Politechnika Wrocławska

Instytut Telekomunikacji, Teleinformatyki i Akustyki

Raport Nr I28/PRE-001/07

PRACA DOKTORSKA

Algorytmy kompensacji warunków transmisyjnych i cech osobniczych mówcy w systemach automatycznego rozpoznawania mowy

Paweł Mrówka

Promotor: dr hab. inż. Ryszard Makowski, prof. PWr

Wrocław, 2007

Dziękuję Panu prof. Ryszardowi Makowskiemu
za cenne wskazówki udzielone mi podczas
realizacji niniejszej pracy.

Spis treści

Wykaz ważniejszych skrótów i oznaczeń

ARM automatyczne rozpoznawanie mowy

DCT discrete cosine transform - dyskretna transformacja kosinusowa

DFT discrete Fourier transform - dyskretna transformacja Fouriera

DWT discrete wavelet transform - dyskretna transformacja falkowa

E-M estymacja typu expectation-maximization

EV Eigenvoices (algorytm)

FIR finite impulse response - skończona odpowiedź impulsowa

GMM Gaussian mixture models - modele wykorzystujące sumy krzywych Gaussa

HMM hidden Markov models - ukryte modele Markowa

IDCT odwrotna DCT

IDFT odwrotna DFT

IIR infinite impulse response - nieskończona odpowiedź impulsowa

IRDCT odwrotna RDCT

LPC linear prediction coefficients - współczynniki prognozy liniowej

MFCC mel frequency cepstral coefficients - melowe współczynniki cepstralne

MLLR maximum likelihood linear transform - transformacja liniowa maksymalnej wiarogodności

PCA principal component analysis - analiza składowych głównych

RDCT DCT z odrzuceniem końcowych współczynników

SA speaker adaptive - adaptujący się do mówcy

SAT	speaker adaptive training - uczenie ukierunkowane na kompensację
SD	speaker dependent - zależny od mówcy
SI	speaker independent - niezależny od mówcy
SNR	signal to noise ratio - stosunek mocy sygnału do mocy szumu
VAD	voice activity detector - detektor obecności sygnału mowy
VTLN	vocal tract length normalization - normalizacja długości toru głosowego
VTS	vector Taylor series - aproksymacja funkcji zniekształceń za pomocą szeregu Taylora

W pracy używano głównie zapisu macierzowego. Macierze oznaczano pogrubionymi wielkimi literami, wektory (domyślnie kolumnowe) pogrubionymi małymi literami, a elementy macierzy oraz inne wielkości skalarne niepogrubionymi małymi literami. Położenie elementu macierzy oznaczano w indeksie dolnym w kolejności: numer wiersza, numer kolumny, bez oddzielania numerów przecinkami. W przypadku numerowania wielkości nie będących elementami macierzy stosowano również indeks dolny, lecz z oddzielaniem numerów przecinkami. Symboli w nawiasach w indeksie górnym używano do oznaczania wielkości w celu zwiększenia liczby dostępnych oznaczeń. Funkcje oznaczano niepogrubionymi literami wielkimi lub małymi.

$\mathbf{A}$	macierz prawdopodobieństw przejść między stanami w HMM
$c^{(rm,p)}_{k,s,\mathscr{J}^{(sp)}}$	początkowe rozpoznawalności izolowanych ramek dla spółgłosek dla mówcy s
$c^{(ro)}_k$	funkcje celu optymalizacji wartości parametrów transformacji widma
$c^{(rs)}_k$	miary oceny rozpoznawalności izolowanych ramek
$\mathbf{D}^{(f)}$	macierz odległości międzyfonemowych
f	częstotliwość
$f^{(max)}$	maksymalna częstotliwość analizowanego sygnału
$f^{(norm)}$	częstotliwość po skalowaniu
$f^{(p)}$	częstotliwość próbkowania
$f^{(v)}$	częstotliwość tonu krtaniowego
$f^{(v,p)}$	częstotliwość progowa tonu krtaniowego w klasyfikacji mówców

$g(f, \boldsymbol{\alpha})$	funkcja skalowania osi częstotliwości z parametrami $\boldsymbol{\alpha}$
$\mathbf{h}_i^{(kan)}$	dyskretne charakterystyki amplitudowe symulowanych zniekształceń transmisyjnych
$h^{(v)}(n)$	odpowiedź impulsowa toru głosowego
$I^{(k)}$	liczba iteracji w iteracyjnym algorytmie poprawiania wyniku rozpoznania
$K^{(g)}$	liczba elementów w banku funkcji skalowania osi częstotliwości
$K^{(h)}$	liczba elementów w banku filtrów
$K^{(kl)}$	liczba klas mówców
$\mathbf{L}$	macierz stosowana w zmodyfikowanej normie Euklidesa
$\mathcal{N}(\cdot\,; \boldsymbol{\mu}, \boldsymbol{\Sigma})$	wielowymiarowy rozkład normalny o wartości oczekiwanej $\boldsymbol{\mu}$ i macierzy kowariancji $\boldsymbol{\Sigma}$
$\mathbf{o}$	wektor współczynników MFCC ramki sygnału bez zniekształceń
$\mathbf{o}^{(norm)}$	wektor współczynników MFCC po normalizacji
$\mathbf{o}^{(zn)}$	wektor współczynników MFCC ramki sygnału zniekształconego
$\mathbf{O}$	sekwencja wektorów $\mathbf{o}$
$p(\cdot)$	funkcja gęstości prawdopodobieństwa
$p_i(\mathbf{o})$	funkcja gęstości prawdopodobieństwa obserwacji wektora $\mathbf{o}$ dla i-tego stanu HMM
$p_i^{(d)}(t)$	rozkład prawdopodobieństwa czasu trwania fonemu i
P	prawdopodobieństwo
$P_i^{(cisz)}$	prawdopodobieństwo fonemu i w ramce ciszy
$P^{(s,l)}$	zlogarytrmowane prawdopodobieństwo optymalnej ścieżki stanów wyznaczonej algorytmem Viterbiego
$P_k^{(wyr)}$	pseudoprawdopodobieństwo określające przynależność rozpoznawanej wypowiedzi do wyrazu k ze słownika
q	stan w modelu języka, stan HMM
$\mathbf{q}$	sekwencja stanów q

$\mathbf{s}$	dyskretne widmo amplitudowe ramki sygnału
$\mathbf{s}^{(norm)}$	dyskretne widmo amplitudowe ramki sygnału po normalizacji
$\mathbf{s}^{(zn)}$	dyskretne widmo amplitudowe ramki sygnału zniekształconego
$u(\cdot)$	funkcja skoku jednostkowego
$\mathbf{U}$	macierz wektorów bazowych w algorytmie EV
$x(n)$	sygnał w dziedzinie czasu
$x^{(v)}(n)$	sygnał pobudzenia głosowego w dziedzinie czasu
$\mathbf{x}^{(i,c)}$	centroid klastra odpowiadającego i-temu elementowi banku
$\mathbf{x}^{(i,d1)}, \mathbf{x}^{(i,d2)}$	elementy dodatkowe i-tego elementu banku
$\boldsymbol{\alpha}^{(g)}, \boldsymbol{\alpha}^{(g,r)}, \boldsymbol{\alpha}^{(g,o)}$	parametry funkcji skalowania osi częstotliwości określające odpowiednio: współrzędne punktów łączenia odcinków, tylko rzędne tych punktów, tylko odcięte tych punktów
$\boldsymbol{\alpha}^{(st)}$	parametry transformacji widma
$\boldsymbol{\beta}$	parametry adaptacji w metodzie EV
$\gamma_{i,t,k}^{(gmm)}(\mathbf{o}_t)$	prawdopodobieństwo przynależnosci wektora $\mathbf{o}_t$ do i-tego stanu HMM i k-tej składowej GMM rozkładu prawdopodobieństwa współczynników MFCC dla tego stanu
$\boldsymbol{\mu}$	wektor wartości oczekiwanej
$\boldsymbol{\mu}^{(sv,sr)}$	średni superwektor w metodzie EV
$\boldsymbol{\pi}$	wektor prawdopodobieństw początkowych stanów w HMM
Θ	statystyczny model języka
σ^2	wariancja
$\boldsymbol{\Sigma}$	macierz kowariancji
$\mathrm{diag}(\mathbf{x})$	macierz przekątniowa z elementami na przekątnej równymi elementom wektora $\mathbf{x}$
$\dim(\mathbf{x})$	wymiar wektora $\mathbf{x}$
$\|\cdot\|_L$	zmodyfikowana norma Euklidesa
$*$	operacja splotu
$\circ$	operacja mnożenia odpowiadających sobie elementów dwóch wektorów

Spis rysunków

Spis tabel

1. Wprowadzenie

Automatyczne rozpoznawanie mowy (ARM) ma na celu zdekodowanie przez maszynę informacji znaczeniowej zawartej w ludzkiej mowie. Dane wejściowe dla systemu ARM stanowi najczęściej cyfrowy sygnał akustyczny, jedno lub wielokanałowy, zarejestrowany w bliskim lub dalekim polu akustycznym. Istnieją jednak systemy wykorzystujące jako dane wejściowe np. sekwencje filmowe ruchu ust mówcy, przy czym takie dodatkowe źródła danych stanowią na ogół tylko uzupełnienie danych akustycznych. Informacja zdekodowana przez system i podana na jego wyjście może być na różnych poziomach złożoności. Najprostszym jest klasyfikacja pojedynczych fragmentów (ramek) sygnału względem zadanego zbioru jednostek językowych (np. allofonów, fonemów, diafonów, sylab). Poziomem wyższym jest rozpoznawanie tych jednostek, lecz przy ich automatycznym wyodrębnianiu z ciągłego sygnału. Kolejnymi poziomami są: rozpoznawanie wyrazów, jako ciągów przyjętych podstawowych jednostek fonetycznych, oraz rozpoznawanie zdań, jako ciągów wyrazów. Sygnał mowy, oprócz informacji czysto językowej, zawiera również informacje prozodyczne, zawarte w intonacji i akcencie, mówiące o stanie emocjonalnym mówcy czy też emocjach, jakie celowo zawarł on w wypowiedzi. Rozpoznawanie informacji prozodycznych może być kolejnym, wyższym poziomem złożoności systemu ARM, ale może być również jedynym zadaniem systemu dedykowanego do tego celu. Ponadto systemy ARM mogą mieć jeszcze inne funkcje, jak np. identyfikacja stanów patologicznych narządu mowy u mówcy.

Przydatność skutecznych systemów ARM nie budzi wątpliwości. Stanowią one znaczne ułatwienie komunikacji człowiek-maszyna, zwłaszcza w przypadkach, gdy inne metody tej komunikacji są niemożliwe lub utrudnione. Można tutaj wymienić np. obsługę urządzeń w czasie prowadzenia pojazdów, systemy bezpieczeństwa aktywowane głosem czy obsługę urządzeń przez osoby niepełnosprawne. Systemy ARM wyręczyć mogą również człowieka w pracach żmudnych i schematycznych takich, jak obsługa telefonicznych systemów informacyjnych czy pisanie dyktowanych tekstów.

Wyróżnić można kilka kryteriów klasyfikacji systemów ARM ze względu na ich cechy funkcjonalne. Poniżej przedstawiono klasyfikację zaproponowaną w [10]:

1. Struktura i złożoność rozpoznawanych wypowiedzi.
    - Izolowane wyrazy. Mały (do 1000 wyrazów) słownik.
    - Sekwencje wyrazów wypowiadane w sposób ciągły. Mały słownik.
    - Zdania. Specjalistyczny (do 10 000 wyrazów) słownik.
    - Mowa dyktowana. Wielki (powyżej 10 000 wyrazów) słownik.
    - Mowa naturalna. Słownik otwarty, nieograniczony.
2. Możliwość użytkowania przez różnych mówców.
    - System zaprojektowany do pracy z jednym danym mówcą (SD - *speaker dependent*).
    - System adaptujący się do danego mówcy (SA - *speaker adaptive*).
    - System zapewniający pracę z wieloma mówcami (SI - *speaker independent*).
3. Warunki pracy i jakość sygnału.
    - Bardzo małe zniekształcenia transmisyjne sygnału. Te same warunki transmisyjne podczas uczenia systemu i jego pracy użytkowej.
    - Różne warunki transmisyjne podczas uczenia systemu i jego pracy użytkowej.
    - Silne zniekształcenia transmisyjne: szum, zniekształcenia liniowe i nieliniowe, nałożenie mowy wielu mówców.
4. Konieczne do zapewnienia zasoby sprzętowe związane ze złożonością obliczeniową systemu.
    - Specjalizowany serwer.
    - Popularny komputer klasy PC.
    - Prosty mikroprocesorowy system sterujący urządzeniem.

Trzeba zaznaczyć, że powyższy podział ma charakter orientacyjny, a granice między typami systemów są często nieostre. W każdym kryterium poszczególne klasy systemów zostały wymienione zgodnie z wzrastającym stopniem trudności ich zaprojektowania.

Problem ARM został podjęty już w latach 50-tych XX wieku. Pierwsze systemy miały na celu rozpoznawanie pojedynczych słów wypowiadanych przez jednego mówcę, przy czym stosowano w tym celu głównie analizę widmową samogłosek. W latach 60-tych rozwijano metody rozpoznawania wzorców słów, zaproponowano w tym celu wykorzystanie algorytmów programowania dynamicznego, np. dynamicznej

transformacji czasowej (DTW - *dynamic time warping*). W latach 70-tych podjęto prace nad systemami niezależnymi od mówcy oraz zaproponowano nowe metody parametryzacji, np. współczynniki prognozy liniowej (LPC - *linear prediction coefficients*). W latach 80-tych zaproponowano wykorzystanie ukrytych modeli Markowa (HMM - *hidden Markov models*), co było jednym z najważniejszych punktów przełomowych w historii badań nad ARM. Zaczęto wykorzystywać również analizę cepstralną podczas parametryzacji sygnału. Począwszy od lat 80-tych rozwijane były systemy rozpoznawania połączonych wyrazów i zdań oraz metody zapewniające skuteczne rozpoznawanie dla wielu mówców i w różnych warunkach transmisyjnych. Lata 90-te przyniosły znaczny postęp w rozwoju systemów rozpoznawania mowy ciągłej i naturalnej [133, 9].

Obecnie istniejące i dostępne komercyjnie systemy rozpoznawania mowy dyktowanej pozwalają osiągnąć skuteczność ponad 95% prawidłowo rozpoznanych wyrazów w przypadku SD lub SA i dobrej jakości sygnału, mają one jednak bardzo dużą złożoność obliczeniową, często wymagającą zastosownia specjalistycznych serwerów. Natomiast dostępne systemy o małym słowniku pozwalają na osiągnięcie podobnej skuteczności, ale już w przypadku SI, w obecności zniekształceń transmisyjnych oraz przy znacznie mniejszym zapotrzebowaniu na zasoby sprzętowe - możliwa jest ich implementacja w stosunkowo prostych systemach sterowania urządzeniami.

Wśród największych ośrodków zajmujących się przez minione 50 lat problemem ARM wymienić można AT&T Bell Labs, MIT, IBM, Cambridge (ogólnodostępny i darmowy system HTK [181]), Microsoft, Nuance. W Polsce ukazało się stosunkowo niewiele publikacji, do ważniejszych najnowszych można zaliczyć: [99, 34, 80, 168, 47, 156, 9].

ARM jest, pomimo kilkudziesięciu lat badań, zagadnieniem, które wciąż nie doczekało się kompleksowego i pełnego rozwiązania. Istniejące systemy zapewniają skuteczność dorównującą człowiekowi, lecz przy jednoczesnych ograniczeniach funkcjonalności takich, jak niewielki słownik, zamknięty zbiór mówców czy brak silnych zniekształceń transmisyjnych. Obecnie prowadzone badania są wielokierunkowe, przy czym wśród najważniejszych kierunków można wymienić: projektowanie coraz lepszych modeli języka na wszystkich poziomach hierarchii (akustycznej, fonetycznej, syntaktycznej i semantycznej), przy czym główny nacisk kładzie się na opracowanie skutecznych modeli na wyższych poziomach, jak również na uzupełnienie sygnału akustycznego innymi źródłami informacji, jak np. ruchy ust mówcy; opracowywanie metod szybkiej adaptacji systemów do zmiennych i trudnych warunków transmisyjnych; integracja systemów ARM z systemami wykorzystującymi mechanizmy sztucznej inteligencji takimi, jak systemy ekspertowe, systemy komunikacji z maszyną za pomocą języka naturalnego, systemy automatycznego tłumaczenia.

Rozpoznawanie komend jest zagadnieniem dobrze znanym, które doczekało się wielu skutecznych rozwiązań, lecz pomimo tego wciąż możliwe jest zaproponowanie

rozwiązań skuteczniejszych. Obecnie większość badań skupia się na systemach rozpoznawania mowy dyktowanej lub naturalnej i metody kompensacji wpływu zmiennych warunków transmisyjnych oraz zmiennych cech osobniczych mówcy opracowywane są pod ich kątem. Metody te nie są możliwe do zaimplementowania w sposób bezpośredni w systemach rozpoznawania komend, gdyż wymagają do przeprowadzenia kompensacji fragmentów mowy o długości co najmniej kilku sekund. Izolowana komenda trwa natomiast najczęściej poniżej 1 sekundy, a nierzadko poniżej 500 ms. Istnieje zatem potrzeba opracowania skutecznych metod kompensacji działających dla bardzo krótkich wypowiedzi. W niniejszej pracy skoncentrowano się na kompensacji wpływu liniowych zniekształceń transmisyjnych i cech osobniczych mówcy w systemie rozpoznawania komend.

1.1. Zniekształcenia transmisyjne i zmienność osobnicza mówców

Dźwięk wytwarzany przez mówcę po wypromieniowaniu przez usta, zanim zostanie wprowadzony do systemu ARM, jest transmitowany złożonym torem. Tor ten stanowi w większości przypadków łańcuch zawierający część akustyczną od ust mówcy do przetwornika elektroakustycznego, następnie część elektryczną do przetwornika analogowo-cyfrowego, a dalej część cyfrową. Części toru może być jednak więcej, np. może występować pośrednie urządzenie rejestrujące lub transmisja analogowa bądź cyfrowa drogą radiową, światłowodową czy też przewodami elektrycznymi. Każda część toru może wnosić specyficzne zniekształcenia liniowe, nieliniowe oraz szum. Stosowany najczęściej model toru transmisyjnego, uwzględniający zniekształcenia liniowe i szum addytywny, dany jest równaniem:

$$x^{(zn)}(n) = x(n) * h^{(zn)}(n) + n(n) \tag{1.1}$$

gdzie $x^{(zn)}(n)$ i $x(n)$ oznaczają odpowiednio sygnał zniekształcony i niezniekształcony, $h^{(zn)}(n)$ - odpowiedź impulsową toru transmisyjnego, a $n(n)$ - szum addytywny. W przypadku modelowania zniekształceń zmiennych w czasie $h^{(zn)}(n)$ jest zależna od czasu. Symbol $*$ oznacza operację splotu.

Mowa ma niezwykle ciekawe własności równoczesnego przenoszenia informacji zarówno o treści wypowiedzi (treść językowa i prozodyczna), jak i o tożsamości mówcy. Z punktu widzenia ARM interesująca jest informacja o treści językowej, natomiast informacja o tożsamości oraz zazwyczaj również informacja o treści prozodycznej jest zbędna i związana z nimi zmienność cech sygnału mowy przyczynia się na ogół do pogorszenia skuteczności systemu. Poza nielicznymi propozycjami (np. [34, 67, 127, 79]) systemy ARM wykorzystują analizę widmową sygnału mowy, a zatem zmienność osobnicza analizowana jest również w dziedzinie widma. Uproszczony model matematyczny wytwarzania mowy jest następujący [40]:

$$x\left(n\right) = x^{(v)}\left(n\right) * h^{(v)}\left(n\right) \tag{1.2}$$

gdzie $x(n)$ oznacza sygnał mowy, $x^{(v)}(n)$ - sygnał pobudzenia głosowego, a $h^{(v)}(n)$ - zmienną w czasie odpowiedź impulsową toru głosowego. Analizując czasowo-częstotliwościową strukturę wielu realizacji fragmentów sygnału mowy odpowiadającym tej samej jednostce fonetycznej wyróżnić można pewne różnice o charakterze systematycznym, związane z systematycznymi zmianami widma pobudzenia i transmitancji toru głosowego. Różnice te określane są jako osobnicze, przy czym dzieli się je na międzyosobnicze - występujące pomiędzy różnymi mówcami oraz wewnątrzosobniczne - występujące dla danego mówcy.

W rozdziale 2.3 opisano bardziej szczegółowo widmową zmienność sygnału mowy związaną ze zniekształceniami transmisyjnymi i cechami osobniczymi, a w dodatku A przedstawiono mechanizm wytwarzania mowy.

1.2. Założenia, cele i teza pracy

Teza pracy:

Możliwe jest zaprojektowanie algorytmu łącznej kompensacji warunków transmisyjnych i cech osobniczych mówcy dla systemu rozpoznawania bardzo krótkich i izolowanych wypowiedzi, charakteryzującego się skutecznością nie mniejszą niż algorytmy znane dotychczas.

Powyższa teza wymaga uzupełnienia w zakresie warunków projektowania i działania systemu ARM, które są następujące:

- Rozpoznawane są bardzo krótkie izolowane wypowiedzi, najczęściej pojedyncze wyrazy, wypowiadane przez różnych mówców (system niezależny od mówcy).
- Szum addytywny i zniekształcenia nieliniowe są na poziomie pozwalającym zaniedbać ich wpływ.
- Zniekształcenia liniowe charakteryzują się łagodnym przebiegiem charakterystyki amplitudowej oraz są wolnozmienne w czasie. Nie występują zniekształcenia całkowicie tłumiące sygnał użyteczny w podpasmach.
- Nie występuje silny pogłos charakteryzujący się ostrymi maksimami charakterystyki amplitudowej.
- Możliwa jest praca z niewielką bazą nagrań, już od 30 mówców w części bazy przeznaczonej do uczenia systemu.

Przyjęte w założeniach warunki akustyczne działania systemu ARM odpowiadają zastosowaniom w pomieszczeniach mieszkalnych, biurowych, dobrze wytłumionych wnętrzach pojazdów. W gorszych warunkach transmisyjnych możliwe jest jednak pewne polepszenie własności sygnału przeprowadzone przed wprowadzeniem go do systemu ARM, np. zredukowanie szumu [145, 129, 12] czy pogłosu [121].

Do osiągnięcia celu naukowego zawartego w tezie pracy wymagane jest wykonanie następujących zadań:

1. Opracowanie narzędzi badawczych w postaci bazy nagrań i kilku wersji systemu ARM.
2. Przegląd znanych rozwiązań oraz ich analiza teoretyczna i eksperymentalna pod kątem przydatności w rozwiązywanym w pracy zagadnieniu.
3. Zaprojektowanie oryginalnej metody kompensacji zniekształceń transmisyjnych i cech osobniczych mówcy.
4. Badania eksperymentalne zaproponowanej metody i analiza ich wyników.

Układ dalszej części pracy jest następujący:

W rozdziale drugim przedstawiono metody ARM w oparciu o statystyczne modele języka, analizę przyczyn występowania zniekształceń transmisyjnych i zmienności cech osobniczych mówcy oraz ich wpływu na działanie systemu ARM, przeprowadzono również przegląd i analizę kilku znanych rozwiązań zagadnienia kompensacji. W rozdziale trzecim przedstawiono modyfikację algorytmu Eigenvoices i jego analizę eksperymentalną. W rozdziale czwartym opisano zaproponowaną oryginalną metodę kompensacji. Rozdział piąty zawiera uzyskane wyniki, a rozdział szósty podsumowanie. W dodatkach zamieszczono opis mechanizmu wytwarzania mowy, opis bazy nagrań i systemu ARM oraz szczegółowe opisy niektórych stosowanych algorytmów.

2. Metody ARM i algorytmy kompensacji

Poniżej przedstawiono podstawy działania systemów ARM opartych o statystyczne modele języka. Podano terminologię związaną z kompensacją w systemach ARM, opisano wpływ zmiennych warunków transmisyjnych i cech osobniczych mówcy na widmo i parametry MFCC sygnału mowy oraz wyjaśniono przyczyny spadku skuteczności systemu spowodowanego tą zmiennością. Przeprowadzono również przegląd rozwiązań znanych z literatury przedmiotu.

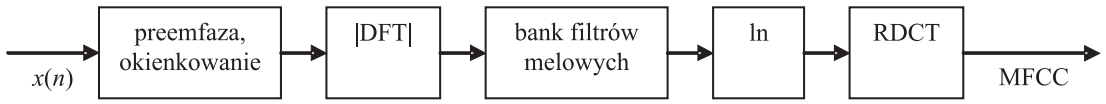
2.1. Rozpoznawanie komend

Zadanie automatycznego rozpoznawania komend polega na rozpoznawaniu izolowanych, bardzo krótkich wypowiedzi, najczęściej pojedynczych wyrazów lub ustalonych grup wyrazów. Niekiedy systemy umożliwiają rozpoznawanie ciągu komend połączonych, wypowiadanych bez przerw. Od systemów rozpoznawania mowy ciągłej różni je mały, zamknięty słownik oraz bardzo prosty model gramatyczny języka, którego rola ogranicza się do narzucania reguł odnośnie kolejności występowania wypowiadanych komend.

2.1.1. Parametryzacja

Celem parametryzacji sygnału mowy na potrzeby ARM jest takie przekształcenie sygnału wejściowego (przebiegu zmian ciśnienia akustycznego), by uzyskać możliwie małą liczbę parametrów zawierających informacje istotne dla systemu, tj. o zawartości fonetycznej sygnału, przy jednoczesnej minimalizacji wrażliwości tych parametrów na zmienność sygnału nieistotną z punktu widzenia ARM. Zaproponowano dotychczas wiele metod parametryzacji, można tutaj wymienić np. LPC [133], PLP (*perceptual linear prediction*) [53], EIH (*ensemble interval histogram*) [133], parametryzację opartą na modelach chaotycznej dynamiki nieliniowej [79, 127, 67], parametryzację metodą siatek o zmiennych parametrach [34], parametryzację opartą o demodulację ciągłą sygnału operatorem Teagera-Kaisera [62], parametryzację wykorzystującą estymację widma metodą MVDR (*minimum variance distortionless response*) [30], metodę *Subband Spectral Centroid Histograms* [41]. Jedną z najczę-

ściej obecnie stosowanych metod parametryzacji są melowe współczynniki cepstralne (MFCC - *mel frequency cepstral coefficients*) [10]. Charakteryzuje się ona zadowalającą skutecznością przy umiarkowanej złożoności obliczeniowej w porównaniu z innymi metodami. Na rys. 2.1 przedstawiono schemat parametryzacji MFCC.



Rys. 2.1. Schemat parametryzacji MFCC.

Matematyczny opis parametryzacji MFCC przedstawiają poniższe równania:

$$x^{(p)}(n) = x(n) - \gamma^{(pre)} \cdot x(n-1) \tag{2.1}$$

$$x^{(pw)}(n) = x^{(p)}(n_0 + n) \cdot w(n), \qquad n = 0, \ldots, N-1 \tag{2.2}$$

$$\mathbf{s} = \left| \mathrm{DFT}\left(\left[x^{(pw)}(0)\ x^{(pw)}(1)\ \ldots\ x^{(pw)}(N-1)\ \underbrace{0\,0\ \ldots\ 0}_{K-N} \right]^T \right) \right| \tag{2.3}$$

$$s_j^{(m)} = \sum_{k=0}^{K/2-1} h_{k,j}^{(mel)} \cdot s_k, \qquad j = 0, \ldots, J-1 \tag{2.4}$$

$$s_j^{(l)} = \ln\left(s_j^{(m)}\right), \qquad j = 0, \ldots, J-1 \tag{2.5}$$

$$o_m = \sum_{j=0}^{J-1} \left(s_j^{(l)} \cdot \cos\left(m \left(j + \frac{1}{2} \right) \frac{\pi}{J} \right) \right), \qquad m = 0, \ldots, M-1 \tag{2.6}$$

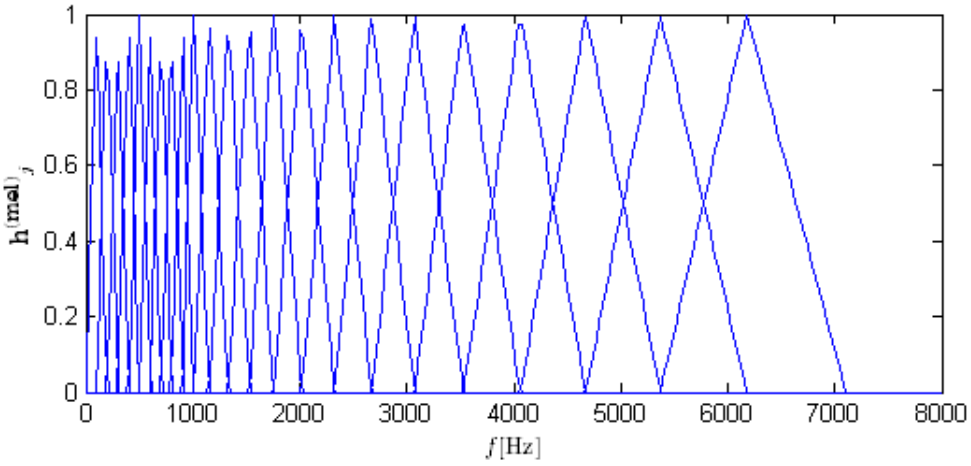
gdzie $x(n)$ oznacza sygnał wejściowy, $x^{(p)}(n)$ - sygnał po zastosowaniu preemfazy ze współczynnikiem $\gamma^{(pre)}$, $x^{(pw)}(n)$ - sygnał po preemfazie i nałożeniu okna czasowego $w(n)$ o długości N próbek, $\mathbf{s}$ - dyskretne widmo amplitudowe ramki sygnału o liczbie prążków K, $\mathbf{s}^{(m)}$ - wartości wyjściowe z J filtrów melowych o charakterystykach amplitudowych $\mathbf{h}_j^{(mel)}$, $\mathbf{o}$ - wektor współczynników MFCC. Równanie (2.6) opisuje dyskretną transformację kosinusową (DCT - *discrete cosine transform*), w której pominięto współczynniki o numerach wyższych od $M-1$. Tranformacja taka będzie dalej oznaczana jako RDCT. Bank filtrów melowych składa się z pasmowo przepustowych filtrów o trójkątnych charakterystykach amplitudowych, zachodzących na siebie, w których częstotliwości środkowe oddalone są o 100 melów. Często jednak nie stosuje się dokładnego przeliczania skali częstotliwości wyrażonej w melach na wyrażoną w hertzach, a położenia częstotliwości środkowych wyznacza się korzystając z funkcji wykładniczej.

Parametryzacja MFCC zawiera elementy modelowania własności percepcyjnych słuchu ludzkiego takie, jak zaniedbanie informacji o widmie fazowym, nieliniowa

skala częstotliwości, uśrednianie mocy sygnału w pasmach oraz nieliniowa reakcja na poziom mocy sygnału. Zastosowanie transformacji RDCT ma na celu zmniejszenie liczby współczynników, ponadto RDCT dokonuje częściowej dekorelacji uzyskanych współczynników, co jest istotne przy modelowaniu ich rozkładów prawdopodobieństwa.

Istnieje wiele modyfikacji algorytmu parametryzacji MFCC. Z ważniejszych można wymienić: zastąpienie operacji logarytmowania potęgowaniem (*root cepstrum*) [142], modyfikacje banku filtrów melowych [95, 86, 189], wprowadzenie zmiennej długości ramki [165]. Często również uwzględnia się w parametryzacji informacje o dynamice zmian w sygnale, poprzez dodanie do współczynników MFCC ich pochodnych po czasie (standardowo pierwszej i drugiej) lub też uwzględniając współczynniki z kilku sąsiednich ramek, a następnie przeprowadzając redukcję całkowitej liczby współczynników tak, aby wyodrębnić takie współczynniki, które zapewniają najlepszą zdolność klasyfikacji jednostek fonetycznych w systemie ARM. Najbardziej znane metody to LDA (*linear discriminant analysis*) i HDA (*heteroscedastic discriminant analysis*) [84].

We wszystkich badaniach przeprowadzonych na potrzeby niniejszej pracy wykorzystywano metodę MFCC z parametrami: $\gamma^{(pre)} = 0.97$, okno Hamminga długości 320 próbek (20 ms przy częstotliwości próbkowania $f^{(p)} = 16$ kHz), przy czym kolejne ramki pobierano co 160 próbek (10 ms), przed wykonaniem DFT sygnał w ramce uzupełniano zerami do długości 512 próbek. Zastosowano bank $J = 23$ filtrów melowych pokrywających pasmo częstotliwości 0 - 7.1 kHz (rys. 2.2), przy czym częstotliwości środkowe pierwszych 10 filtrów oddalone są o 100 Hz, a 13 kolejnych obliczane wg wzoru $f_n^{(sr)} = 1.15 \cdot f_{n-1}^{(sr)}$. Liczba współczynników MFCC wynosiła $M = 15$, nie stosowano pochodnych ani innych współczynników dynamicznych. Przed ustaleniem podanych parametrów metody MFCC przeprowadzono badania wstępne, w których



Rys. 2.2. Bank filtrów melowych.

testowano m. in.: adaptacyjne wyznaczanie parametru $\gamma^{(pre)}$, różną liczbę filtrów melowych, różną liczbę współczynników MFCC, zastosowanie potęgowania zamiast logarytmowania, zastosowanie różnych wag dla współczynników MFCC.

2.1.2. Statystyczne modele języka

ARM z wykorzystaniem wielowarstwowych statystycznych modeli języka opartych na HMM jest obecnie metodą dominującą. Rozwiązania wcześniejsze, rzadko obecnie stosowane, bazowały na porównywaniu wypowiedzi z bazą wzorców i wykorzystaniem odpowiednio skonstruowanych klasyfikatorów. Nowe propozycje, wykorzystujące np. *spiking neural networks* [104, 55] są skuteczne w niektórych zastosowaniach, ale wymagają jeszcze rozwiązania wielu zagadnień.

System wykorzystujący statystyczny model języka Θ ma na celu odwzorowanie sekwencji obserwacji $\mathbf{O} = (\mathbf{o}_0\ \mathbf{o}_1\ ...\ \mathbf{o}_{T-1})$ w sekwencję stanów modelu $\mathbf{q}^{(opt)} = \left(q_0^{(opt)}\ q_1^{(opt)}\ ...\ q_{T-1}^{(opt)}\right)$ tak, by każdej obserwacji był przyporządkowany pewien stan modelu. Zbiór $\mathcal{Q}$ zawiera wszystkie możliwe sekwencje stanów. Obserwacjami $\mathbf{o}_t$ mogą być w szczególności wektory współczynników MFCC. Stany q_t mogą odpowiadać fragmentom jednostek językowych bądź całym jednostkom, takim jak allofony, fonemy czy wyrazy. Jest to zależne od stopnia złożoności modelu. Stosuje się tutaj strukturę hierarchiczną, w której stanowi na poziomie wyższym (np. wyrazowi) odpowiada sekwencja stanów poziomu niższego (np. fonemów). W celu wyznaczenia sekwencji $\mathbf{q}^{(opt)}$ stosuje się Bayesowskie kryterium decyzyjne:

$$\begin{aligned}\mathbf{q}^{(opt)} &= \arg\max_{\mathbf{q}\in\mathcal{Q}} P\left(\mathbf{q}|\mathbf{O},\Theta\right) = \\ &= \arg\max_{\mathbf{q}\in\mathcal{Q}} \frac{P\left(\mathbf{O}|\mathbf{q},\Theta\right)\cdot P\left(\mathbf{q},\Theta\right)}{P\left(\mathbf{O},\Theta\right)} = \arg\max_{\mathbf{q}\in\mathcal{Q}} \left(P\left(\mathbf{O}|\mathbf{q},\Theta\right)\cdot P\left(\mathbf{q},\Theta\right)\right)\end{aligned} \tag{2.7}$$

W przypadku zastosowania dyskretnoczasowego HMM pierwszego rzędu przyjmuje się następujące uproszczenia modelu:

$$P\left(q_t|\mathbf{q}_{t-1} = (q_0\ q_1\ \ldots\ q_{t-1}),\Theta\right) = P\left(q_t|q_{t-1},\Theta\right) \tag{2.8}$$

$$P\left(\mathbf{O}|\mathbf{q},\Theta\right) = \prod_{t=0}^{T-1} P\left(\mathbf{o}_t|q_t,\Theta\right) \tag{2.9}$$

Dodatkowo przyjmuje się stacjonarność modelu:

$$P\left(q_t = j|q_{t-1} = i,\Theta\right) = a_{ij}, \qquad t = 1,\ldots,T-1 \tag{2.10}$$

$$P\left(q_t = i|\Theta\right) = \pi_i, \qquad t = 0 \tag{2.11}$$

$$P\left(\mathbf{o}_t|q_t = i,\Theta\right) = p_i\left(\mathbf{o}_t\right), \qquad t = 0,\ldots,T-1 \tag{2.12}$$

Po zastosowaniu uproszczeń równanie (2.7) przyjmuje postać:

$$\mathbf{q}^{(opt)} = \arg\max_{\mathbf{q}\in\mathscr{Q}} \left(\prod_{t=0}^{T-1} p_{q_t}\left(\mathbf{o}_t\right) \cdot \pi_{q_0} \cdot \prod_{t=1}^{T-1} a_{q_{t-1}q_t} \right) \tag{2.13}$$

W systemach ARM prawdopodobieństwa obserwacji dla danego stanu modelu modeluje się zazwyczaj ciągłymi rozkładami funkcji gęstości prawdopodobieństwa, stąd w równaniu (2.12) wprowadzono oznaczenie tych funkcji jako p_i dla i-tego stanu modelu. Zgodnie z zapisem, wyznaczania prawdopodobieństwa z funkcji gęstości dokonuje się całkując ją po jednostkowym hipersześcianie, którego środek wypada w punkcie obserwacji $\mathbf{o}_t$, a wartość $p_i\left(\mathbf{o}_t\right)$ jest stała i równa wartości w punkcie obserwacji. Formalnie powinno się stosować przejście graniczne przy wymiarach hipersześcianu dążących do zera, lecz w praktyce algorytmy skonstruowane są tak, że nie jest to konieczne. Rozkłady p_i modeluje się najczęściej jako sumy krzywych Gaussa (GMM - *Gaussian mixture models*):

$$p_i\left(\mathbf{o}\right) = \sum_{k=0}^{K-1} c_{i,k} \cdot \mathscr{N}\left(\mathbf{o}; \boldsymbol{\mu}_{i,k}, \boldsymbol{\Sigma}_{i,k}\right) \tag{2.14}$$

gdzie $\mathscr{N}\left(\,\cdot\,; \boldsymbol{\mu}_{i,k}, \boldsymbol{\Sigma}_{i,k}\right)$ oznacza wielowymiarowy rozkład Gaussa o wartości oczekiwanej $\boldsymbol{\mu}_{i,k}$ i macierzy kowariancji $\boldsymbol{\Sigma}_{i,k}$, a $c_{i,k} \in [0;1]$ oznacza wagi rozkładów.

Model języka w przypadku zastosowania opisanych uproszczeń składa się zatem z $\Theta = \left\{\mathbf{A} = [a_{ij}], \pi = [\pi_i], c_{i,k}, \boldsymbol{\mu}_{i,k}, \boldsymbol{\Sigma}_{i,k}\right\}$. Wartości parametrów modelu można wyznaczać różnymi metodami. Dobrze znana metoda Bauma-Welcha jest algorytmem estymacji typu E-M (*expectation-maximization*) i ma na celu maksymalizację funkcji wiarogodności modelu dla zadanego zbioru danych uczących [10, 133]. Jej zaletą jest analityczna postać rozwiązania zadania maksymalizacji funkcji wiarogodności w kolejnych iteracjach. Najpoważniejszą zaś wadą - nieuwzględnianie zdolności klasyfikacji modelu, która to zdolność ma bezpośrednie przełożenie na uzyskiwaną rozpoznawalność. Istnieje wiele metod estymacji parametrów modelu mających na celu maksymalizację rozpoznawalności. Do najpopularniejszych można zaliczyć MMI (*maximal mutual information*) [170, 11], MCE (*minimum classification error*) [172, 68], MPE (*minimum phone error*) [150]. Dla niektórych z nich (MMI, MPE) można uzyskać analityczne postaci rozwiązań zadania optymalizacji w kolejnych iteracjach (tzw. rozszerzony algorytm Bauma-Welcha), choć istnieją tu pewne problemy z jego zbieżnością. Często konieczne jest jednak zastosowanie algorytmów optymalizacji typu gradientowego lub poszukiwań prostych, co stanowi pewną niedogodność. Rozpoznawanie, czyli zadanie znalezienia sekwencji stanów modelu najlepiej „tłumaczącej" daną sekwencję obserwacji, przeprowadzane jest najczęściej z wykorzystaniem algorytmu Viterbiego. W przypadku, gdy liczba stanów jest duża, stosuje się techniki redukcji liczby analizowanych w tym algorytmie ścieżek.

Uproszczony model języka Θ nadaje się do zastosowania w systemach o niewielkim słowniku i prostym modelu gramatycznym, uwzględniającym tylko prawdopodobieństwa występowania danego wyrazu pod warunkiem wystąpienia danego wyrazu poprzedzającego (model 1-gram). Własność ta wynika z zastosowania HMM pierwszego rzędu, a prawdopodobieństwa występowania po sobie danych wyrazów odpowiadają prawdopodobieństwom przejść z ostatniego stanu wyrazu poprzedniego na pierwszy stan wyrazu następnego. W systemach rozpoznawania mowy dyktowanej i naturalnej stosuje się wiele modyfikacji tego modelu. Jedną z najważniejszych jest rozszerzenie modelu gramatycznego do modelu n-gram, gdzie n typowo wynosi 2 lub 3, a więc uwzględnia się dwa lub trzy wyrazy poprzedzające. Jednoetapowe algorytmy poszukiwania najlepszych sekwencji stanów dla modelu n-gram są bardzo złożone obliczeniowo, stąd rozpoznawanie przeprowadzane jest najczęściej kilkuetapowo. W etapie pierwszym uzyskuje się pewien zbiór najlepszych sekwencji stanów za pomocą algorytmu Viterbiego lub innego algorytmu poszukiwania wykorzystującego model 1-gram, w etapach kolejnych uzyskane sekwencje poddaje się wartościowaniu zdefiniowanymi miarami oceny, które z różnymi wagami mogą uwzględniać m.in.: prawdopodobieństwo sekwencji z modelu 1-gram, prawdopodobieństwo modelu n-gram, wartość syntaktyczną, wartość semantyczną [10, 133].

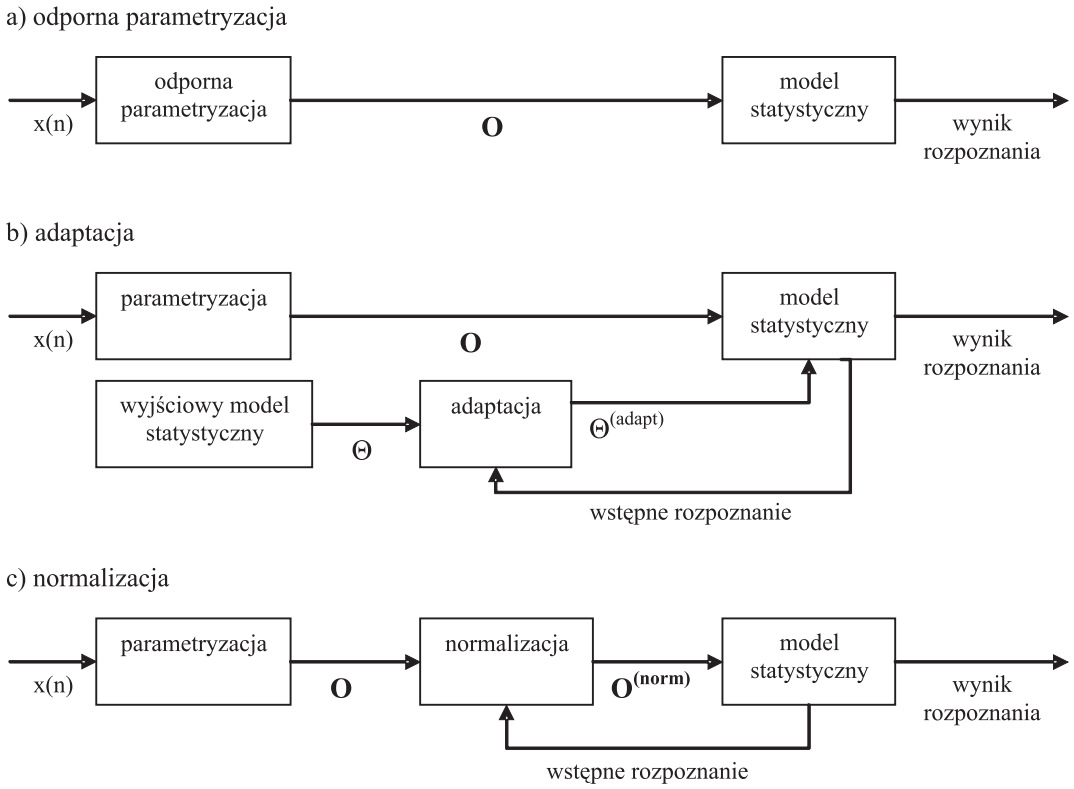
Obecnie prowadzone są intensywne badania nad ulepszeniem modelu języka, przy czym głównie próbuje się rozszerzyć model HMM poprzez zwiększenie jego rzędu, co wiąże się z koniecznością zaprojektowania zmodyfikowanych algorytmów uczenia i rozpoznawania, np. [176, 85]. Do jednych z prostszych, lecz efektywnych modyfikacji należy też dodatkowe modelowanie rozkładów prawdopodobieństwa czasów trwania stanów HMM [133].

Na potrzeby niniejszej pracy zaprojektowane zostały cztery warianty systemu ARM. Punktem wyjścia był model HMM pierwszego rzędu, przy czym w różnych wariantach zastosowano modyfikacje takie, jak modelowanie prawdopodobieństwa czasów trwania stanów, elementy modelu 2-gram, rozpoznawanie dwuetapowe oparte na podziale rozpoznanej sekwencji fonemów na pseudosylaby. Wykorzystywano zarówno algorytm uczenia Bauma-Welcha, jak i specjalnie zaprojektowane algorytmy bazujące na minimalizacji błędu rozpoznania. W rozpoznawaniu zastosowano zmodyfikowane algorytmy Viterbiego. Szczegółowy opis systemu zawiera dodatek C. Pomimo, że system zaprojektowany został do rozpoznawania izolowanych wyrazów, jego konstrukcja umożliwia rozszerzenie jego funkcjonalności do rozpoznawania komend wypowiadanych w sposób ciągły, jest to więc system dość uniwersalny.

2.2. Terminologia: kompensacja, adaptacja, normalizacja, odporna parametryzacja

Problem minimalizacji niekorzystnego wpływu zmienności cech osobniczych mówców i zniekształceń transmisyjnych w systemach ARM rozwiązywany może być na wiele sposobów. Konieczne jest ustalenie terminologii używanej do ich definiowania.

Kompensacja używana będzie jako ogólny termin określający minimalizację niekorzystnego wpływu zmienności cech sygnału mowy na skuteczność systemu ARM. *Adaptacja* oznaczać będzie kompensację polegającą na zmianie wartości parametrów modelu statystycznego przy nie zmienionych wartościach parametrów uzyskanych z sygnału. *Normalizacja* oznaczać będzie kompensację polegającą na zmianie wartości parametrów uzyskanych z sygnału przy nie zmienionych wartościach parametrów modelu. *Odporna (ang. robust) parametryzacja* oznaczać będzie natomiast metody, w których nie następuje transformacja wartości parametrów sygnału czy modelu, lecz sama parametryzacja jest skonstruowana tak, by minimalizować wpływ niekorzystnej zmienności sygnału. Na rys. 2.3 zilustrowano schemat działania przedstawionych wyżej metod.



Rys. 2.3. Rodzaje kompensacji zmienności sygnału mowy.

2.3. Wpływ zmiennych warunków transmisyjnych i cech osobniczych mówcy na widmo sygnału mowy

W części akustycznej toru transmisyjnego zniekształcenia sygnału spowodowane są charakterystyką pomieszczenia oraz zakłóceniami addytywnymi (dalej nazywanymi ogólnie szumem), na które składają się dźwięki inne niż sygnał mowy przeznaczony do rozpoznania. Charakterystyka pomieszczenia może być dokładnie opisana jego odpowiedzią impulsową. Niedogodnością tej formy opisu jest jej duża wrażliwość na zmiany położenia źródła i odbiornika sygnału. Z tego powodu w praktyce częściej stosuje się bardziej ogólny opis własności akustycznych pomieszczenia, poprzez podanie wartości charakteryzujących pogłos takich, jak czasy pojawienia się pierwszych odbić i tłumienie tych odbić, zależny od częstotliwości czas spadku natężenia pola akustycznego o 60 dB po zaniku pobudzenia (RT60), koloryzacja (uwypuklanie w pogłosie danych zakresów częstotliwości).

Charakterystyka kierunkowa mikrofonu oraz charakterystyka rozchodzenia się dźwięku wokół głowy mówcy powinny być również uwzględnione jako źródła zniekształceń, zależne od usytuowania mówcy i mikrofonu w pomieszczeniu. Szum addytywny w części akustycznej toru może mieć charakter stacjonarny lub niestacjonarny (w szczególności impulsowy) oraz może mieć różne widma. Typowe rodzaje szumu spotykane w zagadnieniu ARM to: szum od urządzeń docierający do pomieszczenia, szum uliczny, szum złożony z nałożonych na siebie głosów wielu osób (ang. *babble noise*), muzyka.

Mikrofon może być źródłem znacznych zniekształceń sygnału, często ten element toru ma największy udział we wprowadzonych zniekształceniach. Mikrofony dobrej klasy wnoszą praktycznie tylko zniekształcenia liniowe, natomiast popularne mikrofony niskiej klasy również zniekształcenia nieliniowe, nieraz bardzo znaczne. Spowodowane jest to głównie małym zakresem dynamicznym pracy mikrofonu i łatwością jego przesterowania. Urządzenia analogowego toru elektrycznego takie, jak wzmacniacze, filtry, rejestratory, złącza czy kable są źródłem kolejnych zniekształceń. Szczególnie poważne zniekształcenia wnosi łącze telefoniczne, w którym występują dodatkowe tony o niskich częstotliwościach, addytywny szum stacjonarny, szum impulsowy, filtracja liniowa o zmiennej w czasie charakterystyce amplitudowej i fazowej, zniekształcenia intermodulacyjne, echo. Kanał telefoniczny ponadto ogranicza pasmo sygnału do przedziału 300 Hz - 3.4 kHz [112]. Jeszcze inny charakter mają zniekształcenia wynikające z transmisji analogowej drogą radiową, np. wielodrogowość czy zaniki. Ostatnim elementem toru analogowego jest przetwornik analogowo-cyfrowy, który również może wnosić zniekształcenia liniowe i nieliniowe oraz szum, zwłaszcza szum kwantyzacji.

Sygnał cyfrowy przed wprowadzeniem do systemu ARM może być również narażony na specyficzne zniekształcenia powstałe podczas jego transmisji lub rejestra-

cji. Kompresja sygnału cyfrowego, stosowana w większości współczesnych systemów transmisyjnych, wnosi zniekształcenia o złożonym charakterze. Przykłady zmian w widmie samogłoski polskiej 'a' po transmisji przez siedem różnych kanałów telekomunikacyjnych (POTS, ISDN, G.721, G.723-24kb/s, G.723-40kb/s, GSM, LD-CELP) zawiera praca [139]. O ile kanał analogowy POTS i cyfrowy ISDN powodują właściwie tylko odfiltrowanie składowych widma powyżej 3.4 kHz i poniżej 300 Hz, to w pozostałych systemach widoczne jest wprowadzenie zakłóceń w postaci szumu, zniekształceń struktury harmonicznej sygnału oraz znaczne wytłumienie pasma wokół częstotliwości ok. 2.4 kHz. Częstotliwości niskie i wysokie są, w przeciwieństwie do POTS i ISDN, zachowane.

Model toru transmisyjnego dany równaniem (1.1) charakteryzuje się zadowalającą dokładnością, o ile poziom zniekształceń nieliniowych jest nieduży. Niewielkie zniekształcenia nieliniowe mogą być wtedy modelowane jako składnik szumu addytywnego. Szum addytywny często modeluje się jako biały szum gaussowski, lecz trzeba podkreślić, że w warunkach rzeczywistych rejestrowany szum rzadko ma widmo szumu białego oraz rozkład gaussowski. Najczęściej energia szumu skoncentrowana jest w zakresach niskich częstotliwości, przy czym występują tam na ogół składowe harmoniczne, pochodzące od elementów wirujących urządzeń znajdujących się w otoczeniu czy też przydźwięku sieci energetycznej. W przypadku występowania specyficznych zniekształceń związanych z kompresją sygnału mowy, modelu (1.1) na ogół się nie stosuje. Korzystniejsze jest wtedy zastosowanie specjalnie zaprojektowanych metod parametryzacji oraz struktury systemu [43].

