędów akceptacji dla 21 mówców przy zapewnieniu zerowego błędu odrzucania własnych obiektów pokazuje, że najgorsze wyniki uzyskano kolejno dla mówców #12, #4, #16, #5 oraz #11.



Rys. 6.3: Zależność średniego błędu akceptacji obcych obiektów przy zerowym błędzie odrzucania własnych obiektów od parametrów bramki szumu oraz metryki między histogramami.

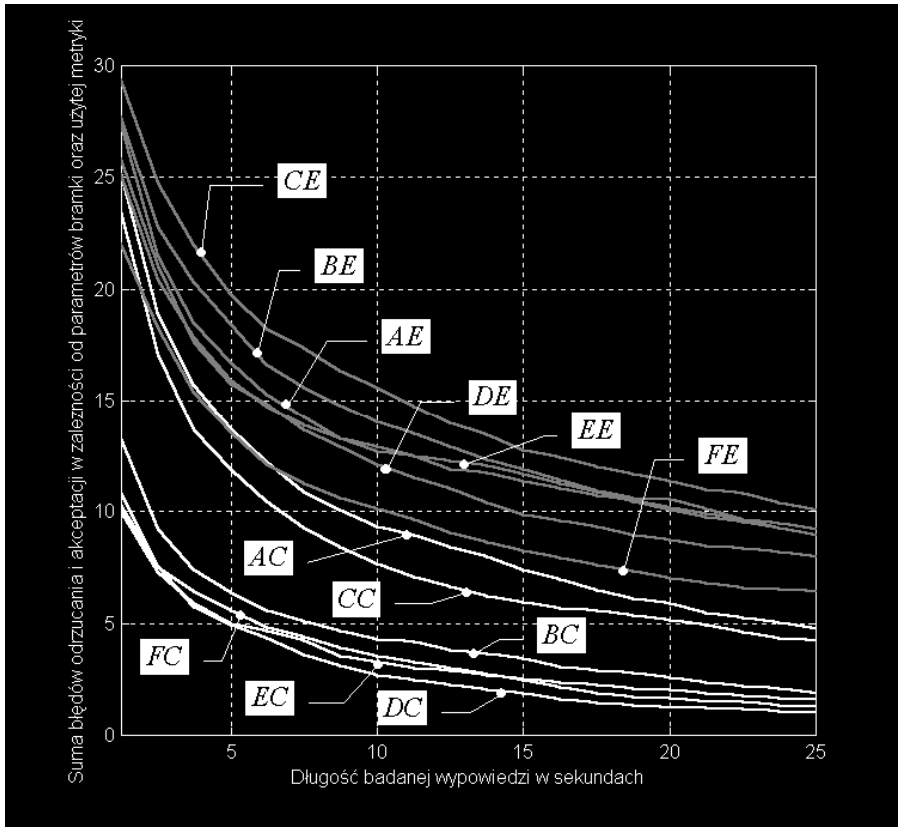


Rys. 6.4: Błędy akceptacji obcych obiektów przy zerowym błędzie odrzucania własnych obiektów dla poszczególnych 21 mówców.

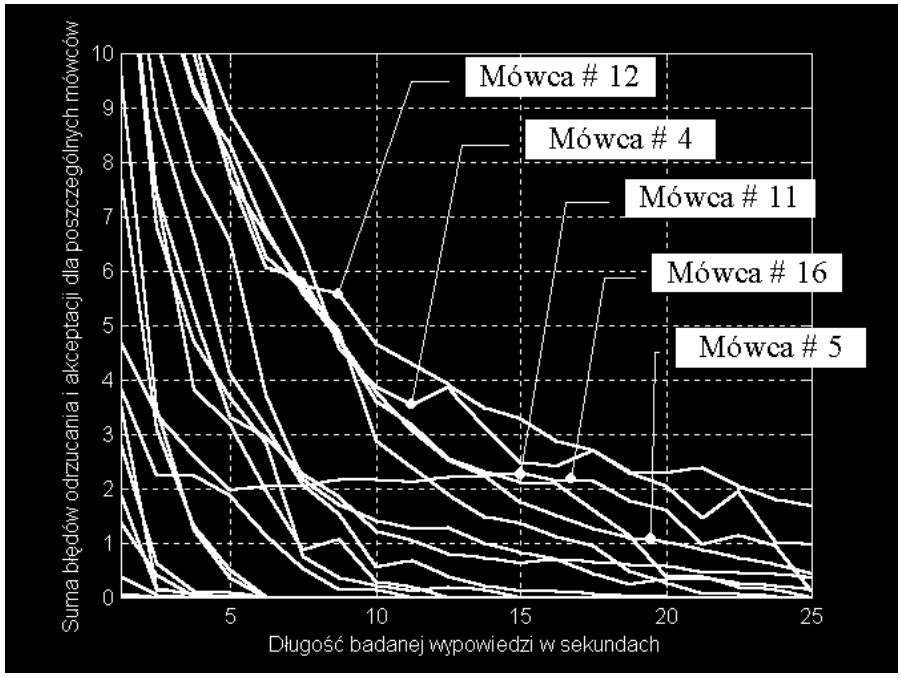
Badania dowodzą, że optymalne warunki identyfikacji, w przypadku gdy kryterium optymalności jest minimalizacja błędu akceptacji, są podobne do tych dla przypadku, gdy kryterium optymalizacji jest minimalizacja błędnego odrzucania. Badania wskazują również na bardzo duży stopień niezależności warunków optymalnej identyfikacji od długości badanej wypowiedzi. Wykresy błędu akceptacji dla poszczególnych mówców (rys. 6.4), w odróżnieniu od ich odpowiedników dla błędu odrzucania (rys. 6.2), wskazują na słabą systematyczność malenia błędu dla rosnących długości wypowiedzi, na podstawie której dokonana jest decyzja o identyfikacji danego mówcy. Niezależnie od sposobu oceny jakości identyfikacji najgorsze wyniki uzyskano, między innymi, dla tej samej trójki mówców: #12, #4 i #16.

6.2.4.3. Suma błędów odrzucania i akceptacji

W tym punkcie podjęto próbę oszacowania jakości identyfikacji na podstawie sumy błędów akceptacji i odrzucania. Próby identyfikacji kolejnych mówców przeprowadzono w warunkach dla nich optymalnych. Oznacza to, że proces identyfikacji jest w tym wypadku równoważny kolejnym weryfikacjom wszystkich mówców należących do populacji. Suma błędu odrzucania i akceptacji jest zależna od progu zaliczania obiektów do klasy, wyrażonego jako minimalna wartość funkcji przynależności uprawniającej do uznania danego obiektu jako należącego do danej klasy (tzw. próg detekcji lub akceptacji). Próg ten można jednoznacznie powiązać z rozmiarem obszaru reprezentującego daną klasę w przestrzeni, przy czym im wyższa jest jego wartość tym mniejszy jest rozmiar tego obszaru. Przy takiej definicji progu wzrost jego wartości oznacza zwiększanie prawdopodobieństwa wystąpienia błędnego odrzucania przy zmniejszaniu prawdopodobieństwa wystąpienia błędu akceptacji. W tym punkcie dla poszczególnych mówców dokonano minimalizacji sumy błędów odrzucania i akceptacji względem wartości progu funkcji przynależności odpowiadających im klas w przestrzeni cech. Poniżej przedstawiono dwa rysunki 6.5, 6.6 analogiczne do tych przedstawionych w punktach 6.2.4.1 oraz 6.2.4.2.



Rys. 6.5: Zależność sumy błędów akceptacji i odrzucania od parametrów bramki szumu oraz metryki odległości między histogramami.



Rys. 6.6: Sumy błędów akceptacji i odrzucania dla poszczególnych 21 mówców.

6.2.5. Wyniki identyfikacji dla zamkniętego zbioru mówców

W opisanych powyżej rozważaniach przyjęto, że zbiór identyfikowanych klas jest zbiorem otwartym. Dlatego też wymagano aby kryterium klasyfikacji nakładało ograniczenia co do minimalnego akceptowalnego stopnia podobieństwa próbki do wzorca[6], co wpływało zarówno na przebieg klasyfikacji jak i na wyniki. W tym punkcie przedstawiono wyniki klasyfikacji uzyskane dla domkniętego zbioru klas przy wykluczeniu możliwości odrzucania obiektu przez wszystkie klasy. Oznacza to, że przestrzeń cech może być całkowicie podzielona pomiędzy rozpatrywane klasy i nie występują w niej obszary nie reprezentujące żadnej klasy. Wbrew oczekiwaniom, w literaturze sprawdzanie jakości metod identyfikacji głosów jest częściej przeprowadzane na podstawie właśnie takiego podziału przestrzeni, chociaż prowadzi to do nieco „zawyżonych" wyników. Dlatego też w celu umożliwienia w miarę rzetelnego porównania zaproponowanej tu metody z innymi metodami opisanymi w literaturze podjęto w tej pracy także próbę oceny jakości identyfikacji przy założeniu wyłączności istnienia ograniczonej liczby klas.

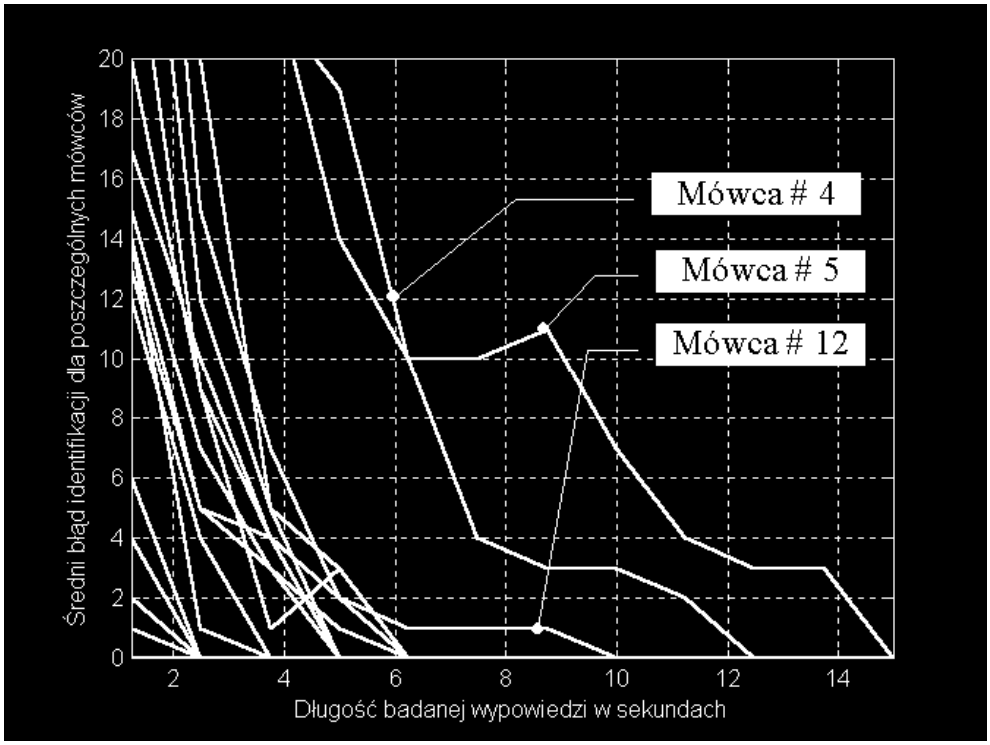
Poniżej podano warunki przeprowadzenia opisanych eksperymentów:

1- jakość sygnału wejściowego: *pok,*
2- przetwarzanie A/C: F_g=10kHz, F_p=22kHz, B=16 bitów,
3- bramka szumu: *różne wartości parametrów bramki (patrz tabela 6.9),*
4- przetwarzanie wstępne: N_E, *d=0.5 sek,*
5- obliczenie histogramów: *hlog,* $L_b=31$,
6- selekcja histogramów: $o_p=0\%$,
7- metryka: *Euklidesowa lub Camberra*
8- metoda klasyfikacji: *pkNN, zakres zmienności parametru k: od 1 do 20.*
9- klasyfikacja grupowa: *Gq=2 do 30,*
10- procedura decyzyjna: *I,*
11- rozmiar populacji: *M=21,*
12- zbiór mówców: *Z,*
13- zależność od tekstu: *indep*.

