POLITECHNIKA LUBELSKA

Paweł Powroźnik

Przetwarzanie sygnałów dźwiękowych mowy polskiej celem rozpoznania stanu emocjonalnego mówcy

MONOGRAFIE

Lublin 2021

Przetwarzanie sygnałów dźwiękowych mowy polskiej celem rozpoznania stanu emocjonalnego mówcy

Monografie – Politechnika Lubelska

Politechnika Lubelska
Wydział Elektrotechniki i Informatyki
ul. Nadbystrzycka 38A
20-618 Lublin

Paweł Powroźnik

Przetwarzanie sygnałów dźwiękowych mowy polskiej celem rozpoznania stanu emocjonalnego mówcy

Wydawnictwo
Politechniki Lubelskiej

Lublin 2021

Recenzenci:
dr hab. inż. Volodymyr Mosorov, prof. Politechniki Łódzkiej
prof. dr hab. Tadeusz Wieczorek, Politechnika Śląska

Publikacja wydana za zgodą Rektora Politechniki Lubelskiej

ISBN: 978-83-7947-493-6

Wydawca: Wydawnictwo Politechniki Lubelskiej
www.biblioteka.pollub.pl/wydawnictwa
ul. Nadbystrzycka 36C, 20-618 Lublin
tel. (81) 538-46-59

Spis treści

1. Wstęp

Komunikacja międzyludzka stanowi podstawę funkcjonowania człowieka w społeczeństwie. Sygnał mowy niesie słuchaczowi nie tylko informacje stricte lingwistyczne, ale również dane związane ze środowiskiem wewnętrznym człowieka, jego nastawieniem do słuchacza czy przeżywanymi emocjami. Poprawna identyfikacja i zrozumienie obu informacji (lingwistycznej i psychologicznej) zdecydowanie poprawia jakość komunikacji.

Rozpoznawanie stanu emocjonalnego mówcy w oparciu o mowę naturalną jest zagadnieniem niezwykle złożonym i skomplikowanym. Jednak przy aktualnym dynamizmie zmian i nowatorskich rozwiązań w funkcjonowaniu interfejsów człowiek – komputer jest to zjawisko coraz bardziej pożądane. Taki stan rzeczy wynika również z rozwoju technologii i narzędzi związanych bezpośrednio z samym przetwarzaniem sygnału mowy, toteż można zaobserwować nowy trend badań skupiających się właśnie na wydobyciu z sygnału dodatkowych, poza lingwistycznych informacji.

Pojawia się coraz więcej publikacji związanych z werbalnymi i niewerbalnymi procesami. Istniejące, uniwersalne systemy rozpoznawania stanów emocjonalnych skupiają się wokół przetwarzania danych dostarczanych poprzez obraz – prace oparte o mimikę twarzy [63], gesty [138] czy stan fizjologiczny organizmu [39]. Jednak to głos jest najłatwiej dostępnym sygnałem. Jest on również nośnikiem informacji wyrażającym stan emocjonalny mówcy.

Badania zaprezentowane w niniejszej pracy skupiają się wokół zagadnień związanych z przetwarzaniem sygnału mowy w celu wydobycia informacji o emocjach, czyli zestawu cech charakterystycznych dla każdego z rodzajów emocji [89] oraz poprawnej jej klasyfikacji. Przeprowadzono eksperymenty zarówno z wykorzystaniem nagrań emocji, odegranych (przez aktorów oraz osoby niezwiązane z tą profesją) jak i nagrań przetworzonych w czasie zbliżonym do rzeczywistego podczas eksperymentów przeprowadzonych w firmie z działem Call-Center. Na potrzeby badań zostały wykorzystane trzy bazy nagrań stanowiące podstawę do prac.

Autor przedstawił nowatorski sposób przetwarzania sygnału mowy niespotykany w publikacjach związanych z identyfikacją emocji. W pracy wyszczególnione zostały zarówno zagadnienia związane z wykorzystanymi parametrami sygnału, jego właściwościami oraz metodami klasyfikacji.

Podstawę do badań stanowiły nagrania zawierające emocje odegrane. Pozwoliło to na opracowanie metod przetwarzania sygnału mowy, które następnie mogły zostać zweryfikowane w środowisku rzeczywistym – Call-Center.

2. Przegląd sposobów przetwarzania sygnału mowy

Emocje to stany psychologiczne oparte na biologii wywołane przez zmiany neurofizjologiczne, różnie związane z myślami, uczuciami, reakcjami behawioralnymi oraz stopniem przyjemności lub niezadowolenia. Obecnie nie ma naukowego konsensusu co do definicji. Emocje często przeplatają się z nastrojem, temperamentem, osobowością, lub usposobieniem.

Już w starożytności filozofowie podejmowali próby zrozumienia, czym są emocje i jaką rolę odgrywają. Jednak to ostanie trzydziestolecie przyniosło burzliwy rozwój badań nad strukturą emocji i ich znaczeniem. Pomimo wzmożonych wysiłków, nie udało się jednoznacznie ustalić, w jaki sposób należy je właściwie interpretować oraz jakie prawa nimi rządzą. Same stany emocjonalną bywają, w zależności od sytuacji, różnie określane: namiętności, uczucia, afekty, wzruszenia, itd. Wzajemne przenikanie stanów emocjonalnych oraz ich złożoność są najczęstszym powodem nieprecyzyjności ich określeń. Podążając za definicją zaproponowaną przez Janusza Reykowskiego [129] proces emocjonalny może być rozumiany jako specyficzna reakcja organizmu, spowodowana gwałtownymi zmianami środowiska wewnętrznego i zewnętrznego. W skład owego procesu wchodzą trzy podstawowe komponenty. Pierwszym z nich jest pobudzenie emocjonalne prowadzące do zmian mobilizacyjnych zachodzących w organizmie. Drugi, ogranicza się do świadomości znaczenia powyższych zmian dla podmiotu. Ostatni komponent związany jest z jakościowymi cechami, mającymi znacznie dla człowieka. Zaburzenia emocjonalne stanowią przykład pierwszego komponentu. Objawiają się one zmianami czynności wegetatywnych i psychicznych, zarówno pod względem intensywności, jak i tempa przebiegu reakcji. Drugi komponent odpowiedzialny jest za odbiór zdarzenia emocjonalnego. Pozytywny proces emocjonalny pobudza czynności odpowiedzialne za podtrzymanie kontaktu z występującą sytuacją, negatywny zaś ukierunkowany jest na przerwanie kontaktu z występującym zdarzeniem. Trzeci z komponentów jest przez Reykowskiego opisywany jako charakteryzujący treść emocji.

Jeden z pionierów naukowej psychologii, według Carolla Izarda i Janusza Reykowskiego, Wilhelm Wundt [53, 128] uważał, że pod pojęciem procesów emocjonalnych ukryty jest rodzaj zjawisk psychologicznych, których mnogość nie pozwala na jednostkowe ich określenie. W oparciu o prace Waltera Cannona oraz Richarda Davidsona można zauważyć, iż Wundt pojmował emocje w trzech wymiarach: napięcie–odprężenie, podniecie–ukojenie oraz przyjemność–przykrość [14, 26].

Z kolei uchodzący za przeciwnika Wundta, amerykański psycholog Edward Bradford Titchener wyróżniał dwa rodzaje stanów uczuciowych:

przykrość oraz przyjemność. Uznawał on istnienie różnych zjawisk emocjonalnych takich jak: uczucia złożone (religijne, moralne, intelektualne), afekty (np. nienawiść, radość) czy nastroje (np. zadowolenie). Bazując na pracy Janusza Reykowskiego, zaproponowany przez Titchenera podział stał się początkiem burzliwych polemik wśród psychologów, którzy kwestionowali różnice pomiędzy afektami i uczuciami prostymi oraz złożonymi [129].

Bazując na dostępnych publikacjach z dziedziny psychologii i psychiatrii [1, 83, 86] można stwierdzić, iż obecnie trudno jest odróżnić i sklasyfikować stany emocjonalne charakteryzujące zjawiska kryjące się pod takimi określeniami jak: wzruszenie, afekt, nastrój czy uczucie. Mała liczba badań empirycznych związanych z tematyką emocji nie pozwoliła dotychczas na jednoznaczne opracowanie kryteriów, różnicujących emocje i stany afektywne czy nastroje [83].

Bazując na pracach Paula Ekmana [32, 33, 34] emocje cechuje wzór ekspresji mimicznej, której brak w przypadku nastrojów. Z kolei Richard Davidson głosi, iż nastrój związany jest z ukierunkowaniem procesów poznawczych, emocja zaś z przebiegiem zachowań [26].

W języku potocznym bardzo często słowa takie jak: „emocje", „uczucia" oraz „nastroje" traktowane są jako synonimy. Jakkolwiek nie są to wyrażenia sprzeczne, jednakże mogą definiować przeciwległe bieguny continuum. Wychodząc od prostych emocji poprzez te bardziej złożone możemy dojść do najbardziej wzniosłych uczuć jak miłość, przyjaźń czy patriotyzm [83]. W publikacji [128] możemy spotkać się z stwierdzeniem, iż emocja związana jest z czynnością ośrodków podkorowych mózgu, przede wszystkim podwzgórza i bezpośrednio sąsiadujących struktur międzymózgowia. Uczucia zaś, postrzegane są jako procesy zachodzące w korze mózgowej.

W oparciu o tradycję językową można stwierdzić, iż termin „uczucia" został zarezerwowany emocjom wyższym takim jak uczucia patriotyczne, uczucia miłości czy nienawiści. Termin „emocja" z kolei, odnosi się do ogółu stanów motywacyjno-fizjologicznych.

Słownik Encyklopedii Psychiatrii [81] traktuje emocje i uczucia jako synonimy opisujące „stosunek podmiotu do ludzi, zjawisk, rzeczy, siebie, swego organizmu i własnego działania". Jednak najczęściej przyjmuje się, iż są to procesy psychiczne wyływające na ustosunkowanie się osób, zjawisk oraz przedmiotów, będące podstawą świadomego i nieświadomego działania [87].

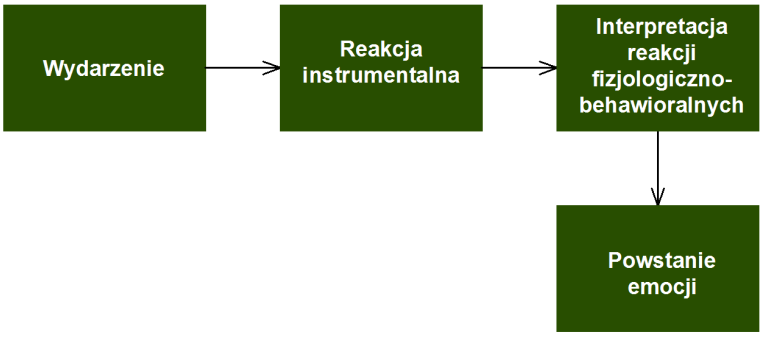
Termin „emocja" pochodzi od łacińskiego czasownika *movere* oznaczającego „poruszyć". Sugeruje zatem pewną skłonność do działania, występującą w każdej z emocji. Powszechnie uważa się, że emocje stanowią element motywacji do działania. Ponadto emocje pojawiają się nagle i zawsze są powiązane z pobudzeniem somatycznym i dużą intensywnością.

Bazując na pracy Anny Grzywa, można wyróżnić dwa znaczenia emocji: utrwalone dyspozycje oraz chwilowe przeżycia. Z kolei na poziomie fenomenologicznym jest to stymulator każdego działania [42]. Znane jest wiele teorii i opracowań dotyczących charakteru emocji, sposobu ich powstawania czy wpływu stanów emocjonalnych na zachowania człowieka. Można tutaj przywołać takich autorów jak Carroll Izard [53], John Broadus Watson [156], Walter Cannon [14], James Papez [109] czy Jana Mazurkiewicza [94, 95].

Jednak największy wpływ na kategoryzację emocji wywarły dwie teorie. Pierwsza została opracowana przez Williama Jamesa i Carla Langea, którzy uważali, iż to obserwacja powoduje zmiany somatyczne, które to powadzą do zachowań behawioralnych. Drugą z nich opracowana przez amerykańskiego psychologa Roberta Plutchika, zakłada podział emocji na podstawowe i złożone. Opierając się na artykule Tima Dalgleish'a można zauważyć, iż Plutchik zaproponował pewną hierarchizację stanów emocjonalnych [25].

2.1. Teoria emocji Jamesa-Langego

Klasyczne teorie emocji mają swój początek w wydanej w 1890 roku monumentalnej pracy Williama Jamesa „Principles of Psychology" i kształtowały się na przestrzeni kolejnych siedemdziesięciu lat. Z kolei koniec klasycyzmu wyznaczony jest książką Elizabeth Duffy „Activation and behavior" opublikowanej w 1962 roku [31]. W swej pracy Rufolf Cardinal stwierdza, iż w rozwoju psychologii emocji rok ten jest wyjątkowy również pod innym względem. Zostały wówczas opublikowane wyniki eksperymentów Stanleya Schachtera i Jerome Singera dając początek atrybucyjnym koncepcjom emocji [15]. Teoria Jamesa-Langego jest zaliczana do grupy obwodowych teorii emocji. Za źródło powstawania stanu emocjonalnego uznaje się dzia-



Rysunek 2.1. Graficzna interpretacja teorii Jamesa-Langego. Opracowanie własne na podstawie [140]

łania, zmiany w narządach wewnętrznych oraz pracy mięśni. Schemat powstawania emocji w oparciu o teorię Jamesa-Langego został przedstawiony na Rysunku 2.1 (patrz s. 10).

Według Williama Jamesa spostrzeżenie danego bodźca wywołuje zmiany fizjologiczne, które w kolejnym etapie są postrzegane przez jednostkę. Obserwacja owych zmian jest z kolei przekształcane w emocję. Opierając się na pracy Jesse Prinza można przyjąć, iż James był przekonany, iż jako pierwsze pojawiają się zmiany fizjologiczne, a w ich następstwie następuje spostrzeżenie emocji [124]. Początkowo James uważał, że zarówno zmiany zachodzące w mięśniach znajdujących się w narządach wewnętrznych, jak i mięśniach poprzecznie prążkowanych odpowiedzialnych za zmiany w mimice twarzy czy powodujące zmiany w mięśniach gładkich prowadzą do powstania emocji. Jednak pod wpływem Langego, swoją uwagę skoncentrował wyłącznie na zmianach mających swoje źródło w autonomicznym układzie nerwowym [140].

James aprobował ideę specyficznego pobudzenia emocjonalnego mówiącą, iż dostrzeżenie zmian fizjologicznych powoduje powstanie różnych uczuć, zatem same zmiany dla każdego rodzaju uczuć muszą być różne. Jednakże w owym czasie nie dysponował badaniami potwierdzającymi powyższą tezę. Dopiero na przełomie lat osiemdziesiątych i dziewięćdziesiątych dwudziestego wieku udało się uzyskać wyniki, które w bezpośredni sposób potwierdzały hipotezę Jamesa. Według Izarda, dla samych autorów teorii to uczucie miało pierwszoplanowe znaczenie, a nie emocje [54]: „[...] uczucie było nie tylko jądrem emocji, lecz także jądrem psychologii”[1]. Stanowiło to istotę myśli i świadomości, dostarczającej poczucia tożsamości osobistej [55].

Według Jamesa–Langego poza podstawowymi emocjami takimi jak strach, radość, smutek czy złość istnieją również emocje bardziej subtelne jak uczucia estetyczne, moralne czy intelektualne. Emocje te są wynikiem odbioru wrażeń wzrokowych i słuchowych. Odczucia pochodzące z wnętrza organizmu, według powyższej teorii, mają znaczenie drugoplanowe [55]. Badacze byli świadomi, iż fizjologiczne pobudzenie spowodowane samą obserwacją obrazów czy słuchaniem muzyki, w oparciu, o które powyższa teoria powstała [140], jest znacznie słabsze, niż w sytuacji realnego zagrożenia [53].

Sformułowana przez Jamesa koncepcja uczuć wyższych jest w pewnym sensie paradoksalna. Pojawienie się emocji wyższych jest następstwem znacznie prostszego mechanizmu sensorycznego, aniżeli w przypadku uczuć podstawowych [140]. Można powiedzieć, że bardziej złożone emocje, po-

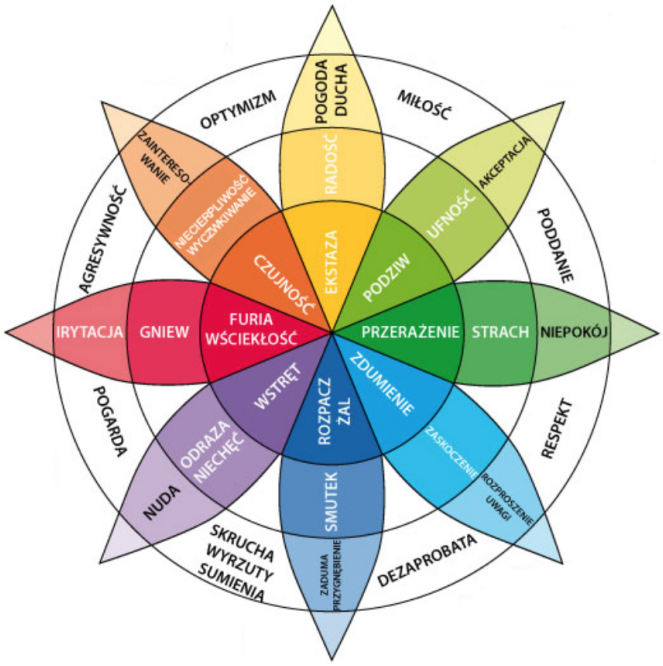
[1] Występujący w języku angielskim związek frazeologiczny feeling of emotion bezpośrednio może być tłumaczony jako odczucie emocji. Jednakże w języku angielskim samo słowo feeling jest raczej używane jako „uczucie”, aniżeli „odczucie”.

wodują szybsze przejście ze stanu rejestracji treści zdarzenia do stanu powstania uczucia. Z kolei w przypadku emocji podstawowych, w pierwszej kolejności pobudzane są eksteroceptory, potem zaś intero i proprioreceptory. Obserwacja pobudzenia interoreceptorów stanowi uczucie podstawowe, eksteroreceptorów – uczucie wyższe [55]. Dla Jamesa emocje są podstawą tworzenia osobowości: „[...] indywidualizm nałożony jest na uczucie, zakamarki uczucia, tajemne, ciemne głębie charakteru – oto jedyne miejsce, gdzie możemy złapać na gorącym uczynku dziejący się fakt: tylko tam możemy spostrzec bezpośrednio, jak odbywają się zjawiska". Emocje są dla Jamesa najważniejszą cechą osobowości i dynamizują zachowanie człowieka [56].

2.2. Psychoewolucyjna teoria Plutchika

Amerykański psycholog R. Plutchik [115] w latach 1960–1980 opracował teorię, na podstawie której wyodrębnił osiem podstawowych emocji stanowiących prototypy wszystkich pozostałych. Do grupy tej zostały zaliczone następujące stany emocjonalne: radość, strach, zdziwienie, akceptacja, smutek, gniew, oczekiwanie oraz obrzydzenie. Prototypy te są wrodzone i mają swoje uzasadnienie w adaptacyjnych zachowaniach mających na celu pomoc w przetrwaniu [68]. Stanowią one swego rodzaju stany idealne, charakteryzujące się tym, że wnioski na ich temat są wyciągane intuicyjne. Wszystkie inne emocje, nie będące stanami podstawowymi występują jako mieszaniny, kombinacje lub związki stanów podstawowych. W efekcie teorii Plutchik opracował swoisty diagram, znany w literaturze jako koło emocji Plutchika. Zostało ono przedstawione na Rysunku 2.2. [115].

Diagram na Rysunku 2.2. (patrz. s. 13) prezentuje stopień podobieństwa pomiędzy podstawowymi emocjami oraz stanami z nich się wywodzącymi. Należy zauważyć, iż emocje podstawowe są wzajemnie wykluczające się i nie mogą być doświadczone jednocześnie. Dlatego też zostały rozmieszczone jako pary biegunów przeciwstawnych. Przykładowo gniew stanowi przeciwieństwo strachu, smutek – radości, niecierpliwość, wyczekiwanie – zaskoczenie. [80]. Nowe stany emocjonalne są tworzone w sytuacji, gdy występujące obok siebie emocje są odczuwane jednocześnie. Połączenie dwóch pierwotnych emocji jest przez Plutchika określane diadą [116]. Na przykład pogarda jest diadą powstałą na skutek odczuwania gniewu oraz odrazy. Diady zbudowane w oparciu o emocje przeciwstawne nie mogą istnieć, z powodu konfliktu zachodzącego między nimi. Badania empiryczne stanowią poparcie powyższego modelu [115].



Rysunek 2.2. Koło emocji Plutchika [115]

Plutchik definiuje główne założenia swojej teorii w oparciu o dziesięć postulatów [115]:

1. Pojęcie emocji odnosi się zarówno do ludzi jak i do zwierząt i ma zastosowanie bez względu na poziom ewolucji.
2. Emocje u rożnych gatunków uzewnętrzniają się w odmienny sposób.
3. Pozwalają one przeciwstawić się zagrożeniom stwarzanym przez nieprzyjazne środowisko.
4. Możliwe jest zidentyfikowanie wspólnych elementów i wzorców u wszystkich gatunków.
5. Liczba emocji podstawowych jest stała.
6. Wszystkie emocje wtórne powstają z kombinacji emocji podstawowych.
7. Pierwotne emocje są stanami idealnymi.
8. Możliwe jest scharakteryzowanie emocji podstawowych za pomocą par biegunowych przeciwstawieństw.
9. Emocje różnią się ze względu na stopień podobieństwa do siebie samych.
10. Każda emocja może występować na różnym poziomie pobudzenia i z różnym natężeniem.

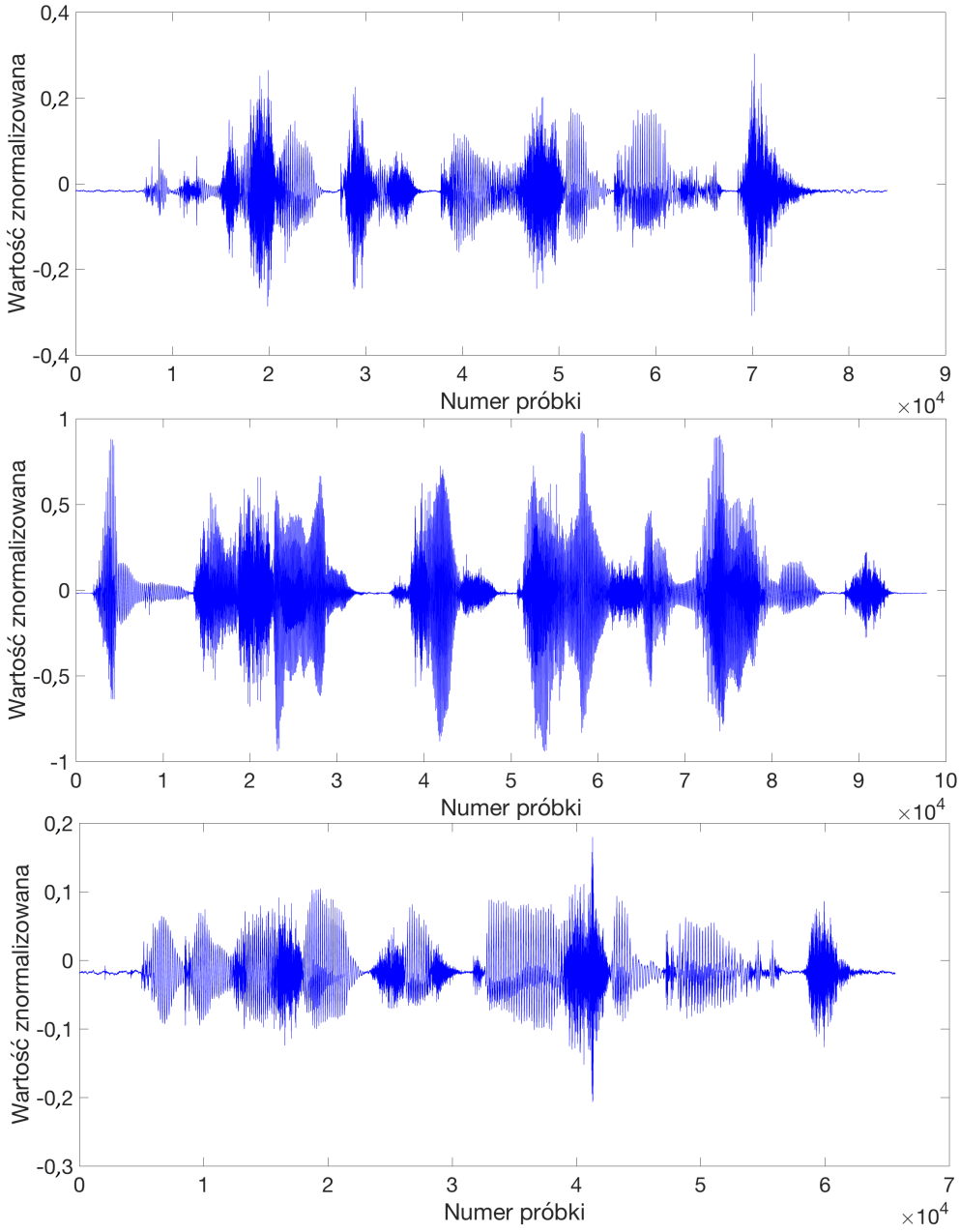
Plutchik twierdził, iż różne zachowania są różnie przez mózg interpretowane i różne obszary odpowiedzialne za emocje są wówczas aktywowane, co z kolei prowadzi do konkretnych zachowań adekwatnych do pobudzającego bodźca. Taki sposób zamiany impulsów na konkretne reakcje jest wielką korzyścią ewolucyjną [116]. Zatem w oparciu o teorię głoszone przez Jamesa-Langego oraz Plutchika możemy sklasyfikować następujące emocje jako podstawowe (pierwotne) [53, 55, 114, 115]: radość, smutek, strach, złość.

Analizując proponowany przez Plutchika model emocji przedstawiony poprzez koło emocji pokazane na Rysunku 2.2. można zauważyć, iż wszystkie emocje pierwotne zbiegają się w jednym punkcie, stanowiącym środek koła. Trudno zaakceptować sytuację, w której człowiek będzie „miotał się" po niezwykle wzbudzanych stanach emocjonalnych, przechodząc ze skrajności w skrajność. Dlatego też postuluje się [82] o dodanie łączącego je stanu neutralnego. Pozwoli nam to na uzyskanie naturalnego wzrostu intensywności emocji wraz z oddaleniem od środka koła.

Od wieków ludzie próbują poprawnie rozpoznać, stan emocjonalny rozmówcy. Staramy się, aby byli oni zainteresowani naszą wypowiedzią. Ma to szczególne znaczenie dla osób pracujących głosem, nauczycieli, wykładowców, telemarketerów, itd. Dlatego też istotnym stanem emocjonalnym, na detekcji którego skupiają się naukowcy, jest znudzenie. Stan ten przez wielu filozofów i psychologów traktowany jest na równi z innymi emocjami pierwotnymi. Pojawia się w pracach zarówno Jamesa-Langego, Plutchika, Cannona-Barda i innych [14, 55, 116, 140]. W wielu pracach traktujących o metodach identyfikacji stanów emocjonalnych jako podstawowe wymieniane są następujące emocje [16, 17, 44, 68, 103, 134, 152]: radość, smutek, strach, złość, znudzenie, stan neutralny.

2.3. Charakterystyka i sposób powstawania sygnału mowy

Mowa jest podstawowym sposobem komunikacji międzyludzkiej. Dodatkowo poza znaczeniem literalnym za jej pomocą przekazywane są emocje, przeżycia artystyczne czy nastroje. Jej wyrazem jest dźwięk, rozumiany w kategoriach fizycznych, jako zmiana położenia cząsteczek ośrodka sprężystego względem ośrodka równowagi [45]. Środowiskiem umożliwiającym powstawanie drgań może być dowolny ośrodek gazowy, stały lub ciekły. Receptorem dźwięku mowy jest ludzie ucho, które reaguje na miejscowe zmiany ciśnienia, powstałego wskutek lokalnego zagęszczania i rozrzedzania powietrza, względem ciśnienia atmosferycznego [45]. W organizmie człowieka generatorem drgań owych cząsteczek jest krtań. Narząd ten zbudowany jest z mięśni, wiązadeł i szkieletu chrzęstnego.



Rysunek 2.3. Oscylogramy dla trzech stanów emocjonalnych: strachu (u góry), radość (pośrodku), stanu neutralnego (na dole)

Wnętrze krtani pokrywa błona śluzowa, w której to umiejscowione są odpowiadające za powstawanie dźwięku dzięki swej zdolności do drgań, fałdy głosowe [117]. Na Rysunku 2.3. został przedstawiony wykres wartości sygnału mowy dla trzech stanów emocjonalnych: strachu (od góry), radości (w środku), stanu neutralnego (od dołu) dla tej samej wypowiedzi w sensie semantycznym. Jak łatwo zauważyć pojawienie się emocji nie tylko zmieniło zakres znormalizowanych wartości próbek występujących w sygnale ale również kształt oscylogramu. Sygnał mowy podobnie jak każdy sygnał definiowany jest jako funkcja czasu *x(t)* przenosząca informację o stanie układu. Ze względu na ową zależność sygnału od czasu, w procesie przetwarzania sygnału mowy wyszczególnianie są parametry czasowe. Wśród nich do najczęściej wyszczególnianych należą parametry opisujące sygnał. Zaliczamy tutaj takie parametry jak: **średnia wartość sygnału** w przedziale czasu określona następującą równością [90]:

$$\bar{x} = \frac{1}{n_2 - n_1} \sum_{n=n_1}^{n_2} x(n), \tag{2.1}$$

gdzie:
n_1 – oznacza numer pierwszej próbki z rozpatrywanego przedziału,
n_2 – numer ostatniej próbki z rozpatrywanego przedziału,
$x(n)$ – oznacza wartość $n - tej$ próbki.

Kolejnym parametrem ekstrahowanym z sygnału mowy jest **mediana wartości sygnału**. Nazywana również przeciętną, wartością środkową czy drugim kwadrylem. Jest uporządkowanym ciągiem wartości sygnału, którego każdy kolejny element musi być większy lub równy poprzedniej wartości. W przypadku gdy liczba rozpatrywanych wartości jest nieparzysta mianem mediany określany jest środkowy wyraz ciągu, w przeciwnym wypadku średnia arytmetyczna dwóch środkowych wyrazów [106].

Wśród parametrów czasowych wyszczególniane jest również **odchylenie standardowe**, mówiące jak bardzo wartość rozpatrywanej próbki różni się od wartości średniej. Parametr ten definiowany jest w następująco [90]:

$$SD = \sqrt{\frac{\sum_{n=0}^{N-1}(x(n) - \bar{x})^2}{N}}, \tag{2.2}$$

gdzie:
$x(n)$ – oznacza wartość $n - tej$ próbki,
N – oznacza liczbę wszystkich próbek,
$\bar{x}$ – wartość średnia.

Równie często stosowanym parametrem czasowym jest **energia sygnału mowy**, definiowana następującym wzorem [90]:

$$E_x = \sum_{n=0}^{N-1} x^2(n), \tag{2.3}$$

gdzie:
$x^2(n)$ – oznacza kwadrat wartość $n-tej$ próbki,
N – oznacza liczbę wszystkich próbek.

Wśród statystycznych parametrów sygnału mowy należy również zaznaczyć istotność **kurtozy**, będącej miarą koncentracji wyników. Wartość ta informuje jak bardzo wyniki obserwacji skoncentrowane są wokół wartości średniej. Innymi słowy kurtoza niesie informację mówiącą jak duży jest „rozrzut" otrzymanych wyników. Parametr ten jest określony następującą równością [90]:

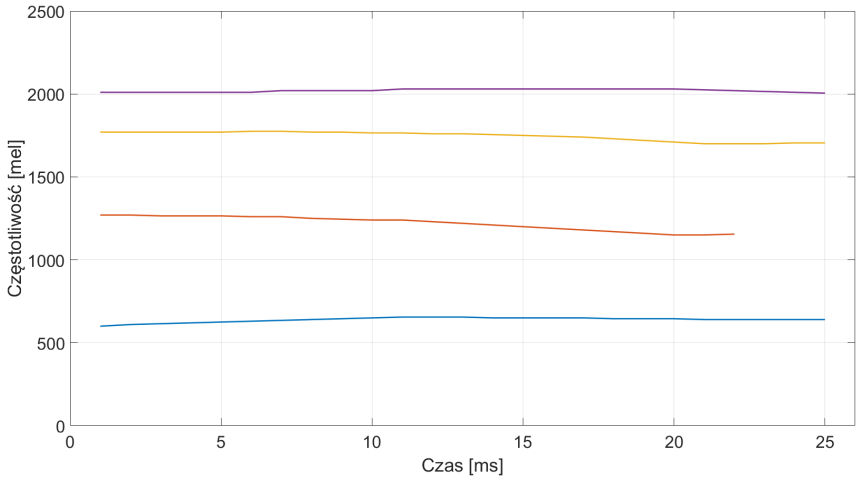
$$K = \frac{n(n-1)}{(n-2)(n-3)(n-4)} \sum_{n=0}^{N-1} \frac{((x(n)-\bar{x})^4}{SD} - \frac{3(n-2)^2}{(n-3)(n-4)}, \tag{2.4}$$

gdzie:
$x(n)$ – oznacza wartość $n-tej$ próbki,
N – oznacza liczbę wszystkich próbek,
$\bar{x}$ – oznacza wartość średnią,
SD – oznacza odchylenie standardowe.

Wśród parametrów związanych z analizą sygnału mowy znaczą rolę odgrywa **ton krtaniowy F0**. Jest to sygnał wytworzony przez drgające struny głosowe i stanowiący pobudzenie dla wszystkich głosek zwartych. Jego częstotliwość odpowiada częstotliwości podstawowej mowy. Zakres częstotliwości F0 jest zależy od płci oraz wieku mówcy. Typowe wartości dla mężczyzn wahają się w przedziale 85–155 Hz, dla kobiet 165–255 Hz, dla dzieci 250–300 Hz. W widmie sygnału obecne są wszystkie harmoniczne częstotliwości podstawowej opadające ok. 6–12 dB/okt [45].

Podczas przejścia przez kanał głosowy człowieka generowana przez struny głosowe fala ulega pewnym modyfikacjom skutkiem czego jest pojawienie się w widmie tonu podstawowego pewnych zniekształceń - lokalnych maksimów, zwanych **formantami**. Częstotliwości zaś związane z nimi określane są częstotliwościami formantowymi [68, 161] oznaczanymi symbolami F1, F2, F3 ... itd. Na Rysunku 2.4., zostały przedstawione przebiegi czasowe podczas wypowiadania samogłoski „a" dla czterech pierwszych forman-

tów. Częstotliwość została przedstawiona w skali mel, szczegółowo opisanej w podrozdziale 2.3.2.



Rysunek 2.4. Przykład przebiegów czasowych dla formantów F1–F4. Opracowanie własne na podstawie [161]

Istotną rolę w zganieniach identyfikacji stanów emocjonalnych w oparciu o sygnał mowy odgrywają parametry opisujące zależności czasowe sygnału mowy, takie jak [92]

- tempo mówienia – informujące o liczbie słów (bądź sylab) wypowiadanych w jednostce czasu,
- liczba fragmentów dźwięcznych w jednostce czasu,
- stosunek mowy dźwięcznej do bezdźwięcznej rozumianej jako stosunek fragmentów wypowiedzi zawierających fragmenty dźwięczne (samogłoski, spółgłoski dźwięczne) oraz fragmenty bezdźwięczne (spółgłoski bezdźwięczne).

Kolejne istotne grupy parametrów stanowiących standardy w zagadnieniach związanych z analizą i przetwarzaniem sygnału mowy stanowią parametry LPC (Liniowe kodowanie predykcyjne – ang. Linear Predictive Coding) oraz MFCC (Parametry mel-cepstralne, ang. Mel-Frequency Cepstral Coefficients) szerzej opisane w poniższych podrozdziałach.

Liniowe kodowanie predykcyjne (LPC)

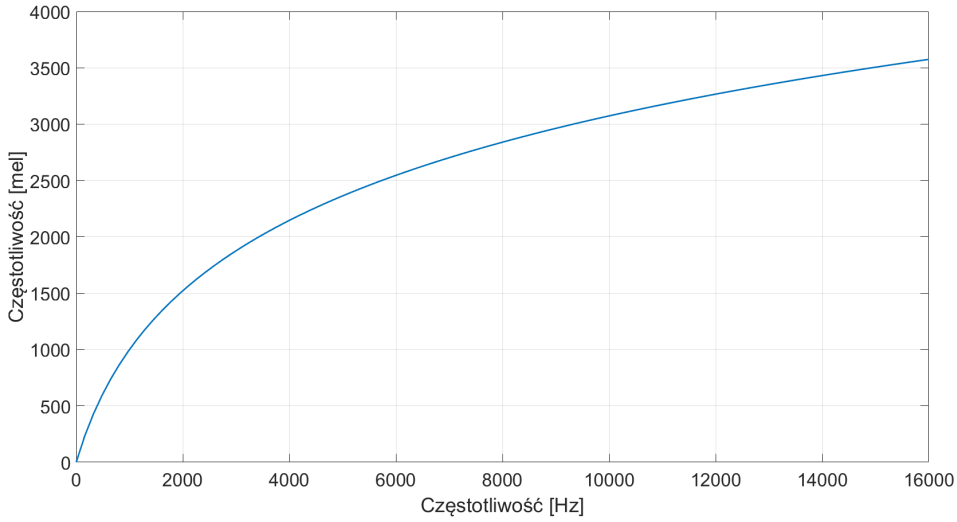
Po raz pierwszy liniowe kodowanie predykcyjne zostało zastosowane w 1966 roku przez Saito i Itakura [84]. Od tego czasu metoda LPC syste-

matycznie zyskiwała na znaczeniu w zagadnieniach związanych z rozpoznawaniem mowy. Idea metody sprowadza się do próby aproksymacji każdej nowej próbki w oparciu o ważoną liniową kombinację próbek ją poprzedzających. Wagi dobierane są w taki sposób aby średniokwadratowy błąd predykcji był jak najmniejszy [141]. Przebieg procesu wyznaczania parametrów liniowego kodowania predykcyjnego został szczegółowo przedstawiony w podrozdziale 3.2.