Często używane są określenia „złe" i „dobre" w odniesieniu do warunków nagrań czy jakości sygnału. Uściślając te pojęcia można powiedzieć, że nagranie „dobre" cechuje się wartością SNR powyżej 30 dB oraz szumem stacjonarnym. Nie występują ponadto zniekształcenia nieliniowe, a charakterystyka zniekształceń liniowych jest stała w czasie i ma łagodny przebieg, tj. nie występuje głębokie tłumienie w wąskich pasmach. W nagraniu „złym" wartość SNR może być poniżej 30 dB, a szum mieć charakter niestacjonarny, w tym impulsowy. Mogą też wystąpić zniekształcenia zmienne w czasie i nieliniowe.

Jedną z przyczyn występowania różnic międzyosobniczych w widmie sygnału są różnice w budowie traktu głosowego. Najważniejszą cechą jego budowy jest długość. W przypadku, gdy trakt głosowy modelowany jest jako sztywna rura złożona z segmentów o różnej średnicy, częstotliwości rezonansowe tej rury, odpowiadające częstotliwościom formantów w mowie, zależą odwrotnie proporcjonalne od długości rury. Bardziej złożone modele uwzględniają oprócz długości także objętości poszczególnych fragmentów toru głosowego, a najdokładniejsze modelują tor głosowy techniką trójwymiarową. Okazuje się jednak ([175, 6]), że zmienności międzyosobniczej uwidaczniającej się w różnicach położenia formantów, nie można z zadowalającą dokładnością tłumaczyć jedynie różnicami w budowie toru głosowego. Równie

ważny jest indywidualny sposób artykulacji głosek przez danego mówcę. Elementem toru głosowego mającym również znaczny wpływ na zmienność międzyosobniczą są struny głosowe, od budowy których zależy w głównej mierze struktura czasowo-częstotliwościowa tonu krtaniowego.

Biorąc pod uwagę różnice w widmach jednostek fonetycznych dla różnych mówców, można ich klasyfikować na różne sposoby, celem wyodrębnienia grup, wewnątrz których różnice te są mniejsze niż w całej populacji. Kryteriami klasyfikacji mogą być: płeć, wiek, miejsce urodzenia/zamieszkania (różnice regionalne - dialekty, gwary), status społeczny, inny język ojczysty czy nawet stan zdrowia (głosy zmienione patologicznie, wady wymowy).

Oprócz międzyosobniczej zmienności widm jednostek fonetycznych, istotna jest również zmienność widm dla jednego mówcy, przy czym rozumie się tutaj zmienność krótkookresową, od wypowiedzi do wypowiedzi, a co najwyżej w okresie kilkunastu dni, np. związaną z nieprzewlekłym stanem chorobowym. Zmienność ta związana jest z różnymi stanami emocjonalnymi, stanami chorobowymi, wpływem otoczenia. Przykładem może być znany efekt Lombarda, polegający m.in. na zmianach częstotliwości formantów i obwiedni widma, wywołanych potrzebą efektywnej komunikacji w środowisku o znacznym poziomie szumu. Specyficzną różnicę wewnątrzosobniczą stanowi również szept. Należy podkreślić, że za różnice między- i wewnątrzosobnicze uważa się systematyczne zmiany w widmie. Dla każdego mówcy występują bowiem także losowe odchylenia tych parametrów związane z niemożliwością artykulacji w identycznie powtarzalny sposób.

Warto zaznaczyć, że różnice osobnicze uwidaczniają się nie tylko w zmianach widm chwilowych jednostek fonetycznych, ważne są również różnice w dynamice zmian widma, związanej z ruchami narządu mowy. Charakterystyka tych ruchów jest cechą indywidualną mówcy. Podstawowym parametrem opisującym dynamikę zmian widma sygnału mowy jest tempo mówienia, wyrażone liczbą pewnych jednostek językowych artykułowanych w jednostce czasu.

W zastosowaniach praktycznych, ze względu na trudność dokładnej identyfikacji i analizy przyczyn zniekształceń transmisyjnych oraz różnic osobniczych, algorytmy dokonują kompensacji skutków, tj. zaistniałych zmian w widmie bądź też zmian parametrów wyznaczonych z tego widma. Oczywiście, można zaproponować również systemy, które wykorzystywałyby pewne dodatkowe informacje o zniekształceniach transmisyjnych czy mówcy, np. o jego płci czy wieku, uzyskane ze źródeł innych niż sygnał mowy, co pozwoliłoby wykorzystać w kompensacji zmian również informację o ich przyczynach.

2.4. Wpływ zmienności sygnału mowy na parametry MFCC oraz skuteczność systemu ARM

Zmienność widma jednostek fonetycznych sygnału mowy, spowodowana zniekształceniami transmisyjnymi i różnicami osobniczymi, skutkuje zmiennością współczynników MFCC. W tej pracy w parametryzacji zastosowano krótkoczasową analizę widma o długości ramki analizy równej 20 ms. Założono, że sygnały i transmitancje są stacjonarne w przedziałach czasu odpowiadających jednej ramce. Zniekształcenia liniowe mogą być modelowane układem opisanym odpowiedzią impulsową. Jeśli odpowiedź ta jest krótsza od długości ramki, to wpływ zniekształceń liniowych na widmo sygnału z zadowalającą dokładnością rozpatrywać można w obrębie jednej ramki. Jeśli natomiast odpowiedź impulsowa jest porównywalna lub dłuższa niż długość ramki, skutkuje to powstaniem zniekształceń międzyramkowych. Są one do pominięcia tylko w przypadku, gdy analizowany jest fragment sygnału stacjonarnego o długości wielu ramek. Sygnał mowy zawiera jednak fragmenty charakteryzujące się szybkimi (o czasie poniżej 20 ms) zmianami w widmie, wtedy zniekształcenia międzyramkowe mogą być znaczne. Większość algorytmów kompensacji współpracujących z systemami opartymi na HMM nie uwzględnia zniekształceń międzyramkowych. Uwzględnienie ich skutkuje znacznym skomplikowaniem zarówno algorytmu kompensacji, jak i modelu statystycznego, co jest nieopłacalne biorąc pod uwagę fakt, iż silne zniekształcenia międzyramkowe są zjawiskiem rzadkim i występują w specyficznych warunkach. Jeśli jest to konieczne, to w takich warunkach stosuje się przetwarzanie wstępne, np. algorytmy usuwania echa linii transmisyjnej oraz echa akustycznego czy algorytmy usuwania pogłosu.

Model zniekształceń transmisyjnych dany równaniem (1.1) uwzględnia zniekształcenia liniowe i szum addytywny. Trzeba tutaj zaznaczyć, że z punktu widzenia algorytmów kompensacji te dwie składowe stanowią pewnego rodzaju stopnie swobody kompensacji i niekoniecznie modelują rzeczywiste przyczyny zniekształceń. Składowa zniekształceń liniowych modeluje liniowe zniekształcenia wewnątrzramkowe, natomiast składowa szumu obejmuje zniekształcenia pozostałe, w tym oprócz szumu addytywnego również efekty zniekształceń nieliniowych i międzyramkowych. Np. w przypadku wystąpienia długiego pogłosu odbicia wczesne, skutkujące filtracją typu grzebieniowego, modelowane są przez składową zniekształceń liniowych, odbicia późne i silnie rozproszone, przenoszące się na kolejne ramki, są traktowane jako szum addytywny.

Poniżej przedstawiono matematyczny opis wpływu zniekształceń transmisyjnych na współczynniki MFCC przy zastosowaniu modelu (1.1) i przy założeniu braku zniekształceń międzyramkowych. Niech $\mathbf{x}$ oznacza dyskretne widmo zespolone ramki sygnału niezniekształconego, $\mathbf{n}$ - dyskretne widmo zespolone szumu addytywnego, a $\mathbf{h}^{(zn)}$ - zespoloną charakterystykę widmową zniekształceń liniowych. Pominięto

wpływ preemfazy, gdyż wnosi ona jedynie systematyczne zmiany w widmie. Pominięto również wpływ okienkowania, zakładając, że związane z nim przeciek oraz wygładzanie widma mają charakter ilościowy, a nie jakościowy. Widmo amplitudowe $\mathbf{s}^{(zn)}$ sygnału zniekształconego jest następujące:

$$\mathbf{s}^{(zn)} = \left|\mathbf{x} \circ \mathbf{h}^{(zn)} + \mathbf{n}\right| \tag{2.15}$$

$$s_k^{(zn)} = s_k \cdot |h_k^{(zn)}| \cdot \sqrt{1 + \left|\frac{n_k}{x_k \cdot h_k^{(zn)}}\right|^2 + 2\left|\frac{n_k}{x_k \cdot h_k^{(zn)}}\right| \cos\left(\arg\left(x_k \cdot h_k^{(zn)}\right) - \arg n_k\right)} \tag{2.16}$$

Operator $\circ$ oznacza mnożenie wektorów przeprowadzane element po elemencie.

Po przeprowadzeniu uśredniania widma w banku filtrów melowych zależności przyjmą postać:

$$s_j^{(m,zn)} = \sum_{k=0}^{K/2-1} \left(h_{k,j}^{(mel)} \cdot s_k \cdot \left(h_j^{(m,zn)} \cdot \sqrt{1 + \left(\frac{n_j^{(m)}}{x_j^{(m)} \cdot h_j^{(m,zn)}}\right)^2 + \epsilon_{j,k}^{(1)}} \right) \right) \tag{2.17}$$

$$s_j^{(m,zn)} = s^{(m)} \cdot h_j^{(m,zn)} \cdot \sqrt{1 + \left(\frac{n_j^{(m)}}{x_j^{(m)} \cdot h_j^{(m,zn)}}\right)^2 + \epsilon_j^{(2)}} \tag{2.18}$$

gdzie dla każdego filtru melowego wyznaczono średnie wartości widma amplitudowego sygnału, szumu i charakterystyki zniekształceń:

$$h_j^{(m,zn)} = \left(\sum_{k=0}^{K/2-1} u\left(h_{k,j}^{(mel)}\right) \cdot |h_k^{(zn)}| \right) / \left(\sum_{k=0}^{K/2-1} u\left(h_{k,j}^{(mel)}\right) \right) \tag{2.19}$$

$$n_j^{(m)} = \left(\sum_{k=0}^{K/2-1} u\left(h_{k,j}^{(mel)}\right) \cdot |n_k| \right) / \left(\sum_{k=0}^{K/2-1} u\left(h_{k,j}^{(mel)}\right) \right) \tag{2.20}$$

$$x_j^{(m)} = \left(\sum_{k=0}^{K/2-1} u\left(h_{k,j}^{(mel)}\right) \cdot |x_k| \right) / \left(\sum_{k=0}^{K/2-1} u\left(h_{k,j}^{(mel)}\right) \right) \tag{2.21}$$

W powyższych równaniach u oznacza funkcję skoku jednostkowego, a K - długość stosowanej transformaty DFT.

Założono, że szum jest niezależny od sygnału, zatem teoretycznie wartość oczekiwana wyrażenia $\cos\left(\arg\left(x_k \cdot h_k^{(zn)}\right) - \arg n_k\right)$ w równaniu (2.16) przyjmuje wartość zero. W rzeczywistej analizie jest ona jednak niezerowa z uwagi na niezerową wariancję estymatora widma. Ten błąd estymacji jest jednym ze składników błędu $\epsilon_{j,k}^{(1)}$, drugi składnik ma natomiast źródło w różnicach widm uśrednianych w filtrach

melowych w stosunku do ich wartości średnich wyznaczonych w filtrach (równania 2.19, 2.20, 2.21). Błąd $\epsilon_j^{(2)}$ jest skumulowanym błędem $\epsilon_{j,k}^{(1)}$ wyznaczonym dla każdego filtru melowego. W przypadku braku szumu błąd $\epsilon_j^{(2)}$ wynosi zero, gdy wewnątrz filtru charakterystyka $|\mathbf{h}^{(zn)}|$ jest stała. W przypadku występowania szumu błąd ten jest tym mniejszy, im widma sygnału, szumu i charakterystyka $|\mathbf{h}^{(zn)}|$ są bardziej stałe wewnątrz danego filtru oraz błąd związany z niezerową wartością $\cos\left(\arg\left(x_k \cdot h_k^{(zn)}\right) - \arg n_k\right)$ jest mniejszy. Następnie wartości wyjściowe z banku filtrów melowych są logarytmowane:

$$h_j^{(l,zn)} = \ln h_j^{(m,zn)} \tag{2.22}$$

$$n_j^{(l)} = \ln n_j^{(m)} \tag{2.23}$$

$$x_j^{(l)} = \ln x_j^{(m)} \tag{2.24}$$

$$s_j^{(l,zn)} = \ln s_j^{(m,zn)} = s_j^{(l)} + h_j^{(l,zn)} + \frac{1}{2}\ln\left(1 + e^{2\left(n_j^{(l)} - x_j^{(l)} - h_j^{(l,zn)}\right)}\right) + \epsilon_j^{(3)} \tag{2.25}$$

Błąd $\epsilon_j^{(3)}$ jest przekształconym wskutek zastosowania nieliniowej operacji logarytmowania błędem $\epsilon_j^{(2)}$. W przypadku, gdy $\epsilon_j^{(2)}$ jest równe zeru, $\epsilon_j^{(3)}$ również przyjmuje wartość zero. Ostatnim krokiem parametryzacji MFCC jest zastosowanie RDCT, która w odróżnieniu od DCT nie jest całkowicie odwracalna.

$$\mathbf{o}^{(h,zn)} = \mathrm{RDCT}\left(\mathbf{h}^{(l,zn)}\right) \tag{2.26}$$

$$\mathbf{o}^{(n)} = \mathrm{RDCT}\left(\mathbf{n}^{(l)}\right) \tag{2.27}$$

$$\mathbf{o}^{(x)} = \mathrm{RDCT}\left(\mathbf{x}^{(l)}\right) \tag{2.28}$$

$$\begin{aligned}\mathbf{o}^{(zn)} &= \mathbf{o} + \mathbf{o}^{(h,zn)} + \\ &+ \mathrm{RDCT}\left(\frac{1}{2}\ln\left(1 + e^{2\left(\mathrm{IRDCT}\left(\mathbf{o}^{(n)}\right) - \mathrm{IRDCT}\left(\mathbf{o}^{(x)}\right) - \mathrm{IRDCT}\left(\mathbf{o}^{(h,zn)}\right)\right)}\right)\right) + \boldsymbol{\epsilon}^{(4)}\end{aligned} \tag{2.29}$$

Błąd $\boldsymbol{\epsilon}^{(4)}$ zawiera w sobie przekształcony za pomocą RDCT błąd $\boldsymbol{\epsilon}^{(3)}$ oraz błąd przybliżenia związany z zastosowaniem transformacji IRDCT.

Jak widać, analiza wpływu zniekształceń transmisyjnych na współczynniki MFCC jest skomplikowana. Nawet przy przyjęciu znaczących uproszczeń, zależność (2.29) jest złożona i ponadto nieliniowa. Analiza statystyczna błędu $\boldsymbol{\epsilon}^{(4)}$, nawet przy przyjęciu prostych modeli statystycznych opisujących sygnał i zniekształcenia, jest analitycznie trudna z uwagi na występujące nieliniowości. W pracach [125, 28] można znaleźć próby zastosowania takiego modelowania w celu poprawy skuteczności kompensacji wpływu zniekształceń transmisyjnych.

We wstępnych badaniach przeprowadzonych na potrzeby niniejszej pracy sprawdzono, że błąd $\boldsymbol{\epsilon}^{(4)}$ jest znaczący i ponadto może mieć niezerową wartość oczekiwaną.

Skutkuje to zmniejszoną skutecznością metod kompensacji bazujących na korekcji wartości współczynników MFCC przeprowadzanej przy wykorzystaniu zależności (2.29) na podstawie oszacowanych wcześniej wartości parametrów opisujących zniekształcenia. W przypadku braku szumu addytywnego z równania (2.29) znika człon nieliniowy, a wpływ zniekształceń liniowych objawia się dodaniem do wektora $\mathbf{o}$ odpowiedniego wektora $\mathbf{o}^{(h,zn)}$, związanego z charakterystyką widmową zniekształceń. Błąd $\boldsymbol{\epsilon}^{(4)}$ w tym przypadku zależy od charakterystyki amplitudowej zniekształceń i jest tym mniejszy, im charakterystyka ta jest bardziej stała wewnątrz filtrów melowych.

Addytywność sygnału i zniekształceń liniowych w dziedzinie cepstrum jest własnością często wykorzystywaną w celu normalizacji wartości parametrów sygnału poprzez odjęcie od wektorów MFCC wektora uśrednionego z całej wypowiedzi. Metoda ta ma jednak dwie zasadnicze wady: wymaga uśredniania parametrów z długiej, co najmniej kilkusekundowej wypowiedzi oraz, co zauważono w pracy [81], jest mało skuteczna w przypadku zniekształceń o mało łagodnej charakterystyce amplitudowej, np. spowodowanej przez pogłos. W takiej sytuacji pojawia się znaczny błąd związany z uśrednianiem w filtrach melowych, wynikający z dużej zmienności widma wewnątrz tych filtrów. Mało stała charakterystyka zniekształceń uwypuklana jest w wąskich pasmach wokół częstotliwości harmonicznych tonu krtaniowego. Na skutek zmian w czasie częstotliwości tego tonu, zmianom ulegają również położenia uwypuklanych pasm, co jest główną przyczyną występowania wspomnianej zmienności widma.

Analiza wpływu cech osobniczych mówcy na współczynniki MFCC jest znacznie bardziej skomplikowana niż analiza wpływu zniekształceń transmisyjnych. Zniekształcenia transmisyjne wpływają jednakowo na cały sygnał mowy, wpływ różnic osobniczych opisywać trzeba natomiast osobno dla poszczególnych jednostek (fonemów, allofonów czy nawet segmentów allofonów). Charakter zmian w widmie jest również bardziej złożony. Model uwzględniający zniekształcenia liniowe i szum addytywny stosować można w pewnym zakresie do zmian w widmie pobudzenia krtaniowego. Jego obwiednia jest bowiem zmienna, a w pobudzeniu dźwięcznym obecność szumu może wiązać się ze stanami chorobowymi lub wiekiem mówcy. Osobnicze zmiany w transmitancji toru głosowego objawiają się jako przesunięcia częstotliwości, zmiany poziomu oraz szerokości pasm formantów. Szczególnie dwa ostatnie czynniki mogą być znaczne. W przypadku kompensacji tych różnic osobno dla poszczególnych jednostek fonetycznych, wystarczające jest zastosowanie odpowiednich filtrów liniowych. Jednak w większości algorytmów kompensacji, zwłaszcza normalizacji, z uwagi na ograniczoną ilość danych, konieczne jest korygowanie zmian dla grup jednostek fonetycznych lub wszystkich jednostek równocześnie. W takim przypadku kompensacja przesunięć częstotliwości formantów dokonywana jest poprzez skalowanie osi częstotliwości, ponieważ dla grup jednostek istnieją pewne systema-

tyczne przesunięcia tych częstotliwości, związane np. z długością toru głosowego. Wpływ takiego skalowania na wartości współczynników MFCC został przeanalizowany w pracy [128]. Pokazano, że w przypadku analizy cepstralnej z melową skalą częstotliwości przy braku uśredniania w banku filtrów, skalowanie osi częstotliwości odpowiada liniowej transformacji przestrzeni parametrów. Uśrednianie w bankach filtrów wprowadza jednak znaczny błąd, co zostało sprawdzone we wstępnym etapie opisywanych badań, tak, że modelowanie za pomocą transformacji liniowej jest nieefektywne.

Różnice w dynamice zmian widma sygnału mowy mają niewielki wpływ na zmienność statycznych współczynników MFCC. Parameryzacja sygnału obejmuje jednak często również współczynniki dynamiczne, na które różnice te mają znaczny wpływ. Łatwo pokazać, że zmiana tempa mówienia, polegająca na a-krotnym jego przyspieszeniu, skutkuje a-krotnym wzrostem wartości pochodnych czasowych współczynników MFCC. W przypadku stosowania współczynników dynamicznych nie bazujących na pochodnych czasowych, opis analityczny wpływu zmian tempa mówienia jest bardzo utrudniony lub niemożliwy.

Zmiany wartości współczynników MFCC spowodowane zniekształceniami transmisyjnymi oraz różnicami osobniczymi powodują spadek skuteczności systemu ARM. Model statystyczny Θ systemu zawiera rozkłady prawdopodobieństwa współczynników MFCC $p_i(\mathbf{o})$ dla każdego stanu i modelu. System można uznać za pewien złożony klasyfikator. Wpływ zmian wartości współczynników MFCC na błąd klasyfikacji prześledzić można analizując klasyfikator uproszczony. Załóżmy, że dany jest prosty klasyfikator Bayesowski, przyporządkowujący skalarną wartość wejściową x do dwóch klas. Rozkład prawdopodobieństwa wartości x dla klasy pierwszej opisany jest rozkładem normalnym $p_1^{(k)} = \mathscr{N}(x; \mu_1, \sigma_1^2)$, a dla klasy drugiej rozkładem $p_2^{(k)} = \mathscr{N}(x; \mu_2, \sigma_2^2)$. Załóżmy, że klasy są równoprawdopodobne. Kryterium klasyfikacji, zapewniające minimalny średni błąd, jest następujące [39]:

$$\frac{p_1^{(k)}(x)}{p_2^{(k)}(x)} \underset{\text{klasa 2}}{\overset{\text{klasa 1}}{\gtrless}} 1 \tag{2.30}$$

Zatem przedziały decyzyjne dla x wyznaczane są przez punkty przecięcia funkcji gęstości prawdopodobieństwa, wyznaczone z równania:

$$p_1^{(k)}(x) = p_2^{(k)}(x) \tag{2.31}$$

$$\frac{1}{\sqrt{2\pi}\sigma_1} \cdot e^{-\frac{1}{2}\cdot\left(\frac{x-\mu_1}{\sigma_1}\right)^2} = \frac{1}{\sqrt{2\pi}\sigma_2} \cdot e^{-\frac{1}{2}\cdot\left(\frac{x-\mu_2}{\sigma_2}\right)^2} \tag{2.32}$$

Można pokazać, że w przypadku $\sigma_1 = \sigma_2$ istnieje jedno rozwiązanie równania (2.32):

$$x_1 = \frac{\mu_1 + \mu_2}{2}, \tag{2.33}$$

a w przypadku $\sigma_1 \neq \sigma_2$ istnieją dwa rozwiązania:

$$x_1 = \frac{\mu_1 + \mu_2}{2} + \frac{\left(\sigma_1^2 + \sigma_2^2\right)\left(\mu_2 - \mu_1\right) + 2\sigma_1\sigma_2\sqrt{\left(\mu_2 - \mu_1\right)^2 + 2\ln\left(\sigma_1/\sigma_2\right)\cdot\left(\sigma_1^2 - \sigma_2^2\right)}}{2\left(\sigma_1^2 - \sigma_2^2\right)} \tag{2.34}$$

$$x_2 = \frac{\mu_1 + \mu_2}{2} + \frac{\left(\sigma_1^2 + \sigma_2^2\right)\left(\mu_2 - \mu_1\right) - 2\sigma_1\sigma_2\sqrt{\left(\mu_2 - \mu_1\right)^2 + 2\ln\left(\sigma_1/\sigma_2\right)\cdot\left(\sigma_1^2 - \sigma_2^2\right)}}{2\left(\sigma_1^2 - \sigma_2^2\right)} \tag{2.35}$$

Załóżmy, że klasy są dobrze rozseparowane, tj. $\delta = \left(\mu_2 - \mu_1\right)^2 / \left(\sigma_1^2 + \sigma_2^2\right) > 0.5$ oraz wariancje rozkładów są zbliżone, tj. $\sigma_2^2 = \Delta \cdot \sigma_1^2$, gdzie $1 < \Delta < 2$. Bez utraty ogólności rozważań załóżmy też, że $\mu_1 < \mu_2$. Powyższe równania można zapisać jako:

$$x_1 = \frac{\mu_1 + \mu_2}{2} + \frac{\mu_2 - \mu_1}{2} \cdot \frac{1 + \Delta + 2\sqrt{\Delta}\cdot\sqrt{1 + \ln\Delta\cdot(\Delta - 1)\,\frac{\sigma_1^2}{\left(\mu_2 - \mu_1\right)^2}}}{1 - \Delta} \tag{2.36}$$

$$x_2 = \frac{\mu_1 + \mu_2}{2} + \frac{\mu_2 - \mu_1}{2} \cdot \frac{1 + \Delta - 2\sqrt{\Delta}\cdot\sqrt{1 + \ln\Delta\cdot(\Delta - 1)\,\frac{\sigma_1^2}{\left(\mu_2 - \mu_1\right)^2}}}{1 - \Delta} \tag{2.37}$$

W przypadku, gdy $\delta = 0.5$ i $\Delta = 2$ równania przyjmą postać:

$$x_1 = \frac{\mu_1 + \mu_2}{2} - 3.21 \cdot \left(\mu_2 - \mu_1\right) \tag{2.38}$$

$$x_2 = \frac{\mu_1 + \mu_2}{2} + 0.21 \cdot \left(\mu_2 - \mu_1\right) \tag{2.39}$$

Można pokazać, że dla $\Delta \to 1^+$ wartości x_1 i x_2 dążą monotonicznie do:

$$\lim_{\Delta \to 1^+} x_1 = -\infty \tag{2.40}$$

$$\lim_{\Delta \to 1^+} x_2 = \frac{\mu_1 + \mu_2}{2} \tag{2.41}$$

Można pokazać również, że dla $\delta \to \infty$ wartości x_1 i x_2 dążą monotonicznie do:

$$\lim_{\delta \to \infty} x_1 = \frac{\mu_1 + \mu_2}{2} + \frac{\mu_2 - \mu_1}{2} \cdot \frac{1}{1 - \sqrt{\Delta}} \tag{2.42}$$

$$\lim_{\delta \to \infty} x_2 = \frac{\mu_1 + \mu_2}{2} + \frac{\mu_2 - \mu_1}{2} \cdot \frac{1}{1 + \sqrt{\Delta}} \tag{2.43}$$

Widać zatem, że próg decyzyjny x_1 można zaniedbać, gdyż wartości funkcji gęstości prawdopodobieństwa, a co za tym idzie błędy klasyfikacji, są dla wartości $x < x_1$ bardzo małe.

Jeśli rozkłady prawdopodobieństwa wartości x zostaną zmienione, tj. zmianie ulegną wartości parametrów $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2$, to zmieni się również położenie progu decyzyjnego. Klasyfikacja przy użyciu poprzedniej wartości progu będzie obarczona większym błędem. Analizując równanie (2.31) można zauważyć, że deterministyczne i odwracalne zniekształcenie liniowe lub nieliniowe nałożone na x skutkuje również przekształceniem wartości progu decyzyjnego za pomocą takiego samego zniekształcenia. Zniekształcenia losowe w postaci szumu addytywnego o wartości oczekiwanej równej zeru, choć nie zmieniają wartości oczekiwanej rozkładów, wpływają jednak na wartość progu decyzyjnego. Wariancja zmiennej x, zakłóconej niezależnym od x addytywnym szumem o wariancji $\sigma_3^2 = \Delta^{(n)} \cdot \sigma_1^2$, rośnie o wartość wariancji szumu. Zatem próg decyzyjny przyjmuje wartość:

$$x_2 = \frac{\mu_1 + \mu_2}{2} + \frac{\mu_2 - \mu_1}{2} \cdot \\ \cdot \frac{1 + \Delta + 2\Delta^{(n)} - 2\sqrt{\left(1 + \Delta^{(n)}\right)\left(\Delta + \Delta^{(n)}\right)} \cdot \sqrt{1 + \ln\left(\frac{\Delta + \Delta^{(n)}}{1 + \Delta^{(n)}}\right) \cdot (\Delta - 1)\,\frac{\sigma_1^2}{(\mu_2 - \mu_1)^2}}}{1 - \Delta} \tag{2.44}$$

Zniekształcenia deterministyczne najskuteczniej można skompensować stosując, o ile istnieje, transformację do nich odwrotną. W praktyce jest to na ogół niemożliwe z uwagi na nieznajomość dokładnej postaci tych zniekształceń. Algorytmy kompensacji wykorzystują więc mechanizmy statystyczne. Najczęściej w sposób iteracyjny estymują parametry kompensacji tak, aby maksymalizować mierzone funkcją wiarogodności dopasowanie parametrów sygnału do modelu statystycznego. Konieczność zastosowania iteracji wynika z naprzemiennie wykonywanych rozpoznania i estymacji parametrów kompensacji (algorytm E-M). Najczęściej też wykorzystywane są jedynie statystyki pierwszego rzędu. W takim przypadku w zaprezentowanym prostym klasyfikatorze możliwa jest jedynie dokładna kompensacja addytywnych zniekształceń liniowych. Szum addytywny o zerowej wartości oczekiwanej i pewne transformacje nieliniowe nie zmieniają wartości oczekiwanej rozkładów, więc kompensacja nie jest przeprowadzana, choć jak zostało to wykazane, wartość progu decyzyjnego ulega zmianie. W celu lepszej kompensacji konieczne jest zatem uwzględnienie statystyk wyższych rzędów, lecz wiąże się to ze znacznym zwiększeniem złożoności obliczeniowej. Rozwiązaniem kompromisowym jest modelowanie rozkładów metodą GMM i kompensacja wartości oczekiwanej dla każdej składowej GMM niezależnie. Lepsze rezultaty kompensacji daje również zastosowanie w wyznaczaniu jej parametrów kryteriów wprost minimalizujących błąd klasyfikacji.

W systemie ARM klasyfikacja jest bardziej złożona, nie występują progi decyzyjne, lecz uwzględniane są wartości funkcji gęstości prawdopodobieństwa. Rozkłady są ponadto wielowymiarowe, a liczba klas wynosi co najmniej kilkadziesiąt. Model statystyczny zawiera też parametry takie, jak prawdopodobieństwa przejść między stanami czy dodatkowe rozkłady prawdopodobieństwa czasów trwania stanów. Ze względu na złożoność obliczeniową, wartości tych parametrów rzadko są modyfikowane podczas kompensacji, choć zmiany cech dynamicznych mowy, takich jak tempo mówienia, sugerują celowość uwzględniania ich w kompensacji.

Adaptacja systemu polega na modyfikacji wartości parametrów modelu statystycznego, a więc najczęściej parametrów opisujących rozkłady prawdopodobieństwa MFCC. Jest to metoda silnie ukierunkowana na minimalizację skutków zniekształceń, a jej parametry najczęściej nie mają jasnego sensu fizycznego. Korzystną własnością tych algorytmów jest natomiast duża elastyczność, gdyż o ile normalizacji poddawane są parametry całego sygnału, tak adaptacja może działać selektywnie w stosunku do parametrów opisujących poszczególne stany modelu. Selektywność taka wymaga jednak dużej ilości danych adaptacyjnych.

Normalizacja może być przeprowadzana na różnych etapach parametryzacji MFCC. Najczęściej modyfikowane są: widmo amplitudowe, wartości wyjściowe z banku filtrów melowych przed lub po logarytmowaniu, współczynniki MFCC. Korzystna wydaje się tu modyfikacja widma amplitudowego, gdyż parametry normalizacji mogą mieć jasny sens fizyczny oraz zredukowany jest błąd wprowadzany w dalszych etapach parametryzacji. Normalizacja w dziedzinie widma amplitudowego ma jednak tę wadę, że bez wprowadzania dodatkowych modeli statystycznych, uzyskanie dokładnych analitycznych rozwiązań zadania maksymalizacji funkcji wiarogodności czy też miary oceny rozpoznania jest niemożliwe. Rozwiązania takie są natomiast możliwe do uzyskania w przypadku kompensacji w dziedzinie współczynników MFCC.

2.5. Przegląd znanych rozwiązań

Poniżej przedstawiono kilka dobrze znanych z literatury przedmiotu rozwiązań zagadnienia kompensacji zniekształceń transmisyjnych i cech osobniczych mówcy.

2.5.1. Vector Taylor Series (VTS) - aproksymacja funkcji zniekształceń za pomocą szeregu Taylora

W algorytmie VTS przyjęto model kanału transmisyjnego dany równaniem (1.1). Normalizacja zniekształceń transmisyjnych odbywa się w dziedzinie zlogarytmowanej widmowej gęstości mocy sygnału. Wpływ zniekształceń na widmo jest w tym przypadku następujący:

$$\mathbf{s}^{(lwgm,zn)} = \mathbf{s}^{(lwgm)} + \mathbf{h}^{(lwgm,zn)} + \ln\left(1 + e^{\mathbf{n}^{(lwgm)} - \mathbf{s}^{(lwgm)} - \mathbf{h}^{(lwgm,zn)}}\right) + \boldsymbol{\epsilon}^{(5)} \tag{2.45}$$

$$\mathbf{s}^{(lwgm,zn)} = \mathbf{s}^{(lwgm)} + r\left(\mathbf{s}^{(lwgm)}, \mathbf{h}^{(lwgm,zn)}, \mathbf{n}^{(lwgm)}\right) + \boldsymbol{\epsilon}^{(5)} \tag{2.46}$$

gdzie $\mathbf{s}^{(lwgm)}$, $\mathbf{s}^{(lwgm,zn)}$ i $\mathbf{n}^{(lwgm)}$ oznaczają zlogarytmowaną widmową gęstość mocy odpowiednio: sygnału niezniekształconego, zniekształconego i szumu. $\mathbf{h}^{(lwgm,zn)}$ oznacza zlogarytmowany kwadrat modułu charakterystyki amplitudowej zniekształceń liniowych. Na błąd $\boldsymbol{\epsilon}^{(5)}$ składają się czynniki analogiczne, co w przypadku błędu $\boldsymbol{\epsilon}^{(3)}$ w równaniu (2.25). Rozkłady gęstości prawdopodobieństwa wektorów $\mathbf{s}^{(lwgm)}$ i $\mathbf{s}^{(lwgm,zn)}$ modelowane są metodą GMM i opisane parametrami: $\boldsymbol{\mu}_k^{(s)}$, $\boldsymbol{\Sigma}_k^{(s)}$, $\boldsymbol{\mu}_k^{(s,zn)}$, $\boldsymbol{\Sigma}_k^{(s,zn)}$, gdzie k oznacza numer składowej GMM. Rozkłady wektorów $\mathbf{n}^{(lwgm)}$ i $\mathbf{h}^{(lwgm,zn)}$ modelowane są pojedynczymi rozkładami normalnymi o parametrach $\boldsymbol{\mu}^{(h,zn)}$, $\boldsymbol{\Sigma}^{(h,zn)}$, $\boldsymbol{\mu}^{(n)}$, $\boldsymbol{\Sigma}^{(n)}$. Długości wektorów zależą od przyjętej liczby pasm uśredniania widmowej gęstości mocy. W celu uzyskania analitycznej postaci zależności między parametrami rozkładów sygnału niezniekształconego i zniekształconego, zaniedbuje się błąd $\boldsymbol{\epsilon}^{(5)}$ oraz rozwija nieliniową funkcję r w szereg Taylora:

$$\begin{aligned}
\mathbf{s}^{(lwgm,zn)} \approx\ & \mathbf{s}^{(lwgm)} + r\left(\mathbf{s}_0^{(lwgm)}, \mathbf{n}_0^{(lwgm)}, \mathbf{h}_0^{(lwgm,zn)}\right) + \\
& + \frac{d}{d\mathbf{s}^{(lwgm)}} r\left(\mathbf{s}_0^{(lwgm)}, \mathbf{n}_0^{(lwgm)}, \mathbf{h}_0^{(lwgm,zn)}\right)\left(\mathbf{s}^{(lwgm)} - \mathbf{s}_0^{(lwgm)}\right) + \\
& + \frac{d}{d\mathbf{n}^{(lwgm)}} r\left(\mathbf{s}_0^{(lwgm)}, \mathbf{n}_0^{(lwgm)}, \mathbf{h}_0^{(lwgm,zn)}\right)\left(\mathbf{n}^{(lwgm)} - \mathbf{n}_0^{(lwgm)}\right) + \\
& + \frac{d}{d\mathbf{h}^{(lwgm,zn)}} r\left(\mathbf{s}_0^{(lwgm)}, \mathbf{n}_0^{(lwgm)}, \mathbf{h}_0^{(lwgm,zn)}\right)\left(\mathbf{h}^{(lwgm,zn)} - \mathbf{h}_0^{(lwgm,zn)}\right)
\end{aligned} \tag{2.47}$$

Za punkty, wokół których rozwijana jest funkcja r, przyjmowane są wartości oczekiwane rozkładów normalnych modelu GMM. Można pokazać, że przy zerowym rzędzie rowinięcia parametry rozkładu sygnału zniekształconego dane są wzorami:

$$\boldsymbol{\mu}_k^{(s,zn)} = \boldsymbol{\mu}_k^{(s)} + r\left(\boldsymbol{\mu}_k^{(s)}, \boldsymbol{\mu}^{(n)}, \boldsymbol{\mu}^{(h,zn)}\right) \tag{2.48}$$

$$\boldsymbol{\Sigma}_k^{(s,zn)} = \boldsymbol{\Sigma}_k^{(s)} \tag{2.49}$$

Przy rozwinięciu rzędu pierwszego wartości oczekiwane dane są równaniem (2.48), natomiast macierze kowariancji są następujące:

$$
\begin{aligned}
\boldsymbol{\Sigma}_k^{(s,zn)} &= \left(\mathbf{I} + \frac{d}{d\mathbf{s}^{(lwgm)}} r^{(1)}\right) \boldsymbol{\Sigma}_k^{(s)} \left(\mathbf{I} + \frac{d}{d\mathbf{s}^{(lwgm)}} r^{(1)}\right)^T + \\
&+ \left(\frac{d}{d\mathbf{n}^{(lwgm)}} r^{(1)}\right) \boldsymbol{\Sigma}^{(n)} \left(\frac{d}{d\mathbf{n}^{(lwgm)}} r^{(1)}\right)^T + \\
&+ \left(\frac{d}{d\mathbf{h}^{(lwgm,zn)}} r^{(1)}\right) \boldsymbol{\Sigma}^{(h,zn)} \left(\frac{d}{d\mathbf{h}^{(lwgm,zn)}} r^{(1)}\right)^T \\
r^{(1)} &= r\left(\boldsymbol{\mu}_k^{(s)}, \boldsymbol{\mu}^{(n)}, \boldsymbol{\mu}^{(h,zn)}\right)
\end{aligned}
\tag{2.50}
$$

gdzie $\mathbf{I}$ oznacza macierz jednostkową. Wartości parametrów rozkładów sygnału niezniekształconego wyznaczane są na podstawie danych z części uczącej bazy nagrań. Wartości parametrów opisujących zniekształcenia liniowe i szum są estymowane w trakcie normalizacji, zgodnie z poniższym algorytmem:

1. Przyjmij początkowe wartości $\boldsymbol{\mu}^{(h,zn)}$, $\boldsymbol{\Sigma}^{(h,zn)}$, $\boldsymbol{\mu}^{(n)}$ i $\boldsymbol{\Sigma}^{(n)}$.
2. Rozwiń funkcję r wokół wartości oczekiwanej każdej składowej GMM $\boldsymbol{\mu}_k^{(s)}$ oraz wokół $\boldsymbol{\mu}^{(h,zn)}$ i $\boldsymbol{\mu}^{(n)}$.
3. Wykorzystując zależności (2.48) oraz (2.49) lub (2.50) przeprowadź jedną iterację algorytmu E-M w celu adaptacji parametrów $\boldsymbol{\mu}^{(h,zn)}$, $\boldsymbol{\Sigma}^{(h,zn)}$, $\boldsymbol{\mu}^{(n)}$ i $\boldsymbol{\Sigma}^{(n)}$.
4. Jeśli nie została osiągnięta zbieżność wartości parametrów (zmiana wartości parametrów względem wartości z poprzedniej iteracji jest większa od zadanego progu), wróć do punktu 2.

Normalizacja zlogarytmowanej widmowej gęstości mocy sygnału zniekształconego dokonywana jest z wykorzystaniem estymatora minimalizującego błąd średniokwadratowy. Estymator ten dany jest poniższym równaniem:

$$
\mathbf{s}^{(lwgm,norm)} = \mathbf{s}^{(lwgm,zn)} - \sum_{k=0}^{K-1} P\left(k|\mathbf{s}^{(lwgm,zn)}\right) \cdot r\left(\boldsymbol{\mu}_k^{(s)}, \boldsymbol{\mu}^{(n)}, \boldsymbol{\mu}^{(h,zn)}\right) \tag{2.51}
$$

gdzie K oznacza liczbę składowych GMM, a $P\left(k|\mathbf{s}^{(lwgm,zn)}\right)$ - prawdopodobieństwo przynależności wektora $\mathbf{s}^{(lwgm,zn)}$ do k-tej składowej, wyznaczane z wykorzystaniem reguły Bayesa. Po normalizacji widma można przeprowadzić dalsze etapy parametryzacji sygnału. W pracach [5, 81, 179, 4] można znaleźć modyfikacje metody VTS.

2.5.2. Wyrównywanie histogramów i rotacja przestrzeni parametrów

Metoda normalizacji, polegająca na wyrównywaniu histogramów parametrów sygnału mowy, została zaproponowana w pracach [23, 108], z tym, że w [108] dodatkowo uzupełniono ją o rotację przestrzeni tych parametrów. Metoda wyrównywania histogramów opiera się na założeniu, że rozkłady prawdopodobieństwa dla

wystarczająco długiego fragmentu sygnału są niezależne od treści wypowiedzi, zależne natomiast od warunków transmisyjnych i cech osobniczych mówcy. Wyrównując histogram przeprowadza się normalizację wszystkich statystyk rozkładu, jest to zatem rozszerzenie metod normalizacji wartości oczekiwanej czy wariancji. Parametrami poddawanymi normalizacji mogą być np. zlogarytmowane wartości wyjściowe z banku filtrów lub współczynniki MFCC. Niech $R\left(o_i\right)$ oznacza dystrybuantę empiryczną odniesienia dla parametru o_i, wyznaczoną na podstawie danych z części uczącej bazy. Niech $R^{(zn)}\left(o_i^{(zn)}\right)$ oznacza dystrybuantę wyznaczoną dla sygnału zniekształconego. Wyrównywanie histogramu dokonywane jest wg zależności:

$$o_i^{(norm)} = R\left(R^{(zn)^{-1}}\left(o_i^{(zn)}\right)\right) \tag{2.52}$$

Histogramy poszczególnych parametrów wyrównywane są niezależnie od siebie. Taka strategia jest słuszna w przypadku założenia statystycznej niezależności parametrów. W praktyce założenie to nie jest spełnione, zwłaszcza w przypadku zlogarytmowanych wartości wyjściowych z banku filtrów. Pewnym rozwiązaniem tego problemu jest rotacja przestrzeni parametrów, przeprowadzana przed wyrównywaniem histogramów w sposób następujący:

$$\mathbf{o}^{(rot)} = \left(\mathbf{P}\mathbf{R}_{\beta^{(rot)}}\mathbf{P}^T + \mathbf{I} - \mathbf{P}\mathbf{P}^T\right) \cdot \mathbf{o}^{(zn)} \tag{2.53}$$

$$\mathbf{R}_{\beta^{(rot)}} = \begin{bmatrix} \cos\beta^{(rot)} & \sin\beta^{(rot)} \\ -\sin\beta^{(rot)} & \cos\beta^{(rot)} \end{bmatrix} \tag{2.54}$$

$$\beta^{(rot)} = \arccos\left(\mathbf{v}^{(1)^T} \cdot \mathbf{v}^{(1,zn)}\right) \tag{2.55}$$

$$\mathbf{P} = \left[\mathbf{v}^{(1)} \quad \frac{\mathbf{v}^{(1,zn)} - \left(\mathbf{v}^{(1)^T} \cdot \mathbf{v}^{(1,zn)}\right)\mathbf{v}^{(1)}}{\left\|\mathbf{v}^{(1,zn)} - \left(\mathbf{v}^{(1)^T} \cdot \mathbf{v}^{(1,zn)}\right)\mathbf{v}^{(1)}\right\|_2}\right] \tag{2.56}$$

gdzie $\mathbf{v}^{(1)}$ jest wektorem własnym macierzy kowariancji wektorów $\mathbf{o}$, odpowiadającym największej wartości własnej, a $\mathbf{v}^{(1,zn)}$ - analogicznym wektorem własnym dla macierzy kowariancji wektorów $\mathbf{o}^{(zn)}$. Wektory własne mają unormowaną do jedności długość. Macierz $\mathbf{P}$ zawiera ortonormalną bazę dwuwymiarowej podprzestrzeni rozpiętej z wykorzystaniem wektorów $\mathbf{v}^{(1)}$ i $\mathbf{v}^{(1,zn)}$, a $\mathbf{R}_{\beta^{(rot)}}$ jest macierzą obrotu w tej podprzestrzeni o kąt $\beta^{(rot)}$ zawarty między wektorami $\mathbf{v}^{(1)}$ i $\mathbf{v}^{(1,zn)}$. Celem całej transformacji jest taka rotacja przestrzeni parametrów sygnału zniekształconego, by właściwy dla niej wektor $\mathbf{v}^{(1,zn)}$ był równoległy do wektora $\mathbf{v}^{(1)}$ dla sygnału niezniekształconego.

Innym istotnym problemem jest zmienny udział fragmentów zawierających ciszę w poddawanych normalizacji sygnałach. Histogramy dla fragmentów mowy i ciszy różnią się znacznie, konieczne jest zatem uwzględnienie ich wzajemnych proporcji.

Można zrealizować to wykorzystując np. detektor obecności sygnału mowy (VAD - *voice activity detector*).

Wyrównywanie histogramów wymaga długich, nawet kilkuminutowych fragmentów sygnału zniekształconego. W pracy [54] zaproponowano normalizację wybranych kwantyli rozkładów prawdopodobieństwa parametrów sygnału, która może być efektywnie przeprowadzona dla krótszych fragmentów sygnału.

2.5.3. Vocal Tract Length Normalization (VTLN) - normalizacja długości toru głosowego

Normalizacja długości toru głosowego ma na celu kompensację systematycznych przesunięć częstotliwości formantów w widmie sygnału mowy różnych mówców. Nazwa metody wiąże się z faktem, że przesunięcia te są skorelowane z długością toru głosowego. Normalizacja dokonywana jest poprzez skalowanie osi częstotliwości:

$$f^{(norm)} = g\left(f, \boldsymbol{\alpha}\right) \tag{2.57}$$

gdzie f i $f^{(norm)}$ oznaczają częstotliwości odpowiednio przed i po skalowaniu, a $g\left(f, \boldsymbol{\alpha}\right)$ jest funkcją, której kształt jest opisany wektorem parametrów $\boldsymbol{\alpha} \in \mathscr{A}$, gdzie zbiór $\mathscr{A}$ zależy od postaci tej funkcji. Funkcja $g\left(f, \boldsymbol{\alpha}\right)$ powinna być rosnąca oraz spełniać warunki $g(0, \boldsymbol{\alpha}) = 0$ i $g\left(f^{(max)}, \boldsymbol{\alpha}\right) = f^{(max)}$, gdzie $f^{(max)}$ to górna granica częstotliwości analizowanego widma. Zaproponowano wiele postaci funkcji $g\left(f, \boldsymbol{\alpha}\right)$, najczęściej spotykane to:

- Liniowa [131, 45, 51]:

  $$g\left(f, \alpha\right) = \alpha \cdot f \tag{2.58}$$

  Funkcja ta nie spełnia warunku granicznego $g\left(f^{(max)}, \alpha\right) = f^{(max)}$. Opisywana jest jednym parametrem $\alpha \in (0; \infty)$.

- Łamana dwusegmentowa [88, 167]:

  $$g\left(f, \alpha\right) = \begin{cases} \alpha \cdot f, & \text{dla } \; f \in [0; f_0] \\ \alpha \cdot f + \frac{f^{(max)} - \alpha \cdot f_0}{f^{(max)} - f_0}\left(f - f_0\right), & \text{dla } \; f \in (f_0; f^{(max)}] \end{cases} \tag{2.59}$$

  Opisywana jest jednym parametrem zmiennym $\alpha \in (0; f^{(max)}/f_0)$ oraz jednym stałym $f_0 \in (0; f^{(max)})$, określającym odciętą punktu łączenia odcinków liniowych.

- Łamana wielosegmentowa. Funkcja, której wykres jest łamaną i spełnia warunki monotoniczności oraz graniczne. Opisywana jest wektorem parametrów zmiennych $\boldsymbol{\alpha}$ i parametrami stałymi, określającymi położenie punktów łączenia odcinków liniowych.

- Biliniowa:

$$g(f,\alpha) = f + \frac{2 \cdot f^{(max)}}{\pi} \operatorname{arc\,tg}\left(\frac{(1-\alpha)\sin\left(\pi \cdot f/f^{(max)}\right)}{1-(1-\alpha)\cos\left(\pi \cdot f/f^{(max)}\right)}\right) \tag{2.60}$$

  Funkcja ta opisuje skalowanie osi częstotliwości występujące przy biliniowym przekształceniu płaszczyzny $\mathscr{Z}$. Opisywana jest jednym parametrem $\alpha \in (0;2)$.

W pracach [32, 184, 103] zaproponowano jeszcze inne postaci funkcji skalującej. Skalowanie osi częstotliwości przeprowadzane powinno być tak, aby zachować wartości amplitudy przesuwanych fragmentów widma.

Skalowanie osi częstotliwości można przeprowadzić bezpośrednio na widmie amplitudowym sygnału [167] lub modyfikując częstotliwości środkowe i pasma filtrów w banku filtrów melowych [131, 88, 167, 184, 51, 186]. Można również modyfikować parametry cepstralne, korzystając z zależności opisujących wpływ skalowania osi częstotliwości na cepstrum [128, 103] lub modyfikować sygnał w dziedzinie czasu tak, by osiągnąć pożądaną modyfikację widma [159].

Estymacja wartości parametrów $\boldsymbol{\alpha}$ może być przeprowadzona na różne sposoby. W [45] zaproponowano estymację na podstawie oszacowanych wartości częstotliwości formantów, metoda ta okazała się jednak zawodna. Najczęściej spotykana jest estymacja mająca na celu maksymalizację pewnej miary oceny rozpoznania, zazwyczaj będącej funkcją wiarogodności. W tym przypadku rozwiązaniem optymalnym byłaby łączna estymacja wartości $\boldsymbol{\alpha}$ i rozpoznanie sekwencji stanów $\mathbf{q}$:

$$\left\{\mathbf{q}^{(opt)}, \boldsymbol{\alpha}^{(opt)}\right\} = \underset{\mathbf{q}\in\mathscr{Q},\boldsymbol{\alpha}\in\mathscr{A}}{\arg\max}\, P\left(\mathbf{q}|\mathbf{O}_{\boldsymbol{\alpha}}, \Theta\right) \tag{2.61}$$

przy czym użyto symbolu $\mathbf{O}_{\boldsymbol{\alpha}}$ w celu podkreślenia faktu, że zmiany wartości $\boldsymbol{\alpha}$ powodują modyfikacje wartości sekwencji wektorów obserwacji $\mathbf{O}$. Łączna optymalizacja jest w praktyce trudno realizowalna z uwagi na dużą złożoność obliczeniową. Stosuje się wiec mniej kosztowne obliczeniowo, choć suboptymalne strategie.

W pierwszej strategii najpierw dokonuje estymacji wartości $\boldsymbol{\alpha}$ tak, by maksymalizować dopasowanie obserwacji $\mathbf{O}_{\boldsymbol{\alpha}}$ do modelu akustycznego, mierzone prawdopodobieństwem warunkowym:

$$\boldsymbol{\alpha}^{(opt)} = \underset{\boldsymbol{\alpha}\in\mathscr{A}}{\arg\max}\, P\left(\mathbf{O}_{\boldsymbol{\alpha}}|\Theta\right) \tag{2.62}$$

Prawdopodobieństwo to można obliczyć wykorzystując metodę *forward-backward* (zob. dodatek C). Następnie przeprowadza się rozpoznanie sekwencji stanów $\mathbf{q}$.