Badania wykazały, iż możliwe jest uzyskanie 100% dokładności identyfikacji nieomal dla każdej z rozpatrywanych wartości parametru k, co zdecydowanie nie zacho-

[6] Minimalna wartość funkcji przynależności.

dziło przy założeniu, że zbiór rozpoznawanych głosów jest zbiorem otwartym. Wstępne badania w przypadku otwartego zbioru rozpoznawanych klas także wskazywały na zdecydowaną przewagę metryki Camberra nad Euklidesowej. Ponadto najlepsze wyniki, podobnie jak w przypadku otwartego zbioru rozpoznawanych klas, uzyskano dla sesji *D*, *E* oraz *F*. Rysunek 6.7 przedstawia procent błędów identyfikacji dla poszczególnych 21 mówców w zależności od długości badanej wypowiedzi.



Rys. 6.7: Błędy identyfikacji dla poszczególnych mówców w przypadku zamkniętego zbioru rozpoznawanych klas.

Należy podkreślić, że nastąpiła ogólna poprawa jakości identyfikacji dzięki założeniu o istnieniu tylko 21 klas (zamknięty zbiór identyfikowanych klas). W szczególności możliwe było uzyskanie zerowych wartości błędów (100% jakość identyfikacji) dla znacznie krótszej wypowiedzi niż w przypadku otwartego zbioru identyfikowanych klas. Najgorsze wyniki były dla mówców #5, #4 i #12. Może się błędnie wydawać, że uzyskane wyniki odpowiadają najbardziej optymalnemu sposobowi

rozdzielania przestrzeni cech między klasami[7], jednak lepszym podziałem przestrzeni cech byłby taki, w którym rozmiary obszarów odpowiadających poszczególnym klasom są zależne od stopnia rozproszenia obiektów w tych klasach. Świadczy o tym chociażby fakt, iż w przypadku niektórych mówców wyniki identyfikacji dla otwartego zbioru rozpoznawanych mówców były lepsze niż dla zbioru zamkniętego.

6.3. Wpływ języka wypowiedzi – badania w zbiorze otwartym

W podrozdziale 6.1.5 badaniom poddano możliwości identyfikacji mówcy, gdy jego model jest stworzony na podstawie wypowiedzi nagrane w innym języku niż tym używanym w fazie testowania. W niniejszym podrozdziale podjęto taką samą próbę identyfikacji lecz przy założeniu, że zbiór identyfikowanych mówców jest zbiorem otwartym. Ponadto badania prowadzono z użyciem 21 osobowej bazy głosów opisanej w podrozdziale 6.2.1. Poniżej podano warunki przeprowadzenia opisanych eksperymentów:

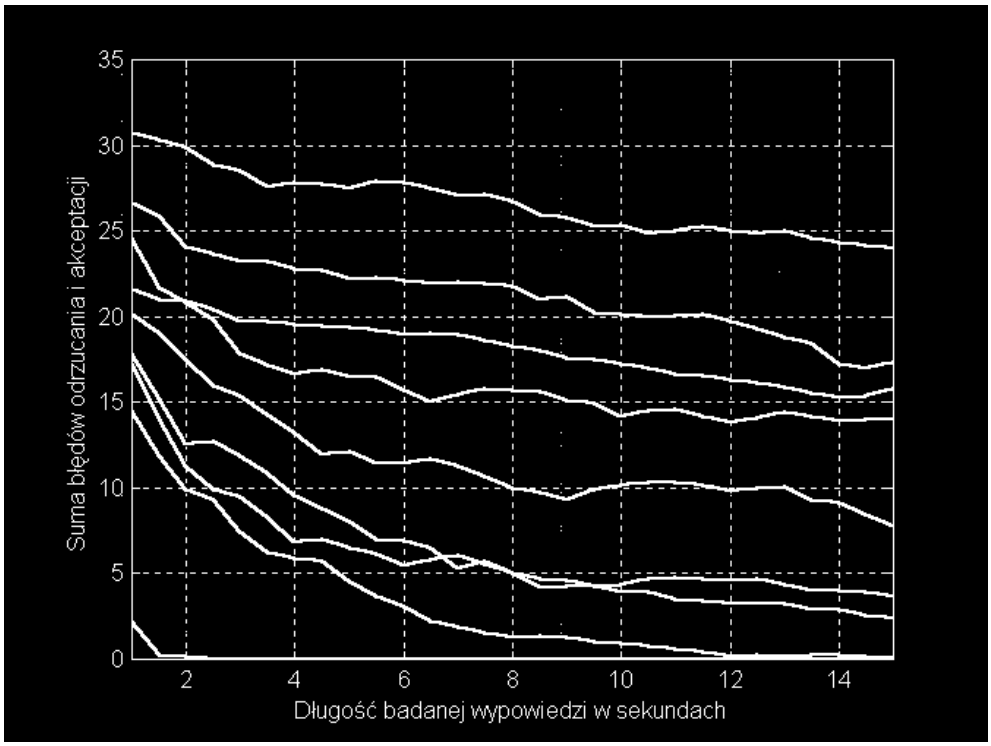
1- jakość sygnału wejściowego: *pok,*
2- przetwarzanie A/C: F_g=10kHz, F_p=22kHz, B=16 bitów,
3- bramka szumu: $P_b=\mu$, $t_b=10ms$,
4- przetwarzanie wstępne: N_E, *d=0.5sek oraz d=2sek,*
5- obliczenie histogramów: *hlog,* $L_b=31$,
6- selekcja histogramów: $o_p=0\%$,
7- metryka: *Camberra,*
8- metoda klasyfikacji: *pkNN, zakres zmienności parametru k od 1 do 20,*
9- klasyfikacja grupowa: *Gq=2 do 30 dla d=0.5 sek, Gq= 2 do 7 dla d=2 sek,*
10- procedura decyzyjna: *W,*
11- rozmiar populacji: *M=21,*
12- zbiór mówców: *O,*
13- zależność od tekstu: *indep*.

Badania zostały przeprowadzone dla mówców o numerach[8]: #1, #4, #5, #6, #9, #10, #11, #12, #13. Jako dane trenujące przyjęto histogramy uzyskane z wypowiedzi nagranych w języku arabskim. Natomiast weryfikacji dokonano na podstawie histogramów obliczonych z wypowiedzi nagranych w innych językach – głównie w języku

[7] Są to wyniki odpowiadające przypadkowi, w którym przestrzeń cech jest równo podzielona między rozpatrywanymi klasami.
[8] Są to osoby posiadają nagrania w dwóch lub trzech językach.

polskim (patrz tabela 6.8). Ocenę jakości weryfikacji dokonano na podstawie minimalnej – względem rozmiaru modelu oraz względem parametru k[9] – sumy błędów akceptacji i odrzucania. Rysunek 6.8 przedstawia otrzymane wyniki odpowiadające następującym warunkom przeprowadzenia weryfikacji: *Gq=1 do 30, d=0.5 sek*.

Można zaobserwować znaczne zróżnicowanie wyników dla poszczególnych mówców. Tabela 6.10 przedstawia wartości błędów dla 15 sekundowej długości wypowiedzi. O ile w przypadku mówców o numerach #2, #6, #10, #11 i #13 można stwierdzić, że możliwe jest przeprowadzenia weryfikacji niezależnie od języka, o tyle nie można wyciągnąć takiego wniosku w przypadku mówców o numerach #4, #5, #9 oraz #12. Należy podkreślić, że gdyby odwrócić role wypowiedzi poszczególnych języków w testowaniu metody – tzn. budować modele głosów np. na podstawie wypowiedzi nagranych w języku polskim i przeprowadzić rozpoznawania na podstawie wypowiedzi zarejestrowanych w języku arabskim – to istnieje możliwość, że uzyskane w ten sposób wyniki będą odmienne od tych wyżej przedstawionych.



Rys. 6.8: Minimalna suma błędów akceptacji i odrzucania dla poszczególnych

[9] Wybrano najlepsze wyniki odpowiadające jednej optymalnej dla konkretnego mówcy wartości parametru k oraz optymalnemu rozmiarowi klasy tego mówcy dla optymalnej wartości k.

mówców w zależności od długości wypowiedzi.

Numer mówcy	#2	#4	#5	#6	#9	#10	#11	#12	#13
Błąd %	0	14,01	23,99	2,41	17,35	7,74	0,10	15,76	3,65

Tabela 6.10: Minimalna suma błędów akceptacji i odrzucania dla 15 sekundowej wypowiedzi.

Wykonano ponadto badania odpowiadające następującym warunkom przeprowadzenia weryfikacji: *Gq= 2 do 7, d=2 sek*. Wyniki uzyskane dla tych warunków były gorsze niż te otrzymane powyżej. Fakt ten wynika ze zmiany długości pojedynczego obiektu (wypowiedzi) z 0.5 sekundy na 2 sekundy. Oznacza to, że im krótsze są badane jednostki mowy tym większe jest ich podobieństwo dla różnych języków.

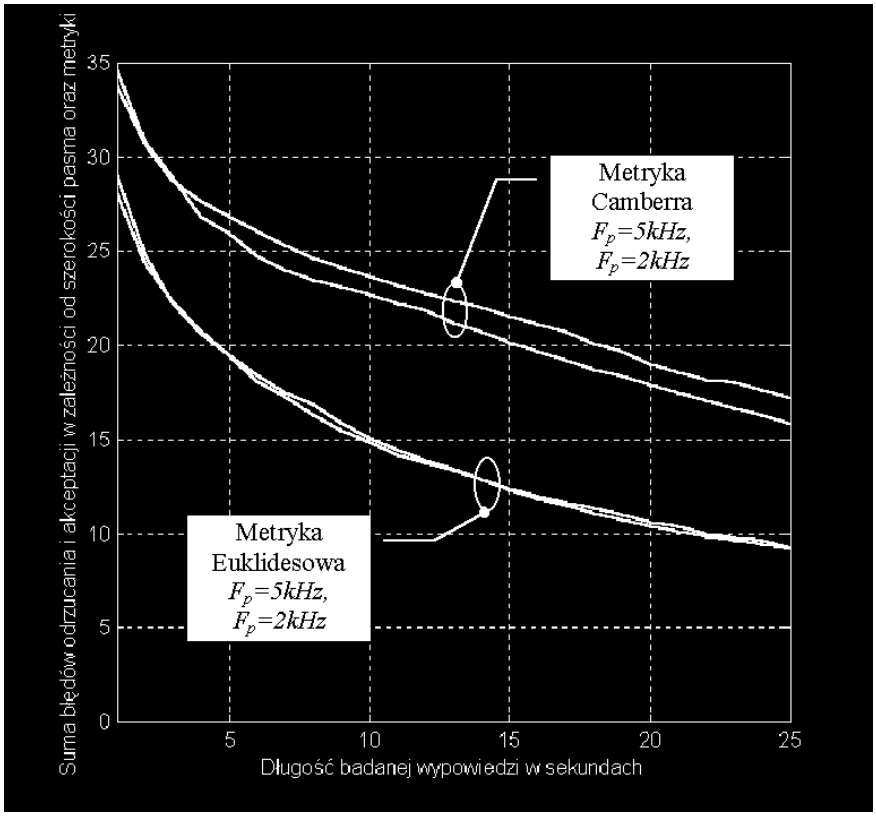
6.4. Wpływ pasma przenoszenia sygnału na jakość rozpoznawania

Celem określenia wpływu pasma przenoszenia sygnału na jakość rozpoznawania wykonano następujący eksperyment. Obniżono częstotliwość próbkowania użytych sygnałów kolejno do F_p=5kHz oraz F_p=2kHz. Towarzyszyło temu obniżenie górnej częstotliwości granicznej sygnałów odpowiednio do 2kHz oraz 1kHz. Dokonano weryfikacji przy następujących warunkach:

1- jakość sygnału wejściowego: *pok,*
2- przetwarzanie A/C: F_p=5kHz oraz F_p=2kHz, B=16 bitów
3- bramka szumu: $P_b=\mu$, $t_b=10ms$,
4- przetwarzanie wstępne: N_E, *d=1sek,*
5- obliczenie histogramów: *hlog,* $L_b=31$,
6- selekcja histogramów: $o_p=0\%$,
7- metryka: *Euklidesowa, Camberra*
8- metoda klasyfikacji: *pkNN, zakres zmienności parametru k od 1 do 20.*
9- klasyfikacja grupowa: *Gq=2 do 25,*
10- procedura decyzyjna: *W,*
11- rozmiar populacji: *M=21,*
12- zbiór mówców: *O,*
13- zależność od tekstu: *indep*.