Badania przeprowadzone pod auspicjami Polskiej Akademii Nauk [67] pokazują, iż najlepsze wyniki w przetwarzaniu sygnału mowy uzyskiwane są w przypadku gdy liczba współczynników LPC wynosi 12. Taką też wartość przyjęto w niniejszych badaniach.

Współczynniki analizy capstralnej w skali mel (MFCC)

Współczynniki MFCC po raz pierwszy zostały opisane w 1980 roku [154] i obecnie są szeroko stosowane jako parametry opisu sygnału mowy [70, 71]. Zależność między częstotliwością wyrażoną w mel i w Hz została przedstawiona na Rysunku 2.5.

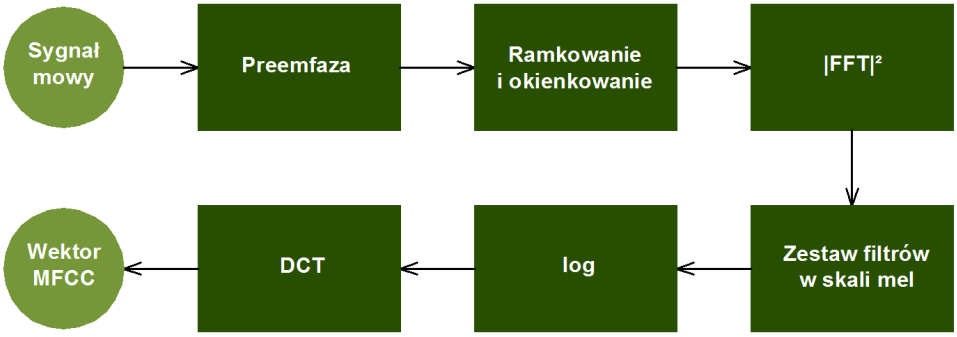


Rysunek 2.5. Zależność między częstotliwością wyrażoną w mel i w Hz. Opracowanie własne na podstawie [161]

Jak widać na Rysunku 2.5. różnica pomiędzy skalami polega na nieliniowym skalowaniu częstotliwości w przypadku skali mel. Zależność pomiędzy

skalami jest określona przy pomocy następującego równania [154]:

$$f_{mel}(f_{Hz}) = 2595 log_{10}(1 + \frac{f_{Hz}}{700}). \tag{2.5}$$



Rysunek 2.6. Schemat wyznaczania parametrów wektora MFCC. Opracowanie własne na podstawie [154].

Na Rysunku 2.6. został przedstawiony schemat wyznaczania parametrów MFCC. Pierwszym krokiem przetwarzania jest osłabienie składowych o niskich częstotliwościach oraz wzmocnienie tych o wysokich. Całość przekształceń odbywa się w oparciu o następującą równość [71]:

$$x'_n = x_n - ax_{n-1}, \tag{2.6}$$

gdzie:
x', x_n – oznacza odpowiednio sygnał po i przed preemfazie. Z kolei jako wartość a przyjmowane jest 0,97 [71].

Kolejnym krokiem przetwarzania sygnału jest ramkowanie oraz okienkowanie. Szerzej opisane w Rozdziale 3. W procesie okienkowania wykorzystywane jest okno Hamminga. Na podstawie każdej z ramek zostaje wyznaczone widmo mocy $|FFT|^2$. Następnie widmo zostaje poddane działaniu filtrów. W literaturze oraz niniejszej pracy najczęściej spotykane są filtry trójkątne, o środkach równomiernie rozłożonych w określonym zakresie częstotliwości. W wyniku działania filtrów otrzymywana jest energia pasma liczona według wzoru [71]:

$$S_m = \sum_{k=1}^{N} |X_r(k)|^2 H_m(k), \tag{2.7}$$

gdzie:

X_r – oznacza widmo ramki,
m – oznacza numer filtra,
H_m – oznacza zestaw filtrów.

W dalszym etapie przekształceń wykorzystywany jest logarytm energii ze względu na podobieństwo w modelowaniu do nieliniowej amplitudowej wrażliwości ucha człowieka.

Bezpośrednio wyznaczeniu wektora współczynników służy dyskretna transformata kosinusowa (DCT), zaś kolejne wartości wyznaczane są w oparciu o następujące równanie [71]:

$$c_i = \sqrt{\frac{2}{M}} \sum_{m=1}^{M} log(S_m) cos(\frac{\pi i}{M}(m - 0,5)), \tag{2.8}$$

gdzie:
M – oznacza liczbę użytych filtrów,
i – oznacza numer współczynnika.

W niniejszych badaniach wykorzystano 12 współczynników MFCC.

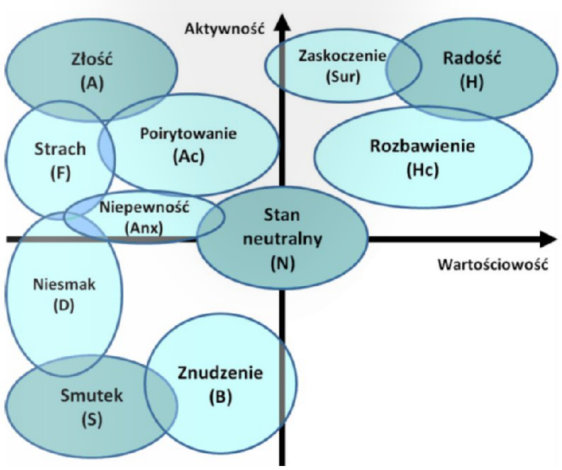
2.4. Przegląd istniejących metod przetwarzania sygnału mowy polskiej

Od kilku lat można zaobserwować dynamiczny rozwój metod, służących do identyfikacji stanu emocjonalnego mówcy. Taka sytuacja ma miejsce, zarówno jeśli chodzi o język polski, jak i inne języki takie jak angielski, niemiecki, hiszpański czy chiński. Istniejące rozwiązania skupiają się przede wszystkim na statystycznych parametrach sygnału mowy polskiej. Znane są publikacje wykorzystujące do identyfikacji narzędzia oparte na algorytmie k-Najbliższych Sąsiadów [70], maszynie wektorów wspierających [57], metodach multimodalnych opracowanych w oparciu o kilka dostępnych współcześnie rozwiązań [68] drzewach decyzyjnych czy ukrytych modelach Markova [72, 159].

W 2008 roku został opracowany na Politechnice Warszawskiej mechanizm, wykorzystujący maszynę wektorów wspierających (ang. Support Vector Machine) do identyfikacji stanu emocjonalnego osoby mówiącej [57]. W podejściu tym badana jest intensywność występowania poszczególnych elementów składowych emocji w przestrzeni trójwymiarowej, w której kolejne płaszczyzny opisane są jako: aktywność, wartościowość i dominacja. W praktyce najczęściej wykorzystywane są płaszczyzny dwuwymiarowe [135, 136]. Na Rysunku 2.7. została przedstawiona jedna z moż-

liwości zrzutowania poszczególnych stanów emocjonalnych na płaszczyznę wartościowość-aktywność.

Naukowcy z Politechniki Warszawskiej przeprowadzili badania w oparciu o dwie grupy nagrań: materiały przygotowane przy współpracy z osobami zawodowo trudniącymi się aktorstwem oraz nagrania mowy spontanicznej. Separacja parametrów służących do badań mowy emocjonalnej nie jest zagadnieniem prostym.



Rysunek 2.7. Rozmieszczenie stanów emocjonalnych na płaszczyźnie wartościowość-aktywność [57]

Nie ma uniwersalnego zbioru parametrów pozwalających na jednoznaczną identyfikację emocji w głosie. Dlatego też najpopularniejsza jest heurystyczna metoda doboru właściwości sygnału mowy [52, 57]. Polega ona na wyodrębnieniu możliwie największej liczby parametrów sygnału, a następnie na eksperymentalnym wyborze tych cech, które jak najdokładniej opisują badane zagadnienie. Najczęściej z sygnału mowy są ekstrahowane następujące parametry [57]:

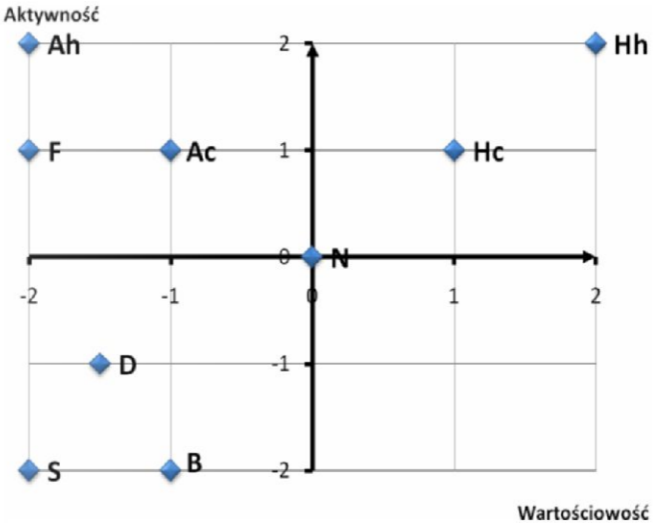
- ton krtaniowy F0 oraz statystyczne parametry sygnału mowy takie jak: wartość średnia, mediana, odchylenie standardowe itp.,
- parametry związane z energią sygnału,
- współczynniki mel-capstrum (ang. MFCC – Mell Frequency Capstral Coefficients),
- parametry opisujące zależności czasowe sygnału mowy, takie jak: tempo mówienia, liczba fragmentów dźwięcznych w jednostce czasu czy stosunek mowy bezdźwięcznej do dźwięcznej,
- wartości częstotliwości formantowych F1–F4,

- parametry opisujące jakość głosu [92].

W celu ograniczenia ilości wykorzystywanych parametrów jak również doboru tych właściwych, najlepiej opisujących badany problem, należy dokonać ich wyboru. Służą temu metody oparte na doświadczeniu naukowców oraz algorytmy selekcji. Do najczęściej stosowanych mogą być zaliczone następujące [41, 92, 135]:

- sekwencyjna płynna selekcja wstępna (ang. Sequential Floating Backward Selection),
- sekwencyjna płynna selekcja postępująca (ang. Sequential Floating Forward Selection),
- sekwencyjna selekcja postępująca (ang. Sequential Forward Selection),
- sekwencyjna selekcja wsteczna (ang. Sequential Backward Selection).

Wyżej wymienione algorytmy selekcji postępującej zaczynają pracę od zbioru pustego, dodając sekwencyjnie te parametry, które pozwolą na jak najlepszą separację klas. Działanie jest kończone w momencie, gdy dodanie kolejnego parametru nie wpływa na wzrost dokładności dyskryminacji klas lub gdy zostanie osiągnięta wymagana liczba parametrów [57]. Algorytmy selekcji wstecznej działają zaczynając od zbioru zawierającego wszystkie wyekstrahowane parametry, a następnie usuwają kolejne do momentu uzyskania zadowalającej separowalności klas. Dodatkowo sprawdzają one poprawność wybieranych parametrów.



Rysunek 2.8. Proponowane rozmieszczenie stanów emocjonalnych na płaszczyźnie wartościowość-aktywność [57]

W badaniach przeprowadzonych przez pracowników Politechniki Warszawskiej zostały wykorzystane dwie bazy nagrań mowy emocjonalnej. Pierwsza, została opracowana przez autorów badania, druga z kolei, zawierała nagrania znajdujące się w Berlińskiej Bazie Nagrań Emocjonalnych (EMO-DB) [12]. W wymienionych wyżej pracach położono nacisk na identyfikację następujących stanów emocjonalnych: stanu neutralnego, radości, smutku oraz znudzenia i strachu. Autorska baza nagrań (BES) została przygotowana w oparciu o audycje radiowe programów rozrywkowych, sportowych oraz publicystycznych transmitowanych za pośrednictwem Polskiego Radia, dzięki czemu poza emocjami zgromadzonymi w Berlińskiej Bazie Danych, autorom udało się sklasyfikować dodatkowo takie stany emocjonalne jak: poirytowanie, rozbawienie, niesmak, zaskoczenie, niepokój czy ekstaza [57]. Liczebność bazy wynosiła 340 nagrań. Na Rysunku 2.8. został przedstawiony proponowany przez autorów z Politechniki Warszawskiej rozkład emocji na płaszczyźnie wartościowość-aktywność.

Poszczególne stany emocjonalne zostały oznaczone następująco:

- N – stan neutralny,
- A – złość,
- H – radość,
- S – smutek,
- B – znudzenie,
- Ac – poirytowanie,
- Hc – rozbawienie,
- D – niesmak,
- Sur – zaskoczenie,
- Anx – niepokój,
- Hh – ekstaza

Podczas opracowywania lokalizacji wyszczególnionych stanów emocjonalnych, przez naukowców z Politechniki Warszawskiej, zostały wykorzystane następujące parametry wyodrębnione z sygnału mowy [57]:

- niedokładność zamknięcia krtani,
- gradienty widmowe,
- częstotliwość F0 oraz jej wartości pochodne,
- liczba zmian dźwięczności przypadająca na jednostkę czasu,
- energia sygnału,
- szerokość oraz lokalizacja pasm formantów F1–F4.

Klasyfikacja poszczególnych stanów emocjonalnych odbywała się z wykorzystaniem maszyny wektorów wspierających oferowaną przez pakiet narzędzi MatLab, umożliwiający przekształcenie wielomianowe 3. stopnia. Dziesięciokrotna walidacja krzyżowa (ang. cross-validation) została wyko-

rzystana jako metoda służąca do testów. Działanie metody polega na wyodrębnieniu dziesięciu podzbiorów, z których dziewięć jest wykorzystywanych jako zbiory uczące, dziesiąty zaś pełni rolę zbioru testowego. Zbiór danych należy przetestować dziesięciokrotnie, tak aby każdy z podzbiorów mógł być wykorzystany w charakterze testowym. Podczas badań została zachowana niezależność mówcy. Otrzymane rezultaty zostały przedstawione w Tabelach 2.1. i 2.2. oraz na Rysunku 2.9.

W przeprowadzonych badaniach poprawność rozpoznawania emocji dla bazy danych BES wyniosła średnio 58,9%, zaś dla Berlińskiej Bazy Nagrań Emocjonalnych osiągnięto około 62,7% skuteczność identyfikacji.

W badaniach prowadzonych na Politechnice Łódzkiej w 2012 roku został opracowany mechanizm klasyfikacji stanu emocjonalnego mówcy oparty na algorytmie k-Najbliższych Sąsiadów (ang. k-Nearest Neightbour) [71].

Tabela 2.1. Wyniki klasyfikacji otrzymane dla bazy danych BES [57] (w %)

Emocje oczekiwane	**Emocje otrzymane na wyjściu**					
	N	**Ac**	**Ah**	**Hc**	**Hh**	**S**
N	**66,68**	10,45	4,48	11,94	0,00	4,48
Ac	13,04	**47,83**	13,04	17,39	6,52	2,17
Ah	10,00	35,00	**27,50**	20,00	5,00	2,50
Hc	20,27	24,32	8,11	**31,08**	6,76	9,46
Hh	4,26	8,51	6,38	21,28	**51,06**	8,51
S	1,52	12,12	3,03	12,12	1,52	**69,70**

Tabela 2.2. Wyniki klasyfikacji otrzymane dla bazy danych EMO-DBS [57] (w %)

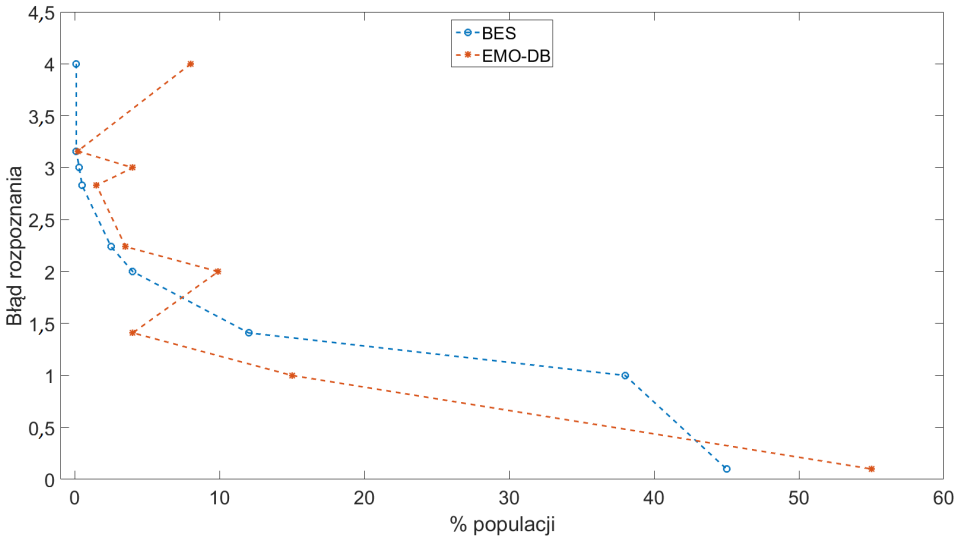
Emocje oczekiwane	**Emocje otrzymane na wyjściu**				
	N	**A**	**H**	**S**	**B**
N	**82,05**	1,28	7,69	2,56	6,41
A	1,60	**79,20**	19,20	0,00	0,00
H	0,00	25,00	**73,53**	0,00	1,47
S	3,23	0,00	0,00	**85,48**	11,29
B	11,25	1,25	7,50	6,25	**73,75**

Zauważono, iż w zależności od artykulacji głosek dźwięcznych i bezdźwięcznych zmienia się ułożenie strun głosowych, co z kolei prowadzi do powstawania tak zwanej częstotliwości podstawowej F0, która to stała się jednym z parametrów rozpatrywanych w badaniach.

Parametr ten jest związany z budową krtani i ma różne zakresy w zależności od płci. Dla mężczyzn częstotliwość podstawowa oscyluje pomiędzy 85 Hz a 155 Hz, u kobiet zaś od 165 Hz do 255 Hz, zaś dla dzieci i niemowląt osiąga jeszcze wyższe wartości.

W badaniach wykorzystane zostały takie parametry związane z częstotliwością podstawową (F0) jak [70, 71]:

- średnia wartość F0,
- standardowe odchylenie częstotliwości podstawowej,
- minimalna i maksymalna wartość F0,
- wartość F0 w kwartylach,
- zakres częstotliwości podstawowej,
- skośność, wartość kurtozy, nachylenie F0 i inne.



Rysunek 2.9. Błąd rozpoznania dla klasyfikacji metodą SVM [57]

W sumie wyszczególnione zostały dwadzieścia cztery parametry, które wykorzystano w dalszej części badań.

Właściwa klasyfikacja odbywała się poprzez wykorzystanie algorytmu k-NN. Jego działanie opiera się o obliczenia odległości w przestrzeni X parametrów pomiędzy nieznaną wartością czynnika x_j, reprezentowanego w następującej postaci [71]:

$$x_j = [x^1, x^2, ..., x^n], \tag{2.9}$$

gdzie:

x^k – oznacza wartość n-tego parametru,
n – oznacza numer parametru,
a wartościami ze zbioru uczącego CU[7]:

$$x_k \in CU, \tag{2.10}$$

$k = 1, 2, 3, ..., I$,
I – oznacza liczbę zbiorów uczących.

Do obliczenia powyższej odległość stosowanych jest wiele metod. Najczęściej jest ona reprezentowana Euklidesową odległością [71]:

$$d(x_k, x_j) = \sum_{i=0}^{n} |x_j^i - x_k^i|. \tag{2.11}$$

Na Rysunku 2.10. (patrz. s. 28) został przedstawiony schemat przetwarzania sygnału mowy opracowany przez pracowników Politechniki Łódzkiej.

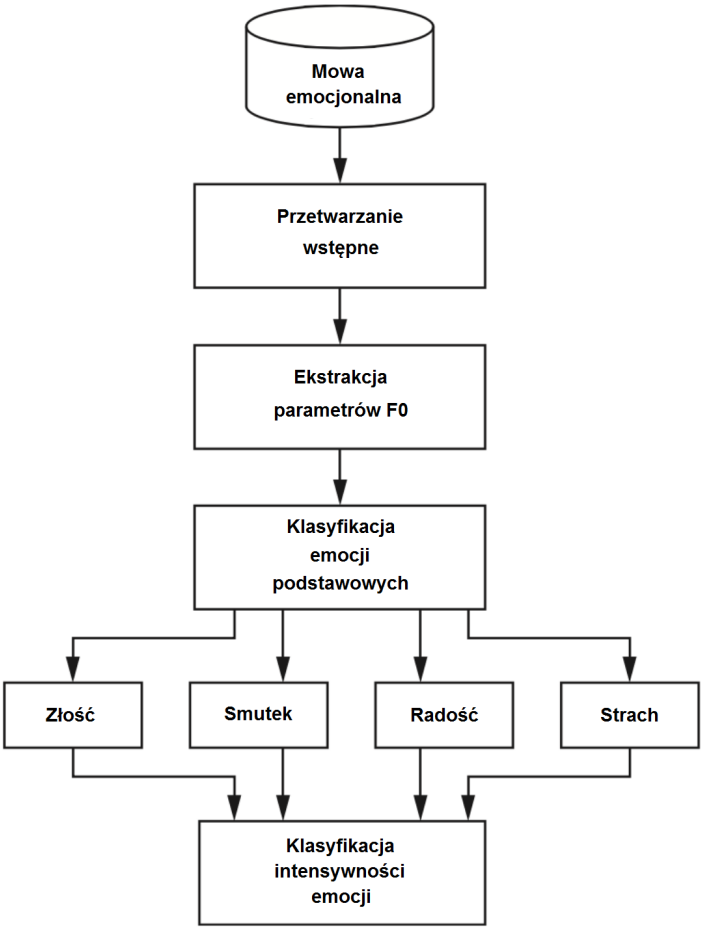
W przeprowadzonych badaniach skupiono się na następujących stanach emocjonalnych: radość, smutek, strach oraz złość. Wzięto również pod uwagę stany o wyższym i niższym zabarwieniu emocjonalnym bazując na kole emocji Plutchika. W rezultacie otrzymano zbiór składający się z dwunastu stanów emocjonalnych. W badaniach została wykorzystana autorska baza nagrań opracowana w Instytucie Elektroniki Medycznej Politechniki Łódzkiej. Uzyskane wyniki zostały pokazane w Tabeli 2.3.

Tabela 2.3. Skuteczność klasyfikacji stanów emocjonalnych przy użyciu algorytmu k-NN [71]

Grupa emocji	**Intensywność emocji**	**Skuteczność identyfikacji**
Złość	Wściekłość	83,8%
	Złość	42,9%
	Irytacja	71,4%
Smutek	Żal	71,4%
	Smutek	42,9%
	Zaduma	62,5%
Radość	Ekstaza	60,0%
	Radość	47,4%
	Pogodność	56,3%
Strach	Strach	57,1%
	Terror	83,8%
	Obawa	71,4%

W przeprowadzonych badaniach średnia skuteczność identyfikacji podstawowych emocji wyniosła około 50% dla stanów podstawowych, dla pozostałych stanów emocjonalnych kształtowała się w granicach 70%.

Naturalnym kierunkiem rozwoju powyższego algorytmu stało się opracowanie multimodalnej metody, wykorzystującej, inne niż częstotliwościowe, właściwości sygnału mowy [70]. Detekcja F0 została zrealizowana przy wykorzystaniu funkcji autokorelacji. Na jej podstawie wyekstrahowane zostały dwadzieścia trzy cechy statystyczne, stanowiące jeden z wektorów wejściowych [70]. Następnie zostały wyznaczone cztery formanty F1–F4, na podstawie których wyodrębniono podstawowe parametry statyczne, uzyskując



Rysunek 2.10. Schemat ideowy przetwarzania sygnału mowy [71]

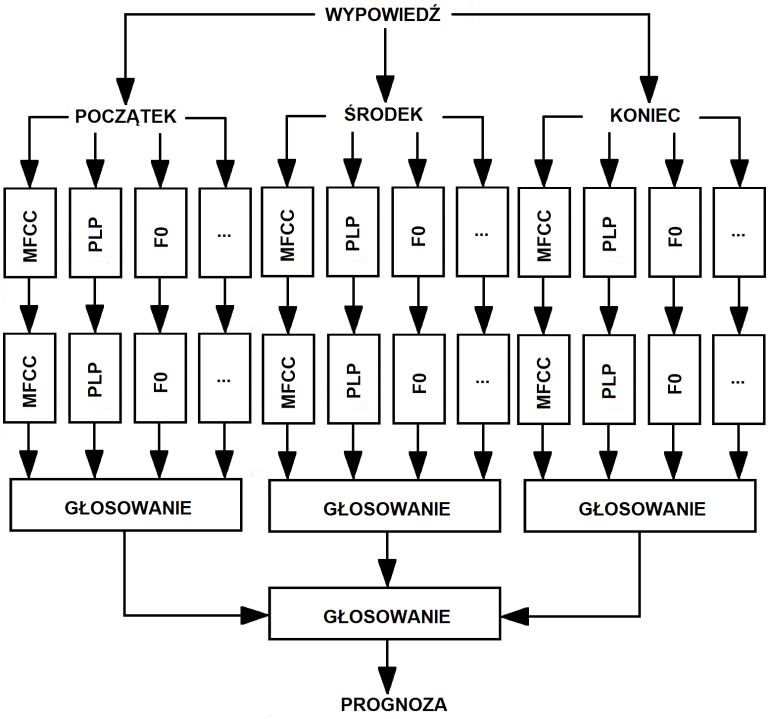
wektor cech, składający się z piętnastu elementów. Standardem w przetwarzaniu sygnału mowy są obecnie współczynniki MFCC. W przeprowadzonych badaniach wyznaczono dwanaście współczynników, z których następnie wyodrębniono sześćdziesiąt parametrów, stanowiących kolejny wektor cech. Liniowe kodowanie predykcyjne (LPC) jest szeroko stosowaną metodą w identyfikacji mowy [64].

W przeprowadzonych badaniach wykorzystano dwanaście parametrów LPC, na podstawie których wyekstrahowano sześćdziesiąt cech stanowiących czwarty wektor wejściowy. Ostatni wektor stanowił zestaw cech oparty o współczynniki PLP (ang. Perceptual Linear Prediction). W Tabeli 2.4. zestawiono wszystkie wykorzystywane przez algorytm parametry.

Tabela 2.4. Zestawienie parametrów wykorzystywanych w multimodalnej metodzie identyfikacji stanów emocjonalnych mówcy [70]

Badany parametr	**Liczba wyekstrahowanych współczynników**	**Dokładność rozpoznania (waga)**
Parametry statystyczne tonu krtaniowego F0	23	57,1%(3)
Energia sygnału	6	52,6%(2)
Parametry statystyczne formantów F1–F4	60	49,7%(1)
Parametry statystyczne dla MFCC	60	75,8%(6)
Parametry statystyczne dla LPC	60	65,0%(4)
Parametry statystyczn dla PLP	60	75,1%(5)

Schemat algorytmu klasyfikacji został przedstawiony na Rysunku 2.11. Pierwszym etapem klasyfikacji jest podział wypowiedzi na trzy części o równej długości: początek, środek i koniec. Następnie każda z nich została przetworzona za pomocą algorytmu k-NN. W rezultacie otrzymano osiemnaście klas (po sześć dla każdej części wypowiedzi). Kolejny etap obejmował przydzielenie wag na zasadzie głosowania. Wartość wagi związana była z dokładnością rozpoznawania przedstawioną w Tabeli 2.3. (patrz. s. 27). Ostatni krok stanowiło głosowanie w obrębie całej wypowiedzi, efektem czego była prognoza [70].

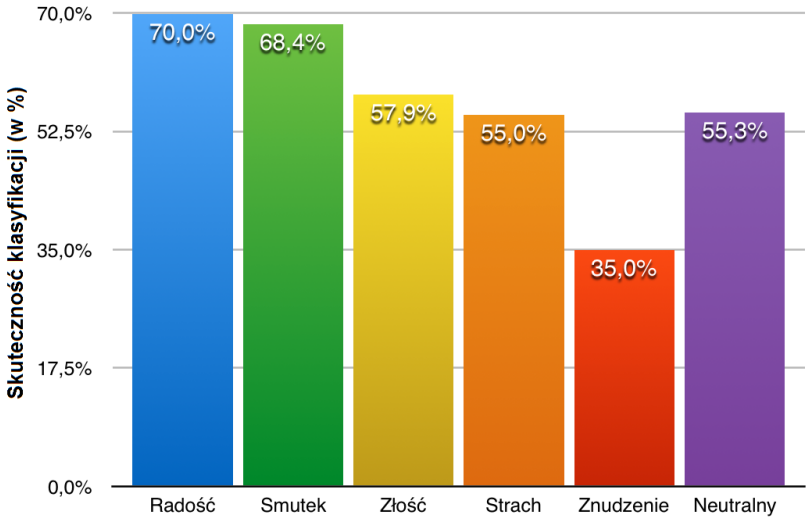


Rysunek 2.11. Proponowany schemat klasyfikacji [71]

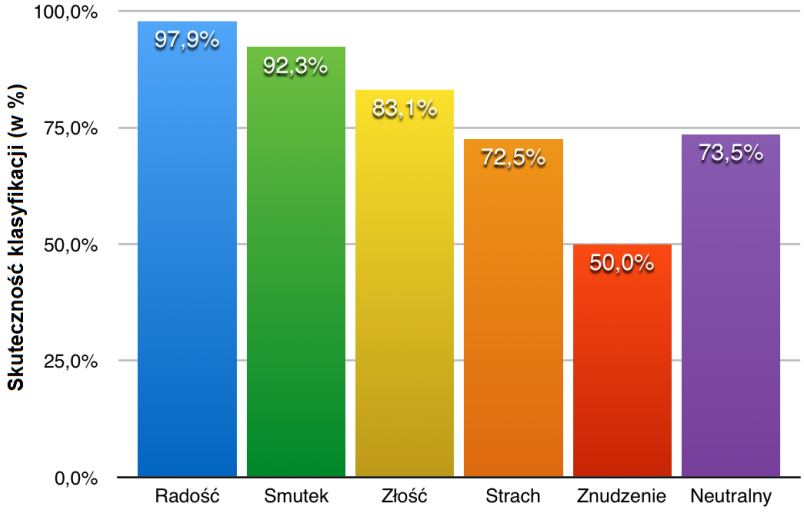
Bazując na przedstawionym wyżej algorytmie przeprowadzone zostały testy na podstawie dwóch grup nagrań. Pierwszą, stanowiła baza emocji odegranych, drugą – baza emocji spontanicznych. Wyniki uzyskane dla obydwu grup zostały przedstawione na Rysunkach 2.12. oraz 2.13. (patrz. s. 31).

Kolejną grupę metod służących do klasyfikacji stanów emocjonalnych stanowią drzewa decyzyjne [21]. W badaniach zostały wykorzystane dwie grupy parametrów. Pierwszą z nich stanowiły czynniki związane z liniową i nieliniową regresją, wyodrębnione z deskryptorów sygnału mowy. Założono, że właściwy opis sygnału mowy stanowić będą parametry takie jak: maksymalne, minimalne i uśrednione wartości sygnał mowy. Drugą grupę stanowiły parametry związane z energią sygnału. W sumie zostało wyodrębnione dziewiętnaście parametrów. W oparciu o wyekstrahowane deskryptory zbudowane zostało drzewo decyzyjne, które na każdym poziomie było w stanie skategoryzować jeden stan emocjonalny. W badaniach skupiono się na emocjach najczęściej występujących w literaturze, tj.: strachu, smutku, radości, złości, znudzeniu oraz stanie neutralnym.

Przeprowadzono badania opierając się na dwóch bazach nagrań. Pierwszą była wspominana już Berlińska Baza Nagrań Emocjonalnych (EMO-DB) [12] drugą autorska baza zawierająca 240 plików dźwiękowych. Badania przeprowadzone zostały dla dwóch przypadków.



Rysunek 2.12. Wyniki klasyfikacji uzyskane dla pierwszej grupy. Opracowanie własne na podstawie [71]



Rysunek 2.13. Wyniki klasyfikacji uzyskane dla drugiej grupy. Opracowanie własne na podstawie [71]

Pierwszy zakładał niezależność mówcy od wypowiadanego tekstu, drugi taką zależność dopuszczał. Najlepsze otrzymane wyniki zostały zaprezentowane w Tabeli 2.5. Średnia skuteczność identyfikacji stanu emocjonalnego mówcy w przypadku binarnych drzew decyzyjnych wyniosła około 72% [21].

Tabela 2.5. Otrzymane wyniki przy wykorzystaniu metody binarnych drzew decyzyjnych [21]

Baza danych	**Zależność mówcy od tekstu**	**Niezależność mówcy (od tekstu)**
Autorska baza	76,30 %	64,18%
EMO-DB	74,39%	72,04%(2)

W literaturze przedmiotu często spotykaną metodą służącą do klasyfikacji stanowią ukryte modele Markova (HMM) (ang. Hidden Markov Models) [22, 23, 37]. Klasyfikacja w oparciu o HMM opiera się o proces decyzyjny Bayesa, w którym prawdopodobieństwo wygenerowania przez sygnał określonego ciągu znaków jest maksymalizowane [125]. Średnia skuteczność identyfikacji uzyskiwana przy wykorzystaniu tego klasyfikatora dla języka polskiego waha się od około 63% do 75% [61]. Jest to zatem wynik porównywalny z innymi opisanymi powyżej.

Konstruowanie efektywnych struktur danych opisujących obiekty o charakterze subuektywnym (emocje) nie jest prostym zadaniem. Najczęstsze modele mowy emocjonalnej obejmują zestaw parametrów wyekstrahowanych bezpośrednio z sygnału mowy, takich jak częstotliwość podstawowa (F0), formant częstotliwości [70], cepstralne współczynniki częstotliwości Mel (MFCC) [3, 58, 85], Linear Predictive Coding (LPC) [18, 98]. Kolejną grupę stanowią modele mowy emocjonalnej oparte na danych uzyskanych ze statystycznych parametrów sygnału mowy [43]. Proponowane są również modele bazujące na parametrach związanych z energią i mocą sygnału mowy oraz modele mowy emocjonalnej przedstawiane w formie graficznej, m.in. spektrogramy [4, 5, 122]. Połączone cechy spektralne i prozodyczne są również brane pod uwagę przy rozpoznawaniu emocji [150].

Ponieważ emocję można wyrazić w sposób łagodny lub intensywny, warto spróbować je odróżnić. W pracy [51] badano intensywność wyrażania emocji. Jako współczynniki ekstrakcji cech wykorzystano częstotliwość i amplitudę sygnału. Bazując na badaniach empirycznych ustalono, że gniew zwiększa szybkość mowy, a smutek ją spowalnia. I odwrotnie, intensywność jest wyższa, jeśli mowa jest bardziej radosna. W badaniach opisano metodę częstotliwości impulsów głosowych (GPF) z dokładnością przewidywania modelu około 83%. Przeprowadzono analizę porównawczą metod GPF

i MFCC z modyfikacjami. Wykazano, że w przypadku konkretnych emocji, które w dużym stopniu wpływają na częstotliwość wypowiedzi (tj. gniewu), lepiej sprawdza się metoda MFCC, natomiast pozostałe emocje lepiej wykrywa zaproponowana w badaniu metoda GPF. Wynika to z faktu, że GPF opiera się na analizie pełnego spektrum sygnału głosowego, a nie tylko w domenie częstotliwości.

Od niedawna stosowane są również podejścia wielomodelowe. Ze względu na komplementarność danych audio i wideo, detekcja stanów emocjonalnych jest analizowana w dwóch etapach, rozpoznając kanał audio jako pierwszą i jako drugą klatkę wideo (w zależności od tego, czy kanał audio zawierał odpowiednią informację) [2]. Wielowarstwowa sieć propagacji wstecznej perceptronu jest używana zarówno do klasyfikacji zarówno głosu, jak i obrazu. Odpowiednio ze współczynnikiem klasyfikacji dla izolowanego sygnału audio około 90%. Połączenie klasyfikacji audio i wideo zwiększa dokładność do około 95%, przy znacznie większej złożoności przetwarzania danych. Podobne podejście do analizy wielomodelowej zostało przedstawione w [13].

Czesto badacze skupiają się na samej separacji stanów emocjonalnych. Jak powszechnie wiadomo występowanie pojedynczej emocji jest w realnych rozmowach sytuacją nad wyraz rzadką. Zazwyczaj przeplatają się one tworząc strukturę łańcucha emocji. Badania nad zagadnieniami tego typu zostały przedstawione w [127]. Pokazane zostało wyodrębnienie cech charakteryzujących stan niepokoju, jako złożenia trzech stanów emocjonalnych: gniewu, strachu i smutku.