W strategii drugiej wykonuje się naprzemiennie dwa kroki (i oznacza numer iteracji):

1. Wyznacz $\mathbf{q}_i$ przy ustalonej wartości $\boldsymbol{\alpha}_{i-1}$:

$$\mathbf{q}_i = \arg\max_{\mathbf{q}\in\mathscr{Q}} P\left(\mathbf{q}|\mathbf{O}_{\boldsymbol{\alpha}_{i-1}}, \Theta\right) \tag{2.63}$$

2. Reestymuj $\boldsymbol{\alpha}_i$ przy ustalonej sekwencji $\mathbf{q}_i$:

$$\boldsymbol{\alpha}_i = \arg\max_{\boldsymbol{\alpha}\in\mathscr{A}} P\left(\mathbf{q}_i|\mathbf{O}_{\boldsymbol{\alpha}}, \Theta\right) \tag{2.64}$$

Iteracje przerywa się po osiągnięcia zbieżności, czyli w momencie, w którym wzrost prawdopodobieństwa maksymalizowanego w równaniach (2.63) i (2.64) względem poprzedniej iteracji jest poniżej zadanego progu.

Idea normalizacji systematycznych zmian w widmie doczekała się również innych rozwiązań, przykładem może być tutaj zastosowanie w parametryzacji transformacji Mellina, dla której liniowe skalowanie osi częstotliwości sygnału nie wpływa na moduł tej transformaty [62].

2.5.4. Algorytm Eigenvoices

Eigenvoices (EV) jest algorytmem adaptacji modelu statystycznego ukierunkowanym na kompensację cech osobniczych mówcy. Pierwowzorem tej metody jest algorytm Eigenfaces, znany w zagadnieniu rozpoznawania twarzy. Można uznać, że EV jest specyficznym algorytmem „miękkiego" podziału mówców na klasy, w którym wartości parametrów modelu statystycznego mówcy są ważoną średnią pewnego zbioru wartości parametrów bazowych. W swojej podstawowej wersji EV umożliwia adaptację wartości oczekiwanych rozkładów normalnych modelu akustycznego [83].

W etapie przygotowawczym algorytmu wykonuje się następujące czynności:

1. Korzystając z części uczącej bazy nagrań wyznacz wartości parametrów modelu języka Θ tak, że wartości oczekiwane rozkładów normalnych $\boldsymbol{\mu}_{s,i,k}$ (s - numer mówcy, i - numer stanu modelu, k - numer składowej GMM) wyznaczane są osobno dla każdego mówcy (model SD), a wartości pozostałych parametrów wspólnie dla wszystkich mówców (model SI).

2. Dla każdego mówcy s utwórz tzw. superwektor poprzez połączenie wszystkich wektorów $\boldsymbol{\mu}_{s,i,k}$ dla tego mówcy:

$$\boldsymbol{\mu}_s^{(sv)} = \left[\boldsymbol{\mu}_{s,0,0}{}^T \ \cdots \ \boldsymbol{\mu}_{s,I-1,K-1}{}^T\right]^T \tag{2.65}$$

3. Przeprowadź analizę składowych głównych (PCA - *principal component analysis*) superwektorów:

$$\boldsymbol{\mu}^{(sv,sr)} = \frac{1}{S}\sum_{s=0}^{S-1}\boldsymbol{\mu}_s^{(sv)} \tag{2.66}$$

$$\boldsymbol{\Sigma}^{(sv)} = \frac{1}{S}\left[\boldsymbol{\mu}_0^{(sv)} \ \cdots \ \boldsymbol{\mu}_{S-1}^{(sv)}\right]\cdot\left[\boldsymbol{\mu}_0^{(sv)} \ \cdots \ \boldsymbol{\mu}_{S-1}^{(sv)}\right]^T - \boldsymbol{\mu}^{(sv,sr)}\cdot\boldsymbol{\mu}^{(sv,sr)^T} \tag{2.67}$$

$$\mathbf{V}^{(sv)}\boldsymbol{\Lambda}^{(sv)}\mathbf{V}^{(sv)^T} = \boldsymbol{\Sigma}^{(sv)} \tag{2.68}$$

gdzie równanie (2.68) opisuje rozkład własny macierzy $\boldsymbol{\Sigma}^{(sv)}$.

4. Wybierz z macierzy $\mathbf{V}^{(sv)}$ R kolumn będących wektorami własnymi, odpowiadającymi największym wartościom własnym z macierzy przekątniowej $\boldsymbol{\Lambda}^{(sv)}$. Niech wektory te będą kolejnymi kolumnami macierzy $\mathbf{U}$.

W etapie adaptacji do nieznanego mówcy wyznacza się R współczynników aproksymacji (wektor $\boldsymbol{\beta}$) w bazie $\mathbf{U}$:

$$\boldsymbol{\mu}^{(sv,adap)} = \boldsymbol{\mu}^{(sv,sr)} + \mathbf{U}\cdot\boldsymbol{\beta} \tag{2.69}$$

Następnie wyznacza się wartości oczekiwane rozkładów normalnych rozkładając superwektor $\boldsymbol{\mu}^{(sv,adap)}$ zgodnie z (2.65).

Współczynniki $\boldsymbol{\beta}$ można wyznaczać na różne sposoby. Najpowszechniejsza jest estymacja metodą maksymalnej wiarogodności:

$$\boldsymbol{\beta}^{(opt)} = \arg\max_{\boldsymbol{\beta}\in\mathscr{R}^R} P\left(\mathbf{O}|\Theta_{\boldsymbol{\beta}}\right) \tag{2.70}$$

Istnieje iteracyjny algorytm typu E-M, mający analityczne rozwiązanie zadania maksymalizacji, pozwalający na wyznaczenie $\boldsymbol{\beta}^{(opt)}$ (zob. rozdział 3.1). Widoczne jest tutaj podobieństwo do metody estymacji parametrów w algorytmie VTLN danej równaniem (2.62). W przypadku EV można również zrealizować strategię podobną do tej danej równaniami (2.63, 2.64), jako szczególny przypadek strategii pierwszej (zob. rozdział 3.1).

W przypadku małej ilości danych adaptacyjnych, można zastosować interpolację modelu zaadaptowanego (SD) z modelem SI. Interpolacja rozumiana jest tutaj jako średnia ważona wartości oczekiwanych z modelu SD i SI.

Istnieje wiele modyfikacji metody EV, z których za najciekawsze można uznać wykorzystanie nieliniowej analizy PCA [96, 97], zastosowanie oddzielnych transformacji dla grup parametrów [164], zastosowanie transformacji pogrupowanych hierarchicznie wg kontekstów fonemów [82]. W [57] uzupełniono EV o rozkłady prawdopodobieństwa współczynników $\boldsymbol{\beta}$ dane *a priori* oraz zastosowano adaptację wariancji rozkładów normalnych modelu. W [72] opisano algorytm adaptacji wykorzystujący

podobną jak w EV analizę PCA, ale użytą do modelowania korelacji między mówcami. W [187] natomiast w podobny sposób badano korelację między różnymi wypowiedziami.

2.5.5. Inne metody

Dużą grupę algorytmów stanowią metody normalizacji cepstralnej, zaprojektowane w celu kompensacji zniekształceń transmisyjnych. Jedną z najbardziej znanych jest *Codeword Dependent Cepstral Normalization* (CDCN) [3], w której stosuje się kwantyzację wektorową przestrzeni parametrów cepstralnych i wyznacza wektory korekcyjne osobno dla każdego wektora kodowego, estymując jednocześnie wartości parametrów opisujących zniekształcenia transmisyjne. Model kanału transmisyjnego w tej metodzie uwzględnia szum addytywny i zniekształcenia liniowe. Inne warianty i modyfikacje tej metody opisane są w pracach [3, 58, 26, 111, 91]. W [1, 21] przedstawiono natomiast modyfikacje prostej metody *Cepstral Mean Normalization* (CMN), polegającej na odejmowaniu od parametrów cepstralnych ich średniej wyznaczonej z całej wypowiedzi.

Zaproponowano wiele algorytmów mających na celu kompensację tylko szumu addytywnego. Metoda *Parallel Model Combination* (PMC) polega na uzupełnieniu modelu akustycznego języka o model szumu stacjonarnego, poprzez wprowadzenie dodatkowych stanów [14, 44, 61]. Problem kompensacji szumu niestacjonarnego poruszony został w [77], gdzie wykorzystano w tym celu algorytm prognozy Kalmana. Zmodyfikowany algorytm Viterbiego, pozwalający na kompensację wpływu szumu impulsowego, zaproponowano w [152]. W [106] natomiast przedstawiono algorytm kompensacji szumu addytywnego o nieznanym widmie. Rozwiązanie to bazuje na modelu akustycznym zawierającym przykłady wpływu różnych rodzajów szumu na parametry sygnału podzielonego na podpasma. Adaptacja dokonywana jest poprzez wybór tych elementów modelu, które najlepiej odpowiadają przetwarzanemu sygnałowi.

Intensywnie rozwijane są obecnie metody brakujących cech (ang. *missing features*) [102, 50, 188, 107, 134, 124], w których sygnał dzielony jest na podpasma i w rozpoznawaniu uwzględniane są tylko te podpasma, których zawartość uznano za wiarygodną. Umożliwia to skuteczne rozpoznawanie mowy nawet w przypadku całkowitego wymazania informacji w pewnych podpasmach.

Pogłos skutkujący znacznymi zniekształceniami międzyramkowymi najczęściej jest traktowany jako składowa szumu addytywnego. W pracy [174] podano natomiast algorytm adaptacji modelu HMM, uwzględniający przyczyny i mechanizm powstawania zniekształceń międzyramkowych.

W algorytmach klasy *Maximum Likelihood Linear Regression* (MLLR), mających na celu kompensację zarówno warunków transmisyjnych, jak i cech osobniczych,

stosowane jest afiniczne przekształcenie przestrzeni parametrów sygnału [98] lub parametrów modelu:

$$\mathbf{x}^{(komp)} = \mathbf{R} \cdot \mathbf{x} + \mathbf{b} \tag{2.71}$$

gdzie $\mathbf{x}$ i $\mathbf{x}^{(komp)}$ oznaczają parametry modelu lub sygnału odpowiednio przed i po kompensacji. W przypadku adaptacji modelu akustycznego, stosować można różne parametry przekształcenia dla różnych grup (klas) stanów modelu [33, 20]. W celu zredukowania liczby parametrów transformacji, których wartości należy wyznaczyć, zastosować można pewne ograniczenia nałożone na strukturę macierzy $\mathbf{R}$ (np. pasmowa, blokowo-przekątniowa) [78, 22] lub przeprowadzić, podobnie jak w algorytmie EV, analizę PCA zbioru macierzy uzyskanych dla różnych mówców [16]. Estymacja wartości parametrów transformacji dokonywana jest zazwyczaj metodą maksymalnej wiarogodności. Zaproponowano jednak również metody mające na celu maksymalizację zdolności klasyfikacji adaptowanego modelu [163, 48]. Inne warianty i modyfikacje metody MLLR znaleźć można w [173, 141].

W algorytmach typu *Maximum a Posteriori* (MAP) dokonywana jest adaptacja wartości parametrów modelu przy wykorzystaniu ich rozkładów prawdopodobieństwa zadanych *a priori* oraz dostępnych parametrów sygnału mówcy adaptowanego. Wśród tego typu metod wymieć można algorytm *Structural* MAP (SMAP) [148], w którym zastosowano hierarchiczną klastryzację parametrów modelu tak, aby poprawić efektywność adaptacji dla krótkich wypowiedzi. W [76, 75] opisano natomiast algorytm adaptacji typu MAP, w którym do wyznaczania rozkładów prawdopodobieństwa *a priori* wykorzystany został odpowiednio skonstruowany model przestrzeni mówców.

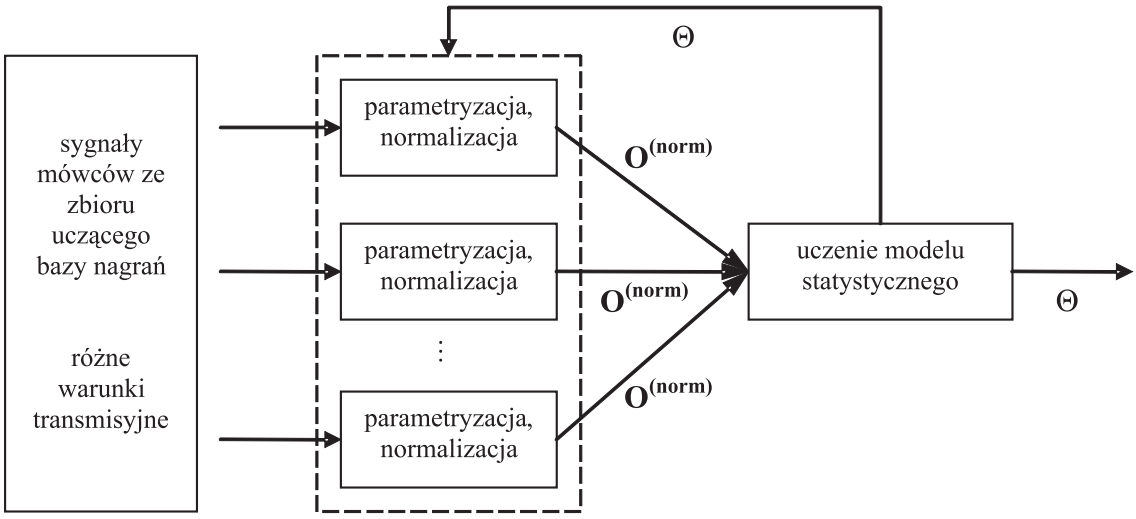
Ważną grupą algorytmów adaptacji są metody, w których stosowany jest podział mówców na klasy. Przykładem mogą być tu algorytmy *Cluster Adaptive Training* (CAT) i *Discriminative* CAT (DCAT) [42, 182], w których wartości parametrów modelu dla danego mówcy wyznacza się jako sumę ważoną parametrów z różnych klas. Inne algorytmy stosujące podział mówców na klasy opisano w [132, 2].

W literaturze znaleźć można również algorytmy normalizacji przestrzeni parametrów sygnału, np. [178], metody wykorzystujące sieci neuronowe, np. [158], czy też metody hybrydowe, łączące w sobie kilka wcześniej zaproponowanych algorytmów, np. [140, 18].

Z metod odpornej parametryzacji wymienić można RASTA [60, 8, 24, 53], polegającą na filtracji trajektorii współczynników cepstralnych filtrem o pasmie przepustowym ok. 2-15 Hz, co ma na celu wyeliminowanie z tych trajektorii składowych nie związanych z artykulacją mowy. Metody *Subband Spectral Centroid Histograms* [41], *Ensemble Interval Histogram* (EIH) [133] czy *Human Factor Cepstral Coefficients* [189] mają natomiast na celu minimalizację wpływu szumu addytywnego na parametry sygnału.

2.5.6. Uczenie systemu ukierunkowane na kompensację

Strategia uczenia systemu ARM ukierunkowana na kompensację (SAT - *speaker adaptive training*) została zilustrowana na rys. 2.4. Model statystyczny systemu Θ uczony jest z wykorzystaniem danych ze zbioru uczącego bazy nagrań, ale po wcześniejszej ich normalizacji. Zbiór uczący zawiera sygnały różnych mówców, ale może zawierać też sygnały zniekształcone w różnych warunkach transmisyjnych. Normalizacja wymaga jednak na ogół istnienia pewnego modelu Θ, zatem uczenie SAT przeprowadza się iteracyjnie, rozpoczynając od modelu SI. Celem tej strategii jest minimalizacja rozrzutu wartości parametrów sygnału dla różnych mówców, a co za tym idzie, zmniejszenie wariancji wyznaczonych w modelu Θ rozkładów prawdopodobieństwa. To z kolei skutkuje zwiększoną zdolnością klasyfikacji modelu, a więc również zwiększoną skutecznością działania systemu ARM. Uczenie SAT można stosować dla różnych metod normalizacji, istotne jest jednak to, by w zarówno w czasie uczenia systemu, jak i w czasie jego pracy użytkowej, stosowana metoda była taka sama.



Rys. 2.4. Uczenie systemu ARM ukierunkowane na kompensację.

2.5.7. Ocena przydatności znanych metod do rozwiązania zagadnienia postawionego w pracy

Przedstawione powyżej algorytmy umożliwiające kompensację liniowych zniekształceń transmisyjnych i cech osobniczych mówcy charakteryzują się następującą prawidłowością: wraz ze wzrostem liczby parametrów, których wartości należy wyznaczyć podczas kompensacji, rośnie jej skuteczność. Kosztem jest jednak konieczność zapewnienia fragmentów sygnału o długości co najmniej kilkunastu sekund. Metody nazywane szybkimi, do których zaliczyć można m.in. EV, VTLN, CAT czy SMAP,

wymagają fragmentów sygnału o długości co najmniej kilku sekund. Ponadto algorytmy szybkie wymagają w swym etapie przygotowawczym uczenia z zastosowaniem dużego zbioru mówców (o liczności powyżej 50). Za wyjątek można uważać tutaj VTLN, którego skuteczność działania nie jest związana z licznością zbioru uczącego.

Większość metod kompensacji bazuje na iteracyjnej modyfikacji wartości parametrów modelu lub sygnału, przy czym niezwykle istotny jest punkt początkowy. W przypadku złego zainicjalizowania algorytmu kompensacja staje się nieskuteczna, a dla krótkich wypowiedzi mała ilość dostępnych danych zwiększa prawdopodobieństwo złej inicjalizacji, czyli inaczej mówiąc - złego pierwszego rozpoznania.

Znane z literatury algorytmy nie nadają się zatem do bezpośredniego zastosowania w systemie rozpoznawania komend. Do eksperymentalnej oceny przydatności znanych metod do rozwiązania zagadnienia postawionego w pracy wybrano metodę EV, uzupełnioną o elementy algorytmu VTS po zaniedbaniu w nim wpływu szumu. W zaproponowanej oryginalnej metodzie kompensacji wykorzystano również elementy metody VTLN.

3. Zmodyfikowany algorytm Eigenvoices

Spośród znanych z literatury algorytmów kompensacji (zob. rozdział 2.5) do eksperymentalnej oceny przydatności w rozwiązywanym w pracy zagadnieniu wybrano algorytm EV (zob. rozdział 2.5.4). Rezultaty podane w literaturze wskazują bowiem, że jest on jednym z najskuteczniejszych w przypadku zadania rozpoznawania krótkich wypowiedzi.

Oryginalny algorytm EV nie umożliwia kompensacji liniowych zniekształceń transmisyjnych. Poniżej zaproponowano modyfikację tego algorytmu, mającą na celu umożliwienie kompensacji takich zniekształceń. Zaproponowano również zastosowanie metody SAT uczenia modelu dla zmodyfikowanego algorytmu EV. Przedstawiono uzyskane wyniki rozpoznawalności izolowanych ramek sygnału mowy oraz przeprowadzono ich analizę.

3.1. Opis algorytmu

Podstawowy opis oryginalnego algorytmu znajduje się w rozdziale 2.5.4. Estymacja wartości współczynników $\boldsymbol{\beta}^{(opt)} \in \mathscr{R}^R$ przeprowadzona może być opisanym niżej iteracyjnym algorytmem typu E-M, w którym naprzemiennie wykonuje się dwa kroki:

1. Wyznacz wartości $\gamma_{i,t,k}^{(gmm)}$ (zob. dodatek C), będące prawdopodobieństwami przynależności wektora obserwacji $\mathbf{o}_t$ do i-tego stanu modelu i k-tej składowej GMM rozkładu prawdopodobieństwa współczynników MFCC dla tego stanu. Rozkłady te opisane są parametrami $\left\{c_{i,k}, \boldsymbol{\mu}_{i,k}^{(adap)}, \boldsymbol{\Sigma}_{i,k}\right\}$, gdzie $\boldsymbol{\mu}_{i,k}^{(adap)}$ są obliczone z (2.69) i (2.65) dla wartości $\boldsymbol{\beta}^{(opt)}$ z poprzedniej iteracji.

2. Reestymuj wartości współczynników $\boldsymbol{\beta}^{(opt)}$, poprzez rozwiązanie układu R równań:

$$\sum_{t=0}^{T-1}\sum_{i=0}^{I-1}\sum_{k=0}^{K-1}\gamma_{i,t,k}^{(gmm)}\left(\mathbf{o}_t - \boldsymbol{\mu}_{i,k}^{(sv,sr)} - \sum_{l=0}^{R-1}\beta_l^{(opt)}\mathbf{u}_{l,i,k}\right)^T \boldsymbol{\Sigma}_{i,k}^{-1}\mathbf{u}_{r,i,k} = 0, \qquad r = 0,\ldots,R-1 \tag{3.1}$$

gdzie $\mathbf{u}_{r,i,k}$ i $\boldsymbol{\mu}_{i,k}^{(sv,sr)}$ oznaczają te części wektorów $\mathbf{u}_r$ i $\boldsymbol{\mu}^{(sv,sr)}$, które odpowiadają wektorowi $\boldsymbol{\mu}_{i,k}$.

Algorytm zatrzymywany jest w momencie, w którym wzrost miary dopasowania modelu statystycznego do danych wejściowych (prawdopodobieństwo warunkowe $P\left(\mathbf{O}|\Theta_{\boldsymbol{\beta}^{(opt)}}\right)$) względem poprzedniej iteracji jest poniżej zadanego progu.

Inną strategię estymacji można uzyskać zastępując w równaniach (3.1) przynależność „miękką" $\gamma_{i,t,k}^{(gmm)}$ przynależnością „twardą" danej obserwacji $\mathbf{o}_t$ do danego stanu i, uzyskaną z algorytmu Viterbiego.

W przeprowadzonych badaniach obliczano rozpoznawalność izolowanych ramek, przy czym zakładano, że przynależność ramek do fonemów jest znana. Każdy stan i odpowiadał jednemu fonemowi. Podczas wstępnych eksperymentów okazało się ponadto, że wystarczające jest zastosowanie modelowania rozkładów prawdopodobieństwa współczynników MFCC dla danego fonemu i danego mówcy za pomocą pojedynczych rozkładów normalnych o przekątniowych macierzach kowariancji. W takim przypadku współczynniki $\boldsymbol{\beta}^{(opt)}$ oblicza się z poniższego układu równań bez konieczności przeprowadzania iteracji.

$$\sum_{n=0}^{N-1}\sum_{i=0}^{I-1}\gamma_{i,n}^{(fon)}\left(\mathbf{o}_n-\boldsymbol{\mu}_i^{(sv,sr)}-\sum_{l=0}^{R-1}\beta_l^{(opt)}\mathbf{u}_{l,i}\right)^T\boldsymbol{\Sigma}_i^{-1}\mathbf{u}_{r,i}=0,\quad r=0,\ldots,R-1 \tag{3.2}$$

gdzie N oznacza liczbę użytych w estymacji izolowanych ramek. Współczynnik przynależności $\gamma_{i,n}^{(fon)}=1$, gdy ramka n należy do fonemu i oraz $\gamma_{i,n}^{(fon)}=0$ w przeciwnym wypadku. W estymacji wykorzystywano po 500 ramek na fonem dla danego mówcy (lub mniej, jeśli nie było tyle dostępnych). We wstępnych eksperymentach sprawdzono, że jest to wartość wystarczająca.

W celu umożliwienia kompensacji zniekształceń liniowych uzupełniono bazę $\mathbf{U}$ o wektory modelujące addytywny wpływ zniekształceń liniowych w dziedzinie MFCC:

$$\mathbf{U}^{(mod)}=\left[\mathbf{U}\ \ \mathbf{e}_1^{(sv)}\ \ \mathbf{e}_2^{(sv)}\ \ \cdots\ \ \mathbf{e}_{R^{(e)}}^{(sv)}\right] \tag{3.3}$$

$$\mathbf{e}_j^{(sv)}=\left[\underbrace{\mathbf{e}_j^T\ \ \mathbf{e}_j^T\ \ \cdots\ \ \mathbf{e}_j^T}_{I\cdot K}\right]^T \tag{3.4}$$

gdzie $\mathbf{e}_j$ oznacza wersor zawierający wartość 1 dla j-tej współrzędnej i 0 dla pozostałych. Wartość $R^{(e)}$ może być co najwyżej równa liczbie współczynników MFCC, w praktyce zniekształcenia liniowe o łagodnym przebiegu charakterystyki amplitudowej wpływają głównie na pierwsze współczynniki MFCC, więc $R^{(e)}$ może być mniejsza od liczby tych współczynników. W celu przywrócenia ortonormalności bazy

rozpinanej przez kolumny macierzy $\mathbf{U}^{(mod)}$ przeprowadzano następującą jej modyfikację:

$$\mathbf{U}^{(mod,ort)} = \left[\mathbf{U} \;\; \mathbf{e}_1^{(sv,ort)} \;\; \mathbf{e}_2^{(sv,ort)} \;\; \cdots \;\; \mathbf{e}_{R^{(e)}}^{(sv,ort)}\right] \tag{3.5}$$

$$\mathbf{e}_j^{(sv,ort)} = \frac{(\mathbf{I} - \mathbf{P}_{j-1})\,\mathbf{e}_j^{(sv)}}{\left\|(\mathbf{I} - \mathbf{P}_{j-1})\,\mathbf{e}_j^{(sv)}\right\|_2} \tag{3.6}$$

gdzie $\mathbf{P}_{j-1}$ jest macierzą rzutu ortogonalnego na podprzestrzeń rozpinaną przez kolumny macierzy $\mathbf{U}$ i $j-1$ wektorów $\mathbf{e}_j^{(sv,ort)}$:

$$\mathbf{P}_j = \left[\mathbf{U} \;\; \mathbf{e}_1^{(sv,ort)} \;\; \cdots \;\; \mathbf{e}_j^{(sv,ort)}\right] \cdot \left[\mathbf{U} \;\; \mathbf{e}_1^{(sv,ort)} \;\; \cdots \;\; \mathbf{e}_j^{(sv,ort)}\right]^T \tag{3.7}$$

Estymacja współczynników $\boldsymbol{\beta}^{(opt)}$ odbywa się w zmodyfikowanym algorytmie tak samo, jak w niezmodyfikowanym, z tym, że teraz liczba tych współczynników wynosi $R^{(mod)} = R + R^{(e)}$.

Zaproponowano, by przed wyznaczeniem wartości parametrów modelu (macierzy $\mathbf{U}^{(mod,ort)}$ i wektora $\boldsymbol{\mu}^{(sv,sr)}$) w zmodyfikowanej metodzie EV zredukować zmienność współczynników MFCC związaną ze zniekształceniami liniowymi. Wykorzystano w tym celu strategię uczenia SAT (zob. rozdział 2.5.6). Normalizacja przeprowadzana była za pomocą uproszczonego algorytmu VTS (zob. rozdział 2.5.1), w którym pominięto wpływ szumu addytywnego, zmieniono dziedzinę normalizacji ze zlogarytmowanych wartości wyjściowych z filtrów melowych na współczynniki MFCC, a zniekształcenia liniowe dla danego mówcy modelowano deterministycznym wektorem $\mathbf{o}^{(h,zn)}$. Rozkładami odniesienia były brzegowe rozkłady prawdopodobieństwa współczynników MFCC mówców ze zbioru uczącego. Zastosowano modelowanie GMM za pomocą $K = 3$ składowych normalnych. Procentowy udział ramek poszczególnych fonemów w danych wykorzystywanych w modelowaniu był taki, jaki występował w nagraniach w bazie. Dla j-tego współczynnika MFCC rozkłady opisane są parametrami $\left\{c_{j,k}^{(o)}, \mu_{j,k}^{(o)}, {\sigma_{j,k}^{(o)}}^2\right\}$. Iteracyjny, składający się z dwóch naprzemiennie wykonywanych kroków, algorytm estymacji wektora $\mathbf{o}^{(h,zn)}$ dla danego mówcy, w przypadku podanych uproszczeń, ma postać:

1. Wyznacz prawdopodobieństwa przynależności danego wektora $\mathbf{o}_n^{(zn)}$ do danych składowych GMM:

$$P\left(n, j, k | o_{j,n}^{(zn)}, o_j^{(h,zn)}\right) = \frac{c_{j,k}^{(o)} \mathcal{N}\left(o_{j,n}^{(zn)}; \mu_{j,k}^{(o)} - o_j^{(h,zn)}, {\sigma_{j,k}^{(o)}}^2\right)}{\sum_{m=0}^{K-1} c_{j,m}^{(o)} \mathcal{N}\left(o_{j,n}^{(zn)}; \mu_{j,m}^{(o)} - o_j^{(h,zn)}, {\sigma_{j,m}^{(o)}}^2\right)} \tag{3.8}$$

2. Reestymuj wektor $\mathbf{o}^{(h,zn)}$:

$$\bar{o}_j^{(h,zn)} = \frac{\sum_{k=0}^{K-1} \frac{1}{{\sigma_{j,k}^{(o)}}^2} \sum_{n=0}^{N-1} \left(o_{j,n}^{(zn)} - \mu_{j,k}^{(o)} \right) P\left(n, j, k | o_{j,n}^{(zn)}, o_j^{(h,zn)} \right)}{\sum_{k=0}^{K-1} \frac{1}{{\sigma_{j,k}^{(o)}}^2} \sum_{n=0}^{N-1} P\left(n, j, k | o_{j,n}^{(zn)}, o_j^{(h,zn)} \right)} \tag{3.9}$$

Algorytm zatrzymywany jest w momencie, w którym wzrost miary dopasowania modelu statystycznego do danych wejściowych (prawdopodobieństwo warunkowe $P\left(\mathbf{O}^{(zn)} | \Theta_{\mathbf{o}^{(h,zn)}}\right)$) względem poprzedniej iteracji jest poniżej zadanego progu. W obliczeniach wykorzystywano ten sam zbiór N wektorów $\mathbf{o}_n^{(zn)}$ dla danego mówcy, który stosowano przy wyznaczaniu rozkładów odniesienia. Kreska nad symbolem oznacza wartość uaktualnioną w bieżącej iteracji.

Normalizację natomiast przeprowadzano następująco dla każdej ramki n:

$$o_{j,n}^{(norm)} = o_{j,n}^{(zn)} - o_j^{(h,zn)}, \quad j = 0, \ldots, R^{(e)} - 1 \tag{3.10}$$

Normalizowane są tylko $R^{(e)}$ pierwsze współczynniki MFCC, co wynika z faktu, że w algorytmie możliwa jest kompensacja zniekształceń liniowych tylko dla tych współczynników.

3.2. Wyniki i wnioski

W tabelach 3.1 i 3.2 oraz na rys. 3.1 przedstawiono wyniki rozpoznawalności izolowanych ramek uzyskane dla zbioru uczącego i testowego bazy nagrań. Opis metodyki pomiaru i zastosowanych miar rozpoznawalności przedstawiono w dodatku E. Podano też wyniki uzyskane zarówno przy braku symulowanych zniekształceń transmisyjnych, jak i w przypadku ich symulowania. Zniekształcenia nakładano losowo dla każdego mówcy jako kaskadowe połączenie jednej z pięciu charakterystyk mikrofonów ($\mathbf{h}_i^{(kan)}$, gdzie $i = 1, \ldots, 5$) i jednej trzech charakterystyk dodatkowych: dwóch liniowych ($\mathbf{h}_6^{(kan)}$ i $\mathbf{h}_7^{(kan)}$) oraz charakterystyki stałej równej 1 (zob. dodatek D). W tabelach i na rysunku podano wartości R i $R^{(e)}$ oznaczające liczby zastosowanych odpowiednich wektorów bazowych. Użyto również następujących skrótów: „zb.ucz." - zbiór uczący, „zb.tst." - zbiór testowy, „znk." - symulacja zniekształceń transmisyjnych, „sam." - wynik rozpoznawalności dla samogłosek, „wsz." - wynik rozpoznawalności dla wszystkich fonemów.

Wyniki rozpoznawalności izolowanych ramek uzyskane w sytuacji, kiedy podczas estymacji wartości parametrów algorytmu kompensacji znana jest przynależność ramek do fonemów, mają na celu wskazanie potencjału przydatności badanego algorytmu. Stanowić one będą punkt odniesienia dla wyników uzyskanych w zaproponowanej dalej w pracy oryginalnej metodzie kompensacji.

Tab. 3.1. Wyniki rozpoznawalności izolowanych ramek dla zmodyfikowanej metody EV. Zastosowano miarę oceny $c_1^{(rs)}$. Czcionką pogrubioną zaznaczano najwyższy wynik w danej kolumnie.

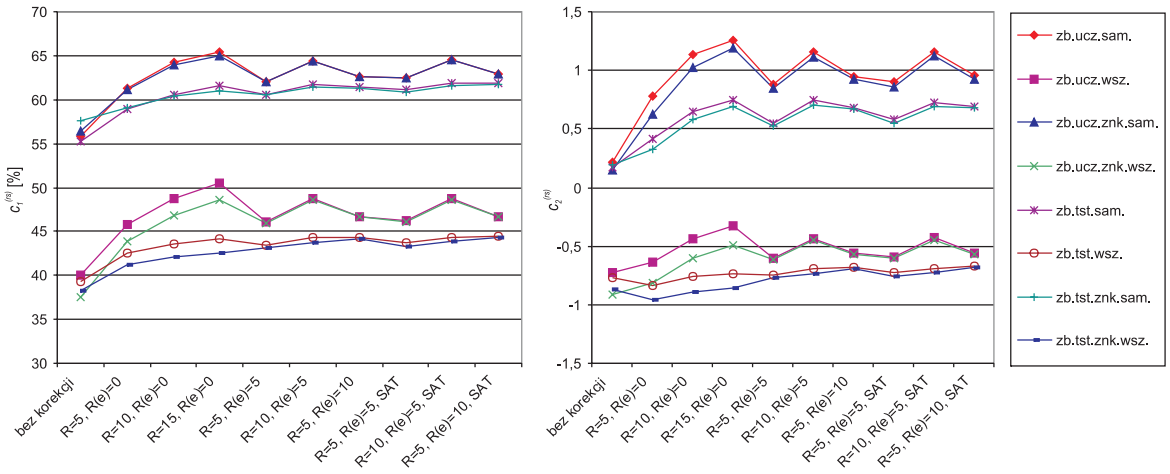
parametry			zb.ucz.		zb.ucz.znk.		zb.tst.		zb.tst.znk.	
R	$R^{(e)}$	SAT	sam.	wsz.	sam.	wsz.	sam.	wsz.	sam.	wsz.
bez kompensacji			55.82	40.00	56.39	37.49	55.27	39.24	57.53	38.24
5	0	n	61.24	45.78	61.14	43.93	59.00	42.48	59.14	41.20
10	0	n	64.30	48.78	63.97	46.85	60.59	43.63	60.38	42.03
15	0	n	**65.36**	**50.50**	**64.91**	48.54	61.53	44.10	61.04	42.52
5	5	n	62.06	46.13	62.02	46.00	60.49	43.45	60.49	43.15
10	5	n	64.41	48.79	64.42	48.55	61.72	**44.41**	61.37	43.79
5	10	n	62.62	46.61	62.62	46.65	61.40	44.27	61.24	44.18
5	5	t	62.45	46.24	62.54	46.09	61.14	43.71	60.89	43.30
10	5	t	64.51	48.79	64.56	**48.56**	**61.88**	44.38	61.60	43.81
5	10	t	62.85	46.68	62.93	46.66	61.83	44.40	**61.76**	**44.31**

Tab. 3.2. Wyniki rozpoznawalności izolowanych ramek dla zmodyfikowanej metody EV. Zastosowano miarę oceny $c_2^{(rs)}$. Czcionką pogrubioną zaznaczano najwyższy wynik w danej kolumnie.

parametry			zb.ucz.		zb.ucz.znk.		zb.tst.		zb.tst.znk.	
R	$R^{(e)}$	SAT	sam.	wsz.	sam.	wsz.	sam.	wsz.	sam.	wsz.
bez kompensacji			0.216	-0.722	0.150	-0.914	0.169	-0.764	0.191	-0.869
5	0	n	0.783	-0.634	0.630	-0.817	0.417	-0.834	0.322	-0.956
10	0	n	1.130	-0.438	1.028	-0.599	0.645	-0.760	0.586	-0.888
15	0	n	**1.256**	**-0.331**	**1.186**	-0.489	**0.749**	-0.735	0.689	-0.860
5	5	n	0.879	-0.599	0.844	-0.614	0.547	-0.746	0.524	-0.774
10	5	n	1.155	-0.432	1.109	-0.452	0.744	-0.694	0.707	-0.731
5	10	n	0.949	-0.564	0.924	-0.572	0.680	-0.679	0.666	-0.688
5	5	t	0.9	-0.593	0.860	-0.608	0.578	-0.730	0.574	-0.761
10	5	t	1.161	-0.430	1.121	**-0.447**	0.720	-0.697	0.696	-0.728
5	10	t	0.955	-0.563	0.926	-0.572	0.694	**-0.672**	**0.679**	**-0.681**

Analizując wyniki można zauważyć, że w przypadku stosowania niezmodyfikowanego algorytmu EV, poprawa rozpoznawalności dla zbioru testowego jest ok. dwukrotnie niższa niż dla zbioru uczącego. Wskazuje to na zbyt małe uogólnienie modelu na zbiór testowy, co wynika z kolei ze zbyt małej liczności zbioru uczącego.

Modyfikacja algorytmu EV, mająca na celu umożliwienie kompensacji zniekształ-



Rys. 3.1. Wyniki rozpoznawalności izolowanych ramek dla zmodyfikowanej metody EV.

ceń liniowych, okazała się skuteczna. W przypadku symulowania zniekształceń transmisyjnych najwyższe poprawy uzyskano (z wyjątkiem rozpoznawalności samogłosek w zbiorze uczącym) dla algorytmu zmodyfikowanego.

Uczenie modelu statystycznego z wykorzystaniem zaproponowanej strategii SAT również przyniosło pozytywny rezultat. Wyniki rozpoznawalności po zastosowaniu SAT były w większości przypadków wyższe niż wyniki uzyskane bez SAT.

Wyniki uzyskane dla metody EV pozwoliły na sprecyzowanie pewnych wymagań (zob. rozdział 4.1), które spełniać powinien projektowany oryginalny algorytm kompensacji.

4. Metoda banków transformacji widma

W tym rozdziale zaprezentowano zaprojektowaną oryginalną metodę kompensacji liniowych zniekształceń transmisyjnych i cech osobniczych mówcy, w której wykorzystano banki transformacji widma sygnału mowy. Opisano postać transformacji widma, metodę optymalizacji wartości jej parametrów i metodę konstrukcji banków transformacji. Podano również sposób implementacji algorytmu kompensacji w systemie ARM.

4.1. Założenia

Uwzględniając analizę znanych z literatury algorytmów kompensacji oraz wyniki uzyskane za pomocą algorytmu EV (zob. rozdział 3.2), postawiono następujące założenia dla projektowanej metody kompensacji, będące uzupełnieniem założeń podanych w rozdziale 1.2:

- Kompensacja przeprowadzana będzie w dziedzinie widma amplitudowego sygnału i składać się będzie z dwóch etapów. W pierwszym przeprowadzane będzie skalowanie osi częstotliwości, mające na celu kompensację różnic osobniczych, związanych z przesunięciami częstotliwości formantów (zob. rozdział 2.3 i dodatek A). W drugim wykonywana będzie filtracja liniowa, której celem będzie normalizacja liniowych zniekształceń transmisyjnych oraz różnic w widmie spowodowanych cechami osobniczymi. Wykorzystanie widma powoduje, że parametry kompensacji mają jasny sens fizyczny, co ułatwia ich analizę oraz projektowanie algorytmów wykonujących na nich operacje takie, jak np. interpolacja czy podział na klasy. Ponadto zastosowanie modyfikacji widma nadaje metodzie kompensacji dużą uniwersalność, gdyż wyznaczanie widma amplitudowego jest wspólnym etapem parametryzacji sygnału mowy dla większości współczesnych systemów ARM.

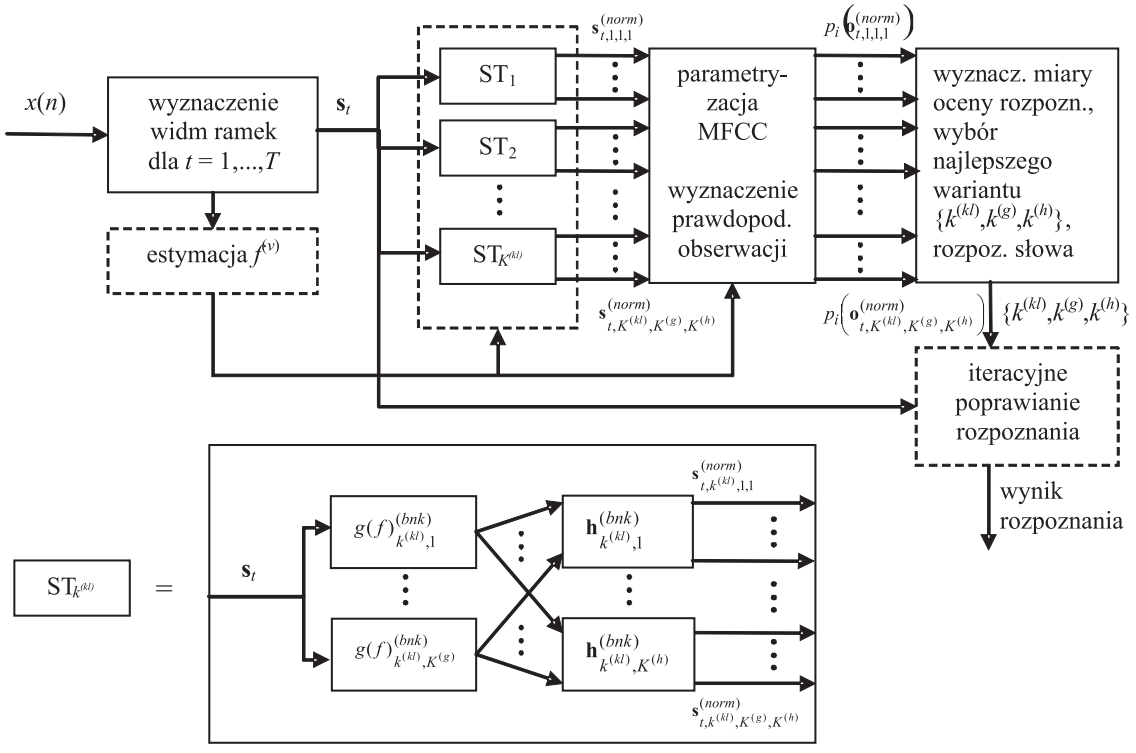
- Zamiast strategii iteracyjnego wyznaczania wartości parametrów kompensacji dla danego mówcy, zastosowana zostanie strategia równoległa z wykorzystaniem banków transformacji widma. Podejście to minimalizuje negatywne

efekty związane ze złą inicjalizacją algorytmów iteracyjnych. Ograniczona iteracyjna reestymacja wartości parametrów przeprowadzona będzie dopiero w końcowej fazie działania algorytmu kompensacji.

- Zastosowany zostanie podział mówców na klasy i strategia SAT uczenia statystycznych modeli języka, w celu zwiększenia zdolności klasyfikacji rozkładów prawdopodobieństwa wchodzących w skład tych modeli.

- Stosowana będzie jedna transformacja widma dla wszystkich fonemów. Wykorzystanie zależności międzyfonemowych w algorytmach kompensacji daje pozytywne rezultaty dla zbioru testowego w przypadku, gdy uczenie algorytmu przeprowadzane jest z zastosowaniem zbioru uczącego o dużej liczności. W sytuacji, kiedy zbiór uczący jest niewielki, zadowalające uogólnienie własności metody na zbiór testowy spodziewane jest dla metod nie wykorzystujących zależności międzyfonemowych.

- Projektowana metoda ma zapewniać osiągniecie rozpoznawalności izolowanych ramek w zbiorze uczącym i testowym nie gorszej niż dla algorytmu EV, dla zbliżonej liczby parametrów kompensacji.

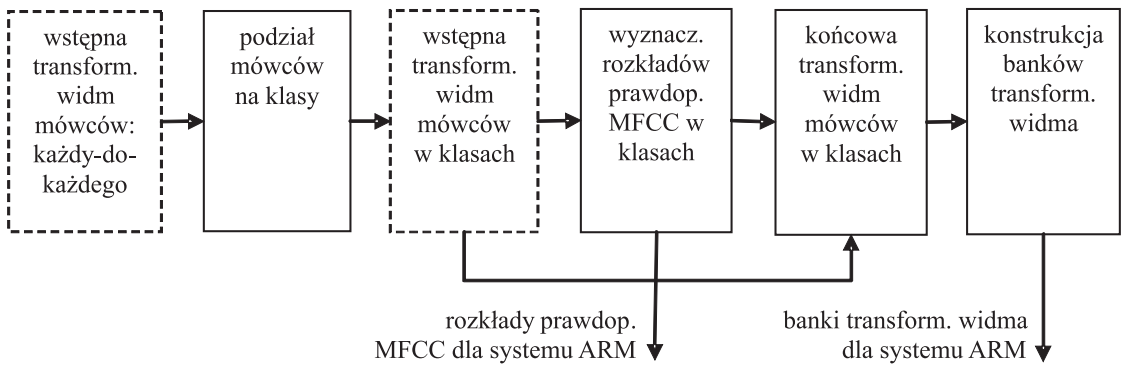


Rys. 4.1. Schemat metody kompensacji liniowych zniekształceń transmisyjnych i cech osobniczych mówcy z zastosowaniem banków transformacji widma.

4.2. Ogólny schemat metody

Ogólny schemat zaproponowanej metody kompensacji przedstawiono na rys. 4.1. Dyskretne widma amplitudowe $\mathbf{s}_t$ ramek sygnału wejściowego $x(n)$ poddawane są normalizacji w bankach transformacji widma ST. Na transformacje w danym banku składa się skalowanie osi częstotliwości liczącym $K^{(g)}$ elementów bankiem funkcji $g(f)^{(bnk)}$ i filtracja liniowa liczącym $K^{(h)}$ elementów bankiem filtrów liniowych o charakterystykach amplitudowych $\mathbf{h}^{(\mathbf{bnk})}$. Liczba transformacji widma dla każdego banku wynosi zatem $K^{(g)} \cdot K^{(h)}$, a liczba wszystkich transformacji $K^{(g)} \cdot K^{(h)} \cdot K^{(kl)}$, ponieważ każda z $K^{(kl)}$ klas mówców ma inny bank ST. Następnie, na podstawie znormalizowanych widm $\mathbf{s}^{(norm)}_{t,k^{(kl)},k^{(g)},k^{(h)}}$, wyznaczane są wektory współczynników MFCC $\mathbf{o}^{(norm)}_{t,k^{(kl)},k^{(g)},k^{(h)}}$. Prawdopodobieństwa obserwacji stanów modelu HMM $p_i\left(\mathbf{o}^{(norm)}_{t,k^{(kl)},k^{(g)},k^{(h)}}\right)$ obliczane są na podstawie osobnych dla każdej klasy rozkładów prawdopodobieństwa współczynników MFCC. Kolejnym etapem jest rozpoznanie wypowiedzi, przy czym wybierana jest ta kombinacja $\left\{k^{(kl)}, k^{(g)}, k^{(h)}\right\}$, która zapewnia najwyższą wartość miary oceny rozpoznania zdefiniowanej w systemie. W celu poprawy wyniku rozpoznania zastosowano dodatkowo iteracyjny algorytm zwiększający różnorodność stosowanych funkcji $g(f)$ i charakterystyk filtrów liniowych w ramach wcześniej wybranego wariantu $\left\{k^{(kl)}, k^{(g)}, k^{(h)}\right\}$. Ponadto zastosowano estymację wartości częstotliwości tonu krtaniowego $f^{(v)}$ w celu zredukowania nakładu obliczeniowego metody, dzięki wstępnemu przyporządkowaniu mówcy do klasy na podstawie wartości $f^{(v)}$.

Na rys. 4.2 przedstawiono ogólny schemat metody konstrukcji banków transformacji widma i metody wyznaczania klas mówców. Przed podziałem na klasy opcjonalnie uwzględniane są transformacje widm mówców, wyznaczane dla każdego mówcy ze zbioru uczącego do każdego innego mówcy z tego zbioru. Po wyznaczeniu klas mówców opcjonalnie wykonywane są transformacje mające na celu zwiększenie zdolności klasyfikacji wyznaczanych w kroku kolejnym rozkładów prawdopodobień-



Rys. 4.2. Schemat metody wyznaczania klas mówców i konstrukcji banków transformacji widma.

stwa współczynników MFCC w klasach. Wartości parametrów tych transformacji służą również jako jedne z parametrów początkowych w procesie wyznaczania końcowych transformacji widm mówców w klasach. Na bazie zbiorów wartości parametrów transformacji końcowych konstruowane są banki transformacji widma.

4.3. Transformacja widma

Poniżej opisano postać dwuetapowej transformacji dyskretnego widma amplitudowego sygnału mowy, parametry charakteryzujące tę transformację oraz metodę optymalizacji wartości parametrów dla danego mówcy.

4.3.1. Postać i parametry transformacji

Przyjęto, że funkcja skalowania osi częstotliwości $f^{(norm)} = g\left(f, \boldsymbol{\alpha}^{(g)}\right)$ jest odcinkami liniowa i charakteryzowana przez wektor parametrów $\boldsymbol{\alpha}^{(g)}$, określający współrzędne punktów łączenia odcinków (rys. 4.3a). Wektor ten można podzielić na wektory $\boldsymbol{\alpha}^{(g,r)}$ i $\boldsymbol{\alpha}^{(g,o)}$, zawierające odpowiednio rzędne i odcięte tych punktów. Wartości współrzędnych ograniczone są tak, by funkcja skalująca była rosnąca. Dodatkowo funkcja spełnia warunki $g(0) = 0$ i $g\left(f^{(max)}\right) = f^{(max)}$, gdzie $f^{(max)}$ to górna granica częstotliwości analizowanego widma. Wartości rzędnych punktów łączenia odcinków są stałe i w implementacji przyjęto $\alpha_1^{(g,r)} = 1.4$ kHz, $\alpha_2^{(g,r)} = 2.3$ kHz i $\alpha_3^{(g,r)} = 4.1$ kHz. Częstotliwości te odpowiadają w przybliżeniu częstotliwościom środkowym 12., 16. i 20. filtru melowego. Wartości odciętych są natomiast wyznaczane na drodze optymalizacji.

Zaproponowana postać funkcji skalującej zapewnia różnowartościowe odwzorowanie osi częstotliwości widma wejściowego w oś częstotliwości widma znormalizowanego. Transformacja widma amplitudowego $\mathbf{s}$ w widmo z przeskalowaną osią częstotliwości $\mathbf{s}^{(skal)}$ przeprowadzana jest następująco:

Dla każdego prążka $n = 0, \ldots, N/2 - 1$ widma $\mathbf{s}^{(skal)}$ oblicz:

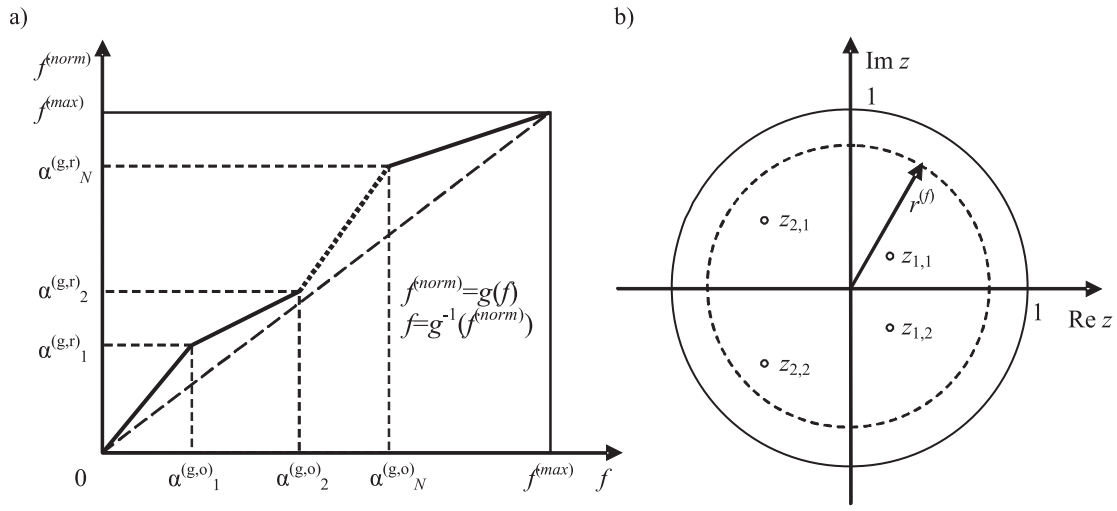
$$n^{(skal,l)} = g^{-1}\left((n - 0.5) \cdot \frac{f^{(p)}}{N}, \boldsymbol{\alpha}^{(g)}\right) \cdot \frac{N}{f^{(p)}} \tag{4.1}$$

$$n^{(skal,p)} = g^{-1}\left((n + 0.5) \cdot \frac{f^{(p)}}{N}, \boldsymbol{\alpha}^{(g)}\right) \cdot \frac{N}{f^{(p)}} \tag{4.2}$$

$$s_n^{(skal)} = \begin{cases} \Big(s_{\left[n^{(skal,l)}\right]} \cdot \left(0.5 + \left[n^{(skal,l)}\right] - n^{(skal,l)}\right) + \\ \quad + s_{\left[n^{(skal,p)}\right]} \cdot \left(0.5 - \left[n^{(skal,p)}\right] + n^{(skal,p)}\right) + \\ \quad + \sum_{i=\left[n^{(skal,l)}\right]+1}^{\left[n^{(skal,p)}\right]-1} s_i \Big) \frac{1}{n^{(skal,p)} - n^{(skal,l)}}, & \text{dla } \left[n^{(skal,l)}\right] \neq \left[n^{(skal,p)}\right] \\ s_{\left[n^{(skal,l)}\right]}, & \text{dla } \left[n^{(skal,l)}\right] = \left[n^{(skal,p)}\right] \end{cases} \tag{4.3}$$

gdzie N oznacza liczbę prążków widma $\mathbf{s}$, a $[\cdot]$ - operację zaokrąglania do najbliższej liczby całkowitej.