Rysunek 6.9 przedstawia minimalne – względem rozmiaru modelu oraz względem wartości parametru k – sumy błędów akceptacji i odrzucania w funkcji długości badanych wypowiedzi dla czterech przypadków weryfikacji w zależności od częstotliwości próbkowania sygnałów oraz użytej metryki. W przeciwieństwie do badań

przeprowadzonych dla częstotliwości próbkowania równej 22kHz (podrozdział 6.2) lepsze wyniki uzyskano dla metryki Euklidesowej niż dla metryki Camberra. Przewaga metryki Euklidesowej w tym wypadku wydaje się być powiązana z różnicami rzędów wartości poszczególnych składowych wektora cech. Jeżeli jedna lub więcej składowych wektora cech przyjmują wartości względnie większe niż pozostałe, to w skutek operacji podniesienia do kwadratu (patrz wzór 2.6) metryka jest zależna wyłącznie od tych składowych. Najczęściej efekt taki jest całkowicie niezgodny z zamierzeniami, gdyż podważa celowość pomiaru pozostałych składowych wektora cech niwelując wnoszoną przez nie informację. Jednak, gdy w rzeczywistości więcej informacji dyskryminującej niosą składowe większego rzędu wektora cech to metryka Euklidesowa okazuje się lepsza od metryki Camberra ponieważ nie 'obniża' ona (poprzez normalizację[10]) moc dyskryminacyjną tych składowych. Zatem wydaje się być pewnym fakt, że w wyniku obniżenia częstotliwości próbkowania tracą znaczenie niektóre składowe histogramu zaś inne (względnie większe składowe) stają się ważniejsze.

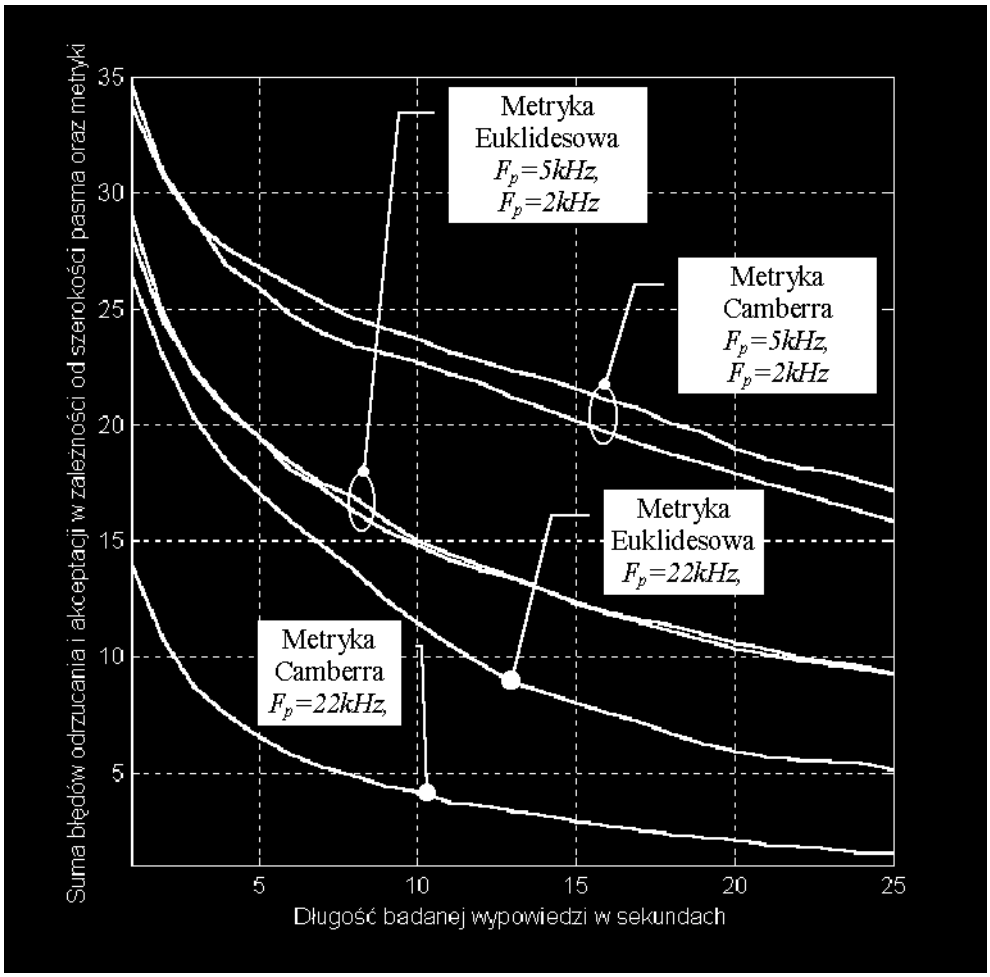


Rys. 6.9: Minimalne sumy błędów akceptacji i odrzucania w zależności od częstotliwości próbkowania sygnałów oraz użytej metryki.

[10] Należy przypomnieć, że metryka Camberra posiada własności samonormalizujące.

Wyniki pokazują także, że obniżenie częstotliwości próbkowania sygnału wpłynęło negatywnie na jakość rozpoznawania. Dla rozpatrywanych przypadków wpływ metryki był jednak zdecydowanie większy. W szczególności w przypadku metryki Camberra można zaobserwować przewagę wyników dla przypadku F_p=5kHz.

W celu pełniejszego przedstawienia zależności błędów od częstotliwości próbkowania sygnałów po ograniczeniu ich pasma przenoszenia na rysunku 6.10 porównywano wykresy z rysunku 6.9 z wynikami uzyskanymi dla częstotliwości próbkowania F_p=20kHz, przy czym pozostałe warunki przeprowadzenia eksperymentu zostały niezmienione. Na tym rysunku widać bezdyskusyjną przewagę wyników dla metryki Camberra, gdy pasmo przenoszenia sygnałów jest równe 10kHz a częstotliwość próbkowania F_p=22kHz.



Rys. 6.10: Porównanie minimalnych sum błędów akceptacji i odrzucania dla różnych częstotliwości próbkowania sygnałów i różnych metryk.

6.5. Badania dla danych wejściowych o jakości telefonicznej

6.5.1. Opis użytej bazy głosów i warunków przeprowadzenia eksperymentów

Wyniki prezentowane w tym podrozdziale uzyskano na podstawie bazy danych głosów zawierającej nagrania dla 50 mówców (25 kobiet, 25 mężczyzn). Jest to baza głosów o jakości telefonicznej, przy czym częstotliwość próbkowania zawartych w niej sygnałów wynosi 10kHz zaś dokładność ich próbkowania 16 bitów/próbkę. Długość nagrań dla pojedynczej osoby wynosiła ok. 1.2 minuty. Jakość rozpoznawania oceniono na podstawie sumy błędów akceptacji i odrzucania. Posługiwanie się tą formą oceny błędów jest szczególnie uzasadniony w przypadku tej bazy, ponieważ zawiera ona stosunkowo dużą liczbę mówców. Zatem całkowite rozdzielenie klas odpowiadających różnym mówcom w przestrzeni cech staje się praktycznie bardzo mało prawdopodobne. Jeśli przykładowo dla danej klasy spośród wszystkich obiektów przestrzeni cech znajdzie się choćby jeden obcy obiekt ze względnie dużą wartością funkcji przynależności[11] do tej klasy, to procent błędów odrzucania własnych obiektów przy zerowej akceptacji obcych obiektów staje się tylko z tego powodu bardzo duży i nie oddaje rzeczywistego, ogólnego obrazu rozdzielności tej klasy od innych. Prawdopodobieństwu wystąpienia powyższej sytuacji sprzyja duża liczba mówców i co zatem idzie duża liczba obiektów w przestrzeni cech. Analogicznie jeśli choć jeden własny obiekt klasy byłby nie reprezentatywny, to pomiar błędu akceptacji obcych obiektów przy zerowym odrzucaniu własnych obiektów także nie będzie miał sensu. Powyższe wnioski zostały zaobserwowane doświadczalnie i z tego względu zdecydowano się na przedstawienie wyników tylko na podstawie sumy błędów akceptacji i odrzucania zminimalizowanej względem rozmiaru danej klasy tzn. względem progu akceptacji (progu detekcji). Innym rozwiązaniem byłoby np. wybór względnie dużej wartości parametru o_p (patrz podrozdział 2.11) pozwalającej na odrzucanie część nie reprezentatywnych obiektów.

Poniżej przedstawiono warunki przeprowadzenia eksperymentów opisanych w tym podrozdziale:

1- jakość sygnału wejściowego: *tel,*
2- przetwarzanie A/C: F_p=10kHz, B=16 bitów,
3- bramka szumu: $P_b=\mu, 1.3\mu, 2\mu$; $t_b=10ms$,
4- przetwarzanie wstępne: *bez normalizacji energii*, $d=0,5sek$,

[11] Może to być w szczególności obiekt nie reprezentatywny lub błędnie zaklasyfikowany.

5- obliczenie histogramów: *hlog,* $L_b=31$,
6- selekcja histogramów: $o_p=0\%$,
7- metryka: *Euklidesowa, Camberra*
8- metoda klasyfikacji: *pkNN, zakres zmienności parametru k od 1 do 20.*
9- klasyfikacja grupowa: *Gq=2 do 30,*
10- procedura decyzyjna: *W,*
11- rozmiar populacji: *M=50,*
12- zbiór mówców: *O,*
13- zależność od tekstu: *indep*.

Na uwagę zasługuje fakt, iż w opisanych w tym punkcie badaniach nie dokonano normalizacji energii sygnału ponieważ były one zarejestrowane takim samym kanałem.

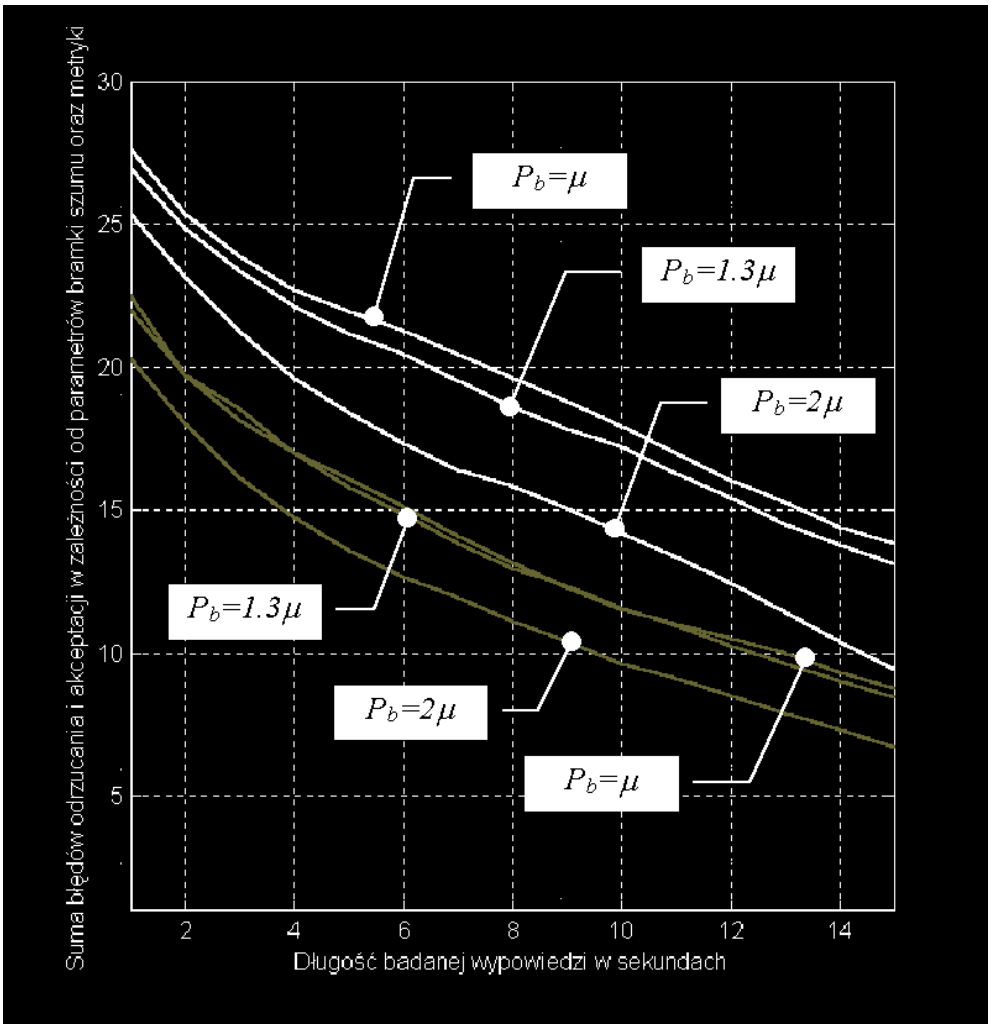
6.5.2. Wyniki weryfikacji

Rysunek 6.11 przedstawia minimalne – względem rozmiaru modelu oraz względem wartości parametru k – sumy błędów akceptacji i odrzucania uśrednione dla całej populacji (50 mówców) w funkcji długości badanej wypowiedzi. Poszczególne wykresy na rysunku przedstawiają uśrednione wartości błędów dla różnych parametrów bramki szumu oraz różnych metryk. Wykresy szare odpowiadają metryce Euklidesowej zaś białe metryce Camberra.