W pracy [29] omówiono problem rozpoznawania emocji w mowie szeptanej, która naturalnie różni się od mowy normalnej. Wykorzystywane zostało rozwiązanie rozszerzone o uczenie transferu cech w oparciu o autoenkoder z redukcją szumów. Takie podejście pozwala na użycie znanego rozwiązania do celów innych niż oryginalne, z dokładnością rozpoznawania co najmniej na poziomie normalnej mowy.

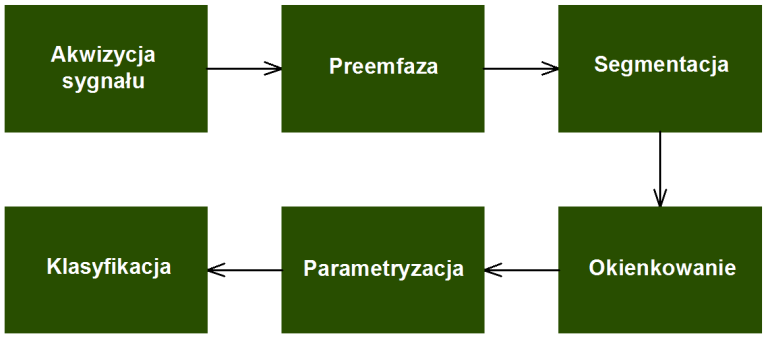
Klasyfikatory wykorzystywane dla szeregowania mowy emocjonalnej są powszechnie znanymi narzędziami, które są od dawna stosowane dla problemów przetwarzania sygnałów mowy oraz identyfikacji innych typów danych. Obejmują one: modele Gaussa (GMM) [123], ukryte modele Markowa (HMM) [119], maszynę wektorów nośnych (SVM) [36], algorytm k-NN, czyli szeroko rozumiane sztuczne sieci neuronowe [65, 88]. Znane są również publikacje, w których w proces identyfikacji emocji zaangażowanych było kilku klasyfikatorów [69].

2.5. Wnioski do rozdziału

1. Na przestrzeni wieków zrodziło się wiele teorii opisujących procesy powstawania i klasyfikowania emocji jak również interpretowania ich przez człowieka.
2. Do najpopularniejszych modeli emocji należy zaliczyć teorię Jamesa-Langego oraz model emocji Plutchika.
3. Opierając się na powyższych teoriach można wyróżnić następujące emocje podstawowe (pierwotne): radość, smutek, strach, złość oraz znudzenie.
4. Z uwagi na brak ostrości przejść z jednego stanu emocjonalnego w drugi należy rozważyć umiejscowienie stanu neutralnego.
5. Do najpopularniejszych klasyfikatorów stanów emocjonalnych mówcy w oparciu o język polski należą: algorytm k–NN, maszyna wektorów wspierających, binarne drzewo decyzyjne oraz ukryte modele Markova.
6. Średnia skuteczność identyfikacji emocji waha się od około 60% do niemal 80%.

3. Etapy przetwarzania sygnału mowy

Wydobycie oczekiwanych informacji z sygnału mowy nie jest możliwe bez wcześniejszego przetworzenia przez system komputerowy, otrzymanych danych. Najogólniej proces ten może zostać podzielony na kilka etapów takich jak: akwizycja, preemfaza, segmentacja czy parametryzacja [163]. Schemat ideowy procesu przetwarzania sygnału mowy został zaprezentowany na Rysunku 3.1.

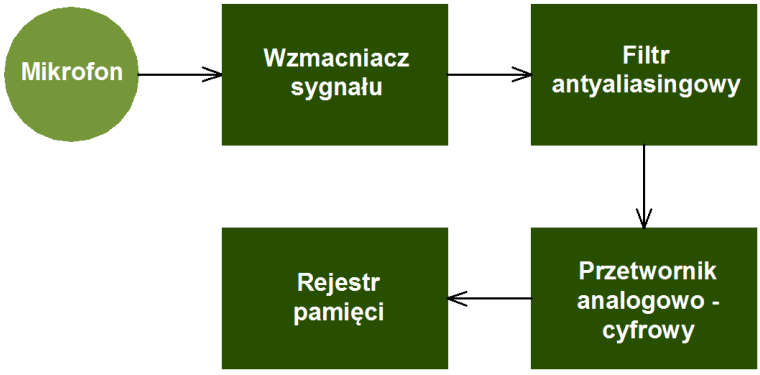


Rysunek 3.1. Schemat ideowy procesu przetwarzania sygnału mowy. Opracowanie własne na podstawie [154]

Przetwarzanie wstępne stanowi pierwszy z etapów zamiany analogowego sygnału mowy na sygnał cyfrowy. Zarejestrowany przy użyciu mikrofonu sygnał zostaje wzmocniony, a następnie poddany procesowi filtracji, którego celem jest redukcja szumów występujących w nagraniu [154]. Wykorzystanie przetwornika analogowo-cyfrowego A/D (ang. Analog/Digital) pozwala przekształcić sygnał analogowy do postaci zdigitalizowanej. Zastosowanie w kolejnym etapie filtra preemfazy skutkuje uwydatnieniem składowych o wyższych częstotliwościach, dzięki czemu sygnał zawiera składowe niosące właściwą informację [154]. Podział nagrania na mniejsze fragmenty, zwane ramkami, sprawia, iż w obrębie ramki segment może być traktowany jako sygnał kwazistacjonarny, co ma szczególne znacznie w przypadku korzystania ze współczynników LPC. Nieciągłości na końcach poszczególnych ramek są usuwane przy użyciu funkcji okna. Tak przetworzony sygnał może zostać poddany procesowi parametryzacji. Jego istotą jest wydobycie z poszczególnych ramek możliwie najwięcej cech sygnału, które w sposób jednoznaczny będą go opisywać. W zależności od badanego zagadnienia oraz sygnału uzyskuje się od kilkunastu do kilkuset parametrów. Bazując na pozyskanych danych dokonywany jest proces klasyfikacji [97].

3.1. Akwizycja sygnału mowy

Proces przetwarzania mowy nie byłby możliwy bez akwizycji sygnału. Jest to pierwszy etap niezbędny w identyfikacji mowy emocjonalnej. Aby rejestrowany za pomocą mikrofonu sygnał analogowy, był poprawnie przetworzony musi zostać przekształcony do postaci cyfrowej. Jak już zostało wspomniane, jest to możliwe przy użyciu przetwornika analogowo – cyfrowego. Zanim jednak dane trafią do przetwornika muszą zostać wzmocnione, a następnie przetworzone przy użyciu filtru antyaliasingowego. Standardowa karta dźwiękowa posiada wszystkie wyżej wymienione elementy. Cały proces rejestracji sygnału został przedstawiony na Rysunku 3.2.



Rysunek 3.2. Schemat akwizycji sygnału mowy. Opracowanie własne na podstawie [154]

Przetwarzanie sygnału polega na jego dyskretyzacji w dziedzinie czasu oraz częstotliwości [130].

3.1.1. Dyskretyzacja czasowa

Częstotliwość próbkowania F_p jest podstawowym parametrem charakteryzującym proces próbkowania sygnału – dyskretyzacji w dziedzinie czasu. Wśród najczęściej występujących częstotliwości wymienia się takie jak: 12 kHz, 16 kHz a nawet 44 kHz [154]. Oczywistym jest, że wyższa rozdzielczość czasowa skutkuje lepszą jakością sygnału cyfrowego. Dokonując wyboru częstotliwości próbkowania, należy pamiętać o twierdzeniu Kottelnikowa-Shannona[2]. Twierdzenie to wprowadza zależność pomiędzy

[2] Bywa również określane jako twierdzenie Nyquista, Nyquista-Kottelnikowa bądź Whittakera-Nyquista-Kottelnikowa-Shannona.

częstotliwościami występującymi w sygnale a częstotliwością jego próbkowania. Zależność ta jest wyrażona następującą nierównością [163]:

$$F_p > 2F_s, \tag{3.1}$$

gdzie:
F_p – oznacza częstotliwość próbkowania sygnału,
F_s – oznacza najwyższą częstotliwość występującą w sygnale.

3.1.2. Dyskretyzacja amplitudowa

Równie ważnym parametrem charakteryzującym sygnał jest liczba bitów, mówiąca o dokładności zapisu wartości próbki. Należy dążyć do kompromisu pomiędzy wiernością zapisu sygnału, przejawiającą się wysoką dyskretyzacją czasową i częstotliwościową, a czasem niezbędnym na przetworzenie sygnału przez systemy komputerowe oraz ilością miejsca niezbędną do przechowywania zgromadzonych danych [154]. Najczęściej stosowana jest następująca ilość bitów 8, 12 oraz 16.

3.1.3. Preemfaza

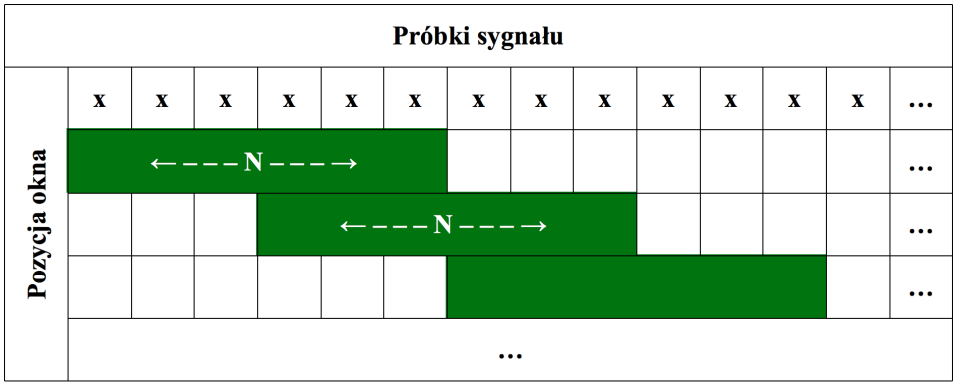
Pożądanym zjawiskiem jest ograniczenie wpływu szumów tła, powstałych na skutek działania sieci elektrycznej. Służy temu tak zwana filtracja preemfazy [6]. Proces redukcji powyższych szumów sprowadza się do ograniczenia szumów o częstotliwości z przedziału 50–60 Hz oraz uwydatnieniu składowych widma o wyższych częstotliwościach. Ponadto proces preemfazy eliminuje zjawisko arytmetyki skończonej precyzji, związane z reprezentacją danych zapisanych w postaci cyfrowej. Szerzej opisane w [154]. Często stosowany w tym zagadnieniu jest górnoprzepustowy filtr o odpowiedzi impulsowej FIR (ang. Finite Impulse Response Hight Pass Filter). Postać równania różnicowego przyjmuje wtedy następującą postać [154]:

$$s[n] = \tilde{s}[n] - a\tilde{s}[n-1], \tag{3.2}$$

gdzie:
$\tilde{s}[n]$ – oznacza sygnał przed filtracją próbki o numerze n,
$s[n]$ – oznacza sygnał cyfrowy po przetworzeniu n-tej próbki
$0,9 \leqslant a \leqslant 1$.
Często za wartość parametru a przyjmuje się 0,95 dzięki czemu wzmocnienie składowych wyższych niż 20 dB jest możliwe.

3.1.4. Segmentacja

Proces segmentacji sygnału *s[n]* (ang. frame blocking) stanowi jeden z pierwszych etapów jego przetwarzania w systemach cyfrowych. W swej idei polega na podziale sygnału na segmenty (ramki) o długości odpowiedniej do dalszych badań. Należy przy tym dążyć do uzyskania kwazistacjonarnego sygnału w obrębie danej ramki. W celu poprawienia jakości sygnału częstym zabiegiem jest nakladkowanie (ang. overlapping). Polega ono na uwzględnieniu powtarzalności próbek w sąsiednich ramkach [79]. Idea procesu nakładkowania została zaprezentowana na Rysunku 3.3.



Rysunek 3.3. Idea procesu nakładkowania. Opracowanie własne na podstawie [79]

Użycie zbyt krótkiej ramki[3] powoduje znaczący wzrost obliczeń koniecznych do przetworzenia sygnału. Ramka zbyt długa[4] sprawia, że sygnał nie może zostać potraktowany jako kwazistacjonarny, co uniemożliwia zastosowanie części metod na etapie parametryzacji sygnału[5] [79]. Ogólnie możliwe jest zastosowanie następującej równości identyfikującej *i–tą* ramkę $\tilde{x}_i[n]$ sygnału *s[n]* [28]:

$$\tilde{x}_i[n] = s[Mi + n], \tag{3.3}$$

gdzie:
$n = 0, 1, 2, ..., N - 1$, N – oznacza liczbę wszystkich próbek,
$i = 0, 1, 2, ..., L - 1$, M – oznacza liczbę wszystkich ramek.

[3] O długości poniżej 10 ms.
[4] O długości powyżej 50 ms.
[5] Np. techniki LPC.

3.1.5. Okienkowanie

Dekompozycja sygnału na segmenty powoduje powstawanie nieciągłości na końcach ramek [79], co skutkuje powstaniem w widmie sygnału składowych o wyższych częstotliwościach. Okienkowanie (ang. windowing) jest procesem przeciwdziałającym takiemu zjawisku. Jego celem jest zatem minimalizacja błędu estymacji funkcji autokorelacji na końcach każdej z ramek. Wygładzenie nieciągłości powoduje usunięcie z widma sygnału fałszywych składowych. Sam proces okienkowania polega na splocie sygnału z funkcją okna [143]. Należy pamiętać, iż proces ten poza usuwaniem nieciągłości, powoduje dodatkowe tłumienie sygnału wejściowego. Mianem okien czasowych są określane funkcje spełniające następujące warunki [20]:

- są symetryczne względem środka przedziału,
- osiągają maksimum w środku przedziału [6],
- są niezerowe w skończonym przedziale czasu.

Innymi słowy okno czasowe opisuje sposób pozyskania próbek z analizowanego sygnału. Jeśli założymy, że w skończonym przedziale czasu, dany jest sygnał *s[n]* wówczas wynik obserwacji takiego impulsu w oknie *w[n]* stanowić będzie funkcja *g[n]*, będące splotem, zdefiniowana następująco [163]:

$$g[n] = s[n] * w[n], \tag{3.4}$$

$n \in (-\infty, +\infty)$.

3.1.6. Okno Hamminga

Zaproponowana przez Richarda W. Hamminga funkcja stanowi szczególny przykład okna czasowego. Funkcja ta została opracowana w celu minimalizacji maksymalnej wartości płatka bocznego i dla każdej próbki sygnału n opisana jest następującą równością [48]:

$$w[n] = \alpha - \beta cos(\frac{2\pi n}{N-1}), \tag{3.5}$$

gdzie:
$\alpha = 0,54$,
$\beta = 1 - \alpha = 0,46$,
N – oznacza liczbę próbek sygnału.

[6] Poza oknem prostokątnym.

3.1.7. Okno Gaussa

Jednym z najczęściej stosowanych funkcji czasowych jest okno Gaussa, które zdefiniowane jest następującym wzorem [48]:

$$w[n] = e^{-\frac{1}{2}\left(\frac{n-\frac{N-1}{2}}{\sigma-\frac{N-1}{2}}\right)^2}, \tag{3.6}$$

gdzie:
N – oznacza liczbę próbek sygnału,
$\sigma \leqslant 0,5$.
Okno Gaussa często znajduje zastosowanie w zagadnieniach wykorzystujących transformatę Fouriera, gdyż jego kształt jest zbliżony do paraboli, co pozwala na swobodne wykorzystanie jej w kwadratowej interpolacji estymacji częstotliwościowej. Ponadto jest to funkcja własna transformacji Fouriera[7], której odchylenie standardowe wyninosi $\frac{N}{2}$.

3.1.8. Okno Dolpha-Czebyszewa

Okno Dolpha-Czebyszewa jest definiowane następująco [93]:

$$w[n] = w_0\left(n - \frac{N-1}{2}\right), \tag{3.7}$$

gdzie:

$$w_0[n] = \frac{1}{N}\sum_{k=0}^{N-1} W_0[k] e^{\frac{i2\pi kn}{N}}, n \in [-\frac{N}{2}, \frac{N}{2}], \tag{3.8}$$

$$W_0[k] = \frac{cos\langle N cos^{-1}[\beta cos\left(\frac{\pi k}{N}\right)]\rangle}{cosh\left(N cosh^{-1}(\beta)\right)}, \tag{3.9}$$

gdzie:
$\beta = cosh[\frac{1}{N} cosh^{-1}\left(10^{\alpha}\right)]$.
Parametr α jest definiowany za pomocą norm Czebyszewa i jest równy logarytmowi stosunku wysokości maksima głównego do bocznych [139].

3.1.9. Okno Blackmana

Jedną z funkcji zwiększających dynamikę sygnału do poziomu około 40 dB jest okno Blackmana. Wzrost dynamiki skutkuje jednak zwiększeniem

[7] W wyniku transformacji Fouriera okna Gaussa dostajemy funkcję Gaussa, dla odpowiednio dobranych parametrów.

listka głównego. Samo okno definiowane jest następująco [47]:

$$w[n] = a_0 - a_1 cos\left(\frac{2\pi n}{N-1}\right) + a_2 cos\left(\frac{4\pi n}{N-1}\right), \tag{3.10}$$

gdzie:
$a_0 = \frac{1-\alpha}{2}, a_1 = \frac{1}{2},$
$a_2 = \frac{\alpha}{2}, \alpha = 0,16.$

3.1.10. Okno Nuttala

Z matematycznego punktu widzenia okno Nuttala stanowi pewne uszczegółowienie okna Blackmana i jest opisane następującą równością [48]:

$$w[n] = a_0 - a_1 cos\left(\frac{2\pi n}{N-1}\right) + a_2 cos\left(\frac{4\pi n}{N-1}\right) - a_3 cos\left(\frac{6\pi n}{N-1}\right), \tag{3.11}$$

gdzie:
$a_0 = 0,355768,$
$a_1 = 0,487396,$
$a_2 = 0,144232,$
$a_3 = 0,012604.$

Opierając się na założeniu, iż dana jest dowolna liczba rzeczywista n, wówczas zarówno funkcja Nuttala jak i jej pierwsza pochodna są ciągłe. Innymi słowy funkcja dąży do 0 dla $n = 0$.

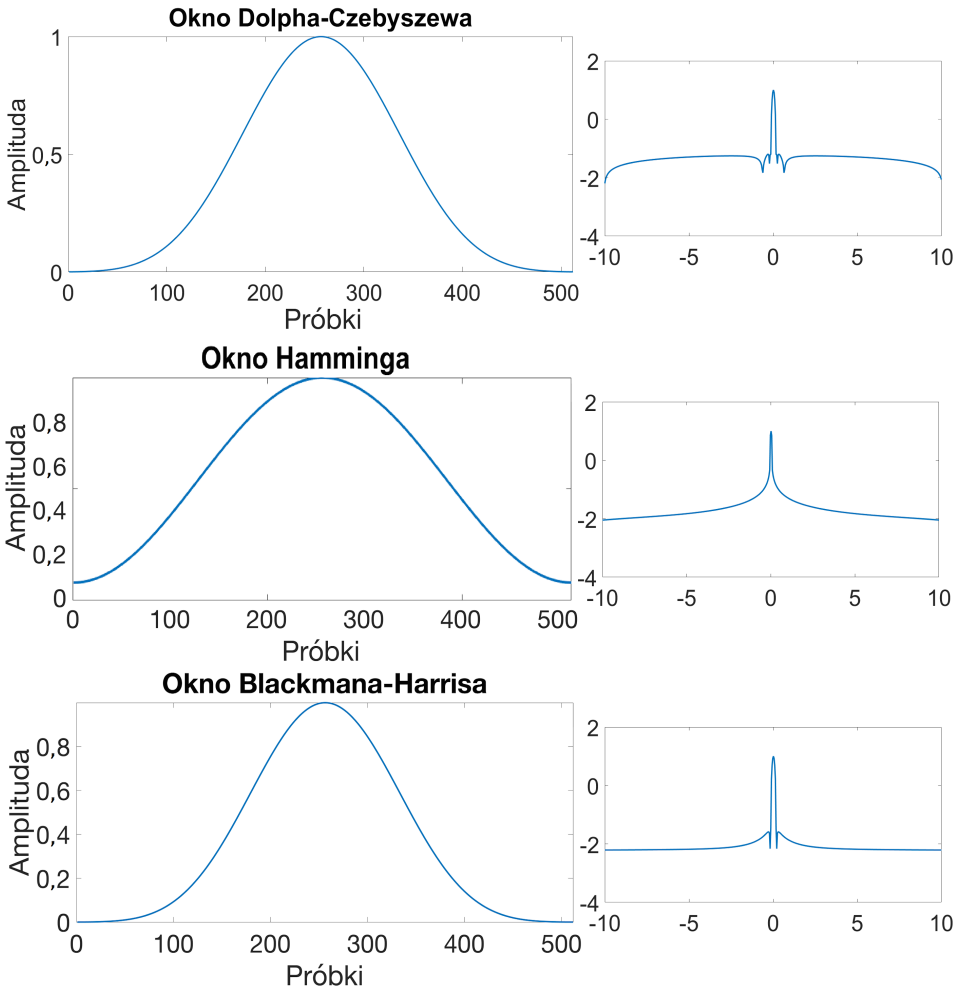
3.1.11. Okno Blackmana-Harrisa

Szczególnym uogólnieniem funkcji Hamminga jest okno Blackmana-Harrisa. Zostało opracowane w celu minimalizacji prążków bocznych poprzez przesunięcie funkcji *sinc*. Okno to jest definiowane w następujący sposób [93]:

$$w[n] = a_0 - a_1 cos\left(\frac{2\pi n}{N-1}\right) + a_2 cos\left(\frac{4\pi n}{N-1}\right) - a_3 cos\left(\frac{6\pi n}{N-1}\right), \tag{3.12}$$

gdzie:
$a_0 = 0,35875,$
$a_1 = 0,48829,$
$a_2 = 0,14128,$
$a_3 = 0,01174.$

Na ogół wybierając funkcję okna dąży się do tego, aby widmo oraz wstęgi boczne były jak najwęższe [154]. Na Rysunku 3.4. zostały przedstawione przykłady funkcji wykorzystywanych okien wraz z ich widmem Fouriera.



Rysunek 3.4. Przykłady wykorzystanych funkcji i widm okien. Opracowanie własne na podstawie [48, 93, 47]

Został zbadany wpływ powyżej opisanych funkcji okna na skuteczność identyfikacji stanu emocjonalnego mówcy. Otrzymane wyniki zostały przedstawione w podrozdziale 5.2.

3.2. Parametryzacja sygnału mowy

Reprezentacja czasowa sygnału mowy nie jest najbardziej efektywnym narzędziem do opisu mowy. Spowodowane jest to dużym rozproszeniem oraz

znaczną redundancją użytecznych informacji na wszystkie próbki sygnału [142]. W celu właściwego przebiegu procesu identyfikacji mowy, należy cechy sygnału wyodrębnić w taki sposób, aby dobrze reprezentować jego własności przy jednoczesnej redukcji wymiarowości.

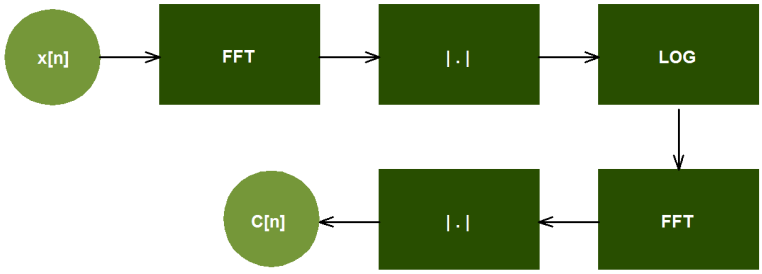
Do najpopularniejszych deskryptorów sygnału mowy można zaliczyć parametry związane ze statystycznymi wartościami sygnału. Można w tej grupie wyodrębnić takie cechy jak: częstotliwość podstawowa, energia i moc sygnału, odchylenie standardowe, mediana, częstość przejść przez zero, wartości formantów F1–F4 i inne.

Analiza cepstralna jest dość szczególna metodą parametryzacji sygnału mowy [125, 107, 7]. Początkowo badania w tym zakresie skupiały się wokół przekształceń sygnału do takich przestrzeni wektorowych, w których działania byłyby równoważne dodawaniu [154]. Dzięki temu, możliwe byłoby wykorzystanie ogólnie znanych systemów liniowych. Bazując na uogólnionej zasadzie superpozycji, metody cepstralne umożliwiają unikatową analizę rozkładu składowych addytywnych.

Jedną z pierwszych opracowanych technik cepstralnych stanowiło opracowane w 1963 roku przez B. P. Bogerta cepstrum rzeczywiste [10]. Może być ono zdefiniowane w następujący sposób [154]:

$$C[n] = |F^{-1} ln(|F(s[n])|)|, \tag{3.13}$$

gdzie:
$s[n]$ – oznacza sygnał wejściowy,
F – oznacza transformatę Fouriera.



Rysunek 3.5. Schemat blokowy wyznaczania cepstrum rzeczywistego sygnału. Opracowanie własne na podstawie [154]

We wzorze 3.13 zastosowany został logarytm modułu widma sygnału. W oparciu o powyższy mechanizm niemożliwe jest zrekonstruowanie sygnału pierwotnego, ze względu na brak informacji o fazie widma [154].

Na Rysunku 3.5. (patrz. s. 43) został przedstawiony schemat wyznaczania cepstrum rzeczywistego sygnału.

Obecnie standardem w przetwarzaniu sygnału mowy są współczynniki MFCC [70] uwzględniające procesy percepcji ludzkiego ucha. Sama procedura wyznaczania współczynników jest zbliżona do schematu przedstawionego na Rysunku 3.5. Dodatkowo należy uwzględnić przekształcenie sygnału na skalę melową. Pierwszy etap stanowi okienkowanie sygnału przy użyciu okna Hamminga. Kolejnym krokiem jest obliczenie FFT oraz podniesienie wartości prążków widma do kwadratu. Wyznaczona zostaje w ten sposób estymata gęstości widmowej sygnału. Kolejny krok stanowi uśrednienie wartości przy pomocy wagowych funkcji o trójkątnym kształcie. Kolejnym krokiem jest zlogarytmowanie estymaty oraz obliczenie transformaty kosinusowej i wyznaczenie kolejnych współczynników w oparciu o następującą równość [70]:

$$c_k = \sqrt{\frac{2}{L}} \sum_{l=0}^{L-1} ln\left(\tilde{Y}(l)\right) cos\left(\frac{\pi k}{L}\left(L + \frac{1}{2}\right)\right), \tag{3.14}$$

gdzie:
$k = 0, 1, 2, ..., q-1$, q – oznacza liczba wyznaczanych współczynników,
L – oznacza liczbę użytych filtrów,
$\tilde{Y}(l)$ – oznacza energię pasma wyznaczaną według następującego wzoru:

$$\tilde{Y}(l) = \sum_{k=1}^{n} |X_r(k)^2| H_m(k), \tag{3.15}$$

gdzie:
m – oznacza numer filtra, X_r – oznacza widmo ramki, H_m – oznacza funkcję okna Hamminga.

Kolejną szeroko stosowaną metodą w zagadnieniach rozpoznawania mowy jest liniowe kodowanie predykcyjne (LPC). Wykorzystuje ono przewidywanie wartości nowych próbek w oparciu o ważoną kombinację n próbek poprzedzających. Wagi dobierane są w taki sposób, aby średniokwadratowy błąd predykcji był jak najmniejszy [68, 71]. Przybliżona wartość aktualnej próbki p_n może być wyrażona następującą równością [68]:

$$p_n = -w_1 p_{n-1} - w_2 p_{n-2} - ... - w_r p_{n-r} + e_n, \tag{3.16}$$

gdzie:
$w_1, w_2, ..w_r$ – oznaczają wagi nazywane współczynnikami predykcji,
e_n – oznacza pozostałość predykcji.

Wagi $w_1, w_2, ...w_r$ są wyznaczane poprzez minimalizację pozostałości predykcji uśrednioną dla N próbek [68]:

$$E : \sum_{n=1}^{N} e_n^2 = \sum_{n=1}^{N} \left(\sum_{k=0}^{r} w_k p_{n-k} \right)^2, w_0 = 1. \tag{3.17}$$

Minimalizacja pozostałości predykcji winna odbywać się poprzez porównanie poszczególnych pochodnych cząstkowych do zera [71]:

$$\frac{\partial E}{\partial p_m} = \sum_{n=1}^{N} 2 p_{n-m} \sum_{k=0}^{r} w_k p_{n-k} = 0, m \in \langle 1, r \rangle, \tag{3.18}$$

W przypadku odwrócenia porządku sumowania otrzymujemy:

$$\sum_{k=0}^{r} S_{mk} w_k = 0, \tag{3.19}$$

gdzie:

$$S_{mk} = \sum_{n=1}^{N} p_{n-m} p_{n-k}. \tag{3.20}$$

Wyznaczone w powyższy sposób współczynniki korelacji S_{mk} służą do obliczenia współczynników predykcji [68]:

$$S_{m0} = - \sum_{k=1}^{r} S_{mk} w_k. \tag{3.21}$$

Powyższa równość w notacji macierzowej przyjmuje następującą postać:

$$\boldsymbol{s_0} = - \left(S * \boldsymbol{w} \right), \tag{3.22}$$

gdzie:
$\boldsymbol{w}, \boldsymbol{s_0}$ – wektory,
S – macierz kwadratowa zawierająca elementy S_{mk}, o rozmiarze r.
Ostatecznie odwracając macierz otrzymujemy:

$$w = -S^{-1} S_0. \tag{3.23}$$

Udowodnione zostało [24], że najkorzystniej jest wyznaczać 12 współczynników LPC. Scharakteryzowane powyżej współczynniki zostały wy-

korzystane do udowodnienia separowalności danych szczegółowo opisanej w rozdziale 5.1.

3.3. Klasyfikatory

Klasyfikacja jest ostatnim etapem przetwarzania sygnału wyszczególnionym na Rysunku 3.1. (patrz. s. 35). W przypadku sygnału mowy do tego rodzaju zadań najczęściej są stosowane takie narzędzia jak: algorytm k-NN [11, 70, 68, 102], maszyna wektorów wspierających [57, 108, 162], metoda ukrytych modeli Markowa [99, 110] czy sztucznych sieci neuronowych. W przeprowadzonych badaniach został wykorzystany ostatni z nich.

Obecnie sztuczne sieci neuronowe są stosowane w dziedzinach, w których klasyczne podejście algorytmiczne się nie sprawdziło, bądź jest niewystarczające. Już w pierwszej połowie dwudziestego wieku naukowcy starali się poznać mechanizmy rządzące zachowaniami i umiejętnościami ludzi i zwierząt. Zadziwiającą zdolnością mózgu jest umiejętność przeprowadzania trafnych dedukcji w oparciu o wiedzę nieprecyzyjną czy niekompletną. Pomimo, że układ nerwowy składa się ze stosunkowo prostych w budowie struktur nerwowych, to jego możliwości są ogromne. Można zatem przyjąć, iż to nie budowa pojedynczych komórek leży u podstaw możliwości tego organu, a sposób połączeń pomiędzy neuronami. Powyższa obserwacja legła u podstaw prac związanych z opracowaniem matematycznego modelu samej komórki nerwowej, jak i całej sztucznej sieci neuronowej. Najogólniej mianem sztucznej sieci neuronowej jest określana grupa połączonych oraz współpracujących ze sobą prostych elementów obliczeniowych ukierunkowanych na przetwarzanie danych. Istotność połączenia pomiędzy poszczególnymi komórkami jest wyrażona poprzez wagę, która może zostać zmodyfikowana w procesie uczenia sieci.

Przewaga SSN nad innymi algorytmicznymi podejściami polega na możliwości generalizacji. Poprawnie nauczona sztuczna sieć neuronowa potrafi poprawnie klasyfikować nieznane wcześniej dane. W najogólniejszym podejściu możemy wyróżnić trzy podstawowe rodzaje SSN: jednokierunkowe, rekurencyjne oraz samo-organizujące. Każda z nich jest charakteryzowana przy pomocy trzech parametrów: modelu neuronu, sposobu połączeń oraz metody uczenia.

3.3.1. Model neuronu McCullocha-Pittsa

Pierwsza definicja sztucznej komórki nerwowej została zaproponowana w 1943 roku przez Warrna McCullocha oraz Waltera Pittsa. Zaproponowany model zakładał, iż do każdego neuronu docierać będą sygnały: $x_1, x_2, ..., x_n$

o wartościach 0 lub 1. Docierające impulsy mogą pochodzić zarówno ze środowiska wewnętrznego jak i zewnętrznego. Ponadto, z każdym wejściem została powiązana pewna jego waga $w_1, w_2, .., w_n$ świadcząca o wpływie danego wejścia na pracę całej komórki. Idea działania polega na obliczeniu sumy ważonej, która to następnie staje się argumentem dla funkcji aktywacji neuronu. Reguła aktywacji neuronu McCullocha-Pittsa jest zdefiniowana następującą równością [1]:

$$y = \begin{cases} 1, \ gdy \sum\limits_{i=1}^{n} x_i * w_i \geqslant T \\ 0, \ gdy \sum\limits_{i=1}^{n} x_i * w_i < T, \end{cases} \tag{3.24}$$

gdzie:
y – oznacza wyjście neuronu,
T – oznacza wartość progową neuronu, po osiągnięciu której komórka nerwowa zaczyna przewodzić impulsy.

Pomimo prostoty w budowie potencjał neuronu McCullocha-Pittsa jest olbrzymi. Przy odpowiednim doborze wag oraz progów możliwa jest realizacja takich funkcji logicznych jak OR, NOR, NOT czy NAND.

3.3.2. Funkcje aktywacji nauronów

W podrozdziale 3.3.1 zostało wspomniane, iż to funkcja aktywacji decyduje o działaniu sztucznej komórki nerwowej. Przy jej pomocy określany jest poziom pobudzenia neuronu. W najogólniejszym przykładzie funkcja aktywacji może mieć następującą postać [30]:

$$y = \phi(\xi), \tag{3.25}$$

gdzie:
ξ – oznacza łączne pobudzenie neuronu wraz z biasem.
Klasyczny model perceptronu przyjmuje, iż funkcja aktywacji ma postać progową, zdefiniowaną następująco [91]:

$$\phi(\xi) = \begin{cases} 1, \ gdy \ \xi \geqslant 0 \\ 0, \ gdy \ \xi < 0, \end{cases} \tag{3.26}$$

Pomimo swej prostoty zdefiniowana powyżej funkcja znajduje zastosowanie we wszelkiego rodzaju problemach decyzyjnych związanych z określeniem przynależności do zbiorów. Bazując na logice matematycznej można traktować sygnału $y = 0$ jako fałsz, z kolei $y = 1$ jako prawdę.

Jednakże nieliniowe sieci neuronowe wymagają użycia bardziej wyrafinowanych postaci funkcji aktywacji. Jedną z częściej stosowanych jest sigmoidalna funkcja aktywacji neuronów, zdefiniowana równością [49]:

$$y = \frac{1}{1 + exp(-\beta\xi)}, \tag{3.27}$$

gdzie:
β – jest współczynnikiem stromości.

Nieliniowa zależność pomiędzy sumarycznym pobudzeniem neuronu a sygnałem wejściowym jest również definiowana przez przyjmującą następującą postać funkcję tangensa hiperbolicznego [73]:

$$y = tanh(\beta\xi) = \frac{exp(\beta\xi) - exp(-\beta\xi)}{exp(\beta\xi) + exp(-\beta\xi)}. \tag{3.28}$$

Skutkiem użycia tangensa hiperbolicznego jako funkcji aktywacji jest otrzymanie wyników z otwartego przedziału $(-1, 1)$. Uzyskanie przedziału zamkniętego obustronnie wymaga zastosowania funkcji sinus, a precyzyjniej mówiąc fragmentu sinusoidy połączonego z asymptotami rozciągającymi dziedzinę funkcji [144]:

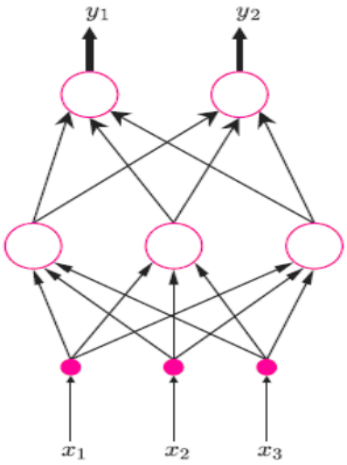
$$y = \begin{cases} -1, \ gdy \ \xi < -\frac{\pi}{2} \\ sin(\beta\xi), \ gdy \ -\frac{\pi}{2} \leqslant \xi \leqslant \frac{\pi}{2} \\ 0, \ gdy \ \xi > \frac{\pi}{2}. \end{cases} \tag{3.29}$$

Wyniki uzyskane przy użyciu wyżej opisanych funkcji zostaną zaprezentowane w rozdziale 5.

3.3.3. Jednokierunkowe sztuczne sieci neuronowe

Jednokierunkowe sztuczne sieci neuronowe (FANN) (ang. Feedforwarded Artificial Neural Networks) są bardzo często stosowane w zagadnieniach klasyfikacyjnych. Ich ogólny schemat został przedstawiony na Rysunku 3.6. Cechą charakterystyczną tej topologii jest brak sprzężenia zwrotnego[8], zatem przez każdy neuron w każdym cyklu (uczenia/przetwarzania) sygnał przechodzi dokładnie jeden raz.

[8] Informacje o wynikach uzyskanych przez sieć są używane w kolejnych iteracjach.