Przedstawiona powyżej transformacja ma tę istotną cechę, że podczas kompresji lub ekspansji odcinków osi częstotliwości nie następuje zmiana amplitudy przesuwanych fragmentów widma. Takie podejście umotywowane jest tym, że skalowanie osi częstotliwości ma na celu przesunięcie częstotliwości formantów, a za zmianę ich amplitudy odpowiedzialny jest kolejny etap transformacji - filtracja liniowa.



Rys. 4.3. Elementy transformacji widma: a) funkcja skalowania osi częstotliwości, b) przykładowy układ zer funkcji transmitancji filtru.

W etapie drugim zastosowano filtr o skończonej odpowiedzi impulsowej (FIR) niskiego rzędu. W algorytmie przeprowadzana jest modyfikacja widma amplitudowego, zatem charakterystyka fazowa filtru nie ma znaczenia. Wystarczające jest więc zastosowanie filtru minimalnofazowego. Na położenie zer funkcji transmitancji nałożono ograniczenie, aby nie leżały zbyt blisko okręgu jednostkowego, co zapobiega zbyt dużemu lokalnemu tłumieniu. W implementacji przyjęto, że liczba zer wynosi 4, a ich maksymalny promień $r^{(f)} = 0.8$. Promienie i kąty określające położenie zer ustalane są na drodze optymalizacji. Przykładowy układ zer filtru na płaszczyźnie $\mathscr{Z}$ pokazano na rys. 4.3b. Filtracja odbywa się w dziedzinie widma dyskretnego poprzez mnożenie widma amplitudowego sygnału i charakterystyki amplitudowej

filtru. Rząd filtru jest tutaj na tyle mały, że nie występują istotne różnice między splotem kołowym, któremu odpowiada mnożenie w dziedzinie widma dyskretnego, a splotem liniowym, stosowanym w filtracji w dziedzinie czasu.

Filtr liniowy opisywany może być, w zależności od potrzeb, różnymi zestawami parametrów:

- Parametry stosowane w optymalizacji. Każdej parze zer funkcji transmitancji filtru $(z_{p,1}, z_{p,2})$ odpowiada w przestrzeni optymalizacji para liczb rzeczywistych $(x_{p,1}, x_{p,2}) \in \mathscr{R}^2$. Transformacja z $(x_{p,1}, x_{p,2})$ do $(z_{p,1}, z_{p,2})$ dana jest zależnościami:

$$y_p = \frac{r^{(f)}}{1 + \exp\left(-x_{p,1}\right)} \tag{4.4}$$

$$\text{jeżeli } 0 \leq x_{p,2} \leq \pi, \text{ to } z_{p,1} = y_p \cdot \exp\left(jx_{p,2}\right) \tag{4.5}$$

$$z_{p,2} = y_p \cdot \exp\left(-jx_{p,2}\right) \tag{4.6}$$

$$\text{jeżeli } x_{p,2} < 0, \text{ to } z_{p,1} = y_p \tag{4.7}$$

$$z_{p,2} = z_{p,1} + \left(\exp\left(x_{p,2}\right) - 1\right)\left(z_{p,1} - r^{(f)}\right) \tag{4.8}$$

$$\text{jeżeli } x_{p,2} > \pi, \text{ to } z_{p,1} = -y_p \tag{4.9}$$

$$z_{p,2} = z_{p,1} - \left(\exp\left(\pi - x_{p,2}\right) - 1\right)\left(r^{(f)} - z_{p,1}\right) \tag{4.10}$$

- Współczynniki równania różnicowego. Parametry te wykorzystywane są do przechowywania filtrów w pamięci komputera oraz jako parametry pośrednie.

  Mając danych P zer zespolonych z_p filtru, $P+1$ współczynniki równania różnicowego $a_i^{(f)}$ wyznacza się z równań:

$$\frac{\prod_{p=0}^{P-1}(z - z_p)}{z^P} = \sum_{i=0}^{P} a_i^{(f,m)} z^{-i} \tag{4.11}$$

$$a_i^{(f)} = \frac{a_i^{(f,m)}}{\sqrt{\sum_{i=0}^{P}\left(a_i^{(f,m)}\right)^2}} \tag{4.12}$$

  Równanie (4.12) ma na celu normalizację współczynników filtru tak, by kwadrat jego charakterystyki amplitudowej całkował się do jedności (w przypadku osi częstotliwości unormowanej tak, że $f^{(p)} = 1$). Taki filtr ma tę własność, że nie zmienia energii filtrowanego nim szumu białego.

Wyznaczanie współczynników $a_i^{(f)}$ z dyskretnej charakterystyki amplitudowej $\mathbf{h}$ dokonywane jest w następujący sposób:

$$\mathbf{h}^{(ceps)} = \mathrm{IDFT}\left(\ln \mathbf{h}\right) \tag{4.13}$$

$$h_n^{(ceps,m)} = \begin{cases} h_n^{(ceps)}, & \text{dla } n = 0 \\ 2 \cdot h_n^{(ceps)}, & \text{dla } n = 1, \ldots, N/2 - 1 \\ 0, & \text{dla } n = N/2, \ldots, N - 1 \end{cases} \tag{4.14}$$

$$\mathbf{h}^{(zesp)} = \exp\left(\mathrm{Re}\left(\mathrm{DFT}\left(\mathbf{h}^{(ceps,m)}\right)\right)\right) \cdot \exp\left(j \cdot \mathrm{Im}\left(\mathrm{DFT}\left(\mathbf{h}^{(ceps,m)}\right)\right)\right) \tag{4.15}$$

$$\mathbf{a}^{(f)} = \left(\mathbf{F}^T\mathbf{F}\right)^{-1}\mathbf{F}^T\mathbf{h}^{(zesp)} \tag{4.16}$$

gdzie N oznacza długość transformaty DFT, a $\mathbf{F}$ jest macierzą transformacji Fouriera o elementach równych $F_{mn} = \exp\left(-2\pi jmn/N\right)$.

W powyższej metodzie odtwarzane jest widmo zespolone $\mathbf{h}^{(zesp)}$ na podstawie widma amplitudowego $\mathbf{h}$, przy wykorzystaniu własności mówiącej, że cepstrum o wartościach zerowych dla ujemnych punktów na osi quefrency odpowiada ciągowi minimalnofazowemu [122]. Błędy związane z dyskretną analizą częstotliwościową, a więc nakładanie się na siebie nieskończonych ciągów cepstralnych, można tutaj zaniedbać. Współczynniki filtru wyznaczane są metodą najmniejszych kwadratów (4.16).

- Dyskretna charakterystyka amplitudowa $\mathbf{h}$ jest stosowana w pewnych operacjach algorytmu kompensacji tam, gdzie ważne jest posiadanie przez parametry jasnego sensu fizycznego lub przeprowadzana jest filtracja w dziedzinie widma. Charakterystykę $\mathbf{h}$ wyznacza się ze współczynników $a_i^{(f)}$ za pomocą DFT, po uprzednim uzupełnieniu wektora współczynników zerami do długości równej liczbie prążków widma ramki sygnału. Ma to na celu uzyskanie charakterystyki o liczbie prążków równej liczbie prążków widma ramki sygnału, co umożliwia przeprowadzenie filtracji w dziedzinie widma.

Dwa etapy transformacji widma można zrealizować w jednym kroku jako działanie macierzowe:

$$\mathbf{s}^{(norm)} = \mathrm{diag}\left(\mathbf{h}\right) \cdot \mathbf{G} \cdot \mathbf{s} \tag{4.17}$$

gdzie $\mathbf{G}$ jest macierzą skalowania osi częstotliwości, wyznaczoną z wykorzystaniem równania (4.3), a $\mathrm{diag}\left(\mathbf{h}\right)$ oznacza macierz przekątniową z elementami na przekątnej równymi elementom $\mathbf{h}$.

4.3.2. Optymalizacja wartości parametrów transformacji dla danego mówcy

Celem optymalizacji wartości parametrów transformacji widma jest maksymalizacja rozpoznawalności izolowanych ramek sygnału danego mówcy. Podejście takie zapewnia niezależność uzyskanych transformacji od zawartości słownika systemu.

We wstępnych badaniach nad metodą optymalizacji zaproponowano, by wartości te wyznaczać poprzez porównanie macierzy transformacji parametrów sygnału, uzyskanej na podstawie parametrów transformacji widma, z macierzą uzyskaną metodą MLLR. Metoda ta nie przyniosła zadowalających rezultatów, do czego przyczyniły się czynniki takie, jak inna struktura porównywanych macierzy i zbyt daleko idące przybliżenia podczas wyznaczania macierzy transformacji parametrów sygnału.

Inną wstępną propozycją było porównywanie rozkładów prawdopodobieństwa parametrów sygnału mówcy odniesienia z rozkładami dla mówcy transformowanego, wyznaczanymi za pomocą odpowiednich transformacji, przy znanych wartościach parametrów transformacji widma. Również ta metoda okazała się nieskuteczna z powodu zbyt dużych przybliżeń zastosowanych przy transformacji rozkładów prawdopodobieństwa i zastosowania w optymalizacji miary podobieństwa między rozkładami nie prowadzącej do wzrostu rozpoznawalności izolowanych ramek.

Rozwiązaniem skutecznym okazała się natomiast optymalizacja funkcji celu ściśle związanych z rozpoznawalnością izolowanych ramek. Punktem wyjścia do zdefiniowania tych funkcji były następujące dwie miary błędu rozpoznania izolowanych ramek dla fonemu i danego mówcy s:

$$c_1^{(br)}\left(s,i,\boldsymbol{\alpha}^{(st)}\right)=\int\limits_{\mathscr{R}^{\dim(\mathbf{o})}} u\left(\ln p_i\left(\mathbf{o}\right)-\max_{\substack{0\le j\le I-1\\ j\neq i}}\ln p_j\left(\mathbf{o}\right)\right)\cdot p_{i,s}^{(mw)}\left(\mathbf{o},\boldsymbol{\alpha}^{(st)}\right)d\mathbf{o} \quad (4.18)$$

$$c_2^{(br)}\left(s,i,\boldsymbol{\alpha}^{(st)}\right)=\int\limits_{\mathscr{R}^{\dim(\mathbf{o})}} \left(\ln p_i\left(\mathbf{o}\right)-\max_{\substack{0\le j\le I-1\\ j\neq i}}\ln p_j\left(\mathbf{o}\right)\right)\cdot p_{i,s}^{(mw)}\left(\mathbf{o},\boldsymbol{\alpha}^{(st)}\right)d\mathbf{o} \quad (4.19)$$

gdzie $\boldsymbol{\alpha}^{(st)}$ oznacza wektor parametrów transformacji widma, na który składają się parametry skalowania osi częstotliwości i parametru filtru liniowego, a u oznacza funkcję skoku jednostkowego. Miara $c_1^{(br)}$ ma na celu wskazanie procentowego udziału ramek, który został prawidłowo przyporządkowany do odpowiadających im fonemów. Miara $c_2^{(br)}$ natomiast uwzględnia również dynamikę błędu klasyfikacji, który wyrażony jest różnicą logarytmu prawdopodobieństwa przynależności danej ramki do odpowiadającego jej fonemu i maksymalnej wartości logarytmu prawdopodobieństwa przynależności ramki do pozostałych fonemów.

Rozpoznawanie izolowanych ramek dokonywane jest na podstawie rozkładów prawdopodobieństwa współczynników MFCC $p_i\left(\mathbf{o}\right)$, nazywanych dalej rozkładami odniesienia. Wyznaczanie wartości parametrów transformacji dokonywane jest w

kilku miejscach algorytmu i w zależności od tego miejsca rozkłady odniesienia mogą być rozkładami dla klasy mówców bądź rozkładami indywidualnymi pewnego mówcy. Rozkłady te są ustalone i nie zależą od bieżących wartości parametrów transformacji widma. Zależą od nich natomiast rozkłady $p_{i,s}^{(mw)}\left(\mathbf{o}, \boldsymbol{\alpha}^{(st)}\right)$, będące bieżącymi rozkładami współczynników MFCC dla fonemu i mówcy s.

W funkcji celu optymalizacji stosowano wartości rozpoznawalności dla zbioru fonemów $\mathscr{J}^{(sm)}$ zawierającego samogłoski i zbioru $\mathscr{J}^{(sp)}$ zawierającego spółgłoski. Nieobciążone i zgodne [130] estymatory wartości miar (4.18) i (4.19), uśrednionych po odpowiednich zbiorach fonemów, dane są zależnością (E.3) (zob. dodatek E). Założono, że maksymalizowana będzie rozpoznawalność samogłosek przy nałożonym ograniczeniu zapobiegającym równoczesnemu jej spadkowi dla spółgłosek. Postąpiono tak, gdyż prawidłowe rozpoznawanie samogłosek jest kluczowe w zadaniu rozpoznawania krótkich wypowiedzi z małego słownika. W konsekwencji zdefiniowano dwa warianty funkcji celu optymalizacji wartości parametrów transformacji widma dla danego mówcy s:

$$c_1^{(ro)}\left(s, \boldsymbol{\alpha}^{(st)}\right) = \begin{cases} c_{1,s,\mathscr{J}^{(sm)}}^{(rm)}, & \text{dla } c_{1,s,\mathscr{J}^{(sp)}}^{(rm)} \geq c_{1,s,\mathscr{J}^{(sp)}}^{(rm,p)} \\ c_{1,s,\mathscr{J}^{(sm)}}^{(rm)} \cdot \beta^{(fk1)^{\left(c_{1,s,\mathscr{J}^{(sp)}}^{(rm,p)} - c_{1,s,\mathscr{J}^{(sp)}}^{(rm)}\right)}}, & \text{dla } c_{1,s,\mathscr{J}^{(sp)}}^{(rm)} < c_{1,s,\mathscr{J}^{(sp)}}^{(rm,p)} \end{cases} \tag{4.20}$$

$$c_2^{(ro)}\left(s, \boldsymbol{\alpha}^{(st)}\right) = \begin{cases} c_{2,s,\mathscr{J}^{(sm)}}^{(rm)}, & \text{dla } c_{2,s,\mathscr{J}^{(sp)}}^{(rm)} \geq c_{2,s,\mathscr{J}^{(sp)}}^{(rm,p)} \\ c_{2,s,\mathscr{J}^{(sm)}}^{(rm)} - \beta^{(fk2)} \cdot \left(c_{2,s,\mathscr{J}^{(sp)}}^{(rm,p)} - c_{2,s,\mathscr{J}^{(sp)}}^{(rm)}\right), & \text{dla } c_{2,s,\mathscr{J}^{(sp)}}^{(rm)} < c_{2,s,\mathscr{J}^{(sp)}}^{(rm,p)} \end{cases} \tag{4.21}$$

gdzie $c_{1,s,\mathscr{J}^{(sp)}}^{(rm,p)}$ i $c_{2,s,\mathscr{J}^{(sp)}}^{(rm,p)}$ oznaczają wartości wyznaczone na początku optymalizacji dla pewnych początkowych wartości parametrów transformacji. W implementacji przyjęto wartości współczynników $\beta^{(fk1)} = 0.87$ i $\beta^{(fk2)} = 2$. W powyższych wzorach wpływ wartości parametrów $\boldsymbol{\alpha}^{(st)}$ na funkcje celu nie jest podany wprost. Wpływ ten uwidacznia się w zmianach wartości estymatorów $c^{(rm)}$, które zależą od bieżących wartości parametrów $\boldsymbol{\alpha}^{(st)}$, co związane jest z kolei ze zmianami współczynników MFCC ramek (na skutek transformacji widma), wykorzystywanych przy obliczeniach wartości $c^{(rm)}$.

Rozważano dwie metody optymalizacji: stochastyczną i deterministyczną. Z uwagi na fakt, że funkcje celu są estymatorami i ich wartości zależą od wylosowanego zbioru ramek, wskazane wydaje się zastosowanie optymalizacji stochastycznej [59, 154, 153], w której zbiór ramek losowany jest w każdej iteracji, a algorytm ma za zadanie znalezienie maksimum wartości oczekiwanej funkcji celu. Przeprowadzono wstępne eksperymenty za pomocą algorytmu SPSA (*Simultaneous Perturbation Stochastic Approximation*) [154], lecz ich wyniki okazały się niezadowalające.

Ostatecznie wybrano deterministyczną metodę optymalizacji, w której zastosowano losowanie jednego zbioru ramek na początku optymalizacji, dostatecznie licznego, by błąd estymacji był dużo mniejszy od uzyskanej poprawy rozpoznawalności oraz by uzyskać dobre uogólnienie, tj. przy innych wylosowanych zbiorach ramek uzyskana poprawa była zachowana. Ze względu na charakter zastosowanych funkcji celu, które mogą być nieciągłe, obszarami stałe i posiadać wiele lokalnych maksimów, zaproponowano dwuetapowy algorytm optymalizacji. W pierwszym etapie stosowany jest algorytm ewolucyjny, którego zadaniem jest znalezienie obszaru, w którym znajduje się globalne maksimum. W drugim etapie używana jest metoda simpleksu Neldera-Meada, której celem jest znalezienie dokładnego położenia maksimum. Szczegóły algorytmu przedstawione są w dodatku F.

4.4. Podział mówców na klasy i wyznaczanie rozkładów prawdopodobieństwa współczynników MFCC w klasach.

Rozpoznawanie mowy, nawet przy zastosowaniu algorytmów kompensacji cech osobniczych i zniekształceń transmisyjnych, jest mało efektywne w przypadku wykorzystywania rozkładów prawdopodobieństwa współczynników MFCC wspólnych dla wszystkich mówców. Wynika to z ich zmniejszonej zdolności klasyfikacji na skutek znacznego rozrzutu wartości współczynników MFCC, związanego ze zmiennością międzyosobniczą. Z kolei użycie jednego rozkładu, pochodzącego od pewnego wybranego mówcy, skutkuje mało efektywnym rozpoznawaniem mowy innych mówców, gdyż algorytmy kompensacji dla krótkich wypowiedzi nie umożliwiają dostatecznie szczegółowej normalizacji współczynników MFCC. Celowe jest zatem przyjęcie rozwiązania pośredniego i zastosowanie rozkładów dla klas mówców.

Przyjmijmy, że w zbiorze uczącym znajduje się N mówców i znane są wartości miary podobieństwa między mówcami m i n, oznaczone jako $d_{mn}^{(mw)}$. Metody wyznaczania tych wartości opisano poniżej. Znana jest też liczba klas $K^{(kl)}$ ($K^{(kl)} < N$), do których przypisani zostaną mówcy. W celu podziału mówców na klasy, z których w każdej określony jest mówca centralny klasy $\nu_k^{(c)}$, dla każdej kombinacji $\omega : \left\{\nu_{\omega,0}^{(c)}, \ldots, \nu_{\omega,k}^{(c)}, \ldots, \nu_{\omega,K^{(kl)}-1}^{(c)}\right\}$, gdzie $\nu_{\omega,k}^{(c)}$ wybierani są spośród N mówców, obliczana jest wartość

$$d_{\omega}^{(mw,sum)} = \sum_{n=0}^{N-1} \left\{ d_{mn}^{(mw)} : m = \underset{k=0\ldots K^{(kl)}-1}{\arg\max} d_{\nu_{\omega,k}^{(c)} n}^{(mw)} \right\} \tag{4.22}$$

a następnie znajduje numer kombinacji optymalnej $\omega^{(opt)}$, dla której wartość $d_{\omega}^{(mw,sum)}$ była największa. Jako mówców centralnych klas przyjmuje się zbiór $\left\{\nu_{\omega^{(opt)},0}^{(c)}, \ldots, \nu_{\omega^{(opt)},K^{(kl)}-1}^{(c)}\right\}$ oraz przyporządkowuje pozostałych mówców do klas tak, że dany mówca należy do tej klasy, dla której ma największą wartość $d^{(mw)}$ do

jej mówcy centralnego. Opisany algorytm ma na celu wybranie mówców centralnych klas mówiących „średnio", czyli takich, dla których sumaryczna miara podobieństwa do innych mówców danej klasy jest wysoka.

4.4.1. Wariant 1. metody wyznaczania klas mówców

Zaproponowano miarę podobieństwa między mówcami w postaci:

$$d_{mn}^{(mw)} = \ln\left(\sum_{i=0}^{I-1} \frac{d_{mn,i}^{(mw,n)}}{\sqrt{d_{mm,i}^{(mw,n)} \cdot d_{nn,i}^{(mw,n)}}}\right) \tag{4.23}$$

$$d_{mn,i}^{(mw,n)} = \int\limits_{\mathscr{R}^{\dim(\mathbf{o})}} p_{i,m}^{(mw)}(\mathbf{o}) \cdot p_{i,n}^{(mw)}(\mathbf{o})\, d\mathbf{o} \tag{4.24}$$

gdzie $p_{i,n}^{(mw)}$ oznacza łączny rozkład prawdopodobieństwa współczynników MFCC dla fonemu i dla mówcy n. W implementacji stosowano rozkłady brzegowe, a te modelowane były z kolei metodą GMM. W takim przypadku równanie (4.24) ma rozwiązanie analityczne (zob. równania C.44 i C.45). W obliczeniach uwzględniano wszystkie fonemy. Zaproponowana miara podobieństwa bazuje na unormowanej korelacji między rozkładami prawdopodobieństwa współczynników MFCC odpowiadających sobie fonemów porównywanych mówców, nie uwzględnia zdolności klasyfikacyjnych tych rozkładów. Przyjmuje wartości z przedziału $(-\infty; 0]$, przy czym wyższa wartość oznacza większe podobieństwo.

Po określeniu klas mówców wyznaczono dla każdej klasy uśrednione rozkłady prawdopodobieństwa współczynników MFCC dla systemu ARM.

4.4.2. Wariant 2. metody wyznaczania klas mówców

Klasy mówców wyznaczano tak, jak w wariancie 1. Następnie przeprowadzano wstępne optymalizacje wartości parametrów transformacji widma wszystkich mówców danej klasy przy rozkładach odniesienia równych rozkładom dla mówcy centralnego klasy. W optymalizacjach tych zastosowano funkcję celu $c_1^{(ro)}$, a wartości początkowe $c_{1,s,\mathscr{J}^{(sp)}}^{(rm,p)}$ wyznaczono dla neutralnych wartości parametrów transformacji widma, tj. przy braku jego modyfikacji.

Rozkłady prawdopodobieństwa współczynników MFCC dla klas wyznaczano uwzględniając uzyskane wstępne transformacje widma. Takie podejście jest wariantem metody SAT uczenia systemu.

4.4.3. Wariant 3. metody wyznaczania klas mówców

Zaproponowano podaną niżej miarę podobieństwa między mówcami, która uwzględnia nie tylko podobieństwo między tymi samymi fonemami u obu porównywanych mówców, ale również podobieństwo do fonemów pozostałych. Dzięki temu brane są pod uwagę zdolności klasyfikacyjne rozkładów prawdopodobieństwa.

$$d_{mn}^{(mw)} = -\sum_{i=0}^{I-1} \frac{b_{mn,i}^{(mw)}}{\sum\limits_{\substack{j=0 \\ j\neq i}}^{I-1} b_{mn,j}^{(mw)}} \tag{4.25}$$

$$b_{mn,i}^{(mw)} = -\ln\left(\int\limits_{\mathscr{R}^{\dim(\mathbf{o})}} \sqrt{p_{i,m}^{(mw)} \cdot p_{i,n}^{(mw)}}\, d\mathbf{o} \right) \tag{4.26}$$

gdzie $b_{mn,i}^{(mw)}$ jest odległością Bhattacharyya między rozkładami $p_{i,m}^{(mw)}$ i $p_{i,n}^{(mw)}$. Odległość Bhattacharyya jest często wykorzystywana do porównywania rozkładów prawdopodobieństwa. Jej wartość odpowiada nachyleniu wykładniczego asymptotycznego spadku średniego błędu klasyfikacji dla równoprawdopodobnych klas, których elementy występują zgodnie z porównywanymi rozkładami prawdopodobieństwa [69, 94]. W implementacji zastosowano tutaj wielowymiarowe rozkłady normalne, a w takim przypadku równanie (4.26) ma następujące rozwiązanie analityczne [94]:

$$\begin{aligned} b_{mn,i}^{(mw)} = {} & \frac{1}{8}\left(\boldsymbol{\mu}_{m,i}^{(mw)} - \boldsymbol{\mu}_{n,i}^{(mw)}\right)^T \cdot \left(\frac{\boldsymbol{\Sigma}_{m,i}^{(mw)} + \boldsymbol{\Sigma}_{n,i}^{(mw)}}{2}\right)^{-1} \cdot \left(\boldsymbol{\mu}_{m,i}^{(mw)} - \boldsymbol{\mu}_{n,i}^{(mw)}\right) + \\ & + \frac{1}{2}\ln\left(\frac{\left|\boldsymbol{\Sigma}_{m,i}^{(mw)} + \boldsymbol{\Sigma}_{n,i}^{(mw)}\right|}{2\sqrt{\left|\boldsymbol{\Sigma}_{m,i}^{(mw)}\right| \cdot \left|\boldsymbol{\Sigma}_{n,i}^{(mw)}\right|}} \right) \end{aligned} \tag{4.27}$$

Miara podobieństwa (4.25) przyjmuje wartości z przedziału $(-\infty; 0]$. Jest również skonstruowana tak, by z równymi wagami uwzględniać wpływ odległości obliczonej dla poszczególnych fonemów i.

Zaproponowano, by podziału na klasy dokonywać po przeprowadzeniu wstępnych transformacji widma mówców typu każdy-do-każdego. Jest to wariant metody SAT. Stosowana podczas podziału na klasy miara podobieństwa $d_{mn}^{(mw)}$ jest wyznaczana z wykorzystaniem nie zmienionych rozkładów prawdopodobieństwa dla mówcy n oraz rozkładów prawdopodobieństwa uzyskanych po transformacji widma mówcy m, przy czym rozkładami odniesienia podczas wyznaczania wartości parametrów tej transformacji są rozkłady dla mówcy n. Jako funkcję celu w optymalizacji parametrów tych transformacji wykorzystano miarę podobieństwa (4.25), gdyż zastosowanie funkcji

$c_1^{(ro)}$ lub $c_2^{(ro)}$ obarczone jest zbyt dużym nakładem obliczeniowym, optymalizacja przeprowadzana jest bowiem dla każdej pary mówców.

Po wyznaczeniu klas mówców, a przed wyznaczeniem rozkładów prawdopodobieństwa współczynników MFCC w klasach, przeprowadzano jeszcze jedną optymalizację, będącą elementem uczenia typu SAT. Cyklicznie, dla każdego mówcy z danej klasy, z wyjątkiem jej mówcy centralnego, wyznaczano wartości parametrów transformacji widma. Rozkładami odniesienia były rozkłady średnie, uzyskane z danych pochodzących od wszystkich mówców danej klasy, z uwzględnieniem aktualnych, wyznaczonych dla nich transformacji widma, w tym również transformacji, której parametry były w danej chwili optymalizowane. Funkcją celu optymalizacji była natomiast suma miar podobieństwa (4.25) rozkładów każdego mówcy z danej klasy, po uwzględnieniu aktualnych transformacji widma, do aktualnego rozkładu odniesienia dla tej klasy.

Rozkłady prawdopodobieństwa współczynników MFCC dla klas wyznaczano uwzględniając transformacje widma z wartościami parametrów wyznaczonymi dla każdego mówcy opisanym powyżej algorytmem.

4.5. Banki transformacji widma

Banki funkcji skalowania osi częstotliwości i banki filtrów liniowych wyznaczano niezależnie od siebie i oddzielnie dla każdej klasy mówców. Danymi wejściowymi algorytmu konstrukcji banków były zbiory wartości parametrów transformacji widma, wyznaczone dla każdego mówcy ze zbioru uczącego. Rozkładami odniesienia podczas wyznaczania wartości tych parametrów były brzegowe rozkłady prawdopodobieństwa współczynników MFCC w klasach, a wartości początkowe $c_{1,s,\mathscr{J}^{(sp)}}^{(rm,p)}$ i $c_{2,s,\mathscr{J}^{(sp)}}^{(rm,p)}$ wyznaczano w wariancie 1. metody wyboru klas mówców przy braku transformacji widma, a w wariantach 2. i 3. przy transformacji widma danego mówcy z wartościami parametrów takimi, jakie użyto przy wyznaczaniu rozkładów współczynników MFCC w klasach. Algorytm konstrukcji banków był taki sam dla filtrów i funkcji skalujących, a zatem dalej element $\mathbf{x}$ będzie oznaczał wektor parametrów $\boldsymbol{\alpha}^{(g,o)}$ dla funkcji skalujących lub charakterystykę amplitudową $\mathbf{h}$ dla filtrów.

4.5.1. Odległość między parametrami transformacji widma

W algorytmie konstrukcji banków transformacji widma konieczne jest określenie odległości między filtrami oraz między funkcjami skalującymi. W pierwszej wersji algorytmu zastosowano odległość Euklidesa między elementami $\mathbf{x}$. Następnie zaproponowano modyfikację tej odległości tak, by była bardziej skorelowana z rozpoznawalnością izolowanych ramek, czyli z kryterium, które zastosowano podczas wyznaczania wartości parametrów transformacji widma. Zaproponowano odległość

opartą na normie kwadratowej z uwagi na jej przystępność dla obliczeń analitycznych. Wstępne eksperymenty wykazały, że odległość taka, pod warunkiem zastosowania transformacji przestrzeni parametrów, daje wyniki porównywalne z wynikami uzyskanymi w przypadku stosowania odległości opartych na normach innych rzędów (testowano rzędy od 0 do 10).

Odległość w przestrzeni unormowanej indukowana jest przez normę elementu przestrzeni. Zaproponowano następującą zmodyfikowaną normę Euklidesa:

$$\|\mathbf{x}\|_L = \sqrt{(\mathbf{V}\cdot\mathbf{x})^T\cdot(\mathbf{V}\cdot\mathbf{x})} = \sqrt{\mathbf{x}^T\cdot\mathbf{T}\cdot\mathbf{x}} = \sqrt{\mathbf{x}^T\cdot\mathbf{L}\mathbf{L}^T\cdot\mathbf{x}} \tag{4.28}$$

gdzie $\mathbf{V}$ jest macierzą pełnego rzędu, transformującą przestrzeń parametrów, $\mathbf{T}$ jest zatem macierzą dodatnio określoną, którą rozłożyć można za pomocą faktoryzacji Choleskiego na dwie macierze trójkątne dolne $\mathbf{L}$. W przypadku funkcji skalującej, macierz $\mathbf{L}$ ma wymiar 3 i wartości wszystkich niezerowych elementów wyznaczane są w opisanej poniżej optymalizacji. W przypadku filtrów natomiast, macierz $\mathbf{L}$ ma wymiar 256 i w celu zmniejszenia liczby parametrów podlegających optymalizacji wyznaczane są wartości tylko tych leżących na głównej przekątnej, składających się na wektor $\mathbf{l}$, podczas gdy pozostałe przyjęto równe zeru. Dodatkowo liczbę parametrów zmniejszono stosując aproksymację wartości $\mathbf{l}$ w bazie kosinusowej:

$$\mathbf{l} = \sum_{i=0}^{\Theta-1}\theta_i\cdot\boldsymbol{\phi}_i \tag{4.29}$$

$$\phi_{ji} = \cos\left(i\cdot j\cdot\pi/L\right), \quad j = 0,\ldots,L-1 \tag{4.30}$$

gdzie L jest długością wektora $\mathbf{l}$. W implementacji przyjęto liczbę współczynników aproksymacji $\Theta = 5$, większa ich liczba nie przynosiła już poprawy uzyskiwanych wyników.

Macierz $\mathbf{L}$ optymalizowano oddzielnie dla każdej klasy mówców i każdego wariantu metody wyznaczania klas mówców i parametrów transformacji widma. Minimalizowano następującą funkcję celu:

$$c^{(l)}(\mathbf{L}) = \sum_{n=0}^{N-1}\sum_{m=0}^{N-1}\left(\|\mathbf{x}_n-\mathbf{x}_m\|_L^2 - \left(c_{n,n}^{(ro,l)} - c_{n,m}^{(ro,l)}\right)\right)^2 \tag{4.31}$$

gdzie $\mathbf{x}_n$ oznacza wektor parametrów transformacji widma dla n-tego spośród N mówców w danej klasie. Wartość $c_{n,m}^{(ro,l)}$ obliczano analogicznie, jak wartość funkcji $c_1^{(ro)}$ lub $c_2^{(ro)}$ (w zależności od wariantu metody optymalizacji wartości parametrów transformacji widma), z tym, że tutaj wyznaczono ją dla mówcy n, którego widmo zostało przekształcone z parametrami $\mathbf{x}_m$. W przypadku, gdy optymalizowano macierz $\mathbf{L}$ dla parametrów funkcji skalującej, wartości parametrów filtrów były ustalone dla każdego mówcy. Podobnie, gdy optymalizowano macierz $\mathbf{L}$ dla parametrów

filtrów, nie zmieniane były wartości parametrów funkcji skalującej. Rozkładami odniesienia przy wyznaczaniu wartości $c_{n,m}^{(ro,l)}$ były rozkłady dla danej klasy mówców.

Ze względu na kwadratową postać funkcji celu $c^{(l)}(\mathbf{L})$, w optymalizacji zastosowano algorytm Levenberga-Marquardta [38]. W tabeli 4.1 przedstawiono uzyskany po optymalizacji macierzy $\mathbf{L}$ średni współczynnik korelacji między wartościami $\|\mathbf{x}_n - \mathbf{x}_m\|_L^2$ i $\left(c_{n,n}^{(ro,l)} - c_{n,m}^{(ro,l)}\right)$ w porównaniu ze współczynnikiem korelacji uzyskanym dla kwadratu normy Euklidesa.

Tab. 4.1. Wpływ optymalizacji macierzy $\mathbf{L}$ na średni współczynnik korelacji między wartościami $\|\mathbf{x}_n - \mathbf{x}_m\|_L^2$ i $\left(c_{n,n}^{(ro,l)} - c_{n,m}^{(ro,l)}\right)$.

średni współczynnik korelacji	kwadrat normy Euklidesa	$\|\cdot\|_L^2$
parametry funkcji skalującej	0.62	0.86
parametry filtru	0.66	0.78

4.5.2. Algorytm konstrukcji banków

Do wyznaczenia elementów banku transformacji widma wykorzystano metodę hierarchicznej klastryzacji z zastosowaniem odległości Warda między klastrami. Hierarchiczna klastryzacja umożliwia osiąganie różnych stopni szczegółowości, natomiast odległość Warda ma tę właściwość, że na każdym etapie pracy algorytmu klastry łączone są tak, by całkowity błąd, związany z przybliżeniem elementów poddanych klastryzacji przez środki (centroidy) klastrów, rósł minimalnie. Niech K oznacza docelową liczbę klastrów, a N ($N > K$) - liczbę wektorów $\mathbf{x}$ poddanych klastryzacji. Schemat algorytmu jest następujący:

1. Inicjalizacja. Z N wejściowych wektorów $\mathbf{x}$ utwórz N jednoelementowych klastrów.

2. Dla każdej pary klastrów (i, j) oblicz odległość Warda:

$$d_{ij}^{(w)} = \sum_{n=0}^{N_i+N_j-2} \left\|\mathbf{x}_n^{(i+j)} - \mathbf{x}^{(i+j,c)}\right\|_L^2 - \sum_{n=0}^{N_i-1} \left\|\mathbf{x}_n^{(i)} - \mathbf{x}^{(i,c)}\right\|_L^2 - \sum_{n=0}^{N_j-1} \left\|\mathbf{x}_n^{(j)} - \mathbf{x}^{(j,c)}\right\|_L^2 \tag{4.32}$$

gdzie indeks górny (i) przy wektorze $\mathbf{x}_n$ oznacza jego przynależność do liczącego N_i elementów klastra i, analogicznie dla indeksów (j) oraz $(i+j)$ (klastra powstałego z połączenia klastrów i i j), a dodatkowy symbol c w indeksie górnym oznacza wektor, będący centroidem odpowiedniego klastra. Centroid

wyznaczano jako średnią arytmetyczną elementów danego klastra, co wynika z zastosowania odległości opartej na normie kwadratowej.

3. Połącz klastry i i j o najmniejszej uzyskanej $d_{ij}^{(w)}$ oraz zmniejsz N o 1.

4. Jeśli $N > K$, to wróć do punktu 2, w przeciwnym wypadku zakończ i za elementy banku przyjmij centroidy uzyskanych klastrów.

W implementacji przyjęto docelową liczbę elementów banków $K = K^{(g)} = K^{(h)} = 4$.

4.5.3. Banki filtrów uwzględniające zniekształcenia transmisyjne

Wyznaczone w podany powyżej sposób banki filtrów nie zapewniały zadowalającej kompensacji zniekształceń wprowadzanych przez tor transmisyjny. Dlatego do algorytmu wprowadzono modyfikację, polegającą na kaskadowym połączeniu kompensacji różnic osobniczych i transmisyjnych. Dany zbiór N charakterystyk amplitudowych $\mathbf{h}_n$ filtrów, stanowiących dane wejściowe do algorytmu konstrukcji banków, został uzupełniony o charakterystyki kompensujące modelowane zniekształcenia liniowe. Uzupełniony zbiór ma postać:

$$\left\{ \mathbf{h}_n \circ \mathbf{h}_i^{(kan)^{-1}} \circ \mathbf{h}_j^{(kan)^{-1}} : \quad n = 0, \ldots, N-1, \;\; i = 0, 1, 2, 3, \;\; j = 0, 6, 7 \right\} \quad (4.33)$$

gdzie operator $\circ$ oznacza mnożenie wektorów przeprowadzane element po elemencie, a $\mathbf{h}_i^{(kan)^{-1}}$ oznacza odwrotność charakterystyki $\mathbf{h}_i^{(kan)}$, modelującej zniekształcenia kanałowe (zob. dodatek D). Wyjątkiem jest $\mathbf{h}_0^{(kan)^{-1}}$, która jest charakterystyką stałą równą 1. Do wyznaczania uzupełnionych zbiorów nie stosowano charakterystyk $\mathbf{h}_4^{(kan)}$ i $\mathbf{h}_5^{(kan)}$, gdyż wykorzystywane są one tylko podczas weryfikacji działania metody kompensacji. Liczba zer funkcji transmitancji, modelujących charakterystyki filtrów, została zwiększona z 4 do 6, gdyż charakterystyki kompensujące zniekształcenia amplitudowe modelowane są za pomocą 6 zer funkcji transmitancji. Współczynniki równania różnicowego $a_i^{(f)}$, wyznaczane z charakterystyk amplitudowych, normalizowano zgodnie z (4.12).

4.5.4. Wyznaczanie elementów dodatkowych banków

Nawet stosując klastryzację hierarchiczną, na dolnych jej poziomach nie można już zwiększyć liczby dostępnych filtrów czy funkcji skalujących, gdyż ograniczona jest ona liczbą mówców w zbiorze uczącym. Większą różnorodność filtrów i funkcji skalujących osiągnięto stosując metodę iteracyjną. Dla każdego uzyskanego w wyniku działania algorytmu z rozdziału 4.5.2 klastra i wyznaczono dwa dodatkowe elementy

$\mathbf{x}^{(i,d1)}$ i $\mathbf{x}^{(i,d2)}$. W etapie rozpoznawania iteracyjnie wyznaczano kombinację liniową elementów dodatkowych i centroidu klastra $\mathbf{x}^{(i,c)}$, stosując metodę poszukiwań prostych, która miała na celu maksymalizację zdefiniowanej w systemie ARM miary oceny rozpoznania (zob. rozdział 4.6). Przedstawiony poniżej algorytm wyznaczania elementów dodatkowych zawiera analogiczną metodę poszukiwań prostych, gdyż powinien odpowiadać algorytmowi stosowanemu w etapie właściwego rozpoznawania.

1. Inicjalizacja. Przyjmij $j = 0$.

   - Jeśli dany klaster zawiera więcej niż 1 element, wyznacz:

     $$\mathbf{x}^{(i,d1)} = \mathbf{x}^{(i,c)} + \sqrt{\lambda} \cdot \mathbf{v} \tag{4.34}$$
     $$\mathbf{x}^{(i,d2)} = \mathbf{x}^{(i,c)} - \sqrt{\lambda} \cdot \mathbf{v} \tag{4.35}$$

     gdzie λ i $\mathbf{v}$ oznaczają odpowiednio największą wartość własną oraz odpowiadający jej wektor własny macierzy kowariancji elementów $\mathbf{x}_n^{(i)}$ danego klastra. Za $\mathbf{B}_0$ przyjmij macierz zawierającą w kolejnych kolumnach elementy $\mathbf{x}^{(i,c)}$, $\mathbf{x}^{(i,d1)}$ i $\mathbf{x}^{(i,d2)}$. Przejdź do punktu 2.

   - Jeśli dany klaster zawiera 1 element, wyznacz:
     - Dla filtrów:

       $$x_m^{(i,d1)} = (0.8 + 0.4 \cdot m/L) \cdot x_m^{(i,c)}, \qquad m = 0, \ldots, L-1 \tag{4.36}$$
       $$x_m^{(i,d2)} = (1.2 - 0.4 \cdot m/L) \cdot x_m^{(i,c)}, \qquad m = 0, \ldots, L-1 \tag{4.37}$$

       gdzie L oznacza długość wektorów $\mathbf{x}$.
     - Dla funkcji skalującej:

       $$\mathbf{x}^{(i,d1)} = 1.1 \cdot \mathbf{x}^{(i,c)} \tag{4.38}$$
       $$\mathbf{x}^{(i,d2)} = 0.9 \cdot \mathbf{x}^{(i,c)} \tag{4.39}$$

       Zakończ algorytm.

2. Zwiększ j o 1. Przyjmij $k = 1$. Powtarzaj punkty 3-6 dla każdego elementu $\mathbf{x}_n^{(i)}$ klastra.

3. Za $\mathbf{E}_k$ przyjmij macierz jednostkową wymiaru 3.

4. Oblicz zmodyfikowaną odległość Euklidesa (zob. rozdział 4.5.1) danego elementu $\mathbf{x}_n^{(i)}$ klastra do $3 + 2(k-1)$ elementów będących kolumnami macierzy $\mathbf{F} = \mathbf{B}_{j-1} \cdot \mathbf{E}_k$. Macierz $\mathbf{E}_k$ zawiera w kolejnych kolumnach współczynniki kombinacji liniowych elementów zawartych w kolumnach macierzy $\mathbf{B}_{j-1}$. Kombinacje te stanowią punkty algorytmu poszukiwań prostych. Numery trzech kolumn

macierzy $\mathbf{F}$, dla których osiągnięto najmniejsze odległości, oznacz zgodnie z rosnącymi odległościami jako r_1, r_2 i r_3. Utwórz dwa nowe punkty poszukiwań prostych, wyznaczając następującą macierz:

$$\mathbf{E}_{k+1} = \left[\mathbf{E}_k \quad \frac{1}{2}\left(\mathbf{e}_{r_1,k} + \mathbf{e}_{r_2,k}\right) \quad \frac{1}{2}\left(\mathbf{e}_{r_1,k} + \mathbf{e}_{r_3,k}\right)\right] \tag{4.40}$$

5. Jeśli $k < I^{(k)}$, to zwiększ k o 1 i wróć do punktu 4.

6. Wyznacz numer kolumny macierzy $\mathbf{F} = \mathbf{B}_{j-1} \cdot \mathbf{E}_{k+1}$ taki, że do elementu zawartego w tej kolumnie zmodyfikowana odległość Euklidesa danego elementu $\mathbf{x}_n^{(i)}$ klastra jest najmniejsza. Zapamiętaj kolumnę macierzy $\mathbf{E}_{k+1}$ o tym numerze i oznacz ją jako $\mathbf{e}_n^{(opt)}$.

7. Wyznacz macierz $\mathbf{B}_j$ rozwiązując układ równań (4.42) metodą najmniejszych kwadratów:

$$\mathbf{B}_j = \left[\mathbf{x}^{(i,c)} \quad \mathbf{B}^{(r)}\right] \tag{4.41}$$

$$\mathbf{B}^{(r)} \cdot \mathbf{E}^{(r)} = \left[\mathbf{x}_0^{(i)} \cdots \mathbf{x}_{N^{(i)}-1}^{(i)}\right] - \mathbf{x}^{(i,c)} \cdot \mathbf{e}^{(r)^T} \tag{4.42}$$

gdzie $\mathbf{E}^{(r)}$ jest macierzą zawierającą w kolejnych kolumnach wektory $\mathbf{e}_n^{(opt)}$, w których pominięto pierwszy element. Pierwsze elementy wektorów $\mathbf{e}_n^{(opt)}$ budują natomiast wektor $\mathbf{e}^{(r)}$. Powyższa operacja ma za zadanie modyfikację elementów dodatkowych przy zachowanej wartości centroidu $\mathbf{x}^{(i,c)}$.

8. Wyznacz sumaryczny błąd w j-tej iteracji

$$\epsilon_j^{(bnk)} = \left\| \mathbf{L}^T \cdot \left(\mathbf{B}^{(r)} \cdot \mathbf{E}^{(r)} - \left[\mathbf{x}_0^{(i)} \cdots \mathbf{x}_{N^{(i)}-1}^{(i)}\right] + \mathbf{x}^{(i,c)} \cdot \mathbf{e}^{(r)^T} \right) \right\|_F \tag{4.43}$$

gdzie indeks F oznacza normę Frobeniusa macierzy.

9. Jeśli $j < I^{(j)}$ i spadek błędu $\epsilon_j^{(bnk)}$ w stosunku do błędu $\epsilon_{j-1}^{(bnk)}$ jest powyżej zadanego progu, to wróć do punktu 2.

10. Jako wyjściowe elementy dodatkowe $\mathbf{x}^{(i,d1)}$ i $\mathbf{x}^{(i,d2)}$ danego klastra wybierz drugą i trzecią kolumnę macierzy $\mathbf{B}_j$.

Zastosowana metoda nie zapewnia ogólnie monotonicznego spadku błędu $\epsilon_j^{(bnk)}$, co wynika z charakteru zastosowanej metody aproksymacji wykorzystującej poszukiwania proste. W praktyce jednak obserwowany jest zadowalający spadek błędu $\epsilon_j^{(bnk)}$ podczas optymalizacji. Liczbę iteracji zastosowaną w implementacji przyjęto $I^{(k)} = 4$ i $I^{(j)} = 8$.

4.6. Rozpoznawanie mowy z wykorzystaniem banków transformacji widma

Opisany poniżej algorytm rozpoznawania mowy z wykorzystaniem banków transformacji widma jest taki sam zarówno w przypadku rozpoznawania komend różnymi wariantami systemu ARM, jak i w przypadku rozpoznawania izolowanych ramek. W każdym zadaniu rozpoznawania stosowana jest odpowiednio zdefiniowana miara oceny rozpoznania, wskazująca na ile „dobry" jest dany wynik (zob. rozdział 4.6.2).

4.6.1. Algorytm rozpoznawania

W pierwszym etapie wyznaczano wartość miary oceny rozpoznania danej wypowiedzi lub zbioru izolowanych ramek dla wszystkich klas mówców i wszystkich kombinacji elementów banków transformacji widma (zob. rys. 4.1). Wybierano klasę i kombinację, dla której uzyskano najwyższą wartość tej miary.

W etapie drugim przeprowadzano iteracyjne poprawianie rozpoznania dla wariantu wybranego w pierwszym etapie, mające na celu maksymalizację miary oceny rozpoznania. Metoda ta jest optymalizacją z ograniczeniami, co zapewnia pozostawanie uzyskiwanych w jej trakcie wartości parametrów transformacji widma w założonych granicach. Podejście takie zapobiega „rozbieganiu" się algorytmu optymalizacji i uzyskiwaniu w konsekwencji błędnych wyników rozpoznania.

1. Inicjalizacja. Przyjmij $k = 0$. Za $\mathbf{B}^{(h)}$ przyjmij macierz zawierającą w kolejnych kolumnach wektory odpowiednio: $\mathbf{x}^{(i,c)}$, $\mathbf{x}^{(i,d1)}$ i $\mathbf{x}^{(i,d2)}$ danego elementu i banku filtrów. Za $\mathbf{B}^{(g)}$ przyjmij macierz zawierającą w kolejnych kolumnach analogiczne wektory dla danego elementu banku funkcji skalowania osi częstotliwości. Macierze $\mathbf{B}^{(h)}$ i $\mathbf{B}^{(g)}$ zawierają zatem elementy bazowe, których kombinacje liniowe (punkty algorytmu poszukiwań prostych) będą wyznaczane. Początkowe współczynniki kombinacji zawarte są w kolumnach następujących macierzy:

$$\mathbf{E}_1^{(h)} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix} \tag{4.44}$$

$$\mathbf{E}_1^{(g)} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \end{bmatrix} \tag{4.45}$$

2. Zwiększ k o 1. Oblicz miarę oceny rozpoznania dla $5 + 4(k-1)$ zbiorów wartości parametrów transformacji widma, wyznaczonych odpowiednio z równań

$\mathbf{F}^{(h)} = \mathbf{B}^{(h)} \cdot \mathbf{E}_k^{(h)}$ i $\mathbf{F}^{(g)} = \mathbf{B}^{(g)} \cdot \mathbf{E}_k^{(g)}$. Każdy zbiór wartości parametrów składa się z parametrów filtru i parametrów funkcji skalującej, które odpowiadają kolumnom macierzy $\mathbf{F}^{(h)}$ i $\mathbf{F}^{(g)}$ o tym samym numerze. Wyznacz numer kolumny macierzy $\mathbf{F}^{(h)}$ i $\mathbf{F}^{(g)}$ (ten sam dla obu) taki, dla którego uzyskano maksymalną wartość miary oceny rozpoznania i oznacz go jako r_1

3. Wyznacz dwa numery kolumn macierzy $\mathbf{F}^{(h)}$ i $\mathbf{F}^{(g)}$ (każdy numer określa tę samą kolumnę w obu macierzach) takie, dla których uzyskano maksymalną wartość miary oceny rozpoznania, ale dla których jednocześnie odpowiadające im kolumny macierzy $\mathbf{E}_k^{(h)}$ są różne od kolumny macierzy $\mathbf{E}_k^{(h)}$ o numerze r_1. Oznacz wybrane numery jako r_2 i r_3.

4. Wyznacz dwa następne numery kolumn macierzy $\mathbf{F}^{(h)}$ i $\mathbf{F}^{(g)}$ (każdy numer określa tę samą kolumnę w obu macierzach) takie, dla których uzyskano maksymalną wartość miary oceny rozpoznania, ale dla których jednocześnie odpowiadające im kolumny macierzy $\mathbf{E}_k^{(g)}$ są różne od kolumny macierzy $\mathbf{E}_k^{(g)}$ o numerze r_1 i ponadto nie są to numery r_2 i r_3. Oznacz wybrane numery jako r_4 i r_5.

5. Jeśli w punktach 3 i 4 nie wybrano czterech numerów kolumn, dobierz numery do czterech takie, dla których uzyskano maksymalną wartość miary oceny rozpoznania, ale które nie zostały wybrane w punktach od 2 do 4.

6. Utwórz cztery nowe punkty poszukiwań prostych wyznaczając następujące macierze:

$$\mathbf{E}_{k+1}^{(h)} = \left[\mathbf{E}_k^{(h)} \quad \frac{\mathbf{e}_{r_1,k}^{(h)} + \mathbf{e}_{r_2,k}^{(h)}}{2} \quad \frac{\mathbf{e}_{r_1,k}^{(h)} + \mathbf{e}_{r_3,k}^{(h)}}{2} \quad \frac{\mathbf{e}_{r_1,k}^{(h)} + \mathbf{e}_{r_4,k}^{(h)}}{2} \quad \frac{\mathbf{e}_{r_1,k}^{(h)} + \mathbf{e}_{r_5,k}^{(h)}}{2} \right] \tag{4.46}$$

$$\mathbf{E}_{k+1}^{(g)} = \left[\mathbf{E}_k^{(g)} \quad \frac{\mathbf{e}_{r_1,k}^{(g)} + \mathbf{e}_{r_2,k}^{(g)}}{2} \quad \frac{\mathbf{e}_{r_1,k}^{(g)} + \mathbf{e}_{r_3,k}^{(g)}}{2} \quad \frac{\mathbf{e}_{r_1,k}^{(g)} + \mathbf{e}_{r_4,k}^{(g)}}{2} \quad \frac{\mathbf{e}_{r_1,k}^{(g)} + \mathbf{e}_{r_5,k}^{(g)}}{2} \right] \tag{4.47}$$

7. Jeśli $k < I^{(k)}$, wróć do punktu 2, w przeciwnym wypadku oblicz wartość miary oceny rozpoznania dla $5+4k$ zbiorów wartości parametrów transformacji widma, wyznaczonych odpowiednio z równań $\mathbf{F}^{(h)} = \mathbf{B}^{(h)} \cdot \mathbf{E}_{k+1}^{(h)}$ i $\mathbf{F}^{(g)} = \mathbf{B}^{(g)} \cdot \mathbf{E}_{k+1}^{(g)}$. Jako końcowy wynik rozpoznania wybierz ten zapewniający maksymalną wartość miary oceny.