Niezależnie od parametrów bramki szumu metryka Euklidesowa pozwala na uzyskanie lepszych wyników. Można to było przywidzieć na podstawie wyników uzyskanych w podrozdziale 6.3. W miarę wzrostu progu bramki wyniki poprawiają się, ponieważ 'wycina się' z sygnałów więcej fragmentów o niższej energii. Ponieważ jakość tych sygnałów jest telefoniczna, to 'wycinane' fragmenty o niższej energii mogą zawierać większe ilości szumu, co tłumaczy poprawę wyników. Należy przypomnieć, że w przypadku sygnałów nagranych w warunkach pokojowych (względnie mniejszy poziom szumu) zależność wyników od progu bramki, gdy jej długość była równa 10ms, była wyraźnie słabsza. Co więcej stwierdzono, że dla tych nagrań wyniki są lepsze dla progu bramki wynoszącego μ niż 1.3μ (przy stałej długości bramki $t_b=10ms$). Na ogół potwierdza to celowość doboru wartości progu bramki w zależności od poziomu szumu towarzyszącego sygnałowi (patrz 4.3.2).

Celem poprawy jakości weryfikacji przeprowadzono ją dla każdego mówcy na warunkach dla niego preferencyjnych, przy czym dotyczyło to łącznie następujących warunków:

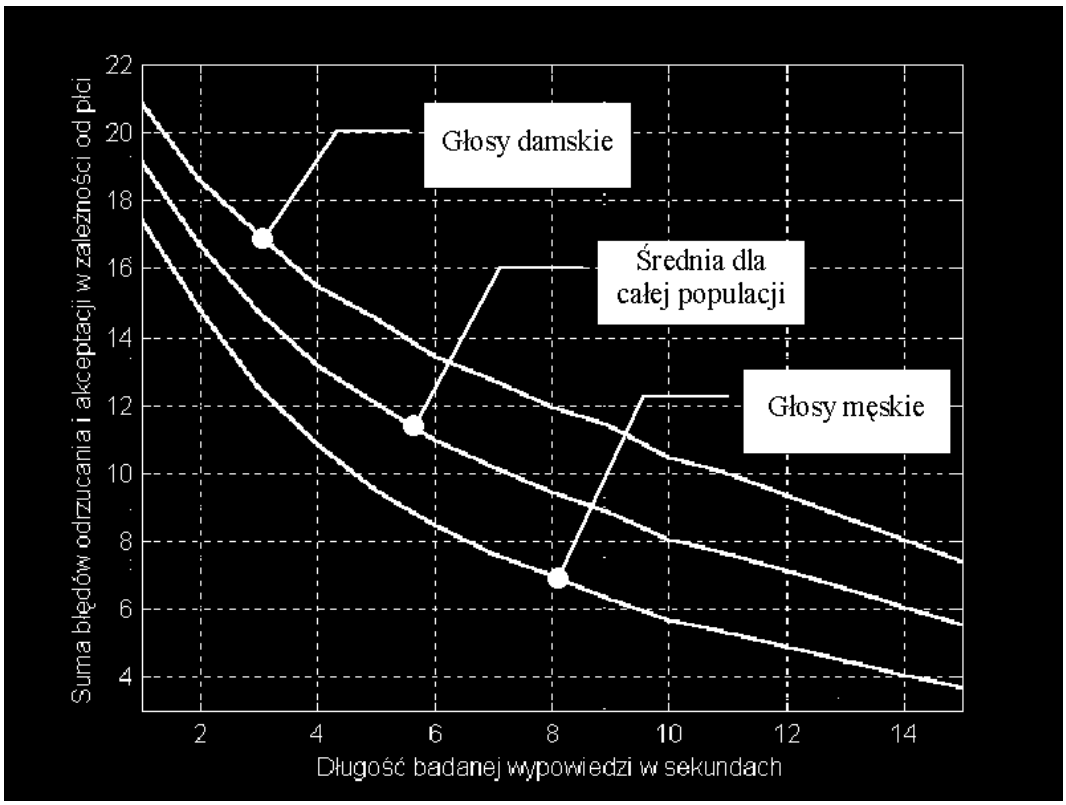
1- parametrów bramki szumu,
2- użytej metryki,
3- wartości parametru *k* metody klasyfikacji *pkNN*,
4- wartości progu akceptacji.



Rys. 6.11: Porównanie minimalnych sum błędów akceptacji i odrzucania dla różnych wartości parametrów bramki szumu oraz różnych metryk.

Rysunek 6.12 przedstawia procent błędów uśrednionych dla całej populacji oraz dla każdej płci z osobna. Zgodnie z wynikami prezentowanymi na tym rysunku można stwierdzić, że wyniki rozpoznawania dla płci męskiej są wyraźnie lepsze. Wyniki te są zgodne z wcześniejszymi badaniami prowadzonymi np. przez Furui (patrz rys. 3.5).

Ogólnie wyniki otrzymane na podstawie tej bazy głosów były gorsze niż te uzyskane na podstawie wcześniej zbadanych baz. Jest to głównie związane z większą liczbą mówców rozpatrywanej populacji oraz z gorszą jakością (telefoniczną) sygnałów wejściowych wykorzystanych w eksperymentach. Negatywny wpływ miały także względnie krótsze dane trenujące (łączna długość nagrań dla jednej osoby)



Rys. 6.12: Średnie sumy błędów akceptacji i odrzucania w zależności od płci mówców.

6.6. Wpływ długości ciągu uczącego na jakość rozpoznawania

Należy przypomnieć, że jedno z głównych założeń prezentowanej rozprawy jest poprawa jakości rozpoznawania w miarę zwiększania długości danych trenujących oraz długości wypowiedzi na podstawie której podejmuje się decyzję o rozpoznawaniu (długości danych testujących). W celu dokładniejszego zbadania granicznych wartości błędów rozpoznawania w funkcji tych długości posłużono się 36 osobowej bazy głosów, w której średnia długość nagrań dla każdego mówcy przekracza 8 minut, przy czym minimalna długość nagrań dla jednej osoby wynosiła ok. 6 minut, a maksymalna ok. 21 minut. Materiał fonetyczny składał się z kilkuset wyrazów lub krótkich zdań nagranych

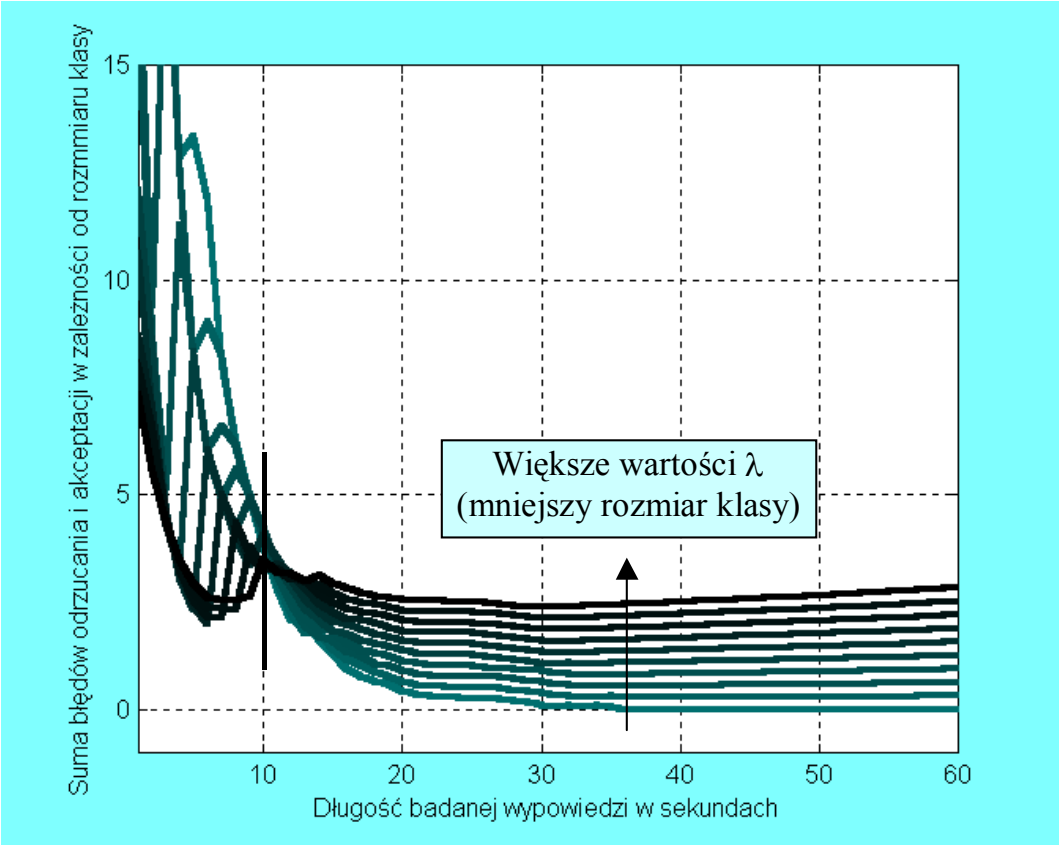
w izolacji[12]. Wśród 36 mówców było 23 mężczyzn, 9 kobiet, 2 chłopcy oraz 2 dziewczyny. Nagrania zarejestrowano w warunkach pokojowych, częstotliwość próbkowania sygnałów wynosiła 16kHz a rozdzielczość amplitudowa 16 bitów/próbkę. Poniżej przedstawiono warunki przeprowadzenia eksperymentu opisanego w tym podrozdziale:

1- jakość sygnału wejściowego: *pok,*
2- przetwarzanie A/C: F_p=16kHz, B=16 bitów,
3- bramka szumu: $P_b=2\mu$; $t_b=10ms$,
4- przetwarzanie wstępne: N_E, $d=1sek$,
5- obliczenie histogramów: *hlog,* $L_b=31$,
6- selekcja histogramów: $o_p=0\%$,
7- metryka: *Euklidesowa, Camberra*
8- metoda klasyfikacji: *pkNN, zakres zmienności parametru k od 1 do 60.*
9- klasyfikacja grupowa: $Gq=2$ *do 60,*
10- procedura decyzyjna: *W,*
11- rozmiar populacji: $M=36$,
12- zbiór mówców: *O,*
13- zależność od tekstu: *indep*.