Rysunek 3.6. Schemat jednokierunkowej SSN

Sieć zbudowana z jednej warstwy neuronów, może reprezentować najprostszą, jednokierunkową SSN. Jeżeli założymy, że zestaw impulsów dla każdego z neuronów będzie taki sam, wówczas odpowiedź sztucznej komórki nerwowej jest definiowana następująco [144]:

$$y^{(m)} = \boldsymbol{X} * \boldsymbol{W}^{(m)} = \sum_{i=1}^{n} x_i * w_i^{(m)}, \tag{3.30}$$

gdzie:
$\boldsymbol{X} =< x_1, x_1, ..., x_n >$ – oznacza wektor wejściowy,
$\boldsymbol{W}^{(m)} =< \boldsymbol{w}_1^{(m)}, \boldsymbol{w}_2^{(m)}, ..., \boldsymbol{w}_n^{(m)} >$ – oznacza wektor wag,
m – oznacza numer neuronu.
Tak opisana sieć będzie w stanie poprawnie identyfikować k klas obiektów. Dzieje się tak, ponieważ wektor wag każdego z neuronów jest w stanie „zapamiętać" jeden z wzorców. Uogólniając, za pomocą równości [144]:

$$Y = X * W_k, \tag{3.31}$$

Y – odpowiedź sieci,

$$\boldsymbol{W_k} = \begin{pmatrix} w_1^{(1)} & w_2^{(1)} & \dots & w_n^{(1)} \\ w_1^{(2)} & w_2^{(2)} & \dots & w_n^{(2)} \\ \vdots & \vdots & \vdots & \ddots \\ w_1^{(k)} & w_2^{(k)} & \dots & w_n^{(k)} \end{pmatrix}$$

możemy opisać kompletny zestaw funkcji realizowanych przez sieć. Opisany powyżej wzór pozwala liniowo odwzorować sygnał $X \in R^n$ w sygnał $Y \in R^k$. Taka transformacja ma szczególne znaczenie praktycznie. Należy pamiętać, iż znaczna część transformacji związanych z przetwarzaniem sygnałów (w tym sygnałów mowy) ma postać liniową [144].

3.3.4. Sieci Kohonena

Pierwsze wzmianki na temat samo-organizujących się sieci datowane są na lata 70-te dwudziestego wieku. Znaczący wpływ na ich rozwój wywarł fiński badacz Teuvo Kohonen, stąd też ten rodzaj SSN często określany jest mianem sieci Kohonena. Do nauki tego rodzaju topologii wykorzystywane jest tak zwane uczenie konkurencyjne [27]. Idea procesu nauki polega na wytworzeniu wzorców wyjścia na podstawie sygnałów wejściowych. Na początku procesu nie istnieją wzorce wyjściowe, znane są jedynie wejścia. Zazwyczaj sieci Kohonena przyjmują formę sieci jednokierunkowych, w których dany jest n wymiarowy wektor wejściowy X, którego elementy połączone są z każdym neuronem. Wektor wag jest zdefiniowany następująco: $W_i =< w_{i1}, w_{i2}, ..., w_{iN} >^T$. Wektory wejściowe stanowią próbę uczącą, podobnie jak w przypadku zwykłych sieci rozpatrywaną w pętli podczas budowy mapy. Proces normalizacji wartości wektora wejściowego poprzedza proces uczenia, co może być zdefiniowane [145]:

$$x_i = \frac{x_k}{\sqrt{\sum_{k=1}^{n} (x_k)^2}}. \tag{3.32}$$

Współzawodnictwo polega na pobudzeniu sieci przy użyciu wektora X.

Zwycięża neuron, którego wagi w najmniejszym stopniu różnią się od odpowiednich składowych wektora wejściowego. Zwycięski, *w-ty* neuron spełnia następującą relację [151]:

$$d(x, w_w) = \min_{1<i<n} d(x, w_i), \tag{3.33}$$

gdzie:
$d(x, w)$ – oznacza odległość pomiędzy wektorami[9].

Kolejnym krokiem jest wyznaczenie wokół zwycięskiej komórki topologicznego sąsiedztwa $S(n)$, którego promień zmniejsza się wraz z upływem czasu. Następnie ma zastosowanie reguła Kohonena, wedle której neuron

[9] W sensie wybranej metryki.

zwycięski oraz te, które znajdują się w jego sąsiedztwie podlegają procesowi adaptacji [145]:

$$w_i(n+1) = w_i(n) + \eta_i(n)[x - w_i(n)], \tag{3.34}$$

gdzie:
$\eta_i(n)$ jest współczynnikiem *i–tej* komórki z sąsiedztwa $S_w(n)$ w chwili k.
Można zauważyć, że wagi współczynników nienależących do sąsiedztwa nie ulegają zmianom, a jego wartość w obrębie sąsiedztwa maleje wraz ze wzrostem odległość od zwycięskiego neuronu. Klasyczny algorytm Kohonena przyjmuje następującą postać [145]:

$$w_i(n+1) = w_i(n) + \eta G(i,x)[x - w_i(n)], \tag{3.35}$$

gdzie:
$G(i,x)$ jest funkcją sąsiedztwa definiowaną poniższym równaniem:

$$G(i,x) = \begin{cases} 1, \ dla \ d(i,w) \ \leqslant \ \lambda \\ 0, \ dla \ d(i,w) \ > \ \lambda, \end{cases} \tag{3.36}$$

gdzie:
$d(i,w)$ stanowi odległość euklidesową pomiędzy neuronem zwycięskim w, a *i-tym* neuronem.

Sąsiedztwo zdefiniowane powyższą równością stanowi tak zwane sąsiedztwo prostokątne. Lepsze rezultaty uczenia pozwala osiągnąć wykorzystanie sąsiedztwa w sensie Gaussa [151, 145, 66], zdefiniowanego następująco [151]:

$$G(i,x) = exp\left(-\frac{d^2(i,w)}{2\lambda^2}\right). \tag{3.37}$$

Sąsiedztwo gaussowskie wymusza lepszą organizację sieci, spowodowaną większym zróżnicowaniem stopnia adaptacji neuronów.

3.3.5. Metody uczenia sztucznych sieci neuronowych

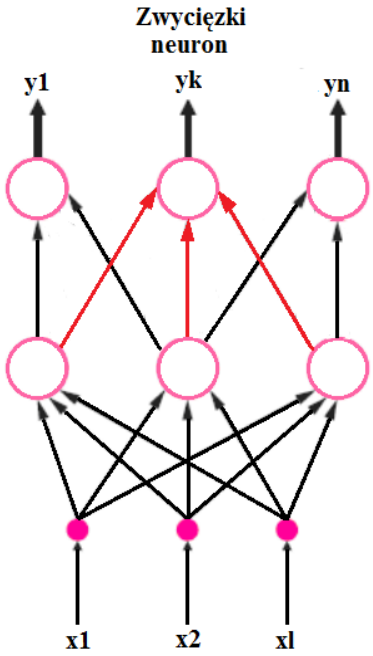
Proces uczenia SSN może przebiegać na dwa sposoby. Pierwszy, związany jest z brakiem informacji na temat oczekiwanych wartości na wyjściu sieci. Jest to tak zwane uczenie nienadzorowane (ang. unsupervised learning). Drugim sposobem jest uczenie sieci z wykorzystaniem nauczyciela (ang. supervised learning). Najczęściej wykorzystywanym w drugim przypadku sposobem nauki jest algorytm wstecznej propagacji błędów.

Uczenie nienadzorowane

Ten rodzaj nauki jest spowodowany brakiem informacji na temat pożądanej odpowiedzi sieci. SSN nie posiada informacji na temat poprawności bądź niepoprawności otrzymanej odpowiedzi, dlatego też wymagana jest każdorazowa analiza reakcji na pobudzenie. Uczenie bez nauczyciela jest możliwe tylko wówczas, gdy wygnały wejściowe posiadają pewne właściwości, takie jak [145]:

- podobieństwo do archetypów – wzorców,
- liczba neuronów składowych powinna przekraczać liczbę wzorców,
- początkowa różnorodność preferencji komórek nerwowych.

Wśród najpopularniejszych metod uczenia nienadzorowanego występuje reguła WTA (ang.Winner Takes All). Jej schemat został przedstawiony na Rysunku 3.7.



Rysunek 3.7. Schemat reguły WTA.

Idea metody WTA sprowadza się do następujących założeń [105]:

- współzawodniczące neurony otrzymują na wejściu identyczne zestawy sygnałów x_j,
- różnice w wyjściu y_j są implikacją różnicy wag poszczególnych neuronów,

- zwycięża neuron najmniej różniący się wagami od aktualnie uczącego wektora,
- aktualizacja wagi występuje wyłącznie dla zwycięskiej komórki nerwowej.

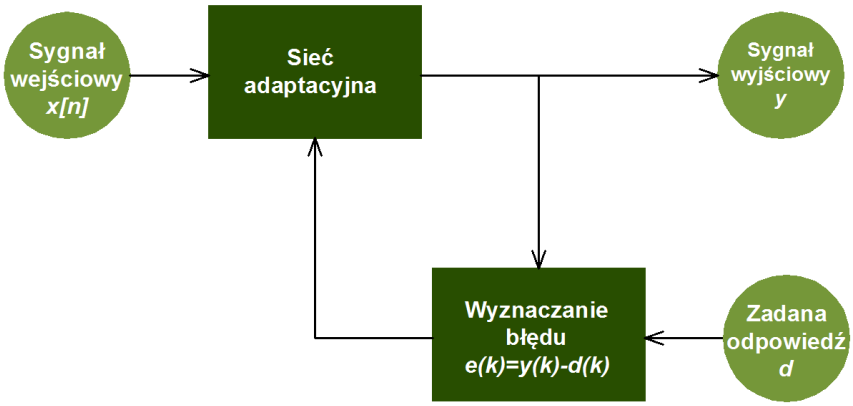
Jedną z wad metody WTA jest możliwość wystąpienia tak zwanych neuronów martwych, którym nigdy nie udało się odnieść zwycięstwa. Prowadzi to do ogólnego spadku efektywność czynnych komórek nerwowych powodujący wzrost globalnego błędu odwzorowań sygnałów.

Uczenie nadzorowane

Proces uczenia nadzorowanego wymaga przygotowania dwóch zestawów danych. Pierwszy, stanowi zbiór, którego zadaniem jest charakterystyka rozważanego problemu zwany jest wektorem uczącym. Drugi zbiór stanowią wektory, zawierające poprawne odpowiedzi na określone pobudzenia sieci. Proces uczenia polega na porównywaniu wartości otrzymanych na wyjściu SSN, z wartościami zawartymi w wektorze oczekiwanych wyników. W przypadku niepoprawności, wyznaczany jest błąd, którego wartość jest przekazywana do wszystkich neuronów. W przypadku uczenia nadzorowanego najczęściej stosowany jest algorytm wstecznej propagacji błędów [144, 145, 157]. Błąd sieci służy do korekcji wag każdego z neuronów. Algorytm kończy swoje działanie w przypadku gdy błąd epoki[10] będzie mniejszy od zadanego lub zostanie osiągnięta zadana liczba epok. Ogólny schemat uczenia z nauczycielem został przedstawiony na Rysunku 3.8. (patrz. s. 54), gdzie dla każdej pary uczącej $< y(k), d(k) >$ wektor błędu jest definiowany w następujący sposób: $e(k) = y(k) - d(k)$, gdzie $y(k)$ – oznacza aktualną odpowiedź sieci na wymuszenie w postaci wektora $x(k)$, a $d(k)$ – oznacza zadany wektor wyjściowy.

Istotą uczenia sztucznych sieci neuronowych jest modyfikacja wag połączeń pomiędzy poszczególnymi neuronami. Zmiana ta ma na celu minimalizację funkcji błędu (kosztu). Powszechnie stosowanym algorytmem uczącym jest algorytm wstecznej propagacji błędów pozwalający na zmianę wag w dowolnej FANN. W trakcie procesu uczenia na wejścia sieci podawane są kolejne wzorce uczące wraz z informacją o oczekiwanych wynikach. Nauka polega na aktualizacji wag poszczególnych neuronów w kolejności odwrotnej niż propagowany sygnał wejściowy. Proces jest powtarzany do momentu osiągnięcia błędu sieci mniejszego od zadanego.

[10] Przetworzenia całego zestawu uczącego.



Rysunek 3.8. Schemat uczenia nadzorowanego

Algorytm ten jest zaliczany do metod gradientowych, dlatego też wymagane jest aby funkcja błędu oraz funkcja aktywacji były różniczkowalne [60]. Uczenie sieci neuronowej polega na inicjalizacja sieci, obliczeniu wartości wyjściowej sieci na podstawie danych oraz wyznaczeniu minimalnej wartości funkcji błędu. Jedną z najpowszechniejszych miar jest funkcja błędu średniokwadratowego Q [9]:

$$Q = \frac{1}{2} \sum_{i=1}^{P} \sum_{j=1}^{K} \left(d_{i,j} - y_{i,j} \right)^2, \tag{3.38}$$

gdzie:
P – oznacza liczbę par uczących,
K – oznacza liczbę wyjść sieci, $y_{i,j}$ – oznacza wartość otrzymaną dla *i-tej* pary uczącej na *j-tym* wyjściu sieci,
$d_{i,j}$ – oznacza wartość oczekiwaną dla *i-tej* pary uczącej na na *i-tym* wyjściu.

Minimalizacja błędu średniokwadratowego opiera się o regułę najszybszego spadku gradientu:

$$\omega_{ij}^{(k)}(n+1) = \omega_{ij}^{(k)}(n) + \eta \left(\nabla_{ij}^{(k)}(n) \right), \tag{3.39}$$

gdzie:
η – oznacza współczynnik uczenia sieci, $\eta \in [0, 1]$,
$\nabla_{ij}^{(k)}$ – oznacza składową gradientu funkcji błędu definiowaną następująco:

$$\nabla_{ij}^{(k)}(n) = \frac{\partial Q(n)}{\partial \omega_{ij}^{(k)}(n)} = \frac{\partial Q(n)}{\partial s_i^{(k)}(n)} \frac{\partial s_i^{(k)}(n)}{\partial \omega_{ij}^{(k)}(n)} = \frac{\partial Q(n)}{\partial s_i^{(k)}(n)} x_j^{(k)}, \tag{3.40}$$

gdzie:
$s_i^{(k)}$ – wartość wzbudzenia[11],
$x_j^{(k)}$ – wartość pobudzenia otrzymaną na *j-tym* wejściu *k-tej* warstwy.

Oznaczając przez [9]:

$$\delta_i^{(k)}(n) = -\frac{1}{2}\frac{\partial Q(n)}{\partial s_i^{(k)}(n)} \tag{3.41}$$

otrzymujemy:

$$\frac{\partial Q(n)}{\partial \omega_{ij}^{(k)}(n)} = -2\delta_i^{(k)} x_j^{(k)}. \tag{3.42}$$

Wówczas równość wyrażona wzorem 3.39 przyjmuje postać:

$$\omega_{ij}^{(k)}(n+1) = \omega_{ij}^{(k)}(n) + 2\eta\delta_i^{(k)} x_j^{(k)}. \tag{3.43}$$

Należy pamiętać, iż sposób wyznaczania $\delta_i^{(k)}$ ze wzoru 3.43 przyjmuje różną postać w zależności od warstwy sieci [9]:

- dla warstwy wyjściowej wyznaczany jest następująco:

$$\begin{aligned}\delta_i^{(L)} &= -\frac{1}{2}\frac{\partial Q(n)}{\partial s_i^{(L)}(n)} = -\frac{1}{2}\frac{\partial \sum_{m=1}^{N} \epsilon_m^{(L)^2}}{\partial s_i^{(L)}(n)} = -\frac{1}{2}\frac{\partial \epsilon_i^{(L)^2}}{\partial s_i^{(L)}(n)} \\ &= -\frac{1}{2}\frac{\partial \left(d_i^{(L)}(n) - y_i^{(L)}(n)\right)^2}{\partial s_i^{(L)}(n)} = -\epsilon_i^{(L)}(n)\frac{\partial y_i^{(L)}(n)}{\partial s_i^{(L)}(n)} \\ &= -\epsilon_i^{(L)}(n) f'\left(s_i^{(L)}(n)\right).\end{aligned} \tag{3.44}$$

- dla pozostałych warstw:

$$\begin{aligned}\delta_i^{(k)} &= -\frac{1}{2}\frac{\partial Q(n)}{\partial s_i^{(k)}(n)} = -\sum_{m=1}^{N_{k+1}}\left(\frac{\partial Q(n)}{\partial s_m^{(k+1)}(n)}\frac{\partial s_m^{(k+1)}(n)}{\partial s_i^{(k)}(n)}\right) \\ &= -\sum_{m=1}^{n_{k+1}}\left(\delta_i^{(k+1)}(n)\,\omega_{mi}^{(k+1)} f'\left(s_i^{(L)}(n)\right)\right) \\ &= -f'\left(s_i^{(L)}(n)\right)\sum_{m=1}^{n_{k+1}}\left(\delta_i^{(k+1)}(n)\,\omega_{mi}^{(k+1)}\right).\end{aligned} \tag{3.45}$$

[11] Suma wartości wejściowych pomnożonych przez wagi.

Z powyższych równań wynika, iż wyznaczanie $\delta_i^{(k)}$ będzie odbywać się w przeciwnej kolejności niż propagowany jest sygnał wejściowy.

Zatem dla *i-tego* neuronu w *k-tej* warstwie błąd może być określony w następujący sposób:

$$\epsilon_i^{(k)} = \sum_{m=1}^{N_{k+1}} \left(\delta_i^{(k+1)}(n)\, \omega_{mi}^{(k+1)}(n) \right). \tag{3.46}$$

Wykorzystując we wzorze 3.45 wyrażenie 3.46 otrzymujemy:

$$\delta_i^{(k)}(n) = \epsilon_i^{(k)}(n)\, f'\left(s_i^{(k)}(n) \right). \tag{3.47}$$

Jeżeli funkcję aktywacji neuronu stanowi funkcja różniczkowalna wówczas algorytm wstecznej propagacji błędów można zapisać w następującej postaci [9] (przy założeniu początkowej losowości wag):

$$\epsilon_i^{(k)}(n) = \begin{cases} \epsilon_i^{(L)}(n), \ dla\ k = L \\ \sum\limits_{m=1}^{N_{k+1}} \delta_m^{(k+1)}(n)\, \omega_{mi}^{(k+1)}(n), \ dla\ k = 1, 2, ..., L-1. \end{cases} \tag{3.48}$$

$$\delta_i^{(k)}(n) = \epsilon_i^{(k)} f'\left(s_i^{(k)}(n) \right), \tag{3.49}$$

$$\omega_{ij}^{(k)}(n+1) = \omega_{ij}^{(k)}(n) + 2\eta \delta_i^{(k)} x_j^{(k)}. \tag{3.50}$$

Wśród najczęściej spotykanych modyfikacji algorytmu wstecznej propagacji błędów są:

- algorytm wstecznej propagacji z adaptacyjną zmianą współczynników uczenia,
- algorytm wstecznej propagacji z momentum,
- algorytm wstecznej propagacji z adaptacyjną zmianą współczynników uczenia i momentum,
- algorytm wstecznej propagacji skalowalnego gradientu.

Wyniki otrzymane przy wykorzystaniu opisanych w niniejszym rozdziale topologii sieciowych i metod nauczania zostaną szczegółowo zaprezentowane i omówione w rozdziale 5.

3.4. Wnioski do rozdziału

1. Funkcja aktywacji neuronu jest bezpośrednio związana z efektywnością sieci neuronowej, jak również jej możliwymi zastosowaniami.

2. Nie można jednoznacznie stwierdzić, która z funkcji aktywacji zagwarantuje najwyższą skuteczność klasyfikacji.
3. Czasowa reprezentacja sygnału mowy nie definiuje jego efektywnej charakterystyki, gdyż nie daje możliwości oceny jego pozostałych parametrów pozwalających na skuteczną klasyfikację.
4. Deskrypcja sygnału mowy powinna odbywać się w oparciu o takie właściwości jak: wartości statystyczne sygnału, formanty, współczynniki MFCC oraz LPC.
5. Do współcześnie stosowanych klasyfikatorów sygnału mowy można zaliczyć: sztuczne sieci neuronowe, algorytm k-NN, metodę ukrytych modeli Markova czy maszynę wektorów podtrzymujących.
6. Wybór właściwej topologii sieci uzależniony jest od danych, jakimi dysponujemy, jak również od rozpatrywanego zagadnienia.
7. Wybór metody uczenia sieci neuronowej zależy od jej topologii oraz dostępnych parametrów i zasobów sprzętowych.

4. Wizualizacja parametrów sygnału mowy polskiej w przestrzeni dwuwymiarowej

Postrzeganie stanów emocjonalnych jest procesem wysoce subiektywnym oraz złożonym i nader często identyczna sytuacja bądź wypowiedź może być zinterpretowana na kilka różnych sposobów. Skutkiem takiego postrzegania emocji jest poszukiwanie rozwiązań pozwalających na wskazanie występowania lub absencji stanów podstawowych oraz stopnia nasilenia każdego z nich [81]. Jest to podejście pomocne w określeniu przeważającego stanu emocjonalnego w całości wypowiedzi, ze względu na możliwą niejednoznaczność emocji.

Według Plutchika emocje podstawowe są to takie stany, które mają znaczenie adaptacyjne, pozwalając gatunkowi na przetrwanie [68]. Autor wyodrębnił osiem takich emocji: strach, smutek, złość, zaskoczenie, radość, ufność oraz wyczekiwanie. W ostatnich latach istotny wpływ na rozwój badań nad mową emocjonalną oraz zagadnieniem identyfikacji emocji miała zorganizowana w 2009 roku w Amsterdamie konferencja ACII[12], w trakcie której jedna z sesji specjalnych została poświęcona zagadnieniom rozpoznawania emocji. Konferencja ta przyczyniła się do wyodrębnienia następujących problemów bezpośrednio związanych z niejednoznacznością mowy emocjonalnej [68]:

1. Podobieństwo stanów emocjonalnych – bardzo trudno określić twarde granice pomiędzy poszczególnymi emocjami leżącymi obok siebie na kole Plutchika (Rysunek 2.2.).
2. Mieszanie się emocji – według Plutchika możliwe jest przeżywanie kilku stanów pierwotnych jednocześnie.
3. Sekwencyjne występowanie emocji – w trakcie wypowiedzi możliwe jest odczuwanie kilku, występujących po sobie stanów emocjonalnych.
4. Maskowanie emocji – próba ukrycia przed rozmówcą przeżywanego stanu emocjonalnego.
5. Konflikt ekspresji – okazywanie stanu emocjonalnego, w nieadekwatny sposób np. płacz ze szczęścia.

Dlatego też niezwykle istotną rolę odgrywa baza nagrań służąca do badań. Ponadto wielce istotne jest odpowiednie przyporządkowanie stanów emocjonalnych do zgromadzonego zbioru sygnałów mowy.

W przypadku niniejszych badań posłużono się trzema zbiorami nagrań mowy emocjonalnej zawierającymi próbki mowy w sześciu stanach emocjonalnych: strach, radość, smutek, złość, znudzenie oraz stan neutralny. Wszystkie próbki poddane zostały ocenie poprzez badanie ankietowe prze-

[12] Affective Computing and Intelligent Interaction.

prowadzone na grupie stu uczestników, którzy mieli za zadanie przyporządkowanie usłyszanej wypowiedzi do jednego z sześciu stanów.

Pierwsza baza nagrań (Baza A) został opracowana, przygotowana i udostępniona przez Zakład Elektroniki Medycznej Politechniki Łódzkiej. Jest to zbiór składający się z 240 nagrań, zawierający wypowiedzi ośmioro osób(czterech kobiet i czterech mężczyzn) zawodowo trudniących się aktorstwem. Wypowiadają oni pięć następujących sentencji:

1. Oni kupili dzisiaj nowy samochód.
2. Jego dziewczyna przylatuje dzisiaj samolotem.
3. Janek był dzisiaj u fryzjera.
4. Ta lampa dzisiaj jest na biurku.
5. Od jutra przestaję się golić.

Powyższa baza nagrań zawiera pliki w formacie '.wav' próbkowanie z częstotliwością 44,1 kHz. Poprawność nagrań oraz przynależność do poszczególnych grup emocjonalnych została potwierdzona przez pięciu ekspertów [67].

Druga baza nagrań (Baza B) została przygotowana w Instytucie Informatyki Politechniki Lubelskiej i zawiera nagrania w sześciu stanach emocjonalnych podobnie jak Baza A. Powyższy zbiór został nagrań został zarejestrowany, dzięki uprzejmość Instytutu Elektrotechniki i Elektrotechnologii Politechniki Lubelskiej, w komorze akustycznej. W badaniach brały udział osoby w wieku 20-30 lat, nie zajmujące się aktorstwem. Uczestnicy badania wypowiadali dokładnie te same sentencje, które zawiera Baza A. Cała baza, której struktura została zaprezentowana w Tabeli 4.1. składa się z 236 nagrań.

Tabela 4.1. Liczba nagrań zawartych w Bazie B

Emocja	**Liczba nagrań (kobiety)**	**Liczba nagrań (mężczyźni)**
złość	14	19
radość	13	27
smutek	10	30
stan neutralny	10	29
znudzenie	15	29
strach	14	26

Trzecia baza nagrań (Baza C) również została opracowana w Instytucie Informatyki Politechniki Lubelskiej jednak nagrania, nie były rejestrowane w komorze akustycznej a w środowisku miejskim (sale wykładowe, ulica, supermarket). W badaniach uczestniczyło trzy kobiety oraz sześciu mężczyzn w wieku 22–31 lat. Wypowiadali oni pięć następujących zdań:

1. Muszę z Tobą porozmawiać.
2. Naprawdę tak myślisz?
3. Nic nie rozumiesz.
4. Piotrek kupił nowy rower.
5. Ta paczka jest już w Krakowie.

Z powodu rejestracji poza komorą akustyczną poza właściwą treścią zawierają one również dźwięki charakterystyczne dla miejsc, w których zostały zarejestrowane (ruch uliczny, rozmowy, szumy, itp.). Cała baza, której struktura została przedstawiona w Tabeli 4.2. zawiera 193 nagrania. Znajdują się w niej nagrania w sześciu stanach emocjonalnych.

Tabela 4.2. Liczba nagrań zawartych w Bazie C

Emocja	**Liczba nagrań (kobiety)**	**Liczba nagrań (mężczyźni)**
złość	6	28
radość	10	20
smutek	12	23
stan neutralny	11	24
znudzenie	8	27
strach	6	18

W przypadku Bazy B oraz Bazy C przynależność nagrań do poszczególnych grup emocji została zweryfikowana za pomocą badań ankietowych. Zostały one przeprowadzone na grupie 95 rezpondentów w wieku 21–31 lat. W rakcie badań uczestnicy po odsłuchaniu nagrania mieli za zadanie przyporządkować je do jednej z sześciu grup identyfikowanych emocji, bądź stwierdzić, iż jednoznaczna identyfikacja nie jest możliwa.

4.1. Separowalność stanów emocjonalnych

Posiadanie samych baz danych zawierających nagrania mowy emocjonalnej jest warunkiem niewystarczającym, aby móc skutecznie identyfikować emocje w mowie w czasie rzeczywistym bądź zbliżonym do rzeczywistego. Pożądane jest jeszcze aby możliwa była separacja poszczególnych emocji. Innymi słowy należy pokazać, że z posiadanych zbiorów nagrań

możliwe będzie wyodrębnienie takich parametrów, które pozwolą na jednoznaczną klasyfikację emocji. W tym celu zastosowane zostało podejście heurystyczne. Z posiadanych nagrań wyodrębnione zostały parametry, które posłużyły później do wizualizacji poszczególnych emocji na płaszczyźnie dwuwymiarowej. Zrezygnowano tutaj z redukcji parametrów mogących mieć tylko nieznaczny wpływ na identyfikację, gdyż celem było pokazanie możliwości separowalności poszczególnych stanów emocjonalnych w zbiorach nagrań. Czas przetwarzania zbiorów nie miał większego znaczenia.

Z posiadanych baz nagrań zostały wyodrębnione zarówno parametry statystyczne sygnału, jak i współczynniki MFCC i LPC[13] oraz inne. Wyboru parametrów, wykorzystywanych w niniejszym badaniu dokonano w oparciu o dostępną literaturę [21, 35, 68, 101, 112]. Całość obliczeń i klasyfikacji została wykonana w środowisku Matlab R 2016. Zbiór wszystkich parametrów służących do separacji danych został przedstawiony w Tabeli 4.3.

Tabela 4.3. Zestawienie deskryptorów sygnału mowy wykorzystywanych w procesie separowalności stanów emocjonalnych

Grupa cech	**Opis cechy**
Energia sygnału	maksymalna wartość energii sygnału minimalna wartość energii sygnału średnia wartość energii sygnału odchylenie standardowe energii sygnału mediana energii sygnału dolny kwartyl energii sygnału górny kwartyl energii sygnału kurtoza energii sygnału współczynnik monotoniczności obwiedni średnia wartość wzrostu obwiedni średnia wartość spadku obwiedni
Współczynniki MFCC	minimalna wartość współczynników (1–12) maksymalna wartość współczynników (1–12) średnia wartość współczynników (1–12) mediana współczynników (1–12) odchylenie standardowe współczynników (1–12)

[13] Po 12 współczynników każdego typu.

Grupa cech	**Opis cechy**
Współczynniki LPC	minimalna wartość współczynników (1–12) maksymalna wartość współczynników (1–12) średnia wartość współczynników (1–12) mediana współczynników (1–12) odchylenie standardowe współczynników (1–12)
Częstotliwość podstawowa F0	średnia wartość F0 maksymalna wartość F0 minimalna wartość F0 odchylenie standardowe F0 zakres F0 dolny kwartyl F0 górny kwartyl F0 kurtoza F0 współczynnik monotoniczności F0 średnia wartość wzrostu F0 minimalna wartość wzrostu F0 maksymalna wartość wzrostu F0 średnia wartość spadku F0 minimalna wartość spadku F0 maksymalna wartość spadku F0
Formanty F1–F4	średnia wartość F1 maksymalna wartość F1 minimalna wartość F1 odchylenie standardowe F1 średnia wartość F2 maksymalna wartość F2 minimalna wartość F2 odchylenie standardowe F2 mediana F2 średnia wartość F3 maksymalna wartość F3 minimalna wartość F3 odchylenie standardowe F3 średnia wartość F4 maksymalna wartość F4 minimalna wartość F4 odchylenie standardowe F4 mediana F0, F1, F2, F3, F4

W sumie dla każdego z nagrań, każdej z baz danych, zostało wyekstrahowane 162 parametry opisujące sygnał: 11 własności związanych z energią sygnału, po 60 charakterystyk dla współczynników MFCC oraz LPC i 31 parametrów w oparciu w częstotliwość podstawową oraz formanty. Wszystkie wymienione wyżej parametry zostały przetworzone z wykorzystaniem metody tSNE pozwalającej na wizualizację dużych danych wielowymiarowych na płaszczyźnie oraz w przestrzeni [147].

4.2. Wizualizacja parametrów sygnału

Wizualizacja danych wielowymiarowych jest istotnym problemem w wielu zagadnieniach nauki, w których konieczna jest ich prezentacja w przestrzeni dwu lub trzy wymiarowej. Zostało opracowane wiele technik służących powyższemu zagadnieniu. Warto tutaj wspomnieć o metodach zaproponowanych przez Chernoffa [19], Keima [75] czy Weinbergera [157]. W niniejszych badaniach wizualizacja danych została oparta o metodę t-SNE (ang. t-Distributed Stochastic Neighbor Embedding) opracowaną przez Geoffreya Hinton and Laurensa van der Maatena [147]. Podstawą zaproponowanej metody stanowi SNE[14] polegająca na zamianie odległości euklidesowej, w przestrzeni wielowymiarowej, pomiędzy dwoma punktami w warunkowe prawdopodobieństwo reprezentacji ich podobieństwa. Podobieństwo punktu x_j do punktu x_i wyrażane jako warunkowe prawdopodobieństwo $p_{j|i}$. Oznacza to, że punkt x_j zostanie oznaczony jako sąsiad punktu x_i jeśli sąsiedzi będą wybierani proporcjonalnie do gęstości rozkładu Gaussa skupionej wokół punktu x_i. Dla punktów występujących blisko siebie w przestrzeni wielowymiarowej $p_{j|i}$ będzie miało relatywnie wysoką wartość. Dla punktów oddalonych od siebie wartość ta będzie dążyć do nieskończoności. Matematycznie warunkowe prawdopodobieństwo $p_{j|i}$ jest definiowane następująco [147]:

$$p_{j|i} = \frac{exp\left(\frac{-||x_i - x_j||^2}{2\sigma_i^2}\right)}{\sum_{k \neq i} exp\left(\frac{-||x_i - x_k||^2}{2\sigma_i^2}\right)}, \tag{4.1}$$

gdzie:
σ – jest odchyleniem standardowym rozkładu Gaussa wokół punktu x_i.
Wartość dla prawdopodobieństwa $p_{i|i}$ została określona jako 0, ponieważ interesuje nas podobieństwo punktu do punktów występujących w jego sąsiedztwie. W analogiczny sposób możliwe jest również wyznaczenie praw-

[14] *Stochastic Neighbor Embedding.*

dopodobieństwa ($q_{j|i}$) dla punktów y_j oraz y_i będącymi odpowiednikami punktów x_j oraz x_i w przestrzeni małowymiarowej [147]:

$$q_{j|i} = \frac{exp\left(-||y_i - y_j||^2\right)}{\sum_{k \neq i} exp\left(-||y_i - y_k||^2\right)}. \tag{4.2}$$

Podobnie jak poprzednio $q_{i|i}$ wynosi 0. Jeżeli punkty y_j oraz y_i dokładnie przekształciły punkty x_j i x_i wówczas warunkowe prawdopodobieństwo $p_{j|i}$ będzie równe $q_{j|i}$. Zadaniem zatem SNE jest minimalizacja niedopasowania pomiędzy $p_{j|i}$ i $q_{j|i}$. Naturalną miarą dokładności odwzorowania $p_{j|i}$ w $q_{j|i}$ wydaje się być dywergencja Kullbacka-Leiblera, która w tym przypadku jest równa stosunkowi entropii do stałej addytywnej [147]. SNE minimalizuje odległość Kullbacka–Leiblera za pomocą metody spadku gradientu. Funkcja kosztu (C) wyrażona jest następująco [148]:

$$C = \sum_i KL\left(P_i||Q_i\right) = \sum_i \sum_j p_{j|i} log_2 \frac{p_{j|i}}{q_{j|i}}, \tag{4.3}$$

gdzie:
P_i – oznacza rozkład prawdopodobieństwa warunkowego dla sąsiedztwa punktu x_i,
Q_i – oznacza rozkład prawdopodobieństwa warunkowego dla sąsiedztwa punktu y_i,
Ponieważ dywergencja Kullbacka-Leiblera nie jest funkcją symetryczną, koszty przekształceń poszczególnych par nie są równe.

Wartość odchylenia standardowego rozkładu Gaussa σ_i jest wyznaczana dla każdego z punktów x_i w przestrzeni wielowymiarowej. Można stwierdzić, że nie istnieje jedna jej wartość dla wszystkich punktów w danym zbiorze, ponieważ gęstość danych może się różnić [147]. W gęstszych rejonach mniejsza wartość σ_i zazwyczaj będzie bardziej odpowiednia. Poszczególne wartość funkcji σ_i wynikają z rozkładu prawdopodobieństwa P_i wyznaczonego dla wszystkich punktów przynależnych do rozpatrywanego zbioru. Wyznaczony rozkład posiada oczywiście pewien poziom entropii, który zmienia się wraz ze zmianami wartości funkcji σ_i. SNE dokonuje binarnego przeszukiwania wartości funkcji σ_i, wraz ze stałym współczynnikiem niepewności, zdefiniowanym następująco [147]:

$$Perp(P_i) = 2^{H(P_i)}, \tag{4.4}$$

gdzie:
$H(P_i)$ – oznacza entropię Shanona zbioru P_i wyrażoną w bitach:

$$H(P_i) = -\sum_j p_{j|i} log_2 p_{j|i}. \tag{4.5}$$

Wspomniana niepewność ($Perp$) może być interpretowana jako efektywna liczba sąsiadów. Typowe jego wartości wynoszą od 5 do 50 [148]. Minimalizacja funkcji kosztu zdefiniowana za pomocą równości 4.3 wyznaczana jest przy użyciu metody spadku gradientu [148]:

$$\frac{\delta C}{\delta y_i} = 2 \sum_j \left(p_{j|i} - q_{j|i} + p_{i|j} - q_{i|j} \right) (y_i - y_j). \tag{4.6}$$

Spadek gradientu jest określany poprzez dodanie aktualnie wyznaczonej wartości gradientu do, zanikającej wykładniczo, sumy poprzednio wyznaczonych wartości, w celu określania zmiany w układzie współrzędnych. W celu poprawy wyników oraz aby uniknąć niskich wartości minimów lokalnych, relatywnie duża wartość współczynnika momentum jest dodawana do wartości gradientu [148]. Matematycznie zmiana wartości gradientu γ jest określana następująco [147]:

$$\gamma^{(t)} = \gamma^{(t-1)} + \eta \frac{\delta C}{\delta \gamma} + \alpha(t) \left(\gamma^{(t-1)} - \gamma^{(t-2)} \right), \tag{4.7}$$

gdzie:
$\gamma^{(t)}$ – oznacza wynik dla iteracji o numerze t,
η – oznacza współczynnik uczenia,
$\alpha(t)$ – oznacza współczynnik momentum dla iteracji o numerze t.