4.6.2. Miary oceny rozpoznania

W przypadku rozpoznawania izolowanych ramek stosowano miary oceny rozpoznania $c_{k,\mathscr{J},\mathscr{S}}^{(rs)}$ (zob. dodatek E).

W zadaniu rozpoznawania komend jako miarę oceny rozpoznania danej wypowiedzi wybierano najwyższą spośród wartości miar oceny $c_k^{(rozp)}$, zdefiniowanych poniżej, uzyskanych dla wyrazów ze słownika (k oznacza numer wyrazu).

- Dla wariantów A i At systemu (zob. rozdziały C.2.2 i C.4.2):

$$c_k^{(rozp)} = P_k^{(wyr)} \cdot P^{(aku)^\kappa} \tag{4.48}$$

  gdzie $P^{(aku)}$ oznacza prawdopodobieństwo ścieżki Viterbiego, podniesione celem znormalizowania do potęgi $1/T$, gdzie T to długość wypowiedzi w ramkach. Wagę κ ustalono eksperymentalnie równą 0.1 dla wariantu A i równą 1.5 dla wariantu At.

- Dla wariantów B i Bt jako miarę $c_k^{(rozp)}$ przyjęto zlogarytmowane prawdopodobieństwo ścieżki Viterbiego, uzyskane za pomocą modelu HMM k-tego wyrazu w słowniku (zob. rozdziały C.3.2 i C.5).

4.6.3. Uczenie SAT systemu ARM

Metody wyznaczania rozkładów prawdopodobieństwa współczynników MFCC zostały omówione w rozdziale 4.4. Wartości pozostałych parametrów systemu wyznaczano również z ukierunkowaniem na rozpoznawanie z zastosowaniem algorytmu kompensacji cech osobniczych i liniowych zniekształceń transmisyjnych. W algorytmie optymalizacji tych wartości zastosowano sprzężenie zwrotne, polegające na cyklicznie przeprowadzanym rozpoznawaniu i wyborze kombinacji klas mówców oraz elementów banków transformacji widma dla każdej wypowiedzi przy bieżących wartościach parametrów systemu i uwzględnianiu tych wyborów w dalszym przebiegu uczenia. W przypadku zastosowania banków filtrów zmodyfikowanych tak, by uwzględniane były modelowane zniekształcenia transmisyjne (zob. rozdział 4.5.3), uczenie systemu odbywało się z zastosowaniem odpowiadających im banków niezmodyfikowanych.

4.6.4. Przyporządkowanie mówcy do klasy na podstawie wartości częstotliwości tonu krtaniowego

Poprawa rozpoznawalności przy wykorzystaniu banków transformacji widma jest kosztowna obliczeniowo, gdyż wymaga przeprowadzenia części parametryzacji sygnału, wyznaczenia prawdopodobieństw obserwacji stanów HMM, zastosowania algorytmu Viterbiego i końcowej klasyfikacji oddzielnie dla każdej transformacji. Dodatkowo kilkanaście transformacji wykonywanych jest w trakcie iteracyjnego poprawiania wyniku rozpoznania. Parametryzacja i wyznaczanie prawdopodobieństw

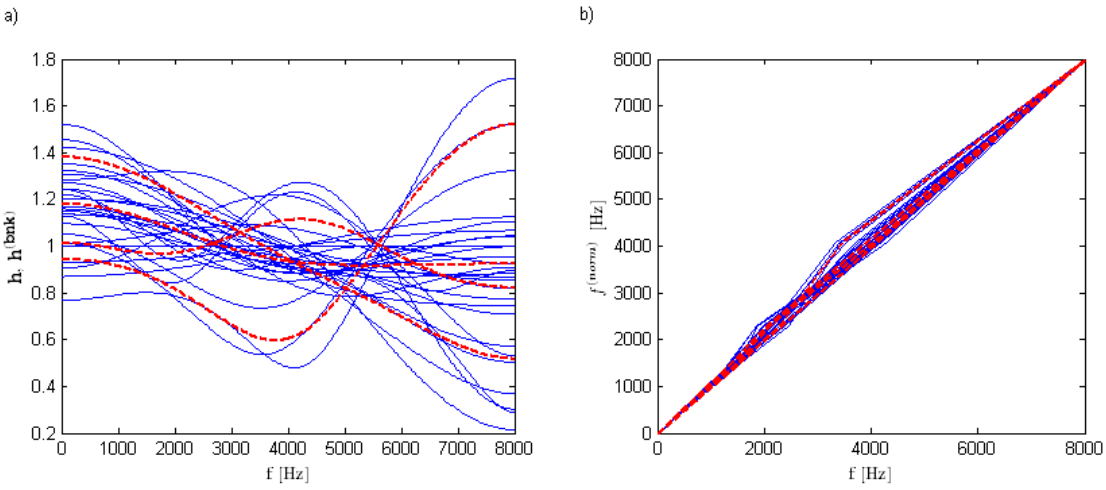
obserwacji stanów HMM są najbardziej kosztownymi obliczeniowo etapami rozpoznawania mowy w przypadku izolowanych słów i małego słownika. Aby częściowo zredukować nakład obliczeniowy zaproponowano, by wybór klasy mówców dla rozpoznawanej wypowiedzi przeprowadzany był na podstawie wartości częstotliwości tonu krtaniowego $f^{(v)}$. Zaobserwowano bowiem, że w przypadku stosowania dwóch klas, algorytmy opisane w rozdziale 4.4 podzieliły mówców w zasadzie względem płci, a w przypadku trzech klas - w dwóch znalazły się głównie głosy męskie, a w pozostałej głównie głosy kobiece.

Algorytm estymacji wartości $f^{(v)}$ powinien charakteryzować się jak najmniejszą złożonością obliczeniową, znikomą w stosunku do złożoności obliczeniowej potrzebnej do wykonania rozpoznania przy użyciu banków transformacji widma. Wprowadzony oryginalny algorytm bazuje na widmie amplitudowym ramki wyznaczanym i tak w trakcie parametryzacji MFCC. Zrezygnowano z dokładniejszych metod korelacyjnych z uwagi na ich zbyt duży dodatkowy koszt obliczeniowy. Szczegóły algorytmu podano w dodatku G.

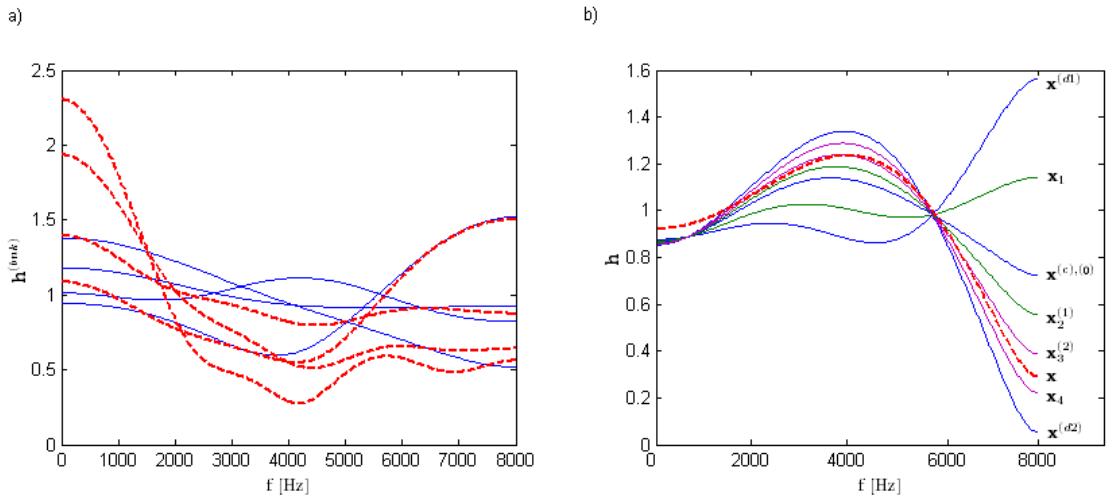
Klasyfikację przeprowadzano poprzez porównanie wykrytej dla danej wypowiedzi wartości $f^{(v)}$ z wartością progu decyzyjnego, równą $f^{(v,p)} = 190$ Hz. W przypadku dwóch klas mówców wybór klasy dokonywany jest wtedy tylko na podstawie $f^{(v)}$, a w przypadku trzech klas, przyporządkowanie głosu uznanego na podstawie wartości $f^{(v)}$ za męski do jednej z dwóch klas zawierających głosy męskie przeprowadzane jest już standardowo.

5. Wyniki rozpoznawania mowy z wykorzystaniem metody banków transformacji widma

Na rys. 5.1 i 5.2 przedstawiono przykłady banków transformacji widma uzyskanych za pomocą metody opisanej w rozdziale 4. Pokazano wyniki dla pierwszej z dwóch klas mówców, przy czym stosowano wariant 3. metody wyznaczania klas mówców i funkcję celu $c_2^{(ro)}$ w optymalizacji wartości parametrów transformacji. Rys. 5.1a przedstawia zbiór filtrów liniowych i wyznaczone na jego podstawie elementy banku, rys. 5.1b analogicznie dla funkcji skalujących, a na rys. 5.2a pokazano różnice między elementami banku filtrów wyznaczonymi bez i z modyfikacją uwzględniającą zniekształcenia transmisyjne. Przykład iteracyjnego ($I^{(k)} = 2$) wyznaczania kombinacji liniowej elementów $\mathbf{x}^{(c)}$, $\mathbf{x}^{(d1)}$ i $\mathbf{x}^{(d2)}$ najbliższej danemu elementowi $\mathbf{x}$ przedstawiono na rys. 5.2b. Kombinacje dodane w iteracji 1. oznaczono jako $\mathbf{x}_1$ i $\mathbf{x}_2$, a w iteracji 2. jako $\mathbf{x}_3$ i $\mathbf{x}_4$. Kombinację najbliższą w danej iteracji elementowi $\mathbf{x}$ oznaczono liczbą w nawiasie umieszczoną w indeksie górnym i wskazującą numer tej iteracji.



Rys. 5.1. Przykłady uzyskanych transformacji widma: a) charakterystyki amplitudowe filtrów dla mówców (lin. ciągłe) i elementy wyznaczonego z nich banku (lin. przerywane), b) funkcje skalujące dla mówców (lin. ciągłe) i elementy wyznaczonego z nich banku (lin. przerywane).



Rys. 5.2. Przykłady uzyskanych transformacji widma: a) elementy banku filtrów wyznaczone bez uwzględniania zniekształceń transmisyjnych (lin. ciągłe) i z ich uwzględnieniem (lin. przerywane), b) iteracyjne wyznaczanie kombinacji liniowej elementów $\mathbf{x}^{(c)}$, $\mathbf{x}^{(d1)}$ i $\mathbf{x}^{(d2)}$ najbliższej danemu elementowi $\mathbf{x}$.

Tab. 5.1. Wyniki rozpoznawalności izolowanych ramek po zastosowaniu transformacji widma dla każdego mówcy. Czcionką pogrubioną zaznaczano najwyższy wynik w danej kolumnie.

li. kl.	war. met. kl.	war. fun. opt.	zbiór uczący $c_1^{(rs)}$ [%] sam.	zbiór uczący $c_1^{(rs)}$ [%] wsz.	zbiór uczący $c_2^{(rs)}$ sam.	zbiór uczący $c_2^{(rs)}$ wsz.	zbiór testowy $c_1^{(rs)}$ [%] sam.	zbiór testowy $c_1^{(rs)}$ [%] wsz.	zbiór testowy $c_2^{(rs)}$ sam.	zbiór testowy $c_2^{(rs)}$ wsz.
bez kompensacji			55.82	40.00	0.216	-0.722	55.27	39.24	0.169	-0.764
2	1	1	62.57	44.15	0.684	-0.542	61.45	42.35	0.552	**-0.644**
2	1	2	61.94	43.51	0.777	-0.539	60.26	41.41	0.667	-0.663
3	1	1	**63.46**	46.12	0.810	-0.455	61.91	**42.36**	0.613	-0.738
3	1	2	63.09	45.66	0.913	**-0.443**	61.42	41.66	0.713	-0.774
2	2	1	62.59	43.88	0.728	-0.553	62.49	42.11	0.638	-0.654
2	2	2	62.26	43.36	0.829	-0.558	61.76	41.75	**0.767**	-0.645
3	2	1	63.39	46.06	0.849	-0.449	62.48	42.18	0.635	-0.733
3	2	2	63.18	45.63	**0.960**	-0.451	61.80	41.71	0.744	-0.738
2	3	1	61.99	45.60	0.633	-0.507	**63.06**	42.09	0.619	-0.704
2	3	2	61.88	45.02	0.747	-0.497	61.98	41.43	0.746	-0.729
3	3	1	62.48	**46.77**	0.770	-0.455	61.69	41.27	0.616	-0.800
3	3	2	62.72	46.23	0.885	-0.447	61.43	40.64	0.740	-0.807

Tab. 5.2. Wyniki rozpoznawalności izolowanych ramek dla **zbioru uczącego** po zastosowaniu banków transformacji widma. **Nie symulowano** zniekształceń transmisyjnych. Wykorzystano banki filtrów **nie uwzględniające** zniekształceń transmisyjnych.

li.	war.	war.	miar.oc.rozp.: $c_1^{(rs)}$ dla sam.				miar.oc.rozp.: $c_1^{(rs)}$ dla wsz.				miar.oc.rozp.: $c_2^{(rs)}$ dla sam.				miar.oc.rozp.: $c_2^{(rs)}$ dla wsz.			
kl.	met.	fun.	$c_1^{(rs)}$ [%]		$c_2^{(rs)}$		$c_1^{(rs)}$ [%]		$c_2^{(rs)}$		$c_1^{(rs)}$ [%]		$c_2^{(rs)}$		$c_1^{(rs)}$ [%]		$c_2^{(rs)}$	
	kl.	opt.	sam.	wsz.	sam.	wsz.	sam.	wsz.	sam.	wsz.	sam.	wsz.	sam.	wsz.	sam.	wsz.	sam.	wsz.
2	1	1	62.69	43.25	0.709	-0.581	61.15	44.58	0.616	-0.524	61.35	42.61	0.778	-0.593	61.16	44.25	0.653	-0.517
2	1	2	62.83	43.01	0.742	-0.586	61.16	44.27	0.671	-0.534	61.62	42.11	0.813	-0.621	61.19	44.00	0.693	-0.521
3	1	1	64.28	44.46	0.830	-0.540	62.36	46.28	0.768	-0.450	62.97	44.40	0.926	-0.530	62.54	45.94	0.811	-0.437
3	1	2	64.38	44.13	0.892	-0.544	62.36	46.13	0.815	-0.447	63.20	43.94	0.964	-0.541	62.49	45.80	0.860	-0.437
2	2	1	62.62	42.68	0.756	-0.605	60.96	44.06	0.669	-0.543	61.51	42.52	0.802	-0.602	60.85	43.75	0.690	-0.530
2	2	2	62.94	42.14	0.781	-0.631	60.71	43.92	0.670	-0.545	61.87	41.84	0.832	-0.633	60.83	43.64	0.698	-0.530
3	2	1	63.90	44.40	0.880	-0.528	62.19	46.30	0.813	-0.446	62.57	44.31	0.962	-0.520	62.32	45.98	0.838	-0.425
3	2	2	64.11	44.30	0.899	-0.524	61.77	46.11	0.807	-0.438	63.06	43.76	0.992	-0.537	62.22	45.79	0.864	-0.419
2	3	1	63.77	42.31	0.762	-0.727	61.32	45.47	0.634	-0.505	62.41	42.03	0.838	-0.733	61.21	45.27	0.638	-0.495
2	3	2	63.48	43.31	0.782	-0.636	61.37	45.54	0.664	-0.499	62.02	42.27	0.839	-0.700	61.21	45.30	0.673	-0.486
3	3	1	64.13	43.46	0.864	-0.669	61.58	46.58	0.718	-0.475	62.64	43.56	0.949	-0.642	61.62	46.18	0.752	-0.453
3	3	2	64.41	43.61	0.906	-0.634	62.28	46.40	0.795	-0.462	63.45	43.59	0.989	-0.643	62.30	46.10	0.820	-0.448

Tab. 5.3. Wyniki rozpoznawalności izolowanych ramek dla **zbioru testowego** po zastosowaniu banków transformacji widma. **Nie symulowano** zniekształceń transmisyjnych. Wykorzystano banki filtrów **nie uwzględniające** zniekształceń transmisyjnych.

li.	war.	war.	miar.oc.rozp.: $c_1^{(rs)}$ dla sam.				miar.oc.rozp.: $c_1^{(rs)}$ dla wsz.				miar.oc.rozp.: $c_2^{(rs)}$ dla sam.				miar.oc.rozp.: $c_2^{(rs)}$ dla wsz.			
kl.	met.	fun.	$c_1^{(rs)}$ [%]		$c_2^{(rs)}$		$c_1^{(rs)}$ [%]		$c_2^{(rs)}$		$c_1^{(rs)}$ [%]		$c_2^{(rs)}$		$c_1^{(rs)}$ [%]		$c_2^{(rs)}$	
	kl.	opt.	sam.	wsz.	sam.	wsz.	sam.	wsz.	sam.	wsz.	sam.	wsz.	sam.	wsz.	sam.	wsz.	sam.	wsz.
2	1	1	60.33	41.28	0.514	-0.683	58.34	42.82	0.395	-0.619	58.98	40.26	0.599	-0.727	58.55	42.44	0.441	-0.602
2	1	2	59.86	39.52	0.479	-0.981	57.89	42.25	0.426	-0.639	58.36	39.78	0.596	-0.752	58.01	42.03	0.458	-0.628
3	1	1	60.80	39.41	0.522	-0.918	57.57	42.32	0.354	-0.691	59.38	39.55	0.624	-0.908	57.71	41.92	0.397	-0.681
3	1	2	61.69	38.43	0.546	-1.076	57.51	42.04	0.431	-0.741	60.00	38.96	0.629	-0.946	57.90	41.49	0.467	-0.704
2	2	1	60.21	40.58	0.562	-0.719	57.92	42.24	0.450	-0.647	59.11	40.64	0.620	-0.701	58.45	41.67	0.525	-0.641
2	2	2	60.77	39.04	0.535	-1.000	58.32	42.34	0.457	-0.638	59.56	40.10	0.632	-0.733	58.62	42.07	0.522	-0.619
3	2	1	61.38	40.26	0.581	-0.850	57.84	42.53	0.436	-0.667	59.62	40.15	0.640	-0.847	58.19	41.90	0.473	-0.647
3	2	2	62.15	39.12	0.600	-1.075	57.86	42.24	0.435	-0.707	60.50	39.92	0.666	-0.823	57.98	41.86	0.483	-0.662
2	3	1	63.71	39.67	0.673	-0.869	60.35	43.67	0.497	-0.612	61.76	39.22	0.755	-0.878	60.17	43.40	0.510	-0.600
2	3	2	63.87	39.01	0.700	-1.022	60.45	43.47	0.552	-0.622	62.07	38.46	0.772	-0.997	60.23	43.25	0.535	-0.604
3	3	1	61.79	38.13	0.544	-1.098	56.29	42.40	0.313	-0.741	59.36	39.01	0.659	-0.906	57.16	42.02	0.382	-0.696
3	3	2	61.86	37.87	0.530	-1.086	56.53	42.11	0.366	-0.738	59.99	39.07	0.651	-0.914	57.42	41.63	0.444	-0.718

Tab. 5.4. Wyniki rozpoznawalności izolowanych ramek dla **zbioru uczącego** po zastosowaniu banków transformacji widma. **Symulowano** zniekształcenia transmisyjne. Wykorzystano banki filtrów **nie uwzględniające** zniekształceń transmisyjnych.

li.	war.	war.	miar.oc.rozp.: $c_1^{(rs)}$ dla sam.				miar.oc.rozp.: $c_1^{(rs)}$ dla wsz.				miar.oc.rozp.: $c_2^{(rs)}$ dla sam.				miar.oc.rozp.: $c_2^{(rs)}$ dla wsz.			
kl.	met.	fun.	$c_1^{(rs)}$ [%]		$c_2^{(rs)}$		$c_1^{(rs)}$ [%]		$c_2^{(rs)}$		$c_1^{(rs)}$ [%]		$c_2^{(rs)}$		$c_1^{(rs)}$ [%]		$c_2^{(rs)}$	
	kl.	opt.	sam.	wsz.	sam.	wsz.	sam.	wsz.	sam.	wsz.	sam.	wsz.	sam.	wsz.	sam.	wsz.	sam.	wsz.
2	1	1	63.40	41.29	0.722	-0.712	61.78	42.58	0.659	-0.656	62.18	41.00	0.796	-0.708	61.96	42.26	0.704	-0.637
2	1	2	63.72	40.97	0.774	-0.760	61.52	42.61	0.673	-0.651	62.67	40.69	0.833	-0.745	61.71	42.42	0.699	-0.637
3	1	1	64.52	41.68	0.813	-0.775	62.19	43.94	0.770	-0.664	63.43	42.27	0.895	-0.740	62.60	43.70	0.794	-0.602
3	1	2	65.13	41.70	0.903	-0.775	62.03	43.79	0.775	-0.669	63.91	41.98	0.974	-0.712	62.83	43.54	0.827	-0.614
2	2	1	63.35	40.89	0.795	-0.730	61.72	42.59	0.733	-0.641	61.98	40.94	0.842	-0.711	61.32	42.26	0.732	-0.625
2	2	2	63.81	40.13	0.796	-0.840	61.40	42.45	0.713	-0.646	63.02	40.57	0.881	-0.734	61.40	42.19	0.740	-0.631
3	2	1	64.39	41.92	0.882	-0.743	62.17	44.28	0.762	-0.619	63.38	42.25	0.983	-0.696	62.24	43.85	0.802	-0.588
3	2	2	65.16	41.57	0.965	-0.774	61.92	44.29	0.824	-0.620	64.07	41.98	1.036	-0.710	62.66	43.95	0.904	-0.550
2	3	1	64.29	39.70	0.733	-0.947	61.12	43.27	0.561	-0.671	63.01	39.74	0.807	-0.946	61.66	42.92	0.602	-0.665
2	3	2	64.27	40.56	0.768	-0.893	61.45	43.57	0.614	-0.645	63.06	40.47	0.849	-0.842	61.74	43.48	0.673	-0.615
3	3	1	64.71	40.97	0.863	-0.878	61.80	44.37	0.711	-0.637	63.22	41.39	0.940	-0.798	62.09	44.16	0.747	-0.618
3	3	2	65.07	40.85	0.926	-0.872	62.02	44.16	0.740	-0.641	63.83	41.19	0.988	-0.834	62.27	43.81	0.795	-0.612

Tab. 5.5. Wyniki rozpoznawalności izolowanych ramek dla **zbioru testowego** po zastosowaniu banków transformacji widma. **Symulowano** zniekształcenia transmisyjne. Wykorzystano banki filtrów **nie uwzględniające** zniekształceń transmisyjnych.

li.	war.	war.	miar.oc.rozp.: $c_1^{(rs)}$ dla sam.				miar.oc.rozp.: $c_1^{(rs)}$ dla wsz.				miar.oc.rozp.: $c_2^{(rs)}$ dla sam.				miar.oc.rozp.: $c_2^{(rs)}$ dla wsz.			
kl.	met.	fun.	$c_1^{(rs)}$ [%]		$c_2^{(rs)}$		$c_1^{(rs)}$ [%]		$c_2^{(rs)}$		$c_1^{(rs)}$ [%]		$c_2^{(rs)}$		$c_1^{(rs)}$ [%]		$c_2^{(rs)}$	
	kl.	opt.	sam.	wsz.	sam.	wsz.	sam.	wsz.	sam.	wsz.	sam.	wsz.	sam.	wsz.	sam.	wsz.	sam.	wsz.
2	1	1	61.79	39.92	0.609	-0.903	59.85	41.77	0.498	-0.703	60.59	39.76	0.695	-0.816	59.69	41.44	0.524	-0.686
2	1	2	62.10	38.41	0.574	-1.152	58.99	41.55	0.504	-0.720	60.16	39.40	0.695	-0.834	59.11	41.37	0.528	-0.699
3	1	1	61.90	38.48	0.594	-1.067	58.43	41.16	0.445	-0.847	60.96	38.41	0.685	-1.050	57.67	40.78	0.393	-0.796
3	1	2	63.23	36.82	0.606	-1.308	58.64	41.14	0.477	-0.837	61.36	38.36	0.733	-1.078	58.67	40.84	0.499	-0.797
2	2	1	62.09	38.49	0.639	-1.019	59.66	41.38	0.573	-0.708	60.68	39.87	0.736	-0.804	59.73	41.03	0.594	-0.699
2	2	2	62.72	37.58	0.649	-1.192	59.35	41.55	0.538	-0.703	61.12	39.35	0.737	-0.830	59.71	41.12	0.595	-0.688
3	2	1	62.85	38.31	0.651	-1.101	58.98	41.50	0.528	-0.811	61.15	39.10	0.729	-1.025	58.45	40.99	0.514	-0.750
3	2	2	64.15	37.23	0.677	-1.321	59.36	41.49	0.550	-0.815	62.12	38.91	0.767	-0.957	58.69	41.09	0.540	-0.744
2	3	1	65.44	37.86	0.721	-1.101	61.33	42.41	0.540	-0.704	63.21	38.36	0.807	-0.979	61.08	42.24	0.561	-0.685
2	3	2	65.42	37.64	0.714	-1.130	61.24	42.51	0.572	-0.701	63.14	38.29	0.788	-1.019	61.01	42.31	0.568	-0.679
3	3	1	63.29	36.33	0.630	-1.285	57.59	41.51	0.380	-0.848	60.80	38.56	0.702	-1.024	58.08	41.44	0.456	-0.760
3	3	2	63.98	36.58	0.666	-1.226	57.99	41.54	0.445	-0.800	61.42	38.13	0.748	-1.044	58.15	41.17	0.492	-0.784

Tab. 5.6. Wyniki rozpoznawalności izolowanych ramek dla **zbioru uczącego** po zastosowaniu banków transformacji widma. **Nie symulowano** zniekształceń transmisyjnych. Wykorzystano banki filtrów **uwzględniające** zniekształcenia transmisyjne.

li.	war.	war.	miar.oc.rozp.: $c_1^{(rs)}$ dla sam.				miar.oc.rozp.: $c_1^{(rs)}$ dla wsz.				miar.oc.rozp.: $c_2^{(rs)}$ dla sam.				miar.oc.rozp.: $c_2^{(rs)}$ dla wsz.			
kl.	met.	fun.	$c_1^{(rs)}$ [%]		$c_2^{(rs)}$		$c_1^{(rs)}$ [%]		$c_2^{(rs)}$		$c_1^{(rs)}$ [%]		$c_2^{(rs)}$		$c_1^{(rs)}$ [%]		$c_2^{(rs)}$	
	kl.	opt.	sam.	wsz.	sam.	wsz.	sam.	wsz.	sam.	wsz.	sam.	wsz.	sam.	wsz.	sam.	wsz.	sam.	wsz.
2	1	1	62.28	42.67	0.676	-0.614	60.73	44.39	0.578	-0.539	60.75	41.86	0.740	-0.633	60.86	44.06	0.617	-0.530
2	1	2	62.17	41.96	0.725	-0.647	60.91	44.06	0.641	-0.547	60.92	41.59	0.777	-0.648	60.91	43.75	0.682	-0.540
3	1	1	63.87	44.32	0.849	-0.539	62.32	46.10	0.771	-0.454	62.61	44.00	0.899	-0.557	61.94	45.85	0.772	-0.442
3	1	2	63.72	43.45	0.881	-0.584	62.04	45.70	0.805	-0.471	62.94	43.66	0.930	-0.566	62.20	45.46	0.847	-0.459
2	2	1	61.95	42.06	0.679	-0.637	60.67	43.75	0.634	-0.558	60.73	41.75	0.715	-0.645	60.55	43.40	0.660	-0.550
2	2	2	62.25	41.80	0.723	-0.655	60.02	43.62	0.623	-0.569	61.32	41.68	0.790	-0.641	60.23	43.22	0.657	-0.554
3	2	1	63.65	44.23	0.863	-0.541	61.82	46.10	0.786	-0.441	62.10	43.60	0.934	-0.568	61.83	45.88	0.799	-0.433
3	2	2	63.51	43.63	0.879	-0.565	61.47	45.80	0.807	-0.454	62.56	43.53	0.958	-0.569	61.75	45.61	0.832	-0.442
2	3	1	62.07	41.79	0.674	-0.749	60.23	44.86	0.550	-0.532	60.41	41.33	0.712	-0.759	60.11	44.58	0.569	-0.520
2	3	2	62.89	42.73	0.700	-0.699	60.58	45.20	0.599	-0.511	61.36	41.45	0.779	-0.766	60.59	44.98	0.622	-0.505
3	3	1	63.25	43.33	0.822	-0.677	61.34	46.23	0.716	-0.489	61.99	43.38	0.891	-0.655	61.07	45.89	0.724	-0.463
3	3	2	63.21	42.45	0.817	-0.710	61.35	45.69	0.731	-0.504	62.16	43.02	0.901	-0.674	61.32	45.47	0.757	-0.481

Tab. 5.7. Wyniki rozpoznawalności izolowanych ramek dla **zbioru testowego** po zastosowaniu banków transformacji widma. **Nie symulowano** zniekształceń transmisyjnych. Wykorzystano banki filtrów **uwzględniające** zniekształcenia transmisyjne.

li.	war.	war.	miar.oc.rozp.: $c_1^{(rs)}$ dla sam.				miar.oc.rozp.: $c_1^{(rs)}$ dla wsz.				miar.oc.rozp.: $c_2^{(rs)}$ dla sam.				miar.oc.rozp.: $c_2^{(rs)}$ dla wsz.			
kl.	met.	fun.	$c_1^{(rs)}$ [%]		$c_2^{(rs)}$		$c_1^{(rs)}$ [%]		$c_2^{(rs)}$		$c_1^{(rs)}$ [%]		$c_2^{(rs)}$		$c_1^{(rs)}$ [%]		$c_2^{(rs)}$	
	kl.	opt.	sam.	wsz.	sam.	wsz.	sam.	wsz.	sam.	wsz.	sam.	wsz.	sam.	wsz.	sam.	wsz.	sam.	wsz.
2	1	1	60.16	40.87	0.509	-0.712	58.22	42.61	0.401	-0.629	58.73	39.88	0.600	-0.751	58.43	42.21	0.433	-0.616
2	1	2	59.84	38.86	0.466	-1.019	57.03	41.60	0.418	-0.673	57.48	38.41	0.566	-0.846	57.21	41.42	0.446	-0.666
3	1	1	60.45	38.96	0.529	-0.908	57.44	42.06	0.369	-0.723	58.54	38.29	0.593	-0.981	57.70	41.76	0.401	-0.696
3	1	2	60.75	37.97	0.524	-1.056	57.53	41.49	0.460	-0.770	59.07	37.83	0.611	-1.024	57.28	40.70	0.449	-0.760
2	2	1	59.76	40.07	0.517	-0.754	58.18	41.98	0.440	-0.652	58.53	39.74	0.593	-0.754	58.33	41.52	0.484	-0.641
2	2	2	60.63	38.73	0.532	-1.019	57.92	41.96	0.456	-0.653	59.17	39.79	0.614	-0.768	58.29	41.78	0.498	-0.634
3	2	1	60.75	39.85	0.557	-0.842	57.83	42.11	0.442	-0.695	59.39	39.21	0.627	-0.860	58.05	41.68	0.476	-0.673
3	2	2	61.70	38.72	0.533	-1.107	57.33	41.95	0.440	-0.758	60.49	39.10	0.646	-0.878	58.40	41.43	0.494	-0.699
2	3	1	62.53	39.26	0.625	-0.892	59.39	43.03	0.463	-0.641	60.13	38.62	0.683	-0.945	59.29	42.95	0.456	-0.622
2	3	2	63.09	38.23	0.589	-1.054	59.70	43.20	0.497	-0.633	61.01	38.32	0.709	-1.004	59.47	42.84	0.512	-0.622
3	3	1	60.75	37.84	0.474	-1.086	55.85	42.09	0.308	-0.731	58.50	39.04	0.602	-0.930	56.81	41.85	0.382	-0.702
3	3	2	60.27	37.56	0.456	-1.089	56.56	41.52	0.366	-0.774	58.76	37.70	0.581	-1.041	57.21	41.16	0.422	-0.736

Tab. 5.8. Wyniki rozpoznawalności izolowanych ramek dla **zbioru uczącego** po zastosowaniu banków transformacji widma. **Symulowano** zniekształcenia transmisyjne. Wykorzystano banki filtrów **uwzględniające** zniekształcenia transmisyjne.

li.	war.	war.	miar.oc.rozp.: $c_1^{(rs)}$ dla sam.				miar.oc.rozp.: $c_1^{(rs)}$ dla wsz.				miar.oc.rozp.: $c_2^{(rs)}$ dla sam.				miar.oc.rozp.: $c_2^{(rs)}$ dla wsz.			
kl.	met.	fun.	$c_1^{(rs)}$ [%]		$c_2^{(rs)}$		$c_1^{(rs)}$ [%]		$c_2^{(rs)}$		$c_1^{(rs)}$ [%]		$c_2^{(rs)}$		$c_1^{(rs)}$ [%]		$c_2^{(rs)}$	
	kl.	opt.	sam.	wsz.	sam.	wsz.	sam.	wsz.	sam.	wsz.	sam.	wsz.	sam.	wsz.	sam.	wsz.	sam.	wsz.
2	1	1	63.74	41.79	0.769	-0.686	61.54	43.86	0.685	-0.565	61.93	41.62	0.833	-0.657	61.61	43.59	0.713	-0.551
2	1	2	63.69	41.76	0.826	-0.688	61.70	43.53	0.745	-0.574	62.49	41.43	0.875	-0.669	61.61	43.16	0.770	-0.566
3	1	1	65.40	42.65	0.920	-0.671	62.98	45.22	0.837	-0.511	64.02	42.93	1.001	-0.624	62.94	44.84	0.861	-0.496
3	1	2	65.29	42.77	0.970	-0.642	62.86	45.19	0.896	-0.521	64.17	42.50	1.029	-0.639	62.86	44.81	0.917	-0.493
2	2	1	63.55	41.54	0.801	-0.681	61.31	43.31	0.736	-0.583	62.30	41.44	0.850	-0.670	61.25	43.01	0.745	-0.570
2	2	2	64.16	40.97	0.848	-0.759	61.43	43.15	0.764	-0.583	62.94	41.36	0.900	-0.680	61.29	42.81	0.765	-0.572
3	2	1	64.80	42.83	0.947	-0.646	62.43	45.43	0.881	-0.511	63.37	42.82	1.029	-0.632	62.68	45.07	0.908	-0.476
3	2	2	65.34	42.41	0.981	-0.704	62.37	45.29	0.905	-0.508	64.04	42.83	1.067	-0.625	62.44	44.92	0.928	-0.469
2	3	1	63.90	41.50	0.778	-0.741	61.10	44.37	0.674	-0.562	62.37	41.29	0.839	-0.768	60.86	44.10	0.682	-0.544
2	3	2	64.28	41.83	0.790	-0.747	61.45	44.55	0.715	-0.548	62.66	41.09	0.867	-0.768	61.48	44.37	0.739	-0.534
3	3	1	64.86	41.51	0.936	-0.803	62.35	45.58	0.831	-0.519	63.25	42.23	1.000	-0.730	62.09	45.21	0.846	-0.500
3	3	2	64.97	41.45	0.925	-0.801	62.51	45.23	0.868	-0.539	63.86	42.37	1.032	-0.713	62.60	44.81	0.885	-0.511

Tab. 5.9. Wyniki rozpoznawalności izolowanych ramek dla **zbioru testowego** po zastosowaniu banków transformacji widma. **Symulowano** zniekształcenia transmisyjne. Wykorzystano banki filtrów **uwzględniające** zniekształcenia transmisyjne.

li.	war.	war.	miar.oc.rozp.: $c_1^{(rs)}$ dla sam.				miar.oc.rozp.: $c_1^{(rs)}$ dla wsz.				miar.oc.rozp.: $c_2^{(rs)}$ dla sam.				miar.oc.rozp.: $c_2^{(rs)}$ dla wsz.			
kl.	met.	fun.	$c_1^{(rs)}$ [%]		$c_2^{(rs)}$		$c_1^{(rs)}$ [%]		$c_2^{(rs)}$		$c_1^{(rs)}$ [%]		$c_2^{(rs)}$		$c_1^{(rs)}$ [%]		$c_2^{(rs)}$	
	kl.	opt.	sam.	wsz.	sam.	wsz.	sam.	wsz.	sam.	wsz.	sam.	wsz.	sam.	wsz.	sam.	wsz.	sam.	wsz.
2	1	1	61.72	40.15	0.613	-0.783	59.15	42.41	0.478	-0.635	59.85	39.91	0.685	-0.766	59.41	42.00	0.516	-0.637
2	1	2	61.66	38.10	0.570	-1.150	58.43	41.67	0.502	-0.697	59.61	39.08	0.681	-0.843	58.34	41.44	0.524	-0.670
3	1	1	61.91	38.17	0.580	-1.079	57.86	41.86	0.406	-0.751	60.36	38.36	0.708	-0.973	57.76	41.52	0.429	-0.720
3	1	2	62.92	37.43	0.619	-1.160	58.35	41.57	0.519	-0.795	60.84	38.09	0.728	-1.011	57.47	41.09	0.483	-0.756
2	2	1	61.70	40.26	0.661	-0.751	59.11	41.96	0.539	-0.665	60.24	40.38	0.733	-0.740	59.33	41.61	0.592	-0.649
2	2	2	62.85	37.92	0.659	-1.102	59.00	41.91	0.547	-0.670	61.25	39.70	0.753	-0.800	59.01	41.48	0.572	-0.655
3	2	1	62.41	38.78	0.648	-0.971	58.60	42.09	0.518	-0.724	60.88	39.46	0.732	-0.884	58.68	41.48	0.536	-0.695
3	2	2	63.98	38.39	0.688	-1.121	58.73	41.93	0.555	-0.757	62.18	38.95	0.769	-0.930	59.16	41.30	0.568	-0.722
2	3	1	64.61	38.98	0.715	-0.924	60.21	42.92	0.541	-0.647	61.88	38.81	0.773	-0.908	60.48	42.77	0.572	-0.627
2	3	2	64.95	37.98	0.686	-1.090	60.49	43.12	0.544	-0.630	61.94	38.20	0.781	-0.979	60.57	42.90	0.555	-0.619
3	3	1	62.80	36.66	0.589	-1.258	57.27	41.99	0.393	-0.755	60.72	38.55	0.736	-0.953	57.53	41.73	0.440	-0.727
3	3	2	63.32	37.27	0.628	-1.137	57.63	41.72	0.455	-0.766	60.93	38.92	0.754	-0.923	57.89	41.39	0.500	-0.746

5.1. Wyniki rozpoznawalności izolowanych ramek

W tabeli 5.1 przedstawiono wyniki rozpoznawalności izolowanych ramek sygnału mowy w przypadku zastosowania optymalnej transformacji widma dla każdego mówcy. W tabelach od 5.2 do 5.9 przedstawiono natomiast wyniki uzyskane z zastosowaniem banków transformacji widma z algorytmem iteracyjnego poprawiania wyniku rozpoznania. W tabelach zastosowano następujące skróty: „li.kl." - liczba klas mówców, „war.met.kl." - wariant metody wyznaczania klas mówców (zob. rozdział 4.4), „war.fun.opt." - wariant funkcji celu użytej w optymalizacji wartości parametrów transformacji widma, wykorzystywanych w konstrukcji banków, 1 oznacza tutaj funkcję $c_1^{(ro)}$, a 2 - funkcję $c_2^{(ro)}$ (zob. rozdział 4.3.2), „miar.oc.rozp." - miara oceny rozpoznania zastosowana podczas wyboru klas mówców i elementów banków transformacji widma (zob. rozdział 4.6.2). Wartości rozpoznawalności wyznaczano dla samogłosek („sam.") lub wszystkich fonemów („wsz."). Metodologia wyznaczania wartości rozpoznawalności izolowanych ramek została opisana w dodatku E. Zniekształcenia transmisyjne symulowano analogicznie, jak w przypadku testowania metody EV (zob. rozdział 3.2).

Porównując wyniki uzyskane dla zmodyfikowanej metody EV (zob. rozdział 3.2) z wynikami przedstawionymi w tabeli 5.1 można stwierdzić, że dla zbioru uczącego metoda EV i zaproponowana metoda transformacji widma dają podobne rezultaty, przy zbliżonej liczbie parametrów, których wartości należy wyznaczyć dla danego mówcy, a która w przypadku metody transformacji widma równa jest 7. Dla mówców ze zbioru testowego wyniki uzyskane metodą transformacji widma okazały się lepsze niż uzyskane metodą EV, z wyjątkiem wartości $c_1^{(rs)}$ dla wszystkich fonemów, która była nieco niższa niż osiągnięta z wykorzystaniem metody EV. Można zatem uznać, że wymagania postawione zaprojektowanej metodzie transformacji widma zasadniczo zostały spełnione.

W przypadku wyników uzyskanych dla banków transformacji widma, zauważyć można następujące fakty:

- W przypadku zbioru uczącego najlepsze rezultaty zapewniło zastosowanie trzech klas mówców, a w przypadku zbioru testowego - dwóch klas. Widać zatem, że w przypadku trzech klas generalizacja statystycznego modelu akustycznego na zbiór testowy była niewystarczająca, co wiąże się z małą licznością zbioru uczącego.

- Najlepsze rezultaty uzyskano w większości stosując wariant 3. i 2. metody wyznaczania klas mówców i rozkładów prawdopodobieństwa współczynników MFCC w klasach. Potwierdza to celowość wprowadzenia metody uczenia typu SAT zastosowanej w tych wariantach.

- W przypadku symulowania zniekształceń transmisyjnych najlepsze rezultaty

uzyskano w większości z zastosowaniem banków filtrów zmodyfikowanych tak, by uwzględniały te zniekształcenia (zob. rozdział 4.5.3), co potwierdza skuteczność zaproponowanej modyfikacji.

- Wyniki uzyskane z wykorzystaniem banków filtrów, w porównaniu z wynikami osiągniętymi dla transformacji optymalnych dla każdego mówcy, są przeważnie nieco wyższe w przypadku wartości rozpoznawalności ($c_1^{(rs)}$ dla sam., $c_1^{(rs)}$ dla wsz., $c_2^{(rs)}$ dla sam. lub $c_2^{(rs)}$ dla wsz.) zgodnej z zastosowaną podczas wyboru klas mówców i elementów banków miarą oceny rozpoznania. Jednocześnie dla pozostałych trzech wartości rozpoznawalności wyniki są na ogół nieco gorsze. Prawidłowość ta jest wyraźniejsza dla zbioru testowego.

Biorąc pod uwagę 16 przypadków odpowiadających wszystkim kombinacjom:

- czterech różnych miar oceny rozpoznania ($c_1^{(rs)}$ dla sam., $c_1^{(rs)}$ dla wsz., $c_2^{(rs)}$ dla sam. i $c_2^{(rs)}$ dla wsz.)
- rozpoznawania bez symulowanych zniekształceń transmisyjnych i z ich symulacją
- rozpoznawania dla mówców ze zbioru uczącego i testowego

sprawdzono, które warianty banków transformacji widma przyniosły najlepsze rezultaty w każdym z przypadków. W tabeli 5.10 podano częstość występowania najlepszych wariantów. Trzy najczęściej występujące oznaczono jako T1, T2 i T3 i zastosowano dalej w rozpoznawaniu komend.

Tab. 5.10. Warianty banków transformacji widma zapewniające najwyższe rozpoznawalności izolowanych ramek.

wariant			liczba	oznaczenie
li.kl.	war.met.kl.	war.fun.opt.	wystąpień	
3	2	2	4	T1
2	3	2	4	T2
2	3	1	4	T3
3	3	1	2	
3	3	2	1	
3	1	1	1	

Tab. 5.11. Wyniki rozpoznawalności komend po zastosowaniu banków transformacji widma. **Nie symulowano** zniekształceń transmisyjnych. W nawiasach podano wyniki uzyskane przy zastosowaniu przyporządkowywania mówców do klas na podstawie $f^{(v)}$. Czcionką pogrubioną zaznaczano najwyższy wynik dla danego wariantu systemu ARM i danego zbioru mówców.

war. sys. ARM	zb. mow.	it. pop. roz.	bez kompensacji	wariant T1	wariant T2	wariant T3	wariant T1 zmodyf.	wariant T2 zmodyf.	wariant T3 zmodyf.
A	ucz.	n	96.71	96.04 (96.66)	96.60 (95.92)	97.56 (96.66)	93.37 (94.84)	94.33 (92.92)	95.35 (93.88)
		t	-	96.04 (96.66)	96.77 (96.20)	**97.68** (96.54)	93.14 (95.30)	93.76 (92.80)	95.40 (94.67)
	test.	n	93.21	96.04 (96.38)	96.04 (94.70)	**96.97** (96.12)	92.59 (92.93)	93.77 (92.26)	95.20 (93.52)
		t	-	96.30 (96.46)	96.22 (94.78)	96.88 (96.21)	92.76 (93.18)	93.68 (93.02)	95.20 (93.78)
At	ucz.	n	97.62	**98.81** (98.13)	97.68 (97.28)	98.30 (97.62)	98.64 (98.13)	97.22 (96.72)	97.28 (96.71)
		t	-	98.76 (98.42)	98.13 (97.56)	98.70 (97.90)	98.53 (98.30)	97.56 (97.16)	97.84 (97.05)
	test.	n	93.98	95.12 (95.04)	96.30 (95.96)	95.79 (95.28)	94.86 (94.70)	95.79 (95.28)	96.55 (95.62)
		t	-	95.37 (95.37)	96.30 (95.96)	96.04 (95.88)	95.12 (95.03)	96.30 (95.79)	**96.72** (95.29)
B	ucz.	n	99.51	99.26 (98.64)	98.98 (98.13)	99.44 (98.36)	98.92 (98.58)	98.52 (97.45)	99.04 (97.73)
		t	-	99.60 (99.32)	99.20 (98.42)	**99.66** (98.36)	99.10 (99.10)	98.36 (97.96)	99.04 (97.73)
	test.	n	98.71	97.05 (96.80)	98.14 (97.56)	98.56 (98.40)	96.88 (96.72)	97.56 (97.48)	98.32 (98.06)
		t	-	97.39 (97.14)	98.32 (97.81)	**98.82** (98.74)	96.72 (96.55)	97.64 (97.72)	98.40 (97.90)
Bt	ucz.	n	98.87	**99.66** (99.32)	98.64 (98.19)	99.09 (98.75)	99.55 (99.32)	97.96 (97.85)	98.41 (98.19)
		t	-	99.55 (99.21)	98.75 (98.41)	99.21 (98.87)	99.21 (99.09)	98.19 (97.96)	98.53 (98.07)
	test.	n	97.14	97.47 (97.47)	98.82 (98.82)	98.82 (98.32)	96.97 (97.14)	98.65 (98.32)	98.48 (97.98)
		t	-	97.81 (97.81)	**98.99** (**98.99**)	98.65 (98.32)	96.80 (96.97)	98.82 (98.65)	98.32 (97.81)

Tab. 5.12. Wyniki rozpoznawalności komend po zastosowaniu banków transformacji widma. **Symulowano** zniekształcenia transmisyjne. W nawiasach podano wyniki uzyskane przy zastosowaniu przyporządkowywania mówców do klas na podstawie $f^{(v)}$. Czcionką pogrubioną zaznaczano najwyższy wynik dla danego wariantu systemu ARM i danego zbioru mówców.

war. sys. ARM	zb. mow.	it. pop. roz.	bez kompensacji	wariant T1	wariant T2	wariant T3	wariant T1 zmodyf.	wariant T2 zmodyf.	wariant T3 zmodyf.
A	ucz.	n	92.63	93.08 (94.50)	94.31 (94.20)	95.12 (94.39)	93.37 (95.29)	94.26 (94.33)	**95.67** (95.39)
		t	-	93.35 (94.88)	94.31 (94.42)	95.24 (94.58)	93.29 (95.35)	93.88 (93.80)	95.48 (95.48)
	test.	n	91.24	94.67 (95.37)	94.61 (94.28)	94.95 (94.98)	93.74 (94.42)	94.86 (93.91)	**96.27** (95.43)
		t	-	94.75 (95.45)	94.70 (94.53)	94.89 (95.18)	93.46 (94.36)	94.50 (94.08)	96.16 (95.20)
At	ucz.	n	92.65	97.09 (96.32)	95.96 (95.54)	96.20 (95.80)	98.34 (97.70)	97.07 (96.86)	97.32 (96.85)
		t	-	97.26 (96.71)	96.54 (96.13)	96.66 (96.26)	**98.42** (98.24)	97.45 (97.17)	97.72 (97.15)
	test.	n	90.29	92.84 (92.37)	94.56 (94.30)	94.30 (94.05)	94.36 (94.33)	95.96 (95.12)	95.73 (95.28)
		t	-	93.44 (93.15)	95.00 (94.78)	94.56 (94.75)	94.72 (94.72)	**96.10** (95.79)	96.07 (95.74)
B	ucz.	n	97.54	97.66 (97.09)	97.56 (97.72)	98.34 (97.73)	98.94 (98.56)	98.36 (97.60)	**98.96** (97.87)
		t	-	97.88 (97.43)	97.60 (97.49)	98.40 (97.85)	98.90 (98.62)	98.02 (97.51)	98.79 (97.66)
	test.	n	97.21	95.37 (95.20)	96.77 (97.14)	97.50 (97.64)	96.41 (96.24)	97.62 (97.67)	**98.34** (98.15)
		t	-	95.99 (95.76)	96.72 (96.94)	97.64 (97.81)	96.41 (96.30)	97.34 (97.56)	98.29 (98.09)
Bt	ucz.	n	97.73	98.30 (97.69)	98.15 (97.96)	98.30 (97.88)	**99.47** (99.10)	98.37 (98.08)	98.56 (98.34)
		t	-	98.56 (98.19)	98.15 (97.96)	98.26 (97.88)	99.32 (99.21)	98.49 (98.11)	98.53 (98.00)
	test.	n	94.89	95.68 (95.57)	97.31 (97.70)	97.03 (97.19)	96.69 (96.63)	97.87 (98.15)	**98.26** (98.04)
		t	-	96.12 (95.96)	97.59 (97.92)	97.25 (97.64)	96.52 (96.97)	98.09 (**98.26**)	97.92 (98.09)

5.2. Wyniki rozpoznawania komend

W tabelach 5.11 i 5.12 przedstawiono wyniki rozpoznawania komend z zastosowaniem banków transformacji widma za pomocą czterech wariantów systemu ARM. Z wyjątkiem wariantu Bt, procedurę uczenia systemu powtarzano dwukrotnie i podano wyniki uśrednione. W przypadku symulowania zniekształceń transmisyjnych nakładano je losowo dla każdej wypowiedzi analogicznie, jak podczas testowania metody EV (zob. rozdział 3.2). Dodatkowo wyniki rozpoznania z symulowanymi zniekształceniami uśredniano z trzech powtórzeń rozpoznawania, z których w każdym przeprowadzano nowe losowanie zniekształceń. Odchylenie standardowe wyników związane z rozpoznawaniem przeprowadzanym wielokrotnie, wyznaczone dla wszystkich badanych wariantów, wyniosło średnio 0.43%. W tabelach użyto następujących skrótów: „war.sys.ARM" - wariant systemu ARM (zob. dodatek C), „zb.mow." - zbiór mówców („ucz." - uczący, „test." - testowy), „it.pop.roz." - iteracyjne poprawianie wyniku rozpoznania (zob. rozdział 4.6). Zastosowane warianty T1, T2 i T3 banków transformacji widma podane zostały w rozdziale poprzednim, „zmodyf." oznacza natomiast, że banki filtrów zostały zmodyfikowane tak, by uwzględniały zniekształcenia transmisyjne (zob. rozdział 4.5.3).