Zbadano jakość weryfikacji dla całej populacji przy jednakowych dla wszystkich wartości parametrów bramki szumu. W wyniku przepuszczeniu sygnałów przez bramkę szumu 'wycięto' z nich fragmentów o łącznie długości równej średnio ok. 48%. Pozostałe fragmenty (ok. 52%) podzielono na takie o długości jednej sekundy ($d=1$) i z każdego fragmentu obliczono jeden histogram amplitudowy, który stanowił pojedynczy obiekt w przestrzeni cech. Minimalna liczba obiektów dla jednego mówcy wynosiła 181, maksymalna 685 a średnia ok. 252 obiektów. Obliczono minimalne – względem wartości parametru k – sumy błędów akceptacji i odrzucania w zależności od dwóch parametrów: długości wypowiedzi oraz rozmiaru modelu weryfikowanego mówcy. Podobnie jak w poprzednich badaniach manipulacji rozmiaru weryfikowanego modelu dokonano poprzez wykluczenie z niego kolejnych obiektów o najniższych wartościach funkcji przynależności. Liczba wykluczonych z klasy obiektów oznaczono symbolem λ. Wartość błędu odrzucania jest wtedy równa liczbie wykluczonych obiektów λ podzielona przez łączną liczbę obiektów należących do weryfikowanej klasy. Suma

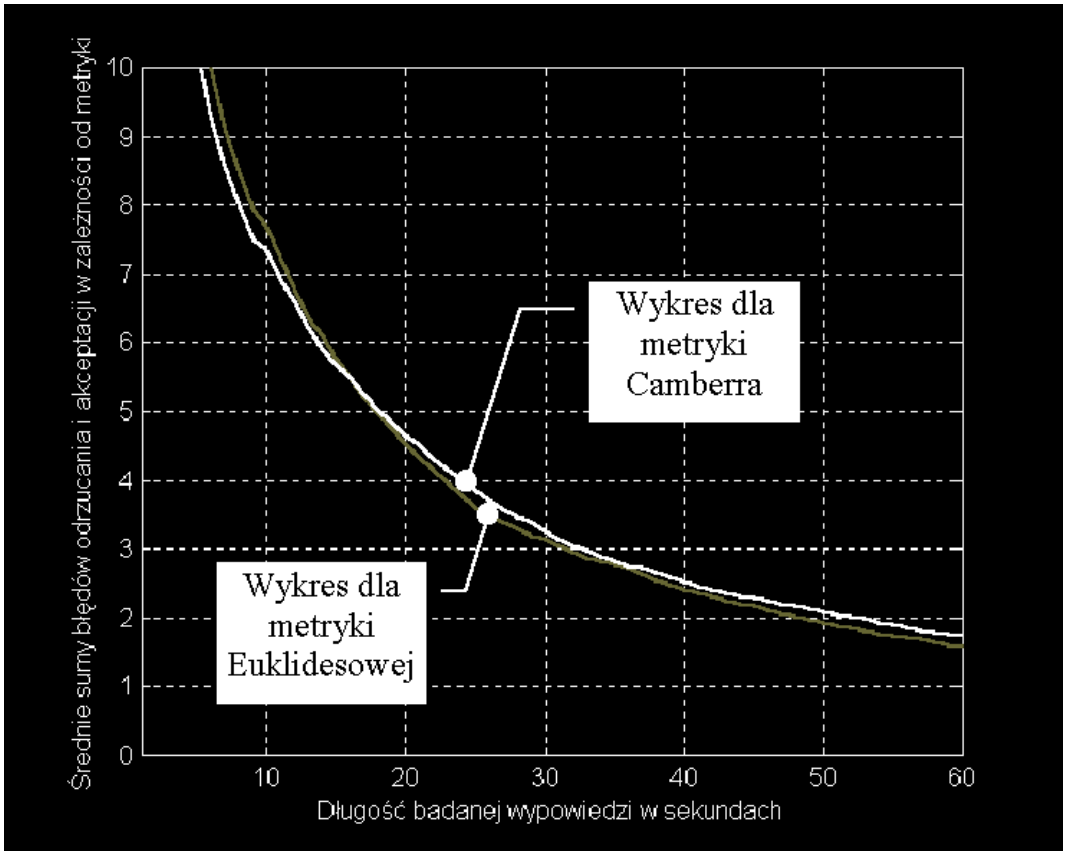
[12] W istocie baza ta została zaprojektowana z myślą o jej zastosowanie w badaniu systemów automatycznego rozpoznawania mowy (treści).

błędów akceptacji i odrzucania obliczono dla wartości λ z zakresu 0 do 10, co w najgorszym przypadku oznacza, że sam procentowy błąd odrzucania własnych obiektów nie przekracza 5.5% (tj. 100*10/181). Rysunek 6.13 przedstawia przykładową rodzinę wykresów błędów dla różnych wartości parametru λ w funkcji długości badanej wypowiedzi, uzyskanych dla jednego modelu (jednego mówcy – głos męski) przy weryfikacji z użyciem metryki Euklidesowej. Wykresy z jaśniejszym kolorem odpowiadają małym wartościom parametru λ (większemu rozmiarowi klasy). Natomiast wykresy ciemniejsze odpowiadają dużym wartościom parametru λ (mniejszemu rozmiarowi klasy). Na podstawie uzyskanych wyników można stwierdzić, że dla małych długości badanych wypowiedzi zalecane jest ograniczenie rozmiaru klas w przestrzeni. Począwszy od pewnego progu długości badanej wypowiedzi (ok. 10 sekund na rysunku 6.13) lepsze wyniki można uzyskać dla większych rozmiarów klas. Należy zauważyć, że począwszy od tej samej wartości długości suma błędów akceptacji i odrzucania staje się praktycznie stała.



Rys. 6.13: Zależność sumy błędów akceptacji i odrzucania od rozmiaru klasy.

Rysunek 6.14 przedstawia średnie (dla całej populacji) wartości minimalnych sum błędów akceptacji i odrzucania w zależności od zastosowanej metryki. Należy zwrócić uwagę na bardzo zbliżone wyniki uzyskane dla różnych metryk, co w świetle uzyskanych wcześnie wyników (roz 6.2, 6.5) uzupełnia obraz związku między jakością rozpoznawania z użyciem różnych metryk (Euklidesowa oraz Camberra) i pasmem przenoszenia rozpatrywanych sygnałów.



Rys. 6.14: Zależność sumy błędów akceptacji i odrzucania od użytej metryki.

6.7. Podsumowanie

Zaprezentowane wyniki ogólnie potwierdzają możliwość rozpoznawania głosów na podstawie histogramów amplitudowych sygnału mowy. Taki opis sygnału dla celu rozpoznawania głosów charakteryzuje się dużym stopniem niezależności od tekstu a także od języka wypowiedzi. Opisane w tym rozdziale doświadczenia zostały przeprowadzone z użyciem kilku baz głosów. Badania prowadzono dla różnych wariantów przetwarzania wstępnego sygnału głównie w zależności od częstotliwości jego

próbkowania oraz parametrów bramki szumu użytej w celu wyboru zdarzeń głosowych na podstawie których dokonywana jest decyzja o rozpoznawaniu. Ponieważ rozdzielczość histogramów użytych w badaniach była mała (dokładnie 31 poziomów amplitud co odpowiada dokładności rzędu 5 bitów/próbkę) w stosunku do dokładności reprezentacji wejściowych sygnałów (16 bitów/próbkę) nie poddano analizie sygnałów próbkowanych z mniejszą dokładnością (np. 12 lub 8 bitów/próbkę). We wszystkich przeprowadzonych badań użyto histogramy z rozdzielczością logarytmiczną. Operowano rożnymi podstawowymi długościami wypowiedzi (wartości parametru *d*) odpowiadającymi pojedynczemu obiektowi w przestrzeni cech. Do badania dłuższych wypowiedzi wykorzystano głównie ideę rozpoznawania grupowego dzieląc je na fragmenty o podstawowej długości *d*. W znacznej części eksperymentów do klasyfikacji aplikowano metodę potencjału najbliższych sąsiadów zarówno z metryką Euklidesową jak i z metryką Camberra. Ponadto zbadano jakość wyników w zależności od parametru *k* tej metody. Większość wyników została zaprezentowana w funkcji długości wypowiedzi na podstawie której dokonywana jest decyzja o rozpoznawaniu. Analizie poddano zarówno możliwości weryfikacji jak i identyfikacji mówców. W tym drugim przypadku badania wykonano głównie dla otwartego zbioru rozpoznawanych klas przy wykorzystaniu, opisanej w rozdziale trzecim, dwuetapowej procedury identyfikacji w otwartych zbiorach. Oceny jakości metody dokonano przede wszystkim na podstawie różnego rodzaju błędów rozpoznawania. Ocenie poddano również możliwość rozpoznawania głosu niezależnie od języka wypowiedzi.

Pomimo, że niektóre użyte bazy głosów zawierały nagrania zarejestrowane w różnych sesjach, odległych w czasie, to jednak nie podjęto szczegółowej analizy jakości rozpoznawania w zależności od długości czasowej między sesjami nagrań. Ponadto nie zbadano możliwości funkcjonowania metody w przypadku, gdy dane trenujące oraz testujące pochodzą z różnych kanałów rejestracyjnych. Do zbadania pozostaje również możliwość aplikacji innych metod klasyfikacji np. opartych na sieciach neuronowych lub na algorytmie kwantyzacji wektorowej bądź binarnej. Badaniom można poddać również możliwość wykorzystania krótkookresowych histogramów amplitudowych do rozpoznawania mówcy metodami zależnymi od tekstu lub rozpoznawania treści mowy. W szczególności mogą to być algorytmy dopasowania próbek do wzorców opartych na programowaniu dynamicznym DTW lub na niejawnych modelach Markowa HMM. Wstępne, zakończone sukcesem, próby aplikacji opisu histogramowego oraz algorytmu DTW w zagadnieniu rozpoznawania mowy oraz mówcy na podstawie ograniczonego

słownika izolowanych wyrazów (10 cyfr) zostały już podjęte przez autora. W przypadku rozpoznawania mówcy wyniki świadczą o możliwości uzyskania lepszych wyników[13] niż tych otrzymanych dla rozpoznawania niezależnego od tekstu. Jednak ze względu na fragmentaryczny charakter tych wyników nie zostały one opisane w niniejszej rozprawie[14].

Zdaniem autora warto podjąć także próbę rozpoznawania mówcy na podstawie zaproponowanego opisu histogramowego oraz parametrów przejść przez zero sygnału mowy [Bas78]. W szczególności interesujące byłoby rozszerzenie zaproponowanego wektora cech o parametrach opisujące rozkłady interwałów między kolejnymi przejściami przez zero sygnału. Istota tego rozszerzenia polega na tym, że histogramowy opis amplitud nie zachowuje informacji o ich 'uporządkowaniu' na osi czasu zaś rozkład interwałów czasowych między kolejnymi przejściami przez zero sygnału dostarcza fragmentaryczny obraz rozmieszczenia amplitud sygnału w czasie. Dodatkowym walorem takiego rozwiązania, gdyby się powiodło, byłaby w dalszym ciągu łatwość ekstrakcji parametrów.

[13] Chodzi tu głównie o możliwość uzyskania takiej samej jakości wyników rozpoznawania lecz z użyciem krótszych danych zarówno trenujących jak i testujących.
[14] Nie było to również przedmiotem badań podjętych w rozprawie.

ENGLISH SUMMARY

ABOUT THE THESIS

Automatic speaker recognition falls under the group of biometric methods of persons' identification. Other techniques from the same group perform the identification task on the bases of fingerprints, hand geometry biometrics, retina scan, face recognition, and DNA analysis. The presented thesis is devoted to a new automatic speaker recognition method based on speech signal representation in time domain by amplitude histograms. The proposed method is called "SPEAKER RECOGNITION BY AMPLITUDE DISCRIMINATION".

HYPOTHESIS AND RESEARCH GOALS

The main hypothesis of the proposed approach is as follows:

"Speech signal representation by long-term amplitude histograms preserves voice-relevant individual characteristics in a degree that allows speaker recognition to be successively performed on it's bases".

In particular, the main emphasis of the thesis is on testing the following question concerning the proposed method:

1. Increasing the training utterances' duration used for building the speaker's model, while retaining their random stimulus content, will result in higher performance.
2. Performing the speaker recognition task on the bases of a sufficiently-long utterance, consisting of random contents (random text), allows obtaining good recognition accuracy irrespective to the following factors:
   - utterances' stimulus content as well as pronunciation quality,
   - utterances' language and semantic aspect,
   - normal environment noise level that accompanies signals' registration,
   - small amplitude resolution[1] of the recording apparatus.

[1] Number of bits per sample.

3. Blind segmentation of utterances (independent of contents and words' boundaries) shouldn't scarify performance.
4. Speaker recognition is also possible when the speech signal bandwidth is limited even to 1 kHz.
5. Speaker recognition can be language-independent. Therefore, it's possible to take the recognition decision on the bases of utterances available in one language (testing data), even though the speakers' models are constructed with the use of utterances of other different language (training data).
6. Recognition is also achievable in open-set recognition mode.

SPEAKER RECOGNITION BY AMPLITUDE DISCRIMINATION

In general, one can divide the speaker recognition task mainly into three stages. The first stage includes acquisition of speech signal, preprocessing and feature extraction. The second stage involves computing the likelihood of similarity (distance) between the extracted feature vectors and those obtained during system training and stored as individual speakers' reference templates or models. In the third stage, a decision, depending on the applied recognition procedure, is made on the bases of one or more similarity result (one for verification procedure or more for identification procedure).