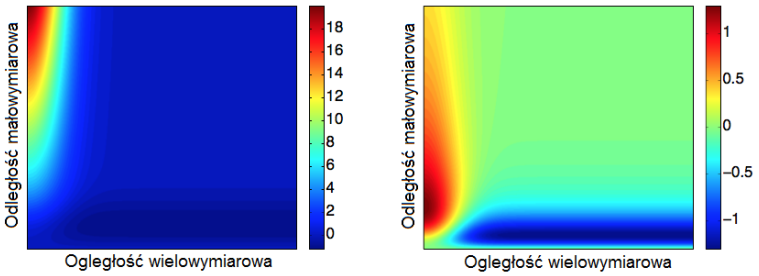
Pomimo niewątpliwie dobrych wyników wizualizacji uzyskiwanych za pomocą opisanej powyżej metody SNE jej optymalizacja stanowi niewątpliwy problem [148]. Dlatego też w 2004 roku opracowana została metoda wykorzystująca rozkład t-Studenta wraz ze wspominają metodą SNE.

Wykorzystując rozkład t-Student z jednym stopniem swobody oraz symetryczność metody SNE, mówiącą że $\forall i, j, p_{ji} = p_{ij}$ oraz $q_{ji} = q_{ij}$, rozkład prawdopodobieństwa zdefiniowany poprzez równanie 4.2 otrzymuje następującą postać [147]:

$$q_{ij} = \frac{\left(1 + ||y_i - y_j||^2\right)^{-1}}{\sum_{k \neq l} \left(1 + ||y_i - y_l||^2\right)^{-1}}. \tag{4.8}$$

Z kolei gradient dywergencji Kullbacka-Leiblera pomiędzy zbiorem P oraz rozkładem prawdopodobieństwa Q wyznaczonego przy wykorzystaniu rozkładu t-Studenta oraz równania 4.8 wyrażony jest następująco [147]:

$$\frac{\delta C}{\delta y_i} = 4 \sum_j (p_{ij} - q_{ij}) (y_i - y_j) \left(1 + ||y_i - y_l||^2\right)^{-1}. \tag{4.9}$$



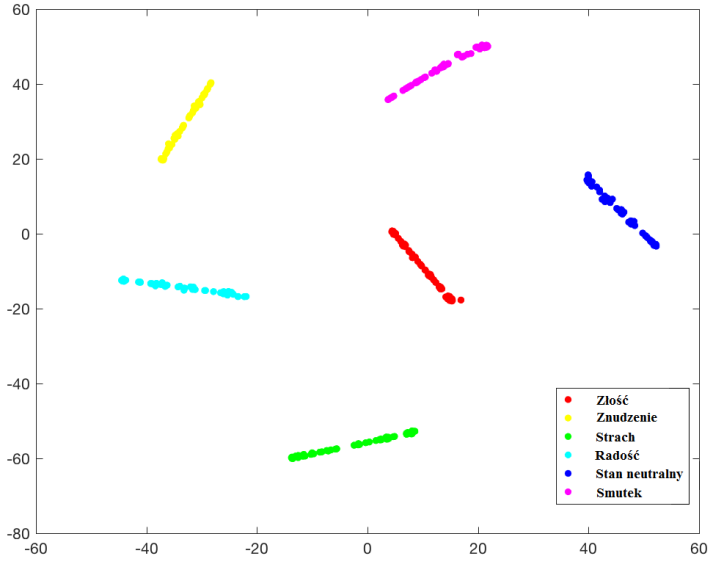
(a) Gradient SNE. (b) Gradient t-SNE.

Rysunek 4.1. Porównanie gradientu SNE i t-SNE. Opracowanie własne na podstawie [147]

Algorithm 1 Schemat działania algorytmu t-SNE

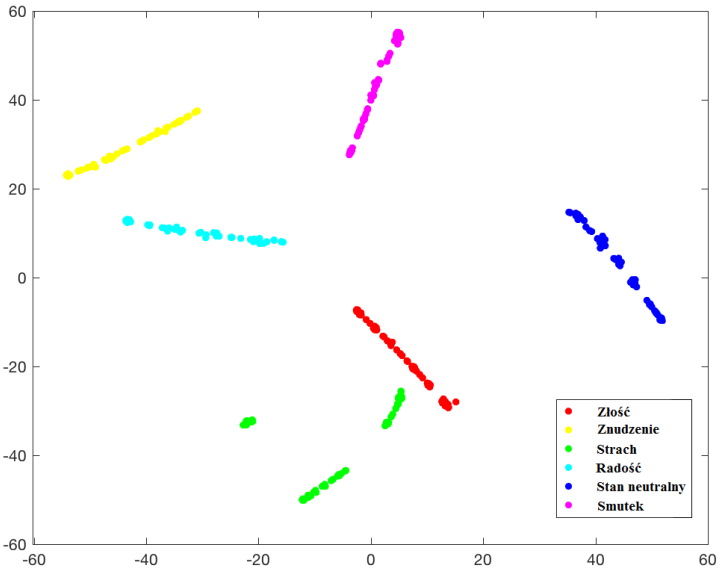
Require: Zbiór danych $X = \{x_1, x_2, x_3, ..., x_n\}$
Parametry funkcji kosztu (niepewność): $Perp$
Parametry optymalizacyjne: liczba iteracji T (ustawiona empirycznie na 1000), współczynnik uczenia u, momentum $\alpha(t)$
Ensure: Zbiór danych małowymiarowych $\gamma^{(T)} = \{y_1, y_2, y_3, ..., y_n\}$
begin
Oblicz prawdopodobieństwo par $p_{j|i}$ wraz z niepewnością $Perp$, w oparciu o równanie 4.1
Ustaw $p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n}$
Ustaw początkową wartość $\gamma^{(0)} = \{y_1, y_2, y_3, ..., y_n\}$
for t = 0 **to** T **do**
Wyznacz q_{ij} przy użyciu równania 4.8
Wyznacz gradient $\frac{\delta C}{\delta y}$ przy użyciu równania 4.9
Wyznacz $\gamma^{(t)} = \gamma^{(t-1)} + \eta \frac{\delta C}{\delta \gamma} + \alpha(t) \left(\gamma^{(t-1)} - \gamma^{(t-2)}\right)$
end for
end

Na Rysunku 4.1. został zaprezentowany gradient pomiędzy dwoma punktami y_i oraz y_j w przestrzeni małowymiarowej, reprezentowany w postaci ich par odległości euklidesowych w przestrzeni wielowymiarowej i małowymiarowej (na przykład w postaci funkcji $||x_i = x_j||$ oraz $||y_i - y_j||$) dla metody SNE oraz t-SNE. Dodatnie wartości gradientu informują o „przyciąganiu" punktów y_i, y_j w przestrzeni małowymiarowej. Ujemne, z kolei stanowią „odpychanie" punktów. Bazując na Rysunku 4.1. można zaobserwować dwie znaczące zalety t-SNE. Po pierwsze gradient t-SNE mocno „odpycha" odległe, w przestrzeni wielowymiarowej, punkty które są reprezentowane poprzez małą odległość w przestrzeni małowymiarowej. SNE ma również podobną własność, jednakże jego oddziaływanie jest nieporównywalnie mniejsze. Po drugie, duże odległości pomiędzy punktami w przestrzeni wielowymiarowej nie są reprezentowane poprzez odległości dążące do nieskończoności w przestrzeni małowymiarowej. Schematycznie działanie algorytmu t-SNE zostało zaprezentowane na Listingu 4.1.

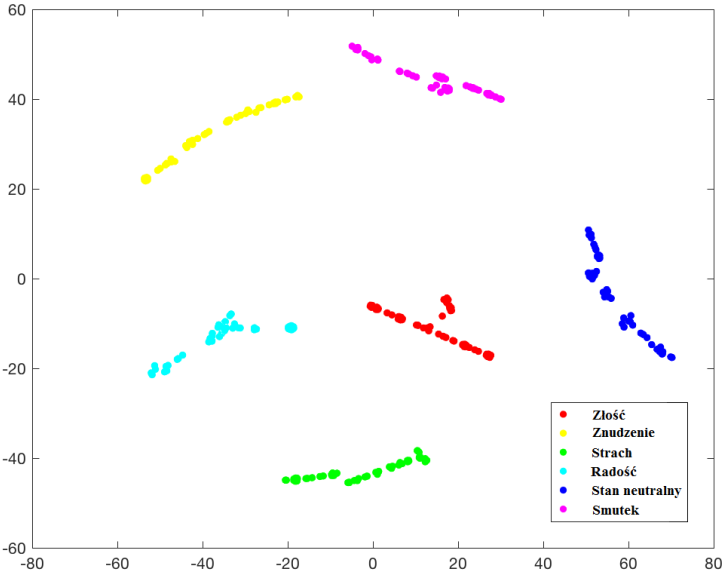


Rysunek 4.2. Separowalność stanów emocjonalnych dla Bazy A

W oparciu o powyższy algorytm oraz parametry opisane w podrozdziale 4.1. dokonano wizualizacji zbiorów nagrań mowy emocjonalnej. Wizualizację tą przedstawiono na Rysunkach 4.2.–4.4. w przestrzeni dwuwymiarowej. Wspomniana wizualizacja miała na celu pokazanie, iż możliwa jest taka separowalność nagrań poszczególnych stanów emocjonalnych, która pozwoli na późniejszą klasyfikację emocji.



Rysunek 4.3. Separowalność stanów emocjonalnych dla Bazy B



Rysunek 4.4. Separowalność stanów emocjonalnych dla Bazy C

Innymi słowy dokonano sprawdzenia czy poszczególne stany emocjonalne nie są ze sobą „wymieszane". Jak widać na Rysunkach 4.2.–4.4. po-

szczególne stany emocjonalne są zgrupowane. Co więcej dla każdej z baz danych nagrań mowy emocjonalnej konkretne stany emocjonalne zlokalizowane są (w przybliżeniu), w tych samych obszarach przestrzeni dwuwymiarowej. Przeprowadzone badania pozwalają stwierdzić, iż zastosowanie odpowiedniego klasyfikatora jak również właściwie dobranej metody przetwarzania danych pozwoli na efektywną klasyfikację stanu emocjonalnego mówcy w czasie zbliżonym do czasu rzeczywistego.

4.3. Wnioski do rozdziału

1. Możliwe jest wykorzystanie podejścia heurystycznego do ekstrakcji zarówno statystycznych jak i formantowych oraz melowych parametrów sygnału mowy.
2. Zastosowanie metody t-SNE jako narzędzia do wizualizacji danych wielowymiarowych pozwala na przedstawienie zabioru nagrań mowy emocjonalnej na płaszczyźnie dwuwymiarowej.
3. Wizualizacje dla poszczególnych baz nagrań pokazują, iż możliwe jest grupowanie poszczególnych stanów emocjonalnych.
4. Dla każdej z baz nagrań mowy emocjonalnej konkretne stany emocjonalne zlokalizowane są w tych samych obszarach przestrzeni dwuwymiarowej.

5. Badania dotyczące detekcji stanu emocjonalnego

Rozpoznawanie stanu emocjonalnego mówcy w oparciu tylko i wyłącznie o parametry wyekstrahowane wyłącznie z sygnału mowy, w dodatku w czasie zbliżonym do czasu rzeczywistego nie jest zagadnieniem trywialnym, jednak jego znacznie systematycznie wzrasta. Związane jest to w pewnym stopniu z dynamizmem zmian zachodzącym w rozwoju systemów bazujących na komunikacji człowiek – komputer (ang. HCI — Human Computer Interaction), gdzie emocje mogą odgrywać pewną znaczącą rolę. W niniejszym rozdziale zaprezentowane zostały autorskie wyniki otrzymane podczas prac nad opisanym powyżej zagadnieniem. Jako klasyfikator zostały wykorzystane sztuczne sieci neuronowe. W zależności od celu badań i etapu prac nad zagadnieniem, zbadane zostały zarówno proste SSN jak i bardziej złożone, wielowarstwowe i wieloneuronowe sieci. Sprawdzony został również wpływ takich czynników jak: funkcja aktywacji neuronów w procesie uczenia SSN, dobór algorytmu uczenia czy wreszcie wybór rodzaju sieci neuronowej jak również jej struktury. Całość symulacji została przeprowadzona w oparciu o środowisko Matlab[15] wraz z pakietem Neural Networks Toolbox.

5.1. Metodyka badań

Przeprowadzone badania skupiały się wokół kilku aspektów. Pierwszym etapem prac były testy pilotażowe mające na celu potwierdzenie bądź negacje tezy mówiącej, iżmożliwa jest skuteczna identyfikacja stanu emocjonalnego mówcy w oparciu o dane pozyskiwane z sygnału mowy polskiej przy pomocy sztucznych sieci neuronowych.

Kolejny aspekt skupiał się wokół zagadnień związanych z doborem odpowiednich metod uczenia SSN oraz funkcji aktywacji neuronów.

Właściwy etap analizy polegał na wykorzystaniu metod spektrograficznych w połączeniu z autorskim sposobem przetwarzania sygnału mowy, w celu opracowania parametrów wejściowych dla SSN jak również dobór odpowiednich własności sztucznych sieci neuronowych.

W badaniach symulacyjnych wykorzystane zostały trzy bazy nagrań emocjonalnej mowy polskiej scharakteryzowane w Rozdziale 4. Każde z nagrań zostało odpowiednio przygotowane do przetwarzania. Pierwszym krokiem była zamiana analogowego sygnału mowy w sygnał dyskretny. Posłużono się tutaj odpowiednio dobraną częstotliwością próbkowania[16]. Zarejestrowany sygnał posiada oczywiście pewne zakłócenia, dlatego też

[15] W zależności od etapu badań w wersji 2015 lub 2016.

[16] W oparciu o twierdzenie Nyquista-Kotielnikova-Shannona.

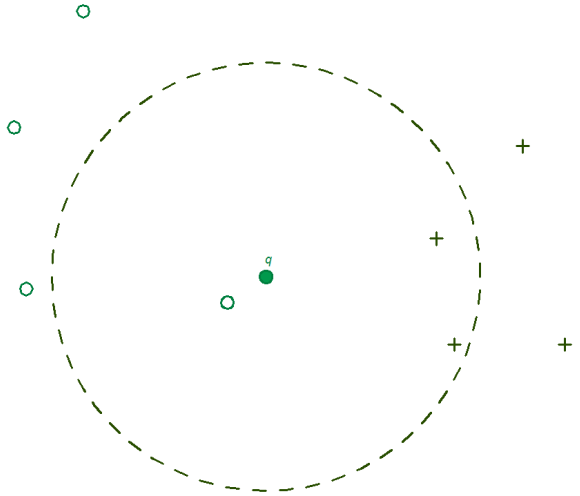
kolejny etap obejmował marginalizację wpływ składowych niepożądanych. W tym celu posłużono się zestawem narzędzi oferowanym przez środowisko Matlab. Odfiltrowane zostały składowe sygnału charanterystyczne dla różnego rodzaju szumów (jak np. szum klimatyzacji, rekuperacji, itp) czy odgłosy ruchu ulicznego (klaksony samochodów). Przetwarzanie wstępne obejmowało również wyrównanie charaktrystki przenoszenia mikrofonu oraz podbicie częstotliwości charakterystycznych dla sygnału mowy. Następnie sygnał był poddawany procesowi normalizacji do przedziału $[-1, 1]$. Opisane powyżej kroki miały miejsce przy wszystkich przeprowadzonych testach. Kolejne działania były uzależnione od aktualnie wykorzystywanej metody przetwarzania sygnału, wykorzystującej albo transformatę Fouriera i metodę spektrogramów lub transformatę falkową wraz ze skalogramami, i zostały dokładnie opisane w odpowiednich podrozdziałach niniejszej pracy.

5.2. Klasyfikacja danych za pomocą algorytmu k-NN

Reguła k najbliższych sąsiadów jest jednym z najpopularniejszych klasyfikatorów mimimaloodległościowych [40]. Jego działanie polega na przypisaniu analizowanej próbki do tej grupy sąsiadów, której występowanie jest najliczniejsze w sąsiedztwie k jej najbliższych sąsiadów [155]. W przypadku, gdy kilka rywalizujących grup jest równo-oddalonych od klasyfikowanej próbki, wówczas przypisane do jednej z klas następuje w sposób arbitralny [40]. Schemat działania reguły k-NN, dla k wynoszącego 3 został zaprezentowany na Rysunku 5.1. Testowa próbka q zostanie przypisana do grupy „krzyżyków" ponieważ wśród jej trzech najbliższych sąsiadów znajdują się dwa obiekty klasy „krzyży" oraz jeden klasy „kółko". Podkreślić należy fakt, iż kolejność badania sąsiadów nie ma znaczenia przy dokonaniu ostatecznej klasyfikacji.

Wśród szeregu zalet powyższego klasyfikatora należy pokreślić zbieżność jego błędu do błędu Bayesa, gdy liczebność zbioru danych dąży do nieskończoności oraz wartość ilorazu k do n dąży do zera [40]. Ponadto algorytm ten zapewnia na ogół akceptowalnie wysoką jakość klasyfikacji przy stosunkowo szybkim uczeniu. Niestety szybkość podejmowania decyzji o przynależności do jednej z klas jest już stosunkowo wolna. Ponadto niezbędne jest przechowywania w pamięci całego zbioru odniesienia.

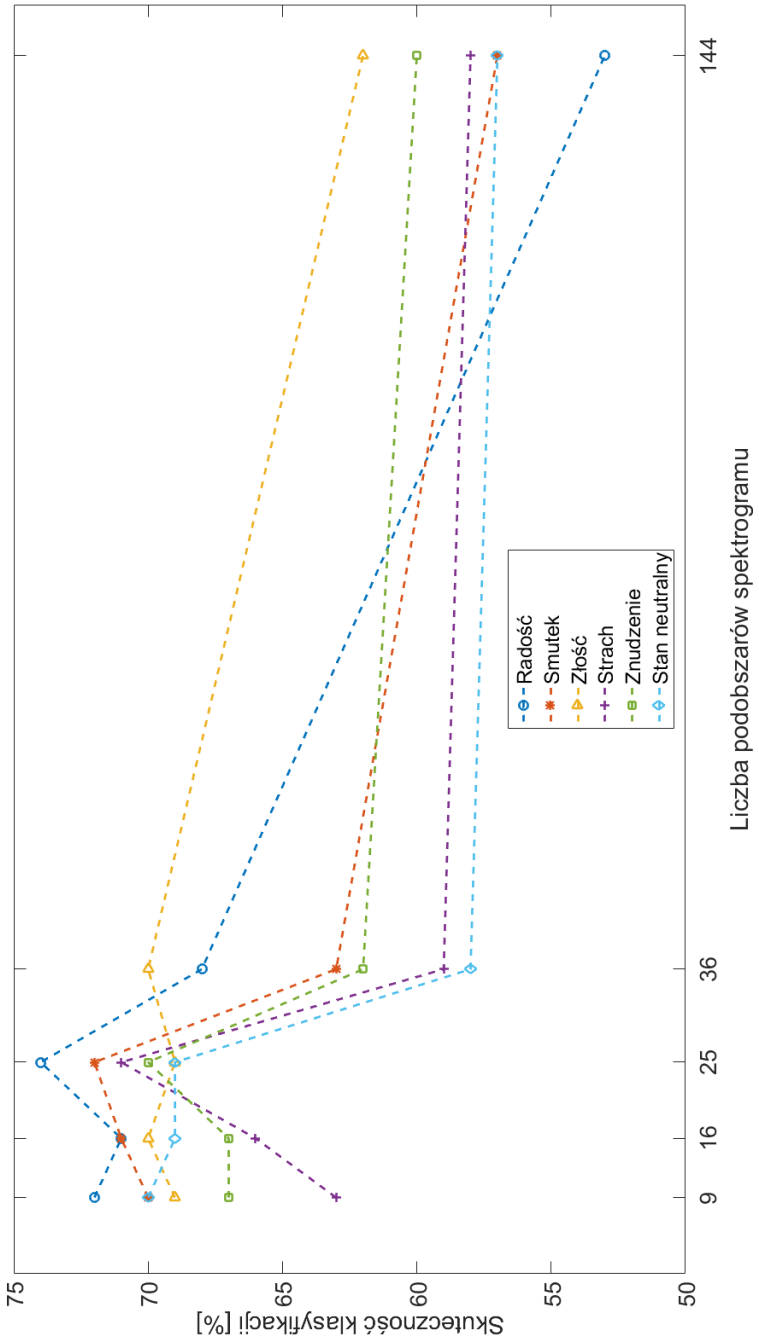
Najszybszą formą reguły k-NN jest reguła 1-NN, przypisującą nieznaną, testową próbkę do klasy jej najbliższego sąsiada. Badania przeprowadzone w trakcie projektu Statlog [77, 96], w ramach którego porównanych zostało kilkanaście klasyfikatorów na ponad 20 dużych bazach danych pokazało, że dla 75% testów najlepszą wartością parametru k była 1.



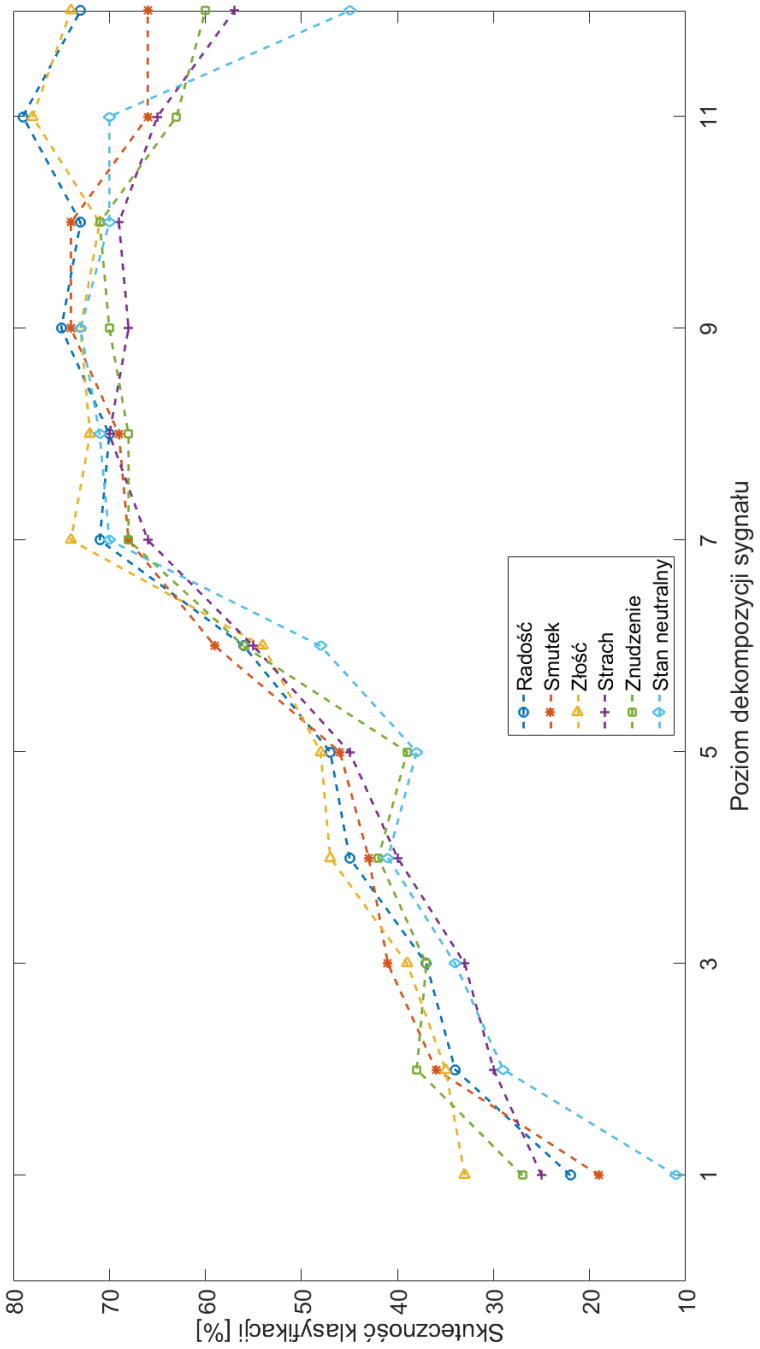
Rysunek 5.1. Algorytm 3-NN. Testowa próbka q zostanie przyporządkowana do klasy „krzyżyków”

Jako punkt odniesienia do eksperymentów wykonanych w ramach niniejszych badań wykorzystane zostały wyniki uzyskane przy użyciu reguły 1-NN. Należy podkreślić, iż jeśli pominięty zostanie etap selekcji cech w regule 1-NN nie występuje etap uczenia. Niestety szumy występujące w badanych próbkach znacząco wpływają na jakość klasyfikacji. Ponadto reguła 1-NN została wykorzystana również do wyznaczenia poziomu dekompozycji sygnału przy wykorzystaniu transformaty falkowej, szerzej opisanej w podrozdziale 5.6. Na Rysunku 5.2. zostały zaprezentowane wyniki skuteczności klasyfikacji przy użyciu metody opartej o spektrogramy. Ze względu na czas przetwarzania danych podział spektrogramu na więcej niż 144 podobszary nie został zbadany. Wpływ poziomu dekompozycji na skuteczność klasyfikacji został z kolei przedstawiony na Rysunku 5.3. W badaniach wykorzystana została falka Haara.

Jak można zaobserwować mała liczba podobszarów (9,16, 25) znacząco wypływa na skuteczność klasyfikacji stanu emocjonalnego mówcy. Należy podkreślić, że większa szczegółowość podziału spektrogramu nie wpływa na poprawę skuteczności identyfikacji. Przy podziale na 144 podobszary otrzymano wyniki znacząco gorsze niż w przypadku podziału na 25 pól. Z wyników zaprezentowanych na Rysunku 5.3. wynika, że najbardziej korzystna, ze względu na skuteczność klasyfikacji, jest dekompozycja sygnału mowy od 7 do 11 poziomu. Powyżej tej granicy otrzymywane wyniki przejawiają tendencję recesywną.



Rysunek 5.2. Wyniki klasyfikacji przy wykorzystaniu reguły 1-NN w zależności od liczby podobszarów spektrogramu



Rysunek 5.3. Wyniki klasyfikacji przy wykorzystaniu reguły 1-NN w zależności od poziomu dekompozycji sygnału

5.3. Badania pilotażowe

Przeprowadzone badania skupiały się wokół następujących stanów emocjonalnych: radość, smutek, strach, złość, znudzenie oraz stan neutralny. Pierwsze badania miały charakter stricte testowy i służyły wyłącznie sprawdzeniu czy wybrany klasyfikator będzie odpowiedni do tego typu zagadnień. W przeprowadzonych eksperymentach posłużono się bazą nagrań mowy emocjonalnej opracowanej przez pracowników Politechniki Łódzkiej, a scharakteryzowanej w rozdziale 4 i oznaczonej poprzez Baza A. Dwustu czterdziesto elementowy zbiór nagrań został podzielony na dwa równoliczne podzbiory, z których jeden stanowił dane uczące SSN, drugi posłużył dotestów. Parametry wejściowe stanowiło pięć wartości. Pierwszą była, zdefiniowana następująco, energia sygnału mowy [163]:

$$Ex = \sum_{n=0}^{N} x^2(n), \tag{5.1}$$

gdzie:
N – oznacza liczbę wszystkich próbek,
n – oznacza aktualnie przetwarzaną próbkę.
$x(n)$ – oznacza wartość n-tej próbki

Drugim parametrem wejściowym była średnia wartość całego sygnału. Matematycznie wartość ta jest opisana w następujący sposób [163]:

$$\bar{x_N} = \lim_{x \to \infty} \frac{1}{2N+1} \sum_{n=-N}^{N} x(n), \tag{5.2}$$

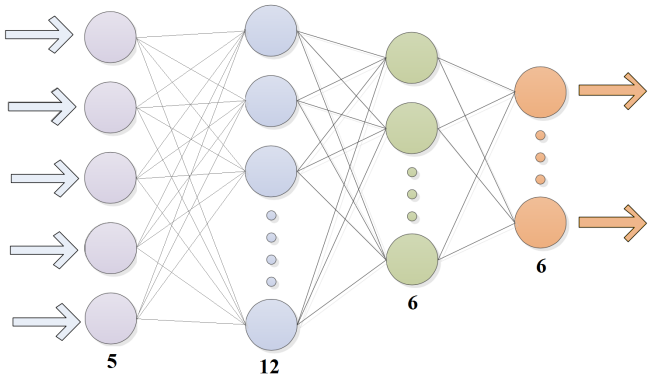
gdzie:
N – oznacza liczbę wszystkich próbek,
n – oznacza aktualnie przetwarzaną próbkę.
$x(n)$ – oznacza wartość n-tej próbki.

Kolejne parametry stanowiły minimalna i maksymalna wartość próbki w przetwarzanym sygnale oraz płeć mówcy. Podstawę do testów stanowiła jednokierunkowa SSN(5-12-6-6)[17] o następujących parametrach:

- pięć neuronów w warstwie wejściowej,
- dwie warstwy ukryte składające się odpowiednio z 12 oraz 6 neuronów,
- sześcioneuronowa warstwa wyjściowa – każdemu z neuronów odpowiada jeden z identyfikowanych stanów emocjonalnych,
- neurony aktowowane funkcją sigmoidalną oraz tangensa hiperbolicznego,

[17] Poszczególne numery odpowiadają liczbie neuronów w kolejnych warstwach.

- sieć uczona za pomocą algorytmu wstecznej propagacji błędów – opisanego szczegółowo w Rozdziale 3,
- liczba epok została ustalona na 1500,
- maksymalny, dopuszczalny błąd wynosił 0,1,
- początkowa wartość wag została wyznaczona przy użyciu metody Monte Carlo i mieściła się w przedziale [0,1].
  Struktura całej sieci została zaprezentowana na Rysunku 5.4.

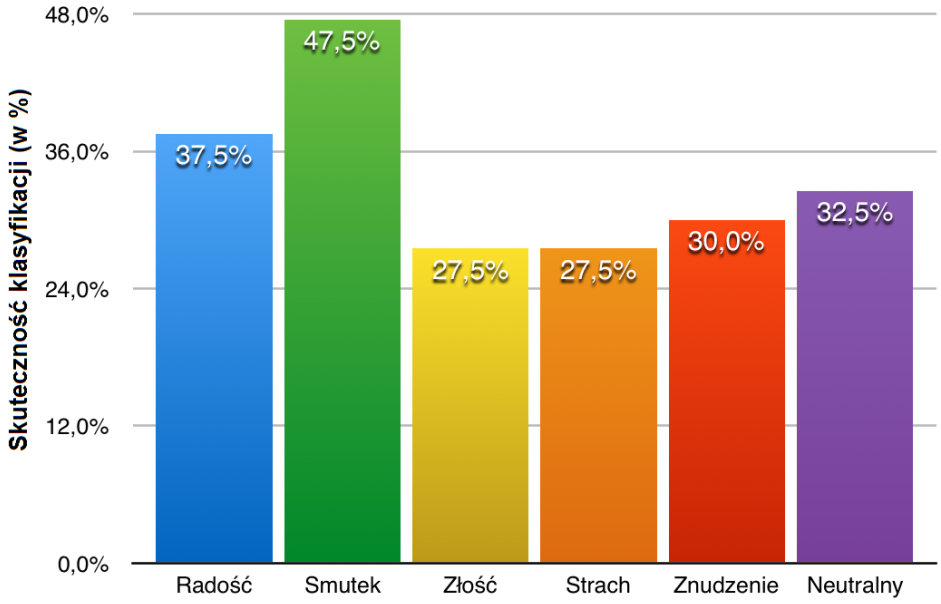


Rysunek 5.4. Struktura wykorzystanej SSN

Podczas badań uzyskano wyniki wahające się w przedziale 27,5–47,5% [120] (w zależności od rozpatrywanego stanu emocjonalnego). Liczba przeprowadzonych symulacji wynosiła 240 (po 40 dla każdego ze stanów emocjonalnych). Całość otrzymanych rezultatów wraz z macierzą pomyłek zostały zaprezentowane odpowiedni na Rysunku 5.5. oraz w Tabeli 5.1.

Tabela 5.1. Macierz pomyłek otrzymana podczas badań z SSN(5-12-6-6) (w %)

Emocje oczekiwane	**Emocje otrzymane na wyjściu**					
	Radość	Smutek	Złość	Strach	Znudzenie	Neutralny
Radość	**37,50**	2,50	30,00	15,00	2,50	2,50
Smutek	2,50	**47,50**	2,50	5,00	12,50	20,00
Złośc	20,00	2,50	**27,50**	25,00	10,00	10,00
Strach	20,00	15,00	25,00	**27,50**	15,00	10,00
Znudzenie	10,00	20,00	12,50	20,00	**30,00**	25,00
Neutralny	2,50	12,50	2,50	7,50	30,00	**32,50**



Rysunek 5.5. Wyniki otrzymane podczas badań z SSN(5-12-6-6)

Łatwo zauważyć, iż otrzymane wyniki są dalekie od oczekiwanych jednakże należy wziąć pod uwagę, że były to tylko badania pilotażowe mające na celu określenie przydatności klasyfikatora. Jako wektor wejściowy zostały wykorzystane nieliczne parametry statystyczne, których liczba nie pozwalała na uzyskanie lepszych rezultatów. Jednakże same badania pozwoliły na dostrzeżenie szeregu problemów. Po pierwsze zostało zauważone, iż niektóre stany emocjonalne są często ze sobą mylone. Po drugie dostrzeżono wpływ płci mówcy na skuteczności identyfikacji stanu emocjonalnego. Po trzecie zauważono konieczność sprawdzenia innych struktur SSN.

5.4. Wpływ algorytmu uczenia SSN na skuteczność identyfikacji stanu emocjonalnego mówcy

Badania pilotażowe pokazały słuszność użycia sztucznych sieci neuronowych jako klasyfikatorów stanów emocjonalnych mówców. Podstawowy problem stanowiła jednak skuteczność identyfikacji emocji. Dlatego też nalżało przyjżeć się procesowi uczenia sieci oraz doborowi paramterów startowych. W związku z powyższym kolejnym krokiem mogącym znacząco wypłynąć na poprawę skuteczności klasyfikacji jest dobór algorytmu uczenia SSN.

Dlatego też kolejne badania skupiały się wokół wspomnianego zagadnienia. Przetestowane zostało jak następujące algorytmy uczenia wpływają na skuteczność klasyfikacji:

- algorytm wstecznej propagacji (GD),
- algorytm wstecznej propagacji z momentum (GDM),
- algorytm wstecznej propagacji z adaptacyjną zmianą współczynnika uczenia (GDA),
- algorytm wstecznej propagacji z adaptacyjną zmianą współczynnika uczenia oraz momentum (GDX),
- algorytm wstecznej propagacji skalowalnego gradientu (SCG).

5.4.1. Algorytm wstecznej propagacji

Szeroko opisany w Rozdziale 3 algorytm wstecznej propagacji można sprowadzić do kilku kroków:

1. Generowanie losowych wartości wag.
2. Podanie wybranego wzorca na wejście sieci.
3. Wyznaczenie odpowiedzi wszystkich neuronów wyjściowych sieci:

$$y_k^n = f\left(\sum_{j=1}^{I} w_{kj}^n y_j^{n-1}\right).$$

4. Wyznaczenie błędów wszystkich neuronów w warstwie wyjściowej:

$$\delta_k^n = z_k - y_k^n.$$

5. Obliczenie błędów w warstwach ukrytych (aby wyznaczyć błąd warstwy h-1, należy znać błąd dla warstwy h, następującej po h-1):

$$\delta_j^{h-1} = \frac{\delta f(u_j^{h-1})}{\delta u_j^{h-1}} \sum_{k=1}^{I} \delta_k^h w_{kj}^h.$$

6. Zmodyfikowanie wartości wag według następującej zależności:

$$w_{ji}^{h-1} = w_{ji}^{h-1} + \eta \delta_j^{h-1} y_i^{h-1}.$$

7. Powrót do punktu 2.

Można zaobserwować, że podstawowy algorytm wstecznej propagacji poza względami optymalizacyjnymi, posiada szereg wad:

- brak pewności osiągnięcia globalnego minimum funkcji błędu,
- występowanie wielu minimów lokalnych,
- konieczność wykonania dużej liczby iteracji,
- groźba wystąpienia oscylacji [144].

5.4.2. Algorytm wstecznej propagacji z momentum

Zastosowanie algorytmu wstecznej propagacji błędów z momentum jest jednym z rozwiązań poprawiających efektywne tempo uczenia bez zagrożenia dla stabilności samego procesu. Idea metody opiera się na wprowadzeniu do procesu uaktualnienia wagi pewnego współczynnika bezwładności. Podobnie jak w przypadku współczynnika uczenia, także współczynnik momentu w najprostszej postaci algorytmu wprowadzany jest jako wartość stała, niezmienna podczas trwania treningu sieci neuronowej, w bardziej wyrafinowanej formie wyrażana jest równaniem [78]:

$$w_{ji}(t) = w_{ji}(t-1) + \eta\delta_j(t-1)y_i(t-1) + \alpha\left(w_{ji}(t-1) - w_{ji}(t-2)\right), \quad (5.3)$$

gdzie:
$\alpha \in (0, 1]$ – współczynnik momentum.

Współczynnik momentum jest niezależny od wartości gradientu oraz uwzględnia poprzednią korekcję wagi [78].

5.4.3. Algorytm wstecznej propagacji ze zmianą adaptacyjną

Pierwotnie współczynnik uczenia sieci neuronowej ustalany był przed rozpoczęciem działania algorytmu wstecznej propagacji błędów. Jego wartość była często dobierana w sposób eksperymentalny, aby dla większości przypadków proces uczenia zbliżał się do funkcji celu. Takie podejście może powodować pewne problemy wynikające z charakterystyki procesu uczenia. Po pierwsze, zbyt duża wartość współczynnika może powodować pomijanie istotnych minimów. Z kolei ustalona zbyt mała wartość może skutkować „utknięciem" w ekstremach lokalnych [132]. Aby zaradzić powyższemu problemowi przyjęło się stosowanie adaptacyjnej wartości wspomnianego współczynnika. Jeżeli aktualny błąd sieci przekracza wartość wyznaczoną w poprzedniej iteracji wartość (oznacza to, że proces uczenia jest daleki od osiągnięcia optimum) współczynnika jest zwiększana, w przeciwnym razie zmniejszana. Takie podejście pozwala na znaczną poprawę parametrów uczenia sieci, a co się z tym wiąże poprawę skuteczności klasyfikacji [153].