Na podstawie uzyskanych wyników podać można następujące spostrzeżenia:

- W przypadku braku symulowania zniekształceń transmisyjnych, w zależności od wariantu systemu ARM, osiągnięto spadek błędu rozpoznania o od 29% do 70% dla zbioru uczącego i od 8% do 65% dla zbioru testowego. W przypadku symulowania zniekształceń spadek ten wynosił odpowiednio: od 41% do 78 % i od 40% do 66%.

- Największą poprawę rozpoznawalności po zastosowaniu banków transformacji widma odnotowano dla wariantów At i Bt systemu, co wynika z faktu, że w tych wariantach dużą wagę w mierze oceny rozpoznania mają prawdopodobieństwa uzyskane bezpośrednio z modelu akustycznego, podczas gdy w wariantach A i B waga ta jest mniejsza. W kompensacji modyfikowane jest bowiem widmo sygnału, a zmiany wywołane tymi modyfikacjami bezpośrednio wpływają na prawdopodobieństwa uzyskane modelu akustycznego.

- W przypadku wariantów systemu A i B najlepsze rezultaty, zarówno dla zbioru uczącego, jak i testowego, osiągnięto stosując wariant T3 banków transformacji widma. W przypadku wariantów At i Bt dla zbioru uczącego najlepszy okazał się bank T1, natomiast dla zbioru testowego dla wariantu At - T3 i T2 odpowiednio przy braku symulowania zniekształceń transmisyjnych i przy ich symulacji, a dla wariantu Bt - T2.

- Przy braku symulowania zniekształceń transmisyjnych najlepsze rezultaty uzy-

skano, poza jednym przypadkiem, stosując banki niezmodyfikowane, a w przypadku symulacji zniekształceń najbardziej efektywne okazały się, zgodnie z oczekiwaniami, banki zmodyfikowane.

- Biorąc pod uwagę najlepsze rezultaty uzyskane w przypadku standardowej metody przyporządkowywania mówców do klas i braku symulacji zniekształceń transmisyjnych, w 5 na 8 przypadkach iteracyjne poprawianie wyniku rozpoznania przyniosło pozytywny rezultat. W przypadku symulowania zniekształceń pozytywny rezultat osiągnięto w zaledwie 2 na 8 przypadków. Stosując natomiast przyporządkowywanie mówców do klas na podstawie wartości $f^{(v)}$, zarówno przy braku symulacji, jak i w przypadku symulowania zniekształceń transmisyjnych, zastosowanie iteracyjnego poprawiania dało pozytywny wynik w 7 na 8 przypadków.

- Najlepsze rezultaty uzyskane z zastosowaniem przyporządkowywania mówców do klas na podstawie wartości $f^{(v)}$ były niższe od rezultatów uzyskanych dla przyporządkowywania standardowego średnio o 0.43% i 0.29% (podano spadki wartości bezwzględnych rozpoznania) odpowiednio: przy braku symulacji zniekształceń transmisyjnych i w przypadku ich symulowania. Spadki te są zatem relatywnie niewielkie.

6. Podsumowanie

Celem niniejszej pracy było zaprojektowanie algorytmu efektywnej kompensacji warunków transmisyjnych i cech osobniczych mówcy dla systemu rozpoznawania bardzo krótkich i izolowanych wypowiedzi. Aby ten cel osiągnąć zostały zrealizowane następujące zadania:

1. Opracowano narzędzia badawcze w postaci systemu ARM (zob. dodatek C) bazującego na parametryzacji MFCC i statystycznych modelach języka wykorzystujących HMM. Zaprojektowano cztery warianty takiego systemu, w których wprowadzonymi przez Autora oryginalnymi elementami są m.in.:

   - algorytm VAD oparty na analizie zmian w czasie energii w podpasmach sygnału,
   - algorytm wykrywania granic pseudolylab i modyfikacja algorytmu Viterbiego tak, aby uwzględniane były w nim te granice,
   - metoda rozpoznawania dwuetapowego bazująca na analizie sekwencji pseudosylab,
   - algorytmy uczenia modelu statystycznego systemu z zastosowaniem funkcji celu uwzględniających zdolności klasyfikacji tego systemu.

   Ponadto, na potrzeby prowadzonych badań, wykonano zasadniczą część prac związanych z przygotowaniem bazy nagrań mowy polskiej bnITTA (zob. dodatek B), w tym: wykonanie części nagrań, opracowanie oryginalnych narzędzi do półautomatycznej segmentacji i etykietyzacji, przeprowadzenie segmentacji i etykietyzacji części nagrań.

2. Przeprowadzono analizę przyczyn występowania zniekształceń transmisyjnych i zmienności cech osobniczych mówcy, jak również ich wpływu na parametry MFCC oraz skuteczność działania systemu ARM opartego na statystycznych modelach języka (zob. rozdział 2).

3. Dokonano przeglądu znanych z literatury przedmiotu rozwiązań zagadnienia kompensacji cech osobniczych i zniekształceń transmisyjnych oraz przeprowa-

dzono analizę teoretyczną, a w przypadku algorytmu EV również eksperymentalną, ich przydatności w rozwiązywanym zagadnieniu (zob. rozdział 2.5 i 3).

4. Zaproponowano oryginalną modyfikację algorytmu EV poszerzającą możliwość jego stosowania również do kompensacji liniowych zniekształceń transmisyjnych oraz zaproponowano zastosowanie wariantu metody SAT uczenia modelu dla zmodyfikowanego algorytmu EV (zob. rozdział 3).

5. Zaprojektowano oryginalną metodę kompensacji liniowych zniekształceń transmisyjnych i cech osobniczych mówcy opartą na bankach transformacji widma sygnału (zob. rozdział 4), a jej istotniejszymi elementami są:

   - transformacja widma wykorzystująca skalowanie osi częstotliwości i filtrację liniową,
   - algorytm wyznaczania optymalnych wartości parametrów transformacji dla danego mówcy,
   - algorytm podziału mówców na klasy i wyznaczania rozkładów prawdopodobieństwa współczynników MFCC w klasach,
   - algorytm konstrukcji banków transformacji,
   - algorytm rozpoznawania mowy z wykorzystaniem banków transformacji widma i iteracyjnego poprawiania wyniku rozpoznania,
   - metoda przyporządkowywania mówców do klas na podstawie wartości częstotliwości tonu krtaniowego, szacowanej za pomocą zaproponowanego szybkiego algorytmu.

Na podstawie wyników przeprowadzonych badań nie można wskazać jednego wariantu zaproponowanej metody banków transformacji widma, który zapewniałby najlepsze rezultaty jednocześnie we wszystkich zastosowanych wariantach pomiaru rozpoznawalności komend i izolowanych ramek sygnału mowy. Miarodajną ocenę skuteczności algorytmu kompensacji zapewniają wyniki rozpoznawalności komend, uzyskane dla mówców ze zbioru testowego w warunkach symulacji zniekształceń transmisyjnych. W tym przypadku najskuteczniejsza okazała się kompensacja z zastosowaniem dwóch klas mówców wybranych za pomocą trzeciego wariantu algorytmu podziału mówców na klasy oraz banków filtrów wyznaczonych metodą mającą na celu polepszenie ich zdolności kompensacji zniekształceń transmisyjnych.

Osiągnięta z zastosowaniem tego wariantu metody transformacji widma poprawa rozpoznawalności izolowanych ramek dla mówców ze zbioru testowego, mierzona czterema różnymi miarami oceny, była średnio o 88% wyższa w stosunku do poprawy osiągniętej za pomocą algorytmu EV, przy zachowaniu zbliżonych warunków pomiaru. Natomiast spadek błędu rozpoznania komend dla mówców ze zbioru testowego, przy symulowaniu zniekształceń transmisyjnych, wyniósł w zależności od

wariantu systemu ARM od 39% do 60% w przypadku standardowej metody przyporządkowywania mówców do klas i zastosowania iteracyjnego poprawiania wyniku rozpoznania. Przyporządkowywanie mówców do klas na podstawie wartości częstotliwości tonu ktraniowego skutkowało spadkiem rozpoznawalności bezwzględnej średnio o 0.33%.

Wyniki poprawy rozpoznawalności komend i izolowanych ramek, uzyskane za pomocą zaproponowanej przez Autora oryginalnej metody banków transformacji widma, są satysfakcjonujące. Metoda ta cechuje się również dużą uniwersalnością, gdyż jej działanie sprowadza się do modyfikacji widma amplitudowego ramek sygnału mowy, którego wyznaczanie jest etapem wspólnym dla większości współczesnych systemów ARM. W celu dostosowania metody do użycia jej w danym systemie wymagane jest jedynie sformułowanie, charakterystycznej dla tego systemu, miary oceny rozpoznania wypowiedzi. Pozytywne wyniki osiągnięte dla czterech zaprojektowanych w niniejszej pracy wariantów systemu ARM sugerują, że z dużym prawdopodobieństwem można je uogólnić także na inne systemy.

Bibliografia

[1] A. Acero, X. Huang, "Augmented Cepstral Normalization for Robust Speech Recognition," w *Proc. IEEE Workshop on Automatic Speech Recognition*, 1995.

[2] A. Acero, X. Huang, "Speaker and Gender Normalization for Continuous-Density Hidden Markov Models," w *Proc. ICASSP*, 1996.

[3] A. Acero, "Acoustical and Environmental Robustness in Automatic Speech Recognition," praca doktorska, Carnegie Mellon University, Pittsburgh, 1990.

[4] M. Afify, "Accurate Compensation in the Log-Spectral Domain for Noisy Speech Recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 13, no. 3, May 2005.

[5] M. Afify, O. Siohan, "Sequential Estimation with Optimal Forgetting for Robust Speech Recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 12, no. 1, January 2004.

[6] L. Apostol, P. Perrier, M. Baciu, C. Segebarth, P. Badin, "Using the Formant/Cavity Affiliation to Study the Inter-Speaker Variability: Assessment from MRI Data," w *Proc. 5th Speech Production Seminar*, 2000.

[7] J. Arabas, *Wykłady z algorytmów ewolucyjnych.* WNT, Warszawa, 2004.

[8] C. Avendano, S. van Vuuren, H. Hermansky, "Data-Based RASTA-Like Filter Design for Channel Normalization in ASR," w *Proc. ICSLP*, 1996.

[9] C. Basztura, *Rozmawiać z komputerem.* Wydawnictwo Format, Wrocław, 1992.

[10] C. Becchetti, L. P. Ricotti, *Speech Recognition. Theory and C++ Implementation.* John Wiley & Sons, 1999.

[11] A. Ben-Yishai, D. Burshtein, "A Discriminative Training Algorithm for Hidden Markov Models," *IEEE Trans. on Speech and Audio Processing*, vol. 12, no. 3, May 2004.

[12] L. Benaroya, F. Bimbot, G. Gravier, R. Gribonval, "Experiments in Audio Source Separation with One Sensor for Robust Speech Recognition," *Speech Communication*, vol. 48, 2006.

[13] J. Bilmes, "What HMMs Can Do," University of Washington, Tech. Rep. UWEETR-2002-0003, 2002.

[14] C. Cerisara, L. Rigazio, J.-C. Junqua, "α-Jacobian Environmental Adaptation," *Speech Communication*, vol. 42, 2004.

[15] R. Chelouah, P. Siarry, "Genetic and Nelder–Mead Algorithms Hybridized for a More Accurate Global Optimization of Continuous Multiminima Functions," *European Journal of Operational Research*, vol. 148, 2003.

[16] K.-T. Chen, H.-M. Wang, "Eigenspace-Based Linear Transformation Approach for Rapid Speaker Adaptation," w *Proc. ISCA Tutorial and Research Workshop on Adaptation Methods for Speech Recognition*, 2001.

[17] S. S. Chen, P. DeSouza, "Speaker Adaptation by Correlation (ABC)," w *Proc. Eurospeech*, 1997.

[18] S.-M. Chi, Y.-H. Oh, "Lombard Effect Compensation and Noise Suppression for Noisy Lombard Speech Recognition," w *Proc. ICSLP*, 1996.

[19] J.-T. Chien, "Adaptive Hierarchy of Hidden Markov Models for Transformation-Based Adaptation," *Speech Communication*, vol. 36, 2002.

[20] J.-T. Chien, C.-H. Huang, "Aggregate a Posteriori Linear Regression Adaptation," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 14, no. 3, May 2006.

[21] J.-T. Chien, H.-C. Wang, L.-M. Lee, "Estimation of Channel Bias for Telephone Speech Recognition," w *Proc. ICSLP*, 1996.

[22] X. Cui, A. Alwan, "Robust Speaker Adaptation by Weighted Model Averaging Based on the Minimum Description Length Criterion," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 15, no. 2, February 2007.

[23] A. de la Torre, A. M. Peinado, J. C. Segura, J. L. Perez-Cordoba, M. C. Benitez, A. J. Rubio, "Histogram Equalization of Speech Representation," *IEEE Trans. on Speech and Audio Processing*, vol. 13, no. 3, May 2005.

[24] J. de Veth, L. Boves, "Phase Corrected Rasta for Automatic Speech Recognition over the Phone," w *Proc. ICASSP*, 1997.

[25] S. Deligne, S. Dharanipragada, R. Gopinath, B. Maison, P. Olsen, H. Printz, "A Robust High Accuracy Speech Recognition System for Mobile Applications," *IEEE Trans. on Speech and Audio Processing*, vol. 10, no. 8, November 2002.

[26] S. Deligne, R. Gopinath, "Robust Speech Recognition with Multi-Channel Codebook Dependent Cepstral Normalization (MCDCN)," w *Proc. ASRU*, 2001.

[27] L. Deng, D. Yu, A. Acero, "A Bidirectional Target-Filtering Model of Speech Coarticulation and Reduction: Two-Stage Implementation for Phonetic Recognition," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 14, no. 1, January 2006.

[28] L. Deng, J. Droppo, A. Acero, "Enhancement of Log Mel Power Spectra of Speech Using a Phase-Sensitive Model of the Acoustic Environment and Sequential Estimation of the Corrupting Noise," *IEEE Trans. on Speech and Audio Processing*, vol. 12, no. 2, March 2004.

[29] O. Deshmukh, C. Y. Espy-Wilson, A. Salomon, J. Singh, "Use of Temporal Information: Detection of Periodicity, Aperiodicity, and Pitch in Speech," *IEEE Trans. on Speech and Audio Processing*, vol. 13, no. 5, September 2005.

[30] S. Dharanipragada, U. H. Yapanel, B. D. Rao, "Robust Feature Extraction for Continuous Speech Recognition Using the MVDR Spectrum Estimation Method," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 15, no. 1, January 2007.

[31] D. Dimitriadis, P. Maragos, "Continuous Energy Demodulation Methods and Application to Speech Analysis," *Speech Communication*, vol. 48, 2006.

[32] P. L. Dognin, "A Bandpass Transform for Speaker Normalization," praca doktorska, University of Pittsburgh, 2003.

[33] S.-J. Doh, R. M. Stern, "Inter-Class MLLR for Speaker Adaptation," w *Proc. ICASSP*, 2000.

[34] J. Dulas, "Metoda siatek o zmiennych parametrach w zastosowaniu do rozpoznawania fonemów mowy polskiej," praca doktorska, Politechnika Opolska, 2002.

[35] G. Evangelopoulos, P. Maragos, "Multiband Modulation Energy Tracking for Noisy Speech Detection," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 14, no. 6, November 2006.

[36] F. A. Everest, *The Master Handbook of Acoustics. Fourth Edition.* McGraw-Hill, 2001.

[37] W. Findeisen, J. Szymanowski, A. Wierzbicki, *Teoria i metody obliczeniowe optymalizacji.* PWN, Warszawa, 1980.

[38] R. Fletcher, *Practical Methods of Optimization. Volume 1 and 2.* John Wiley & Sons, 1981.

[39] L. E. Franks, *Teoria sygnałów.* PWN, Warszawa, 1975.

[40] S. Furui, *Digital Speech Processing, Synthesis, and Recognition. Second Edition, Revised and Expanded.* Marcel Dekker, New York, 2001.

[41] B. Gajic, K. K. Paliwal, "Robust Speech Recognition in Noisy Environments Based on Subband Spectral Centroid Histograms," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 14, no. 2, March 2006.

[42] M. J. F. Gales, "Cluster Adaptive Training of Hidden Markov Models," *IEEE Trans. on Speech and Audio Processing*, vol. 8, no. 4, July 2000.

[43] A. Gallardo-Antolín, C. Peláez-Moreno, F. D. de María, "Recognizing GSM Digital Speech," *IEEE Trans. on Speech and Audio Processing*, vol. 13, no. 6, November 2005.

[44] Y. Gong, "A Method of Joint Compensation of Additive and Convolutive Distortions for Speaker-Independent Speech Recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 13, no. 5, September 2005.

[45] E. B. Gouvea, R. M. Stern, "Speaker Normalization Through Formant-Based Warping of the Frequency Scale," w *Proc. Eurospeech*, 1997.

[46] S. Grocholewski, *Baza nagrań sygnałów mowy CORPORA. Instrukcja użytkowania.* Politechnika Poznańska, Instytut Informatyki, Poznań, 1997.

[47] S. Grocholewski, "Statystyczne podstawy systemu ARM dla języka polskiego," rozprawa habilitacyjna, Politechnika Poznańska, 2001.

[48] A. Gunawardana, W. Byrne, "Discriminative Speaker Adaptation with Conditional Maximum Likelihood Linear Regression," w *Proc. Eurospeech*, 2001.

[49] T. Hain, P. C. Woodland, G. Evermann, M. J. F. Gales, X. Liu, G. L. Moore, D. Povey, *et al.*, "Automatic Transcription of Conversational Telephone Speech," *IEEE Trans. on Speech and Audio Processing*, vol. 13, no. 6, November 2005.

[50] S. Harding, J. Barker, G. J. Brown, "Mask Estimation for Missing Data Speech Recognition Based on Statistics of Binaural Interaction," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 14, no. 1, January 2006.

[51] R. Hariharan, O. Viikki, "An Integrated Study of Speaker Normalization and HMM Adaptation for Noise Robust Speaker-Independent Speech Recognition," *Speech Communication*, vol. 37, 2002.

[52] S. Haykin, *Adaptive Filter Theory. Second Edition.* Prentice-Hall, 1991.

[53] H. Hermansky, N. Morgan, A. Bayya, P. Kohn, "RASTA-PLP Speech Analysis," US West Advanced Technologies with International Computer Science Institute, Tech. Rep. TR-91-069, 1991.

[54] F. Hilger, H. Ney, "Quantile Based Histogram Equalization for Noise Robust Large Vocabulary Speech Recognition," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 14, no. 3, May 2006.

[55] M. Holmberg, D. Gelbart, U. Ramacher, W. Hemmert, "Automatic Speech Recognition with Neural Spike Trains," w *Proc. Eurospeech*, 2005.

[56] C. Huang, T. Chen, S. Li, E. Chang, J. Zhou, "Analysis of Speaker Variability," w *Proc. Eurospeech*, 2001.

[57] C.-H. Huang, J.-T. Chien, H.-M. Wang, "A New Eigenvoice Approach to Speaker Adaptation," w *Proc. International Symposium on Chinese Spoken Language Processing*, 2004.

[58] J. M. Huerta, "Alignment-Based Codeword-Dependent Cepstral Normalization," *IEEE Trans. on Speech and Audio Processing*, vol. 10, no. 7, October 2002.

[59] D. G. Humphrey, J. R. Wilson, "A Revised Simplex Search Procedure for Stochastic Simulation Response-Surface Optimization," w *Proc. of the 1998 Winter Simulation Conference*, 1998.

[60] J.-W. Hung, L.-S. Lee, "Optimization of Temporal Filters for Constructing Robust Features in Speech Recognition," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 14, no. 3, May 2006.

[61] J.-W. Hung, J.-L. Shen, L.-S. Lee, "New Approaches for Domain Transformation and Parameter Combination for Improved Accuracy in Parallel Model Combination (PMC) Techniques," *IEEE Trans. on Speech and Audio Processing*, vol. 9, no. 8, November 2001.

[62] T. Irino, R. D. Patterson, "Segregating Information About the Size and Shape of the Vocal Tract Using Time-Domain Auditory Model: The Stabilised Wavelet-Mellin Transform," *Speech Communication*, vol. 36, 2002.

[63] T. Irino, R. D. Patterson, "A Dynamic Compressive Gammachirp Auditory Filterbank," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 14, no. 6, November 2006.

[64] C. R. Jankowski Jr., H. H. Vo, R. P. Lippmann, "A Comparison of Signal Processing Front Ends for Automatic Word Recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 3, no. 4, July 1995.

[65] W. Jassem, *Podstawy fonetyki akustycznej.* PWN, Warszawa, 1973.

[66] H. Jiang, "Confidence Measures for Speech Recognition: A Survey," *Speech Communication*, vol. 45, 2005.

[67] M. T. Johnson, R. J. Povinelli, A. C. Lindgren, J. Ye, X. Liu, K. M. Indrebo, "Time-Domain Isolated Phoneme Classification Using Reconstructed Phase Spaces," *IEEE Trans. on Speech and Audio Processing*, vol. 13, no. 4, July 2005.

[68] B.-H. Juang, W. Chou, C.-H. Lee, "Minimum Classification Error Rate Methods for Speech Recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 5, no. 3, May 1997.

[69] T. Kailath, "The Divergence and Bhattacharyya Distance Measures in Signal Selection," *IEEE Trans. on Communication Technology*, vol. 15, no. 1, February 1967.

[70] M. Karnjanadecha, S. A. Zahorian, "Signal Modeling for High-Performance Robust Isolated Word Recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 9, no. 6, September 2001.

[71] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory.* Prentice-Hall, 1993.

[72] P. Kenny, G. Boulianne, P. Ouellet, P. Dumouchel, "Speaker Adaptation Using an Eigenphone Basis," *IEEE Trans. on Speech and Audio Processing*, vol. 12, no. 6, November 2004.

[73] M. Kepesi, L. Weruaga, "Adaptive Chirp-Based Time–Frequency Analysis of Speech Signals," *Speech Communication*, vol. 48, 2006.

[74] A. Kiełbasiński, H. Schwetlick, *Numeryczna algebra liniowa.* WNT, Warszawa, 1992.

[75] D. K. Kim, N. S. Kim, "Maximum a Posteriori Adaptation of HMM Parameters Based on Speaker Space Projection," *Speech Communication*, vol. 42, 2004.

[76] D. K. Kim, N. S. Kim, "Rapid Online Adaptation Using Speaker Space Model Evolution," *Speech Communication*, vol. 42, 2004.

[77] N. S. Kim, "Feature Domain Compensation of Nonstationary Noise for Robust Speech Recognition," *Speech Communication*, vol. 37, 2002.

[78] Y. Kim, "Maximum-Likelihood Affine Cepstral Filtering (MLACF) Technique for Speaker Normalization," w *Proc. Eurospeech*, 2001.

[79] I. Kokkinos, P. Maragos, "Nonlinear Speech Analysis Using Models for Chaotic Systems," *IEEE Trans. on Speech and Audio Processing*, vol. 13, no. 6, November 2005.

[80] P. Kłosowski, "Usprawnienie procesu rozpoznawania mowy w oparciu o fonetykę i fonologię języka polskiego," praca doktorska, Politechnika Śląska, 2000.

[81] T. T. Kristjansson, "Speech Recognition in Adverse Environments: A Probabilistic Approach," praca doktorska, University of Waterloo, 2002.

[82] R. Kuhn, F. Perronnin, P. Nguyen, J.-C. Junqua, L. Rigazio, "Very Fast Adaptation with a Compact Context-Dependent Eigenvoice Model," w *Proc. ICASSP*, 2001.

[83] R. Kuhn, J.-C. Junqua, P. Nguyen, N. Niedzielski, "Rapid Speaker Adaptation in Eigenvoice Space," *IEEE Trans. on Speech and Audio Processing*, vol. 8, no. 6, November 2000.

[84] N. Kumar, A. G. Andreou, "Heteroscedastic Discriminant Analysis and Reduced Rank HMMs for Improved Speech Recognition," *Speech Communcation*, vol. 26, no. 4, 1998.

[85] H.-K. J. Kuo, Y. Gao, "Maximum Entropy Direct Models for Speech Recognition," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 14, no. 3, May 2006.

[86] C. Lee, D. Hyun, E. Choi, J. Go, C. Lee, "Optimizing Feature Extraction for Speech Recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 11, no. 1, January 2003.

[87] K.-S. Lee, "MLP-Based Phone Boundary Refining for a TTS Database," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 14, no. 3, May 2006.

[88] L. Lee, R. Rose, "A Frequency Warping Approach to Speaker Normalization," *IEEE Trans. on Speech and Audio Processing*, vol. 6, no. 1, January 1998.

[89] K. Li, M. N. S. Swamy, M. O. Ahmad, "An Improved Voice Activity Detection Using Higher Order Statistics," *IEEE Trans. on Speech and Audio Processing*, vol. 13, no. 5, September 2005.

[90] X. Li, J. Malkin, J. A. Bilmes, "A High-Speed, Low-Resource ASR Back-End Based on Custom Arithmetic," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 14, no. 5, September 2006.

[91] F. H. Liu, R. M. Stern, X. Huang, A. Acero, "Efficient Cepstral Normalization for Robust Speech Recognition," w *Proc. ARPA Speech and Natural Language Workshop*, 1993.

[92] G. Mahe, A. Gilloire, L. Gros, "Correction of Voice Timbre Distortions in Telephone Networks: Method and Evaluation," *Speech Communication*, vol. 43, 2004.

[93] W. Majewski, "Aural-Perceptual Voice Recognition of Original Speakers and Their Imitators," *Archives of Acoustics*, vol. 30, no. 4 (Supplement), 2005.

[94] B. Mak, , E. Barnard, "Phone Clustering Using the Bhattacharyya Distance," w *Proc. ICSLP*, 1996.

[95] B. K.-W. Mak, Y.-C. Tam, P. Q. Li, "Discriminative Auditory-Based Features for Robust Speech Recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 12, no. 1, January 2004.

[96] B. Mak, J. T. Kwok, S. Ho, "Kernel Eigenvoice Speaker Adaptation," *IEEE Trans. on Speech and Audio Processing*, vol. 13, no. 5, September 2005.

[97] B. K.-W. Mak, R. W.-H. Hsiao, S. K.-L. Ho, J. T. Kwok, "Embedded Kernel Eigenvoice Speaker Adaptation and Its Implication to Reference Speaker Weighting," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 14, no. 4, July 2006.

[98] R. J. Mammone, X. Zhang, R. P. Ramachandran, "Robust Speaker Recognition. A Feature-Based Approach," *IEEE Signal Processing Magazine*, vol. 13, no. 5, September 1996.

[99] K. Marasek, "Large Vocabulary Continuous Speech Recognition System for Polish," *Archives of Acoustics*, vol. 24, no. 4, 2003.

[100] K. Markov, J. Dang, S. Nakamura, "Integration of Articulatory and Spectrum Features Based on the Hybrid HMM/BN Modeling Framework," *Speech Communication*, vol. 48, 2006.

[101] M. Marzinzik, B. Kollmeier, "Speech Pause Detection for Noise Spectrum Estimation by Tracking Power Envelope Dynamics," *IEEE Trans. on Speech and Audio Processing*, vol. 10, no. 2, February 2002.

[102] J. McAuley, J. Ming, D. Stewart, P. Hanna, "Subband Correlation and Robust Speech Recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 15, no. 5, September 2005.

[103] J. McDonough, T. Schaaf, A. Waibel, "Speaker Adaptation with All-Pass Transforms," *Speech Communication*, vol. 42, 2004.

[104] D. Mercier, R. Seguier, "Spiking Neurons (STANNs) in Speech Recognition," w *Proc. 3rd WSES International Conference on Neural Networks and Applications*, 2002.

[105] C. Meyer, H. Schramm, "Boosting HMM Acoustic Models in Large Vocabulary Speech Recognition," *Speech Communication*, vol. 28, 2006.

[106] J. Ming, "Noise Compensation for Speech Recognition with Arbitrary Additive Noise," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 14, no. 3, May 2006.

[107] J. Ming, P. Jancovic, F. J. Smith, "Robust Speech Recognition Using Probabilistic Union Models," *IEEE Trans. on Speech and Audio Processing*, vol. 10, no. 6, September 2002.

[108] S. Molau, "Normalization in the Acoustic Feature Space for Improved Speech Recognition," praca doktorska, Rheinisch-Westfalischen Technischen Hochschule, Aachen, 2003.

[109] D. S. Moore, *The Practice of Business Statistics: Using Data for Decisions*. W. H. Freeman & Co., 2002.

[110] P. J. Moreno, B. Raj, R. M. Stern, "A Vector Taylor Series Approach for Environment-Independent Speech Recognition," w *Proc. ICASSP*, 1996.

[111] P. J. Moreno, B. Raj, E. Gouvea, R. M. Stern, "Multivariate-Gaussian-Based Cepstral Normalization for Robust Speech Recognition," w *Proc. ICASSP*, 1995.

[112] P. J. Moreno, R. M. Stern, "Sources of Degradation of Speech Recognition in the Telephone Network," w *Proc. ICASSP*, 1994.

[113] P. Mrówka, R. Makowski, "Channel and Speaker Variety Compensation Using Modified Eigenvoices Algorithm," w *Proc. ICSES*, 2004.

[114] P. Mrówka, R. Makowski, "Otwarty system do badań wariantów algorytmów rozpoznawania mowy polskiej," w *XI KOWBAN*, 2004.

[115] P. Mrówka, R. Makowski, "Poprawa jakości rozpoznawania mowy poprzez transformację widma sygnału danego mówcy," w *XII KOWBAN*, 2005.

[116] P. Mrówka, R. Makowski, "Improvement in Short Utterances Recognition via Employment of Spectral Transformations Banks," w *Proc. ICSES*, 2006.

[117] P. Mrówka, R. Makowski, "Modifications of Spectral Transformations Banks Method for Speech Recognition Improvement," w *Proc. ICSES*, 2006.

[118] P. Mrówka, R. Makowski, "Normalization of Speaker Individual Characteristics and Compensation of Linear Transmission Distortions in Command Recognition Systems," *Archives of Acoustics*, w recenzji.

[119] M. Naito, L. Deng, Y. Sagisaka, "Speaker Clustering for Speech Recognition Using Vocal Tract Parameters," *Speech Communication*, vol. 36, 2002.

[120] A. Nakamura, "Restructuring Gaussian Mixture Density Functions in Speaker-Independent Acoustic Models," *Speech Communication*, vol. 36, 2002.

[121] T. Nakatani, K. Kinoshita, M. Miyoshi, "Harmonicity-Based Blind Dereverberation for Single-Channel Speech Signals," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 15, no. 1, January 2007.

[122] A. V. Oppenheim, R. W. Schafer, *Cyfrowe przetwarzanie sygnałów.* WKŁ, Warszawa, 1979.

[123] M. Padmanabhan, S. Dharanipragada, "Maximizing Information Content in Feature Extraction," *IEEE Trans. on Speech and Audio Processing*, vol. 13, no. 4, July 2005.

[124] K. J. Palomaki, G. J. Brown, J. P. Barker, "Techniques for Handling Convolutional Distortion with Missing Data Automatic Speech Recognition," *Speech Communication*, vol. 43, 2004.

[125] Y. Pan, A. Waibel, "The Effects of Room Acoustics on MFCC Speech Parameter," w *Proc. ICSLP*, 2000.

[126] J. Park, H. Ko, "Effective Acoustic Model Clustering via Decision-Tree with Supervised Learning," *Speech Communication*, vol. 46, 2005.

[127] V. Pitsikalis, I. Kokkinos, P. Maragos, "Nonlinear Analysis of Speech Signals: Generalized Dimensions and Lyapunov Exponents," w *Proc. Eurospeech*, 2003.

[128] M. Pitz, H. Ney, "Vocal Tract Normalization Equals Linear Transformation in Cepstral Space," *IEEE Trans. on Speech and Audio Processing*, vol. 13, no. 5, September 2005.

[129] C. Plapous, C. Marro, P. Scalart, "Improved Signal-to-Noise Ratio Estimation for Speech Enhancement," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 14, no. 6, November 2006.

[130] A. Plucińska, E. Pluciński, *Probabilistyka.* WNT, Warszawa, 2000.

[131] A. Potamianos, S. Narayanan, "Robust Recognition of Children's Speech," *IEEE Trans. on Speech and Audio Processing*, vol. 11, no. 6, November 2003.

[132] E. J. Pusateri, T. J. Hazen, "Rapid Speaker Adaptation Using Speaker Clustering," w *Proc. ICSLP*, 2002.

[133] L. Rabiner, B.-H. Juang, *Fundamentals of Speech Recognition.* Prentice-Hall, 1993.

[134] B. Raj, M. L. Seltzer, R. M. Stern, "Reconstruction of Missing Features for Robust Speech Recognition," *Speech Communication*, vol. 43, 2004.

[135] J. Ramirez, J. C. Segura, C. Benitez, "A New Kullback–Leibler VAD for Speech Recognition in Noise," *IEEE Signal Processing Letters*, vol. 11, no. 2, February 2004.

[136] J. Ramirez, J. C. Segura, C. Benitez, A. de la Torre, A. Rubio, "A New Adaptive Long-Term Spectral Estimation Voice Activity Detector," w *Proc. Eurospeech*, 2003.

[137] G. Riccardi, D. Hakkani-Tur, "Active Learning: Theory and Applications to Automatic Speech Recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 13, no. 4, July 2005.

[138] A. J. Robinson, G. D. Cook, D. P. W. Ellis, E. Fosler-Lussier, S. J. Renals, D. A. G. Williams, "Connectionist Speech Recognition of Broadcast News," *Speech Communication*, vol. 37, 2002.

[139] T. M. Rutkowski, "Reducing Environmental and Transmission Interference to Improve Automatic Speaker Recognition," praca doktorska, Politechnika Wrocławska, 2002.

[140] S. Sagayama, Y. Kato, M. Nakai, H. Shimodaira, "Jacobian Approach to Joint Adaptation to Noise, Channel and Vocal Tract Length," w *Proc. ICASSP*, 2002.

[141] G. Saon, "A Non-Linear Speaker Adaptation Technique Using Kernel Ridge Regression," w *Proc. ICASSP*, 2006.

[142] R. Sarilaya, J. H. L. Hansen, "Analysis of the Root-Cepstrum for Acoustic Modeling and Fast Decoding in Speech Recognition," w *Proc. Eurospeech*, 2001.

[143] H. G. Schuster, *Chaos deterministyczny. Wprowadzenie.* PWN, Warszawa, 1993.

[144] J.-L. Schwartz, C. Abry, L.-J. Boe, L. Menard, N. Vallee, "Asymmetries in Vowel Perception, in the Context of the Dispersion–Focalisation Theory," *Speech Communication*, vol. 45, 2005.

[145] J. W. Seok, K. S. Bae, "Speech Enhancement with Reduction of Noise Components in the Wavelet Domain," w *Proc. ICASSP*, 1997.

[146] G. Shi, M. M. Shanechi, P. Aarabi, "On the Importance of Phase in Human Speech Recognition," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 14, no. 5, September 2006.

[147] T. Shimamura, H. Kobayashi, "Weighted Autocorrelation for Pitch Extraction of Noisy Speech," *IEEE Trans. on Speech and Audio Processing*, vol. 9, no. 7, October 2001.

[148] K. Shinoda, C.-H. Lee, "A Structural Bayes Approach to Speaker Adaptation," *IEEE Trans. on Speech and Audio Processing*, vol. 9, no. 3, March 2001.

[149] J. Silva, S. Narayanan, "Average Divergence Distance as a Statistical Discrimination Measure for Hidden Markov Models," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 14, no. 3, May 2006.

[150] K. C. Sim, M. J. F. Gales, "Minimum Phone Error Training of Precision Matrix Models," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 14, no. 3, May 2006.

[151] J. Sirigos, N. Fakotakis, G. Kokkinakis, "A Hybrid Syllable Recognition System Based on Vowel Spotting," *Speech Communication*, vol. 38, 2002.

[152] M. Siu, A. Chan, "A Robust Viterbi Algorithm Against Impulsive Noise with Application to Speech Recognition," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 14, no. 6, November 2006.

[153] J. C. Spall, "Multivariate Stochastic Approximation Using a Simultaneous Perturbation Gradient Approximtion," *IEEE Trans. on Automatic Control*, vol. 37, no. 3, March 1992.

[154] J. C. Spall, "Adaptive Stochastic Approximation by the Simultaneous Perturbation Method," *IEEE Trans. on Automatic Control*, vol. 45, no. 10, October 2000.

[155] J. J. Sroka, L. D. Braida, "Human and Machine Consonant Recognition," *Speech Communication*, vol. 45, 2005.

[156] P. Staroniewicz, "Zastosowanie zoptymalizowanych niejawnych modeli Markowa do opisu przejść międzyfonemowych dla potrzeb automatycznego rozpozmnawania mowy ciągłej," praca doktorska, Politechnika Wrocławska, 2001.

[157] P. Staroniewicz, "Automatic Segmentation Based on Spectral Characteristics of Speech Signal," *Archives of Acoustics*, vol. 30, no. 4 (Supplement), 2005.

[158] N. Strom, "Speaker Adaptation by Modeling the Speaker Variation in a Continuous Speech Recognition System," w *Proc. ICSLP*, 1996.

[159] D. Sundermann, A. Bonafonte, H. Ney, "Time Domain Vocal Tract Length Normalization," w *Proc. 4th IEEE International Symposium on Signal Processing and Information Technology*, 2004.

[160] J. Szabatin, *Podstawy teorii sygnałów*. WKŁ, Warszawa, 2000.

[161] R. Tadeusiewicz, *Sygnał mowy*. WKŁ, Warszawa, 1988.

[162] D. T. Toledano, L. A. H. Gomez, L. V. Grande, "Automatic Phonetic Segmentation," *IEEE Trans. on Speech and Audio Processing*, vol. 11, no. 6, November 2003.

[163] S. Tsakalidis, V. Doumpiotis, W. Byrne, "Discriminative Linear Transforms for Feature Normalization and Speaker Adaptation in HMM Estimation," *IEEE Trans. on Speech and Audio Processing*, vol. 13, no. 3, May 2005.

[164] Y. Tsao, S.-M. Lee, L.-S. Lee, "Segmental Eigenvoice with Delicate Eigenspace for Improved Speaker Adaptation," *IEEE Trans. on Speech and Audio Processing*, vol. 13, no. 3, May 2005.

[165] V. Tyagi, H. Bourlard, C. Wellekens, "On Variable-Scale Piecewise Stationary Spectral Analysis of Speech Signals for ASR," *Speech Communication*, vol. 48, 2006.

[166] P. Veprek, M. S. Scordilis, "Analysis, Enhancement and Evaluation of Five Pitch Determination Techniques," *Speech Communication*, vol. 37, 2002.

[167] L. Welling, H. Ney, S. Kanthak, "Speaker Adaptive Modeling by Vocal Tract Normalization," *IEEE Trans. on Speech and Audio Processing*, vol. 10, no. 6, September 2002.

[168] R. Wielgat, "Zastosowanie znaczników fonetycznych, nieliniowej transformacji czasowej i niejawnych modeli Markowa do rozpoznawania izolowanych słów mowy polskiej," praca doktorska, Politechnika Łódzka, 2001.

[169] R. Wielgat, "Improving Speech Recognition Accuracy Using HMM and DTW Methods," w *Proc. ICSES*, 2004.

[170] P. Woodland, D. Povey, "Large Scale Discriminative Training for Speech Recognition," w *Proc. ISCA ITRW Automatic Speech Recognition: Challenges for the Millenium*, 2000.

[171] B.-F. Wu, K.-C. Wang, "Robust Endpoint Detection Algorithm Based on the Adaptive Band-Partitioning Spectral Entropy in Adverse Environments," *IEEE Trans. on Speech and Audio Processing*, vol. 13, no. 5, September 2005.

[172] J. Wu, Q. Huo, "An Environment-Compensated Minimum Classification Error Training Approach Based on Stochastic Vector Mapping," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 14, no. 6, November 2006.

[173] X. Wu, Y. Yan, "Speaker Adaptation Using Constrained Transformation," *IEEE Trans. on Speech and Audio Processing*, vol. 12, no. 2, March 2004.

[174] H. Yamamoto, T. Nishimoto, S. Sagayama, "Frame-by-Frame HMM Adaptation for Reverberant Speech Recognition," w *Proc. Special Workshop in Maui (SWIM)*, 2004.

[175] C.-S. Yang, H. Kasuya, "Speaker Individualities of Vocal Tract Shapes of Japanese Vowels Measured by Magnetic Resonance Images," w *Proc. ICSLP*, 1996.

[176] K. Yao, K. K. Paliwal, T.-W. Lee, "Generative Factor Analyzed HMM for Automatic Speech Recognition," *Speech Communication*, vol. 45, 2005.

[177] K. Yao, E. Visser, O.-W. Kwon, T.-W. Lee, "A Speech Processing Front-End with Eigenspace Normalization for Robust Speech Recognition in Noisy Automobile Environments," w *Proc. Eurospeech*, 2003.

[178] L. Yao, D. Yu, T. Huang, "A Unified Spectral Transformation Adaptation Approach for Robust Speech Recognition," w *Proc. ICSLP*, 1996.

[179] U. H. Yapanel, J. H. L. Hansen, "Towards an Intelligent Acoustic Front-End for Automatic Speech Recognition: Built-In Speaker Normalization (BISN)," w *Proc. ICASSP*, 2005.

[180] N. B. Yoma, C. Molina, J. Silva, C. Busso, "Modeling, Estimating, and Compensating Low-Bit Rate Coding Distortion in Speech Recognition," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 14, no. 1, January 2006.

[181] S. Young, G. Evermann, T. Hain, D. Kershaw, J. Odell, D. Ollason, D. Povey, *et al.*, *The HTK Book*. Cambridge University Engineering Department, 2002.

[182] K. Yu, M. J. F. Gales, "Discriminative Cluster Adaptive Training," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 14, no. 5, September 2006.

[183] Y.-S. Yun, Y.-H. Oh, "A Segmental-Feature HMM for Continuous Speech Recognition Based on a Parametric Trajectory Model," *Speech Communication*, vol. 38, 2002.

[184] P. Zhan, M. Westphal, M. Finke, A. Waibel, "Speaker Normalization and Speaker Adaptation - A Combination for Conversational Speech Recognition," w *Proc. Eurospeech*, 1997.

[185] J. Zhang, W. Ward, B. Pellom, "Phone Based Voice Activity Detection Using Online Bayesian Adaptation with Conjugate Normal Distributions," w *Proc. ICASSP*, 2002.

[186] R. Zhao, Z. Wang, "Robust Speech Recognition Based on Spectral Adjusting and Warping," w *Proc. ICASSP*, 2005.

[187] B. Zhou, J. H. L. Hansen, "Rapid Discriminative Acoustic Model Based on Eigenspace Mapping for Fast Speaker Adaptation," *IEEE Trans. on Speech and Audio Processing*, vol. 13, no. 4, July 2005.

[188] D. Zhu, S. Nakamura, K. K. Paliwal, R. Wang, "Maximum Likelihood Sub-Band Adaptation for Robust Speech Recognition," *Speech Communication*, vol. 47, 2005.

[189] T. Zieliński, P. Gajda, M. Stachura, R. Wielgat, D. Król, T. Woźniak, S. Grabias, "Application of Human Factor Cepstral Coefficients to Robust Recognition of Pathological Pronunciation in Noisy Environment," w *Proc. ICSES*, 2006.

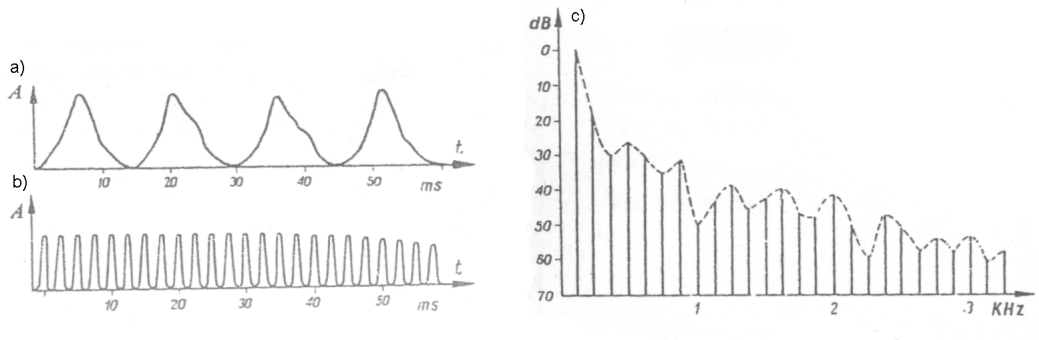
[190] T. P. Zieliński, *Od teorii do cyfrowego przetwarzania sygnałów*. Uczelniane Wydawnictwa Naukowo-Dydaktyczne AGH, 2002.

Dodatki

A. Mechanizm wytwarzania mowy

Między- i wewnątrzosobnicza zmienność widm jednostek fonetycznych związana jest bezpośrednio z mechanizmem wytwarzania mowy, którego uproszczony opis podano poniżej.

Powietrze z płuc przetłaczane jest przez krtań, gdzie znajdują się wiązadła (struny) głosowe. W przypadku, kiedy są one luźne, powietrze przepływa swobodnie, choć przepływ ten ma charakter turbulentny. Gdy struny głosowe są napięte, przepływające powietrze wprawia je w drgania, czego skutkiem jest zmiana w czasie powierzchni szczeliny (głośni), przez którą przetłaczane jest powietrze (rys. A.1). W efekcie zmian prędkości objętościowej przepływającego przez głośnię powietrza powstaje dźwięk określany jako ton krtaniowy. Człowiek nie jest w stanie świadomie kontrolować bezpośrednio drgań strun głosowych, może jednak wpływać na ich napięcie, a w konsekwencji na częstotliwość tonu krtaniowego ($f^{(v)}$). Zakres możliwych do uzyskania przez człowieka wartości $f^{(v)}$ mieści się w przedziale 80 - 1000 Hz. Każdy człowiek ma pewien charakterystyczny dla siebie zakres wartości $f^{(v)}$ oraz schematy przebiegów tej wartości, występujące w wypowiedziach o danych cechach intonacyjnych. Dla głosów kobiecych wartość $f^{(v)}$ jest przeciętnie dwa razy wyższa niż dla głosów męskich [65, 161].

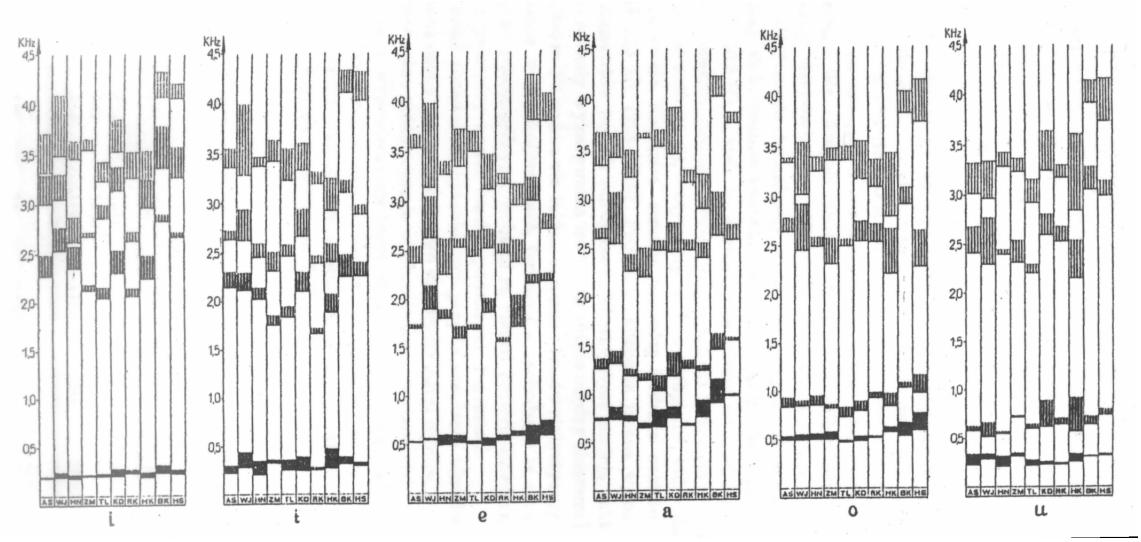


Rys. A.1. Ton krtaniowy: a) zmiana w czasie powierzchni głośni dla niskiego głosu męskiego, b) zmiana w czasie powierzchni głośni dla wysokiego głosu żeńskiego, c) widmo amplitudowe tonu krtaniowego dla średnio wysokiego głosu męskiego [65].

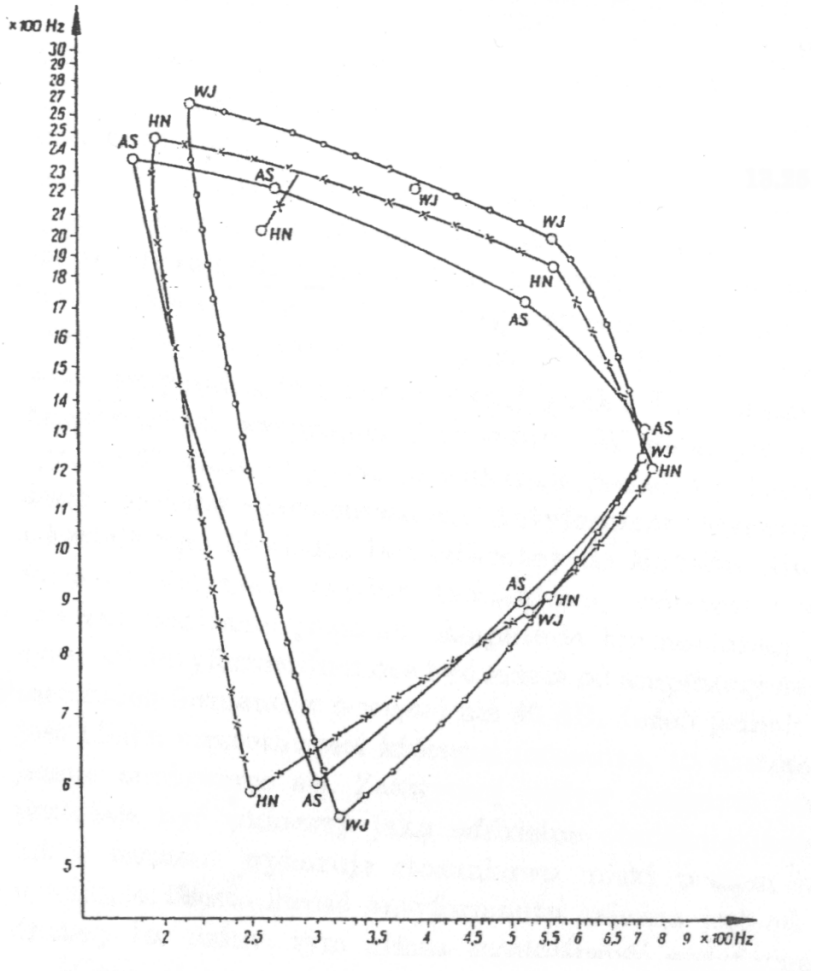
Części toru głosowego powyżej strun głosowych stanowią elementy układu rezonansowego, kształtującego widmo sygnału mowy. Części toru poniżej strun głosowych są tutaj zazwyczaj pomijane, a ich wpływ uwzględniany jest w charakterystyce widmowej tonu krtaniowego. W przypadku samogłosek powietrze przepływa swobodnie przez jamę ustną, a jama nosowa jest zamknięta. W przypadku spółgłosek nosowych przepływ możliwy jest natomiast tylko przez jamę nosową, podczas gdy jama ustna stanowi zamkniętą komorę rezonansową. Podczas wytwarzania spółgłosek bocznych język zamyka środkową część toru głosowego, a przepływ powietrza nie odbywa się wzdłuż osi języka, jak w przypadku samogłosek, lecz po jego bokach. W przypadku spółgłosek trących w torze głosowym występuje przewężenie powodujące silnie turbulentny przepływ powietrza, a co za tym idzie, powstawanie dodatkowego pobudzenia o charakterze szumowym. W przypadku spółgłosek zwartych natomiast, dodatkowe pobudzenie ma charakter impulsowy, związany z zamknięciem toru głosowego, a następnie jego nagłym otwarciem [65, 161].