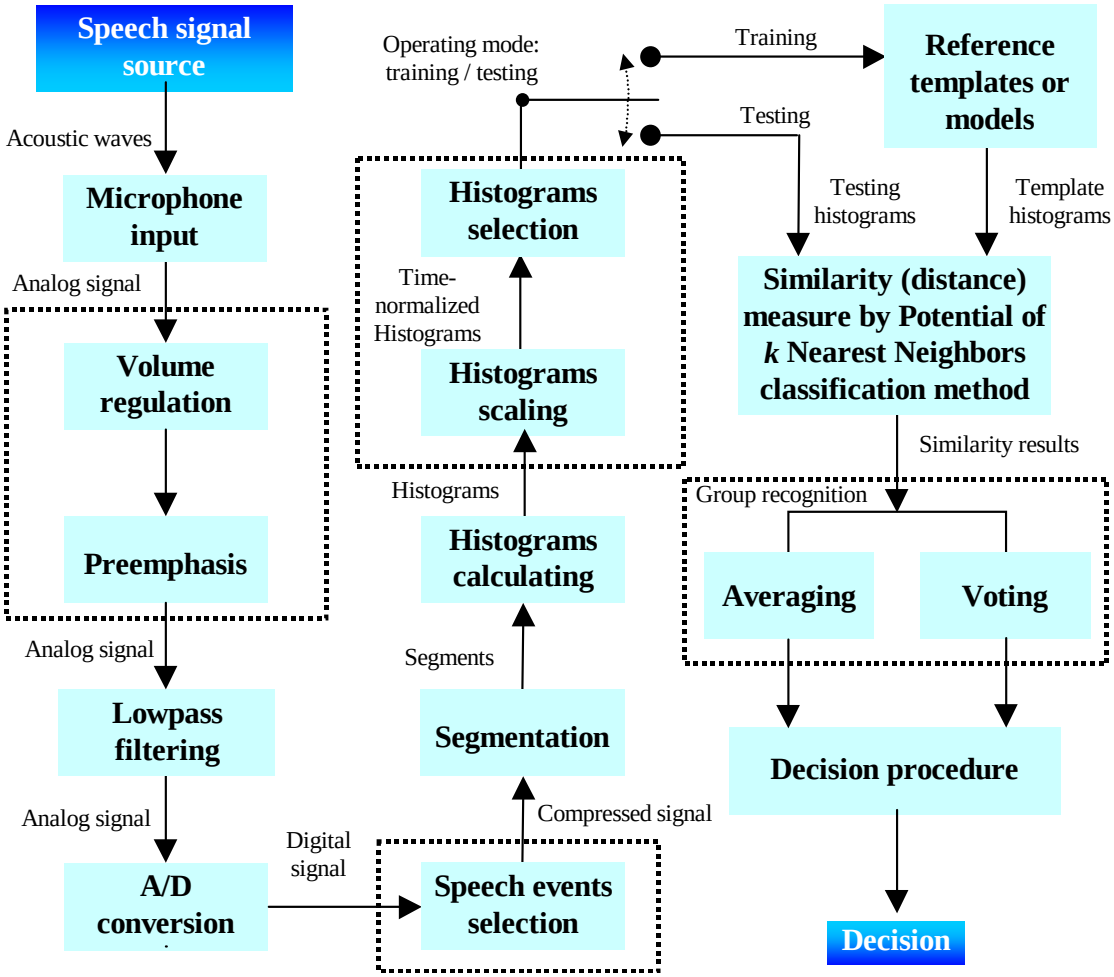
The philosophy of the new method "Speaker Recognition by Amplitude Discrimination" relies on a novel representation of the speech signal by it's *amplitude histograms*. Such histograms are obtained from time-domain speech signal by simple segregation of instant amplitudes into a number of levels given as the assumed *histogram-levels' vector*. The number of boundary levels (the dimension of the levels' vector) is called the *histogram's resolution*. In an effort to perform the recognition task on the bases of some selected speech fragments that possess relatively higher gain level, input signal may be passed through a noise gate that selects these fragments. Instead of silencing the unselected speech fragments, they are excluded from the signal while the remaining fragments (the selected ones) are linked together resulting in a compressed signal form.

In addition to speech representation by amplitude histograms, a new template matching method built as a combination of two classification methods (namely *the k nearest neighbors kNN* and *the potential functions* methods) and called "THE

POTENTIAL OF THE *k* NEAREST NEIGHBORS *pkNN*", was proposed as the pattern identification method.

A full diagram of the discussed method is presented below. Optional bloks are denoted by a doted frames. Prior to feature extraction (histograms calculating) the analog input signal can preprocessed as follows:

1. Volume regulation or gain normalization to a predefined level.
2. In order to raise the signals to noise ratio SNR, the input signal may be preemphasised by emphasizing the higher-frequency components. This operation may also be done after analog to digital conversion.
3. Lowpass filtering before analog to digital conversion (antialiasing filtering).
4. Analog to digital conversion.
5. Speech events selection using a noise gate.



In the second recognition stage, amplitude histograms are calculated from speech portions obtained after 'blind' segmentation. Further these histograms can be length-normalized. Length normalization can be achieved by dividing the histograms values by the number of samples in the segment that corresponds to that histogram. This operation is used instead of time alignment of different repetitions of the same word or phrase before comparing them for similarity. It can be also used in order to allow comparing speech signals obtained using different sampling rates. In some cases selection of representative histogram samples is required so as to improve the recognition accuracy. For this purpose a selection algorithm was also proposed.

In the training phase of the system the extracted histograms are used to build reference templates or models. During testing phase the calculated histograms are matched with the template ones using a pattern matching technique called "The Potential of the *k* Nearest Neighbors *pkNN*". A group recognition procedure is also

suggested to take recognition decisions on bases of more than one pattern (one histogram). In particular, similarity results for individual histograms may by averaged or used in a voting decision scheme. The decision procedure takes into consideration several aspects related to the system's decision mode (identification or verification) and decision alternatives.

THESIS – CONTENTS & ORGANIZATION

The thesis is divided into seven chapters. The first chapter "INTRODUCTION" introduces the main hypothesis and assumptions of the proposed novel approach to the speaker recognition problem (sec. 1.1). It includes a brief discussion of the dimensions of the speaker recognition problem (sec. 1.2). The thesis organization is also presented in this chapter (sec. 1.3).

Chapter 2, entitled "AUTOMATIC SPEAKER RECOGNITION – SYSTEM STRUCTURE" covers in details the general framework of the automatic speaker recognition systems. In this chapter, speech analysis and representation methods are outlined (sec 2.2) with special emphasis on speech signal key parameters, significant from the speaker's voice-relevant characteristics point of view (sec 2.3). Further, the question of choosing optimum (discriminative) features and parameters' ranking is addressed in section 2.4. Sections 2.5 and 2.6 are devoted to speech acquisition, preprocessing and segmentation, while section 2.7 discusses the feature extraction issues. Beyond, vector quantization technique in general and binary quantization technique in particular are explained in section 2.8. Moreover, the use of various metrics for computing the distance between feature vectors (similarity measure) is considered in sections 2.9. Besides, three classification methods based on minimal distance conception are covered in section 2.10. These are *the k nearest neighbors kNN*, *the potential of the k nearest neighbors pkNN* and *the nearest mean NM* methods. In addition, an algorithm for feature vectors selection, designed to be used with the *potential of the k nearest neighbors* method, is proposed in section 2.11. Furthermore, section 2.12 is devoted to a very important discussion concerning recognition errors and their dependence on the manner in which the features' space is divided and on the type of the considered speaker set, being either open or closed. A comparison of the classification methods, presented in section 2.10, is included in section 2.13. Next, the idea of patterns' group recognition is introduced and explained in section 2.14. Finally, a measure of speakers' classes separation degree in the features' space is suggested in section 2.15.

Chapter 3, entitled "CLASSIFICATION AND COMPARISON OF AUTOMATIC SPEAKER RECOGNITION METHODS", addresses several topics related to the differences between various approaches to speaker recognition and gives a general overview of the current achieved results. In particular classification criterions of automatic speaker recognition systems are reviewed in section 3.2. This includes system classification depending on the decision procedure (identification or verification – sec. 3.2.1), the population size and type of speakers' set (open or closed speakers' set – sec. 3.2.2) or text and language dependence (sec. 3.2.3). In addition text-independent speaker recognition methods, based on long-term statistical analysis, are considered separately in section 3.3. Further, an overview of the current existing methods with a summary of the achieved results are provided in section 3.4. Moreover, the dilemma of comparing the performance of different methods is outlined in section 3.5. Subjective speaker recognition techniques are also described in section 3.6. Finally, a summary of the chapter is given in section 3.7.

Chapter 4, "REPRESENTATION OF SPEECH SIGNAL BY LONG–TERM AMPLITUDE HISTOGRAMS", is devoted to the formal theory of amplitude histograms as a new form of speech signal representation. Section 4.1 includes the mathematical basic of calculating different types of histograms as well as some additional characterization of their dependence on the signal's gain, phase and bandwidth. Further, a survey of polish language speech sounds representation via speech histograms, with some examples, is provided in section 4.2. In addition, other variants of histograms' extraction, namely histograms of signal's derivatives and selective histograms, are intorduced in section 4.3. The formers are obtained, in general, after signal preemphasis (sec. 4.3.1), while the later involve passing the speech signal through a noise gate in order to select some speech fragments that posses relatively higher gain level. Further, a discussion of histograms' robustness to white noise degradation of the signal, is reported in section 4.4. Finally, a simple implementation scheme for hardware amplitude histograms' obtaining is proposed and analyzed in section 4.5.

Chapter 5, entitled "SPEAKER RECOGNITION ON THE BASES OF LONG–TERM AMPLITUDE HISTOGRAMS", presents a full description of the proposed method. To avoid information redundancy, some of the method's components which are described in chapter 2 or 4, were excluded from discussion included this chapter. An overview of the recognition conditions is given in section 5.2. This includes signal preprocessing variants (sec 5.2.1), histograms' type (sec. 5.2.2), distance measure (metric – sec. 5.2.3)

and classification variants (sec. 5.2.4). Further, a summary of all the recognition conditions is included in section 5.3. In section 5.4, the ability of the histograms to discriminate voices of different speakers is initially tested and estimated by means of a proposed discrimination measure. In particular, the dependence of the histograms' discriminative power on the considered speech signal duration (sec. 5.4.1), the used metric (sec. 5.4.2) and the type of the applied histograms (sec 5.4.3, 5.4.4), is primarily investigated. Finally, a description of the "SPEAKER RECOGNITION BY AMPLITUDE DISCRIMINATION" method, on the background of other speaker recognition approaches, is considered in section 5.5.

Chapter 6, entitled "RELATED EXPERIMENTS AND RESULTS" provides very comprehensive details of the conducted experiments and achieved results. Section 6.1 is completely devoted to experiments carried out with the use of classical classification methods: *the nearest mean NM* (sec. 6.1.2) and *the k nearest neighbors kNN* (sec. 6.1.3). Moreover, this sections includes the description of the used speech material (sec. 6.1.1), an analysis of the recognition accuracy with respect to the duration of the testing utterance (sec. 6.1.4) as well as the results of a successive attempt to perform language-independent speaker recognition. All results presented in section 6.1 involve a small closed speakers' set. Further, in section 6.2 more extensive experiments', based on continuous speech data recordings, collected for 21 speakers in different languages (mainly Arabic and Polish), are reported. In particular, section 6.2.1 describes the used speech data and the recording environments while section 6.2.2 specifies the applied signal preprocessing operations. Next, a summary of the experiments is given in section 6.2.3. The results obtained for the open-speakers' set experiments are contained in sections 6.2.4. The recognition accuracy is estimated by the following recognition errors:

- the percentages of false acceptance decisions while the percentage of false rejection decisions is equal to zero (sec 6.2.4.1),
- the percentages of false rejection decisions while the percentage of false acceptance decisions is equal to zero (sec 6.2.4.1),
- the percentage of the minimum sum of false acceptance and false rejection decisions obtained for optimum speakers' class size in features' space.

In general, the above mentioned errors are obtained for one or more optimum classification conditions. The results for the closed speakers' set are covered separately

in section 6.2.5. A relatively more comprehensive analysis of language-independent speaker recognition is reported in section 6.3. Moreover, the dependence of the method's performance on speech signal bandwidth is considered in section 6.4. Besides, results accomplished for telephony speech, obtained for 50 speakers (25 female, 25 male), are presented in section 6.5. The description of the used speech material is included in section 6.5.1, while the results are reported in section 6.5.2. Method's performance, estimated using another speech data base registered for 36 speakers, is further investigated in terms of training data duration. Final conclusions are given in the last section of the chapter 6.7.