5.4.4. Algorytm wstecznej propagacji gradientu sprzężonego

W algorytmie gradientów sprzężonych poszukiwania minimum odbywa się nie w oparciu o malejącą wartość gradientu, a w taki sposób aby aktualna wartość była w sprzężona z poprzednio wyznaczoną. Matematycznie, kierunek zmian, jest to wyrażone w następującej postaci [132]:

$$p_t = -\nabla E(\overrightarrow{w_t}) + \textstyle\sum_{j=0}^{t-1} \beta_t p_j. \quad (5.4)$$

$E(w_t) = \sum_{k=1}^{N} \sum_{i=1}^{m} [r_{ki}(w)]^2$ – oznacza błąd popełniany przez *i-te* wyjście sieci dla *k-tego* zestawu uczącego, z m wejściami,
$\beta = \frac{\nabla J_w^{(j+1)} - \nabla J_w^{(j)} * J_w^{(j+1)}}{J_w^{(j)^2}}$, gdzie J oznacza jakobian,
p_j – ciąg j wzajemnie sprzężonych kierunków, tworzących bazę R^n.

Jak można zaobserwować zachodzi tutaj konieczność uwzględniania wszystkich poprzednich zmian kierunku poszukiwań gradientu, dlatego też, uwzględniając warunek ortogonalności, wzór 5.4 upraszcza się do następującej postaci [132]:

$$p_t = -\nabla E(\overrightarrow{w_t}) + \beta_t p_j. \tag{5.5}$$

Widać, że nowa wartość gradientu zależy tylko od poprzednio wyznaczonego kierunku oraz współczynnika sprzężenia β. Istnieje wiele metod wyznaczania owego współczynnika [100]. Wśród najczęściej wykorzystywanych należy wymienić: Fletchera-Reevesa [38] czy wykorzystywana w niniejszej pracy metoda Polaka-Ribiere'a [118, 149]:

$$\beta = \frac{[\nabla E(\overrightarrow{w_k})]^T (\nabla E(\overrightarrow{w_k}) - \nabla E(\overrightarrow{w_{k-1}}))}{[\nabla E(\overrightarrow{w_{k-1}})]^T \nabla E(\overrightarrow{w_{k-1}})}. \tag{5.6}$$

Pewną niedogodnością opisanej metody jest możliwość wystąpienia błędów zaokrągleń po przekroczeniu pewnej, zazwyczaj dość dużej, liczby epok. Jednakże, ze względu na szybką zbieżność, algorytm gradientów sprzężonych jest szeroko stosowany [132].

Kolejnym czynnikiem mającym znaczny wpływ na skuteczność klasyfikacji oraz szybkość uczenia sztucznych sieci neuronowych jest wykorzystana funkcja aktywacji. W podrozdziale 3.6.2. zostały opisane funkcje aktywacji wykorzystane w niniejszych badaniach. Podstawę do testów stanowiła podobnie jak w przypadku badań pilotażowych, jednokierunkowa SSN(5-12-6-6) o parametrach wymienionych poniżej:

- pięć neuronów w warstwie wejściowej,
- dwie warstwy ukryte składające się odpowiednio z 12 oraz 6 neuronów,
- sześcioneuronowa warstwa wyjściowa – każdemu z neuronów odpowiada jeden z identyfikowanych stanów emocjonalnych,
- funkcja aktywacji neuronów: sigmoidalna, liniowa, znormalizowana funkcja wykładnicza (Softmax), tangens hiperboliczny,
- sieć uczona za pomocą różnych wersji algorytmu wstecznej propagacji błędów – opisany szczegółowo powyżej,
- liczba epok została ustalona na 1500 i maksymalny błąd wynosił: 0,1,
- początkowe wartości wag wyznaczone losowo z zakresu [0,1].

Zostały przeprowadzone symulacje, w których wykorzystano Bazę A [121]. Połowa nagrań posłużyła jako dane uczące, druga połowa jako dane testowe. Parametry wejściowe stanowiły wartości opisane w podrozdziale dotyczącym badań pilotażowych. Przeprowadzone zostało 600 symulacji. Najlepsze wyniki otrzymano w przypadku wykorzystania funkcji tangensa hiperbolicznego jako funkcji aktywacji i uczeniu sieci za pomocą metody GDM. Otrzymane wyniki zostały przedstawione w Tabelach 5.2.–5.6.

Tabela 5.2. Wyniki otrzymane dla SSN(5-12-6-6) i algorytmu GD (w %)

Emocje oczekiwane	**Emocje otrzymane na wyjściu**		
	Funkcja liniowa	Funkcja sigmoidalna	Tangens hiperboliczny
Radość	32,50	37,50	35,00
Smutek	42,50	47,50	50,00
Złość	25,00	27,50	32,50
Strach	27,50	27,50	30,00
Znudzenie	32,50	30,00	30,00
Neutralny	32,50	32,50	32,50

Tabela 5.3. Wyniki otrzymane dla SSN(5-12-6-6) i algorytmu GDA (w %)

Emocje oczekiwane	**Emocje otrzymane na wyjściu**		
	Funkcja liniowa	Funkcja sigmoidalna	Tangens hiperboliczny
Radość	35,00	40,00	37,50
Smutek	42,50	42,50	45,00
Złość	27,50	27,50	25,00
Strach	32,50	35,00	32,50
Znudzenie	30,00	30,00	30,00
Neutralny	32,50	35,00	35,00

Łatwo zauważyć, iż wyniki rzędu 30% są dalekie od satysfakcjonujących, jednakże pokazano, iż w przypadku wykorzystania standardowego algorytmu wstecznej propagacji błędów najskuteczniejszą funkcją aktywacji neuronów jest tangens hiperboliczny. Należy podkreślić, że stopień identyfikacji smutku przy użyciu stosunkowo trywialnej struktury sztucznej sieci neuronowej jest znacząco lepszy, niż pozostałych stanów emocjonalnych.

Tabela 5.4. Wyniki otrzymane dla SSN(5-12-6-6) i algorytmu GDM (w %)

Emocje oczekiwane	**Emocje otrzymane na wyjściu**		
	Funkcja liniowa	Funkcja sigmoidalna	Tangens hiperboliczny
Radość	37,50	37,50	42,50
Smutek	45,00	47,50	52,50
Złość	25,00	27,50	30,00
Strach	32,50	30,00	27,50
Znudzenie	30,00	32,50	32,50
Neutralny	35,00	35,00	35,00

Podobna sytuacja ma miejsce w przypadku algorytmu wstecznej propagacji błędów z adaptacyjną zmianą współczynnika uczenia sieci. Należy jednak zwrócić tutaj uwagę na fakt, iż średni czas uczenia algorytmem GDA w porównaniu do klasycznego algorytmu GD (patrz. Rysunek 5.6. patrz s. 84) był o rząd wielkości mniejszy. Również w tym wypadku najskuteczniejszą klasyfikację gwarantowała tangens hiperboliczny jako funkcja aktywacji neuronów, a identyfikacja smutku wypadła zdecydowanie lepiej niż pozostałych stanów emocjonalnych.

Nieznacznie lepszy poziom klasyfikacji w stosunku do algorytmów GD (Tabela 5.2. patrz s. 81) oraz GDA (Tabela 5.3. patrz s. 81) zapewniała sztuczna sieć neuronowa, która była uczona przy wykorzystaniu algorytmu wstecznej propagacji błędów z współczynnikiem momentum. Jednak w tym przypadku, podobnie jak miało to miejsce dla algorytmu GD, średni czas uczenia SSN nierzadko przekraczał 1000s. Najgorzej w tym wypadku przebiegła klasyfikacja złości oraz znudzenia osiągając wyniki nieprzekraczające 30%. Podobnie jak dla sieci uczonych algorytmami GD oraz GDA również dla SSN uczonych GDM (Tabela 5.4.) najwyższą skuteczność klasyfikacji osiągnięto dla sieci, której neurony były aktywowane tangensem hiperbolicznym. Zdecydowanie najwyższy poziom identyfikacji stanu emocjonalnego mówcy osiągnięty został przy identyfikacji smutku.

Użycie algorytmu wstecznej propagacji błędów z adaptacyjną zmianą współczynników uczenia oraz momentum zagwarantowała tylko nieznacznie lepszy poziom identyfikacji stanów emocjonalnych. Podobnie jak przy poprzednich algorytmach najłatwiej rozpoznawalny był smutek, z kolei złość oraz znudzenie były stanami, których klasyfikacja przebiegła najgorzej.

Algorytm GDX (Tabela 5.5.) okazał się być jednym z najszybciej działających metod uczenia sztucznych sieci neuronowych. Średni czas nauki wyniósł niespełna 15 s.

Algorytm wstecznej propagacji skalowalnego gradientu okazał się być zdecydowanie najszybciej działającą metodą nauki SSN. Średni czas potrzebny do jednokrotnego nauczenia SSN przyjmował wartości poniżej 10 s. Jednakże skuteczność klasyfikacji stanów emocjonalnych przez sieć uczoną algorytmem SCG (Tabela 5.6.) jest nieznacznie gorsza, niż w przypadku metody GDX czy GDM. Również w przypadku SSN nauczanych przy wykorzystaniu algorytmu SCG najlepsze wyniki klasyfikacji uzyskiwane były, gdy neurony aktywowane były przy użyciu tangensa hiperbolicznego. Dlatego też w dalszych badaniach to właśnie ta funkcja aktywacji była stosowana. Również w tym przypadku stanem emocjonalnym najlepiej identyfikowanym był smutek, najgorzej zaś klasyfikacja przebiegała w przypadku złości.

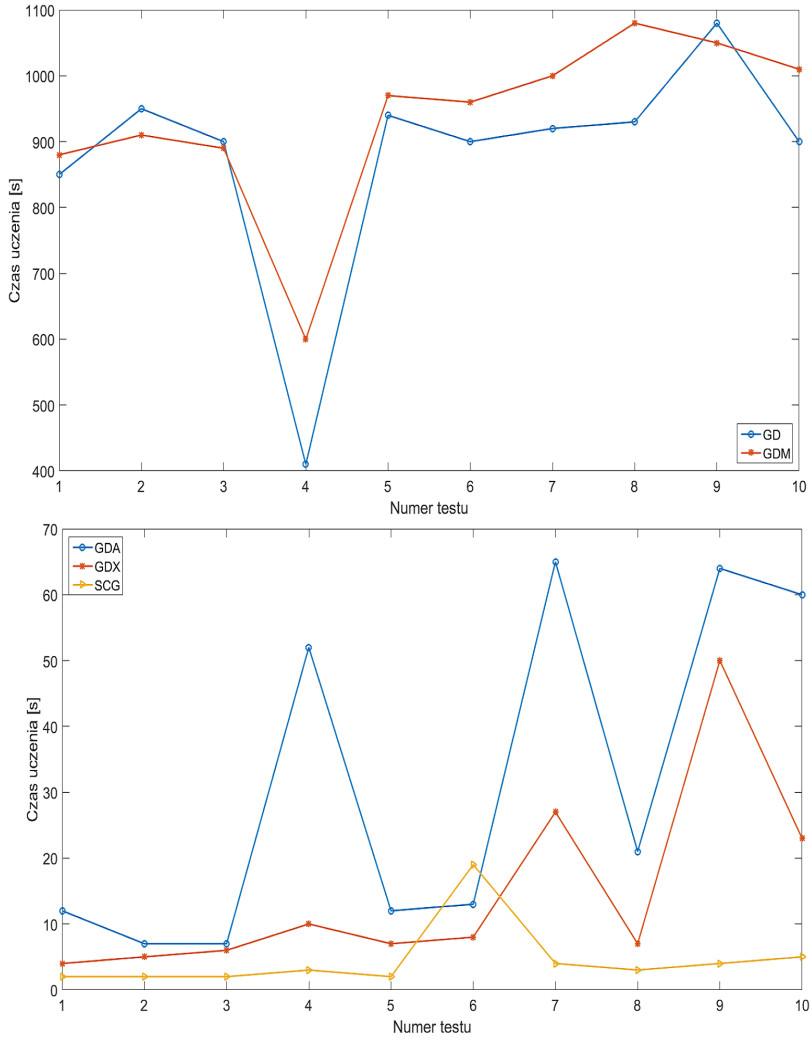
Tabela 5.5. Wyniki otrzymane dla SSN(5-12-6-6) i algorytmu GDX (w %)

Emocje oczekiwane	**Emocje otrzymane na wyjściu**		
	Funkcja liniowa	Funkcja sigmoidalna	Tangens hiperboliczny
Radość	37,50	40,00	42,50
Smutek	47,50	47,50	52,50
Złość	25,00	30,00	30,00
Strach	35,00	32,50	30,00
Znudzenie	27,50	27,50	25,00
Neutralny	32,50	30,00	32,50

Tabela 5.6. Wyniki otrzymane dla SSN(5-12-6-6) i algorytmu SCG (w %)

Emocje oczekiwane	**Emocje otrzymane na wyjściu**		
	Funkcja liniowa	Funkcja sigmoidalna	Tangens hiperboliczny
Radość	35,00	32,50	40,00
Smutek	42,50	42,50	50,00
Złość	27,50	27,50	27,50
Strach	27,50	27,50	25,00
Znudzenie	30,00	30,00	32,50
Neutralny	32,50	35,00	32,50

Pokazany na Rysunku 5.6. średni czas uczenia został zmierzony w przypadku, gdy metodą aktywacji neuronów była funkcja sigmoidalna. Można zaobserwować znaczne różnice w czasie potrzeby do nauki SSN za pomocą poszczególnych algorytmów. Należy podkreślić, iż czas niezbędny aby sztuczna sieć neuronowa potrafiła poprawnie klasyfikować dane w przypadku metody GD oraz GDM jest o rząd wielkości większy niż w przypadku pozostałych metod. Jednocześnie skuteczność klasyfikacji jest porównywalna.



Rysunek 5.6. Porównanie czasów uczenia dla różnych algorytmów

5.5. Wykorzystanie spektrogramów w procesie przetwarzania sygnału mowy polskiej

Metody czasowo – częstotliwościowe są szeroko stosowane w zagadnieniach związanych z przetwarzaniem mowy. Częstym zjawiskiem jest dodatkowy ich podział na reprezentację czas – częstotliwość oraz czas – skala [79]. Czasowo – częstotliwościowe metody pozwalają na estymację widma sygnału mowy w krótkim przedziale czasowym. Wspomniane przybliżenie jest dokonywane w oparciu o dane pozyskiwane za pomocą metody przesuwnego okna[18].

Krótkoczasowa transformata Fouriera (STFT) oraz metody spektrograficzne odgrywają szczególną rolę w zagadnieniach przetwarzania sygnału mowy. Metody te są zaliczane do przetwarzania sygnału mowy w przestrzeni czas – częstotliwość [76]. Ciągła krótkoczasowa transformata Fouriera może stanowić szczególny przypadek transformaty Gabora [163]. Dla sygnału ciągłego $x(t)$, w dziedzinie częstotliwości, wspomniana transformata jest definiowana następująco [79]

$$STFT_x^F(t,f) = e^{-j2\pi ft} \int_{-\inf}^{+\inf} X(\Theta) W^*(\Theta - f) e^{j2\pi\Theta t} d\Theta. \tag{5.7}$$

Natomiast w dziedzinie czasu STFT definiuje następująca równość:

$$STFT_x^F(t,f) = \int_{-\inf}^{+\inf} x(\tau) w^*(\tau - t) e^{j2\pi f\tau} d\tau, \tag{5.8}$$

gdzie:
$w(t)$ – oznacza funkcję okna w widmie Fouriera $W(\Theta)$,
$X(\Theta)$ – oznacza widmo analizowanego sygnału,
„$*$” – oznacza sprzężenie zespolone.
Bazując na równaniu 5.8 wykonywane są obliczenia, które wykorzystują przekształcenie Fouriera dla fragmentów sygnału wejściowego pozyskiwanych przy użyciu okna $w(t)$. W dziedzinie częstotliwości transformata Fouriera jest równoważna [20]:

1. Odwrotnemu przekształceniu Fouriera wyznaczonemu dla fragmentu widma sygnału $W(\Theta)$ pozyskane poprzez użycie w dziedzinie częstotliwości okna $W(\Theta - f)$.

[18] Opisanej w Rozdziale 2.

2. Przemieszczeniu w dziedzinie częstotliwości sygnału wyznaczonego w oparciu o punkt 1 do częstotliwości zerowej poprzez przemnożenie sygnału przez $e^{j2\pi f\tau}$.

W cyfrowym przetwarzaniu sygnałów szczególne znaczne ma następująca postać równania 5.8 [79]:

$$STFT(n,k) = \sum_{m=0}^{N} w(m)x(n-m)e^{-j\left(\frac{2\pi}{n}k\right)m}, \tag{5.9}$$

gdzie:
$n = 0, N, 2N, ..., M - N$,
$k = 0, 1, 2, ..., N - 1$,
M – oznacza liczbę analizowanych próbek. Z kolei związek pomiędzy krótkoczasową transformatą Fouriera oraz spektrogramem $S(n,k)$ definiuje następująca zależność [79]:

$$S(n,k) = |STFT(n,k)|^2. \tag{5.10}$$

5.5.1. Dobór parametrów spektrogramu

Nie jest możliwa efektywna identyfikacja stanu emocjonalnego mówcy bez odpowiednio dobranych parametrów spektrogramu. Wśród tych własności należy wymieć: szerokość i funkcję okna czy rozdzielczość czasowo-częstotliwościową. Należy zauważyć, iż najlepszą rozdzielczość czasową można uzyskać przy nakładowaniu rzędu $N - 1$, gdzie N oznacza liczbę próbek w sygnale. Jak łatwo zauważyć taki rodzaj nakładkowania – przesuwanie okna tylko o jedną próbkę, powoduje silny wzrost liczby niezbędnych obliczeń. Dlatego też w badaniach wykonanych na potrzeby niniejszej pracy wykorzystane zostały nakładowanie wynoszące 50% szerokości okna. Sam dobór odpowiedniej szerokości okna również jest zagadnieniem dość złożonym. Najlepsza efektywność jest uzyskiwana gdy stosunek szerokości okna w dziedzinie częstotliwośći do szerokości okna w dziedzicie czasu[19] był równy stosunkowi przyrostu częstotliwości do czasu, w którym dany przyrost miał miejsce [79]:

$$\frac{A}{B} = \frac{\Delta f}{\Delta t}, \tag{5.11}$$

[19] Obie wartości liczone w sensie średniokwadratowym [126].

gdzie:

$$A = \sqrt{\frac{1}{E} \int_{-\inf}^{+\inf} f^2 |W(f)^2| df}, \tag{5.12}$$

$$B = \sqrt{\frac{1}{E} \int_{-\inf}^{+\inf} t^2 |w(t)^2| dt}, \tag{5.13}$$

$$E = \sqrt{\int_{-\inf}^{+\inf} |w(t)|^2 dt} = \sqrt{\int_{-\inf}^{+\inf} |W(f)|^2 df}. \tag{5.14}$$

Problem doboru rodzaju funkcji okna jak i jej parametrów powinien stanowić swoisty kompromis pomiędzy jakością sygnału uzyskiwaną na wyjściu, a czasem niezbędnym do wykonania obliczeń. Należy również zauważyć, iż sam dobór funkcji okna jest pewnym kompromisem pomiędzy szerokością listka głównego, poziomem pierwszego listka bocznego oraz szybkością zmian poziomów listków bocznych wraz ze wzrostem częstotliwości. A zatem jest to kompromis pomiędzy dokładnością wartości amplitudy oraz częstotliwością [121].

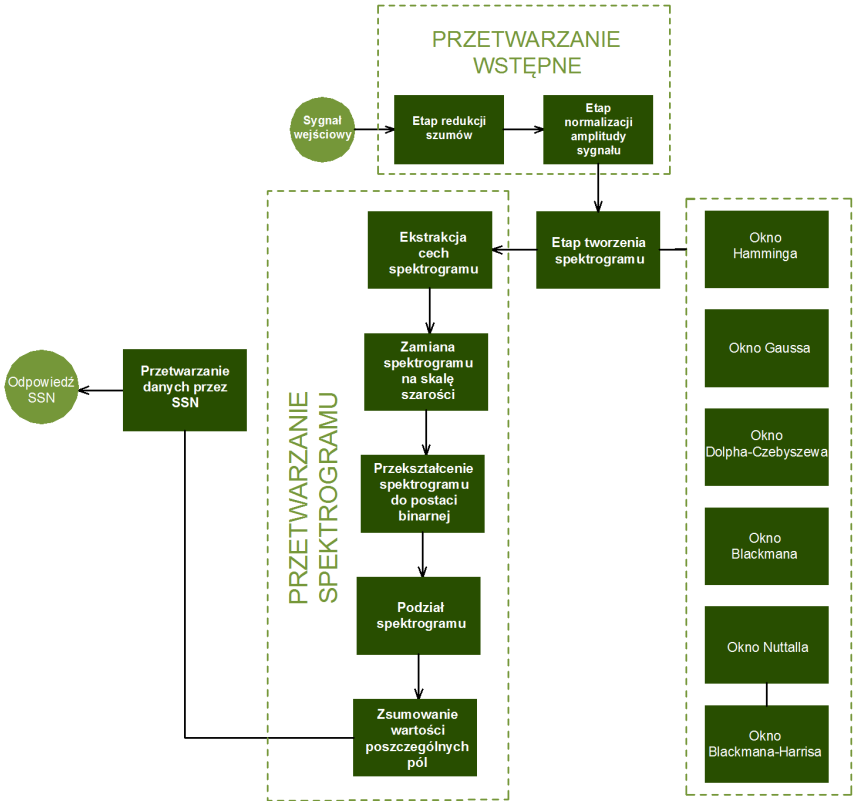
5.5.2. Ekstrakcja cech spektrogramu

Nieodzownym elementem przetwarzania danych zakodowanych w postaci spektrogramu jest ich ekstrakcja. W ramach niniejszych badań etap ten został podzielony na kilka zadań. Głównym elementem etapu ekstrakcji cech ze spektrogramu było utworzenie zestawu danych stanowiących wektor wejściowy dla SSN. Proces ekstrakcji cech przebiegał następująco:

1. Przedstawienie spektrogramu w skali odcieni szarości (0–255).
2. Przekształcenie otrzymanego spektrogramu w postać binarną. Wartości poniżej progu zyskały wartość 0, powyżej – 1. Przeprowadzony został szereg eksperymentów mający na celu wyznaczenie najlepszej wartości progu. Zbadany został zakres od 100 do 200. Najlepsze wyniki zostały osiągnięte gdy wartość progu wynosiła 155.
3. Obraz uzyskany poprzez zamianę spektrogramu do obrazu binarnego został podzielony odpowiednio na 9, 16, 25, 36 oraz 144 fragmenty. Dla każdego z podziałów przeprowadzone zostały oddzielne eksperymenty.

4. Wartości w poszczególnych obszarach zostały zsumowane stając się wektorem wejściowym dla sztucznej sieci neuronowej, której struktura uzależniona była od podziału spektrogramu.

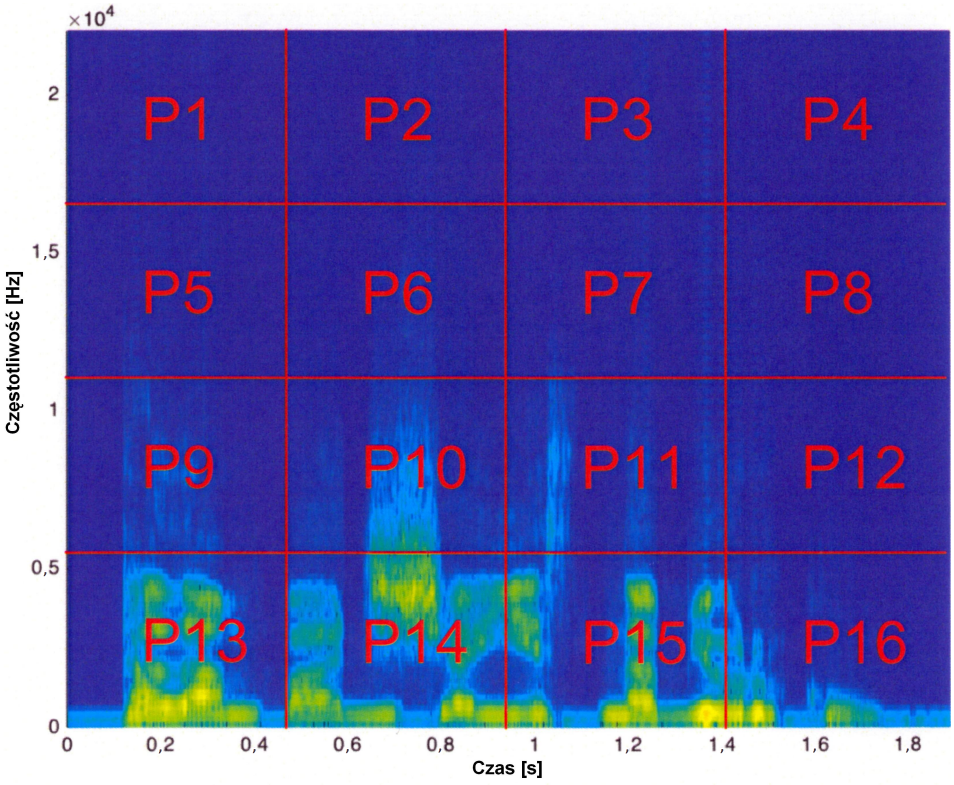
Całość procesu przetwarzania sygnału oraz klasyfikacji została przedstawiona na Rysunku 5.7. Z kolei przykład podziału spektrogramu został zaprezentowany na Rysunku 5.8. (patrz s. 89). W przeprowadzonych badaniach został ponadto sprawdzony wpływ funkcji okna na skuteczność identyfikacji stanu emocjonalnego mówcy. Wykorzystane zostały metody okienkowania opisane w Rozdziale 3.



Rysunek 5.7. Schemat przetwarzania sygnału mowy z wykorzystaniem spektrogramów

5.5.3. Zastosowane sztuczne sieci neuronowe

W niniejszym rozdziale przedstawiono wyniki symulacji, które zostały przeprowadzone w środowisku Matlab 2016 z wykorzystaniem pakietu Neural Network Toolbox. Użyte struktury sztucznych sieci neuronowych były ściśle związane z podziałem spektrogramu opisanym powyżej. W każdym



Rysunek 5.8. Przykład podziału spektrogramu (dla lepszej czytelności został przedstawiony spektrogram w skali barwnej)

z eksperymentów wykorzystane zostały czterowarstwowe SSN, w których ostatnia warstwa zbudowana była z sześciu neuronów odpowiadającym sześciu identyfikowanym stanom emocjonalnym. Warstwa wejściowa składała się z tylu neuronów jaka była liczba obszarów wydzielonych ze spektrogramu oraz dodatkowego neuronu przechowującego informację o płci mówcy lub bias. Ilość neuronów w warstwach ukrytych dobierana była eksperymentalnie. W szczególności struktura SSN prezentowała się następująco:

- podział spektrogramu na 9 obszarów:
  - 10 neuronów w warstwie wejściowej,
  - 20 neurony w pierwszej warstwie ukrytej,
  - 40 neurony w drugiej warstwie ukrytej,
  - 6 neuronów w warstwie wyjściowej.
- podział spektrogramu na 16 obszarów:

  - 17 neuronów w warstwie wejściowej,
  - 34 neurony w pierwszej warstwie ukrytej,
  - 68 neurony w drugiej warstwie ukrytej,
  - 6 neuronów w warstwie wyjściowej.
- podział spektrogramu na 25 obszarów:
  - 26 neuronów w warstwie wejściowej,
  - 52 neurony w pierwszej warstwie ukrytej,
  - 104 neurony w drugiej warstwie ukrytej,
  - 6 neuronów w warstwie wyjściowej.
- podział spektrogramu na 36 obszarów:
  - 37 neuronów w warstwie wejściowej,
  - 74 neurony w pierwszej warstwie ukrytej,
  - 148 neurony w drugiej warstwie ukrytej,
  - 6 neuronów w warstwie wyjściowej.
- podział spektrogramu na 144 obszarów:
  - 145 neuronów w warstwie wejściowej,
  - 290 neurony w pierwszej warstwie ukrytej,
  - 580 neurony w drugiej warstwie ukrytej,
  - 6 neuronów w warstwie wyjściowej.

We wszystkich badaniach jako funkcji aktywacji użyty został tangens hiperboliczny. Sieć, ze względu na szybkość działania, uczona była przy wykorzystaniu algorytmu GDX. Nauka odbywała się do momentu osiągnięcia maksymalnie dopuszczalnego błędu, wynoszącego 0,05 lub do 2000 epok.

5.5.4. Inicjalizacja wag neuronów

Ustawienie początkowej wartości wag dla wszystkich neuronów stanowi jeden z istotnych problemów wpływających na skuteczność i czas uczeniasztucznych sieci neuronowych. Środowisko Matlab dostarcza kilka metod umożliwiających inicjalizację wag zbudowanej sztucznej sieci neuronowej. Wśród najczęsciej stosowanych należy wymienić funkcję $RAND$ pozwalającą na wygenerowanie losowych wartości oraz funkcję $RANDN$ umożliwiającą generacje znormalizowanych wartości losowych. Ponadto możliwe jest wygenerowanie wartość znormalizowanych do 1 przy pomocy funkcji $RANDNC$. W badaniach przeprowadzonych na potrzeby niniejszej pracy jako funkcję inicjalizacji wag wybrana została metoda $INITNW$ (Nguyen-Widrow [104]). Metoda szczegółowo została opisana zarówno w dokumentacji środowiska Matlab. W przeprowadzonych badaniach zdecydowano się na wykorzystanie algorytmu Nguyen-Widrow od inicjalizacji wag neuronów ze względu na pojawiający się niewielki błąd uczenia [111].

W badaniach wykorzystane zostały dwie bazy nagrań: Baza A oraz Baza B. Pięćdziesiąt procent nagrań zawartych w pierwszej z nich posłużyło jako dane uczące. Pozostałe pliki dźwiękowe wykorzystywane były do testów. Jednocześnie został sprawdzony wpływ funkcji okna wykorzystywanej podczas opracowywania spektrogramu na skuteczność identyfikacji stanu emocjonalnego mówcy. Uzyskane wyniki zostały zaprezentowane w Tabelach 5.7. oraz 5.8. W badaniach związanych z wpływem funkcji okna na skuteczność identyfikacji stanu emocjonalnego mówcy przeprowadzone zostało 2136 testy. Najniższa uzyskana wartość klasyfikacji wynosiła 71,88%, najwyższa 87,50%. Z kolei w dalszych badaniach (związane ze strukturą SSN) wykonano 1780 symulacji, wykorzystując okno Dolpha-Czebyszeva jako funkcję okna.

Tabela 5.7. Macierz pomyłek - wartość średnia dla wszystkich eksperymentów (w %)

Emocje oczekiwane	**Emocje otrzymane na wyjściu**					
	Radość	Smutek	Złość	Strach	Znudzenie	Neutralny
Radość	**83,34**	2,67	5,12	3,64	2,98	2,25
Smutek	1,65	**78,12**	2,65	1,58	8,11	7,89
Złośc	4,75	3,75	**80,04**	5,03	3,11	3,32
Strach	5,55	1,88	7,03	**79,52**	2,13	3,89
Znudzenie	1,00	8,12	1,02	0,94	**79,69**	9,23
Neutralny	0,80	8,23	4,42	2,22	8,98	**75,35**

Tabela 5.8. Wpływ funkcji okna na skuteczność identyfikacji stanu emocjonalnego przy wykorzystaniu SSN(26-52-104-6) (w %)

Typy okien	**Emocje otrzymane na wyjściu**					
	Radość	Smutek	Złość	Strach	Znudzenie	Neutralny
Hamminga	84,38	76,04	79,17	80,21	79,17	73,96
Gaussa	78,13	73,96	75,00	71,88	75,00	71,88
Dolpha-Czebyszeva	**87,50**	**82,29**	**83,33**	**84,38**	**82,29**	**78,13**
Blackmana	82,29	77,08	78,13	77,08	79,17	76,04
Nuttalla	83,33	79,17	82,29	81,25	80,21	75,00
Blackmana-Harrisa	84,38	80,17	82,29	82,29	80,21	77,08

Jak łatwo zauważyć najlepszą skuteczność, bez względu na zastosowane okno, uzyskano dla radości, najgorszą dla stanu neutralnego. Jest to związane z częstotliwością podstawową, która dla radości zdecydowanie różni się od pozostałych stanów emocjonalnych [146]. Średnia wartość amplitudy dla poszczególnych emocji [113] również mogła mieć wpływ na otrzymane wyniki. Należy zauważyć (Tabela 5.7.), iż radość była najrzadziej mylonym stanem emocjonalnym, z kolei stan neutralny często był mylony ze znudzeniem oraz smutkiem. Najlepsze wyniki zostały uzyskane przy wykorzystaniu sieci neuronowej składającej się z 26 neuronów w warstwie wejściowej, 52-neuronowej pierwszej warstwie ukrytej, 104-neuronowej drugiej warstwie ukrytej oraz 6 neuronów wyjściowych. Najskuteczniejszą funkcją okna okazała się być funkcja Dolpha-Czebyszewa (Tabela 5.8.), dla której skuteczność identyfikacji stanów emocjonalnych sięgnęła niemal 88% w przypadku rozpoznawania radości. Jest to związane z kształtem okna oraz jego skutecznością eliminowania przecieku danych.

Warto zauważyć, iż okno Dolpha-Czebyszewa jest efektem optymalizacji, w której to ograniczona została wysokość listków bocznych przy jednoczesnej minimalizacji szerokości listka głównego, co nie ma miejsca w przypadku pozostałych okien poddanych analizie [122]. Można również zauważyć, że wraz ze wzrostem liczby podziałów spektrogramu skuteczność identyfikacji, zmienia się w niewielkim stopniu, z kolei czas potrzebny na przetworzenie pojedynczego nagrania ulega dość znacznym zmianom. Szczegółowe wyniki uzyskane podczas badań zostały zaprezentowane w poniższych tabelach.

Tabela 5.9. Wyniki uzyskane przy wykorzystaniu SSN(10-20-40-6) (w %)

Emocje oczekiwane	**Emocje otrzymane na wyjściu**					
	Radość	Smutek	Złość	Strach	Znudzenie	Neutralny
Kobiety	78,13	73,96	75,00	71,88	77,08	71,88
Mężczyźni	84,38	80,21	82,29	82,21	80,21	77,08
Kobiety i mężczyźni	82,29	77,08	78,13	77,08	79,17	76,04

Na podstawie wyników zaprezentowanych w Tabeli 5.9. można stwierdzić, że wykorzystanie spektrogramów w procesie identyfikacji stanu emocjonalnego mówcy przynosi oczekiwane rezultaty. Należy podkreślić, iż użycie sieci neuronowej o nieznacznych rozmiarach pozwoliło na klasyfikację emocji na poziomie ponad 70%.

Tabela 5.10. Wyniki uzyskane przy wykorzystaniu SSN(17-34-68-6) (w %)

Emocje oczekiwane	**Emocje otrzymane na wyjściu**					
	Radość	Smutek	Złość	Strach	Znudzenie	Neutralny
Kobiety	72,92	76,04	77,08	81,25	72,92	76,048
Mężczyźni	82,29	80,21	81,25	83,33	75,00	79,17
Kobiety i meżczyźni	79,17	79,17	80,21	84,38	73,96	76,04

Średnia skuteczność identyfikacji w przypadku SSN (10-20-40-6) wyniosła 75%. Widać również, że rozpoznawanie emocji generowanych przez mężczyzn było nieznacznie lepsze (o około 4%) niż w przypadku kobiet.

Zaprezentowane w Tabeli 5.10. wyniki eksperymentów pozwalają przypuszczać, że wzrost liczby podobszarów spektrogramów, a tym samym zmiana struktury sztucznej sieci neuronowej na sieć o większej licznie neuronów pozwala osiągnąć lepsze wyniki. Średnia skuteczność klasyfikacji stanów emocjonalnych w obu grupach (kobiety i mężczyźni) wyniosła ponad 78%. Również w przypadku SSN (17-34-68-6) poprawność klasyfikacji w przypadku mężczyzn była lepsza dla kobiet.

Jednak wzrost rozmiaru SSN spowodował powiększenie się powyższej różnicy do poziomu ok. 6,4%. Podkreślić należy, że została przekroczona 80% skuteczność klasyfikacji w przypadku mężczyzn, co wydaje się być wynikiem w pełni satysfakcjonującym.

Tabela 5.11. Wyniki uzyskane przy wykorzystaniu SSN(26-52-104-6) (w %)

Emocje oczekiwane	**Emocje otrzymane na wyjściu**					
	Radość	Smutek	Złość	Strach	Znudzenie	Neutralny
Kobiety	81,25	76,04	72,92	77,08	76,04	72,92
Mężczyźni	87,50	82,29	83,33	84,38	82,29	78,13
Kobiety i meżczyźni	84,38	76,04	79,17	80,21	79,17	73,96

W Tabeli 5.11. zostały zaprezentowane wyniki uzyskane przy użyciu sztucznej sieci neuronowej zbudowanej z 26 neuronów zlokalizowanych w warstwie wejściowej, 52 oraz 104 neuronach w dwóch kolejnych warstwach ukrytych i warstwie wyjściowej złożonej z 6 neuronów. Jest to struktura sieci, która pozwoliła na osiągnięcie najlepszych wyników w przeprowadzonych badaniach wykorzystujących spektrogramy.