W równaniu (1.2) sygnał pobudzenia $x^{(v)}$ w przypadku samogłosek odpowiada tonowi krtaniowemu. W przypadku spółgłosek trących pobudzenie ma charakter szumu, a w przypadku zwartych - sygnału impulsowego. Jeśli spółgłoska jest dźwięczna, występuje dodatkowo pobudzenie tonem krtaniowym. W tej sytuacji $x^{(v)}$ reprezentuje łącznie wszystkie występujące pobudzenia. Trzeba jednak zaznaczyć, że nie jest to prosta suma wspomnianych sygnałów pobudzenia, gdyż trzeba uwzględnić „przeniesienie" pobudzenia szumowego w miejsce powstawania tonu krtaniowego, a więc transmitancję toru głosowego pomiędzy miejscami generowania pobudzeń. Jest to niedogodność spowodowana dużą prostotą modelu, lecz mimo to stosowanie go daje zadowalające wyniki. Na $h^{(v)}$ składa się transmitancja toru głosowego od miejsca pobudzenia do ust oraz charakterystyka promieniowania ust. Nieraz charakterystykę promieniowania ust wyróżnia się oddzielnym członem dodanym do równania (1.2). Dodatkowo, w szczegółowej analizie, można uwzględniać również transmitancję związaną z rozchodzeniem się dźwięku wokół głowy mówcy, charakterystykę pomieszczenia, mikrofonu i innych elementów toru transmisyjnego.

Formantami nazywane są pasma częstotliwości sygnału mowy, w których występuje uwypuklenie poziomu widmowej gęstości mocy. Formant opisywany jest najczęściej za pomocą trzech parametrów: częstotliwości środkowej, szerokości pasma i amplitudy. Na ogół podaje się tylko wartości pierwszego parametru. Przykłady zmienności między- i wewnątrzosobniczej w przypadku formantów sześciu samogłosek polskich przedstawiają rys. A.2 i A.3.



Rys. A.2. Zakresy formantów samogłosek polskich dla 10 różnych mówców [65].



Rys. A.3. Pętle formantowe F_1F_2 (pierwszego i drugiego formantu) samogłosek polskich dla 3 różnych mówców [65].

B. Baza nagrań sygnałów mowy

W pracy korzystano z dwóch baz nagrań sygnałów mowy polskiej. Poniżej przedstawiono ich podstawowe charakterystyki, podział na część uczącą i testową oraz przyjęty zbiór jednostek fonetycznych.

B.1. CORPORA

Baza CORPORA [46], zrealizowana w Instytucie Informatyki Politechniki Poznańskiej, cechuje się następującymi parametrami:

- Liczba mówców: **37**
- Liczba zestawów nagrań: **45**
- Zawartość pojedynczego zestawu nagrań:
  - 33 głoski
  - 200 imion
  - 10 cyfr
  - 8 komend
  - 114 zdań
- Parametry sygnału cyfrowego: częstotliwość próbkowania 16 kHz, rozdzielczość 16 bitów.
- Warunki nagrań: Nagrania wykonano w pomieszczeniach biurowych przy pomocy mikrofonu pojemnościowego, a w przypadku jednego zestawu - dynamicznego. Rejestracja prowadzona była w polu bliskim.
- Segmentacja i etykietyzacja: Wszystkie wypowiedzi są posegmentowane i zetykietyzowane. Przyjęty zbiór jednostek fonetycznych opisano niżej. Minimalny skok w segmentacji wynosił 5 ms.

B.2. bnITTA

Baza bnITTA jest otwartym projektem realizowanym w Instytucie Telekomunikacji, Teleinformatyki i Akustyki Politechniki Wrocławskiej. Autor jest jednym z głównych wykonawców tej bazy. Poniżej podano parametry dotyczące tylko tej części bazy, którą wykorzystano w pracy:

- Liczba mówców: **16**
- Liczba zestawów nagrań: **37**
- Zawartość pojedynczego zestawu nagrań:
  - 223 wyrazy (w tym cyfry i komendy jak w bazie CORPORA) - **poddane segmentacji i etykietyzacji**
  - 10 komend wielowyrazowych (w tym komenda jak w bazie CORPORA)
  - 37 głosek
  - 60 trzywyrazowych ciągów składających się z cyfr i znaku „+”
  - 5 zdań zawierających ten sam wyraz w różnych kontekstach
  - zawierający 510 słów ciąg zdań w postaci tekstu informatycznego
  - zawierający 475 słów ciąg zdań w postaci tekstu prasowego
- Parametry sygnału cyfrowego: częstotliwość próbkowania 48 kHz, rozdzielczość 16 bitów.
- Warunki nagrań: Większość nagrań wykonano w pomieszczeniu studyjnym, kilka zestawów w pomieszczeniu biurowym. W większości nagrań używano mikrofonu dynamicznego Shure PG48. Kilka zestawów zarejestrowano z użyciem przypinanego mikrofonu elektretowego. Rejestracja prowadzona była w polu bliskim. Używano magnetofonu cyfrowego DAT. Poziom tła akustycznego był w większości nagrań niski, w kilku średni.
- Segmentacja i etykietyzacja: Część nagrań poddano segmentacji i etykietyzacji. Przyjęty zbiór jednostek fonetycznych opisano niżej. Minimalny skok w segmentacji wynosił 1 ms. Wykorzystano napisane przez autora narzędzie do półautomatycznej segmentacji i etykietyzacji.

B.3. Podział bazy

Podział bazy na część uczącą i testową oraz statystyki wieku i płci mówców przedstawiono w tabeli B.1.

Tab. B.1. Podział i statystyki bazy nagrań. Podano liczby mówców, a w nawiasach liczby zestawów nagrań. Nagrania jednego mówcy znalazły się zarówno w części uczącej, jak i testowej, przy czym różniły się znacznie warunkami akustycznymi.

baza	wiek/płeć	cała baza	zbiór uczący	zbiór testowy
CORPORA	9-15 lat	4(6)	2(4)	2(2)
	20-29 lat	20(20)	11(11)	9(9)
	30-49 lat	8(12)	4(7)	4(5)
	50-70 lat	5(7)	3(4)	2(3)
bnITTA	20-29 lat	10(27)	7(17)	4(10)
	30-49 lat	2(4)	1(2)	1(2)
	50-70 lat	4(6)	2(4)	2(2)
ogółem	głosy męskie	41(66)	23(39)	19(27)
	głosy kobiece i dziecięce	12(16)	7(10)	5(6)
	ogółem	**53(82)**	**30(49)**	**24(33)**

B.4. Przyjęty zestaw fonemów

Jako podstawową jednostkę fonetyczną przyjęto fonem. W tabeli B.2 przedstawiono zastosowany w pracy zestaw fonemów i pseudofonemów, wraz z zaznaczeniem relacji w stosunku do podziałów używanych w bazach nagrań. Przyjęto jednoznakowe symbole fonemów wzorowane na stosowanych w bazie CORPORA. Kolumny tabeli B.2 zawierają kolejno: **(1)** - symbol fonemu stosowany w pracy, **(2)** - symbol fonemu w notacji SAMPA, **(3)** - przykład występowania, **(4)** - relację do zestawu fonemów w bazie CORPORA, **(5)** - relację do zestawu fonemów w bazie bnITTA.

Zawarte w kolumnach **(4)** i **(5)** relacje objaśniają kilka różnic w zestawach fonemów, a także metodach segmentacji i etykietyzacji, używanych w pracy i w obu bazach. W bazie CORPORA występują fonemy 'ą' i 'ę'. Zostały one rozdzielone w połowie i fragmenty początkowe zaliczone zostały do fonemów 'o' i 'e' (relacja nr 2), a końcowe do fonemu 'N' (relacja nr 3). Segmentacja fonemów 'p', 't', 'k', 'c', 'ć' 'C' w tej bazie włącza do nich fragment zwarcia. W pracy przyjęto, że zwarcia oznacza się osobnymi pseudofonemami, tak więc wymienione fonemy dzielono na fragmenty zwarcia i fragmenty plozji+aspiracji/afrykacji w ten sposób, że do zwarcia zaliczano te ramki danej realizacji fonemu, dla których energia była mniejsza od progu wynoszącego 0.2 średniej energii całej realizacji fonemu (relacja nr 5), pozostałe ramki uznawano za plozję+aspirację/afrykację (relacja nr 4). Podobnie rozdzielano fonemy 'Z', 'Ż', 'Ź' na fragmenty zwarcia dźwięcznego (relacja nr 7) i plozji+afrykacji (relacja nr 6) z tym, że tutaj do zwarcia zaliczano te ramki danej realizacji fonemu, dla których stosunek mocy sygnału w paśmie 2-8 kHz do mocy sygnału w paśmie

Tab. B.2. Przyjęty w pracy zestaw fonemów i pseudofonemów. Opis w tekście.

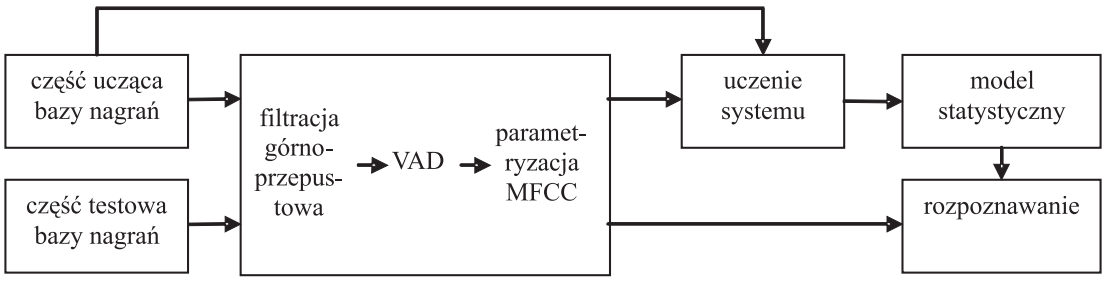
(1)	(2)	(3)	(4)	(5)
a	a	p*a*t	1	1
e	e	t*e*st	2	1
o	o	p*o*t	2	1
u	u	p*u*k	1	1
i	i	t*i*k	1	1
y	I	t*y*p	1	1
p	p	*p*at	4	1
b	b	*b*at	1	8
t	t	*t*est	4	1
d	d	*d*ym	1	8
k	k	*k*at	4	1
g	g	*g*en	1	8
c	ts	*c*oś	4	1
Z	dz	*dz*won	6	1
ć	ts’	*ć*wicz	4	1
Ź	dz’	*dź*więk	6	1
C	tS	*cz*yn	4	1
Ż	dZ	*dż*in	6	1
f	f	*f*in	1	1
w	v	*w*aga	1	1

(1)	(2)	(3)	(4)	(5)
s	s	*s*yk	1	1
z	z	*z*ez	1	1
S	S	*sz*yk	1	1
ż	Z	*ż*yto	1	1
ś	s’	*ś*wit	1	1
ź	z’	*ź*le	1	1
h	x	*h*ak	1	1
m	m	*m*ak	1	1
n	n	*n*asz	1	1
ń	n’	ko*ń*	1	1
N	N	go*n*g	3	1
k	l	*l*uk	1	1
r	r	*r*ak	1	1
ł	w	*ł*uk	1	1
j	j	*j*ak	1	1
@		pseudofonem zwarcia dźwięcznego	7	1
#		pseudofonem zwarcia bezdźwięcznego i ciszy	5	9

0.1-2 kHz był mniejszy niż 0.2. W fonemach ’b’, ’d’, ’g’ nie uwzględniano podziału na zwarcie i plozję, stąd w przypadku danych z bazy bnITTA łączono osobno w niej wyróżnione fragmenty zwarcia i plozji (relacja nr 8). Relacja nr 9 oznacza łączne traktowanie fragmentów ciszy i zwarcia bezdźwięcznego, relacja nr 1 natomiast, że granice fonemów są zgodne z segmentacją w bazie.

C. System ARM

Zaprojektowany jako narzędzie badawcze system ARM charakteryzuje się niewielkim, zamkniętym słownikiem, zawierającym następujące wyrazy: „zero”, „jeden”, „dwa”, „trzy”, „cztery”, „pięć”, „sześć”, „siedem”, „osiem”, „dziewięć”, „kropka”, „spacja”, „przecinek”, „nie”, „źle”, „nowa linia”, „cofnij”, „stop”. Zaproponowano cztery warianty systemu, różniące się modyfikacjami modelu statystycznego, jak również metod uczenia i rozpoznawania. Wszystkie warianty bazowały na parametryzacji MFCC i modelowaniu HMM. Warto zaznaczyć, że rozpoznawaniu podlegało 18 wyrazów, lecz w pewnych etapach uczenia systemu wykorzystywano wszystkie dostępne w bazie nagrania. Ponadto uwzględnione zostały różnice w sposobie wypowiadania zawartych w słowniku wyrazów, np. „tSy” i „Cy”, tak, że rzeczywista liczba wyrazów w słowniku wynosiła 22. System użyty w badaniach jest względnie prosty, a jego złożoność obliczeniowa pozwala na przeprowadzenie w zadowalającym czasie wielokrotnie powtarzanych procedur uczenia i rozpoznawania. Jednocześnie zastosowane w systemie rozwiązania umożliwiają łatwą i naturalną rozbudowę jego słownika, umożliwienie rozpoznawania wyrazów połączonych, wprowadzenie bardziej złożonych metod parametryzacji. Uzyskane z wykorzystaniem systemu wyniki potwierdzające skuteczność zaproponowanej metody kompensacji można więc z dużym prawdopodobieństwem uogólnić na systemy bardziej złożone. Rys. C.1 przedstawia główne bloki funkcjonalne systemu.



Rys. C.1. Ogólny funkcjonalny schemat systemu ARM.

C.1. Filtracja wstępna i wykrywanie obecności sygnału mowy

Pierwszą operacją wykonywaną na sygnale jest filtracja górnoprzepustowym filtrem Czebyszewa typu II, rzędu 6, o minimalnym tłumieniu w pasmie zaporowym równym 60 dB i pasmie przejściowym 100-200 Hz. Jej celem jest tłumienie sygnału w pasmie częstotliwości nie zawierającym użytecznych informacji z punktu widzenia ARM, zawierającym natomiast często znaczne zakłócenia.

W celu wybrania z zarejestrowanego sygnału fragmentów zawierających istotne z punktu widzenia dalszej analizy informacje, stosowany jest algorytm wykrywania obecności sygnału mowy (VAD). Algorytm ten ma wybrać zatem fragmenty zawierające mowę, a odrzucać ciszę. Wykrywane są również fragmenty zwarć przed bezdźwięcznymi spółgłoskami zwartymi i zwarto-trącymi, choć nie wszystkie warianty systemu wykorzystują te informacje. Uniwersalny algorytm VAD powinien umożliwiać detekcję w obecności szumu addytywnego, choć w badaniach przeprowadzanych w tej pracy wykorzystywano sygnały nie zaszumione. Przyjęto, że zmiany widma szumu w czasie są znacznie wolniejsze niż zmiany widma sygnału mowy.

W literaturze znaleźć można wiele propozycji algorytmów VAD, np. [135, 89, 35, 171, 101, 136, 185]. Najskuteczniejsze okazują się metody oparte na analizie zmian energii sygnału w czasie. W niniejszej pracy zaproponowano oryginalny algorytm VAD, wykorzystujący analizę energetyczną sygnału. Kolejne etapy działania tego algorytmu są następujące:

1. Wyodrębnij z sygnału wejściowego sygnały $x_1^{(f)}$, $x_2^{(f)}$, $x_3^{(f)}$ i $x_4^{(f)}$, odpowiadające podpasmom o częstotliwościach granicznych 0-1 kHz, 1-2 kHz, 2-4 kHz i 4-8 kHz. Wykorzystano tutaj dyskretną transformację falkową (DWT - *discrete wavelet transform*) o strukturze oktawowej, w której zastosowano falkę Daubechies rzędu 9. Z punktu widzenia analizy energetycznej, DWT z falkami ortogonalnymi ma tę zaletę, że nie występuje przeciek energii między podpasmami, tj. sumując energię sygnałów w podpasmach uzyskuje się energię sygnału pełnopasmowego.

2. W każdym podpasmie b wyznacz energię bieżącą sygnału wg zależności:

$$e_b(n) = \alpha_b^{(e)} e_b(n-1) + \left(1 - \alpha_b^{(e)}\right) \frac{1}{M_b^{(e)}} \sum_{m=0}^{M_b^{(e)}-1} {x_b^{(f)}}^2 (n-m) \qquad \text{(C.1)}$$

   gdzie $M_b^{(e)}$ dobrano tak, by uśredniane było 10 ms sygnału, a zatem $M_1^{(e)} = 20$, $M_2^{(e)} = 20$, $M_3^{(e)} = 40$, $M_4^{(e)} = 80$. Współczynnik $\alpha_1^{(e)}$ przyjęto równy 0.9, a pozostałe: $\alpha_2^{(e)} = \alpha_1^{(e)}$, $\alpha_3^{(e)} = {\alpha_1^{(e)}}^{1/2}$ i $\alpha_4^{(e)} = {\alpha_1^{(e)}}^{1/4}$.

3. W każdym podpasmie b wyznacz bieżące estymaty wartości oczekiwanej $\mu_b^{(n)}$ i odchylenia standardowego $\sigma_b^{(n)}$ poziomu mocy szumu. Uaktualnianie estymat

odbywa się wg poniższych zależności w przedziałach czasu nie zawierających mowy. Wykrywanie tych przedziałów opisane będzie dalej.

$$\mu_b^{(n)}(n) = \alpha_b^{(n)} \mu_b^{(n)}(n-1) + \left(1 - \alpha_b^{(n)}\right) e_b(n) \tag{C.2}$$

$$\sigma_b^{(n)}(n) = \alpha_b^{(n)} \sigma_b^{(n)}(n-1) + \left(1 - \alpha_b^{(n)}\right) \left| e_b(n) - \mu_b^{(n)}(n) \right| \tag{C.3}$$

gdzie współczynnik $\alpha_1^{(n)}$ przyjęto równy 0.99, a pozostałe: $\alpha_2^{(n)} = \alpha_1^{(n)}$, $\alpha_3^{(n)} = \alpha_1^{(n)^{1/2}}$ i $\alpha_4^{(n)} = \alpha_1^{(n)^{1/4}}$.

4. W każdym podpasmie b wyznacz współczynniki obecności mowy w_b, przyjmujące wartości z przedziału [0;1] i obliczane ze wzoru:

$$w_b(n) = \begin{cases} 0, & \text{dla } e_b^{(norm)}(n) \leq T^{(d)} \vee e_b(n) < T_b^{(abs)} \\ \frac{e_b^{(norm)}(n) - T^{(d)}}{T^{(g)} - T^{(d)}}, & \text{dla } T^{(d)} < e_b^{(norm)}(n) < T^{(g)} \\ 1, & \text{dla } e_b^{(norm)}(n) \geq T^{(g)} \end{cases} \tag{C.4}$$

$$e_b^{(norm)}(n) = \frac{e_b(n) - \mu_b^{(n)}(n)}{\sigma_b^{(n)}(n)} \tag{C.5}$$

gdzie progi przyjęto równe $T^{(d)} = 3$, $T^{(g)} = 6$, $T_1^{(abs)} = 10^{-3}$, $T_2^{(abs)} = 10^{-3}$, $T_3^{(abs)} = 10^{-4}$, $T_4^{(abs)} = 10^{-4.5}$.

5. Wyznacz globalny współczynnik obecności mowy $w^{(glob)}$ jako wygładzoną średnią ze współczynników obecności mowy z czterech podpasm b:

$$w^{(glob,nw)}(n) = \frac{1}{4}\left(w_1(n) + w_2(n) + \frac{1}{2}\left(w_3(2n) + w_3(2n-1)\right) + \right.$$
$$\left. + \frac{1}{4}\left(w_4(4n) + w_4(4n-1) + w_4(4n-2) + w_4(4n-3)\right)\right) \tag{C.6}$$

$$w^{(glob)}(n) = \frac{1}{M^{(w)}} \sum_{m=0}^{M^{(w)}-1} w^{(glob,nw)}(n-m) \tag{C.7}$$

gdzie $M^{(w)} = 20$, co odpowiada uśrednianiu fragmentu o długości 10 ms.

6. Jako fragmenty zawierające mowę oznacz te, dla których $w^{(glob)}$ przekracza zadany próg, równy 0.18. Nie zaznaczaj fragmentów mowy krótszych niż 10 ms oraz przerw we fragmentach mowy krótszych niż 20 ms.

Estymacja parametrów szumu wymaga znajomości granic fragmentów sygnału nie zawierających mowy. Są one wykrywane za pomocą detektora dźwięczności, czyli

uporządkowanej i wolnozmiennej w czasie struktury harmonicznej widmowej gęstości mocy sygnału. Jako fragmenty nie zawierające mowy oznaczane są te, które znajdują się w odległości większej niż 300 ms od końców wykrytych fragmentów dźwięcznych. Podejście takie wymaga, by pojawiały się odpowiednio częste przerwy w mowie, co w przypadku systemu rozpoznającego izolowane słowa jest łatwe do spełnienia. W celu wyznaczenia fragmentów dźwięcznych dokonywano następujących operacji:

1. Do analizy wybierz sygnał stanowiący podpasmo sygnału wejściowego o częstotliwościach granicznych 0-2 kHz, uzyskany w trakcie wyznaczania DWT, a następnie przeprowadź analizę widmową tego sygnału przy następujących wartościach parametrów: długość ramki analizy 128 próbek (32 ms), skok ramki 64 próbki (16 ms), okno Hamminga, długość transformaty 256 próbek. Oblicz kwadrat modułu widma każdej ramki n, oznaczony dalej jako $\mathbf{s}_n^{(wgm)}$. Uzyskane widma wygładź wg zależności:

$$\mathbf{s}_n^{(wgm,f)} = \alpha^{(s)}\mathbf{s}_{n-1}^{(wgm,f)} + \left(1-\alpha^{(s)}\right)\mathbf{s}_n^{(wgm)} \tag{C.8}$$

   gdzie $\alpha^{(s)}$ przyjęto równe 0.7.

2. Dla zakresu częstotliwości od 156.25 do 1875 Hz wykryj maksima widma $\mathbf{s}_n^{(wgm,f)}$, a następnie każdemu wykrytemu maksimum przypisz wartość amplitudy wg zależności:

$$a_{i,n}^{(wgm)} = s_{k_i,n}^{(wgm,f)} - \frac{1}{2}\left(s_{k_i-1,n}^{(wgm,f)} + s_{k_i+1,n}^{(wgm,f)}\right) \tag{C.9}$$

   gdzie i oznacza numer wykrytego maksimum, a k_i - numer prążka widma, przy którym ono wystąpiło.

3. Jeśli wartość $a_{i,n}^{(wgm)}$ przekracza próg równy $8 \cdot 10^{-6}$, oblicz skorygowaną częstotliwość wykrytego maksimum ze wzoru:

$$f_{i,n}^{(m)} = \left(-\frac{s_{k_i-1,n}^{(wgm,f)} - s_{k_i+1,n}^{(wgm,f)}}{4 \cdot a_{i,n}^{(wgm)}} + k_i\right) \cdot 15.625 \ \text{[Hz]} \tag{C.10}$$

   Wzory (C.9) i (C.10) wynikają z interpolacji funkcją kwadratową, zastosowanej do punktu wykrytego maksimum i dwóch punktów z nim sąsiadujących.

4. Wartość wyjściową detektora dźwięczności $v\,(n)$ wyznacz następująco:

$$v^{(t)}\left(x, T^{(x,d)}, T^{(x,g)}\right) = \begin{cases} 0, & \text{dla } x \le T^{(x,d)} \\ \frac{x-T^{(x,d)}}{T^{(x,g)}-T^{(x,d)}}, & \text{dla } T^{(x,d)} < x < T^{(x,g)} \\ 1, & \text{dla } x \ge T^{(x,g)} \end{cases} \tag{C.11}$$

$$f_{i,n}^{(m,c1)} = \min_j \left| f_{i,n}^{(m)} - f_{j,n-1}^{(m)} \right| \tag{C.12}$$

$$f_{i,n}^{(m,c2)} = \min_j \left| f_{i,n}^{(m)} - f_{j,n-2}^{(m)} \right| \tag{C.13}$$

$$\tag{C.14}$$

$$\begin{aligned} v\left(n\right) = \sum_i v^{(t)} \left(a_{i,n}^{(wgm)}, T^{(a,d)}, T^{(a,g)} \right) \cdot \\ \cdot v^{(t)} \left(f_{i,n}^{(m,c1)}, T^{(f1,d)}, T^{(f1,g)} \right) \cdot v^{(t)} \left(f_{i,n}^{(m,c2)}, T^{(f2,d)}, T^{(f2,g)} \right) \end{aligned} \tag{C.15}$$

gdzie progi przyjęto równe $T^{(a,d)} = 8 \cdot 10^{-6}$, $T^{(a,g)} = 20 \cdot 10^{-6}$, $T^{(f1,d)} = 5$ Hz, $T^{(f1,g)} = 10$ Hz, $T^{(f2,d)} = 10$ Hz, $T^{(f2,g)} = 20$ Hz.

5. Następnie wygładź $v\left(n\right)$:

$$v^{(f)}\left(n\right) = \alpha^{(v)} v^{(f)}\left(n-1\right) + \left(1 - \alpha^{(v)}\right) v\left(n\right) \tag{C.16}$$

gdzie $\alpha^{(v)} = 0.8$ oraz wyznacz zmienny w czasie próg $v^{(tr)}$, którego przekroczenie traktowane będzie jako wystąpienie fragmentu dźwięcznego w sygnale:

$$v^{(f,d)}\left(n\right) = \alpha^{(v,d)} v^{(f,d)}\left(n-1\right) + \left(1 - \alpha^{(v,d)}\right) v\left(n\right) \tag{C.17}$$

$$v^{(tr)}\left(n\right) = v^{(f,d)}\left(n\right) + T^{(v)} \tag{C.18}$$

gdzie $\alpha^{(v,d)} = 0.997$, a $T^{(v)} = 0.2$.

Na rys. C.2 zaprezentowano przykład działania algorytmu VAD.

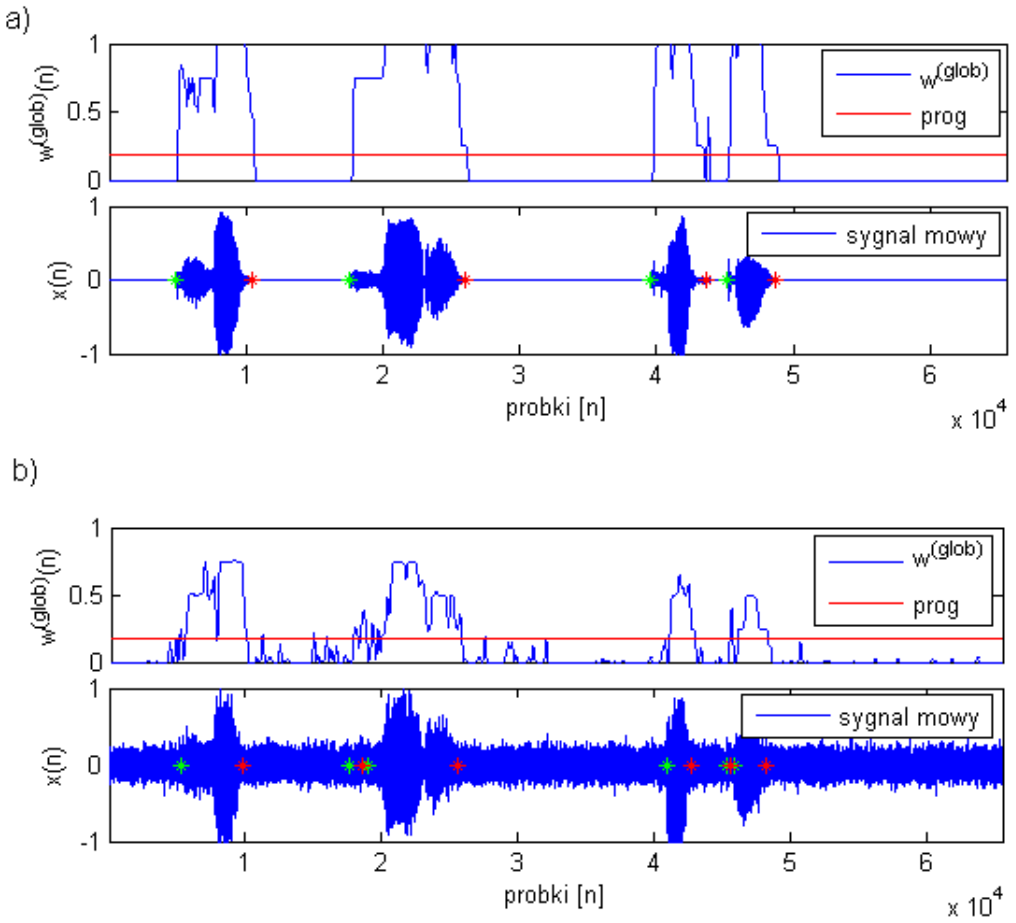
C.2. Wariant A systemu

W wariancie A systemu każdemu fonemowi odpowiada jeden stan HMM. Zastosowano jedną wspólną macierz prawdopodobieństw przejść międzyfonemowych i wektor prawdopodobieństw fonemów początkowych dla wszystkich wyrazów. Rozpoznawanie jest kilkuetapowe: najpierw wyznacza się sekwencję fonemów, następnie dzieli się ją sekwencję pseudosylab, a tę z kolei porównuje się z wzorcami wyrazów w słowniku. Wykorzystano również rozkłady prawdopodobieństwa czasów trwania fonemów. Ogólny schemat wariantu A systemu przedstawiono na rys. C.3.

C.2.1. Uczenie

Rozkłady prawdopodobieństwa współczynników MFCC

Rozkłady prawdopodobieństwa współczynników MFCC dla fonemów wyznaczano na podstawie segmentacji i etykietyzacji dostępnej w bazie nagrań. Ograniczono się

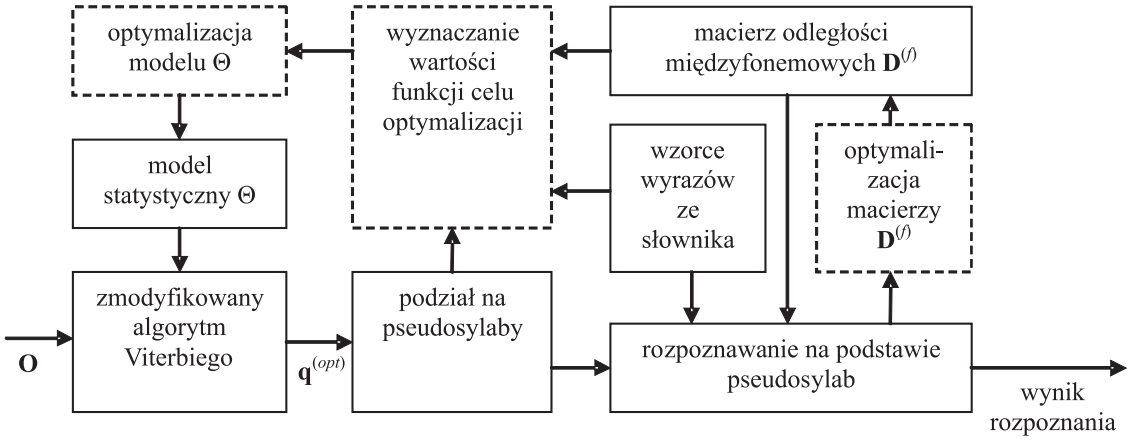


Rys. C.2. Przykład działania algorytmu VAD w przypadku: a) sygnału nie zaszumionego, b) sygnału z dodanym szumem zarejestrowanym wewnątrz samochodu jadącego autostradą. Sygnał zawiera wyrazy „trzy", „zero", „kropka" wypowiadane przez dwie różne osoby.

do modelowania rozkładów brzegowych. Wykorzystano podany niżej algorytm E-M estymacji wartości parametrów modelu GMM dla i-tego fonemu i m-tego współczynnika MFCC:

1. Wyznacz centroidy dla dostępnego w estymacji zbioru N wartości współczynników MFCC $\{o_{m,i,n} : n = 0, \ldots, N-1\}$. Liczba centroidów równa jest liczbie składowych GMM. Zastosowano iteracyjny algorytm Lloyda.

2. Zainicjalizuj wartości oczekiwane składowych GMM $\mu_{m,i,k}$ wartościami wyznaczonych centroidów, zainicjalizuj wagi $c_{m,i,k}$ i wariancje $\sigma^2_{m,i,k}$ na podstawie zbiorów wartości współczynników przynależnych do danego centroidu.



Rys. C.3. Ogólny schemat wariantu A systemu ARM. Bloki zaznaczone linią przerywaną występują tylko w etapie uczenia.

3. Iteracyjnie estymuj wartości parametrów GMM zgodnie z zależnościami:

$$\bar{c}_{m,i,k} = \frac{1}{N} \sum_{n=0}^{N-1} P\left(k|o_{m,i,n}\right) \tag{C.19}$$

$$\bar{\mu}_{m,i,k} = \frac{\sum_{n=0}^{N-1} P\left(k|o_{m,i,n}\right) \cdot o_{m,i,n}}{\sum_{n=o}^{N-1} P\left(k|o_{m,i,n}\right)} \tag{C.20}$$

$$\bar{\sigma}^2_{m,i,k} = \frac{\sum_{n=0}^{N-1} P\left(k|o_{m,i,n}\right) \cdot \left(o_{m,i,n} - \mu_{m,i,k}\right)^2}{\sum_{n=0}^{N-1} P\left(k|o_{m,i,n}\right)} \tag{C.21}$$

gdzie prawdopodobieństwo $P\left(k|o_{m,i,n}\right)$ przynależności wartości $o_{m,i,n}$ do k-tej spośród K składowych GMM wyznaczane jest z reguły Bayesa:

$$P\left(k|o_{m,i,n}\right) = \frac{c_{m,i,k} \cdot \mathcal{N}\left(o_{m,i,n}; \mu_{m,i,k}, \sigma^2_{m,i,k}\right)}{\sum_{l=0}^{K-1} c_{m,i,l} \cdot \mathcal{N}\left(o_{m,i,n}; \mu_{m,i,l}, \sigma^2_{m,i,l}\right)} \tag{C.22}$$

Kreska nad symbolem oznacza wartość uaktualnianą w bieżącej iteracji, brak kreski - wartość z poprzedniej iteracji. Algorytm zatrzymywany jest wtedy, gdy wzrost miary dopasowania modelu statystycznego do danych wejściowych (prawdopodobieństwa $P\left(\left\{o_{m,i,n} : n = 0, \ldots, N-1\right\} \mid \left\{c_{m,i,k}, \mu_{m,i,k}, \sigma^2_{m,i,k}\right\}\right)$) względem poprzedniej iteracji jest poniżej zadanego progu.

Stosowano $K = 6$ składowych GMM. W celu przyspieszenia dalszych obliczeń rozkłady prawdopodobieństwa zostały stablicowane.

Rozkłady prawdopodobieństwa czasów trwania fonemów

Rozkłady prawdopodobieństwa czasów trwania fonemów $p_i^{(d)}(t)$ modelowano jako sumy rozkładów logarytmiczno-normalnych z uwagi na fakt, że czas trwania nie może być ujemny. Wykorzystano w tym celu tę samą metodę estymacji wartości parametrów, co w przypadku GMM, z tym, że danymi wejściowymi były tutaj zlogarytmowane czasy trwania fonemów, uzyskane z części uczącej bazy nagrań. Wyznaczone wartości parametrów rozkładów normalnych dla danych zlogarytmowanych odpowiadają wartościom parametrów rozkładów logarytmiczno-normalnych dla danych niezlogarytmowanych. Stosowano od dwóch do siedmiu składowych logarytmiczno-normalnych.

Prawdopodobieństwa obserwacji fonemów w ramce ciszy

Algorytm VAD wykrywa fragmenty zwarć bezdźwięcznych w wypowiedziach. Prawdopodobieństwo obserwacji fonemu i w ramkach obejmujących te fragmenty (ramkach ciszy) oznaczono jako $P_i^{(cisz)}$ i podstawiano za $p_i(\mathbf{o})$. Prawdopodobieństwa te wyznaczano w trakcie optymalizacji wartości parametrów modelu. Tylko w ramkach ciszy występuje niezerowe prawdopodobieństwo pseudofonemu ciszy '#', we fragmentach oznaczonych przez VAD jako mowa, równe jest ono zeru.

Zmodyfikowany algorytm Viterbiego

Algorytm Viterbiego jest stosowany w celu znalezienia sekwencji stanów $\mathbf{q}^{(opt)}$ odpowiadającej sekwencji obserwacji $\mathbf{O}$ przy zadanym modelu $\Theta = \left\{\mathbf{A}, \boldsymbol{\pi}, p_i(\mathbf{o}), p_i^{(d)}(t), P_i^{(cisz)}\right\}$. Algorytm ten jest wykorzystywany zarówno w etapie uczenia systemu, jak i podczas jego pracy użytkowej. Ilość stanów modelu wynosi I, a długość sekwencji obserwacji - T. Opis podstawowej wersji algorytmu, w której nie uwzględniono prawdopodobieństw czasów trwania fonemów, można znaleźć np. w [133]. Poniżej opisano algorytm uwzględniający te prawdopodobieństwa oraz działający na wartościach zlogarytmowanych.

1. Wyznacz dla $i, j = 0, \ldots, I-1$ wartości zlogarytmowane:

$$a_{ij}^{(l)} = \ln a_{ij} \tag{C.23}$$

$$\pi_i^{(l)} = \ln \pi_i \tag{C.24}$$

$$p_i^{(l)}(\mathbf{o}_t) = \ln\left(p_i(\mathbf{o}_t)\right), \qquad t = 0, \ldots, T-1 \tag{C.25}$$

$$p_i^{(d,l)}(t) = \eta \cdot \ln\left(p_i^{(d)}(t)\right) - (t-1) \cdot \ln a_{ii}, \qquad t = 1, \ldots, T^{(d,max)} \tag{C.26}$$

$$p_i^{(r,d,l)}(t) = \eta \cdot \ln\left(1 - \sum_{l=1}^{t} p_i^{(d)}(l)\right) - (t-1) \cdot \ln a_{ii}, \quad t = 1, \ldots, T^{(d,max)} \tag{C.27}$$

gdzie $T^{(d,max)}$ oznacza maksymalny uwzględniany przez rozkłady czas trwania fonemu, równy 50 ramek. Dla $t > T^{(d,max)}$ przyjmuje się $p_i^{(d,l)}(t) = p_i^{(d,l)}\left(T^{(d,max)}\right)$ i $p_i^{(r,d,l)}(t) = p_i^{(r,d,l)}\left(T^{(d,max)}\right)$.

2. Inicjalizacja (dla $i = 0, \ldots, I-1$):

$$\delta_{i,0}^{(l)} = \pi_i^{(l)} + p_i^{(l)}(\mathbf{o}_0) \tag{C.28}$$

$$\psi_{i,0} = 0 \tag{C.29}$$

$$\xi_{i,0} = 1 \tag{C.30}$$

3. Rekurencja. Powtarzaj dla $t = 1, \ldots, T-1$:

$$\psi_{i,t} = \underset{0 \le j \le I-1}{\arg\max} \left(\delta_{j,t-1}^{(l)} + a_{ji}^{(l)} + \bar{p}_j^{(d,l)} \right), \qquad i = 0, \ldots, I-1 \tag{C.31}$$

$$\bar{p}_j^{(d,l)} = \begin{cases} p_j^{(d,l)}(\xi_{j,t-1}), & \text{dla } j \neq i \\ p_i^{(r,d,l)}(\xi_{i,t-1}), & \text{dla } j = i \end{cases} \tag{C.32}$$

$$\delta_{i,t}^{(l)} = \begin{cases} \delta_{\psi_{i,t},t-1}^{(l)} + a_{\psi_{i,t}i}^{(l)} + p_i^{(l)}(\mathbf{o}_t) + p_{\psi_{i,t}}^{(d,l)}\left(\xi_{\psi_{i,t},t-1}\right), & \text{dla } \psi_{i,t} \neq i \\ \delta_{i,t-1}^{(l)} + a_{ii}^{(l)} + p_i^{(l)}(\mathbf{o}_t), & \text{dla } \psi_{i,t} = i \end{cases}, i = 0, \ldots, I-1 \tag{C.33}$$

$$\xi_{i,t} = \begin{cases} 1, & \text{dla } \psi_{i,t} \neq i \\ \xi_{i,t-1} + 1, & \text{dla } \psi_{i,t} = i \end{cases}, \quad i = 0, \ldots, I-1 \tag{C.34}$$

4. Zakończenie:

$$\bar{\delta}_{i,T-1}^{(l)} = \delta_{i,T-1}^{(l)} + p_i^{(d,l)}(\xi_{i,T-1}), \quad i = 0, \ldots, I-1 \tag{C.35}$$

$$P^{(s,l)} = \max_{0 \le i \le I-1} \bar{\delta}_{i,T-1}^{(l)} \tag{C.36}$$

$$q_{T-1}^{(opt)} = \underset{0 \le i \le I-1}{\arg\max}\, \bar{\delta}_{i,T-1}^{(l)} \tag{C.37}$$

gdzie $P^{(s,l)}$ oznacza zlogarytmowane prawdopodobieństwo znalezionej sekwencji stanów.

5. Wyznacz sekwencję stanów:

$$q_t^{(opt)} = \psi_{q_{t+1}^{(opt)},t+1}, \quad t = T-2, T-3, \ldots, 0 \tag{C.38}$$

Współczynnik $\eta > 0$ jest wagą nałożoną na prawdopodobieństwa czasów trwania fonemów i został przyjęty równy 3 na podstawie wstępnych badań. Równania (C.26) i (C.27) zawierają składnik korygujący rozkłady prawdopodobieństwa tych czasów, związany z wartością prawdopodobieństwa a_{ii} przejścia fonemu na samego siebie.

W drugim etapie rozpoznawania system dzieli sekwencję fonemów na skoncentrowane wokół samogłosek pseudosylaby, dlatego istotne jest, żeby algorytm Viterbiego podawał prawidłowe rozpoznanie pozycji występowania samogłosek. Błędy takie, jak pominięcie samogłoski czy rozpoznanie dwóch samogłosek jako jednej, są trudne do skorygowania po wyznaczeniu pseudosylab i prowadzą najczęściej do błędnej klasyfikacji rozpoznawanego słowa. W przypadku stosowania algorytmu Viterbiego, nawet uwzględniającego prawdopodobieństwa czasów trwania fonemów, często dochodzi do „zlania" się w rozpoznaniu dwóch samogłosek w jedną, gdy fragment je rozdzielający jest krótki lub niewyraźny, a sąsiadujące z nim samogłoski należą do tego samego fonemu bądź fonemów podobnych z punktu widzenia odległości międzyfonemowej. Aby zminimalizować wspomniane zjawiska, wprowadzono mechanizm wstępnej segmentacji polegającej na wykrywaniu granic pseudosylab. Jego działanie przedstawia się następująco:

1. Wyznacz prawdopodobieństwa występowania samogłosek i fonemu 'j' dla ramek $t = 0, \ldots, T-1$:

$$P_t^{(samo)} = \frac{\sum\limits_{i \text{ dla samogłosek i fonemu 'j'}} p_i\left(\mathbf{o}_t\right)}{\sum\limits_{i=0}^{I-1} p_i\left(\mathbf{o}_t\right)} \tag{C.39}$$

2. Wyznacz z $P_t^{(samo)}$ średnią krótkoterminową $P_t^{(samo,kt)}$ i długoterminową $P_t^{(samo,dt)}$:

$$P_t^{(samo,kt)} = \frac{1}{M^{(p)}} \sum_{m=0}^{M^{(p)}-1} P_{t-m}^{(samo)} \tag{C.40}$$

$$P_t^{(samo,dt)} = (1-\alpha^{(p)}) P_{t-1}^{(samo,dt)} + \alpha^{(p)} P_t^{(samo,kt)} \tag{C.41}$$

   gdzie $M^{(p)} = 5$, a $\alpha^{(p)} = 0.3$. W powyższych równaniach uwzględniono następujące warunki początkowe: $M^{(p)} = t+1$ dla $t < 5$ i $P_0^{(samo,dt)} = P_0^{(samo)}$.

3. Wyznacz maksima lokalnych trajektorii $P_t^{(samo,kt)}$ i zachowaj do dalszej analizy te maksima, których wartość przekroczyła próg równy 0.5.

4. Określ przedziały występowania samogłosek następującą metodą: dla każdego wykrytego w punkcie 3. maksimum wyznacz przedział czasu, w którym trajektoria $P_t^{(samo,kt)}$ przyjmuje wartości większe niż wartość danego maksimum pomniejszona o stałą 0.2. Zachodzące na siebie przedziały połącz. Przedziały pozostałe po wyodrębnieniu przedziałów samogłoskowych uznawane są za przedziały braku obecności samogłosek.

5. Przeanalizuj każdy przedział samogłoskowy w celu sprawdzenia, czy nie nastąpiło „zlanie" się dwóch samogłosek w jedną.

   (a) Przejdź do kolejnych etapów analizy, jeżeli długość wykrytego przedziału samogłoskowego jest dłuższa niż 200 ms. W przeciwnym wypadku zaklasyfikuj przedział jako pojedynczą samogłoską.

   (b) Wyznacz trajektorię $P_t^{(samo,r)}$ będącą różnicą między $P_t^{(samo,kt)}$ i trajektorią łączącą maksima wykryte w punkcie 3. odcinkami linii prostych.

   (c) Określ wartość maksymalną trajektorii $P_t^{(samo,r)}$ i jeśli przekracza ona próg równy 0.05 oraz jednocześnie wartość $P_t^{(samo,dt)}$ jest większa niż wartość $P_t^{(samo,kt)}$ w tym punkcie, to staje się on potencjalnym punktem rozdziału analizowanego przedziału samogłoskowego.

   (d) Rozdziel przedział samogłoskowy, gdy długości przedziałów powstałych po rozdzieleniu, równe τ_1 i τ_2, nie różnią się zbyt istotnie. Warunek ten jest spełniony, gdy współczynnik proporcji długości $\tau^{(p)}$, dany poniższym równaniem, nie przekracza wartości 3.

$$\tau^{(p)} = \frac{\tau_1}{\tau_2} + \frac{\tau_2}{\tau_1} \tag{C.42}$$

   (e) Ustal granice przedziałów samogłoskowych po przeprowadzonym podziale w punktach odpowiadających jednej ramce przed i jednej ramce za wyznaczonym punktem podziału.

6. Po rozdzieleniu przedziałów samogłoskowych rozpocznij procedurę szukania przedziałów do podziału od początku, tj. wróć do punktu 5. Powtarzaj procedurę tak długo, aż nie nastąpi żaden podział.

Uzyskane w powyżej opisanej metodzie granice przedziałów samogłoskowych służą do wprowadzenia następujących modyfikacji działania algorytmu Viterbiego:

1. Znalezienie w każdym przedziale rozdzielającym przedziały samogłoskowe punktu minimum trajektorii $P_t^{(samo,kt)}$. W punkcie tym wymuszana jest zmiana stanu, o ile bieżący stan odpowiada fonemowi samogłoskowemu. Wymuszenie następuje poprzez dodanie do zlogarytmowanego prawdopodobieństwa przejścia na ten sam fonem $a_{ii}^{(l)}$ kary równej (-100).

2. Ustalenie zlogarytmowanego prawdopodobieństwa obserwacji $p_i^{(l)}(\mathbf{o}_t)$ dla fonemów spółgłoskowych równego (-100) w punktach maksimów lokalnych trajektorii $P_t^{(samo,kt)}$, o ile wartości tych maksimów są większe od 0.9.

3. W celu zminimalizowania wpływu błędów powstających na skutek niskiej rozpoznawalności fragmentów wybrzmiewania samogłosek kończących wypowiedź, przeprowadza się skracanie wypowiedzi. Odrzucane są jej końcowe ramki (maksymalnie 5) od miejsca, w którym spełnione są następujące warunki:

(a) Wartość średnia trajektorii $P_t^{(samo,kt)}$, obliczona z ramek występujących po analizowanym punkcie, jest większa niż 0.1.

(b) Wśród ramek występujących po analizowanym punkcie istnieje co najmniej jedna, w której $P_t^{(samo,dt)} - P_t^{(samo,kt)} > 0.05$

Wyznaczenie sekwencji pseudosylab

Uzyskany z algorytmu Viterbiego ciąg fonemów poddawany jest kompresji, tj. ciągi tych samych fonemów zastępowane są jednym fonemem. Ponadto usuwany jest pseudofonem ciszy i zwarcia bezdźwięcznego '#' oraz pseudofonem zwarcia dźwięcznego '@'. Powstały ciąg fonemów dzieli się 5-elementowe pseudosylaby. Centralną pozycję w pseudosylabach zajmuje samogłoska, po której lewej i prawej stronie umieszczane są spółgłoski zawarte między sąsiednimi samogłoskami, np.:

- „zero" → „- z e r -" „- r o - -"
- „nowalińja" → „- n o w -" „- w a l - „ „- l i ń j" „ń j a - -"

Macierz odległości międzyfonemowych

Macierz odległości międzyfonemowych $\mathbf{D}^{(f)}$ umożliwia określenie podobieństwa dowolnych dwóch fonemów. Im wartość odległości jest większa, tym fonemy są mniej podobne do siebie. Elementy $d_{ij}^{(f)}$ macierzy $\mathbf{D}^{(f)}$ wyznaczano na postawie rozkładów prawdopodobieństwa współczynników MFCC w następujący sposób:

$$d_{ij}^{(f)} = -\ln\left(\frac{d_{ij}^{(f,n)}}{\sqrt{d_{ii}^{(f,n)} \cdot d_{jj}^{(f,n)}}}\right) \tag{C.43}$$

$$d_{ij}^{(f,n)} = \int\limits_{\mathscr{R}^{\dim(\mathbf{o})}} p_i(\mathbf{o}) \cdot p_j(\mathbf{o})\, d\mathbf{o} \tag{C.44}$$

gdzie $p_i(\mathbf{o})$ oznacza rozkład prawdopodobieństwa współczynników MFCC dla i-tego fonemu. W przypadku, gdy łączne rozkłady $p_i(\mathbf{o})$ faktoryzowane są na rozkłady brzegowe, a te modelowane są metodą GMM, równanie (C.44) ma następujące rozwiązanie analityczne:

$$d_{ij}^{(f,n)} = \prod_{m=0}^{\dim(\mathbf{o})-1} \sum_{k=0}^{K-1} \sum_{l=0}^{K-1} c_{m,i,k} \cdot c_{m,j,l} \cdot \mathscr{N}\left(0; \mu_{m,i,k} - \mu_{m,j,l}, \sigma_{m,i,k}^2 + \sigma_{m,j,l}^2\right) \tag{C.45}$$

gdzie $\left\{c_{m,i,k}, \mu_{m,i,k}, \sigma_{m,i,k}^2\right\}$ oznaczają parametry k-tej składowej GMM dla fonemu i i współczynnika MFCC m.

W celu zwiększenia przejrzystości zapisu wyrażeń podanych poniżej, wprowadzono oznaczenie $d_{i,j}^{(f)}$ równoznaczne oznaczeniu $d_{ij}^{(f)}$.

Funkcja celu optymalizacji wartości parametrów modelu

W procesie optymalizacji wartości parametrów modelu statystycznego wykorzystywana jest funkcja celu, określająca poprawność rozpoznania wypowiedzi ze zbioru uczącego bazy nagrań. Bazuje ona na odległości międzysylabowej. Niech $i_j^{(r)}$ oznacza j-ty licząc od lewej fonem w i-tej sylabie wypowiedzi rozpoznanej przez system, a $i_j^{(w)}$ - j-ty od lewej fonem w i-tej sylabie wzorca prawidłowego rozpoznania tej wypowiedzi. Odległość między sylabami dana jest zależnością:

$$d^{(syl)}_{i^{(r)},i^{(w)}} = 2 \cdot d^{(f)}_{i_3^{(r)},i_3^{(w)}} + d^{(syl,l)}_{i^{(r)},i^{(w)}} + d^{(syl,p)}_{i^{(r)},i^{(w)}} \tag{C.46}$$

gdzie $d^{(syl,l)}_{i^{(r)},i^{(w)}}$ oznacza odległość między odpowiadającymi sobie częściami sylab po lewej stronie samogłoski, a $d^{(syl,p)}_{i^{(r)},i^{(w)}}$ - po stronie prawej. Zasady obliczania tych odległości zawiera tabela C.1. Zastosowano następujące oznaczenia: '-' - brak fonemu na danej pozycji, '*' - obecność fonemu na danej pozycji. Podano przypadki dla lewej strony pseudosylaby, dla strony prawej stosuje się te zasady symetrycznie.