The closing chapter, entitled "SUMMARY" provides an overview of the work conducted during the research (sec. 7.1), gives a survey of the proved hypothesis and assumptions, postulated at the beginning of the thesis (sec. 7.2), discusses the accomplished accuracy in terms of various recognition conditions and indicates the optimum encountered ones (sec 7.3). Further, section 7.4 indicates the advantages and disadvantages of the designed speaker recognition method. In the final section 7.5, suggestions for future related-research are mentioned. Several topics from chapter 7 are also referred to in following part of this summary.

The bibliography, referred to in the thesis, follows the last chapter. An appendix, briefly describing the software functions, written in MATLAB language and used for method realization and verification, is further included at the end of the thesis.

RESEARCH AND WORK PERFORMED

The following tasks, pertaining to the introduced thesis, were particularly realized:

1. An original representation of speech signal based on long-term amplitude histograms was proposed for use in speaker recognition tasks. Besides, a straight forward implementation scheme for hardware amplitude histograms' obtaining was proposed.
2. An algorithm for speech events selection, on the bases of which the recognition decision is performed, was build by employing the idea of noise gate filtering, used

in general for noise reduction by silencing the speech signal's fragments that doesn't correspond to speech.

3. An original pattern matching algorithm, being a hybrid connection of the idea of the *k nearest neighbors kNN* and *potential functions* methods, and employing group recognizing scheme by averaging or voting, was designed as part of the proposed speaker recognition method.
4. In view of the considered pattern matching approach, a feature selection procedure, was suggested for eliminating non-representative histogram samples.
5. Several speech data bases were created for use in research realization and method's performance verification attempts.
6. Dedicated procedures, for realizing each elements of the proposed method, were written in MATLAB language. The designed software enables speaker recognition performing with .WAV format data input files.
7. The designed method was tested during several comprehensive experiments. The method's accuracy was estimated by means of diverse recognition errors. Experiments were conducted using a variety of speaker recognition data bases and aimed to find the optimum recognition conditions of the proposed method and estimated the recognition quality.

A SURVEY OF EXPERIMENTS' RESULTS

All the hypothesizes, assumed at the beginning of the thesis, were verified as correct. The dependence of the recognition accuracy, on the following recognition conditions, was particularly examined:

1. Quality of speech material.

The best results' quality appeared for speech material registered in low noise environments (room environments). With regard to these results, accuracy accomplished for telephony speech was relatively lower because of the higher noise level, large number of speakers considered as well as shorter training data.

2. Signals' sampling parameters.

A general tendency of accuracy decrease was observed as the signals' bandwidth was more limited. The histograms resolution used in the experiments corresponds to

approximately 5 bits per sample resolution. Therefor there was no need to examine the influence of signals' sampling resolution on the obtained recognition quality.

3. Speech events selection.

Experiment revealed that better results are achieved for selective histograms. The optimum noise gate parameters were found to be as follows:

- Decay time 10ms, threshold μ^2 - for office or room environment speech.
- Decay time 10ms, threshold 2μ - for telephony speech.

4. Speech segmentation – duration '*d*' of utterance corresponding to a single pattern in features' space.

Research involving the following *d* values: 0.5 sec., 1 sec., 2 sec. and 10 sec, ascertained definitive results' independence of parameter *d*, under the condition that the joint duration of the testing utterance is constant.

5. Type of histograms.

It was proved that logarithmically spaced amplitude levels of histograms delivers clearer images of the signal and though results in better recognition accuracy. An adequate number of levels for these histograms was found to be 31.

6. Duration of testing utterances used in recognition phase.

Most of accomplished results were presented as a function of the test-utterances' duration. Except for a few exceptions, an absolute error reduction was registered for longer duration of test-utterances.

7. Used metric.

Two metric functions were considered in the experiments: Euclidean and Camberr distance measures. The former proved to be a significantly better when used with speech signals sampled at 10, 5 or 2kHz, while the later definitively prevailed for signals sampled at 44 or 22kHz. Results for both metrics, achieved for signals sampled at 16kHz sampling rate, were found to be approximately identical.

8. Pattern matching technique.

[2] μ: average absolute amplitude of the noise gate input signal.

For small population size and closed speakers' set classic pattern matching methods (*kNN & NM*) were found to be sufficiently effective. However for large population and open-speakers' set these methods didn't give satisfactory results. Meanwhile, the proposed "Potential of *k* nearest neighbors *pkNN*" pattern matching technique permitted obtaining relatively much better results.

9. Parameter *k* of the "Potential of the *k* nearest neighbor" method.

Experiments revealed that the optimum vales of the *k* parameter belonged to interval 1 to 5. In addition, the optimum *k* values were found to be basically dependent on the number of patterns (objects) in the features' space that corresponds to a single speaker.

10. Speakers' sex.

A comparison of accuracy with respect to the speakers' sex indicated that results for male speakers are better by a factor of roughly 5%.

11. Open/Closed – speakers' set.

In the identification experiments, a very consequential and almost unconditional accuracy improvement for closed-speakers' set, compared to that of the open-speakers' set, was detected.

12. Text and language dependence.

All experiments where undertaken independently of the utterances' text (contents). The speech repertoire of all the used data bases was random and very extensive. Efforts, aiming to examine the possibility of language-independent speaker recognition, were successfully performed for both closed and open-speakers' set.

ADVANTAGES & DISADVANTAGES OF THE NEW APPROACH

Advantages:

- Very simple feature vectors that are easy to extract
- The possibility of straight forward implementation of hardware obtaining of the features vectors.
- Robustness of feature vectors against white noise degradation,
- Processing only selected fragments of speech signal enables performing the recognition task in real time and reduces the memory size required,required for speakers' models storage.
- Text and language independence.
- Possibility of applying for large populations and open-speakers' set.
- Ability to use in normal conditions (room environment or telephony speech).

Disadvantages:

- Obtainment of very good recognizing accuracy implies using test-utterances relatively long in duration (5-15 seconds dependent on other recognition conditions).
- Achieving good results requires also relatively long training data (2-3 minutes).
- Partial dependence of accuracy on channel characteristics and registering apparatus, particularly on signal's gain and bandwidth as well as microfon dynamics.
- Optimum recognition conditions are dependent on the considered speaker, which is a problem of identification systems only.
- Accuracy is differentiated for different speakers and additionally depends on the speaker's sex.

SUGGESTED DIRECTIONS OF FUTURE-RELATED RESEARCH

Although the presented thesis includes a wide investigation of the performance of the new proposed methods, some questions related to this approach remain to be answered. In particular the author suggests examining the following problems:

1. Optimization of the size of the features' space by clustering techniques.
2. Possibility of channel-independent recognition (recognition of utterances obtained from one channel on the bases of speakers' models built using speech collected through another channel).
3. Estimation of accuracy in terms of periods between recording sessions.
4. Comparing the effectiveness of the proposed method with other approaches[3].
5. Extending the proposed feature vector by considering zero crossing parameters.
6. Application of other classification methods, for example neural networks, vector quantization or binary quantization.

An interesting solution is the possibility of performing text-dependent speaker recognition, based on amplitude histograms representation of speech signal and HMM or DTW pattern matching techniques. In addition, amplitude histograms may be used in the same scheme for speech recognition task.

[3] Taking into consideration the comparing limitations, presented in chapter 3.

Bibliografia.

[Abs97] *Special issue on Automated Biometric Systems* Proceedings of the IEEE Vol. 85, No. 9, September 1997.

[Ajz76] Ajzerman M.A., Brawerman E.M., Rozonoer L.I., *Rozpoznawanie obrazów – metoda funkcji potencjalnych*. WNT, Warszawa, 1976.

[Amb98] Ambardar A., Borghesani C., *Mastering DSP Concepts Using MATLAB*. Prentice Hall, New Jersey, 1998.

[Ass94] Assaleh K.T., Mammone R.J., *New LP-Derived Features for Speaker Identification*. IEEE Trans. on Speech and Audio Processing, Vol. 2, No. 4, October 1994.

[Ata72] Atal B.S., *Automatic Speaker Recognition Based on Pitch Contours*. J. Acoust. Soc. Amer., Vol. 52 No. 6 (part 2), 1972.

[Ata74] Atal B.S., *Effectiveness of Linear Prediction Characteristics of Speech Wave for Automatic Speaker Identification and Verification*. J. Acoust. Soc. Amer., Vol. 55, No. 6, June 1972.

[Ata76] Atal B.S., *Automatic Recognition of Speaker from Their Voices*. Proceedings of the IEEE Vol. 64, No. 4 April 1997.

[Bar96] Barnwell III T.P., Nayebi K., Richardson C.H., *Speech Coding – A Computer Laboratory Textbook*. John Wiley & Sons, New York, 1996.

[Bara94] Baranowski J., Kalinowski B., Nosal Z., *Układy elektroniczne część III: układy i systemy cyfrowe*. WNT, Warszawa, 1994.

[Bas78] Basztura Cz., Majewski W., *Zastosowanie długoterminowej analizy przejść przez zero sygnału mowy do automatycznego rozpoznawania mówców*. Archiwum Akustyki Vol. 13, Zeszyt 1, 1978.

[Bas81] Basztura Cz., Majewski W., *wpływ wybranych parametrów kanału telefonicznego na identyfikację głosu*. Archiwum Akustyki Vol. 16, Zeszyt 4, 1981.

[Bas87] Basztura Cz., Majewski W., Jurkiewicz J., *Automatyczne rozpoznawanie głosów w zbiorach otwartych*. Archiwum Akustyki Vol. 22, Zeszyt 4, 1987.

[Bas88] Basztura Cz., Majewski W., Jurkiewicz J., *Automatic Voice Recognition in Open Sets*. Archives of Acoustics, Vol. 13, 1988.

[Bas91] Basztura Cz., Zuk J., *Image Similarity Functions in Non-Parametric Algorithms of Voice Identification*. Archiwum Akustyki Vol. 16, Zeszyt 2, 1991.

[Bas91a] Basztura Cz., *Experiments of Automatic Speaker Recognition in Open Sets*. Speech Communication, Vol. 10, 1991.

[Bas92] Basztura Cz., *Rozmawiać z komputerem.* Wydawnictwo Format, Wrocław 1992.

[Bas93] Basztura Cz., *Jak rozmawiać z komputerem ludzkim głosem.* Wydawnictwo Format, Wrocław 1993.

[Bas95] Basztura Cz., Baściuk K., *Aplikacyjne i obiektywne uwarunkowania identyfikacji i weryfikacji głosów w badaniach fonoskopijnych*. I Krajowa Konferencja Głosowa Komunikacja Człowiek-Komputer, Wrocław, październik 1995.

[Bie81] Bieńkowski P., *Reguły Generacji Sygnałów Reprezentujących Statystyczne Rozkłady Parametrów Sygnału Mowy*. Rozprawa doktorska, Instytut Telekomunikacji i Akustyki, Politechnika Wrocławska, Wrocław, 1981.

[Bol70] Bolt R.H. et al., *Speaker Identification by Speech Spectrograms: A Scientists' View of it's Reliability for Legal Purposes* . J. Acoust. Soc. Amer. 47, 1970.

[Bol73] Bolt R.H. et al., *Speaker Identification by Speech Spectrograms: Some Further Observations*. J. Acoust. Soc. Amer. Vol. 54, No. 2, 1970.

[Cam97] Campbell JR., Joseph P., *Speaker Recognition: A Tutorial*. Proceedings of the IEEE Vol. 85, No. 9, September 1997.

[Che93] Chen M.S., Lin P.H., Wang H.C., *Speaker Identification Based on a Matrix Quantization Method*. IEEE Trans. on Signal Processing, Vol. 41, No. 1, January 1993.

[Che95] Che C.W., Lin Q., *Speaker recognition Using HMM with Experiments on the YOHO database*. Proceeding of EUROSPEECH, Madrid, Spain, September 1995.

[Cro81] Crochiere R.E., *Interpolation and Decimation of Digital Signals – A Tutorial Review*. Proceedings of the IEEE, Vol. 69, No. 3, March 1981.

[Del93] Deller J.R., Proakis J., Hansen J., *Discrete-Time Processing of Speech Signals.* Macmillan Publishing Company, New York, 1993.