Tabela 5.12. Wyniki uzyskane przy wykorzystaniu SSN(37-74-148-6) (w %)

Emocje oczekiwane	**Emocje otrzymane na wyjściu**					
	Radość	Smutek	Złość	Strach	Znudzenie	Neutralny
Kobiety	79,17	72,29	75,00	68,75	71,88	69,79
Mężczyźni	77,08	75,00	73,96	77,08	71,88	75,00
Kobiety i mężczyźni	73,96	70,83	72,92	80,21	70,83	71,88

Średnia skuteczność identyfikacji stanu emocjonalnego mówcy dla obu płci wyniosła powyżej 78,8%. W przypadku grupy mężczyzn osiągnięty rezultat oscylował w okolicy 83%, dla kobiet wynik ten wyniósł niemal 76%.

W oparciu o wyniki pokazane w Tabeli 5.12. można stwierdzić, że wzrost liczby podobszarów powyżej poziomu 36 nie powoduje poprawy osiąganych wyników klasyfikacji. Średnia skuteczność klasyfikacji wyniosła 73,4% co jest wynikiem o ponad 5% niższym niż w przypadku SSN (26-52-104-6). Również w badaniach z użyciem SSN(37-74-148-6) wyższą skuteczność uzyskano dla grupy mężczyzn.

Tabela 5.13. Wyniki uzyskane przy wykorzystaniu SSN(145-290-580-6) (w %)

Emocje oczekiwane	**Emocje otrzymane na wyjściu**					
	Radość	Smutek	Złość	Strach	Znudzenie	Neutralny
Kobiety	67,71	65,63	64,58	65,63	67,71	64,58
Mężczyźni	72,29	70,86	72,92	71,88	70,86	67,71
Kobiety i mężczyźni	71,88	67,71	69,79	67,71	68,75	66,67

W powyższej tabeli zostały zaprezentowane wyniki uzyskane dla podziału spektrogramu na 144 podobszary. Średnia skuteczność klasyfikacji przy wykorzystaniu SSN(145-290-580-6) wyniosła niespełna 69% co jest wynikiem zdecydowanie odbiegającym od rezultatów uzyskanych przy podziale spektrogramu na mniejszą liczbę podobszarów. We wszystkich eksperymentach zdecydowanie najgorzej identyfikowanym stanem emocjonalnym był stan neutralny. Wynika to z faktu, iż jest on często był mylony ze smutkiem oraz znudzeniem, które charakteryzują się podobnym pasmem częstotliwości. Z kolei najwyższe skuteczności identyfikacji uzyskano dla radości oraz strachu. W przeprowadzonych badaniach skuteczniej identyfikowane były stany emocjonalne mężczyzn niż kobiet. Różnice w klasyfikacji były rzędu kilku procent. Biorąc pod uwagę, iż człowiek jest w stanie poprawnie

sklasyfikować stan emocjonalny nieznanej mu osoby w około 60% przypadków [133] można stwierdzić, iż uzyskane rezultaty (szczególnie w przypadku SSN(26-52-104-6)) są w pełni satysfakcjonujące.

5.6. Wykorzystanie skalogramów w procesie przetwarzania sygnału mowy polskiej

Poza spektrogramami oraz krótkoczasową transformatą Fouriera duże znaczenie w zagadnieniach związanych z przetwarzaniem sygnałów odgrywają w ostaniach latach metody oparte o transformatę falkową. W niniejszym podrozdziale zostaną przedstawione wyniki uzyskane poprzez wykorzystanie wyżej wymienionego sposobu do opracowania skalogramu, który następnie zostanie poddany kolejnym etapom przetwarzania. Cały proces będzie miał na celu uzyskanie jak najlepszych wyników klasyfikacji.

5.6.1. Transformata falkowa

Rozwinięcie w szereg Fouriera nie daje informacji o zachowaniu funkcji. Niedokładna jest równiej jej wartość aproksymowana w otoczeniu punktu $x = 0$. Z tego powodu została opracowana, przez Jeana Morleta oraz Alexa Grossmana, ciągła transformata falkowa (CWT) dla jednowymiarowych sygnałów wyrażonych w następującej postaci [8]:

$$x(t) \in L^2(R), \tag{5.15}$$

przyjmuje następującą formę [62]:

$$w(a,b) = \frac{1}{\sqrt{a}} \int_{-\inf}^{+\inf} x(t)\Psi^* \frac{t-b}{a} dt, \tag{5.16}$$

gdzie:
$*$ – oznacza sprzężenie funkcji zespolonej,
$a, (a > 0)$ – oznacza parametr skali,
b – oznacza parametr przesunięcia,
Ψ – oznacza falkę macierzystą.
Innymi słowy CTW jest miarą zienności funkcji $ft)$ w otoczeniu b o rozmiarze proporcjonalnym do a.

Pomimo wielu zalet ciągłej transformaty falkowej częściej w zagadnieniach związanych z przetwarzaniem sygnałów jest wykorzystywana jej dyskretna postać [131]. Jednym z powodów takiego stanu rzeczy może być prostota dyskretnej transformaty falkowej (DWT) w porównaniu z jej wersją

ciągłą. Ponadto systemy komputerowe przetwarzają dyskretną postać sygnału, a zatem użycie DWT wydaje się być naturalnym wyjściem [131]. Dyskretna transformata falkowa jest zdefiniowana w następujący sposób [131]:

$$DWT(m,n) = \frac{1}{\sqrt{a^m}} \sum_n s(k)\Psi(a^{-m}n - bk), \tag{5.17}$$

gdzie:
$S(k)$ – oznacza sygnał wejściowy,
Ψ – oznacza falkę macierzystą,
$a, (a > 0)$ – oznacza parametr skali,
b – oznacza parametr przesunięcia.

Jedną z podstawowych zalet dyskretnej transformaty falkowej w porównaniu z opisaną wcześniej krótkoczasową transformatą Fouriera jest fakt, że DWT zapewnia dokładne i niezakłócone informacje o czasie co stanowi istotne udogodnienie w przypadku przetwarzania sygnałów [160].

Pierwszym krokiem pracy algorytmu DWT jest porówanie falki macierzystej z początkiem sygnału. Wyliczony współczynnik odzwierciedla podobieństwo aktualnego fragmentu przetwarzanego sygnału do falki macierzystej. Kolejnym krokiem jest wybór następnego fragmentu sygnału (zmiana parametru b) i ponowne porównanie go z falką. Następnie następuje przeskalowanie falki (zmiana parametru a) i powtórzenie czynności opisanych powyżej [158].

Działanie transformaty falkowej oraz funkcji skalującej przypomina sposób pracy filtra pasmowego, gdzie funkcja skalowalna jest skorelowana z filtrem dolnoprzepustowym, z kolei funkcja macierzysta z górnoprzepustowym. Pierwszy wyznacza aproksymację, drugi natomiast tworzy splot z badanym sygnałem obliczając podobieństwo współczynników. Owo powiązanie jest definiowane za pomocą następującej równości [131]:

$$h_H(n) = (-1)^n h_L(N - n), \tag{5.18}$$

gdzie:
h_H – oznacza filtr górnoprzepustowy,
h_L – oznacza filtr dolnoprzepustowy.

W przeprowadzonych badaniach zostały rozważone dwa rodzaje funkcji macierzystych: funkcji Haara oraz funkcji kapelusza Meksykańskiego na dwóch poziomach dekompozycji 7 oraz 9. Poziomy dekompozycji zostały wyznaczone w oparciu o eksperymenty opisane w podrozdziale 5.2. Za kryterium selekcji do dalszych badań posłużono się średnim czasem przetwarza-

nia pojedynczego skalogramu oraz średnim czasem uczenia SSN. Otrzymane wyniki zostały zaprezentowane na Rysunkach 5.10. oraz 5.11.

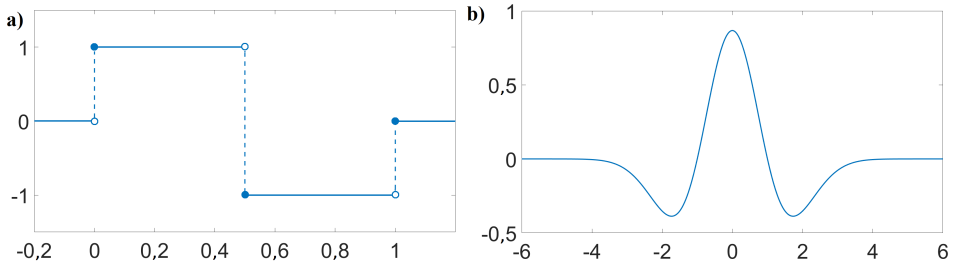
Funkcja Haara jest zdefiniowana w następujący sposób [163]:

$$H(t) = \begin{cases} 0 & \text{dla } t < 0 \\ 1 & \text{dla } 0 \leqslant t < 0,5 \\ -1 & \text{dla } 0,5 \leqslant t < 1 \\ 0 & \text{dla } t \geqslant 1. \end{cases} \tag{5.19}$$

Z kolei funkcję kapelusza meksykańskiego opisuje poniższe równanie [50]:

$$\psi(t) = \frac{2}{\sqrt{3}} \pi^{\frac{-1}{4}} (1 - t^2) e^{\frac{-t^2}{2}}. \tag{5.20}$$

Przykłady przebiegu opisanych powyżej funkcji macierzystych zostały przedstawione na Rysunku 5.9.



Rysunek 5.9. Przykład przebiegu funkcji Haara (a) oraz funkcji kapelusza meksykańskiego (b)

Na Rysunku 5.10. przedstawione zostały kryteria selekcji dla falek 7. rzędu. Należy zauważyć, iż podział skalogramu na mniej niż 175 podobszarów powodował skuteczność klasyfikacji na zdecydowanie niższym poziomie niż w przypadku wygenerowanie większej liczby danych wejściowych. Podział na 210 podobszarów i więcej nie powodował znacznego wzrostu skuteczności klasyfikacji, dlatego też do dalszych badań został użyty podział skalogramu na 210 pól. Również wartość ta jest znacząca w przypadku średniego czasu przetwarzania pojedynczego nagrania (od momentu wczytania pliku, poprzez wygenerowanie parametrów wejściowych do sieci neuronowej, po etap klasyfikacji). W przypadku podziału na 245 podobszarów wartość ta jest

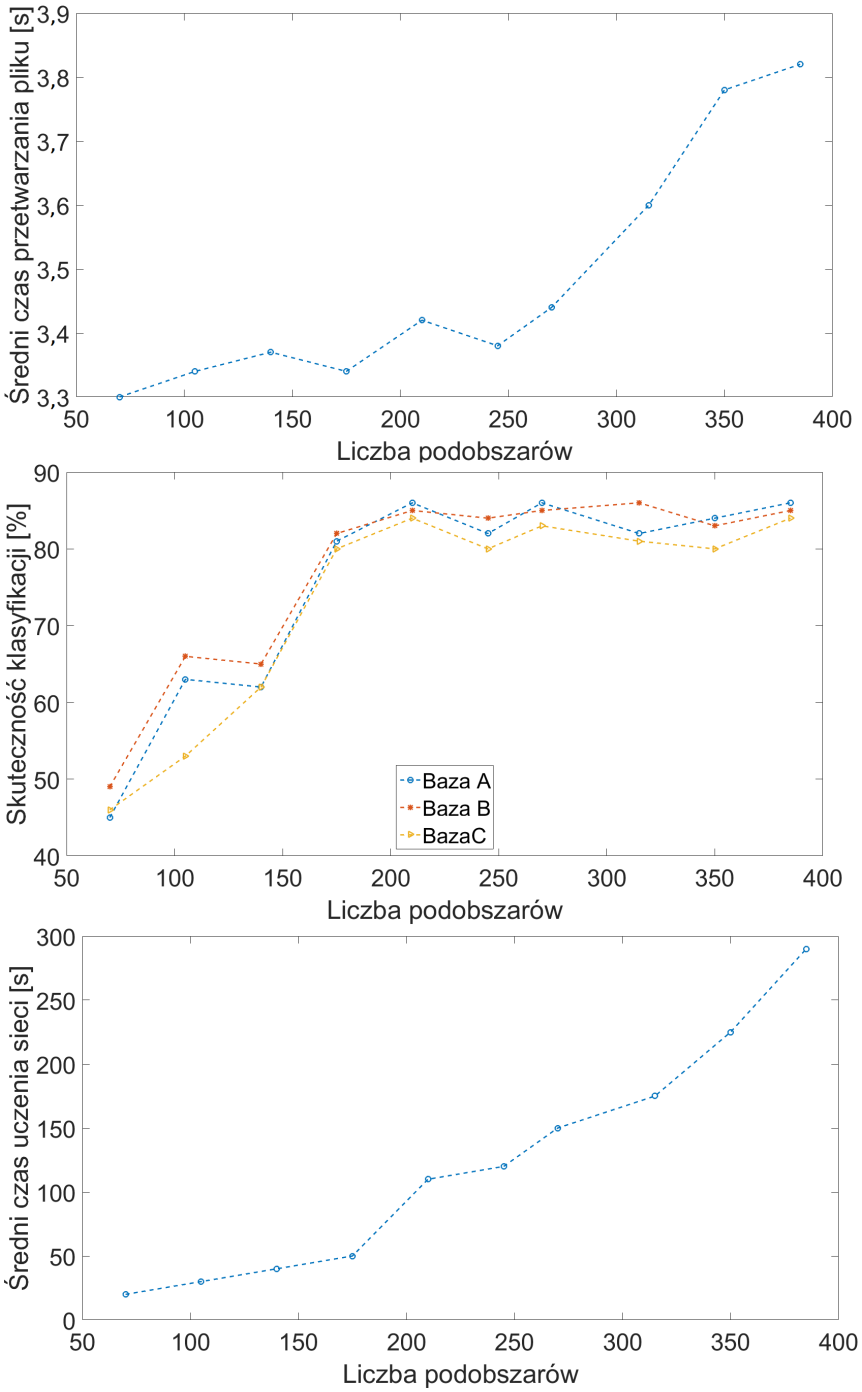
nieznacznie niższa jednak średni czas potrzebny do nauki sieci jest już większy i wprost proporcjonalny do liczby podobszarów.

W przypadku dekompozycji sygnału do 9. poziomu najlepsze wyniki zostały uzyskane przy podziale skalogramu na 360 pól. W prawdzie średni czas przetwarzania pojedynczego nagrania jest w tym przypadku zdecydowanie wyższy niż dla podziału na 270 podobszarów (patrz Rysunek 5.11. s. 100), jednak w tym przypadku skuteczność klasyfikacji jest zauważalnie niższa. Podobnie jak dla sygnału dekompowanego do 7 poziomu średni czas niezbędny do uczenia sieci jest wprost proporcjonalny do liczby podobszarów skalogramu. Ponieważ proces uczenia sztucznych sieci neuronowych odbywał się jednokrotnie czas niezbędny do nauczenia sieci nie miał znaczenia przy dokonywaniu wyborów.

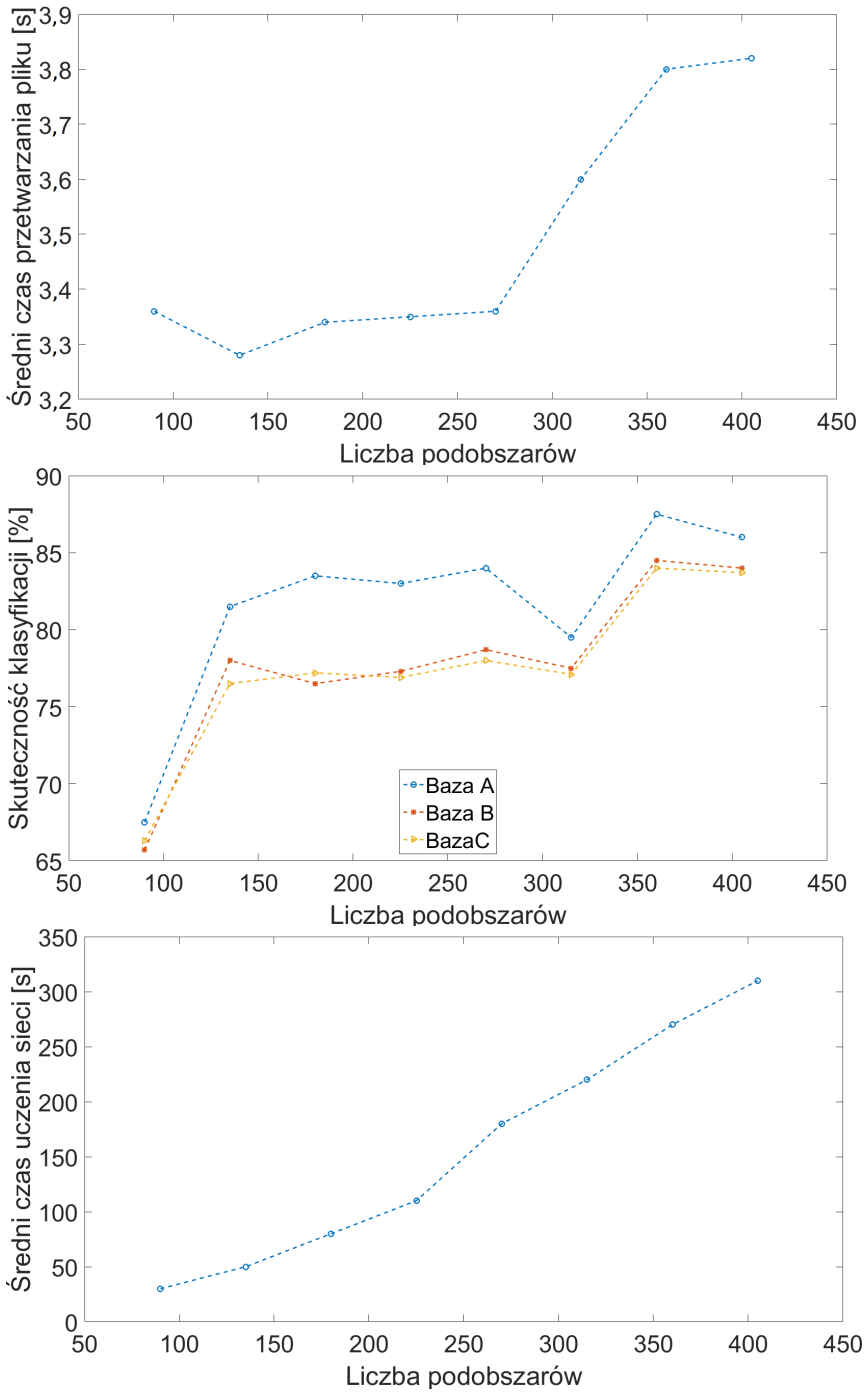
W badaniach zostały wykorzystane wszystkie bazy nagrań opisane w Rozdziale 4. (Baza A, Baza B oraz Baza C). Pierwsza z nich posłużyła jako wektor uczący, dwie pozostałe jako dane testowe. Dzięki czemu została zachowana niezależność od mówców oraz od wypowiadanego tekstu. Początkowe etapy przetwarzana sygnału były identyczne jak w przypadku badań opisanych w podrozdziale dotyczącym spektrogramów. Podstawową różnicą było opracowanie skalogramu i następnie jego przetwarzanie. Cały proces został zobrazowany na Rysunku 5.12. (patrz. s. 101). W trakcie prowadzonych prac wykonane zostało 2145 symulacji.

Pierwszy krok przetwarzania skalogramu skupiał się na jego przetworzeniu do obrazu w skali szarości, a następnie do postaci binarnej. Zastosowany zastał tutaj podobny sposób wyznaczania progu jak w przypadku skalogramów. Zostały zbadane wartości z zakresu 50–200. Najlepsze efekty uzyskane były dla progu wynoszącego 100. Zatem wszystkie wartości poniżej 100 zostały zamienione na 0, powyżej na 1. Kolejnym krokiem był podział skalogramu na mniejsze obszary. Przykład podziału został przedstawiony na Rysunku 5.13. Podział dokonywany był w taki sposób ażeby linie podziału pokrywały się z poziomem użytych falek. W przypadku wykorzystania falek 7. rzędu skalogram był dzielony odpowiednio na: 70, 105, 140, 175, 210, 245, 270, 315, 350 i 385 obszarów. Dla falek rzędu 9. było to odpowiednio: 90, 135, 180, 225, 270, 315, 360, 405 podobszarów.

Najlepsze wyniki zostały uzyskane przy podziale spkalogramu na 210 elementów, w przypadku falki 7. rzędu oraz na 360 dla 9-cio. rzędowych falek. Struktura wykorzystanych sztucznych sieci neuronowych była uzależniona od aktualnie wykorzystywanego podziału skalogramu. W przypadku badań nad falkami 7. rzędu warstwa wejściowa SSN zbudowana była z 211 neuronów. Podobnie jak w przypadku skalogramów dodatkowe wejście przechowywało informację o płci mówcy lub biasie.

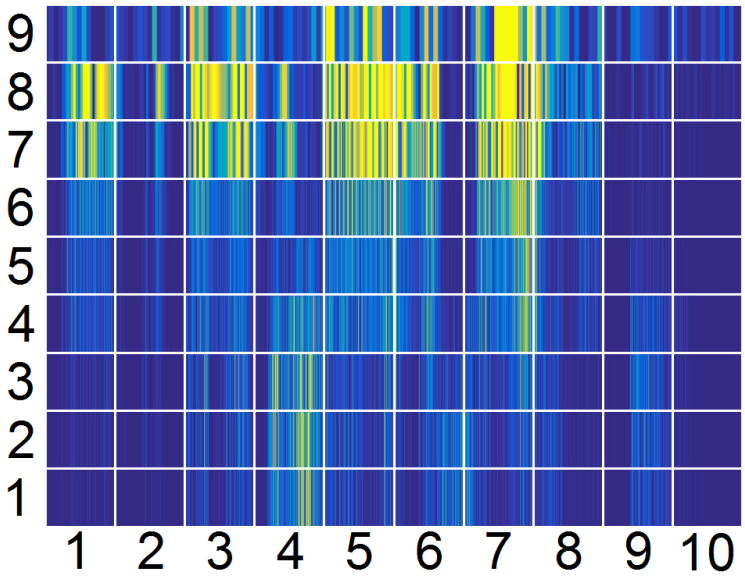


Rysunek 5.10. Kryteria selekcji dla falek 7. rzędu



Rysunek 5.11. Kryteria selekcji dla falek 9. rzędu

Rysunek 5.12. Schemat przetwarzania sygnału mowy z wykorzystaniem transformaty falkowej



Rysunek 5.13. Przykład podziału skalogramu. Dla lepszej czytelności przedstawiony w skali barwnej

Sieć posiadała również dwie warstwy ukryte, z których pierwsza liczyła 844 neurony, a druga 422. Warstwę wejściową stanowiło 6 neuronów odpowiadających za identyfikację konkretnych stanów emocjonalnych mówcy.

W przypadku dekompozycji sygnału przy użyciu falki 9. rzędu sieć zbudowana była z 361 neuronów zlokalizowanych w warstwie wejściowej, 1444 neuronów w pierwszej warstwie ukrytej, 722 w drugie warstwie ukrytej oraz 6-cio neuronowej warstwy wyjściowej.

Wszystkie badania prowadzone były z wykorzystaniem algorytmu wstecznej propagacji błędów z momentum oraz adaptacyjną zmianą współczynników uczenia (GDX). Liczba epok została określona na 2000, a dopuszczały poziom błędu na 0,1. Jako funkcję aktywacji neuronów wykorzystany został tangens hiperboliczny. Całość symulacji podobnie jak w przypadku prac z wykorzystaniem spektrogramów została przeprowadzona w środowisku Matlab 2016 z pakietem Neural Networks. Wszystkie SSN były uczone przy wykorzystaniu 50% nagrań pochodzących z Bazy A. Zdecydowano się na ten zbiór nagrań ze względu na ich jakość. W trakcie badań przeprowadzone zostały 712 symulacje.

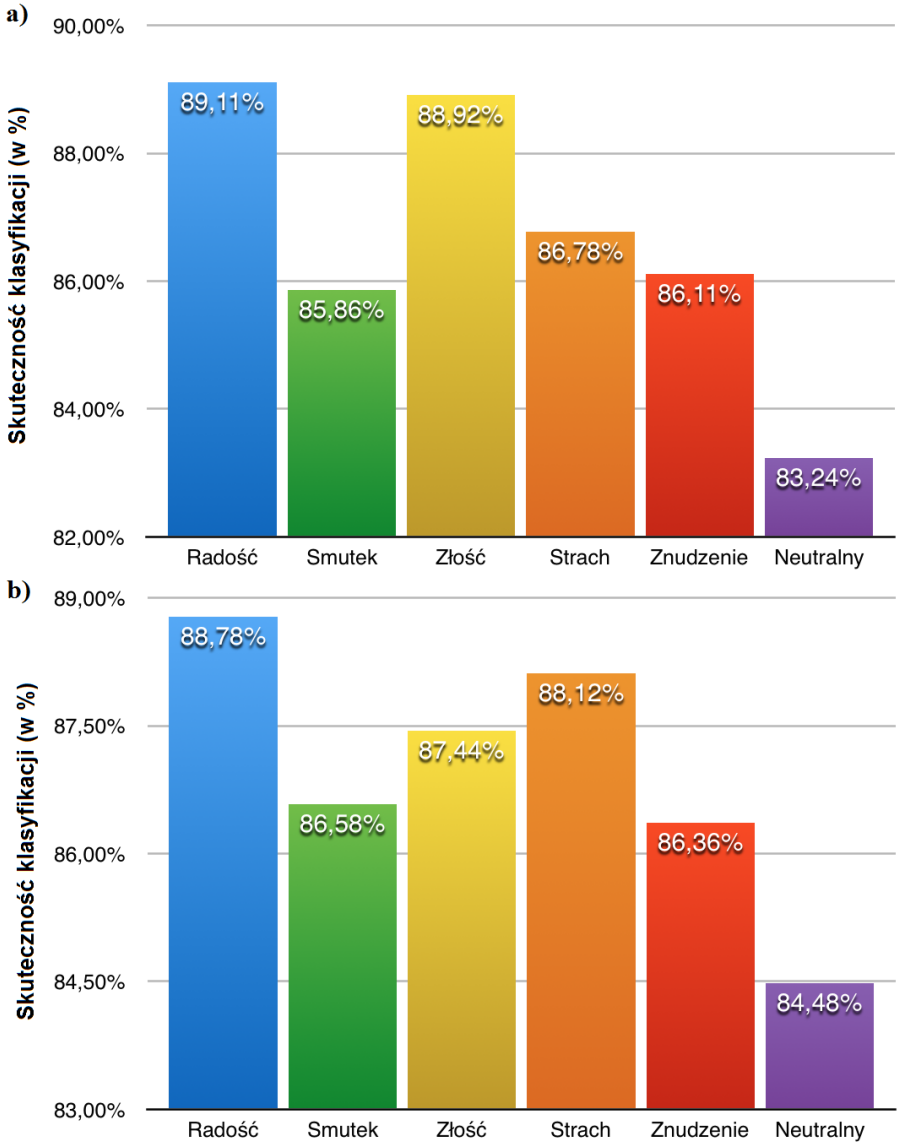
Podział skalogramu oraz określenie poziomu energii w każdym z podobszarów pozwolił na utworzenie wektora wejściowego dla SSN. Podczas wszystkich eksperymentów skuteczność identyfikacji stanów emocjonalnych mówcy wyniosła około 85%. Nieznacznie lepsze wyniki zostały osiągnięte w przypadku dekompozycji sygnału 9. rzędu. Jednakże czas potrzebny na przetwarzanie danych był znacząco wyższy niż w przypadku falek 7. rzędu. Otrzymane rezultaty zostały przedstawione na Rysunkach 5.14.–5.15. (patrz s. 103 i 104).

Tabela 5.14. Macierz pomyłek dla dekompozycji sygnału 7. rzędu (w %)

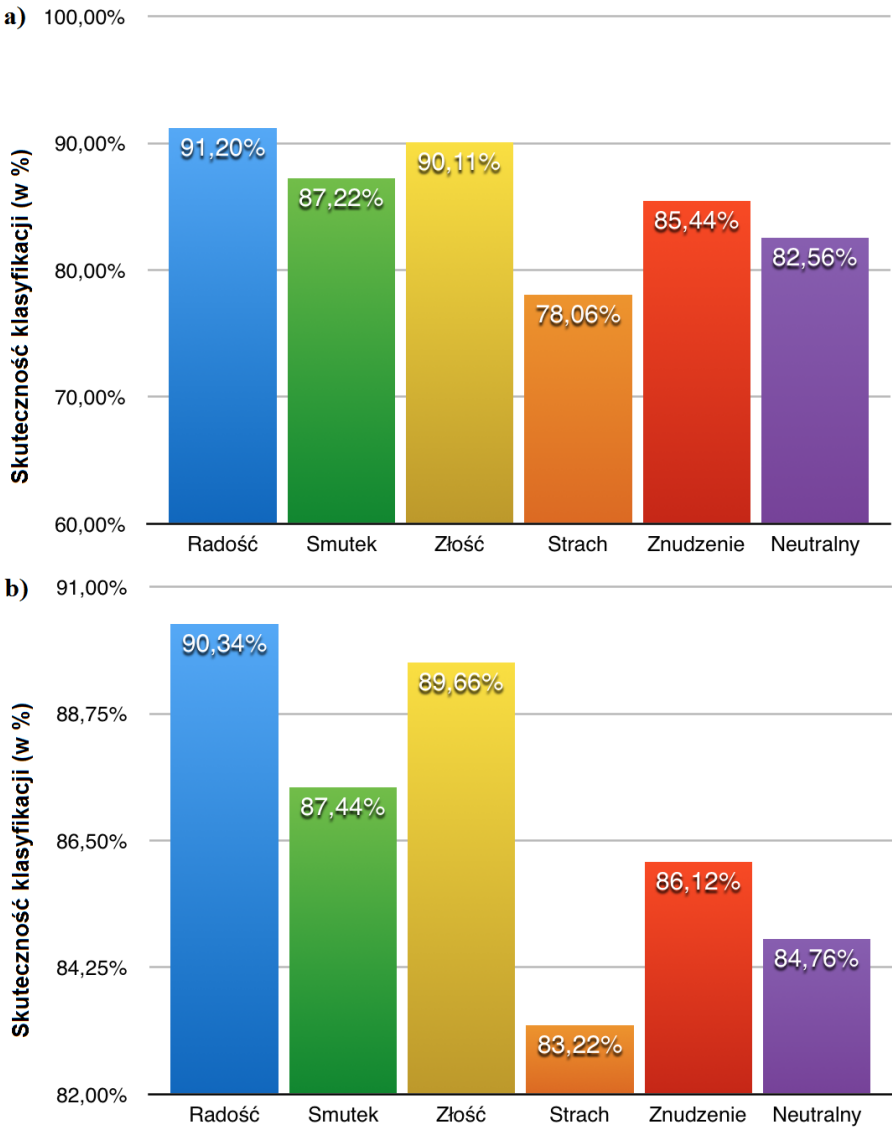
Emocje oczekiwane	**Emocje otrzymane na wyjściu**					
	Radość	Smutek	Złość	Strach	Znudzenie	Neutralny
Radość	**87,95**	1,88	2,98	3,96	2,12	1,11
Smutek	1,56	**86,22**	1,24	1, 86	4,24	4,88
Złośc	4,12	2,02	**88,11**	2,96	1,34	1,45
Strach	3,87	1,88	3,54	**87,50**	2,12	1,09
Znudzenie	1,56	3,64	1,78	1,56	**86,24**	5,22
Neutralny	0,80	5,68	2,46	1,88	5,32	**83,86**

Średnia skuteczność identyfikacji stanów emocjonalnych mówcy dla falek 7. rzędu dekompozycji wynosiła 86,65%, dla 9. poziomu 86,88%. Należy zauważyć, iż poziomy skuteczności są bardzo zbliżone, przy znacznie

większych kosztach obliczeń w przypadku dekompozycji sygnału 9. rzędu. Wymaga uwagi fakt poziomu identyfikacji radości i złości sięgający 90% bez względu na poziom dekompozycji oraz rodzaj użytej falki. Najgorsza skuteczność została osiągnięta w przypadku stanu neutralnego, który jest mylony ze znudzeniem oraz smutkiem.



Rysunek 5.14. Otrzymane wyniki dla dekompozycji sygnału 7. rzędu. Dla falek a) Haara, b) kapelusz meksykański



Rysunek 5.15. Otrzymane wyniki dla dekompozycji sygnału 9. rzędu. Dla falek a) Haara, b) kapelusz meksykański

Najbardziej od pozostałych wyników odbiega poziom identyfikacji strachu w przypadku dekompozycji falką 9. poziomu. Otrzymano tutaj odpowiednio 78,06% w przypadku falki Haara oraz 83,22% w przypadku falki kapelusza meksykańskiego. Wartości te są najniższymi uzyskanymi podczas niniejszych badań.

Należy zauważyć, iż w całym toku badań nie udało się do końca wyeliminować zjawiska notorycznego mylenia niektórych stanów emocjonalnych. Bazując na macierzach pomyłek, przedstawionych na w Tabelach 5.14. oraz 5.15. można zauważyć, iż jeśli pominięte zostaną wyniki uzyskane dla identyfikacji strachu, znaczące różnice pomiędzy poziomami poprawnej klasyfikacji dla falek 7. oraz 9. rzędu. Być może należałoby dla każdego ze stanów emocjonalnych ustalić inne poziomy progów podczas przetwarzania skalogramów do postaci binarnej. Należy podkreślić, iż wyniki uzyskane podczas badań symulacyjnych pozwalają przypuszczać, iż badania eksperymentalne zakończą się sukcesem i możliwa będzie identyfikacja stanu emocjonalnego mówcy języka polskiego, w czasie zbliżonym do rzeczywistego.

Tabela 5.15. Macierz pomyłek dla dekompozycji sygnału 9. rzędu (w %)

Emocje oczekiwane	**Emocje otrzymane na wyjściu**					
	Radość	Smutek	Złość	Strach	Znudzenie	Neutralny
Radość	**90,77**	1,85	1,98	2,47	0,87	2,06
Smutek	1,50	**87,33**	1,76	1,64	3,33	4,44
Złośc	3,08	1,50	**89,89**	2,64	1,44	1,45
Strach	6,24	1,95	6,23	**80,83**	2,68	2,07
Znudzenie	1,46	3,28	1,64	1,52	**88,78**	3,32
Neutralny	1,20	5,24	2,37	1,87	5,66	**83,66**

5.6.2. Badania eksperymentalne

Wyniki badań przeprowadzonych w oparciu o posiadane bazy nagrań emocjonalnej mowy polskiej spowodowały, iż możliwe stało się sprawdzenie poprawności działania opracowanej metody przetwarzania sygnału mowy w środowisku Call Center. Dzięki uprzejmości jednej z firm świadczących usługi telekomunikacyjne zostały przeprowadzone badania w czasie rzeczywistych rozmów z kientami.

Środowisko badawcze stanowił dyktafon firmy „OLIMPUS" podłączony do komputera marki „Macbook Pro" z zainstalowanym środowiskiem Matlab 2016.

Bazując na doświadczeniach pracowników działu Call-Center przyjęto założenie, iż najbardziej nacechowanym emocjonalnie fragmentem wypowiedzi jest jej początek zaraz po grzecznościowym zwrocie na powitanie. Przyjęto zatem, iż analizowany będzie fragment wypowiedzi mający swój początek w trzeciej sekundzie rozmowy i trwający kolejne pięć sekund. Również dostępne publikacje z dziedziny psychologii emocji oraz segmentacji

mowy potwierdzają, iż to początkowe fragmenty wypowiedzi są najbardziej nacechowane emocjonalnie [59, 46, 74, 137].

Przeprowadzone zostały badania wykorzystujące zarówno techinki przetwarzania spektrogramów jak i skalogramów. W przypadku pierwszych z nich do badań zostały zastosowane sztuczne sieci neuronowe wykorzystujące podział spektrogramu na 26 podobszarów[20]. Jako funkcję okna została używa funkcja Dolpha-Czebyszewa. Jako zbiór uczący zostały wykorzystane nagrania z Bazy A. W ramach badań przeprowadzone zostały 338 próby klasyfikacji stanów emocjonalnych.