Tab. C.1. Zasady obliczania odległości międzysylabowych.

lewa str. sylaby $i^{(w)}$	lewa str. sylaby $i^{(r)}$	$d^{(syl,l)}_{i^{(r)},i^{(w)}}$
- -	- -	0
- -	-*	$d^{(f)}_{i_2^{(r)},'\#'}$
- -	**	$d^{(f)}_{i_1^{(r)},'\#'} + d^{(f)}_{i_2^{(r)},'\#'}$
-*	- -	$d^{(f)}_{i_2^{(w)},'\#'}$
-*	-*	$d^{(f)}_{i_2^{(r)},i_2^{(w)}}$
-*	**	$\min\left(d^{(f)}_{i_1^{(r)},i_2^{(w)}}, d^{(f)}_{i_2^{(r)},i_2^{(w)}}\right) + \min\left(d^{(f)}_{i_1^{(r)},i_2^{(r)}}, d^{(f)}_{i_j^{(r)},'\#'}\right)$ $j = \arg\max_{j=1,2} d^{(f)}_{i_j^{(r)},i_2^{(w)}}$
**	- -	$d^{(f)}_{i_1^{(w)},'\#'} + d^{(f)}_{i_2^{(w)},'\#'}$
**	-*	$\min\left(d^{(f)}_{i_1^{(w)},i_2^{(r)}}, d^{(f)}_{i_2^{(w)},i_2^{(r)}}\right) + \min\left(d^{(f)}_{i_1^{(w)},i_2^{(w)}}, d^{(f)}_{i_j^{(w)},'\#'}\right)$ $j = \arg\max_{j=1,2} d^{(f)}_{i_j^{(r)},i_2^{(w)}}$
**	**	$d^{(f)}_{i_1^{(r)},i_1^{(w)}} + d^{(f)}_{i_2^{(r)},i_2^{(w)}}$

Dla pojedynczej rozpoznanej wypowiedzi sumuje się odległości $d^{(syl)}$ między kolejnymi sylabami rozpoznanymi i odpowiadającymi im sylabami wzorca. W przypadku, gdy liczba sylab we wzorcu i wyrazie rozpoznawanym jest różna, znajdywana jest taka kombinacja dopasowania sylab do siebie, dla której sumaryczna odległość

międzysylabowa jest najmniejsza. W przypadku takim do odległości dodawana jest również kara za różną ilość sylab, wynosząca 30 za każdą pominiętą lub dodaną sylabę. Dodaje się ponadto karę za każdą pozostawioną spółgłoskę, tj. nie przypisaną do żadnej sylaby, równą średniej odległości tej spółgłoski od pozostałych fonemów.

Funkcja celu optymalizacji wyznaczana jest jako suma odległości wyznaczonych dla wszystkich wypowiedzi należących do słownika i wszystkich mówców ze zbioru uczącego bazy nagrań.

Optymalizacja wartości parametrów $\mathbf{A}$, $\boldsymbol{\pi}$, $p_i^{(d)}(t)$ i $P_i^{(cisz)}$

Poprzedzająca optymalizację inicjalizacja wartości w macierzy $\mathbf{A}$ i wektorze $\boldsymbol{\pi}$ dokonywana jest na podstawie danych o segmentacji zawartych w bazie nagrań, z których wyznacza się częstości przejść międzyfonemowych. Inicjalizacja wartości prawdopodobieństw obserwacji fonemów w ramce ciszy $P_i^{(cisz)}$ polega na zadaniu prawdopodobieństwa równego 0.99 dla fonemu ciszy, a dla pozostałych fonemów równych wartości prawdopodobieństwa takich, aby ich suma dla wszystkich fonemów wynosiła 1.

Optymalizacja odbywa się na zasadzie minimalizacji zdefiniowanej wyżej funkcji celu. Z uwagi na własności tej funkcji, która jest obszarami stała, zastosowano metodę poszukiwań prostych. W każdej iteracji następuje losowanie kolejności zmian wartości elementów macierzy $\mathbf{A}$, wektora $\boldsymbol{\pi}$ i prawdopodobieństw $P_i^{(cisz)}$. Następnie wartości te zmieniane są w pewnym zakresie i równocześnie zmieniane są w sposób proporcjonalny wartości wszystkich elementów od nich uzależnionych, co wynika z konieczności sumowania się do jedności prawdopodobieństw $P_i^{(cisz)}$, prawdopodobieństw w wektorze $\boldsymbol{\pi}$ oraz prawdopodobieństw w wierszach macierzy $\mathbf{A}$. Zmiany wartości elementów wyznaczają zatem zmienne kierunki optymalizacji. Pomiar funkcji celu odbywa się w punktach równo rozłożonych w danym zakresie zmian i zapamiętywane jest położenie punktu, dla którego wartość funkcji celu jest najmniejsza. Zakres poszukiwań dla każdego elementu zmieniany jest z iteracji na iterację w zależności od uzyskanych rezultatów. Co pewną liczbę iteracji następuje powrót do początkowych wartości zakresów, co ma na celu uniknięcie zatrzymania się optymalizacji w lokalnym minimum funkcji celu. Optymalizacja zatrzymywana jest po przeprowadzeniu zadanej liczby iteracji.

Optymalizacja rozkładów prawdopodobieństwa czasów trwania fonemów $p_i^{(d)}$ wykonywana jest po zakończeniu każdej iteracji optymalizacji wartości elementów $\mathbf{A}$, $\boldsymbol{\pi}$ i $P_i^{(cisz)}$. Wykorzystuje się tę samą, co opisana powyżej, metodę, a zmianie podlegają współczynniki liniowego skalowania osi czasu dla rozkładów logarytmiczno-normalnych.

Optymalizacja macierzy odległości międzyfonemowych

Wartości elementów macierzy $\mathbf{D}^{(f)}$ podlegają ograniczonej optymalizacji tak, by maksymalizować końcową rozpoznawalność systemu (zob. rozdział C.2.2). Jest

ona przeprowadzana po wyznaczaniu wartości parametrów modelu statystycznego. Metodą poszukiwań prostych wyznaczane są nieujemne wartości wektora wag $\mathbf{w}^{(f)}$, który służy do następującej modyfikacji macierzy $\mathbf{D}^{(f)}$:

$$\mathbf{D}^{(f,opt)} = \mathrm{diag}\left(\mathbf{w}^{(f)}\right) \cdot \mathbf{D}^{(f)} \cdot \mathrm{diag}\left(\mathbf{w}^{(f)}\right) \tag{C.47}$$

C.2.2. Rozpoznawanie

Rozpoznawanie z wykorzystaniem odległości międzysylabowych polega na wyznaczeniu dla rozpoznawanej wypowiedzi pseudoprawdopodobieństw $P_k^{(wyr)}$, określających jej podobieństwo do k-tego wyrazu ze słownika, a następnie wyborze jako rozpoznanego tego wyrazu k, dla którego $P_k^{(wyr)}$ przyjęło największą wartość. Wartości $P_k^{(wyr)}$ obliczano następująco:

1. Dla danej rozpoznawanej wypowiedzi wyznacz sekwencję fonemów za pomocą zmodyfikowanego algorytmu Viterbiego, a następnie przeprowadź jej podział na pseudolylaby. W dalszej analizie uwzględnij tylko te wyrazy ze słownika, dla których liczba sylab w ich wzorcach różni się o co najwyżej jedną w stosunku do liczby sylab w wyrazie rozpoznawanym.

2. Dla każdej sylaby $i^{(r)}$ w wyrazie rozpoznawanym wyznacz odległość do każdej sylaby wzorcowej $i^{(w)}$ metodą taką, jak przy obliczaniu funkcji celu optymalizacji wartości parametrów modelu statystycznego. Następnie uzyskane w ten sposób odległości zamień na pseudoprawdopodobieństwa wg zależności:

$$P^{(syl)}_{i^{(r)},i^{(w)}} = \frac{\left(d^{(syl)}_{i^{(r)},i^{(w)}} - \log\left(1-\beta^{(syl)}\right) \cdot \sum\limits_{l} d^{(syl)}_{i^{(r)},l^{(w)}}\right)^{-1}}{\sum\limits_{m}\left(d^{(syl)}_{i^{(r)},m^{(w)}} - \log\left(1-\beta^{(syl)}\right) \cdot \sum\limits_{l} d^{(syl)}_{i^{(r)},l^{(w)}}\right)^{-1}} \tag{C.48}$$

   gdzie $\beta^{(syl)}$ jest współczynnikiem „rozmywającym" pseudoprawdopodobieństwa i przyjmuje wartość od 0 (brak „rozmycia") do 1 (pełne „rozmycie", tj. wszystkie pseudoprawdopodobieństwa są równe). Sumowania we wzorze (C.48) wykonuje się po wszystkich sylabach wzorcowych występujących słowniku.

3. Określ pseudoprawdopodobieństwo wyrazu k korzystając z poniższej zależności:

$$P_k^{(wyr)} = \left(\max_{\left\{j_k^{(syl,r)},j_k^{(syl,w)}\right\}} \prod_i P^{(syl)}_{j^{(syl,r)}_{k,i},j^{(syl,w)}_{k,i}}\right)^{\frac{\left(\gamma^{(syl)}\right)^{\left|N^{(syl,r)}-N^{(syl,w)}\right|}}{\max\left(N^{(syl,r)},N^{(syl,w)}\right)}} \tag{C.49}$$

gdzie $\left\{j_k^{(syl,r)}, j_k^{(syl,w)}\right\}$ oznacza j-tą kombinację dopasowania sylab wyrazu rozpoznawanego do sylab wzorca wyrazu k. Dla danej kombinacji j wartości $j_{k,i}^{(syl,r)}$ oraz $j_{k,i}^{(syl,w)}$ oznaczają odpowiednio numer sylaby wyrazu rozpoznawanego i numer sylaby wzorcowej występującej na i-tej pozycji tej kombinacji. $N^{(syl,r)}$ oznacza liczbę sylab w wyrazie rozpoznawanym, $N^{(syl,w)}$ - w wyrazie wzorcowym, a $\gamma^{(syl)} > 1$ oznacza współczynnik kary za różną ilość sylab w wyrazie wzorcowym i rozpoznawanym.

Parametry $\beta^{(syl)} = 10^{-7}$ i $\gamma^{(syl)} = 3.4$ dobrano w trakcie wstępnych badań algorytmu tak, by uzyskać maksymalną skuteczność działania systemu.

C.3. Wariant B systemu

W wariancie B systemu przetwarzanie wstępne i VAD działają tak samo, jak w wariancie A. Od algorytmu Viterbiego konstrukcja jest zmieniona. Dla każdego wyrazu w słowniku tworzony jest model HMM mający tyle stanów, z ilu fonemów składa się dany wyraz. Kolejnym stanom przyporządkowane są rozkłady prawdopodobieństwa $p_i(\mathbf{o})$, odpowiadające kolejnym fonemom w wyrazie. Możliwe jest tylko przejście stanu na siebie oraz na stan następny. Nie wykorzystuje się wprowadzonego w wariancie A mechanizmu wyznaczania granic pseudosylab, wykorzystywane są natomiast prawdopodobieństwa obecności fonemów w ramce określonej przez VAD jako ramka ciszy. Rozpoznanie dokonywane jest bez podziału na pseudosylaby, wykorzystywane są natomiast zlogarytmowane prawdopodobieństwa $P^{(s,l)}$ sekwencji stanów znalezione algorytmem Viterbiego.

C.3.1. Uczenie

Rozkłady prawdopodobieństwa $p_i(\mathbf{o})$ i $p_i^{(d)}(t)$ wyznaczane są analogicznie, jak dla wariantu A. Podczas uczenia wariantu B systemu nie jest konieczna optymalizacja wartości elementów macierzy $\mathbf{A}$ i wektora $\boldsymbol{\pi}$. Prawdopodobieństwa przejść fonemu na siebie lub na fonem następny zależą tutaj tylko od rozkładów $p_i^{(d)}(t)$. Optymalizowane są zatem tylko prawdopodobieństwa $P_i^{(cisz)}$ oraz skale modyfikujące rozkłady $p_i^{(d)}(t)$. Optymalizacja dokonywana jest analogicznym algorytmem, jak w wariancie A z tym, że tutaj maksymalizowana jest następująca funkcja celu:

$$c^{(opt,B)} = \sum_{n=1}^{N} w^{(opt,B)}\left(\frac{1}{T_n}\left(P_{k_n,n}^{(s,l)} - \max_{m \neq k_n} P_{m,n}^{(s,l)}\right)\right) \tag{C.50}$$

$$w^{(opt,B)}(x) = \begin{cases} x, & \text{dla } x < 0.1 \\ 0.1, & \text{dla } x \geq 0.1 \end{cases} \tag{C.51}$$

gdzie N oznacza liczbę wypowiedzi w zbiorze uczącym, T_n - długość wypowiedzi n, k_n - numer prawidłowego wyrazu, który zawiera wypowiedź n, a $P_{k,n}^{(s,l)}$ - zlogarytmowane prawdopodobieństwo sekwencji stanów znalezione algorytmem Viterbiego dla wypowiedzi n uzyskane dla HMM wyrazu k.

C.3.2. Rozpoznawanie

Dla danej rozpoznawanej wypowiedzi wyznaczano zlogarytmowane prawdopodobieństwa $P_k^{(s,l)}$ sekwencji stanów znalezione algorytmem Vitebiego dla każdego k-tego modelu HMM, odpowiadającego k-temu wyrazowi w słowniku. Jako wyraz rozpoznany wybierano ten, dla którego prawdopodobieństwo to było największe.

C.4. Wariant At systemu

Wariant At różni się od wariantu A następującymi modyfikacjami:

- Zastosowano modelowanie fonemów za pomocą dwóch lub trzech stanów HMM. Po zrezygnowaniu z pseudofonemu zwarcia bezdźwięcznego '@' liczba fonemów wyniosła 36, a liczba stanów HMM - 101.
- Z uwagi na fakt, że wielostanowe modelowanie fonemów umożliwia w pewnym zakresie modelowanie prawdopodobieństwa czasów ich trwania (poprzez wartości elementów w macierzy prawdopodobieństw przejść między stanami), zrezygnowano z bezpośredniego modelowania prawdopodobieństwa czasów trwania fonemów.
- Zrezygnowano z wykrywania granic pseudosylab i związanej z tym modyfikacji algorytmu Viterbiego.
- Algorytm VAD stosowany jest do wykrywania tylko początku i końca wypowiedzi. Zwarcia bezdźwięczne modelowane są za pomocą odpowiednich stanów HMM fonemów. W związku z tym nie stosuje się prawdopodobieństw obserwacji fonemów w ramkach ciszy.
- Zastosowano elementy modelowania 2-gram w odniesieniu do fonemów. W zmodyfikowanym algorytmie Viterbiego nie są dopuszczalne trójfonemowe sekwencje, które nie występują w słowniku.

C.4.1. Uczenie

Wielostanowe modele fonemów

Topologie zastosowanych modeli fonemów przedstawiono w tabeli C.2. We wszystkich modelach możliwe są tylko przejścia stanu na siebie lub na stan następny. Modele są niezależne od kontekstu. Motywacją użycia modeli dwustanowych był fakt, że samogłoski zwarte są krótkie i zawierają dwa główne segmenty fonetyczne: zwarcie i plozję. Aspiracja w języku polskim jest zjawiskiem rzadkim i występuje głównie na końcach wyrazów. W przypadku spółgłosek zwartych dźwięcznych model nie uwzględnia możliwości pominięcia zwarcia, natomiast w przypadku bezdźwięcznych można je pominąć. Podobnie postąpiono w przypadku dźwięcznych i bezdźwięcznych spółgłosek zwarto-trących. Model ciszy przyjęto taki sam jak model dla spółgłosek zwartych bezdźwięcznych.

Tab. C.2. Topologie wielostanowych modeli fonemów.

topologia	fonemy
liczba stanów: 3 stan początkowy: 1 stan końcowy: 3	'a' 'e' 'i' 'o' 'u' 'y' 'j' 'l' 'ł' 'r' 'f' 'w' 's' 'z' 'ś' 'ź' 'S' 'ż' 'h' 'Z' 'Ż' 'Ź' 'm' 'n' ń' 'N'
liczba stanów: 3 stan początkowy: 1 lub 2 stan końcowy: 3	'c' 'ć' 'C'
liczba stanów: 2 stan początkowy: 1 stan końcowy: 2	'b' 'd' 'g'
liczba stanów: 2 stan początkowy: 1 lub 2 stan końcowy: 2	'p' 't' 'k' '#'

Rozkłady prawdopodobieństwa współczynników MFCC dla każdego stanu modelowano metodą GMM przy pomocy sumy $K = 6$ wielowymiarowych rozkładów normalnych o przekątniowych macierzach kowariancji. Dla każdego fonemu estymowano wartości niezerowych elementów macierzy prawdopodobieństw przejść międzystanowych i wektora prawdopodobieństw stanów początkowych. Zastosowano trójetapową metodę wyznaczania wartości parametrów modeli:

1. Wykorzystując segmentację i etykietyzację dostępną w bazie nagrań wyznaczono wartości parametrów rozkładów prawdopodobieństwa współczynników MFCC. W przypadku modelu trzystanowego każdą realizację fonemu dzielono

na trzy równe fragmenty i rozkłady dla stanów modelu wyznaczano z wykorzystaniem danych z odpowiednich fragmentów. W przypadku modeli dwustanowych realizację fonemu dzielono na dwa fragmenty, przy czym pierwszy stanowił 3/4 długości fonemu, a drugi 1/4. Wyjątkiem był model ciszy, gdzie zastosowano podział na dwa równe fragmenty.

2. Dokonano estymacji wartości elementów macierzy prawdopodobieństw przejść międzystanowych i wektora prawdopodobieństw stanów początkowych dla każdego fonemu. Reestymowano ponadto parametry rozkładów prawdopodobieństwa współczynników MFCC. Estymację przeprowadzano z wykorzystaniem izolowanych realizacji fonemów, zgodnie z segmentacją bazy nagrań. Zastosowano opisany poniżej algorytm Bauma-Welcha.

   Załóżmy, że dostępny jest zbiór N sekwencji obserwacji $\{\mathbf{O}_0, \ldots, \mathbf{O}_{N-1}\}$, długość sekwencji n-tej wynosi T_n. Celem jest estymacja wartości parametrów modelu $\Theta = \left\{\mathbf{A}, \boldsymbol{\pi}, c_{i,k}, \boldsymbol{\mu}_{i,k}, \boldsymbol{\Sigma}_{i,k}\right\}$. Algorytm Bauma-Welcha [133] jest iteracyjnym algorytmem typu E-M, a każda iteracja składa się z następujących kroków:

   (a) Przeprowadź procedurę *forward* dla każdej sekwencji n:
   Inicjalizacja:

$$\alpha_{n,i,0} = \pi_i p_i\left(\mathbf{o}_{n,0}\right), \quad i = 0, \ldots, I-1 \tag{C.52}$$

   Indukcja:

$$\alpha_{n,i,t} = \left(\sum_{j=0}^{I-1} \alpha_{n,j,t-1} a_{ji}\right) p_i\left(\mathbf{o}_{n,t}\right), \quad i = 0, \ldots, I-1, \; t = 1, \ldots, T_n - 1 \tag{C.53}$$

   (b) Przeprowadź procedurę *backward* dla każdej sekwencji n:
   Inicjalizacja:

$$\beta_{n,i,T_n-1} = 1, \quad i = 0, \ldots, I-1 \tag{C.54}$$

   Indukcja:

$$\beta_{n,i,t} = \sum_{j=0}^{I-1} a_{ij} p_j\left(\mathbf{o}_{n,t+1}\right) \beta_{n,j,t+1}, \quad i = 0, \ldots, I-1, \; t = T_n - 2, \ldots, 0 \tag{C.55}$$

(c) Uaktualnij wartości parametrów modelu:

$$\gamma_{n,i,t} = \frac{\alpha_{n,i,t}\beta_{n,i,t}}{\sum_{j=1}^{I-1} \alpha_{n,j,t}\beta_{n,j,t}}, \quad \begin{array}{l} i = 0, \ldots, I-1, \quad t = 0, \ldots, T_n - 1, \\ n = 0, \ldots, N-1 \end{array} \tag{C.56}$$

$$\gamma^{(gmm)}_{n,i,t,k} = \gamma_{n,i,t} \frac{c_{i,k}\mathcal{N}\left(\mathbf{o}_{n,t}; \boldsymbol{\mu}_{i,k}, \boldsymbol{\Sigma}_{i,k}\right)}{\sum_{m=0}^{K-1} c_{i,m}\mathcal{N}\left(\mathbf{o}_{n,t}; \boldsymbol{\mu}_{i,m}, \boldsymbol{\Sigma}_{i,m}\right)}, \quad \begin{array}{l} n = 0, \ldots, N-1, \\ i = 0, \ldots, I-1, \\ t = 0, \ldots, T_n - 1, \\ k = 0, \ldots, K-1 \end{array} \tag{C.57}$$

$$\bar{\pi}_i = \frac{1}{N} \sum_{n=0}^{N-1} \gamma_{n,i,0}, \quad i = 0, \ldots, I-1 \tag{C.58}$$

$$\bar{a}_{ij} = \frac{\sum_{n=0}^{N-1} \sum_{t=0}^{T_n-2} \frac{\alpha_{n,i,t} a_{ij} p_j(\mathbf{o}_{n,t+1}) \beta_{n,j,t+1}}{\sum_{k=0}^{I-1} \alpha_{n,k,t}\beta_{n,k,t}}}{\sum_{n=0}^{N-1} \sum_{t=0}^{T_n-2} \gamma_{n,i,t}}, \quad i, j = 0, \ldots, I-1 \tag{C.59}$$

$$\bar{c}_{i,k} = \frac{\sum_{n=0}^{N-1} \sum_{t=0}^{T_n-1} \gamma^{(gmm)}_{n,i,t,k}}{\sum_{n=0}^{N-1} \sum_{t=0}^{T_n-1} \gamma_{n,i,t}}, \quad \begin{array}{l} i = 0, \ldots, I-1, \\ k = 0, \ldots, K-1 \end{array} \tag{C.60}$$

$$\bar{\boldsymbol{\mu}}_{i,k} = \frac{\sum_{n=0}^{N-1} \sum_{t=0}^{T_n-1} \gamma^{(gmm)}_{n,i,t,k} \mathbf{o}_{n,t}}{\sum_{n=0}^{N-1} \sum_{t=0}^{T_n-1} \gamma^{(gmm)}_{n,i,t,k}}, \quad \begin{array}{l} i = 0, \ldots, I-1, \\ k = 0, \ldots, K-1 \end{array} \tag{C.61}$$

$$\bar{\boldsymbol{\Sigma}}_{i,k} = \frac{\sum_{n=0}^{N-1} \sum_{t=0}^{T_n-1} \gamma^{(gmm)}_{n,i,t,k} \left(\mathbf{o}_{n,t} - \boldsymbol{\mu}_{i,k}\right) \cdot \left(\mathbf{o}_{n,t} - \boldsymbol{\mu}_{i,k}\right)^T}{\sum_{n=0}^{N-1} \sum_{t=0}^{T_n-1} \gamma^{(gmm)}_{n,i,t,k}}, \quad \begin{array}{l} i = 0, \ldots, I-1, \\ k = 0, \ldots, K-1 \end{array} \tag{C.62}$$

Kreski nad symbolami oznaczają wartości uaktualnione w danej iteracji. W implementacji zastosowano dodatkowo skalowanie prawdopodobieństw [133] i inne mechanizmy zwiększające stabilność numeryczną algorytmu. Algorytm zatrzymywany jest w momencie, gdy wzrost miary dopasowania modelu statystycznego do danych wejściowych (prawdopodobieństwa $P\left(\{\mathbf{O}_0, \ldots, \mathbf{O}_{N-1}\} | \Theta\right)$) względem poprzedniej iteracji jest poniżej zadanego progu.

Warto zauważyć, że wykorzystywane przez niektóre algorytmy kompensacji prawdopodobieństwo sekwencji obserwacji $\mathbf{O}_n$ dla danych wartości parametrów modelu statystycznego Θ, dane jest następującą zależnością:

$$P\left(\mathbf{O}_n|\Theta\right) = \sum_{i=0}^{I-1} \alpha_{n,i,t}\beta_{n,i,t}, \quad \text{dla dowolnego } \; t = 0, \ldots, T_n - 1 \tag{C.63}$$

3. Zastosowano tzw. uczenie zanurzone (ang. *embedded training*) w celu umożliwienia algorytmowi uczenia systemu znalezienia optymalnych, z punktu widzenia metody estymacji, granic międzyfonemowych. Reestymacja wartości parametrów modelu dokonywana była metodą Bauma-Welcha z wykorzystaniem sekwencji trójfonemowych. Okno obejmujące trzy kolejne fonemy było przesuwane wzdłuż wypowiedzi ze skokiem co jeden fonem. Granice skrajnych fonemów w sekwencji nadal były zadane przez segmentację bazy nagrań, reestymowano tylko wartości parametrów dla fonemu środkowego, z wyjątkiem sytuacji, gdy skrajne fonemy były fonemami początkowymi lub końcowymi wypowiedzi, wtedy reestymowano również wartości parametrów ich modeli. Stosowano sekwencje trójfonemowe, a nie sekwencje fonemów dla całej wypowiedzi, w celu uniknięcia sytuacji prowadzących do powstania w procesie uczenia nowych segmentacji zawierających błędy grube, tj. takie, w których nowa segmentacja nie pokrywa się z faktycznie występującymi granicami fonemów, nawet uwzględniając tolerancję kilkudziesięciu milisekund.

Pełna macierz $\mathbf{A}$ i wektor $\boldsymbol{\pi}$, stosowane w wariancie At systemu, zbudowane są z odpowiednich macierzy prawdopodobieństw przejść międzystanowych i wektorów prawdopodobieństw stanów początkowych dla poszczególnych fonemów, zmodyfikowanych przez prawdopodobieństwa przejść międzyfonemowych. Prawdopodobieństwa przejść międzyfonemowych i prawdopodobieństwa fonemów początkowych optymalizowane są natomiast analogiczną metodą jak w wariancie A. Stosowana jest też ta sama, co w wariancie A, macierz odległości międzyfonemowych i przeprowadzana jest analogiczna jej optymalizacja.

C.4.2. Rozpoznawanie

Zaprojektowano zmodyfikowany algorytm Viterbiego, który nie zezwala na przejście międzyfonemowe, jeśli dana sekwencja trzech kolejnych fonemów nie występuje w słowniku. Taka modyfikacja wymusiła konieczność przetwarzania w algorytmie Viterbiego $K = 8$ najlepszych sekwencji stanów. Pewne sekwencje o wysokim prawdopodobieństwie mogą być bowiem zabronione w trakcie działania algorytmu i w takim przypadku rozpatrywane są sekwencje alternatywne. Trzeba zaznaczyć, że

odnoszą się one do sekwencji fonemów, a nie sekwencji stanów poszczególnych fonemów, w związku z czym w obrębie jednego fonemu, dla stanów tego fonemu, nie następuje zmiana ustalonego zestawu K najlepszych sekwencji.

Podział na pseudosylaby i pozostałe etapy rozpoznawania są analogiczne, jak w wariancie A.

C.5. Wariant Bt systemu

W wariancie Bt wprowadzono następujące zmiany w stosunku do wariantu B:

- Zastosowano modelowanie fonemów analogiczne, jak w wariancie At. Zrezygnowano z bezpośredniego modelowania prawdopodobieństwa czasów trwania fonemów.
- Podobnie jak w wariancie At, algorytm VAD stosowany jest do wykrywania tylko początku i końca wypowiedzi. Nie stosuje się prawdopodobieństw obserwacji fonemów w ramkach ciszy.
- Macierze prawdopodobieństw przejść międzystanowych i wektory prawdopodobieństw stanów początkowych dla modeli HMM wyrazów skonstruowane są na podstawie danych pochodzących z modeli fonemów.

Wariant Bt nie wymaga dodatkowego uczenia. Pozostałe etapy rozpoznawania są analogiczne, jak w wariancie B.

D. Charakterystyki symulowanych zniekształceń transmisyjnych

Liniowe zniekształcenia transmisyjne symulowano za pomocą charakterystyk amplitudowych pięciu popularnych mikrofonów oraz dwóch dodatkowych, nakładanych kaskadowo charakterystyk liniowych, symulujących zmiany w charakterystykach kierunkowych mikrofonów, związane z wzajemnym położeniem mikrofonu i mówcy. Charakterystyki liniowe mogą symulować dodatkowo wpływ akustyki pomieszczenia czy też pewne zmiany osobnicze, związane np. ze stanami emocjonalnymi mówcy, wpływającymi na obwiednię widma amplitudowego sygnału mowy. Charakterystyki symulowanych zniekształceń przygotowano w następujący sposób:

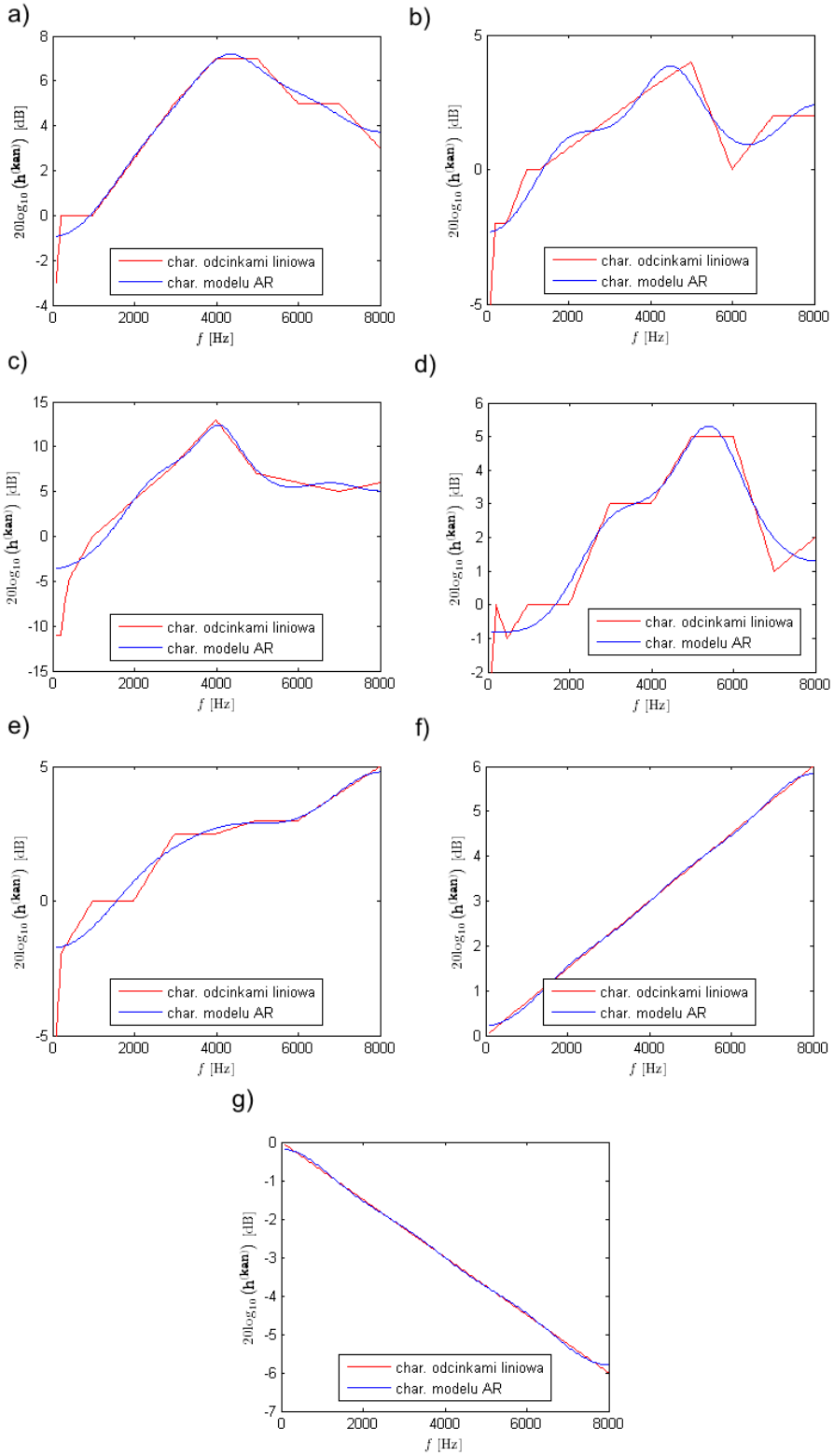
1. Dane z wykresów udostępnianych przez producentów mikrofonów zapisano w postaci charakterystyk odcinkami liniowych.

2. Obliczono podniesione do kwadratu widma amplitudowe charakterystyk wyznaczonych w punkcie 1. Przeprowadzając na nich odwrotną DFT wyznaczono funkcje autokorelacji dla każdej charakterystyki.

3. Za pomocą algorytmu Levinsona-Durbina dla każdej charakterystyki wyznaczono wartości współczynników modelu AR (ang. *auto regressive*) rzędu 7. Zastosowanie modelu AR wynikało z konieczności zapewnienia zgodności typu filtrów kompensujących zniekształcenia, uzyskanych poprzez odwrócenie funkcji transmitancji typu AR (a zatem będących filtrami FIR), z typem filtrów liniowych stosowanych w zaproponowanej metodzie kompensacji.

4. Stosując DFT wyznaczono charakterystyki amplitudowe $\mathbf{h}_i^{(kan)}$ na podstawie wartości współczynników modeli AR.

Na rys. D.1 przedstawiono charakterystyki symulowanych zniekształceń transmisyjnych $\mathbf{h}_i^{(kan)}$ oraz wyjściowe charakterystyki odcinkami liniowe.



Rys. D.1. Charakterystyki amplitudowe symulowanych zniekształceń transmisyjnych. Mikrofony: a) Shure PG48 ($\mathbf{h}_1^{(kan)}$), b) Shure PG58 ($\mathbf{h}_2^{(kan)}$), c) Skytronik ($\mathbf{h}_3^{(kan)}$), d) Shure SM58 ($\mathbf{h}_4^{(kan)}$), e) Shure SM86 ($\mathbf{h}_5^{(kan)}$). Charakterystyki liniowe: f) +6dB/8kHz ($\mathbf{h}_6^{(kan)}$), g) -6dB/8kHz ($\mathbf{h}_7^{(kan)}$).

E. Metodologia pomiaru rozpoznawalności izolowanych ramek

Rozpoznawalność izolowanych ramek oceniano przy pomocy dwóch zdefiniowanych poniżej miar. Miara oznaczona indeksem dolnym 1 wskazuje, jaki procent ramek został prawidłowo przyporządkowany do odpowiadających im fonemów, natomiast miara oznaczona indeksem 2 uwzględnia również dynamikę błędu klasyfikacji, wyrażonego różnicą logarytmu prawdopodobieństwa przynależności danej ramki do odpowiadającego jej fonemu i maksymalnej wartości logarytmu prawdopodobieństwa przynależności ramki do pozostałych fonemów.

Punktem wyjścia do zdefiniowania wspomnianych wyżej miar były następujące miary oceny rozpoznania pojedynczej ramki:

$$c_1^{(rr)}\left(\mathbf{o}_{n,i,s}\right) = u\left(\ln p_i\left(\mathbf{o}_{n,i,s}\right) - \max_{\substack{0\le j\le I-1\\ j\neq i}} \ln p_j\left(\mathbf{o}_{n,i,s}\right)\right) \tag{E.1}$$

i

$$c_2^{(rr)}\left(\mathbf{o}_{n,i,s}\right) = \ln p_i\left(\mathbf{o}_{n,i,s}\right) - \max_{\substack{0\le j\le I-1\\ j\neq i}} \ln p_j\left(\mathbf{o}_{n,i,s}\right) \tag{E.2}$$

gdzie u oznacza funkcję skoku jednostkowego, $\mathbf{o}_{n,i,s}$ - wektor współczynników MFCC dla mówcy s i n-tej ramki i-tego z I rozpoznawanych fonemów, a $p_i\left(\mathbf{o}\right)$ - zadany w systemie ARM rozkład prawdopodobieństwa współczynników MFCC dla fonemu i.

W celu wyznaczenia miary oceny uśrednionej dla danego mówcy s, dla każdego fonemu i losowano $N_i = 500$ ramek lub mniej, jeśli nie było tyle dostępnych. Wstępne badania wykazały, że taka liczba jest wystarczająca. Zbiór wektorów $\mathbf{o}$ dla wylosowanych ramek oznaczano jako $\mathscr{C}_{s,i}$. Estymator średniej rozpoznawalności izolowanych ramek dla mówcy s i zbioru fonemów $\mathscr{J}$ dany jest wzorem:

$$c_{k,s,\mathscr{J}}^{(rm)} = \frac{1}{\bar{\mathscr{J}}} \sum_{i\in\mathscr{J}} \frac{1}{\bar{\mathscr{C}}_{s,i}} \sum_{\mathbf{o}\in\mathscr{C}_{s,i}} c_k^{(rr)}\left(\mathbf{o}\right) \tag{E.3}$$

Kreska nad symbolami zbiorów oznacza ich moc. Wyniki rozpoznawalności podawano dla zbioru $\mathscr{J}$ zawierającego wszystkie fonemy lub tylko samogłoski. Rozpoznawalność izolowanych ramek dla systemu obliczano uśredniając rozpoznawalności uzyskane dla zbioru mówców $\mathscr{S}$:

$$c^{(rs)}_{k,\mathscr{J},\mathscr{S}} = \frac{1}{\bar{\mathscr{S}}} \sum_{s \in \mathscr{S}} c^{(rm)}_{k,s,\mathscr{J}} \tag{E.4}$$

Zbiór mówców $\mathscr{S}$ zawiera mówców ze zbioru uczącego bazy nagrań lub ze zbioru testowego.

W przypadku porównywania rozpoznawalności uzyskanej przed i po zastosowaniu algorytmów kompensacji istotne jest zbadanie, czy uzyskana dla danego mówcy poprawa bądź pogorszenie średniej rozpoznawalności jest statystycznie istotne. W tym celu, oprócz porównywania estymat wartości oczekiwanych rozpoznawalności $c^{(rm)}_{k,s,\mathscr{J}}$, konieczne jest również uwzględnienie ich wariancji. Wariancje te szacowano przy pomocy metody *bootstrap* [109]. Dla każdego zbioru $\mathscr{C}_{s,i}$ wyznaczano $N^{(boot)} = 500$ zbiorów $\mathscr{C}^{(boot)}_{s,i,b}$, przy czym każdy zbiór b zawierał elementy wybrane ze zbioru $\mathscr{C}_{s,i}$ poprzez losowanie ze zwracaniem. Liczba elementów w każdym zbiorze $\mathscr{C}^{(boot)}_{s,i,b}$ była taka sama, jak liczba elementów w wyjściowym zbiorze $\mathscr{C}_{s,i}$. Następnie obliczano $N^{(boot)}$ wartości estymatora:

$$c^{(rm,boot)}_{k,s,\mathscr{J},b} = \frac{1}{\bar{\mathscr{J}}} \sum_{i \in \mathscr{J}} \frac{1}{\bar{\mathscr{C}}^{(boot)}_{s,i,b}} \sum_{\mathbf{o} \in \mathscr{C}^{(boot)}_{s,i,b}} c^{(rr)}_{k}(\mathbf{o}) \tag{E.5}$$

Własnością metody *bootstrap* jest to, że dla dużych $N^{(boot)}$ rozkłady wartości $c^{(rm,boot)}_{k,s,\mathscr{J},b}$ odpowiadają rzeczywistym rozkładom szacowanej rozpoznawalności. Można je uznać za rozkłady normalne, co sprawdzono we wstępnych badaniach za pomocą odpowiedniego testu statystycznego [130]. Zatem rozkład rozpoznawalności izolowanych ramek dla danego mówcy można opisać oszacowanymi parametrami $\left\{c^{(rm)}_{k,s,\mathscr{J}}, \sigma^{(rm)2}_{k,s,\mathscr{J}}\right\}$, gdzie wariancja $\sigma^{(rm)2}_{k,s,\mathscr{J}}$ wyznaczana jest z $N^{(boot)}$ elementowej próby $c^{(rm,boot)}_{k,s,\mathscr{J},b}$ za pomocą estymatora wariancji dla rozkładu normalnego.

We wstępnych badaniach sprawdzono za pomocą testu na równość wariancji rozkładów normalnych [130], że dla danego mówcy wariancję przed kompensacją $\sigma^{(rm)2}_{k,s,\mathscr{J}}$ można uznać za równą wariancji po kompensacji $\sigma^{(rm,komp)2}_{k,s,\mathscr{J}}$. Ostatecznie zmianę oszacowanej rozpoznawalności po kompensacji $c^{(rm,komp)}_{k,s,\mathscr{J}}$ w stosunku do rozpoznawalności przed kompensacją $c^{(rm)}_{k,s,\mathscr{J}}$ uznawano za istotną wtedy, gdy stosując test na równe wartości oczekiwane rozkładów normalnych o równych wariancjach (test t) [130] na poziomie istotności 0.95 odrzucano hipotezę o równych wartościach $c^{(rm)}_{k,s,\mathscr{J}}$ i $c^{(rm,komp)}_{k,s,\mathscr{J}}$. W przypadku, gdy nie było podstaw do odrzucenia takiej hipotezy, przymowano $c^{(rm,komp)}_{k,s,\mathscr{J}} = c^{(rm)}_{k,s,\mathscr{J}}$.

F. Hybrydowy algorytm optymalizacji wartości parametrów transformacji widma

Algorytmy ewolucyjne [7] są skuteczne w przypadku maksymalizacji funkcji celu nieciągłych, obszarami stałych, o wielu lokalnych maksimach, a więc takich, dla których algorytmy gradientowe, czy nawet algorytmy poszukiwań prostych, są nieskuteczne. Algorytmy ewolucyjne mają możliwość silnej eksploracji przestrzeni optymalizacji oraz opuszczania lokalnych maksimów. Na ogół nie umożliwiają jednak precyzyjnego określenia punktu maksimum. Dlatego zaproponowano algorytm hybrydowy. Podobną metodę można znaleźć np. w [15].

W zaproponowanym algorytmie pierwszy etap optymalizacji przeprowadzany jest z wykorzystaniem następującego algorytmu ewolucyjnego:

1. Wylosuj populację początkową $N^{(pop)}$ osobników. Zainicjalizuj wartości charakteryzujących ich parametrów $\boldsymbol{\alpha}^{(st)}$, losując je wg rozkładów normalnych o parametrach $\left\{\mu_i^{(\alpha,ini)}, {\sigma_i^{(\alpha,ini)}}^2\right\}$, gdzie i oznacza numer parametru. Oblicz wartości funkcji celu dla wszystkich osobników z populacji.

2. Wytypuj $N^{(pow)}$ osobników o najmniejszej wartości funkcji celu i przeznacz je do późniejszego skasowania.

3. Wylosuj metodą ruletki $N^{(pow)}$ spośród $N^{(pop)}$ osobników do późniejszego rozmnożenia. Osobnik losowany jest z prawdopodobieństwem proporcjonalnym do wartości jego funkcji celu pomniejszonej o minimalną wartość funkcji celu osobników w populacji. Ten sam osobnik może być losowany wielokrotnie.

4. Skopiuj każdego osobnika przeznaczonego do rozmnożenia i zmodyfikuj wartości jego parametrów, dodając do nich zaburzenia losowane wg rozkładów normalnych o parametrach $\left\{0, {\sigma_{i,t}^{(\alpha,mod)}}^2\right\}$.

5. Zastąp osobniki przeznaczone do skasowania osobnikami zmodyfikowanymi i oblicz dla nich wartość funkcji celu.

6. Jeśli nie jest spełniony jest warunek stopu, wróć do punktu 2. W przeciwnym wypadku wybierz osobnika o najwyższej funkcji celu i podaj wartości jego parametrów jako wartości wyjściowe pierwszego etapu optymalizacji.

W implementacji liczby osobników wynosiły $N^{(pop)} = 16$ i $N^{(pow)} = 8$. Warunkiem zatrzymania było wykonanie 50 iteracji. Parametry rozkładów normalnych stosowanych podczas inicjalizacji przyjęto następujące:

- Dla parametrów skalowania osi częstotliwości wartości oczekiwane równały się wartościom neutralnym, przy których oś częstotliwości nie jest zmieniana, a odchylenia standardowe przyjęto równe 30, 60 i 120 odpowiednio dla parametrów $\alpha_1^{(g,o)}$, $\alpha_2^{(g,o)}$ i $\alpha_3^{(g,o)}$.
- Dla parametrów filtru liniowego wartości oczekiwane przyjęto równe zeru, a odchylenia standardowe równe 0.9.

Odchylenia standardowe rozkładów stosowanych do modyfikacji wartości parametrów osobników były zmniejszane w końcowych iteracjach, co miało na celu zwiększenie zdolności eksploatacji algorytmu. Zmian tych dokonywano wg zależności:

$$\sigma_{i,t}^{(\alpha,mod)} = \begin{cases} \sigma_{i,0}^{(\alpha,mod)}, & \text{dla } t = 0, \ldots, T^{(sr)} \\ \sigma_{i,0}^{(\alpha,mod)} \frac{T^{(sk)} - t + \beta^{(sr)}\left(t - T^{(sr)}\right)}{T^{(sk)} - T^{(sr)}}, & \text{dla } t = T^{(sr)} + 1, \ldots, T^{(sk)} - 1 \end{cases} \tag{F.1}$$

gdzie $T^{(sr)}$ oznacza iterację, od której zaczyna się zmniejszanie odchylenia standardowego, a $T^{(sk)}$ - całkowitą liczbę iteracji. W implementacji przyjęto $T^{(sr)} = 25$, $\beta^{(sr)} = 0.25$, a początkowe odchylenia standardowe $\sigma_{i,0}^{(\alpha,mod)}$:

- Dla parametrów skalowania osi częstotliwości 10, 20 i 40 odpowiednio dla parametrów $\alpha_1^{(g,o)}$, $\alpha_2^{(g,o)}$ i $\alpha_3^{(g,o)}$.
- Dla parametrów filtru liniowego 0.3.

W drugim etapie zastosowano algorytm simpleksu Neldera-Meada [37], będący algorytmem poszukiwań prostych.

G. Algorytm estymacji wartości częstotliwości tonu krtaniowego

Algorytm estymacji wartości częstotliwości tonu krtaniowego $f^{(v)}$ powinien mieć małą złożoność obliczeniową i korzystać z wyników pośrednich, uzyskanych podczas parametryzacji sygnału. Nie jest wymagana bardzo wysoka (< 1 Hz) precyzja estymacji oraz nie jest potrzebne wyznaczanie trajektorii $f^{(v)}$ w wypowiedzi (a więc dokładna estymacja $f^{(v)}$ w każdej ramce), a jedynie obliczenie jednej wartości dla całej wypowiedzi. Znane z literatury [29, 147, 166] algorytmy okazały się zbyt kosztowne obliczeniowo lub zbyt mało precyzyjne. Uwzględniając podane założenia, zaproponowano algorytm wykorzystujący obliczone podczas parametryzacji MFCC widmo amplitudowe sygnału. Poniżej podano schemat algorytmu:

1. Dla każdej ramki t danej wypowiedzi pobierz widmo amplitudowe $\mathbf{s}_t$, obliczone w trakcie parametryzacji MFCC.

   (a) Wyznacz minima i maksima w widmie $\mathbf{s}_t$, analizując jego prążki w zakresie $m = 3, \ldots, 94$, co odpowiada zakresowi częstotliwości od 93.75 Hz do 2937.5 Hz. Niech wektor $\mathbf{n}^{(f,min)}$ zawiera numery prążków, dla których wykryto minima, a $\mathbf{n}^{(f,max)}$ - maksima.

   (b) Dla każdej pary sąsiadujących ze sobą maksimów $n_i^{(f,max)}$ i $n_{i+1}^{(f,max)}$ wyznacz różnicę:

   $$n_i^{(df)} = n_{i+1}^{(f,max)} - n_i^{(f,max)} \tag{G.1}$$

   oraz wagę:

   $$n_i^{(dfw)} = \min\left(s_{n_i^{(f,max)},t}, s_{n_{i+1}^{(f,max)},t}\right) - s_{n_i^{(f,min)},t} \tag{G.2}$$

   gdzie $n_i^{(f,min)}$ zawiera numer prążka odpowiadającego minimum położonemu pomiędzy maksimami o numerach prążków $n_i^{(f,max)}$ i $n_{i+1}^{(f,max)}$.

(c) Wyznacz wartości:

$$f_t^{(v,t)} = \frac{\sum_{i=0}^{I-1} n_i^{(df)} \cdot n_i^{(dfw)}}{\sum_{i=0}^{I-1} n_i^{(dfw)}} \cdot \frac{f^{(p)}}{M} \tag{G.3}$$

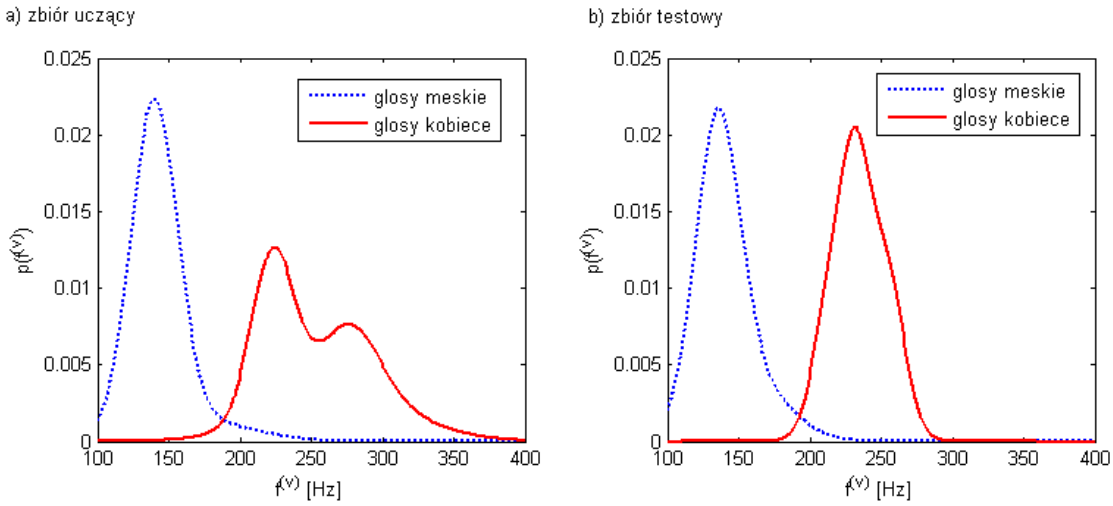
$$f_t^{(vw,t)} = \frac{1}{I} \sum_{i=0}^{I-1} n_i^{(dfw)} \tag{G.4}$$

gdzie I oznacza liczbę przeanalizowanych i zapamiętanych par maksimów, $f^{(p)}$ - częstotliwość próbkowania, a M - długość stosowanej do wyznaczenia widma transformaty DFT.

2. Znajdź maksymalną wartość $f_t^{(vw,t)}$ i oznacz jako $f^{(vw,max)}$. Wyznacz końcową wartość $f^{(v)}$ dla wypowiedzi jako średnią z tych $f_t^{(v,t)}$, dla których odpowiadające im $f_t^{(vw,t)} > f^{(vw,max)} \cdot T^{(vw)}$, gdzie próg $T^{(vw)}$ przyjęto równy 0.7.

Badania wykazały, że dokładność estymacji $f^{(v)}$ przez opisany algorytmu jest satysfakcjonująca. Nie dokonuje się w nim oddzielnie detekcji ramek dźwięcznych/bezdźwięcznych, gdyż rolę tą spełnia analiza przebiegu wag $f_t^{(vw,t)}$. Wagi te przyjmują wartości wysokie dla ramek o uporządkowanym widmie, zawierającym duża liczbę wyraźnych maksimów, a zatem widmie charakterystycznym dla fragmentów dźwięcznych.

W celu określenia wartości progu decyzyjnego $f^{(v,p)} = 190$ Hz dla algorytmu wyboru klas, aproksymowano (za pomocą sumy trzech krzywych Gaussa) rozkłady



Rys. G.1. Rozkłady prawdopodobieństwa $f^{(v)}$.

prawdopodobieństwa $f^{(v)}$ dla głosów męskich i kobiecych (rys. G.1), przy czym wykorzystano wszystkie wypowiedzi dostępne w bazie nagrań. Wartość progu wyznaczono korzystając z danych ze zbioru uczącego w miejscu przecięcia się wykresów aproksymowanych rozkładów, co odpowiada Bayesowskiemu kryterium decyzyjnemu przy założeniu, że obie klasy są równoprawdopodobne [39]. Widoczny na rys. G.1a dwumodalny rozkład prawdopodobieństwa $f^{(v)}$ dla głosów kobiecych w zbiorze uczącym wynika z obecności w tym zbiorze również głosu dziecięcego, charakteryzującego się wysoką wartością $f^{(v)}$. Niewielka liczność tego zbioru przyczyniła się do wyraźnego uformowania drugiego maksimum w rozkładzie prawdopodobieństwa wartości $f^{(v)}$.

W tabeli G.1 podano wyniki klasyfikacji mówców na podstawie wartości $f^{(v)}$.

Tab. G.1. Wyniki klasyfikacji mówców na podstawie wartości $f^{(v)}$.

błąd klasyfikacji [%]	głosy męskie	głosy kobiece	średnio
zbiór uczący	4.40	2.55	4.02
zbiór testowy	3.13	0.73	2.69