[Dod85] Doddington G.R., *Speaker Recognition – Identifying People by Their Voices*. Proceedings of the IEEE Vol. 76, September 1997.

[Dud73] Duda O., Hart P., *Pattern Classification and Scene Analysis*. John Wiley & Sons, New York 1973.

[Fla70] Flanagan J.L., *Speech Analysis, Synthesis and Perception*. Springer Verlag, Berlin, Heidelberg, New Jork 1970.

[Fur79] Furui S., *New Techniques for Automatic Speaker Verification Using Telephone Speech*. J. Acoust. Soc. Amer., Suppl. 1, No. 66, 1979.

[Fur81] Furui S., *Cepstral Analysis Technique for Automatic Speaker Verification*. IEEE Trans. on ASSP, Vol. 29, 1981

[Fur89] Furui S., *Digital Speech Processing, Synthesis, and Recognition.* Marcel Dekker, New York, 1989.

[Fur91] Furui S., *Speaker-Dependent-Feature Extraction, Recognition and Processing Techniques*. Speech Communication, Vol. 10, 1991.

[Fur94] Furui S., Matsui T.E., *Phoneme Level Voice Individuality Used in Speaker Recognition*. Proceedings of ICSLP 94, vol. 3, 1994.

[Fur96] Furui S., *An overview of speaker recognition technology*. In Automatic Speech and Speaker Recognition Advanced Topics, edited by Chin-Hui Lee, Frank K. Soong, Kuldip K. Paliwal, Kluwer Academic Publishers, 1996.

[Gol69] Gold B., Rabiner L.R., *Parallel Processing Techniques for Estimating Pitch Periods of Speech in the Time Domain*. J. Acoust. Soc. Amer., Vol. 46, No. 2 (part 2), 1979.

[Hig86] Higgins A., Wohlford R.E., *A New method of Text Independent Speaker Recognition*. Proceedings of ICASSP86, Tokyo, Japan, 1986.

[Hig91] Higgins A., Bahler L., Porter J., *Voice Verification Using Randomized Phrase Prompting*. Digital Signal Processing, Vol. 1, 1991.

[Hig96] Higgins A., Bahler L., Porter J., *Voice Identification Using Non-Parametric Density Matching*. In Automatic Speech and Speaker Recognition Advanced Topics, edited by Chin-Hui Lee, Frank K. Soong, Kuldip K. Paliwal, Kluwer Academic Publishers, 1996.

[Ife96] Ifeachor E.C., Jervis B.W., *Digital Signal Processing: A Practical Approach*. Addison-Wesley Publishing Company, New York, 1996.

[Jaj93] Jajuga K., *Statystyczna analiza wielowymiarowa*. PWN, Warszawa, 1993

[Jay74] Jayant S.N., *Digital Coding of Speech Waveforms: PCM, DPCM, and DM Quantizers,* Proceedings of the IEEE, Vol. 62, No. 5, May 1974.

[Jua82] Juang J.H., Wong D.Y., Gray A.H., *Distortion Performance of Vector Quantization for LPC Voice Coding*. IEEE Trans. ASSP, Vol. 30, No. 2, April 1982.

[Jua92] Juang B.H., Katagiri S., *Discriminative Learning for Minimum Error Classification*. IEEE Trans. on Signal Processing, Vol. 40, 1992.

[Jua93] Juang B.H., Rabiner L.R., *Fundamentals of Speech Recognition.* Prentice Hall, New Jersey, 1993.

[Jun98] Jung B.H., Chen T., *The Past, Present and Future of Speech Processing*. IEEE Signal Processing Magazine, Vol. 15, No. 3, 1998.

[Kle95] Kleijn W.B., Paliwal K.K. (Editors), *Speech Coding and Synthesis*. Elsevier, Amsterdam, 1995.

[Kos73] Kosiel U., *Statistical Analysis of Speaker-dependent Differences in the Long-Term Average Spectrum of Polish Speech*. In Speech Analysis and Perception, vol. 3, W. Jassem (editor), Polish Academy of Sciences, Warsaw, Poland: PWN-Polish Scientific Publishers, 1973.

[Lee96] Lee C.H., Soong F.K., Paliwal K.K., (editors) *Automatic Speech and Speaker Recognition Advanced Topics*. Kluwer Academic Publishers, 1996.

[Mak75] Makhoul J., *Linear Prediction: A Tutorial Review*. Proceedings of the IEEE Vol. 63 No. 4 April 1975.

[Mar77] Markel J.D., Oshika B.T., Gray Jr.A.H., *Long-Term Feature Averaging for Speaker Recognition*. IEEE Trans. on ASSP, Vol. 25, No. 4, 1977.

[Mar79] Markel J.D., Davis S.B., *Long-Term Feature Averaging for Speaker Recognition*. IEEE Trans. on ASSP, Vol. 27, No. 1, 1979.

[Mat93] Matsui T., Furui S., *Concatenated Phoneme Models for Text Variable Speaker Recognition*. Proceedings of ICASSP93, April 1993.

[Mat94] Matsui T., Furui S., *Comparison of Text-Independent Speaker Recognition Methods Using VQ-Distortion and Discrete/Continuous HMM's*. IEEE Trans. on Speech and Audio Processing, Vol. 2, No. 3, July 1994.

[Mat95] Matsui T., Kanno T., Furui S., *Speaker Recognition Using HMM Composition in Noisy Environments*. Proceeding of EUROSPEECH, Madrid, Spain, September 1995.

[MD79] Markel J.D., Davis S.B., *Text-Independent Speaker Recognition from a Large Linguistically Unconstrained Time-Spaced Data Base*. IEEE Trans. on ASSP, Vol. 27, No. 1, 1979.

[Mic96] Michalewicz Z., *Algorytmy genetyczne + struktury danych = programy ewolucyjne*. Warszawa, 1996.

[Mro96] Mrozek B., Mrozek Z., *MATLAB – Uniwersalne środowisko do obliczeń naukowo-technicznych*. PLJ, Warszawa 1996.

[Nis98] Proceedings of the NIST Speaker Recognition Workshop, Maryland, USA, March/April 1998

[Pol54] Pollack I., Pickett J.M., Sumby W.H., *The Identification of Speakers by Voice*. J. Acoust. Soc. Amer., Vol. 26, 1954.

[Pru66] Pruzansky S., Bricker P.D., *Effect of Stimulus Content and Duration on Talker Identification*. J. Acoust. Soc. Amer., Vol. 40, No. 6, 1966.

[Rab78] Rabiner L.R., Schafer R.W., *Digital Processing of Speech Signals*. Prentice Hall, Englewood Cliffs, New Jersey, 1978.

[Rab89] Rabiner L.R., *A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition*. Proceedings of the IEEE , Vol. 77, No. 2, 1989.

[Rey95] Reynolds D., Carlson B., *Text-Independent Speaker verification Using Decoupled and Integrated Speaker and Speech Recognizers*. Proceeding of EUROSPEECH, Madrid, Spain, September 1995.

[Rey95a] Reynolds D., *Speaker Identification and Verification Using Gaussian Mixture Speaker Models*. Speech Communication, Vol. 17, 1995.

[Rey95b] Reynolds D., Rose R., *Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models*., IEEE Trans. on Speech and Audio Processing, Vol. 3, No. 1, January 1995.

[Rey96] Reynolds D., M.I.T. Lincoln Laboratory Site Presentation in Speaker Recognition Workshop. (http://jaguar.ncsl.nist.gov/speaker/).

[Ros73] Rosenberg A.E., *Listeners Performance in Speaker Verification Tasks*. IEEE Trans. on Audio Electroacoustics, Vol. 21, 1973.

[Ros76] Rosenberg A.E., *Automatic Speaker Verification: A Review*. Proceedings of the IEEE Vol. 64 No. 4 April 1997.

[Ros91] Rosenberg A.E., Soong F.K. *Recent Research in Automatic Speaker Recognition*. In *Advances in Speech Signal Processing,* edited by Furui S., Sondhi M., Marcel Dekker, New York, 1991.

[Rut96] Ruth A.D., Stearns S.D., *Signal Processing Algorithms in MATLAB*, Prentice Hall, New Jersey, 1996.

[Sai85] Saito S., Nakata K., *Fundamentals of Speech Signal Processing*. Academic Press, Tokyo, Japan, 1985.

[Sas93] Sasal W., *Układy scalone serii UCY74LS i UCY74S – parametry i zastosowania*. WKiŁ, Warszawa, 1993.

[Sch75] Schafer R.W., Rabiner L.R., *Digital Representation of Speech Signals*. Proceedings of the IEEE, Vol. 63, No. 4, April 1975s.

[Sch92] Schalkoff R., *Pattern recognition: statistical, structural and neural approaches*. John Wiley & Sons, New York, 1992.

[Sho98] Shomali A., Kapusta M., *Rozpoznawanie głosów na podstawie dyskryminacji amplitudowej sygnału mowy*. Półrocznik Automatyki, Tom 3, Zeszyt 1, AGH - Kraków 1998.

[Sho99] Shomali A., Kapusta M., Al-Hadidi M., *Identyfikacja głosów na podstawie długo-okresowego histogramu amplitudowego sygnału mowy*. Półrocznik Automatyki, Tom 2, Zeszyt 2, AGH - Kraków, 1999.

[Sho99a] Shomali A., Kapusta M., Al-Hadidi M., *Weryfikacja mówcy na podstawie długookresowego histogramu amplitudowego sygnału mowy*. Materiały konferencyjne I Ogólnopolskie Warsztaty Doktoranckie, Istebna-Zaolzie, październik, 1999.

[Sho99b] Shomali A., Kapusta M., Gajer M., *Zastosowanie niejawnych modeli Markowa w systemach automatycznego rozpoznawania mowy*. Kwartalnik Elektrotechnika i Elektronika - AGH, zeszyt 3, Kraków, 1999.

[Sho99c] Shomali A., Kapusta M., Al-Hadidi M., *Kodowanie sygnału mowy za pomocą jego ekstremów czasowych* (w opracowaniu).

[Soo85] Soong F.K., Rosenberg L.R., Rabiner L.R., *A Vector Quantization Approach to Speaker Recognition*. Proceeding of ICASSP85, 1985.

[Ste96] Stearns S.D., David R.A., *Signal Processing Algorithms in MATLAB*. Prentice Hall, New Jersey, 1996.

[Sza90] Szabatin J., *Podstawy Teorii Sygnałów*. WKiŁ, Warszawa 1990.

[Tad88] Tadeusiewicz R., *Sygnał Mowy*. WKiŁ, Warszawa 1988.

[Tad92] Tadeusiewicz R., *Systemy Wizyjne Robotów Przemysłowych*. WNT, Warszawa 1992.

[Tad93] Tadeusiewicz R.: *Sieci neuronowe*. Akademicka Oficyna Wydawnicza RM, Warszawa, 1993.

[Tis91] Tishby N.Z., *On the Application of Mixture AR Hidden Markov Models to Text Independent Speaker Recognition*. IEEE Trans. on Signal Processing, Vol. 39, No. 3, March 1991.

[Vas96] Vaseghi S.V., *Advanced Signal Processing and Digital Noise Reduction*. Wiley - Teubner Communications, New York, 1996.

[Wal98] Walkowiak T., *Metoda rozpoznawania mówcy oparta o przetwarzanie perceptualne i sieci neuronowe typu RBF*. Rozprawa doktorska, Instytut Cybernetyki Technicznej – Politechnika Wrocławska, Wrocław, 1998.

[Wil71] William S., Mohn Jr., *Two Statistical Feature Evaluation Techniques Applied to Speaker Identification*. IEEE Trans. on Computers Vol. 20, No. 9, September 1971.

[Wol72] Wolf J.J., *Efficient Acoustic Parameters for Speaker Recognition*. J. Acoust. Soc. Amer., Vol. 51 No. 6 (part 2), March 1972.

[Yua99] Yuan Z.X., Xu B.L., Yu C.Z., *Binary Quantization of Feature Vectors for Robust Text-Independent Speaker Identification*. IEEE Trans. on ASSP, Vol. 7, No. 1, January 1999.

[Zil98] Zilovic M.S., Ramachandran R.P., Mammone R.J., *Speaker Identification Based on the Use of Robust Cepstral Features Obtained form Pole-Zero Transfer Functions*. IEEE Trans. on ASSP, Vol. 6, No. 3, May 1998.

[Zis96] Zissman M.A., *Comparison of four approaches to automatic language identification of telephone speech.* Proceedings of the IEEE Vol. 4 No. 1, January 1996.