Tabela 5.16. Liczba nagrań przetworzonych podczas badań eksperymentalnych wykorzystanych w badaniach z użyciem metody spektrogramów

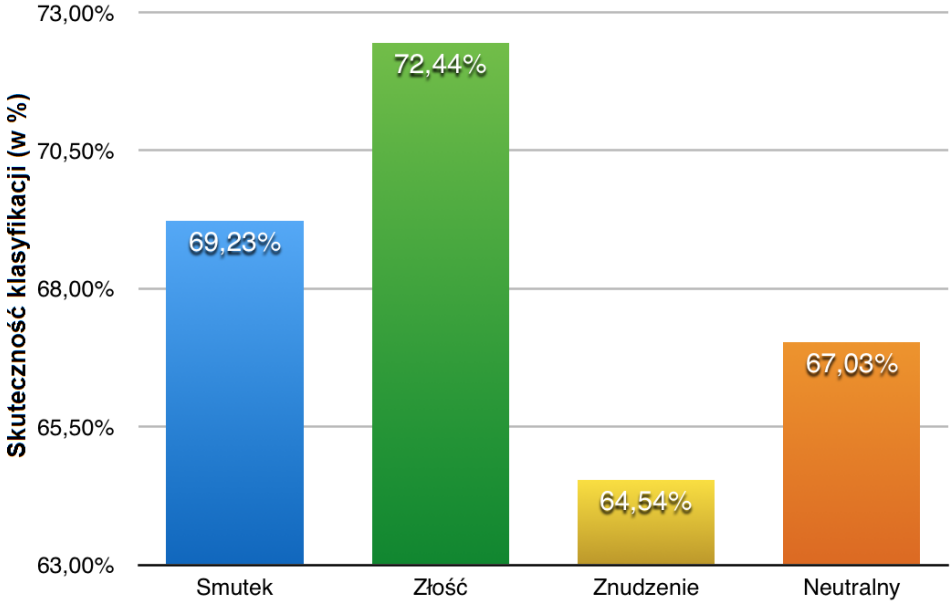
Emocja	**Liczba nagrań (kobiety)**	**Liczba nagrań (mężczyźni)**
Smutek	28	11
Złość	43	55
Znudzenie	65	45
Stan neutralny	67	24

Niestety nie wszystkie rodzaje rozpoznawanych emocji udało się zarejestrować podczas rozmów. Szczegółowy wykaz został przedstawiony w Tabeli 5.16. Poprawność klasyfikacji została następnie zweryfikowana przez 83 osobową grupę respondentów, którzy po odsłuchaniu zarejestrowanej wypowiedzi określali jej przynależność od jednego ze stanów emocjonalnych. Szczegółowe wyniki zostały zaprezentowane na Rysunku 5.16. oraz w Tabeli 5.17.

Tabela 5.17. Macierz pomyłek przy wykorzystaniu metody spektrogramów (w %)

Emocje oczekiwane	**Emocje otrzymane na wyjściu**			
	Smutek	Złość	Znudzenie	Neutralny
Smutek	**69,23**	5,13	12,82	12,72
Złośc	8,16	**72,44**	11,12	8,28
Znudzenie	19,09	3,64	**64,54**	12,73
Neutralny	18,68	2,20	11,91	**67,03**

[20] Szczegółowe informacje o SSN zostały zawarte w podrozdziale 5.2.



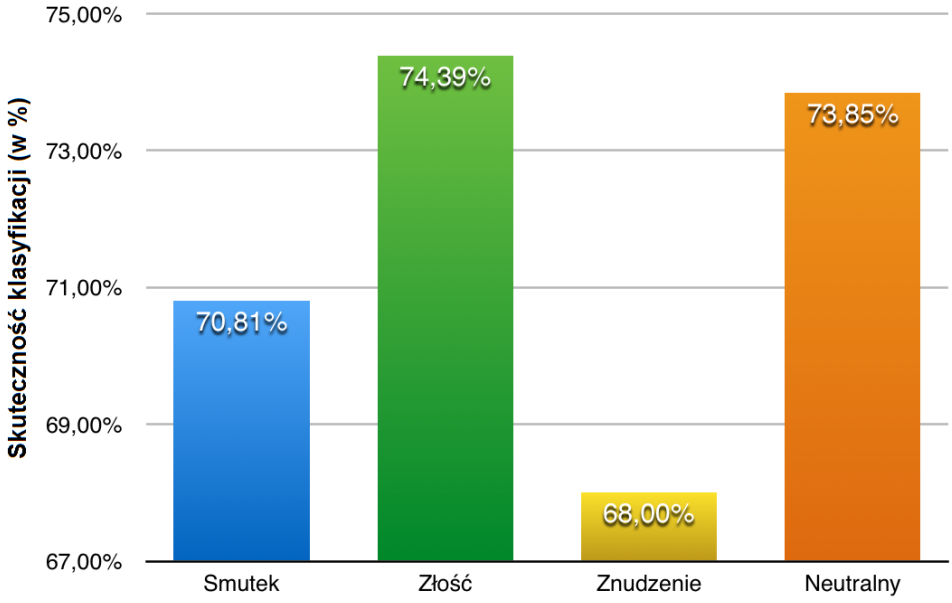
Rysunek 5.16. Otrzymane wyniki przy wykorzystaniu metody spektrogramów (w %)

Jak widać otrzymane wyniki oscylują w przedziale 64–72% poprawnie sklasyfikowanych emocji. Biorąc pod uwagę przeprowadzone w 2003 roku badania K. Scherer, który to udowodnił, że człowiek jest w stanie jedynie w 60% poprawnie zinterpretować stan emocjonalny nieznanej mu osoby [133], można przyjąć, iż otrzymane wyniki są zadowalające.

Drugą grupę badań stanowiły, te przeprowadzone przy wykorzystaniu metod związanych z dyskretną transformatą falkową oraz skalogramami. W eksperymentach zastosowana została 7. rzędu falka Haara. Podział skalogramu ograniczono do 210 podobszarów. Otrzymane wyniki zostały przedstawione na Rysunku 5.17. oraz z Tabeli 5.18.

Tabela 5.18. Macierz pomyłek przy wykorzystaniu metody skalogramów (w %)

Emocje oczekiwane	**Emocje otrzymane na wyjściu**			
	Smutek	Złość	Znudzenie	Neutralny
Smutek	**70,81**	4,17	12,50	12,50
Złośc	6,09	**74,39**	10,98	8,54
Znudzenie	19,00	4,00	**68,00**	9,00
Neutralny	15,38	3,08	7,69	**73,85**



Rysunek 5.17. Otrzymane wyniki przy wykorzystaniu metody skalogramów (w %)

Podobnie jak w przypadku badań nad pierwszą grupą nagrań nie udało się zarejestrować radości oraz strachu. Szczegółowe informacje na temat nagrań zostały zaprezentowane w Tabeli 5.19. W trakcie przeprowadzonych badań wykonano 271 klasyfikacji.

Tabela 5.19. Liczba nagrań przetworzonych podczas badań eksperymentalnych wykorzystanych w badaniach z użyciem metody skalogramów

Emocja	**Liczba nagrań (kobiety)**	**Liczba nagrań (mężczyźni)**
Smutek	11	13
Złość	38	44
Znudzenie	52	48
stan neutralny	26	39

Jak widać użycie metod wykorzystujących DWT pozwoliło na wzrost skuteczności rozpoznawanych emocji o kilka punktów procentowych. Otrzymane rezultaty pozwalają sądzić, iż dalsze badania pozwolą na jeszcze skuteczniejszą identyfikację stanów emocjonalnych w oparciu o sygnał mowy naturalnej.

5.7. Wnioski do rozdziału

1. Badania pilotażowe pokazały, iż możliwe jest skuteczne identyfikowanie stanów emocjonalnych mówców w opraciu o sygnał mowy polskiej.
2. Pokazano zależności pomiędzy różnymi wariantami algorytmu wstecznej propagacji błędów, a czasem ucznia oraz skutecznością identyfikacji stanów emocjonalnych.
3. Pokazany został wpływ funkcji aktywacji neuronów na skuteczność identyfikacja stanu emocjonalnego.
4. Udowodniono wpływ funkcji okna na skuteczność klasyfikacji SSN.
5. Zaprezentowano autorski mechanizm przetwarzania sygnału mowy oparty o spektrogramy pozwalający na identyfikację stanu emocjonalnego mówcy na poziomie 80%.
6. Przedstawiono modyfikację metody opartej o skalogramy, wykorzystującą dyskretną transformatę falkową pozwalającą na wzrost skuteczności identyfikacji do poziomu 85%.
7. Badania eksperymentalne pokazały, iż opracowany sposób przetwarzania sygnału mowy w czasie zbliżonym do czasu rzeczywistego wydaje się być wystarczająco skuteczny. Uzyskane rezultaty oscylowały w przedziale 64–74% co jest wynikiem zadowalającym w odniesieniu do [135].
8. Badania eksperymentalne pokazały, iż w środowisku naturalnym bardziej skuteczna okazuje się metoda przetwarzania sygnału oparta o dyskretną transformatę Fouriera.

6. Wnioski końcowe

1. Bazując na przeglądzie literatury podmiotu można stwierdzić, iż obecnie klasyfikacja emocji ogranicza się do następujących stanów emocjonalnych: radość, smutek, strach, złość, znudzenie oraz stan neutralny.
2. Zaprezentowana charakterystyka powstawania sygnału mowy, pozwala na jego rejestrację przy użyciu ogólnodostępnych narzędzi.
3. Bazując na parametrach wyekstrahowanych z sygnału mowy możliwe jest pokazanie separowalności stanów emocjonalnych.
4. Nie można jednoznacznie stwierdzić, że zastosowanie istniejących modeli i sposobów przetwarzania sygnału mowy, skutecznych od identyfikacji emocji w takich językach jak angielski, niemiecki, chiński czy arabski pozwolą na uzyskanie satysfakcjonujących rezultatów w przypadku sygnału mowy polskiej.
5. Pokazano wykorzystanie nowatorskiej metody przetwarzania sygnału, niespotykanej w literaturze przedmiotu, pozwalającej na skuteczną klasyfikację stanu emocjonalnego mówcy w oparciu o język polski.
6. Założenia przyjęte na potrzeby przeprowadzonych badań nie wpływają znacząco na jakość zaproponowanej metody przetwarzania sygnału.
7. Przeprowadzone badania pozwalają sądzić, iż wykorzystanie skalogramów i spektrogramów jak również innych narzędzi do klasyfikacji stanów emocjonalnych wyrażanych w językach innych niż polski pozwoli również osiągnąć satysfakcjonujące rezultaty.
8. Pomimo, iż opisane metody nie zapewniają 100% skuteczności klasyfikacji to w środowisku laboratoryjnym pozwalają na uzyskanie ponad 90% skuteczności klasyfikacji przy około 75% efektywności uzyskanej w rzeczywistym środowisku pracy.
9. Pokazane zostało, że rodzaj użytego klasyfikatora nie ma radykalnego wpływu na skuteczność klasyfikacji, gwarantem jest sposób przetwarzania sygnału mowy.

Literatura

[1] H. T. Abraham. (Physio)logical circuits: the intellectual origins of the mcculloch-pitts neural networks. *Journal of the history of the behavioural science*, pp. 3–25, 2002.

[2] S. A. Al-agha, H. H. Saleh, R. F. Ghani. Multi-model emotion expression recognition for lip synchronization. *Proceedings of IT-ELA, Baghdad, Iraq*, pp. 171–177, 2020.

[3] N. Anithadevi, P. Gokul, S. M. Nandan, R. Magesh, S. Shiddharth. Automated speech recognition system for speaker emotion classification. *Proc. ICCCS*, 2020.

[4] A. M. Badshah, J. Ahmad, N. Rahim, S. W. Baik. Speech emotion recognition from spectrograms with deep convolutional neural network. *2017 International Conference on Platform Technology and Service (PlatCon)*, pp. 1–5, 2017.

[5] A. M. Badshah, N. Rahim, N. Ullah, J. Ahmad, K. Muhammad, M. Y. Leeand S. Kwon, S. W. Baik. Deep features-based speech emotion recognition for smart affective services. *Multimedia Tools and Applications*, 78(5), pp. 5571–5589, 2017.

[6] S. Banriskhem, K. Mahadeva. Speech/music classification using speech specific features. *Digital Signal Processing*, 48, pp. 71–83, 2016.

[7] C. Baszutra. *Rozmawiać z komputerem.* Wydawnictwo Prac Naukowych Format, Wrocław, 1992.

[8] D. Białasiewicz. *Falki i aproksymacje.* Wydawnictwo Naukowo-Techniczne, Warszawa, 2000.

[9] C. Bishop. Neural Network for Pattern Recognition. Oxford, 2005.

[10] J. W. Bogert, B. P. Healy, M. J. Tukey. The quefrency alanysis of time series for echoes: Cepstrum, pseudo-autocovariance, cross-cepstrum and saphe cracking. *Proceedings of the Symposium on Time Series Analysis*, pp. 209–243, 1963.

[11] G. Bombatkar, A Bhoyar. Emotion recognition using speech processing using k-nearest neighbor algorithm. *International Journal of Engineering Research and Applications*, pp. 68–71, 2014.

[12] A. Burkhard, F. Paeschke. A database of german emotional speech. *Proceedings Interspeech*, *Lisbon, Portugal*, 2005.

[13] L. Cai, J. Dong, M. Wei. Multi-modal emotion recognition from speech and facial expression based on deep learning. *Proceedings of CAC, Shanghai, China*, pp. 5726–5729, 2020.

[14] W.B Cannon. The James-Lange theory of emotions: a critical examination and an alternative theory. *Journal of Nonverbal Behaviour*, 6, pp. 238–251, 1982.

[15] R. N. Cardnial. Psychological basis of emotions. *Emotion and motivation*, pp. 1–10, 2003.

[16] I. Castro-Vale, M. Severo. Emotion recognition ability test using jacfee photos: A validity/reliability study of a war veterans' sample and their offspring. *PLoS ONE*, 7(10), 2015.

[17] A. Catalan, M. Gonzalez de Artaza. Differences in facial emotion recognition between first episode psychosis, borderline personality disorder and healthy controls. *PLoS ONE*, 11(7), 2016.

[18] A. Chamoli, A. Semwal, N. Saikia. Detection of emotion in analysis of speech using linear predictive coding techniques (lpc). *Proceedings of the International Conference on Inventive Systems and Control (ICISC-2017)*, pp. 1–4, 2017.

[19] H. Chernoff. Use of faces to represent points in k–dimensional space graphically. *American Statistical Association Journal*, 68(342), pp. 361–368, 1973.

[20] M. Chmaj, T. Lankosz. Akwizycja i przetwarzanie sygnałów cyfrowych. Kraków, s. 22–76, 2011.

[21] J. Cichosz, K. Ślot. Emotion recognition in speech signal using emotion extracting binary decision trees. *Institute of Electronics, Technical University of Lodz, Poland, Technical Report*, 2010.

[22] K. Cichosz, J. Ślot. Low-dimensional feature space derivation for emotion recognition. *Proceedings Interspeech, Lisbon, Portugal*, 2005.

[23] Z. Ciota. Emotion recognition on the basis of human speech. *Applied Electromagnetics and Communications, ICECom*, 2005.

[24] Z. Ciota. Metody przetwarzania sygnałów akustycznych w komputerowej analizie mowy. Akademicka Oficyna Wydawnicza EXIT, Warszawa, 2010.

[25] T. Dalgleish. The emotional brain. *Nature*, 5, pp. 582–589, 2004.

[26] R. J. Davidson, N.A. Fox. Asymmetrical brain activity discriminates between positive versus negative affective stimuli in human infants. *Science*, 218, pp. 1235–1237, 1982.

[27] M.deLuna-Ortega, C.Mora-Gonzalez. Analysis of Kohonen's neural network with application to speech recognition. *MICAI 2009. Workshop Computer Vision and Pattern Recognition – WCVPR*, 2009.

[28] J. R. Deller. Discrete Time Processing of Speech Signal. Prentice Hall, New Jersey, 1993.

[29] J. Deng, S. Fruhholz, Z. Zhang, B. Schuller. Recognizing emotions from whispered speech based on acoustic feature transfer learning. *IEEE Access*, 5, pp. 5235–5246, 2017.

[30] N. Duch, W. Jankowski. Transfer functions: hidden possibilities for better neural networks. *ESANN'2001 proceedings – European Symposium on Artificial Neural Networks Bruges (Belgium)*, 2001.

[31] E. Duffy. Activation and behavior. Wiley, London, 1962.

[32] P. Ekman, J.R. Davidson. Natura emocji. Nastroje, emocje, cechy. GWP, Gdańsk, 1998.

[33] P. Ekman, J.R. Davidson. Natura emocji. Podstawowe Zagadnienia. GWP, Gdańsk, 1998.

[34] P. Ekman, W.V. Friesen. Felt, false, and miserable smiles. *Journal of Nonverbal Behavior*, 6, pp. 238–251, 1982.

[35] F. Espinoza-Cuadros, F. Fernandez-Pozo. Speech signal and facial image processing for obstructive sleep apnea assessment. *Computational and Mathematical Methods in Medicine*, pp. 1–13, 2015.

[36] V. Fernandes, L. Mascarehnas, C. Mendonca, A. Johnson, R. Mishra. Speech emotion recognition using mel frequency cepstral coefficient and svm classifier. *Proceedings of the 2018 International Conference on System Modeling and Advancement in Research Trends*, pp. 200–204, 2018.

[37] F. Fersini, E. Messina, E. Arosio, G. Archetti. Audio-based emotion recognition in judicial domain: A multilayer support vector machines approach. *Machine Learning and Data Mining in Pattern Recognition 6th International Conference, MLDM 2009*, pp. 594–602, 2009.

[38] R. Fletcher, C. Reeves. Function minimization by conjugate gradients. *Computer Journal*, 1, pp. 149–154, 1964.

[39] M. Garay, I. Cearreta, N. Miguel Lopez, J. Fajardo. Assistive technology and affective mediation. *An Interdisciplinary Journal on Humans in ICT Environments*, pp. 55–83, 2006.

[40] S. Grabowski. Konstrukcja klasyfikatorów minimalnoodległościowych o strukturze sieciowej. Praca doktorska, Politechnika Łódźka, Łódź, 2003.

[41] M. Grimm, K. Kroschel. Support vector regression for automatic recognition of spontaneous emotions in speech. *IEEE International Conference on Acoustics, Speech and Signal Processing – ICASSP*, 2007.

[42] A. Grzywa. Tendencyjność myślenia. Lublin, 1995.

[43] T. S. Gunawan, M. F. Alghifari, M. A. Morshidi, M. Kartiwi. A review on emotion recognition algorithms using speech analysis. *Indonesian Journal of Electrical Engineering and Informatics*, 6(1), pp. 12–20, 2018.

[44] H. Gunes. Automatic, dimensional and continuous emotion recognition. *International Journal of Synthetic Emotions*, 1, pp. 68–99, 2010.

[45] A. Głowacz. Komputerowe techniki analizy informacji zawartej w sygnałach akustycznych maszyn elektrycznych dla celów disgnostyki stanów przedawaryjnych. Praca doktorska, Akademia Górniczo Hutnicza, Kraków, 2013.

[46] U Hareli, S. Hess. The social signal value of emotions. *Cognition & Emotion*, pp. 385–389, 2012.

[47] J. Harris, R. Frederic. On the use of windows for harmonic analysis with the discrete Fourier transform. *Proceedings of the IEEE*, 66, pp. 51–83, 1978.

[48] R. Heinzel, G. Rudiger, A. Schilling. Spectrum and spectral density estimation by the discrete fourier transform (dft), including a comprehensive list of window functions and some new flat-top windows (technical report). *Max Planck Institute (MPI) fur Gravitationsphysik*, pp. 28–36, 2002.

[49] H. Hikawa. A new digital pulse-mode neuron with adjustable activation function. *IEEE Transactions on Neural Networks*, 14, pp. 236–242, 2003.

[50] L Hong-Yo, F. Hai-Liang. General formula for finding mexican hat wavelets by virtue of dirac's representation theory and coherent state. *Optical Letters*, 31(1), pp. 407–409, 2006.

[51] N. Hossain, M. Naznin. Finding emotion from multi-lingual voice data. *Proceedings of COMPSAC,* Madrid, Spain, pp. 408–417, 2020.

[52] B. Igraś, M. Ziółko. Baza nagrań mowy emocjonalnej. *Studia Informatica*, 112(2), pp. 67–77, 2013.

[53] C. E. Izard. The face of emotion. New York, 1971.

[54] C. E. Izard. The substrates and functions of emotion feeling. William James and current emotion theory. P*ersonality and Social Psychology Bulletin*, pp. 626–635, 1990.

[55] W. James. Principles of psychology. New York, 1890.

[56] W. James. Doświadczenia religinje. Warszawa, 1958.

[57] M. Janicki, A. Turkot. Rozpoznawanie stanu emocjonalnego mówcy z wykorzystaniem maszyny wektorów wspierających. *KSTiT*, Bydgoszcz, 2008.

[58] T. P. Jiang, H. Fu, H. Tao. Speech emotion recognition using deep convolutional neural network and simple recurrent unit. *Engineering Letters*, 27(4), pp. 1–6, 2019.

[59] S. Johar. Emotion, affect and personality in speech. *Springer Briefs in Speech Technology*, pp. 9–15, 2016.

[60] H. Jurkiewicz. Nieeuklidesowe sieci neuronowe. Praca doktorska, UMK w Toruniu, 2009.

[61] Ł. Juszkiewicz. Speech emotion recognition system for social robots. J*ournal of Automation, Mobile Robotics & Intelligent Systems*, 7(4), pp. 59–65, 2013.

[62] I. Józefczyk. Dyskretna transformata falkowa dla wybranego modelu symulacyjnego sygnału wibroakustycznego. *Diagnostyka*, 34, pp. 137–141, 2005.

[63] R. Kalioub, P. Robinson. Mind reading machines automated inference of cognitive mental states from video. *IEEE International Conference on Systems, Man and Cybernetics*, pp. 682–688, 2004.

[64] N. Kamaruddin, A. Wahab. Driver behavior analysis through speech emotion understanding, *Intelligent Vehicles Symposium (IV) IEEE*, pp. 238–243, 2010.

[65] B. C. Kamble. Speech recognition using artificial neural network. *International Journal of Computing, Communication and Instrumentation Engineering*, 3(1), pp. 1–4, 2016.

[66] B. C. Kamble. Speech recognition using artificial neural network – a review. *International Journal of Computing, Communications and Instrumentation Engg. (IJCCIE)*, 3(1), pp. 1–4, 2016.

[67] A. Kaminska, D. Sapiński, T. Pelikant. Polish emotional natural speech database. *Conference: Signal Processing Symposium 2015*, At Debe, 2015.

[68] D. Kamińska. Rozpoznawanie emocji na podstawie mowy naturalnej. Praca doktorska, Politechnika Łódźka, 2014.

[69] D. Kamińska. Emotional speech recognition based on the committee of classifiers. *Entropy*, 21(10), pp. 1–17, 2019.

[70] D. Kamińska, A. Palikant. Zastosowanie mulimodalnej klasyfikacji w rozpoznawaniu stanów emocjonalnych na podstawie mowy spontanicznej. *IAPGOŚ*, 3, pp. 36–39, 2012.

[71] D. Kamińska, A. Pelikant. Recegnition of human emotions from a speech signal based on plutchnik's model. *INTL International Journal of Electronic and Telecommunications*, 58(2), pp. 165–170, 2012.

[72] C. Kang, C Hang, B. Lee, S. Youn, D. Lee. Speaker dependent emotion recognition using speech signals. *International Conference of Spoken Language Archive*, pp. 1–4, 2000.

[73] B. Karlik, V. Olgac. Performance analysis of various activation functions in generalized mlp architectures of neural networks. I*nternational Journal of Artificial Intelligence And Expert Systems*, 66, pp. 111–122, 2014.

[74] M Kashem. Speech Processing. Chapter 13. Dhaka University, Dhaka, 2004.

[75] D. A. Keim. Designing pixel-oriented visualization techniques: Theory and applications. *IEEE Transactions on Visualization and Computer Graphics*, 6, pp. 59–78, 2000.

[76] K. H. Kim, E. H. Hyu. Speech emotion recognition using eigen-fft in clean and noisy environment. *16th IEEE International Conference on Robots and Human Interactive Communication*, 2007.

[77] R. King, C. Feng. Statlog: Comparison of classi cation algorithms on large real-worl problems. *Applied Artificial Intelligence*, 9(3), pp. 289–333, 1995.

[78] M. Kolasa, M. Drechny. Wybrane aspekty zastosowania sztucznej sieci neuronowej w systemie sterowania oświetleniem ulicznym. *Przegląd elektrotechniczny*, 3, pp. 36–39, 2017.

[79] E. Konratowski. Czasowo-częstotliwościowa analiza dragń z wykorzystaniem metody overlapping. *Logistyka*, 3, pp. 3104–3110, 2014.

[80] B. Kort, R. Reilly. An effective model of interplay between emotions and learning: Reengineering education pedagogy — building a learning companion. *Proceedings of the IEEE International Conference on Advanced Learning Technologies*, 2001.

[81] L. Korzeniowski, S. Pużyński. Encyklopedyczny słownik psychiatrii. PZWL, Warszawa, 1986.

[82] Z. Kowalczuk, M. Czubenko. Xemotion – obliczeniowy model emocji dedykowany dla inteligentnych systemów decyzyjnych. *Pomiary Automatyka Robotyka*, 17(2), pp. 60–65, 2013.

[83] K. Kucharska-Pietura. Trudności definicyjne i klasyfikacyjne zjawisk emocjonalnych. *Wiadomości Psychiatryczne*, 5(2), pp. 131–135, 2002.

[84] M. Kłaczyński. Zjawiska wibroakustyczne w kanale głosowym człowieka. Praca doktorska, Akademia Górniczo Hutnicza, Kraków, 2007.

[85] S. Lalitha, D. Geyasruti, R. Narayanan. Emotion detection using mfcc and cepstrum features. *Proceedia Computer Science*, 70, pp. 29–35, 2015.

[86] P. J. Lang. The varieties of emotional experience: A meditation on James-Lange theory. *Psychological Review*, 2, pp. 211–221, 1994.

[87] J. LeDoux. The emotional brain, fear, and the amygdala. *Cellular and Molecular Neurobiology*, 23, pp. 727–738, 2003.

[88] K. H. Lee, D. H. Kim. Design of a convolutional neural network for speech emotion recognition. *PICTC, Jeju, Korea (South)*, pp. 1332–1335, 2020.

[89] J. Leppanen. Emotional information processing in mood disorders: a review of behavioral and neuroimaging findings. *Current opinion in psychiatry*, 19, pp. 34–39, 2006.

[90] A. Lopatka, J. Kotus, K. Czyzewski. Detection, classification and localization of acoustic events in the presence of background noise for acoustic surveillance of hazardous situations. *Multimedia Tools and Applications*, 75(17), pp. 10407—10439, 2016.

[91] C. Lu. Circuit design of an adjustable neuron activation function and its derivative. *Electronic Letters*, 36, pp. 553 – 555, 2002.

[92] M. Lugger, B. Yang. The relevance of voice quality features in speaker independent emotion recognition. *IEEE International Conference on Acoustics, Speech and Signal Processing – ICASSP*, 2007.

[93] P. Lynch. The Dolph-Chebyshev window: A simple optimal filter. *America Meteorological Society Journal of the Online*, 125, pp. 655–660, 2011.

[94] J. Mazurkiewicz. Zarys fizjologicznej teorii uczuć. PZWL, Warszawa, 1930.

[95] J. Mazurkiewicz. Wstęp do psychofizjologii normalnej. Ewolucja aktywności korowo–psychicznej. PZWL, Warszawa, 1950.

[96] D. Michie, D. Gelhalter. Machine Learning, Neural and Statistical Classification. New York, 1994.

[97] F. Mitsugi, H. Stryczewska. Application of optical wave microphone to gliding arc discharge. *Przegląd Elektrotechniczny*, pp. 105–108, 2012.

[98] O. A. Mohammad, M. Elhadef. Arabic speech emotion recognition method based on lpc and ppsd. *Proceedings of ICCAKM, Dubai, United Arab Emirates*, pp. 31–36, 2021.

[99] R. Mohammad, M. Ainon. Natural speaker-independent arabic speech recognition system based on hidden Markov models using sphinx tools. *Computer and Communication Engineering*, pp. 98–113, 2010.

[100] M. Moller. A scaled conjugate gradient algorithm for fast supervised learning. *Neural Networks*, 6(4), pp. 525–533, 1993.

[101] G. Muhammad. Automatic speech recognition using interlaced derivative pattern for cloud based healthcare system. *Cluster Computing*, 18(2), pp. 795–802, 2015.

[102] K. Naess, A. Livescu. Articulatory feature classification using nearest neighbors. *Proceedings Interspeech 2011*, Florence, Italy, 2011.

[103] A. Newen, A. Welpinghus. Emotion recognition as pattern recognition: The relevance of perception. *Mind Lang*, 30, pp. 187–208, 2015.

[104] D. Nguyen, B. Widrow. Improving the learning speed of 2-layer neural networks by choosing initial values of the adaptive weights. *Proceedings of the International Joint Conference on Neural Networks*, pp. 21–26, 1990.

[105] N. Nutalid. Evolving Probabilistic Spiking Neural Networks for Modelling and Pattern Recognition of Spatio-temporal Data on the Case Study of Electroencephalography (EEG) Brain Data. Praca doktorska, Auckland University of Technology, Auckland, 2012.

[106] S. Ombach, J. Kloeden, P. Cyganowski. From Elementary Probability to Stochastic Differential Equations with Maple. 2002.

[107] A. V. Oppenheim. Sygnały cyfrowe przetwarzanie i zasotsowanie. Warszawa, 1982.

[108] D. Padrell-Sendra, J. Martin-Iglesias. Support vector machines for continuous speech recognition. 1*4th European Signal Processing Conference (EUSIPCO 2006)*, *Florence, Italy*, 2006.

[109] J. W. Papez. A proposed mechanism of emotion. *Archives of Neurological Psychiatry*, 38, pp. 725–743, 1937.

[110] D. B. Paul. Speech recognition using hidden Markov models. *The Lincoln Laboratory Journal*, pp. 41–62, 2013.

[111] A. Pavelka, A. Prochazka. Algorithms for initialization of neural network weights. *Proceedings of the Conference Technical Computing*, 2004.

[112] B.L. Peeters, G. Giordano. The timbre toolbox: Extracting audio descriptors from musical signals. *The Journal of the Acoustical Society of America*, 130(2902), pp. 2902–2916, 2011.

[113] C. Pfitzinger, H. R. Kaernbach. Amplitude and amplitude variation of emotional speech. *Interspeech*, 2008.

[114] R. Plutchik. What is an emotion? *The Journal of Psychology Interdisciplinary and Applied*, 61(2), pp. 295–303, 1965.

[115] R. Plutchik. Emotion: A Psychoevolutionary Synthesis. New York, 1980.

[116] R. Plutchik. The nature of emotions. *American Scientist*, 89, pp. 344–350, 2001.
[117] M. Śliwińska-Kowalska. Głos narzędziem pracy. Łódź, 1999.
[118] E. Polak. Computational methods in optimisation: a unified approach. 1971.
[119] R. Povinelli, M. Johnson, A. C. Lindgren, J. Ye. Analysis of emotion recognition system through speech signal using knn, gmm and svm classifier. *Journal of Electronics and Communication Engineering*, 10(2), pp. 55–61, 2015.
[120] P. Powroźnik. Polish emotional speech recognition using artifical neural networks. *Advances in Science and Technology Research Journal (ASTRJ)*, 8(24), pp. 24–27, 2014.
[121] P. Powroźnik, D. Czerwiński. Effectiveness comparison on an artificial neural networks to identify polish emotional speech. *Przegląd elektrotechniczny*, 92(7), pp. 45–48, 2016.
[122] P. Powroźnik, D. Czerwiński. Spectram methods in polish emotional speech recognition. *Advances in Science and Technology Research Journal (ASTRJ)*, 10(32), pp. 73–81, 2016.
[123] C. Prakash, V. B. Gaikwad, R. R. Singh, O. Prakash. Time series classification using gaussian mixture models of reconstructed phase spaces. *IEEE Transactions on Knowledge and Data Engineering*, 16(6), pp. 779–783, 2004.
[124] J. Prinz. Thinking about feeling: Contemporary philosophers on emotions. *Emotions embodied*, pp. 44–59, 2004.
[125] L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *IEEE proceedings*, 77, pp. 257–287, 1989.
[126] R. Rak, A. Majkowski. Czasowo-częstotliwościowa analiza sygnałów. *Encyklopedia Przeglądu Elektrotechnicznego*, pp. 515–520, 2004.
[127] R.A. Rammohan, J. Medikonda, D.I. Pothiyil. Speech signal-based modelling of basic emotions to analyse compound emotion: Anxiety. *Proceedings of DISCOVER, Udupi, India*, pp. 218–223, 2020.
[128] J. Reykowski. Eksperymentalna psychologia emocji. KiW, Warszawa, 1974.
[129] J. Reykowski. Procesy emocjonalne. Motywacja. Osobowość. PWN, Warszawa, 1992.
[130] F. Reynolds, D. Quatieri. Speaker verification using adapted gaussian mixture models. *Digital Signal Processing*, pp. 19–41, 2000.
[131] B. Rozengał. Zastosowanie transformaty falkowej do wykrywania zwarć w linii dwustronnie zasilanej. *Elektrotechnika*, 109, pp. 93–110, 2012.
[132] A. Rusiecki. Algorytmy uczenia sztucznych sieci snuronowych odporne na błędy danych. Praca doktorska, Politechnika Wrocławska, 2007.
[133] K. Scherer. Vocal communication of emotions. A Review of Research. *Paradigms in Speech Communication*, 40, pp. 227–256, 2003.
[134] P. Schmid, M. Schmid-Mast. Mood effects on emotion recognition. *Motivation and Emotion*, 34, pp. 288–292, 2010.

[135] B. Schuller, S. Reiter. Speaker independent speech emotion recognition by ensemble calssification. *Proceedings of the 2005 IEEE International Conference on Multimedia and Expo, ICME 2005, July 6–9, 2005, Amsterdam, The Netherlands*, 2005.

[136] S. Schuller, D. Seppi. Towards more reality in the recognition of emotional speech. *IEEE International Conference on Acoustics, Speech and Signal Processing – ICASSP*, 2007.

[137] J Shwati. Emotion, Afects and Personality in Speech. The Bias in Language and Paralanguage. Springer, Switzerland, 2016.

[138] M. Silva, A.P. Madurapperuma, P.R. Marasinghe, A. Osano. A multi-agent based interactive system towards childs emotion performances quantified through affective body gestures. *International Conference on Pattern Recognition*, pp. 1236–1239, 2006.

[139] J. O. Smith III. Spectral Audio Signal Processing. Second Edition. W3K Publishing, Stanford, 2011.

[140] J. Strelau. Psychologia ogólna. Tom 2. Gdańsk, 2007.

[141] K. Szostek. Rozpoznawanie mowy metodami niejawnych modeli Markowa HMM. Praca doktorska, Akademia Górniczo Hutnicza, Kraków, 2006.

[142] A. Tadeusiewicz, R. Izworski. Metody komputerowej ekstrakcji parametrów dystynktywnych z ciągłego sygnału mowy polskiej. *Archiwum akustyki*, 18(3), 1983.

[143] R. Tadeusiewicz. Sygnał mowy. Warszawa, 1988.

[144] R. Tadeusiewicz. Sztuczne sieci neuronowe. Kraków, 1993.

[145] R. Tadeusiewicz. Elementarne wprowadzenie do techniki sieci neuronowych z przykładowymi programamami. Warszawa, 1998.

[146] L. Thompson, W. F. Balkwill. Decoding speech prosody in five languages. *Semiotica*, 158, pp. 407–424, 2006.

[147] L. J. van der Maaten, J. E. Hinton. Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research*, 9, pp. 2579–2605, 2008.

[148] L. J. van der Maaten. Accelerating t-sne using tree-based algorithms. *Journal of Machine Learning Research*, 15, pp. 3221–3245, 2014.

[149] P.P. van der Smagt. Minimization methods for training feedforward neural networks. *Neural Networks*, 1, pp. 1–11, 1994.

[150] V. V. R. Vegesna, K. Gurugubelli, A. K. Vuppala. Application of emotion recognition and modification for emotional telugu speech recognition. *Mobile Networks and Applications*, 24(1), pp. 193–201, 2019.

[151] R. Venkateswarlu, R. Kumari. Novel approach for speech recognition by using self organized maps. *International Journal of Computer Science and Information Technology*, 3(4), pp. 199–210, 2011.

[152] C. Vinola, K. Vimaladevi. A survey on human emotion recognition approaches, databases and applications. *Electronic Letters on Computer Visions and Image Analysis*, 2(14), pp. 24–44, 2015.

[153] J.K Vogl, T. P. Mangis. Accelerating the convergence of the back propagation method. *Biological Cybernetics*, 59, pp. 256–264, 1988.

[154] P. Walendowski. Zastosowanie sieci neuronowych typu SVM do rozpoznawania mowy. Praca doktorska, Politechnika Wrocławska, Wrocław, 2008.

[155] J. Wang, P. Neskovic. Improving nearest neighbor rule with a simple adaptive distance measures. *Pattern Recognition Letters*, 28(2), pp. 207–213, 2007.

[156] J. B. Watson. Behaviorism. Norton, New York, 1924.

[157] K. Weinberger, F. Sha. Learning a kernel matrix for nonlinear dimensionality reduction. *In Proceedings of the 21st International Confernence on Machine Learning*, 2004.

[158] W. Wójcik. Application of fibre-optic flame monitoring systems to diagnostics of combustion process in power boilers. *Bulletin of the Polish Academy of Sciences Technical Sciences*, 52(2), pp. 177–195, 2008.

[159] S. Yacoub, S. Simske. Recognition of emotions in interactive voice response systems. *Eurospeech*, pp. 1–4, 2003.

[160] W. Yohanes, R. Ser. Discrete wavelet transform coefficients for emotion recognition from eeg signals. *Conference proceedings IEEE engineering in medicine and biology society*, pp. 2251–2254, 2012.

[161] P. Zelazko, i inni. Zastosowanie algorytmu dtw jako narzędzia w identyfikacji mówcy. *Problemy kryminalistyki*, 280(2), pp. 53–57, 2013.

[162] S. Zhang. Structured Support Vector Machines for Speech Recognition. Praca doktorska, Cambridge University, 2014.

[163] T. P. Zieliński. Cyfrowe przetwarzanie sygnałów. Od teorii do zastosowań. Warszawa, 2009.